B. J. K. Kleijn

The frequentist theory of Bayesian statistics

September 2, 2022

Springer

Preface

As a frequentist, I cannot think of a better statistical tool than the Bayesian posterior. Whether in parameter or density estimation, hypothesis testing, uncertainty quantification or decision theoretic questions, there are always examples of priors with posteriors that satisfy *frequentist* criteria of optimality. A good example of an optimal Bayesian answer to a frequentist question, arises in the (apparently straightforward) estimation of a multivariate normal mean of dimension three or higher, based on an *i.i.d.* sample of observations: it was shown in the late 1950's that there exists a family of so-called *super-efficient* estimators (*e.g.* the famous James-Stein estimator), that outperform the sample mean and all other unbiased estimators when compared in mean-squared error. It was shown in the 1970's (and can be expected based on the so-called *complete class theorem*) that there exist so-called *empirical Bayes estimators* that display a James-Stein-type of super-efficiency.

Criteria for frequentist optimality are often formulated in terms of large-sample behaviour, and most examples of posteriors with good frequentist properties concern forms of large-sample convergence. For example, in the fourth chapter, we consider the Bernstein-von Mises theorem that establishes asymptotic normality of the posterior in smooth, parametric models and shows that Bayesian credible sets approximate optimal frequentist confidence sets asymptotically. Indeed, such a correspondence between credible sets and confidence sets is possible also with finite amounts of data, if one is willing to enlarge credible sets in a suitable way: it is shown in the second chapter that if the posterior concentrates a certain, lower-bounded amount of mass around the true value of the parameter in expectation, then *enlargements of credible sets of a certain credible level are exact confidence sets of a chosen confidence level, with finite amounts of data*. This construction is used in the eleventh chapter to find confidence sets for the community assignment in sparse versions of the two-community stochastic block model.

But such finite-sample correspondences are rare: in part II, we consider various forms of large-sample posterior convergence from the frequentist perspective in non-parametric models, and again we find examples of priors that induce frequentist optimality of procedures based on their posterior distributions. Invariably Bayesian procedures can be shown to display suitable forms of asymptotic optimality in great

generality and with relative ease, but only on a model subset of prior probability one. Typically a prior null-set of possible exceptions is left, which spoils optimality for the frequentist. To strengthen such Bayesian optimality properties to the corresponding forms of frequentist optimality, the prior must induce a weak form of contiguity (called *remote contiguity*) between the sequences of true data distributions and of (local) prior predictive distributions. Remote contiguity and its applications for the frequentist validity of Bayesian limits are analysed in generality in the seventh chapter, and some pointers are given regarding the potential applications of remote contiguity in other problems.

The aim of this book is two-fold: a first goal is to provide a mathematically sound and complete general framework to analyse Bayesian procedures and their conversion to frequentist methods. A second goal is to illustrate the frequentist optimality of Bayesian methods; more specifically, to give suitable conditions for priors that give rise to frequentist optimality of posterior-based procedures, with examples of priors that satisfy those conditions. In both these respects, this book aims to be comprehensive, at the expense of other aspects that are covered in other works; particularly, this book does not attempt to illustrate computational matters (the interested reader is referred, for example, to Neal (1993) [202] and Robert (2001) [218]), nor does it approach Bayesian statistics from a decision-theoretic/classification-oriented perspective (as provided by Ripley (1996) [219]), nor does it give a purely Bayesian overview (see, particularly, Berger (1985) [19] and Bernardo and Smith (1993) [25]), nor does it constitute a review of examples and applications with an emphasis on translation of non-parametric posterior asymptotics to frequentism (see, *e.g.*, Ghosh and Ramamoorthi [111], Ghosal and van der Vaart [110]). Regarding other sources that illuminate related subjects, we mention in particular the entry-level discussion of frequentist asymptotic statistics (with non- and semi-parametric elaborations) in van der Vaart (1998) [248]; a high-level Bourbaki-inspired text is found in Le Cam (1986) [179], which develops a general mathematical framework for decision theory, dealing with Bayesian statistics as an important area of its application. For a more down-to-earth version of this work, applied mostly to efficient estimation in parametric models, the interested reader is referred to Le Cam and Yang (1990) [183].

The present book has grown out of a set of lecture notes that were first written for a lecture series in Bayesian statistics at the University of Amsterdam in the spring of 2007 and updated in the years since. The lectures were aimed (initially) at first-year MSc.-students in statistics, probability, mathematics and related fields like economics, computer science and physics. Over the years, these lectures have evolved into a one-semester course Bayesian statistics and frequentist optimality for third-year BSc.-students in mathematical statistics and related disciplines, in the form of lectures and exercise classes based on the material of part I. The course's goal is for students to understand the basic properties of Bayesian statistical methods; to know how analogous frequentist methods compare; to understand frequentist efficient estimation and its relation to the Bernstein-von-Mises theorem; to be able to apply this knowledge to statistical questions and to know the extent (and limitations) of conclusions based thereon. More concretely, the course covers the material

of the first four chapters of part I. All Bayesian methods are presented side-by-side with analogous frequentist methods and their criteria for optimality. Also discussed are the standard ways of choosing and constructing parametric prior distributions, by objective, subjective and empirical standards. Of course, model misspecification is a distinctly frequentist issue that is especially acute in parametric setting, so a serious note of caution cannot be omitted; accordingly, part I concludes with a fifth chapter that discusses posterior behaviour in misspecified parametric and semi-parametric models.

Ideally the BSc.-course would be followed by an MSc.-level course on the material covered in part II. Part II covers non-parametric Bayesian methods, again with a special emphasis on frequentist convergence properties and asymptotic optimality. The sixth chapter reviews asymptotic estimation, introduces convergence of posterior distributions and explains Doob's Bayesian consistency theorem. It then turns to frequentist posterior consistency with Schwartz's theorem and posterior rates of convergence with the Ghosal-Ghosh-van der Vaart theorem. As such, the sixth chapter provides an overview of the theory underpinning a large part of the existing frequentist literature and most concrete examples in non-parametric Bayesian statistics.

The seventh chapter generalizes the frequentist theory of the sixth chapter: where Schwartz's theorem poses conditions of uniform testability and lower bounds for prior mass in Kullback-Leibler neighbourhoods, chapter eight relaxes these to a less demanding, Bayesian form of testability and the requirement of remote contiguity. The central conclusion is that the existence of Bayesian tests is equivalent with Doob's form of Bayesian posterior consistency, and that remote contiguity promotes that conclusion to frequentist forms of posterior consistency. The resulting theorems for posterior convergence and uncertainty quantification are fully general (in that they permit non-*i.i.d.* (*e.g.* Markov chains of) observations, sample-size dependent parameter spaces and priors, data in non-standard forms like random graphs, *etcetera*) and make possible the asymptotic interpretation of (enlarged) credible sets as consistent confidence sets, generalizing the most important inferential consequence of the Bernstein-von Mises theorem to non-parametric setting.

The eighth chapter introduces the Dirichlet process and the family of Polya tree processes, which describe so-called inverse systems of random histograms. Inverse systems of random histograms provide an attractive way to specify distributions on the space probability measures and represent the essence of Bayesian non-parametric statistics: without referring to infinite-dimensional parametrizations, inverse limit systems define random probability distributions directly, in a way that is computationally accessible by construction. As such, inverse limit priors form the backbone for a large part of the existing literature on Bayesian non-parametric statistics and modern machine learning. In the eighth chapter, we discuss systems of random histograms and their coherence, explore the above examples, point out their conjugacy and tailfreeness and we prove that the corresponding posteriors are consistent. All of that is done under the (initially unproven) assumption that the relevant inverse systems correspond to well-defined limiting probability distributions on the space of all probability distributions. However the matter of proving such existence is a notoriously difficult mathematical problem that dominates the second half of

the chapter. Chapter eight shows that inverse systems of random histograms fall in one of three distinct classes: if probability mass is spread equitably enough over the random histogram, the limit is a Borel probability distribution for the topology of total variation with a support that is a dominated sub-model; if the distribution of probability mass is somewhat more pronounced, the limit is a Borel probability distribution for Prokhorov's topology of weak convergence and the support can be unrestricted; and if the distribution of probability mass is extremely disparate, the limit exists only on a zero-dimensional compactification of the sample space. These three 'phases' of inverse limit systems are characterized precisely in corresponding existence theorems for the inverse limits of random histograms in the second half of chapter eight.

As is apparent from chapter six onward and amplified further in chapter seven, the existence of certain sequences of hypothesis testing procedures is of essential importance to determine whether posterior distributions converge and how fast this convergence occurs. The ninth chapter analyses the existence of test sequences from various perspectives: it is investigated under which conditions Schwartz's uniform tests exist, what changes for pointwise tests and how this relates to the existence of Bayesian test sequences of chapter seven. The answers to these existence questions come in the form of fully general equivalences which characterize pairs of hypotheses that are testable and pairs that are not, without posing conditions on the model under consideration. Besides being of fundamental value for a better understanding of what statistics can be expected to achieve and what not, the fact that it is often possible to calculate or approximate posteriors concretely enables a practical method to construct tests and model selection criteria. With a prior that induces remote contiguity, such constructions are even interpretable along frequentist lines.

The tenth and eleventh chapters provide applications of the theory discussed in preceding chapters: the tenth chapter applies the theory of the sixth chapter to the errors-in-variables model for non-parametric regression, based on traditional constructions with function-space parametrizations of the model covered by controlled numbers of Hellinger balls, leading to uniform test sequences and priors placing mass in their centre points. In the eleventh chapter, we consider the more modern question of community detection in networks: given a graph with two communities of vertices known to be more highly connected within than between communities, we analyse the precise way in which the posterior concentrates around the true community structure. The graph is assumed to be an inhomogeneous Erdős-Rényi graph, with edges that occur with degrees of sparsity for which probabilists have shown detection to be only just possible consistently in the large-graph limit. Moreover, derived inequalities for posterior concentration enable the conversion of Bayesian credible sets into exact frequentist confidence sets not just in the large-graph limit (as in chapter six), but also for graphs with a finite number of vertices (as in chapter two).

An attempt has been made to make part I of this book as self-contained as possible. Although some basic experience with standard statistical methods is assumed, the mathematical aspects and optimality theory of frequentist statistical tools are explained in detail, alongside their Bayesian analogues (although some of proofs

that can be found elsewhere are omitted). For completeness appendix B summarizes the necessary elements of measure theory, with an emphasis on conditional distributions and some elaboration on Martingale convergence and Daniell-Kolmogorov existence of stochastic processes. Because asymptotic statistics essentially revolves around convergence, topology plays a central role (particularly) in the second part of this book, far more central than in most other books on mathematical statistics (with the exception of [179]). All definitions, lemmas and theorems (even in part I) are made with this topological foundation in mind. Appendix C collects basic topological definitions, properties and theorems (without proofs) and looks in particular at topologies on spaces of (probability) measures. Several specific topological subjects are discussed in some more detail: uniform spaces, Polish spaces, inverse limit spaces, function spaces, vector spaces and locally convex spaces receive extra attention and Radon measures are discussed in some detail. The most practical criteria and propositions for stochastic convergence are summarized in appendix C.9. Extra attention also goes to approximation of probability measures by means of contiguity, as discussed in subsection C.10.

For corrections to early versions of this book and corrections to the exercises, I thank Chris Muris, Audrius Jukonis, Stefano Rizelli and Mike Derksen. I thank my co-authors, Aad van der Vaart for his supervision of the work in chapter 5, Harm de With for his part in the developments of chapter 8, and Jan van Waaij for the collaboration that led to the results of chapter 11. I thank Jan van Mill for the discussions on zero-dimensional (Polish) spaces. I thank Peter Bickel for sharing his insights into the nature of statistics and reasoning under uncertainty, and the role of mathematics therein. And most of all, I thank my wife and my family for their unwavering support during the writing of this book.

Bas Kleijn
Amsterdam, September 2022

Contents

Part I
Parametric Bayesian statistics

Chapter 1
Introduction

The goal of statistical inference is to understand, describe and estimate (aspects of) the randomness of measured data. Quite naturally this invites the assumption that the data represents a sample from an unknown but fixed probability distribution.

1.1 Frequentist statistics

Any frequentist inferential procedure relies on three basic ingredients: the data, a model and an estimation procedure. The central assumption in frequentism is that the data has a definite but unknown, underlying distribution to which all inference pertains. The *data* is a measurement or observation which we denote by Y, taking values in a corresponding sample space.

Definition 1.1.1. The *sample space* for an observation Y is a measurable space $\mathscr{Y}$ with σ-*algebra* $\mathscr{B}$ (see definition B.1.5) containing all values that Y can take upon measurement.

Measurements and data can take any form, ranging from *categorical data* (sometimes referred to as *nominal data* where the sample space is simply a (usually finite) set of points or labels with no further mathematical structure), *ordinal data* (also known as *ranked data*, where the sample space is endowed with an total ordering), to *interval data* (where in addition to having an ordering, the sample space allows one to compare differences or distances between points), to *ratio data* (where we have all the structure of the real line). Moreover Y can collect the results of a number of measurements, so that it takes its values in the form of a vector (think of an experiment involving repeated, stochastically independent measurements of the same quantity, leading to a so-called independent and identically distributed (or *i.i.d.*) sample). The data Y can even be *functional data* which takes its values in a space of functions or in other infinite-dimensional spaces, for example, in the statistical study of continuous-time time-series. Y may even be a *random graph*, as in the stochastic block model of chapter 11.

The sample space $\mathscr{Y}$ is assumed to be a measurable space to enable the consideration of probability measures on $\mathscr{Y}$, formalizing the uncertainty in measurement of Y. As was said in the opening words of this chapter, frequentist statistics hinges on the assumption that there exists a probability measure $P_0 : \mathscr{B} \to [0,1]$ on the sample space $\mathscr{Y}$ representing the *"true distribution of the data"*:

$$Y \sim P_0 \tag{1.1}$$

Hence from the frequentist perspective, statistics revolves around the central question: "What does the data make clear about P_0?", which may be considered in parts by questions like, "From the data, what can we say about the mean of P_0?", "Based on the data that we have, how sharp can we formulate hypotheses concerning the value of the variance of P_0?", *etcetera*.

The second ingredient of a statistical procedure is a model, which contains all explanations under consideration of the randomness in Y. (See proposition B.2.6.)

Definition 1.1.2. A (frequentist) statistical *model* $\mathscr{P}$ is a collection of probability measures $P : \mathscr{B} \to [0,1]$ on the sample space $(\mathscr{Y}, \mathscr{B})$. The distributions P are called *model distributions*. For every sample space $(\mathscr{Y}, \mathscr{B})$, the collection $\mathscr{M}^1(\mathscr{Y}, \mathscr{B})$ of *all* probability distributions is called the *full model* (sometimes referred to as the *full non-parametric model*).

The model $\mathscr{P}$ contains the candidate distributions for Y that the statistician finds "reasonable" explanations of the uncertainty he observes (or expects to observe) in Y. As such, it constitutes a choice of the statistician analyzing the data rather than a given. From a more mathematical perspective we observe that a model $\mathscr{P}$ on $(\mathscr{Y}, \mathscr{B})$ is a subset of the space $\mathscr{M}(\mathscr{Y}, \mathscr{B})$ of all bounded, signed measures $\mu : \mathscr{B} \to \mathbb{R}$ (that is, all countably additive, real-valued set functions) that are of finite total variation. Equipped with the total-variational *norm* (see appendix B, definition B.2.5), $\mu \mapsto \|\mu\|$, $\mathscr{M}(\mathscr{Y}, \mathscr{B})$ is a Banach space [82], in which the full model can be characterized by,

$$\mathscr{M}^1(\mathscr{Y}, \mathscr{B}) = \left\{ P \in \mathscr{M}(\mathscr{Y}, \mathscr{B}) : P \geq 0, P(\mathscr{Y}) = 1 \right\}.$$

Often, we describe models as families of probability densities rather than distributions.

Definition 1.1.3. If there exists a σ-finite measure $\mu : \mathscr{B} \to [0, \infty]$ such that for all $P \in \mathscr{P}$, $P \ll \mu$, we say that the model is *dominated* (notation: $\mathscr{P} \ll \mu$).

The Radon-Nikodym theorem (see theorem B.3.10) guarantees that we may represent a dominated probability measure P in terms of a *probability density function* $p = dP/d\mu : \mathscr{Y} \to [0, \infty)$ that satisfies $\int_A p(y)\, d\mu(y) = P(A)$ for all $A \in \mathscr{B}$. For dominated models, it makes sense to adopt a slightly different mathematical perspective: if μ dominates $\mathscr{P}$, we map $\mathscr{P}$ to the space of all μ-integrable functions $L_1(\mu)$ by means of the Radon-Nikodym mapping.

Example 1.1.4. Suppose that $\mathscr{Y}$ is countable (and let $\mathscr{B}$ be the powerset of $\mathscr{Y}$): then the measure μ that puts mass one at every point in $\mathscr{Y}$, also known as the *counting*

measure on $\mathscr{Y}$, is σ-finite and dominates every other (bounded) measure on $\mathscr{Y}$. Consequently, any model on $(\mathscr{Y},\mathscr{B})$ can be represented in terms of elements p in the Banach space $L_1(\mu)$, more commonly denoted as ℓ_1,

$$\ell_1 = \Big\{(f_1,f_2,\ldots) \in [0,1]^\infty : \sum_{i\geq 1} |f_i| < \infty\Big\}. \tag{1.2}$$

where it is noted that $p_i \geq 0$ and $\|p\| = \sum_i p_i = 1$ for all P in the set Λ of all probability measures on $(\mathscr{Y},\mathscr{B})$.

In case the sample space is not discrete, the full model is not dominated by a σ-finite measure (see exercise 1.6.3). However, suppose that a σ-finite measure μ on the sample space is given. The Radon-Nikodym mapping maps every μ-dominated model $\mathscr{P}$ to a subset of,

$$\mathscr{M}^1(\mu) = \Big\{p \in L_1(\mu) \,:\, p \geq 0, \int_{\mathscr{Y}} p(y)\,d\mu(y) = 1\Big\}.$$

For the following proposition, the model is endowed with a metric in the form of the total-variational

Proposition 1.1.5. *The mapping between a model $\mathscr{P}$ dominated by a σ-finite measure μ and its $L_1(\mu)$-representation is an isometry: for all $p_1, p_2 \in \mathscr{P}$,*

$$\|P_1 - P_2\| = \tfrac{1}{2}\int_{\mathscr{Y}} \big|p_1(y) - p_2(y)\big|\,d\mu(y) = \int_{\mathscr{Y}} \big(p_1(y) - p_2(y)\big)_+\,d\mu(y).$$

(The proof is given in exercises 1.6.2, 4.4.7.) Note that a dominating measure is not unique, so there are many L_1-representations of $\mathscr{P}$. The most common way of representing a statistical model is a description in terms of a parametrization.

Definition 1.1.6. A model $\mathscr{P}$ is *parametrized* with *parameter space* Θ, if there exists a surjective map $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$, called the parametrization of $\mathscr{P}$.

Parametrizations are motivated by the context of the statistical question and the parameter θ usually has a clear interpretation when viewed in this context. The formulation of parametric models constitutes the modelling step of statistics: to the statistician, it transforms the data from a mere list of numbers to an informative (but noisy) representation of an underlying truth.

Definition 1.1.7. A parametrization of a statistical model $\mathscr{P}$ is said to be *identifiable*, if the map $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$ is injective.

Injectivity of the parametrization means that for all $\theta_1, \theta_2 \in \Theta$, $\theta_1 \neq \theta_2$ implies that $P_{\theta_1} \neq P_{\theta_2}$: no two different parameter values θ_1 and θ_2 give rise to the same distribution. Clearly, in order for $\theta \in \Theta$ to serve as a useful representation for the candidate distributions P_θ, identifiability is a first requirement. Other common conditions on the map $\theta \mapsto P_\theta$ are continuity (for example, with respect to the total-variational norm, or through requiring continuity of all maps $\theta \mapsto P_\theta g$, with g any bounded

measurable function), differentiability (the definition of which may involve technical subtleties in case Θ is infinite-dimensional) and other smoothness conditions (*e.g.* definition 4.1.12).

Remark 1.1.8. Although strictly speaking ambiguous, it is commonplace to refer to both $\mathscr{P}$ and the parametrizing space Θ as "the model". This practice is not unreasonable in view of the fact that, in practice, almost all models are parametrized in an identifiable way, so that there exists a bijective correspondence between Θ and $\mathscr{P}$. Here, reference to the model always concerns $\mathscr{P}$ while Θ is always called the parameter space.

An assumption often made in frequentist statistics is that the true distribution of the data is a model distribution.

Definition 1.1.9. A model $\mathscr{P}$ is said to be *well-specified* if it contains the true distribution of the data P_0, *i.e.*

$$P_0 \in \mathscr{P}. \tag{1.3}$$

If (1.3) does not hold, the model is said to be misspecified.

Clearly if $\mathscr{P}$ is parametrized by Θ, (1.3) implies the existence of a point $\theta_0 \in \Theta$ such that $P_{\theta_0} = P_0$; if, in addition, the model is identifiable, the parameter value θ_0 is unique.

If the full non-parametric model is used, (1.3) holds trivially. However, for smaller models, (1.3) has the status of an assumption on the unknown quantity of interest P_0 and may as such be hard to justify. The reason for (the somewhat odd and certainly very contentious) assumption (1.3) lies in the interpretation of statistical conclusions: an estimate of a parameter is of value if that parameter can be attributed to the "true" distribution of the data. If, on the other hand, one assumes that the model is mis-specified, parameter estimates may reflect aspects of the true distribution but cannot be associated with the true distribution of the data directly any more.

The model we use in a statistical procedure constitutes a *choice* rather than a given: presented with a particular statistical problem, different statisticians may choose to use different models. The only condition is that (1.3) is satisfied, which is why we have to choose the model in a "reasonable way" given the nature of Y. When choosing the model, two considerations compete: on the one hand, small models are easy to handle mathematically and statistically and parameters usually have clear interpretions, on the other hand, for large models, assumption (1.3) is more realistic since they have a better chance of containing P_0 (or at least approximate it more closely). The amount of data available plays a crucial role: if we have a limited sample, simple models have a better chance of leading to sensible results, while an abundance of data enables more sophisticated forms of statistical analysis. In this respect the most important distinction is made in terms of the dimension of the model.

Definition 1.1.10. A model $\mathscr{P}$ is said to be *parametric* of *dimension* d, if there exists an identifiable parametrization $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$, where $\Theta \subset \mathbb{R}^d$ with non-empty interior, $\mathring{\Theta} \neq \varnothing$.

The requirement regarding the interior of Θ in definition 1.1.10 ensures that the dimension d really concerns Θ and not just the dimension of the space $\mathbb{R}^d$ (in which Θ could otherwise be a lower-dimensional subset).

Example 1.1.11. The *normal model* for a single, real measurement Y, is the collection of all normal distributions on $\mathbb{R}$, *i.e.*

$$\mathscr{P} = \left\{N(\mu,\sigma^2) : (\mu,\sigma) \in \Theta\right\}$$

where the parametrizing space Θ equals $\mathbb{R}\times(0,\infty)$. The map $(\mu,\sigma)\mapsto N(\mu,\sigma^2)$ is surjective and injective, *i.e.* the normal model is a two-dimensional, identifiable parametric model. Moreover, the normal model is dominated by the Lebesgue measure on the sample space $\mathbb{R}$ and can hence be described in terms of Lebesgue-densities:

$$p_{\mu,\sigma}(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}.$$

Note that for any fixed $y\in\mathscr{Y}$, the dependence $\Theta\to\mathbb{R} : (\mu,\sigma)\mapsto p_{\mu,\sigma}(y)$ is continuous on all of Θ. So if (μ_n,σ_n) converges to (μ,σ) in Θ, then $p_n(y) := p_{\mu_n,\sigma_n}(y)$ converges to $p(y) := p_{\mu,\sigma}(y)$. Then total-variational distance between the distributions P_n and P (associated with the densities p_n and p respectively) satisfies,

$$\|P_n - P\| = \tfrac{1}{2}\int_{\mathscr{Y}} \left|p_n(y) - p(y)\right| d\mu(y) \to 0.$$

by proposition 1.1.5 and Scheffé's lemma (see corollary C.9.9). Conclude that the parametrization $\Theta\to\mathscr{P} : (\mu,\sigma)\mapsto P_{\mu,\sigma}$ is continuous with respect to the total-variational metric on $\mathscr{P}$.

Definition 1.1.12. If there is no finite-dimensional Θ that parametrizes $\mathscr{P}$, then $\mathscr{P}$ is called a *non-parametric model*.

For instance, the full model $\mathscr{M}^1(\mathscr{Y},\mathscr{B})$ is non-parametric unless the sample space contains only a finite number of points.

Example 1.1.13. Let $\mathscr{Y}$ be a finite set containing $n\geq 1$ points $y_1,y_2,\ldots,y_n$ and let $\mathscr{B}$ be the *power-set* $2^{\mathscr{Y}}$ of $\mathscr{Y}$. Any probability measure $P:\mathscr{B}\to[0,1]$ on $(\mathscr{Y},\mathscr{B})$ is absolutely continuous with respect to the *counting measure* on $\mathscr{Y}$ (see example B.2.8). The density of P with respect to the counting measure is a map $p:\mathscr{Y}\to\mathbb{R}$ such that $p\geq 0$ and

$$\sum_{i=1}^{n} p(y_i) = 1.$$

As such, P can be identified with an element of the so-called *simplex* S_n in $\mathbb{R}^n$, defined as follows

$$S_n = \left\{p=(p_1,\ldots,p_n)\in\mathbb{R}^n \,:\, p_i\geq 0, \sum_{i=1}^{n} p_i = 1\right\}. \tag{1.4}$$

This leads to an identifiable parametrization $S_n \to \mathscr{P} : p \mapsto P$ of the full model on $(\mathscr{Y}, \mathscr{B})$, of dimension $n-1$. Note that S_n has empty interior in $\mathbb{R}^n$, but can be brought in one-to-one correspondence with a compact set in $\mathbb{R}^{n-1}$ with non-empty interior by the embedding:

$$\Big\{(p_1,\dots,p_{n-1}) \in \mathbb{R}^{n-1} \,:\, p_i \geq 0, \sum_{i=1}^{n-1} p_i \leq 1\Big\} \to S_n :$$
$$(p_1,\dots,p_{n-1}) \mapsto \Big(p_1,\dots,p_{n-1}, 1-\sum_{i=1}^{n-1} p_i\Big).$$

1.2 Frequentist estimation

The third ingredient of a frequentist inferential procedure is an estimation method. Clearly not all statistical problems involve an explicit estimation step and of those that do, not all estimate the distribution P_0 directly. Nevertheless, one may regard the problem of point-estimation in the model $\mathscr{P}$ as prototypical.

Definition 1.2.1. A *point-estimator* (or *estimator*) for P_0 (in $\mathscr{P}$) is a map $\widehat{P} : \mathscr{Y} \to \mathscr{P}$, representing our "best guess" $\widehat{P}(Y)$ in $\mathscr{P}$ for P_0 based on the data Y (and other known quantities).

Note that a point-estimator is a *statistic*: since a point-estimator must be calculable in practice, it may depend only on information that is *known* to the statistician after he has performed the measurement realized as $Y = y$. Also note that a point-estimator is a stochastic quantity: $\widehat{P} = \widehat{P}(Y)$ depends on Y and is hence random. Upon measurement of Y resulting in a realisation $Y = y$, the realisation of the estimator is an *estimate* $\widehat{P}(y)$, a definite point in $\mathscr{P}$. If the model is parametrized, one may define a point-estimator $\widehat{\theta} : \mathscr{Y} \to \Theta$ for θ_0, from which we obtain $\widehat{P} = P_{\widehat{\theta}}$ as an estimator for P_0. In that case the continuity requirement we impose on the map $\theta \mapsto P_\theta$ guarantees that $\theta \to \theta_0$ implies $P_\theta \to P_{\theta_0}$. If the model is identifiable, estimation of θ_0 in Θ is equivalent to estimation of P_0 in $\mathscr{P}$.

Aside from estimates for the distribution P_0, one is often interested in estimating only certain aspects of P_0.

Example 1.2.2. Suppose that a bank tries to assess market risk for an asset: they have the asset on the books for price x but tomorrow's market will say that it is worth a price X, distributed according to an unknown P_0. To assess the risk of holding the position until tomorrow, the absolute return $X - x$ is of importance. Of course, the bank would prefer to have a reliable estimate for P_0 (and thus for the distribution of $X - x$) but that question is often too hard to answer and reliability cannot be guaranteed. Instead, the bank will resort to a simplification by focussing on the aspect of the distribution P_0 that they find most important for their risk assessment. A popular notion in this context is a quantity called value-at-risk: given a time-horizon (in this case, tomorrow) and a significance level $\alpha \in (0,1)$ (often chosen

equal to 0.05 or 0.01), value-at-risk q is defined as the maximal $q < 0$ at which,

$$P_0(X - x < q) \leq \alpha.$$

To interpret q, note that losses exceeding value-at-risk occur on only an expected fraction α of all trading days. In statistical terms, q is a quantile of P_0.

Another example occurs in parametric models: if the dimension d of a parametric model is greater than one, we may choose to estimate only one component of θ (called the *parameter of interest*) and disregard other components (called *nuisance parameters*). More generally, we may choose to estimate certain properties of P_0 (*e.g.* its expectation, variance) rather than P_0 itself and in many cases, direct estimation of the property of interest of P_0 is more efficient than estimation through $\widehat{P}$.

Example 1.2.3. Consider a model $\mathscr{P}$ consisting of distributions on $\mathbb{R}$ with finite expectation and define the functional $e : \mathscr{P} \to \mathbb{R}$ by the expectation $e(P) = PX$. Suppose that we are interested in the expectation $e_0 = e(P_0)$ of the true distribution. Obviously, based on an estimator $\widehat{P}$ for P_0 we may define an estimator,

$$\widehat{e} = \int_{\mathbb{R}} x d\widehat{P}(x) \tag{1.5}$$

to estimate e_0. For instance, assume that X is integrable under P_0 and $Y = (X_1, \ldots, X_n)$ collects the results of an *i.i.d.* experiment with $X_i \sim P_0$ marginally (for all $1 \leq i \leq n$), then the *empirical expectation* of X, defined simply as the *sample-average* of X,

$$\mathbb{P}_n X = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

provides an estimator for e_0. (Note that the sample-average is also of the form (1.5) if we choose as our point-estimator for P_0 the empirical distribution $\widehat{P} = \mathbb{P}_n$, see example B.2.10.) The *law of large numbers* guarantees that $\mathbb{P}_n X$ converges to e_0 almost-surely as $n \to \infty$ (*consistency*, as in definition 6.1.1), and (if X is quadratically integrable) the *central limit theorem* asserts that this convergence proceeds at *rate* $n^{-1/2}$ (as in definition 4.1.3) and that the *limit distribution* (see definition 4.1.5) is zero-mean normal with $P_0(X - P_0 X)^2$ as its variance. Many parametrizations $\theta \mapsto P_\theta$ are such that (components of) θ coincide with expectations. Often, other properties of P_0 can also be related to expectations: for example, if $X \in \mathbb{R}$, the probabilities $F_0(s) = P_0(X \leq s) = P_0 1\{X \leq s\}$ $(s \in \mathbb{R})$ can be estimated by the so-called *empirical distribution function*,

$$\mathbb{F}_n(s) = \frac{1}{n} \sum_{i=1}^{n} 1\{X_i \leq s\},$$

i.e. as the empirical expectation of the function $x \mapsto 1\{x \leq s\}$. This leads to a step-function with n jumps of size $1/n$ at sample-points, which estimates the distribution function F_0. Generalizing, any property of P_0 that can be expressed in terms

of an expectation of a P_0-integrable function of X, $P_0 g$, is estimable by the corresponding empirical expectation, $\mathbb{P}_n g$. (With regard to the estimator $\mathbb{F}_n$, the convergence $\mathbb{F}_n(s) \to F_0(s)$ does not only hold pointwise but even uniform in s, *i.e.* $\sup_{s\in\mathbb{R}} |\mathbb{F}_n(s) - F_0(s)| \to 0$ almost-surely, *c.f.* the *Glivenko-Cantelli theorem*.)

To estimate a probability distribution (or any of its properties or parameters), many different estimators may exist. Therefore, the use of any particular estimator constitutes (another) *choice* made by the statistician analyzing the problem. Whether such a choice is a good or a bad one depends on *optimality criteria*, which are either dictated by the particular nature of the problem (see section 2.5 which extends the purely inferential point of view), or based on more generically desirable properties of the estimator. (This explains the use of the rather vague qualification "best guess" in definition 1.2.1.)

Example 1.2.4. To illustrate what we mean by "desirable properties", note the following. When estimating P_0 one may decide to use an estimator $\widehat{P}$ because it has the property that it is close to the true distribution of Y in total variation: there exist small constants $\varepsilon > 0$ and $\alpha > 0$ such that for all $P \in \mathscr{P}$,

$$P\big(\|\widehat{P}(Y) - P\| < \varepsilon\big) > 1 - \alpha,$$

i.e. if $Y \sim P$, then $\widehat{P}$ lies close to P with high P-probability. Note that we formulate this property "for all P in the model": since $P_0 \in \mathscr{P}$ is unknown, the only way to guarantee that this property holds under P_0, is to prove that it holds for all $P \in \mathscr{P}$ (provided that (1.3) holds). By contrast, for Bayesians any claim concerning points P in the model is acceptable if it is true almost-everywhere in $\mathscr{P}$ with respect to the prior measure.

A popular method of estimation that satisfies common optimality criteria in many (but certainly not all, see [183]) problems is maximum-likelihood estimation.

Definition 1.2.5. Suppose that the model $\mathscr{P}$ is dominated by a σ-finite measure μ and parametrized through μ-densities by $\theta \mapsto p_\theta \in L_1(\mu)$. The *likelihood principle* (see [218] for an elaborate overview and spirited argument in favour) says that all information implied by data Y concerning the parameter θ is contained in the *likelihood-function* $\theta \mapsto p_\theta(Y)$ (note that this defines a *random* function $\theta \to [0,\infty]$). Accordingly, one can define $\hat{\theta} \in \Theta$ as an estimator for the true parameter value θ_0 by maximization,

$$p_{\hat{\theta}}(Y) = \sup_{\theta\in\Theta} p_\theta(Y).$$

So $\hat{\theta}$ is the point in the parameter space for which the likelihood-function evaluated in Y, $\Theta \to [0,\infty] : \theta \mapsto p_\theta(Y)$ attains its maximum. This defines the *maximum-likelihood estimator* (or *MLE*) $\hat{\theta}$ for θ_0.

Remark 1.2.6. The MLE $\widehat{P} = P_\theta$ does not depend on the dominating measure μ chosen to define the densities $p_\theta = dP_\theta/d\mu$.

A word of caution is in order: mathematically, the above "definition" of the MLE begs questions of existence and uniqueness: regarding $\theta \mapsto p_\theta(Y)$ as a (random) map on the parameter space, there may not be any point in $\mathscr{P}$ where the likelihood takes on its supremal value (with P_0-probability one), nor is there any guarantee that such a maximal point is unique (with P_0-probability one). Additional model properties are needed to alleviate these two issues. (For example, if the likelihood depends continuously on an parameter from a compact parameter space, existence is guaranteed.)

The above is only a very brief and rather abstract overview of the basic framework of frequentist statistics, highlighting the central premise that a true underlying distribution P_0 for Y exists. It makes clear, however, that frequentist inference concerns itself primarily with the stochastics of the random variable Y and not with the *context* in which Y resides. Other than the fact that the model has to be chosen "reasonably" based on the nature of Y, frequentist inference does not involve any information regarding the background of the statistical problem in its procedures unless one chooses to use such information explicitly (see, for example, remark 2.2.21 on penalized maximum-likelihood estimation). In Bayesian statistics the use of background information is an integral part of the procedure unless one chooses to disregard it: by the definition of a prior measure, the statistician may express that he believes in certain points of the model more strongly than others. This thought is elaborated on further in section 1.3 (*e.g.* example 1.3.1).

Similarly, results of estimation procedures are sensitive to the context in which they are used: two statistical experiments may give rise to the same model formally, but the estimator used in one experiment may be totally unfit for use in the other experiment.

Example 1.2.7. For example, if we interested in a statistic that predicts the rise or fall of a certain share-price on the stock market based on its value over the past week, the estimator we use does not have to be a very conservative one: we are interested primarily in its long-term performance and not in the occasional mistaken prediction. However, if we wish to predict the rise or fall of white-bloodcell counts in an HIV-patient based on last week's counts, overly optimistic predictions lead to tragic consequences, and far more conservative statistical methods are called for.

Although in the above example, data and models are very similar, the estimator used in the medical application should be much more conservative than the estimator used in the stock-market problem. The purely statistical aspects of both questions are the same, but the context in which inference is expressed calls for different approaches. Such considerations form the motivation for statistical decision theory, as explained further in section 2.5.

1.3 Bayesian statistics

Bayesian statistics provides an alternative approach to statistical questions, named after Rev. Thomas Bayes, the author of "*An essay towards solving a problem in the doctrine of chances*" published posthumously in 1763 [13]. Bayes considered a number of probabilistic questions in which data and parameters are treated on equal footing. The Bayesian procedure itself is explained in detail in chapter 2 and further chapters explore its properties. In this section we have the more modest goal of illustrating the conceptual differences with frequentist statistical analysis.

In Bayesian statistics, data and model form two factors of the same space, *i.e.* no formal distinction is made between measured quantities Y and parameters θ. One may envisage the process of generating a realized observation $Y = y$ as two draws, one draw from Θ to select a value of θ, and a subsequent draw from the model distribution P_θ to arrive at the realization $Y = y$. This perspective may seem rather strange in view of the definitions made in section 1.1, but in [13], Bayes gives examples in which this perspective is perfectly reasonable (see subsection 2.1.2). An element P_θ of the model is interpreted simply as the distribution of Y *given* the parameter value θ, *i.e.* as the conditional distribution of $Y|\theta$. The joint distribution of (Y, θ) then follows upon specification of the marginal distribution of θ on Θ, which is called the *prior*. Based on the joint distribution for the data Y and the parameter θ, straightforward conditioning on Y gives rise to a conditional distribution $\Pi(\cdot|Y)$ for the parameter $\theta|Y$ called the *posterior distribution* on the model Θ. Hence, given the model, the data and a prior distribution, the Bayesian procedure leads to a posterior distribution that incorporates the information provided by the data. All statistical questions are then answered using the posterior. For example, what a frequentist would call point-estimation of the underlying distribution with the posterior expectation,

$$P^{\Pi|Y}(A) = \int_\Theta P_\theta(A)\, d\Pi(\theta|Y),$$

(for all measurable A), is called *prediction* by Bayesians, who refer to $P^{\Pi|Y}$ as the *posterior predictive distribution*.

Often in applications, the nature of the data and the background of the problem suggest that certain values of θ are more "likely" than others, even before any measurements are done. The model $\mathscr{P}$ describes possible probabilistic explanations of the data and, in a sense, the statistician believes more strongly in certain explanations than in others. This is illustrated by the following example, which is due to L. Savage (1961) [222].

Example 1.3.1. Consider the following three statistical experiments:

1. A lady who drinks milk in her tea claims to be able to tell which was poured first, the tea or the milk. In ten trials, she determines correctly whether it was tea or milk that entered the cups first.

2. A music expert claims to be able to tell whether a page of music was written by Haydn or by Mozart. In ten trials conducted, he correctly determines the composer every time.
3. A drunken friend says that he can predict the outcome of a fair coin-flip. In ten trials, he is right every time.

Let us analyze these three experiments in a frequentist fashion, *e.g.* we assume that the trials are independent and possess a definite Bernoulli distribution, *c.f.* (1.1). In all three experiments, $\theta_0 \in \Theta = [0,1]$ is the per-trial probability that the person gives the right answer. We test their respective claims posing the hypotheses (see exercise 1.6.5):

$$H_0: \quad \theta_0 = \tfrac{1}{2}, \qquad H_1: \quad \theta_0 > \tfrac{1}{2}.$$

The total number of successes out of ten trials is a sufficient statistic for θ and we use it as our test-statistic, noting that its distribution is binomial with $n = 10$, $\theta = \theta_0$ under H_0. Given the data Y with realization y of ten correct answers, applicable in all three examples, we reject H_0 at p-value $2^{-10} \approx 0.1\%$. So there is strong evidence to support the claims made in all three cases. Note that there is no difference in the frequentist analyses: formally, all three cases are treated exactly the same.

Yet intuitively (and also in every-day practice), one would be inclined to treat the three claims on different footing: in the second experiment, we have no reason to doubt the expert's claim, whereas in the third case, the friend's condition makes his claim less than plausible. In the first experiment, the validity of the lady's claim is hard to guess beforehand. The outcome of the experiments would be as expected in the second case and remarkable in the first. In the third case, one would either consider the friend extremely lucky, or begin to doubt the fairness of the coin being flipped.

The above example convincingly makes the point that in our intuitive approach to statistical issues, we include *all* knowledge we have, even resorting to strongly biased estimators if the model does not permit a non-biased way to incorporate it. The Bayesian approach to statistics allows us to choose priors that reflect this subjectivity: from the outset, we attach more prior mass to parameter-values that we deem more likely, or that we believe in more strongly. In the above example, we would choose a prior that concentrates more mass at high values of θ in the second case and at low values in the third case. In the first case, the absence of prior knowledge would lead us to remain objective, attaching equal prior weights to high and low values of θ. Although the frequentist's testing procedure can be adapted to reflect subjectivity, the Bayesian procedure incorporates it rather more naturally through the choice of a prior.

Subjectivist Bayesians view the above as an advantage; objectivist Bayesians and frequentists view it as a disadvantage. Subjectivist Bayesians argue that personal beliefs are an essential part of statistical reasoning, deserving of a explicit role in the formalism and interpretation of results. Objectivist Bayesians and frequentists reject this thought because scientific reasoning should be devoid of any personal beliefs or interpretation (see section 3.2). So the above freedom in the choice of the prior is also the Achilles heel of Bayesian statistics: fervent frequentists and objectivist

Bayesians take the point of view that the choice of prior is an undesirable source of ambiguity, rather than a welcome way to incorporate "expert knowledge" as in example 1.3.1. After all, if the subjectivist Bayesian does not like the outcome of his analysis, he can just go back and change the prior to obtain a different outcome. Similarly, if two subjectivist Bayesians analyze the same data they may reach completely different conclusions, depending on the extent to which their respective priors differ.

To a certain extent such ambiguity is also present in frequentist statistics, since frequentists have the freedom to choose biased point-estimators. For example, the use of either a maximum-likelihood or penalized maximum-likelihood estimator leads to differences, the size of which depends on the relative sizes of likelihood and penalty. Indeed, through the maximum-a-posteriori Bayesian point-estimator (see definition 2.2.20), one can demonstrate that the log-prior-density can be viewed as a penalty term in a penalized maximum-likelihood procedure, *c.f.* remark 2.2.21. But the way in which subjectivity is expressed in the Bayesian setting is integral to the approach and completely natural.

A second difference in philosophy between frequentist and Bayesian statisticians arises as a result of the fact that the Bayesian procedure does not require that we presume the existence of a "true, underlying distribution" P_0 of Y (compare with (1.1)). The subjectivist Bayesian views the model with (prior or posterior) distribution as his own, subjective explanation of the uncertainty in the data. For that reason, subjectivists prefer to talk about their (prior or posterior) *"belief"* concerning parameter values rather than implying objective validity of their assertions. On the one hand, such a point of view makes intrinsic ambiguities surrounding statistical procedures explicit; on the other hand, one may wonder about the relevance of strictly personal belief in a scientific tradition that emphasizes universality of reported results.

The philosophical debate between Bayesians and frequentist has raged with varying intensity for decades, but remains undecided to this date. In practice, the choice for a Bayesian or frequentist estimation procedure is usually not motivated by philosophical considerations, but by far more practical issues, such as ease of computation and implementation, common custom in the relevant field of application, specific expertise of the researcher or other forms of simple convenience. More recent developments [12] suggest that the philosophical debate will be put to rest in favour of more practical considerations as well. In later chapters it is demonstrated how Bayesian and frequentist statistical limits are related in the large-sample asymptotic regime.

1.4 The frequentist analysis of Bayesian methods

Since this point has the potential to cause great confusion, we emphasize the following: this text presents Bayesian statistics from a hybrid perspective, *i.e.* we consider Bayesian techniques but analyze them in frequentist setting and with frequentist methods.

We take the frequentist point of view with regard to the data, *e.g.* assumption (1.1); we distinguish between sample space and model and we do not adhere to subjectivist interpretations of results (although their perspective is discussed in the main text). On the other hand, we endow the model with a prior probability measure and calculate the posterior distribution, *i.e.* we use concepts and definitions from Bayesian statistics. This enables us to assess Bayesian methods on equal footing with frequentist statistical methods and extends the range of interesting questions. Note, however, that the derivation of expression (2.13) for the posterior, for example, is the result of subjectivist Bayesian assumptions on data and model. Since these assumptions are at odds with the frequentist perspective, we shall take (2.13) as a *definition* rather than a derived form (see subsection 2.1.4).

Much of the material covered in this book does not depend on any particular philosophical point of view, especially when the subject matter is purely mathematical. Nevertheless, it is important to realize when philosophical issues may come into play and there will be points where this is the case. In particular when discussing asymptotic properties of Bayesian procedures, adoption of assumption (1.1) is instrumental (lacking a limit point representing the (frequentist) true distribution of the data, any form of asymptotic convergence would be much less meaningful).

1.5 Markov-chain Monte-Carlo simulation [EMPTY]

1.6 Exercises

1.6.1. Let $Y \in \mathscr{Y}$ be a random variable with unknown distribution P_0. Let $\mathscr{P}$ be a model for Y, dominated by a σ-finite measure μ and parametrized by $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$. Assume that the *maximum-likelihood estimator* $\hat{\theta}$ (see definition 1.2.5) is well-defined, P_0-almost-surely. Show that if ν is another σ-finite measure dominating $\mathscr{P}$ and we calculate the likelihood using ν-densities, then the associated MLE is equal to $\hat{\theta}$. Conclude that the MLE does not depend on the dominating measure used, *c.f.* remark 1.2.6.

1.6.2. Prove proposition 1.1.5.

1.6.3. Let $\mathscr{Y} = \mathbb{R}$ with σ-*algebra* $\mathscr{B}$. Show that if $\mathscr{B}$ is the usual Borel σ-algebra, then $\mathscr{M}^1(\mathscr{Y}, \mathscr{B})$ is not dominated. Also show, that if $\mathscr{B}$ is generated by the collection of all half-open intervals $(x, y]$, where $x, y \in \mathbb{Z}$, $x < y$, then $\mathscr{M}^1(\mathscr{Y}, \mathscr{B})$ is dominated.

1.6.4. Although customarily the model is defined first and estimators follow, it is possible to reverse the order: suppose that we have a certain fixed estimator in mind, how should we choose the model in order for the fixed estimator to perform?

More explicitly, consider a data vector $Y = (X_1, \ldots, X_n)$ that forms an *i.i.d.* sample from a unknown distribution P_0 on $\mathbb{R}$. We are interested in estimation of the quantity $\psi = P_0 g(X)$, assumed to be finite, where $g : \mathbb{R} \to \mathbb{R}$ is a given measurable function

defined on the sample space for X. Examples: if g is the identity map, then ψ is the expectation of X; if $g = (X - P_0X)^2$, then ψ is the variance of X; if $g = 1\{X \leq x\}$ for some $x \in \mathbb{R}$, then $\psi = F_0(x)$, the value of the distribution function associated with P_0 at x. In such cases, estimation of ψ by the sample mean $\widehat{\psi}_n := \mathbb{P}_n g$ appears sensible for large n.

a. A minimal requirement for $\widehat{\psi}_n$ to make sense as an estimator for ψ, is *consistency* (see also definition 6.1.1): we say that the estimators $\widehat{\psi}_n$ are consistent if we are able to estimate ψ with arbitrarily high precision when we raise the amount of data used (that is, the number n) high enough. Based on consistency, characterize the largest model $\mathscr{P}$ in which estimation of ψ by $\widehat{\psi}_n$ makes sense. What can be said of $\mathscr{P}$ if g is a bounded function?

To analyze the behaviour of $\widehat{\psi}_n$ in some more detail, consider the following.

b. Restrict the model $\mathscr{P}$ further based on a more detailed criterion for convergence: $\widehat{\psi}_n$ is said to *converge to ψ at rate $n^{-1/2}$* (see also definition 4.1.3), if for any $\varepsilon > 0$ there is an $M > 0$, such that,

$$\sup_{n \geq 1} P_0^n \left(n^{1/2} |\widehat{\psi}_n - \psi| > M \right) < \varepsilon,$$

as $n \to \infty$. *Hint: consider the central limit theorem, which says that (under a certain integrability condition), the $n^{1/2}$-rescaled differences converge weakly to a normal distribution. Next note that any weakly convergent sequence is uniformly tight (see definition C.7.14) and use the Heine-Borel characterization of compactness in $\mathbb{R}$ to finish the argument.*

c. Compare the property under *b.* above with example 1.2.4 and state in words how the quality of $\widehat{\psi}_n$ as an estimator for ψ improves as $n \to \infty$.

1.6.5. In the three experiments of example 1.3.1, describe a test for hypotheses H_0 and H_1 at level $\alpha \in (0,1)$, for example the likelihood ratio test. Calculate the p-value of the realization of 10 successes and 0 failures (in 10 Bernoulli trials according to H_0).

Chapter 2
Bayesian basics

In this chapter, we consider the basic definitions and properties of Bayesian statistical and decision-theoretic methods. We derive the posterior distribution from data, model and prior and we discuss how the posterior should be viewed if one assumes the frequentist point of view of section 1.1. In section 2.2 we consider point estimators derived from the posterior and in section 2.3 we discuss confidence sets and credible sets. Section 2.4 discusses the Neyman-Pearson theory of hypothesis testing, as well as a brief introduction to the Le Cam's theory of asymptotically optimal test sequences and, of course, posterior odds and Bayes factors. Section 2.5 concludes the chapter with a discussion of minimax risk and Bayes risk, with their respective versions of decision theory. Throughout the chapter the explicit goal is to consider frequentist methods side-by-side with the Bayesian procedures, for comparison and reference. In chapter 7 we consider the condition that enable frequentist interpretation of Bayesian methods in the large-sample limit.

2.1 Bayes's rule, prior and posterior distributions

In this section, we introduce the basic definitions and procedures in Bayesian statistics. Formalizing the Bayesian procedure can be done in several ways. We start this section with considerations that are traditionally qualified as being of a "subjectivist" nature: in subsection 2.1.1 we derive the relation between data, model and prior on the one hand and the posterior on the other, based on Bayes's Rule without reference to the frequentist's "true distribution of the data". To stay clear on what a Bayesian means when we speaks of a model, we consider the support of a prior (see subsection 2.1.3) and consider a prototypical example usually referred to as Bayes's Billiard in subsection 2.1.2. In subsection 2.1.4 we revert to the "frequentist" point of view through an assumption relating the "true distribution of the data" to the prior predictive distribution (see definition 2.1.4).

2.1.1 Bayes's rule

The Bayesian framework does not just view the data Y as a random variable but casts the parameter in that form as well. The parameter space Θ is assumed to be a measurable space, with σ-algebra $\mathscr{G}$ and, rather than just taking on fixed values θ as in the frequentist case, the parameter is represented by a random variable ϑ taking values in Θ. We assume that on the product-space $\mathscr{Y} \times \Theta$ (with product σ-algebra $\sigma(\mathscr{B} \times \mathscr{G})$) we have a probability measure,

$$\Pi^* : \sigma(\mathscr{B} \times \mathscr{G}) \to [0,1]. \tag{2.1}$$

The probability measure Π^* provides a joint probability distribution for (Y, ϑ), where Y is the observation and ϑ (the random variable associated with) the parameter of the model.

Implicitly the choice for the measure Π^* defines the model in Bayesian context, by the possibility to condition the distribution of Y on fixed values $\vartheta = \theta$ in Θ. The *conditional distribution* for $Y|\vartheta$ (see appendix B.4) describes the distribution of the observation Y *given* the parameter ϑ. As such, the distributions for $Y|\vartheta = \theta$ can be identified with the elements P_θ of what was referred to as a parametrized model $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ in chapter 1.

Definition 2.1.1. The distribution of the data Y conditional on the parameter ϑ (*c.f.* definition B.4.4) is a regular conditional distribution,

$$\Pi_{Y|\vartheta} : \mathscr{B} \times \Theta \to [0,1], \tag{2.2}$$

which describes the *model distributions*.

(see definition B.4.5). Since conditional probabilities are defined almost-surely with respect to the marginal (see definition B.4.4), the Bayesian notion of a *model* is represented only up to null-sets of the marginal distribution of ϑ: we may add to or remove from the model at will, as long as we make sure that the changes have prior measure equal to zero: in the Bayesian perspective, the model itself is a Π-almost-sure concept.

Definition 2.1.2. The marginal distribution $\Pi : \mathscr{G} \to [0,1]$ for ϑ is called the *prior*.

The prior is interpreted in the subjectivist's philosophy as the "degree of belief" attached to subsets of the model *a priori*, that is, before any observation has been made or incorporated in the calculation. It is important to note that Π^* is usually constructed by choice of a prior measure Π for ϑ and model distributions $\theta \mapsto P_\theta$,

$$\Pi^*(B \times G) = \int_G \Pi(B|\vartheta = \theta)\, d\Pi(\theta) = \int_G P_\theta(B)\, d\Pi(\theta),$$

for all $B \in \mathscr{B}$ and $G \in \mathscr{G}$ (where $\theta \to P_\theta(B)$ is assumed to be $\mathscr{G}$-measurable for all $B \in \mathscr{B}$). Central to the Bayesian framework is the conditional distribution for ϑ given Y, called the posterior. The transition from prior to posterior represents

the way in which "prior belief" is turned into "posterior belief" (concerning the parameter) based on the data. The posterior is interpreted as a data-amended version of the prior, that is to say, the subjectivist's original "degree of belief" corrected by observation of Y through conditioning. Below we define the posterior in conjunction with the marginal distribution for the data, the so-called prior predictive distribution.

Definition 2.1.3. The conditional distribution $\Pi_{\vartheta|Y} : \mathscr{G} \times \mathscr{Y} \to [0,1]$ for $\vartheta|Y$ is called the *posterior* distribution.

The definition of the posterior is almost-sure with respect to the marginal *data distribution* P^{Π} (see definition B.4.3 and the concluding remarks of subsection B.4).

Definition 2.1.4. The marginal distribution $P^{\Pi} : \mathscr{B} \to [0,1]$ for Y is called the *prior predictive distribution*. If, in the above, one replaces the prior by the posterior, the resulting distribution for Y is referred to as the *posterior predictive distribution*.

In the subjectivist philosophy, the prior predictive distribution describes a subjectivist's expectations concerning the observation Y based only on model and prior, *i.e.* before involving the data or realizations thereof. Given model and prior, the prior predictive distribution is of mixture form.

Lemma 2.1.5. *The prior predictive P^{Π} can be expressed in terms of the prior and the model distributions as follows,*

$$P^{\Pi}(Y \in B) = \int_{\Theta} P_{\theta}(B)\, d\Pi(\theta), \tag{2.3}$$

for all $B \in \mathscr{B}$.

The probability measure P^{Π} is called "predictive" because given the model distributions and the prior weights we assign them, their weighted average represents our belief regarding the distribution of the observation Y. Such belief, held prior to observation, forms a prediction for the distribution of Y.

The Bayesian symmetry between observation and parameter invites an identity expressing its essence. Bayes's Rule relates model distributions, prior, posterior and prior predictive distribution through $\Pi(\theta \in G|Y \in B)\,\Pi(Y \in B) = \Pi(Y \in B|\theta \in G)\,\Pi(\theta \in G)$, for all $B \in \mathscr{B}$ and $G \in \mathscr{G}$ (see proposition B.4.2). The following theorem restates this fact in terms of the concepts we have introduced above, in a property which is sometimes referred to as a *disintegration* of the joint measure on model times sample space: (2.4) should be viewed as a double-sided version of definition B.4.4.

Theorem 2.1.6. *Posterior, prior predictive, model distributions and prior are related through Bayes's Rule,*

$$\int_B \Pi(G|Y = y)\, dP^{\Pi}(y) = \int_G P_{\theta}(B)\, d\Pi(\theta), \tag{2.4}$$

for all $B \in \mathscr{B}$ and $G \in \mathscr{G}$.

Proof. Equality (2.4) follows since both sides are equal to $\Pi^*(B \times G)$, *c.f.* definition B.4.4.

Note that, given model and prior, property (2.4) characterizes the posterior, up to re-definition on null sets of the prior predictive distribution P^Π. Consequently, we may turn this theorem around and use property (2.4) as the *defining* property of the posterior.

Definition 2.1.7. Given model and prior, any map $\pi : \mathscr{G} \times \mathscr{Y} \to [0,1]$ such that $y \mapsto \pi(G,y)$ is measurable for all $G \in \mathscr{G}$ and such that π satisfies,

$$\int_B \pi(G,y)\, dP^\Pi(y) = \int_G P_\theta(B)\, d\Pi(\theta), \tag{2.5}$$

for all $B \in \mathscr{B}$ and $G \in \mathscr{G}$, is called a *version of the posterior*.

Unfortunately property (2.5) does not imply that π is a *regular* conditional probability, so we are left with an equivalence in which property 2 of definition B.4.5 remains as a condition.

Proposition 2.1.8. *A map $\pi : \mathscr{G} \times \mathscr{Y} \to [0,1]$ is a regular version of the posterior iff $\pi : y \mapsto \pi(G,y)$ is $\mathscr{B}$-measurable for all $G \in \mathscr{G}$, satisfies (2.5), and $G \mapsto \pi(G,y)$ is a (probability) measure on $\mathscr{G}$ for P^Π-almost-all $y \in \mathscr{Y}$.*

Remark 2.1.9. For statistical questions that only involve a finite number of posterior probabilities, regularity of the posterior is not a requirement: if we test for hypotheses $\Theta_0 \subset \Theta$ versus $\Theta_1 = \Theta \setminus \Theta_0$ with posterior odds or Bayes factors (see section 2.4), countable additivity or other measure-like properties are not required, only the posterior probabilities $\Pi(\Theta_0|Y)$ and $\Pi(\Theta_1|Y)$ play a role. Similarly, for the definition of a credible set $D(Y)$ (of level α) only the posterior probability $\Pi(\vartheta \in D(Y)|Y)$ matters (in that it has to be greater than or equal to $1-\alpha$, see section 2.3), without using measure-like properties of the posterior. By contrast, the definition of the posterior predictive distribution (see definition 2.2.2) or the conditional Bayes solution to a decision-theoretic question (as in definition 2.5.15), for example, do refer to the posterior as an almost-surely defined measure and require a regular posterior distribution.

The following expression for the posterior in a dominated model implies regularity: assuming that the model $\mathscr{P}$ is dominated, the posterior can be expressed in terms of model densities. Since most statistical models are defined as families of densities (*e.g.* Lebesgue-densities on $\mathbb{R}$ or $\mathbb{R}^n$) this accessible form of the posterior is used very often in practice and examples.

Theorem 2.1.10. *Assume that the model $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ is dominated by a σ-finite measure μ on $(\mathscr{Y}, \mathscr{B})$ with densities $p_\theta = dP_\theta/d\mu$. Then the posterior can be expressed as,*

$$\Pi(\,\vartheta \in G\,|\,Y\,) = \int_G p_\theta(Y)\, d\Pi(\theta) \;\Big/\; \int_\Theta p_\theta(Y)\, d\Pi(\theta), \tag{2.6}$$

for all $G \in \mathscr{G}$. This version of the posterior is regular.

Proof. Since the model is dominated, the prior predictive distribution has a density with respect to μ, because for every $B \in \mathscr{B}$,

$$P^{\Pi}(B) = \int_{\Theta} P_\theta(B)\, d\Pi(\theta) = \int_{\Theta} \int_B p_\theta(y)\, d\mu(y)\, d\Pi(\theta)$$
$$= \int_B \Big(\int_{\Theta} p_\theta(y)\, d\Pi(\theta) \Big)\, d\mu(y).$$

in accordance with the Fubini and Radon-Nikodym theorems. The prior predictive density $p^{\Pi} : \mathscr{Y} \to \mathbb{R}$ is therefore equal to the denominator on the *r.h.s.* of (2.6). Let $B \in \mathscr{B}$ and $G \in \mathscr{G}$ be given. Substituting (2.6) into the *l.h.s.* of (2.5), we obtain,

$$\int_B \Pi(G|Y=y)\, dP^{\Pi}(y) = \int_B \Big(\int_G p_\theta(Y)\, d\Pi(\theta) \Big/ \int_{\Theta} p_\theta(Y)\, d\Pi(\theta) \Big)\, dP^{\Pi}(y)$$
$$= \int_B \int_G p_\theta(y)\, d\Pi(\theta)\, d\mu(y) = \int_G P_\theta(B)\, d\Pi(\theta).$$

According to theorem 2.1.6, (2.6) is a version of the posterior and property 3 of definition B.4.5 is satisfied. Property 1 of definition B.4.5 follows from Fubini's theorem (which guarantees measurability of the *r.h.s.* of (2.6)). Since $P^{\Pi}(p^{\Pi} > 0) = 1$, the denominator in (2.6) is non-zero P^{Π}-almost-surely and the posterior is well-defined (as a map $\mathscr{G} \to [0,1]$), P^{Π}-almost-surely. In addition, for all y such that $p^{\Pi}(y) > 0$ and any sequence (G_n) of disjoint, $\mathscr{G}$-measurable sets,

$$\Pi\Big(\vartheta \in \bigcup_{n\geq 1} G_n \,\Big|\, Y = y \Big) = (p^{\Pi}(y))^{-1} \int_{\cup_n G_n} p_\theta(y)\, d\Pi(\theta)$$
$$= (p^{\Pi}(y))^{-1} \int \sum_{n\geq 1} 1_{\{\theta \in G_n\}}\, p_\theta(y)\, d\Pi(\theta)$$
$$= \sum_{n\geq 1} (p^{\Pi}(y))^{-1} \int_{G_n} p_\theta(y)\, d\Pi(\theta) = \sum_{n\geq 1} \Pi(\,\vartheta \in G_n \,|\, Y = y\,),$$

by monotone convergence. We have established that on an event of P^{Π}-measure one, this version of the posterior is well-defined and countably additive, so that also property 2 of definition B.4.5 holds. Conclude that (2.6) is a regular version of the posterior.

In the rest of part I and most of part II, we shall hardly concern ourselves with regularity of posteriors: in all parametric and most non-parametric settings explored here and in the literature, the model is dominated or it is a Polish space (see theorem B.4.7), either of which implies existence of regular posteriors. But in part II we shall also encounter topological circumstances (from rather compelling theoretical arguments based on certain weak model topologies) and questions regarding regularity will resurface there.

To demonstrate that it is easy to define a model (with prior) that theorem B.4.7 does *not* cover, consider the following example.

Example 2.1.11. Suppose that the sample space is $\mathbb{R}$ and the model $\mathscr{P}$ consists of all *discrete probability measures*, of the form (see also example B.2.9):

$$P = \sum_{j=1}^{m} p_j \delta_{x_j}, \tag{2.7}$$

for some $m \geq 1$, with $(p_1, \ldots, p_m)$ in the simplex S_m (see example 1.1.13) and $x_1, \ldots, x_m \in \mathbb{R}$. A suitable prior for this model exists (if one is willing to allow $m = \infty$): distributions drawn from a so-called *Dirichlet process prior* (see section 8.2) are of the form (2.7) with probability one. There is no σ-finite dominating measure for this model (not even if we restrict to measures of the form (2.7) with $m = 1$, see exercise 1.6.3) and the model can not be represented by a family of densities, *c.f.* definition 1.1.3. Definition (2.6) cannot be used in this case. We have to resort to definition 2.1.3 in order to make sense of the posterior distribution and existence of a version of the posterior that displays regularity is a concern in this case.

This model $\mathscr{P}$ can also be used as a parametrizing space for a so-called *mixture model* $\mathscr{P}'$ of distributions on $\mathbb{R}$. For a fixed probability distribution F with Lebesgue density $f : \mathbb{R} \to \mathbb{R}$ and any probability distribution P on $\mathbb{R}$, define the *convolution* f_P as follows,

$$f_P(x) = \int f(x-y)\, dP(y),$$

for (Lebesgue-almost-all) $x \in \mathbb{R}$. Note that f_P is a Lebesgue probability density on $\mathbb{R}$ (due to Fubini's theorem), describing the distribution of the random variable $Z = X + Y$, for some (X, Y) that are independent and marginally, $X \sim P$ and $Y \sim F$. If we let P be from the model $\mathscr{P}$ above, convolution defines a map from $\mathscr{P}$ to a new model $\mathscr{P}'$ of densities, $\mathscr{P} \to \mathscr{P}' : P \mapsto f_P$, of the form,

$$f_P(x) = \sum_{j=1}^{m} p_j f(x_j - x),$$

where P is as in (2.7), a convex combination of m *clusters* in which the intra-cluster variability is described by the density f: the model describes observation of a randomly selected cluster location $X = x_j$ with random (*e.g.* noisy) displacement $Y \sim F$.

2.1.2 Bayes's billiard

To many who have been introduced to statistics from the frequentist point of view, treating the parameter θ for the model as a random variable ϑ seems somewhat unnatural because the frequentist role for the parameter is entirely different from that of the data. The following example demonstrates that in certain situations the Bayesian point of view is not unnatural at all.

Example 2.1.12. In the posthumous publication of "An essay towards solving a problem in the doctrine of chances" in 1763 [13], Thomas Bayes included an example of a situation in which the above, subjectivist perspective arises quite naturally. It involves a number of red balls and one white ball placed on a table and has become known in the literature as *Bayes's billiard*.

We consider the following experiment: unseen by the statistician, someone places n red balls and one white ball on a billiard table of length 1. The statistician will be reported the number K of red balls that is closer to the cushion than the white ball (K plays the role of the *data* in this example) and is asked to give a distribution reflecting his beliefs concerning the position of the white ball X (X plays the role of the parameter) based on K. Calling the distance between the white ball and the bottom cushion of the table X and the distances between the red balls and the bottom cushion Y_i, $(i = 1,\ldots,n)$, it is known to the statistician that their joint distribution is:

$$(X;Y_1,\ldots,Y_n) \sim U[0,1]^{n+1}, \tag{2.8}$$

i.e. all balls are placed independently and uniformly. This distribution gives rise both to the *model* (for K) and to the *prior* (for X). Prior knowledge concerning X (*i.e.* without knowing the observed value $K = k$) offers little information: the best that can be said is that $X \sim U[0,1]$, *i.e.* the prior is uniform. The question is how this distribution for X changes when we incorporate the observation $K = k$, that is, when we use the observation to arrive at *posterior* beliefs.

Since for every i, Y_i and X are independent *c.f.* (2.8), we have model distributions that give rise to,

$$P(Y_i \leq X | X = x) = P(Y_i \leq x) = x,$$

for any $x \in [0,1]$. So for each of the red balls, determining whether it lies closer to the cushion than the white ball amounts to a Bernoulli experiment with parameter x. Since in addition the positions $Y_1,\ldots,Y_n$ are independent, counting the number K of red balls closer to the cushion than the white ball amounts to counting "successes" in a sequence of independent Bernoulli experiments. We conclude that K has a binomial distribution $\mathrm{Bin}(n;x)$, *i.e.*,

$$P(K = k | X = x) = \frac{n!}{k!(n-k)!} x^k (1-x)^{n-k}.$$

It is possible to obtain the density for the distribution of X conditional on $K = k$ from the above display using Bayes's Rule:

$$p(x|K = k) = P(K = k|X = x)\frac{p(x)}{P(K = k)}, \tag{2.9}$$

but in order to use it, we need the two marginal densities $p(x)$ (the prior density) and $P(K = k)$ (the prior predictive density) in the fraction. From (2.8) it is known that $p(x) = 1$ and $P(K = k)$ can be obtained by integrating,

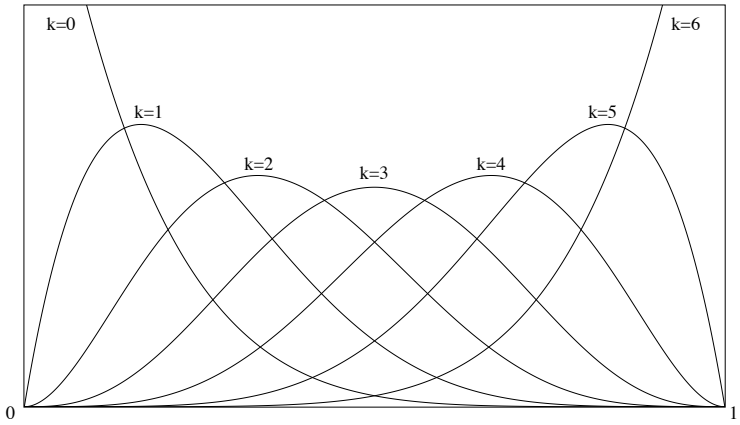


Fig. 2.1 Posterior densities for the position X of the white ball, given the number k of red balls closer to the cushion of the billiard (out of a total of $n = 6$ red balls). For the lower values of k, the white ball is close to the cushion with high probability, since otherwise more red balls would probably lie closer to the cushion. This is reflected by the posterior density for $X|K = 1$, for example, by the fact that it concentrates much of its mass close to $x = 0$.

$$P(K = k) = \int_0^1 P(K = k|X = x)\, p(x)\, dx.$$

Substituting in (2.9), we find:

$$p(x|K = k) = \frac{P(K = k|X = x)\, p(x)}{\int_0^1 P(K = k|X = x)\, p(x)\, dx} = B(n,k)\, x^k (1-x)^{n-k},$$

where $B(n,k)$ is a normalization factor. The x-dependence of the density in the above display reveals that $X|K = k$ is distributed according to a Beta-distribution, $B(k+1, n-k+1)$, so that the normalization factor $B(n,k)$ must equal $B(n,k) = \Gamma(n+2)/\Gamma(k+1)\Gamma(n-k+1)$.

This provides the statistician with distributions reflecting his beliefs concerning the position of the white ball for all possible values k for the observation K. Through conditioning on $K = k$, the prior distribution of X is changed into the posterior for X: if a relatively small number of red balls is closer to the cushion than the white ball (*i.e.* in case k is small compared to n), then the white ball is probably close to the cushion; if k is relatively large, the white ball is probably far from the cushion (see figure 2.1). The illustration on the cover of this book appears in [13] and is Bayes's own version of his Billiard, complete with Beta-density drawn along the bottom edge.

2.1.3 The Bayesian view of the model

Based on the definitions of subsection 2.1.1 a remark is in order with regard to the notion of the *model* in Bayesian statistics: if, for a subset $\mathscr{P}_1 \subset \mathscr{P}$, the prior assigns mass zero, then for all practical purposes $\mathscr{P}_1$ does not play a role since omission of $\mathscr{P}_1$ from $\mathscr{P}$ does not influence the posterior. As long as the model is parametric, *i.e.* $\Theta \subset \mathbb{R}^d$, we can always use priors that dominate the Lebesgue measure, ensuring that any $\mathscr{P}_1$ of prior measure zero has Lebesgue measure zero in Θ and can therefore be thought of as negligibly small. However, in non-parametric models null-sets of the prior and posterior may be much larger than expected intuitively.

Example 2.1.13. Taking the above argument to the extreme, consider a normal location model $\mathscr{P} = \{N(\theta,1) : \theta \in \mathbb{R}\}$ with a prior $\Pi = \delta_{\theta_1}$ (see example B.2.9), for some $\theta_1 \in \mathbb{R}$, defined on the Borel σ-algebra $\mathscr{B}$. Then the model is dominated by the Lebesgue measure and the posterior takes the form:

$$\Pi(\vartheta \in B|Y) = \int_B p_\theta(Y)\,d\Pi(\theta) \Big/ \int_\Theta p_\theta(Y)\,d\Pi(\theta) = \frac{p_{\theta_1}(Y)}{p_{\theta_1}(Y)}\Pi(B) = \Pi(B).$$

for any $B \in \mathscr{B}$. In other words, the posterior *equals* the prior, concentrating all its mass in the point θ_1. Even though we started out with a model that suggests estimation of location, effectively the model consists of only one point due to the degeneracy of the prior. In subjectivist terms, the prior belief is fully biased towards θ_1, leaving no room for amendment by the data when we condition to obtain the posterior.

This example raises the question which part of the model proper $\mathscr{P}$ plays a role in the Bayesian approach. From a topological perspective it is helpful to make the following definition.

Definition 2.1.14. In addition to $(\Theta,\mathscr{G},\Pi)$ being a probability space, let $(\Theta,\mathscr{T})$ be a topological space and assume that $\mathscr{G}$ contains the Borel σ-algebra $\mathscr{B}$ corresponding to the topology $\mathscr{T}$. The *support*support!of a measure $\operatorname{supp}(\Pi)$ of the prior Π is defined as the smallest closed set S such that $\Pi(\Theta \setminus S) = 0$.

It is tempting to equate the support of a prior to the set described by the following intersection.

$$S = \bigcap \big\{G \in \mathscr{G} : G \text{ closed},\ \Pi(\Theta \setminus G) = 0\big\}. \tag{2.10}$$

Perhaps somewhat surprisingly, the validity of this identification is hard to establish: for any $(\Theta,\mathscr{G})$ as in definition C.1.18, S is measurable, in fact, S is (an intersection of closed sets so S is) closed. Since the Borel σ-algebra is generated by the open sets, $S \in \mathscr{B} \subset \mathscr{G}$. To show that $\Pi(\Theta \setminus S) = 0$, poses extra conditions on the space Θ; the following lemma covers a large (but not exhaustive) class of models.

Proposition 2.1.15. *In addition to $(\Theta,\mathscr{G},\Pi)$ being a probability space, let $(\Theta,\mathscr{T})$ be a topological space and assume that $\mathscr{G}$ contains the Borel σ-algebra $\mathscr{B}$ corresponding to the topology $\mathscr{T}$. If $\mathscr{T}$ is second countable, then $S = \operatorname{supp}(\Pi)$.*

Proof. Consider the complement $V = \Theta \setminus S$. We can write,

$$V = \bigcup \{U \in \mathscr{G} : U \text{ open}, \Pi(U) = 0\}. \tag{2.11}$$

The set V is open and contains every open subset of Π-measure zero. Because the topology is second countable, V can be written as a *countable* union of open sets $\{U_k : k \geq 1\}$ of Π-measure zero. Therefore, $\Pi(V) = \Pi(\cup_{k\geq 1} U_k) \leq \sum_{k\geq 1} \Pi(U_k) = 0$ and we conclude that $\Pi(S) = 1$.

This implies, for example, that the support of Π is of the form (2.10) if $(\Theta, \mathscr{T})$ is a separable metrizable space. However not all parameter spaces have Polish complements and a suitable generalization exists [227]: the *Radon property* of a Borel measure (see definition C.8.1) is enough to fix the support problem in a very direct way, for parameter spaces that only have to be Hausdorff.

Proposition 2.1.16. *In addition to $(\Theta, \mathscr{G}, \Pi)$ being a probability space, let $(\Theta, \mathscr{T})$ be a Hausdorff topological space and assume that $\mathscr{G}$ is the Borel σ-algebra $\mathscr{B}$ corresponding to the topology $\mathscr{T}$. If Π is a Radon measure, then $S = \text{supp}(\Pi)$.*

Proof. Let $\mathscr{V}$ denote the collection of all open U in Θ with $\Pi(U) = 0$ and label with a set I: $\mathscr{V} = \{U_\alpha : \alpha \in I\}$. Then the set V defined in (2.11) can be written as $V = \cap\{U_\alpha : \alpha \in I\}$. For every compact $K \subset V$, there exists a *finite* subset $J \subset I$ such that $K \subset \cup\{U_\beta : \beta \in J\}$. It follows that $\Pi(K) \leq \sum_{\beta\in J} \Pi(U_\beta) = 0$ and by inner regularity,

$$\Pi(V) = \sup\{\Pi(K) : K \text{ compact}, K \subset V\} = 0.$$

Example 2.1.17. In example 2.1.13, the model $\mathscr{P}$ consists of all normal distributions of the form $N(\theta, 1)$, $\theta \in \mathbb{R}$, but the support of the prior $\text{supp}(\Pi)$ equals the singleton $\{N(\theta_1, 1)\} \subset \mathscr{P}$.

A well-defined support gives a topological answer to the question where in the parameter space prior mass is concentrated. This suggests that we can distribute mass throughout the parameter space in an equitable fashion: for every $\theta \in \text{supp}(\Pi)$ and every open neighbourhood U of θ, $\Pi(\theta) > 0$. The suggestion extends to the hope that, if we choose a prior Π of full support, $\text{supp}(\Pi) = \Theta$, somehow every point in the model is involved in the subsequent Bayesian analysis.

But of course, we know that the model may be redefined up to null-sets of Π, without influencing the posterior, that is, with equivalence of subsequent Bayesian analyses (*c.f.* the Π-almost-sure nature of the identification (2.2)). When we think of this issue in a model with a countable, discrete parameter space, it does not lead to any ambiguity: the only null-set of a prior with full support is the empty set, in that case. In the setting of a parametric $\Theta \subset \mathbb{R}^d$ with a prior that is absolutely continuous with respect to Lebesgue measure, full support implies ambiguity only on Lebesgue null-sets, which we can still think of as negligible in an intuitively acceptable way. In both cases, the support of the prior is a reasonable substitute for the vague, prior-almost-sure notion of a Bayesian model introduced after definition 2.1.1.

When the model is non-parametric, the support of the prior can become a very misleading intuition regarding the model subset on which the prior is concentrated:

in the following example, a prior Π of full support is constructed on the space of all probability measures on a countably infinite sample space $\mathscr{X}$. Π has null-sets, however, that are very large (*e.g.* containing all p that assign mass to more than a finite number of points in $\mathscr{X}$).

Example 2.1.18. Consider again the full model on a countable sample space $\mathscr{X} = \{i : i \geq 1\}$, as in example 1.1.4, and represent it as the ℓ_1-subset (see definition 1.2),

$$S_\infty = \{p \in \ell_1 : p_i \geq 0, \Sigma_{i\geq 1} p_i = 1\}.$$

Also define the subsets $S_{\infty,k} \subset S_\infty$, $(k \geq 1)$, $S_{\infty,k} = \{p \in S_\infty : p_i = 0, i \geq k\}$, $S_{\infty,0} = \cup\{S_{\infty,k} : k \geq 1\}$ (all p with finite support) and $\mathscr{N} = \{p \in S_\infty : p_i > 0, i \geq 1\}$ (all p with full support). Note that $\mathscr{N}$ can be thought of as describing a generic point in S_∞ (and this is made rigorous when one remarks that $\mathscr{N}$ is *residual* in S_∞ in the Baire sense, see after definition C.4.4). For all $k \geq 1$, place priors Π_k of full support on the finite-dimensional simplices (see (1.4)) that the $S_{\infty,k}$ describe (and embed in S_∞). Define a prior Π on S_∞ based on a sequence (λ_k) such that $\lambda_k > 0$ for all $k \geq 1$ and $\sum_k \lambda_k = 1$,

$$\Pi(A) = \sum_{k\geq 1} \lambda_k \Pi_k(A).$$

It is noted that the normed space ℓ_1 is separable, so the support S of Π is well-defined and coincides with (2.10). To find S, let $\varepsilon > 0$ and $p \in S_\infty$ be given. There exists a $k \geq 1$ such that $\sum_{i>k} p_i < \frac{1}{2}\varepsilon$. Therefore, there exists a $q \in S_{\infty,k}$ and an ℓ_1-neighbourhood U of q in $S_{\infty,k}$ such that for all $q' \in U$, $\|p - q'\| < \varepsilon$. Therefore,

$$\Pi\big(\{q \in S_\infty : \|p - q\| < \varepsilon\}\big) \geq \lambda_k \Pi_k(U) > 0.$$

Conclude that $S = S_\infty$, that is, Π is of full support. Nevertheless, $\mathscr{N} \cap S_{\infty,k} = \varnothing$ for all $k \geq 1$, so $\Pi(\mathscr{N}) = 0$. The set $\mathscr{N}$ is a null-set of Π: any Bayesian analysis with this prior involves support S_∞ but $S_\infty \setminus \mathscr{N}$ is equally deserving to be called 'the model' from the Bayesian perspective.

The prior-almost-sure nature of the Bayesian perspective on statistical models is problematic from a frequentist point of view: if a frequentist makes room for all of S_∞ as a model for the observation Y (example 2.1.18), then the assumption $Y \sim P_0 \in S_\infty$ is meant to include more than just the the p with finite support, the $p \in \mathscr{N}$ are part of the assumption too. Because the posterior is ultimately a measure-theoretic rather than a topological definition, the conceptual confusion about where a prior places its mass, lies at the heart of many (all?) examples of inconsistency of Bayesian methods in non-parametric models (see *e.g.* [98]).

2.1.4 A frequentist's view of the posterior

So far, we have not discussed the details of the data Y, we have treated Y completely abstractly. In this section we consider, firstly, the relation between the frequentist

distribution of Y (the "true" P_0) and the Bayesian distribution of Y (the marginal P^{Π}). Secondly, we consider samples of independent, repeated measurements of a random variable X. We shall see that the Bayesian way to describe data and statistical experiments is in contradiction with the frequentist assumption.

The derivation of the posterior in subsection 2.1.1 does not refer to any *"true, underlying distribution of the data"* but it does involve a marginal distribution for Y, the prior predictive distribution of definition 2.1.4. If one adopts the frequentist framework to analyze Bayesian tools like the posterior, a discrepancy arises since P_0 and P^{Π} are two distributions for the data Y that are not equal (for a striking instance of the discrepancy, see remark 2.1.19 below). To the frequentist, P^{Π} is a side-product of the Bayesian construction that has no realistic interpretation. There is, however, a clear technical issue: all definitions and derivations in subsection 2.1.1 are almost-sure with respect to the prior predictive distribution. To ensure that all of this continues to make sense after we adopt assumption (1.1) we require that P^{Π} dominates P_0 [154]:

$$P_0 \ll P^{\Pi}. \tag{2.12}$$

In that case, null-sets of P^{Π} are also null-sets of P_0, so that all P^{Π}-almost-sure statements and definitions are also P_0-almost-sure. In particular, expression (2.6) for the posterior in a dominated model satisfies the regularity condition not only P^{Π}- but also P_0-almost-surely, if we assume (2.12). We shall adopt the frequentist philosophy to analyse Bayesian tools, *i.e.* we assume (1.1) and (2.12).

In many experiments or observations, the data consists of a sample of n repeated, stochastically independent measurements of the same quantity (an *i.i.d.* sample). To accommodate this and other situations where the data is gathered sequentially, we assume that we observe data X^n taking values in measurable spaces $(\mathscr{X}_n, \mathscr{B}_n)$ for all $n \geq 1$, and we consider parametrized models $\Theta \to \mathscr{P}_n : \theta \mapsto P_{\theta,n}$. The frequentist assumes that there is some sequence of probability measures $(P_{0,n})$ such that $X^n \sim P_{0,n}$ for all $n \geq 1$, and often, that there exists a $\theta_0 \in \Theta$, such that, $P_{0,n} = P_{\theta_0,n}$ for all $n \geq 1$. In the case of *i.i.d.* data from a measurable space $(\mathscr{X}, \mathscr{B})$, $X^n = (X_1, \ldots, X_n) \in \mathscr{X}^n$ with Θ some collection $\mathscr{P}$ of probability measures P on $(\mathscr{X}, \mathscr{B})$ and parametrization $\mathscr{P} \to \mathscr{P}_n : P \mapsto P^n$. Assuming a well-specified model $\mathscr{P}$ implies the existence of some $P_0 \in \mathscr{P}$ such that $P_{0,n} = P_0^n$ for all $n \geq 1$.

For Bayesians $(\Theta, \mathscr{G}, \Pi)$ is a measurable space and $\Theta \to \mathscr{P}_n : \theta \mapsto P_{\theta,n}(B)$ must be measurable for all $B \in \mathscr{B}_n$. But because Bayesians do not entertain the concept of a 'true' distribution of the data, they express assumptions concerning the data *only* through model distributions. Particularly for the *i.i.d.* assumption, the Bayesian assumes *conditional independence* of the observations, given $\vartheta = \theta$:

$$\Pi_{X^n|\vartheta}(X_1 \in A_1, \ldots, X_n \in A_n \,|\, \vartheta = \theta) = \prod_{i=1}^{n} \Pi_{Y|\vartheta}(X_i \in A_i \,|\, \vartheta = \theta) = \prod_{i=1}^{n} P_\theta(A_i),$$

for all $(A_1, \ldots, A_n) \in \mathscr{A}^n$ and Π-almost all θ. Similarly we see that the prior predictive distribution for *i.i.d.* data takes the form:

$$P_n^{\Pi}(X_1 \in A_1, \ldots, X_n \in A_n) = \int_{\Theta} \prod_{i=1}^{n} P_\theta(A_i)\, d\Pi(\theta).$$

The posterior is now a solution for Bayes's Rule in the following form,

$$\int_A \Pi(B|X_1 = x_1, \ldots, X_n = x_n)\, dP_n^{\Pi}(x_1, \ldots, x_n) = \int_B \prod_{i=1}^{n} P_\theta(A_i)\, d\Pi(\theta),$$

where $A = A_1 \times \ldots \times A_n$, $B \in \mathscr{G}$. Assuming that the model $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ for the marginal distributions is dominated by a σ-finite measure μ on $\mathscr{X}$, the above can also be expressed in terms of μ-densities $p_\theta = dP_\theta/d\mu$. Using theorem 2.1.10 we obtain the following expression for the posterior distribution:

$$\Pi(\vartheta \in B|X_1, X_2, \ldots, X_n) = \int_B \prod_{i=1}^{n} p_\theta(X_i)\, d\Pi(\theta) \Big/ \int_\Theta \prod_{i=1}^{n} p_\theta(X_i)\, d\Pi(\theta), \quad (2.13)$$

for any $B \in \mathscr{G}$. Since $P_0(p_0(X) > 0) = 1$, the assumption that $(X_1, \ldots, X_n) \sim P_0^n$ allows us to rewrite this expression with likelihood ratios,

$$\Pi(\vartheta \in B|X_1, X_2, \ldots, X_n) = \int_B \prod_{i=1}^{n} \frac{p_\theta}{p_0}(X_i)\, d\Pi(\theta) \Big/ \int_\Theta \prod_{i=1}^{n} \frac{p_\theta}{p_0}(X_i)\, d\Pi(\theta), \quad (2.14)$$

P_0^n-almost-surely. In a dominated model, the Radon-Nikodym derivative (see theorem B.3.10) of the posterior with respect to the prior is the likelihood function, normalized to be a probability density function:

$$\frac{d\Pi(\cdot|X_1, \ldots, X_n)}{d\Pi}(\theta) = \prod_{i=1}^{n} p_\theta(X_i) \Big/ \int_\Theta \prod_{i=1}^{n} p_\theta(X_i)\, d\Pi(\theta), \quad (2.15)$$

P_n^{Π}-almost-surely, and under (2.12), also P_0^n-almost-surely. The latter fact explains why such strong relations exist between Bayesian and maximum-likelihood methods. Indeed, the proportionality of the posterior density and the likelihood provides a useful qualitative picture of the posterior as a measure that concentrates on regions in the model where the likelihood is relatively high. This may serve as a direct, Fisherian motivation for the use of Bayesian methods in a frequentist context, *c.f.* section 1.4.

Remark 2.1.19. Note that the prior predictive distribution for *i.i.d.* data is itself *not a product distribution* but a mixture of product distributions! This illustrates the discrepancy between P_0 and P^{Π} quite clearly: while the true distribution of the data describes an *i.i.d.* random vector, the prior predictive distribution describes a random vector that is just *exchangeable* (in accordance with De Finetti's theorem (see theorem A.2.2)).

Remark 2.1.20. For the frequentist to use Bayesian tools, *e.g.* a posterior calculated using (2.13), he has to assume condition (2.12). In the context of *i.i.d.* samples, that requirement takes the form,

$$P_0^n \ll P_n^\Pi, \text{ (for all } n \geq 1).$$

2.1.5 From prior to posterior

To conclude the section, consider the following *recipe* for the Bayesian analysis of a data set, illustrated with a very simple parametric example.

(i) Based on the background of the data Y, the statistician chooses a model $\mathscr{P}$ of "reasonable" candidate distributions for the data (usually with some parametrization $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$).
(ii) A prior measure Π on $\mathscr{P}$ is chosen, reflecting "belief" concerning these candidates, see chapter 3, (usually as a probability measure on Θ).
(iii) Based on definition 2.1.3, on expression (2.6) or in the case of an *i.i.d.* sample, on (2.13), we calculate the posterior as a function of the data Y.
(iv) We observe a realization of the data $Y = y$ and use it to calculate a realisation of the posterior.

The statistician may then infer properties of the parameter θ from the posterior $\Pi(\cdot|Y = y)$. One important point: when reporting the results of any statistical procedure, one is obliged to reveal all relevant details concerning the methods followed and the data. So when making inference on θ, the statistician should report on the nature of the sample used and his choice of model, and in the Bayesian case, should always report his choice of prior as well, with a clear motivation.

Example 2.1.21. To illustrate the above "recipe" with a concrete example, consider the one-dimensional parametric model $\mathscr{P}$ consisting of exponential distributions:

$$\mathscr{P} = \{\text{Exp}(\theta) : \theta \in \Theta = (0,\infty)\}.$$

Lebesgue measure dominates the model and densities take the form $p_\theta(x) = \theta \exp(-\theta x)$, for $x \geq 0$. Assume that the data consists of n observations, (conditionally) independent and identically distributed. As a prior on the model, we take another exponential distribution with density $\pi(\theta) = \exp(-\theta)$ (for $\theta \in \Theta$). The posterior density relative to Lebesgue measure on Θ takes the form,

$$d\Pi(\theta|X_1,\ldots,X_n) = C(X_1,\ldots,X_n)\Big(\prod_{i=1}^n \theta\, e^{-\theta X_i}\, 1_{\{X_i \geq 0\}}\Big) e^{-\theta}\, d\theta$$

where $C(X_1,\ldots,X_n)$ denotes the (data-dependent) normalization factor that makes the posterior a probability measure. We calculate,

$$d\Pi(\theta|X_1,\ldots,X_n) = C(X_1,\ldots,X_n)\, \theta^n\, e^{-\theta(1+\Sigma_i X_i)}\, 1_{\{X_{(1)} \geq 0\}}\, d\theta$$

(where $X_{(1)} = \min_i X_i$). Since,

$$\int_0^\infty \theta^n e^{-\alpha\theta}\, d\theta = \frac{n!}{\alpha^{n+1}},$$

we see that $C(X_1,\ldots,X_n)$ must be equal to $(1+\sum_i X_i)^{n+1}/n!$. So for any measurable $A \subset \Theta$, the posterior probability is given by:

$$\Pi(\vartheta \in A | X_1,\ldots,X_n) = \frac{1}{n!}\Big(1+\sum_{i=1}^n X_i\Big)^{n+1} 1_{\{X_{(1)}\geq 0\}} \int_A \theta^n e^{-\theta(1+\sum_{i=1}^n X_i)}\, d\theta.$$

Note that the posterior density collapses to zero (and no longer describes a probability distribution!) if $X_i < 0$ for some $1 \leq i \leq n$. As Bayesians, we insist that the data must be compatible with the model, we require that $\Pi^*(X_i \geq 0) = P_n^\Pi(X_i \geq 0) = 1$. As frequentists we involve the underlying distribution P_0, requiring that $P_0(X \geq 0) = 1$ so that the posterior is well-defined P_0^n-almost-surely. More generally, P^Π dominates Lebesgue measure, so $P_0 \ll P^\Pi$ as long as P_0 has a density with respect to Lebesgue measure.

2.2 Bayesian point estimators

When considering questions of statistical estimation, the outcome of a frequentist procedure is of a different nature than the outcome of a Bayesian procedure: a point-estimator (the frequentist outcome) gives a *point in the model* whereas the posterior (the Bayesian outcome) is a *distribution on the model*. A first question, then, concerns the manner in which to compare the two. We assume the frequentist philosophy to analyse Bayesian methodology (*c.f.* subsection 2.1.4) and, in this section, we derive point-estimators from the posterior distribution in various ways: we consider the posterior predictive distribution, as well as the parametric posterior mean and the maximum-a-posteriori estimator. In later sections we approach the matter from the opposite perspective: every point-estimator has a *sampling distribution*, which can be compared with the posterior because both are distributions on the model or the parameter space. This is the view that gives rise to the Bernstein-von Mises theorem of chapter 4.

2.2.1 Posterior predictive distribution

We think of a Bayesian point-estimator as a point in the model around which posterior mass is accumulated most, a point around which the posterior distribution is concentrated in some way. As such, any reasonable Bayesian point-estimator should represent the *"location"* of the posterior distribution. However there is no unique definition for the "location" of a distribution and, accordingly, there are many different ways to define Bayesian point-estimators.

Remark 2.2.1. Arguably, there are distributions for which even the *existence* of a "location" is questionable. For instance, consider the convex combination of point-masses $P = \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_{+1}$ on $(\mathbb{R}, \mathscr{B})$. Reasonable definitions of location, like the mean and the median of P, all assign as the location of P the point $0 \in \mathbb{R}$. Yet small neighbourhoods of 0 do not receive *any* P-mass, so 0 can hardly be viewed as a point around which P concentrates its mass. The problem is not of a mathematical nature, it is conceptual: when we think of the "location" of a distribution we normally think of *unimodal distributions* which have unambiguous "locations". However, it is common practice to formulate the notion for all distributions by the same definitions.

One quantity that is often used to represent a distribution's location is its expectation. This motivates the first and most Bayesian definition of a posterior-based point-estimator: the posterior predictive distribution.

Definition 2.2.2. Consider a statistical problem involving data Y taking values in a sample space $(\mathscr{Y}, \mathscr{B})$ and a model $(\mathscr{P}, \mathscr{G})$ with prior Π. Assume that all the maps $\mathscr{P} \to [0,1] : P \mapsto P(B)$, $(B \in \mathscr{B})$ are measurable with respect to $\mathscr{G}$ and that the posterior $\Pi(\cdot\,|Y)$ is a regular conditional distribution. The *posterior predictive distribution*) is a data-dependent set-function $\hat{P} : \mathscr{B} \times \mathscr{Y} \to [0,1]$, defined by,

$$\hat{P}(B, y) = \int_{\mathscr{P}} P(B)\, d\Pi(P\,|\,Y = y), \tag{2.16}$$

for every event $B \in \mathscr{B}$, almost surely. (Notation: usually we suppress y-dependence and write $\hat{P} : \mathscr{B} \to [0,1] : B \mapsto \hat{P}(B)$.)

Remark 2.2.3. The qualification "almost surely" in the formulation of proposition 2.2.2 has distinct explanations for Bayesians and for frequentists: for the Bayesian, the data Y is marginally distributed according to the prior predictive distribution, so it is with respect to null sets of P^{Π} that "almost surely" is to be interpreted in that case. By contrast, the frequentist assumes that $Y \sim P_0$, so he is forced to adopt assumption (2.12) and the interpretation of "almost surely" refers to null sets of P_0 in that case. This dual use of the phrase "almost surely" re-occurs in many places below.

This probability measure $\hat{P}$ is called "predictive" because of the following, Bayesian interpretation: with a prior Π on a model $\mathscr{P}$ and an observation $Y = y$ of the data, we calculate a posterior $\Pi(\cdot|Y = y)$. If we were to conduct the same experiment again with new (independent) observation Y', we would use the posterior $\Pi(\cdot|Y = y)$ as our new prior and we would predict the distribution of Y' to be $\hat{P}$. This is clearly different from the frequentist interpretation, in which $\hat{P}$ is an estimator for P_0.

Proposition 2.2.4. *The posterior predictive distribution $\hat{P} : \mathscr{B} \to [0,1]$ is a probability measure, almost surely.*

Proof. Since we are assuming that the posterior is a regular conditional distribution, $\hat{P}$ is defined almost-surely as a map $\mathscr{B} \to [0,1]$. Let $F \in \mathscr{B}$ denote the event that $\hat{P}$ is well-defined and let $y \in F$ be given. Clearly, for all $B \in \mathscr{B}$, $0 \leq \hat{P}(B) \leq 1$.

Let $(B_i)_{i\geq 1} \subset \mathscr{B}$ be any sequence of disjoint events. Since $(P,i) \mapsto P(B_i)$ is non-negative and measurable, Fubini's theorem (or monotone convergence) applies in the third equality below:

$$\begin{aligned}\hat{P}\Big(\bigcup_{i\geq 1} B_i\Big) &= \int_{\mathscr{P}} P\Big(\bigcup_{i\geq 1} B_i\Big)\, d\Pi(P\,|\,Y=y) = \int_{\mathscr{P}} \sum_{i\geq 1} P(B_i)\, d\Pi(P\,|\,Y=y) \\ &= \sum_{i\geq 1} \int_{\mathscr{P}} P(B_i)\, d\Pi(P\,|\,Y=y) = \sum_{i\geq 1} \hat{P}(B_i),\end{aligned}$$

which proves countable additivity of $\hat{P}$ for all $y \in F$, that is, almost-surely.

Although we refer to $\hat{P}$ as a point-estimator (see definition 1.2.1), generically $\hat{P}$ does not lie in $\mathscr{P}$ as the following theorem shows.

Theorem 2.2.5. *On $\mathscr{M}(\mathscr{Y},\mathscr{B})$ with a locally convex topology, let the posterior be a Radon probability measure with support $\mathscr{P}$ that is a bounded subset of $\mathscr{M}_1^+(\mathscr{Y},\mathscr{B})$, almost-surely. Then $\hat{P}$ lies in the closed convex hull of $\mathscr{P}$, almost-surely.*

Proof. By assumption, there is an event $F \in \mathscr{B}$ of P^{Π}-(or P_0-)probability one, such that the posterior $\Pi(\cdot|Y=y)$ is a well-defined Radon measure for every $y \in F$. Fix some $y \in F$: the map $\hat{P}$ of (2.16) is then a probability measure in $\mathscr{M}_1^+(\mathscr{Y},\mathscr{B})$. Let $p_\alpha : \mathscr{M}(\mathscr{Y},\mathscr{B}) \to \mathbb{R}$, $(\alpha \in \mathscr{A})$ denote a family of semi-norms generating the uniformity for the locally convex space $\mathscr{M}(\mathscr{Y},\mathscr{B})$ (*c.f.* [50], Ch. II, § 4, No. 1, Corollary to prop. 1). For any neighbourhood U of $\hat{P}$, there exists an $\alpha \in \mathscr{A}$ and an $\varepsilon > 0$ such that the p_α-neighbourhood $V = \{P \in \mathscr{P} : p_\alpha(P-\hat{P}) < \varepsilon\}$ is contained in U. By assumption $\mathscr{P}$ is bounded, so there exists a constant $s_\alpha > 0$ such that $s_\alpha = \sup_{P\in\mathscr{P}} p_\alpha(P) < \infty$. Choose some $0 < \delta < \frac{1}{6} s_\alpha^{-1} \varepsilon$. Since the posterior for $Y = y$ is a Radon measure, there is a compact $K \subset \mathscr{P}$ such that $\Pi(K|Y=y)) > 1-\delta$. Condition the posterior for $Y = y$ on K and write, for every $B \in \mathscr{B}$,

$$\hat{P}_K(B) = \int P(B)\, d\Pi(P\,|\,K, Y=y) = \frac{1}{\Pi(K\,|\,Y=y)} \int_K P(B)\, d\Pi(P\,|\,Y=y).$$

A proof following that of proposition 2.2.4 shows that the map $\hat{P}_K : \mathscr{B} \to [0,1]$ is a probability measure in $\mathscr{M}_1^+(\mathscr{Y},\mathscr{B})$. By the triangle inequality,

$$\begin{aligned}p_\alpha(\hat{P}-\hat{P}_K) &\leq p_\alpha\Big(\int_{\mathscr{P}\setminus K} P\, d\Pi(P\,|\,Y=y)\Big) + \frac{\delta}{1-\delta} p_\alpha\Big(\int_K P\, d\Pi(P\,|\,Y=y)\Big) \\ &\leq \int_{\mathscr{P}\setminus K} p_\alpha(P)\, d\Pi(P\,|\,Y=y) + \frac{\delta}{1-\delta} \int_K p_\alpha(P)\, d\Pi(P\,|\,Y=y) \\ &\leq 3 s_\alpha \delta < \tfrac{1}{2}\varepsilon.\end{aligned}$$

(Note that the restriction of p_α to $\mathscr{P}$ is bounded and continuous, so approximating measures with *finite* support exist (*c.f.* [48], Ch. III, § 4, No. 4, Theorem 1) and the triangle inequality for p_α gives rise to the Jensen-like second inequality in the above

display.) Since K is compact, there exists an $N \geq 1$ and $P_1, \ldots, P_N$ in K such that their open p_α-neighbourhoods $V_i = \{P \in K : p_\alpha(P - P_i) < \frac{1}{2}\varepsilon\}$ form a finite cover of K. Through definition of $C_{i+1} = V_{i+1} \setminus V_i$ (for all $1 \leq i \leq N$, with $C_1 = V_1$) this cover generates a finite measurable partition $\{C_1, \ldots, C_N\}$ of K. For $1 \leq i \leq N$, define $\lambda_i = \Pi(C_i \,|\, K, Y = y)$ and note that,

$$\begin{aligned} p_\alpha\Big(\hat{P}_K - \sum_{i=1}^N \lambda_i P_i\Big) &= p_\alpha\Big(\sum_{i=1}^N \int_{C_i} (P - P_i)\, d\Pi(P \,|\, K, Y = y)\Big) \\ &\leq \sum_{i=1}^N p_\alpha\Big(\int_{C_i} (P - P_i)\, d\Pi(P \,|\, K, Y = y)\Big) \\ &\leq \sum_{i=1}^N \int_{C_i} p_\alpha(P - P_i)\, d\Pi(P \,|\, K, Y = y) \leq \tfrac{1}{2}\varepsilon. \end{aligned}$$

Consequently,

$$p_\alpha\Big(\hat{P} - \sum_{i=1}^N \lambda_i P_i\Big) \leq p_\alpha(\hat{P} - \hat{P}_K) + p_\alpha\Big(\hat{P}_K - \sum_{i=1}^N \lambda_i P_i\Big) < \varepsilon,$$

We have shown that any neighbourhood U of $\hat{P}$ has non-empty intersection with the convex hull of $\mathscr{P}$, almost surely. Conclude that $\hat{P}$ lies in the closed convex hull of $\mathscr{P}$, almost surely.

Consider this theorem with four locally convex topologies: $\mathscr{T}_C$, $\mathscr{T}_1$, $\mathscr{T}_\infty$ and $\mathscr{T}_{TV}$. In all four cases, the boundedness condition for $\mathscr{P}$ is satisfied trivially (because all (semi-)norms for these locally convex spaces are bounded by total-variation, which is bounded by one on $\mathscr{M}_1^+(\mathscr{Y}, \mathscr{B})$).

Corollary 2.2.6. *If $\mathscr{Y}$ is a Polish space with $\mathscr{B}$ its Borel σ-algebra, and we consider $\mathscr{M}(\mathscr{Y}, \mathscr{B})$ in the $\mathscr{T}_C$ topology, $\hat{P}$ lies in the closed convex hull of the posterior support, almost surely.*

Proof. The space $\mathscr{M}(\mathscr{Y}, \mathscr{B})$ with the $\mathscr{T}_C$ topology is Polish (see proposition C.9.4). Thinking of $\mathscr{M}(\mathscr{Y}, \mathscr{B})$ as a parameter space, conditioning on $Y = y$ gives rise to a posterior that is a Borel measure with a regular version, according to theorem B.4.7. Based on the fact that any Borel measure on a Polish space is Radon, theorem 2.2.5 applies.

In the case of the total-variational topology, the Radon property is more difficult to establish. A general condition is domination of the support of the posterior.

Corollary 2.2.7. *If we consider $\mathscr{M}(\mathscr{Y}, \mathscr{B})$ in the $\mathscr{T}_{TV}$ topology and the support $\mathscr{P}$ of the posterior is dominated, almost surely, then $\hat{P}$ lies in the closed convex hull of the posterior support, almost surely.*

Proof. For an event $F \in \mathscr{B}$ of P^Π-(or P_0-)probability one and any $y \in F$, the posterior $\Pi(\cdot|Y = y)$ is a well-defined Borel probability measure supported on $\mathscr{P}$, with

corresponding $\hat{P}$ in $\mathscr{M}_1^+(\mathscr{Y},\mathscr{B})$. The dominated subset $\mathscr{P}$ is separable in the (metric) $\mathscr{T}_{TV}$ topology and so is the completion L of its linear span (viewed as a subspace of $\mathscr{M}(\mathscr{Y},\mathscr{B})$). So L is Polish and $\Pi(\cdot|Y=y)$ can also be viewed as a Radon measure on L, supported on $\mathscr{P}\subset L$. The proof of theorem 2.2.5 continues with L in the role of $\mathscr{M}(\mathscr{Y},\mathscr{B})$.

In the topologies $\mathscr{T}_1$ and $\mathscr{T}_\infty$, the Radon property is even harder to establish, because metrizability is no longer guaranteed.

Corollary 2.2.8. *If we consider $\mathscr{M}(\mathscr{Y},\mathscr{B})$ in the $\mathscr{T}_1$, or in the $\mathscr{T}_\infty$ topology and the support $\mathscr{P}$ of the posterior is dominated and metrizable, almost surely, then $\hat{P}$ lies in the closed convex hull of the posterior support, almost surely.*

Proof. The proof is almost identical to that of corollary 2.2.7: here, the subset $\mathscr{P}$ is separable because it is dominated and metrizable by assumption, so the completion L of its linear span is Polish.

The extra metrizability condition is by-passed in the case of the inverse limit distributions of appendix D, which have the Radon property by construction, and examples of Pólya tree distributions that are $\mathscr{T}_1$-Radon are given there.

2.2.2 Posterior mean

In many practical situations the model $\mathscr{P}$ is parametric and a different form of "averaging over the model" applies.

Definition 2.2.9. Let the model $\mathscr{P}$ have a parametrization $\Theta\to\mathscr{P}:\theta\mapsto P_\theta$, where Θ is a closed, convex subset of $\mathbb{R}^d$. Let Π be a Borel prior on Θ with posterior $\Pi(\cdot\,|Y)$. If ϑ is integrable with respect to the posterior almost-surely, then the *posterior mean* is defined

$$\hat{\theta}_1(Y)=\int_\Theta \theta\, d\Pi(\,\theta\,|\,Y\,)\in\Theta, \tag{2.17}$$

almost-surely.

In definition 2.2.9 closed-convexity of Θ is a condition, otherwise there is no guarantee that $\hat{\theta}_1(Y)\in\Theta$ (meaningless because it would leave $P_{\hat{\theta}_1}$ undefined).

Example 2.2.10. In example 2.1.21 the posterior takes the form:

$$\Pi(\vartheta\in A|X_1,\ldots,X_n)=\frac{1}{n!}\Big(1+\sum_{i=1}^n X_i\Big)^{n+1}1_{\{X_{(1)}\geq 0\}}\int_A \theta^n e^{-\theta(1+\Sigma_{i=1}^n X_i)}\,d\theta.$$

Assuming that $P_0(X\geq 0)=1$, we omit the indicator for $X_{(1)}\geq 0$ and write the posterior mean of definition 2.2.9 as follows:

$$\begin{aligned}\hat{\theta}_1(Y) &= \int_\Theta \theta \, d\Pi(\,\theta\,|\,Y\,) = \frac{1}{n!}\Big(1+\sum_{i=1}^n X_i\Big)^{n+1} \int_0^\infty \theta^{n+1} e^{-\theta(1+\sum_{i=1}^n X_i)}\, d\theta \\ &= \frac{1}{n!}\Big(1+\sum_{i=1}^n X_i\Big)^{-1} \int_0^\infty \zeta^{n+1} e^{-\zeta}\, d\zeta = (n+1)\Big(1+\sum_{i=1}^n X_i\Big)^{-1},\end{aligned} \tag{2.18}$$

where we have used that $\int_0^\infty \zeta^{n+1} e^{-\zeta}\, d\zeta = \Gamma(n+2) = (n+1)!$.

From a frequentist perspective, it is worth noting the import of the *factorization theorem*, which says that the parameter-dependent factor in the likelihood is a function of the data only through a so-called sufficient statistic $T(Y)$ (a statistic is *sufficient* if the conditional distribution $Y|T$ does not depend on the parameter).

Theorem 2.2.11. *Let $\mathscr{P}$ be a dominated parametric model with parametrization $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$ and densities $p_\theta : \mathscr{Y} \to [0,\infty)$. Then the parameter-dependence of the likelihood $\Theta \to [0,\infty) : \theta \mapsto p_\theta(Y)$ is expressed in terms of a sufficient statistic: there exist functions $g : \Theta \times \mathbb{R}^k \to [0,\infty)$, $h : \mathscr{Y} \to [0,\infty)$ and a sufficient statistic $T : \mathscr{Y} \to \mathbb{R}^k$ such that for all $\theta \in \Theta$,*

$$p_\theta(Y) = g\big(\theta, T(Y)\big)h(Y),$$

almost-surely.

First note that, based on the factorization theorem, the most practical way to obtain a sufficient statistic (in a dominated model) is a close look at the parameter-dependence of the likelihood function. Second note that, based on theorem 2.1.10, the posterior is a function of the data only through the likelihood, and $h(Y)$ cancels as a factor in both numerator and denominator. Therefore the posterior is a function of the data Y only through a sufficient statistic $T(Y)$. Sufficient statistics often (but not always) also satisfies the following property.

Definition 2.2.12. A statistic $T(Y)$ is *complete* for the model $\mathscr{P}$, if, for all measurable real-valued f, the assertion that, for all $\theta \in \Theta$, $P_\theta f(T(Y)) = 0$ implies $P_\theta(f(T(Y)) = 0) = 1$.

Sufficiency and completeness are important for unbiased estimation of parameters with respect to mean-squared error, due to the *optimality theorem of Lehmann-Scheffé* (For a proof, see Lehmann and Casella (1998) [169].)

Theorem 2.2.13. *(Lehmann-Scheffé)*
Let $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ be a parametrized model for data Y with sufficient and complete statistic $T(Y)$. Any unbiased, quadratically integrable estimator $\hat{\theta}(Y)$ is a function of Y only through $T(Y)$, if and only if, for any other unbiased, quadratically integrable estimator $\hat{\eta}(Y)$,

$$P_\theta\big(\hat{\theta}(Y) - \theta\big)^2 \le P_\theta\big(\hat{\eta}(Y) - \theta\big)^2,$$

for all $\theta \in \Theta$.

Conclude that unbiased quadratically integrable estimators are optimal in mean square error, if and only if, they depend on the data only through a sufficient and complete statistic. This has the following immediate consequence for estimators derived from a posterior.

Corollary 2.2.14. *If $T(Y)$ is sufficient and complete for $\mathscr{P}$, then any point-estimator $\hat{\theta}$ based on the posterior that is unbiased and quadratically integrable, is optimal in the sense of theorem 2.2.13.*

The direct usefulness of corollary 2.2.14 is limited, because the presence of the prior tends to cause a bias for point-estimators based on posteriors. However, such bias can be controlled or even eliminated if one chooses the prior by methods from what is called *empirical Bayes* (see, *e.g.*, subsection 3.4.2). In that case corollary 2.2.14 applies and directly proves optimality in the sense of theorem 2.2.13. Frequentist performance of Bayesian methods in (smooth) parametric estimation problems is considered again in chapter 4 and reaches the conclusion (theorem 4.2.1) that the posterior generally gives rise to optimal point estimators also in an asymptotic sense (within the much wider class of all *regular estimators*, see definition 4.1.10).

2.2.3 Small-ball and formal Bayes estimators

Since there are multiple ways of defining the location of a distribution, there are multiple ways of obtaining point-estimators from the posterior distribution. A straightforward alternative for the posterior averages of the previous subsection, is given in the following definition which requires that the model is one-dimensional.

Definition 2.2.15. Let Θ be a closed, non-empty subset of $\mathbb{R}$ and let $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ be a parametric model with Borel prior Π on Θ and posterior $\Pi(.|Y)$. The *posterior median* is defined by,

$$\tilde{\theta}(Y) = \inf\{s \in \Theta : \Pi(\vartheta \leq s|Y) \geq 1/2\},$$

almost-surely.

Thus the posterior median represents the smallest value for θ such that the posterior mass to its left is greater than or equal to $1/2$. This definition simplifies drastically in case the posterior has a continuous, (strictly) monotone distribution function: in that case the above definition reduces to the perhaps more familiar definition as the (unique) point $\tilde{\theta} \in \Theta$ where $\Pi(\vartheta \leq \tilde{\theta}\,|\,Y) = 1/2$. In some situations, the posterior median offers an advantage over the posterior mean since its definition does not depend on integrability requirements and because of robustness against perturbation of the tails of the posterior.

Another alternative is decision-theoretic in essence (see section 2.5), that is, one takes the perspective in which an assessment of *loss* is inherent. Suppose that we consider estimation in a metric model $(\mathscr{P}, d)$ and we quantify errors in estimation

as follows: if the true distribution of the data is P_0 and we estimate that it is P, then we incur a loss (to be specified further by the context of the problem) that is a monotone increasing function $\ell : [0,\infty) \mapsto [0,\infty)$ of the distance $d(P_0,P)$. If we assume that the posterior concentrates its mass around P_0 (as well as possible) then the estimator that minimizes the *expected loss* relative to the posterior optimizes the so-called *Bayesian risk function*,

$$r(\Pi,\theta') = \int_\Theta P_\theta \ell(d(\theta,\theta'))\, d\Pi(\theta),$$

for given prior Π and estimator θ'.

Definition 2.2.16. Let $\mathscr{P}$ be a model with metric $d : \mathscr{P} \times \mathscr{P} \to \mathbb{R}$ and a Borel prior Π on $\mathscr{G}$. Any monotone increasing function $\ell : [0,\infty) \mapsto [0,\infty)$ is called a *loss-function* if it is such that $\ell(0) = 0$. (Often ℓ is assumed to be convex or (semi-)continuous.) Provided it exists, the formal Bayes estimator is a minimizer $\tilde{P}$ of the function,

$$\mathscr{P} \to \mathbb{R} : P \mapsto \int_{\mathscr{P}} \ell(d(Q,P))\, d\Pi(Q\,|\,Y),$$

over the model $\mathscr{P}$, defined almost-surely.

Note that definition 2.2.16 retains its form when expressed in terms of a formal Bayes estmator $\tilde{\theta}$ for a parameter $\theta \in \Theta$.

Theorem 2.2.17. *Let the model $\mathscr{P}$ be dominated and parametrized by a metric space (Θ,d). If the formal Bayes estimator $\tilde{\theta}$ is well-defined, it minimizes the Bayesian risk function:*

$$r(\Pi,\tilde{\theta}) = \inf_{\theta' \in \Theta} r(\Pi,\theta').$$

Proof. Rewrite the Bayesian risk function for the formal Bayes estimator:

$$\begin{aligned} r(\Pi,\tilde{\theta}) &= \int_\Theta P_\theta \ell(d(\theta,\tilde{\theta}))\, d\Pi(\theta) = \int_\Theta \int_{\mathscr{Y}} \ell(d(\theta,\tilde{\theta}(y)))\, dP_\theta(y)\, d\Pi(\theta) \\ &= \int_{\mathscr{Y}} \int_\Theta \ell(d(\theta,\tilde{\theta}(y)))\, p_\theta(y)\, d\Pi(\theta)\, d\mu(y) \\ &= \int_{\mathscr{Y}} \Big(\int_\Theta p_\theta(y)\, d\Pi(\theta) \Big) \int_\Theta \ell(d(\theta,\tilde{\theta}(y)))\, d\Pi(\theta|Y=y)\, d\mu(y). \end{aligned}$$

where we use the Radon-Nikodym theorem (see theorem B.3.10), Fubini's theorem (see theorem B.3.9) and the definition of the posterior, *c.f.* (2.13). Using the prior predictive distribution (2.3), we rewrite the Bayesian risk function further:

$$r(\Pi,\tilde{\theta}) = \int_{\mathscr{Y}} \int_\Theta \ell(d(\theta,\tilde{\theta}(y)))\, d\Pi(\theta|Y=y)\, dP^\Pi(y). \tag{2.19}$$

By assumption, the formal Bayes estimator $\tilde{\theta}$ exists. Since $\tilde{\theta}$ satisfies

$$\int_\Theta \ell(d(\theta,\tilde{\theta}(y)))\, d\Pi(\theta\,|\,Y=y) = \inf_{\theta' \in \Theta} \int_\Theta \ell(d(\theta,\theta'))\, d\Pi(\theta\,|\,Y=y)$$

for P^{Π}-almost all $y \in \mathscr{Y}$, we obtain

$$r(\Pi, \tilde{\theta}) = \int_{\mathscr{Y}} \inf_{\theta' \in \Theta} \int_{\Theta} \ell(d(\theta, \theta')) \, d\Pi(\theta | Y = y) \, dP^{\Pi}(y)$$
$$\leq \inf_{\theta \in \Theta} \int_{\mathscr{Y}} \int_{\Theta} \ell(d(\theta, \theta')) \, d\Pi(\theta | Y = y) \, dP^{\Pi}(y) = \inf_{\theta' \in \Theta} r(\Pi, \theta').$$

One estimator of this type is defined in the following intuitively reasonable way.

Definition 2.2.18. Let the data Y with model $\mathscr{P}$, metric d and prior Π be given. Suppose that the σ-algebra on which Π is defined contains the Borel σ-algebra. For given $\varepsilon > 0$, the *small-ball estimator* is defined to be the maximizer $\tilde{P}$ of the function

$$P \mapsto \Pi(B_d(P, \varepsilon) \,|\, Y), \tag{2.20}$$

over the model, where $B_d(P, \varepsilon)$ is the d-ball in $\mathscr{P}$ of radius ε centred on P. Provided that such a maximizer exists and is unique, it is defined almost-surely.

Note that this is simply the formal Bayes estimator for the loss function $\ell(d) = 1\{d \geq \varepsilon\}$. Existence of a small-ball estimator $\tilde{P}$ therefore implies optimality in the sense that,

$$\Pi(d(P, \tilde{P}) \geq \varepsilon \,|\, Y) = \inf_{Q \in \mathscr{P}} \Pi(d(P, Q) \geq \varepsilon \,|\, Y).$$

Remark 2.2.19. Similarly to definition 2.2.18, for a fixed value p such that $1/2 < p < 1$, we may define a Bayesian point estimator as the centre point of the smallest d-ball with posterior mass greater than or equal to p (if it exists and is unique).

Suitable conditions for the existence of small-ball estimators form the subject of exercise 2.6.13.

2.2.4 The maximum-a-posteriori estimator

If the posterior is dominated by a σ-finite measure ν, the posterior density with respect to ν can be used as a basis for defining Bayesian point estimators.

Definition 2.2.20. Let $\mathscr{P}$ be a model with parametrization $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$ and a prior Π on Θ. Assume that the posterior is almost-surely absolutely continuous with respect to a σ-finite measure ν on Θ, with ν-density $\Theta \to [0, \infty) : \theta \mapsto \pi(\theta | Y)$. The *maximum-a-posteriori estimator* (or *MAP-estimator*, or *posterior mode*) $\hat{\theta}_2(Y)$ for θ is defined as a point in the model where the posterior density takes on its maximal value:

$$\pi(\hat{\theta}_2 | Y) = \sup_{\theta \in \Theta} \pi(\theta | Y). \tag{2.21}$$

Provided that such a point exists and is unique almost-surely, the MAP-estimator is defined almost-surely.

The MAP-estimator has a serious weak point: a different choice of dominating measure ν leads to a different MAP estimator! In fact, a change of ν is equivalent to a change of prior distribution (which does not correspond to an explicit, formal change of Bayesian 'belief' because the prior remains unchanged). A MAP-estimator is therefore not fully specified unless we indicate which dominating measure was used to define the posterior density. Often the Lebesgue measure is used without further comment, or objective measures (see section 3.2) are used. Another option is to use the prior measure as the dominating measure, in which case the MAP estimator equals the maximum-likelihood estimator.

Remark 2.2.21. There is an interesting connection between (Bayesian) MAP-estimation and (frequentist) maximum-likelihood estimation. Referring to formula (2.13) we see that in an *i.i.d.* experiment with parametric model, the MAP-estimator maximizes:

$$\Theta \to \mathbb{R} : \theta \mapsto \prod_{i=1}^{n} p_\theta(X_i)\,\pi(\theta),$$

where it is assumed that the model is dominated and that the prior has a density π with respect to the Lebesgue measure ν. If the prior had been uniform, the last factor would have dropped out and maximization of the posterior density *is* maximization of the likelihood. Therefore, differences between ML and MAP estimators are entirely due to non-uniformity of the prior. Subjectivist interpretation aside, prior non-uniformity has an interpretation in the frequentist setting as well, through what is called *penalized maximum likelihood estimation* (see, for example, van de Geer (2000) [103]): Bayes's rule applied to the posterior density $\pi_n(\theta|X_1,\ldots,X_n)$ gives:

$$\log \pi_n(\theta|X_1,\ldots,X_n) = \log \prod_{i=1}^{n} p_\theta(X_i) + \log \pi(\theta) + D(X_1,\ldots,X_n),$$

where D is a (θ-independent) normalization constant. The first term equals the log-likelihood and the logarithm of the prior plays the role of a penalty term when maximizing over θ. Hence, maximizing the posterior density over the model Θ can be identified with maximization of a penalized likelihood over Θ. So defining a penalized MLE $\hat{\theta}_n$ with the logarithm of the prior density $\theta \mapsto \log \pi(\theta)$ in the role of the penalty, the MAP-estimator coincides with $\hat{\theta}_n$. The above offers a direct connection between Bayesian and frequentist methods of point-estimation. As such, it provides a frequentist interpretation of the prior as a penalty in the ML procedure.

All Bayesian point estimators defined above as maximizers or minimizers over the model suffer from the usual existence and uniqueness issues associated with extrema. However, there are straightforward methods to overcome such issues. We illustrate using the MAP-estimator. Questions concerning the existence and uniqueness of MAP-estimators should be compared to those of the existence and uniqueness of *M-estimators* in frequentist statistics. Although it is hard to formulate conditions of a general nature to guarantee that the MAP-estimator exists, often one can use the following lemma to guarantee existence.

Lemma 2.2.22. *Consider a parametrized model $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$; If the parameter space Θ is compact and the posterior density $\theta \mapsto \pi(\theta|Y)$ is upper-semi-continuous, then the MAP-estimator exists almost surely.*

To prove uniqueness one has to be aware of various possible problems among which, for instance, identifiability of the model (see section 1.1, in particular definition 1.1.7).

Example 2.2.23. Assuming that $P_0(X \geq 0) = 1$, the posterior density in example 2.1.21 has the form:

$$\pi(\theta|X_1,\ldots,X_n) = \frac{1}{n!}\Big(1+\sum_{i=1}^{n} X_i\Big)^{n+1}\theta^n\, e^{-\theta(1+\Sigma_{i=1}^n X_i)},$$

P_0^n-almost-surely, where $\theta > 0$. Setting the θ-derivative to zero, we find that the MAP-estimator is given by:

$$\hat{\theta}_2(Y) = n\Big(1+\sum_{i=1}^{n} X_i\Big)^{-1}.$$

The MAP-estimator is similar to the maximum likelihood estimator (equal to $n(\Sigma_i X_i)^{-1}$) and the posterior mean (equal to $(n+1)(1+\Sigma_i X_i)^{-1}$), see (2.18)). Although it is possible technically that these three estimators differ substantially, in many (*e.g.* unimodal) cases the maximum of the posterior density lies in the bulk that determines the posterior mean as well, and MAP and posterior mean are close. If, in addition, the influence of the prior is relatively small because the likelihood function peaks very sharply at its maximum, the maximum-likelihood estimator is expected to be close too. Note that differences between these three estimators become negligible in the limit $n \to \infty$.

2.3 Confidence sets and credible sets

Besides point-estimation, frequentist statistics has several other inferential techniques at its disposal. The two most prominent are the analysis of confidence intervals and the testing of statistical hypotheses. In the next section, we consider frequentist testing of hypotheses, in this section, we discuss frequentist confidence sets and their Bayesian counterparts, called credible sets.

2.3.1 Frequentist confidence sets

Assume that we have a model $\mathscr{P}$ parametrized by an identifiable parameter θ in a parameter set Θ, assuming that the true distribution of the data $Y \sim P_0$ belongs to the model, that is, $P_0 = P_{\theta_0}$ for some $\theta_0 \in \Theta$. The inferential goal is to use the data Y

to define an model subset $C(Y)$ that contains θ_0 with "high" probability. The word "high" requires quantification in terms of a level α, called the confidence level. Let $\mathscr{C}$ denote a class of subsets of Θ (for example, with $\Theta = \mathbb{R}$ we often choose $\mathscr{C}$ equal to the class of all closed intervals in Θ, or if $\Theta = \mathbb{R}^d$ we could take the class of all ellipsoids in Θ).

Definition 2.3.1. Let $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$ be an identifiable parametrization. Choose a *confidence level* $\alpha \in (0,1)$. Let $C_\alpha : \mathscr{Y} \to \mathscr{C}$ describe a data-dependent subset of Θ. Then C_α is a *confidence set* for θ of *confidence level* α, if $\{y \in \mathscr{Y} : \theta \in C_\alpha(y)\}$ is $\mathscr{B}$-measurable for every $\theta \in \Theta$ and C_α solves the equation,

$$P_\theta\big(\theta \in C_\alpha(Y)\big) \geq 1 - \alpha, \tag{2.22}$$

for all $\theta \in \Theta$.

Measurability of the events $\{\theta \in C_\alpha(Y)\}$ is rarely problematic, a technical matter to be addressed at the model-specific level. The data-dependence of $C_\alpha(Y)$ is meant to express the requirement that $C_\alpha(Y)$ is a *statistic* (as defined below definition 1.2.1). Conceptually a confidence set can be compared to a point estimator $\hat{P} : \mathscr{Y} \to \mathscr{P}$: however, rather than focussing on a data-dependent *point* in Θ, a data-dependent *subset* in Θ is to inform us about the parameter.

Clearly confidence sets are not unique and small confidence sets are more informative than large ones. For example, the constant assignment $C_\alpha(y) = \Theta$ for all $y \in \mathscr{Y}$ is a confidence set for any level $\alpha \in (0,1)$, but it does not have any informative value. If, for some confidence level α, we have two different procedures of finding confidence sets, leading to sets C_α and D_α of confidence level α respectively, and $C_\alpha \subset D_\alpha$, P_θ-almost-surely for all θ, then C_α is preferred over D_α.

Example 2.3.2. Let $X^n = (X_1, \ldots, X_n)$ be an *i.i.d.* sample from a normal distribution $P_0 = N(\mu, \sigma^2)$ with known variance $\sigma^2 > 0$ and unknown $\mu \in \mathbb{R}$. As is well-known, the sample average is normally distributed,

$$\frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu_0, \sigma_n^2), \tag{2.23}$$

with a variance $\sigma_n^2 = \sigma^2/n$. If we adopt the sample average as an estimator $\hat{\mu}(X^n)$ for μ, we can rephrase as follows:

$$P_0^n\left(\frac{\hat{\mu}_n(X^n) - \mu_0}{\sigma_n} \leq x\right) = \Phi(x),$$

for all $x \in \mathbb{R}$, where Φ denotes the distribution function of the standard normal distribution. Consequently,

$$P_0^n\left(\hat{\mu}_n - \frac{\sigma x}{\sqrt{n}} < \mu_0 \leq \hat{\mu}_n + \frac{\sigma x}{\sqrt{n}}\right) = \Phi(x) - \Phi(-x).$$

Fixing some confidence level $\alpha > 0$, we solve for $x_{\alpha/2}$ in the equation $\Phi(x_{\alpha/2}) - \Phi(-x_{\alpha/2}) = 1 - \alpha$ to arrive at the conclusion that the interval,

$$C_\alpha = \left[\hat{\mu}_n - \frac{\sigma x_{\alpha/2}}{\sqrt{n}}, \hat{\mu}_n + \frac{\sigma x_{\alpha/2}}{\sqrt{n}}\right]$$

is a level-α confidence set for the parameter μ.

As in example 2.3.2, confidence intervals for a parameter θ are often derived from estimators for θ: in the case of example 2.3.2 the sample average estimates μ and it is the distribution of the sample average around μ that determines the confidence interval.

Definition 2.3.3. Let $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$ be an identifiable model with measurable parameter space $(\Theta, \mathscr{G})$. Let $\hat{\theta} : \mathscr{Y} \to \Theta$ be a measurable estimator for the parameter θ and assume that $Y \sim P_0$. Then the distribution of $\hat{\theta}(Y)$ over Θ, characterised by probabilities,

$$\mathscr{G} \to [0,1] : G \mapsto P_0\big(\hat{\theta}(Y) \in G\big),$$

is called the *sampling distribution* of the estimator (under P_0).

If we assume that $P_0 = P_{\theta_0}$ for some $\theta_0 \in \Theta$, and we realise that the randomness in $\hat{\theta}(Y)$ occurs due to $Y \sim P_{\theta_0}$, we expect the sampling distribution of $\hat{\theta}(Y)$ to depend on θ_0; in fact, it is exactly this dependence that allows us to draw statistical conclusions. Moreover, the sampling distribution gives concrete meaning to the amount of *uncertainty* surrounding estimation with $\hat{\theta}$. In example 2.3.2, the location of the sampling distribution for $\hat{\mu}$ is μ and σ_n determines the probabilities that differences $|\hat{\mu} - \mu|$ exceed x for all $x > 0$. To summarize this estimator-based perspective: a confidence set expresses how much uncertainty remains concerning the true value of a parameter after estimation.

Generally, sampling distributions are not available except for the simplest estimators in the simplest models so in most cases, confidence sets have to be approximated in some way. The most popular approximation applies with very large samples, based on the central limit. To accommodate this approximation, we define sequences of confidence sets that reach the required confidence level in the limit $n \to \infty$.

Definition 2.3.4. For every $n \geq 1$, let $X^n \sim P_{0,n}$ be data taking values in sample spaces $(\mathscr{X}_n, \mathscr{B}_n)$, with models $\mathscr{P}_n$ and identifiable parametrizations $\Theta \to \mathscr{P}_n : \theta \mapsto P_{\theta,n}$. Choose a confidence level $\alpha \in [0,1)$. Random subsets $C_{\alpha,n} : \mathscr{X}_n \to \mathscr{C}$ of Θ such that $\{x^n \in \mathscr{X}_n : \theta \in C_{\alpha,n}(x^n)\}$ is $\mathscr{B}_n$-measurable for every $\theta \in \Theta$ and,

$$\liminf_{n\to\infty} P_{\theta,n}\big(\theta \in C_{\alpha,n}(X^n)\big) \geq 1 - \alpha, \tag{2.24}$$

for all $\theta \in \Theta$, are called *asymptotic confidence sets* of *asymptotic confidence level* (or *coverage*) α.

Desirable properties of *sequences* of confidence sets are expressed by limits, for example, coverage with high probability (where we take $\alpha = 0$ in (2.24)). However,

the constant choice $C_n(X^n) = \Theta$ shows that coverage alone expresses only part of what a confidence set is to supply: in addition to coverage we want confidence sets to be as small as possible in order to be *informative*. To re-phrase, another desirable property is *non-coverage* of values of the parameter other than the true one.

Definition 2.3.5. In the setting of definition 2.3.4, asymptotic confidence sets $C_n(X^n)$, $n \geq 1$, that satisfy,

$$\lim_{n\to\infty} P_{\theta,n}\big(\theta \notin C_n(X^n)\big) = 0, \tag{2.25}$$

for all $\theta \in \Theta$, are said to be *(asymptotically) consistent*. If the sets $C_n(X^n)$ satisfy,

$$\lim_{n\to\infty} P_{\theta,n}\big(\theta' \in C_n(X^n)\big) = 0, \tag{2.26}$$

for all $\theta, \theta' \in \Theta$ such that $\theta' \neq \theta$, they are said to be *(asymptotically) informative*.

Confidence levels can be re-introduced in both limits (2.25) and (2.26) by the refinements, (for $\theta, \theta' \in \Theta$, $\theta' \neq \theta$),

$$P_{\theta,n}\big(\theta \notin C_n(X^n)\big) = o(a_n), \quad P_{\theta',n}\big(\theta \in C_n(X^n)\big) = o(a_n),$$

depending on a *level sequence* (a_n), $a_n \downarrow 0$. As it turns out, example 2.3.2 applies in an approximate form in all asymptotic cases where the central limit theorem applies, as the following example demonstrates.

Example 2.3.6. Let $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$ be an identifiable, parametric model for measurements X and assume that $X^n = (X_1, \ldots, X_n)$ is an *i.i.d.* sample from a distribution P_{θ_0} for some $\theta_0 \in \Theta$. Suppose that there exists a measurable function $f : \mathbb{R} \to \mathbb{R}$, such that $P_\theta f(X) = \theta$ and $P_\theta f(X)^2 < \infty$ for all $\theta \in \Theta$. Moreover, we assume that for some known constant $S > 0$, $\sigma^2(\theta) = P_\theta(f(X) - \theta)^2 \leq S^2$, for all $\theta \in \Theta$. Consider the sample-average $\hat{\theta}_n(X^n) = n^{-1}\sum_{i--1}^{n} f(X_i)$. According to the *central limit theorem*, estimators that are sample averages for such f have sampling distributions that converge weakly to normal distributions. Choose a confidence level $\alpha \in (0,1)$ and note that,

$$P^n_{\theta_0}\left(-\frac{\sigma(\theta_0)\,x_{\alpha/2}}{\sqrt{n}} < \hat{\theta}_n - \theta_0 \leq \frac{\sigma(\theta_0)\,x_{\alpha/2}}{\sqrt{n}}\right) \to 1-\alpha, \tag{2.27}$$

as $n \to \infty$. Define $C_{\alpha,n}$ by

$$C_{\alpha,n} = \left[\hat{\theta}_n - \frac{S x_{\alpha/2}}{\sqrt{n}}, \hat{\theta}_n + \frac{S x_{\alpha/2}}{\sqrt{n}}\right].$$

Then $P^n_{\theta_0}(\theta_0 \in C_{\alpha,n}) \to 1-\alpha$, so for any $\alpha' > \alpha$, $P^n_{\theta_0}(\theta_0 \in C_{\alpha,n}) \geq 1-\alpha'$ if n is large enough. Note that if we had not used S but $\sigma(\theta_0)$ instead, the θ_0-dependence of $\sigma(\theta_0)$ would violate the requirement that $C_{\alpha,n}$ be a statistic: since the true value θ_0 of θ is unknown, so is $\sigma(\theta_0)$. Substituting the (known) upper-bound S for $\sigma(\theta_0)$ enlarges the $\sigma(\theta_0)$-interval that follows from (2.27), while eliminating the

θ_0-dependence. In a practical situation one would not assume that there is some upper bound $S > 0$, but substitute $\sigma(\theta_0)$ by an estimator $\hat{\sigma}_n(X^n)$ (a practice known as *studentization*, after the Student t-distribution one obtains upon plugging in $\hat{\sigma}_n(X^n)$ for finite $n \geq 1$ with X_i that are marginally normal; refer to the case of example 2.3.2 if σ^2 had not been known.) Since the asymptotics of the studentized version are equal to those of the version based on $\sigma(\theta_0)$, studentization does not change the conclusions we based on (2.27), that is,

$$C'_{\alpha,n} = \left[\hat{\theta}_n(X^n) - \frac{\hat{\sigma}_n(X^n)\, x_{\alpha/2}}{\sqrt{n}}, \hat{\theta}_n(X^n) + \frac{\hat{\sigma}_n(X^n)\, x_{\alpha/2}}{\sqrt{n}} \right].$$

are asymptotic confidence intervals of any level $\alpha' > \alpha$. If we only wish to stress the asymptotic behaviour of the width of confidence sets, a third asymptotic alternative is to define intervals of width $M_n n^{-1/2}$ around $\hat{\theta}_n$, for any M_n that diverge to infinity very slowly, which are asymptotically consistent and asymptotically informative.

The derivation of asymptotic confidence sets in the above example can be generalized quite far: firstly the central limit theorem generalizes to a multi-variate central limit theorem, and secondly, the *delta method* (see Chapter 3 in [248] for an overview) permits generalization to estimators for differentiable functions of expectations. Combined with results for asymptotic optimality of estimators for smooth parameters (see chapter 4), this leads to so-called Wald-type confidence ellipsoids of (4.4), which are viewed as optimal, in that they are based on the sampling distribution of so-called best-regular estimators (see the definition below theorem 4.1.17).

2.3.2 Bayesian credible sets

The Bayesian analogs of confidence sets are called credible sets and are derived from the posterior distribution. The rationale behind the definition of credible sets is exactly the same one that motivated confidence sets: we look for a subset D of the model that is as small as possible while receiving a certain minimal probability. However, here the notion of "probability" is not based on the sampling distribution of an estimator, but on the posterior distribution.

Definition 2.3.7. Let $(\Theta, \mathscr{G})$ be a measurable space parametrizing an identifiable model $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$ for data $Y \in \mathscr{Y}$, with prior Π. Choose an $\alpha \in (0,1)$. Let $D_\alpha : \mathscr{Y} \to \mathscr{G}$ describe a data-dependent, measurable subset of Θ. Then D_α is a *credible set* of *credible level* α for ϑ if it solves the equation,

$$\Pi\big(\vartheta \in D_\alpha(Y) \,\big|\, Y \big) \geq 1 - \alpha, \tag{2.28}$$

almost-surely.

Note that the posterior does not have to be a regular conditional probability in this definition, since it does not rely on countable additivity of the posterior. To find

credible sets in examples one starts by calculating the posterior distribution from the prior and the data and, based on that, derives a subset $D_\alpha(Y)$ such that (2.28) is satisfied. From a frequentist perspective, credible sets are *statistics* since they are defined based entirely on the posterior (which is a statistic itself). A credible set is sometimes referred to as a *credible region*, or, if D is an interval in a one-dimensional parameter space, a *credible interval*. Like with confidence sets, we can extend this definition to the asymptotic regime.

Definition 2.3.8. For every $n \geq 1$, let X^n be data taking values in sample spaces $(\mathscr{X}_n, \mathscr{B}_n)$, with models $\mathscr{P}_n$ and identifiable parametrizations $\Theta \to \mathscr{P}_n : \theta \mapsto P_{\theta,n}$. Choose a sequence of credible levels (α_n), $\alpha_n \in (0,1)$, $\alpha_n \downarrow 0$. A sequence of data-dependent subsets $D_{\alpha,n}(X^n) \in \mathscr{G}$ that solves,

$$\Pi\big(\vartheta \in D_{\alpha,n}(X^n) \,\big|\, X^n\big) = o(\alpha_n), \tag{2.29}$$

almost-surely, is called a *sequence of asymptotic credible sets* of levels (α_n).

Definition 2.3.7 suffices to capture the concept of a credible set, but offers too much freedom in the choice of D: given a level $\alpha > 0$, many sets will satisfy (2.28), just like confidence sets can be chosen in many different ways. Note that, also here, we prefer smaller sets over large ones: if, for some level α, two different level-α credible sets F_α and G_α are given, both satisfying (2.28) and $F_\alpha \subset G_\alpha$ then F_α is preferred over G_α. If the posterior is dominated with density $\theta \mapsto \pi(\theta|Y)$, we can be more specific. We define, for every $k \geq 0$, the data-dependent level-sets,

$$D(Y,k) = \big\{\theta \in \Theta : \pi(\theta|Y) \geq k\big\}, \tag{2.30}$$

and consider so-called HPD-sets (for *highest posterior density*).

Definition 2.3.9. Let $(\Theta, \mathscr{G})$ a measurable space parametrizing a model $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$ for data $Y \in \mathscr{Y}$, with prior Π. Assume that the posterior is almost-surely dominated by a σ-finite measure μ on $(\Theta, \mathscr{G})$, with density $\pi(\,\cdot\,|Y) : \Theta \to [0,\infty)$. Choose $\alpha \in (0,1)$. A *level-α HPD credible set* for ϑ is the subset $D_\alpha(Y) = D(Y, k_\alpha(Y))$, where,

$$k_\alpha(Y) = \sup\big\{k \geq 0 : \Pi(\vartheta \in D(Y,k)|Y) \geq 1-\alpha\big\}.$$

Note that HPD credible sets depend on the choice of dominating measure: if we had chosen to use a different measure μ, HPD credible sets would have changed. In fact, among all credible sets of level α, the HPD credible set $D_\alpha(Y)$ has minimal μ-measure, almost-surely. (See exercise 2.6.17.)

2.3.3 Enlarged credible sets as confidence sets

Consider a statistical experiment in which we observe data X^n of some fixed size n (*e.g.* an n-point *i.i.d.* or Markov sample) and assume, for the moment, that the

parameter space Θ_n for the model $\mathscr{P}_n = \{P_{\theta_n,n} : \theta_n \in \Theta_n\}$ is *finite*. (An example of such a situation is found in chapter 11, where we study a random graph X^n with n vertices in two disjoint communities, and the parameter θ_n assigns all vertices to one or the other community.) We assume also a prior Π_n on Θ_n, such that, with growing n, an arbitrarily large fraction of the posterior mass ends up in the singleton $\{\theta_n\}$ with high $P_{\theta_n,n}$-probability. Then it is clear that *any* sequence of credible sets $D_n(X^n)$ of credible levels $1-\gamma_n$ with $\liminf_n \gamma_n > 0$, will contain θ_n with high $P_{\theta_n,n}$-probability as $n \to \infty$. So, asymptotically, the credible sets $D_n(X^n)$ are also consistent confidence sets.

This asymptotic argument can be made precise also for fixed values of n: if a credible set $D_n(X^n)$ receives posterior mass $1-\gamma$ and the probability of finding the posterior weight of the singleton $\{\theta_n\}$ *above* γ is high, then the probability that $\theta_n \in D_n(X^n)$, is equally high. In fact, the argument also holds in case the parameter space is not finite or discrete: suppose that, for any $\theta_n \in \Theta_n$, there exists a measurable $B(\theta_n) \subset \Theta_n$ such that $\theta_n \in B(\theta_n)$, with posterior mass above γ with high $P_{\theta_n,n}$-probability. Then $B(\theta_n)$ intersects credible sets $D_n(X^n)$ of credible levels $1-\gamma$ with high $P_{\theta_n,n}$-probability. Consequently, the unknown θ_n lies in the *union of all* $B(\theta_n)$*'s that intersect* $D_n(X^n)$, with high $P_{\theta_n,n}$-probability. So if we know that the posterior tends to concentrate mass in neighbourhoods $B(\theta_n)$ of the truth θ_n in Θ_n, then we may enlarge credible sets to obtain confidence sets (see [158]).

Lemma 2.3.10. *Fix $n \geq 1$ and some prior Π_n on Θ_n, let $\theta_n \in \Theta_n$ and $X^n \sim P_{\theta_n,n}$ be given. Let $B(\theta_n) \subset \Theta_n$ be a subset with expected posterior probability that is lower-bounded,*

$$P_{\theta_n,n}\Pi\big(B(\theta_n) \,\big|\, X^n\big) \geq 1-\beta, \tag{2.31}$$

for some $0 < \beta < 1$. For any $0 < \gamma < 1$ and any credible set $D(X^n) \subset \Theta_n$ of level $1-\gamma$,

$$P_{\theta_n}\big(B(\theta_n) \cap D(X^n) \neq \varnothing\big) \geq 1 - \frac{\beta}{1-\gamma}.$$

Proof. We first prove that for every $0 < r < 1$,

$$P_{\theta_n,n}\big(\Pi(B(\theta_n)|X^n) \geq r\big) \geq 1 - \frac{\beta}{1-r},$$

by contradiction: let $\delta > 0$ be given and define the event,

$$E = \big\{x_n \in \mathscr{X}_n : \Pi\big(B(\theta_n) \,\big|\, X^n = x^n\big) \geq r\big\}.$$

Suppose that $P_{\theta_n,n}(E) \leq 1-\beta/(1-r)-\delta$. Then,

$$P_{\theta_n,n}\Pi(B(\theta_n)|X^n) \leq P_{\theta_n,n}(E) + r(1-P_{\theta_n,n}(E)) \leq 1-\beta-\delta(1-r) < 1-\beta, \tag{2.32}$$

which contradicts the assumption that $P_{\theta_n,n}\Pi(B(\theta_n)|X^n) \geq 1-\beta$. Since this holds for every $\delta > 0$, we have $P_{\theta_n,n}(E) \geq 1-\beta/(1-r)$. Choose $r > \gamma$. As $D(X^n)$ has posterior mass at least $1-\gamma$, $B(\theta_n)$ and $D(x^n)$ cannot be disjoint for $x^n \in E$. So,

$$P_{\theta_n,n}\big(B(\theta_n)\cap D(X^n)\neq\varnothing\big)\geq P_{\theta_n,n}(E)\geq 1-\frac{\beta}{1-\gamma},$$

which proves the assertion.

We formulate the more practical versions of the above lemma in the form of the following two corollaries.

Corollary 2.3.11. *Fix $n\geq 1$ and some prior Π_n on a (discrete) parameter space Θ_n, let $\theta_n\in\Theta_n$ and $X^n\sim P_{\theta_n,n}$ be given. Suppose that for some $0<\beta<1$,*

$$P_{\theta_n,n}\Pi\big(\{\theta_n\}\,\big|\,X^n\big)\geq 1-\beta, \tag{2.33}$$

Then for any $0<\gamma<1$ and any credible set $D(X^n)\subset\Theta_n$ of level $1-\gamma$,

$$P_{\theta_n,n}\big(\theta_n\in D(X^n)\big)\geq 1-\frac{\beta}{1-\gamma}.$$

Proof. The assertion follows directly from lemma 2.3.10 upon the choice $B(\theta_n)=\{\theta_n\}$, for all $\theta_n\in\Theta_n$.

The lower bound on posterior mass in any point of the parameter space is reasonable only in the context of *discrete* parameter spaces. In non-discrete parameter spaces, a different version of the argument is needed. For that, consider a *metric* parameter space (Θ_n,d_n). Then $B(\theta_n)$ is defined as a d_n-ball around θ_n of some radius, large enough to guarantee that (2.31) holds.

Definition 2.3.12. Fix $n\geq 1$. For any credible set $D(X^n)$ and radius $r_n>0$, we define the *(d_n-)enlargement* $C(X^n)$ of $D(X^n)$ of radius r_n, to be,

$$C(X^n)=\big\{\theta_n\in\Theta_n:\exists_{\eta_n\in D_n(X^n)},d_n(\theta_n,\eta_n)\leq r_n\big\},$$

i.e. the union of all radius-r_n d_n-balls centred on points in $D(X^n)$.

Corollary 2.3.13. *Fix $n\geq 1$ and some prior Π_n on a (discrete) parameter space Θ_n, let $\theta_n\in\Theta_n$ and $X^n\sim P_{\theta_n,n}$ be given. Suppose that for some $0<\beta<1$,*

$$P_{\theta_n,n}\Pi\big(\{\eta_n:d_n(\theta_n,\eta_n)\leq r_n\}\,\big|\,X^n\big)\geq 1-\beta, \tag{2.34}$$

Then for any $0<\gamma<1$ and any credible set $D(X^n)\subset\Theta_n$ of level $1-\gamma$, the d_n-enlargement $C(X^n)$ satisfies,

$$P_{\theta_n,n}\big(\theta_n\in C(X^n)\big)\geq 1-\frac{\beta}{1-\gamma},$$

i.e. $C(X^n)$ is a confidence set of said level.

Proof. The assertion follows directly from lemma 2.3.10 upon the choice $B(\theta_n)=\{\eta_n\in\Theta_n:d_n(\eta_n,\theta_n)\leq r_n\}$, for all $\theta_n\in\Theta_n$.

One might expect the relation between Bayesian and frequentist uncertainty quantification to involve some type of proportionality between credible and confidence levels, not just asymptotically but also at finite sample sizes. Somewhat surprisingly, it emerges that the finite-sample confidence level of a credible set depends mostly on the expected amount of mis-placed posterior probability and less on the credible level. Note that it is important that the lower bound (2.34) is *sharp*: unnecessarily large values of β or r_n cause unnecessarily high credible levels and lead to unnecessarily conservative enlargement radii.

Note that the credible sets $D(X^n)$ or their enlargements $C(X^n)$ are *exact confidence regions at finite sample sizes*. Compare this with, for example, the Wald-type confidence ellipsoids of (4.4) which are approximate confidence sets motivated by the large-sample limit behaviour of likelihood functions in smooth parametric models. In the latter category of models, posterior asymptotic behaviour and the relationship between credible sets and Wald-type sets is studied in chapter 4. The above corollaries require only finite amounts of data, and leave the parameter space largely unrestricted, *e.g.* non-parametric and n-dependent. In section 7.7 we consider asymptotic enlargement of credible sets again, leaving room for non-metric neighbourhoods $B(\theta_n)$ and θ_n-dependent radii r_n and foregoing the requirement on concentration of posterior mass. In chapter 11 we use corollaries 2.3.11 and 2.3.13 to derive confidence sets for community assignment vectors in a two-community stochastic block model.

2.3.4 Asymptotic confidence balls from converging posteriors

The argument of the previous subsection can also be analysed in the large-sample limit $n \to \infty$. Using the notation of the previous subsection, if we can show that posteriors satisfy (see definition 6.4.1),

$$P_{\theta_n,n}\Pi\big(\Theta_n \setminus B(\theta_n) \,\big|\, X^n\big) = O(\beta_n),$$

for $\beta_n \to 0$, and we consider credible sets $D(X^n)$ of levels $1-\gamma_n$, that do not go to zero (too fast). Then the enlargements $C(X^n)$ satisfy,

$$P_{\theta_n,n}\big(\theta_n \in C(X^n)\big) \geq 1 - \frac{\beta_n}{1-\gamma_n} \to 1.$$

Moreover in metric parameter spaces, the radii of the enlargements can be controlled. Again, we assume that (Θ_n, d_n) are metric spaces. Denote balls in Θ_n as follows,

$$B(\theta_n, r_n) = \{\eta_n \in \Theta_n : d_n(\eta, \theta_n) \leq r_n\},$$

In case we define credible (or confidence) balls, centre point θ_n and radius r_n are data-dependent, chosen such as to satisfy definition (2.28) (or 7.7.2). Given a certain credible level $1-\gamma$ and a posterior distribution $\Pi(\cdot|X^n)$, there exists a data-

dependent *minimal* radius,

$$\hat{r}_n(\gamma) = \inf\{\, r > 0 : \exists_{\theta_n \in \Theta_n} \Pi(B(\theta_n, r)|X^n) \geq 1-\gamma \},$$

the infimum of radii for which credible balls of said level exist. Below, we formulate a theorem (see [155]) that assumes posterior convergence at a rate r_n and asserts that corresponding enlargements of credible balls of (near-)minimal radii are asymptotically consistent confidence balls.

Theorem 2.3.14. *Suppose that* $0 < \gamma \leq 1$ *and for some radii* $r_n > 0$*,*

$$\Pi\big(B(\theta_n, r_n) \;\big|\; X^n\big) \xrightarrow{P_{\theta_n,n}} 1. \tag{2.35}$$

Let $B(\hat{\theta}_n, \hat{r}_n)$ *be level-*$1-\gamma$ *credible balls of near-minimal radii* $\hat{r}_n = (1+o(1))\hat{r}_n(\gamma)$*. Then with high* $P_{\theta_n,n}$*-probability,* $\hat{r}_n \leq (1+o(1))r_n$*, and the enlargements* $C_n(X^n) = B(\hat{\theta}_n, \hat{r}_n + r_n) \subset B(\hat{\theta}_n, 2(1+o(1))r_n)$ *of the credible balls* $B(\hat{\theta}_n, \hat{r}_n)$ *satisfy,*

$$P_{\theta_n,n}\big(\, \theta_n \in C_n(X^n)\,\big) \to 1,$$

i.e. the $C_n(X^n)$ *are asymptotically consistent confidence balls of radii (arbitrarily close to)* $2r_n$*.*

Proof. Let $n \geq 1$ be given. For every $\theta_n \in \Theta$, let $r_n(\theta_n, X^n)$ denote the infimal radius of balls in Θ_n centred on θ_n of posterior mass at least $1-\gamma$. Define $\hat{\theta}_n(X^n)$ as the centre point of a credible ball $B(\hat{\theta}_n, \hat{r}_n)$ of level $1-\gamma$, with near-minimal radius $\hat{r}_n$,

$$\hat{r}_n \leq (1+o(1)) \inf\{r_n(\theta_n, x^n) : \theta_n \in \Theta_n\}.$$

Note that by definition of $B(\hat{\theta}_n, \hat{r}_n)$ as a credible set (and by the frequentist interpretation of almost-sureness, *c.f.* remark 2.2.3),

$$P_{\theta_n,n}\big(\, \Pi(B(\hat{\theta}_n, \hat{r}_n)|X^n) \geq 1-\gamma\,\big) = 1,$$

for all $n \geq 1$. Posterior convergence implies that, for large enough n, the ball $B(\theta_n, r_n)$ is a credible ball of level $1-\gamma$, with high $P_{\theta_n,n}$-probability. Therefore,

$$\hat{r}_n \leq (1+o(1))\, r_n(\theta_{0,n}, X^n) \leq (1+o(1))\, r_n,$$

with high $P_{\theta_n,n}$-probability. Again based on posterior convergence, the balls $B(\theta_n, r_n)$ satisfy,

$$P_{\theta_n,n}\big(\, \Pi(B(\theta_n, r_n)|X^n) > \gamma\,\big) \to 1.$$

Conclude that, with high $P_{\theta_n,n}$-probability,

$$B(\theta_n, r_n) \cap B(\hat{\theta}_n, \hat{r}_n) \neq \varnothing,$$

implying asymptotic coverage of θ_n for the enlargements $C_n(X^n)$, with high $P_{\theta_n,n}$-probability.

It is noted that this construction does *not* enable *rate-adaptivity* of the confidence balls $C(X^n)$ (see [125, 54, 239]): for the construction of the enlarged sets $C(X^n)$, a *known rate* r_n is required. If such a rate is dependent on the underlying data-distribution, estimation of r_n is problematic [239].

2.4 Testing hypotheses, posterior odds and Bayes factors

Having discussed confidence sets and credible sets in the previous section, we now turn to the related subject of hypothesis testing. We start with a discussion of the Neyman-Pearson framework and the famous lemma concerning the optimal test for testing one distribution versus another. Next we consider tests of uniform testing power and minimax optimality, as well as asymptotic testing. In the last subsection we consider posterior odds and Bayes factors, as well as Bayesian test functions.

2.4.1 Neyman-Pearson tests

Assume that we have data $Y \in \mathscr{Y}$ and a parametrized model $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ such that $Y \sim P_{\theta_0}$ for some $\theta_0 \in \Theta$. For simplicity, we assume that $\Theta \subset \mathbb{R}$ whenever the dimension of Θ is of importance. In statistical testing the *hypotheses* are mutually exclusive speculations concerning the distribution of the data. The model contains all distributions the frequentist takes into account as candidates for P_0, so hypotheses are formulated in terms of a partition of the model (or its parametrization space) into two disjoint subsets.

Definition 2.4.1. Testing of hypotheses proceeds through choice of a model subset Θ_0 corresponding to the so-called *null hypothesis* H_0 and it's complement $\Theta_1 = \Theta \setminus \Theta_0$, called the *alternative hypothesis* H_1. We distinguish between so-called *simple* hypotheses which consist of a single point in Θ and *composite* hypotheses which consist of bigger subsets in Θ.

Based on the frequentist assumption that there is a true value θ_0 of the parameter, the simplest, most intuitive question one can ask regarding H_0 and H_1, is which of the two is deemed most likely to contain P_0 given the data Y. But that is not the most popular frequentist testing tool: in the *Neyman-Pearson testing procedure*, H_0 and H_1 do *not* have symmetric roles. The goal of Neyman-Pearson hypothesis testing is not to choose one or the other, but to find out *whether or not the data contains "enough" evidence to reject H_0 as a likely explanation when compared to explanations offered by the alternative*. To paraphrase: the outcome of the procedure is acceptance of H_1 or not, and never leads to the conclusion that we accept H_0. So Neyman-Pearson testing differs from *symmetric testing*, in which H_0 and H_1 play interchangeable roles and we make a choice for one or the other based on the data, as in subsections 2.4.3 and 2.4.4.

In the Neyman-Pearson paradigm, one usually departs from a *test statistic* $T(Y) \in \mathbb{R}^d$, displaying different behaviour depending on whether the data Y is distributed according to a distribution in H_0 or a distribution in H_1. To make the distinction, one defines a *critical set* $K \subset \mathbb{R}^d$ such that $P_\theta(T \in K)$ is "small" for all $\theta \in \Theta_0$ and $P_\theta(T \notin K)$ is "small" for all $\theta \in \Theta_1$. What "small" probabilities are in this context is quantified by the so-called *significance level* $\alpha \in (0,1)$.

Definition 2.4.2. Let $\Theta \to \mathscr{P} : \theta \to P_\theta$ be an identifiable, parametrized model for a sample Y. Formulate two hypotheses H_0 and H_1 by introducing a two-set partition $\{\Theta_0, \Theta_1\}$ of the model Θ:

$$H_0 : \quad \theta_0 \in \Theta_0, \qquad H_1 : \quad \theta_0 \in \Theta_1.$$

We say that a test for these hypotheses based on a *test-statistic* $T : \mathscr{Y} \to \mathbb{R}^d$ with *critical set* $K \subset \mathbb{R}^d$ is of *significance level* $\alpha \in (0,1)$ if the *power function* $\pi : \Theta \to [0,1]$, defined by

$$\pi(\theta) = P_\theta\big(T(Y) \in K\big),$$

is uniformly small over Θ_0:

$$\sup_{\theta \in \Theta_0} \pi(\theta) \leq \alpha. \tag{2.36}$$

From the above definition we arrive at the conclusion that if $Y = y$ and $T(y) \in K$, hypothesis H_0 is improbable enough to be rejected, since H_0 forms an "unlikely" explanation of observed data (at said significance level). The degree of "unlikeliness" can be quantified in terms of the so-called *p-value*, which is the lowest significance level at which the realised value of the test statistic $T(y)$ would have led us to reject H_0.

Of course there is the possibility that our decision is wrong and H_0 is actually true but $T(y) \in K$ nevertheless, so that our rejection of the null hypothesis is unwarranted. This is called a *type-I error*; a *type-II error* is made when we do *not* reject H_0 while H_0 is not true. The significance level α represents a fixed upper bound for the probability of a type-I error, *c.f.* (2.36). Clearly, there is a test that never makes a type-I error (any T with $K = \varnothing$, which never leads to rejection of H_0, a valid Neyman-Pearson test for any significance level $\alpha \in (0,1)$), but the type-II error probability equals one. If, for some significance level $\alpha \in (0,1)$, we have two different test that satisfy the Type-I error bound (2.36), we prefer the test with minimal Type-II error probability. Ideally we look for a pair (T,K) satisfying (2.36), that minimizes $P_\theta(T(Y) \notin K)$ for all $\theta \in \Theta_1$. However, generically such *uniformly most-powerful tests* do not exist, because of the possibility that some pair (T,K) is most powerful over some subset of Θ_1, while some other pair (T',K') is most powerful over some other subset of Θ_1. We consider the Neyman-Pearson approach to testing in some more detail in the following example in the context of normally distributed data.

Example 2.4.3. Consider a model $\mathscr{P}$ of normal distributions $N(\mu,\sigma^2)$ with unknown location $\mu \in \mathbb{R}$ and known variance $\sigma^2 > 0$. Let $X^n = (X_1, \ldots, X_n)$ be an

i.i.d. sample from a normal distribution $P_0 = N(\mu, \sigma^2)$ for some $\mu \in \mathbb{R}$. By choosing some location $\mu_0 \in \mathbb{R}$, we formulate null- and alternative hypotheses,

$$H_0: \quad \mu = \mu_0, \qquad H_1: \quad \mu \neq \mu_0,$$

We also choose a significance level $\alpha \in (0,1)$. As we have seen in example 2.3.2 and exercise 2.6.14, the sample average is normally distributed, $\hat{\mu}_n(X^n) = n^{-1}\sum_i X_i \sim N(\mu, \sigma^2/n)$. Note that,

$$P_\mu^n\Big(\sqrt{n}\big(\hat{\mu}_n(X^n) - \mu\big) \leq \sigma x\Big) = \Phi(x),$$

for all $x \in \mathbb{R}$, where Φ denotes the distribution function of the standard normal distribution. Re-write to obtain,

$$P_\mu^n\left(-\frac{\sigma x}{\sqrt{n}} < \hat{\mu}_n(X^n) - \mu \leq \frac{\sigma x}{\sqrt{n}}\right) = \Phi(x) - \Phi(-x),$$

for $x > 0$. So under the null-hypothesis,

$$P_{\mu_0}^n\left(\mu_0 - \frac{x_{\alpha/2}\,\sigma}{\sqrt{n}} < \hat{\mu}_n(X^n) \leq \mu_0 + \frac{x_{\alpha/2}\,\sigma}{\sqrt{n}}\right) = 1 - \alpha,$$

if we choose the quantiles $x_{\alpha/2}$ like in example 2.3.2. Hence the null-hypothesis makes it improbable to observe $|\hat{\mu}_n(X^n) - \mu_0| > n^{-1/2}\sigma x_{\alpha/2}$, which gives rise to the following definition of the critical set $K_{\alpha,n}$,

$$K_{\alpha,n} = \mathbb{R} \setminus \left[\mu_0 - \frac{x_{\alpha/2}\,\sigma}{\sqrt{n}}, \mu_0 - \frac{x_{\alpha/2}\,\sigma}{\sqrt{n}}\right],$$

enabling us to formulate our decision on rejection of the null hypothesis,

(i) if $\hat{\mu}_n(X^n) \in K_{\alpha,n}$, we reject H_0,
(ii) if $\hat{\mu}_n(X^n) \notin K_\alpha$, we do not reject H_0,

at significance level α. We re-iterate the warning regarding interpretation: under case *(ii)*, we do *not* draw the conclusion that H_0 is accepted: the data does not provide enough evidence to reject the null hypothesis but that does not imply that we accept it.

Note the behaviour of the procedure with varying sample-size: keeping the significance level fixed, the width of the critical sets $K_{\alpha,n}$ is of order $O(n^{-1/2})$, so smaller and smaller critical sets can be used as more information concerning the distribution P_0 (read, data) comes available. Conversely, if we keep the critical set fixed, the probability for a Type-I error decreases (exponentially) with growing sample-size. Analogous to example 2.3.6, it is common practice to use *asymptotic* Neyman-Pearson tests (usually based on the *central limit theorem*), because sampling distributions are rarely available in closed form.

2.4.2 Randomized tests and the Neyman-Pearson lemma

According to the Neyman-Pearson approach, tests are only considered if they satisfy (2.36) and are optimal if, in addition, they maximize testing power over the alternative uniformly. Optimality in this sense is sometimes not achievable, so one wonders if, mathematically, anything can be said at all. The Neyman-Pearson lemma answers this question in the affirmative for problems where the null- and alternative hypotheses each contain a single distribution.

To formulate the Neyman-Pearson lemma, however, we have to generalize the testing procedure slightly: as it turns out the existence of an optimal test can only be guaranteed if we allow for a *randomization* of our decision.

Definition 2.4.4. Let $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ be a parametrized model for data Y taking values in a measurable sample space $(\mathscr{Y}, \mathscr{B})$, distributed according to P_θ for some $\theta \in \Theta$. Formulate two hypotheses H_0 and H_1 for θ based on a two-set partition $\{\Theta_0, \Theta_1\}$ of the model $\mathscr{P}$:

$$H_0 : \quad \theta \in \Theta_0, \qquad H_1 : \quad \theta \in \Theta_1.$$

A *test function* ϕ is a measurable map $\phi : \mathscr{Y} \to [0,1]$ used in the following procedure called a *randomized test*: given Y, we reject H_0 with probability $\phi(Y)$, and otherwise we do not reject H_0. The *power function* associated with the test function ϕ is given by $\pi : \Theta \to [0,1] : \theta \mapsto P_\theta \phi(Y)$.

To explain the randomized test procedure differently, view it as follows: after we observe $Y = y$, we determine $\phi(y)$ and draw an (independent) $V \sim U[0,1]$: if $V \leq \phi(y)$, we reject H_0, if $V > \phi(y)$, we do not. Note that if we use the test function $\phi(Y) = 1\{T(Y) \in K\}$, the randomized test reduces to the original (non-random) procedure of rejecting H_0 if $T(Y) \in K$. Clearly the probability for Type-I error when using the randomized procedure equals $\pi(\theta)$ (for $\theta \in \Theta_0$) and the probability for Type-II error equals $1 - \pi(\theta)$ (for $\theta \in \Theta_1$). When we fix a significance level α, we require that ϕ gives rise to a Type-I error probability that is bounded by α, uniformly over Θ_0, *c.f.* (2.36). Among test functions ϕ that satisfy (2.36), we look for a test that minimizes the Type-II error probability $1 - \pi(\theta)$ for values of θ in Θ_1.

Existence of *uniformly most powerful* randomized tests of a fixed significance level cannot be guaranteed (again, some ϕ may be optimal for certain values of θ in Θ_1, while some other ϕ' is optimal for other values). However, if both null and alternative hypothesis are simple, a (randomized) optimal test exists by the famed Neyman-Pearson lemma [170].

Lemma 2.4.5. *Suppose the model is $\mathscr{P} = \{P_{\theta_0}, P_{\theta_1}\}$ and write $p_{\theta_0} : \mathscr{Y} \mapsto \mathbb{R}$ and $p_{\theta_1} : \mathscr{Y} \mapsto \mathbb{R}$ for the densities of P_{θ_0} and P_{θ_1} relative to some σ-finite measure μ. Choose a significance level $\alpha \in (0,1)$ and consider a test of the form,*

$$\phi(y) = \begin{cases} 1 & \text{if } p_{\theta_1}(y) > c\, p_{\theta_0}(y) \\ \gamma(y) & \text{if } p_{\theta_1}(y) = c\, p_{\theta_0}(y) \\ 0 & \text{if } p_{\theta_1}(y) < c\, p_{\theta_0}(y) \end{cases}, \tag{2.37}$$

where the measurable function $\gamma : \mathscr{Y} \to [0,1]$ *and the constant* $c \in [0,\infty]$ *form a solution to the equation:*

$$P_{\theta_0}\phi(Y) = \alpha,$$

The following two assertions concern the hypotheses,

$$H_0 : \quad \theta = \theta_0, \qquad H_1 : \quad \theta = \theta_1.$$

(i.) If a test of the form (2.37) has significance level α *then it is most powerful among all tests of level* α.

(ii.) If a test ϕ' *is most powerful, then* ϕ' *is of the form (2.37) for some* $\gamma(x)$ *and some c, almost-surely with respect to both* P_{θ_0} *and* P_{θ_1}.

Proof. See Lehmann and Cassela (2005) [170].

The lemma is often used in conjunction with some condition on the model (or its likelihood function) to extend this point-vs-point version to composite hypotheses which are more interesting from a practical point of view.

Example 2.4.6. Suppose that we consider a random variable X drawn from a normal distribution $N(\theta,1)$ where $\theta \in \Theta = \{-1,1\}$. Fixing a significance level $\alpha \in (0,1)$, we consider a test of the form (2.37) for the hypotheses,

$$H_0 : \theta = -1, \quad H_1 : \theta = +1. \tag{2.38}$$

A simple calculation shows that $p_{+1}(X)/p_{-1}(X) = e^{2X}$, so, with Φ denoting the distribution function for the standard normal distribution,

$$P_{-1}\phi(X) = P_{-1}\big(p_{+1}(X) > c\,p_{-1}(X)\big) = P_{-1}\big(e^{2X} > c\big)$$

$$= P_{-1}(X > \tfrac{1}{2}\log c) = 1 - \Phi(\tfrac{1}{2}\log c + 1),$$

where we have used that X is distributed continuously (so that the middle term in (2.37) does not play a role and any γ will do), and that $P_{-1}(X \le x) = \Phi(x+1)$. So to find c, we solve $1 - \Phi(\frac{1}{2}\log c + 1) = \alpha$, so that the Neyman-Pearson procedure for testing the hypotheses (2.38) has the form,

(i) if $X > x_{1-\alpha} - 1$, we reject H_0, and,
(ii) if $X \le x_{1-\alpha} - 1$, we do not see enough evidence in the data to reject H_0,

at significance level α.

2.4.3 Symmetric and asymptotic testing

Two points remain, the first being an asymptotic perspective on testing: just like we often study limits of sequences of estimators and conditions for their optimality, we are interested also in sequences of tests.

Definition 2.4.7. For every $n \geq 1$, let $X^n \sim P_{0,n}$ be data taking values in sample spaces $(\mathscr{X}_n, \mathscr{B}_n)$, with models $\mathscr{P}_n$ and identifiable parametrizations $\Theta \to \mathscr{P}_n : \theta \mapsto P_{\theta,n}$. Choose a significance level $\alpha \in [0,1)$. Test functions $\phi_n : \mathscr{X}_n \to [0,1]$ such that,

$$\limsup_{n\to\infty} P_{\theta,n}\phi_n(X^n) \leq \alpha, \tag{2.39}$$

for all $\theta \in \Theta_0$, are called *asymptotic tests* of *asymptotic significance level* α. The *power sequence* of the test sequence (ϕ_n), $\pi_n : \Theta \to [0,1]$ is defined by:

$$\pi_n(\theta) = P_{\theta,n}\phi_n,$$

representing the $P_{\theta,n}$-probability of rejecting H_0.

The quality of the test sequence depends on the behaviour of the power sequence on Θ_0 and Θ_1: we could follow the Neyman-Pearson paradigm again, choose an α, restrict to those ϕ_n that satisfy (2.39) and prefer test sequences that have high power on the alternative in the limit $n \to \infty$. But here, we re-formulate to accommodate symmetric roles for null- and alternative hypotheses and we change the procedure accordingly: we reject H_0 and accept H_1 (resp. accept H_0 and reject H_1) randomly with probability $\phi_n(X^n)$ (resp. $1 - \phi_n(X^n)$). We also discard the significance level and simply require convergence to the ideal power function (*e.g.* $\pi(\theta) = 0$ for $\theta \in \Theta_0$ and $\pi(\theta) = 1$ for $\theta \in \Theta_1$) in the limit $n \to \infty$.

Definition 2.4.8. In the setting of definition 2.4.7, a test sequence (ϕ_n) that satisfies,

$$\lim_{n\to\infty} P_{\theta,n}\phi_n(X^n) = 0, \tag{2.40}$$

for all $\theta \in \Theta_0$, and

$$\lim_{n\to\infty} P_{\theta,n}\big(1 - \phi_n(X^n)\big) = 0, \tag{2.41}$$

for all $\theta \in \Theta_1$, are said to be *(asymptotically) consistent*.

With this new definition, let us consider a sequential version of the likelihood ratio test of the Neyman-Pearson lemma, lemma 2.4.5.

Example 2.4.9. Suppose that we are given two sequences (P_n) and (Q_n) of distributions for data X^n taking values in measurable spaces $(\mathscr{X}_n, \mathscr{B}_n)$ for all $n \geq 1$. We hypothesise that either $X^n \sim P_n$ or $X^n \sim Q_n$ and wish to determine statistically which is true. This is the setting of the Neyman-Pearson lemma, so a test based on *likelihood ratios* dP_n/dQ_n seems reasonable (define $\mu_n = P_n + Q_n$ and write $p_n = dP_n/d\mu_n$ and $q_n = dQ_n/d\mu_n$ for the Radon-Nikodym derivatives):

$$\phi_n(X^n) = 1\{p_n(X^n) < q_n(X^n)\}.$$

Then,

$$
\begin{aligned}
&P_n\phi_n + Q_n(1-\phi_n) \\
&\quad = \int_{\mathscr{X}_n} \big(p_n(x^n)1\{p_n(x^n) < q_n(x^n)\} + q_n(x^n)1\{p_n(x^n) \geq q_n(x^n)\}\big)\, d\mu_n(x^n) \\
&\quad \leq \int_{\mathscr{X}_n} \sqrt{p_n(x^n)q_n(x^n)}\, d\mu_n(x^n) = 1 - \frac{1}{2}\int_{\mathscr{X}_n} \big(\sqrt{p_n(x^n)} - \sqrt{q_n(x^n)}\big)^2\, d\mu_n(x^n) \\
&\quad \leq 1 - H^2(P_n, Q_n).
\end{aligned}
$$

Moreover, for any other choice of test function ψ_n, $P_n\phi_n + Q_n(1-\phi_n) \leq P_n\psi_n + Q_n(1-\psi_n)$. So if the Hellinger distances $H(P_n, Q_n) \to 1$, a consistent test for (P_n) versus (Q_n) exists (namely the likelihood ratio test). This conclusion is very general and emphasizes the fundamental role that the Hellinger metric plays in mathematical statistics.

In case $X^n = (X_1, \ldots, X_n) \in \mathscr{X}^n$ represents an *i.i.d.* sample, P_n and Q_n are n-fold product measures of distributions $P_{1,n}$ and $Q_{1,n}$ on $\mathscr{X}$: $P_n = P_{1,n}^n$, $Q_n = Q_{1,n}^n$. According to exercise 4.4.1, $H^2(P_n, Q_n) \leq nH^2(P_{1,n}, Q_{1,n})$ for all $n \geq 1$. The above upper bound suggests that, for the existence of a test that consistently distinguishes between $P_{1,n}$ and $Q_{1,n}$, it is necessary that $H^2(P_{1,n}, Q_{1,n}) \geq hn^{-1/2}$ for some $h > 0$, and this is indeed the case (see [236], or note that $H^2(P_{1,n}, Q_{1,n}) = O(n^{-1/2})$ implies *contiguity*, *c.f.* definition 7.2.1 and the remark that follows it).

Depending on the subsets Θ_0 and Θ_1, there is a question whether a consistent test sequence for the pair exists or not. The answer in the case of *i.i.d.* sampling, which is given in chapter 9, characterizes those Θ_0, Θ_1 that can be tested consistently, as precisely those subsets that can be written as countable unions of 'closed' sets (where the relevant model topology requires further discussion).

Given two test sequences, we may compare them through the limits of their type-I and type-II errors.

Definition 2.4.10. Let (ϕ_n) and (ψ_n) be two test sequences for Θ_0 versus Θ_1. Let $\theta \in \Theta_0$ (resp. $\eta \in \Theta_1$) be given. We say that (ϕ_n) is *asymptotically more powerful* than (ψ_n) at $\theta \in \Theta_0$ (resp. $\eta \in \Theta_1$), if,

$$
\lim_{n\to\infty} P_{\eta,n}\phi_n \leq \lim_{n\to\infty} P_{\theta,n}\psi_n, \quad \big(\text{resp. } \lim_{n\to\infty} P_{\eta,n}\phi_n \geq \lim_{n\to\infty} P_{\theta,n}\psi_n\big). \tag{2.42}
$$

If (2.42) holds for *all* points in Θ, the test sequence (ϕ_n) is said to be uniformly asymptotically more powerful than (ψ_n). If one can show that this holds for all test sequences (ψ_n), then (ϕ_n) is said to be *uniformly asymptotically most powerful*.

This ordering of test sequences is not complete: it is quite possible that (ϕ_n) is asymptotically more powerful than (ψ_n) on a subset of Θ, whereas on its complement in Θ, (ψ_n) is asymptotically more powerful. As a result, the existence of uniformly asymptotically most powerful test sequences is problematic and no generalization of the Neyman-Pearson lemma exists for composite hypotheses, not even when required only asymptotically.

To counter such problems we can choose to evaluate testing power for a test sequence (ϕ_n) *uniformly* over Θ_0 and Θ_1, adding *maximal* type-I and -II error probabilities.

Definition 2.4.11. Given two disjoint model subsets Θ_0, Θ_1, the *uniform testing power* (π_n) for Θ_0 versus Θ_1 of a test sequence (ϕ_n) is given by,

$$\pi_n = \sup_{\theta \in \Theta 0} P_{\theta,n}\phi_n + \sup_{\eta \in \Theta_1} P_{\eta,n}(1-\phi_n), \tag{2.43}$$

Clearly there is a stronger, uniform version of consistency too, which incorporates a testing rate quite naturally.

Definition 2.4.12. Given a sequence (a_n), $a_n > 0$, $a_n \to 0$, we say that a test sequence for hypotheses Θ_0, Θ_1 is *uniformly consistent* at *uniform testing rate* a_n, if

$$\sup_{\theta \in \Theta_0} P_{\theta,n}\phi_n + \sup_{\eta \in \Theta_1} P_{\eta,n}(1-\phi_n) = o(a_n).$$

A test sequence is simply *uniformly consistent* if it is uniformly consistent at some rate.

Again, depending on the subsets Θ_0 and Θ_1, there arises the question whether a uniformly consistent test sequence for the pair exists or not. The answer in the case of *i.i.d.* sampling, which is given in part II, characterizes those Θ_0, Θ_1 that can be tested consistently, as precisely those subsets that can be *separated uniformly* (where the relevant model uniformity requires further discussion).

For fixed $n \geq 1$, one wonders about the existence of a test of optimal uniform testing power, *i.e.* a test function $\phi_n : \mathscr{X}_n \to [0,1]$ such that,

$$\sup_{\theta \in \Theta_0} P_{\theta,n}\phi_n + \sup_{\eta \in \Theta_1} P_{\eta,n}(1-\phi_n) = \inf_{\psi} \Big(\sup_{\theta \in \Theta_0} P_{\theta,n}\psi + \sup_{\eta \in \Theta_1} P_{\eta,n}(1-\psi) \Big), \tag{2.44}$$

where the infimum runs over all measurable $\psi : \mathscr{X}_n \to [0,1]$. In that case, the test (ϕ_n) is said to be *minimax optimal* (and the test sequence is said to be *minimax optimal* if this holds for every $n \geq 1$). The existence of such tests under certain convexity, continuity and compactness conditions is a consequence of the so-called *minimax theorem* (see theorem 2.5.6 in the next section). (The following construction is due to Le Cam and is discussed in more detail in [179], section 16.4.)

Lemma 2.4.13. *(Minimax Hellinger tests)*
Let $\mathscr{P}$ be a model for data $Y \in \mathscr{Y}$ and let $\mathscr{P}', \mathscr{P}'' \subset \mathscr{P}$ be model subsets with convex hulls C' and C'' in $\mathscr{M}^1(\mathscr{Y}, \mathscr{B})$, separated by non-zero Hellinger distance:

$$H(C', C'') = \inf_{P \in C', Q \in C''} H(P,Q) > 0. \tag{2.45}$$

Then there exists a test function $\phi : \mathscr{Y} \to [0,1]$ such that,

$$\sup_{P \in \mathscr{P}'} P\phi + \sup_{Q \in \mathscr{P}''} Q(1-\phi) \leq 1 - H^2(C', C''),$$

called the minimax Hellinger tests for $\mathscr{P}_0$ versus $\mathscr{P}_1$.

Proof. For this proof, we refer to theorem 2.5.6. Define $\Theta = C' \times C''$ and let Δ be the set of all measurable $\psi : \mathscr{Y} \to [0,1]$, which form a convex subset of the space F of all measurable $f : \mathscr{Y} \to \mathbb{R}$. Let E denote the vector space of all finite, signed measures on $(\mathscr{Y}, \mathscr{B})$ and note that Θ is a convex subset of the vector space G of all finite, signed measures on $\mathscr{Y} \times \mathscr{Y}$. The bi-linear form $B(P, \psi) = P\psi$ places E and F in *dual correspondence* (see definition C.7.4). Note that Δ is weakly bounded and closed (due to the bi-polar theorem, theorem C.7.7), in the weakly complete space F, so according to proposition C.7.5, Δ is compact (compare with the Banach-Alaoglu theorem of Banach space theory). Consider the risk function,

$$R((P,Q),\psi) = P\psi + Q(1-\psi),$$

defined for all $(P,Q) \in C' \times C''$ and test functions $\psi \in \Delta$, which is concave in (P,Q) and convex in ψ. The risk function depends on $\psi \in \Delta$ in a $\sigma(F,E)$-continuous way. According to the minimax theorem, there exists a test function ϕ such that,

$$\begin{aligned} &\sup_{P \in C'} P\phi + \sup_{Q \in C''} Q(1-\phi) \\ &\qquad = \inf_{\psi \in \Delta} \Big(\sup_{P \in C'} P\psi + \sup_{Q \in C''} Q(1-\psi) \Big) = \sup_{P \in C', Q \in C''} \inf_{\psi \in \Delta} \big(P\psi + Q(1-\psi) \big). \end{aligned}$$

The right-hand side permits (P,Q)-dependent choices for the test functions ψ, enabling the (sharp) Neyman-Pearson-type bound of example 2.4.9:

$$\sup_{P \in C'} P\phi + \sup_{Q \in C''} Q(1-\phi) \leq \sup_{P \in C', Q \in C''} (1 - H^2(P,Q)).$$

Minimax tests separating two disjoint Hellinger balls will play a prominent role in the posterior estimation theorems of chapter 6. For that reason we develop the above bound a bit further in the special case of *i.i.d.* data.

Corollary 2.4.14. *Let $X^n = (X_1, \ldots, X_n)$ be an i.i.d. sample, $X^n \sim P^n$ for some single-observation distribution P in a model $\mathscr{P}$. Let $\mathscr{P}', \mathscr{P}''$ be two subsets of $\mathscr{P}$, with convex hulls C' and C'' in $\mathscr{M}^1(\mathscr{Y}, \mathscr{B})$ that are separated in Hellinger distance. Then there exists a test function $\phi_n : \mathscr{X}_n \to [0,1]$ such that,*

$$\sup_{P \in \mathscr{P}'} P^n\phi_n + \sup_{Q \in \mathscr{P}''} Q^n(1-\phi_n) \leq \big(1 - H^2(C', C'')\big)^n.$$

Proof. The proof revolves around an argument that shows that the convexity restrictions for the product space in which P^n and Q^n live, lead to a suitable factorization of the Hellinger bound. The details can be found in [179], section 16.4, particularly Lemma 2.

Note that any Hellinger ball in $\mathscr{M}^1(\mathscr{X}, \mathscr{B})$ is convex, so with *i.i.d.* data and hypotheses $\mathscr{P}_0$ and $\mathscr{P}_1$ that fit inside two disjoint Hellinger balls at non-zero Hellinger distance from each other, a uniformly consistent test sequence of exponential uniform testing rate (note that $(1-h^2)^n \leq e^{-nh^2}$ for all $h > 0$). Indeed, since corol-

lary 2.4.14 is formulated for each fixed value of $n \geq 1$, we may use the same construction for n-dependent hypotheses $\mathscr{P}_{0,n}$ and $\mathscr{P}_{1,n}$: as long as the Hellinger distances between the corresponding convex hulls $H(C_n', C_n'')$ decreases to zero more slowly than $n^{-1/2}$, there exists a uniformly consistent test sequence.

2.4.4 Posterior odds and Bayes factors

Bayesian hypothesis testing treats null and alternative hypotheses *symmetrically*. This poses an immediate conceptual difference with the most common frequentist methods (*e.g.* the Neyman-Pearson procedure) of hypothesis testing. It also leaves a lot of room for great philosophical disagreement between frequentist and Bayesian views, in which neither side leaves room for the conceptual starting points of the other. Therefore any direct comparison between Bayesian and frequentist testing is difficult (see, however, [12]). In a frequentist analysis of Bayesian testing methods, true comparison is only possible with symmetric forms of frequentist testing.

In the Bayesian perspective, the subsets Θ_0 and Θ_1 of the parameter space have posterior and prior probabilities which are used directly to formulate the test: based on the proportions between those probabilities, we shall decide which hypothesis is the preferred one, based on the following definitions.

Definition 2.4.15. Let $(\Theta, \mathscr{G})$ a measurable space parametrizing a model $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$ for data $Y \in \mathscr{Y}$, with prior $\Pi : \mathscr{G} \to [0,1]$. Let $\{\Theta_0, \Theta_1\}$ be a measurable partition of Θ such that $\Pi(\Theta_0) > 0$ and $\Pi(\Theta_1) > 0$. The *prior* and *posterior odds ratios* in favour of Θ_0 are defined by $\Pi(\Theta_0)/\Pi(\Theta_1)$ and $\Pi(\Theta_0|Y)/\Pi(\Theta_1|Y)$ respectively. The *Bayes factor* in favour of Θ_0 is defined to be

$$B = \frac{\Pi(\Theta_0|Y)}{\Pi(\Theta_1|Y)} \frac{\Pi(\Theta_1)}{\Pi(\Theta_0)}.$$

When doing Bayesian hypothesis testing, we have a choice of which ratio to use and that choice will correspond directly with a choice for subjectivist or objectivist philosophies. In the subjectivist's view, the posterior odds ratio has a clear interpretation: if

$$\frac{\Pi(\Theta_0|Y)}{\Pi(\Theta_1|Y)} > 1,$$

then the probability of $\vartheta \in \Theta_0$ is greater than the probability of $\vartheta \in \Theta_1$. Hence, if the posterior odds ratio exceeds one the subjectivist adopts H_0 rather than H_1; if, on the other hand, the posterior odds ratio lies below one, then the subjectivist accepts H_1 and rejects H_0. The objectivist would object to this practice, saying that the relative prior weights of Θ_0 and Θ_1 can introduce a heavy bias in favour of one or the other in this approach (upon which the subjectivist would answer that that is exactly what he had in mind). The objectivist would prefer to use a criterion that is less dependent on the prior weights of Θ_0 and Θ_1. We look at a very simple example to illustrate the point.

Example 2.4.16. Let Θ be a parameter space that consists of only two points, θ_0 and θ_1 and let $\Theta_0 = \{\theta_0\}$, $\Theta_1 = \{\theta_1\}$, corresponding to simple null and alternative hypotheses H_0, H_1. Denote the prior by Π and assume that both $\Pi(\{\theta_0\}) > 0$ and $\Pi(\{\theta_1\}) > 0$. By Bayes's rule, the posterior weights of Θ_0 and Θ_1 are

$$\Pi(\vartheta \in \Theta_i | Y) = \frac{p_{\theta_i}(Y)\Pi(\Theta_i)}{p_{\theta_0}(Y)\Pi(\Theta_0) + p_{\theta_1}(Y)\Pi(\Theta_1)},$$

for $i = 0, 1$. Therefore, the posterior odds ratio takes the form:

$$\frac{\Pi(\vartheta \in \Theta_0 | Y)}{\Pi(\vartheta \in \Theta_1 | Y)} = \frac{p_{\theta_0}(Y)\Pi(\Theta_0)}{p_{\theta_1}(Y)\Pi(\Theta_1)},$$

and the Bayes factor equals the likelihood ratio:

$$B = \frac{p_{\theta_0}(Y)}{p_{\theta_1}(Y)}.$$

The objectivist prefers the Bayes factor to make a choice between two hypotheses: if $B > 1$ the objectivist adopts H_0 rather than H_1; if, on the other hand, $B < 1$, then the objectivist adopts H_1 rather than H_0. Note that the choice that results from this objective Bayesian testing procedure is identical to choice one makes based on the symmetric likelihood-ratio procedure of example 2.4.9.

We see that the Bayes factor does not depend on the prior weights of Θ_0 and Θ_1 but the posterior odds ratio does. Indeed, suppose we stack the prior odds heavily in favour of Θ_0, by choosing $\Pi(\Theta_0) = 1 - \varepsilon$ and $\Pi(\Theta_1) = \varepsilon$ (for some small $\varepsilon > 0$). Even if the likelihood ratio $p_{\theta_0}(Y)/p_{\theta_1}(Y)$ is much smaller than one (but greater than $\varepsilon/1-\varepsilon$), the subjectivist's criterion favours H_0. In that case, the data clearly advocates hypothesis H_1 but the prior odds force adoption of H_0.

In example 2.4.16 the Bayes factor is independent of the choice of the prior. In general, the Bayes factor is not completely independent of the prior, but it does not depend on the relative prior weights of Θ_0 and Θ_1.

Lemma 2.4.17. *Let $(\Theta, \mathscr{G})$ a measurable space parametrizing a model $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$ for data $Y \in \mathscr{Y}$, with prior $\Pi : \mathscr{G} \to [0,1]$. Let $\{\Theta_0, \Theta_1\}$ be a partition of Θ such that $\Pi(\Theta_0) > 0$ and $\Pi(\Theta_1) > 0$. Then the Bayes factor B in favour of Θ_0 does not depend on the prior odds ratio.*

Proof. For any prior such that $\Pi(\Theta_0) > 0$ and $\Pi(\Theta_1) > 0$,

$$\Pi(A) = \Pi(A|\Theta_0)\,\Pi(\Theta_0) + \Pi(A|\Theta_1)\,\Pi(\Theta_1), \tag{2.46}$$

for all $A \in \mathscr{G}$. In other words, Π is decomposed as a convex combination of two probability measures on Θ_0 and Θ_1 respectively. The Bayes factor is then rewritten (see (2.4)):

$$B = \frac{\Pi(\Theta_0|Y)}{\Pi(\Theta_1|Y)}\frac{\Pi(\Theta_1)}{\Pi(\Theta_0)} = \frac{\Pi(Y|\Theta_0)}{\Pi(Y|\Theta_1)},$$

where, in a dominated model,

$$\Pi(Y|\Theta_i) = \int_{\Theta_i} p_\theta(Y)\, d\Pi(\theta|\Theta_i),$$

for $i = 0, 1$. In terms of the decomposition (2.46), B depends on $\Pi(\cdot|\Theta_0)$ and $\Pi(\cdot|\Theta_1)$, but not on $\Pi(\Theta_0)$ and $\Pi(\Theta_1)$.

So the difference between Bayes factors and posterior odds is exactly the bias introduced by non-zero prior odds; as such, it represents directly the difference between objectivist and subjectivist Bayesian philosophies.

Example 2.4.18. Consider data $X^n = (X_1, \ldots, X_n)$ that form an *i.i.d.* sample from a uniform distribution $U[\theta, \theta + 1]$, with $\theta \in \Theta = [-1, 1]$. We formulate hypotheses,

$$H_0 : \theta \geq 0, \quad H_1 : \theta < 0.$$

and, to show how prior odds influence posterior odds but not Bayes factors, we use a prior with a Lebesgue density of the form,

$$\pi(\theta) = \lambda\, 1\{\theta < 0\} + (1 - \lambda)\, 1\{\theta \geq 0\},$$

for some $0 < \lambda < 1$ (where it is noted that $\lambda = 0$ or $\lambda = 1$ would *not* be valid choices). Consequently, the prior odds in favour of Θ_0 are $1 - 1/\lambda$. The likelihood is given by,

$$p_\theta(X_1, \ldots, X_n) = \prod_{i=1}^{n} 1\{\theta \leq X_i \leq \theta + 1\},$$

and the posterior density (relative to the Lebesgue measure on $\Theta = [-1, 1]$) is proportional to,

$$\begin{aligned} \pi(\theta|X_1, \ldots, X_n) \propto\ & \lambda\, 1\{\theta < 0\}\, 1\{\theta \leq X_{(1)}\}\, 1\{X_{(n)} \leq \theta + 1\} \\ & + (1 - \lambda) 1\{\theta \geq 0\}\, 1\{\theta \leq X_{(1)}\}\, 1\{X_{(n)} \leq \theta + 1\}, \end{aligned}$$

where $X_{(1)}$ and $X_{(n)}$ denote first and last order statistics of the sample respectively. To calculate the posterior odds we do not need the normalization factor in the posterior and we see immediately that,

$$\frac{\Pi\big(\theta \geq 0 \,\big|\, X_1, \ldots, X_n\big)}{\Pi\big(\theta < 0 \,\big|\, X_1, \ldots, X_n\big)} = \frac{1 - \lambda}{\lambda} \frac{\int_0^1 1\{X_{(n)} - 1 \leq \theta \leq X_{(1)}\}\, d\theta}{\int_{-1}^0 1\{X_{(n)} - 1 \leq \theta \leq X_{(1)}\}\, d\theta}$$

Note the proportionality to the prior odds: the Bayes factor B is equal to only the latter fraction in the expression on the right-hand side of the above display and is insensitive to the subjective choice for λ.

To conclude this section we make the following important remark.

Remark 2.4.19. The condition that both Θ_0 and Θ_1 receive prior mass strictly above zero is important since Bayes factors and odds ratios are based on conditioning of

ϑ. Bayesian hypothesis testing is sensible *only* if both Θ_0 and Θ_1 receive non-zero prior mass. This remark plays a role particularly when comparing a *simple* null hypothesis to an alternative, as illustrated in exercise 2.6.19.

2.5 Decision theory and classification

Many practical problems require that we make an observation and based on the outcome, make a decision of some kind: when treating patients, diagnostic variables lead to diagnoses; in financial markets, analysis of data leads to decisions that aim to optimize positioning; *etcetera*. In this section, we look at problems of this nature, first from a frequentist perspective and then with the Bayesian approach.

Practical problems like those described above involve optimality criteria prescribed by the *context* of the problem. For example, any statistical procedure meant to assist in medical diagnosis, should reflect that the misdiagnosis of a serious illness has far more serious consequences than that of a case of the cold. The methods of *statistical inference* that we have discussed thus far concentrate only on the stochastic description of the observations: the accuracy of an estimation procedure, coverage probabilities for confidence intervals or the probability of Type-I and type-II errors in testing procedures. By contrast, *statistical decision theory* formalizes optimality of decision-taking in terms of the contextual consequences of (right or wrong) decisions.

In statistical *decision theory* the nomenclature is slightly different from that introduced earlier. We consider a system that is in an unknown *state* $\theta \in \Theta$, where Θ is called the *state space*. The observation Y still takes its values in a measurable *sample space* $(\mathscr{Y}, \mathscr{B})$ and is still considered stochastic. Its distribution $P_\theta : \mathscr{B} \to [0,1]$ is a function of the state θ of the system. The observation does not reveal the state of the system completely or with certainty. Based on the observation Y, we take a *decision* $a \in \mathscr{A}$ (or perform an *action* a, as some prefer to say), where $\mathscr{A}$ is the called the *decision space*. For each state θ of the system there may be an optimal decision but since observation of Y does not give us the state θ of the system with certainty, the decision is stochastic and may be suboptimal. The goal of statistical decision theory is to arrive at a rule that decides in the best possible way given only the data Y.

If $a \in \mathscr{A}$ is defined as a function of the state θ, the above does not add anything new to the approach we were already following: aside from the names, the concepts introduced here are those used in the usual problem of statistically estimating $a(\theta)$ based on data $Y \sim P_\theta$. What sets decision theory apart is the formal introduction of the decision a and the associated notion of optimality, the loss-function.

Definition 2.5.1. Any bounded function $L : \Theta \times \mathscr{A} \to \mathbb{R}$ is a *loss-function*.

(Technical note: the assumption that losses are bounded is mathematically convenient and hardly poses limitations in applications. Although unbounded loss functions are also of interest, here, we include boundedness in the definition. Additional

properties like measurability, continuity, convexity, *etcetera* are assumed later.) The loss-function has the following interpretation: if a decision a is taken while the state of the system is θ, then a loss $L(\theta,a)$ is incurred. To illustrate, in systems where observation of the state is direct (*i.e.* $Y=\theta$) and non-stochastic, the optimal decision $a(\theta)$ given the state θ is any value of a that minimizes the loss $L(\theta,a)$. The current problem is more difficult because the state θ is unknown and can not be observed directly; all we have is the P_θ-distributed observation Y.

Definition 2.5.2. Let $\mathscr{A}$ be a measurable space with σ-algebra $\mathscr{H}$. A measurable $\delta : \mathscr{Y} \to \mathscr{A}$ is called a *decision rule*.

A decision-rule is a prescribed procedure to arrive at a decision $\delta(y)$, for any possible realisation of the observation $Y=y$. We denote the collection of all decision rules under consideration by Δ. Clearly our goal will be to find decision rules in Δ that "minimize the loss" in an appropriate sense.

Definition 2.5.3. The *risk-function* $R : \Theta \times \Delta \to \mathbb{R}$ is defined as the expected loss under $Y \sim P_\theta$ when using δ,

$$R(\theta,\delta) = \int L(\theta,\delta(Y))\, dP_\theta. \tag{2.47}$$

For any given decision problem, the *risk family* R is the collection of all risk functions,

$$R = \big\{ \theta \mapsto R(\theta,\delta) \,:\, \delta \in \Delta \big\}.$$

The above basic ingredients of decision-theoretic problems play a role in both the frequentist and Bayesian analysis. We consider the frequentist approach first and then look at decision theory from a Bayesian perspective.

2.5.1 Frequentist decision theory

Assuming the perspective of the frequentist, we suppose that $Y \sim P_{\theta_0}$ for some state $\theta_0 \in \Theta$ and would like to assess any decision rule δ according to the risk $P_{\theta_0} L(\theta_0,\delta)$ at θ_0. But θ_0 is unknown, so we are forced to consider *all* values of θ and look at the risk-function.

Definition 2.5.4. Let the state-space Θ, states P_θ, ($\theta \in \Theta$), decision space $\mathscr{A}$ and loss L be given. Choose $\delta_1, \delta_2 \in \Delta$. The decision rule δ_1 is *risk-better* than δ_2, if

$$\forall_{\theta\in\Theta} : \quad R(\theta,\delta_1) \le R(\theta,\delta_2), \tag{2.48}$$

and there exists some $\theta \in \Theta$ for which this inequality is strict. A decision rule $\delta \in \Delta$ is *inadmissible* if there exists a $\delta' \in \Delta$ that is risk-better than δ; a decision rule $\delta \in \Delta$ is *admissible*, if it is not inadmissible.

It is clear that the definition of risk-better decision-rules is intended to order decision rules: if the risk-function associated with a decision-rule is relatively small,

then that decision rule is preferable. Note, however, that the *ordering* we impose by definition 2.5.4 is *partial* rather than *complete*: pairs δ_1, δ_2 of decision rules may exist such that neither δ_1 nor δ_2 is risk-better than the other. This is due to the fact that δ_1 may perform better for values of θ in some $\Theta_1 \subset \Theta$, while δ_2 performs better in $\Theta_2 = \Theta \setminus \Theta_1$, resulting in a situation where neither is risk-better.

Note that admissibility of a decision rule δ does not imply any sort of uniform optimality among rules in Δ. By straightforward logical negation, a decision rule δ is admissible, if for all $\delta' \in \Delta$, there exists a $\theta_0 \in \Theta$ such that $R(\theta_0, \delta) < R(\theta_0, \delta')$, or $R(\theta, \delta) \leq R(\theta, \delta')$ for all $\theta \in \Theta$. That means that, in comparison with any other rule, an admissible rule matches risk everywhere or risk-outperforms in at least one value of the parameter (while it may risk-underperform elsewhere in Θ). This leaves room, for example, for families of risk functions that do not contain *any* pair for which one is risk-better than the other, implying that all associated decision rules are admissible. We come back to admissibility in relation to Bayesian decision theory in subsection 2.5.3.

It is important to find a way to compare risk functions (and thereby decision rules) in a θ-independent way and thus arrive at a complete ordering among decision rules. This motivates the following definition.

Definition 2.5.5. *(Minimax decision principle)* Let the state-space Θ, states P_θ, ($\theta \in \Theta$), decision space $\mathscr{A}$ and loss L be given. The function

$$\Delta \to \mathbb{R} : \delta \mapsto \sup_{\theta \in \Theta} R(\theta, \delta)$$

is called the *minimax risk*. Let $\delta_1, \delta_2 \in \Delta$ be given. The decision rule δ_1 is *minimax-preferred* to δ_2, if

$$\sup_{\theta \in \Theta} R(\theta, \delta_1) \leq \sup_{\theta \in \Theta} R(\theta, \delta_2).$$

If $\delta^M \in \Delta$ minimizes $\delta \mapsto \sup_\theta R(\theta, \delta)$ then δ^M is called a *minimax decision-rule*.

One of the corner stones of decision theory is the so-called *minimax theorem* which guarantees the existence of minimax decision rules under very general conditions.

Theorem 2.5.6. *Let Θ be a convex subset of a vector space and let Δ be a convex compact subset of a locally convex space. Assume that $\delta \mapsto R(\theta, \delta)$ is concave continuous as a map on Δ, for every $\theta \in \Theta$; and that $\theta \mapsto R(\theta, \delta)$ is convex as a map on Θ, for every δ. Then there exists a minimax decision rule $\delta^M \in \Delta$,*

$$\sup_{\theta \in \Theta} R(\theta, \delta^M) = \inf_{\delta \in \Delta} \sup_{\theta \in \Theta} R(\theta, \delta) = \sup_{\theta \in \Theta} \inf_{\delta \in \Delta} R(\theta, \delta). \tag{2.49}$$

Proof. See Sion (1958) [233] and Strasser (1985) [236], p. 239.

Since many loss-functions used in practice satisfy the convexity requirements, the minimax theorem has broad applicability in statistical decision theory and many other fields, particularly econometrics. Note, however, that the minimax theorem holds only for *convex* Δ. In other words, if we want to guarantee the existence of

an minimax-optimal decision rule, we are forced to consider convex combinations of decision rules. Unless $\mathscr{A}$ is a convex set, convex combinations of decision rules have no interpretation, so we adjust the definition of δ slightly.

Definition 2.5.7. Let $(\mathscr{A},\mathscr{H})$ be a measurable space. A map δ that associates a random variable $\delta|Y=y$ taking values in $\mathscr{A}$ with every possible realisation of the data $Y=y$, is called a *randomised decision rule*.

(Technical note: such maps $(A,y)\mapsto P(\delta\in A|Y=y)$ have to be *Markov kernels*, that is, satisfy requirements 1 and 2 of definition B.4.5. See also the notion of a *transition*, as used in [175, 179].) The decision procedure is adapted by randomization: having seen $Y=y$ realised, we draw a random point in $\mathscr{A}$ from the distribution of $\delta|Y=y$, and accordingly, the *risk function* for a randomized decision rule is defined as,

$$R(\theta,\delta)=\int_{\mathscr{Y}}\int_{\mathscr{A}}L(\theta,\delta)\,dP(\delta|Y=y)\,dP_\theta(y) \tag{2.50}$$

We define the *maximal risk family* for a given decision problem as the collection of risk functions associated with *all* randomized decision rules. Since spaces of Markov kernels are convex (verify that definition B.4.5 defines a convex set), the maximal risk family is a convex set of risk functions defined on Θ. In the maximal family of risk functions (or, in any convex family) the minimax theorem formulates conditions that imply the existence of minimax-optimal *randomized* decision rules.

Example 2.5.8. To be able to use the minimax theorem, we assume that Θ is convex in a locally convex space (for example, Θ could be (a convex subset of) the convex set of all probability density functions p in the normed space of all (signed) Lebesgue densities on $[0,1]$) and we assume that the *loss function is convex*. The most technical issue is a suitable choice for the topology on the space of Markov kernels, for which both the compactness and continuity requirements are met. For a loss function that is bounded and continuous in δ and a compact space $\mathscr{A}$, Prokhorov's weak topology on factors of the product space $\Pi\{P(\delta\in\cdot|Y=y):y\in\mathscr{Y}\}$ makes the space of all Markov kernels compact by Tychonov's theorem (after definition C.2.7), while the dependence $P(\delta\in\cdot|Y=\cdot)\mapsto R(\theta,\delta)$ is continuous for every θ. (The product space contains all Markov kernels and the product topology means the following: $P(\delta\in\cdot|Y=\cdot)\to Q(\delta\in\cdot|Y=\cdot)$ whenever $P(\delta\in\cdot|Y=y)$ converges weakly to $Q(\delta\in\cdot|Y=y)$, for every $y\in\mathscr{Y}$. Compare with the topology of pointwise convergence, as after definition C.6.1.) Then the minimax theorem asserts the existence of a possibly randomized minimax decision rule δ^M, such that $\sup_{p\in\Theta}R(p,\delta^M)=\inf_\delta\sup_{p\in\Theta}R(p,\delta^M)$. (See exercise 2.6.24.)

Remark 2.5.9. One important remark concerning the use the minimax decision principle remains: considering (2.49), we see that the minimax principle chooses the decision rule that minimizes the *maximum* of the risk $R(\cdot,\delta)$ over Θ. As such, the minimax criterion takes into account *only* the worst-case scenario and prefers decision rules whose worst case compares well to the worst cases of other decision rules. In practical problems, that means that the minimax principle tends to take a rather pessimistic (or, more neutrally, conservative) perspective on decision problems.

To conclude, we demonstrate that the decision-theoretic approach can also be used to formulate estimation problems in a generalized way, if we choose the decision space $\mathscr{A}$ equal to the state-space Θ.

Example 2.5.10. (Decision theoretic L_2-estimation) Let $Y \sim N(\theta_0, 1)$ for some unknown $\theta_0 \in \Theta$, an (bounded or unbounded) interval in $\mathbb{R}$. Choose $\mathscr{A} = \Theta$ and $L : \Theta \times \Theta \to \mathbb{R}$ equal to the quadratic difference,

$$L(\theta, a) = (\theta - a)^2,$$

a choice referred to as an *L_2-loss*. (If boundedness is a concern, we may always replace any $L \geq 0$ by $L \wedge 1$, without changing the minimization question materially.) Consider the decision-space

$$\Delta = \{\delta_c : \mathscr{Y} \to \mathscr{A} \,:\, \delta_c(y) = c\,y,\ c \geq 0\}.$$

Note that Δ plays the role of a family of estimators for θ_0 here. The risk-function takes the form:

$$\begin{aligned} R(\theta, \delta_c) &= \int L(\theta, \delta_c(Y))\, dP_\theta = \int_{\mathbb{R}} (\theta - cy)^2 dN(\theta, 1)(y) \\ &= \int_{\mathbb{R}} \big(c(\theta - y) + (1-c)\theta\big)^2 dN(\theta, 1)(y) \\ &= \int_{\mathbb{R}} \Big(c^2(y-\theta)^2 + 2c(1-c)\theta(\theta - y) + (1-c)^2\theta^2\Big) dN(\theta, 1)(y) \\ &= c^2 + (1-c)^2\theta^2. \end{aligned}$$

It follows that δ_1 is risk-better than all δ_c for $c > 1$, so that for all $c > 1$, δ_c is inadmissible. But c may lie in $[0, 1)$ as well, and ordering in the uniform sense of (2.48) does not apply to any of the corresponding δ_c. To see this, note that $R(\theta, \delta_1) = 1$ for all θ, whereas for $c < 1$ and some $\theta > 1/(1-c)$, $R(0, \delta_c) < 1 < R(\theta, \delta_c)$. Indeed, for any $0 \leq c_1 < c_2 \leq 1$, δ_{c_1} and δ_{c_2} are incompatible. Therefore, all the decision rules δ_c, $0 \leq c \leq 1$ are admissible.

The minimax criterion does give rise to a preference. However, in order to guarantee its existence, we have to change the original problem because, as things stand, risk functions are unbounded. One way to control the problem is to bound the parameter space: let $M > 0$ be given and assume that $\Theta = [-M, M]$. The minimax risk for δ_c is then given by

$$\sup_{\theta \in \Theta} R(\theta, \delta_c) = c^2 + (1-c)^2 M^2,$$

which is minimal iff $c = M^2/(1+M^2)$, *i.e.* the minimax decision rule for this problem (or, since we are using decision theory to estimate a parameter in this case, the *minimax estimator* in Δ with respect to L_2-loss) is therefore,

$$\delta^M(Y) = \frac{M^2}{1+M^2} Y.$$

Note that if we let $M \to \infty$, this estimator for θ converges to the MLE for said problem.

Remark 2.5.11. Example 2.5.10 also offers opportunity to make the point that admissibility is not sufficient for an estimator to be a 'good' (or even sensible) estimator: consider the same decision problem, with a family of 'decision rules' $\hat{\theta} : \mathscr{Y} \to \Theta$ that contains stochastic estimators like $\hat{\theta}(Y) = Y$ and the estimators above, as well as the deterministic family of all data-independent, fully biased estimators $\hat{\theta}_{\theta_0}(y) = \theta_0$, for all $y \in \mathscr{Y}$ and $\theta_0 \in \Theta$. Note that $L(\theta_0, \hat{\theta}_{\theta_0}(Y)) = 0$ with P_{θ_0}-probability one, so for every $\theta_0 \in \Theta$, $R(\theta_0, \hat{\theta}_{\theta_0}) = 0$. However, $R(\theta_0, \hat{\theta}_{\theta_1})$ for $\theta_1 \neq \theta_0$, as well as $R(\theta_0, \hat{\theta})$ for a stochastic estimators $\hat{\theta}(Y)$ are strictly greater than 0. That implies that all the estimators $\hat{\theta}_{\theta_0}$ are admissible!

2.5.2 Bayesian decision theory

Returning to remark 2.5.9, by comparison, Bayesian decision theory presents a more balanced perspective because instead of maximizing the risk function over Θ, the Bayesian has the prior to integrate over Θ. Optimization of the resulting integral takes into account more than just the worst case, so that the resulting decision rule is based on a less pessimistic perspective than the minimax decision rule.

Definition 2.5.12. Let the state-space Θ, states P_θ, ($\theta \in \Theta$), decision space $\mathscr{A}$ and loss L be given. Additionally, assume that $(\Theta, \mathscr{G})$ is a measurable space with prior $\Pi : \mathscr{G} \to \mathbb{R}$, and that $\theta \mapsto R(\theta, \delta)$ is measurable for every δ. The map r,

$$r(\Pi, \delta) = \int_\Theta R(\theta, \delta)\, d\Pi(\theta), \tag{2.51}$$

is called the *Bayesian risk function*. Let $\delta_1, \delta_2 \in \Delta$ be given. The decision rule δ_1 is *Bayes-preferred* to δ_2, if

$$r(\Pi, \delta_1) \leq r(\Pi, \delta_2).$$

If $\delta^\Pi \in \Delta$ minimizes $\delta \mapsto r(\Pi, \delta)$, *i.e.*

$$r(\Pi, \delta^\Pi) = \inf_{\delta \in \Delta} r(\Pi, \delta). \tag{2.52}$$

then δ^Π is called a *Bayes rule* for the prior Π. The quantity $\inf_\delta r(\Pi, \delta)$ is called the *Bayes risk* (for the prior Π).

The relative pessimism of the minimax decision rule is an expression of the following comparison of the respective criteria.

Proposition 2.5.13. *Let $Y \in \mathscr{Y}$ denote data in a decision theoretic problem with state space Θ, decision space $\mathscr{A}$ and loss $L : \Theta \times \mathscr{A} \to \mathbb{R}$. For any prior Π and all*

$\delta : \mathscr{Y} \to \mathscr{A}$,

$$r(\Pi, \delta) \leq \sup_{\theta \in \Theta} R(\theta, \delta),$$

i.e. any Bayesian risk function is upper bounded by minimax risk.

The proof of this proposition follows from the fact that the minimax risk is an upper bound for the integrand in the Bayesian risk function.

Example 2.5.14. (see example 2.5.10) Let $\Theta = \mathbb{R}$ and $Y \sim N(\theta_0, 1)$ for some unknown $\theta_0 \in \Theta$. Choose the loss-function $L : \Theta \times \mathscr{A} \to \mathbb{R}$ and the decision space Δ as in example 2.5.10. We choose a prior $\Pi = N(0, \tau^2)$ (for some $\tau > 0$) on Θ. Then the Bayesian risk function is given by:

$$\begin{aligned} r(\Pi, \delta_c) &= \int_\Theta R(\theta, \delta_c)\, d\Pi(\theta) = \int_{\mathbb{R}} \left(c^2 + (1-c)^2\theta^2\right) dN(0, \tau^2)(\theta) \\ &= c^2 + (1-c)^2\tau^2, \end{aligned}$$

which is minimal iff $c = \tau^2/(1+\tau^2)$. The (unique) Bayes rule for this problem and corresponding Bayes risk are therefore,

$$\delta^\Pi(Y) = \frac{\tau^2}{1+\tau^2} Y, \qquad r(\Pi, \delta^\Pi) = \frac{\tau^2}{1+\tau^2}.$$

In the Bayesian case, there is no need for a compact parameter space Θ, since we do not maximize but integrate over Θ.

In the above example, we could find the Bayes rule by straightforward optimization of the Bayesian risk function, because the class Δ was rather restricted. If we extend the class Δ to contain *all* non-randomized decision rules, the problem of finding the Bayes rule seems to be far more complicated at first glance. However, as we shall see in theorem 2.5.16, the following definition turns out to be the solution to this question.

Definition 2.5.15. *(The conditional Bayes decision principle)* Let the state-space Θ, states P_θ, ($\theta \in \Theta$), decision space $\mathscr{A}$ and loss L be given. In addition, assume that $(\Theta, \mathscr{G})$ is a measurable space with prior $\Pi : \mathscr{G} \to \mathbb{R}$, and that $\theta \mapsto L(\theta, \delta)$ is measurable for every decision rule δ. We define $\delta^* : \mathscr{Y} \to \mathscr{A}$ to be such that for P^Π-almost-all $y \in \mathscr{Y}$,

$$\int_\Theta L(\theta, \delta^*(y))\, d\Pi(\theta | Y = y) = \inf_{a \in \mathscr{A}} \int_\Theta L(\theta, a)\, d\Pi(\theta | Y = y). \tag{2.53}$$

Pointwise for almost-all y, the decision rule $\delta^*(y)$ is assumed to minimizes the *posterior* expected loss. This defines the decision rule δ^* implicitly as a point-wise minimizer, which raises the usual questions concerning existence and uniqueness, of which little can be said in any generality. However, if existence (and measurability) of δ^* is established, δ^* is Bayes-risk optimal.

Theorem 2.5.16. *Assume that $(\Theta, \mathscr{G})$ and $(\mathscr{A}, \mathscr{H})$ are measurable spaces, with prior $\Pi : \mathscr{G} \to \mathbb{R}$ on Θ and let L be a measurable loss function. If the decision rule $\delta^* : \mathscr{Y} \to \mathscr{A}$ is well-defined and measurable, then δ^* is a Bayes rule.*

Proof. Denote the class of all decision rules for this problem again by Δ. According to theorem 2.1.6 (more particularly, exercise 2.6.7, based on (2.4)) for any measurable decision rule $\delta : \mathscr{Y} \to \mathscr{A}$,

$$\begin{aligned} r(\Pi, \delta) &= \int_\Theta R(\theta, \delta)\, d\Pi(\theta) = \int_\Theta \int_{\mathscr{Y}} L(\theta, \delta(y))\, dP_\theta(y)\, d\Pi(\theta) \\ &= \int_{\mathscr{Y}} \int_\Theta L(\theta, \delta(y))\, d\Pi(\theta|Y = y)\, dP^\Pi(y). \end{aligned}$$

By assumption, the conditional Bayes decision rule δ^* exists. Since δ^* satisfies (2.53) point-wise for all $y \in \mathscr{Y}$, we have

$$\int_\Theta L(\theta, \delta^*(y))\, d\Pi(\theta|Y = y) \leq \inf_{\delta \in \Delta} \int_\Theta L(\theta, \delta(y))\, d\Pi(\theta|Y = y).$$

Substituting this in (2.19), we obtain

$$\begin{aligned} r(\Pi, \delta^*) &\leq \int_{\mathscr{Y}} \inf_{\delta \in \Delta} \int_\Theta L(\theta, \delta(y))\, d\Pi(\theta|Y = y)\, dP^\Pi(y) \\ &\leq \inf_{\delta \in \Delta} \int_{\mathscr{Y}} \int_\Theta L(\theta, \delta(y))\, d\Pi(\theta|Y = y)\, dP^\Pi(y) = \inf_{\delta \in \Delta} r(\Pi, \delta). \end{aligned}$$

which proves that δ^* is a Bayes rule.

It is noted that randomization of the decision is not needed when optimizing with respect to the Bayes risk. The conditional Bayes decision rule is non-randomized and optimal.

2.5.3 Admissibility and the complete class theorem

It is important to understand the relationship between Bayes rules and *admissibility*: below we consider the so-called *complete class theorem* ([249, 250], see also [218]), which roughly says that, for any decision problem, any admissible decision rule δ is a Bayes rule for some prior distribution. The implication is that the frequentist looking for admissible decision rules, needs to consider only those decision rules that minimize posterior expected loss for some prior.

Definition 2.5.17. A subset C of a set Δ of decision functions is a *complete class* if for any $\delta \in \Delta \setminus C$, we can find an element in C that is risk-better than δ. A complete class C is a *minimal complete class*, if it contains no proper subset that is a complete class.

Clearly, admissible decision rules are in any complete class, but more is true.

Proposition 2.5.18. *If a minimal complete class C exists, C is equal to the set of all admissible decision rules.*

Proof. Note that if C is a minimal complete class and $\delta \in C$ is not admissible, then there is a $\delta' \in \Delta$ that is risk-better than δ. Since C is minimal, δ' does not lie in C. Then there is a $\delta'' \in C$ that is risk-better than δ', and by extension, also risk-better than δ. The latter result is not possible because C is minimal. Conclude that C is a subset of the admissible decision rules.

So if the set of all admissible decision rules is complete, it is minimal complete.

For the formulation of the theorem below, we assume that Θ is compact and all risk functions $R(\cdot, \delta)$ are continuous. The associated risk family R is a subset of $C(\Theta)$, which we view as a Banach space relative to the sup-norm. (For the proof below, we assume only the theory of locally convex spaces, and leave aside aspects of the proof that depend on the ordered vector-space structure. (See the introductory paragraph of appendix C.7.))

Theorem 2.5.19. *(Complete class theorem)*
Let the parameter space Θ be compact and let the risk family R be convex. Assume that all risk functions in R are continuous in θ. Then, any admissible decision rule $\delta \in \Delta$ is a Bayes rule for some Borel prior probability measure on Θ.

Proof. Let δ be an admissible (randomized) decision rule. Due to the admissibility of δ, the set of risk functions that are risk-better than $R(\cdot, \delta)$ is empty, In the linearly shifted convex family $R - R(\cdot, \delta)$, the risk function for δ is shifted to zero. The closure of $R - R(\cdot, \delta)$ is a closed convex subset that meets the negative quadrant only in the origin, due to admissibility. According to the Hahn-Banach theorem (in its geometric form, theorem C.7.11) there exists a continuous linear functional I on $C(\Theta)$ such that $I \geq 0$ on the closure of $R - R(\cdot, \delta)$, which implies $I(R(\cdot, \delta)) \leq I(R(\cdot, \delta'))$ for any δ'. The Riesz representation theorem, theorem C.8.2, says that there exists a finite Borel measure μ on Θ such that,

$$I(f) = \int_\Theta f(\theta)\, d\mu(\theta),$$

for any continuous function f on Θ. (This argument does not establish the positivity of μ! For that, a version of the Hahn-Banach theorem specific to Riesz spaces is required, that separates (non-strictly) any individual point in the lower bound of a closed convex set from the rest of that set. For more, see [48], chapter I.) After normalization, $\Pi = \mu / \|\mu\|$ defines a Borel probability measure, such that,

$$r(\Pi, \delta) = \int_\Theta R(\theta, \delta)\, d\Pi(\theta) \leq \int_\Theta R(\theta, \delta')\, d\Pi(\theta) = r(\Pi, \delta')$$

for every δ'. So δ is the Bayes rule for the prior Π.

Regarding the convexity condition for the risk family, recall the example of the *maximal risk family*, which is convex.

Not only are there conditions so that admissible decision rules are Bayes rules, one of these conditions also guarantees that Bayes rules are admissible.

Theorem 2.5.20. *Let Θ be a Hausdorff topological space with a Radon prior probability measure Π of full support. Assume also that for every $\delta \in \Delta$, $\theta \mapsto R(\theta,\delta)$ is continuous and $r(\Pi,\delta) < \infty$. Then any decision rule δ' that is a Bayes rule for Π, is an admissible decision rule.*

Proof. Let δ' be a Bayes rule for a prior Π of full support. Suppose that δ' is not admissible: then there exists a $\delta \in \Delta$ such that, for all $\theta \in \Theta$,

$$R(\theta,\delta) \leq R(\theta,\delta'),$$

and for some parameter value $\theta_0 \in \Theta$ and some $\varepsilon > 0$, $R(\theta_0,\delta) - R(\theta_0,\delta') = -\varepsilon < 0$. By continuity of the risk functions, there exists an open neighbourhood U of θ_0 such that,

$$R(\theta,\delta) - R(\theta,\delta') \leq -\tfrac{1}{2}\varepsilon,$$

for all $\theta \in U$. Since Π has full support, $\Pi(U) > 0$, and,

$$\begin{aligned} r(\Pi,\delta) - r(\Pi,\delta') &= \int_\Theta \big(R(\theta,\delta) - R(\theta,\delta')\big) d\Pi(\theta) \\ &= \int_{\Theta\setminus U} \big(R(\theta,\delta) - R(\theta,\delta')\big) d\Pi(\theta) + \int_U \big(R(\theta,\delta) - R(\theta,\delta')\big) d\Pi(\theta) \\ &\leq -\tfrac{1}{2}\varepsilon\Pi(U) < 0, \end{aligned}$$

and δ' is not a Bayes rule for Π.

Theorems 2.5.19 and 2.5.20 immediately raises two further questions: the first question is whether the compactness and continuity conditions are necessary? The existence of a probability measure Π depends on the Riesz representation and one wonders how much the Riesz-Markov-Kakutani generalization, theorem C.8.7, can add to this, or how this type of theorem can be stretched by sequential versions on σ-compact spaces, for example. Indeed there exist many generalized complete class theorems (see, *e.g.* Robert (2001) [218]), and with the right definition of what a 'generalized' Bayes rule amounts to, admissible decision rules on non-compact parameter spaces can be represented as such. The second question, of course, is under which conditions Bayes rules form a complete class (see Le Cam (1955) [172]).

2.5.4 Frequentist versus Bayesian classification

Many decision-theoretic questions take the form of a *classification* problem: under consideration is a population Ω of objects that each belong to one of a finite number of classes $\mathscr{A} = \{1,2,\ldots,L\}$. The class K of the object is the unknown quantity of interest. Observing a vector Y of its features, the goal is to *classify* the object, *i.e.*

estimate which class it belongs to. We formalize the problem in decision-theoretic terms: the population is a probability space $(\Omega, \mathscr{F}, P)$; both the *feature vector* and the class of the object are random variables, $Y : \Omega \to \mathscr{Y}$ and $K : \Omega \to \mathscr{A}$ respectively. The state-space in a classification problem equals the decision space $\mathscr{A}$: the class can be viewed as a "state" in the sense that the distribution $P_{Y|K=k}$ of Y *given* the class $K = k$ depends on k.

A decision rule (or *classifier*, as it is usually referred to in the context of classification problems) is based on the feature vector Y and classifies the object in class $\delta(Y)$, *i.e.* the classifier maps features to classes by means of a map $\delta : \mathscr{Y} \to \mathscr{A}$. Such a δ can be viewed equivalently as a finite partition of the feature-space $\mathscr{Y}$: for every $k \in \mathscr{A}$, we define

$$\mathscr{Y}_k = \{y \in \mathscr{Y} : \delta(y) = k\}$$

and note that if $k \neq l$, then $\mathscr{Y}_k \cap \mathscr{Y}_l = \varnothing$ and $\mathscr{Y}_1 \cup \mathscr{Y}_2 \cup \ldots \cup \mathscr{Y}_L = \mathscr{Y}$. The partition of the feature space is such that if $Y = y \in \mathscr{Y}_k$ for certain $k \in \mathscr{A}$, then δ classifies the object in class k. Depending on the context of the classification problem, a loss-function $L : \mathscr{A} \times \mathscr{A} \to \mathbb{R}$ is defined (see the examples in the introduction to this section). Without context, the most natural loss function in a classification problem is,

$$L(k,l) = 1_{\{k \neq l\}}.$$

i.e. we incur a loss equal to one for each *misclassification*. Using the minimax decision principle, we look for a classifier $\delta^M : \mathscr{Y} \to \mathscr{A}$ that minimizes:

$$\delta \mapsto \sup_{k \in \mathscr{A}} \int_{\mathscr{Y}} L(k, \delta(y))\, dP(y|K = k) = \sup_{k \in \mathscr{A}} P\big(\delta(Y) \neq K \,\big|\, K = k\big),$$

i.e. the minimax decision principle prescribes that we minimize the probability of misclassification *uniformly over all classes* (consequently focussing *only* on the misclassification probability for the *worst-case value of* $k \in \mathscr{A}$).

In a Bayesian context, we need a prior on the state-space, which equals $\mathscr{A}$ in classification problems. Note that if known (or estimable, as in many practical circumstances), the marginal probability distribution for K is to be used as the prior for the state k, in accordance with definition 2.1.2. In the absence of information on the marginal distribution of K, ignorance can be represented by equal prior weights $1/L$ for all values of K. Here, we assume that the probabilities $P(K = k)$ are known and use them to define the prior density with respect to the counting measure on the (finite) space $\mathscr{A}$:

$$\pi(k) = P(K = k).$$

The Bayes rule $\delta^* : \mathscr{Y} \to \mathscr{A}$ for this classification problem is defined as the minimizer of

$$\delta \mapsto \sum_{k \in \mathscr{A}} L(k, \delta(y))\, d\Pi(k|Y = y) = \sum_{k=1}^{L} \Pi\big(\delta(y) \neq K \,\big|\, Y = y\big)$$

for every $y \in \mathscr{Y}$. According to theorem 2.5.16, the classifier δ^* minimizes the Bayes risk, which in this situation is given by:

$$\begin{aligned} r(\Pi, \delta) &= \sum_{k \in \mathscr{A}} R(k, \delta)\, \pi(k) = \sum_{k \in \mathscr{A}} \int_{\mathscr{Y}} L(k, \delta(y))\, dP(y|K=k)\, \pi(k) \\ &= \sum_{k \in \mathscr{A}} P\big(K \neq \delta(Y) \;\big|\; K = k\big)\, P(K=k) = P\big(K \neq \delta(Y)\big). \end{aligned}$$

Summarizing, the Bayes rule δ^* minimizes the *overall probability of misclassification*, rather than the worst-case that the minimax classifier focusses on. Readers interested in the statistics of classification and its applications are encouraged to read B. Ripley's "Pattern recognition and neural networks" (1996) [219].

To close the chapter, the following remark is in order: when we started our comparison of frequentist and Bayesian methods, we highlighted the conflict of philosophies. But now that we have seen some of the differences in more detail by considering estimation, confidence sets, testing and decision theory in both schools, we can be more specific. Statistical problems can be solved in both schools and often the differences are smaller that one might fear (especially in the large-sample limit, see part II). Whether one chooses for a Bayesian or frequentist statistical method is usually not determined by deeply held philosophical beliefs, but by much more practical considerations. Perhaps the classification example of this subsection illustrates this point most clearly: if one is concerned about correct classification for objects in the most difficult class, one should opt for the minimax decision rule. If, on the other hand, one wants to minimize the overall misclassification probability, one should choose to adopt the conditional Bayes decision rule, with a prior for k that equals the marginal for K (or approximates it well). In other words, depending on the risk to be minimized (minimax risk and Bayes risk are different!) one arrives at different classifiers.

More generally the (subjective) choice of a prior can either form a benefit or a drawback, depending on the needs: on the one hand, frequentist methods can claim a form of objectivity that is appreciated, for example, in most scientific and medical settings. On the other, choice of a prior offers (admittedly subjective) control over a statistical procedures through bias. If well-chosen, such bias can be of great intuitive statistical value, like the subjective bias of example 1.3.1. Bayesian methods are popular in forensic science because the freedom to choose a prior leaves room to incorporate background information and common-sense. From a more technical point of view, bias may be required for regularization purposes (like a penalty in frequentist terms, see remark 2.2.21). Prior bias may even be guided in a data-dependent way, *e.g.* when we employ *empirical Bayesian* methods to optimize a procedure (as in section 3.4).

Another reason to use one or the other may be computational advantages or useful theoretical results that exist for one school but have no analogue in the other. Philosophical preference should not play a role in the choice for a statistical procedure, practicality and usefulness should (and usually do).

2.6 Exercises

2.6.1. CALIBRATION
A physicist prepares for measurement of a physical quantity Z in his laboratory. To that end, he installs a measurement apparatus that will give him an outcome of the form $Y = Z + e$ where e is a measurement error due to the inaccuracy of the apparatus, assumed to be stochastically independent of Z. Note that if the expectation of e equals zero, long-run sample averages converge to the expectation of Z; if the expected error $Pe \neq 0$, on the other hand, averaging does not cancel out the resulting bias. The manufacturer of the apparatus says that e is normally distributed with known variance $\sigma^2 > 0$. The mean θ of this normal distribution depends on the way the apparatus is installed and thus requires calibration. The following questions pertain to the calibration procedure.

The physicist decides to conduct the following steps to calibrate his measurement. First, he makes certain that the apparatus receives no input signal, $Z = 0$. Then he repeats measurement of Y, generating an *i.i.d.* sample of size n, which amounts to an *i.i.d.* sample from the distribution of e used to estimate the unknown mean θ. The physicist expects that the expected error *Pe* lies close to zero.

a. Explain why, from a subjectivist point of view, the choice $\theta \sim N(0, \tau^2)$ forms a suitable prior in this situation. Explain the role of the parameter $\tau^2 > 0$.
b. With the choice of prior as in part *a.*, calculate the posterior density for θ.
c. Interpret the influence of τ^2 on the posterior, taking into account your answer under part *a. Hint: take limits* $\tau^2 \downarrow 0$ *and* $\tau^2 \uparrow \infty$ *in the expression you have found under* b.
d. What is the influence of the sample size n? Show that the particular choice of the constant τ^2 becomes irrelevant in the large-sample limit $n \to \infty$.

2.6.2. Let $X_1, \ldots, X_n$ be an *i.i.d.* sample from the uniform distribution $U[0, \theta]$, with unknown parameter $\theta \in \Theta = (1, \infty)$. As a prior for θ, choose the Pareto distribution with exponent $\alpha > 0$. Calculate the posterior density for θ with respect to the Lebesgue measure on $(1, \infty)$.

2.6.3. Let $X_1, \ldots, X_n$ be an *i.i.d.* sample from the Poisson distribution Poisson(λ), with unknown parameter $\lambda > 0$. As a prior for λ, let $\lambda \sim \Gamma(2, 1)$. Calculate the posterior density for λ with respect to the Lebesgue measure on $[0, \infty)$.

2.6.4. Let $X_1, \ldots, X_n$ be an *i.i.d.* sample from a binomial distribution Bin(k, θ), with known $k \geq 1$ and unknown $\theta \in \Theta = [0, 1]$. As a prior for θ, use a beta distribution, $\theta \sim \beta(2, 2)$. Calculate the posterior density for θ with respect to the Lebesgue measure on $[0, 1]$.

2.6.5. Let $X_1, \ldots, X_n$ be an *i.i.d.* sample from a normal distribution $N(0, \sigma^2)$, with unknown $\sigma^2 > 0$. We define the prior for the variance σ^2 implicitly, by stating that the *inverse* $1/\sigma^2$ is distributed according to a $\Gamma(\alpha, \beta)$ distribution. Calculate the posterior density for σ^2 with respect to the Lebesgue measure on $[0, \infty)$.

2.6.6. Let $(\mathscr{P},\mathscr{F},\Pi)$ be a model with prior for *i.i.d.* $X_1,\ldots,X_n$ taking values in a sample space $(\mathscr{X},\mathscr{B})$. Suppose that the model is dominated by a σ-finite measure μ on $(\mathscr{X},\mathscr{B})$ and that the prior is dominated by a σ-finite measure ν on $(\mathscr{P},\mathscr{G})$. Show that if μ' is another σ-finite measure on $(\mathscr{X},\mathscr{B})$, such that $\mathscr{P} \ll \mu' \ll \mu$, and ν' is another σ-finite measure on $(\mathscr{P},\mathscr{G})$, such that $\Pi \ll \nu' \ll \nu$, then the MAP estimator $\tilde{\theta}_2$ does *not* change with μ' (compare with exercise 1.6.1), but $\tilde{\theta}_2$ *does* change with ν'.

2.6.7. Use theorem 2.1.6 to show that, for any $\mathscr{B}\times\mathscr{G}$-measurable $\varphi : \mathscr{Y}\times\Theta \to [0,\infty]$,

$$\int_{\mathscr{Y}}\int_{\Theta}\varphi(y,\theta)\,d\Pi(\theta|Y=y)\,dP^{\Pi}(y) = \int_{\Theta}\int_{\mathscr{Y}}\varphi(y,\theta)\,dP_\theta(y)\,d\Pi(\theta).$$

Hint: in the case that we have a dominated model, (2.13) provides a more explicit form of the posterior, which makes the following simplified proof possible. Use definitions (2.51) and (2.47), and theorems B.3.10, B.3.9 to arrive at,

$$\begin{aligned}\int_{\Theta}\int_{\mathscr{Y}}\varphi(y,\theta)\,dP_\theta(y)\,d\Pi(\theta) &= \int_{\mathscr{Y}}\int_{\Theta}\varphi(y,\theta)\,p_\theta(y)\,d\Pi(\theta)\,d\mu(y)\\ &= \int_{\mathscr{Y}}\Big(\int_{\Theta}p_\theta(y)\,d\Pi(\theta)\Big)\int_{\Theta}\varphi(y,\theta)\,d\Pi(\theta|Y=y)\,d\mu(y),\end{aligned}$$

With (2.3), rewrite the expression,

$$\int_{\Theta}\int_{\mathscr{Y}}\varphi(y,\theta)\,dP_\theta(y)\,d\Pi(\theta) = \int_{\mathscr{Y}}\int_{\Theta}\varphi(y,\theta)\,d\Pi(\theta|Y=y)\,dP^{\Pi}(y).$$

In the non-dominated case, rely on a monotone sequence of simple approximations for the $\sigma(\mathscr{B}\times\mathscr{G})$-measurable φ, monotone convergence of integrals and approximation in Π^*-measure of $A\in\sigma(\mathscr{B}\times\mathscr{G})$ by finite unions of (disjoint) rectangles to use Bayes's Rule (2.4) directly.

2.6.8. In the model of exercise 2.6.2, calculate the maximum-likelihood estimator, the posterior mean and the maximum-a-posteriori estimator.

2.6.9. In the model of exercise 2.6.3, calculate the maximum-likelihood estimator, the posterior mean and the maximum-a-posteriori estimator.

2.6.10. In the model of exercise 2.6.4, calculate the maximum-likelihood estimator, the posterior mean and the maximum-a-posteriori estimator.

2.6.11. In the model of exercise 2.6.5, calculate the maximum-likelihood estimator, the posterior mean and the maximum-a-posteriori estimator.

2.6.12. Consider the following questions in the context of exercise 2.6.3, after exercise 2.6.9.

a. Let $n\to\infty$ both in the MLE and MAP estimator and conclude that the difference vanishes in the limit, P_λ-almost-surely.

b. Following remark 2.2.21, explain the difference between ML and MAP estimators exclusively in terms of the prior.
c. Consider and discuss the choice of prior $\lambda \sim \Gamma(2,1)$ twice, once in a qualitative, subjectivist Bayesian fashion, and once following the frequentist interpretation of the log-prior-density.

2.6.13. Let $Y \sim P_0$ denote the data and $\mathscr{P}$ a model with metric d. Suppose that $\mathscr{P}$ is endowed with a prior defined on the Borel σ-algebra induced by the metric topology. Assume that $P_0 \ll P^{\Pi}$ and that $\mathscr{P}$ is compact. The following questions pertain to the small-ball estimators defined in definition 2.2.18 and remark 2.2.19. We assume that the posterior distribution is such that for all $\varepsilon > 0$ and all $P \in \mathscr{P}$, the (topological) boundary of the ball $B_d(P,\varepsilon)$ receives mass equal to zero: $\Pi(\partial B_d(P,\varepsilon)|Y) = 0$, P_0-almost surely.

a. Show that, for any $p \in (1/2,1)$ and large enough $\varepsilon > 0$, the small-ball estimator $\hat{P}$ of exists, P_0-almost-surely.
b. Show that for any two measurable model subsets $A, B \subset \mathscr{P}$,

$$\big|\,\Pi(A\,|\,Y\,) - \Pi(B\,|\,Y\,)\,\big| \leq \Pi(A \cup B\,|\,Y\,) - \Pi(A \cap B\,|\,Y\,),$$

P_0-almost-surely.
c. Show that for every $\varepsilon > 0$, the map $P \mapsto \Pi(\,B_d(P,\varepsilon)\,|\,Y\,)$ is continuous, P_0-almost-surely.
d. Show that for every $\varepsilon > 0$, the small-ball estimator of definition 2.2.18 exists.
e. Let some $p \in (1/2,1)$ be given. Suppose that $\varepsilon > 0$ denotes some radius for which there exists a ball $B_d(P,\varepsilon) \subset \mathscr{P}$ of posterior probability greater than or equal to p. Show that, if both $\hat{P}_1$ and $\hat{P}_2$ are centre points of such balls, then $d(\hat{P}_1, \hat{P}_2) < 2\varepsilon$, P_0-almost-surely.

2.6.14. Let $X^n = (X_1, \ldots, X_n)$ be an *i.i.d.* sample from the normal distribution $N(\mu, \sigma^2)$ for certain $\mu \in \mathbb{R}$, $\sigma^2 > 0$. Show that the sample average is distributed according to the normal distribution,

$$\mathbb{P}_n X \sim N(\mu, \sigma_n^2),$$

with variance $\sigma_n^2 = \sigma^2/n$.

2.6.15. Let Y be normally distributed with known variance $\sigma^2 > 0$ and unknown location θ. As a prior for θ, choose $\Pi = N(0,\tau^2)$. Let $\alpha \in (0,1)$ be given. Using the posterior density with respect to the Lebesgue measure, express the level-α HPD credible set in terms of Y, σ^2, τ^2 and quantiles of the standard normal distribution. Consider the limit $\tau^2 \to \infty$ and compare with level-α confidence intervals centred on the ML estimate for θ.

2.6.16. Let $Y \sim \mathrm{Bin}(n;p)$ for known $n \geq 1$ and unknown $p \in (0,1)$. As a prior for p, choose $\Pi = \mathrm{Beta}(\frac{1}{2},\frac{1}{2})$. Calculate the posterior distribution for the parameter p. Using the Lebesgue measure on $(0,1)$ to define the posterior density, give a level-α credible interval $D_\alpha(Y)$ for p, using quantiles of beta-distributions. Give an equation that characterizes the choice of quantiles for which $D_\alpha(Y)$ is an HPD credible set.

2.6.17. Let $(\Theta, \mathscr{G})$ a measurable space parametrizing a model $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$ for data $Y \in \mathscr{Y}$, with prior Π. Assume that the posterior is regular and dominated by a σ-finite measure μ on $(\Theta, \mathscr{G})$, with density $\pi(\cdot|Y) : \Theta \to [0, \infty)$. Show that if the HPD credible set $D_\alpha(Y)$ satisfies $\Pi(D_\alpha(Y)|Y) = 1 - \alpha$, then $D_\alpha(Y)$ has minimal μ-measure among all credible sets of level α, almost-surely.

2.6.18. Let Θ be a subset of $\mathbb{R}$ and let $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$ describe an identifiable parametrization of the model $\mathscr{P}$ for an *i.i.d.* sample $X_1, \ldots, X_n$, and assume that there exists a $\theta_0 \in \Theta$ such that P_{θ_0} is the marginal distribution for each of the X_i. Let θ and θ' with $\theta' > \theta$ from Θ be given and consider the hypotheses,

$$H_0 : \quad \theta_0 = \theta, \qquad H_1 : \theta_0 = \theta',$$

Given a significance level $\alpha \in (0, 1)$, write down the Neyman-Pearson test for H_0 versus H_1 (see lemma 2.4.5), in each of the following cases,

a. for all $\theta \in [0, 1]$, $P_\theta = \text{Bernoulli}(\theta)$;
b. for all $\theta \in (0, \infty)$, $P_\theta = \text{Poisson}(\theta)$;
c. for all $\theta \in [0, 1]$, $P_\theta = \text{Bin}(\theta, k)$ for some known integer $k \geq 1$.

2.6.19. Consider a dominated model $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ for data Y, where $\Theta \subset \mathbb{R}$ is an interval. For certain $\theta_0 \in \Theta$, consider the simple null-hypothesis and alternative:

$$H_0 : \quad \theta = \theta_0, \qquad H_1 \quad : \theta \neq \theta_0.$$

Show that if the prior Π is absolutely continuous with respect to the Lebesgue measure on Θ, then the posterior odds ratio in favour of the hypothesis H_0 equals zero, almost surely.
[Remark: conclude that calculation of posterior odds ratios makes sense only *if both hypotheses receive non-zero prior mass. Otherwise, the statistical question we ask is rendered invalid* ex ante *by our beliefs concerning θ, as expressed through the choice of the prior. (See example 2.1.13.)]*

2.6.20. Let $X_1, \ldots, X_n$ be an *i.i.d.* sample from a binomial distribution $\text{Bin}(\theta, k)$, for some known integer $k \geq 1$ and an unknown parameter $\theta \in \Theta = [0, 1]$. Let the prior Π for θ be a Beta distribution $\text{Beta}(\alpha, \beta)$, with certain parameters $\alpha, \beta > 1$.

a. Calculate the posterior distribution for θ.
b. Write down the equations that determine the two end-points of the HPD credible interval (based on the density of the posterior relative to Lebesgue measure on Θ), for given credible level $\delta \in (0, 1)$.

Consider the hypotheses,

$$H_0 : \quad \theta \leq \tfrac{1}{2}, \qquad H_1 : \quad \theta > \tfrac{1}{2}.$$

c. Give the prior odds, posterior odds and Bayes factor for the hypotheses H_0 and H_1.

2.6.21. PRISONER'S DILEMMA
Two men have been arrested on the suspicion of burglary and are held in separate cells awaiting interrogation. The prisoners have been told that burglary carries a maximum sentence of x years. However, for any prisoner who confesses, the sentence is reduced to y years (where $0 < y < x$).

Guilty of the crime he is accused of, our prisoner contemplates whether to confess to receive a lower sentence, or to deny involvement in the hope of escaping justice. If he keeps his mouth shut and so does his partner in crime, they will both walk away free. If he keeps his mouth shut but his partner talks, he gets the maximum sentence. If he talks, he will always receive a sentence of y years and the other prisoner receives y or x years depending on whether he confessed or not himself. To talk or not to talk, that is the question.

There is no data in this problem, so we set θ equal to 1 or 0, depending on whether the other prisoner talks or not. Our prisoner can decide to talk ($t = 1$) or not ($t = 0$). The loss function $L(\theta, t)$ equals the prison term for our prisoner. In the absence of data, risk and loss are equal.

a. Calculate the minimax risk for both $t = 0$ and $t = 1$. Argue that the minimax-optimal decision for our prisoner is to confess.

As argued in section 2.5, the minimax decision can be overly pessimistic. In the above, it assumes that the other prisoner will talk and chooses t accordingly.

The Bayesian perspective balances matters depending on the chance that the other prisoner will confess when interrogated. This chance finds its way into the formalism as a prior for the trustworthiness of the other prisoner. Let $p \in [0,1]$ be the probability that the other prisoner confesses, *i.e.* $\Pi(\theta = 1) = p$ and $\Pi(\theta = 0) = 1 - p$.

b. Calculate the Bayesian risks for $t = 0$ and $t = 1$ in terms of x, y and p. Argue that the Bayes rule for our prisoner is as follows: if $y/x > p$ then our prisoner does not confess, if $y/x < p$, the prisoner confesses. If $y/x = p$, the Bayes decision criterion does not have a preference.

So, depending on the degree to which our prisoner trusts his associate and the ratio of prison terms, the Bayesian draws his conclusion. The latter is certainly more sophisticated and perhaps more realistic, but it requires that our prisoner quantifies his trust in his partner in the form of a prior distribution.

2.6.22. Consider a classification problem based on a probability space $(\Omega, \mathscr{F}, P)$, measurable feature vector $Y : \Omega \to \mathscr{Y}$ and class $K : \Omega \to \mathscr{K}$, where both $\mathscr{Y}$ and $\mathscr{K}$ are Polish spaces. Assume that $\mathscr{A} = \mathscr{K}$ and the loss function $L : \mathscr{K} \times \mathscr{K} \to \mathbb{R}$ is such that $L(k,l) \geq 0$ with equality if $k = l$.

a. Show that if the σ-algebra generated by K is contained in the σ-algebra generated by Y, there exists a non-randomized classifier $\delta : \mathscr{Y} \to \mathscr{K}$ with the following property: if we define model distributions P_k as conditional distributions for Y given $K = k$ (*c.f.* definition 2.1.1), then the risk function $k \mapsto R(k, \delta)$ is almost-surely minimal (that is: $R(k, \delta) = 0$, for all k in a measurable subset $F \subset \mathscr{K}$ such that $P(K(\omega) \in F) = 1$.

Interpret as follows: if 'Y contains all information about K', we can reconstruct K from Y, making perfect classification possible.

This point can be extended, because for classification we adopt an essentially Bayesian view of the class/state/parameter k in the problem, when we assume that $K \in \mathscr{K}$ has a distribution (acting as a prior for k). Below, we assume that $\mathscr{K}$ is $\mathbb{R}$ with K quadratically integrable, and we specify to Bayesian risk functions with quadratic loss.

b. Show that if $L(k,l) = (k-l)^2$, then the conditional expectation $E_P[K|Y] : \mathscr{Y} \to \mathbb{R}$ of K given Y defines a classifier that minimizes the Bayesian risk function. *Hint: $E_P[K|Y]$ is the orthogonal projection of K onto the closed subspace of square-integrable, Y-measurable functions $L^2(\Omega, \sigma(Y), P)$ in $L^2(\Omega, \mathscr{F}, P)$.*

The conditional expectation $E_P[K|Y]$ represents the closest (L^2-)approximation of K we can make with a function that can only depend on Y, representing another perspective on the 'information that Y contains about K'.

c. Show explicitly that with quadratic loss like above, $E_P[K|Y] : \mathscr{Y} \to \mathbb{R}$ coincides with the conditional Bayes decision rule δ^* of definition 2.5.15.

2.6.23. Consider a decision problem in which we have a state space $\Theta = \{\theta_0, \theta_1\}$ and a decision space $\mathscr{A} = \{a_1, a_2, a_3, a_4, a_5\}$, with the following loss function:

$$L(\theta_0, a_1) = 0,\ L(\theta_0, a_2) = 3,\ L(\theta_0, a_3) = 1,\ L(\theta_0, a_4) = 3,\ L(\theta_0, a_5) = 4,$$
$$L(\theta_1, a_1) = 4,\ L(\theta_1, a_2) = 4,\ L(\theta_1, a_3) = 0,\ L(\theta_1, a_4) = 0,\ L(\theta_1, a_5) = 1,$$

a. Consider a prior such that $\Pi(\theta_0) = 4/5$ and $\Pi(\theta_1) = 1/5$. Although there is no data in this problem, there is still the question which decisions minimize the Bayesian risk function under this prior. Find all Bayes rules.

2.6.24. Make the reasoning of example 2.5.8 precise and formulate an existence theorem for a minimax decision rule. *Hint: find a topology (using definition C.7.14) such that for every y, the space of all distributions for $\delta|Y = y$ is compact; use Tychonov's theorem to conclude that the product is compact; show that the space of all Markov kernels is closed in the product space; show that the map taking Markov kernels into risk functions is continuous.*

Some example exam problems

2.6.25. Consider an experiment in which we observe a single X distributed according to a binomial distribution $P_{n,p} = \mathrm{Bin}(n,p)$. We assume that n is known, and for the unknown parameter $p \in [0,1]$, we have three possible prior choices,

$$\Pi_0 = \mathrm{Beta}(\tfrac{1}{2}, \tfrac{1}{2}), \quad \Pi_1 = U[0,1], \quad \Pi_2 = \mathrm{Beta}(0,0).$$

a. With the three priors Π_0, Π_1, Π_2, calculate the three posteriors. Give the associated posterior means $\tilde{p}_0(X)$, $\tilde{p}_1(X)$, $\tilde{p}_2(X)$ and posterior variances.

b. With the three priors Π_0, Π_1, Π_2, calculate the three MAP-estimators $\hat{p}_0(X)$, $\hat{p}_1(X)$, $\hat{p}_2(X)$ and calculate their biases.

Take the decision-theoretic perspective on estimation by the choice $\Theta = [0,1]$ for the space in which our decisions take values, and a quadratic loss function $L : [0,1]^2 \to [0,\infty) : (p_1, p_2) \mapsto (p_1 - p_2)^2$.

c. With the three priors Π_0, Π_1, Π_2, calculate the three formal Bayes estimators $p'_0(X)$, $p'_1(X)$, $p'_2(X)$.

For $i \in \{1,2,3\}$, suppose we are interested in estimators $T(X)$ that minimize the quantity,

$$s_i = \int_0^1 P_{n,p}\big(T(X) - p\big)^2 \, d\Pi_i(p),$$

d. Which estimator has minimal s_i: $\tilde{p}_i$, $\hat{p}_i$ or p'_i? Explain why.

2.6.26. In this problem, we calculate posterior distributions.

a. Define $\mathscr{P} = \{N(\theta, \sigma^2) : \theta \in \Theta = [0,\infty)\}$, the model for normal distributions of unknown, positive location θ and known variance $\sigma^2 > 0$. We assume that the data $X_1, \ldots, X_n$ form an *i.i.d.* sample from P_{θ_0} for some $\theta_0 \in \mathbb{R}$. As a prior on Θ, we use the exponential distribution $\mathrm{Exp}(\lambda)$ for some $\lambda > 0$ (which has density $\pi(\theta) = \lambda \exp(-\lambda\,\theta)$ for $\theta \geq 0$ and $\pi(\theta) = 0$ for $\theta < 0$). Calculate the posterior for θ. *Hint: express the normalization constant in terms of the distribution function Φ for the standard normal distribution.*

b. Consider a single-observation X from a uniform distribution $U[0,\theta]$ with unknown $\theta > \beta$, for some known constant $\beta \in \mathbb{R}$. As a prior Π for θ, we use a Pareto distribution with parameters $\alpha > 0$, β. *Hint: recall that the Pareto distribution has a Lebesgue density $\pi(\theta) = (\alpha\beta^\alpha)/\theta^{\alpha+1}$ for all $\theta > \beta$ and $\pi(\theta) = 0$ for $\theta \leq \beta$.* Calculate the posterior distribution for θ.

2.6.27. A series of $n \geq 1$ independent Bernoulli trials $X_1, \ldots, X_n$ is performed, with unknown success probability $\theta \in [0,1]$: for all $1 \leq i \leq n$, $P(X_i = 1) = \theta$. We assign a beta-prior with parameters $a > 0, b > 0$ for θ. Denote $S = \sum_{i=1}^n X_i$.

a. Calculate the posterior for θ. Also give the posterior mean and variance, as a function of a, b and S.

b. Suppose we plan to perform the experiment again, independently, and shall observe a total number of successes T. Based on the outcome S of the original series, the Bayesian predicts the distribution of T. Based on your answer under part *a.*, give this prediction.
Hint: express your answer for $P(T = k|S)$ as a fraction $B(a_1,b_1)/B(a_2,b_2)$ of two normalization constants for beta-densities, times $\binom{n}{k}$, for $a_1, a_2, b_1, b_2 > 0$ that depend on a, b, k and S.

Assume again only the original series with outcome S has been observed. Instead of an independent second series, we observe one more Bernoulli trial Y, which is independent of the $X_1, \ldots, X_n$, with success probability that is only half as large as that for the first n observations, $P(Y = 1) = \theta/2$.

c. Write down the likelihood function $\theta \mapsto L(\theta; X_1, \ldots, X_n, Y)$ for observation of the original series and Y.

Suppose that we observe $Y = 0$ and we take a prior with parameters $a = b = 1$.

d. Show that, with $Y = 0$, the posterior density of θ relative to Lebesgue density on $[0,1]$, is given by,

$$\pi(\theta|S) = C\big(2\theta^S - \theta^{S+1}\big)(1-\theta)^{n-S}$$

where the normalization constant is,

$$C^{-1} = 2\frac{\Gamma(S+1)\Gamma(n-S+1)}{\Gamma(n+2)} - \frac{\Gamma(S+2)\Gamma(n-S+1)}{\Gamma(n+3)}.$$

Hint: Note that $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$.

2.6.28. Let the data $Y \in [0,\infty)$ be distributed according to an exponential distribution $\mathrm{Exp}(\theta)$, for some $\theta > 0$. As a prior for θ, we use a Gamma-distribution $\Gamma(\alpha,\beta)$ with hyperparameters $\alpha > 0$, $\beta > 0$.

a. Calculate the posterior distribution for θ, given Y.

Assume that, next, we observe Y' which is another draw from $\mathrm{Exp}(\theta)$, stochastically independent of Y. (Mind, all conditional on the same draw $\vartheta = \theta$ from the prior.)

b. Show that the posterior predictive distribution for Y', given Y has a Lebesgue density given by,

$$p(y'|Y=y) = (\alpha+1)\frac{(\beta+y)^{\alpha+1}}{(\beta+y+y')^{\alpha+2}}.$$

Hint: for all $x > 0$, $\Gamma(x+1) = x\Gamma(x)$.

c. Show (explicitly, by integration) that $p(y'|Y=y)$ above describes a probability density with respect to Lebesgue measure on $[0,\infty)$.

2.6.29. Consider the parametric model $\mathscr{P}$ consisting of geometric distributions $\mathscr{P} = \{\mathrm{Geo}(\theta) : \theta \in \Theta = (0,1)\}$ for data $X_1, \ldots, X_n \in \{0,1,2,\ldots\}$. Counting measure dominates the model and densities take the form $p_\theta(k) = \theta(1-\theta)^k$ for $k \in \{0,1,2,\ldots\}$. As a prior Π on $(0,1)$ we take a $\mathrm{Beta}(\alpha,\beta)$-distribution with known parameters $\alpha, \beta > 0$. The Lebesgue density of Π is given by,

$$\pi(\theta) = B(\alpha,\beta)^{-1}\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

for $\theta \in (0,1)$, where $B(\alpha,\beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$ is the normalization constant.

a. Calculate the density for the posterior with respect to Lebesgue measure.
b. Express the posterior mean $\hat{\theta}_1(X_1, \ldots, X_n)$ as a fraction of normalization constants $B(\alpha',\beta')$ for certain data-dependent values of α', β'. Use the relation $\Gamma(x+1) = x\Gamma(x)$, valid for all $x \geq 0$ to show that,

$$\hat{\theta}_1(X_1,\ldots,X_n) = \frac{\alpha+n}{\alpha+\beta+n+\sum_i X_i}.$$

c. Calculate also the *maximum-likelihood* estimator $\hat{\theta}_{ML}$ for θ. Show that in the limit $n \to \infty$, $\hat{\theta}_1 - \hat{\theta}_{ML}$ goes to zero.

2.6.30. In this problem sentences are presented from which one or more words have been left out: in each part below, give the missing word (or *words*, as in parts *a.* and *e.*).

a. Given a parametrized model $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ for data Y, with closed, ... parameter space Θ, it is assumed that ϑ is ..... with respect to the posterior in order to define the posterior mean $\hat{\theta} = \int \theta \, d\Pi(\theta|Y)$.
b. Given a parametrized model $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$, the parametrization is said to be *identifiable* if the map $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$ is .....
c. Given a parametrized model $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ for data Y the conditional distribution for Y given $\theta \in \Theta$ is called a ..... distribution.
d. Given a parametrized model $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ for data $Y \sim P_0$ and an estimator $\hat{\theta} : \mathscr{Y} \to \Theta$, the distribution for $\hat{\theta}(Y)$ under P_0 is called the ..... distribution.
e. Suppose a prior has been chosen for which the prior odds ratio is not equal to one. The ..... Bayesian prefers Bayes factors over posterior odds ratios, while the ..... Bayesian prefers posterior odds ratios over Bayes factors.

2.6.31. In the following it is required that you give short but accurate expressions for definitions, (in)equalities or other notions from the theory of Bayesian statistics.

a. Let $(\mathscr{Y},\mathscr{B})$ be a measurable sample space and let a model $\mathscr{P} = \{p_\theta : \theta \in \Theta\}$ for data $Y \in \mathscr{Y}$ be given, with a measurable parameter space $(\Theta,\mathscr{G})$ and a prior $\Pi : \mathscr{G} \to [0,1]$. The posterior $\Pi(\cdot|Y)$, prior predictive P^Π, model distributions P_θ and prior Π are related through an equality called *Bayes's Rule*. Give this equality.
b. Frequentist use of Bayesian tools requires that Bayesian definitions make sense under frequentist assumptions: given the frequentist view that the data Y has distribution $Y \sim P_0$, state the technical condition that guarantees Bayesian definitions are also viable as frequentist definitions.
c. Often the data is known/assumed to have been generated as an *independent and identically distributed (i.i.d.) sample*. How does one express the *i.i.d.* property in the Bayesian framework?
d. Describe in one or two sentences (but with accuracy!) how a posterior odds ratio differs from the corresponding Bayes factor.
e. Let a parametrized model $\mathscr{P} = \{p_\theta : \theta \in \Theta\}$ for data Y with $\Theta \subset \mathbb{R}^d$ be given. Suppose that we choose a prior Π and that the posterior distribution is dominated by Lebesgue measure on Θ, with density $\Theta \to [0,\infty) : \theta \mapsto \pi(\theta|Y)$. Give the definition of the HPD credible set of credible level $\alpha \in (0,1)$.

2.6.32. Consider the parametric model $\mathscr{P}$ consisting of exponential distributions $\mathscr{P} = \{\mathrm{Exp}(\theta) : \theta \in \Theta = (0,\infty)\}$. Lebesgue measure dominates the model and densities take the form $p_\theta(x) = \theta \exp(-\theta x)$, for $x \geq 0$. Assume that the data $X_1,\ldots,X_n$

form an *i.i.d.* sample from some $P_0 = P_{\theta_0} \in \mathscr{P}$. As a prior Π on the model we take a $\Gamma(\alpha,\beta)$-distribution with known parameters $\alpha,\beta > 0$. The Lebesgue density of Π is given by,

$$\pi(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\,\theta^{\alpha-1}\,e^{-\beta\,\theta},$$

for the parameter $\theta > 0$.

a. Calculate the density for the posterior with respect to the Lebesgue measure.
b. Calculate the *maximum-a-posteriori* estimator $\hat{\theta}_{MAP}$ for θ. Calculate also the *maximum-likelihood* estimator $\hat{\theta}_{ML}$ for θ.
c. Using the *law of large numbers*, show that the difference between $\hat{\theta}_{MAP}$ and $\hat{\theta}_{ML}$ goes to zero as $n \to \infty$.
d. Let $D(X^n)$ be a credible interval $[\theta_0(X^n), \theta_1(X^n)]$. Using that the posterior density $\theta \mapsto \pi(\theta|X^n)$ is unimodal (that means: is increasing on an interval $(-\infty,\theta^*]$ and decreasing on $[\theta^*,\infty)$) write down two equations for $\theta_0(X^n)$ and $\theta_1(X^n)$ in order for $D(X^n)$ to be a HPD credible set of credible level α. (There is *no need* to solve these equations or to substitute your answer from part a.)

For the last part of this problem, consider the hypotheses,

$$H_0: \quad 0 < \theta \leq 1, \qquad H_1: \quad \theta > 1.$$

e. Write down the Bayes factor in favour of hypothesis H_0. (There is *no need* to solve or simplify the resulting integrals.)

2.6.33. In this problem, we consider minimax and Bayesian decision theory.

a. Give the definitions of the following notions: loss-function, risk-function, minimax risk, minimax decision rule.
b. State the minimax theorem.
c. Give the definitions of the following notions: Bayesian risk function, Bayes risk, Bayes rule.
d. State the conditional Bayes decision principle.
e. Under the assumption that the model is dominated, prove that the conditional Bayes decision rule is a Bayes rule.
f. Without the assumption that the model is dominated, prove that the conditional Bayes decision rule is a Bayes rule. *Hint: use the disintegration equality.*

2.6.34. A radiation measurement device reports cumulative counts X of radiation over a certain time interval. The count X is assumed to be Poisson distributed with mean $\lambda > 0$. Over the time interval, the observed total count was $x = 23$. For the parameter λ, a Gamma prior $\Gamma(\alpha,\beta)$ with mean 2 and variance 0.5 is chosen.

a. Calculate the posterior $\Pi(\cdot|X)$ for λ given X. Then substitute $X = 23$ to find the realized posterior.
b. Show that the posterior mean $\hat{\lambda}(X)$ given X, is a weighted average of X and a function $f(\alpha,\beta;X)$. Give f.

c. Let $c > 0$ be given. Draw a precise graph of the loss function $\ell : (0,\infty) \to \mathbb{R}$,

$$\mu \mapsto \ell(\mu,\lambda) = e^{c(\mu-\lambda)} - c(\mu-\lambda) - 1,$$

around the point $\mu = \lambda$. Does this loss function reflect, either (A) that under-estimation of the parameter is worse than over-estimation? or (B) that over-estimation of the parameter is worse than under-estimation? Motivate your answer.

d. Show that the estimator $\tilde{\lambda} > 0$ defined as the minimizer of expected loss under the posterior, that is, the (X-dependent) value $\tilde{\lambda}$ such that,

$$\int_0^\infty \ell(\tilde{\lambda},\lambda)\, d\Pi(\lambda|X) = \inf_{\mu>0} \int_0^\infty \ell(\mu,\lambda)\, d\Pi(\lambda|X),$$

is given by,

$$\tilde{\lambda}(X) = -\frac{1}{c} \log \int_0^\infty e^{-c\lambda}\, d\Pi(\lambda|X)$$

e. Without making any calculations, show that the Bayesian risk of $\tilde{\lambda}$ is lower than or equal to the Bayesian risk of $\hat{\lambda}$.

2.6.35. Consider a decision problem involving a states θ from a state space Θ labelling probability distributions P_θ for data Y to be observed, based on which we take a decision a from a decision space $\mathscr{A}$. The loss function is denoted $L : \Theta \times \mathscr{A} \to [0,\infty)$. When we adopt the Bayesian perspective, we endow Θ with (a σ-algebra and) a prior Π.

We take $\Theta = \mathbb{R}$, $Y \in \mathbb{R}$, $P_\theta = N(\theta,1)$ for all θ and a Borel prior $\Pi = N(0,\tau^2)$ for some $\tau^2 > 0$.

a. What is the posterior for the state θ?

Take $\Theta = \mathscr{A}$, denote the decision rule by $\delta : \mathbb{R} \to \mathscr{A}$ and take as a loss function:

$$L(\theta_1,\theta_2) = a\,\theta_1^2 + 2b\,\theta_1\theta_2 + c\,\theta_2^2$$

for some constants $a,b,c \geq 0$.

b. Based on the conditional Bayes decision principle, derive the Bayes rule δ^* and give the Bayes risk.

c. Explain why the constant a does not make an appearance in the expression for δ^*.

2.6.36. Consider a state ϑ that can take values in $\Theta = \{\theta_1,\theta_2\}$ only, and an action a that lies in $\mathscr{A} = \{a_1,a_2\}$, with a loss function $L : \Theta \times \mathscr{A} \to \mathbb{R}$,

$$L(\theta_1,a_1) = L(\theta_2,a_2) = 0, \quad L(\theta_1,a_2) = 5, \quad L(\theta_2,a_1) = 10.$$

As a prior for the state ϑ, we choose $\Pi(\vartheta = \theta_1) = \eta$, $\Pi(\vartheta = \theta_2) = 1-\eta$, for some fixed $\eta \in (0,1)$. We observe a normal random variable $X \in \mathbb{R}$, distributed according to $P_{\theta_1} = N(0,1)$ if $\vartheta = \theta_1$, or according to $P_{\theta_2} = N(1,1)$ if $\vartheta = \theta_2$.

a. Calculate the posterior probabilities $\Pi(\vartheta = \theta_1|X)$ and $\Pi(\vartheta = \theta_2|X)$.
b. Let $\delta : \mathbb{R} \to \mathscr{A}$ be a decision rule; give the Bayesian risk function $r(\Pi, \delta)$. *Hint: calculate first* $R(\theta_1, \delta)$ *and* $R(\theta_2, \delta)$.
c. Calculate the expected posterior loss under the action a_1; then calculate also the expected posterior loss under action a_2.

For each realization $X = x$ of the observation, we can now compare the expected posterior losses associated with actions a_1 and a_2. Let $\delta_c : \mathbb{R} \to \mathscr{A}$ be the decision rule defined by,

$$\delta_c(x) = \begin{cases} a_1, & \text{if } x \leq c \\ a_2, & \text{if } x > c \end{cases}$$

for some choice of the threshold $c \in \mathbb{R}$.

d. State the *conditional Bayes risk principle*. How is it related to the notion of a *Bayes rule*?
e. For which value of c is δ_c a Bayes rule?

Chapter 3
Choice of the prior

Bayesian procedures have received much criticism from frequentists, often focussing on the choice of the prior as an undesirable source of ambiguity. The answer of the subjectivist that the prior represents the "belief" of the statistician or "expert knowledge" pertaining to the measurement's randomness elevates this ambiguity to a matter of principle, thus setting the stage for a heated debate between "pure" Bayesians and "pure" frequentists concerning the philosophical merits of either school within statistics. As said, the issue is complicated further by the fact that the Bayesian procedure does not refer to any "true distribution" P_0 of the observation (see section 2.1), providing another point of fundamental disagreement for fanatics to lock horns over. Leaving the philosophical argument to others, we shall try to discuss the choice of a prior at a more conventional, practical level and point to mathematical properties that choices for the prior have.

In this chapter we consider priors for parametric models from various points of view. In section 3.1, we discuss priors that emphasize the subjectivist's "belief". In section 3.2 we construct priors with the express purpose *not* to emphasize any part of the model more than others, as advocated by objectivist Bayesians. Hierarchical prior construction and Bayesian modelling are the subject of section 3.3, and methods that choose priors by frequentist means (commonly known as *empirical Bayes* methods) form the subject of section 3.4. Because it is mathematically desirable and computationally advantageous to have closed-form expressions for posterior distributions, so-called *conjugacy* of families of distributions over the parameter space is considered in section 3.5. Special attention goes to the Dirichlet distributions of section 3.6 because they describe a conjugate family of probability distributions on spaces of probability measures (rather than parametrizing spaces). As will become clear in the course of the chapter, the choice of a prior is highly dependent on the model under consideration, as well as on the purpose of the analysis.

All of the material of this chapter applies only in parametric models. To find a suitable prior for a non-parametric model can be surprisingly difficult. The concept of a measure on an infinite-dimensional space has technical subtleties that do not play a role in parametric models (*e.g.* the lack of default dominating measures like the counting measure on a discrete space or Lebesgue measure on $\mathbb{R}^d$). One con-

struction stands out as completely natural, however, built from refining partitions and coherent collections of priors like those of section 3.6: see section 8.2 for the definition of a conjugate family of priors on non-parametric models called *Dirichlet process priors*.

3.1 Subjective priors

As was explained in chapters 1 and 2, all statistical procedures require the statistician to make certain choices, *e.g.* for model and method of inference. The subjectivist chooses the model as a collection of stochastic explanations of the data that he finds "reasonable", based on criteria no different from those frequentists and objectivist Bayesians would use.

3.1.1 Motivation for the subjectivist approach

Bayesians then proceed to choose a prior that assigns mass to all parts of the model, usually in such a manner that the support of the prior covers all of the model itself. But even after the support is fixed, there is a large collection of possible priors left to be considered, leading to different posterior distributions. The objectivist Bayesian will choose from those possibilities a prior that is "homogeneous" (in some suitable sense), in the hope of achieving *unbiased inference*. The subjectivist, however, chooses his prior such as to emphasize parts of the model that he believes in stronger than others, thereby introducing a bias in his inferential procedure explicitly. Such a prior is called a *subjective prior*, or *informative prior*. The reason for this approach is best explained by examples like 1.3.1, which demonstrate that intuitive statistical reasoning is not free of bias either.

Subjectivity finds its mathematical expression when high prior "belief" is translated into "relatively large" amounts of assigned prior mass to certain regions of the model. However, there is no clear rule directing the exact fashion in which prior mass is to be distributed. From a mathematical perspective this is a rather serious shortcoming, because it leaves us without a precise definition of the subjectivist approach. Often the subjectivist will have a reasonably precise idea about his "beliefs" at the roughest level (*e.g.* concerning partitions of the model into a few subsets) but none at more detailed levels. When the parameter space Θ is unbounded this lack of detail becomes acute, given that the tail of the prior is hard to fix by subjective reasoning, yet highly influential for the inferential conclusions based on its posterior. In practice, a subjectivist will often choose his prior without mathematical precision. He considers the problem, interprets the parameters in his model and chooses a prior to reflect all the (background) information at his disposition, ultimately filling in remaining details in an ad-hoc manner. It is worthwhile to mention that studies have been conducted focussed on the ability of people to make a realistic guess at a

probability distribution: they have shown that without specific training or practice, people tend to be overconfident in their assessments, assigning too much mass to possibilities they deem most likely and too little to others [3]. This suggests that people tend to formulate their "beliefs" on a deterministic basis and vary around their established opinions only slightly (by adding only a little bit of "noise") when asked to give a realistic probabilistic perspective. (For more concerning the intricacies of choosing subjective prior distributions, see Berger (1985) [19].)

Remark 3.1.1. For this reason it is imperative that a subjectivist prior choice is *fully* described alongside inferential conclusions based upon it. Reporting on methods is important in any statistical setting, but if chosen methods lead to express bias, explanation is even more important. Indeed, not only the prior but also the reasoning leading to its choice should be reported, because in a subjectivist setting, the motivation for the choice of a certain prior is *an intrinsic part of the statistical analysis*.

3.1.2 Methods for the construction of subjective priors

If the model Θ is one-dimensional and the parameter θ has a clear interpretation, it is often not exceedingly difficult to find a reasonable prior Π expressing the subjectivist's *"belief"* concerning the value of θ.

Example 3.1.2. If one measures the speed of light *in vacuo* c (a physical constant, approximately equal to 299792458 m/s), the experiment will be subject to random perturbations outside the control of the experimenter. For example, imperfections of the vacuum in the experimental equipment, small errors in timing devices, electronic noise and countless other factors may influence the resulting measured speed Y. We model the perturbations collectively as a normally distributed error $e \sim N(0, \sigma^2)$ where σ is known as a characteristic of the experimental setup. The measured speed is modelled as $Y = c + e$, *i.e.* the model $\mathscr{P} = \{N(c, \sigma^2) : c > 0\}$ is used to infer on c. Based on experiments in the past the experimenter knows that c has a value close to $3 \cdot 10^8$ m/s, so he chooses his prior to reflect this: a normal distribution located at 300000000 m/s with a standard deviation of (say) 1000000 m/s will do. The latter choice is arbitrary, just like the choice for a *normal* distribution over other possible error distributions.

The situation changes when the parameter has a higher dimension, $\Theta \subset \mathbb{R}^d$: first of all, interpretability of each of the d components of $\theta = (\theta_1, \theta_2, \ldots, \theta_d)$ can be far from straightforward, so that concepts like prior "belief" or "expert knowledge" become inadequate guidelines for the choice of a prior. Additionally, the choice for a prior in higher-dimensional models also involves choices concerning the dependence structure between parameters.

Remark 3.1.3. Often, subjectivist inference employs exceedingly simple, parametric models for the sake of interpretability of the parameter (and to be able to choose a prior accordingly). Most frequentists would object to such choices for their obvious

lack of realism, since they view the data as being generated by a "true, underlying distribution", usually assumed to be an element of the model. By contrast, the subjectivist does not have the ambition to be strictly realistic and calls for interpretability instead: to the subjectivist, inference is a personal rather than a universal matter. As such, the preference for simple parametric models is a matter of subjective interpretation rather than an assumption concerning reality or realistic distributions for the data.

When confronted with the question which subjective prior to use on a higher-dimensional model, it is often of help to define the prior in several steps based on a choice for the dependence structure between various components of the parameter. Consider a the parameter that lies in $\mathbb{R}^d$ (endow $\mathbb{R}$ with the (Borel) σ-algebra $\mathscr{B}$) and suppose that the subjectivist can formulate a reasonable distribution for the first component θ_1, provided he can think about the other components $\theta_2,\ldots,\theta_d$ as being fixed at any value: this prescribes the *conditional prior distribution* $\Pi_{\theta_1|\theta_2,\ldots,\theta_d}$, of θ_1 given the other components. Next suppose that a reasonable subjective prior for the second component may be found, conditional on $\theta_3,\ldots,\theta_d$. This amounts to specification of the conditional distribution $\Pi_{\theta_2|\theta_3,\ldots,\theta_d}$. If we continue like this, eventually defining the marginal prior Π_{θ_d} for the last component θ_d, we have found a dependent prior for the full parameter θ, because for all $A_1,\ldots,A_d \in \mathscr{B}$,

$$\begin{aligned}&\Pi(\theta_1 \in A_1,\ldots,\theta_d \in A_d) = \\ &\quad \Pi(\theta_1 \in A_1|\theta_2 \in A_2,\ldots,\theta_d \in A_d) \times \ldots \times \Pi(\theta_{d-1} \in A_{d-1}|\theta_d \in A_d)\,\Pi(\theta_d \in A_d).\end{aligned}$$

Example 3.1.4. We consider a certain species of monkey and are interested in estimation of the average weight w and length l for a certain sub-population. Suppose that we observe an *i.i.d.* sample of pairs (W_i,L_i), $(1 \le i \le n)$, weight and length of the i-th monkey drawn from the sub-population. Let's assume that the sample size n is small and we want to use our limited amount of data in the most efficient way. In the choice for a prior for (w,l), the (subjective) Bayesian will look for external information that informs the estimation of (w,l) beyond what the data itself has to offer: let's assume we have prior knowledge that derives from the (much larger) population of all monkeys of this species: according to the subjectivist's expert knowledge, weight w and length l of this species are related (approximately) through a power-law relationship $w = K l^\alpha$, for some known $K,\alpha > 0$.

As our model for (W,L), we choose products of Gamma distributions with shape parameter $k = 3$: $(W,L) \sim \Gamma(w/3,3) \times \Gamma(l/3,3)$, so that the model distribution $(W,L)|(w,l)$ has expectation (w,l). Note that to the frequentist, this suggests estimation by sample means $(\overline{W}_n,\overline{L}_n)$. A subjective prior for (w,l) is now defined as follows: *given* the length l, we specify (for some fixed choice $\lambda > 0$),

$$w|l \sim \Gamma\big(\lambda^{-1} K l^\alpha,\lambda\big),$$

so that the conditional prior expectation for $w|l$ is $K l^\alpha$. For l, we choose a marginal prior of the form,

$$l \sim \Gamma(\lambda^{-1}\ell,\lambda),$$

where ℓ is an approximate mean length for a monkey of this species. This has an effect that can be explained in two ways: firstly the bias that the prior introduces clearly shrinks the estimate towards the curve $w = K l^{\alpha}$, permitting it to "correct" for variance that the sample means might display because they are based on a rather small number of observations. A second view is that the chosen prior lets the observed lengths L_i play a role not only for the estimation of l but also that of w, using that, according to the curve, a monkey of length L_i usually has weight $K L_i^{\alpha}$. To interpret λ, note that prior variances of l and $w|l$ are equal to ℓ^2/λ and $K^2 l^{2\alpha}/\lambda$ respectively, *i.e.* λ is inversely proportional to uncertainty expressed in the prior (higher values of λ bias the prior (and hence also the posterior) more to the prior expectation $(\ell, K\ell^{\alpha})$).

So in the above example, observed lengths become informative for the mean weight w and observed weights become informative for the mean length l, through the choice of a subjective prior. This extra inferential aspect is the strength of subjective Bayesian statistics and it enables a wealth of modelling options. Clearly a frequentist would also be able to shrink his estimates towards the expected curve for (w,l), but his philosophy, like that of the objective Bayesian, tells him not to.

The construction indicated here is reminiscent of that of a so-called *hyperprior*, which is discussed in section 3.5. The difference is, that components of θ all occur in the definition of model distributions P_θ, whereas hyperparameters do not. Note that it is important to choose a parametrization of the model in which the independence between θ_i and $(\theta_1, \ldots, \theta_{i-1})$, *given* $(\theta_{i+1}, \ldots, \theta_d)$, is plausible for all $i \geq 1$.

In certain situations, the subjectivist has more factual information at his disposal when defining the prior for his analysis. In particular, if a probability distribution on the model reflecting the subjectivist's "beliefs" can be found by other statistical means, it can be used as a prior. Suppose the statistician is planning to measure a quantity Y and infer on a model $\mathscr{P}$; suppose also that this experiment repeats or extends an earlier analysis. From the earlier analysis, the statistician may have obtained a posterior distribution on $\mathscr{P}$. For the new experiment, this posterior may serve as a prior.

Example 3.1.5. Let $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$ be a parametrized model for *i.i.d.* $X_1, X_2, \ldots, X_n$ with prior measure $\Pi_1 : \mathscr{G} \to [0,1]$. Let the model be dominated (see definition 1.1.3), so that the posterior $\Pi_1(\cdot|X_1, \ldots, X_n)$ satisfies (2.15). Suppose that this experiment has been conducted, with the sample realised as $(X_1, X_2, \ldots, X_n) = (x_1, x_2, \ldots, x_n)$. Next, consider a new, independent experiment in which a quantity X_{n+1} is measured (with the same model). As a prior Π_2 for the new experiment, we use the (realised) posterior of the earlier experiment, *i.e.* for all $G \in \mathscr{G}$,

$$\Pi_2(G) = \Pi_1(G|X_1 = x_1, \ldots, X_n = x_n).$$

The posterior for the second experiment then satisfies:

$$d\Pi_2(\theta|X_{n+1}) = \frac{p_\theta(X_{n+1})\, d\Pi_1(\theta|X_1 = x_1, \ldots, X_n = x_n)}{\int_\Theta p_\theta(X_{n+1})\, d\Pi_1(\theta|X_1 = x_1, \ldots, X_n = x_n)}$$
$$= \frac{p_\theta(X_{n+1}) \prod_{i=1}^{n} p_\theta(x_i) d\Pi_1(\theta)}{\int_\Theta p_\theta(X_{n+1}) \prod_{j=1}^{n} p_\theta(x_j)\, d\Pi_1(\theta)} \tag{3.1}$$

The latter form is comparable to the posterior that would have been obtained if we had conducted a single experiment with an *i.i.d.* sample $X_1, X_2, \ldots, X_{n+1}$ of size $n+1$ and prior Π_1. In that case, the posterior would have been of the form:

$$\Pi(\cdot\,|X_1, \ldots, X_{n+1}) = \frac{\prod_{i=1}^{n+1} p_\theta(X_i)\, d\Pi_1(\theta)}{\int_\Theta \prod_{j=1}^{n+1} p_\theta(X_j)\, d\Pi_1(\theta)}, \tag{3.2}$$

i.e. the only difference is the fact that the posterior $\Pi_1(\cdot\,|X_1 = x_1, \ldots, X_n = x_n)$ is realised. As such, we may interpret independent consecutive experiments as a single, interrupted experiment and the posterior $\Pi_1(\cdot\,|X_1, \ldots, X_n)$ can be viewed as an intermediate result.

3.2 Non-informative priors

Objectivist Bayesians argee with frequentists that the "beliefs" of the statistician analyzing a given measurement should play a minimal role in the methodology. Obviously, the model choice already introduces a bias, but rather than embrace this necessity and expand upon it like subjectivists, they seek to keep the remainder of the procedure unbiased. In particular, they aim to use priors that do not introduce additional information (in the form of prior "belief") in the procedure. Subjectivists introduce their "belief" by concentrating prior mass in certain regions of the model; correspondingly, objectivists prefer priors that are "homogeneous" in an appropriate sense.

3.2.1 Uniform priors

At first glance, one may be inclined to argue that a prior is *objective* (or *non-informative*) if it is uniform over the parameter space: if we are inferring on parameter $\theta \in [0,1]$ and we do not want to favour any part of the model over any other, we would choose the Lebesgue measure on $[0,1]$ for a prior. Attempts to

minimize the amount of subjectivity introduced by the prior therefore focus on uniformity (argumentation that departs from the Shannon entropy in discrete probability spaces reaches the same conclusion (see, for example, Ghosh and Ramamoorthi (2003) [111], p. 47)). The original references on Bayesian methods (*e.g.* Bayes (1763) [13], Laplace (1774) [165]) use uniform priors as well. But there are several problems with this approach: first of all, one must wonder how to extend such reasoning when $\theta \in \mathbb{R}$ (or any other unbounded subset of $\mathbb{R}$). In that case, Lebesgue measure is infinite and cannot be normalized to a probability measure. Any attempt to extend Π to such unbounded models as a probability measure would eventually lead to inhomogeneity, *i.e.* go at the expense of the unbiasedness of the procedure.

The compromise some objectivists are willing to make, is to relinquish the interpretation that subjectivists give to the prior: they do not express any prior "degree of belief" in $A \in \mathscr{G}$ through the subjectivist statement that the (prior) probability of finding $\vartheta \in A$ equals $\Pi(A)$. Although they maintain the Bayesian interpretation of the posterior, they view the prior as a mathematical definition rather than a philosophical concept. Then, the following definition can be made without further reservations.

Definition 3.2.1. Given a model $(\Theta, \mathscr{G})$, a prior measure $\Pi : \mathscr{G} \to [0, \infty]$ such that $\Pi(\Theta) = \infty$ is called an *improper* prior.

Note that any dependence on the normalization factor for a prior cancels in the expression for the posterior, *c.f.* (2.4) or (2.6): any finite multiple of a (bounded) prior is equivalent to the original prior as far as the posterior is concerned. However, this argument does not extend to the improper case: integrability problems or other infinities may ruin the procedure, even to the point where the posterior measure becomes infinite or ill-defined. So not just the philosophical foundation of the Bayesian approach is lost, mathematical integrity of the procedure can no longer be guaranteed either. When confronted with an improper prior, the entire procedure must be checked for potential problems. In particular, one must verify that the posterior is a well-defined *probability* measure. (Throughout this book we use only priors that are probability measures.)

But even if one is willing to accept that objectivity of the prior requires that we restrict attention to models on which "uniform" probability measures exist (*e.g.* with Θ a bounded subset of $\mathbb{R}^d$), a more fundamental problem exists: the very notion of uniformity is dependent on the parametrization of the model, which is problematic because the parametrization is a *subjective* choice: the result is that two objectivist Bayesians may insist on uniformity each in their own chosen parametrization, and reach a subjective disagreement on what is objectively bona fide! To see this we look at a model that can be parametrized in two ways and we consider the way in which uniformity as seen in one parametrization manifests itself in the other parametrization. Suppose that we have a d-dimensional parametric model $\mathscr{P}$ with two different parametrizations, on $\Theta_1 \subset \mathbb{R}^d$ and $\Theta_2 \subset \mathbb{R}^d$ respectively,

$$\phi_1 : \Theta_1 \to \mathscr{P}, \qquad \phi_2 : \Theta_2 \to \mathscr{P} \tag{3.3}$$

both of which are bijective. Assume that $\mathscr{P}$ is a measurable space with σ-algebra $\mathscr{G}$. Require that ϕ_1 and ϕ_2 are Borel-to-$\mathscr{G}$ measurable. Assume that their inverses ϕ_1^{-1} and ϕ_2^{-1} are measurable as well (or assume that $\mathscr{P}$ is a Souslin space and use the remark following theorem C.4.10). Assuming that Θ_1 is bounded, we consider the uniform prior Π_1 on Θ_1, *e.g.* the normalized Lebesgue measure $\Pi_1(A) = \mu(\Theta_1)^{-1}\mu(A)$, for all $A \in \mathscr{B}_1$. This induces a prior Π_1' on $\mathscr{P}$: for all $B \in \mathscr{G}$,

$$\Pi_1'(B) = (\Pi_1 \circ \phi_1^{-1})(B). \tag{3.4}$$

In turn, this induces a prior Π_1'' on Θ_2: for all $C \in \mathscr{B}_2$,

$$\Pi_1''(C) = (\Pi_1' \circ (\phi_2^{-1})^{-1})(C) = (\Pi_1' \circ \phi_2)(C) = \big(\Pi_1 \circ (\phi_1^{-1} \circ \phi_2)\big)(C).$$

Even though Π_1 is uniform, generically Π_1'' is *not*, because, effectively, we are mapping (a subset of) $\mathbb{R}^d$ to $\mathbb{R}^d$ by $\phi_2^{-1} \circ \phi_1 : \Theta_1 \to \Theta_2$. (Differentiable counterparts to such measurable re-coordinatizations are used extensively in differential geometry, where a smooth manifold is parametrized in various ways by sets of maps called *charts*.)

Example 3.2.2. Consider the model of all normal distributions centred on the origin with unknown variance between 0 and 1. We may parametrize this model in many different ways, but we consider only the following two:

$$\phi_1 : (0,1) \to \mathscr{P} : \tau \mapsto N(0,\tau), \qquad \phi_2 : (0,1) \to \mathscr{P} : \sigma \mapsto N(0,\sigma^2). \tag{3.5}$$

Although used more commonly than ϕ_1, parametrization ϕ_2 is not special in any sense: both parametrizations describe exactly the same model. Now, suppose that we choose to endow the first parametrization with a uniform prior Π_1, equal to the Lebesgue measure μ on $(0,1)$. By (3.4), this induces a prior on $\mathscr{P}$. Let us now see what this prior looks like if we consider $\mathscr{P}$ parametrized by σ: for any constant $C \in (0,1)$ the point $N(0,C)$ in $\mathscr{P}$ is the image of $\tau = C$ and $\sigma = \sqrt{C}$, so the relation between τ and corresponding σ is given by

$$\tau(\sigma) = (\phi_2^{-1} \circ \phi_1)(\sigma) = \sigma^2.$$

Since Π_1 equals the Lebesgue measure, we find that the density of Π_1'' with respect to the Lebesgue measure equals:

$$\pi_1''(\sigma)\,d\sigma = \pi_1(\tau(\sigma))\left|\frac{d\tau}{d\sigma}\right|(\sigma)\,d\sigma = 2\sigma\,d\sigma.$$

This density is non-constant and we see that Π_1'' is non-uniform. In a subjectivist sense, the prior Π_1'' places higher prior "belief" on values of σ close to 1 than on values close to 0.

From the above argument and example 3.2.2, we see that uniformity of the prior is entirely dependent on the parametrization: what we call "uniform" in one parametrization, may be highly non-uniform in another: what is deemed objective in one

parametrization can turn out to be highly subjective in another. What matters is the model $\mathscr{P}$ itself and *not* its parametrization in terms of any specific parameter. A notion of uniformity intrinsic to $\mathscr{P}$ would resolve the matter in a parametrization-independent way, but spaces of probability measures do not come with such a notion automatically.

3.2.2 Jeffreys prior and reference priors

Once it is clear that uniformity on any parametrizing space does not have intrinsic meaning in the model $\mathscr{P}$, the very definition of objectivity in terms of uniformity of the prior is void. A subjectivist can use any parametrization to formulate his prejudice but an objectivist has to define his notion of "objectivity" regardless of the parametrization used. Therefore, the emphasis is shifted: instead of looking for uniform priors, we look for priors that are well-defined on $\mathscr{P}$ and declare them objective. For differentiable parametric models, a construction from Riemannian geometry can be used to define a parameterisation-independent prior (see Jeffreys (1946), (1961) [136, 137]) if we interpret the Fisher information as a Riemannian metric on the model (as first proposed by Rao (1945) [214] and extended by Efron (1975) [87]; for an overview, see Amari (1990) [4]) and use the square-root of its determinant as a density with respect to the Lebesgue measure.

Definition 3.2.3. Let $\Theta \subset \mathbb{R}^d$ be open and let $\mathscr{P}$ be a dominated model with identifiable, differentiable parametrization $\Theta \to \mathscr{P}$. Assume that for every $\theta \in \Theta$, the score-function $\dot{\ell}_\theta$ is square-integrable with respect to P_θ. The *Jeffreys prior* Π has the square root of the determinant of the Fisher information $I_\theta = P_\theta \dot{\ell}_\theta \dot{\ell}_\theta^T$ as its density with respect to the Lebesgue measure on Θ:

$$d\Pi(\theta) = \sqrt{\det(I_\theta)}\, d\theta. \tag{3.6}$$

The expression for Jeffreys prior has the appearance of being highly dependent on the parametrization of $\mathscr{P}$ in terms of $\theta \in \Theta$. However, the form (3.6) of this prior is the *same* in *any* other parametrization related in a smooth way (a property referred to sometimes as *covariance* with respect to diffeomorphisms). In other words, no matter which (smooth) parametrization we use to calculate Π *c.f.* (3.6), the induced measure Π' on $\mathscr{P}$ is always the same one. As such, Jeffreys prior is a measure defined intrinsically on $\mathscr{P}$.

Example 3.2.4. We calculate the density of Jeffreys prior in the normal model of example 3.2.2. The score-function with respect to the parameter σ in parametrization ϕ_2 of $\mathscr{P}$ is given by:

$$\dot{\ell}_\sigma(X) = \frac{1}{\sigma}\left(\frac{X^2}{\sigma^2} - 1\right).$$

The Fisher information (which is a 1×1-matrix in this case), is then given by:

$$I_\sigma = P_\sigma \dot{\ell}_\sigma \dot{\ell}_\sigma = \frac{1}{\sigma^2} P_\sigma \Big(\frac{X^2}{\sigma^2} - 1 \Big)^2 = \frac{2}{\sigma^2}$$

Therefore, the density for Jeffreys prior Π takes the form

$$d\Pi(\sigma) = \frac{\sqrt{2}}{\sigma} d\sigma,$$

for all $\sigma \in \Theta_2 = (0,1)$. A similar calculation using the parametrization ϕ_1 shows that, in terms of the parameter τ, Jeffreys prior takes the form:

$$d\Pi(\tau) = \frac{1}{\sqrt{2}\tau} d\tau,$$

for all $\tau \in \Theta_1 = (0,1)$. That both densities give rise to the same measure on $\mathscr{P}$ is the assertion of the following lemma.

Lemma 3.2.5. *(Parameterization-independence of Jeffreys prior) Suppose that the conditions of definition 3.2.3 are satisfied and that ϕ_1 and ϕ_2 are two parametrizations related through a diffeomorphism (i.e. $\phi_1^{-1} \circ \phi_2$ and $\phi_2^{-1} \circ \phi_1$ are differentiable bijections). Then the densities (3.6), calculated in coordinates ϕ_1 and ϕ_2 induce the same measure on $\mathscr{P}$, the Jeffreys prior.*

Proof. Since the Fisher information can be written as:

$$I_{\theta_1} = P_{\theta_1}(\dot{\ell}_{\theta_1} \dot{\ell}_{\theta_1}^T),$$

and the score $\dot{\ell}_{\theta_1}(X)$ is defined as the gradient of $\theta_1 \mapsto \log p_{\theta_1}(X)$ with respect to θ_1, the change of parametrization $\theta_1(\theta_2) = (\phi_1^{-1} \circ \phi_2)(\theta_2)$ induces a transformation of the form,

$$I_{\theta_2} = S_{1,2}(\theta_2) I_{\theta_1(\theta_2)} S_{1,2}(\theta_2)^T,$$

on the Fisher information matrix, where $S_{1,2}(\theta_2)$ is the total derivative of $\theta_2 \mapsto \theta_1(\theta_2)$ in the point θ_2 of the model. Therefore,

$$\begin{aligned}
\sqrt{\det I_{\theta_2}}\, d\theta_2 &= \sqrt{\det(S_{1,2}(\theta_2) I_{\theta_1(\theta_2)} S_{1,2}(\theta_2)^T)}\, d\theta_2 \\
&= \sqrt{\det(S_{1,2}(\theta_2))^2} \sqrt{\det(I_{\theta_1(\theta_2)})}\, d\theta_2 \\
&= \sqrt{\det(I_{\theta_1(\theta_2)})}\, \big|\det(S_{1,2}(\theta_2))\big|\, d\theta_2 = \sqrt{\det(I_{\theta_1})}\, d\theta_1
\end{aligned}$$

i.e. the form of the density of the Jeffreys prior is such that reparametrization leads exactly to the Jacobian for the transformation of $d\theta_2$ to $d\theta_1$.

Ultimately, the above construction derives from the fact that the Fisher information I_θ (or, in fact, any Hessian of a twice-differentiable convex function) can be viewed as a Riemann metric on the "smooth manifold" $\mathscr{P}$. The definition of a measure with Lebesgue density (3.6) is then a standard construction of a measure on the manifold in differential geometry.

Example 3.2.6. To continue with the normal model of examples 3.2.2 and 3.2.4, we note that $\sigma(\tau) = \sqrt{\tau}$, so that $d\sigma/d\tau(\tau) = 1/(2\sqrt{\tau})$. As a result,

$$\sqrt{\det I_{\theta_2}}\, d\theta_2 = \frac{\sqrt{2}}{\sigma}\, d\sigma = \frac{\sqrt{2}}{\sigma(\tau)} \left|\frac{d\sigma}{d\tau}\right| (\tau)\, d\tau = \frac{1}{\sqrt{2}\tau}\, d\tau = \sqrt{\det(I_{\theta_1})}\, d\theta_1,$$

which verifies the assertion of lemma 3.2.5 explicitly.

Other constructions and criteria for the construction of non-informative priors exist: very popular is the use of so-called *reference priors*, as introduced in Lindley (1956) [185] and rediscovered in Bernardo (1979) [24] (see also Berger and Bernardo (1992) [20]). By defining principle, a reference prior is required to maximize the *Kullback-Leibler divergence* of the prior relative to the posterior. To motivate this condition, we have to look at information theory, from which Kullback-Leibler divergence emerges as a (popular but by no means unique) way to quantify the notion of the "amount of information" contained in a probability distribution. Sometimes called the *Shannon entropy*, the Kullback-Leibler divergence of the counting measure with respect to a distribution P in discrete probability spaces,

$$S(P) = \sum_{\omega \in \Omega} p(\omega) \log(p(\omega)),$$

can be presented as such convincingly (see Bolzmann (1895, 1898) [44], Shannon (1948) [230]). For lack of a default dominating measure, the argument does not extend formally to continuous probability spaces but is generalized regardless.

Definition 3.2.7. A *reference prior* Π on a dominated, parametrized model $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$ for an observation Y is a maximizer of the so-called *Lindley entropy*,

$$S_L = \int \int \log\Big(\frac{\pi(\theta|Y=y)}{\pi(\theta)}\Big) d\Pi(\theta|Y=y)\, dP^{\Pi}(y),$$

which measures the prior-predictive expectation of the Kullback-Leibler divergence of the prior with respect to the posterior.

Note that Bayes's Rule (2.4) (see also exercise 2.6.7) allows us to rewrite the Lindley entropy in the form,

$$S_L = \int \int \log\Big(\frac{\pi(\theta|Y=y)}{\pi(\theta)}\Big) dP_\theta(y)\, d\Pi(\theta),$$

Usually, the derivation of a reference prior [24] is performed in the large-sample limit where the posterior for a sufficiently smooth model becomes asymptotically normal, in accordance with the Bernstein-von Mises theorem of chapter 4. For certain models, Jeffreys prior emerges as a reference prior.

For an overview of various objective methods of constructing priors, the reader is referred to Kass and Wasserman (1995) [141]. When using non-informative priors, however, the following general warning should be heeded.

Remark 3.2.8. In many models, non-informative priors, including Jeffreys prior and reference priors, are improper.

3.3 Hierarchical priors

Consider again the problem of estimating the mean of a single, normally distributed observation Y with known variance. The model consists of all normal distributions $P_\theta = N(\theta, \sigma^2)$, where $\theta \in \mathbb{R}$ is unknown and $\sigma^2 > 0$ is known. Imposing a normal prior on the parameter θ, $\Pi = N(0, \tau^2)$, for some choice of $\tau^2 > 0$, we calculate the posterior distribution,

$$\Pi(\theta \in A|Y) = N\left(\frac{\tau^2}{\sigma^2+\tau^2}Y, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\right)(A), \tag{3.7}$$

for every $A \in \mathscr{B}$. As long as sufficient expert knowledge is available, subjectivist choices for a certain value of τ^2 can be motivated, as in example 3.1.2. But in situations where no prior belief on the parameter θ is available, or if the parameter itself does not have a clear interpretation, there is no subjectivist way forward, even though a choice for τ^2 is required. This leaves various options: we may express our ignorance concerning τ^2 by choosing a prior on objectivist grounds, or by considering (more and more homogeneous but still normal) priors in the limit $\tau \to \infty$, motivated by the approximate unbiasedness of resulting estimators.

Remark 3.3.1. However, from a statistical perspective there exists a better way to deal with the uncertainty in τ^2: since τ^2 is not known, we estimate its value from the data.

In this section and the next, we consider this answer from the Bayesian and from the frequentist's angle respectively, giving rise to procedures known as *hierarchical Bayesian modelling* and *empirical Bayesian estimation*.

3.3.1 Hyperparameters and hyperpriors

First we turn to the Bayesian answer to remark 3.3.1: the Bayesian views a parameter to be estimated as just another random variable in the probability model. In case we want to estimate the parameter for a family of priors, then that parameter is to be included in the probability space from the start. Going back to the example with which we started this section, this means that we still use normal distributions $P_\theta = N(\theta, \sigma^2)$ to model the uncertainty in the data Y, supply $\theta \in \mathbb{R}$ with a prior $\Pi_1 = N(0, \tau^2)$ and then proceed to choose a another prior Π_2 for $\tau^2 \in (0, \infty)$:

$$Y|\theta, \tau^2 = Y|\theta \sim P_\theta = N(\theta, \sigma^2), \quad \theta|\tau^2 \sim \Pi_1 = N(0, \tau^2), \quad \tau^2 \sim \Pi_2,$$

Note that the parameter τ^2 has no direct bearing on the model distributions: conditionally on θ, $Y|\theta$ is independent of τ^2. The hierarchical Bayesian approach gives rise to priors that intermediate between subjective and objective philosophies. The subjectivist makes a definite, informed choice for τ^2 while the objectivist keeps himself as uncommitted as possible: if Π_2 is chosen highly concentrated around one point in the model resembling a degenerate measure, the procedure will be close to subjective; if Π_2 is spread widely and is far from degenerate, the procedure will be less biased and closer to objective.

More importantly, the flexibility gained through introduction of Π_2 offers a much wider freedom of modelling. Particularly we may add several levels of parameters, building up a *hierarchy* of priors for parameters of priors. Such structures are used to express detailed subjectivist beliefs, much in the way graphical models are used to build intricate dependency structures for observed data. The origins of the hierarchical approach go back, at least, to Lindley and Smith (1972) [186].

Definition 3.3.2. Let the data Y be random in $(\mathscr{Y},\mathscr{B})$. A *hierarchical Bayesian model* for Y consists of a collection of probability measures $\mathscr{P}=\{P_\theta:\theta\in\Theta_0\}$, with $(\Theta_0,\mathscr{G}_0)$ measurable and endowed with a prior $\Pi:\mathscr{G}_0\to[0,1]$ built up in the following way: for some $k\geq 1$, we introduce measurable spaces $(\Theta_i,\mathscr{G}_i)$, $i=1,2,\ldots,k$, conditional priors

$$\mathscr{G}_i\times\Theta_{i+1}\to[0,1]:(G,\theta_{i+1})\mapsto\Pi_i(G|\theta_{i+1}),$$

for $i=0,\ldots,k-1$ and a marginal $\Pi_k:\mathscr{G}_k\to[0,1]$ on Θ_k. The prior for the original parameter θ is then defined by,

$$\Pi(\theta\in G)=\int_{\Theta_1\times\ldots\times\Theta_k}\Pi_0(\theta\in G|\theta_1)\,d\Pi(\theta_1|\theta_2)\ldots d\Pi(\theta_{k-1}|\theta_k)\,d\Pi_k(\theta_k),\quad(3.8)$$

for all $G\in\mathscr{G}_0$. The parameters $\theta_1,\ldots\theta_k$ and the priors $\Pi_1,\ldots,\Pi_2$ are called *hyperparameters* and their *hyperpriors*.

It is worth mentioning that the same hierarchical structure is used in so-called *graphical models* to model detailed dependence structures for higher-dimensional observations: if $Y=(Y_1,\ldots,Y_d)$, the joint distribution may be constructed in several steps from the conditional distributions $Y_i|Y_{i+1}\ldots Y_d$. A graphical model is defined when these conditionals are chosen from (usually parametric) families, leading to statistical questions regarding estimation, testing and uncertainty quantification for the parameter. The resulting models are sometimes referred to as *Bayesian belief networks* due to the analogy with hierarchical priors. However the conceptual difference is clear: components of Y are observed while components of the parameter θ and hyperparameters are not.

Definition 3.3.2 is also very close to the general Bayesian model that incorporates all components of the parameter as modelling parameters, as in example 3.1.4. What distinguishes hierarchical modelling from the general situation is the independence of $Y|\theta$ from higher (hyper)components of the parameter. This distinction is repeated

at higher levels in the hierarchy, *i.e.* levels are separate from one another through the conditional independence of $\theta_i|\theta_{i+1}$ from $\theta_{i+2},\dots,\theta_k$.

Remark 3.3.3. The hierarchy indicated in definition 3.3.2 inherently loses interpretability as we ascend in level. One may be able to give a viable interpretation to the parameter θ and to the hyperparameter θ_1, but higher-level parameters $\theta_2,\theta_3,\dots$ become harder and harder to understand heuristically. Since the interpretation of the hierarchy requires a subjective motivation of the hyperpriors, interpretability of each level is imperative, or left as a non-informative choice. In practice, Bayesian hierarchical models are rarely more than a few levels deep ($k = 2,3$) and the last hyperprior Π_k is often chosen by objective criteria.

3.3.2 Hierarchical prior construction in an example

To illustrate the rather formal definitions of the previous subsection, we consider a very basic example of Bayesian modelling with hierarchical priors in some detail.

Example 3.3.4. We observe the number Y of surviving offspring from a bird's litter and aim to estimate the number of eggs the bird has laid: the bird lays $N \geq 0$ eggs, distributed according to a Poisson distribution with parameter $\lambda > 0$. For the particular species of bird in question, the Poisson rate λ is not known exactly: the uncertainty in λ can be modelled in many ways; here we choose to model it by a *Gamma*-distribution $\Gamma(\alpha,\beta)$ (with density denoted $p_{\alpha,\beta}$), where α and β are chosen to reflect our imprecise knowledge of λ as well as possible. Each of the eggs then comes out, producing a viable chick with known probability $p \in [0,1]$, independently. Hence, the total number Y of surviving chicks from the litter is distributed according to a binomial distribution, conditional on N,

$$Y|N,\lambda = Y|N \sim \mathrm{Bin}(N,p), \quad N|\lambda \sim \mathrm{Poisson}(\lambda), \quad \lambda \sim \Gamma(\alpha,\beta).$$

The posterior distribution is obtained as follows: conditional on $N = n$, the probability of finding $Y = k$ is binomial,

$$P(Y = k|N = n) = \binom{n}{k} p^k (1-p)^{n-k},$$

so Bayes's rule tells us that the posterior is given by:

$$P(N = n|Y = k) = \frac{P(N = n)}{P(Y = k)} \binom{n}{k} p^k (1-p)^{n-k}.$$

Since $\sum_{n\geq 0} P(N = n|Y = k) = 1$ for every k, the marginal $P(Y = k)$ (that is, the denominator or normalization factor for the posterior given $Y = k$) can be read off once we have the expression for the numerator. We therefore concentrate on the marginal for $N = n$, ($n \geq 0$):

$$P(N=n)=\int_{\mathbb{R}} P(N=n|\lambda)\,p_{\alpha,\beta}(\lambda)\,d\lambda=\frac{\beta^{\alpha}}{\Gamma(\alpha)}\int_0^{\infty}\frac{e^{-\lambda}\,\lambda^{n}}{n!}\lambda^{\alpha-1}\,e^{-\beta\lambda}\,d\lambda.$$

The integral is solved using the normalization constant of the $\Gamma(\alpha+n,\beta+1)$-distribution:

$$\int_0^{\infty} e^{-(\beta+1)\lambda}\lambda^{\alpha+n-1}\,d\lambda=\frac{\Gamma(\alpha+n)}{(\beta+1)^{\alpha+n}}.$$

Substituting and using the identity $\Gamma(\alpha+1)=\alpha\Gamma(\alpha)$, we find:

$$\begin{aligned}P(N=n)&=\frac{\Gamma(\alpha+n)}{\Gamma(\alpha)}\frac{1}{n!}\frac{\beta^{\alpha}}{(\beta+1)^{\alpha+n}}\\&=\frac{1}{n!}\Big(\frac{\beta}{\beta+1}\Big)^{\alpha}(\beta+1)^{-n}\prod_{l=1}^{n}(\alpha+l-1)\end{aligned}\tag{3.9}$$

Although not in keeping with the subjective argumentation of the introduction to this example, for simplicity we consider $\alpha=\beta=1$ and find that in that case,

$$P(N=n)=\Big(\frac{1}{2}\Big)^{(n+1)}.$$

The posterior for $N=n$ given $Y=k$ then takes the form:

$$P(N=n|Y=k)=\frac{1}{2^n}\binom{n}{k}p^k(1-p)^{n-k}\Big/\sum_{m\geq 0}\frac{1}{2^m}\binom{m}{k}p^k(1-p)^{m-k}.$$

The eventual form of the posterior illustrates that the hierarchy contributes only to the construction of the prior: in case we choose $\alpha=\beta=1$, the posterior we find from the hierarchical approach does not differ from the posterior that we would have found if we had started from the model that incorporates a geometric prior for N,

$$Y|N\sim\text{Bin}(N,p),\quad N\sim\text{Geo}(1/2).$$

Indeed, even if we leave α and β free, the marginal distribution for N we found in (3.9) is none other than the prior (3.8) for this problem.

The hierarchical approach to prior construction allows for greater freedom and a more solid foundation to *motivate* the choice for certain prior over other possibilities. This point is all the more significant in light of remark 3.1.1: the motivation of a subjectivist choice for the prior is part of the statistical analysis rather than an external aspect of the procedure. Hierarchical Bayesian modelling helps to refine and justify motivations for subjectivist priors.

3.4 Empirical priors

More unexpected is the frequentist perspective on remark 3.3.1, which goes by the general name *empirical Bayes*: point-estimate τ^2 first based on available data and then perform the Bayesian analysis with the estimate as a "plug-in" for the unknown τ^2. Critical notes can be placed with the philosophical foundations for this practice, since it appears to combine the methods of two contradictory schools of statistics. Be that as it may, the method is used routinely based on its practicality: ultimately justification comes from the subjectivist who does not reject frequentist methods to obtain expert knowledge on his parameters, or from the frequentist who wants to de-bias posteriors.

There are two problems however: of course common sense tells us that it is crucial for any statistical analysis that we first obtain a certain feeling for the statistical problem by inspection of the data, before making decisions on how to analyse it. However, this frequentist form of "expert knowledge" is at odds with another common-sense practical rule: good statistical practice requires that one may not use the data to decide which statistical method to use for the analysis of the same data. The rationale behind this dictum is the potential for introduction of bias in the analysis. The first problem, then, is that posteriors for data-dependent priors have the potential to be biased in complicated and unpredictable ways.

This warning is customarily ignored in the literature: it is common practice to calculate the posterior $\Pi(\cdot|Y,\tau)$ for fixed values of a hyperparameter τ^2 and subsequently substitute "plug-in" estimates $\hat{\tau}(Y)$ based on the same data Y. The resulting quantity,

$$\Pi(\,\vartheta \in B\,|\,Y,\hat{\tau}(Y)\,) = \int_B p_\theta(Y)\,d\Pi_{\hat{\tau}(Y)}(\theta) \Big/ \int_\Theta p_\theta(Y)\,d\Pi_{\hat{\tau}(Y)}(\theta), \qquad (3.10)$$

is not a posterior: the random probability measure (3.10) has little to do with the conditional distribution of $\theta|Y$, because the definition of the posterior, Bayes's rule in the form (2.4), does not leave room for any data-dependence of the prior.

To avoid these problems, a data-dependent prior should not depend on the *same data* Y that is used later to derive the posterior distribution for $\theta|Y$: ideally one splits the available data into two independent parts, making any data-driven choice for the prior based on one sub-sample and performing the analysis proper with the other. Independence between the sub-samples guarantees the absence of bias. In fact the subjective prior choice of example 3.1.4 is motivated in this way: a natural "split of the sample" occurs in situations where we analyse data pertaining to individuals from a larger population. It is often reasonable to assume that hyperparameters can be estimated from the population and that we can use the estimates as hyperparameters for the prior choice to analyse the data for the individuals. Typically, one has data consisting of *i.i.d.* $X_1,\ldots,X_n$ that can be split into two independent, smaller *i.i.d.* samples (which poses the interesting question which fraction of the data is to be used for estimation of hyperparameters like τ^2 and how much data should be involved in the calculation of the posterior). But even after splitting of the sample into

Y and Y', a possible source of problems remains: uncertainty quantification based on an empirical-Bayes posterior accounts for the uncertainty due to the random nature of Y, but not that of Y', potentially leading to credible sets that are too small. In case we do not split the sample, the problem is aggravated by the fact that the amount of data available for calculation of the posterior is larger, leading to credible sets that are smaller, suggesting even less uncertainty.

A sophisticated application of the *empirical Bayes* idea is the estimation of hyperparameters by maximum-likelihood applied to the prior predictive distribution. Recall that the prior expectation of the distribution of the data (see definition 2.1.4) "predicts" the marginal distribution of the data. This prediction may be reversed to decide which value for the hyperparameter leads to the best explanation of the observed data based on the likelihood principle. More formally, denote the data by Y (taking values in a measurable space $(\mathscr{Y},\mathscr{B})$) and denote the model by $\mathscr{P} = \{P_\theta : \theta \in \Theta_0\}$. Consider a family of priors parametrized by a hyperparameter $\eta \in H$, $\{\Pi_\eta : \eta \in H\}$. For every η, the prior predictive distribution P_η is given by:

$$P_\eta(A) = \int_\Theta P_\theta(A)\, d\Pi_\eta(\theta),$$

for all $A \in \mathscr{B}$. In this way we obtain a new model for the observation Y, given by $\mathscr{P}' = \{P_\eta : \eta \in H\}$, contained in the closed convex hull of the original model (see theorem 2.2.5). Note that this new model is parametrized by the hyperparameter; hence if we close our eyes to the rest of the problem and we follow the maximum-likelihood procedure for estimation of η in this new model, we find the value of the hyperparameter that best explains the observation Y. Assuming that the model $\mathscr{P}'$ is dominated, with densities $\{p_\eta : \eta \in H\}$, the maximum-likelihood estimate is found as the point $\hat{\eta}(Y) \in H$ such that

$$p_{\hat{\eta}}(Y) = \sup_{\eta \in H} p_\eta(Y).$$

under the assumptions of existence and uniqueness, by the usual methods for maximum-likelihood estimation.

Definition 3.4.1. The estimator $\hat{\eta}(Y)$ is called the *ML-II estimator*, provided it exists and is unique.

Remark 3.4.2. There is one caveat that applies to the ML-II approach: in case the data Y consists of an *i.i.d.*-distributed sample, the prior predictive distribution describes the sample as exchangeable, but not *i.i.d.*! Hence, comparison of prior predictive distributions with the data suffer from the objection raised in remark 2.1.19. The frequentist who assumes that the true, underlying distribution P_0^n of the sample is *i.i.d.*, has to keep in mind that the ML-II model is misspecified.

3.4.1 Model selection with empirical methods

A situation where empirical Bayes methods are often used, is in *model selection*: suppose that there are several models $\mathscr{P}_1, \mathscr{P}_2, \ldots$ with priors $\Pi_1, \Pi_2, \ldots$, each of which may serve as a reasonable explanation of the data, depending on an unknown parameter $K \in \{1, 2, \ldots\}$. The choice to use model-prior pair $(\mathscr{P}_k, \Pi_k)$ in the determination of the posterior is made after estimation of K. Where the Bayesian chooses a hyperprior for the hyperparameter K, frequentist ways to estimate K leads to empirical Bayes methods.

Example 3.4.3. Consider the situation where we are provided with an *i.i.d.* sequence of specimens from a population that is divided into an unknown, finite number of classes K. All we know about the classes is that they occur with equal probabilities in the population. For each specimen the (random) class L is unknown, all we observe is a real-valued X, where it is assumed that $X|L = l$ is normally distributed conditional on the class l, with unknown mean $\mu_l \in \mathbb{R}$ and known variance (which we normalize to 1). Then each observation X is distributed according to a *discrete mixture* of normal distributions,

$$X|K, \mu \sim P_{K;\mu_1,\ldots,\mu_K} = \frac{1}{K} \sum_{l=1}^{K} N(\mu_l, 1),$$

where the components of $\mu = (\mu_1, \ldots, \mu_K) \in \mathbb{R}^K$ satisfy $\mu_1 < \ldots < \mu_K$, to maintain identifiability of the parametrization. For every value of $K \geq 1$, we have a model of the form,

$$\mathscr{P}_K = \left\{ P_{K;\mu_1,\ldots,\mu_K} : (\mu_1, \ldots, \mu_K) \in \mathbb{R}^K, \mu_1 < \ldots < \mu_K \right\}$$

Each of these models can be endowed with a prior Π_K on $\mathbb{R}^K$, for example, by drawing an *i.i.d.* sample $\mu'_1, \ldots, \mu'_K$ from the standard normal distribution and ordering the results: $\mu_k = \mu'_{(k)}$, $(1 \leq k \leq K)$.

At this point, a Bayesian would choose a hyperprior Π_2 for the discrete hyperparameter $K \geq 1$ and proceed to calculate the posterior using *all models* $\mathscr{P}_k$, weighed by the prior masses $\Pi_2(K = k)$ for all $k \geq 1$, in accordance with the methods of section 3.3. Alternatively, the Bayesian can (split the sample and) estimate K with some estimator $\hat{K}$ (like a frequentist would), to analyse the posterior using *only one model* $\mathscr{P}_{\hat{K}}$, *c.f.* the methods of this section. To estimate K various methods exist: inspection of the data may reveal which number of classes is most appropriate if there are clearly separated peaks in the observations. Otherwise, posterior odds (based on the right prior for K, see below) or frequentist *clustering methods* exist to estimate K.

But estimation of K to select one of the models $\mathscr{P}_K$ is a difficult statistical problem: maximization of the likelihood with an unbounded number of classes picks a number of classes equal to (or in the order of) the sample-size, simply because assigning each data-point its own class leads to the largest likelihood function. A

similar phenomenon arises in regression, where it is called *overfitting*: if we allow regression polynomials of arbitrary degree, maximization of the likelihood fits the data perfectly by choosing a polynomial of degree equal to the sample-size. The fit is perfect, residuals are zero and any associated measure for quality (like R^2) will reflect this. But we are no longer doing statistics because we are not distinguishing signal from noise (in fact, we have interpreted all noise as signal and the fit reflects this). Using such a fit in practice (for example to predict the distribution of another independent sample) leads to bad results because even though the second sample has the same underlying signal, noise differs and the fit does not anticipate this. Both for estimation of the number of mixture components and of the degree of a regression polynomial, one would like to have a sensible way to *regularize* the estimate for the number of clusters K, and then estimate $\mu_1, \ldots, \mu_k$.

In such questions of *model selection*, penalized likelihood criteria are employed which favour smaller choices for K over larger ones. Note that it is not clear, neither intuitively nor mathematically, how the penalty should depend on K or on the sample-size, nor which proportionality between penalty and likelihood is appropriate. A well-known standard choice comes in the form of the so-called *Akaike information criterion (AIC)* for model selection [229]: it argues for maximization of the (k-dependent) likelihood *minus* twice the dimension of the k'th parameter space (here $2k$), motivated from information theory and large sample sizes. The Bayesian faces the same problem when he chooses a prior for K: if he assigns too much prior weight to the higher-dimensional models, his estimators (or, equivalently, the bulk of the resulting posterior's mass) will get the chance to "run off" to infinity with growing samplesize, indicating inconsistency from over-fitting. The so-called *Bayesian information criterion (BIC)* [229] weighs the AIC penalty by the logarithm of the sample size, motivated by the Bernstein-von Mises limit of chapter 4, maximizing likelihood minus $2k\log(n)$. Indeed, the correspondence between the frequentist's necessity for a penalty in maximum-likelihood methods on the one hand, and the Bayesian's need for a prior expressing sufficient bias for the lower-dimensional model choices on the other, is explained in remark 2.2.21.

It is difficult to indicate which regularization method is preferred, as long as the argument is to be made for each sample-size $n \geq 1$ separately. Matters organise themselves in the large-sample limit, where one would like to select the model *consistently*: if we observe larger and larger *i.i.d.* samples $X^n = (X_1, X_2, \ldots, X_n)$, with each X_i distributed like X above marginally, for some unobserved value $K = k$, we would like to have a model selection method that selects the correct number of clusters k *with probability growing to one as* $n \to \infty$.

Example 3.4.4. In part II, we shall see that in the model of example 3.4.3, consistent selection of K is possible, if we restrict the model to consist of an upper-bounded number of clusters and the locations μ_i all lie in a fixed, compact subset of $\mathbb{R}$, with some fixed minimal distance between them. We can summarize these requirements in terms of a single integer $M \geq 1$ such that, $1 \leq K \leq M$ and,

$$\mathscr{P}_l = \left\{ P_{l;\mu_1,\ldots,\mu_l} : (\mu_1, \ldots, \mu_l) \in [-M, M]^l, \mu_1 < \ldots < \mu_l, \mu_{i+1} - \mu_i > 1/M \right\}.$$

Then any convex combination of priors Π_l that are of full support on their respective submodels $\mathscr{P}_l$,

$$\Pi = \sum_{l=1}^{M} \pi_l \, \Pi_l,$$

for $0 < \pi_1, \ldots, \pi_M < 1$ such that $\sum_{l=1}^{M} \pi_l = 1$, will lead to a sequence of posteriors on $\mathscr{P} = \cup_{l=1}^{M} \mathscr{P}_l$ that concentrate all mass in the correct component $\mathscr{P}_k$ with probability growing to one as $n \to \infty$: consequently, posterior odds can be used to model select consistently. The restriction that all classes are represented in equal numbers in the population is not necessary (although consistent selection with posterior odds requires a minimal value $1/M$ for each of the fractions). And the question also arises, what if we use upper bounds M_n that grow larger with growing sample-size n? How fast can M_n go to infinity, while still achieving a consistent posterior?

3.4.2 Bias and the James-Stein estimator

What is clear in the clustering and regression examples, is that model selection can also be viewed as *correction of a bias* inherent to our estimation method, a bias towards models with a high number of clusters or high order of a regression polynomial. Such views are particularly fruitful in the Bayesian case, because often, Bayesian point estimators that are expectations with respect to the posterior like the posterior predictive distribution $P^{\Pi|Y}$, (but also the posterior mean of a parameter) can be decomposed into an unbiased, consistent estimate $\hat{P}(Y)$ and a *bias* ascribed to the prior, like the prior predictive distribution P^{Π} (but also the prior mean of a parameter),

$$P^{\Pi|Y} = (1-\lambda)\,\hat{P}(Y) + \lambda\, P^{\Pi}. \tag{3.11}$$

Refer to decomposition (8.8) for an example in the context of the Dirichlet process prior and posterior. If with growing samplesize n we have $\lambda_n \to 0$, then the posterior predictive distribution follows the unbiased, frequentist estimate asymptotically (and will be consistent if $\hat{P}_n$ is). There are also cases where λ_n does not go to zero and bias persists in the limit, leading to *inconsistency* of the Bayesian estimator (see [65] for examples with so-called Gibbs-type priors).

With empirical Bayes methods to estimate which prior predictive distribution (or prior mean parameter value) is most appropriate, however, the inherent prior bias in (3.11) may be repairable: if we use the data to de-bias the prior predictive distribution P^{Π} itself, such problems can be mitigated or eliminated altogether.

Example 3.4.5. (Univariate normal mean) Consider the simpler case of $X_1, X_2, \ldots$ that are *i.i.d.*-$N(\theta, \sigma^2)$-distributed (with known $\sigma^2 > 0$) and a normal, non-central prior for the parameter $\theta \in \mathbb{R}$, that is, $\Pi = N(\alpha, \tau^2)$ for some $\alpha \in \mathbb{R}$ and $\tau^2 > 0$. The posterior distribution is again a normal distribution (see section 3.5) and it is easily seen that the posterior mean is,

$$\begin{aligned}\overline{\theta}_n(X_1,\ldots,X_n) &= \frac{\sigma^2\alpha+\tau^2\sum_{i=1}^n X_i}{\sigma^2+n\tau^2}\\ &= \frac{n\tau^2}{\sigma^2+n\tau^2}\overline{X}_n+\frac{\sigma^2}{\sigma^2+n\tau^2}\alpha.\end{aligned} \tag{3.12}$$

Note that the sample average $\overline{X}_n$ is an unbiased, consistent estimator for the location θ, while the prior expectation α introduces a bias. As $n\to\infty$, the difference between the posterior mean and the sample average goes to zero, so we conclude that the bias introduced by the prior vanishes asymptotically.

To *de-bias* $\overline{\theta}_n$ not only in the large-sample limit, we use empirical Bayes and estimate α from an independent *i.i.d.* sample $X'_1,\ldots,X'_n$, with the sample average,

$$\hat{\alpha}_n(X'_1,\ldots,X'_n)=\overline{X}'_n,$$

then both contributions in (3.12) are unbiased *at finite values of n*. Such *de-biasing* of the posterior mean with empirical methods gives rise to point-estimators that are optimal according to theorem 2.2.13 if they are quadratically integrable.

So, in case that posterior predictive distribution or mean has a bias, empirical Bayes methods can be used to correct. That idea is applied somewhat unexpectedly in the following example.

Example 3.4.6. (Multivariate normal mean) Suppose that $d\geq 3$ and we consider a data vector $Y=(Y_1,\ldots,Y_d)$ with components Y_i that are modelled as independent, and distributed according to a multivariate normal distribution with a covariance matrix that is a known multiple ($\sigma^2>0$) of the identity,

$$Y_i|\theta\sim N(\theta_i,\sigma^2),$$

for each $1\leq i\leq d$. A moment's thought shows that the ML estimator for θ based only on Y, is given by $\hat{\theta}_{ML}(Y)=Y$. A prior for the parameter $\theta\in\mathbb{R}^d$ is chosen as follows: we view the components $(\theta_1,\ldots,\theta_d)$ as an *i.i.d.* sample from the one-dimensional normal distribution $N(\mu,\tau^2)$ with hyperparameters $\mu\in\mathbb{R}$ and $\tau^2\geq 0$ ($\tau^2=0$ corresponding to a prior distribution degenerate at $\theta=\mu$). The prior predictive distribution for the data vector Y, given μ,τ^2, has Lebesgue density,

$$p^{\Pi}_{\mu,\tau^2}(y_1,\ldots,y_d)=\prod_{j=1}^d\frac{1}{\sqrt{2\pi}(\sigma^2+\tau^2)^{1/2}}\exp\Big(-\frac{1}{2}\frac{(y_j-\mu)^2}{\sigma^2+\tau^2}\Big).$$

The ML-II method prescribes that we maximize the prior predictive likelihood $p^{\Pi}_{\mu,\tau^2}(Y'_1,\ldots,Y'_d)$ based on an independent copy Y' of Y (in principle, but with $Y'=Y$ commonly). Analysing $(\mu,\tau^2)\mapsto\log p^{\Pi}_{\mu,\tau^2}$, we find that μ- and τ^2-derivatives are given by:

$$\frac{\partial}{\partial\mu}\log p^{\Pi}_{\mu,\tau^2}(Y') = \frac{1}{\sigma^2+\tau^2}\sum_{j=1}^{d}(Y'_i-\mu),$$
$$-2\frac{\partial}{\partial\tau^2}\log p^{\Pi}_{\mu,\tau^2}(Y') = \frac{d}{\sigma^2+\tau^2} - \frac{d\big(s_d^2(Y')+(\overline{Y'}-\mu)^2\big)}{(\sigma^2+\tau^2)^2},$$

with $\overline{Y'} = d^{-1}\sum_j Y'_i$ and $s_d^2(Y') = d^{-1}\sum_j (Y'_j-\overline{Y'})^2$. Solving for μ and substituting, we find ML-II estimates for the hyperparameters,

$$\hat{\mu}(Y') = \overline{Y'}, \quad \hat{\tau}^2(Y') = \max\{0, s_d^2(Y') - \sigma^2\},$$

(with $\hat{\tau}^2(Y') = 0$ signifying degeneracy at $\theta = \hat{\mu}(Y')$). Essentially, the resulting empirical prior imposes, for each component θ_j, $1 \le j \le d$, a bias towards the average value $\overline{Y'}$ of the observed components Y'_1 through Y'_d with two distinct cases. When $s_d^2(Y') \ge \sigma^2$, differences between (observed) components are relatively large and the prior is normal located at $\hat{\mu}(Y')$ with a variance that adds with σ^2 to the observed $s_d^2(Y')$; when $s_d^2(Y') < \sigma^2$, differences between observed components $Y'_1,\ldots,Y'_d$ are relatively small, indicating that their average $\overline{Y'}$ may be informative for estimation of the means θ_i of individual components of Y. In the first case, the bias formulated by the prior expectation $\overline{Y'}$ is mitigated by a prior variance that leaves room for doubt; in the second case, the prior is concentrated all the way on $\delta_{\theta=\overline{Y'}}$ (as in example 2.1.13). This is reflected in the posterior: when $s_d^2(Y') \ge \sigma^2$, the empirical Bayes posterior for the j-th component of θ is (see (3.7)),

$$\begin{aligned}\Pi(\theta\in A|Y,\mu=\hat{\mu}(Y'),\tau^2=\hat{\tau}^2(Y')) \\ = N_d\big(\hat{\lambda}(Y')\overline{Y'} + (1-\hat{\lambda}(Y'))Y, (1-\hat{\lambda}(Y'))\sigma^2\big)(A),\end{aligned}$$

where $\hat{\lambda}(Y') = \sigma^2/s_d^2(Y')$; when $s_d^2(Y') < \sigma^2$, all posterior mass is *shrunk* into one point (see example 2.1.13). Re-combining both cases, we decompose the empirical Bayes posterior expectation in the form ($1 \le j \le d$),

$$\tilde{\theta}_j(Y;Y') = \big(1-\hat{\kappa}(Y')\big)\overline{Y'} + \hat{\kappa}(Y')Y_j, \tag{3.13}$$

where $\hat{\kappa}(Y') = \max\{0, 1-\sigma^2/s_d^2(Y')\}$. As said, it is customary not to split the sample and use Y also in the role of Y'. In that case write $\tilde{\theta}(Y) = (1-\hat{\kappa}(Y))\overline{Y} + \hat{\kappa}(Y)Y$.

It came as a great surprise that the empirical Bayes estimator $\tilde{\theta}(Y)$ outperforms the maximum-likelihood estimate $\hat{\theta}(Y)$ and all other unbiased estimators for the problem with respect to expected squared-error loss $L(\theta,\theta') = \|\theta-\theta'\|^2 = \Sigma_j(\theta_j - \theta'_j)^2$ (see Efron and Morris (1973) [86]). In fact, a slightly different estimator that shrinks the unbiased estimate towards the sample average was written down without reference to any Bayesian methods in Stein's 1956 work [234]: for $d \ge 3$, what is now known as the *James-Stein estimator* is the *shrinkage estimator*,

$$\hat{\theta}_{JS}(Y) = \Big(1 - \frac{(d-2)\sigma^2}{s_d^2}\Big)(Y-\overline{Y}) + \overline{Y}. \tag{3.14}$$

It was shown in James and Stein (1961) [133] that $\hat{\theta}_{JS}$ is risk-better than the usual estimates: shockingly, for all $\theta \in \mathbb{R}^d$,

$$P_\theta \|\hat{\theta}_{JS} - \theta\|^2 \leq P_\theta \|\hat{\theta} - \theta\|^2,$$

with *strict* inequality for most values of θ (for proof, see corollary 4.7.2 in Lehmann and Casella (1998) [169]). The Lehmann-Scheffé theorem even amplifies this to the assertion that the James-Stein estimator is risk-better than any unbiased estimator, *i.e.* unbiased estimators are *inadmissable* (see definition 2.5.4 and example 2.5.10). It was recognized in [86] that point-estimators that result from empirical Bayes posteriors introduce the type of bias that the James-Stein estimator has and, correspondingly, outperform unbiased estimators in the given example. Indeed, if we use an unbiased estimator for $\sigma^2/(\sigma^2+\tau^2)$ rather than the ML-II estimate (see problem 4.7.1 in [169]), the resulting empirical Bayes estimator coincides with the James-Stein estimator. Because the ML estimator for this problem is optimal with respect to mean squared-error within the class of unbiased estimators, *c.f.* theorem 2.2.13, it is called *efficient*. Correspondingly, the James-Stein and empirical Bayes estimators are called *superefficient*.

To provide some counterweight to that remarkable conclusion, let us consider some of the drawbacks of shrinkage estimation: first of all, the improvement occurs *only* if we assess performance using the d-dimensional mean squared-error. For example, it can be shown that the James-Stein estimator estimates individual components of θ with *larger* errors than the non-shrunk estimate [169]. Secondly, the squared-error *Bayes* risk function (combine definition 2.5.12 and example 2.5.10) for shrinkage estimators is generally *higher* than that of their non-shrunk versions (for a discussion in the context of the multivariate mean problem discussed above, see example 4.7.3 in [169]). This may be explained by the fact that risk functions of estimators with fixed points of shrinkage tend to display wild fluctuations around the point of shrinkage (although this phenonemon is well-understood only if the model is one-dimensional [171]). Finally, posterior variance tends to be shrunk as well (see (3.7)), which leads to credible sets (and confidence sets based on the James-Stein estimator) that are too small. This can be understood from the fact that the empirical Bayes posterior does not account for the inaccuracies in the estimation of τ^2, it only quantifies the uncertainty in the subsequent estimation of θ, thus underestimating the overall error. Notwithstanding their practical usefulness, this is perhaps the most serious short-coming of shrinkage estimators: improved estimation accuracy comes at the cost of impaired uncertainty quantification and other forms of inference beyond point estimation.

3.5 Conjugate families

In this section, we consider a type of prior choice that is motivated primarily by mathematical convenience, rather than philosophy or statistical utility. Recall that

if we model the data with normal distributions of known variance but unknown location θ and we supply θ with a normal prior, then the posterior for θ is again a normal distribution. Since the calculation of the posterior is tractable, any choice for the parameters of the normal prior can immediately be updated to values for location and variance of the normal posterior upon observation of $Y = y$. Not only does this signify ease of manipulation in calculations with the posterior, it also reduces the computational burden dramatically since numerical integration or simulation from the posterior is no longer necessary.

3.5.1 Basic definition with an example

The subject of this section revolves around the following definition.

Definition 3.5.1. Let $(\mathscr{P}, \mathscr{A})$ be a measurable model for an observation $Y \in \mathscr{Y}$. Let M denote a collection of probability distributions on $(\mathscr{P}, \mathscr{A})$. The set M is called a *conjugate family* for the model $\mathscr{P}$, if the posterior based on a prior from M again lies in M:

$$\Pi \in M \quad \Rightarrow \quad \Pi(\cdot|Y = y) \in M, \tag{3.15}$$

for almost all $y \in \mathscr{Y}$.

(Like before, the phrase "almost all" in the above definition refers to the prior predictive distribution for Bayesians, and to the true P_0 for frequentists, taking into account condition (2.12).) Such structure was first proposed by Raiffa and Schlaifer (1961) [212]. Their method for the prior choice is usually classified as objectivist because it does not rely on subjectivist notions and is motivated without reference to outside factors.

Remark 3.5.2. Often in the literature, a prior is refered to as a *"conjugate prior"* if the posterior is of the same form. This is somewhat misleading, since it is the family M that is closed under conditioning on the data Y, a property that depends on the model and on M itself, but *not* on the particular $\Pi \in M$.

Example 3.5.3. Consider an experiment in which we observe n independent Bernoulli trials and consider the total number of successes, $Y \sim \text{Bin}(n, p)$ with unknown parameter $p \in [0, 1]$,

$$P_p(Y = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

For the parameter p we choose a prior $p \sim \text{Beta}(\alpha, \beta)$ from the Beta-family, for some $\alpha, \beta > 0$,

$$d\Pi(p) = B(\alpha, \beta)\, p^{\alpha-1}(1-p)^{\beta-1}\, dp,$$

where $B(\alpha, \beta) = \Gamma(\alpha+\beta)/(\Gamma(\alpha)\Gamma(\beta))$ normalizes Π. Then the posterior density with respect to the Lebesgue measure on $[0, 1]$ is proportional to:

$$d\Pi(p|Y) \propto p^Y (1-p)^{n-Y}\, p^{\alpha-1}(1-p)^{\beta-1}\, dp = p^{\alpha+Y-1}(1-p)^{\beta+n-Y-1}\, dp,$$

We conclude that the posterior again lies in the Beta-family, with parameters equal to a data-amended version of those of the prior, as follows:

$$\Pi(\cdot|Y) = \text{Beta}(\alpha + Y, \beta + n - Y).$$

So the family of Beta-distributions is a conjugate family for the binomial model. Depending on the available amount of prior information on θ, the prior's parameters may be chosen on subjective grounds. However, in the absence thereof, the parameters α, β suffer from the same ambiguity that plagues the parameter τ^2 featuring in the example with which we opened this section.

Example 3.5.3 indicates a strategy to find conjugate families for a given parametrized, dominated model $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$. Customarily, we view densities $y \mapsto p_\theta(y)$ as functions of the outcome $Y = y$ but they are functions of the parameter θ as well and their dependence $\theta \mapsto p_\theta(y)$ determines which prior densities $\theta \mapsto \pi(\theta)$ preserve their functional form when multiplied by the likelihood $p_\theta(Y)$ to yield the posterior density.

3.5.2 Exponential families

Although we shall encounter an example of a conjugate family for a non-parametric model in subsection 8.2.2, conjugate families are mostly part of parametric statistics. Many models are so-called exponential families, for which conjugate families of priors are found readily.

Definition 3.5.4. A Lebesgue-dominated collection of probability measures $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ (with densities p_θ) is called a k-parameter *exponential family*, if there exists a $k \geq 1$ such that for all $\theta \in \Theta$,

$$p_\theta(x) = \exp\Big(\sum_{i=1}^{k} \eta_i(\theta)\, T_i(x) - B(\theta)\Big) h(x), \tag{3.16}$$

where h and T_i, $i = 1, \ldots, k$, are statistics and B, η_i, $i = 1, \ldots, k$ are real-valued functions on Θ.

Any exponential family can be parametrized such that the exponent in (3.16) is linear in the parameter: by the mapping $\Theta \to H : \eta_i = \eta_i(\theta)$ (a bijection if the original parametrization is identifiable), taking Θ into $H = \eta(\Theta)$ and B into $A(\eta) = B(\theta(\eta))$, any exponential family can be rewritten in its so-called canonical form.

Definition 3.5.5. An exponential family $\mathscr{P} = \{P_\eta : \eta \in H\}$, $H \subset \mathbb{R}^k$ is said to be in its *canonical representation*, if

$$p_\eta(x) = \exp\Big(\sum_{i=1}^{k} \eta_i\, T_i(x) - A(\eta)\Big)\, h(x). \tag{3.17}$$

In addition, $\mathscr{P}$ is said to be *of full rank* if the interior of $H \subset \mathbb{R}^k$ is non-void, *i.e.* $\mathring{H} \neq \varnothing$.

Although they are parametric models, exponential families are versatile modelling tools and have properties that are mathematically tractable; many common models, like the Bernoulli-, normal-, binomial-, Gamma-, Poisson-models, *etcetera*, can be rewritten in the form (3.16). To give an example of a type of parameter that cannot be accommodated in an exponential family, consider models in which the support of model distributions is parameter-dependent, like the family of all uniform distributions on $\mathbb{R}$, or the parameter that describes the domain offset in the Pareto-model.

The statistical practicality stems primarily from the fact that for an exponential family of full rank, the statistics T_i, $i = 1, \ldots, k$ are sufficient and complete, enabling the use of the Lehmann-Scheffé theorem, (theorem 2.2.13) for minimal-variance unbiased estimation. Their versatility can be understood in many ways, *e.g.* by the Pitman-Koopman-Darmois theorem (see, Jeffreys (1961) [137]; or Robert (2001) [218]): a family of distributions with parameter-independent supports is exponential, if and only if in the models describing its *i.i.d.* samples, there exist sufficient statistics whose dimension remains bounded asymptotically.

Example 3.5.6. The model of all normal distributions $\mathscr{P} = \{N(\mu, \sigma^2) : \theta \in \mathbb{R}, \sigma^2 > 0\}$ on $\mathbb{R}$ forms an exponential family. To see this, write $\theta = (\mu, \sigma^2) \in \Theta = \mathbb{R} \times (0, \infty)$ and rewrite the usual parametrization in the form (3.16), as follows,

$$\begin{aligned} p_{\mu,\sigma^2}(x) &= (2\pi)^{-1/2}\sigma^{-1} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big) \\ &= (2\pi)^{-1/2} \exp\Big(-\frac{1}{2\theta_2}x^2 + \frac{\theta_1^2}{\theta_2}x - \frac{\theta_1}{2\theta_2} - \frac{1}{2}\log\theta_2\Big), \end{aligned}$$

and, comparing with (3.16), we read off,

$$\eta_1(\theta) = \frac{\theta_1}{\theta_2}, \quad \eta_2(\theta) = -\frac{1}{2\theta_2}, \quad B(\theta) = \frac{\theta_1^2}{2\theta_2} + \frac{1}{2}\log\theta_2,$$
$$T_1(x) = x, \quad T_2(x) = x^2, \quad h(x) = (2\pi)^{-1/2}.$$

The map $\eta : \Theta \to H : \theta \mapsto (\eta_1, \eta_2)(\theta)$ takes the original parameter into the canonical parameter $\eta \in H = \mathbb{R} \times (-\infty, 0)$. Note that the inverse of η takes the form,

$$(\theta_1, \theta_2)(\eta) = \Big(-\frac{\eta_1}{2\eta_2}, -\frac{1}{2\eta_2}\Big),$$

from which we deduce that,

$$A(\eta) = B(\theta(\eta)) = -\frac{\eta_1^2}{4\eta_2} + \frac{1}{2}\log\big(-\frac{1}{2\eta_2}\big),$$

for the new normalization. Expressed in these new parameters η, the density takes the form (3.17). Note that $H = \mathbb{R} \times (-\infty, 0)$ has non-empty interior, so the normal

model is an exponential family of full rank. In case we had started with the model $\mathscr{P} = \{N(\theta,\theta) : \theta > 0\}$, for example, the analysis would have been largely analogous; however, the latter $\mathscr{P}$ is *not* of full rank.

Presently our interest lies with the following theorem which says that if a model $\mathscr{P}$ constitutes an exponential family, there exists a conjugate family of priors for $\mathscr{P}$.

Theorem 3.5.7. *Let $\mathscr{P}$ be a model that can be written as an exponential family, c.f. definition 3.5.4. In its canonical parametrization (3.17), $\mathscr{P}$ and the family of distributions $\Pi_{\mu,\lambda}$, defined by Lebesgue probability densities*

$$\pi_{\mu,\lambda}(\eta) = K(\mu,\lambda)\exp\Big(\sum_{i=1}^{k}\eta_i\mu_i - \lambda A(\eta)\Big), \tag{3.18}$$

(where $\mu \in \mathbb{R}^k$ and $\lambda \in \mathbb{R}$ are such that $0 < K(\mu,\lambda) < \infty$), is a conjugate family for $\mathscr{P}$: the posterior associated with prior $\Pi_{\mu,\lambda}$ is $\Pi_{\mu+T(X),\lambda+1}$.

Proof. Parametrize $\mathscr{P}$ as in (3.17). Choosing a prior on H of the form (3.18), we find that the posterior again takes the form (3.18),

$$\pi(\eta|X) \propto \exp\Big(\sum_{i=1}^{k}\eta_i(\mu_i + T_i(X)) - (\lambda+1)A(\eta)\Big)$$

(the factor $h(X)$ arises both in numerator and denominator of (2.6) and is η-independent, so that it cancels). The data-amended versions of the parameters μ and λ that emerge from the posterior are therefore given by:

$$(\mu + T(X), \lambda + 1),$$

and we conclude that the distributions $\Pi_{\mu,\lambda}$ form a conjugate family for $\mathscr{P}$.

3.6 Dirichlet priors

In this section we consider any sample space $\mathscr{X}$ of finite cardinal and priors on the space of all probability distributions on $\mathscr{X}$. Not only does this serve as the full model for a random observation that can take only a finite number of values and the parameter space for the corresponding multinomial distributions, it also serves as the building block for the construction of a class of priors on non-parametric models, as illustrated by the Dirichlet process priors of section 8.2.

Let $\mathscr{X} = \{1,2,\ldots,k\}$ (with its powerset $2^{\mathscr{X}}$ as a σ-algebra) and consider the collection $M^1(\mathscr{X})$ of all probability measures on $\mathscr{X}$. Every $P \in M^1(\mathscr{X})$ has a density $p : \mathscr{X} \to [0,1]$ (with respect to the counting measure on $\mathscr{X}$) and we denote $p_i = p(i) = P(\{i\})$, so that for every $A \in 2^{\mathscr{X}}$, $P(A) = \sum_{l\in A} p_l$. Therefore, the space $M^1(\mathscr{X})$ can be parametrized as follows,

$$M^1(\mathscr{X}) = \Big\{P : 2^{\mathscr{X}} \to [0,1] : \sum_{i=1}^{k} p_i = 1,\, p_i \geq 0,\, (1 \leq i \leq k)\Big\},$$

and is in bijective correspondence with the *simplex* in $\mathbb{R}^k$ (see example 1.1.13). We are interested the following family of distributions on $M^1(\mathscr{X})$, which generalize the family of Beta distributions.

Definition 3.6.1. *(Finite-dimensional Dirichlet distribution)* Let $\mu = (\mu_1, \ldots, \mu_k)$ with $\mu_i \geq 0$ for all $1 \leq i \leq k$. A vector $p = (p_1, \ldots, p_k)$ satisfying $p_i \geq 0$ for all $1 \leq i \leq k$ and $\Sigma_i p_i = 1$, is said to have a *Dirichlet distribution* D_μ with parameter μ, if the density π for p satisfies:

$$\pi(p) = \frac{\Gamma\left(\sum_{l=1}^{k} \mu_l\right)}{\Gamma(\mu_1)\ldots\Gamma(\mu_k)} p_1^{\mu_1 - 1}\, p_2^{\mu_2 - 1} \ldots p_k^{\mu_k - 1}.$$

If $\mu_l = 0$ for some l, $1 \leq l \leq k$, then we set $D_\mu(p_l = 0) = 1$ marginally and we treat the remaining components of p as $(k-1)$-dimensional.

Example 3.6.2. Consider the case where $k = 2$ (so that $p_2 = 1 - p_1$): in that case, the density of the Dirichlet distribution takes the form:

$$\pi(p_1, p_2) = \frac{\Gamma(\mu_1 + \mu_2)}{\Gamma(\mu_1)\Gamma(\mu_2)} p_1^{\mu_1 - 1} (1 - p_1)^{\mu_2 - 1},$$

i.e. p_1 has a Beta distribution $B(\mu_1, \mu_2)$.

We also note the following two well-known facts on the Dirichlet distribution (proofs can be found in [111]).

Lemma 3.6.3. *(Gamma-representation of D_μ)*
If $Z_1, \ldots, Z_k$ are independent and each marginally Γ-distributed $Z_i \sim \Gamma(\mu_i, 1)$, $1 \leq i \leq k$, then with $S = \sum_{i=1}^{k} Z_i$,

$$\left(\frac{Z_1}{S}, \ldots, \frac{Z_k}{S}\right) \sim D_\mu, \tag{3.19}$$

i.e. the normalized vector has a Dirichlet distribution and is independent of S.

Lemma 3.6.3 shows that we may think of a D_μ-distributed vector as the L^1-projection of a vector composed of k independent, Γ-distributed components, onto the space $M^1(\mathscr{X})$ of probability distributions.

Lemma 3.6.4. *Let $\mathscr{X}$ be a finite point-set. If the density $p : \mathscr{X} \to [0,1]$ of a distribution P is distributed according to a Dirichlet distribution with parameter μ, $p \sim D_\mu$, then for any partition $\{A_1, \ldots, A_m\}$ of $\mathscr{X}$, the vector of probabilities $(P(A_1), P(A_2), \ldots, P(A_m))$ has a Dirichlet distribution,*

$$\big(P(A_1), P(A_2), \ldots, P(A_m)\big) \sim D_{\mu'},$$

where the parameter μ' is given by:

$$(\mu'_1,\ldots,\mu'_m) = \Big(\sum_{l\in A_1}\mu_l,\ldots,\sum_{l\in A_m}\mu_l\Big). \tag{3.20}$$

The identification (3.20) in lemma 3.6.4 suggests that we adopt a slightly different perspective on the definition of the Dirichlet distribution: we view μ as a *bounded measure* on $\mathscr{X}$, so that $P\sim D_\mu$, if and only if, for every partition $(A_1,\ldots,A_m)$,

$$\big(P(A_1),\ldots,P(A_m)\big)\sim D_{(\mu(A_1),\ldots,\mu(A_m))}. \tag{3.21}$$

Property (3.21) serves as the point of departure of the generalization to the non-parametric model, because it does not depend on the finite nature of $\mathscr{X}$ (see definition (8.5)).

Definition 3.6.5. Let $\mathscr{X}$ be a finite point-set; the *Dirichlet family* $\mathscr{D}(\mathscr{X})$ is defined to be the collection of all Dirichlet distributions on $M^1(\mathscr{X})$, *i.e.* $\mathscr{D}(\mathscr{X})$ consists of all D_μ with μ a bounded measure on $\mathscr{X}$.

Properties of Dirichlet distributions now follow and are listed as direct consequences in the following lemma.

Lemma 3.6.6. *Let μ be a bounded measure on a finite point-set $\mathscr{X}$ and let $B\subset\mathscr{X}$ be given. Then, if $\mu(B)=0$, then $P(B)=0$, D_μ-almost-surely; if $\mu(B)>0$, then $P(B)>0$, D_μ-almost-surely, and the D_μ-expectation of P is,*

$$\int P(B)\,dD_\mu(P) = \frac{\mu(B)}{\mu(\mathscr{X})}.$$

Proof. Consider the partition (B_1,B_2) of $\mathscr{X}$, where $B_1=B$, $B_2=\mathscr{X}\setminus B$. According to (3.21),

$$\big(P(B_1),P(B_2)\big)\sim D_{(\mu(B),\mu(\mathscr{X})-\mu(B))},$$

so that $P(B)\sim\text{Beta}(\mu(B),\mu(\mathscr{X})-\mu(B))$. Stated properties then follow from the properties of Beta distributions.

The following property of Dirichlet distributions describes two independent Dirichlet-distributed quantities in convex combination, which form a new Dirichlet-distributed quantity if mixed by means of an independent Beta-distributed parameter.

Lemma 3.6.7. *Let $\mathscr{X}$ be a finite point-set and let μ_1, μ_2 be two measures on $(\mathscr{X},2^{\mathscr{X}})$. Let (P_1,P_2) be independent and marginally distributed as*

$$P_1\sim D_{\mu_1},\quad P_2\sim D_{\mu_2}.$$

Furthermore, let λ be independent of P_1,P_2 and marginally distributed according to $\lambda\sim\text{Beta}(\mu_1(\mathscr{X}),\mu_2(\mathscr{X}))$. Then the convex combination $\lambda P_1+(1-\lambda)P_2$ again has a Dirichlet distribution with base measure $\mu_1+\mu_2$:

$$\lambda P_1+(1-\lambda)P_2\sim D_{\mu_1+\mu_2}.$$

The reason to choose Dirichlet distributions on $M^1(\mathscr{X})$ rather than some other parametric family, is the fact that they are conjugate for the multinomial model, which amounts to conjugacy of the Dirichlet family for the full model of *i.i.d.* observations in $\mathscr{X}$.

Theorem 3.6.8. *Let $\mathscr{X}$ be a finite sample space and let $X_1, \ldots, X_n$ denote an i.i.d. sample of observations in $\mathscr{X}$. The Dirichlet family $\mathscr{D}(\mathscr{X})$ is conjugate for the full model: if the prior equals D_μ, the posterior is a Dirichlet distribution D_{μ_n} with,*

$$\mu_n = \mu + \sum_{i=1}^{n} \delta_{X_i}, \tag{3.22}$$

as a base measure.

Proof. The posterior can be written as in (2.15) with the likelihood taking the form:

$$P \mapsto \prod_{i=1}^{n} p_{X_i} = \prod_{l=1}^{k} p_l^{N_l},$$

where N_l denotes the number of X_i equal to l, for all $1 \leq l \leq k$. Multiplying by the prior density for $\Pi = D_\mu$, we find that the posterior density is proportional to,

$$\pi(p|X_1, \ldots, X_n) \propto \pi(p) \prod_{i=1}^{n} p_{X_i} \propto \prod_{l=1}^{k} p_l^{N_l} \prod_{l=1}^{k} p_l^{\mu_l - 1} = \prod_{l=1}^{k} p_l^{\mu_l + N_l - 1},$$

which is again a Dirichlet density (but with changed base measure). Since the posterior is a probability distribution, we know that the normalization factor follows suit. Note that we may view N_l as the density of the measure,

$$N_l = \sum_{i=1}^{n} 1\{X_i = l\} = \sum_{i=1}^{n} \delta_{X_i}(\{l\}),$$

for every $1 \leq l \leq k$. So the posterior is the Dirichlet distribution D_{μ_n}, with base measure (3.22).

The posterior predictive distribution for a single, new observation is therefore given by,

$$\begin{aligned} P^{\Pi|X^n}(A) &= \int P(A)\, d\Pi(P|X^n) = \int P(A)\, dD_{\mu_n}(P) = \frac{\mu_n(A)}{\mu_n(\mathscr{X})} \\ &= \frac{\mu(A) + \sum_{i=1}^{n} \delta_{X_i}(A)}{\mu(\mathscr{X}) + n} = (1 - \lambda_n) \frac{1}{n} \sum_{i=1}^{n} 1(X_i \in A) + \lambda_n \frac{\mu(A)}{\mu(\mathscr{X})}, \end{aligned} \tag{3.23}$$

with $\lambda_n = \mu(\mathscr{X})(\mu(\mathscr{X}) + n)^{-1}$. Here, $\mathbb{P}_n(A) = n^{-1} \sum_i 1(X_i \in A)$ is an unbiased, consistent estimator for the true probability of $X \in A$, while $\mu(A)/\mu(\mathscr{X})$ represents the location of prior bias. The strength of this bias is controlled by $\mu(\mathscr{X})$, which

serves to controls how concentrated D_μ is around its location: if $\mu(\mathscr{X})$ is large compared to n, prior bias is strongly represented while the unbiased, purely-data-based estimator $\mathbb{P}_n$ is more muted; if $\mu(\mathscr{X})$ is small compared to n, prior bias becomes is less pronounced and posterior predictive distribution adopts more of the purely data-based estimator $\mathbb{P}_n$. As $n \to \infty$, $\lambda_n \to 0$ and $\mathbb{P}_n$ overwhelms all prior bias. At finite n, the empirical choice $\hat{\mu}_n(X^n) = \mathbb{P}_n$ *de-biases* the posterior predictive distribution in the sense that $P^n P^{\Pi|X^n}(A) = P(A)$ for every $A \subset \mathscr{X}$.

3.7 Exercises

3.7.1. A PROPER JEFFREYS PRIOR
Let X be a random variable, distributed $\mathrm{Bin}(n;p)$ for known n and unknown $p \in (0,1)$. Calculate Jeffreys prior for this model, identify a standard family of probability distributions that this prior would belong to, if it were normalized as a probability distribution.

3.7.2. JEFFREYS AND UNIFORM PRIORS
Let $\mathscr{P}$ be a model parametrized according to some mapping $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$. Assuming differentiability of this map, Jeffreys prior Π takes the form (3.6). In other parametrizations, the *form* of this expression remains the same, but the actual dependence on the parameter changes. This makes it possible that there exists another parametrization of $\mathscr{P}$ such that Jeffreys prior is *equal* to the uniform prior. We shall explore this possibility below.
For each of the following models in their 'standard' parametrizations $\theta \mapsto P_\theta$, find a parameter $\eta = \eta(\theta)$ with parameter space $H = \eta(\Theta)$, such that the density of the Jeffreys prior, expressed in terms of η, is constant. Also express model distributions in η-dependent form.

a. The model of all Poisson distributions,

$$P_\lambda(X = k) = p_\lambda(k) = e^{-\lambda}\frac{\lambda^k}{k!},$$

for $k \geq 0$, with unknown $\lambda > 0$.

b. The models of all $\Gamma(k,\theta)$-distributions with Lebesgue densities,

$$p_{k,\theta}(x) = \frac{1}{\Gamma(k)\theta^k}\, x^{k-1} \exp(-x/\theta),$$

for $x \geq 0$, with known $k > 0$ and unknown $\theta \in (0,\infty)$.

c. The model of all binomial distributions,

$$P_\theta(X = k) = p_\theta(k) = \binom{n}{k}\theta^k(1-\theta)^{n-k},$$

for $k \geq 0$, with known $n \geq 1$ and unknown $\theta \in (0,1)$.

To conclude, prove the following:

d. If a parametrization η like above exists and H is unbounded, Jeffreys prior is improper (in all parametrizations).

3.7.3. OPTIMALITY OF UNBIASED BAYESIAN POINT ESTIMATORS
Let $\mathscr{P}$ be a dominated, parametric model, parametrized identifiably by $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$, for some $\Theta \subset \mathbb{R}^k$. Assume that $(X_1,\ldots,X_n) \in \mathscr{X}^n$ form an *i.i.d.* sample from a distribution $P_0 = P_{\theta_0} \in \mathscr{P}$, for some $\theta_0 \in \Theta$. Let Π be a prior on Θ and denote the posterior by $\Pi(\cdot|X_1,\ldots,X_n)$. Assume that $T : \mathscr{X}^n \to \mathbb{R}^m$ is a sufficient statistic for the model $\mathscr{P}$.

a. Use the factorization theorem to show that the posterior depends on the data only through the sufficient statistic $T(X_1,\ldots,X_n)$.
b. Let $\hat{\theta}_n : \mathscr{X}^n \to \Theta$ denote a point-estimator derived from the posterior. Use *a.* above to argue that there exists a function $\tilde{\theta}_n : \mathbb{R}^m \to \Theta$, such that,

$$\hat{\theta}_n(X_1,\ldots,X_n) = \tilde{\theta}_n(T(X_1,\ldots,X_n)).$$

Bayesian point-estimators share this property with other point-estimators that are derived from the likelihood function, like the maximum-likelihood estimator and penalized versions thereof. Next, assume that T is complete, that $P_0^n(\hat{\theta}_n)^2 < \infty$ and that $\hat{\theta}_n$ is *unbiased*, *i.e.* $P_0^n\hat{\theta}_n = \theta_0$.

c. Apply the Lehmann-Scheffé theorem to prove that, for any other unbiased estimator $\hat{\theta}'_n : \mathscr{X}^n \mapsto \Theta$,

$$P_0^n(\hat{\theta}_n - \theta_0)^2 \leq P_0^n(\hat{\theta}'_n - \theta_0)^2.$$

The message of this exercise is, that Bayesian point-estimators that happen to be unbiased and quadratically integrable, are automatically L_2-*optimal* in the class of all unbiased estimators for θ. They share this remarkable property with maximum-likelihood estimators.

3.7.4. CONJUGATE MODEL-PRIOR PAIRS
In this exercise, conjugate model-prior pairs $(\mathscr{P},\Pi)$ are provided. In each case, we denote the parameter we wish to estimate by θ and assume that other parameters have known values. Let X denote a single-observation.
In each case, derive the posterior distribution to prove conjugacy and identify the X-dependent transformation of parameters that takes prior into posterior.

a. $X|\theta \sim N(\theta,\sigma^2)$ and $\theta \sim N(\mu,\tau^2)$, with known $\sigma^2 > 0$ and some choice for $\tau^2 > 0$.
b. $X|\theta \sim \text{Poisson}(\theta)$ and $\theta \sim \Gamma(\alpha,\beta)$, with some choice for $\alpha,\beta > 0$.
c. $X|\theta \sim \Gamma(\rho,\theta)$ and $\theta \sim \Gamma(\alpha,\beta)$, with known $\rho > 0$ and some choice for $\alpha,\beta > 0$.
d. $X|\theta \sim \text{Bin}(n;\theta)$ and $\theta \sim \beta(\alpha,\beta)$, with known $n \geq 1$ and some choice for $\alpha,\beta > 0$.

e. $X|\theta \sim N(\mu, \theta^{-1})$ and $\theta \sim \Gamma(\alpha, \beta)$, with known $\mu > 0$ and some choice for $\alpha, \beta > 0$.
f. $X|\theta_1, \ldots, \theta_k \sim M_k(n; \theta_1, \ldots, \theta_k)$ with known $k, n \geq 1$ and $\theta \sim D_\alpha$, where M_k denotes the multinomial distribution for n observations drawn from k classes with probabilities $\theta_1, \ldots, \theta_k$ and D_α is a Dirichlet distribution on the simplex in $\mathbb{R}^k$ (see definition 3.6.1; this proves again theorem 3.6.8).

3.7.5. In this exercise, we generalize the setup of example 3.3.4 to multinomial rather than binomial context. Let $k \geq 1$ be known. Consider an observed random variable Y and an unobserved $N = 1, 2, \ldots$, such that, conditionally on N, Y is distributed multinomially over k classes, while N has a Poisson distribution with hyperparameter $\lambda > 0$,

$$Y|N \sim M_k(N; p_1, p_2, \ldots, p_k), \qquad N \sim \text{Poisson}(\lambda).$$

Determine the prior predictive distribution of Y, as a function of the hyperparameter λ.

3.7.6. Let $X_1, \ldots, X_n$ form an *i.i.d.* sample from a Poisson distribution Poisson(θ) with unknown $\theta > 0$. As a family of possible priors for the Bayesian analysis of this data, consider exponential distributions $\theta \sim \Pi_\lambda = \text{Exp}(\lambda)$, where $\lambda > 0$ is a hyperparameter.

a. Calculate the prior predictive distribution for X.
b. Give the ML-II estimate $\hat{\lambda}$ for λ.
c. With the estimated hyperparameter, give the posterior distribution $\theta|X_1, \ldots, X_n$.
d. Calculate the posterior mean. Compare its data-dependence to that of the posterior mean we would have obtained if we had not made an empirical choice for the hyperparameter, but a fixed choice.

3.7.7. Let $X_1, \ldots, X_n$ form an *i.i.d.* sample from a binomial distribution Bin$(N; p)$, for known N and unknown $p \in [0, 1]$. For the parameter p we take a prior $p \sim \beta(\alpha, \beta)$ with hyperparameters $\alpha, \beta > 0$.

a. Show that the family of β-distributions is conjugate for binomial data.
b. Using (standard expressions for) the expectation and variance of β-distributions, give the posterior mean and variance in terms of the original α and β chosen for the prior and the data.
c. Calculate the prior predictive distribution and discuss the steps one would perform in the ML-II procedure to estimate p.

Some example exam problems

3.7.8. In 1814, Laplace asked the question, *"What is the probability p that the sun comes up tomorrow?"* In the following, we illustrate his Bayesian answer. The data

is based on binary $X_1,\ldots,X_n \in \{0,1\}$, denoting whether the sun came up ($X_i = 1$), or did not come up ($X_i = 0$) on day $1 \leq i \leq n$ in the observation period: the statistician observes only the total number of times the sun came up $Y := \sum_{i=1}^{n} X_i$. We assume that the $X_1,\ldots,X_n$ form an *i.i.d.* sample from the Bernoulli(p)-distribution, where $n \geq 1$ is known and $p \in [0,1]$ is the unknown parameter of interest.

a. Give the model distributions as densities with respect to the counting measure $q_p(k) := P(Y = k|p)$ for $0 \leq k \leq n$ and parameter $p \in [0,1]$.

Regarding the prior Π for p, we make the objectivist's choice and pick a uniform distribution: $p \sim U[0,1]$.

b. Calculate the posterior for p, given Y. To which parametric family of distributions does the posterior belong? In terms of the standard parametrization of this family, give the (Y-dependent) values of the parameters.
c. Give the posterior mean $\hat{p}_{1,n}$.

A subjectivist would argue that the above uniform prior ignores well-established prior knowledge concerning the parameter p: on all days outside the observation period, the sun has always come up. According to subjectivist standards, the prior for p should reflect that piece of information. *(Hint: A random variable Z has a beta distribution Beta(a,b) with $a,b > 0$, if $Z \in [0,1]$ and $P_{a,b}(Z \leq z) = B(a,b)^{-1} \int_0^z x^{a-1}(1-x)^{b-1}\,dx$, with normalization $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$.)*

d. Calculate the posterior mean $\hat{p}_{2,n}$ for a prior $\Pi = \text{Beta}(a,b)$, where the hyperparameters $a,b > 0$ are not fixed yet. Make a choice for the values of the parameters a,b that express the above, subjectivist expert knowledge.
e. Show that, regardless of the choice for the hyperparameters a,b, the difference between $\hat{p}_{1,n}$ and $\hat{p}_{2,n}$ goes to zero as the length of the observation period $n \to \infty$.

3.7.9. For some $n \geq 1$, let $X_1,\ldots,X_n$ form an *i.i.d.* sample from a Poisson distribution Poisson(θ), for some $\theta > 0$.

a. Give the Jeffreys prior for this model. Is this prior proper?
b. Give the posterior distribution for θ based on the Jeffreys prior of part *a.* Indicate to which standard family of distributions this posterior belongs and give the associated parameter values in terms of the sample size n and the observations $X_1,\ldots,X_n$.
c. Based on the posterior of part *b.*, give the posterior mean $\hat{\theta}_n$. View $\hat{\theta}_n$ as a point estimator for θ and determine what its bias is.

Assume that the sample mean $\overline{X}_n$ is a sufficient and complete statistic in the Poisson model.

d. Based on the conditions for the theorem of Lehmann-Scheffé, argue that the point estimator $\hat{\theta}'_n$, defined to be equal to posterior mean $\hat{\theta}_n$ minus its bias, is the unique minimal-variance unbiased estimator for θ.

3.7.10. Consider the so-called Galenshore distribution for Y with parameters $a > 0$ and $\theta > 0$, which is defined by the Lebesgue density:

$$p_{a,\theta}(y) = \frac{2}{\Gamma(a)}\theta^{2a}y^{2a-1}e^{-\theta^2 y^2},$$

for $y > 0$.

a. Let Y be distributed according to a Galenshore distribution with known a and unknown θ. Show that the family of Galenshore distributions for θ is a conjugate family in this model. Given Y and a Galenshore prior with hyperparameters $c, d > 0$, give the Galenshore parameters for the posterior.

3.7.11. In this problem, we consider exponential families.

a. What is the general form of a k-parameter exponential family. Express your answer by characterization of a collection of densities.
b. Give the canonical form of an exponential family. When do we say that an exponential family has *full rank*?

(Caution: in the following two parts, use the general or canonical form of an exponential family and do not choose some example.)

c. Give an exponential family in general, use the canonical form to write down a collection of distributions on the parameter space and show that this collection forms a conjugate family.
d. Given an exponential family in general, calculate the parameters of the posterior given a prior from the conjugate family of part *c.*.

Let $\mathscr{P}$ be a $k = 2$-parameter model of Lebesgue densities $p_{\alpha,\beta}$ with $\alpha, \beta > 0$, of the form,

$$p_{\alpha,\beta}(x) = \begin{cases} 0, & \text{if } x < \beta \\ \frac{\alpha\beta}{x^{\alpha+1}}, & \text{if } x \geq \beta \end{cases}$$

e. Is $\mathscr{P}$ an exponential family?

3.7.12. Consider a model in which we observe a sample of $X_1, \ldots, X_n$ that are independent but not identically distributed: for each X_i, there is a Binomial distribution for the sum of n Bernoulli trials with an i-dependent success-probability θ_i ($1 \leq i \leq n$). The prior for $\theta = (\theta_1, \ldots, \theta_n)$ will have a hyperparameter η,

$$X_i \,\big|\, \theta, \eta = X_i \,\big|\, \theta_i \sim \text{Bin}(n, \theta_i).$$

The parameters $\theta_1, \ldots, \theta_k$ form an *i.i.d.* sample from a Beta-prior with parameters equal to $\eta \in [0,1]$ and $(1-\eta)$. That means the vector of all θ_i has an n-fold product distribution,

$$(\theta_1, \ldots, \theta_n) \,\big|\, \eta \sim \text{Beta}(\eta, 1-\eta)^n.$$

For the hyperparameter η, we impose a uniform hyperprior,

$$\eta \sim U[0,1].$$

a. Show that the posterior mean for $\sum_i \theta_i$ equals,

$$\int \sum_{i=1}^{n} \theta_i \, d\Pi(\theta|X_1,\ldots,X_n) = \frac{n}{n+1}\left(\bar{X}_n + \hat{\eta}_n\right)$$

where $\hat{\eta}_n$ denotes the posterior mean for η. *(Hint: Start your calculation as if there were a fixed value of η. The result is interpreted as being conditional on η, and integration with respect to the posterior for η yields the result required.)*

Chapter 4
The Bernstein-von Mises theorem

Throughout the preceding chapters, we have occasionally looked at the behaviour of statistical methods for estimation, testing, uncertainty quantification and decision taking in the asymptotic limit, *i.e.* when the sample size goes to infinity. The asymptotic regime of statistical methods provides approximations to hard-to-obtain finite-sample results: while most finite-sample calculations are intractable even in the simplest models, the analysis of the large-sample limit often remains possible. The asymptotic answer may then be used as an *approximation* to the finite-sample answer. That perspective also dominates the developments of part II of this book, which deals with non-parametric models.

In this chapter we consider the large-sample behaviour of posterior distributions on *smooth parametric models* for *i.i.d.* sequences of data. Here, smoothness roughly says that we assume a dominated model parametrized by $\theta \in \Theta$, requiring that the dependence $\theta \mapsto \log p_\theta(X^n)$ of the likelihood function on the parameter is differentiable (see, however, definition 4.1.12). The frequentist asymptotic theory of estimation, testing and uncertainty quantification in smooth parametric models is well-understood: if X^n is distributed *i.i.d.*-P_{θ_0} for some true value θ_0 of the (k-dimensional) parameter, and the estimators $\hat{\theta}_n(X^n)$ belong to the family of so-called *regular* estimators (see definition 4.1.10), then the $n^{1/2}$-rescaled differences between estimators $\hat{\theta}_n(X^n)$ and θ_0 converge weakly to a limit described by Hajék's 1970 convolution theorem. Accordingly, the best possible regular estimators are those that satisfy,

$$\sqrt{n}\big(\hat{\theta}_n(X^n) - \theta_0\big) \xrightarrow{P_{\theta_0}\text{-w.}} N_k(0, I_{\theta_0}^{-1}),$$

where I_{θ_0} denotes the Fisher information at θ_0. Estimators with this limiting behaviour are called *efficient* and the limit distribution gives rise to Wald-type confidence ellipsoids centred on efficient estimators (see definition (4.4)), as well as test sequences that separate θ_0 from complements of ellipsoids of radii proportional to $(1 + o(1))n^{-1/2}$.

In section 4.2 we consider the Bernstein-von Mises theorem, which asserts that the sequence of posteriors on a smooth parametric model converges in total variation to a sequence of normal distributions centred on efficient point-estimators, with

covariance $(nI_{\theta_0})^{-1}$:

$$\sup_B \left| \Pi\big(\vartheta \in B \mid X_1, \ldots, X_n\big) - N(\hat{\theta}_n, (nI_{\theta_0})^{-1})(B) \right| \xrightarrow{P_{\theta_0}} 0, \tag{4.1}$$

where $(\hat{\theta}_n)$ denotes any efficient estimator sequence. The limit (4.1) concerns a relatively strong form of convergence, and correspondingly, permits refinement to the level of *uncertainty quantification*: sequences of *credible sets* are approximations of Wald-type, *efficient confidence sets* asymptotically. We compare this to asymptotic uncertainty quantification as in subsection 2.3.4, where enlargements of credible balls were shown to be asymptotically interpretable as confidence balls. To conclude this chapter we take a brief excursion to non-parametric setting: we consider the Bernstein-von Mises theorem for *semi-parametric* estimation problems (where a model of distributions $P_{\theta,\eta}$, ($\theta \in \Theta$ (parametric), $\eta \in H$ (non-parametric)) is proposed for the estimation of (only) the *parameter of interest* θ, in the presence of a *nuisance parameter* η).

Although the name of the central theorem of this chapter refers to the historical work of Bernstein (1917) [15] and von Mises (1931) [195], it is Le Cam (1953) [171] who truly deserves the credit for the present-day, general formulation. Certainly the most useful reference for this subject is Le Cam and Yang (1990) [183]. A version of the Bernstein-von Mises theorem based on Le Cam's inequality (see subsection 7.1.3) can be found in Le Cam (1986) [179].

4.1 Efficient estimation in smooth parametric models

First we consider frequentist estimation in smooth parametric models and state Hajék's convolution theorem, which characterizes *efficiency* of estimation. This paves the way for the Bernstein-von Mises theorem of the next section, which asserts that posterior distributions in smooth parametric models concentrate in an asymptotically normal way around efficient point-estimators. Essential to the development of efficient estimation are two concepts: smoothness of the model and regularity of the estimator. When properly defined and then combined, smoothness and regularity describe a notion of statistical optimality comparable (and related) to estimators that achieve minimal mean-squared error within the family of unbiased estimators in sense of Lehmann-Scheffé, *c.f.* theorem 2.2.13. By contrast, the analysis given here is, on the one hand, strictly asymptotic, but on the other, not limited to unbiased estimators. Before we specify to this setting, however, we briefly digress to introduce some generalities concerning asymptotic estimation.

4.1.1 Asymptotic statistics

The study of the asymptotic regime of an estimation procedure is interesting for two reasons. Firstly, as was mentioned in the introductory words of this chapter, asymptotic results provide approximations to exact values: while exact finite-sample calculations are usually intractable, the analysis of their large-sample limits often remains possible. Secondly, if we have several possible estimation procedures available, asymptotic large-sample behaviour provides ways to compare their performance. For example, to choose between two consistent estimation procedures, one can consider rate of convergence and other properties of limit distributions that characterise the degree of concentration (like asymptotic variance or asymptotic risk). In this subsection, we provide some aspects of asymptotic point estimation that are important for this and following chapters. It should be noted that this discussion is not intended to be comprehensive, nor is it stretched to full generality. For a more comprehensive presentation, the reader is referred to some of the excellent books devoted entirely to asymptotic statistics, like Ibragimov and Has'minskii (1981) [131], Le Cam and Yang (1990) [183] and van der Vaart (1998) [248].

The (decidedly frequentist) notion of *consistency* of a sequence of estimators based on a growing sample X^n (taking values in spaces $\mathscr{X}_n$) and a well-specified model $\mathscr{P}$, means that the sequence converges to the true distribution of the data as the size of the sample goes to infinity.

Definition 4.1.1. A sequence $\hat{P}_n : \mathscr{X}_n \to \mathscr{P}$ of estimators in a metric model $(\mathscr{P}, d)$ is said to be *consistent in a point* $P_0 \in \mathscr{P}$, if:

$$d(\hat{P}_n(X^n), P_0) \xrightarrow{P_0} 0.$$

and simply *consistent* if this holds for all points in $\mathscr{P}$.

To generalize to models parametrized by θ in a topological parameter space Θ containing a true parameter θ_0, consistency of a sequence of estimators $\hat{\theta}_n(X^n) \in \Theta$ means that $\hat{\theta}_n$ converges to θ_0. The definition of consistency can be strengthened to *almost-sure consistency*, by requiring that $d(\hat{P}_n(X^n), P_0)$ converges to zero (or $\hat{\theta}_n(X^n)$ to θ_0) P_0-almost-surely.

Example 4.1.2. Let $\mathscr{P}$ be a model for distributions on $\mathbb{R}$, parametrized by a location parameter $\theta \in \mathbb{R}$ and a parameter Q in a (possibly non-parametric) family H of distributions in $M^1(\mathbb{R})$, such that $\int x dQ(x) = 0$: $P_{\theta,Q}(B) = Q(B - \theta)$. That is, θ parametrizes the expectation of $P_{\theta,Q}$, while Q describes how probability is distributed around that point. We observe *i.i.d.* samples $X^n = (X_1, X_2, \ldots, X_n)$ with single-observation distribution P_{θ_0,Q_0} for some $\theta_0 \in \mathbb{R}$ and some $Q_0 \in H$. According to the *law of large numbers*, sample averages are almost-surely consistent estimators for the location θ_0:

$$\hat{\theta}_n(X^n) = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P_{\theta_0,Q_0}\text{-a.s.}} \theta_0.$$

An estimator that is consistent in a metric model may be analysed further by appraisal of its rate of convergence and limit distribution.

Definition 4.1.3. Let $\hat{P}_n : \mathscr{X}_n \to \mathscr{P}$ be a sequence of estimators in a metric model $(\mathscr{P}, d)$. Given $P_0 \in \mathscr{P}$, any sequence r_n such that,

$$r_n^{-1}\, d(\hat{P}_n(X^n), P_0) = O_{P_0}(1), \tag{4.2}$$

is an upper bound to the *rate of convergence* of the estimator sequence $\hat{P}_n$ with respect to the metric d, at P_0.

(For most estimators in most models, the rate of convergence is the same for all P_0; see, however, examples like 4.1.8 to emphasize that that feature is not generic). The rate of convergence thus describes the scaling necessary to have metric differences between $\hat{P}_n$ and P_0 that are distributed in a non-degenerate way, yet remain bounded in probability. Similarly, in a metric parametrizing space (Θ, d) the rate of convergence is such that $r_n^{-1} d(\hat{\theta}_n(X^n), \theta_0) = O_{P_{\theta_0}}(1)$.

Example 4.1.4. Consider the model and estimator of example 4.1.2. If we assume that H consists of (a subset of) all $Q \in M^1(\mathbb{R})$ such that $\int x^2\, dQ(x) < \infty$, then the rescaled differences $n^{1/2}\big(\hat{\theta}_n(X^n) - \theta_0\big)$ converge weakly due to the *central limit theorem*. Accordingly, the sequence $n^{1/2}\|\hat{\theta}_n - \theta_0\|$ is *uniformly tight* and (4.2) holds with rate $r_n = n^{-1/2}$.

Heightening the level of detail one step further, we require that the sequence of estimators, when centred on its point of convergence and rescaled by the rate, converges weakly to a non-degenerate distribution over the (localised) model.

Definition 4.1.5. Let $\hat{P}_n$ be a sequence of estimators in a metric model $(\mathscr{P}, d)$. Given $P_0 \in \mathscr{P}$ and rate sequence r_n, we say that $\hat{P}_n$ has *limit distribution* L_{P_0} at P_0, if,

$$r_n^{-1}\,(\hat{P}_n - P_0) \xrightarrow{P_0\text{-w.}} L_{P_0}, \tag{4.3}$$

where L_{P_0} is a non-degenerate Borel probability measure on $\mathscr{P}$.

In the parametric case, we say that $\hat{\theta}_n$ converges to θ_0 at rate r_n with non-degenerate limit distribution L_{θ_0} on $\Theta \subset \mathbb{R}^k$ if $r_n^{-1}(\hat{\theta}_n - \theta_0)$ converges weakly to L_{θ_0} on Θ under P_{θ_0}.

Example 4.1.6. Consider again the model and estimator of examples 4.1.2 and 4.1.4. Again assuming that H consists of (a subset of) all $Q \in M^1(\mathbb{R})$ such that $\int x^2\, dQ(x) < \infty$, the *central limit theorem* implies that,

$$\sqrt{n}\big(\hat{\theta}_n(X^n) - \theta_0\big) \xrightarrow{P_{\theta_0,Q_0}\text{-w.}} N(0, \sigma^2(Q_0)),$$

where the variance $\sigma^2(Q_0) = P_{\theta_0,Q_0}(X - \theta_0)^2$ is equal to the variance of Q_0. Accordingly, the estimators $\hat{\theta}_n$ are consistent at rate $n^{-1/2}$ and have a normal limit distribution with expectation 0 and (Q_0-dependent) variance σ^2. Note that a smooth function of the expectation, a parameter ψ that can be written as a (known) differentiable function $g(\theta)$ of the expectation θ, is estimable by $\hat{\psi}_n = g(\hat{\theta}_n)$, and according to the *delta rule*,

$$\sqrt{n}\big(\hat{\psi}_n(X^n) - \psi_0\big) \xrightarrow{P_{\theta_0,Q_0}\text{-w.}} N(0, g'(\theta_0)^2\sigma^2(Q_0)).$$

See van der Vaart (1998) [248] for much more on asymptotic statistics.

4.1.2 Asymptotic optimality in smooth parametric estimation

The concept of efficiency has its origin in Fisher's 1920's claim of asymptotic optimality of the maximum-likelihood estimator in differentiable parametric models. Here, optimality of the ML estimate means that they are *asymptotically consistent* achieve optimal $n^{-1/2}$-rate of convergence and have a limit distribution of minimal variance. In 1930's and –40's, Fisher's ideas on optimality in differentiable models were sharpened and elaborated upon. To illustrate, consider the following classical result from M-estimation (which can be found as theorem 5.23 in [248]).

Theorem 4.1.7. *Let Θ be open in $\mathbb{R}^k$ and assume that $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ is a Lebesgue-dominated model for i.i.d. data $X_1, X_2, \ldots$, with densities $p_\theta : \mathscr{X} \to \mathbb{R}$ such that $\theta \mapsto \log p_\theta(x)$ is differentiable at θ_0 for all $x \in \mathscr{X}$, with derivative (or score function) $\dot{\ell}_\theta(x)$. Assume that there exists a function $\dot{\ell} : \mathscr{X} \to \mathbb{R}$ such that $P_0\dot{\ell}^2 < \infty$ and,*

$$\big|\log p_{\theta_1}(x) - \log p_{\theta_2}(x)\big| \leq \dot{\ell}(x)\,\|\theta_1 - \theta_2\|,$$

for all θ_1, θ_2 in an open neighbourhood of θ_0. Furthermore, assume that $\theta \mapsto P_{\theta_0}\log p_\theta$ has a second-order Taylor expansion around θ_0 of the form,

$$P_{\theta_0}\log p_\theta = P_{\theta_0}\log p_{\theta_0} + \tfrac{1}{2}(\theta - \theta_0)^T I_{\theta_0}(\theta - \theta_0) + o(\|\theta - \theta_0\|^2),$$

with non-singular I_{θ_0}. If $(\hat{\theta}_n)$ is a consistent estimator sequence satisfying,

$$\mathbb{P}_n \log p_{\hat{\theta}_n} \geq \sup_{\theta\in\Theta} \mathbb{P}_n \log p_\theta - o_{P_{\theta_0}}(n^{-1}),$$

then $(\hat{\theta}_n)$ is asymptotically linear,

$$n^{1/2}(\hat{\theta}_n - \theta_0) = n^{-1/2}\sum_{i=1}^{n} I_{\theta_0}^{-1}\dot{\ell}_{\theta_0}(X_i) + o_{P_{\theta_0}}(1)$$

In particular, $n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{\theta_0\text{-w.}} N(0, I_{\theta_0}^{-1})$.

The last assertion of theorem 4.1.7 says that the (near-)maximum-likelihood estimators $(\hat{\theta}_n)$ are asymptotically consistent, converge at rate $n^{-1/2}$ and have the inverse Fisher information $I_{\theta_0}^{-1}$ as the covariance matrix for their (normal) limit distribution. At this stage of the discussion, we do not have an argument to show that this asymptotic behaviour is in any sense optimal. Nevertheless, let us take the opportunity to illustrate briefly how asymptotic behaviour translates into inference on θ by considering associated asymptotic confidence sets.

Recall definition 2.3.4 and example 2.3.6: an *asymptotic confidence set* is an approximate confidence set that is derived not from an exact sampling distribution, but from approximations implied by limit distributions, *e.g.* from the normal distribution $N(0, I_{\theta_0}^{-1})$ in the above theorem. To demonstrate, first suppose that the model is one-dimensional and satisfies the conditions of theorem 4.1.7. Denoting quantiles of the standard normal distribution by ξ_α, we see from the last assertion of the theorem that:

$$P_{\theta_0}^n\left(-\xi_\alpha I_{\theta_0}^{-1/2} < n^{1/2}(\hat{\theta}_n - \theta_0) \leq \xi_\alpha I_{\theta_0}^{-1/2}\right) \to 1 - 2\alpha,$$

If the Fisher information were known, this would give rise immediately to a confidence interval: the above display implies that,

$$\left[\hat{\theta}_n - n^{-1/2}\,\xi_\alpha\, I_{\theta_0}^{-1/2}\,,\, \hat{\theta}_n + n^{-1/2}\,\xi_\alpha\, I_{\theta_0}^{-1/2}\right]$$

has asymptotic coverage probability $1-2\alpha$. Since the Fisher information is not known exactly, we substitute an estimator for it, for example the sample variance S_n^2, to arrive at a *studentized* version of the above, which has the same asymptotic coverage and can therefore be used as an asymptotic confidence interval. But we could also have chosen to "plug in" the estimator $\hat{\theta}_n$ for θ_0 in the expression for the Fisher information to arrive at an estimate $I_{\hat{\theta}_n}$. To generalize to higher-dimensional $\Theta \subset \mathbb{R}^k$, recall that if Z has a k-dimensional multivariate normal distribution $N_k(0, \Sigma)$, then $Z^T \Sigma^{-1} Z$ possess a χ^2-distribution with k degrees of freedom. Denoting quantiles of the χ_k^2-distribution by $\chi_{k,\alpha}^2$, we find that so-called *Wald-type confidence sets*, ellipsoids of the form,

$$C_\alpha(X_1, \ldots, X_n) = \left\{\theta \in \Theta : n(\theta - \hat{\theta}_n)^T I_{\hat{\theta}_n}(\theta - \hat{\theta}_n) \leq \chi_{k,\alpha}^2\right\}, \tag{4.4}$$

have minimal Lebesgue measures among sets with coverage probabilities converging to $1-\alpha$.

4.1.3 Regular and irregular estimator sequences

Theorem 4.1.7 requires a rather large number of smoothness properties of the model: log-densities are required to be differentiable and Lipschitz and the *Kullback-Leibler divergence* must display a second-order expansion with non-singular second derivative matrix. These conditions are not only there to reflect model smoothness, they also guarantee that the ML estimator displays a property known as *regularity*. (The conditions listed are usually referred to as "regularity conditions".) The prominence of regularity in the context of optimality questions was not fully appreciated until in 1951, J. Hodges discovered an estimator that displayed *superefficiency* with regard to the asymptotic rate of convergence.

Example 4.1.8. (Hodges's shrinkage estimator)
Suppose that we estimate a parameter $\theta \in \Theta = \mathbb{R}$ with an estimator sequence $(\hat{\theta}_n)$,

satisfying limiting behaviour described by,

$$n^{1/2}(\hat{\theta}_n - \theta) \xrightarrow{P_\theta\text{-w.}} L_\theta,$$

for some laws L_θ, for all $\theta \in \Theta$. In addition, we define a so-called *shrinkage estimator*,

$$S_n(X^n) = \begin{cases} \hat{\theta}_n(X^n), & \text{if } |\hat{\theta}_n(X^n)| \geq n^{-1/4} \\ 0, & \text{if } |\hat{\theta}_n(X^n)| < n^{-1/4}. \end{cases}$$

The estimator S_n has a *bias* towards 0: any realization of $\hat{\theta}_n$ that is close enough to 0 is "shrunk" to 0 fully. One shows quite easily that S_n has the same asymptotic behaviour as $\hat{\theta}_n$ as long as $\theta \neq 0$, *i.e.* $n^{1/2}(S_n - \theta) \xrightarrow{P_\theta\text{-w.}} L_\theta$ if $\theta \neq 0$. But if $\theta = 0$, $\varepsilon_n^{-1}(S_n - 0) \xrightarrow{P_{\theta=0}\text{-w.}} 0$ for *any* sequence $\varepsilon_n > 0$, $\varepsilon_n \downarrow 0$. In other words, the asymptotic quality of S_n is as good as that of $\hat{\theta}_n$, and *strictly better* if $\theta = 0$. In a next step we could improve on S_n, by constructing a version of S_n that displays shrinkage in another point. Generalisation of this construction to other estimators and other models essentially says that *any* estimator sequence can be improved upon in a strict sense, at least in one point, through some form of shrinkage. Essentially this argument makes all estimators *inadmissible*.

Remark 4.1.9. In one-dimensional models [171], asymptotic superefficiency comes at a price, paid in terms of the behaviour of risk functions in neighbourhoods of the point of shrinkage and superefficiency can only be achieved on a subset of Lebesgue measure zero. In models of dimension three or higher, this restriction does not apply, as demonstrated by the non-asymptotic risk improvement of the James-Stein estimator over the ML estimator (see subsection 3.4.2).

So at certain points in the parameter space, Hodges's shrinkage estimators estimate with a rate of convergence that is strictly faster than that of the MLE and other estimators like it, while estimating the parameter with identical asymptotics for all other points in the model. In 1951, Hodges's superefficiency indicated that Fisher's 1920's claim was false without further refinement and that a comprehensive understanding of optimality in differentiable estimation problems remained elusive.

Hodges's example shows that any estimator sequence can be improved upon in at least one point of the model, which invalidates the question for an optimal estimator. To leave room for a notion of optimality, Hodges's shrinkage estimator has to be excluded from the class of eligible estimators. To prepare the relevant definition heuristically, note that, given Hodges's counterexample, it is not enough to specify the way that an estimator sequence converges *pointwise*; we must restrict the behaviour of estimators over ($n^{-1/2}$-)neighbourhoods rather than allow the type of wild variations that make Hodges's example possible.

Definition 4.1.10. Let $\Theta \subset \mathbb{R}^k$ be open. An estimator sequence (T_n) for the parameter θ is said to be *regular* at θ if, for all $h \in \mathbb{R}^k$,

$$n^{1/2}\Big(T_n - \big(\theta + n^{-1/2}h\big)\Big) \xrightarrow{P_n\text{-w.}} L_\theta,$$

where $P_n = P_{\theta+n^{-1/2}h}$. T_n is said to be *regular* if it is regular in all $\theta \in \Theta$, and *irregular* if there is a $\theta \in \Theta$ where T_n is not regular.

The point of definition 4.1.10 is the requirement that the limit law is independent of h, indicating that limiting behaviour is insensitive to perturbation of the parameter of size $n^{-1/2}h$. Typical examples of regular estimators are sample-means that estimate expectations (provided a second moment exists), while typical non-regular estimators are shrinkage estimators (like those of example 4.1.8) and estimators like $\hat{\theta}_n = \max\{X_i : 1 \leq i \leq n\}$ for the parameter θ that represents the upper bound of the support for the distribution of a bounded, real-valued random variable X.

Example 4.1.11. (Hodges's shrinkage estimator, cont.)
To demonstrate that Hodges's shrinkage estimators are *irregular*, consider the case that $\theta = 0$: if we assume that $\hat{\theta}_n$ is regular at $\theta = 0$, then,

$$n^{1/2}\left(\hat{\theta}_n - \frac{h}{\sqrt{n}}\right) \xrightarrow{P_n\text{-w.}} L_0,$$

for some limit distribution L_0. Since $n^{1/2}|\hat{\theta}_n(X^n)|$ stays below M_n with high probability, for any $M_n \to \infty$, $|\hat{\theta}_n| \leq n^{-1/4}$ with high probability. That means that $S_n(X^n) = 0$ with high probability, so that,

$$n^{1/2}\left(S_n - \frac{h}{\sqrt{n}}\right) = h,$$

which is not h-independent and can not be of the form L_0': $S_n(X^n)$ is not regular at $\theta = 0$.

4.1.4 Local asymptotic normality and the convolution theorem

The second ingredient we need, is a proper definition for what "model smoothness" means. The property in question was formulated in [174]: rather than require differentiability of likelihood functions *etcetera*, their local behaviour is described directly in terms of random variables playing the role of *score functions*. The "local" aspect of the definition stems from the n-dependent re-coordinatization in terms of the *local parameter* $h = n^{1/2}(\theta - \theta_0)$. (In the following we assume that the sample is *i.i.d.*, although usually the definition is extended to more general, dependent models for the data and applies to models for autoregressive time-series, random walks on finite state spaces, *etcetera* [179]).

Definition 4.1.12. (Local asymptotic normality (LAN), [174])
Let $\Theta \subset \mathbb{R}^k$ be open, parametrizing a model $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ for *i.i.d.* data $X_1, X_2, \ldots$ that is dominated by a σ-finite measure with densities p_θ. The model is said to be *locally asymptotically normal (LAN)* at θ_0 if, for any converging sequence $h_n \to h$ in $\mathbb{R}^k$:

$$\log \prod_{i=1}^{n} \frac{p_{\theta_0+n^{-1/2}h_n}}{p_{\theta_0}}(X_i) = h^T \Gamma_{n,\theta_0} - \tfrac{1}{2} h^T I_{\theta_0} h + o_{P_{\theta_0}}(1), \tag{4.5}$$

for random vectors Γ_{n,θ_0} such that $\Gamma_{n,\theta_0} \xrightarrow{P_{\theta_0}\text{-w.}} N_k(0, I_{\theta_0})$.

Typical parameters for which the LAN-expansion (7.14) holds are the parameters θ (or $\eta(\theta)$) in exponential families of definition 3.5.4, and typical examples of parameters that are *not* LAN are domain boundaries, like those of a uniform distribution on an interval or those in exponential or Pareto models. The LAN property formulates a notion of smoothness in parameter dependence and it is useful to formulate sufficient conditions based on differentiability of the density $\theta \mapsto p_\theta(x)$ at θ_0 for every x.

Proposition 4.1.13. *Let $\Theta \subset \mathbb{R}^k$ be open, parametrizing a dominated model $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ for i.i.d. data $X_1, X_2, \ldots \in \mathscr{X}$ with densities $p_\theta : \mathscr{X} \to [0,\infty)$. Assume that the map $\theta \mapsto \sqrt{p_\theta(x)}$ is continuously differentiable for every x. If elements of the matrix $I_\theta = P_\theta \dot{\ell}_\theta \dot{\ell}_\theta^T$ are finite and depend on θ continuously, then the model is LAN with respect to θ, with,*

$$\Gamma_{n,\theta_0} = n^{-1/2} \sum_{i=1}^{n} \dot{\ell}_{\theta_0}(X_i).$$

Proof. See of lemma 7.6 and theorem 7.2 in [248].

But local asymptotic normality can be achieved under weaker conditions; well known is the following property, best characterized as Hadamard differentiability of square-roots of model densities relative to the $L_2(P_0)$ norm.

Definition 4.1.14. (Differentiability in quadratic mean (DQM))
Let $\Theta \subset \mathbb{R}^k$ be open. A dominated model $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ for *i.i.d.* data $X_1, X_2, \ldots$ with densities p_θ is said to be *differentiable in quadratic mean* at $\theta_0 \in \Theta$, if there exists a score function $\dot{\ell}_{\theta_0} \in L_2(P_{\theta_0})$ such that:

$$\int \left(p_{\theta_0+h}^{1/2} - p_{\theta_0}^{1/2} - \tfrac{1}{2} h^T \dot{\ell}_{\theta_0} p_{\theta_0}^{1/2} \right)^2 d\mu = o\left(\|h\|^2\right),$$

as $h \to 0$.

Proposition 4.1.15. *A dominated model $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ for i.i.d. data $X_1, X_2, \ldots$ is DQM at θ_0, if and only if, it is LAN at θ_0.*

Proof. For a proof of the forward implication, see theorem 2 in section 17.3 of [179], or proposition 1 in section 7.2 of [183]. For a proof of the converse, see proposition 2 in section 17.3 of [179], or proposition 3 in section 7.2 of [183].

In many situations, it is quite straightforward to demonstrate the LAN property directly, in *i.i.d.* context usually through application of the *central limit theorem* for Γ_{n,θ_0} and the *law of large numbers* for the term that is second order in h.

Local asymptotic normality of the model and regularity of the estimator sequence come together in the following theorem which describes the foundation for the convolution theorem that follows: the models $\mathscr{P}_n = \{P_\theta^n : \theta \in \Theta\}$ for the *i.i.d.* samples X^n have a "limiting model" (for a fully developed theory of this type of limits of experiments, see Le Cam (1964) [175]; see also [179, 183] and [242]) that describes a single-observation for a normal distribution with unknown location, and regular sequences of estimators (T_n) are matched with a random variable T in the limiting model, in an asymptotically unbiased way.

Theorem 4.1.16. *Let $\Theta \subset \mathbb{R}^k$ be open; let $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ be LAN at θ_0 with non-singular Fisher information I_{θ_0}. Let (T_n) be regular estimators in the "localized models" $\{P_{\theta_0 + n^{-1/2}h} : h \in \mathbb{R}^k\}$. Then there exists a (randomized) statistic T in the normal location model $\{N_k(h, I_{\theta_0}^{-1}) : h \in \mathbb{R}^k\}$ such that $T - h \sim L_{\theta_0}$ for all $h \in \mathbb{R}^k$.*

Proof. See theorems 7.10, 8.3 and 8.4 in [248], or the more elaborate corollary 7.4.23 in [242].

Theorem 4.1.16 provides every regular estimator sequence with a limit in the form of a statistic in a very simple statistical experiment involving only a single $N_k(h, I_{\theta_0}^{-1})$-distributed observation X with unknown location h: the (weak) limit distribution that describes the local asymptotics of the sequence (T_n) under $P_{\theta_0 + n^{-1/2}h}$ *equals* the distribution of T under h, for all $h \in \mathbb{R}^k$. Moreover, regularity of the sequence (T_n) implies that under $N_k(h, I_{\theta_0}^{-1})$, the distribution of T relative to h is independent of h, an invariance usually known as *equivariance-in-law*. The class of equivariant-in-law estimators for location in the model $\{N_k(h, I_{\theta_0}^{-1}) : h \in \mathbb{R}^k\}$ is fully known: for any equivariant-in-law estimator T for h, there exists a probability distribution M such that $T \sim N_k(h, I_{\theta_0}^{-1}) * M$. The most straightforward example is $T = X$, for which $M = \delta_0$. This argument gives rise to the following central result in the theory of efficient estimation.

Theorem 4.1.17. *(Convolution theorem (Hajék, 1970) [121])*
Let $\Theta \subset \mathbb{R}^k$ be open and let $\{P_\theta : \theta \in \Theta\}$ be LAN at θ_0 with non-singular Fisher information I_{θ_0}. Let (T_n) be a regular estimator sequence with limit distribution L_{θ_0}. Then there exists a probability distribution M_{θ_0} such that,

$$L_{\theta_0} = N_k(0, I_{\theta_0}^{-1}) * M_{\theta_0},$$

In particular, if L_{θ_0} has a covariance matrix Σ_{θ_0}, then $\Sigma_{\theta_0} \geq I_{\theta_0}^{-1}$.

(for $k \times k$-matrices, the inequality $\Sigma_{\theta_0} \geq I_{\theta_0}^{-1}$ means that for all $v \in \mathbb{R}^k$, $v^T(\Sigma_{\theta_0} - I_{\theta_0}^{-1})v \geq 0$.) The occurrence of the inverse Fisher information as an optimal lower-bound in asymptotic context is finally explained here: the estimator T is unbiased so it satisfies the Cramér-Rao lower bound for asymptotic variance in the limiting model $\{N_k(h, I_{\theta_0}^{-1}) : h \in \mathbb{R}^k\}$. Convolution of $N_k(0, I_{\theta_0}^{-1})$ with any distribution M raises its variance unless M is degenerate: the last assertion of the convolution theorem says that, within the class of regular estimates, asymptotic variance is lower-bounded by the inverse Fisher information. A regular estimator that is optimal in

this sense, is called *best-regular* (or sometimes, *efficient*); an example is the ML estimator of theorem 4.1.7. Anderson's lemma below broadens the notion of optimality, in the sense that best-regular estimators outperform other regular estimators with respect to many loss functions.

Definition 4.1.18. A *sub-convex loss-function* is a map $\ell : \mathbb{R}^k \to [0,\infty)$ such that the level sets $\{x \in \mathbb{R}^k : \ell(x) \leq c\}$ are closed, convex and symmetric around the origin.

Examples of subconvex loss-functions are many and include, for example, the common choices $\ell(x) = \|x\|^p$, $p \geq 1$.

Lemma 4.1.19. *(Anderson's lemma)*
For any $k \geq 1$, any sub-convex loss function ℓ, any probability distribution M on $\mathbb{R}^k$ and any k-dimensional covariance matrix Σ,

$$\int \ell \, dN_k(0,\Sigma) \leq \int \ell \, d(N_k(0,\Sigma) * M).$$

Proof. A proof of Anderson's lemma can be found, for instance, in [131].

Based on Anderson's lemma, we see that the extent of the convolution theorem is greater than mere optimality with respect to some specific loss function, efficiency concerns all sub-convex loss functions. To conclude we mention the following equivalence, which characterizes efficiency concisely in terms of a weakly converging sequence.

Proposition 4.1.20. *In a LAN model, estimators (T_n) for θ are best-regular, if and only if, the (T_n) are asymptotically linear, i.e. for all θ in the model,*

$$n^{1/2}(T_n - \theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_\theta^{-1} \dot{\ell}_\theta(X_i) + o_{P_\theta}(1). \tag{4.6}$$

The random sequence of $n^{-1/2}$-rescaled sums on the *r.h.s.* of (4.6) is denoted by Δ_{n,θ_0} in theorem 4.2.1. Coming back to theorem 4.1.7, we see that under stated conditions, a consistent *MLE* $(\hat{\theta}_n)$ *is best-regular*, finally giving substance to Fisher's claim. Referring to the discussion on confidence sets with which we opened this section, we now know that in a LAN model confidence sets of the form (4.4), based on best-regular estimators $(\hat{\theta}_n)$, enjoy a similar form of optimality: according to the convolution theorem, the asymptotic sampling distributions of best-regular estimator sequences are all the same and sharpest among asymptotic sampling distributions for regular estimators.

4.2 Le Cam's Bernstein-von Mises theorem

To address the question of efficiency in smooth parametric models from a Bayesian perspective, we turn to the Bernstein-von Mises theorem. The first results concerning limiting normality of a posterior distribution date back to Laplace (1820) [166].

Later, Bernstein (1917) [15] and von Mises (1931) [195] proved results to a similar extent. Le Cam used the term 'Bernstein-von Mises theorem' in 1953 [171] and proved its assertion in greater generality. Walker (1969) [251] and Dawid (1970) [63] gave extensions to these results and Bickel and Yahav (1969) [27] proved a limit theorem for posterior means. Below we follow Le Cam and Yang (1990) [183].

The (proof of the) Bernstein-von-Mises theorem depends crucially on local asymptotic normality of the model at θ_0. A quick sketch of the proof can be given as follows. Suppose that the prior has a Lebesgue density that is continuous and strictly positive at θ_0. Also assume that the posterior concentrates in neighbourhoods of θ_0 of sizes decreasing as $n^{-1/2}$. Then it makes sense to consider the posterior density for the *local parameter* $h = \sqrt{n}(\theta - \theta_0)$, with Lebesgue-density:

$$\pi_n(h \,|\, X_1, X_2, \ldots, X_n) = \prod_{i=1}^n p_{\theta_0 + h/\sqrt{n}}(X_i)\, \pi(\theta_0 + h/\sqrt{n}) \Big/ \int \prod_{i=1}^n p_{\theta_0 + h'/\sqrt{n}}(X_i)\, \pi(\theta_0 + h'/\sqrt{n})\, dh',$$

almost-surely. Continuity of the Lebesgue density π of the prior at θ_0 implies that $\pi(\theta_0 + h/\sqrt{n})$ converges to the constant $\pi(\theta_0)$, which is strictly positive by assumption. This makes it plausible that upon substitution of the likelihood expansion (4.5), the posterior density converges to:

$$\frac{\prod_{i=1}^n p_{\theta_0 + h/\sqrt{n}}(X_i)\, dh}{\int \prod_{i=1}^n p_{\theta_0 + h'/\sqrt{n}}(X_i)\, dh'} \approx \frac{e^{h^T \Delta_{n,\theta_0} - \frac{1}{2} h^T I_{\theta_0} h}\, dh}{\int e^{h'^T \Delta_{n,\theta_0} - \frac{1}{2} h'^T I_{\theta_0} h'}\, dh'} \to \frac{dN(h, I_{\theta_0}^{-1})(X)\, dh}{\int dN(h', I_{\theta_0}^{-1})(X)\, dh'} \tag{4.7}$$

(in a suitable sense with respect to P_0). Here X is an observation in the normal limit model $\{N(h, I_{\theta_0}^{-1}) : h \in \mathbb{R}^k\}$. The *r.h.s.* of the last display equals $dN(X, I_{\theta_0}^{-1})(h)$ and is the posterior based on a sample consisting only of X and the Lebesgue prior on H for the limit model.

4.2.1 Conditions and consequences of the Bernstein-von Mises theorem

The Bernstein-von Mises theorem has been formulated in many different forms; the most general form is as follows [171, 183].

Theorem 4.2.1. *(Bernstein-von Mises)*
Assume that $\Theta \subset \mathbb{R}^k$ is open and that the model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is identifiable and dominated. Suppose $X_1, X_2, \ldots$ forms an i.i.d. sample from P_{θ_0} for some $\theta_0 \in \Theta$. Assume that the model is locally asymptotically normal at θ_0 with non-singular Fisher information I_{θ_0}. Furthermore suppose that the prior Π_Θ has a Lebesgue density that is continuous and strictly positive at θ_0 and that for every $\varepsilon > 0$, there exists a test sequence (ϕ_n) such that,

$$P^n_{\theta_0}\phi_n \to 0, \qquad \sup_{\|\theta-\theta_0\|>\varepsilon} P^n_\theta(1-\phi_n) \to 0.$$

Then posteriors converge to a normal distribution in total variation,

$$\left\| \Pi(h \in \cdot | X_1, \ldots, X_n) - N(\Delta_{n,\theta_0}, I^{-1}_{\theta_0}) \right\| \xrightarrow{P_0} 0,$$

centred on $\Delta_{n,\theta_0} = \sqrt{n}(\hat{\theta}_n - \theta_0)$*, where* $\hat{\theta}_n$ *is any best-regular estimator sequence.*

Proof. For a proof, see theorem 4.2.4, as well as the misspecified theorems in chapter 5.

Since the total-variational distance $\|N(\mu,\Sigma) - N(\nu,\Sigma)\|$ is bounded by a multiple of $\|\mu - \nu\|$, we find that the assertion of the Bernstein-von-Mises theorem can also be formulated with $\sqrt{n}(\hat{\theta}_n - \theta_0)$ replacing Δ_{n,θ_0}. Using the invariance of total-variation under rescaling and shifts, this leads to (4.1). In particular, according to theorem 4.1.7 and equivalence (4.6), the maximum-likelihood estimator is best-regular under stated smoothness conditions on the (log-)likelihood. This serves to motivate the often-heard statement that "Posterior means coincide with maximum-likelihood estimators asymptotically". In figure 4.1, Bernstein-von Mises-type of convergence of the posterior is demonstrated with a graphical/numerical example. Also displayed in figure 4.1 are the *MAP-estimator* of definition 2.2.20 and the ML estimator. Here, the MLE is efficient so it forms a possible centring sequence for the limiting sequence of normal distributions in the assertion of the Bernstein-von Mises theorem. Furthermore it is noted that the posterior concentrates more and more sharply, reflecting the n^{-1}-proportionality of the variance of its limiting sequence of normals. It is perhaps a bit surprising in figure 4.1 to see limiting normality obtain already at such relatively low values of the sample size n. It cannot be excluded that this is merely a manifestation the normality of the underlying model, but onset of normality of the posterior appears to happen at unexpectedly low values of n also in other smooth, parametric setting. It suggests that asymptotic conclusions based on the Bernstein-von Mises limit accrue validity fairly rapidly, for n in the order of several hundred to several thousand *i.i.d.* replications of the observation, at least, in well-behaved simple cases.

The uniformity in the assertion of the Bernstein-Von Mises theorem over model subsets B implies that it holds also for model subsets that are random. In particular, given some $0 < \alpha < 1$, it is noted that the (Lebesgue-)smallest sets $C_\alpha(X_1, \ldots, X_n)$ such that,

$$N_{\hat{\theta}_n, (nI_{\theta_0})^{-1}}\big(C_\alpha(X_1, \ldots, X_n)\big) \geq 1 - \alpha,$$

are ellipsoids of the form (4.4). Since posterior coverage of C_α converges to the *l.h.s.* in the above display, in accordance with the Bernstein-Von Mises limit, we see that the C_α are asymptotic credible sets of posterior coverage $1 - \alpha$. Conversely, a sequence $(D_n(X_1, \ldots, X_n))$ of *credible sets* of coverage $1 - \alpha$, is a sequence of sets that have *asymptotic confidence level* (arbitrarily close to) $1 - \alpha$ and credible sets of minimal Lebesgue measure coincide with Wald-type confidence sets asymptotically.

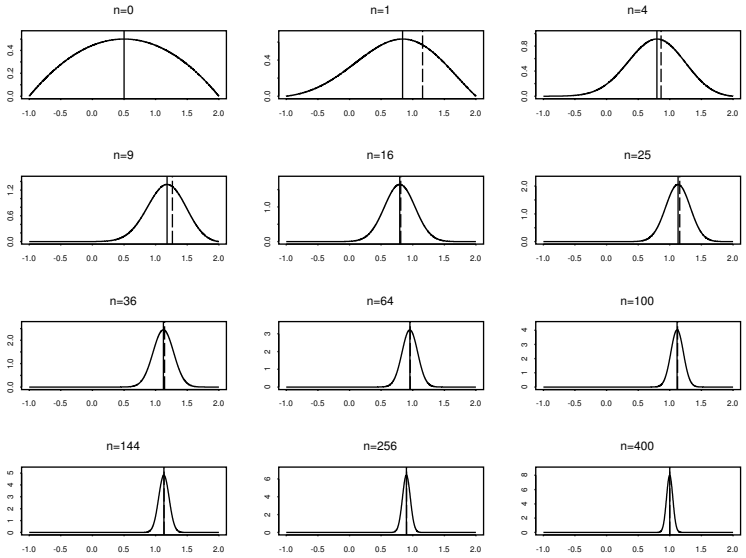


Fig. 4.1 Convergence of the posterior density. The samples used for calculation of the posterior distributions consist of n observations; the model consists of all normal distributions with mean between -1 and 2 and variance 1 and has a polynomial prior, shown in the first ($n = 0$) graph. For all sample sizes, the *maximum a posteriori* and maximum likelihood estimators are indicated by a vertical line and a dashed vertical line respectively. (From Kleijn (2003))

The above approximation in terms of uncertainty quantification gives rise to an identification in smooth, parametric models between inference based on frequentist best-regular point-estimators and inference based on Bayesian posteriors. In a practical sense, it eliminates the need to estimate θ and the Fisher information I_θ at θ to arrive at asymptotic confidence sets, if we have an approximation of the posterior distribution of high enough quality (*e.g.* from MCMC simulation), provided the Bernstein-von Mises theorem holds.

Remark 4.2.2. The asymptotic identification of credible and confidence sets is partially anticipated by theorem 2.3.14, solely on the basis of posterior concentration: in the proof of the Bernstein-von Mises theorem below, it becomes clear that the conditions of theorem 4.2.1 imply that, the posterior converges at rate $n^{-1/2}$, *i.e.* for any sequence $M_n \to \infty$,

$$\Pi\big(B(\theta_0, n^{-1/2}M_n) \;\big|\; X_1, \ldots, X_n\big) \xrightarrow{P_0} 1,$$

(see lemma 4.2.8), implying that condition (2.35) is satisfied. Following theorem 2.3.14, the radius-$n^{-1/2}M_n$ enlargements $C_n(X^n)$ of credible sets $D_n(X^n)$ (for any credible level $\gamma > 0$) are asymptotically consistent confidence sets. If we let

$\gamma_n \downarrow 0$ slowly enough, then the enlargements are approximated well by balls of radius $n^{-1/2}M_n$ centred on a point where the (approximately Gaussian) posterior density peaks, *e.g.* the MAP estimator. If, like in most situations and in figure 4.1, the MAP and ML estimators converge asymptotically and are best-regular, these balls decrease in radius at rates arbitrarily close to the (asymptotically optimal) rate $n^{-1/2}$. Such a sequence of balls asymptotically includes the Wald ellipsoids that are the optimal in smooth parametric setting for large enough n. Theorem 2.3.14 demonstrates what remains of that optimality if we do not use smoothness or parametric aspects and we maintain only posterior concentration as a condition. Note that if we had proved posterior concentration of the more specific form,

$$\Pi\big(n(\theta-\theta_0)^T I_{\theta_0}(\theta-\theta_0) \leq \chi_n^2 \;\big|\; X_1,\ldots,X_n\big) \xrightarrow{P_0} 1,$$

for certain constants $\chi_n^2 > 0$, then enlarged confidence regions $C(X^n)$ would have the ellipsoid form of Wald's optimal confidence regions (4.4).

To conclude let us briefly reflect on the conditions of theorem 4.2.1: local asymptotic normality and non-singularity of the associated Fisher information are minimal smoothness conditions. They also arise in theorem 4.1.7 and form the backdrop for any discussion of efficiency. More significant is the required existence of a "consistent" test sequence: what is required is that, asymptotically, we can distinguish P_0 from any complement of a θ-neighbourhood around θ_0 in a uniform way. One should compare this condition with the requirement of consistency of near-maximizers of the likelihood in theorem 4.1.7. Test conditions of the type given also play a central role in the developments of chapters 6 and 9.

4.2.2 Proof of the Bernstein-von Mises theorem

Below we divide the proof of the Bernstein-Von Mises theorem in two parts, with a requirement of local $n^{-1/2}$-sized consistency for the posterior in between. In a separate lemma, we show that a score-test fills in the gap between local and global consistency. To maintain the connection with chapter 5, we give the proof of the Bernstein-Von Mises theorem based on a smoothness property that is slightly stronger than local asymptotic normality.

Definition 4.2.3. We say that a parametric model $\mathscr{P}$ is *stochastically LAN* (sLAN) at θ_0, if the LAN property of definition 4.1.12 is satisfied for every *random* sequence (h_n) that is bounded in probability, *i.e.* for all $h_n = O_{P_0}(1)$:

$$\log \prod_{i=1}^n \frac{p_{\theta_0+n^{-1/2}h_n}}{p_{\theta_0}}(X_i) - h_n^T \Gamma_{n,\theta_0} + \tfrac{1}{2} h_n^T I_{\theta_0} h_n = o_{P_{\theta_0}}(1), \tag{4.8}$$

for random vectors Γ_{n,θ_0} such that $\Gamma_{n,\theta_0} \xrightarrow{\theta_0\text{-w.}} N_k(0, I_{\theta_0})$.

Theorem 4.2.4. *Let the sample $X_1, X_2, \ldots$ be distributed i.i.d.-P_0. Let $\Theta \subset \mathbb{R}^k$ be open, let $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ be stochastically LAN at θ_0 with non-singular Fisher information I_{θ_0} and let the prior Π on Θ be Lebesgue dominated with continuous, non-zero density. Furthermore, assume that for every sequence of origin-centred balls $(K_n) \subset \mathbb{R}^d$ with radii $M_n \to \infty$, we have:*

$$\Pi_n\big(h \in K_n \;\big|\; X_1, \ldots, X_n\big) \xrightarrow{P_0} 1. \tag{4.9}$$

Then posteriors converge to normal distributions in total variation:

$$\sup_B \Big| \Pi_n\big(h \in B \;\big|\; X_1, \ldots, X_n\big) - N_{\Delta_{n,\theta_0}, I_{\theta_0}^{-1}}(B) \Big| \xrightarrow{P_0} 0, \tag{4.10}$$

where $\Delta_{n,\theta_0} = \sqrt{n}(\hat{\theta}_n - \theta_0)$ for any best-regular estimator-sequence $\hat{\theta}_n$.

Proof. The proof is split into two parts: in the first part, we prove the assertion conditional on a compact neighbourhood K of 0 in Θ, and in the second part we diagonalize based on a sequence (K_n) with $\cup_n K_n = \mathbb{R}^k$ to prove (5.8). Throughout the proof we denote the posterior for h given $X_1, X_2, \ldots, X_n$ by Π_n and the normal distribution $N_{\Delta_{n,\theta_0}, I_{\theta_0}^{-1}}$ by Φ_n (for Δ_{n,θ_0}, see proposition 4.1.20). For $K \subset \mathbb{R}^k$, conditional versions are denoted Π_n^K and Φ_n^K respectively (assuming that $\Pi_n(K) > 0$ and $\Phi_n(K) > 0$, of course).

Let $K \subset \Theta$ be a ball centered on the origin in $\mathbb{R}^k$. For every open neighbourhood $U \subset \Theta$ of θ_0, $\theta_0 + n^{-1/2} K \subset U$ for large enough n. Since θ_0 is an internal point of Θ, we can define, for large enough n, the random functions $f_n : K \times K \to \mathbb{R}$ by:

$$f_n(g,h) = \Big(1 - \frac{\phi_n(h)}{\phi_n(g)} \frac{s_n(g)}{s_n(h)} \frac{\pi_n(g)}{\pi_n(h)}\Big)_+,$$

where $\phi_n : K \to \mathbb{R}$ is the Lebesgue density of the (randomly located) distribution $N_{\Delta_{n,\theta_0}, I_{\theta_0}^{-1}}$, $\pi_n : K \to \mathbb{R}$ is the Lebesgue density of the prior for the centred and rescaled parameter h and $s_n : K \to \mathbb{R}$ equals the likelihood product:

$$s_n(h) = \prod_{i=1}^n \frac{p_{\theta_0 + h/\sqrt{n}}}{p_{\theta_0}}(X_i).$$

Since the model is stochastically LAN by assumption, we have for every random sequence $(h_n) \subset K$:

$$\log s_n(h_n) = \sqrt{n} h_n (\mathbb{P}_n - P_0)\dot{\ell}_{\theta_0} - \tfrac{1}{2} h_n^T I_{\theta_0} h_n + o_{P_0}(1),$$

$$\log \phi_n(h_n) = -\tfrac{1}{2}(h_n - \Delta_{n,\theta_0})^T I_{\theta_0}(h_n - \Delta_{n,\theta_0}) + \text{constant}.$$

For any two sequences $(h_n), (g_n) \subset K$, $\pi_n(g_n)/\pi_n(h_n) \to 1$ as $n \to \infty$. Combining this with the above display we see that:

$$\begin{aligned}
&\log \frac{\phi_n(h_n)}{\phi_n(g_n)} \frac{s_n(g_n)}{s_n(h_n)} \frac{\pi_n(g_n)}{\pi_n(h_n)} \\
&\quad = -\sqrt{n} h_n (\mathbb{P}_n - P_0) \dot{\ell}_{\theta_0} + \tfrac{1}{2} h_n^T I_{\theta_0} h_n + \sqrt{n} g_n (\mathbb{P}_n - P_0) \dot{\ell}_{\theta_0} - \tfrac{1}{2} g_n^T I_{\theta_0} g_n + o_{P_0}(1) \\
&\qquad - \tfrac{1}{2} (h_n - \Delta_{n,\theta_0})^T I_{\theta_0} (h_n - \Delta_{n,\theta_0}) + \tfrac{1}{2} (g_n - \Delta_{n,\theta_0})^T I_{\theta_0} (g_n - \Delta_{n,\theta_0}) \\
&\quad = o_{P_0}(1)
\end{aligned}$$

as $n \to \infty$ by proposition 4.1.20. Since $x \mapsto (1 - e^x)_+$ is continuous on $\mathbb{R}$, we conclude that for every pair of random sequences $(g_n, h_n) \subset K \times K$:

$$f_n(g_n, h_n) \xrightarrow{P_0} 0, \quad (n \to \infty).$$

For fixed, large enough n, P_0^n-almost-sure continuity of $(g,h) \mapsto \log s_n(g)/s_n(h)$ on $K \times K$ is guaranteed by the stochastic LAN-condition. Each of the locations Δ_{n,θ_0} for Φ_n is tight, so $(g,h) \mapsto \phi_n(g)/\phi_n(h)$ is continuous on all of $K \times K$, P_0^n-almost-surely. Continuity (in a neighbourhood of θ_0) and positivity of the prior density guarantee that this holds for $(g,h) \mapsto \pi_n(g)/\pi_n(h)$ as well. We conclude that for large enough n, the random functions f_n are continuous on $K \times K$, P_0^n-almost-surely. Application of lemma 4.2.5 then leads to the conclusion that,

$$\sup_{g,h \in K} f_n(g,h) \xrightarrow{P_0} 0, \quad (n \to \infty). \tag{4.11}$$

Since K contains a neighbourhood of 0, $\Phi_n(K) > 0$ is guaranteed. Let Ξ_n denote the event that $\Pi_n(K) > 0$. Let $\eta > 0$ be given and based on that, define the events:

$$\Omega_n = \left\{ \omega : \sup_{g,h \in K} f_n(g,h) \leq \eta \right\}.$$

Consider the expression (recall that the total-variation norm is bounded by 2):

$$P_0^n \left\| \Pi_n^K - \Phi_n^K \right\| 1_{\Xi_n} \leq P_0^n \left\| \Pi_n^K - \Phi_n^K \right\| 1_{\Omega_n \cap \Xi_n} + 2 P_0^n (\Xi_n \setminus \Omega_n). \tag{4.12}$$

As a result of (4.11) the latter term is $o(1)$ as $n \to \infty$. The remaining term on the *r.h.s.* can be calculated as follows:

$$\begin{aligned}
\tfrac{1}{2} P_0^n \left\| \Pi_n^K - \Phi_n^K \right\| 1_{\Omega_n \cap \Xi_n} &= \tfrac{1}{2} P_0^n \int \left(1 - \frac{d\Phi_n^K}{d\Pi_n^K} \right)_+ d\Pi_n^K \, 1_{\Omega_n \cap \Xi_n} \\
&= \tfrac{1}{2} P_0^n \int_K \left(1 - \phi_n^K(h) \frac{\int_K s_n(g) \pi_n(g) dg}{s_n(h) \pi_n(h)} \right)_+ d\Pi_n^K(h) \, 1_{\Omega_n \cap \Xi_n} \\
&= \tfrac{1}{2} P_0^n \int_K \left(1 - \int_K \frac{s_n(g) \pi_n(g) \phi_n^K(h)}{s_n(h) \pi_n(h) \phi_n^K(g)} d\Phi_n^K(g) \right)_+ d\Pi_n^K(h) \, 1_{\Omega_n \cap \Xi_n}.
\end{aligned}$$

Note that for all $g,h \in K$ we have $\phi_n^K(h)/\phi_n^K(g) = \phi_n(h)/\phi_n(g)$ since, on K, ϕ_n^K differs from ϕ_n only by a normalisation factor. We use Jensen's inequality (with respect to the Φ_n^K-expectation) for the (convex) function $x \mapsto (1-x)_+$ to derive:

$$\tfrac{1}{2}P_0^n\big\|\Pi_n^K - \Phi_n^K\big\|1_{\Omega_n\cap\Xi_n} \le \tfrac{1}{2}P_0^n \int \Big(1 - \frac{s_n(g)\pi_n(g)\phi_n(h)}{s_n(h)\pi_n(h)\phi_n(g)}\Big)_+ d\Phi_n^K(g)\, d\Pi_n^K(h) 1_{\Omega_n\cap\Xi_n}$$

$$\le \tfrac{1}{2}P_0^n \int \sup_{g,h\in K} f_n(g,h) 1_{\Omega_n\cap\Xi_n} d\Phi_n^K(g)\, d\Pi_n^K(h) \le \tfrac{1}{2}\eta.$$

Combination with (4.12) shows that for all compact $K \subset \mathbb{R}^d$ containing a neighbourhood of 0,

$$P_0^n\big\|\Pi_n^K - \Phi_n^K\big\|1_{\Xi_n} \to 0.$$

Now let (K_m) be a sequence of origin-centred balls in $\mathbb{R}^k$ with radii $M_m \to \infty$. For each $m \ge 1$, the above display holds, so if we choose a sequence of balls (K_n) that traverses the sequence K_m slowly enough, convergence to zero can still be guaranteed. Moreover, the corresponding events $\Xi_n = \{\omega : \Pi_n(K_n) > 0\}$ satisfy $P_0^n(\Xi_n) \to 1$ as a result of (4.9). We conclude that there exists a sequence of radii (M_n) such that $M_n \to \infty$ and

$$P_0^n\big\|\Pi_n^{K_n} - \Phi_n^{K_n}\big\| \to 0, \tag{4.13}$$

(where it is understood that the conditional probabilities on the *l.h.s.* are well-defined on sets of probability growing to one). Combining (4.9) and lemma 4.2.7, we then use lemma 4.2.6 to conclude that:

$$P_0^n\big\|\Pi_n - \Phi_n\big\| \to 0,$$

which implies (4.10).

The proof of theorem 4.2.4 makes use of the following three lemmas. For their formulation, it is not necessary that the sample is *i.i.d.*, and we denote the true data-distributions by $P_{0,n}$.

Lemma 4.2.5. *Let (f_n) be a sequence of random functions $K \to \mathbb{R}$, where K is compact. Assume that for large enough $n \ge 1$, f_n is continuous $P_{0,n}$-almost-surely. Then the following are equivalent:*

(i) Uniform convergence in probability:

$$\sup_{h\in K}\big|f_n(h)\big| \xrightarrow{P_{0,n}} 0,$$

(ii) Convergence along any random sequence $(h_n) \subset K$ in probability:

$$f_n(h_n) \xrightarrow{P_{0,n}} 0,$$

as $n \to \infty$.

Proof. ((ii)⇒(i), by contradiction.) Assume that there exist $\delta, \varepsilon > 0$ such that:

$$\limsup_{n\to\infty} P_{0,n}\big(\sup_{h\in K}\big|f_n(h)\big| > \delta\big) = \varepsilon.$$

Since the functions f_n are continuous $P_{0,n}$-almost-surely, there exists (with $P_{0,n}$-probability one) a sequence $(\tilde{h}_n)$ such that for every $n \geq 1$, $\tilde{h}_n \in K$ and

$$\big|f_n(\tilde{h}_n)\big| = \sup_{h\in K}\big|f_n(h)\big|.$$

Consequently, for this particular random sequence in K, we have:

$$\limsup_{n\to\infty} P_{0,n}\Big(\big|f_n(\tilde{h}_n)\big| > \delta\Big) = \varepsilon > 0.$$

which contradicts (ii). *((i)⇒(ii).)* Conversely, given a random sequence $(h_n) \subset K$, and for every $\delta > 0$,

$$P_{0,n}\Big(\sup_{h\in K}\big|f_n(h)\big| > \delta\Big) \geq P_{0,n}\Big(\big|f_n(h_n)\big| > \delta\Big).$$

Given (i), the *l.h.s.* converges to zero and hence so does the *r.h.s.*.

The next lemma shows that given two sequences of probability measures, a sequence of balls that grows fast enough can be used conditionally to calculate the difference in total-variational distance, even when the sequences consist of random measures.

Lemma 4.2.6. *Let (Π_n) and (Φ_n) be two sequences of random probability measures on $\mathbb{R}^k$. Let (K_n) be a sequence of subsets of $\mathbb{R}^k$ such that*

$$\Pi_n(\mathbb{R}^k \setminus K_n) \xrightarrow{P_{0,n}} 0, \quad \Phi_n(\mathbb{R}^k \setminus K_n) \xrightarrow{P_{0,n}} 0. \tag{4.14}$$

Then

$$\big\|\Pi_n - \Phi_n\big\| - \big\|\Pi_n^{K_n} - \Phi_n^{K_n}\big\| \xrightarrow{P_{0,n}} 0. \tag{4.15}$$

Proof. Let K, a measurable subset of $\mathbb{R}^k$ and $n \geq 1$ be given and assume that $\Pi_n(K) > 0$ and $\Phi_n(K) > 0$. Then for any measurable $B \subset \mathbb{R}^k$ we have:

$$\begin{aligned}
\big|\Pi_n(B) - \Pi_n^K(B)\big| &= \Big|\Pi_n(B) - \frac{\Pi_n(B\cap K)}{\Pi_n(K)}\Big| \\
&= \big|\Pi_n\big(B\cap(\mathbb{R}^k\setminus K)\big) + \big(1-\Pi_n(K)^{-1}\big)\Pi_n(B\cap K)\big| \\
&\leq \Pi_n\big(B\cap(\mathbb{R}^k\setminus K)\big) + \Pi_n(\mathbb{R}^k\setminus K)\Pi_n^K(B) \leq 2\Pi_n(\mathbb{R}^k\setminus K).
\end{aligned}$$

and hence also:

$$\Big|\big(\Pi_n(B) - \Pi_n^K(B)\big) - \big(\Phi_n(B) - \Phi_n^K(B)\big)\Big| \leq 2\big(\Pi_n(\mathbb{R}^k\setminus K) + \Phi_n(\mathbb{R}^k\setminus K)\big). \tag{4.16}$$

As a result of the triangle inequality, we then find that the difference in total-variation distances between Π_n and Φ_n on the one hand and Π_n^K and Φ_n^K on the other is bounded above by the expression on the right in the above display (which is independent of B).

Define A_n, B_n to be the events that $\Pi_n(K_n) > 0$, $\Phi_n(K_n) > 0$ respectively. On $\Xi_n = A_n \cap B_n$, $\Pi_n^{K_n}$ and $\Phi_n^{K_n}$ are well-defined probability measures. Assumption (4.14) guarantees that $P_0^n(\Xi_n)$ converges to 1. Restricting attention to the event Ξ_n in the above upon substitution of the sequence (K_n) and using (4.14) for the limit of (4.16) we find (4.15), where it is understood that the conditional probabilities on the *l.h.s.* are well-defined with probability growing to 1.

To apply the above lemma in the concluding steps of the proof of theorem 4.2.4, rate conditions for both posterior and limiting normal sequences are needed. The rate condition (4.9) for the posterior is assumed and the following lemma demonstrates that its analog for the sequence of normals is satisfied when the sequence of centre points Δ_{n,θ_0} is uniformly tight.

Lemma 4.2.7. *Let K_n be a sequence of balls centred on the origin with radii $M_n \to \infty$. Let (Φ_n) be a sequence of normal distributions (with fixed covariance matrix V) located at the random points $(\Delta_n) \subset \mathbb{R}^k$. If the sequence Δ_n is uniformly tight, then:*

$$\Phi_n(\mathbb{R}^k \setminus K_n) = N_{\Delta_n,V}(H \in \mathbb{R}^k \setminus K_n) \xrightarrow{P_{0,n}} 0.$$

Proof. Let $\delta > 0$ be given. Uniform tightness of the sequence (Δ_n) implies the existence of a constant $L > 0$ such that:

$$\sup_{n \geq 1} P_{0,n}(\|\Delta_n\| \geq L) \leq \delta.$$

For all $n \geq 1$, call $A_n = \{\|\Delta_n\| \geq L\}$, $A_n^c = \{\|\Delta_n\| < L\}$. Let $\mu \in \mathbb{R}^k$ be given. Since $N(\mu, V)$ is tight, for every given $\varepsilon > 0$, there exists a constant L' such that $N_{\mu,V}(H \in B(\mu, L')) \geq 1 - \varepsilon$ (where $B(\mu, L')$ defines a ball of radius L' around the point μ. Assuming that $\mu \leq L$, $B(\mu, L') \subset B(0, L + L')$ so that with $M = L + L'$, $N_{\mu,V}(H \in B(0, M)) \geq 1 - \varepsilon$ for all μ such that $\|\mu\| \leq L$. Choose $N \geq 1$ such that $M_n \geq M$ for all $n \geq N$. Let $n \geq N$ be given. Then:

$$P_{0,n}\big(\Phi_n(\mathbb{R}^k \setminus B(0, M_n)) > \varepsilon\big) \leq \delta + P_{0,n}\Big(\big\{N_{\Delta_n,V}(H \notin B(0, M_n)) > \varepsilon\big\} \cap A_n^c\Big) \quad (4.17)$$

Note that on the complement of A_n, $\|\Delta_n\| < L$, so:

$$N_{\Delta_n,V}(H \notin B(0, M_n)) \leq 1 - N_{\Delta_n,V}(H \in B(0, M)) \leq 1 - \inf_{\|\mu\| \leq L} N_{\mu,V}(H \in B(0, M)) \leq \varepsilon,$$

and we conclude that the last term on the *r.h.s.* of (4.17) equals zero.

Aside from a slightly stronger smoothness property in the form of the stochastic LAN condition, theorem 4.2.4 appears to require more than theorem 4.2.1, in the

sense that it requires posterior consistency at rate $n^{-1/2}$ rather than the (fixed) tests for consistency. The following lemma asserts that, assuming smoothness, the latter condition is enough to satisfy the former. Its proof is based on the construction of a score test that fills in the "gap" left between the fixed-alternative tests and the growing alternative $\|\theta - \theta_0\| \geq n^{-1/2} M_n$.

Lemma 4.2.8. *Assume that $\Theta \subset \mathbb{R}^k$ is open and that the model $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ is identifiable and dominated. Assume that the model is locally asymptotically normal at θ_0 with non-singular Fisher information I_{θ_0} and that the prior is Lebesgue dominated with continuous with a density that is non-zero at θ_0. Furthermore, suppose that there exists a test sequence (ϕ_n) such that,*

$$P^n_{\theta_0} \phi_n \to 0, \qquad \sup_{\|\theta - \theta_0\| > \varepsilon} P^n_\theta (1 - \phi_n) \to 0.$$

Then the posterior converges at rate $n^{-1/2}$, i.e. for every sequence $M_n \to \infty$,

$$\Pi\big(\|\theta - \theta_0\| \geq n^{-1/2} M_n \;\big|\; X_1, \ldots, X_n \big) \xrightarrow{P_0} 0.$$

Proof. A proof is given in theorem 5.3.1, section 5.3), in the more general, misspecified situation.

4.3 Semi-parametric Bernstein-von Mises theorems [EMPTY]

4.4 Exercises

4.4.1. Let $(\mathscr{X}, \mathscr{B})$ be a measurable space with probability measures P, Q. Show that, for any $n \geq 1$, $H^2(P^n, Q^n) \leq n H^2(P, Q)$.

4.4.2. Assume that $n^{1/2}(\hat{\theta}_n - \theta_0) \sim N(0, I^{-1}_{\theta_0})$. Show that the ellipsoids (4.4) are of minimal Lebesgue measure among all subsets of asymptotic coverage $1 - \alpha$.

4.4.3. Consider Hodges's estimators S_n of example 4.1.8. Show that, for any rate sequence (ε_n), $\varepsilon_n \downarrow 0$, $\varepsilon_n^{-1}(S_n - 0) \xrightarrow{0\text{-w.}} 0$.

4.4.4. Let $\Theta = (0, \infty)$ and let $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ be the model of Poisson distributions P_θ with means θ. Let the data be an *i.i.d.* sample from P_{θ_0} for some $\theta_0 \in \Theta$. Show that this model is LAN for all θ_0.

4.4.5. Let $\Theta = \mathbb{R}$ and let $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ be the model of normal distributions $N(\theta, 1)$ of unit variance with means θ. Let the data be an *i.i.d.* sample from P_θ for some $\theta \in \Theta$. Show that this model is LAN for all θ.

4.4.6. Let f be a Lebesgue density on $\mathbb{R}$ that is symmetric around the origin. Define the model $\mathscr{P} = \{P_{\mu,\sigma} : \mu \in \mathbb{R}, \sigma \in (0, \infty)\}$ by densities $f_{\mu,\sigma}(x) = \sigma^{-1} f((x - \mu)/\sigma)$. Show that the Fisher information matrix is diagonal.

4.4.7. Let P and Q be probability measures on a measurable space $(\mathscr{X},\mathscr{B})$,

a. Show that there exists a σ-finite measure μ such that $P,Q \ll \mu$.
b. Using Radon-Nikodym derivatives $p = dP/d\mu$ and $q = dQ/d\mu$, prove that,

$$\sup_{B\in\mathscr{B}} \big|P(B) - Q(B)\big| = \frac{1}{2}\int \big|p-q\big|\,d\mu.$$

c. Show that, for any sequence (Q_n) of probability measures on $(\mathscr{X},\mathscr{B})$, there exists a probability measure P that dominates all Q_n, $(n \geq 1)$.
d. Use the completeness of $L_1(\mathscr{X},\mathscr{B},P)$ to show that the metric space $\mathscr{M}(\mathscr{X},\mathscr{B})$ of all probability measures on $(\mathscr{X},\mathscr{B})$ is complete in the topology of total variation.

4.4.8. Let $\Theta = (0,\infty)$ and $\mathscr{P} = \{N(0,\theta^2) : \theta \in \Theta\}$. Let Π be a Lebesgue dominated prior on Θ, with continuous, non-zero Lebesgue density. Show that this model satisfies the conditions of the Bernstein-von Mises theorem 4.2.1. Find the problematic range of parameter values in this model. *(Hint: calculate the Fisher information, find a problematic limit for it and describe the effect on the limiting sequence of normal distributions for parameter values close to the problematic limit.)*

4.4.9. Approximation in measure from within by compact subsets has a deep background in analysis. Central is the notion of a *Radon measure* (see definition C.8.1). Show that any probability measure on a Polish space is Radon. *Hint: recall that Polish spaces are Lindelöf: every open cover of an open subset has a countable subcover. (This statement generalizes to continuous images of Polish spaces, known as Souslin spaces.)*

4.4.10. Show that the Borel measure μ of the *Riesz representation theorem* is a *Radon measure* (see definition C.8.1).

4.4.11. Prove the following: for $\theta \in \Theta = \mathbb{R}$, let $F_\theta(x) = (1 - e^{-(x-\theta)}) \vee 0$ be the standard exponential distribution function located at θ. Assume that $X_1, X_2, \ldots$ form an *i.i.d.* sample from F_{θ_0}, for some θ_0. Let Π be a Lebegues dominated prior on Θ with continuous, non-zero density. Then the associated posterior distribution satisfies, with $h = n(\theta - \theta_0)$,

$$\sup_A \Big| \Pi_n\big(h \in A \;\big|\; X_1,\ldots,X_n\big) - \mathrm{Exp}^-_{n(\hat\theta_n-\theta_0)}(A) \Big| \xrightarrow{\theta_0} 0,$$

where $\hat\theta_n = X_{(1)}$ is the maximum likelihood estimate for θ_0 and Exp^-_a denotes the standard negative exponential distribution located at a. *(NB: This is an example of an* irregular *estimation problem: clearly the model does not depend on θ in a differentiable way. Inspection of the assertion shows that the rate of convergence is n^{-1} rather than $n^{-1/2}$, the rate of convergence in regular situations. In addition, the limiting shape of the posterior is not normal but exponential.)*

4.4.12. Show the following: let (X_n) be a sequence of real random variables. If the sequence (X_n) is almost-surely bounded (*i.e.* there exists a constant that bounds all X_n almost-surely), then convergence of X_n in probability and convergence of X_n in expectation are equivalent.

Some example exam problems

4.4.13. This problem concerns the frequentist theory of efficient estimation and the Bernstein-von Mises theorem. Denote the model by $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$, for some open parameter space $\Theta \subset \mathbb{R}$ (note: we specialize to *one-dimensional* parameter spaces here).

a. Assume that $\mathscr{P}$ is a model for *i.i.d.* data $X_1, \ldots, X_n \in \mathscr{X}$, that is dominated with densities $p_\theta : \mathscr{X} \to \mathbb{R}$ for all $\theta \in \Theta$. State the definition of *local asymptotic normality* of the model. What is the usual form of the term linear in the local parameter h on the right-hand side (in terms of the score $\dot{\ell}_\theta(x) = \partial/\partial\theta \log p_\theta(x)$)?
b. State Hajék's convolution theorem. Discuss the roles of the two main conditions, local asymptotic normality and regularity (that is, formulate what these two conditions require at a heuristic level). Explain that the assertion implies that there exists a lower bound for asymptotic variance and give this lower bound.
c. Give the Bernstein-von Mises theorem. Discuss its conditions regarding the model and the prior. Explain what justifies the often-heard phrase, "The posterior centres on the maximum-likelihood estimator asymptotically."
d. Explain why the Bernstein-von Mises theorem enables the interpretation of credible sets as asymptotic confidence sets.

Chapter 5
Model misspecification

Generally speaking, statistical analysis requires a choice of a *model*, which may not include the frequentist true distribution of the data. Throughout most of what has preceded, we have assumed that the model $\mathscr{P}$ is *well-specified*, *c.f.* definition 1.1.9. In asymptotic context, well-specification translates into the assumption that for every $n \geq 1$, the true distribution $P_{0,n}$ of the sample X^n lies in the n-th model $\mathscr{P}_n$. In the more specific situation that these models are parametrized with the help of a single parameter space Θ by maps $\Theta \to \mathscr{P}_n : \theta \mapsto P_{\theta,n}$, well-specification is expressed through the stronger assumption that there exists a $\theta_0 \in \Theta$ such that $P_{0,n} = P_{\theta_0,n}$ for all $n \geq 1$. Assumptions of this nature, which concern the unknown quantity of interest θ_0 directly, are accepted as an article of faith in most frequentist statistical procedures (and often difficult or impossible to verify, even asymptotically, through tests based on the data (see chapter 9, particularly, examples 9.4.12–9.4.14)).

In the proofs of theorems, it is rarely a problem if there is no single $\theta_0 \in \Theta$ to explain all $P_{0,n}$, because often one can prove exactly the same for n-dependent $\theta_{0,n}$ such that $P_{0,n} = P_{\theta_{0,n},n}$. But what happens to our statistical procedures in the far-worse case when,

$$P_{0,n} \notin \mathscr{P}_n, \tag{5.1}$$

the true distribution of the data does not even lie in the model? The smaller the models $\mathscr{P}_n$, the more stringent the assumption that the model is well-specified. Especially when we consider a parametric models, when $\Theta \subset \mathbb{R}^k$, chances are that the models we have for the true distribution of the data are misspecified, *c.f.* (5.1). Commonly ignored in practice, this fact implies that many statistical procedures are carried out with misspecified models. Theorems assuming well-specification are used regardless, seldom leading to significant problems, which raises the question: "Why? What can be said about the reliability of statistical tools in misspecified situations?"

When we dissociate the definition of the model from sufficient assumptions on P_0 for our tools to work, we explore the maximal extent of their applicability properly. The goal is to state model assumptions and assumptions on the true distribution of the data separately: in the case of an *i.i.d.* sample, for example, we would specify

a model $\mathscr{P}$ and assume given *i.i.d.* observations $X_1, \ldots, X_n$, marginally distributed according to some unknown single-observation distribution P_0. We would then formulate conditions for P_0 (rather than restrict the model and assume it to be well-specified) in order for the assertion of the theorem to hold. Ideally, these conditions are satisfied not only by the model distributions but also by a large, non-parametric set of other distributions, so that the misspecified theorem generalises its well specified version and describes in detail the consequences of misspecification.

5.1 Misspecification in smooth parametric models

In this chapter, we address the misspecification question in the particular, parametric case of the Bernstein-von Mises theorem of chapter 4. The main conclusion we shall draw, is that in the asymptotic limit, the posterior distribution of a parameter in misspecified LAN parametric models is still approximated well by a random normal distribution, but Bayesian *credible sets cease to be valid as approximate confidence sets if the model is misspecified*. We obtain the result under conditions that are comparable to those in the well-specified situation: uniform testability against fixed alternatives and sufficient prior mass in neighbourhoods of the point of convergence. The rate of convergence is considered in detail, with special attention for the existence and construction of suitable test sequences.

We do not discuss the asymptotic behaviour of posterior distributions in misspecified *non-parametric* models; a discussion (of less specificity than that of the Bernstein-von Mises theorem) is found in Kleijn (2004) [150] and Kleijn and van der Vaart (2006) [151].

5.1.1 Misspecified maximum likelihood estimation

A class of point estimators that generalises relatively easily to the misspecified situation is that of M-estimators (see van de Geer (2000) [103] and van der Vaart (1998, 1996) [248, 247]). Consider a smooth parametric model $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ of single-observation distributions for *i.i.d.* samples $X^n(X_1, \ldots, X_n)$, $n \geq 1$. We denote the true single-observation distribution by P_0 and do *not* assume that $P_0 \in \mathscr{P}$. An M-estimator is a (near-)maximiser $\hat{\theta}_n$ of the function $M_n : \Theta \to \mathbb{R}$ with,

$$M_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} m_\theta(X_i),$$

for some P_0-integrable $m_\theta(x)$. Assuming that $\mathscr{P}$ is a dominated model with probability densities p_θ, the *maximum-likelihood estimator* is the M-estimator for the choice $m_\theta(x) = \log p_\theta(x)$. Under certain, rather stringent conditions (see, for example, [248], section 5.2), $\hat{\theta}_n$ converges to the maximum θ^* of the function $\theta \mapsto$

$P_0 m_\theta(X)$. In the maximum-likelihood case, the estimators $\hat{\theta}_n$ converge (in P_0-probability or P_0-almost-surely) to the point $\theta^* \in \Theta$ that minimises the so-called *Kullback-Leibler divergence* of P_θ with respect to P_0:

$$\theta \mapsto -P_0 \log \frac{p_\theta}{p_0}, \tag{5.2}$$

over the model Θ. (The P_0-almost-sure existence and uniqueness of θ^* are non-trivial conditions for the model $\mathscr{P}$ and true P_0.) The fact that θ^* does not correspond to the true distribution P_0 directly is inconsequential: the maximum-likelihood procedure defines the 'best' approximation of P_0 within $\mathscr{P}$ to be the point of minimal Kullback-Leibler divergence (note that other choices for $x \mapsto m_\theta(x)$ would lead to different ways of 'projecting' P_0 onto $\mathscr{P}$). The asymptotic behaviour of the maximum-likelihood procedure is postponed to lemmas 5.2.3 and 5.2.4, but we note here that under regularity conditions (see [248], sections 5.3 and 5.5), maximum likelihood estimators $\hat{\theta}_n$ converge to θ^* in an asymptotically normal way,

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{P_0\text{-w.}} N_{0, V_{\theta^*}^{-1} I_{\theta^*} V_{\theta^*}^{-1}}, \tag{5.3}$$

where V_{θ^*} is the second-order coefficient in the Taylor expansion of the Kullback-Leibler divergence (assumed non-singular) and $I_{\theta^*} = P_0 \dot{\ell}_{\theta^*} \dot{\ell}_{\theta^*}^T$ denotes the Fisher information at θ^* (see, for instance, Huber (1967) [130]). In the well-specified case, V_{θ^*} equals I_{θ^*} and the asymptotic variance reduces to a single instance of the inverse Fisher information, but that cancellation does not occur in the misspecified case.

5.1.2 The misspecified Bernstein-von Mises theorem

Consistency of posterior distributions and asymptotic normality of the posterior mean under misspecification have been considered in Berk (1966, 1970) [21, 22] and Bunke and Milhaud (1998) [53]. The behaviour of the full posterior distribution was studied in Kleijn and van der Vaart (2004) [152]. Here we follow the latter and derive the asymptotic normality of the full posterior distribution in the misspecified situation under conditions comparable to those obtained in the well-specified case of section 4.2. We focus on dominated models for *i.i.d.* data where the posterior distribution follows (2.13) and we assume that the observations are sampled from a density p_0 that is not necessarily of the form p_{θ_0} for some θ_0. It is shown that the Bernstein-von Mises assertion (4.1) can be extended to this situation, in the form,

$$\sup_B \left| \Pi\big(\vartheta \in B \,\big|\, X_1, \ldots, X_n \big) - N(\hat{\theta}_n, (nV_{\theta^*})^{-1})(B) \right| \xrightarrow{P_0} 0, \tag{5.4}$$

where θ^* is the parameter value minimizing the Kullback-Leibler divergence $\theta \mapsto P_0 \log(p_0/p_\theta)$, V_{θ^*} is minus the second derivative matrix of this map, and $\hat{\theta}_n$ are suitable estimators.

Remark 5.1.1. According to (5.3), maximum likelihood estimators in the misspecified model are asymptotically normal with mean zero and covariance matrix given by $\Sigma_{\theta^*} = V_{\theta^*}(P_0\dot{\ell}_{\theta^*}\dot{\ell}_{\theta^*}^T)^{-1}V_{\theta^*}$. The corresponding *Wald-type confidence sets* (see (4.4)) for the misspecified parameter take the form $\hat{\theta}_n + \Sigma_{\theta^*}^{1/2}C/\sqrt{n}$ for C a central set in the Gaussian distribution. Because the covariance matrix V_{θ^*} appearing in the misspecified Bernstein-von Mises theorem is *not* the sandwich covariance matrix, credible sets of posterior probability $1-\alpha$ *do not correspond* to the misspecified Wald sets. Although they are correctly centered, they may have the wrong width, and are in general not $1-\alpha$-confidence sets. To make the consequences of the mismatch between the asymptotic covariance matrix $V_{\theta^*}^{-1}P_0(\dot{\ell}_{\theta^*}\dot{\ell}_{\theta^*}^T)V_{\theta^*}^{-1}$ and limiting covariance matrix $V_{\theta^*}^{-1}$ explicit, consider the following example.

Example 5.1.2. Let P_θ be the normal distribution with mean θ and variance 1, and let the true distribution possess mean zero and variance $\sigma^2 > 0$. Then $\theta^* = 0$, $P_0\dot{\ell}_{\theta^*}^2 = \sigma^2$ and $V_{\theta^*} = 1$. It follows that the radius of the $1-\alpha$-Bayesian credible set is $z_\alpha/\sqrt{n}$, whereas a $1-\alpha$-confidence set around the mean has radius $z_\alpha\sigma/\sqrt{n}$. Depending on $\sigma^2 \leq 1$ or $\sigma^2 > 1$, the credible set can have coverage arbitrarily close to 0 or 1.

So credible sets may over- or under-cover as confidence sets, depending on the true distribution of the observations and the model and to extreme amounts.

This chapter's presentation is split into two parts: in section 5.2 we derive normality of the posterior given that it shrinks at a $\sqrt{n}$-rate of posterior convergence (theorem 5.2.2). We actually state this result for the general situation of locally asymptotically normal (LAN) models, and next specify to the *i.i.d.* case. Next in section 5.3 we discuss results guaranteeing the desired rate of convergence, where we first show sufficiency of existence of certain tests (theorem 5.3.1), and next construct appropriate tests (theorem 5.3.5).

5.2 Posterior limit distribution

Throughout the presentation of the misspecified Bernstein-von Mises theorem and its consequences, we denote the model parametrizations by $\theta \mapsto P_\theta^{(n)}$ and the corresponding random variables by $X^{(n)}$ (when possibly non-*i.i.d.*) and by $\theta \mapsto P_\theta^n$ and X^n (when *i.i.d.*), deviating from the notations $\theta \mapsto P_{\theta,n}$ and X^n used elsewhere in this book, for typographic reasons.

5.2.1 Posterior asymptotic normality in smooth models

Let Θ be an open subset of $\mathbb{R}^k$ parametrizing statistical models $\{P_\theta^{(n)} : \theta \in \Theta\}$. For simplicity, we assume that for each n there exists a single measure that dominates

all measures $P_\theta^{(n)}$ as well as a "true measure" $P_0^{(n)}$, and we assume that there exist densities $p_\theta^{(n)}$ and $p_0^{(n)}$ such that the maps $(\theta, x) \mapsto p_\theta^{(n)}$ are measurable. Generalizing definition 4.1.12, we consider models satisfying a smoothness condition of the following type.

Definition 5.2.1. We say that a misspecified parametric model $\mathscr{P}$ is *stochastically LAN* (sLAN) at an inner point $\theta^* \in \Theta$ and relative to a given norming rate $\delta_n \to 0$, if there exist random vectors Δ_{n,θ^*} and non-singular matrices V_{θ^*} such that the sequence Δ_{n,θ^*} is bounded in P_0-probability and for every compact set $K \subset \mathbb{R}^k$,

$$\sup_{h \in K} \left| \log \frac{p_{\theta^* + \delta_n h}^{(n)}}{p_{\theta^*}^{(n)}}(X^{(n)}) - h^T V_{\theta^*} \Delta_{n,\theta^*} + \tfrac{1}{2} h^T V_{\theta^*} h \right| \to 0, \tag{5.5}$$

in $P_0^{(n)}$-probability.

The prior measure Π on Θ is assumed to be a probability measure with Lebesgue-density π, continuous and positive on a neighbourhood of a given point θ^*. Priors satisfying these criteria assign enough mass to (sufficiently small) balls around θ^* to allow for optimal rates of convergence of the posterior if certain regularity conditions are met (see section 5.3). Like before, the posterior based on an observation $X^{(n)}$ is denoted $\Pi(\cdot\,|X^{(n)})$: for every Borel set A,

$$\Pi\big(\vartheta \in A \,\big|\, X^{(n)}\big) = \int_A p_\theta^{(n)}(X^{(n)}) \pi(\theta)\, d\theta \,\Big/\, \int_\Theta p_\theta^{(n)}(X^{(n)}) \pi(\theta)\, d\theta. \tag{5.6}$$

We stress that both definition (5.5) and the assertion of theorem 5.2.2 below involve convergence *in* $P_0^{(n)}$*-probability*, that is, with respect to the true distribution of the data.

Theorem 5.2.2. *Assume that definition 5.2.1 holds at some $\theta^* \in \Theta$ and let the prior Π be Lebesgue absolutely continuous with a continuous density that is strictly positive in θ^*. Furthermore, assume that for every sequence of constants $M_n \to \infty$,*

$$P_0^{(n)} \Pi\big(\|\vartheta - \theta^*\| > \delta_n M_n \,\big|\, X^{(n)}\big) \to 0. \tag{5.7}$$

Then the sequence of posteriors converges to a sequence of normal distributions in total variation:

$$\sup_B \left| \Pi\big((\vartheta - \theta^*)/\delta_n \in B \,\big|\, X^{(n)}\big) - N_{\Delta_{n,\theta^*}, V_{\theta^*}^{-1}}(B) \right| \xrightarrow{P_0} 0. \tag{5.8}$$

Proof. The proof is identical to that of theorem 4.2.1, with a few small changes: the local parameter h is now defined with the help of the rate δ_n. Throughout the proof we denote the posterior for $H = (\vartheta - \theta^*)/\delta_n$ given $X^{(n)}$ by $\Pi_n(\cdot|X^{(n)})$ which follows from that for θ by $\Pi_n(H \in B|X^{(n)}) = \Pi((\vartheta - \theta^*)/\delta_n \in B|X^{(n)})$ for all Borel sets B. Furthermore, we denote the normal distribution $N_{\Delta_{n,\theta^*}, V_{\theta^*}^{-1}}$ by Φ_n. For a compact subset $K \subset \mathbb{R}^k$ such that $\Pi_n(H \in K|X^{(n)}) > 0$, we define a conditional

version Π_n^K of Π_n by $\Pi_n^K(B|X^{(n)}) = \Pi_n(B\cap K|X^{(n)})/\Pi_n(K|X^{(n)})$. Similarly we defined a conditional measure Φ_n^K corresponding to Φ_n. Then, following the proof of theorem 4.2.1, it is noted that,

$$\begin{aligned}
&\log\frac{\phi_n(h_n)}{\phi_n(g_n)}\frac{s_n(g_n)}{s_n(h_n)}\frac{\pi_n(g_n)}{\pi_n(h_n)}\\
&\quad= (g_n-h_n)^T V_{\theta^*}\Delta_{n,\theta^*} + \tfrac{1}{2}h_n^T V_{\theta^*}h_n - \tfrac{1}{2}g_n^T V_{\theta^*}g_n + o_{P_0}(1)\\
&\quad\quad - \tfrac{1}{2}(h_n-\Delta_{n,\theta^*})^T V_{\theta^*}(h_n-\Delta_{n,\theta^*}) + \tfrac{1}{2}(g_n-\Delta_{n,\theta^*})^T V_{\theta^*}(g_n-\Delta_{n,\theta^*})\\
&\quad= o_{P_0}(1),
\end{aligned}$$

as $n\to\infty$. The rest of the proof is identical.

Condition (5.7) fixes the rate of convergence of the posterior distribution to be that occuring in the LAN property. Sufficient conditions to satisfy (5.7) in the case of *i.i.d.* observations are given in section 5.3.

5.2.2 Posterior asymptotic normality in the i.i.d. case

Consider the situation that the observation is a vector $X^{(n)} = (X_1,\ldots,X_n)$ and the model consists of n-fold product measures $P_\theta^{(n)} = P_\theta^n$, where the components P_θ are given by densities p_θ such that the maps $(\theta,x)\mapsto p_\theta(x)$ are measurable and $\theta\mapsto p_\theta$ is smooth (in the sense of lemma 5.2.3). Assume that the observations form an *i.i.d.* sample from a distribution P_0 with density p_0 relative to a common dominating measure. Assume that the Kullback-Leibler divergence of the model relative to P_0 is finite and minimized at $\theta^*\in\Theta$, *i.e.*:

$$-P_0\log\frac{p_{\theta^*}}{p_0} = \inf_{\theta\in\Theta} -P_0\log\frac{p_\theta}{p_0} < \infty. \tag{5.9}$$

In this situation we set $\delta_n = n^{-1/2}$ and use $\Delta_{n,\theta^*} = \sqrt{n}V_{\theta^*}^{-1}(\mathbb{P}_n - P_0)\dot\ell_{\theta^*}$ as the centering sequence (where $\dot\ell_{\theta^*}$ denotes the score function of the model $\theta\mapsto p_\theta$ at θ^*).

Lemmas that establish the LAN expansion (5.5) usually assume a well-specified model, whereas current interest requires local asymptotic normality in misspecified situations. To that end we consider the following lemma which gives sufficient conditions.

Lemma 5.2.3. *If the function $\theta\mapsto\log p_\theta(X_1)$ is differentiable at θ^* in P_0-probability with derivative $\dot\ell_{\theta^*}(X_1)$ and:*

(i) there is an open neighbourhood U of θ^* and a $L^2(P_0)$-function m_{θ^*} such that for all $\theta_1,\theta_2\in U$:

$$\left|\log\frac{p_{\theta_1}}{p_{\theta_2}}\right| \le m_{\theta^*}\|\theta_1-\theta_2\|, \quad (P_0-a.s.), \tag{5.10}$$

(ii) the Kullback-Leibler divergence with respect to P_0 has a 2nd-order Taylor-expansion around θ^*:

$$-P_0 \log \frac{p_\theta}{p_{\theta^*}} = \tfrac{1}{2}(\theta-\theta^*)V_{\theta^*}(\theta-\theta^*) + o(\|\theta-\theta^*\|^2), \quad (\theta \to \theta^*), \tag{5.11}$$

where V_{θ^*} is a positive-definite $k \times k$-matrix,

then (5.5) holds with $\delta_n = n^{-1/2}$ and $\Delta_{n,\theta^*} = \sqrt{n}V_{\theta^*}^{-1}(\mathbb{P}_n - P_0)\dot{\ell}_{\theta^*}$. Furthermore, the score function is bounded as follows:

$$\|\dot{\ell}_{\theta^*}(X)\| \leq m_{\theta^*}(X), \quad (P_0 - a.s.). \tag{5.12}$$

Finally, we have:

$$P_0\dot{\ell}_{\theta^*} = \frac{\partial}{\partial\theta}\big[P_0 \log p_\theta\big]_{\theta=\theta^*} = 0. \tag{5.13}$$

Proof. Using lemma 19.31 in Van der Vaart (1998) [248] for $\ell_\theta(X) = \log p_\theta(X)$, the conditions of which are satisfied by assumption, we see that for any sequence (h_n) that is bounded in P_0-probability:

$$\sqrt{n}(\mathbb{P}_n - P_0)\Big(\sqrt{n}\big(\ell_{\theta^*+(h_n/\sqrt{n})} - \ell_{\theta^*}\big) - h_n^T\dot{\ell}_{\theta^*}\Big) \xrightarrow{P_0} 0. \tag{5.14}$$

Hence, we see that,

$$n\mathbb{P}_n \log \frac{p_{\theta^*+h_n/\sqrt{n}}}{p_{\theta^*}} - \sqrt{n}h_n^T(\mathbb{P}_n - P_0)\dot{\ell}_{\theta^*} - nP_0 \log \frac{p_{\theta^*+h_n/\sqrt{n}}}{p_{\theta^*}} = o_{P_0}(1).$$

Using the second-order Taylor-expansion (5.11):

$$P_0 \log \frac{p_{\theta^*+h_n/\sqrt{n}}}{p_{\theta^*}} - \frac{1}{2n}h_n^T V_{\theta^*}h_n = o_{P_0}(1),$$

and substituting the log-likelihood product for the first term, we find (5.5).

Regarding the centering sequence Δ_{n,θ^*} and its relation to the maximum-likelihood estimator, we note the following lemma concerning the limit distribution of maximum-likelihood sequences.

Lemma 5.2.4. *Assume that the model satisfies the conditions of lemma 5.2.3 with non-singular V_{θ^*}. Then a sequence of estimators $\hat{\theta}_n$ such that $\hat{\theta}_n$ converges to θ^* in P_0-probability and,*

$$\mathbb{P}_n \log p_{\hat{\theta}_n} \geq \sup_\theta \mathbb{P}_n \log p_\theta - o_{P_0}(n^{-1}),$$

satisfies the asymptotic expansion:

$$\sqrt{n}(\hat{\theta}_n - \theta^*) = \frac{1}{\sqrt{n}}\sum_{i=1}^n V_{\theta^*}^{-1}\dot{\ell}_{\theta^*}(X_i) + o_{P_0}(1). \tag{5.15}$$

Proof. See van der Vaart (1998) [248], p. 54.

As noted earlier, lemma 5.2.4 implies that for consistent maximum-likelihood estimators the distribution of $\sqrt{n}(\hat{\theta}_n - \theta^*)$ has a normal limit with mean zero and covariance,

$$\Sigma = V_{\theta^*}^{-1} P_0(\dot{\ell}_{\theta^*} \dot{\ell}_{\theta^*}^T) V_{\theta^*}^{-1}.$$

More important for present purposes, however, is the fact that according to (5.15), the differences between $\sqrt{n}(\hat{\theta}_n - \theta^*)$ and Δ_{n,θ^*} go to zero in probability. The Bernstein-von Mises assertion (5.8) can also be formulated as in (5.4), which demonstrates the usual interpretation of the Bernstein-von Mises theorem most clearly: the sequence of posteriors resembles more-and-more closely a sequence of "sharpening" normal distributions centred at the maximum-likelihood estimators. More generally, any sequence of estimators satisfying (5.15) (any *best-regular* estimator sequence) may be used to centre the normal limit sequence. The conditions for lemma 5.2.4 are close to the conditions of the above Bernstein-von Mises theorem. As we have seen, in the well-specified situation the Lipschitz condition (5.10) can be replaced by the condition of differentiability in quadratic mean.

5.2.3 Asymptotic normality of point-estimators

Having discussed the posterior distributional limit, a natural question concerns the asymptotic properties of point-estimators derived from the posterior, like the posterior mean and median.

Based on the Bernstein-von Mises assertion (5.8) alone, one sees that any functional $f : \mathscr{P} \mapsto \mathbb{R}$, continuous relative to the total-variational norm, when applied to the sequence of posterior laws, converges to f applied to the normal limit distribution. Another general consideration follows from a generic construction of point-estimates from posteriors and demonstrate that posterior consistency at rate δ_n implies frequentist consistency at rate δ_n.

Theorem 5.2.5. *Let $X_1, \ldots, X_n$ be distributed i.i.d.-P_0 and let $\Pi(\cdot|X_1, \ldots, X_n)$ denote a sequence of posterior distributions on Θ that satisfies (5.7). Then there exist Bayesian point-estimators $\hat{\theta}_n$ such that:*

$$\delta_n^{-1}(\hat{\theta}_n - \theta^*) = O_{P_0}(1), \tag{5.16}$$

i.e. $\hat{\theta}_n$ is consistent and converges to θ^* at rate δ_n.

Proof. Define $\hat{\theta}_n$ to be the centre of a smallest ball that contains posterior mass at least $1/2$ (see remark 2.2.19). Because the ball around θ^* of radius $\delta_n M_n$ contains posterior mass tending to 1, the radius of a smallest ball must be bounded by $\delta_n M_n$ and the smallest ball must intersect the ball of radius $\delta_n M_n$ around θ^* with probability tending to 1. This shows that $\|\hat{\theta}_n - \theta^*\| \leq 2\delta_n M_n$ with probability tending to one.

This general point is more appropriate in non-parametric context and the above existence theorem does not pertain to the most widely-used Bayesian point-estimators. Asymptotic normality of the posterior mean in a misspecified model has been analysed in Bunke and Milhaud (1998) [53]; here, we generalize their discussion and prove asymptotic normality and efficiency for a class of point-estimators defined by a general loss function, which includes the posterior mean and median.

Let $\ell : \mathbb{R}^k \to [0,\infty)$ be a loss-function with the following properties: ℓ is continuous and satisfies, for every $M > 0$,

$$\sup_{\|h\| \leq M} \ell(h) \leq \inf_{\|h\| > 2M} \ell(h),$$

with strict inequality for some M. Furthermore, we assume that ℓ is subpolynomial, *i.e.* for some $p > 0$,

$$\ell(h) \leq 1 + \|h\|^p. \tag{5.17}$$

Define the estimators $\hat{\theta}_n$ as the formal Bayes estimators (see definition 2.2.16) that minimize,

$$t \mapsto \int \ell\big(\sqrt{n}(t-\theta)\big)\, d\Pi(\theta|X_1,\ldots,X_n).$$

Theorem 5.2.6. *Assume that the model satisfies (5.5) for some $\theta^* \in \Theta$ and that the conditions of theorems 5.3.1 are satisfied. Let $\ell : \mathbb{R}^k \to [0,\infty)$ be a loss-function with the properties listed and assume that $\int \|\theta\|^p \, d\Pi(\theta) < \infty$. Then under P_0, the sequence $\sqrt{n}(\hat{\theta}_n - \theta^*)$ converges weakly to the minimizer of,*

$$t \mapsto Z(t) = \int \ell(t-h)\, dN_{X, V_{\theta^*}^{-1}}(h),$$

where $X \sim N(0, V_{\theta^*}^{-1} P_0(\dot{\ell}_{\theta^*} \dot{\ell}_{\theta^*}^T) V_{\theta^*}^{-1})$, provided that any two minimizers of this process coincide almost-surely. In particular, if the loss function is subconvex (e.g. $\ell(x) = \|x\|^2$ or $\ell(x) = \|x\|$, giving the posterior mean and median), then $\sqrt{n}(\hat{\theta}_n - \theta^*)$ converges weakly to X under P_0.

Proof. The theorem can be proved along the same lines as theorem 10.8 in [248]. The main difference is in proving that, for any $M_n \to \infty$,

$$U_n := \int_{\|h\| > M_n} \|h\|^p \, d\Pi_n(h|X_1,\ldots,X_n) \xrightarrow{P_0} 0. \tag{5.18}$$

Here, abusing notation, we write $d\Pi_n(h|X_1,\ldots,X_n)$ to denote integrals relative to the posterior distribution of the local parameter $h = \sqrt{n}(\theta - \theta^*)$. Under misspecification a new proof is required, for which we extend the proof of theorem 5.3.1 below.

Once (5.18) is established, the proof continues as follows. The variable $\hat{h}_n = \sqrt{n}(\hat{\theta}_n - \theta)$ is the maximizer of the process $t \mapsto \int \ell(t-h)\, d\Pi_n(h|X_1,\ldots,X_n)$. Then $\hat{h}_n = O_{P_0}(1)$. Fix some compact set K and for given $M > 0$ define the processes

$$t \mapsto Z_{n,M}(t) = \int_{\|h\| \le M} \ell(t-h)\, d\Pi_n(h|X_1,\ldots,X_n)$$
$$t \mapsto W_{n,M}(t) = \int_{\|h\| \le M} \ell(t-h)\, dN_{\Delta_n, V_{\theta^*}^{-1}}(h)$$
$$t \mapsto W_M(t) = \int_{\|h\| \le M} \ell(t-h)\, dN_{X, V_{\theta^*}^{-1}}(h)$$

Since $\sup_{t\in K, \|h\|\le M} \ell(t-h) < \infty$, $Z_{n,M} - W_{n,M} = o_{P_0}(1)$ in $\ell^\infty(K)$ by theorem 5.2.2. Since Δ_n converges weakly to X under P_0^n, the *continuous mapping theorem* implies that $W_{n,M} \xrightarrow{P_0\text{-w.}} W_M$ in $\ell^\infty(K)$. Because ℓ has subpolynomial tails, integrable with respect to $N_{X,V_{\theta^*}^{-1}}$, $W_M \xrightarrow{P_0} Z$ in $\ell^\infty(K)$ as $M \to \infty$. Thus $Z_{n,M} \xrightarrow{P_0\text{-w.}} W_M$ in $\ell^\infty(K)$, for every $M > 0$, and $W_M \xrightarrow{P_0} Z$ as $M \to \infty$. We conclude that there exists a sequence $M_n \to \infty$ such that $Z_{n,M_n} \xrightarrow{P_0\text{-w.}} Z$. The limit (5.18) implies that $Z_{n,M_n} - Z_n = o_{P_0}(1)$ in $\ell^\infty(K)$ and we conclude that $Z_n \xrightarrow{P_0\text{-w.}} Z$ in $\ell^\infty(K)$. Due to the continuity of ℓ, $t \mapsto Z(t)$ is continuous almost surely. This, together with the assumed unicity of maxima of these sample paths, enables the argmax theorem (corollary 5.58 in [248]) and we conclude that $\hat{h}_n \xrightarrow{P_0\text{-w.}} \hat{h}$, where $\hat{h}$ is the minimizer of $Z(t)$.

For the proof of (5.18) we adopt the notation of theorem 5.3.1. The tests ω_n employed there can be taken nonrandomized without loss of generality (otherwise replace them for instance by $1_{\omega_n > 1/2}$) and then $U_n\omega_n$ tends to zero in probability because ω_n does so. Thus (5.18) is proved once it is established that, with $\varepsilon_n = M_n/\sqrt{n}$, the sequences,

$$d_n = P_0^n(1-\omega_n)\, 1_{\Omega\setminus\Xi_n} \int_{\varepsilon_n \le \|\theta-\theta^*\| < \varepsilon} n^{p/2}\|\theta-\theta^*\|^p\, d\Pi(\theta \mid X_1,\ldots,X_n)$$
$$d_n' = P_0^n(1-\omega_n)\, 1_{\Omega\setminus\Omega_n} \int_{\|\theta-\theta^*\| \ge \varepsilon} n^{p/2}\|\theta-\theta^*\|^p\, d\Pi(\theta \mid X_1,\ldots,X_n)$$

go to zero. We can use bounds as in the proof of theorem 5.3.1, but instead of (5.23), we arrive at the bounds,

$$d_n \le \frac{e^{n(a_n^2(1+C)-D\varepsilon^2)}}{\Pi\big(B(a_n,\theta^*;P_0)\big)} n^{p/2} \int \|\theta-\theta^*\|^p\, d\Pi(\theta),$$
$$d_n' \le K' e^{-\frac{1}{2}nD\varepsilon_n^2} \sum_{j=1}^{\infty} n^{p/2}(j+1)^{d+p}\varepsilon_n^p e^{-nD(j^2-1)\varepsilon_n^2},$$

both of which tend to zero. The last assertion of the theorem follows, because for a subconvex loss function the process Z is minimized uniquely by X, as a consequence of lemma 4.1.19).

5.3 Rate of convergence

In a Bayesian context, the rate of convergence is defined as the maximal pace at which balls around the point of convergence can be shrunk to radius zero while still capturing a posterior mass that converges to one asymptotically (see definition 6.4.1). Current interest lies in the fact that the Bernstein-von Mises theorem of the previous section formulates condition (5.7) (with $\delta_n = n^{-1/2}$),

$$\Pi\big(\,\|\vartheta - \theta^*\| \geq M_n/\sqrt{n} \;\big|\; X_1, \ldots, X_n\,\big) \xrightarrow{P} 0,$$

for all $M_n \to \infty$. As we shall see in chapter 6, a convenient way of establishing the above is through the condition that suitable test sequences exist. As was shown in a well-specified context in Ghosal *et al.* (2000) [106] and under misspecification in Kleijn and Van der Vaart (2006) [151], the most important requirement for convergence of the posterior at a certain rate is the existence of a test-sequence that separates the point of convergence from the complements of balls shrinking at said rate. In chapters 6 and 7, we extend this point further.

This is also the approach we follow here: we show that the sequence of posterior probabilities in the above display converges to zero in P_0-probability if a test sequence exists that is suitable in the sense given above (see the proof of theorem 5.3.1). However, under the regularity conditions that were formulated to establish local asymptotic normality under misspecification in the previous section, more can be said: not complements of shrinking balls, but fixed alternatives are to be suitably testable against P_0, thus relaxing the testing condition considerably. Locally, the construction relies on score-tests to separate the point of convergence from complements of neighbourhoods shrinking at rate $1/\sqrt{n}$, using Bernstein's inequality to obtain exponential power. The tests for fixed alternatives are used to extend those local tests to the full model.

In this section we prove that a prior mass condition and suitable test sequences suffice to prove convergence at the rate required for the Bernstein-von Mises theorem formulated in section 5.2. The theorem that begins the next subsection summarizes the conclusion. Throughout the section we consider the *i.i.d.* case, with notation as in subsection 5.2.2.

5.3.1 Posterior rate of convergence

With use of theorem 5.3.5, we formulate a theorem that ensures $\sqrt{n}$-rate of convergence for the posterior distributions of smooth, testable models with sufficient prior mass around the point of convergence. The testability condition is formulated using measures Q_θ, defined by,

$$Q_\theta(A) = P_0\Big(\frac{p_\theta}{p_{\theta^*}} 1_A\Big),$$

for all $A \in \mathscr{A}$ and all $\theta \in \Theta$. Note that all Q_θ are dominated by P_0 and that $Q_{\theta^*} = P_0$. Also note that if the model is well-specified, then $P_{\theta^*} = P_0$ and $Q_\theta = P_\theta$ for all θ. Therefore the use of Q_θ instead of P_θ to formulate the testing condition is relevant only in the misspecified situation (see Kleijn and Van der Vaart (2006) [151] for more on this subject). The proof of theorem 5.3.1 makes use of Kullback-Leibler neighbourhoods of θ^* of the form:

$$B(\varepsilon, \theta^*; P_0) = \Big\{ \theta \in \Theta : -P_0 \log \frac{p_\theta}{p_{\theta^*}} \leq \varepsilon^2, P_0 \Big(\log \frac{p_\theta}{p_{\theta^*}} \Big)^2 \leq \varepsilon^2 \Big\}, \tag{5.19}$$

for some $\varepsilon > 0$.

Theorem 5.3.1. *Assume that the model $\mathscr{P}$ satisfies the smoothness conditions of lemma 5.2.3, where in addition, it is required that $P_0(p_\theta / p_{\theta^*}) < \infty$ for all θ in a neighbourhood of θ^* and $P_0(e^{s m_{\theta^*}}) < \infty$ for some $s > 0$. Assume that the prior possesses a density that is continuous and positive in a neighbourhood of θ^*. Furthermore, assume that $P_0(\dot{\ell}_{\theta^*} \dot{\ell}_{\theta^*}^T)$ is invertible and that for every $\varepsilon > 0$ there exists a sequence of tests (ϕ_n) such that:*

$$P_0^n \phi_n \to 0, \qquad \sup_{\{\theta : \|\theta - \theta^*\| \geq \varepsilon\}} Q_\theta^n (1 - \phi_n) \to 0. \tag{5.20}$$

Then the posterior converges at rate $1/\sqrt{n}$, i.e. for every sequence (M_n), $M_n \to \infty$:

$$\Pi \big(\theta \in \Theta : \|\theta - \theta^*\| \geq M_n / \sqrt{n} \,\big|\, X_1, X_2, \ldots, X_n \big) \xrightarrow{P_0} 0.$$

Proof. Let (M_n) be given, and define the sequence (ε_n) by $\varepsilon_n = M_n / \sqrt{n}$. According to theorem 5.3.5 there exists a sequence of tests (ω_n) and constants $D > 0$ and $\varepsilon > 0$ such that (5.25) holds. We use these tests to split the P_0^n-expectation of the posterior measure as follows:

$$\begin{aligned} &P_0^n \Pi \big(\|\vartheta - \theta^*\| \geq \varepsilon_n \,\big|\, X_1, X_2, \ldots, X_n \big) \\ &\qquad \leq P_0^n \omega_n + P_0^n (1 - \omega_n) \Pi \big(\|\vartheta - \theta^*\| \geq \varepsilon_n \,\big|\, X_1, X_2, \ldots, X_n \big). \end{aligned}$$

The first term is of order $o(1)$ as $n \to \infty$ by (5.25). Given a constant $\varepsilon > 0$ (to be specified later), the second term can be decomposed as:

$$\begin{aligned} &P_0^n (1 - \omega_n) \Pi \big(\|\vartheta - \theta^*\| \geq \varepsilon_n \,\big|\, X_1, X_2, \ldots, X_n \big) \\ &\quad = P_0^n (1 - \omega_n) \Pi \big(\|\vartheta - \theta^*\| \geq \varepsilon \,\big|\, X_1, X_2, \ldots, X_n \big) \\ &\qquad + P_0^n (1 - \omega_n) \Pi \big(\varepsilon_n \leq \|\vartheta - \theta^*\| < \varepsilon \,\big|\, X_1, X_2, \ldots, X_n \big). \end{aligned} \tag{5.21}$$

Given two constants $M, M' > 0$ (also to be specified at a later stage), we define the sequences (a_n), $a_n = M \sqrt{\log n / n}$ and (b_n), $b_n = M' \varepsilon_n$. Based on a_n and b_n, we define two sequences of events:

$$\Xi_n = \Big\{ \int_\Theta \prod_{i=1}^n \frac{p_\theta}{p_{\theta^*}}(X_i)\, d\Pi(\theta) \le \Pi\big(B(a_n,\theta^*;P_0)\big) e^{-na_n^2(1+C)} \Big\},$$
$$\Omega_n = \Big\{ \int_\Theta \prod_{i=1}^n \frac{p_\theta}{p_{\theta^*}}(X_i)\, d\Pi(\theta) \le \Pi\big(B(b_n,\theta^*;P_0)\big) e^{-nb_n^2(1+C)} \Big\}.$$

The sequence (Ξ_n) is used to split the first term on the *r.h.s.* of (5.21) and estimate it as follows:

$$\begin{aligned} &P_0^n(1-\omega_n)\Pi\big(\|\vartheta-\theta^*\| \ge \varepsilon \,\big|\, X_1,X_2,\ldots,X_n\big) \\ &\quad \le P_0(\Xi_n) + P_0^n(1-\omega_n)\, 1_{\Omega\setminus\Xi_n}\, \Pi\big(\|\vartheta-\theta^*\| \ge \varepsilon \,\big|\, X_1,X_2,\ldots,X_n\big). \end{aligned}$$

According to lemma 5.3.3, the first term is of order $o(1)$ as $n\to\infty$. The second term is estimated further with the use of lemmas 5.3.3, 5.3.4 and theorem 5.3.5: for some $C>0$,

$$\begin{aligned} &P_0^n(1-\omega_n)\, 1_{\Omega\setminus\Xi_n}\, \Pi\big(\|\vartheta-\theta^*\| \ge \varepsilon \,\big|\, X_1,X_2,\ldots,X_n\big) \\ &\quad \le \frac{e^{na_n^2(1+C)}}{\Pi\big(B(a_n,\theta^*;P_0)\big)} \int_{\{\theta:\|\theta-\theta^*\|\ge\varepsilon\}} Q_\theta^n(1-\omega_n)\, d\Pi(\theta) \\ &\quad \le \frac{e^{n(a_n^2(1+C)-D\varepsilon^2)}}{\Pi\big(B(a_n,\theta^*;P_0)\big)} \Pi\big(\|\vartheta-\theta^*\| \ge \varepsilon\big). \end{aligned} \tag{5.22}$$

Note that $a_n^2(1+C) - D\varepsilon^2 \le -a_n^2(1+C)$ for large enough n, so that:

$$\frac{e^{n(a_n^2(1+C)-D\varepsilon^2)}}{\Pi\big(B(a_n,\theta^*;P_0)\big)} \le K^{-1} e^{-na_n^2(1+C)} (a_n)^{-k} \le \frac{1}{M^{k/2}K} (\log n)^{-k/2} n^{-M^2(1+C)+\frac{k}{2}},$$

for large enough n, using (5.24). A large enough choice for the constant M then ensures that the expression on the *l.h.s.* in the next-to-last display is of order $o(1)$ as $n\to\infty$.

The sequence (Ω_n) is used to split the second term on the *r.h.s.* of (5.21) after which we estimate it in a similar manner. Again the term that derives from 1_{Ω_n} is of order $o(1)$, and

$$\begin{aligned} &P_0^n(1-\omega_n)\, 1_{\Omega\setminus\Omega_n}\, \Pi\big(\varepsilon_n \le \|\theta-\theta^*\| < \varepsilon \,\big|\, X_1,X_2,\ldots,X_n\big) \\ &\quad \le \frac{e^{nb_n^2(1+C)}}{\Pi\big(B(b_n,\theta^*;P_0)\big)} \sum_{j=1}^J \int_{A_{n,j}} Q_\theta^n(1-\omega_n)\, d\Pi(\theta), \end{aligned}$$

where we have split the domain of integration into spherical shells $A_{n,j}$, $(1\le j\le J$, with J the smallest integer such that $(J+1)\varepsilon_n > \varepsilon)$: $A_{n,j} = \big\{\theta : j\varepsilon_n \le \|\theta-\theta^*\| \le \big((j+1)\varepsilon_n\big)\wedge\varepsilon\big\}$. Applying theorem 5.3.5 to each of the shells separately, we obtain:

$$
\begin{aligned}
&P_0^n(1-\omega_n)\,1_{\Omega\setminus\Omega_n}\,\Pi\big(\,\varepsilon_n \le \|\vartheta-\theta^*\| < \varepsilon \;\big|\; X_1,X_2,\ldots,X_n\big)\\
&\quad=\sum_{j=1}^{J} e^{nb_n^2(1+C)} \sup_{\theta\in A_{n,j}} Q_\theta^n(1-\omega_n)\,\frac{\Pi(A_{n,j})}{\Pi\big(B(b_n,\theta^*;P_0)\big)}\\
&\quad\le\sum_{j=1}^{J} e^{nb_n^2(1+C)-nDj^2\varepsilon_n^2}\,\frac{\Pi\big(\,\|\vartheta-\theta^*\|\le (j+1)\varepsilon_n\big)}{\Pi\big(B(b_n,\theta^*;P_0)\big)}.
\end{aligned}
$$

For a small enough ε and large enough n, the sets $\big\{\,\theta : \|\theta-\theta^*\|\le (j+1)\varepsilon_n\,\big\}$ all fall within the neighbourhood U of θ^* on which the prior density π is continuous. Hence π is uniformly bounded by a constant $R>0$ and we see that: $\Pi\{\,\theta : \|\theta-\theta^*\|\le (j+1)\varepsilon_n\,\}\le RV_k(j+1)^k\varepsilon_n^k$, where V_k is the Lebesgue-volume of the k-dimensional ball of unit radius. Combining this with (5.24), there exists a constant $K'>0$ such that, with $M'<\sqrt{D/2(1+C)}$:

$$
\begin{aligned}
&P_0^n(1-\omega_n)\,1_{\Omega\setminus\Omega_n}\,\Pi\big(\,\varepsilon_n \le \|\vartheta-\theta^*\| < \varepsilon \;\big|\; X_1,\ldots,X_n\big)\\
&\qquad\le K' e^{-\frac{1}{2}nD\varepsilon_n^2}\sum_{j=1}^{\infty}(j+1)^k e^{-nD(j^2-1)\varepsilon_n^2},
\end{aligned}
\tag{5.23}
$$

for large enough n. The series is convergent and we conclude that this term is also of order $o(1)$ as $n\to\infty$.

Consistent uniform testability of the type (5.20) is a relatively weak requirement, inspired by Schwartz' conditions for non-parametric consistency in well-specified setting (see section 6.3). To demonstrate its usefulness, we show in the next theorem that suitable tests exist as soon as the parameter set is compact and the model is suitably continuous in the parameter.

Theorem 5.3.2. *Assume that Θ is compact and that θ^* is a unique point of minimum of $\theta\mapsto -P_0\log p_\theta$. Furthermore assume that $P_0(p_\theta/p_{\theta^*})<\infty$ for all $\theta\in\Theta$ and that the map,*

$$
\theta\mapsto P_0\Big(\frac{p_\theta}{p_{\theta_1}^s\,p_{\theta^*}^{1-s}}\Big),
$$

is continuous at θ_1 for every s in a left neighbourhood of 1, for every θ_1. Then there exist tests ϕ_n satisfying (5.20). A sufficient condition is that for every $\theta_1\in\Theta$ the maps $\theta\mapsto p_\theta/p_{\theta_1}$ and $\theta\mapsto p_\theta/p_{\theta^*}$ are continuous in $L_1(P_0)$ at $\theta=\theta_1$.

Proof. For given $\theta_1\neq\theta^*$ consider the tests,

$$
\phi_{n,\theta_1}=1\{\mathbb{P}_n\log(p_0/q_{\theta_1})<0\}.
$$

Because $\mathbb{P}_n\log(p_0/q_{\theta_1})\to P_0\log(p_0/q_{\theta_1})$ in P_0^n-probability by the *law of large numbers*, and $P_0\log(p_0/q_{\theta_1})=P_0\log(p_{\theta^*}/p_{\theta_1})>0$ for $\theta_1\neq\theta^*$ by the definition of θ^* we have that $P_0^n\phi_{n,\theta_1}\to 0$ as $n\to\infty$. By Markov's inequality we have that,

$$Q_\theta^n(1-\phi_{n,\theta_1}) = Q_\theta^n\left(e^{sn\mathbb{P}_n \log(p_0/q_{\theta_1})} > 1\right)$$
$$\leq Q_\theta^n e^{sn\mathbb{P}_n \log(p_0/q_{\theta_1})} = \left(Q_\theta (p_0/q_{\theta_1})^s\right)^n = \rho(\theta_1,\theta,s)^n,$$

for $\rho(\theta_1,\theta,s) = \int p_0^s q_{\theta_1}^{-s} q_\theta \, d\mu$. It is known [151] that the Hellinger transform (see also [236] and [179]) $s \mapsto \rho(\theta_1,\theta_1,s)$ tends to $P_0(q_{\theta_1} > 0) = P_0(p_{\theta_1} > 0)$ as $s \uparrow 1$ and has derivative from the left equal to $P_0 \log(q_{\theta_1}/p_0) 1_{q_{\theta_1}>0} = P_0 \log(p_{\theta_1}/p_{\theta^*}) 1_{p_{\theta_1}>0}$ at $s = 1$. We have that either $P_0(p_{\theta_1} > 0) < 1$ or $P_0(p_{\theta_1} > 0) = 1$ and $P_0 \log(p_{\theta_1}/p_{\theta^*}) 1_{p_{\theta_1}>0} = P_0 \log(p_{\theta_1}/p_{\theta^*}) < 0$ (or both). In all cases there exists $s_{\theta_1} < 1$ arbitrarily close to 1 such that $\rho(\theta_1,\theta_1,s_{\theta_1}) < 1$. By assumption the map $\theta \mapsto \rho(\theta_1,\theta,s_{\theta_1})$ is continuous at θ_1. Therefore, for every θ_1 there exists an open neighbourhood G_{θ_1} such that,

$$r_{\theta_1} = \sup_{\theta \in G_{\theta_1}} \rho(\theta_1,\theta,s_{\theta_1}) < 1.$$

The set $\{\theta \in \Theta : \|\theta - \theta^*\| \geq \varepsilon\}$ is compact and hence can be covered with finitely many sets of the type G_{θ_1}, say G_{θ_i} for $1 = 1,\ldots,l$. We now define

$$\phi_n = \max_{i=1,\ldots,l} \phi_{n,\theta_i}.$$

This test satisfies

$$P_0^n \phi_n \leq \sum_{i=1}^l P_0^n \phi_{n,\theta_i} \to 0,$$
$$Q_\theta^n(1-\phi_n) \leq \max_{i=1}^l Q_\theta^n(1-\phi_{n,\theta_i}) \leq \max_{i=1}^l r_{\theta_i}^n \to 0,$$

uniformly in $\theta \in \cup_{i=1}^l G_{\theta_i}$. Therefore the tests ϕ_n satisfy the requirements. To prove the last assertion we write $\rho(\theta_1,\theta,s) = P_0(p_\theta/p_{\theta_1})^s (p_\theta/p_{\theta^*})^{1-s}$. Continuity of the maps $\theta \mapsto (p_\theta/p_{\theta_1})$ and $\theta \mapsto (p_\theta/p_{\theta^*})$ in $L_1(P_0)$ can be seen to imply the required continuity of the map $\theta \mapsto \rho(\theta_1,\theta,s)$.

Beyond compactness it appears impossible to give mere qualitative sufficient conditions, like continuity, for consistent testability. For "natural" parametrizations it ought to be true that distant parameters (outside a given compact) are the easy ones to test for (and a test designed for a given compact ought to be consistent even for points outside the compact), but this depends on the structure of the model. Alternatively, many models allow either approximation by compacts from within, or a suitable compactification in which the preceding result can be applied, but we omit a discussion.

In the proof of theorem 5.3.1, lower bounds in probability on the denominators of posterior probabilities are needed, as provided by the following lemma.

Lemma 5.3.3. *For given $\varepsilon > 0$ and $\theta^* \in \Theta$ such that $P_0 \log(p_0/p_{\theta^*}) < \infty$ define $B(\varepsilon,\theta^*;P_0)$ by (5.19). Then for every $C > 0$ and probability measure Π on Θ:*

$$P_0^n\Big(\int_\Theta \prod_{i=1}^n \frac{p_\theta}{p_{\theta^*}}(X_i)\,d\Pi(\theta) \le \Pi\big(B(\varepsilon,\theta^*;P_0)\big)e^{-n\varepsilon^2(1+C)}\Big) \le \frac{1}{C^2 n\varepsilon^2}.$$

Proof. This lemma can be found as lemma 7.1 in Kleijn and Van der Vaart (2006) [151], and follows essentially the same steps as lemma 6.4.6.

Moreover, the prior mass of the Kullback-Leibler neighbourhoods $B(\varepsilon,\theta^*;P_0)$ can be lower-bounded if we make the regularity assumptions for the model used in section 5.2 and the assumption that the prior has a Lebesgue density that is well-behaved at θ^*.

Lemma 5.3.4. *Under the smoothness conditions of lemma 5.2.3 and assuming that the prior density π is continuous and strictly positive in θ^*, there exists a constant $K>0$ such that the prior mass of the Kullback-Leibler neighbourhoods $B(\varepsilon,\theta^*;P_0)$ satisfies:*

$$\Pi\big(B(\varepsilon,\theta^*;P_0)\big) \ge K\varepsilon^k, \tag{5.24}$$

for small enough $\varepsilon>0$.

Proof. As a result of the smoothness conditions, we have, for some constants $c_1,c_2>0$ and small enough $\|\theta-\theta^*\|$,

$$-P_0\log(p_\theta/p_{\theta^*}) \le c_1\|\theta-\theta^*\|^2, \quad P_0(\log(p_\theta/p_{\theta^*}))^2 \le c_2\|\theta-\theta^*\|^2.$$

Defining $c=(1/c_1\wedge 1/c_2)^{1/2}$, this implies that for small enough $\varepsilon>0$, $\{\theta\in\Theta:\|\theta-\theta^*\|\le c\varepsilon\}\subset B(\varepsilon,\theta^*;P_0)$. Since the Lebesgue-density π of the prior is continuous and strictly positive in θ^*, we see that there exists a $\delta'>0$ such that for all $0<\delta\le\delta'$: $\Pi\big(\theta\in\Theta:\|\theta-\theta^*\|\le\delta\big)\ge\frac{1}{2}V_k\pi(\theta^*)\,\delta^k>0$. Hence, for small enough ε, $c\varepsilon\le\delta'$ and we obtain (5.24) upon combination.

5.3.2 Suitable test sequences

In this subsection we prove that the existence of test sequences (under misspecification) of uniform exponential power for complements of shrinking balls around θ^* versus P_0 (as needed in the proof of theorem 5.3.1), is guaranteed whenever asymptotically consistent test-sequences exist for complements of *fixed* balls around θ^* versus P_0 and the conditions of lemmas 5.2.3 and 5.3.6 are met.

Theorem 5.3.5. *Assume that the conditions of lemma 5.2.3 are satisfied, where in addition, it is required that $P_0(p_\theta/p_{\theta^*})<\infty$ for all θ in a neighbourhood of θ^* and $P_0(e^{sm_{\theta^*}})<\infty$ for some $s>0$. Furthermore, suppose that $P_0\dot{\ell}_{\theta^*}\dot{\ell}_{\theta^*}^T$ is invertible and for every $\varepsilon>0$ there exists a sequence of test functions (ϕ_n), such that:*

$$P_0^n\phi_n\to 0, \qquad \sup_{\{\theta:\|\theta-\theta^*\|\ge\varepsilon\}} Q_\theta^n(1-\phi_n)\to 0.$$

Then for every sequence (M_n) such that $M_n \to \infty$ there exists a sequence of tests (ω_n) such that for some constants $D > 0$, $\varepsilon > 0$ and large enough n:

$$P_0^n \omega_n \to 0, \quad Q_\theta^n (1 - \omega_n) \leq e^{-nD(\|\theta - \theta^*\|^2 \wedge \varepsilon^2)}, \tag{5.25}$$

for all $\theta \in \Theta$ such that $\|\theta - \theta^*\| \geq M_n/\sqrt{n}$.

Proof. Let (M_n) be given. We construct two sequences of tests: one sequence to test P_0 versus $\{Q_\theta : \theta \in \Theta_1\}$ with $\Theta_1 = \{\theta \in \Theta : M_n/\sqrt{n} \leq \|\theta - \theta^*\| \leq \varepsilon\}$, and the other to test P_0 versus $\{Q_\theta : \theta \in \Theta_2\}$ with $\Theta_2 = \{\theta : \|\theta - \theta^*\| > \varepsilon\}$, both uniformly with exponential power (for a suitable choice of ε). We combine these sequences to test P_0 versus $\{Q_\theta : \|\theta - \theta^*\| \geq M_n/\sqrt{n}\}$ uniformly with exponential power.

For the construction of the first sequence, a constant $L > 0$ is chosen to truncate the score-function component-wise (*i.e.* for all $1 \leq k \leq d$, $(\dot{\ell}^L_{\theta^*})_k = 0$ if $|(\dot{\ell}_{\theta^*})_k| \geq L$ and $(\dot{\ell}^L_{\theta^*})_k = (\dot{\ell}_{\theta^*})_k$ otherwise) and we define:

$$\omega_{1,n} = 1\big\{\|(\mathbb{P}_n - P_0)\dot{\ell}^L_{\theta^*}\| > \sqrt{M_n/n}\big\},$$

Because the function $\dot{\ell}_{\theta^*}$ is square-integrable, we can ensure that the matrices $P_0(\dot{\ell}_{\theta^*}\dot{\ell}^T_{\theta^*})$, $P_0(\dot{\ell}_{\theta^*}(\dot{\ell}^L_{\theta^*})^T)$ and $P_0(\dot{\ell}^L_{\theta^*}(\dot{\ell}^L_{\theta^*})^T)$ are arbitrarily close (for instance in operator norm) by a sufficiently large choice for the constant L. We fix such an L throughout the proof.

By the *central limit theorem* $P_0^n \omega_{1,n} = P_0^n\big(\|\sqrt{n}(\mathbb{P}_n - P_0)\dot{\ell}^L_{\theta^*}\|^2 > M_n\big) \to 0$. Turning to $Q_\theta^n(1 - \omega_{1,n})$ for $\theta \in \Theta_1$, we note that for all θ:

$$\begin{aligned} Q_\theta^n\Big(\|(\mathbb{P}_n - P_0)\dot{\ell}^L_{\theta^*}\| \leq \sqrt{M_n/n}\Big) &= Q_\theta^n\Big(\sup_{v \in S} v^T(\mathbb{P}_n - P_0)\dot{\ell}^L_{\theta^*} \leq \sqrt{M_n/n}\Big) \\ &\leq \inf_{v \in S} Q_\theta^n\Big(v^T(\mathbb{P}_n - P_0)\dot{\ell}^L_{\theta^*} \leq \sqrt{M_n/n}\Big), \end{aligned}$$

where S is the sphere of unity in $\mathbb{R}^k$. With the choice $v = (\theta - \theta^*)/\|\theta - \theta^*\|$ as an upper bound for the *r.h.s.* in the above display, we note that:

$$\begin{aligned} &Q_\theta^n\Big((\theta - \theta^*)^T(\mathbb{P}_n - P_0)\dot{\ell}^L_{\theta^*} \leq \sqrt{M_n/n}\|\theta - \theta^*\|\Big) \\ &= Q_\theta^n\Big((\theta^* - \theta)^T(\mathbb{P}_n - \tilde{Q}_\theta)\dot{\ell}^L_{\theta^*} \geq (\theta - \theta^*)^T(\tilde{Q}_\theta - \tilde{Q}_{\theta^*})\dot{\ell}^L_{\theta^*} - \sqrt{M_n/n}\|\theta - \theta^*\|\Big), \end{aligned}$$

where we have used the notation (for all $\theta \in \Theta_1$ with small enough $\varepsilon > 0$) $\tilde{Q}_\theta = \|Q_\theta\|^{-1}Q_\theta$ and the fact that $P_0 = Q_{\theta^*} = \tilde{Q}_{\theta^*}$. By straightforward manipulation, we find:

$$\begin{aligned} &(\theta - \theta^*)^T\big(\tilde{Q}_\theta - \tilde{Q}_{\theta^*}\big)\dot{\ell}^L_{\theta^*} \\ &\quad = \frac{1}{P_0(p_\theta/p_{\theta^*})}(\theta - \theta^*)^T\Big(P_0\big((p_\theta/p_{\theta^*} - 1)\dot{\ell}^L_{\theta^*}\big) + \big(1 - P_0(p_\theta/p_{\theta^*})\big)P_0\dot{\ell}^L_{\theta^*}\Big). \end{aligned}$$

In view of lemma 5.3.6 and conditions (5.10), (5.11), $(P_0(p_\theta/p_{\theta^*}) - 1)$ is of order $O(\|\theta - \theta^*\|^2)$ as $(\theta \to \theta^*)$, which means that if we approximate the above display

up to order $o(\|\theta-\theta^*\|^2)$, we can limit attention on the *r.h.s.* to the first term in the last factor and equate the first factor to 1. Furthermore, using the differentiability of $\theta \mapsto \log(p_\theta/p_{\theta^*})$, condition (5.10) and lemma 5.3.6, we see that:

$$\begin{aligned} &P_0\left\|\Big(\frac{p_\theta}{p_{\theta^*}}-1-(\theta-\theta^*)^T\dot\ell_{\theta^*}\Big)\dot\ell^L_{\theta^*}\right\| \\ &\quad\le P_0\left\|\Big(\frac{p_\theta}{p_{\theta^*}}-1-\log\frac{p_\theta}{p_{\theta^*}}\Big)\dot\ell^L_{\theta^*}\right\| + P_0\left\|\Big(\log\frac{p_\theta}{p_{\theta^*}}-(\theta-\theta^*)^T\dot\ell_{\theta^*}\Big)\dot\ell^L_{\theta^*}\right\|, \end{aligned}$$

which is $o\big(\|\theta-\theta^*\|\big)$. Also note that since $M_n\to\infty$ and for all $\theta\in\Theta_1$, $\|\theta-\theta^*\|\ge M_n/\sqrt{n}$, $-\|\theta-\theta^*\|\sqrt{M_n/n}\ge -\|\theta-\theta^*\|^2(M_n)^{-1/2}$. Summarizing the above and combining with the remark made at the beginning of the proof concerning the choice of L, we find that for every $\delta>0$, there exist choices of $\varepsilon>0$, $L>0$ and $N\ge 1$ such that for all $n\ge N$ and all θ in Θ_1:

$$\begin{aligned} (\theta-\theta^*)^T\big(\tilde Q_\theta-\tilde Q_{\theta^*}\big)\dot\ell^L_{\theta^*} &- \sqrt{M_n/n}\|\theta-\theta^*\| \\ &\ge (\theta-\theta^*)^T P_0\big(\dot\ell_{\theta^*}\dot\ell^T_{\theta^*}\big)(\theta-\theta^*)-\delta\|\theta-\theta^*\|^2. \end{aligned}$$

We denote $\Delta(\theta)=(\theta-\theta^*)^TP_0(\dot\ell_{\theta^*}\dot\ell^T_{\theta^*})(\theta-\theta^*)$ and since $P_0(\dot\ell_{\theta^*}\dot\ell^T_{\theta^*})$ is strictly positive definite by assumption, its smallest eigenvalue c is greater than zero. Hence, $-\delta\|\theta-\theta^*\|^2\ge-\delta/c\,\Delta(\theta)$. and there exists a constant $r(\delta)$ (depending only on the matrix $P_0(\dot\ell_{\theta^*}\dot\ell^T_{\theta^*})$ and with the property that $r(\delta)\to 1$ if $\delta\to 0$) such that:

$$Q^n_\theta(1-\omega_{1,n})\le Q^n_\theta\Big((\theta^*-\theta)^T(\mathbb{P}_n-\tilde Q_\theta)\dot\ell^L_{\theta^*}\ge r(\delta)\Delta(\theta)\Big),$$

for small enough ε, large enough L and large enough n, demonstrating that the type-II error is bounded above by the (unnormalized) tail probability $Q^n_\theta(\bar W_n\ge r(\delta)\Delta(\theta))$ of the mean of the variables $W_i=(\theta^*-\theta)^T(\dot\ell^L_{\theta^*}(X_i)-\tilde Q_\theta\dot\ell^L_{\theta^*})$, $(1\le i\le n)$. so that $\tilde Q_\theta W_i=0$. The random variables W_i are independent and bounded since:

$$|W_i|\le\|\theta-\theta^*\|\big(\|\dot\ell^L_{\theta^*}(X_i)\|+\|\tilde Q_\theta\dot\ell^L_{\theta^*}\|\big)\le 2L\sqrt{d}\|\theta-\theta^*\|.$$

The variance of W_i under $\tilde Q_\theta$ is expressed as follows:

$$\mathrm{Var}_{\tilde Q_\theta}W_i=(\theta-\theta^*)^T\Big(\big)\tilde Q_\theta\big(\dot\ell^L_{\theta^*}(\dot\ell^L_{\theta^*})^T\big)-\tilde Q_\theta\dot\ell^L_{\theta^*}\tilde Q_\theta(\dot\ell^L_{\theta^*})^T\Big)(\theta-\theta^*).$$

Using that $P_0\dot\ell_{\theta^*}=0$ (see (5.13)), the above can be estimated like before, with the result that there exists a constant $s(\delta)$ (depending only on (the largest eigenvalue of) the matrix $P_0\dot\ell_{\theta^*}\dot\ell^T_{\theta^*}$ and with the property that $s(\delta)\to 1$ as $\delta\to 0$) such that:

$$\mathrm{Var}_{\tilde Q_\theta}(W_i)\le s(\delta)\Delta(\theta),$$

for small enough ε and large enough L. We apply Bernstein's inequality to obtain:

$$Q_\theta^n(1-\omega_{1,n}) = \|Q_\theta\|^n \tilde{Q}_\theta^n\big(W_1+\ldots+W_n \geq nr(\delta)\Delta(\theta)\big)$$
$$\leq \|Q_\theta\|^n \exp\Big(-\frac{1}{2}\frac{r(\delta)^2 n\Delta(\theta)}{s(\delta)+\frac{3}{2}L\sqrt{d}\|\theta-\theta^*\|r(\delta)}\Big). \tag{5.26}$$

The factor $t(\delta) = r(\delta)^2(s(\delta)+\frac{3}{2}L\sqrt{d}\|\theta-\theta^*\|r(\delta))^{-1}$ lies arbitrarily close to 1 for sufficiently small choices of δ and ε. As for the n-th power of the norm of Q_θ, we use lemma 5.3.6, (5.10) and (5.11) to estimate the norm of Q_θ as follows:

$$\begin{aligned}\|Q_\theta\| &= 1+P_0\log\frac{p_\theta}{p_{\theta^*}}+\tfrac{1}{2}P_0\Big(\log\frac{p_\theta}{p_{\theta^*}}\Big)^2+o(\|\theta-\theta^*\|^2)\\ &\leq 1+P_0\log\frac{p_\theta}{p_{\theta^*}}+\tfrac{1}{2}(\theta-\theta^*)^T P_0\big(\dot{\ell}_{\theta^*}\dot{\ell}_{\theta^*}^T\big)(\theta-\theta^*)+o(\|\theta-\theta^*\|^2)\\ &\leq 1-\tfrac{1}{2}(\theta-\theta^*)^T V_{\theta^*}(\theta-\theta^*)+\tfrac{1}{2}u(\delta)\Delta(\theta),\end{aligned} \tag{5.27}$$

for some constant $u(\delta)$ such that $u(\delta)\to 1$ if $\delta\to 0$. Because $1+x\leq e^x$ for all $x\in\mathbb{R}$, we obtain, for sufficiently small $\|\theta-\theta^*\|$:

$$Q_\theta^n(1-\omega_{1,n}) \leq \exp\Big(-\frac{n}{2}(\theta-\theta^*)^T V_{\theta^*}(\theta-\theta^*)+\frac{n}{2}\big(u(\delta)-t(\delta)\big)\Delta(\theta)\Big). \tag{5.28}$$

Note that $u(\delta)-t(\delta)\to 0$ as $\delta\to 0$ and $\Delta(\theta)$ is upper bounded by a multiple of $\|\theta-\theta^*\|^2$. Since V_{θ^*} is assumed to be invertible, we conclude that there exists a constant $C>0$ such that for large enough L, small enough $\varepsilon>0$ and large enough n:

$$Q_\theta^n(1-\omega_{1,n}) \leq e^{-Cn\|\theta-\theta^*\|^2}. \tag{5.29}$$

Concerning the range $\|\theta-\theta^*\|>\varepsilon$, an asymptotically consistent uniform test-sequence of P_0 versus Q_θ exists by assumption, and it is shown in chapter 9 that this implies the existence of tests $\omega_{2,n}$ of uniformly *exponential* testing power. The sequence (ψ_n) is defined as the maximum of the two sequences defined above: $\psi_n = \omega_{1,n}\vee\omega_{2,n}$ for all $n\geq 1$, in which case $P_0^n\psi_n \leq P_0^n\omega_{1,n}+P_0^n\omega_{2,n}\to 0$ and:

$$\begin{aligned}\sup_{\theta\in A_n} Q_\theta^n(1-\psi_n) &= \sup_{\theta\in\Theta_1} Q_\theta^n(1-\psi_n)\vee\sup_{\theta\in\Theta_2} Q_\theta^n(1-\psi_n)\\ &\leq \sup_{\theta\in\Theta_1} Q_\theta^n(1-\omega_{1,n})\vee\sup_{\theta\in\Theta_2} Q_\theta^n(1-\omega_{2,n}).\end{aligned}$$

A suitable choice for the constant $D>0$ lead to (5.25).

The following lemma is used in the proof of theorem 5.3.5 to control the behaviour of $\|Q_\theta\|$ in neighbourhoods of θ^*.

Lemma 5.3.6. *Assume that $P_0(p_\theta/p_{\theta^*})$ and $-P_0\log(p_\theta/p_0)$ are finite for all θ in a neighbourhood U' of θ^*. Furthermore, assume that there exist a measurable function m such that,*

$$\Big|\log\frac{p_\theta}{p_{\theta^*}}\Big| \leq m\|\theta-\theta^*\|, \quad (P_0-a.s.). \tag{5.30}$$

for all $\theta \in U'$ and such that $P_0(e^{sm}) < \infty$ for some $s > 0$. Then,

$$P_0\Big|\frac{p_\theta}{p_{\theta^*}} - 1 - \log\frac{p_\theta}{p_{\theta^*}} - \frac{1}{2}\Big(\log\frac{p_\theta}{p_{\theta^*}}\Big)^2\Big| = o\big(\|\theta-\theta^*\|^2\big).$$

Proof. The function $R(x)$ defined by $e^x = 1 + x + \frac{1}{2}x^2 + x^2 R(x)$ increases from $-\frac{1}{2}$ in the limit $(x \to -\infty)$ to ∞ as $(x \to \infty)$, with $R(x) \to R(0) = 0$ if $(x \to 0)$. We also have $|R(-x)| \le R(x) \le e^x/x^2$ for all $x > 0$. The *l.h.s.* of the assertion of the lemma can be written as,

$$P_0\Big(\log\frac{p_\theta}{p_{\theta^*}}\Big)^2\Big|R\Big(\log\frac{p_\theta}{p_{\theta^*}}\Big)\Big| \le \|\theta-\theta^*\|^2 P_0\big(m^2 R(m\|\theta-\theta^*\|)\big).$$

The expectation on the *r.h.s.* of the above display is bounded by $P_0 m_\theta^2 R(\varepsilon m_\theta)$ if $\|\theta-\theta^*\| \le \varepsilon$. The functions $m^2 R(\varepsilon m)$ are dominated by e^{sm} for sufficiently small ε and converge pointwise to $m^2 R(0) = 0$ as $\varepsilon \downarrow 0$. The lemma then follows from the *dominated convergence theorem*, theorem B.3.8.

5.4 Model selection with the BIC criterion [EMPTY]

5.5 Exercises [EMPTY]

Part II
Non-parametric Bayesian statistics

Chapter 6
Asymptotic posterior concentration

Although the subject matter of part I is almost exclusively parametric, the presentation is such that most of the definitions (and some of their consequences) generalize to a setting where the parameter belongs to an infinite-dimensional space. In such cases both prior and posterior measures are probability measures on a space Θ that lacks a lot of structure we take for granted in subsets of $\mathbb{R}^k$. Two differences play a central role. Firstly, no (locally) finite translation-invariant measures exist on infinite-dimensional spaces, so there is *no analogue of Lebesgue measure*, and consequently, no canonical way of thinking about density functions for priors and posteriors for infinite-dimensional parameters. Secondly, the *topology* on Θ is not fixed by default, there is no 'natural choice' like that of the unique norm topology for subspaces of $\mathbb{R}^k$. Topology on infinite-dimensional spaces is far more diverse and far more influential than in finite-dimensional setting. In many ways, having to choose a topology is an advantage because we can choose a topology that suits our statistical purposes the most naturally. For example, to formulate necessary and sufficient conditions for consistent estimation, the Le Cam-Schwartz theorem (see theorem 9.1.1) focusses on the choice of a weak topology stronger than Prokhorov's but weaker than total-variation. But technically, variations in the refinement level of the model topology complicate matters, especially if we choose to equip the model with a strong topology (*e.g.* Hellinger/total-variation) and require the posterior to concentrate accordingly (as we do in this chapter).

There are two direct ways in which the topology on Θ plays a role for Bayesian procedures: firstly, the topology determines the Borel σ-algebra that usually defines the domain for prior and posterior measures on infinite-dimensional spaces. Secondly, we shall be interested in *asymptotic posterior concentration*: as argued in subsection 4.1.1, statistical procedure are expected to become more-and-more precise as the amount of data grows, and ideally, one would like full precision in the limit where that amount goes to infinity. In the case of a posterior on a well-specified model with parameter space Θ, precision means what we expect to find posterior mass concentrating 'around' the true $\theta_0 \in \Theta$. The topology fixes what is meant by the word 'around' and is of great influence regarding the strength of the conclusion that the posterior concentrates around θ_0 asymptotically.

The theorems of this chapter concern *i.i.d.* data in the form of samples $X^n = (X_1, X_2, \ldots, X_n)$ drawn from P_0^n and a model $\mathscr{P}$ of single-observation distributions. We consider only *metric models* $(\mathscr{P}, d)$, almost exclusively with the metric d equal to the Hellinger metric or a metric that is topologically equivalent. The main goal is to establish three classical results regarding posterior concentration, namely Doob's theorem, Schwartz's theorem and the Ghosh-Ghosal-van der Vaart theorem. More complex dependence structures for the sample and more varied model topologies are discussed in chapter 7.

6.1 Posterior concentration and model topology

Consistency is incontestable as an asymptotic criterion from the frequentist point of view, but is not free of controversy in Bayesian statistics. Specifically, the subjectivist Bayesian point of view does not attach value to any special point of convergence P_0 because no 'underlying' or 'true' distribution for the sample $X_1, X_2, \ldots$ is assumed within the subjectivist paradigm. The notion of 'merging' is perhaps closer to the subjectivist's philosophy: given two different priors Π_1 and Π_2 on a model Θ, merging is said to occur if the total-variation distance between the posterior predictive distributions goes to zero (see Blackwell and Dubins (1962) [37] and, for an overview, Ghosh and Ramamoorthi (2003) [111]). Relations between merging and posterior consistency as defined below are discussed in Diaconis and Freedman (1986) [71].

6.1.1 Posterior consistency

We start by defining what frequentist consistency means for a sequence of posterior distributions. We consider sequentially observed (possibly non-*i.i.d.*) data, non-dominated models and priors or parameter spaces that may depend on the sample size (see remark A.0.1 for general conventions).

Definition 6.1.1. The posteriors $\Pi(\cdot|X^n)$ are *consistent at* $\theta \in \Theta$ if for every neighbourhood U of θ,

$$\Pi(U|X^n) \xrightarrow{P_{\theta,n}} 1. \tag{6.1}$$

The posteriors are said to be *consistent* if this holds for all $\theta \in \Theta$. A weaker form of posterior convergence is *Bayesian consistency*, when (6.1) holds for Π-almost-all $\theta \in \Theta$. We say that the posterior is *almost-surely consistent* if convergence occurs almost-surely with respect to some coupling for the sequence $(P_{\theta_0,n})$.

For example, in the common case of a metric model $(\mathscr{P}, d)$ of single-observation distributions for *i.i.d.* data $X_1, \ldots, X_n \sim P_0^n$, consistency is equivalent to the condition that for every $\varepsilon > 0$:

$$\Pi(d(P,P_0) \geq \varepsilon \,|\, X_1, X_2, \ldots, X_n) \xrightarrow{P_0\text{-a.s.}} 0, \tag{6.2}$$

since the above display is the complement of an open ball and every open neighbourhood of P_0 contains an open ball centred on P_0.

Proposition 6.1.2. *Assume that Θ is a completely regular space. The posterior is consistent at $\theta_0 \in \Theta$, if and only if,*

$$\int f(\theta)\, d\Pi(\theta|X^n) \xrightarrow{P_{\theta_0,n}} f(\theta_0), \tag{6.3}$$

for every bounded, continuous $f : \Theta \to \mathbb{R}$.

Proof. Assume (6.1). Let $f : \Theta \to \mathbb{R}$ be bounded and continuous (with $M > 0$ such that $|f| \leq M$). Let $\eta > 0$ be given and let $U \subset \Theta$ be a neighbourhood of θ_0 such that $|f(\theta) - f(\theta_0)| < \eta$ for all $\theta \in U$. Integrate f with respect to the (assumed to be regular and $P_{\theta_0,n}$-almost-surely well-defined) posterior and to δ_{θ_0}:

$$\begin{aligned}
&\left| \int f(\theta)\, d\Pi(\theta|X^n) - f(\theta_0) \right| \\
&\quad \leq \int_{\Theta \setminus U} |f(\theta) - f(\theta_0)|\, d\Pi(\theta|X^n) + \int_U |f(\theta) - f(\theta_0)|\, d\Pi(\theta|X^n) \\
&\quad \leq 2M\, \Pi(\Theta \setminus U|X^n) \quad + \sup_{\theta \in U} |f(\theta) - f(\theta_0)|\, \Pi(U|X^n) \leq \eta + o_{P_{\theta_0,n}}(1),
\end{aligned}$$

as $n \to \infty$, so that (6.3) holds. Conversely, assume (6.3). Let U be an open neighbourhood of θ_0. Because Θ is completely regular (see definition C.2.3), there exists a continuous $f : \Theta \to [0,1]$ such that $f = 1$ at $\{\theta_0\}$ and $f = 0$ on $\Theta \setminus U$. Then,

$$\Pi(U|X^n) \geq \int f(\theta)\, d\Pi(\theta|X^n) \xrightarrow{P_{\theta_0,n}} \int f(\theta)\, d\delta_{\theta_0}(P) = 1.$$

Consequently, (6.1) holds.

Metrisable spaces are uniform spaces; a topological vector space is uniform, if and only if, it is completely regular and subspaces of completely regular spaces are completely regular. So the above implies the following corollary immediately, in the common case of a metric model $(\mathscr{P}, d)$ for *i.i.d.* data.

Corollary 6.1.3. *On metric models $(\mathscr{P}, d)$, (6.1), (6.2) and (6.3) are equivalent.*

As becomes clear in chapter 9 the most convenient choice here is not the canonical one: both Prokhorov's weak and the total-variational/Hellinger topologies are metric and attractive intuitively, but the natural model topology for the study of frequentist consistency in *i.i.d.* setting (see the *Le Cam-Schwartz theorem*, theorem 9.1.1) is a uniform topology $\mathscr{T}_\infty$ *stronger than Prokhorov's* and *weaker than total-variation*. Without restrictions on the model $\mathscr{P}$, the topology $\mathscr{T}_\infty$ is non-metrizable (note that it is not even first-countable, in general).

6.1.2 Consistency of Bayesian point-estimators

Point-estimators derived from a consistent Bayesian procedure are consistent themselves under some mild conditions. We reiterate that the notion of a point-estimator is not an entirely natural extension to the Bayesian framework: for example, if the model is non-convex (and hardly any model is), the posterior predictive distribution of definition 2.1.4 lies outside the model. Similarly, perfectly well-defined posteriors may lead to ill-defined point-estimators due to integrability issues or non-existence of maximisers, which become more severe as the model becomes more complicated.

Here, we endow a single-observation model $\mathscr{P}$ (again, with data $X_1, X_2, \ldots$ taking values in a measurable space $(\mathscr{X}, \mathscr{B})$ that are distributed *i.i.d.*-P_0) with the total-variational topology and corresponding Borel σ-algebra.

Theorem 6.1.4. *Assume that the Borel prior Π and underlying distribution $P_0 \in \mathscr{P}$ are such that the sequence of posteriors is consistent (P_0-almost-surely). Then the posterior predictive distributions $\hat{P}_n$ are (P_0-almost-surely) consistent point-estimators for P_0 with respect to total-variation.*

Proof. Note that the domain of definition of the map $P \mapsto \|P - P_0\|$ extends to the convex hull $\mathrm{co}(\mathscr{P})$ of $\mathscr{P}$ in $\mathscr{M}^1(\mathscr{X}, \mathscr{B})$. Since $P \mapsto \|P - P_0\|$ is convex by virtue of the triangle inequality, Jensen's inequality (see, *e.g.* theorem 10.2.6 in Dudley (1989) [81]) says that the posterior mean $\hat{P}_n$ satisfies:

$$\|\hat{P}_n - P_0\| = \left\| \int_{\mathscr{P}} P \, d\Pi(P \,|\, X_1, \ldots, X_n) - P_0 \right\| \leq \int_{\mathscr{P}} \|P - P_0\| \, d\Pi(P \,|\, X_1, \ldots, X_n).$$

Since the posteriors $\Pi(\cdot \,|\, X_1, \ldots, X_n)$ converge weakly to P_0 (P_0-almost-surely) and the map $P \mapsto \|P - P_0\|$ is bounded and continuous in the total-variational topology, we conclude that the *r.h.s.* in the above display converges to the expectation of $\|P - P_0\|$ under the limit law δ_{P_0} (P_0-almost-surely), which equals zero. Hence $\hat{P}_n$ converges to P_0 in total variation (P_0-almost-surely).

More generally, given an arbitrary convex metric d on the model $\mathscr{P}$, theorem 6.1.4 can be proved if the metric d is convex and bounded on $\mathscr{P}$. Similar arguments demonstrate consistency for other classes of point estimators derived from a consistent sequence of posterior distributions, for example the *formal Bayes estimators* of subsection 2.2.3.

6.2 Bayesian consistency and Doob's theorem

In this section, we concentrate on a sufficient condition for *Bayesian consistency*, a form of posterior consistency that holds for all $P_0 \in \mathscr{P}$ except (perhaps) in a model subset that is a *null-set of the prior*. The first and perhaps most famous consistency theorem in Bayesian statistics is that given by Doob (1949) [80].

Theorem 6.2.1. *(Doob's theorem)*
For all $n \geq 1$, let $(X_1, X_2, \ldots, X_n) \in \mathscr{X}^n$ be i.i.d., with a single-observation model $\mathscr{P}$. Suppose $\mathscr{X}$ and $\mathscr{P}$ are Polish spaces and that $P \mapsto P(A)$ is Borel measurable for every Borel set $A \subset \mathscr{X}$. Then for any Borel prior Π the posterior is consistent at P, P-almost-surely, for Π-almost-all P.

Proof. The proof of this theorem is an application of Doob's martingale convergence theorem. We prove (a more general version of) this theorem in chapter 9.

The measurability condition is relatively minor, a bound on the level of refinement of the model topology (which is satisfied by Prokhorov's weak model topology already). For the sample space $\mathscr{X}$, the requirement of Polishness is not too stringent, so let us focus on the requirement that $\mathscr{P}$ is Polish. First of all, for many models and parameter spaces, Polishness is easily achieved (*e.g.* as open or closed (indeed, G_δ) subsets of larger Polish spaces, like $\mathbb{R}^k$, Banach spaces or M_+^b sets). More generally, one notices that, although many statistical models are not *complete* metric spaces, we may argue that we can replace $\mathscr{P}$ by its *completion* $\hat{\mathscr{P}}$ as long as we define a prior $\hat{\Pi}$ on the Borel σ-algebra of the completion as $\hat{\Pi}(B) = \Pi(B \cap \mathscr{P})$, making the difference $\hat{\mathscr{P}} \setminus \mathscr{P}$ a null-set of $\hat{\Pi}$. In order to argue like this, one has to prove that $\mathscr{P}$ is a Borel measurable subset of $\hat{\mathscr{P}}$.

Separability is a different matter. To relate to parametrizing spaces immediately (indeed, theorem 6.2.1 is also true if $\mathscr{P}$ is a *Souslin space*, see Le Cam (1986) [179]), note that function spaces like $L^1(\mu)$- and $L^\infty(\mu)$-spaces for the Lebesgue measure on $\mathbb{R}$, for example, are closely related but the former is separable while the latter is not [82]. Similarly, smoothness classes display diversity: while *Sobolev spaces* are separable, the closely related *Hölder spaces* are not. Regarding the Hellinger/total-variational topology on spaces of single-observation distributions, it is noted that, according to proposition 8.9.1, $\mathscr{P}$ is separable, if and only if, $\mathscr{P}$ is dominated and $\mathscr{B}$ is generated countably.

The above delineates the realm of applicability of Doob's theorem more concretely than the characterization of the model as a Polish space: in practical terms, we require a metric d on $\mathscr{P}$ that is strong enough to guarantee measurability of $P \to P(A)$, while not so strong as to ruin separability. One of the natural formulations is in terms of families of densities, with the Hellinger or total-variational metric, for data that takes its values in a Polish space.

Corollary 6.2.2. *For all $n \geq 1$, let $\mathscr{X}$ be Polish and let $(X_1, X_2, \ldots, X_n) \in \mathscr{X}^n$ be i.i.d., with a single-observation model $\mathscr{P}$ that is dominated and endowed with the Hellinger/total-variational topology. Assume that $\mathscr{P}$ is complete, or that $\mathscr{P}$ is a Borel measurable subset of $\hat{\mathscr{P}}$. Then for any Borel prior Π, the posterior is consistent at P, P-almost-surely, for Π-almost-all P.*

Proof. All functions $P \mapsto P(A)$ are continuous with respect to the Hellinger/total-variational topology, so they are Borel measurable as functions on $\mathscr{P}$. Then apply theorem 6.2.1 to the Borel measure $\hat{\Pi}$.

The other natural formulation is with the full model $M^1(\mathscr{X})$ and Prokhorov's weak topology, for data that takes values in a Polish space. It is noted that in chapter 8, priors of full support with respect to Prokhorov's weak topology are constructed.

Corollary 6.2.3. *For all $n \geq 1$, let $\mathscr{X}$ be Polish and let $(X_1, X_2, \ldots, X_n) \in \mathscr{X}^n$ be i.i.d., with the full single-observation model $M^1(\mathscr{X})$, endowed with Prokhorov's weak topology. Then for any Borel prior Π, the posterior is consistent at P, P-almost-surely, for Π-almost-all P.*

Proof. All functions $P \mapsto P(A)$ are measurable with respect to the Borel σ-algebra associated with Prokhorov's weak topology (see exercise 6.6.1). According to theorem C.8.9, $M_+^b(\mathscr{X})$ is a Polish space; the space $M^1(\mathscr{X})$ is a closed subset and hence Polish as well. Then apply theorem 6.2.1.

For (most) Bayesians Doob's theorem is more than enough: *c.f.* the last remarks before example 2.1.18, for the Bayesian 'the model' is defined only up to prior null-sets. To illustrate this point intuitively, we consider it first from the parametric perspective: for an open $\Theta \subset \mathbb{R}^k$ with continuous $\Theta \to \mathscr{P} : \theta \mapsto P_\theta$, and a prior that dominates the Lebesgue measure on Θ, the above theorem leaves room for posterior *inconsistency* only on subsets of Lebesgue measure zero. A popular view is, that consistency theorems like the above show that "the data always overrides prior beliefs asymptotically".

However, this note of optimism relies heavily on finite-dimensional intuition and, more particularly, Lebesgue measure. There is absolutely no implication that analogous expectations are justified in non-parametric context. Indeed, Doob's theorem becomes highly problematic in such models: the theorem stays true exactly as stated, it simply means something else than what finite-dimensional intuition suggests. Strictly speaking, only frequentists recognize consistency problems: Doob's proof says nothing about specific points in the model, *i.e.* given a particular $P_0 \in \mathscr{P}$ underlying the sample, Doob's theorem does not give conditions that can be checked to see whether the Bayesian procedure will be consistent at this particular P_0: it is always possible that P_0 belongs to the null-set for which inconsistency occurs. That such null-sets may be large, is clear from example 2.1.18 and that, indeed, this may lead to grave problems in non-parametric situations, becomes apparent when we consider the counterexamples given by Freedman (1963,1965) [97, 98] and Diaconis and Freedman (1986) [71, 72]. Non-parametric examples of inconsistency in Bayesian regression are found in Cox (1993) [61] and Diaconis and Freedman (1998) [74]. Basically what is shown is that the null-set on which inconsistency occurs in Doob's theorem can be rather large in non-parametric situations. Some authors are tempted to present the above as definitive proof of the fact that Bayesian statistics are unfit for non-parametric estimation problems. More precise is the statement that not every choice of prior is suitable, raising the question that will entertain us for the rest of this chapter and next: under which conditions on model and prior, can we expect frequentist forms of consistency to hold? We come back to Freedman's counterexamples in subsection 6.5.1.

6.3 Schwartz's posterior consistency theorem

Fortunately a theorem exists that provides sufficient conditions for consistency at a *specific* point $P_0 \in \mathscr{P}$. Requiring these conditions to hold for *every* $P_0 \in \mathscr{P}$, makes Schwartz's theorem below a *frequentist* consistency guarantee for the posterior that is valid in both in parametric and in non-parametric setting.

Theorem 6.3.1. *(Schwartz (1965))*
For all $n \geq 1$, let $(X_1, X_2, \ldots, X_n) \in \mathscr{X}^n$ be i.i.d. $-P_0$, where P_0 lies in a dominated model $\mathscr{P}$. Let U denote an open neighbourhood of P_0 in $\mathscr{P}$. If,

(i) there exist measurable $\phi_n : \mathscr{X}^n \to [0,1]$, such that,

$$P_0^n \phi_n = o(1), \quad \sup\{Q^n(1-\phi_n) : Q \in \mathscr{P} \setminus U\} = o(1), \tag{6.4}$$

(ii) and Π is a Kullback-Leibler (KL-)prior, i.e. for all $\delta > 0$,

$$\Pi\Big(P \in \mathscr{P} : -P_0 \log \frac{p}{p_0}(X) < \delta\Big) > 0, \tag{6.5}$$

then $\Pi(U|X^n) \xrightarrow{P_0\text{-a.s.}} 1$.

The condition of domination in the above theorem is strictly speaking redundant, the theorem is true without it. (In subsection 6.3.1, replace p/p_0 by the Radon-Nikodym derivative dP/dP_0 throughout and change the third equality in (6.8) into less-or-equal.) However, the Kullback-Leibler divergence that plays a role in the second condition, is very sensitive to mismatches in the supports of model distributions: if $P_0(p(X) = 0) > 0$ or vice versa, the logarithm diverges. So even though the theorem holds in non-dominated models in principle, in practice mismatch of null-sets interferes with the formulation of the prior condition. This fact forms the basis for one of the counterexamples of subsection 6.5.2. Historical counterexamples (see Diaconis and Freedman [71], for example) also fail the *lower bound for prior mass* in Kullback-Leibler neighbourhoods of the true distribution (see also Barron *et al.* (1999) [16]). In chapter 7, condition 6.5 is generalized to a weakened form of *contiguity*, relieving Schwartz's theorem of its sensitivity to problems of this type.

Comparing condition (6.4) with complete regularity (definition C.2.3), one notices conceptually similar roles for separating functions and test functions: the sequence of test functions in (6.4) "separates" the singleton $\{P_0\}$ from the alternative, as a stochastic, uniform limit. Requiring existence of such test sequences, is the condition that there must be some statistical procedure that allows us to "separate" the true distribution of the data from the complement of any open neighbourhood. This central testing condition of Schwartz's theorem is related directly to posterior asymptotic behaviour in chapter 7. It is also noted that the choice of the topology on $\mathscr{P}$ can be adapted to the existence question for tests, given the model. The latter possibility is considered explicitly in chapter 9.

6.3.1 Proof of Schwartz's theorem

Proof. Define V to be the complement of the open U around P_0 in $\mathscr{P}$. We start by splitting the n-th posterior measure of V with the test function ϕ_n and taking the limes superior:

$$\begin{aligned}\limsup_{n\to\infty} &\Pi\big(V|X_1,\ldots,X_n\big)\\ &\leq \limsup_{n\to\infty} \Pi\big(V|X_1,\ldots,X_n\big)(1-\phi_n) + \limsup_{n\to\infty} \Pi\big(V|X_1,\ldots,X_n\big)\phi_n.\end{aligned} \tag{6.6}$$

For given $\eta > 0$ (to be fixed at a later stage) we consider the subset $K_\eta = \{P \in \mathscr{P} : -P_0 \log(p/p_0) \leq \eta\}$. For every $P \in K_\eta$, the *law of large numbers* says that:

$$\left| \mathbb{P}_n \log \frac{p}{p_0} - P_0 \log \frac{p}{p_0} \right| \to 0, \quad (P_0 - a.s.).$$

Hence for every $\alpha > \eta$ and all $P \in K_\eta$, there exists an $N \geq 1$ such that for all $n \geq N$, $\prod_{i=1}^n (p/p_0)(X_i) \geq e^{-n\alpha}$, P_0^n-almost-surely. This can be used to lower-bound the denominator in the expression for the posterior P_0^n-almost-surely as follows:

$$\begin{aligned}\liminf_{n\to\infty} e^{n\alpha} \int_{\mathscr{P}} \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi(P) &\geq \liminf_{n\to\infty} e^{n\alpha} \int_{K_\eta} \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi(P)\\ &\geq \int_{K_\eta} \liminf_{n\to\infty} e^{n\alpha} \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi(P) \geq \Pi(K_\eta),\end{aligned}$$

where we use *Fatou's lemma* (see lemma B.3.7) to obtain the second inequality. Since by assumption, $\Pi(K_\eta) > 0$ we see that the first term on the *r.h.s.* of (6.6) can be estimated as follows:

$$\begin{aligned}&\limsup_{n\to\infty} \Pi(V|X_1,\ldots,X_n)(1-\phi_n)(X_1,\ldots,X_n)\\ &\quad= \limsup_{n\to\infty} \frac{\displaystyle\int_V \prod_{i=1}^n \frac{p}{p_0}(X_i)\,(1-\phi_n)(X_1,\ldots,X_n)\, d\Pi(P)}{\displaystyle\int_{\mathscr{P}} \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi(P)}\\ &\quad\leq \frac{\limsup_{n\to\infty} e^{n\alpha} \displaystyle\int_V \prod_{i=1}^n (p/p_0)(X_i)\,(1-\phi_n)(X_1,\ldots,X_n)\, d\Pi(P)}{\liminf_{n\to\infty} e^{n\alpha} \displaystyle\int_{\mathscr{P}} \prod_{i=1}^n (p/p_0)(X_i)\, d\Pi(P)}\\ &\quad\leq \frac{1}{\Pi(K_\eta)} \limsup_{n\to\infty} f_n(X_1,\ldots,X_n),\end{aligned} \tag{6.7}$$

where we use P_0^∞-almost-surely defined $f_n : \mathscr{X}^n \to [0,\infty]$,

$$f_n(X_1,\ldots,X_n) = e^{n\alpha}\int_V \prod_{i=1}^n \frac{p}{p_0}(X_i)\,(1-\phi_n)(X_1,\ldots,X_n)\,d\Pi(P).$$

Fubini's theorem and the fact that the test-sequence can be assumed to be uniformly exponential (see lemma 6.3.2) guarantee that there exists a constant $\beta > 0$ such that for large enough n,

$$\begin{aligned} P_0^\infty f_n = P_0^n f_n &= e^{n\alpha}\int_V P_0^n\Big(\prod_{i=1}^n \frac{p}{p_0}(X_i)\,(1-\phi_n)(X_1,\ldots,X_n)\Big)\,d\Pi(P) \\ &\leq e^{n\alpha}\int_V P^n(1-\phi_n)\,d\Pi(P) \leq e^{-n(\beta-\alpha)}. \end{aligned} \tag{6.8}$$

We choose η strictly below β and can then choose α such that $\eta < \alpha < 1/2(\beta+\eta)$. Markov's inequality can be used to show that:

$$P_0^\infty\big(f_n > e^{-\frac{n}{2}(\beta-\eta)}\big) \leq e^{n(\alpha-\frac{1}{2}(\beta+\eta))}.$$

Hence the series $\sum_{n=1}^\infty P_0^\infty(f_n > \exp -\frac{n}{2}(\beta-\eta))$ converges and the *first Borel-Cantelli lemma* (lemma B.2.11) then leads to the conclusion that:

$$0 = P_0^\infty\Big(\bigcap_{N=1}^\infty \bigcup_{n\geq N}\big\{f_n > e^{-\frac{n}{2}(\beta-\eta)}\big\}\Big) \geq P_0^\infty\Big(\limsup_{n\to\infty}\big(f_n - e^{-\frac{n}{2}(\beta-\eta)}\big) > 0\Big)$$

Since $f_n \geq 0$, we see that $f_n \to 0$, $(P_0 - a.s.)$, which we substitute in (6.7).

We estimate the last term on the *r.h.s.* of (6.6) with an argument similar to that used above for the functions f_n. Note that $P_0^n\Pi(V|X_1,\ldots,X_n)\phi_n \leq P_0^n\phi_n \leq e^{-nC}$ for some positive constant C, according to lemma 6.3.2. Markov's inequality and the first Borel-Cantelli lemma suffice to show that:

$$\phi_n\Pi\big(V|X_1,\ldots,X_n\big) \xrightarrow{P_0\text{-a.s.}} 0. \tag{6.9}$$

Combination of (6.7) and (6.9) proves that (6.6) equals zero.

The following lemma says that the existence of uniformly consistent test sequences implies the existence of uniformly consistent test sequences *at exponential rate*.

Lemma 6.3.2. *Suppose that for some $P_0 \in \mathscr{P}$, $V \subset \mathscr{P}$ there exists a sequence of test functions (ϕ_n) such that:*

$$P_0^n\phi_n \to 0, \quad \sup_{Q\in V} Q^n(1-\phi_n) \to 0.$$

Then there exists a sequence of tests (ω_n) and positive constants C, D such that:

$$P_0^n\omega_n \leq e^{-nC}, \quad \sup_{Q\in V} Q^n(1-\omega_n) \leq e^{-nD} \tag{6.10}$$

Proof. The proof is an application of Hoeffding's inequality, see proposition 9.3.1.

6.4 Posterior convergence at a rate

Recalling the formulation of posterior consistency given in (6.2), we define the rate of convergence for a consistent sequence of posteriors as the fastest rate ε_n with which we can let balls $B_d(P_0, \varepsilon_n)$ shrink to radius zero, while still capturing posterior masses that converges to one in the limit $n \to \infty$. We formalise this as follows.

Definition 6.4.1. Let $\mathscr{P}$ be a model with metric d and Borel prior Π. Assume that $X_1, X_2, \ldots$ are *i.i.d.*-P_0, for some $P_0 \in \mathscr{P}$. Let the sequence ε_n be such that $\varepsilon_n > 0$ and $\varepsilon_n \downarrow 0$. We say that the sequence of posterior measures $\Pi(\cdot \,|\, X_1, X_2, \ldots, X_n)$ *converges (at least) at rate* ε_n at P_0, if for all sequences $M_n \to \infty$:

$$\Pi\big(\, d(P, P_0) \geq M_n \varepsilon_n \,|\, X_1, X_2, \ldots, X_n \,\big) \xrightarrow{P_0} 0, \tag{6.11}$$

To demonstrate how this definition relates to the rate of convergence for derived point-estimators like the posterior predictive distribution, assume that the sequence of posteriors satisfies (6.11). With the sequence ε_n, we define moreover the point estimators $\tilde{P}_n$ as (near-)maximisers in the model of the maps:

$$P \mapsto \Pi\big(\, B(P, \varepsilon_n) \,|\, X_1, \ldots, X_n \,\big),$$

where $B(P, \varepsilon) \subset \mathscr{P}$ is the *d-ball* of radius ε around P in the model.

Proposition 6.4.2. *Assuming (6.11), for every sequence $M_n \to \infty$ the estimator sequence $\tilde{P}_n$ satisfies,*

$$P_0^n\big(\, d(\tilde{P}_n, P_0) \leq 2 M_n \varepsilon_n \,\big) \to 1. \tag{6.12}$$

As a result $\tilde{P}_n$ converges to P_0 with respect to d (at least) at rate ε_n.

Proof. Let $\tilde{P}_n$ like above be given. By definition of a near-maximiser:

$$\begin{aligned} \Pi\big(\, B(\tilde{P}_n, M_n \varepsilon_n) \,|\, X_1, \ldots, X_n \,\big) &\geq \sup_{P \in \mathscr{P}} \Pi\big(\, B(P, M_n \varepsilon_n) \,|\, X_1, \ldots, X_n \,\big) - o_{P_0}(1) \\ &\geq \Pi\big(\, B(P_0, M_n \varepsilon_n) \,|\, X_1, \ldots, X_n \,\big) - o_{P_0}(1). \end{aligned}$$

Because the first term on the *r.h.s.* of the above display converges to one (according to (6.11)) and the second to zero in P_0-probability, the *l.h.s.* converges to one in P_0-probability. Since $B(\tilde{P}_n, M_n \varepsilon_n) \cap B(P_0, M_n \varepsilon_n) = \varnothing$ if $d(\tilde{P}_n, P_0) > 2 M_n \varepsilon_n$, the fact that the total posterior mass of the model does not exceed one guarantees that $d(\tilde{P}_n, P_0) \leq 2 M_n \varepsilon_n$ with P_0-probability growing to one as $n \to \infty$, demonstrating that ε_n bounds the rate of convergence.

A proof that does not differ in an essential way from the above can be given for the centre point of the d-ball of minimal radius containing posterior mass $p > 1/2$ (see exercise 6.6.2). Note that, for any $M_n \to \infty$, balls of radii $2 M_n \varepsilon_n$ centred on $\tilde{P}_n$ are *asymptotically consistent confidence balls*. (Compare the above proof with corollary 2.3.13 and with theorem 2.3.14.)

The possibility to construct point estimator sequences from posterior distributions converging at the same rate (*e.g.* $\tilde{P}_n$ above), implies that limitations on the rate of convergence (arising in particular in non-parametric estimation problems, see (6.21) below, for example) derived for point estimation, apply unabated to Bayesian rates. This argument applies to other asymptotic performance criteria as well.

6.4.1 The Ghosal-Ghosh-van der Vaart theorem

With regard to sufficient conditions for the defining property (6.11) of posterior rates of convergence, we note Le Cam (1973) [177] and Ibragimov and Has'minskii (1981) [131], who prove that under regularity conditions, posteriors on parametric models achieve $\sqrt{n}$-rate of convergence (mostly along the lines of the proof of theorem 5.3.1). Le Cam (1986) [179] considers rates of convergence of formal Bayes estimators, based on unpublished work using what is now know as Le Cam's inequality (see Le Cam (197X) [178] and subsection 7.1.3). Historically, the two main references dealing with Bayesian rates of convergence in non-parametric models are Ghosal, Ghosh and Van der Vaart (2000) [106] and Shen and Wasserman (2001) [231], and many examples have been collected in Ghosal and van der Vaart (2017) [110]. We postpone further discussion of the literature to the introduction of section 7.5.

Again, we assume a (non-parametric) model $\mathscr{P}$ with metric d and prior Π. To formulate the main theorem of this subsection we define, for every $\varepsilon > 0$,

$$K(P_0,\varepsilon) = \Big\{ P \in \mathscr{P} : -P_0 \log \frac{p}{p_0} \leq \varepsilon^2, P_0 \Big(\log \frac{p}{p_0} \Big)^2 \leq \varepsilon^2 \Big\}. \tag{6.13}$$

This allows us to formulate a more specific version of Schwartz's Kullback-Leibler condition (6.5), in the form of (6.14).

Theorem 6.4.3. *Suppose that for a sequence ε_n with $\varepsilon_n > 0$, $\varepsilon_n \downarrow 0$ and $n\varepsilon_n^2 \to \infty$, some $P_0 \in \mathscr{P}$, and a sequence (Π_n) of priors, the following two conditions hold:*

(i) Π_n is a Ghosal-Ghosh-van der Vaart (GGV-)prior i.e. there exists a constant $C > 0$ such that:

$$\Pi_n\big(K(P_0,\varepsilon_n)\big) \geq e^{-nC\varepsilon_n^2}. \tag{6.14}$$

(ii) There exists a sequence ϕ_n of test-functions ϕ_n and a constant $L > 0$ such that:

$$P_0^n \phi_n \to 0, \qquad \sup_{P : d(P,P_0) \geq \varepsilon_n} P^n(1-\phi_n) \leq e^{-nL\varepsilon_n^2}. \tag{6.15}$$

Then for a sufficiently large $M > 0$,

$$P_0^n \Pi\big(d(P,P_0) \geq M\varepsilon_n \,|\, X_1,\ldots,X_n \big) \to 0. \tag{6.16}$$

(Note that the assertion establishes convergence in P_0-expectation, which implies convergence in P_0-probability because the posterior is bounded.) The rate theorem given here is a variation on theorem 2.1 in Ghosal, Ghosh and Van der Vaart (2000) [106]; their version is different in two respects. First of all they express the testing condition through a sufficient condition based on the model's entropy numbers. We come back to this point in subsection 6.4.3. Secondly, they restrict attention to a sequence of models $\mathscr{P}_n$ that grows in Π_n-measure to the full model $\mathscr{P}$ sufficiently fast,

$$\Pi_n\big(\mathscr{P} \setminus \mathscr{P}_n\big) \leq e^{-nL'\varepsilon_n^2}. \tag{6.17}$$

The submodels $(\mathscr{P}_n)$ are then used to express the entropy condition and referred to as a *sieve* that approximates $\mathscr{P}$ quickly enough with growing n, *c.f.* (6.17). This separation between submodels of controlled entropy and complements of bounded prior mass is due to Barron (1988) [9] and Barron *et al.* (1999) [14].

In subsections 6.4.3 and 6.4.4, we analyse conditions (6.15) and (6.14) separately. First, we prove theorem 6.4.3.

6.4.2 Proof of the Ghosal-Ghosh-van der Vaart theorem

Proof. Define, for every $\eta > 0$, $A(\eta) = \{P \in \mathscr{P} : d(P, P_0) \geq \eta\}$. The expectation in (6.16) can be decomposed using the tests ϕ_n; for every $n \geq 1$ and every $M > 1$, we have:

$$\begin{aligned} &P_0^n \Pi\big(A(M\varepsilon_n) \,\big|\, X_1, \ldots, X_n\big) \\ &\quad = P_0^n\, \phi_n(X)\, \Pi\big(A(M\varepsilon_n) \,\big|\, X_1, \ldots, X_n\big) + P_0^n (1-\phi_n)(X)\, \Pi\big(A(M\varepsilon_n) \,\big|\, X_1, \ldots, X_n\big). \end{aligned}$$

We estimate the terms on the right-hand side separately. Due to the first inequality in (6.15), the first term converges to zero. To estimate the second term, we (assume that $\mathscr{P}$ is dominated, in a non-essential way, see remarkrem:dominationcondition below) and substitute (2.14) to obtain,

$$\begin{aligned} &P_0^n \Pi\big(A(M\varepsilon_n) \,\big|\, X_1, \ldots, X_n\big)(1-\phi_n)(X) \\ &\quad = P_0^n \Big[\int_{A(M\varepsilon_n)} \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi(P)\, (1-\phi_n)(X) \Big/ \int_{\mathscr{P}} \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi(P) \Big] \end{aligned} \tag{6.18}$$

in which the denominator can be lower-bounded by application of lemma 6.4.6, since by assumption (6.14), $\Pi(K(P_0, \varepsilon_n)) > 0$. Let Ω_n be the subset in $\mathscr{X}^n$ for which the inequality between left- and right-hand sides in the following display holds:

$$\int_{\mathscr{P}} \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi(P) \geq \int_{K(P_0,\varepsilon_n)} \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi(P) \geq e^{-(1+L')n\varepsilon_n^2} \Pi(K(P_0, \varepsilon_n)), \tag{6.19}$$

as in (6.26), with $L' > 0$ as yet unspecified. Decomposing the P_0^n-expectation in (6.18) into separate integrals over Ω_n and $\mathscr{X}^n \setminus \Omega_n$, we find:

$$\begin{aligned} &P_0^n \Pi\big(A(M\varepsilon_n) \,\big|\, X_1,\ldots,X_n\big)(1-\phi_n) \\ &\qquad \le P_0^n \Pi\big(A(M\varepsilon_n) \,\big|\, X_1,\ldots,X_n\big)(1-\phi_n) 1_{\Omega_n} + P_0^n(\mathscr{X}^n \setminus \Omega_n). \end{aligned}$$

Note that $P_0^n(\mathscr{X}^n \setminus \Omega_n) = o(1)$ as $n \to \infty$ according to (6.26). The first term is estimated as follows:

$$\begin{aligned} &P_0^n \Pi\big(A(M\varepsilon_n) \,\big|\, X_1,\ldots,X_n\big)(1-\phi_n)\, 1_{\Omega_n} \\ &\quad \le \frac{e^{(1+L')n\varepsilon_n^2}}{\Pi(K(P_0,\varepsilon_n))} P_0^n\Big((1-\phi_n)\int_{A(M\varepsilon_n)} \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi(P)\Big) \\ &\quad \le \frac{e^{(1+L')n\varepsilon_n^2}}{\Pi(K(P_0,\varepsilon_n))} \int_{A(M\varepsilon_n)} P^n\big(1-\phi_n\big)\, d\Pi(P) \\ &\quad \le e^{(1+L')n\varepsilon_n^2} \frac{\Pi(A(M\varepsilon_n))}{\Pi(K(P_0,\varepsilon_n))} \sup_{P\in A(M\varepsilon_n)} P^n\big(1-\phi_n\big), \end{aligned} \tag{6.20}$$

where we have substituted (6.19) and used the positivity of the integrand, applied Fubini's theorem and bounded the integrand by its supremum over $A(M\varepsilon_n)$. Application of the second inequality in (6.15) gives:

$$P_0^n \Pi\big(A(M\varepsilon_n) \,\big|\, X_1,\ldots,X_n\big)(1-\phi_n) \le e^{(1+L'+C-M^2L)n\varepsilon_n^2} + o(1).$$

Hence, for all $L' > 0$ there exists a constant $M > 0$ such that the above expression converges to zero. This leads us to conclude that:

$$P_0^n \Pi\big(A(M\varepsilon_n) \,\big|\, X_1,\ldots,X_n\big) \to 0, \qquad (n \to \infty).$$

for sufficiently large $M > 0$.

6.4.3 *Entropy numbers and uniform test sequences*

Recall that the *packing number* $D(\varepsilon, \mathscr{P}, d)$ of a space $\mathscr{X}$ with metric d is defined as the maximal number of points in $\mathscr{P}$ such that the d-distance between all pairs is at least ε. This number is related to the so-called *covering number* $N(\varepsilon, \mathscr{P}, d)$ which is defined as the minimal number of d-balls of radius ε needed to cover $\mathscr{P}$, by the following inequalities: $N(\varepsilon, \mathscr{P}, d) \le D(\varepsilon, \mathscr{P}, d) \le N(\varepsilon/2, \mathscr{P}, d)$ (see exercise 6.6.3). Instead of condition (6.15), [106] imposes the following alternative condition in terms of these so-called *entropy numbers* relative to the Hellinger metrics on the models $\mathscr{P}_n$.

(ii') The ε-packing numbers $D(\varepsilon, \mathscr{P}_n, H)$ for the models $\mathscr{P}_n$ satisfy:

$$D(\varepsilon_n, \mathscr{P}_n, H) \leq e^{n\varepsilon_n^2}, \tag{6.21}$$

for large enough n, where ε_n is the rate sequence.

Entropy condition (6.21) implies the existence of a uniform sequence of test functions (see Le Cam (1973,1986) [177, 179] and Birgé (1983,1984) [32, 33]), as is shown below.

Recall that lemma 2.4.13 asserts the existence of minimax Hellinger tests between convex subsets separated by non-zero Hellinger distance. Let us consider two Hellinger balls B, V in $\mathscr{M}^1(\mathscr{X}, \mathscr{B})$ (which are convex, see exercise 6.6.4) at non-zero Hellinger distance in $\mathscr{P}$. For for every $n \geq 1$, there exists a test function $\phi_n : \mathscr{X}_n \to [0,1]$ such that,

$$\sup_{P \in B} P^n \phi_n + \sup_{Q \in V} Q^n (1 - \phi_n) \leq e^{-nH^2(B,V)}. \tag{6.22}$$

the *minimax Hellinger tests*. In the construction of tests suitable for theorem 6.4.3, the role of B is taken by $\{P_0\}$, but the complements of open neighbourhoods are not convex. In order to apply the above anyway, we *cover the alternative* with Hellinger balls (as first suggested in Le Cam (1973) [177]) and combine the individual tests between those balls and $\{P_0\}$ into a single test for the non-convex alternative. The *number* of balls needed in the cover (N below) then becomes a factor diminishing the testing power.

Lemma 6.4.4. *Fix $n, N \geq 1$ and $B, V_1, \ldots, V_N \subset \mathscr{P}$ and let V be any subset of $\cup_{i=1}^N V_i$. If there exist test functions ϕ_i, $(1 \leq i \leq N)$, such that (6.22) holds for all V_i, $(1 \leq i \leq N)$, then there exists a test function ψ such that,*

$$\sup_{P \in B} P^n \psi + \sup_{Q \in V} Q^n (1 - \psi) \leq N e^{-n \min\{H^2(B,V_i) : 1 \leq i \leq N\}}. \tag{6.23}$$

Proof. If the theorem holds for $\cup_{i=1}^N V_i$, then it holds for any subset thereof; so without loss of generality we assume that $V = \cup_{i=1}^N V_i$. Define $\psi = \max\{\phi_i : 1 \leq i \leq N\}$, then for any $P \in B$,

$$P^n \psi \leq \sum_{i=1}^N P^n \phi_i \leq N \max\{P^n \phi_i : 1 \leq i \leq N\}$$

and for every $Q \in V$,

$$\begin{aligned} Q^n (1 - \psi) &= Q^n \{\min(1 - \phi_i) : 1 \leq i \leq N\} \\ &\leq \sum_{i=1}^N Q^n (1 - \phi_i) \leq N \max\{Q^n (1 - \phi_i) : 1 \leq i \leq N\}. \end{aligned}$$

Combination leads to (6.23).

Now it is clear how the existence of test sequences (ϕ_n) as in (6.15) follows from the entropy condition (6.21). For some $M > 2$, define $B_n = \{P_0\}$, $V_n = \{P \in \mathscr{P}_n : H(P,P_0) \geq M\varepsilon_n\}$ and $N = N(\varepsilon_n, V_n, H) \leq N(\varepsilon_n, \mathscr{P}_n, H) \leq D(\varepsilon_n, \mathscr{P}_n, H) \leq e^{n\varepsilon^2}$. The *convex cover* $V_1, \ldots, V_N$ of V_n consists of Hellinger balls of radius ε_n centred in V_n, each of which satisfies $H(B,V_i) \geq (M-1)\varepsilon_n$ by virtue of the triangle inequality. Lemma 6.4.4 then says there exits a test sequence (ϕ_n) such that, for every $n \geq 1$,

$$\sup_{P \in B} P^n \psi + \sup_{Q \in V_n} Q^n(1-\psi) \leq N e^{-n(M-1)\varepsilon_n^2} \leq e^{-n(M-2)\varepsilon_n^2},$$

which is enough for (6.15). This way, rates of convergence ε_n are determined by upper bounds on entropy numbers of the type (6.21).

The sufficient condition for existence of suitable tests that [106] employs, the entropy condition (6.21), has to be proved in individual cases, however. To illustrate, if $\mathscr{P}_n = \mathscr{P}$ does not change with growing sample size, then finiteness of all Hellinger entropy numbers implies that $\mathscr{P}$ is totally bounded for the Hellinger metric. Hence the completion of $\mathscr{P}$ is Hellinger compact, which limits the applicability of the above somewhat. If $\mathscr{P}$ is σ-*compact* and Hellinger covering numbers do not grow too fast, the argument can be generalized.

Example 6.4.5. In certain infinite-dimensional spaces (like Sobolev balls, VC-classes or classes of monotone functions, see for example [247, 110]), entropy numbers have been calculated. Suppose that we have a parametrized model $\Theta \to \mathscr{P} : \theta \to P_\theta$ for single-observations and Θ is one of these examples, a subspace of a normed space with norm $\|\cdot\|$, such that, for some sequence η_n,

$$\log N(\eta_n, \Theta, \|\cdot\|) \leq n\eta_n^2, \tag{6.24}$$

and also that there exist two constants $K > 0$, $\alpha > 0$ such that for all $\theta_1, \theta_2 \in \Theta$ that are close enough, the Hellinger metric H is related to $\|\cdot\|$ through,

$$H(P_{\theta_1}, P_{\theta_2}) \leq K\|\theta_1 - \theta_2\|^\alpha. \tag{6.25}$$

Then $\|\cdot\|$-balls of radius η in Θ are mapped into H-balls of radius $K\eta^\alpha$, so that $\mathscr{P}$ has a cover of H-balls of radius ε of order (upper-bounded by) $N((\varepsilon/K)^{1/\alpha}, \Theta, \|\cdot\|)$. Therefore, condition (6.21) for the rate ε_n is determined by the entropy bound (6.24) for the parametrizing space. So if the model has a parameter space of the special kind for which entropy bounds can be calculated, and a relation of the type (6.25) (or slightly more involved, see (10.26) in chapter 10, for example) applies, then the existence of uniform test sequences can be guaranteed. Unfortunately, the number of examples of parameter spaces with known entropy bounds is somewhat limited [247].

We come back to this point and alternatives in chapters 7 and 9.

6.4.4 Lower bounds on prior mass

To conclude this section we give the lemma needed in the proof of theorem 6.4.3 to lower-bound the denominator of the posterior in probability, leading to *lower bounds on prior mass* locally around the true distribution of the data. This lemma, presently more a technical afterthought than an integral part of the theory, will be drawn to the foreground when we discuss remote contiguity in chapter 7.

Lemma 6.4.6. *Let $\varepsilon > 0$ and $P_0 \in \mathscr{P}$ be given and let $K(P_0, \varepsilon)$ be defined as in (6.13). If $\Pi(K(P_0,\varepsilon)) > 0$, then for every $L > 0$:*

$$P_0^n\Big(\int_{K(P_0,\varepsilon)} \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi(P) \le e^{-n\varepsilon^2(1+L)} \Pi(K(P_0,\varepsilon))\Big) \le \frac{1}{nL^2\varepsilon^2}. \tag{6.26}$$

Proof. Write Π' for $\Pi(\cdot|K(P_0,\varepsilon))$, the prior conditioned on $K(P_0,\varepsilon)$. By Jensen's inequality,

$$\log \int \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi'(P) \ge \sum_{i=1}^n \int \log \frac{p}{p_0}(X_i)\, d\Pi'(P).$$

Therefore, for any $L > 0$,

$$\begin{aligned} &P_0^n\Big(\int \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi'(P) \le e^{-n\varepsilon^2(1+L)})\Big) \\ &\quad \le P_0^n\Big(\sqrt{n}\big(\mathbb{P}_n - P_0\big) \int \log \frac{p}{p_0}\, d\Pi'(P) \\ &\qquad \le -\sqrt{n}(1+L)\varepsilon^2 - \sqrt{n}P_0 \int \log \frac{p}{p_0}\, d\Pi'(P)\Big). \end{aligned} \tag{6.27}$$

With the help of Fubini's theorem and the definition of $K(P_0,\varepsilon)$, we see that the *r.h.s.* above is bounded by $-\sqrt{n}L\varepsilon^2$. Note that the variance of the integrated log-likelihood is bounded with Jensen's inequality,

$$\mathrm{Var}\Big(\int \log \frac{p}{p_0}\, d\Pi'(P)\Big) \le P_0 \int \Big(\log \frac{p}{p_0}\Big)^2 d\Pi'(P), \tag{6.28}$$

and Chebyshev's inequality then implies that the *r.h.s.* of (6.27) is bounded by $(nL^2\varepsilon^2)^{-2}$ times the *r.h.s.* of (6.28). The definition of $K(P_0,\varepsilon)$ then shows that the assertion holds.

6.5 Frequentist counterexamples

To demonstrate that the assertion of Doob's Bayesian consistency theorem can be much weaker than expected in non-parametric setting, first Schwartz and then

Freedman constructed counterexamples in the early 1960's which are illustrated in subsection 6.5.1. In their time, Freedman's counterexamples and subsequent examples of *posterior inconsistency* established a widespread conviction that Bayesian methods were unfit for frequentist non-parametric purposes: examples of problematic posterior behaviour in non-parametric setting continued to captivate [71, 72, 61, 73, 74, 100, 101], while Schwartz's theorem received only limited (but steadily growing) amounts of attention [105]: subsequent frequentist theorems (*e.g.* by Barron (1988) [9], Barron-Schervish-Wasserman (1999) [14], Ghosal-Ghosh-van der Vaart (2000) [106], Shen-Wasserman (2001) [231], Walker (2004) [252] and Walker-Lijoi-Prünster (2007) [255], Kleijn-Zhao (2019) [156], Kleijn (2021) [157] and many others) have extended the applicability of theorem 6.3.1 but not its essence, the combination of a testing (or sufficient entropy) condition with a lower bound for local prior mass around the true distribution for the data.

Regarding the material of sections 6.3 and 6.4, finding counterexamples is straightforward: certainly Schwartz's classical theorem and work of Ghosal, Ghosh and van der Vaart [106] have been very influential and form the frequentist backbone for the literature on frequentist non-parametric Bayesian statistics since 2000; but, as demonstrated insubsection 6.5.2, there are very simple parametric models in which the Kullback-Leibler priors of Schwartz or their more specific Ghosal-Ghosh-van der Vaart variations *do not exist*. The fact that posterior consistency obtains without problems in those examples, serves as one of the motivations for the generalizations we discuss in chapter 7.

6.5.1 Freedman's counterexamples

The first examples of unexpected posterior inconsistency are due to Schwartz (1961) [225], but it was Freedman (1963) [97] who made the point famous with a simple non-parametric counterexample, discussed in detail as example 6.5.1 below. In Freedman (1965) [98] it was even shown that, without further conditions on the prior, inconsistency is generic in a topological sense (see theorem 6.5.2).

Example 6.5.1. (Freedman (1963) [97])
Consider a sample $X_1, X_2, \ldots$ of random positive integers. Denote the space of all probability distributions on $\mathbb{N}$ by Λ and assume that the sample is *i.i.d.*-P_0, for some $P_0 \in \Lambda$. For any $P \in \Lambda$, write $p(i) = P(\{X = i\})$ for all $i \geq 1$. The total-variational and weak topologies on Λ are equivalent (defined, $P \to Q$ if $p(i) \to q(i)$ for all $i \geq 1$). Let $Q \in \Lambda \setminus \{P_0\}$ be given. To arrive at a prior with P_0 in its support, leading to a posterior that concentrates on Q, we consider sequences (P_m) and (Q_n) such that $Q_m \to Q$ and $P_m \to P_0$ as $m \to \infty$. The prior Π places masses $\alpha_m > 0$ at P_m and $\beta_m > 0$ at Q_m ($m \geq 1$), so that P_0 lies in the support of Π. A careful construction of the distributions Q_m that involves P_0, guarantees that the posterior satisfies,

$$\frac{\Pi(\{Q_m\}|X^n)}{\Pi(\{Q_{m+1}\}|X^n)} \xrightarrow{P_0\text{-a.s.}} 0,$$

that is, posterior mass is shifted further out into the tail as n grows to infinity, forcing all posterior mass that resides in $\{Q_m : m \geq 1\}$ into arbitrarily small neighbourhoods of Q. In a second step, the distributions P_m and prior weights α_m are chosen such that the likelihood at P_m grows large for high values of m and small for lower values as n increases, so that the posterior mass in $\{P_m : m \geq 1\}$ also accumulates in the tail. However, the prior weights α_m may be chosen to decrease very fast with m, in such a way that,

$$\frac{\Pi(\{P_m : m \geq 1\}|X^n)}{\Pi(\{Q_m : m \geq 1\}|X^n)} \xrightarrow{P_0\text{-a.s.}} 0,$$

thus forcing all posterior mass into $\{Q_m : m \geq 1\}$ as n grows. Combination of the previous two displays leads to the conclusion that for every neighbourhood U_Q of Q,

$$\Pi(U_Q|X^n) \xrightarrow{P_0\text{-a.s.}} 1,$$

so the posterior is inconsistent. Other choices of the weights α_m that place more prior mass in the tail *do* lead to consistent posterior distributions.

Some objected to Freedman's counterexample, because knowledge of P_0 is required to construct the prior that causes inconsistency. So it was possible to argue that Freedman's counterexample amounted to nothing more than a demonstration that unfortunate circumstances could be created, probably not a fact of great concern in any generic sense.

To strengthen Freedman's point one would need to construct a prior of full support without explicit knowledge of P_0. In the setting of example 6.5.1, denote the space of all distributions on Λ by $\pi(\Lambda)$. Note that since Λ is Polish, so is $\pi(\Lambda)$ and so is the product $\Lambda \times \pi(\Lambda)$.

Theorem 6.5.2. *(Freedman's posterior inconsistency theorem)*
Let $X_1, X_2, \ldots$ form an sample of i.i.d.-P_0 random integers, let Λ denote the space of all distributions on $\mathbb{N}$ and let $\pi(\Lambda)$ denote the space of all Borel probability measures on Λ, both in Prohorov's weak topology. The set of pairs $(P_0, \Pi) \in \Lambda \times \pi(\Lambda)$ such that for all open $U \subset \Lambda$,

$$\limsup_{n\to\infty} P_0^n \Pi(U|X^n) = 1,$$

is residual.

Proof. See Freedman (1965) [98] and Le Cam (1986) [179].

And so, the set of pairs $(P_0, \Pi) \in \Lambda \times \pi(\Lambda)$ for which the limiting behaviour of the posterior is acceptable to the frequentist, is *meagre* in $\Lambda \times \pi(\Lambda)$. The proof is based on example 2.1.18. The question arises, what is the conclusion we draw from Freedman's objections of inconsistency? (See [71, 72, 74, 100, 101] and Le Cam's comment [180]). Leaving constructions with intentional pathology aside, it is theorem 6.5.2 that poses the real challenge to non-parametric Bayesian statistics. However, its message is quite encouraging when interpreted correctly: meagreness in the sense of theorem 6.5.2 means that there is a condition missing. Not all priors

are fit for frequentist purpose, indeed a (topologically large) subset of priors are not. The remaining priors, those that *are* useful to the frequentist, form a (topologically small) subset, characterized by a property. Freedman failed to recognize that his result was indicative of the next step in the theoretical development: Schwartz's Kullback-Leibler condition (6.5) provides exactly such a property (which may explain the tone of [180]).

6.5.2 Counterexamples: Schwartz and GGV conditions

In this subsection, it is shown that there exist very simple parametric models in which no prior satisfies Schwartz's Kullback-Leibler condition (6.5), and similarly, that there are very simple parametric models in which Schwartz's Kullback-Leibler condition may be satisfied, but no prior satisfies the Ghosal-Ghosh-van der Vaart condition (6.14).

Example 6.5.3. Consider $X_1, X_2, \ldots$ that are *i.i.d.*-P_0 with Lebesgue density $p_0 : \mathbb{R} \to \mathbb{R}$ supported on an interval of known width (say, 1) but unknown location. Parametrize in terms of a continuous density η on $[0,1]$ with $\eta(x) > 0$ for all $x \in [0,1]$ and a location $\theta \in \mathbb{R}$: $p_{\theta,\eta}(x) = \eta(x-\theta)\, 1_{[\theta,\theta+1]}(x)$. A moment's thought makes clear that if $\theta \neq \theta'$,

$$-P_{\theta,\eta} \log \frac{p_{\theta',\eta'}}{p_{\theta,\eta}} = \infty,$$

for all η, η'. Therefore Kullback-Leibler neighbourhoods do not have any extent in the θ-direction and *no prior is a Kullback-Leibler prior in this model*.

Example 6.5.4. Consider an *i.i.d.* sample of integers $X_1, X_2, \ldots$ from a heavy-tailed distribution P_a, $(a \geq 1)$, defined by,

$$p_a(k) = P_a(X = k) = \frac{1}{Z_a} \frac{1}{k^a (\log k)^3} \tag{6.29}$$

for all $k \geq 2$, with $Z_a = \sum_{k \geq 2} k^{-a} (\log k)^{-3} < \infty$. As it turns out (see exercise 6.6.5), for $a = 1$, $b > 1$,

$$-P_a \log \frac{p_b}{p_a} < \infty, \quad P_a \left(\log \frac{p_b}{p_a} \right)^2 = \infty. \tag{6.30}$$

Therefore, Schwartz's KL-condition (6.5) for the prior for the parameter a can be satisfied but there exists no prior such that (6.14) is satisfied for all P_0 in the model, that is, there is no *Ghosal-Ghosh-van der Vaart prior*. In fact, if we change the third power of the log-factor in the denominator of (6.29) to a square, Schwartz's KL-priors also do not exist.

Nonetheless, in chapter 7 it is shown that with generic choices for the priors, the posteriors of both examples 6.5.3 and 6.5.4 are appropriately consistent (see examples **??** and **??**).

Simple as the above examples are, they are indicative of a more general problem (that is clearly difficult to contain through the choice of priors on non-parametric models and explains the multitudes of sufficient conditions that non-parametric applications of Schwartz's theorem and the GGV-theorem to specific models often entail): for any P_0 it is possible to find distributions P with density ratios p/p_0 that vary 'wildly enough' to cause log-likelihood ratios $\log p/p_0$ to lose integrability or square-integrability. Originating in early analyses of posterior inconsistency [14, 71], the phenomenon of *data-tracking* [254] sketches a similar qualitative picture of situations where posterior consistency fails. The notion of remote contiguity of chapter 7 defines a precise way in which variations of p/p_0 may be bounded to guarantee consistency, which also covers examples like 6.5.3 and 6.5.4.

6.6 Exercises

6.6.1. Let $\mathscr{X}$ be Polish and let $A \subset \mathscr{X}$ be Borel measurable. Show that the function $P \mapsto P(A)$ is measurable with respect to the Borel σ-algebra associated with Prokhorov's weak topology on $M^1(\mathscr{X})$.

6.6.2. Assuming (6.11), show that for every sequence $M_n \to \infty$, the centre point $\tilde{P}'_n$ of the d-ball of minimal radius containing posterior mass $p > 1/2$ satisfies,

$$P_0^n\big(d(\tilde{P}'_n, P_0) \leq 2M_n\varepsilon_n\big) \to 1. \tag{6.31}$$

As a result $\tilde{P}'_n$ converges to P_0 with respect to d (at least) at rate ε_n.

6.6.3. Let $(\mathscr{P}, d)$ be a metric model. Prove that packing and covering numbers satisfy,

$$N(\varepsilon, \mathscr{P}, d) \leq D(\varepsilon, \mathscr{P}, d) \leq N(\varepsilon/2, \mathscr{P}, d),$$

for all $\varepsilon > 0$.

6.6.4. Although the word 'ball' is associated with convexity intuitively, the *metric balls* that play such a prominent role in this chapter and elsewhere are *not guaranteed to be convex*.

a. Let $(\mathscr{X}, \mathscr{B})$ be a measurable space. Show that a Hellinger ball in $\mathscr{M}^1(\mathscr{X}, \mathscr{B})$ is convex. *Hint: consider* $H^2(P, Q)$. *See lemma 3 of section 16.4 of [179].*
b. Give an example of a convex metric space $(\mathscr{X}, d)$ with metric balls that are not convex.

6.6.5. In example 6.5.4, prove that (6.30) holds. Also show that if we change the third power of the log-factor in the denominator of (6.29) to a square, KL-priors also do not exist.

Chapter 7
Frequentist validity of Bayesian limits

In this chapter, we re-develop the Bayesian theory of non-parametric statistics of the previous chapter from the ground up. The calculations presented in chapter 6 are more than adequate if one is willing to restrict attention to the "traditional" setting for examples in non-parametric statistics, where the data forms an *i.i.d.* sample from a distribution in a model of Hellinger entropy with known upper-bound, following the path set out in subsections 6.4.3 and 6.4.4.

The setting for present-day non-parametric statistical challenges is more general, though: data in the computer age is not only of very large scale, it is often of much more complex structure than that of an *i.i.d.* data set. *Dependence* among data points not only occurs in stochastic processes, like the *time-series* that typify data in financial markets, but also in random walks and branching processes that occur on graphs, like the widely used world-wide-web sampling technique of *webcrawling*. Non-parametric models for complex data like those that arise from questions in machine learning and network science are usually not compatible with the technical formulation of subsections 6.4.3 and 6.4.4. (The iterated maps of so-called *deep neural networks* and the highly dependent *preferential attachment model* for random graphs are two examples, and so is the community-detection problem of chapter 11.)

Below we re-examine for which priors Bayesian limits are limits valid in the frequentist sense: is Schwartz's Kullback-Leibler condition perhaps a manifestation of a more general notion? The argument leads to other questions for which insightful answers have been elusive: why is Doob's theorem completely different from Schwartz's? Why does weak consistency in the full non-parametric model (*e.g.* with the Dirichlet process prior [94], or more modern variations [64]) reside in a corner of its own (with *tailfreeness* [98, 92] as sufficient property of the prior), apparently unrelated to posterior consistency in either Doob's or Schwartz's views? And to extend the scope further, what can be said about hypothesis testing, classification, model selection, *etcetera*? Given that the Bernstein-von Mises theorem cannot be expected to hold in any generality outside parametric setting [61, 101], what relationship exists between credible sets and confidence sets?

The central property to enable frequentist interpretation of posterior asymptotics is defined as *remote contiguity* in section 7.2. It expresses a weakened form of Le Cam's contiguity, relating the true distribution of the data to localized prior predictive distributions. Where Schwartz's Kullback-Leibler neighbourhoods represent a choice for the localization appropriate when the sample is *i.i.d.*, remote contiguity generalises the notion to include non-*i.i.d.* samples and sample-size-dependent model/prior pairs. We then see how Doob's prior-almost-sure consistency is strengthened to reach Schwartz's frequentist conclusion, or how a test that is consistent prior-almost-surely becomes a test that is consistent in *all* points of the model, or how a Bayesian credible set can be "enlarged" to serve as a frequentist confidence set asymptotically. The latter point extends the main implication of the Bernstein-von Mises theorem, theorem 4.2.1, to non-parametric models provided the prior induces remote contiguity.

In section 7.1 we concentrate on an inequality that relates testing to posterior concentration and indicates the relation with Le Cam's inequality. Section 7.2 introduces remote contiguity and the analogue of Le Cam's First Lemma. In section 7.4, frequentist theorems on the asymptotic behaviour of posterior distributions are proved, on posterior consistency, on rates of convergence, on model selection with posterior odds and on the conversion of credible sets to confidence sets. Section 7.8 formulates the conclusions. The central condition of testability is analysed further in chapter 9. Application to community detection in random graphs follows in chapter 11.

7.1 Posterior concentration and asymptotic tests

First we consider a lemma that relates concentration of posterior mass in certain model subsets to the existence of test sequences that distinguish between those subsets. More precisely, it is shown that the expected posterior mass outside a model subset V with respect to the local prior predictive distribution over a model subset B, is upper bounded (roughly) by the testing power of *any* statistical test for the hypotheses B versus V: if a test sequence exists, the posterior will concentrate its mass appropriately.

7.1.1 Bayesian test sequences

We follow Schwartz and consider asymptotic testing; however, we define test sequences immediately in Bayesian context by involving priors from the outset. We consider sequentially observed, (possibly non-*i.i.d.*) samples X^n, distributed according to $P_{\theta_0,n}$ for some $\theta_0 \in \Theta$, within the model $\theta \to P_{\theta,n}$. (More generally, we refer to appendix A for the notation and conventions assumed through this chapter.)

Definition 7.1.1. Given priors (Π_n) on the measurable space $(\Theta, \mathscr{G})$, model subsets $(B_n), (V_n) \subset \mathscr{G}$ and $a_n \downarrow 0$, a sequence of $\mathscr{B}_n$-measurable maps $\phi_n : \mathscr{X}_n \to [0,1]$ is called a *Bayesian test sequence for B_n versus V_n (under Π_n) of power a_n*, if,

$$\int_{B_n} P_{\theta,n}\phi_n \, d\Pi_n(\theta) + \int_{V_n} P_{\theta,n}(1-\phi_n)\, d\Pi_n(\theta) = o(a_n). \tag{7.1}$$

We say that (ϕ_n) is a *Bayesian test sequence for B_n versus V_n (under Π_n)* if (7.1) holds for some $a_n \downarrow 0$.

Note that if we have sequences (C_n) and (W_n) such that $C_n \subset B_n$ and $W_n \subset V_n$ for all $n \geq 1$, then a Bayesian test sequence for (B_n) versus (V_n) of power a_n is a Bayesian test sequence for (C_n) versus (W_n) of power (at least) a_n.

Lemma 7.1.2. *For any $B, V \in \mathscr{G}$ and any measurable $\phi : \mathscr{X} \to [0,1]$,*

$$\int_B P_\theta \Pi(V|X)\, d\Pi(\theta) \leq \int_B P_\theta \phi \, d\Pi(\theta) + \int_V P_\theta(1-\phi)\, d\Pi(\theta). \tag{7.2}$$

Proof. Due to Bayes's Rule (A.4) and monotone convergence,

$$\int (1-\phi(X))\, \Pi(V|X)\, dP^\Pi = \int_V P_\theta(1-\phi(X))\, d\Pi(\theta).$$

Accordingly,

$$\int_B P_\theta(1-\phi)\, \Pi(V|X)\, d\Pi(\theta) \leq \int (1-\phi)\, \Pi(V|X)\, dP^\Pi = \int_V P_\theta(1-\phi)\, d\Pi(\theta).$$

Inequality (7.2) follows from the fact that $\Pi(V|X) \leq 1$.

So the mere existence of a test sequence is enough to guarantee posterior concentration, a fact expressed in n-dependent form through the following proposition.

Proposition 7.1.3. *Let $(\mathscr{X}_n, \mathscr{B}_n)$, $(\Theta_n, \mathscr{G}_n)$, $(\mathscr{P}_n)$ and (Π_n) be given. Given sequences $(B_n), (V_n) \subset \mathscr{G}_n$ and $a_n, b_n, c_n \downarrow 0$ such that $a_n = o(b_n \wedge c_n)$ and,*

$$\Pi_n(B_n) \geq b_n > 0, \quad \Pi_n(V_n) \geq c_n > 0,$$

the following are equivalent:

(i) there exists a Bayesian test sequence for B_n versus V_n of power a_n,
(ii) mutually expected posterior weights vanish as follows,

$$P_n^{\Pi_n|B_n}\Pi(V_n|X^n) = o(a_n b_n^{-1}), \quad P_n^{\Pi_n|V_n}\Pi(B_n|X^n) = o(a_n c_n^{-1}), \tag{7.3}$$

for all $n \geq 1$.

Proof. Assume *(i)*. Then (and analogously for V_n),

$$P_n^{\Pi_n|B_n}\Pi(V_n|X^n) = b_n^{-1} \int_{B_n} P_{\theta,n}\Pi(V_n|X^n)\, d\Pi_n(\theta) = o(a_n b_n^{-1}).$$

Assume *(ii)*. Without loss of generality, assume that $\Pi_n(B_n \cup V_n) = 1$ for all $n \geq 1$. Then,

$$b_n P_n^{\Pi_n|B_n} \Pi(V_n|X^n) + c_n P_n^{\Pi_n|V_n} \Pi(B_n|X^n) = o(a_n),$$

so $\phi_n(X^n) = \Pi(V_n|X^n)$ defines a Bayesian test sequence of power a_n.

We come back to the equivalence of Bayesian test existence and posterior concentration in subsection 7.1.2, as well as in section 7.4. To illustrate how proposition 7.1.3 relates to frequentist posterior concentration and how this involves remote contiguity, consider model subsets $V_n = V$ that are all equal to the complement of a neighbourhood U of P_0. The subsets $B_n = B$ are thought of as being even closer to the $P_{0,n}$, in such a way that the expectations of the random variables $X^n \mapsto \Pi(V|X^n)$ under $P_n^{\Pi_n|B_n}$ "dominate" their expectations under $P_{0,n}$ in a suitable way. Then sufficiency of prior mass b_n given testing power a_n, is enough to assert that $P_{0,n}\Pi(V|X^n) \to 0$. Remote contiguity makes this notion of domination precise.

7.1.2 Existence of Bayesian test sequences

Lemma 7.1.2 and proposition 7.1.3 require the existence of test sequences of the Bayesian type. That question is unfamiliar, frequentists are used to test sequences for pointwise or uniform testing, *e.g.* those of subsection 6.4.3. Another example is formed by complements of *weak* neighbourhoods, which are testable uniformly as we shall see in chapter 9.

Requiring the existence of a Bayesian test sequence *c.f.* (7.1) is quite different. We shall illustrate this point in various ways below. First of all the existence of a Bayesian test sequence is linked directly to behaviour of the posterior itself.

Theorem 7.1.4. *Let $(\Theta, \mathscr{G}, \Pi)$ be given and assume that there is a coupling $X \in X^\infty$ with distribution P_θ and marginals $X^n \sim P_{\theta,n}$ for every $\theta \in \Theta$ and $n \geq 1$. For any $B, V \in \mathscr{G}$ with $\Pi(B) > 0, \Pi(V) > 0$, the following are equivalent:*

(i) there are $\mathscr{B}_n$-msb. $\phi_n : \mathscr{X}_n \to [0,1]$ such that for Π-almost-all $\theta \in B, \theta' \in V$,

$$\phi_n(X^n) \xrightarrow{P_\theta\text{-a.s.}} 0, \quad \phi_n(X^n) \xrightarrow{P_{\theta'}\text{-a.s.}} 1,$$

(ii) there are $\mathscr{B}_n$-msb. $\phi_n : \mathscr{X}_n \to [0,1]$ such that for Π-almost-all $\theta \in B, \theta' \in V$,

$$P_{\theta,n}\phi_n \to 0, \quad P_{\theta',n}(1-\phi_n) \to 0,$$

(iii) there are $\mathscr{B}_n$-msb. $\phi_n : \mathscr{X}_n \to [0,1]$ such that,

$$\int_B P_{\theta,n}\phi_n \, d\Pi(\theta) + \int_V P_{\theta,n}(1-\phi_n) \, d\Pi(\theta) \to 0,$$

(iv) for Π-almost-all $\theta \in B$, $\theta' \in V$,

$$\Pi(V|X^n) \xrightarrow{P_{\theta,n}\text{-a.s.}} 0, \qquad \Pi(B|X^n) \xrightarrow{P_{\theta',n}\text{-a.s.}} 0.$$

Proof. Condition *(i)* implies *(ii)* trivially and *(ii)* implies *(iii)* by dominated convergence. Assume *(iii)* and note that by lemma 7.1.2,

$$\int P_{\theta,n}\Pi(V|X^n)\, d\Pi(\theta|B) \to 0.$$

With the coupling X of the observations X^n, martingale convergence in $L^1(\mathscr{X}^\infty \times \Theta)$ (relative to the probability measure Π^* defined by $\Pi^*(A \times B) = \int_B P_\theta(A)\, d\Pi(\theta)$ for measurable $A \subset \mathscr{X}^\infty$ and $B \subset \Theta$), shows there is a measurable $g : \mathscr{X}^\infty \to [0,1]$ such that,

$$\int P_\theta \big|\Pi(V|X^n) - g(X)\big|\, d\Pi(\theta|B) \to 0.$$

So $\int P_\theta g(X)\, d\Pi(\theta|B) = 0$, implying that $g = 0$, P_θ-almost-surely for Π-almost-all $\theta \in B$. Using martingale convergence again (now in $L^\infty(\mathscr{X}^\infty \times \Theta)$), conclude $\Pi(V|X^n) \to 0$, P_θ-almost-surely for Π-almost-all $\theta \in B$, from which *(iv)* follows. Choose $\phi(X^n) = \Pi(V|X^n)$ to conclude that *(i)* follows from *(iv)*.

The interpretation of this theorem is gratifying to supporters of the likelihood principle and pure Bayesians: distinctions between model subsets are Bayesian testable, if and only if, they are picked up by the posterior asymptotically, if and only if, there exists a pointwise test for B versus V that is Π-almost-surely consistent.

For a second, more frequentist way to illustrate how basic the existence of a Bayesian test sequences is, consider a parameter space (Θ, d) which is a metric space with fixed Borel prior Π and d-consistent estimators $\hat{\theta}_n : \mathscr{X}_n \to \Theta$ for θ. Then for every $\theta_0 \in \Theta$ and $\varepsilon > 0$, there exists a pointwise test sequence (and hence, by dominated convergence, also a Bayesian test sequence) for $B = \{\theta \in \Theta : d(\theta, \theta_0) < \frac{1}{2}\varepsilon\}$ versus $V = \{\theta \in \Theta : d(\theta, \theta_0) > \varepsilon\}$. This approach is followed in example 7.6.4 on random walks, see the definition of the test following inequality (7.26).

A third perspective on the existence of Bayesian tests arises from Doob's argument. From our present perspective, we note that theorem 9.5.1 implies a proof of Doob's consistency theorem through the following existence result on Bayesian test sequences. (Note: here and elsewhere in *i.i.d.* setting, the parameter space Θ is the single-observation model $\mathscr{P}$, θ is the single-observation distribution P and $\theta \mapsto P_{\theta,n}$ is $P \mapsto P^n$.)

Proposition 7.1.5. *Consider a model $\mathscr{P}$ of single-observation distributions P for i.i.d. data $(X_1, X_2, \ldots, X_n) \sim P^n$, $(n \geq 1)$. Assume that $\mathscr{P}$ is a Polish space with Borel prior Π. For any Borel set V there is a Bayesian test sequence for V versus $\mathscr{P} \setminus V$ under Π.*

Proof. We prove this theorem in chapter 9. (See [179], section 17.1, proposition 1 with the indicator for V; see also [51].)

Doob's theorem is recovered when we let V be the complement of any open neighbourhood U of P_0. Comparing with conditions for the existence of uniform tests,

Bayesian tests are quite abundant: whereas uniform testing relies on the minimax theorem (forcing convexity, compactness and continuity requirements into the picture), Bayesian tests exist quite generally (at least, for Polish parameters with *i.i.d.* data).

The fourth perspective on the existence of Bayesian tests concerns a direct way to construct a Bayesian test sequence of optimal power, based on the fact that we are really only testing *barycentres* against each other: let priors (Π_n) and $\mathscr{G}$-measurable model subsets B_n, V_n be given. For given tests (ϕ_n) and power sequence a_n, write (7.1) as follows:

$$\Pi_n(B_n)\, P_n^{\Pi_n|B_n}\phi_n(X^n) + \Pi_n(V_n)\, P_n^{\Pi_n|V_n}(1-\phi_n(X^n)) = o(a_n),$$

and note that what is required here, is a (weighted) test of $(P_n^{\Pi_n|B_n})$ versus $(P_n^{\Pi_n|V_n})$. The likelihood-ratio test of example 2.4.9 (denote the density for $P_n^{\Pi_n|B_n}$ with respect to $\mu_n = P_n^{\Pi_n|B_n} + P_n^{\Pi_n|V_n}$ by $p_{B_n,n}$, and similar for $P_n^{\Pi_n|V_n}$),

$$\phi_n(X^n) = 1_{\{\Pi_n(V_n)\, p_{V_n,n}(X^n) > \Pi_n(B_n)\, p_{B_n,n}(X^n)\}},$$

is optimal and has power $\|\Pi_n(B_n)\, P_n^{\Pi_n|B_n} \wedge \Pi_n(V_n)\, P_n^{\Pi_n|B_n}\|$. This proves the following useful proposition that re-expresses power in terms of the relevant Hellinger transform (see Remark 1 of section 16.4 in Le Cam (1986) [179]).

Proposition 7.1.6. *Fix $n \geq 1$ and let a prior (Π_n) and measurable model subsets B_n, V_n be given. There exists a test function $\phi_n : \mathscr{X}_n \to [0,1]$ such that,*

$$\begin{aligned}\int_{B_n} P_{\theta,n}\phi_n\, d\Pi_n(\theta) + \int_{V_n} P_{\theta,n}(1-\phi_n)\, d\Pi_n(\theta) \\ \leq \int \Big(\Pi_n(B_n)\, p_{B_n,n}(x)\Big)^{\alpha} \Big(\Pi_n(V_n)\, p_{V_n,n}(x)\Big)^{1-\alpha} d\mu_n(x),\end{aligned} \tag{7.4}$$

for any $0 \leq \alpha \leq 1$.

Proposition 7.1.6 generalises proposition 7.1.5 and makes Bayesian tests available with a sharp bound on the power under fully general conditions. For the connection with minimax tests, we note the following. If $\{P_{\theta,n} : \theta \in B_n\}$ and $\{P_{\theta,n} : \theta \in V_n\}$ are *convex* sets (and the Π_n are Radon measures, *e.g.* in Polish parameter spaces), then,

$$H\big(P_n^{\Pi_n|B_n}, P_n^{\Pi_n|V_n}\big) \geq \inf\{H(P_{\theta,n}, P_{\theta',n}) : \theta \in B_n, \theta' \in V_n\}.$$

Combination with (7.4) for $\alpha = 1/2$, implies that the minimax upper bound in *i.i.d.* cases, *c.f.* proposition **??**, remains valid:

$$\int_{B_n} P^n\phi_n\, d\Pi_n(P) + \int_{V_n} Q^n(1-\phi_n)\, d\Pi_n(Q) \leq \sqrt{\Pi_n(B_n)\,\Pi_n(V_n)}\, e^{-n\varepsilon_n^2}, \tag{7.5}$$

where $\varepsilon_n = \inf\{H(P,Q) : P \in B_n, Q \in V_n\}$. Given $a_n \downarrow 0$, any pointwise test ϕ_n that satisfies (7.1) *for all* probability measures Π_n on Θ, is a (weighted) minimax test for B_n versus V_n of power a_n.

Note that the above enhances the role that the prior plays in the frequentist discussion of the asymptotic behaviour of the posterior: the prior is not only important in requirements like (6.5), but can also be of influence in the testing condition: where testing power is relatively weak, prior mass should be scarce to compensate and where testing power is strong, prior mass can be plentiful. To make use of this, one imposes *upper bounds on prior mass* in certain hard-to-test subsets of the model (as opposed to *lower bounds* like (6.5)). See example 7.6.4 on random-walk data. In the Hellinger-geometric view, the prior determines whether the local prior predictive distributions $P_n^{\Pi_n|B_n}$ and $P_n^{\Pi_n|V_n}$ lie close together or not in Hellinger distance, and thus to the *r.h.s.* of (7.4) for $\alpha = 1/2$.

7.1.3 Le Cam's inequality

Referring to the argument following proposition 7.1.3, one way of guaranteeing that the expectations of $X^n \mapsto \Pi(V|X^n)$ under $P_n^{\Pi|B_n}$ approximate those under $P_{0,n}$, is to choose $B_n = \{\theta \in \Theta : \|P_{\theta,n} - P_{\theta_0,n}\| \leq \delta_n\}$, for some sequence $\delta_n \to 0$, because in that case, $|P_{0,n}\psi - P_n^{\Pi|B_n}\psi| \leq \|P_{0,n} - P_n^{\Pi|B_n}\| \leq \delta_n$, for any random variable $\psi : \mathscr{X}_n \to [0,1]$. Without fixing the definition of the sets B_n, one may use this step to specify inequality (7.2) further:

$$\begin{aligned} P_{0,n}\Pi(V_n|X) \leq{}& \left\|P_{0,n} - P_n^{\Pi|B_n}\right\| \\ &+ \int P_{\theta,n}\phi_n \, d\Pi_n(\theta|B_n) + \frac{\Pi_n(V_n)}{\Pi_n(B_n)} \int P_{\theta,n}(1-\phi_n)\, d\Pi_n(\theta|V_n), \end{aligned} \tag{7.6}$$

for B_n and V_n such that $\Pi_n(B_n) > 0$ and $\Pi_n(V_n) > 0$. Le Cam's inequality (7.6) is used, for example, in the proof of the Bernstein-von Mises theorem, see lemma 2 in section 8.4 of [183]. A less successful application pertains to non-parametric posterior rates of convergence for *i.i.d.* data, in an unpublished paper [178]. Rates of convergence obtained in this way are suboptimal: Le Cam qualifies the first term on the right-hand side of (7.6) as a *"considerable nuisance"* and concludes that *"it is unclear at the time of this writing what general features, besides the metric structure, could be used to refine the results"*, (see [179], end of section 16.6). In [259], Le Cam relates the posterior question to dimensionality restrictions [177, 231, 106] and reiterates, *"And for Bayes risk, I know that just the metric structure does not catch everything, but I don't know what else to look at, except calculations."*

7.2 Remote contiguity

Le Cam's notion of contiguity (see Le Cam (1960) [174]) describes an asymptotic version of absolute continuity, applicable to sequences of probability measures in a limiting sense. A condensed overview of the most basic characterizations of contiguity and some essential references are found in appendix C.10. In this section we weaken the property of contiguity in a way that is suitable to promote Π-almost-everywhere Bayesian limits to frequentist limits that hold everywhere.

7.2.1 Definition and criteria for remote contiguity

The notion of "domination" left undefined in the argument following proposition 7.1.3 is made rigorous here.

Definition 7.2.1. Given measurable spaces $(\mathscr{X}_n, \mathscr{B}_n)$, $n \geq 1$ with two sequences (P_n) and (Q_n) of probability measures and a sequence $\rho_n \downarrow 0$, we say that Q_n is ρ_n-*remotely contiguous* with respect to P_n, notation $Q_n \lhd \rho_n^{-1} P_n$, if,

$$P_n \phi_n(X^n) = o(\rho_n) \quad \Rightarrow \quad Q_n \phi_n(X^n) = o(1), \tag{7.7}$$

for every sequence of $\mathscr{B}_n$-measurable $\phi_n : \mathscr{X}_n \to [0,1]$.

Note that for a sequence (Q_n) that is a_n-remotely contiguous with respect to (P_n), there exists no test sequence that distinguishes between P_n and Q_n with power of order $o(a_n)$. Note also that given two sequences (P_n) and (Q_n), contiguity $P_n \lhd Q_n$ is equivalent to remote contiguity $P_n \lhd a_n^{-1} Q_n$ for all $a_n \downarrow 0$. Given sequences $a_n, b_n \downarrow 0$ with $a_n = O(b_n)$, b_n-remote contiguity implies a_n-remote contiguity of (P_n) with respect to (Q_n).

Example 7.2.2. Let $\mathscr{P}$ be a model for the distribution of a single-observation in *i.i.d.* samples $X^n = (X_1, \ldots, X_n)$. Let P_0, P and $\varepsilon > 0$ be such that $-P_0 \log(dP/dP_0) < \varepsilon^2$. The *law of large numbers* implies that for large enough n,

$$\frac{dP^n}{dP_0^n}(X^n) \geq e^{-\frac{n}{2}\varepsilon^2}, \tag{7.8}$$

with P_0^n-probability one. Consequently, for large enough n and for any $\mathscr{B}_n$-measurable sequence $\psi_n : \mathscr{X}_n \to [0,1]$,

$$P^n \psi_n \geq e^{-\frac{1}{2}n\varepsilon^2} P_0^n \psi_n. \tag{7.9}$$

Therefore, if $P^n \phi_n = o(\exp(-\frac{1}{2}n\varepsilon^2))$ then $P_0^n \phi_n = o(1)$. Conclude that for every $\varepsilon > 0$, the Kullback-Leibler neighbourhood $\{P : -P_0 \log(dP/dP_0) < \varepsilon^2\}$ consists of model distributions for which the sequence (P_0^n) of product distributions are $\exp(-\frac{1}{2}n\varepsilon^2)$-remotely contiguous with respect to (P^n).

Criteria for remote contiguity are given in the lemma below; note that, here, we give sufficient conditions, rather than necessary and sufficient, as in Le Cam's First Lemma, lemma C.10.2 or . (For the precise, Q_n-almost-sure definition of $(dP_n/dQ_n)^{-1}$, see appendix A.)

Lemma 7.2.3. *Given probability measures (P_n), (Q_n) on measurable spaces $(\mathscr{X}_n, \mathscr{B}_n)$ and $a_n \downarrow 0$, $Q_n \lhd a_n^{-1} P_n$, if any of the following hold:*

(i) for any bounded, $\mathscr{B}_n$-measurable $T_n : \mathscr{X}_n \to [0,1]$, $a_n^{-1} T_n \xrightarrow{P_n} 0$ implies $T_n \xrightarrow{Q_n} 0$,
(ii) for any $\varepsilon > 0$, there is a $\delta > 0$ such that $Q_n(dP_n/dQ_n < \delta a_n) < \varepsilon$, for large enough n,
(iii) there is a $b > 0$ such that $\liminf_n b a_n^{-1} P_n(dQ_n/dP_n > b a_n^{-1}) = 1$,
(iv) for any $\varepsilon > 0$, there is a constant $c > 0$ such that $\|Q_n - Q_n \wedge c a_n^{-1} P_n\| < \varepsilon$, for large enough n,
(v) under Q_n every subsequence of $(a_n (dP_n/dQ_n)^{-1})$ has a weakly convergent subsequence.

Remark 7.2.4. The proof of this lemma actually shows that (*(i)* or *(iv)*) implies remote contiguity; that (*(ii)* or *(iii)*) implies *(iv)* and that *(v)* is equivalent to *(ii)*.

Proof. Assume *(i)*. Let $\phi_n : \mathscr{X}_n \to [0,1]$ be given and assume that $P_n \phi_n = o(a_n)$. By Markov's inequality, for every $\varepsilon > 0$, $P_n(a_n^{-1} \phi_n > \varepsilon) = o(1)$. By assumption, it now follows that $\phi_n \xrightarrow{Q_n} 0$. Because $0 \le \phi_n \le 1$ the latter conclusion is equivalent to $Q_n \phi_n = o(1)$. Conclude that $Q_n \lhd a_n^{-1} P_n$. Next, assume *(iv)*. Let $\varepsilon > 0$ and $\phi_n : \mathscr{X}_n \to [0,1]$ be given. There exist $c > 0$ and $N \ge 1$ such that for all $n \ge N$,

$$Q_n \phi_n < c a_n^{-1} P_n \phi_n + \frac{\varepsilon}{2}.$$

If we assume that $P_n \phi_n = o(a_n)$ then there is a $N' \ge N$ such that $c a_n^{-1} P_n \phi_n < \varepsilon/2$ for all $n \ge N'$. Consequently, for every $\varepsilon > 0$, there exists an $N' \ge 1$ such that $Q_n \phi_n < \varepsilon$ for all $n \ge N'$. Conclude that $Q_n \lhd a_n^{-1} P_n$. To show that *(ii)* $\Rightarrow$ *(iv)*, let $\mu_n = P_n + Q_n$ and denote μ_n-densities for P_n, Q_n by $p_n, q_n : \mathscr{X}_n \to \mathbb{R}$. Then, for any $n \ge 1$, $c > 0$,

$$\begin{aligned} \left\| Q_n - Q_n \wedge c a_n^{-1} P_n \right\| &= \sup_{A \in \mathscr{B}_n} \left(\int_A q_n \, d\mu_n - \int_A q_n \, d\mu_n \wedge \int_A c a_n^{-1} p_n \, d\mu_n \right) \\ &\le \sup_{A \in \mathscr{B}_n} \int_A (q_n - q_n \wedge c a_n^{-1} p_n) \, d\mu_n \qquad (7.10) \\ &= \int 1\{q_n > c a_n^{-1} p_n\} (q_n - c a_n^{-1} p_n) \, d\mu_n. \end{aligned}$$

Note that the right-hand side of (7.10) is bounded above by $Q_n(dP_n/dQ_n < c^{-1} a_n)$. To show that *(iii)* $\Rightarrow$ *(iv)*, it is noted that, for all $c > 0$ and $n \ge 1$,

$$0 \le \int c a_n^{-1} P_n(q_n > c a_n^{-1} p_n) \le Q_n(q_n > c a_n^{-1} p_n) \le 1,$$

so (7.10) goes to zero if $\liminf_{n\to\infty} c\,a_n^{-1} P_n(dQ_n/dP_n > c\,a_n^{-1}) = 1$. To prove that *(v)* $\Leftrightarrow$ *(ii)*, note that Prohorov's theorem says that weak convergence of a subsequence within any subsequence of $a_n(dP_n/dQ_n)^{-1}$ under Q_n (see appendix A) is equivalent to the asymptotic tightness of $(a_n(dP_n/dQ_n)^{-1} : n \geq 1)$ under Q_n, *i.e.* for every $\varepsilon > 0$ there exists an $M > 0$ such that $Q_n(a_n(dP_n/dQ_n)^{-1} > M) < \varepsilon$ for all $n \geq 1$. This is equivalent to *(ii)*.

To conclude this section, we specify the definition of remote contiguity slightly further.

Definition 7.2.5. Given measurable spaces $(\mathscr{X}_n, \mathscr{B}_n)$, $(n \geq 1)$ with two sequences (P_n) and (Q_n) of probability measures and sequences $\rho_n, \sigma_n > 0$, $\rho_n, \sigma_n \to 0$, we say that Q_n is ρ_n-to-σ_n remotely contiguous with respect to P_n, notation $\sigma_n^{-1} Q_n \lhd \rho_n^{-1} P_n$, if,

$$P_n\phi_n(X^n) = o(\rho_n) \quad \Rightarrow \quad Q_n\phi_n(X^n) = o(\sigma_n),$$

for every sequence of $\mathscr{B}_n$-measurable $\phi_n : \mathscr{X}_n \to [0,1]$.

Like definition 7.2.1, definition 7.2.5 allows for reformulation similar to lemma 7.2.3, *e.g.* if for some sequences ρ_n, σ_n like in definition 7.2.5,

$$\left\| Q_n - Q_n \wedge \sigma_n \rho_n^{-1} P_n \right\| = o(\sigma_n),$$

then $\sigma_n^{-1} Q_n \lhd \rho_n^{-1} P_n$. We leave the formulation of other sufficient conditions to the reader.

Example 7.2.6. Inequality (7.9) in example 7.2.2 implies that $b_n^{-1} P_0^n \lhd a_n^{-1} P^n$, for any $a_n \leq \exp(-n\alpha^2)$ with $\alpha^2 > \frac{1}{2}\varepsilon^2$ and $b_n = \exp(-n(\alpha^2 - \frac{1}{2}\varepsilon^2))$. It is noted that this implies that $\phi_n(X^n) \xrightarrow{Q_n\text{-a.s.}} 0$ for any $\phi_n : \mathscr{X}_n \to [0,1]$ such that $P_n\phi_n(X^n) = o(\rho_n)$ (more generally, this holds whenever $\sum_n \sigma_n < \infty$, as a consequence of the *first Borel-Cantelli lemma* (lemma B.2.11)).

7.3 Remote contiguity for Bayesian limits

The relevant applications in the context of Bayesian limit theorems concern remote contiguity of the sequence of true distributions $P_{\theta_0,n}$ with respect to local prior predictive distributions $P_n^{\Pi_n|B_n}$, where the sets $B_n \subset \Theta$ are such that,

$$P_{\theta_0,n} \lhd a_n^{-1} P_n^{\Pi_n|B_n}, \tag{7.11}$$

for some rate $a_n \downarrow 0$. In the case of *i.i.d.* data, Barron [9] introduces strong and weak notions of *merging* of $P_{\theta_0,n}$ with (non-local) prior predictive distributions P_n^{Π}. The weak version imposes condition *(ii)* of lemma 7.2.3 for all exponential rates simultaneously. *Strong merging* (or *matching* [8]) coincides with Schwartz's almost-sure limit, while *Weak merging* (and weak matching) are viewed as limits in probability.

By contrast, if we have a specific rate a_n in mind, the relevant stochastic mode of convergence for remote contiguity is not almost-sure convergence or even convergence in probability, but convergence with respect to *Prokhorov's weak topology*: namely, according to lemma 7.2.3-*(v)*, (7.11) holds if inverse likelihood ratios Z_n have a weak limit Z when re-scaled by a_n,

$$Z_n = (dP_n^{\Pi_n|B_n}/dP_{\theta_0,n})^{-1}(X^n), \quad a_n Z_n \xrightarrow{P_{\theta_0,n}\text{-w.}} Z.$$

But condition (7.11) can also be written out, for example to the requirement that for some constant $\delta > 0$,

$$P_{\theta_0,n}\Big(\int \frac{dP_{\theta,n}}{dP_{\theta_0,n}}(X^n)\, d\Pi_n(\theta|B_n) < \delta\, a_n \Big) \to 0,$$

with the help of lemma 7.2.3-*(ii)*. This allows us to reformulate lemma 6.4.6 as follows.

Proposition 7.3.1. *Consider a model $\mathscr{P}$ of single-observation distributions P for i.i.d. data $(X_1, X_2, \ldots, X_n) \sim P^n$, $(n \geq 1)$, with priors (Π_n). Let $\varepsilon_n > 0$, $\varepsilon_n \downarrow 0$ and $P_0 \in \mathscr{P}$ be given and let $B_n = B(\varepsilon_n; P_0)$ be defined as in (6.13). Assuming $\Pi_n(B_n) > 0$, we have,*

$$P_0^n \lhd e^{-n\varepsilon_n^2(1+\delta)} P_n^{\Pi_n|B_n}$$

for any $\delta > 0$.

The next proposition should be viewed in light of Le Cam and Yang (1988) [181], which considers properties like contiguity, convergence of experiments and local asymptotic normality in situations of statistical information loss. To make the present case compatible, we think of (remote) contiguity for probability measures that arise as marginals for the data X^n when information concerning the (Bayesian random) parameter θ is unavailable.

Proposition 7.3.2. *Let $\theta_0 \in \Theta$ and a prior $\Pi : \mathscr{G} \to [0,1]$ be given. Let B be a measurable subset of Θ such that $\Pi(B) > 0$. Assume that for some $a_n \downarrow 0$, the family,*

$$\left\{ a_n \Big(\frac{dP_{\theta,n}}{dP_{\theta_0,n}} \Big)^{-1}(X^n) : \theta \in B, n \geq 1 \right\},$$

is uniformly tight under $P_{\theta_0,n}$. Then $P_{\theta_0,n} \lhd a_n^{-1} P_n^{\Pi|B}$.

Proof. For every $\varepsilon > 0$, there exists a constant $\delta > 0$ such that,

$$P_{\theta_0,n}\left(a_n \Big(\frac{dP_{\theta,n}}{dP_{\theta_0,n}} \Big)^{-1}(X^n) > \frac{1}{\delta} \right) < \varepsilon,$$

for all $\theta \in B$, $n \geq 1$. For this choice of δ, condition *(ii)* of lemma 7.2.3 is satisfied for all $\theta \in B$ simultaneously, and *c.f.* the proof of said lemma, for given $\varepsilon > 0$, there exists a $c > 0$ such that,

$$\|P_{\theta_0,n} - P_{\theta_0,n} \wedge c\, a_n^{-1} P_{\theta,n}\| < \varepsilon, \tag{7.12}$$

for all $\theta \in B$, $n \geq 1$. Now note that for any $A \in \mathscr{B}_n$,

$$\begin{aligned} 0 \leq & P_{\theta_0,n}(A) - P_{\theta_0,n}(A) \wedge c\, a_n^{-1} P_n^{\Pi|B}(A) \\ & \leq \int \big(P_{\theta_0,n}(A) - P_{\theta_0,n}(A) \wedge c\, a_n^{-1} P_{\theta,n}(A)\big) d\Pi(\theta|B). \end{aligned}$$

Taking the supremum with respect to A, we find the following inequality in terms of total variational norms,

$$\big\|P_{\theta_0,n} - P_{\theta_0,n} \wedge c\, a_n^{-1} P_n^{\Pi|B}\big\| \leq \int \big\|P_{\theta_0,n} - P_{\theta_0,n} \wedge c\, a_n^{-1} P_{\theta,n}\big\| d\Pi(\theta|B).$$

Since the total-variational norm is bounded and $\Pi(\cdot|B)$ is a probability measure, Fatou's lemma says that,

$$\begin{aligned} \limsup_{n\to\infty} & \big\|P_{\theta_0,n} - P_{\theta_0,n} \wedge c\, a_n^{-1} P_n^{\Pi|B}\big\| \\ & \leq \int \limsup_{n\to\infty} \big\|P_{\theta_0,n} - P_{\theta_0,n} \wedge c\, a_n^{-1} P_{\theta,n}\big\| d\Pi(\theta|B), \end{aligned}$$

and the *r.h.s.* equals zero *c.f.* (7.12). According to condition *(iv)* of lemma 7.2.3 this implies the assertion.

To re-establish contact with the notion of merging, note the following. If remote contiguity of the type (7.11) can be achieved for a sequence of subsets (B_n), then it also holds for any sequence of sets (*e.g.* all equal to Θ, in Barron's case) that contain the B_n but at a rate that differs proportionally to the fraction of prior masses.

Lemma 7.3.3. *For all $n \geq 1$, let $B_n \subset \Theta$ be such that $\Pi_n(B_n) > 0$ and C_n such that $B_n \subset C_n$ with $c_n = \Pi_n(B_n)/\Pi_n(C_n) \downarrow 0$, then,*

$$P_n^{\Pi_n|B_n} \lhd c_n^{-1}\, P_n^{\Pi_n|C_n}.$$

Also, if for some sequence (P_n), $P_n \lhd a_n^{-1}\, P_n^{\Pi_n|B_n}$ then $P_n \lhd a_n^{-1} c_n^{-1}\, P_n^{\Pi_n|C_n}$.

Proof. Fix $n \geq 1$. Because $B_n \subset C_n$, for every $A \in \mathscr{B}_n$, we have,

$$\int_{B_n} P_{\theta,n}(A)\, d\Pi(\theta) \leq \int_{C_n} P_{\theta,n}(A)\, d\Pi(\theta),$$

so $P_n^{\Pi_n|B_n}(A) \leq \Pi_n(C_n)/\Pi_n(B_n)\, P_n^{\Pi_n|C_n}(A)$. So if for some sequence $\phi_n : \mathscr{X}_n \to [0,1]$, we have $P_n^{\Pi_n|C_n}\phi_n(X^n) = o(\Pi_n(B_n)/\Pi_n(C_n))$, then the $P_n^{\Pi_n|B_n}$-expectations of $\phi_n(X^n)$ are $o(1)$, proving the first claim. If $P_n^{\Pi_n|C_n}\phi_n(X^n) = o(a_n \Pi_n(B_n)/\Pi_n(C_n))$, then $P_n^{\Pi_n|B_n}\phi_n(X^n) = o(a_n)$ and, hence, $P_n\phi_n(X^n) = o(1)$.

So when considering possible choices for the sequence (B_n), smaller choices lead to rates a_n that go to zero more slowly, rendering (7.7) applicable to more sequences

of test functions. This advantage is to be balanced against later requirements that $\Pi_n(B_n)$ may not decrease too fast.

7.3.1 Remote contiguity, examples in regression

To better understand the counterexamples of subsection 6.5.1, notice the high sensitivity of remote contiguity to the existence of subsets of the sample spaces assigned probability zero under some model distributions, while the true probability is non-zero. More generally, remote contiguity is sensitive to subsets E_n assigned fast-decreasing probabilities under local prior predictive distributions $P_n^{\Pi_n|B_n}(E_n)$, while the probabilities $P_{\theta_0,n}(E_n)$ remain high, which is what definition 7.2.1 expresses. The rate $a_n \downarrow 0$ helps to control the likelihood ratio (compare to the unscaled limits of likelihood ratios that play a central role in the theory of *convergence of statistical experiments* [179, 183, 242]), conceivably enough to force uniform tightness in many non-parametric situations.

To compare contiguity and its remote analogue in the context of (parametric and non-parametric) Bayesian regression, consider the following example.

Example 7.3.4. Let $\mathscr{F}$ denote a class of functions $\mathbb{R} \to \mathbb{R}$. We consider samples $X^n = ((X_1,Y_1),\ldots,(X_n,Y_n))$, $(n \geq 1)$ of points in $\mathbb{R}^2$, assumed to be related through,

$$Y_i = f_0(X_i) + e_i,$$

for some unknown $f_0 \in \mathscr{F}$, where the errors are *i.i.d.* standard normal $e_1,\ldots,e_n \sim N(0,1)^n$ and independent of the *i.i.d.* covariates $X_1,\ldots,X_n \sim P^n$, for some ancillary distribution P on $\mathbb{R}$. Assume that $\mathscr{F} \subset L^2(P)$ and that $Pf_0(X) = 0$ for all $f \in \mathscr{F}$. We use the L^2-norm $\|f\|_{P,2}^2 = \int f^2\,dP$ to define a metric d on $\mathscr{F}$, $d(f,g) = \|f-g\|_{P,2}$. Given a parameter $f \in \mathscr{F}$, denote the sample distributions as $P_{f,n}$. We distinguish two cases: *(a)* the case of linear regression, where $\mathscr{F} = \{f_\theta : \mathbb{R} \to \mathbb{R} : \theta \in \Theta\}$, where $\theta = (a,b) \in \Theta = \mathbb{R}^2$ and $f_\theta(x) = ax+b$; *(b)* the case of non-parametric regression, where we do not restrict $\mathscr{F}$ beforehand; and *(c)* a case where we replace the parameter f by non-parametric point-estimators $\hat{f}_n : \mathscr{X}_n \to L^2(P)$, like replacing a nuisance parameter by an estimate to obtain an approximate distribution (*e.g.* *profile likelihood*, see Murphy and van der Vaart (2000) [201]).

Let Π be a Borel prior Π on $\mathscr{F}$ and place remote contiguity in context by assuming, for the moment, that for some $\rho > 0$, there exist $0 < r < \rho$ and $\tau > 0$, as well as Bayesian tests ϕ_n for $B = \{f \in \mathscr{F} : \|f-f_0\|_{P,2} < r\}$ versus $V = \{f \in \mathscr{F} : \|f-f_0\|_{P,2} \geq \rho\}$ under Π of power $a_n = \exp(-\frac{1}{2}n\tau^2)$. If this is the case, we may assume that $r < \frac{1}{2}\tau$ without loss of generality. Suppose also that Π has a support in $L^2(P)$ that contains all of $\mathscr{F}$.

Example 7.3.5. Let us concentrate on case *(b)* first: a bit of manipulation casts the a_n-rescaled likelihood ratio for $f \in \mathscr{F}$ in the following form,

$$a_n^{-1}\frac{dP_{f,n}}{dP_{f_0,n}}(X^n) = a_n^{-1}\prod_{i=1}^n \frac{e^{-\frac{1}{2}(Y_i-f(X_i))^2}}{e^{-\frac{1}{2}(Y_i-f_0(X_i))^2}} = e^{-\frac{1}{2}\sum_{i=1}^n \left(2e_i(f-f_0)(X_i)+(f-f_0)^2(X_i)-\tau^2\right)}, \tag{7.13}$$

under $X^n \sim P_{f_0,n}$. The exponent is controlled by the *law of large numbers*,

$$\frac{1}{n}\sum_{i=1}^n \left(2e_i(f-f_0)(X_i)+(f-f_0)^2(X_i)-\tau^2\right) \xrightarrow{P_{f_0}^\infty\text{-a.s.}} \|f-f_0\|_{P,2}^2-\tau^2.$$

Hence, for every $\varepsilon > 0$ there exists an $N(f,\varepsilon) \geq 1$ such that the exponent in (7.13) satisfies the upper bound,

$$\sum_{i=1}^n \left(2e_i(f-f_0)(X_i)+(f-f_0)^2(X_i)-\tau^2\right) \leq n(\|f-f_0\|_{P,2}^2-\tau^2+\varepsilon^2),$$

for all $n \geq N(f,\varepsilon)$. Since $\Pi(B) > 0$, we may condition Π on B, choose $\varepsilon = \frac{1}{2}\tau$ and use Fatou's inequality to find that,

$$\liminf_{n\to\infty} e^{\frac{1}{2}n\tau^2}\frac{dP_n^{\Pi|B}}{dP_{f_0,n}}(X^n) \geq \liminf_{n\to\infty} e^{\frac{1}{4}n\tau^2} = \infty,$$

$P_{f_0}^\infty$-almost-surely. Consequently, for any choice of δ,

$$P_{f_0,n}\left(\frac{dP_n^{\Pi|B}}{dP_{f_0,n}}(X^n) < \delta\, e^{-\frac{1}{2}n\tau^2}\right) \to 0,$$

and we conclude that $P_{f_0,n} \lhd e^{\frac{1}{2}n\tau^2} P_n^{\Pi|B}$.

Example 7.3.6. To analyse case *(c)* next, we consider rates $a_n = \exp(-\frac{1}{2}n\tau_n^2)$ in remote-contiguity statement (7.13) with $f = \hat{f}_n(X^n)$. Assume that the estimators $\hat{f}_n$ are $L^2(P)$-consistent at rate ε_n, *i.e.* $\varepsilon_n^{-2}\|\hat{f}_n - f_0\|_{P,2}^2 = O_P(1)$. Also assume that,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \left(e_i(\hat{f}_n-f_0)(X_i)\right) = O_{P_{f_0,n}}(1),$$

and,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \left((\hat{f}_n-f_0)^2(X_i) - \|\hat{f}_n-f_0\|_{P,2}^2\right) = O_{P_{f_0,n}}(1),$$

See subsection 3.4.3 (and particularly 3.4.3.2) in [246]; through the use of *maximal inequalities*, the latter discusses in great detail how smoothness assumptions on the space $\mathscr{F}$ allow verification of uniform versions of these convergence statements, which imply the versions above. For example, if $\mathscr{F}$ is a bounded subset of $C^\alpha[0,1]^d$, the space of all α-times differentiable functions on $[0,1]^d$ (with the Hölder norm), then the least-squares estimator $\hat{f}_n$ converges to f_0 at L^2-rate,

$$\varepsilon_n = n^{\frac{\alpha}{2\alpha+d}}.$$

For any τ such that $\varepsilon_n = o(\tau_n)$, we then find that,

$$P_{f_0,n} \lhd e^{\frac{1}{2}n\tau_n^2} P_{\hat{f}_n,n},$$

demonstrating that remote contiguity also applies where the approximation of one sequence by another is required, with possible application outside of Bayesian limits.

Example 7.3.7. As for case *(a)*, one has the choice of using a prior like above, but also to proceed differently: expression (7.13) can be written in terms of a local parameter $h \in \mathbb{R}^k$ which, for given θ_0 and $n \geq 1$, is related to θ by $\theta = \theta_0 + n^{-1/2}h$. For $h \in \mathbb{R}^2$, we write $P_{h,n} = P_{\theta_0+n^{-1/2}h,n}$, $P_{0,n} = P_{\theta_0,n}$ and rewrite the likelihood ratio (7.13) as follows,

$$\frac{dP_{h,n}}{dP_{0,n}}(X^n) = e^{\frac{1}{\sqrt{n}}\sum_{i=1}^n h\cdot\ell_{\theta_0}(X_i,Y_i) - \frac{1}{2}h\cdot I_{\theta_0}\cdot h + R_n}, \tag{7.14}$$

where $\ell_{\theta_0} : \mathbb{R}^2 \to \mathbb{R}^2 : (x,y) \mapsto (y - a_0x - b_0)(x,1)$ is the score function for θ, $I_{\theta_0} = P_{\theta_0,1}\ell_{\theta_0}\ell_{\theta_0}^T$ is the Fisher information matrix and $R_n \xrightarrow{P_{\theta_0,n}} 0$. Assume that I_{θ_0} is non-singular and note the central limit,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \ell_{\theta_0}(X_i,Y_i) \xrightarrow{P_{\theta_0,n}\text{-w.}} N_2(0, I_{\theta_0}),$$

which expresses local asymptotic normality of the model, *c.f.* definition 4.1.12, and implies that for any fixed $h \in \mathbb{R}^2$, $P_{h,n} \lhd P_{0,n}$, proving remote contiguity at any rate.

Clearly a proof of contiguity puts requirements on the likelihood of a relatively stringent nature compared to the requirements posed by remote contiguity. The LAN example relies on quite subtle argumentation that is natural in parametric context, but cannot be expected to generalise to the same powerful extent in non-parametric setting (notwithstanding successes in semi-parametric statistics). In non-parametric cases, a less delicate argument is required and remote contiguity appears to provide it in an accessible way (through relatively straightforward analysis of weak limits of a_n-re-scaled inverse likelihood ratios).

Remote contiguity also applies in more irregular situations: example 6.5.3 does not admit KL priors, but satisfies the requirement of remote contiguity. (Choose η equal to the uniform density for simplicity.)

Example 7.3.8. Consider again the model of example 6.5.3, where we observe an *i.i.d.* sample from the uniform distribution on $[\theta, \theta+1]$, for unknown $\theta \in \mathbb{R}$. The model is parametrized in terms of distributions P_θ with Lebesgue densities of the form $p_\theta(x) = 1_{[\theta,\theta+1]}(x)$, for $\theta \in \Theta = \mathbb{R}$. Pick a prior Π on Θ with a continuous and strictly positive Lebesgue density $\pi : \mathbb{R} \to \mathbb{R}$ and, for some rate $\delta_n \downarrow 0$, choose

$B_n = (\theta_0, \theta_0 + \delta_n)$. Note that for any $\alpha > 0$, there exists an $N \geq 1$ such that for all $n \geq N$, $(1-\alpha)\pi(\theta_0)\delta_n \leq \Pi(B_n) \leq (1+\alpha)\pi(\theta_0)\delta_n$. Note that for any $\theta \in B_n$ and $X^n \sim P^n_{\theta_0}$, $dP^n_\theta / dP^n_{\theta_0}(X^n) = 1\{X_{(1)} > \theta\}$, and correspondingly,

$$\begin{aligned} \frac{dP_n^{\Pi|B_n}}{dP^n_{\theta_0}}(X^n) &= \Pi_n(B_n)^{-1} \int_{\theta_0}^{\theta_0+\delta_n} 1\{X_{(1)} > \theta\}\, d\Pi(\theta) \\ &\geq \frac{1-\alpha}{1+\alpha} \frac{\delta_n \wedge (X_{(1)} - \theta_0)}{\delta_n}, \end{aligned}$$

for large enough n. As a consequence, for every $\delta > 0$ and all $a_n \downarrow 0$,

$$P^n_{\theta_0}\left(\frac{dP_n^{\Pi|B_n}}{dP^n_{\theta_0}}(X^n) < \delta\, a_n \right) \leq P^n_{\theta_0}\big(\delta_n^{-1}(X_{(1)} - \theta_0) < (1+\alpha)\delta\, a_n \big),$$

for large enough $n \geq 1$. Since $n(X_{(1)} - \theta_0)$ has an exponential weak limit under $P^n_{\theta_0}$, we choose $\delta_n = n^{-1}$, so that the *r.h.s.* in the above display goes to zero. So $P_{\theta_0,n} \lhd a_n^{-1} P_n^{\Pi_n|B_n}$, for any $a_n \downarrow 0$. Conclude that with these choices for Π and B_n, (7.11) holds, for any a_n.

7.4 Posterior concentration

In this section new frequentist theorems are formulated involving the convergence of posterior distributions. First we give a basic proof for posterior consistency assuming existence of suitable test sequences and remote contiguity of true distributions $(P_{\theta_0,n})$ with respect to local prior predictive distributions. Then it is not difficult to extend the proof to the case of posterior rates of convergence in metric topologies. With the same methodology it is possible to address questions in Bayesian hypothesis testing and model selection: if a Bayesian test to distinguish between two hypotheses exists and remote contiguity applies, frequentist consistency of the Bayes Factor can be guaranteed. We conclude with a theorem that uses remote contiguity to describe a general relation that exists between credible sets and confidence sets, provided the prior induces remotely-contiguous local prior predictive distributions.

We start with posterior consistency, *c.f.* definition 6.1.1 and proposition 6.1.2. The formulation is has the generality of remark A.0.1 and the theorem applies to non-*i.i.d.* data, and with n-dependent models and priors.

Theorem 7.4.1. *Assume that for all $n \geq 1$, the data $X^n \sim P_{\theta_0,n}$ for some $\theta_0 \in \Theta$. Fix a prior $\Pi : \mathscr{G} \to [0,1]$ and assume that for given $B, V \in \mathscr{G}$ with $\Pi(B) > 0$ and $a_n \downarrow 0$,*

(i) there exist Bayesian tests ϕ_n for B versus V,

$$\int_B P_{\theta,n}\phi_n\, d\Pi(\theta) + \int_V P_{\theta',n}(1-\phi_n)\, d\Pi(\theta') = o(a_n), \tag{7.15}$$

(ii) the sequence $P_{\theta_0,n}$ satisfies $P_{\theta_0,n} \lhd a_n^{-1} P_n^{\Pi|B}$.

Then $\Pi(V|X^n) \xrightarrow{P_{\theta_0,n}} 0$.

Proof. Choose $B_n = B$, $V_n = V$ and use proposition 7.1.3 to see that $P_n^{\Pi|B}\Pi(V|X^n)$ is upper bounded by $\Pi(B)^{-1}$ times the *l.h.s.* of (7.15) and, hence, is of order $o(a_n)$. Condition *(ii)* then implies that $P_{\theta_0,n}\Pi(V|X^n) = o(1)$, which is equivalent to $\Pi(V|X^n) \xrightarrow{P_{\theta_0,n}} 0$ since $0 \leq \Pi(V|X^n) \leq 1$, $P_{\theta_0,n}$-almost-surely, for all $n \geq 1$.

These conditions may be interpreted as follows: theorem 9.5.1 lends condition *(i)* a distinctly Bayesian interpretation: it requires a Bayesian test to set V apart from B with testing power a_n. Lemma 7.1.2 translates this into the (still Bayesian) statement that the posteriors for V go to zero in $P_n^{\Pi|B}$-expectation. Condition *(ii)* is there to promote this Bayesian point to a frequentist one through (7.7).

One of the first questions we have, is how Freedman's inconsistent posteriors relate to the above. Since test sequences of exponential power exist to separate complements of weak neighbourhoods, *c.f.* proposition A.0.6, Freedman's inconsistencies must violate the requirement of remote contiguity in theorem 7.4.1.

Example 7.4.2. As noted already, the space Λ of examples 1.1.4, 2.1.18 and 6.5.1 is a Polish space; in particular Λ is metric and second countable, so the subspace $\mathscr{N}$ contains a countable dense subset D. For $Q \in D$, let V be the set of all prior probability measures on Λ with finite support, of which one point is Q and the remaining points lie in Λ_0. The proof of the theorem in [98] that asserts that the set of consistent pairs (P_0, Π) is of the first category in $\Lambda \times \pi(\Lambda)$ departs from the observation that if P_0 lies in $\mathscr{N}$ and we use a prior from V, then,

$$\Pi(\{Q\}|X^n) \xrightarrow{P_0\text{-a.s.}} 1,$$

(in fact, as is shown below, with P_0^∞-probability one there exists an $N \geq 1$ such that $\Pi(\{Q\}|X^n) = 1$ for all $n \geq N$). The proof continues to assert that V lies dense in $\pi(\Lambda)$, and, through sequences of continuous extensions involving D, that posterior inconsistency for elements of V implies posterior inconsistency for all Π in $\pi(\Lambda)$ with the possible exception of a set of the first category.

From the present perspective it is interesting to view the inconsistency of elements of V in light of the conditions of theorem 7.4.1. Define, for some bounded $f : \mathbb{N} \to \mathbb{R}$ and $\varepsilon > 0$, two subsets of Λ,

$$B = \{P : |Pf - P_0 f| < \tfrac{1}{2}\varepsilon\}, \quad V = \{P : |Pf - P_0 f| \geq \varepsilon\}.$$

Proposition A.0.6 asserts the existence of a uniform test sequence for B versus V of exponential power. With regard to remote contiguity, for an element Π of V with support of order $M+1$, write,

$$\Pi = \beta\,\delta_Q + \sum_{m=1}^{M} \alpha_m \delta_{P_m},$$

where $\beta + \sum_m \alpha_m = 1$ and $P_m \in \Lambda_0$ $(1 \leq m \leq M)$. Without loss of generality, assume that ε and f are such that Q does not lie in B. Consider,

$$\frac{dP_n^{\Pi|B}}{dP_0^n}(X^n) = \frac{1}{\Pi(B)} \int_B \frac{dP^n}{dP_0^n}(X^n)\, d\Pi(P) \leq \frac{1}{\Pi(B)} \sum_{m=1}^{M} \alpha_m \frac{dP_m^n}{dP_0^n}(X^n).$$

For every $1 \leq m \leq M$, there exists a $k(m)$ such that $P_m(X = k(m)) = 0$, and the probability of the event E_n that none of the $X_1, \ldots, X_n$ equal $k(m)$ is $(1 - P_0(X = k(m)))^n$. Note that E_n is also the event that $dP_m^n/dP_0^n(X^n) > 0$.

Hence for every $1 \leq m \leq M$ and all X in an event of P_0^∞-probability one, there exists an $N_m \geq 1$ such that $dP_m^n/dP_0^n(X^n) = 0$ for all $n \geq N_m$. Consequently, for all X in an event of P_0^∞-probability one, there exists an $N \geq 1$ such that $dP_n^{\Pi|B}/dP_0^n(X^n) = 0$ for all $n \geq N$. Therefore, condition *(ii)* of lemma 7.2.3 is not satisfied for any sequence $a_n \downarrow 0$. A direct proof that (7.7) does not hold for any a_n is also possible: given the prior $\Pi \in V$, define,

$$\phi_n(X^n) = \prod_{m=1}^{M} 1_{\{\exists_{1 \leq i \leq n} : X_i = k(m)\}}.$$

Then the expectation of ϕ_n with respect to the local prior predictive distribution equals zero, so $P_n^{\Pi|B}\phi_n = o(a_n)$ for any $a_n \downarrow 0$. However, $P_0^n \phi_n(X^n) \to 1$, so the prior Π does *not* give rise to a sequence of prior predictive distributions $(P_n^{\Pi|B})$ with respect to which (P_0^n) is remotely contiguous, for any $a_n \downarrow 0$.

A proof of a theorem very close to Schwartz's theorem is now possible. Consider condition *(i)* of theorem 6.3.1: a well-known argument based on Hoeffding's inequality guarantees the existence of a uniform test sequence *of exponential power* whenever a uniform test sequence test sequence exists, so Schwartz equivalently assumes that there exists a $D > 0$ such that,

$$P_0^n \phi_n + \sup_{Q \in \mathscr{P} \setminus U} Q^n(1 - \phi_n) = o(e^{-nD}). \tag{7.16}$$

We vary slightly and assume the existence of a Bayesian test sequence of exponential power. In the following theorem, let $\mathscr{P}$ denote a Hausdorff space of single-observation distributions on $(\mathscr{X}, \mathscr{B})$ with Borel prior Π.

Corollary 7.4.3. *For all $n \geq 1$, let $(X_1, X_2, \ldots, X_n) \sim P_0^n$ for some $P_0 \in \mathscr{P}$. Let U denote an open neighbourhood of P_0 and define $K(\varepsilon) = \{P \in \mathscr{P} : -P_0 \log(dP/dP_0) < \varepsilon^2\}$. If,*

(i) there exist $\varepsilon > 0$, $D > 0$ and a sequence of measurable $\psi_n : \mathscr{X}^n \to [0,1]$, such that,

$$\int_{K(\varepsilon)} P^n \psi_n \, d\Pi(P) + \int_{\mathscr{P} \setminus U} Q^n(1 - \psi_n)\, d\Pi(Q) = o(e^{-nD}),$$

(ii) and $\Pi(K(\varepsilon)) > 0$ for all $\varepsilon > 0$,

then $\Pi(U|X^n) \xrightarrow{P_0\text{-}a.s.} 1$.

Proof. A prior Π satisfying condition *(ii)* guarantees that $P_0^n \ll P_n^\Pi$ for all $n \geq 1$, *c.f.* the remark preceding proposition A.0.7. Choose ε such that $\varepsilon^2 < D$. Recall that for every $P \in K(\varepsilon)$, the exponential lower bound (7.8) for likelihood ratios of dP^n/dP_0^n exists. Hence the limes inferior of $\exp(\frac{1}{2}n\varepsilon^2)(dP^n/dP_0^n)(X^n)$ is greater than or equal to one with P_0^∞-probability one. Then, with the use of Fatou's lemma and the assumption that $\Pi(K(\varepsilon)) > 0$,

$$\liminf_{n\to\infty} \frac{e^{nD}}{\Pi(K(\varepsilon))} \int_{K(\varepsilon)} \frac{dP_\theta^n}{dP_{\theta_0}^n}(X^n)\, d\Pi(\theta) \geq 1,$$

with $P_{\theta_0}^\infty$-probability one, showing that sufficient condition *(ii)* of lemma 7.2.3 holds. Conclude that,

$$P_0^n \lhd e^{nD} P_n^{\Pi|K(\varepsilon)},$$

and use theorem 7.4.1 to see that $\Pi(U|X^n) \xrightarrow{P_{\theta_0,n}} 1$.

Example 7.4.4. As an example of the tests required under condition *(i)* of corollary 7.4.3, consider $\mathscr{P}$ in the Hellinger topology, assuming totally-boundedness. Let U be the Hellinger-ball of radius 4ε around P_{θ_0} of example 7.4.6 and let V be its complement. The Hellinger ball $B_H(\varepsilon)$ in equation (7.18) contains the set $K(\varepsilon)$. Alternatively we may consider the model in any of the weak topologies $\mathscr{T}_n$: let $\varepsilon > 0$ be given and let U denote a weak neighbourhood of the form $\{P \in \mathscr{P} : |(P^n - P_0^n)f| \geq 2\varepsilon\}$, for some bounded measurable $f : \mathscr{X}_n \to [0,1]$, as in proposition A.0.6. The set B of proposition A.0.6 contains a set $K(\delta)$, for some $\delta > 0$. Both these applications were noted by Schwartz in [226].

7.4.1 Consistency of Bayesian point estimators

As we know from proposition 6.1.2, if Θ is a Hausdorff, completely regular space, the posterior is consistent at $\theta_0 \in \Theta$, if and only if,

$$\int f(\theta)\, d\Pi(\theta|X^n) \xrightarrow{P_{\theta_0,n}} f(\theta_0),$$

for every bounded, continuous $f : \Theta \to \mathbb{R}$. Proposition 6.1.2 is used to prove consistency of frequentist point-estimators derived from the posterior, more generally than before in subsection 6.1.2.

Example 7.4.5. Consider a model $\mathscr{P}$ of single-observation distributions P on $(\mathscr{X}, \mathscr{B})$ for *i.i.d.* data $(X_1, X_2, \ldots, X_n) \sim P^n$, $(n \geq 1)$. Assume that the true distribution of the data is $P_0 \in \mathscr{P}$ and that the model topology is Prohorov's weak topology or stronger. Then for any bounded, continuous $g : \mathscr{X} \to \mathbb{R}$, the map,

$$f : \mathscr{P} \to \mathbb{R} : P \mapsto \big|(P - P_0)g(X)\big|,$$

is continuous. Assuming that the posterior is weakly consistent at P_0,

$$\big|P_1^{\Pi_n|X^n} g - P_0 g\big| \leq \int \big|(P - P_0)g\big|\, d\Pi(P|X^n) \xrightarrow{P_{\theta_0}} 0, \tag{7.17}$$

so posterior predictive distributions are consistent point estimators in Prohorov's weak topology. Replacing the maps g by bounded, measurable maps $\mathscr{X} \to \mathbb{R}$ and assuming posterior consistency in $\mathscr{T}_1$, one proves consistency of posterior predictive distributions in $\mathscr{T}_1$ in exactly the same way. Taking the supremum over measurable $g : \mathscr{X} \to [0,1]$ in (7.17) and assuming that the posterior is consistent in the total variational topology, posterior predictive distributions are consistent in total variation as frequentist point estimators.

7.4.2 Posterior concentration and Hellinger entropy

Referring to the convexity requirement in proposition **??** on minimax tests, it is noted that questions concerning consistency require the existence of tests in which at least one of the two hypotheses is a non-convex set, typically the complement of a neighbourhood. Imposing the model $\mathscr{P}$ to be of bounded entropy with respect to the Hellinger metric allows construction of such tests, based on the uniform tests of proposition **??**. Below, we apply well-known constructions for the uniform tests in Schwartz's theorem from the frequentist literature [177, 32, 33, 106] to the construction of Bayesian tests. Due to relations that exist between metrics for model parameters and the Hellinger metric in many examples and applications, the material covered here is widely applicable in (non-parametric) models for *i.i.d.* data.

Example 7.4.6. Consider a model $\mathscr{P}$ of distributions P for *i.i.d.* data $X^n \sim P^n$, $(n \geq 1)$ and, in addition, suppose that $\mathscr{P}$ is totally bounded with respect to the Hellinger distance. Let $P_0 \in \mathscr{P}$ and $\varepsilon > 0$ be given, denote $V(\varepsilon) = \{P \in \mathscr{P} : H(P_0, P) \geq 4\varepsilon\}$, $B_H(\varepsilon) = \{P \in \mathscr{P} : H(P_0, P) < \varepsilon\}$. There exists an $N(\varepsilon) \geq 1$ and a cover of $V(\varepsilon)$ by H-balls $V_1, \ldots, V_{N(\varepsilon)}$ of radius ε and for any point Q in any V_i and any $P \in B_H(\varepsilon)$, $H(Q,P) > 2\varepsilon$. According to proposition 7.1.6 with $\alpha = 1/2$ and (7.5), for each $1 \leq i \leq N(\varepsilon)$ there exists a Bayesian test sequence $(\phi_{i,n})$ for $B_H(\varepsilon)$ versus V_i of power (upper bounded by) $\exp(-2n\varepsilon^2)$. Then, for any subset $B' \subset B_H(\varepsilon)$,

$$
\begin{aligned}
P_n^{\Pi|B'}\Pi(V|X^n) &\le \sum_{i=1}^{N(\varepsilon)} P_n^{\Pi|B'}\Pi(V_i|X^n) \\
&\le \frac{1}{\Pi(B')}\sum_{i=1}^{N(\varepsilon)}\Big(\int_{B'} P^n\phi_n\,d\Pi(P)+\int_{V_i}P^n(1-\phi_n)\,d\Pi(P)\Big) \qquad (7.18)\\
&\le \sum_{i=1}^{N(\varepsilon)}\sqrt{\frac{\Pi(V_i)}{\Pi(B')}}\exp(-2n\varepsilon^2),
\end{aligned}
$$

which is smaller than or equal to $e^{-n\varepsilon^2}$ for large enough n.

To balance entropy and prior mass differently in Hellinger separable models, Barron (1988) [9] and Barron *et al.* (1999) [14] formulate an alternative condition that is based on the Radon property that any prior on a Polish space has.

Example 7.4.7. Consider a model $\mathscr{P}$ of distributions P for *i.i.d.* data $X^n \sim P^n$, $(n \ge 1)$, with priors (Π_n). Assume that the model $\mathscr{P}$ is Polish in the Hellinger topology. Let $P_0 \in \mathscr{P}$ and $\varepsilon > 0$ be given; for a fixed $M > 1$, define $V = \{P \in \mathscr{P} : H(P_0,P) \ge M\varepsilon\}$, $B_H = \{P \in \mathscr{P} : H(P_0,P) < \varepsilon\}$. For any sequence $\delta_m \downarrow 0$, there exist compacta $K_m \subset \mathscr{P}$ such that $\Pi(K_m) \ge 1-\delta_m$ for all $m \ge 1$. For each $m \ge 1$, K_m is Hellinger totally bounded so there exists a Bayesian test sequence $\phi_{m,n}$ for $B_H(\varepsilon)\cap K_m$ versus $V(\varepsilon)\cap K_m$. Since,

$$
\begin{aligned}
&\int_{B_H} P^n\phi_n\,d\Pi(P)+\int_V Q^n(1-\phi_n)\,d\Pi(Q)\\
&\qquad\le \int_{B_H\cap K_m} P^n\phi_{m,n}\,d\Pi(P) \quad + \int_{V\cap K_m} Q^n(1-\phi_{m,n})\,d\Pi(Q)+\delta_m,
\end{aligned}
$$

and all three terms go to zero, a diagonalization argument confirms the existence of a Bayesian test for B_H versus V. To control the power of this test and to generalise to the case where $\varepsilon = \varepsilon_n$ is n-dependent, more is required: as we increase m with n, the prior mass $\delta_{m(n)}$ outside of $K_n = K_{m(n)}$ must decrease fast enough, while the order of the cover must be bounded: if $\Pi_n(K_n) \ge 1-\exp(-L_1 n\varepsilon_n^2)$ and the Hellinger entropy of K_n satisfies $\log N(\varepsilon_n, K_n, H) \le L_2 n\varepsilon_n^2$ for some $L_1, L_2 > 0$, there exist $M > 1$, $L > 0$, and a sequence of tests (ϕ_n) such that,

$$
\int_{B_H(\varepsilon_n)} P^n\phi_n\,d\Pi(P)+\int_{V(\varepsilon_n)} Q^n(1-\phi_n)\,d\Pi(Q) \le e^{-Ln\varepsilon_n^2},
$$

for large enough n. (For related constructions, see Barron (1988) [9], Barron *et al.* (1999) [14] and Ghosal, Ghosh and van der Vaart (2000) [106].)

7.5 Rates of posterior concentration

A significant extension to the theory on posterior convergence is formed by results concerning posterior convergence in metric spaces *at a rate*. Minimax rates of convergence for (estimators based on) posterior distributions were considered more or less simultaneously in Ghosal-Ghosh-van der Vaart [106] and Shen-Wasserman [231]. Both propose an extension of Schwartz's theorem to posterior rates of convergence [106, 231] and apply Barron's sieve idea with a well-known entropy argument [32, 33] to a shrinking sequence of Hellinger neighbourhoods and employs a more specific, rate-related version of the Kullback-Leibler condition (6.5) for the prior. Both appear to be inspired by contemporary results regarding Hellinger rates of convergence for sieve MLE's, as well as on Barron-Schervish-Wasserman [14], which concerns posterior consistency based on controlled bracketing entropy for a sieve, up to subsets of negligible prior mass, following ideas that were first laid down in [9]. It is remarked already in [14] that their main theorem is easily re-formulated as a rate-of-convergence theorem, with reference to [231]. More recently, Walker, Lijoi and Prünster [255] have added to these considerations with a theorem for Hellinger rates of posterior concentration in models that are separable for the Hellinger metric, with a central condition that calls for summability of square-roots of prior masses of covers of the model by Hellinger balls, based on analogous consistency results in Walker [252]. More recent is [156], which shows that alternative, less stringent versions of the prior conditions of [106, 231] exist, if one is willing to be more specific about model conditions.

Here we apply Bayesian testability and remote contiguity conditions to prove (frequentist) posterior convergence at a rate.

Theorem 7.5.1. *Assume that for all $n \geq 1$, the data $X^n \sim P_{\theta_0,n}$ for some $\theta_0 \in \Theta$. Fix priors $\Pi_n : \mathscr{G} \to [0,1]$ and assume that for given $B_n, V_n \in \mathscr{G}$ with $\Pi_n(B_n) > 0$ and $a_n, b_n \downarrow 0$ such that $a_n = o(b_n)$,*

(i) there are Bayesian tests $\phi_n : \mathscr{X}_n \to [0,1]$ such that,

$$\int_{B_n} P_{\theta,n}\phi_n \, d\Pi_n(\theta) + \int_{V_n} P_{\theta,n}(1-\phi_n) \, d\Pi_n(\theta) = o(a_n), \tag{7.19}$$

(ii) The prior mass of B_n is lower-bounded, $\Pi_n(B_n) \geq b_n$,
(iii) The sequence $P_{\theta_0,n}$ satisfies $P_{\theta_0,n} \lhd b_n a_n^{-1} P_n^{\Pi_n|B_n}$.

Then $\Pi(V_n|X^n) \xrightarrow{P_{\theta_0,n}} 0$.

Proof. Proposition 7.1.3 says that $P_n^{\Pi_n|B_n}\Pi(V_n|X^n)$ is of order $o(b_n^{-1}a_n)$. Condition *(iii)* then implies that $P_{\theta_0,n}\Pi(V_n|X^n) = o(1)$, which is equivalent to $\Pi(V_n|X^n) \xrightarrow{P_{\theta_0,n}} 0$ since $0 \leq \Pi(V_n|X^n) \leq 1$, $P_{\theta_0,n}$-almost-surely for all $n \geq 1$.

To connect with the literature we interpret lower bound (7.20) again, reformulating lemma 6.4.6 as a statement of remote contiguity.

Lemma 7.5.2. *For all $n \geq 1$, assume that $(X_1, X_2, \ldots, X_n) \in \mathscr{X}^n \sim P_0^n$ for some $P_0 \in \mathscr{P}$ and let $\varepsilon_n \downarrow 0$ be given. Let B_n be as in example 7.5.4. Then, for any priors Π_n such that $\Pi_n(B_n) > 0$,*

$$P_{\theta_0,n}\left(\int \frac{dP_\theta^n}{dP_{\theta_0}^n}(X^n)\, d\Pi_n(\theta|B_n) < e^{-cn\varepsilon_n^2} \right) \to 0,$$

for any constant $c > 1$.

Example 7.5.3. To apply theorem 7.5.1, consider again the situation of a uniform distribution with an unknown location, as in examples 6.5.3 and 7.3.8. Take V_n equal to $\{\theta : \theta - \theta_0 > \varepsilon_n\}$ $\{\theta : \theta_0 - \theta > \varepsilon_n\}$ respectively, with $\varepsilon_n = M_n/n$ for any $M_n \to \infty$. It is noted that, for every $0 < c < 1$, the likelihood ratio test,

$$\phi_n(X^n) = 1\{dP_{\theta_0+\varepsilon_n,n}/dP_{\theta_0,n}(X^n) > c\} = 1\{X_{(1)} > \theta_0 + \varepsilon_n\},$$

satisfies $P_\theta^n(1-\phi_n)(X^n) = 0$ for all $\theta \in V_n$, and if we choose $\delta_n = 1/2$ and $\varepsilon_n = M_n/n$ for some $M_n \to \infty$, $P_\theta^n \phi_n \leq e^{-M_n+1}$ for all $\theta \in B_n$, so that,

$$\int_{B_n} P_\theta^n \phi_n(\, d\Pi(\theta) + \int_{V_n} P_\theta^n(1-\phi_n)\, d\Pi(\theta) \leq \Pi(B_n)\, e^{-M_n+1},$$

Using lemma 7.1.2, we see that $P_n^{\Pi|B_n}\Pi(V_n|X^n) \leq e^{-M_n+1}$. Based on the conclusion of example **??** above, remote contiguity implies that $P_{\theta_0}^n \Pi(V_n|X^n) \to 0$. Treating the case $\theta < \theta_0 - \varepsilon_n$ similarly, we conclude that the posterior is consistent at (any ε_n slower than) rate $1/n$.

Example 7.5.4. Let us briefly review the conditions of [14, 106, 231] in light of theorem 7.5.1: let $\varepsilon_n \downarrow 0$ denote the Hellinger rate of convergence we have in mind, let $M > 1$ be some constant and define,

$$V_n = \{P \in \mathscr{P} : H(P, P_0) \geq M\varepsilon_n\},$$
$$B_n = \{P \in \mathscr{P} : -P_0 \log dP/dP_0 < \varepsilon_n^2,\, P_0 \log^2 dP/dP_0 < \varepsilon_n^2\}.$$

If $\varepsilon = \varepsilon_n$ with $\varepsilon_n \downarrow 0$ and $n\varepsilon_n^2 \to \infty$, and the model's Hellinger entropy is upper-bounded by $\log N(\varepsilon_n, \mathscr{P}, H) \leq Kn\varepsilon_n^2$ for some $K > 0$, the construction of example 7.4.6 extends to tests that separate $V_n = \{P \in \mathscr{P} : H(P_0, P) \geq 4\varepsilon_n\}$ from $B_n = \{P \in \mathscr{P} : H(P_0, P) < \varepsilon_n\}$ asymptotically, with power $\exp(-nL\varepsilon_n^2)$ for some $L > 0$. (See also the so-called *Le Cam dimension* of a model [177] and Birgé's rate-oriented work [32, 33].) It is worth pointing out at this stage that posterior inconsistency due to the phenomenon of 'data tracking' [14, 254], whereby weak posterior consistency holds but Hellinger consistency fails, can only be due to failure of the testing condition in the Hellinger case.

Note that B_n is contained in the Hellinger ball of radius ε_n around P_0, so (7.19) holds. New in [106, 231] is the condition for the priors Π_n,

$$\Pi_n(B_n) \geq e^{-Cn\varepsilon_n^2}, \tag{7.20}$$

for some $C > 0$. With the help of lemmas 7.5.2 and 7.2.3-*(ii)*, we conclude that,

$$P_0^n \lhd e^{cn\varepsilon_n^2} P_n^{\Pi|B_n}, \tag{7.21}$$

for any $c > 1$. If we choose M such that $DM^2 - C > 1$, theorem 7.5.1 proves that $\Pi(V_n|X^n) \xrightarrow{P_0} 0$, *i.e.* the posterior is Hellinger consistent at rate ε_n.

Note that the argument also extends to models that are Hellinger separable: in that case (7.18) remains valid, but with $N(\varepsilon) = \infty$. The mass fractions $\Pi(V_i)/\Pi(B')$ become important (we point to strong connections with Walker's theorem [252, 255]). Here we see the balance between prior mass and testing power for Bayesian tests, as intended by the remark that closes the subsection on the existence of Bayesian test sequences in section 7.1.

Certain (simple, parametric) models do not allow the definition of priors that satisfy (7.20), and alternative less restrictive choices for the sets B_n are possible under mild conditions on the model [156].

7.5.1 Remote contiguity and the LAN condition

To conclude we consider remote contiguity under the condition that the model is LAN (see definition 4.1.12 and LeCam (1960) [174]).

Lemma 7.5.5. *Assume that the model satisfies LAN condition (4.5) with non-singular I_{θ_0} and that the prior Π for θ has a Lebesgue-density $\pi : \mathbb{R}^d \to \mathbb{R}$ that is continuous and strictly positive in all of Θ. For given $H > 0$, define the subsets $B_n = \{\theta \in \Theta : \theta = \theta_0 + n^{-1/2}h, \|h\| \leq H\}$. Then,*

$$P_{0,n} \lhd c_n^{-1} P_n^{\Pi|B_n}, \tag{7.22}$$

for any $c_n \downarrow 0$.

Proof. According to lemma 3 in section 8.4 of Le Cam and Yang (1990) [183], $P_{\theta_0,n}$ is contiguous with respect to $P_n^{\Pi|B_n}$. That implies the assertion.

Note that for some $K > 0$, $\Pi(B_n) \geq b_n := K(H/\sqrt{n})^d$. Assume again the existence of Bayesian tests for $V = \{\theta \in \Theta : \|\theta - \theta_0\| > \rho\}$ (for some $\rho > 0$) versus B_n (or some B such that $B_n \subset B$), of power $a_n = \exp(-\frac{1}{2}n\tau^2)$ (for some $\tau > 0$). Then $a_n b_n^{-1} = o(1)$, and, assuming (7.22), theorem 7.5.1 implies that $\Pi(\|\theta - \theta_0\| > \rho|X^n) \xrightarrow{P_{\theta_0,n}} 0$, so consistency is straightforwardly demonstrated.

The case becomes somewhat more complicated if we are interested in optimality of parametric rates: following the above, a logarithmic correction arises from the lower bound $\Pi(B_n) \geq K(H/\sqrt{n})^d$ when combined in the application of theorem 7.5.1. To alleviate this, we adapt the construction somewhat: define $V_n = \{\theta \in \Theta : \|\theta - \theta_0\| \leq M_n n^{-1/2}\}$ for some $M_n \to \infty$ and B_n like above. Under the condition

that there exists a uniform test sequence for any *fixed* $V = \{\theta \in \Theta : \|\theta - \theta_0\| > \rho\}$ versus B_n (see, for example, [**?**]), uniform test sequences for V_n versus B_n of power $e^{-K'M_n^2}$ exist, for some $k' > 0$. Alternatively, assume that the Hellinger distance and the norm on Θ are related through inequalities of the form,

$$K_1\|\theta - \theta'\| \leq H(P_\theta, P_{\theta'}) \leq K_2\|\theta - \theta'\|,$$

for some constants $K_1, K_2 > 0$. Then cover V_n with rings,

$$V_{n,k} = \left\{\theta \in V_n : \frac{(M_n + k - 1)}{\sqrt{n}} \leq \|\theta - \theta_0\| \leq \frac{(M_n + k)}{\sqrt{n}}\right\},$$

for $k \geq 1$ and cover each ring with balls $V_{n,k,l}$ of radius $n^{-1/2}$, where $1 \leq l \leq L_{n,k}$ and $L_{n,k}$ the minimal number of radius-$n^{-1/2}$ balls needed to cover $V_{n,k}$, related to the *Le Cam dimension* [177]. With the B_n defined like above, and the inequality,

$$\begin{aligned}\int P_{\theta,n}\Pi(V_{n,k,l}|X^n)\, d\Pi_n(\theta|B_n) \\ \leq \sup_{\theta\in B_n} P_{\theta,n}\phi_{n,k,l} + \frac{\Pi_n(V_{n,k,l})}{\Pi_n(B_n)} \sup_{\theta\in V_{n,k,l}} P_{\theta,n}(1-\phi_{n,k,l}),\end{aligned}$$

where the $\phi_{n,k,l}$ are the uniform minimax tests for B_n versus $V_{n,k,l}$ of lemma **??**, of power $\exp(-K'(M_n + k - 1)^2)$ for some $K' > 0$. We define $\phi_{n,k} = \max\{\phi_{n,k,l} : 1 \leq l \leq L_{n,k}\}$ for $V_{n,k}$ versus B_n and note,

$$\int P_{\theta,n}\Pi(V_{n,k}|X^n)\, d\Pi_n(\theta|B_n) \leq \left(L_{n,k} + \frac{\Pi_n(V_{n,k})}{\Pi_n(B_n)}\right) e^{-K(M_n+k-1)^2},$$

where the numbers $L_{n,k}$ are upper bounded by a multiple of $(M_n + k)^d$ and the fraction of prior masses $\Pi_n(V_{n,k})/\Pi_n(B_n)$ can be controlled without logarithmic corrections when summing over k next.

7.6 Consistent hypothesis testing with Bayes factors

The Neyman-Pearson paradigm notwithstanding, hypothesis testing and classification concern the same fundamental statistical question, to find a procedure to choose one subset from a given partition of the parameter space as the most likely to contain the parameter value of the distribution that has generated the data observed. Asymptotically one wonders whether choices following such a procedure focus on the correct subset with probability growing to one.

From a somewhat shifted perspective, we argue as follows: no statistician can be certain of the validity of specifics in his model choice and therefore always runs the risk of biasing his analysis from the outset. Non-parametric approaches alleviate his concern but imply greater uncertainty within the model, leaving the statistician with

the desire to select the correct (sub)model on the basis of the data before embarking upon the statistical analysis proper (for a recent overview, see [241]). The issue also makes an appearance in asymptotic context, where over-parametrized models leave room for inconsistency of estimators, requiring regularization [34, 35, 52].

Model selection describes all statistical methods that attempt to determine from the data which model to use. (Take for example sparse variable selection, where one projects out the majority of covariates prior to actual estimation, and the model-selection question is which projection is optimal.) Methods for model selection range from simple rules-of-thumb, to cross-validation and penalization of the likelihood function. Here we propose to conduct the frequentist analysis with the help of a posterior: when faced with a (dichotomous) model choice, we let the so-called Bayes factor formulate our preference. For an analysis of hypothesis testing that compares Bayesian and frequentist views, see [12]. An objective Bayesian perspective on model selection is provided in [257].

Definition 7.6.1. For all $n \geq 1$, let the model be parametrized by maps $\theta \mapsto P_{\theta,n}$ on a parameter space $(\Theta, \mathscr{G})$ with priors $\Pi_n : \mathscr{G} \to [0,1]$. Consider disjoint, measurable $B, V \subset \Theta$. For given $n \geq 1$, we say that the *Bayes factor for testing B versus V*,

$$F_n = \frac{\Pi(B|X^n)}{\Pi(V|X^n)} \frac{\Pi_n(V)}{\Pi_n(B)},$$

is consistent for testing B versus V, if for all $\theta \in V$, $F_n \xrightarrow{P_{\theta,n}} 0$ and for all $\theta \in B$, $F_n^{-1} \xrightarrow{P_{\theta,n}} 0$.

7.6.1 Frequentist model selection with posteriors

Let us first consider this from a purely Bayesian perspective: for fixed prior Π and *i.i.d.* data, theorem 9.5.1 says that the posterior gives rise to consistent Bayes factors for B versus V in a Bayesian (that is, Π-almost-sure) way, iff a Bayesian test sequence for B versus V exists. If the parameter space Θ is Polish and the maps $\theta \mapsto P_\theta(A)$ are Borel measurable for all $A \in \mathscr{B}$, proposition 7.1.5 says that any Borel set V is Bayesian testable versus $\Theta \setminus V$, so in Polish models for *i.i.d.* data, model selection with Bayes factors is Π-almost-surely consistent for all Borel measurable $V \subset \Theta$.

The frequentist requires strictly more, however, so we employ remote contiguity again to bridge the gap with the Bayesian formulation.

Theorem 7.6.2. *For all $n \geq 1$, let the model be parametrized by maps $\theta \mapsto P_{\theta,n}$ on a parameter space with $(\Theta, \mathscr{G})$ with priors $\Pi_n : \mathscr{G} \to [0,1]$. Consider disjoint, measurable $B, V \subset \Theta$ with $\Pi_n(B), \Pi_n(V) > 0$ such that,*

(i) There exist Bayesian tests for B versus V of power $a_n \downarrow 0$,

$$\int_B P^n \phi_n \, d\Pi_n(P) + \int_V Q^n (1 - \phi_n) \, d\Pi_n(Q) = o(a_n),$$

(ii) For every $\theta \in B$, $P_{\theta,n} \lhd a_n^{-1} P_n^{\Pi_n|B}$, *and for every* $\theta \in V$, $P_{\theta,n} \lhd a_n^{-1} P_n^{\Pi_n|V}$.

Then the Bayes factor for B *versus* V *is consistent.*

Note that the second condition of theorem 9.6.3 can be replaced by a local condition: if, for every $\theta \in B$, there exists a sequence $B_n(\theta) \subset B$ such that $\Pi_n(B_n(\theta)) \geq b_n$ and $P_{\theta,n} \lhd a_n^{-1} b_n P_n^{\Pi_n|B_n}$, then $P_{\theta,n} \lhd a_n^{-1} P_n^{\Pi_n|B}$ (as a consequence of lemma 7.3.3 with $C_n = B$).

7.6.2 Goodness-of-fit Bayes factors for random walks

Consider the asymptotic consistency of goodness-of-fit tests for the transition kernel of a Markov chain with posterior odds or Bayes factors. Bayesian analyses of Markov chains on a finite state space are found in [237] and references therein. Consistency results *c.f.* [252] for random walk data are found in [108]. Large-deviation results for posterior distributions are derived in [206, 85]. The examples below are based on ergodicity for remote contiguity and Hoeffding's inequality for uniformly ergodic Markov chains [193, 113] to construct suitable tests. We first prove the analogue of Schwartz's construction in the case of an ergodic random walk.

Let $(S, \mathscr{S})$ denote a measurable state space for a discrete-time, stationary Markov process P describing a random walk $X^n = \{X_i \in S : 0 \leq i \leq n\}$ of length $n \geq 1$ (conditional on a starting position X_0). The chain has a Markov transition kernel $P(\cdot|\cdot) : \mathscr{S} \times S \to [0,1]$ that describes $X_i|X_{i-1}$ for all $i \geq 1$.

Led by Pearson's approach to goodness-of-fit testing, we choose a finite partition $\alpha = \{A_1, \ldots, A_N\}$ of S and 'bin the data' in the sense that we switch to a new process Z^n taking values in the finite state space $S_\alpha = \{e_j : 1 \leq j \leq N\}$ (where e_j denotes the j-th standard basis vector in $\mathbb{R}^n$), defined by $Z^n = \{Z_i \in S_\alpha : 0 \leq i \leq n\}$, with $Z_i = (1\{X_i \in A_1\}, \ldots, 1\{X_i \in A_N\})$. The process Z^n forms a stationary Markov chain on S_α with distribution $P_{\alpha,n}$. The model is parametrized in terms of the convex set Θ of $N \times N$ Markov transition matrices p_α on the finite state space S_α,

$$p_\alpha(k|l) = P_{\alpha,n}(Z_i = e_k | Z_{i-1} = e_l) = P(X_i \in A_k | X_{i-1} \in A_l), \tag{7.23}$$

for all $0 \leq i \leq n$ and $1 \leq k, l \leq N$. We assume that $P_{\alpha,n}$ is ergodic with equilibrium distribution that we denote by π_α, and $\pi_\alpha(k) := \pi_\alpha(Z = k)$. We are interested in Bayes factors for goodness-of-fit type questions, given a parameter space consisting of transition matrices.

Example 7.6.3. Assume that the true transition kernel P_0 gives rise to a matrix $p_0 \in \Theta$ that generates an ergodic Markov chain Z^n. Denote the true distribution of Z^n by $P_{0,n}$ and the equilibrium distribution by π_0 (with $\pi_0(k) := \pi_0(Z = k)$). For given $\varepsilon > 0$, define,

$$B' = \Big\{ p_\alpha \in \Theta : \sum_{k,l=1}^{N} -p_0(l|k)\pi_0(k) \log \frac{p_\alpha(l|k)}{p_0(l|k)} < \varepsilon^2 \Big\}.$$

Assume that $\Pi(B') > 0$. According to the ergodic theorem, for every $p_\alpha \in B'$,

$$\frac{1}{n}\sum_{i=1}^{n} \log \frac{p_\alpha(Z_i|Z_{i-1})}{p_0(Z_i|Z_{i-1})} \xrightarrow{P_{0,n}\text{-a.s.}} \sum_{k,l=1}^{N} p_0(l|k)\pi_0(k) \log \frac{p(l|k)}{p_0(l|k)},$$

(compare with the rate-function in the large-deviation results in [206, 85]) so that, for large enough n,

$$\frac{dP_{\alpha,n}}{dP_{0,n}}(Z^n) = \prod_{i=1}^{n} \frac{p_\alpha(Z_i|Z_{i-1})}{p_0(Z_i|Z_{i-1})} \geq e^{-\frac{n}{2}\varepsilon^2},$$

$P_{0,n}$-almost-surely. Just like in Schwartz's proof [226], in proposition 8.4.1 and in example 7.3.4, the assumption $\Pi(B') > 0$ and Fatou's lemma imply remote contiguity because,

$$P_{0,n}\Big(\int \frac{dP_{\alpha,n}}{dP_{0,n}}(Z^n)\, d\Pi(p_\alpha|B') < e^{-\frac{n}{2}\varepsilon^2} \Big) \to 0.$$

So lemma 7.2.3 says that $P_{0,n} \lhd \exp(\frac{n}{2}\varepsilon^2) P_n^{\Pi|B'}$.

However, exponential remote contiguity will turn out not to be enough for goodness-of-fit tests below, unless we impose stringent model conditions. Instead, we shall resort to local asymptotic normality for a sharper result.

Example 7.6.4. We formulate goodness-of-fit hypotheses in terms of the joint distribution for two consecutive steps in the random walk. Like Pearson, we fix some such distribution P_0 and consider hypotheses based on differences of 'bin probabilities' $p_\alpha(k,l) = p_\alpha(k|l)\pi_\alpha(l)$,

$$\begin{aligned} H_0 &: \max_{1\leq k,l\leq N} \big| p_\alpha(k,l) - p_0(k,l) \big| < \varepsilon, \\ H_1 &: \max_{1\leq k,l\leq N} \big| p_\alpha(k,l) - p_0(k,l) \big| \geq \varepsilon, \end{aligned} \tag{7.24}$$

for some fixed $\varepsilon > 0$. The sets B and V are defined as the sets of transition matrices $p_\alpha \in \Theta$ that satisfy hypotheses H_0 and H_1 respectively. We assume that the prior is chosen such that $\Pi(B) > 0$ and $\Pi(V) > 0$.

Endowed with some matrix norm, Θ is compact and a Borel prior on Θ can be defined in various ways. For example, we may assign the vector $(p_\alpha(\cdot|1), \ldots, p_\alpha(\cdot|N))$ a product of Dirichlet distributions. Conjugacy applies and the posterior for p_α is again a product of Dirichlet distributions [237]. For an alternative family of priors, consider the set $\mathscr{E}$ of N^N $N \times N$-matrices E that have standard basis vectors e_k in $\mathbb{R}^N$ as columns. Each $E \in \mathscr{E}$ is a deterministic Markov transition matrix on S_α and $\mathscr{E}$ is the extremal set of the polyhedral set Θ. According to Choquet's theorem, every transition matrix p_α can then be written in the form,

$$p_\alpha = \sum_{E\in\mathscr{E}} \lambda_E E, \tag{7.25}$$

for a (non-unique) combination of $\lambda_\mathscr{E} := \{\lambda_E : E \in \mathscr{E}\}$ such that $\lambda_E \geq 0, \sum_\mathscr{E} \lambda_E = 1$. If $\lambda_E > 0$ for all $E \in \mathscr{E}$, the resulting Markov chain is ergodic and we denote the corresponding distributions for Z^n by $P_{\alpha,n}$. Any Borel prior Π' (*e.g.* a Dirichlet distribution) on the simplex S_{N^N} in $\mathbb{R}^{N^N}$ is a prior for $\lambda_\mathscr{E}$ and induces a Borel prior Π on Θ. Note that all non-ergodic transition matrices lie in the boundary $\partial\Theta$, so if we choose Π' such that $\Pi(\mathring{\Theta}) = 1$, ergodicity may be assumed in all prior-almost-sure arguments. This is true for any Π' that is absolutely continuous with respect to the ($N^N - 1$-dimensional) Lebesgue measure on S_{N^N} (for example when we choose Π' equal to a Dirichlet distribution). Note that if the associated density is continuous and strictly positive, $\Pi(B) > 0$ and $\Pi(V) > 0$.

We intend to use theorem 9.6.3 with B and V defined by H_0 and H_1, so we first demonstrate that a Bayesian test sequence for B versus V exists, based on a version of Hoeffding's inequality valid for random walks [113]. First, define, for given $0 < \lambda_n \leq N^{-N}$ such that $\lambda_n \downarrow 0$,

$$S'_n := \left\{\lambda_\mathscr{E} \in S^{N^N} : \lambda_E \geq \lambda_n / N^{N-1}, \text{for all } E \in \mathscr{E}\right\},$$

and denote the image of S'_n under (7.25) by S_n. Note that if $\Pi(\partial\Theta) = 0$, then $\pi_{S,n} := \Pi(\Theta \setminus S_n) \to 0$.

Now fix $n \geq 1$ for the moment. Recalling the nature of the matrices E, we see that for every $1 \leq k,l \leq N$, $p_\alpha(k|l)$ as in equation (7.25) is greater than or equal to λ_n. Consequently, the corresponding Markov chain satisfies condition (A.1) of Glynn and Ormoneit [113] (closely related to the notion of uniform ergodicity [193]): starting in any point X_0 under a transition from S_n, the probability that X_1 lies in $A \subset S_\alpha$ is greater than or equal to $\lambda_n \phi(A)$, where ϕ is the uniform probability measure on S_α. This mixing condition enables a version of Hoeffding's inequality (see theorem 2 in [113]): for any $\lambda_\mathscr{E} \in S'_n$ and $1 \leq k,l \leq N$, the transition matrix of equation (7.25) is such that, with $\hat{p}_n(k,l) = n^{-1}\sum_i 1\{Z_i = k, Z_{i-1} = l\}$,

$$P_{\alpha,n}\left(\hat{p}_n(k,l) - p_\alpha(k,l) \geq \delta\right) \leq \exp\left(-\frac{\lambda_n^2(n\delta - 2\lambda_n^{-1})^2}{2n}\right). \tag{7.26}$$

Now define for a given sequence $\delta_n > 0$ with $\delta_n \downarrow 0$ and all $n \geq 1, 1 \leq k,l \leq N$,

$$\begin{aligned}
B_n &= \{p_\alpha \in \Theta : \max_{k,l} \left|p_\alpha(k,l) - p_0(k,l)\right| < \varepsilon - \delta_n\},\\
V_{k,l} &= \{p_\alpha \in \Theta : \left|p_\alpha(k,l) - p_0(k,l)\right| \geq \varepsilon\},\\
V_{+,k,l,n} &= \{p_\alpha \in \Theta : p_\alpha(k,l) - p_0(k,l) \geq \varepsilon + \delta_n\},\\
V_{-,k,l,n} &= \{p_\alpha \in \Theta : p_\alpha(k,l) - p_0(k,l) \leq -\varepsilon - \delta_n\}.
\end{aligned}$$

Note that if Π' is absolutely continuous with respect to the Lebesgue measure on S^{N^N}, then $\pi_{B,n} := \Pi(B \setminus B_n) \to 0$ and $\pi_{n,k,l} := \Pi(V_{k,l} \setminus (V_{+,k,l,n} \cup V_{-,k,l,n})) \to 0$.

If we define the test $\phi_{+,k,l,n}(Z^n) = 1\{\hat{p}_n(k,l) - p_0(k,l) \geq \varepsilon\}$, then for any $p_\alpha \in B_n \cap S_n$,

$$\begin{aligned} P_{\alpha,n}\phi_{+,k,l,n}(Z^n) &\leq P_{\alpha,n}\big(\hat{p}_n(k,l) - p_\alpha(k,l) \geq \delta_n\big) \\ &\leq \exp\Big(-\frac{\lambda_n^2(n\delta_n - 2\lambda_n^{-1})^2}{2n}\Big). \end{aligned}$$

If on the other hand, p_α lies in the intersection of $V_{+,n,k,l}$ with S_n, we find,

$$\begin{aligned} P_{\alpha,n}(1-\phi_{+,n,k,l}(Z^n)) &= P_{\alpha,n}\big(\hat{p}_n(k,l) - p_\alpha(k,l) < -\delta_n\big) \\ &\leq \exp\Big(-\frac{\lambda_n^2(n\delta_n - 2\lambda_n^{-1})^2}{2n}\Big). \end{aligned}$$

Choosing the sequences δ_n and λ_n such that $n\delta_n^2\lambda_n^2 \to \infty$, we also have $\lambda_n^{-1} = o(n\delta_n)$, so the exponent on the right is smaller than or equal to $-\frac{1}{8}n\lambda_n^2\delta_n^2$.

So if we define $\phi_n(Z^n) = \max_{k,l}\{\phi_{-,k,l,n}(Z^n), \phi_{+,k,l,n}(Z^n)\}$,

$$\begin{aligned} &\int_B P_{\alpha,n}\phi_n\, d\Pi(p_\alpha) + \int_V Q_{\alpha,n}(1-\phi_n)\, d\Pi(q_\alpha) \\ &\quad\leq \int_{B\cap S_n} P_{\alpha,n}\phi_n\, d\Pi(p_\alpha) + \int_{V\cap S_n} Q_{\alpha,n}(1-\phi_n)\, d\Pi(q_\alpha) + \Pi(\Theta\setminus S_n) \\ &\quad\leq \int_B \sum_{k,l=1}^N P_{\alpha,n}(\phi_{-,k,l,n} + \phi_{+,k,l,n})\, d\Pi(p_\alpha) \\ &\qquad + \sum_{k,l=1}^N \Big(\int_{V_{-,k,l}} Q_{\alpha,n}(1-\phi_{-,k,l,n})\, d\Pi(q_\alpha) \\ &\qquad + \int_{V_{+,k,l}} Q_{\alpha,n}(1-\phi_{+,k,l,n})\, d\Pi(q_\alpha)\Big) \\ &\qquad + \sum_{k,l=1}^N \Pi\big(V_{n,k,l}\setminus(V_{+,n,k,l}\cup V_{+,n,k,l})\big) + \Pi(\Theta\setminus S_n) + \Pi(B\setminus B_n) \\ &\quad\leq 2N^2 e^{-\frac{1}{8}n\lambda_n^2\delta_n^2} + \pi_{B,n} + \pi_{S,n} + \sum_{k,l=1}^N \pi_{n,k,l}. \end{aligned}$$

So if we choose a prior Π' on S^{N^N} that is absolutely continuous with respect to Lebesgue measure, then (ϕ_n) defines a Bayesian test sequence for B versus V.

Because we have not imposed control over the rates at which the terms on the *r.h.s.* go to zero, remote contiguity at exponential rates is not good enough. Even if we would restrict supports of a sequence of priors such that $\pi_{B,N} = \pi_{S,n} = \pi_{n,k,l} = 0$, the first term on the *r.h.s.* is sub-exponential. To obtain a rate sharp enough, we note that the chain Z^n is positive recurrent, which guarantees that the dependence $p_\alpha \to dP_{\alpha,n}/dP_{0,n}$ is locally asymptotically normal [128, 114]. According to lemma 7.5.5, this implies that local prior predictive distributions based on $n^{-1/2}$-neighbourhoods of p_0 in Θ are c_n-remotely contiguous to $P_{0,n}$ for *any* rate c_n, if the prior has full

support. If we require that the prior density π' with respect to Lebesgue measure on S^{N^N} is continuous and strictly positive, then we see that there exists a constant $\pi > 0$ such that $\pi'(\lambda) \geq \pi$ for all $\lambda \in S^{N^N}$, so that for every $n^{-1/2}$-neighbourhood B_n of p_0, there exists a $K > 0$ such that $\Pi(B_n) \geq b_n := K n^{-N^N/2}$. Although local asymptotic normality guarantees remote contiguity at arbitrary rate, we still have to make sure that $c_n \to 0$ in lemma 7.5.5, *i.e.* that $a_n = o(b_n)$. Then the remark directly after theorem 9.6.3 shows that condition *(ii)* of said theorem is satisfied.

The above leads to the following conclusion concerning goodness-of-fit testing *c.f.* (7.24).

Proposition 7.6.5. *Let X^n be a stationary, discrete time Markov chain on a measurable state space $(S, \mathscr{S})$. Choose a finite, measurable partition α of S such that the Markov chain Z^n is ergodic. Choose a prior Π' on S^{N^N} absolutely continuous with respect to Lebesgue measure with a continuous density that is everywhere strictly positive. Assume that,*

(i) $n\lambda_n^2 \delta_n^2 / \log(n) \to \infty$,
(ii) $\Pi(B \setminus B_n), \Pi(\Theta \setminus S_n) = o(n^{-(N^N/2)})$,
(iii) $\max_{k,l} \Pi(V_{k,l} \setminus (V_{+,k,l,n} \cup V_{-,k,l,n})) = o(n^{-(N^N/2)})$.

Then for any choice of $\varepsilon > 0$, the Bayes factors F_n are consistent for H_0 versus H_1.

To guarantee ergodicity of Z^n one may use an empirical device, *i.e.* we may use an independent, finite-length realization of the random walk X^n to find a partition α such that for all $1 \leq k, l \leq N$, we observe some m-step transition from l to k. An interesting generalisation concerns a hypothesized Markov transition kernel P_0 for the process X^n and partitions α_n (with projections p_{0,α_n} as in (7.23)), chosen such that α_{n+1} refines α_n for all $n \geq 1$. Bayes factors then test a sequence of pairs of hypotheses (7.24) centred on the p_{0,α_n}. The arguments leading to proposition 7.6.5 do not require modification and the rate of growth N_n comes into the conditions of proposition 7.6.5.

Example 7.6.4 demonstrates the enhancement of the role of the prior as intended by the remark that closes the subsection on the existence of Bayesian test sequences in section 7.1: where testing power is relatively weak, prior mass should be scarce to compensate and where testing power is strong, prior mass should be plentiful. A random walk for which mixing does not occur quickly enough does not give rise to (7.26) and alternatives for which separation decreases too fast lose testing power, so the difference sets of proposition 7.6.5 are the hard-to-test parts of the parameter space and conditions *(ii)–(iii)* formulate how scarce prior mass in these parts has to be.

7.7 Confidence sets from credible sets

The Bernstein-von Mises theorem [183] asserts that the posterior for a smooth, finite-dimensional parameter converges in total variation to a normal distribution

centred on an efficient estimate with the inverse Fisher information as its covariance, if the prior has full support. The methodological implication is that Bayesian credible sets derived from such a posterior can be reinterpreted as asymptotically efficient confidence sets. This parametric fact begs for the exploration of possible non-parametric extensions but Freedman discourages us [101] with counterexamples (see also [61]) and concludes that: *"The sad lesson for inference is this. If frequentist coverage probabilities are wanted in an infinite-dimensional problem, then frequentist coverage probabilities must be computed."*

In recent years, much effort has gone into calculations that address the question whether non-parametric credible sets can play the role of confidence sets nonetheless. The focus lies on well-controlled examples in which both model and prior are Gaussian so that the posterior is conjugate and analyse posterior expectation and variance to determine whether credible metric balls have asymptotic frequentist coverage (for examples, see Szabó, van der Vaart and van Zanten [239] and references therein). Below, we change the question slightly and do not seek to justify the use of credible sets as confidence sets; from the present perspective it appears more natural to ask in which particular fashion a credible set is to be transformed in order to guarantee the transform is a confidence set, at least in the large-sample limit.

In previous subsections, we have applied remote contiguity after the concentration inequality to control the $P_{\theta_0,n}$-expectation of the posterior probability for the alternative V through its $P_n^{\Pi|B_n}$-expectation. In the discussion of the coverage of credible sets that follows, remote contiguity is applied to control the $P_{\theta_0,n}$-probability that θ_0 falls outside the prospective confidence set through its $P_n^{\Pi|B_n}$-probability. The theorem below then follows from an application of Bayes's rule (A.4). Credible levels provide the sequence a_n.

Definition 7.7.1. Let $(\Theta,\mathscr{G})$ with prior Π, denote the sequence of posteriors by $\Pi(\cdot|\cdot) : \mathscr{G} \times \mathscr{X}_n \to [0,1]$. Let $\mathscr{D}$ denote a collection of measurable subsets of Θ. A *sequence of credible sets* (D_n) of *credible levels* $1-a_n$ (where $0 \leq a_n \leq 1$, $a_n \downarrow 0$) is a sequence of set-valued maps $D_n : \mathscr{X}_n \to \mathscr{D}$ such that $\Pi(\Theta \setminus D_n(x)|x) = o(a_n)$ for $P_n^{\Pi_n}$-almost-all $x \in \mathscr{X}_n$.

Definition 7.7.2. For $0 \leq a \leq 1$, a set-valued map $x \mapsto C(x)$ defined on $\mathscr{X}$ such that, for all $\theta \in \Theta$, $P_\theta(\theta \notin C(X)) \leq a$, is called a confidence set of level $1-a$. If the levels $1-a_n$ of a sequence of confidence sets $C_n(X^n)$ go to 1 as $n \to \infty$, the $C_n(X^n)$ are said to be asymptotically consistent.

Definition 7.7.3. Let D be a (credible) set in Θ and let $B = \{B(\theta) : \theta \in \Theta\}$ denote a collection of model subsets such that $\theta \in B(\theta)$ for all $\theta \in \Theta$. A model subset C' is said to be (a confidence set) associated with D under B, if for all $\theta \in \Theta \setminus C'$, $B(\theta) \cap D = \varnothing$. The intersection C of all C' like above equals $\{\theta \in \Theta : B(\theta) \cap D \neq \varnothing\}$ and is called the minimal (confidence) set associated with D under B (see Fig 7.1).

Example 7.7.7 makes this construction explicit in uniform spaces and specializes to metric context.

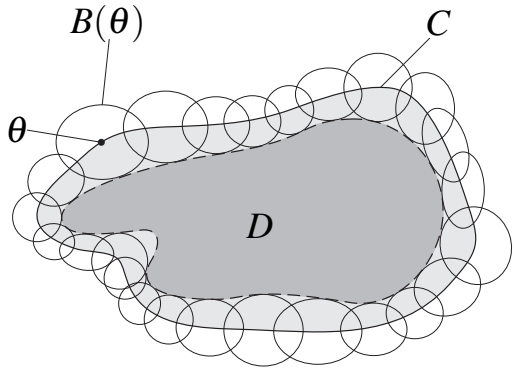


Fig. 7.1 The relation between a credible set D and its associated (minimal) confidence set C under B in Venn diagrams: the extra points θ in the associated confidence set C not included in the credible set D are characterized by non-empty intersection $B(\theta) \cap D \neq \varnothing$.

Theorem 7.7.4. *Let $\theta_0 \in \Theta$ and $0 \leq a_n \leq 1$, $b_n > 0$ such that $a_n = o(b_n)$ be given. Choose priors Π_n and let D_n denote level-$(1-a_n)$ credible sets. Furthermore, for all $\theta \in \Theta$, let $B_n = \{B_n(\theta) \in \mathscr{G} : \theta \in \Theta\}$ denote a sequence such that,*

(i) $\Pi_n(B_n(\theta_0)) \geq b_n$,
(ii) $P_{\theta_0,n} \lhd b_n a_n^{-1} P_n^{\Pi_n | B_n(\theta_0)}$.

Then any confidence sets C_n associated with the credible sets D_n under B_n are asymptotically consistent, i.e. for all $\theta_0 \in \Theta$,

$$P_{\theta_0,n}\big(\theta_0 \in C_n(X^n)\big) \to 1. \tag{7.27}$$

Proof. Fix $n \geq 1$ and let D_n denote a credible set of level $1 - o(a_n)$, defined for all $x \in F_n \subset \mathscr{X}_n$ such that $P_n^{\Pi_n}(F_n) = 1$. For any $x \in F_n$, let $C_n(x)$ denote a confidence set associated with $D_n(x)$ under B. Due to definition 7.7.3, $\theta_0 \in \Theta \setminus C_n(x)$ implies that $B_n(\theta_0) \cap D_n(x) = \varnothing$. Hence the posterior mass of $B(\theta_0)$ satisfies $\Pi(B_n(\theta_0)|x) = o(a_n)$. Consequently, the function $x \mapsto 1\{\theta_0 \in \Theta \setminus C_n(x)\}\, \Pi(B(\theta_0)|x)$ is $o(a_n)$ for all $x \in F_n$. Integrating with respect to the n-th prior predictive distribution and dividing by the prior mass of $B_n(\theta_0)$, one obtains,

$$\frac{1}{\Pi_n(B_n(\theta_0))} \int 1\{\theta_0 \in \Theta \setminus C_n\}\, \Pi(B_n(\theta_0)|X^n)\, dP_n^{\Pi_n} \leq \frac{a_n}{b_n}.$$

Applying Bayes's rule in the form (A.4), we see that,

$$P_n^{\Pi_n|B_n(\theta_0)}\big(\theta_0 \in \Theta \setminus C_n(X^n)\big) = \int P_{\theta,n}\big(\theta_0 \in \Theta \setminus C_n(X^n)\big)\, d\Pi_n(\theta|B_n) \leq \frac{a_n}{b_n}.$$

By the definition of remote contiguity, this implies asymptotic coverage *c.f.* (7.27).

Proof. (corollary 7.7.5)
Define $a_n = \exp(-C'n\varepsilon_n^2)$, $b_n = \exp(-Cn\varepsilon_n^2)$, so that the D_n are credible sets of level $1 - o(a_n)$, the sets B_n of example 7.5.4 satisfy condition *(i)* of theorem 7.7.4 and $b_n a_n^{-1} = \exp(cn\varepsilon_n^2)$ for some $c > 0$. By (7.21), we see that condition *(ii)* of theorem 7.7.4 is satisfied. The assertion now follows.

This refutes Freedman's lesson, showing that the asymptotic identification of credible sets and confidence sets in smooth parametric models (the main inferential implication of the Bernstein-von Mises theorem) generalises to the above form of asymptotic congruence in non-parametric models. The fact that this statement holds in full generality implies very practical ways to obtain confidence sets from posteriors, calculated, simulated or approximated. A second remark concerns the confidence levels of associated confidence sets. In order for the assertion of theorem 7.7.4 to be specific regarding the confidence level (rather than just resulting in asymptotic coverage), we re-write the last condition of theorem 7.7.4 as follows,

(ii') $c_n^{-1} P_{\theta_0,n} \lhd b_n a_n^{-1} P_n^{\Pi_n|B_n(\theta_0)}$,

so that the last step in the proof of theorem 7.7.4 is more specific; particularly, assertion (7.27) becomes,

$$P_{\theta_0,n}\big(\theta \in D_n(X^n)\big) = o(c_n),$$

i.e. the confidence level of the sets $D_n(X^n)$ is $1 - Kc_n$ asymptotically (for some constant $K > 0$ and large enough n).

The following corollary that specializes to the *i.i.d.* situation is immediate (see example 7.7.8). Let $\mathscr{P}$ denote a model of single-observation distributions, endowed with the Hellinger or total-variational topology.

Corollary 7.7.5. *For $n \geq 1$ assume that $(X_1, X_2, \ldots, X_n) \in \mathscr{X}^n \sim P_0^n$ for some $P_0 \in \mathscr{P}$. Let Π_n denote Borel priors on $\mathscr{P}$, with constant $C > 0$ and rate sequence $\varepsilon_n \downarrow 0$ such that (7.20) is satisfied. Denote by D_n credible sets of level $1 - \exp(-C'n\varepsilon_n^2)$, for some $C' > C$. Then the confidence sets C_n associated with D_n under radius-ε_n Hellinger-enlargement are asymptotically consistent.*

Note that in the above corollary,

$$\mathrm{diam}_H(C_n(X^n)) = \mathrm{diam}_H(D_n(X^n)) + 2\varepsilon_n,$$

P_0^n-almost surely. If, in addition to the conditions in the above corollary, tests satisfying (7.19) with $a_n = \exp(-C'n\varepsilon_n^2)$ exist, the posterior is consistent at rate ε_n and sets $D_n(X^n)$ have diameters decreasing as ε_n, *c.f.* theorem 7.5.1. In the case ε_n is the minimax rate of convergence for the problem, the confidence sets $C_n(X^n)$ attain rate-optimality [187]. Rate-adaptivity [125, 54, 239] is not possible like this because a definite, non-data-dependent choice for the B_n is required.

7.7.1 Credible/confidence sets in metric spaces

First we come back to the remark following proposition **??**, concerning shrinking confidence balls centred on small-ball estimators.

Proposition 7.7.6. *Let ...*

Proof. Denote ...

When enlarging credible sets to confidence sets using a collection of subsets B as in definition 7.7.3, measurability of confidence sets is guaranteed if $B(\theta)$ is *open* in Θ for all $\theta \in \Theta$.

Example 7.7.7. Let $\mathscr{G}$ be the Borel σ-algebra for a uniform topology on Θ, like the weak and metric topologies of appendix **??**. Let W denote a symmetric entourage and, for every $\theta \in \Theta$, define $B(\theta) = \{\theta' \in \Theta : (\theta, \theta') \in W\}$, a neighbourhood of θ. Let D denote any credible set. A confidence set associated with D under B is any set C' such that the complement of D contains the W-enlargement of the complement of C'. Equivalently (by the symmetry of W), the W-enlargement of D does not meet the complement of C'. Then the minimal confidence set C associated with D is the W-enlargement of D. If the $B(\theta)$ are all open neighbourhoods (*e.g.* whenever W is a symmetric entourage from a fundamental system for the uniformity on Θ), the minimal confidence set associated with D is open. The most common examples include the Hellinger or total-variational metric uniformities, but weak topologies (like Prohorov's or $\mathscr{T}_n$-topologies) and polar topologies are uniform too.

Example 7.7.8. To illustrate example 7.7.7 with a customary situation, consider a parameter space Θ with parametrization $\theta \mapsto P_\theta^n$, to define a model for *i.i.d.* data $X^n = (X_1, \ldots, X_n) \sim P_{\theta_0}^n$, for some $\theta_0 \in \Theta$. Let $\mathscr{D}$ be the class of all pre-images of Hellinger balls, *i.e.* sets $D(\theta, \varepsilon) \subset \Theta$ of the form,

$$D(\theta, \varepsilon) = \left\{ \theta' \in \Theta : H(P_\theta, P_{\theta'}) < \varepsilon \right\},$$

for any $\theta \in \Theta$ and $\varepsilon > 0$. After choice of a Kullback-Leibler prior Π for θ and calculation of the posteriors, choose D_n equal to the pre-image $D(\hat{\theta}_n, \hat{\varepsilon}_n)$ of a (*e.g.* the one with the smallest radius, if that exists) Hellinger ball with credible level $1 - o(a_n)$, $a_n = \exp(-n\alpha^2)$ for some $\alpha > 0$. Assume, now, that for some $0 < \varepsilon < \alpha$, the W of example 7.7.7 is the Hellinger entourage $W = \{(\theta, \theta') : H(P_\theta, P_{\theta'}) < \varepsilon\}$. Since Kullback-Leibler neighbourhoods are contained in Hellinger balls, the sets $D(\hat{\theta}_n, \hat{\varepsilon}_n + \varepsilon)$ (associated with D_n under the entourage W), is a sequence of asymptotic confidence sets, provided the prior satisfies (6.5). If we make ε vary with n, neighbourhoods of the form B_n in example 7.5.4 are contained in Hellinger balls of radius ε_n, and in that case,

$$C_n(X^n) = D(\hat{\theta}_n, \hat{\varepsilon}_n + \varepsilon_n),$$

is a sequence of asymptotic confidence sets, provided that the prior satisfies (7.20).

7.8 Conclusions

We list and discuss the main conclusions of this chapter below.

Frequentist validity of Bayesian limits
There exists a systematic way of taking Bayesian limits into frequentist ones, if priors satisfy an extra condition relating true data distributions to localized prior predictive distributions. This extra condition generalises Schwartz's Kullback-Leibler condition and amounts to a weakened form of contiguity, termed *remote contiguity*.

For example regarding consistency with *i.i.d.* data, Doob shows that a Bayesian form of posterior consistency holds without any real conditions on the model. To the frequentist, 'holes' of potential inconsistency remain, in null-sets of the prior. Remote contiguity 'fills the holes' and elevates the Bayesian form of consistency to the frequentist one. Similarly, prior-almost-surely consistent tests are promoted to frequentist consistent tests and Bayesian credible sets are converted to frequentist confidence sets.

The nature of Bayesian test sequences
The existence of a Bayesian test sequence is equivalent to consistent posterior convergence in the Bayesian, prior-almost-sure sense. In theorems above, a Bayesian test sequence thus represents the Bayesian limit for which we seek frequentist validity through remote contiguity. Bayesian test sequences are more abundant than the more familiar uniform test sequences. Aside from prior mass requirements arising from remote contiguity, *the prior should assign little weight where testing power is weak and much where testing power is strong*, ideally.

Example 7.6.4 illustrates the influence of the prior when constructing a test sequence. Aside from the familiar lower bounds for prior mass that arise from remote contiguity, existence of Bayesian tests also poses upper bounds for prior mass.

Systematic analysis of complex models and datasets
Although many examples have been studied on a case-by-case basis in the literature, the systematic analysis of limiting properties of posteriors in cases where the data is dependent, or where the model, the parameter space and/or the prior are sample-size dependent, requires generalisation of Schwartz's theorem and its variations, which the formalism presented here provides.

To elaborate, given the growing interest in the analysis of dependent datasets gathered from networks (*e.g.* by *webcrawlers* that random walk linked webpages), or from time-series/stochastic processes (*e.g.* financial data of the high-frequency type), or in the form of high-dimensional or even functional data (biological, financial, medical and meteorological fields provide many examples), the development of new Bayesian methods involving such aspects benefits from a simple, insightful, systematic perspective to guide the search for suitable priors in concrete examples.

To illustrate the last point, let us consider consistent community detection in stochastic block models [204, 30]. Bayesian methods have been developed for consistent selection of the number of communities [124], for community detection with a controlled error-rate with a growing number of communities [57] and for consistent community detection using empirical priors [238]. A moment's thought on the discrete nature of the community assignment vector suggests a sequence of uniform priors, for which remote contiguity (of $B_n = \{P_{0,n}\}$) is guaranteed (at any rate) and prior mass lower bounded by $b_n = K_n! K_n^{-n}$ (where K_n is the number of communities at 'sample size' n). It would be interesting to see under which conditions a Bayesian test sequence of power $a_n = o(b_n)$ can be devised that tests the true assignment vector versus all alternatives (in the sparse regime [66, 2, 199]). Rather than apply a Chernoff bound like in [57], one would probably have to start from the probabilistic [199] or information-theoretic [2] analyses of respective algorithmic solutions in the (very closely related) planted bi-section model. If a suitably powerful test can be shown to exist, theorem 7.5.1 proves frequentist consistency of the posterior.

Methodology for uncertainty quantification
Use of a prior that induces remote contiguity allows one to convert credible sets of a calculated, simulated or approximated posterior into asymptotically consistent confidence sets, in full generality. This extends the main inferential implication of the Bernstein-von Mises theorem to non-parametric models without smoothness conditions.

The latter conclusion forms the most important and practically useful aspect of this book.

7.9 Exercises [EMPTY]

Chapter 8
Inverse limit priors and posteriors

In the non-parametric Bayesian examples of chapter 6, formulation of model and prior proceeds through parametrization: given a subset Θ of a (usually infinite-dimensional Banach or Hilbert) space with Borel probability measure Π on Θ, the model $\mathscr{P}$ arises as the image of a map $\Theta \to M^1(\mathscr{X}) : \theta \mapsto P_\theta$ that is measurable with respect to a suitable σ-algebra on $M^1(\mathscr{X})$. The intrinsically Bayesian *inverse limit distributions* we study in this chapter provide a direct (that is, non-parametrized) formulation of Borel probability distributions on $M^1(\mathscr{X})$ or subspaces thereof. Based on a refining family $\mathscr{A}$ of partitions $\alpha = (A_1, \ldots, A_N)$ of $\mathscr{X}$ with prescribed distributions Π_α for the histograms $(P(A_1), \ldots, P(A_N))$, one would like to discuss a random element $P \sim \Pi$ in $M^1(\mathscr{X})$ with the property that for each partition α, the marginal distribution of $(P(A_1), \ldots, P(A_N))$ matches the prescribed Π_α. Central in the argument is the condition of *coherence*, which says that if A is the disjoint union of two sets A_1, A_2 in $\mathscr{X}$, $P(A_1) + P(A_2)$ and $P(A_1 \cup A_2)$ must be distributed the same.

After some introductory remarks concerning random histograms, we start the chapter with a discussion of two standard families of inverse limit distributions that are used in non-parametric Bayesian statistics (see, for example [94, 95, 107] and many more) and machine learning (see, *e.g.* [40] and others), called the Dirichlet process distributions and the Pólya tree distributions. We consider both in some detail, but refer to other sources for a more broad discussion of their applications.

The issue we focus on here, is the mathematical questions of existence: given a family $\mathscr{A}$ of partitions α with a coherent system of histogram distributions Π_α, *does there exist* a corresponding random P in $M^1(\mathscr{X})$? The matter is non-trivial and many authors have given conditions for the existence of Dirichlet and Pólya tree processes (some of which more appropriate and accurate than others). In this chapter we focus on existence of general inverse limit distributions as Borel and Radon probability measures for various topologies on $M^1(\mathscr{X})$, and we show that the support and approximative properties of an inverse limit distribution vary accordingly: inverse limit distributions that are Borel for Prokhorov's weak topology can have a support that covers all of $M^1(\mathscr{X})$, while inverse limit distributions that are also Radon for the so-called Le Cam-Schwartz topology are necessarily sup-

ported on L^1-subspaces. From the numerical perspective, it is important to note that approximations in terms of histograms $(P(A_1),\ldots,P(A_N))$ are uniformly controlled by histograms of the mean measure in the latter case, but not in the former.

8.1 Random histograms

In this chapter, we let $\mathscr{X}$ be a *topological* sample space; ordinarily $\mathscr{X}$ is a space like $\mathbb{R}$ or $\mathbb{R}^d$, but *functional data* [213] takes its values in spaces of functions, curves, manifolds or distributions (in principle, and in some sampled form in practice), which we assume to be Hausdorff completely regular, or more specifically Polish. Dirichlet process priors have been formulated for functional data as well, and the matter of Bochner-Kolmogorov existence has been raised (see Petrone, Guindani and Gelfand (2009) [207]).

Throughout this introductory section, however, we equate $\mathscr{X}$ to $\mathbb{R}$ for concreteness. The space $\mathscr{X}$ has a metrizable topology $\mathscr{T}$ (with a countable basis) and, correspondingly, a (countably generated) Borel σ-algebra we denote by $\mathscr{B}$. We consider a collection $\mathscr{A}$ of partitions α of $\mathscr{X}$ that consist of a finite number N of Borel sets: $\alpha=\{A_1,\ldots,A_N\}$. (We let $N:\mathscr{A}\to\mathbb{N}$ denote the map that associates with any finite partition α its cardinal.) Initially we shall think of $\mathscr{A}$ as the collection of *all* finite measurable partitions of $\mathscr{X}$, but later we restrict to smaller collections.

Note that for any Borel probability measure $P\in M^1(\mathscr{X})$ on $\mathscr{X}$, there exists a map on $\mathscr{A}$,

$$\alpha\mapsto P_\alpha=\big(P(A_1),\ldots,P(A_{N(\alpha)})\big).$$

that takes any finite, measurable partition α of $\mathscr{X}$ into the *(α-)histogram* associated with P. Note that $P(A_1)+\ldots+P(A_{N(\alpha)})=1$, so P_α is an element of the *simplex* $S_{N(\alpha)}$ (see example 1.1.13). If $\alpha,\beta\in\mathscr{A}$ and β refines α, denote $\alpha=\{A_1,\ldots,A_{N(\alpha)}\}$, $\beta=\{B_1,\ldots,B_{N(\beta)}\}$, and for every $1\le i\le N(\alpha)$, let $J_{\alpha\beta}(i)\subset\{1,\ldots,N(\beta)\}$ be such that $A_i=\cup_{j\in J_{\alpha\beta}(i)}B_j$. By (finite) additivity of the measure P (see definition B.2.1), we have,

$$P(A_i)=P\big(\cup_{j\in J_{\alpha\beta}(i)}B_j\big)=\sum_{j\in J_{\alpha\beta}(i)}P(B_j),\tag{8.1}$$

for all $1\le i\le N(\alpha)$, so the histograms P_α and P_β are related through summation of probabilities for components that are unified when partitions coarsen. Figure 8.1 illustrates how a mixture of three Beta-distributions is mapped to eight, increasingly refined histograms, by repeated subdivisions of the interval $[0,1]$ (a so-called dyadic tree of partitions, see section 8.3). To summarize, any probability measure $P\in M^1(\mathscr{X})$ defines a collection of histograms $\{P_\alpha:\alpha\in\mathscr{A}\}$ related through (8.1) which, conversely, are enough to reconstruct P if $\mathscr{A}$ is rich enough (if $\mathscr{A}$ contains all finite measurable partitions, then the α of the form $\{A,\mathscr{X}\setminus A\}$ are clearly enough; if $\mathscr{A}$ contains α that, jointly, contain all elements of a ring that generates

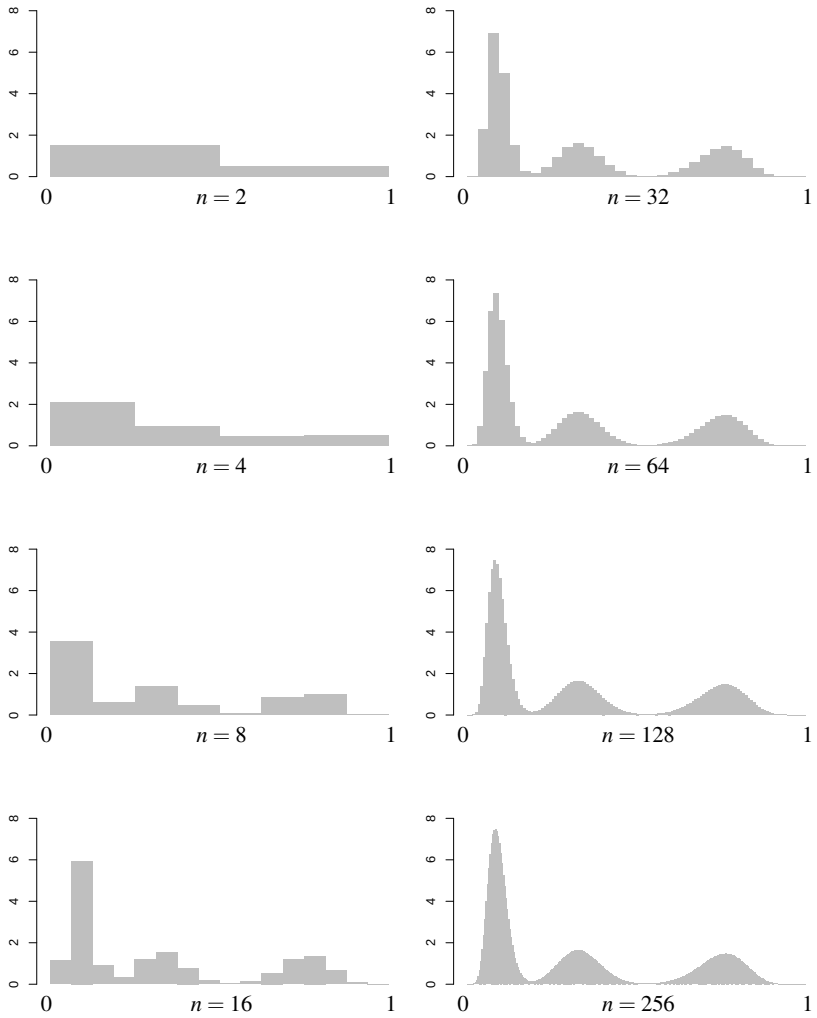


Fig. 8.1 Histograms associated with a dyadic tree of refining partitions of the interval $[0,1]$ for the mixture of Beta-distributions $\frac{1}{2}\text{Beta}(10,100) + \frac{1}{4}\text{Beta}(20,40) + \frac{1}{4}\text{Beta}(30,10)$. Randomization of the number of components, mixing constants or Beta-parameters would result in *random histograms*.

the Borel σ-algebra, then the P_α fix P uniquely, through Carathéodory's extension, theorem B.2.3).

Example 8.1.1. Let P be the normal distribution $N(\mu,\sigma^2)$ on (the Borel σ-algebra) on $\mathbb{R}$, for certain $\mu \in \mathbb{R}$, $\sigma^2 > 0$. Let $\mathscr{A}$ consist of all partitions of $\mathbb{R}$ of the form,

$$\alpha = \big\{(-\infty, a_1], (a_1, a_2], \ldots, (a_{M-1}, a_M], (a_M, \infty)\big\}, \tag{8.2}$$

where $M \geq 0$, $a_1 < a_2 < \ldots < a_M$ with $a_1, \ldots, a_M \in \mathbb{Q}$. Then the α-histogram of P is given by,

$$P_\alpha = \big(\Phi_{\mu,\sigma^2}(a_1), \Phi_{\mu,\sigma^2}(a_2) - \Phi_{\mu,\sigma^2}(a_1), \ldots, 1 - \Phi_{\mu,\sigma^2}(a_M)\big) \in S_{M+1}, \tag{8.3}$$

where $\Phi_{\mu,\sigma^2} : \mathbb{R} \to [0,1]$ denotes the distribution function of the normal distribution $N(\mu, \sigma^2)$. The collection of all histograms of the form (8.3) is rich enough to fix P uniquely to be the normal distribution $N(\mu, \sigma^2)$, because the ring $\mathscr{R}$ consisting of the empty set and all finite unions of half-open intervals $(a, b]$ with $a, b \in \mathbb{Q}$, $a < b$ generates the Borel σ-algebra on $\mathbb{R}$.

Based on example 8.1.1, it is clear that a parametric family of probability distributions corresponds to a parametric collection of histograms. By extension, a prior on a parametric family corresponds to a collection of *random histograms*.

Example 8.1.2. Let Π be a prior probability distribution on $\Theta = \mathbb{R} \times (0, \infty)$, and let $\theta = (\mu, \sigma^2) \in \Theta$ parametrize the family of all normal distributions on $\mathbb{R}$: $P_\theta = N(\mu, \sigma^2)$. Then, for every $\theta \in \Theta$ and every α of the form (8.2), $P_{\theta,\alpha}$ is a histogram of the form (8.3) and if we regard $\theta \sim \Pi$ as random, then for every $\alpha \in \mathscr{A}$, the resulting histogram P_α is a *random histogram* with a distribution Π_α on $S_{N(\alpha)}$.

To generalize this construction, we consider a model $\mathscr{P}$ with a prior or posterior distribution Π on $\mathscr{P}$. Given a collection $\mathscr{A}$ of partitions α, the mapping $(\alpha, P) \mapsto P_\alpha$ induces a collection of random histograms, as formalized in the following theorem.

Theorem 8.1.3. *Let $(\mathscr{P}, \mathscr{G})$ be a measurable model of probability distributions on a measurable space $(\mathscr{X}, \mathscr{B})$, with probability distribution $\Pi : \mathscr{G} \to [0,1]$ and let $\mathscr{A}$ denote a collection of partitions. Assume that for every $\alpha \in \mathscr{A}$ and every $A \in \alpha$, $P \mapsto P(A)$ is $\mathscr{G}$-measurable. Then, for every $\alpha \in \mathscr{A}$, the map $(\alpha, P) \mapsto P_\alpha$ induces a random histogram,*

$$P_\alpha = \big(P(A_1), \ldots, P(A_{N(\alpha)})\big) \sim \Pi_\alpha, \tag{8.4}$$

where Π_α is a probability distribution on $S_{N(\alpha)}$. If $\alpha, \beta \in \mathscr{A}$ and β refines α, then the distribution of P_α follows from that of P_β through summation as in (8.1).

The central question of this chapter concerns the converse of the above theorem: suppose that we specify a collection $\mathscr{A}$ of partitions α such that the set of all sets in all partitions $\alpha \in \mathscr{A}$ generates $\mathscr{B}$ and suppose that we provide distributions Π_α for all random histograms P_α.

Definition 8.1.4. Let $(\mathscr{P}, \mathscr{G})$ be a measurable model of probability distributions on a measurable space $(\mathscr{X}, \mathscr{B})$. Let $\mathscr{A}$ be a collection of $\mathscr{B}$-measurable partitions of $\mathscr{X}$, partially ordered by refinement. For every $\alpha \in \mathscr{A}$, let Π_α be a distribution as in (8.4). Assume that the resulting system of random histograms $(P_\alpha : \alpha \in \mathscr{A})$ is *coherent* in the sense that if $\alpha, \beta \in \mathscr{A}$ and β refines α, then the distribution Π_α of P_α follows from Π_β through summation as in (8.1). If there exists a probability

distribution Π on $(\mathscr{P},\mathscr{G})$ with restrictions Π_α for all $\alpha \in \mathscr{A}$, then Π is called the *inverse limit measure* associated with the *inverse system of measures* described by the random histograms $(P_\alpha : \alpha \in \mathscr{A})$.

With these definitions, we rephrase the central question as follows: under which conditions does an inverse system of random histograms have an inverse limit Π? The inverse limit construction provides ways to define and discuss priors and posteriors on non-parametric Bayesian models (for example, the Dirichlet and Pólya-tree families of distributions discussed below) that are nonetheless essentially parametric in nature (because each of the simplices $S_{N(\alpha)}$ is finite-dimensional) and hence directly accessible to numerical approximation (choose a partition α that is fine enough for the purposes and compute). The existence question has answers that are (perhaps) somewhat more complex than expected and are topological rather than measure-theoretical in nature.

8.2 Dirichlet priors and posteriors

In this section we first consider the so-called *Dirichlet process* defined by a random, finitely-additive, Borel probability set-function $P : \mathscr{B} \to [0,1]$ on the real line. An attempt is made to give a simple proof of its σ-additivity, but that proof fails and we postpone a demonstration that P is almost-surely a probability *measure* until a later stage. In the second subsection we explore some of its properties in a Bayesian setting, particularly we show that with *i.i.d.* data, a Dirichlet prior gives rise to a Dirichlet posterior and we consider the resulting asymptotic composition of the posterior predictive distribution from a frequentist perspective.

8.2.1 The Dirichlet process

Although there are other ways to construct the Dirichlet family of distributions (see, for example, Blackwell and MacQueen (1973) [39]) here we depart from some non-zero, bounded, positive Borel measure $\mu : \mathscr{B} \to [0,\infty)$ on $\mathscr{X} = \mathbb{R}$ and define Dirichlet-distributed (recall section 3.6) random histograms $P_\alpha \sim D_{(\mu(A_1),\ldots,\mu(A_m))}$, for all finite partitions α of $\mathbb{R}$ into Borel measurable subsets $A_1,\ldots,A_N \subset \mathbb{R}$. Because of lemma 3.6.4 (particularly equation (3.20)), (finite) additivity of μ guarantees that if another partition β refines α, the distribution of P_α follows from that of P_β through summation as in (8.1). Using the Daniell-Kolmogorov existence theorem for stochastic processes (see theorem B.6.3), it is shown that there exits a *coupling* $P \sim D_\mu$ for the random variables $\{P(A) : A \in \mathscr{B}\}$. This construction results in a random mapping $P : \mathscr{B} \to [0,1]$ that is finitely additive with D_μ-probability one.

Theorem 8.2.1. *(Existence of the Dirichlet process)*
Given a non-zero, bounded, positive Borel measure μ on $\mathbb{R}$, there exists a unique

probability measure D_μ on the space of bounded, positive, finitely-additive Borel set-functions on $\mathbb{R}$, such that for $P \sim D_\mu$ and every $\mathscr{B}$-measurable partition $(A_1, \ldots, A_k)$ of $\mathbb{R}$,

$$\big(P(A_1), \ldots, P(A_k)\big) \sim D_{(\mu(A_1),\ldots,\mu(A_k))}. \tag{8.5}$$

Proof. Let $k \geq 1$ and $B_1, \ldots, B_k \in \mathscr{B}$ be given; note that here the B_i's are arbitrary and do not form a partition of $\mathbb{R}$, in the sense that they may intersect and their union may not cover all of $\mathbb{R}$. First, we fix the marginal distribution for $(P(B_1), \ldots, P(B_k))$ in terms of that of a partition: through the indicators 1_{B_i} we define 2^k new sets $A_{\nu_1 \ldots \nu_k}$, with $\nu_1, \ldots, \nu_k \in \{0,1\}$, as follows:

$$1_{A_{\nu_1 \ldots \nu_k}}(x) = \prod_{i=1}^{k} 1_{B_i}^{\nu_i}(x)\,(1 - 1_{B_i}(x))^{1-\nu_i},$$

for all $x \in \mathbb{R}$. Then the collection $\{A_{\nu_1 \ldots \nu_k} : \nu_i \in \{0,1\}, 1 \leq i \leq k\}$ *does* form a partition α of $\mathbb{R}$ and from (8.5) we have the histogram marginals,

$$P_\alpha = \big(P(A_{\nu_1 \ldots \nu_k}) : \nu_i \in \{0,1\}, 1 \leq i \leq k\big) \sim D_{(\mu(A_{\nu_1 \ldots \nu_k}) : \nu_i \in \{0,1\}, 1 \leq i \leq k)}.$$

The distribution of the vector $(P(B_1), \ldots, P(B_k))$ then follows from summing appropriately over the ν_i's:

$$\big(P(B_1), \ldots, P(B_k)\big) = \left(\sum_{\{\nu : \nu_1 = 1\}} P(A_{\nu_1 \ldots \nu_k}), \ldots, \sum_{\{\nu : \nu_k = 1\}} P(A_{\nu_1 \ldots \nu_k}) \right), \tag{8.6}$$

in accordance with (3.20). This defines all finite-dimensional marginal distributions as needed in the Daniell-Kolmogorov theorem. To arrive at the underlying probability space $(\Omega, \mathscr{F}, \Pi)$, we have to verify the Kolmogorov consistency conditions (K1) and (K2) of theorem B.6.3, which is a straighforward (albeit somewhat tedious) exercise that is done explicitly in Ferguson (1973,1974) [94, 95]: it is seen readily that omission of one of the B_i's reduces the number of components in the partition α by a factor 2 and components sum appropriately; a permutation of the B_i's amounts to an analogous permutation of the binary indices ν_i, resulting in the required permutation of the components of the finite-dimensional marginal distributions. Conclude that there exists a probability space $(\Omega, \mathscr{F}, D_\mu)$ on which the stochastic process $\{P(A) : A \in \mathscr{B}\}$ can be represented with finite dimensional marginals *c.f.* (8.6). Clearly the resulting random set-function P is finitely additive with D_μ-probability one. Uniqueness of D_μ is trivial: for any other probability measure D' on the space of bounded, positive, finitely-additive Borel set-functions on $\mathbb{R}$, there exits (a cylinder set on which D_μ and D' differ and thence) a Borel set A such that the marginal distribution for $(P(A), P(\mathscr{X} \setminus A))$ under D' differs from (8.5).

The resulting process distribution D_μ is called the *Dirichlet process distribution* with *base measure* μ. The conclusion of theorem 8.2.1 is somewhat disappointing because we were hoping to define random probability *measures*, not just probability set-functions that are finitely additive. The random quantity P is merely the

sample-path of Kolmogorov's stochastic process. What remains, is to demonstrate D_μ-almost-sure *countable* additivity of P. We follow the historical proof [94, 95] but note beforehand that it contains a mistake. (Finding the mistake is left to the reader as exercise 8.10.1.) Although the assertion of proposition 8.2.2 is true, it is surprisingly hard to formulate a correct proof. Existence of inverse limit distributions is discussed comprehensively in sections 8.5–8.9.

Proposition 8.2.2. *Given a bounded, positive Borel base measure μ on $\mathbb{R}$, the Dirichlet process distribution D_μ is concentrated entirely on the subspace of probability measures,*

$$D_\mu\big(P \in M^1(\mathbb{R})\big) = 1,$$

i.e. P is countably additive, D_μ-almost-surely.

Proof. Let (B_n) be a sequence in $\mathscr{B}$ that decreases to $\varnothing$. Since μ is countably additive, $\mu(B_n) \to 0$, according to the continuity theorem for measures (theorem B.2.7). Therefore, there exists a subsequence $(B_{n_j})_{j\geq 1}$ such that $\sum_j \mu(B_{n_j}) < \infty$. For fixed $\varepsilon > 0$, using Markov's inequality,

$$\sum_{j\geq 1} \Pi\big(P(B_{n_j}) > \varepsilon\big) \leq \sum_{j\geq 1} \frac{1}{\varepsilon} \int P(B_{n_j})\, dD_\mu(P) = \frac{1}{\varepsilon} \sum_{j\geq 1} \frac{\mu(B_{n_j})}{\mu(\mathbb{R})} < \infty,$$

according to lemma 3.6.6. From the *first Borel-Cantelli lemma* (lemma B.2.11), we see that, for every $\varepsilon > 0$,

$$D_\mu\Big(\limsup_{j\to\infty} \{P(B_{n_j}) > \varepsilon\}\Big) = D_\mu\left(\bigcap_{J\geq 1} \bigcup_{j\geq J} \{P(B_{n_j}) > \varepsilon\}\right) = 0,$$

which shows that $\lim_j P(B_{n_j}) = 0$, D_μ-almost-surely. Since, by D_μ-almost-sure finite additivity of P,

$$D_\mu\big(P(B_n) \geq P(B_{n+1}) \geq \ldots\big) = 1,$$

we conclude that $\lim_n P(B_n) = 0$, D_μ-almost-surely. Again using theorem B.2.7, P is countably additive, D_μ-almost-surely.

8.2.2 Conjugacy of the Dirichlet family

Although we have not reached the conclusion that the Dirichlet process distributions live on the subspace of all probability measures in $\mathbb{R}$, that conclusion is true as we shall see later. For now we assume that $P \in M^1(\mathbb{R})$, D_μ-almost-surely, and explore the consequences from a Bayesian perspective. In particular, we show that with *i.i.d.* data, a Dirichlet prior gives rise to a Dirichlet posterior (*e.g.* the Dirichlet family of distributions is *conjugate* for any of the full models for *i.i.d.* samples $X_1, \ldots, X_n$, see definition 3.5.1).

Theorem 8.2.3. *Fix $n \geq 1$. Let $X_1, \ldots, X_n$ be an i.i.d. sample of observations in $\mathbb{R}$. Let μ be a bounded, positive Borel base measure on $\mathbb{R}$ with associated Dirichlet process distribution D_μ, used as a prior for the distribution of a single-observation X_i. For any measurable $C \subset M^1(\mathbb{R})$ the posterior probability is given by,*

$$\Pi\big(P \in C \;\big|\; X_1, \ldots, X_n\big) = D_{\mu + \sum_{i=1}^n \delta_{X_i}}(C),$$

almost-surely.

This theorem has implications for weak consistency of posteriors of Dirichlet priors for *i.i.d.* data. Although the theorem is correct as stated, the proof we give below (which is customary in the literature), is somewhat careless regarding the data-dependence of the posterior distribution. We return to this point in section 8.4, which addresses weak consistency more comprehensively.

Proof. Denote $\mu_n = \mu + \sum_{i=1}^n \delta_{X_i}$. Let $\alpha = (A_1, \ldots, A_k)$ be a Borel measurable partition of $\mathbb{R}$ and consider the *cylinderset*,

$$\big\{P \in M^1(\mathbb{R}) : (P(A_1), \ldots, P(A_k)) \in B\big\}, \tag{8.7}$$

where B lies in the k-fold product σ-algebra of the Borel σ-algebra on $[0,1]$. The marginal prior for the histogram P_α is the Dirichlet distribution $D_{(\mu(A_1),\ldots,\mu(A_k))}$, and the model for the data is multinomial with likelihood,

$$L_\alpha(P(A_1), \ldots, P(A_k); X_1, \ldots, X_n) = \prod_{i=1}^n \prod_{j=1}^k P(A_j)^{1\{X_i \in A_j\}}.$$

The proof of theorem 3.6.8 can now be followed to the conclusion that the posterior for $(P(A_1), \ldots, P(A_k))$ is Dirichlet again, $D_{(\mu_n(A_1),\ldots,\mu_n(A_k))}$. Finite unions of sets of the form (8.7) form a ring that generates the (Daniell-Kolmogorov) domain $\mathscr{F}$ for Dirichlet process distributions, so equality of the posterior to D_{μ_n} for all cylindersets implies equality of the posterior to D_{μ_n}, according to the Carathéodory extension, theorem B.2.3.

This theorem makes the Dirichlet process a very practical tool for data analysis with the full, non-parametric model: one chooses a finite partition α of the real line that provides sufficiently high resolution in a bounded subset of interest (with a single complementary set that can contain 'outliers', often chosen such that it contains none of the data points). One then takes some base measure μ (for example, some 'best guess' $Q \in M^1(\mathbb{R})$ the frequentist has for true distribution of the data, normalized with a constant $\Lambda > 0$ of a size that reflects the degree of conviction behind the above 'best guess': $\mu = \Lambda Q$). One may then proceed to calculate the posterior distribution for the α-histogram directly as $D_{(\mu_n(A_1),\ldots,\mu_n(A_k))}$.

This simple perspective is also expressed through Bayesian point estimators for the distribution of a single-observation X_i. Again, let $X_1, X_2, \ldots$ be an *i.i.d.* sample of observations in $\mathbb{R}$ and let μ be a bounded, positive Borel measure on $\mathbb{R}$ with associated Dirichlet process prior D_μ. Let $B \in \mathscr{B}$ be given. Following the steps in (3.23), the posterior predictive distribution (see definition 2.2.2), is then given by:

$$P^{D_\mu|X^n}(B) = \frac{\mu(\mathbb{R})}{\mu(\mathbb{R})+n} P^{D_\mu}(B) + \frac{n}{\mu(\mathbb{R})+n} \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}(B),$$

almost-surely, where P^{D_μ} denotes the prior predictive distribution. With reference to decompositions (3.11) and (3.13), we see that the posterior predictive distribution can be viewed as a convex combination of the empirical distribution $\mathbb{P}_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$ and a bias term P^{D_μ} (equal to Q in the above 'best-guess' scenario),

$$P^{D_\mu|X^n} = \lambda_n P^{D_\mu} + (1-\lambda_n)\mathbb{P}_n, \tag{8.8}$$

almost-surely. As a result, we see that,

$$\left\| P^{D_\mu|X^n} - \mathbb{P}_n \right\|_{TV} = \lambda_n \left\| P^{D_\mu} - \mathbb{P}_n \right\| \leq \lambda_n,$$

almost-surely. (Note that $\lambda_n = \Lambda/(\Lambda+n)$ in the above 'best-guess' scenario, so that larger values for Λ correspond to stronger bias.) Since $\lambda_n \to 0$ as $n \to \infty$, the difference between the sequence of posterior predictive distributions and the sequence of empirical distributions converges to zero in total variation almost surely, as we let the samplesize grow to infinity. Note that the estimator $\mathbb{P}_n$ for P, based on *i.i.d.* data $X^n = (X_1, \ldots, X_n) \sim P^n$, is almost-surely $\mathscr{T}_1$-consistent by the *law of large numbers*, a fact that also underpins the existence results of chapter 9.

8.3 Pólya tree priors and posteriors

Pólya tree distributions generalize the family of Dirichlet distributions, maintaining most of its attractive properties (like tail-freeness, as we shall see in section 8.4) and adding others (like the fact that certain Pólya tree distributions are supported on dominated subspaces of $M^1(\mathscr{X})$). Here we give only the briefest of introductions, for much more on Pólya tree distributions, see [162, 191, 167, 168] and the overviews in [111, 110].

The Pólya tree distribution is defined through a sequence of refining partitions of $\mathbb{R}$ or the interval $[0,1]$, where in each step, every set in the preceding partition is split in two subsets. To accommodate the resulting dyadic tree of refinements, we define the following: for every $m \geq 0$, we denote by $\mathscr{E}_m$ the set of all binary sequences ε of length m (and we denote the empty binary sequence formally as $\varepsilon_\varnothing$, forming the only element of the set denoted $\mathscr{E}_0$). We also define the set $\mathscr{E} = \cup_{m\geq 0}\mathscr{E}_m$ of all *finite binary sequences* (including the empty one). For any two binary sequences $\varepsilon \in \mathscr{E}_m$, $\varepsilon' \in \mathscr{E}_{m'}$, we write $\varepsilon\varepsilon'$ for the concatenation in $\mathscr{E}_{m+m'}$. In particular, for any $\varepsilon \in \mathscr{E}_m$, $\varepsilon 0$ ($\varepsilon 1$) in $\mathscr{E}_{m+1}$ appends a zero (one) to ε. Also note that $\varepsilon_\varnothing \varepsilon = \varepsilon\varepsilon_\varnothing = \varepsilon$ for all $\varepsilon \in \mathscr{E}$. We write out $\varepsilon \in \mathscr{E}_m$ as $\varepsilon = e_1 \ldots e_m$, for $e_i \in \{0,1\}$.

We use $\mathscr{E}$ to organise a refining sequence $\mathscr{A} = \{\alpha_n : n \geq 0\}$ of partitions, $\alpha_0 = \{\mathscr{X}\}$, $\alpha_1 = \{A_0, A_1\}$, $\alpha_2 = \{A_{00}, A_{01}, A_{10}, A_{11}\}$, *etcetera*, into a *dyadic tree*, defining $\alpha_n = \{A_\varepsilon : \varepsilon \in \mathscr{E}_n\}$ and for all $\varepsilon \in \mathscr{E}$,

$$A_\varepsilon = A_{\varepsilon 0} \cup A_{\varepsilon 1}. \tag{8.9}$$

For reasons given in section 8.6, we only look at refinement through intersection with open sets and their complements, *i.e.* for every $\varepsilon \in \mathscr{E}$ either $A_{\varepsilon 0}$ or $A_{\varepsilon 1}$ equals $A_\varepsilon \cap U$ for some open U in $\mathscr{X}$ (we say that the partitions α_n are *generated by a basis* on $\mathscr{X}$, see definition 8.6.2). Also, we require that for every open subset U of $\mathscr{X}$, there exists an $\varepsilon \in \mathscr{E}$ such that $A_\varepsilon \subset U$ (we say that $\mathscr{A}$ *resolves* the topology of $\mathscr{X}$, see definition 8.6.1). The resulting sequence of partitions $\mathscr{A} = \{\alpha_n : n \geq 1\}$ is smaller than the collection of all Borel measurable partitions that was used in theorem 8.2.1 to define the Dirichlet process. Indeed, *if* a random probability measure $P \sim \Pi$ as in theorem 8.1.3 *exists*, then random histograms as in (8.4) are defined for all Borel measurable partitions. However, to define an inverse system and prove existence, smaller families of partitions are possible and, as in this case of the Pólya tree processes, more practical.

Example 8.3.1. A typical example of a *dyadic tree* of partitions starts with $\mathscr{X} = [0,1]$ (or $(0,1)$) and partitions α_m, $m \geq 1$, consisting of 2^m intervals of the forms (l,u), $[l,u)$, $(l,u]$ or $[l,u]$, where $l = u - 2^{-m}$ and $u = 2^{-m}k$, $k = 1,2\ldots,2^m$: with every step in the sequence of refinements, every existing interval is bi-sected at the mid-point. Such partitions are generated by a basis and resolve $\mathscr{X}$.

Example 8.3.2. We also specify a dyadic tree of partitions of $\mathbb{R}$. Let $\mathscr{E} = \cup_{m\geq 0}\mathscr{E}_m$ denote the set of all *finite binary sequences* and define a refining sequence of partitions $\mathscr{A} = \{\alpha_m : m \geq 0\}$ into intervals based on a strictly increasing positive sequence $0 = a_0 < a_1 < a_2 < \ldots$, $a_m \to \infty$, as follows: $A_\varnothing = \mathbb{R}$, $A_0 = (-\infty, -a_0)$, $A_1 = [a_0, \infty)$, $A_{00} = (-\infty, -a_1)$, $A_{01} = [-a_1, a_0)$, $A_{10} = [a_0, a_1]$, $A_{11} = (a_1, \infty)$, and we continue splitting the outer-most intervals like this: for $m \geq 2$, $\varepsilon = 0\ldots0 \in \mathscr{E}_m$, $\varepsilon'_m = 1\ldots1 \in \mathscr{E}_m$ the elements $A_{\varepsilon_m 0}, A_{\varepsilon_m 1}, A_{\varepsilon'_m 0}, A_{\varepsilon'_m 1} \in \alpha_m$ are given by $A_{\varepsilon_m 0} = (-\infty, -a_m)$, $A_{\varepsilon_m 1} = [-a_m, -a_{m-1})$, $A_{\varepsilon'_m 0} = [a_{m-1}, a_m]$, $A_{\varepsilon'_m 1} = (a_m, \infty)$ (and, of course, suitable dyadic refinement into intervals of the intervening sets in the partitions α_m). Such partitions are generated by a basis and resolve $\mathbb{R}$.

To arrive at random histogram distributions for the Pólya tree, we define for every $\varepsilon \in \mathscr{E}$, a so-called *splitting variable* $V_{\varepsilon 0}$ (and $V_{\varepsilon 1} = 1 - V_{\varepsilon 0}$) taking values in $[0,1]$ such that,

(i) for any $\varepsilon, \varepsilon' \in \mathscr{E}$, $V_{\varepsilon 0}$ is independent of $V_{\varepsilon' 0}$;
(ii) for every $\varepsilon \in \mathscr{E}$, there exist $\beta_{\varepsilon 0}, \beta_{\varepsilon 1} > 0$ such that $V_{\varepsilon 0}$ has a Beta$(\beta_{\varepsilon 0}, \beta_{\varepsilon 1})$ distribution.

The splitting variables $V_{\varepsilon 0}$ are interpreted as random fractions that determine how much of the probability mass of A_ε goes to $A_{\varepsilon 0}$ and how much remains for $A_{\varepsilon 1}$, in accordance with (8.9):

$$V_{\varepsilon 0} = P(A_{\varepsilon 0} | A_\varepsilon), \quad V_{\varepsilon 1} = P(A_{\varepsilon 1} | A_\varepsilon).$$

Consequently, for every $m \geq 1$, $\varepsilon = e_1 \ldots e_m \in \mathscr{E}_m$,

$$P(A_\varepsilon) = V_{e_1} V_{e_1 e_2} \dots V_{e_1 \dots e_m} = \prod_{l=1}^{m} V_{e_1 \dots e_l},$$

which fixes the histogram probability measures Π_{α_m} on the *simplex* $S_{\alpha_m} = S_{2^m}$ for all $m \geq 1$,

$$\big(P(A_\varepsilon) : \varepsilon \in \mathscr{E}_m\big) \sim \Pi_{\alpha_m}. \tag{8.10}$$

By construction, the Π_{α_m} are such that refinement and coarsening of partitions (corresponding to relations of the type (8.1)) are accommodated properly.

Example 8.3.3. Let us list some possible choices for the system of partitions and the parameters β_ε, $\varepsilon \in \mathscr{E}$ with the properties of the random distributions that correspond to the coherent system of measures they induce.

(i) Consider $\mathscr{X} = [0,1]$ with a *dyadic tree* of partitions as defined in example 8.3.1. If for all $\varepsilon \in \mathscr{E}$,

$$\beta_\varepsilon = \beta_{\varepsilon 0} + \beta_{\varepsilon 1},$$

then the resulting Pólya tree histograms are those of the Dirichlet process on $[0,1]$ with Lebesgue measure as its base measure.

(ii) Again considering $\mathscr{X} = [0,1]$ with a *dyadic tree* of partitions as defined in example 8.3.1, choose $\beta_\varepsilon = 1$ for all $\varepsilon \in \mathscr{E}$. With probability one, the resulting random probability measure P is mutually singular with respect to Lebesgue measure, but $P(\{x\}) = 0$ for all $x \in [0,1]$.

(iii) Again with $\mathscr{X} = [0,1]$ partitioned like above, choose $\beta_\varepsilon = m^2$ for all $\varepsilon \in \mathscr{E}_m$, all $m \geq 1$. As we shall see in sections 8.7 and 8.9, this choice induces a distribution on $M^1([0,1])$ that is Radon with respect to the total-variational topology.

(iv) To define also a counterexample, consider $\mathscr{X} = \mathbb{R}$ with a *dyadic tree* of partitions as defined in example 8.3.2, and for all $m \geq 0$, $\varepsilon \in \mathscr{E}_m$,

$$\beta_{\varepsilon 0} = \cos\big(\tfrac{1}{2}\pi x(\varepsilon)\big), \quad \beta_{\varepsilon 1} = \sin\big(\tfrac{1}{2}\pi x(\varepsilon)\big),$$

(see example C.4.8 for the definition of the *Cantor mid-point function* $x : \mathscr{E} \to [0,1]$). It is shown in example 8.6.22 that Pólya tree random histograms defined in this way form a coherent system that does not lead to a Borel probability measure on $M^1(\mathbb{R})$ with Prokhorov's weak topology. A similar construction can be based on $\mathscr{X} = (0,1)$ with a dyadic tree of partitions as in example 8.3.1.

Distributions P as in *(ii)* are called *singular continuous* and have distribution functions $F : [0,1] \to [0,1]$ that are continuous but are, in every point, either constant or non-differentiable. Examples can be supported, for example, on the fractal subset $\mathscr{C} \subset [0,1]$ of example C.4.8 homeomorphic to the *Cantor space*, which is nowhere dense, uncountable and has Lebesgue measure zero. Any compact metrizable space is a continuous image of the Cantor space (see appendix C.4, after theorem C.4.9), and any continuous injection of the Cantor set into $\mathbb{R}^d$ with image C has a continuous inverse $f : C \to \{0,1\}^{\mathbb{N}}$ in homeomorphic correspondence with $\mathscr{C}$, so any Lebesgue absolutely continuous probability measure on an continuous, injective image C of the Cantor space induces a singular continuous probability measure on $[0,1]$.

In sections 8.5–8.9 we explore the existence of Pólya tree processes as Borel probability measures on $M^1(\mathscr{X})$, in Prokhorov's weak topology, the Le Cam-Schwartz topology and the total-variational topology. Instrumental in any further study of the Pólya tree process are the following properties of the random histograms defined by the above: for every $m \geq 1$ and every $\varepsilon \in \mathscr{E}_m$,

$$G(A_\varepsilon) := \int_{S_{\alpha_m}} P(A_\varepsilon)\, d\Pi_{\alpha_m} = \prod_{l=1}^{m} \frac{\beta_{e_1 \ldots e_l}}{\beta_{e_1 \ldots e_{l-1}0} + \beta_{e_1 \ldots e_{l-1}1}}, \tag{8.11}$$

by independence of the variables $V_{\varepsilon 0}$ and expectations of the Beta-distributions. The variance of $P(A_\varepsilon)$ is expressed in terms of the parameters β with the help of the quadratic expectations,

$$\begin{aligned} \int_{S_{\alpha_m}} P(A_\varepsilon)^2\, d\Pi_{\alpha_m} &= \int_{S_{\alpha_m}} \prod_{l=1}^{m} V_{e_1 \ldots e_l}^2\, d\Pi_{\alpha_m} \\ &= \prod_{l=1}^{m} \frac{\beta_{e_1 \ldots e_l}(\beta_{e_1 \ldots e_l} + 1)}{(\beta_{e_1 \ldots e_{l-1}0} + \beta_{e_1 \ldots e_{l-1}1})(\beta_{e_1 \ldots e_{l-1}0} + \beta_{e_1 \ldots e_{l-1}1} + 1)}, \end{aligned}$$

based on independence of the variables $V_{\varepsilon 0}$ and variances of the Beta-distributions.

8.3.1 Discrete and dominated Pólya tree distributions

The versatility of the family of Pólya tree distributions becomes clear when we consider the variety of manifestations that make their appearance: as it turns out, a draw from a Pólya tree distribution can be discrete (that is, a countable convex combination of random point-masses) almost surely, or it can be dominated by the mean measure almost-surely, depending on the coefficients $(\beta_{\varepsilon 0}, \beta_{\varepsilon 1})$, $(\varepsilon \in \mathscr{E})$. In this subsection, we consider both possibilities in some detail.

Example 8.3.4. We first note that the family of Pólya tree distributions and the family of Dirichlet distributions have a non-empty intersection (see Ferguson (1974) [95], Lavine (1992) [167]).

Proposition 8.3.5. *Let $\beta_{\varepsilon 0}, \beta_{\varepsilon 1} > 0$ be given for all $\varepsilon \in \mathscr{E}$. If $\beta_{\varepsilon 0} + \beta_{\varepsilon 1} = \beta_\varepsilon$ for all $\varepsilon \in \mathscr{E}$, then for every $m \geq 1$, the random α_m-histogram (8.10) coincides with the corresponding random α_m-histogram (8.5) of the Dirichlet process.*

It follows from the above proposition, that if Pólya tree random histograms with $\beta_{\varepsilon 0} + \beta_{\varepsilon 1} = \beta_\varepsilon$ for all $\varepsilon \in \mathscr{E}$ give rise to a well-defined probability measure Π on $M^1(\mathscr{X})$, then Π is also a Dirichlet process distribution and *any* Borel measurable partition α gives rise to a random histogram *c.f.* (8.4) that is of the Dirichlet form (8.5).

Kingman (1967, 1975) [148, 149] studies certain families of random measures that have a property called *complete randomness*.

Definition 8.3.6. Let $(\mathscr{X},\mathscr{B})$ be any measurable space. A positive random measure $\nu \sim \Pi$ is called a *completely random measure* if, for any finite collection of disjoint sets $A_1,\ldots,A_n \in \mathscr{B}$, the measures $\nu(A_1),\ldots,\nu(A_n)$ are independent. If, for some completely random measure $\nu \sim \Pi$, $\Pi(\nu(\mathscr{X}) < \infty) = 1$, then $\nu/\nu(\mathscr{X})$ is a random probability measure called a *normalized completely random measure*.

Definition 8.3.7. For any positive random measure $\nu \sim \Pi$ on $(\mathscr{X},\mathscr{B})$ and any $t > 0$, define the *cumulant* $\lambda_t : \mathscr{B} \to [0,\infty]$, by,

$$\lambda_t(A) = \log \int e^{t\nu(A)}\, d\Pi(\nu).$$

The following property of completely random measures is now immediate.

Proposition 8.3.8. *If $\nu \sim \Pi$ is a completely random measure, λ_t is a positive measure for any $t > 0$.*

Proof. If $\{A_i : i \geq 1\} \subset \mathscr{B}$ is a countable collection of disjoint measurable sets and $A = \cup_{i\geq 1} A_i$, then,

$$\begin{aligned}\lambda_t(A) &= \log \int e^{t\nu(A)}\, d\Pi(\nu) = \log \int e^{t\sum_{i\geq 1}\nu(A_i)}\, d\Pi(\nu)\\ &= \log \int \prod_{i\geq 1} e^{t\nu(A_i)}\, d\Pi(\nu) = \log \prod_{i\geq 1} \int e^{t\nu(A_i)}\, d\Pi(\nu) = \sum_{i\geq 1} \lambda_t(A_i),\end{aligned}$$

and clearly, $\lambda_t(\varnothing) = 0$.

Based on the Gamma-representation of the Dirichlet distributions (see lemma 3.6.3), we see that Dirichlet random probability measures are of the type discussed by Kingman.

Proposition 8.3.9. *Let $\mathscr{X}$ be a Polish space and let $\mathscr{B}$ denote its Borel σ-algebra. If μ is a positive base-measure and the Dirichlet process defines a random probability measure $P \sim D_\mu$, then P is a normalized completely random measure.*

Proof. For any disjoint $A_1,\ldots,A_n \in \mathscr{B}$ we may define $A_0 = \mathbb{R} \setminus \cup_{i=1}^n A_i$, so that $A_0, A_1, \ldots, A_n$ form a disjoint Borel measurable partition of $\mathscr{X}$. Lemma 3.6.3 asserts that there exist independent $Z_0,\ldots,Z_n$ with $Z_i \sim \Gamma(\mu(A_i),1)$ for all $i = 0,\ldots,n$ such that, $(P(A_0),\ldots,P(A_n))$ and $(Z_0/S,\ldots,Z_n/S)$ have the same (finite-dimensional Dirichlet) distribution.

The completely random process $\{Z_(A) : A \in \mathscr{B}\}$ that is defined through the random histograms,

$$(Z(A_1),\ldots,Z(A_n)) \sim \prod_{i=1}^n \Gamma(\mu(A_i),1),$$

for all finite measurable partitions $A_1,\ldots,A_n$ is called the *Gamma process* with base measure μ and the Dirichlet process could also be called the normalized Gamma process. Many other constructions of this type are thinkable. The important property

to keep in mind here, is the way in which refinement and coarsening of partitions is expressed through the histogram distributions: if $\beta = \{B_1, \ldots, B_m\}$ refines $\alpha = \{A_1, \ldots, A_n\}$, then the distributions of the random histograms associated with β and α must express this: if we impose that the distributions for $(Z(A_1), \ldots, Z(A_n))$ and $(Z(B_1), \ldots, Z(B_m))$ are such that, in distribution,

$$Z(A_i) = \sum_{j \in J_{\alpha\beta}(i)} Z(B_j),$$

then $S = \sum_i Z(A_i) = \sum_j Z(B_j)$ in distribution, and the corresponding random histograms satisfy (8.1). Since the $Z(B_i)$'s are independent, the requirement is that the distributions of all components of partitions are *infinitely divisible* (see Steutel and van Harn (2004) [235]). Infinite divisibility is expressed in the case of the Gamma process by the fact that the sum of two independent random variables $Z_1 \sim \Gamma(\mu_1, 1)$ and $Z_2 \sim \Gamma(\mu_2, 1)$ has a $\Gamma(\mu_1 + \mu_2, 1)$ distribution. In the special case where $\mathscr{X} = [0, \infty)$ the theory of *stochastic processes with independent increments* is particularly well-developed, and in the present context, gives rise to various families of priors (*e.g. compound Poisson process*, σ*-stable process*, *extended* and *generalized Gamma processes*, *Beta-process*, *etcetera*). We refer to a very extensive literature, summarized well in Bayesian non-parametric setting in James, Lijoi and Prünster (2009) [135] and in Ghosal and Van der Vaart (2017) [110].

The most important property of completely random measures is the fact that they describe random measures that are *discrete* with probability one (see example B.2.9) [148, 149]. Below, $D(\mathscr{X})$ denotes the subset of all *discrete measures* on $\mathscr{X}$.

Theorem 8.3.10. *(Kingman (1967, 1975) [148, 149])*
Let $\nu \sim \Pi$ be a completely random measure with cumulant measures λ_t for $t > 0$. If all λ_t are σ-finite then with Π-probability one, ν satisfies the decomposition,

$$\nu = \beta + \nu_f + \nu_r,$$

where β is a non-random, non-degenerate, σ-finite measure on $(\mathscr{X}, \mathscr{B})$; ν_f is a discrete measure supported on a fixed, countable subset $\mathscr{X}' \subset \mathscr{X}$ where $\nu_f(\{x\})$ and $\nu_f(\{x'\})$ are independent if $x, x' \in \mathscr{X}'$, $x \neq x'$; and ν_r is a random element of $D(\mathscr{X})$ that is independent of ν_r.

As it turns out, σ-finiteness of λ_t is equivalent with the existence of countably many $C_1, C_2, \ldots \in \mathscr{B}$ such that $\Pi(\nu(C_i) < \infty) > 0$, for every $i \geq 1$, and the set of fixed atoms $\mathscr{X}'$ is the set of atoms of λ_t (and σ-additivity of λ_t implies countability of $\mathscr{X}'$). The random discrete measure ν_r can be described as a Poisson point-process on $\mathscr{X} \times (0, \infty]$ with an intensity measure with certain precisely circumscribed properties (for details, see [148, 149] or, for example, [110]). The measure β appears in λ_t as the t-linear contribution: $\lambda_t(A) = t\beta(A) + \ldots$. So a completely random measure with cumulant measures that have no $t\beta$-terms ($\beta = 0$) and no fixed atoms ($\nu_f = 0$) are characterized as purely of the form ν_r, that is, random discrete measures. Clearly, if ν is a completely random measure of this type that is bounded Π-almost-surely,

then the corresponding normalized completely random measure is a random probability measure that is discrete Π-almost-surely (see proposition 8.3.11 below).

This discreteness property is one of the most prominent characteristics of Dirichlet process distributions (see Blackwell (1973) [38]). Below a proof is given that is not based (directly) on infinite divisibility or Kingman's theorem but uses a disintegration argument due to Berk and Savage (1979) [23], illustrated further in James (2003) [134], that departs from the conjugacy of the Dirichlet family, theorem 8.2.3.

Proposition 8.3.11. *Let μ be a bounded measure on $\mathscr{X}$ and assume that the Dirichlet process defines a random probability measure $P \sim D_\mu$. Then $D_\mu\big(P \in D(\mathscr{X})\big) = 1$.*

Proof. A probability measure P is discrete if and only if all its mass is contained in singletons, that is, $P(\{x \in \mathbb{R} : P(\{x\}) > 0\}) = 1$. If we consider $P \sim D_\mu$, then, by Bayes's Rule in the form of disintegration (2.4) (see also exercise 2.6.7), we may condition on X rather than P to obtain,

$$\begin{aligned} \int P(P(\{X\}) > 0)\, dD_\mu(P) &= \int_{\mathbb{R}} \int 1_{\{(x,P):P(\{x\})>0\}}\, d\Pi(P|X=x)\, dP^{\Pi}(x) \\ &= \int_{\mathbb{R}} \int 1_{\{(x,P):P(\{x\})>0\}}\, dD_{\mu+\delta_x}(P)\, dP^{D_\mu}(x). \end{aligned} \tag{8.12}$$

where P^{D_μ} denotes the prior predictive distribution. Since,

$$P(\{x\}) \sim \text{Beta}\big(\mu(\{x\}) + 1, \mu(\mathbb{R}) - \mu(\{x\})\big),$$

if $P \sim D_{\mu+\delta_x}$, we have $P(\{x\}) > 0$, $D_{\mu+\delta_x}$-almost-surely, so that the inner integral in (8.12) evaluates to one, for every $x \in \mathbb{R}$. Since $P(\{x \in \mathbb{R} : P(\{x\}) > 0\}) \leq 1$, that is possible only if $D_\mu\big(P \in D(\mathbb{R})\big) = 1$.

If we combine propositions 8.3.5 and 8.3.9 with theorem 8.3.10 or theorem 8.3.11, we see that there are Pólya tree random probability distributions that are Π-almost-surely discrete. The following example, which is also part of the Pólya tree family, shows exactly the opposite character, a random probability measure that is dominated by Lebesgue measure with Π-probability one.

Example 8.3.12. Kraft (1964) [162] devises random histograms on $\mathscr{X} = [0,1]$ as follows: we take $Z_{1,1} = 1/2$ and for every $n \geq 2$, we consider sets of independent random variables $\{Z_{n,k} :, k = 1,2,\ldots,2^n - 1\}$, some of which are inherited from the previous set: $Z_{n,k} = Z_{n-1,l}$ (if $k = 2l$, $1 \leq l \leq 2^{n-1} - 1$); and some of which are drawn independently (if $k = 1,3,5,\ldots,2^n - 1$). At every level $n \geq 1$, we obtain a set of $2^n - 1$ independent random variables in this way, and we impose only that $0 \leq Z_{n,k} \leq 1$ for all (n,k) and that all expectations are equal to $1/2$. Denote the probability space for the corresponding *coupling* by $(\Omega, \mathscr{F}, \Pi)$ and let $([0,1], \mathscr{B}, \mu)$ denote the interval $[0,1]$, regarded as a probability space with Borel σ-algebra and Lebesgue measure. Based on the sets of $Z_{n,k}$, we give a pointwise definition of a

random distribution function $F_n : [0,1] \to [0,1]$ for every $n \geq 1$, through a recurrence relation:

$$F_1(0) = 0, \quad F_1(1/2) = Z_{1,1} = \frac{1}{2}, \quad F_1(1) = 1,$$
$$F_n\Big(\frac{k}{2^n}\Big) = \big(1 - Z_{n,k}\big)F_n\Big(\frac{k-1}{2^n}\Big) + Z_{n,k}F_n\Big(\frac{k+1}{2^n}\Big),$$

with linearly interpolated distributions functions,

$$F_n(x) = F_n\Big(\frac{k}{2^n}\Big) + 2^n\Big(x - \frac{k}{2^n}\Big)\Big(F_n\Big(\frac{k+1}{2^n}\Big) - F_n\Big(\frac{k}{2^n}\Big)\Big),$$

for $2^{-n}k < x < 2^{-n}(k+1)$. Although organised differently, Pólya tree distributions defined on $[0,1]$ with partitions $\alpha_n = \{[0,2^{-n}], (2^{-n}l, 2^{-n}(l+1)] : l = 1,2,\ldots,2^n - 1\}$, and $V_{\varepsilon 0} \sim \text{Beta}(\beta_{\varepsilon 0}, \beta_{\varepsilon 0})$ for some $\beta_{\varepsilon 0} > 0$ are of this type (if we condition on $P_{\alpha_1}(A_{1,1}) = P_{\alpha_1}(A_{1,2})$, or let $\beta_{\varnothing 0} \to \infty$). The sequence of random distribution functions F_n describe random histograms on $[0,1]$ and the question Kraft answered, concerns a sufficient condition for the F_n to describe a limiting random distribution function F that is dominated by μ.

Clearly, for all $n \geq 1$ and $2^{-n}k < x < 2^{-n}(k+1)$, the derivative of F_n exists and is equal to,

$$f_n(x) = \frac{dF_n}{d\mu(x)}(x) = 2^n\Big(F_n\Big(\frac{k+1}{2^n}\Big) - F_n\Big(\frac{k}{2^n}\Big)\Big),$$

and since Lebesgue measure of the sets $\{2^{-n}k : 0 \leq k \leq 2^n\}$ equals zero,

$$F_n(A) = \int_A dF_n(x) = \int_A f_n(x)\, d\mu(x),$$

for every Borel measurable subset A of $[0,1]$. Kraft shows the following.

Theorem 8.3.13. *If, for given random $F_n \sim \Pi$ like above,*

$$\sup_{n \geq 1} \int \Big(\int_0^1 f_n(x)^2\, dx\Big)\, d\Pi < \infty, \tag{8.13}$$

then the probability densities f_n converge to a random probability density f in $L^1([0,1], \mathscr{B}, \mu)$, Π-almost-surely.

Proof. On the product probability space $(\Omega \times [0,1], \sigma(\mathscr{F} \times \mathscr{B}), \Pi \times \mu)$, the random variables $f_n : \Omega \times [0,1] \to [0,\infty)$ form a Martingale, since the expectations of the random variables $Z_{n,k}$ are equal to $1/2$. Martingale convergence implies that there exists a random variable $f : \Omega \times [0,1] \to [0,\infty)$ such that $f_n \to f$, $\Pi \times \mu$-almost-surely. Denote the supremum in condition (8.13) by K. By the square-integrability of f_n with respect to Π, and the Cauchy-Schwartz and Chebyshev inequalities,

$$\int_{f_n>M} f_n(\omega,x)\,d(\Pi\times\mu)(\omega,x) = \int_0^1\int_\Omega f_n(\omega,x)\,1\{f_n(\omega,x)>M\}\,d\Pi(\omega)\,d\mu(x)$$
$$\leq \int_0^1 \left(\int_\Omega f_n(\omega,x)^2\,d\Pi(\omega)\right)^{1/2} \Pi\big(f_n(\omega,x)>M\big)^{1/2}\,d\mu(x)$$
$$\leq \frac{1}{M}\int_0^1\int_\Omega f_n(\omega,x)^2\,d\Pi(\omega)\,d\mu(x) \leq \frac{K}{M},$$

by Fubini's theorem (see theorem B.3.9). Note that the right-hand side is independent of n, so the left-hand side goes to zero uniformly in n, as $M\to\infty$. Conclude that the Martingale f_n is uniformly integrable and hence $f_n\to f$ in $L^1(\Pi\times\mu)$, which amounts to (see [162] for the details),

$$\int_0^1 \big|f_n(\omega,x)-f(\omega,x)\big|\,d\mu(x)\to 0,$$

for Π-almost-all $\omega\in\Omega$.

Let us consider condition 8.13 in some more detail: for any $x\in[0,1]$, there exists a binary sequence $(\varepsilon_i(x))$, such that $x=\sum_{i=1}^\infty 2^{-i}\varepsilon_i(x)$, and for all $x\in[0,1]$ such that $x\neq 2^{-1}k$ for any $n\geq 1$, $0\leq k\leq 2^n$, there exists a sequence $(k_i(x))$ such that $2^{-i}k_i(x)<x<2^{-i}(k_i(x)+1)$ for all $i\geq 1$. This defines $(\varepsilon_i(x))$, $(k_i(x))$ for Lebesgue-almost-all $x\in[0,1]$ and for those x,

$$f_n(x)=\prod_{i=1}^n 2\big(Z_{n,k_i(x)+1}\big)^{1-\varepsilon_i(x)}\big(1-Z_{n,k_i(x)}\big)^{\varepsilon_i(x)}.$$

Consequently, for every x like above,

$$\int f_n(x)^2\,d\Pi=\prod_{i=1}^n 4\left(\int Z^2_{n,k_i(x)+1}\,d\Pi\right)^{1-\varepsilon_i(x)}\left(\int(1-Z_{n,k_i(x)})^2\,d\Pi\right)^{\varepsilon_i(x)},$$

Kraft uses this form to show that condition 8.13 is satisfied whenever,

$$\sum_{n\geq 1}\max\{\mathrm{Var}_\Pi(Z_{n,k}):1\leq k\leq 2^n-1\}<\infty. \tag{8.14}$$

To explain this condition at the heuristic level, one could say that condition (8.14) imposes that the distributions for the random variables $Z_{n,k}$ (which play the same role as the splitting variables $V_{\varepsilon 0}$ in the Pólya tree contruction), should concentrate around their expectations $1/2$ sufficiently fast to 'spread out' the probability mass in the previous random histogram equitably enough for the resulting random probability measure F to be dominated by μ.

We shall see condition (8.14) again in subsection 8.7.1, where it is shown that Pólya tree distributions Π of this type are Radon measures with respect to the Le Cam-Schwartz topology on $M^1([0,1])$ (*c.f.* theorem 8.7.1), which implies that the support of Π is dominated by the expected measure G (see proposition 8.8.6). One verifies easily that $G=\mu$ in Kraft's example. It is worth noting at this point

that according to the Dunford-Pettis theorem (see theorem C.7.12), uniform integrability characterizes relative compactness of norm-bounded sets of measures in the Le Cam-Schwartz topology, emphasizing the connection with the Radon property.

The conclusion of this subsection is that the Pólya tree family of random probability distributions has various widely different manifestations. In sections 8.6, 8.7 and 8.8 we shall see that this is ultimately due to the very same existence issues we have been avoiding so far: depending on whether random histograms give rise to a Borel probability measure on $M^1(\mathscr{X})$ for the Le Cam-Schwartz or Prokhorov's weak topology, domination plays a role or not.

8.4 Tailfreeness and weak posterior consistency

To assess the asymptotic behaviour of posteriors that are based on priors of the types discussed in sections 8.2 and 8.3, we take a step back and first consider the much simpler situation of *i.i.d.* data from a finite sample space, as in section 3.6: subsection 8.4.1 demonstrates posterior consistency in total variation for priors of full support. Extending this, subsection 8.4.2 concerns weak consistency of posteriors for priors that have Freedman's tailfreeness property [98, 92], like the Dirichlet process priors section 8.2 and the Pólya tree priors of section 8.3.

8.4.1 Posterior consistency with finite sample spaces

First consider the situation where we observe an *i.i.d.* sample of random variables $X_1, X_2, \ldots$ taking values in a space $\mathscr{X}_N$ consisting of a *finite* number of points N. Writing $\mathscr{X}_N$ as the set of integers $\{1,\ldots,N\}$, we note that the space $M^1(\mathscr{X}_N)$ of all probability measures P on the measurable space $(\mathscr{X}_N, 2^{\mathscr{X}_N})$ with the total-variational metric $(P,Q) \mapsto \|P-Q\|$, is in isometric correspondence with the simplex S_N (see example 1.1.13) equipped with the L^1-norm $(p,q) \mapsto \|p-q\| = \Sigma_{k=1}^N |p_k - q_k|$ that S_N inherits from $\mathbb{R}^N$ (here, $k \mapsto p_k$ denotes the density of $P \in M^1(\mathscr{X}_N)$ with respect to the counting measure). We also define $M' = \{P \in M^1(\mathscr{X}_N) : P(\{k\}) > 0, 1 \le k \le N\}$ (and the corresponding subset $S' = \{p \in S_N : p(k) > 0, 1 \le k \le N\}$ in S_N).

Proposition 8.4.1. *If the data is an i.i.d. sample of $\mathscr{X}_N$-valued random variables, then for any $n \ge 1$, any Borel prior $\Pi : \mathscr{G} \to [0,1]$ of full support on $M^1(\mathscr{X}_N)$, any $P_0 \in M^1(\mathscr{X}_N)$ and any ball total-variational ball B around P_0, there exists an $\varepsilon' > 0$ such that,*

$$P_0^n \lhd e^{\frac{1}{2}n\varepsilon^2} P_n^{\Pi|B}, \tag{8.15}$$

for all $0 < \varepsilon < \varepsilon'$.

Proof. By the inequality $\|P-Q\| \leq -P\log(dQ/dP)$, the ball B around P_0 contains all sets of the form $K(\varepsilon) = \{P \in M' : -P_0 \log(dP/dP_0) < \varepsilon\}$, for some $\varepsilon' > 0$ and all $0 < \varepsilon < \varepsilon'$. Fix such an ε. Because the mapping $P \mapsto -P_0 \log(dP/dP_0)$ is continuous on M' and M' is dense in M, there exists an open neighbourhood U of P_0 in M such that $U \cap M' \subset K(\varepsilon)$. Since both M' and U are open and Π has full support, $\Pi(K(\varepsilon)) \geq \Pi(U \cap M') > 0$. With the help of example 7.2.2, we see that for every $P \in K(\varepsilon)$,

$$e^{\frac{1}{2}n\varepsilon^2} \frac{dP^n}{dP_0^n}(X^n) \geq 1,$$

for large enough n, P_0-almost-surely. Fatou's lemma again confirms that condition *(ii)* of lemma 7.2.3 is satisfied. Conclude that assertion (8.15) holds.

Together with a uniform test sequence of exponential power, this leads to posteriors that are consistent almost-surely, analogous to Schwartz's theorem 6.3.1.

Theorem 8.4.2. *Let $\mathscr{X}_N$ be a sample space containing a finite number of points and let $X_1, X_2, \ldots$ be an i.i.d. sample of observations in $\mathscr{X}_N$, distributed according to some distribution $P_0 \in M^1(\mathscr{X}_N)$. Endow $M^1(\mathscr{X}_N)$ with the total-variational metric. For any prior Π on $M^1(\mathscr{X}_N)$ that is of full support, the posterior distribution is consistent, almost-surely.*

Proof. Define for given $\delta > 0$, consider the hypotheses,

$$B = \{P \in M : \|P - P_0\| < \delta\},\ V = \{Q \in M : \|Q - P_0\| > 2\delta\}.$$

Noting that $M^1(\mathscr{X}_N)$ is compact (or with the help of the simplex representation S_N) one sees that entropy numbers of $M^1(\mathscr{X}_N)$ are bounded, so the construction of example 7.4.6 shows that uniform tests of exponential power e^{-nD} (for some $D > 0$) exist for B versus V. Application of proposition 8.4.1 shows that the choice for an $0 < \varepsilon < \varepsilon'$ small enough, guarantees that $\Pi(V|X^n)$ goes to zero in P_0^n-probability. Conclude that the posterior resulting from a prior Π of full support on $M^1(\mathscr{X}_N)$ is consistent in total variation.

8.4.2 Tailfreeness and weak consistency

Fix a finite, measurable partition α of a sample space $\mathscr{X}$, denote its cardinal by $N(\alpha)$ and its elements by $A_1, \ldots, A_{N(\alpha)}$. For every $n \geq 1$, denote by $\sigma_{\alpha,n}$ the finite σ-algebra $\sigma(\alpha^n) \subset \mathscr{B}^n$ generated by products of the form $A_{i_1} \times \cdots \times A_{i_n} \subset \mathscr{X}^n$, with $1 \leq i_1, \ldots, i_n \leq N$. Take $N = N(\alpha)$ and identify the finite sample space $\mathscr{X}_N$ of subsection 8.4.1 with an arbitrary *basis* of orthogonal unit-vectors $\{e_1, \ldots, e_N\}$ in $\mathbb{R}^N$. Write $\mathscr{X}_\alpha$ for this space $\mathscr{X}_N$ and define the projection $\varphi'_\alpha : \mathscr{X} \to \mathscr{X}_\alpha$ by,

$$\varphi'_\alpha(x) = \big(1\{x \in A_1\}, \ldots, 1\{x \in A_N\}\big). \tag{8.16}$$

Note that if $\alpha, \beta \in \mathscr{A}$ and β refines α, then (in the notation defines before (8.1)),

$$1_{A_i}(x) = 1_{\cup_{j\in J_{\alpha\beta}(i)}B_j}(x) = \sum_{j\in J_{\alpha\beta}(i)} 1_{B_j}(x),$$

for all $x \in \mathscr{X}$. Therefore, $\varphi_{\alpha\beta} : \mathscr{X}_\beta \to \mathscr{X}_\alpha$ that map unit vectors $e_j \in \mathscr{X}_\beta$ to $e_i \in \mathscr{X}_\alpha$ if $j \in J_{\alpha\beta}(i)$, define maps such that $(\mathscr{X}_\alpha, \varphi_{\alpha\beta})$ is an inverse system (for now only) in the set-theoretic sense (see definition C.5.1).

We view $\mathscr{X}_\alpha$ (respectively $\mathscr{X}_\alpha^n$) as a probability space, and identify its *powerset* σ_α with the σ-algebra $\sigma_{\alpha,1}$ in $\mathscr{B}$ (respectively $\sigma_{\alpha,n}$ in $\mathscr{B}^n$). Probability measures on $\mathscr{X}_\alpha$ are denoted $P_\alpha : \sigma_\alpha \to [0,1]$ and we identify such P_α with elements of $S_{N(\alpha)}$ as we did in subsection 8.4.1. We also define the map $\varphi_{*\alpha} : M^1(\mathscr{X}) \to S_{N(\alpha)}$,

$$\varphi_{*\alpha}(P) = \big(P(A_1), \ldots, P(A_N)\big),$$

that projects Borel probability measures on $\mathscr{X}$ onto their α-histograms (which can be identified with their restrictions to the finite σ-algebra $\sigma_{\alpha,1}$). Given some $P \in M^1(\mathscr{X})$, the projection ϕ_α maps the $\mathscr{X}$-valued random variable X defined by $P(X \in A) = P(A)$ for all $A \in \mathscr{B}$ to a $\mathscr{X}_\alpha$-valued random variable $Z_\alpha = \varphi'_\alpha(X)$ with distribution $P(Z_\alpha \in A) = P_\alpha(A) = (\varphi_{*\alpha}(P))(A)$ for all $A \in \sigma_\alpha$. When this projection is applied to an *i.i.d.* sample of $n \geq 1$ observations, we obtain ${Z_\alpha}^n = (\varphi'_\alpha(X_1), \ldots, \varphi'_\alpha(X_n))$ in ${\mathscr{X}_\alpha}^n$.

A special class of inverse limit priors is the class of tailfree priors [98, 92], defined as follows.

Definition 8.4.3. Let $(\mathscr{P}, \mathscr{G})$ be a measurable model on a measurable space $(\mathscr{X}, \mathscr{B})$, with partitions $\alpha \in \mathscr{A}$ and random histograms distributed according to Π_α, $\alpha \in \mathscr{A}$. Such an inverse system of measures is said to be *tailfree* , if for all α, β such that β refines α, the following holds: the random vector $(P(B_j|A_k) : 1 \leq k \leq N(\alpha), j \in J_{\alpha\beta}(k))$ is independent of $(P(A_1), \ldots, P(A_{N(\alpha)}))$. If a tailfree inverse system of measures has an inverse limit measure Π, then Π is said to be *tailfree*.

Although somewhat technical in its formulation, explicit control of the choice for the Π_α renders this definition quite feasible in practice. The usefulness of tailfreeness derives from the following property.

Proposition 8.4.4. *If an inverse system Π_α, $\alpha \in \mathscr{A}$, is tailfree and has an inverse limit prior Π with posterior $\Pi(\cdot|X^n)$ given $n \geq 1$ and $X^n \in \mathscr{X}^n$, then the posterior is tailfree and the mapping $X^n \mapsto \Pi(A|X^n)$ is $\sigma_{\alpha,n}$-measurable for every $\alpha \in \mathscr{A}$ and $A \in \alpha$.*

Proof. See Freedman (1963) [97] or Ghosal and van der Vaart (2017) [110], theorems 3.14 and 3.15.

Let Π_α denote a Borel prior on $S_{N(\alpha)}$. By definition 2.1.7, the posterior on $S_{N(\alpha)}$ given ${Z_\alpha}^n = {z_\alpha}^n \in \mathscr{X}_\alpha^n$ is a Borel measure denoted $\Pi_\alpha(\cdot|{Z_\alpha}^n = {z_\alpha}^n)$, which satisfies, for all $A \in \sigma_{N(\alpha),n}$ and any Borel set V in $S_{N(\alpha)}$,

$$\int_A \Pi_\alpha(V|{Z_\alpha}^n = {z_\alpha}^n)\, dP_n^{\Pi_\alpha}({z_\alpha}^n) = \int_V P_\alpha^n(A)\, d\Pi_\alpha(P_\alpha).$$

In the model for the original *i.i.d.* sample X^n, Bayes's rule takes the form, for all $A' \in \mathscr{B}_n$ and all Borel sets V' in $M^1(\mathscr{X})$,

$$\int_{A'} \Pi(V'|X^n = x^n)\, dP_n^{\Pi}(x^n) = \int_{V'} P^n(A')\, d\Pi(P),$$

defining the posterior for P. Now specify that V' is the pre-image $\varphi_{*\alpha}^{-1}(V)$ of a Borel measurable V in $S_{N(\alpha)}$: if the prior Π is tailfree, then proposition 8.4.4 says that the data-dependence of the posterior for such a V', $X^n \mapsto \Pi(V'|X^n)$, is measurable with respect to $\sigma_{\alpha,n}$. So there exists a function $g_n : \mathscr{X}_\alpha^n \to [0,1]$ such that,

$$\Pi(V'|X^n = x^n) = g_n(\varphi'_\alpha(x_1), \ldots, \varphi'_\alpha(x_n)),$$

for P_n^{Π}-almost-all $x^n \in \mathscr{X}^n$. Then, for given $A' \in \sigma_{\alpha,n}$ (with corresponding A in the Borel σ-algebra on $S_{N(\alpha)}^n$),

$$\begin{aligned}\int_{A'} \Pi(V'|X^n)\, dP_n^{\Pi} &= \int P^n(1_{A'}(X^n)\, \Pi(V'|X^n))\, d\Pi(P) \\ &= \int P_\alpha^n\big(1_A(Z_\alpha{}^n)\, g_n(Z_\alpha{}^n)\big)\, d\Pi_\alpha(P_\alpha) = \int_A g_n(Z_\alpha{}^n)\, dP_n^{\Pi_\alpha},\end{aligned}$$

while also,

$$\int_{V'} P^n(A')\, d\Pi(P) = \int_V P_\alpha^n(A)\, d\Pi_\alpha(P_\alpha).$$

This shows that $Z_\alpha{}^n \mapsto g_n(Z_\alpha{}^n)$ is a version of the posterior $\Pi_\alpha(\cdot|Z_\alpha{}^n)$ on $S_{N(\alpha)}$. In other words, we can write $\Pi(V'|X^n) = \Pi_\alpha(V|\varphi'_\alpha(X^n)) = \Pi_\alpha(V|Z_\alpha{}^n)$, P_n^{Π}-almost-surely. To summarize, for tailfree priors, posteriors $\Pi_\alpha(\cdot|Z_\alpha{}^n)$ on $S_{N(\alpha)}$ based on projected data $Z_\alpha{}^n$) are projections of the full posterior $\Pi(\cdot|X^n)$.

This implies that for tailfree priors, questions of posterior consistency or concentration are distributed over the inverse system by the histogram projections $\varphi_{*\alpha}$. Denote the true distribution of a single observation from the *i.i.d.* sample X^n by $P_0 \in M^1(\mathscr{X})$. For any V' of the form $\varphi_{*\alpha}^{-1}(V)$ for some α and a neighbourhood V of $P_{0,\alpha} = \varphi_{*\alpha}(P_0)$ in $S_{N(\alpha)}$, the question whether $\Pi(V'|X^n)$ converges to one in P_0-probability reduces to the question whether $\Pi(V|Z_\alpha^n)$ converges to one in $P_{0,\alpha}$-probability. Remote contiguity then only has to hold as in subsection 8.4.1.

Alternatively, note directly that, because $X^n \mapsto \Pi(V'|X^n)$ is $\sigma_{\alpha,n}$-measurable for any $V' = \varphi_{*\alpha}^{-1}(V)$, remote contiguity (as in definition 7.2.1) needs to hold *only* for $\phi_n : \mathscr{X}^n \to [0,1]$ that are measurable with respect to $\sigma_{\alpha,n}$ (rather than $\mathscr{B}^n$) for every $n \geq 1$. That conclusion again reduces the remote contiguity requirement necessary for the consistency of the posterior for the parameter $(P(A_1), \ldots P(A_N))$ to that of the posterior applicable to data from the finite sample space $\mathscr{X}_\alpha$, like in subsection 8.4.1. Full supports of the priors Π_α then guarantee remote contiguity for exponential rates as required in condition *(ii)* of theorem 7.4.1. The uniform tests of exponential power for weak neighbourhoods, as in proposition A.0.6, complete the proof that tailfree priors lead to consistent posterior distributions. Hence consistency of the posteriors $\Pi_\alpha(\cdot|Z_\alpha^n)$ for all α implies that the full posterior $\Pi(\cdot|X^n)$ is

consistent in the *inverse limit topology* that the inverse system of topological spaces $(S_\alpha : \alpha \in \mathscr{A})$ induces on $M^1(\mathscr{X})$.

Theorem 8.4.5. *Let Π_α be a tailfree, coherent inverse system of priors of full support on the simplices $S_{N(\alpha)}$. Then the posterior is consistent in the inverse limit topology, i.e. for every $P_0 \in M^1(\mathscr{X})$,*

$$\Pi(U|X^n) \xrightarrow{P_0\text{-a.s.}} 1,$$

where U is any inverse limit neighbourhood of P_0.

Proof. Recall that the collection of all finite intersections of sets of the form $\varphi_{*\alpha}^{-1}(U')$, where $\alpha \in \mathscr{A}$ and U' is a neighbourhood of $\varphi_{*\alpha}(P_0)$ in $S_{N(\alpha)}$, forms a basis for the inverse limit topology on $M^1(\mathscr{X})$ (see proposition C.5.4). Hence for any inverse limit neighbourhood U of P_0 there exists an α and a neighbourhood U' of $\varphi_{*\alpha}(P_0)$ in $S_{N(\alpha)}$ such that $\varphi_{*\alpha}(U') \subset U$. Because under the assumptions,

$$\Pi_\alpha(U'|{Z_\alpha}^n) \xrightarrow{P_{0,\alpha}\text{-a.s.}} 1,$$

according to theorem 8.4.2, and,

$$\begin{aligned}\Pi(U|X^n = x^n) &\geq \Pi(\varphi_{*\alpha}(U')|X^n = x^n)\\ &= \Pi_\alpha(U'|\varphi'_\alpha(x_1), \ldots, \varphi'_\alpha(x_n)) = \Pi_\alpha(U'|{Z_\alpha}^n = {z_\alpha}^n),\end{aligned}$$

for all $x^n \in \mathscr{X}^n$, the assertion is proved.

Of course, this begs the question what this inverse limit topology on $M^1(\mathscr{X})$ is and this depends on the specific nature of the partitions in the inverse system $(S_\alpha : \alpha \in \mathscr{A})$. As we shall see in subsection 8.6.1, there are various possibilities: if we assume that $\mathscr{X}$ is compact, Polish and the inverse system consists of partitions that are *generated by the basis* (see definition 8.6.2 below) then the inverse limit topology is *Prokhorov's weak topology*. However, if $\mathscr{X}$ is Polish but not compact with such an inverse system, then the inverse limit topology is the *vague topology* (see definition C.9.1). If the inverse system consists of all measurable partitions (see definition **??** then the inverse limit topology is the *Le Cam-Schwartz topology* (see definitions C.9.5 and C.9.5).

Proposition 8.4.6. *The inverse limit topology $\mathscr{T}$ that the inverse system $(S_\alpha : \alpha \in \mathscr{A})$ induces on $M^1(\mathscr{X})$ contains the vague topology. If $\mathscr{X}$ is not compact, $\mathscr{T}$ does not contain Prokhorov's weak topology.*

Proof. Let $f : \mathscr{X} \to \mathbb{R}$ be continuous with compact support.

In the particular case of the Dirichlet process prior, coherency of the inverse system defined by (8.5) is satisfied (*c.f.* the proof of theorem 8.2.1) and full support of the base measure μ implies full support for all Π_α.

As we shall see in the next section, a collection $\mathscr{A}$ of partitions α that are obtained by intersection with elements from a (countable) topological basis for $\mathbb{R}$ is rich enough to render the inverse limit topology equal to Prokhorov's weak topology on $M^1(\mathscr{X})$. If we assume the existence of the Dirichlet prior as a given for the moment, the above proves the following corollary.

Corollary 8.4.7. *Given a bounded, positive Borel measure μ on $\mathbb{R}$ of full support, the posterior associated with the Dirichlet prior D_μ is consistent in Prokhorov's weak topology.*

8.5 Existence of inverse limit measures

Existence of the Dirichlet process prior of subsection 8.2.1, defined in terms of the random histograms (8.5), was left as an open issue following lemma 8.2.2 (see exercise 8.10.1). In subsection 8.5.1 we discuss various existence results from stochastic analysis as well as from the Bayesian non-parametric literature. In subsection 8.5.2 we consider a necessary and sufficient condition for the existence of limits for inverse systems of random histograms in detail, which we use in subsequent sections to prove the existence of the Dirichlet and Pólya-tree families of priors, and to consider their supports in Prokhorov's weak topology, the Le Cam-Schwartz topology and the total-variational topology in subsequent sections.

8.5.1 A variety of existence results

The literature on Bayesian non-parametric statistics and the literature on stochastic analysis have formulated a wide variety of conditions for the existence of inverse limit measures, more or less independently. First explorations of the subject in stochastic analysis date back to the 1950's, starting with Bochner (1955) [41] and Choksi (1958) [58], who formulate the classical Bochner-Kolmogorov conditions for existence of a random distribution function on $\mathbb{R}$. To formalize earlier developments and to state Bochner's theorem, the following definitions are given first (see also definition C.5.2). The index set I is assumed to be a directed set with order relation $\leq$.

Definition 8.5.1. Let $(\mathscr{Y}_\alpha, \varphi_{\alpha\beta})$ be an *inverse system of topological spaces* with *inverse limit* $\mathscr{Y}$ and continuous projections $\varphi_\alpha : \mathscr{Y} \to \mathscr{Y}_\alpha$. Denote the Borel σ-algebra's on the spaces $\mathscr{Y}_\alpha$ by $\mathscr{B}_\alpha$ for all $\alpha \in I$, and the Borel σ-algebra on $\mathscr{Y}$ by $\mathscr{B}$. A collection of Borel measures $\mu_\alpha \in M(\mathscr{Y}_\alpha, \mathscr{B}_\alpha)$ is said to be a *(coherent) inverse system of measures*, if for all $\alpha, \beta \in I$ with $\alpha \leq \beta$, $\mu_\alpha = \mu_\beta \circ \varphi_{\alpha\beta}^{-1}$.

This definition is elaborated upon in proposition 8.5.6. The following common condition on the inverse system of spaces $\mathscr{Y}_\alpha$ is there to guarantee that the inverse limit space $\mathscr{Y}$ is not empty.

Definition 8.5.2. An inverse system of topological spaces $(\mathscr{Y}_\alpha, \varphi_{\alpha\beta})$ is said to be *sequentially maximal* if for every increasing sequence $\alpha_1 \leq \alpha_2 \leq \ldots$ in I and $y_n \in \mathscr{Y}_{\alpha_n}$ such that $\varphi_{\alpha_n\alpha_{n+1}}(y_{n+1}) = y_n$, there exists a $y \in \mathscr{Y}$ such that $\varphi_{\alpha_n}(y) = y_n$ for all $n \geq 1$.

Bochner's theorem (see Bochner (1955), [41]) can then be stated as follows.

Theorem 8.5.3. *Let $(\mathscr{Y}_\alpha, \varphi_{\alpha\beta})$ be an inverse system of Hausdorff topological spaces with inverse limit $\mathscr{Y}$. Assume that $(\mathscr{Y}_\alpha, \varphi_{\alpha\beta})$ is sequentially maximal and let Π_α, $\alpha \in I$, be a coherent system of inner-regular probability measures on the spaces $\mathscr{Y}_\alpha$. Let $\mathscr{B}$ denote the σ-algebra on $\mathscr{Y}$, generated by the collection of all $\varphi_\alpha^{-1}(\mathscr{B}_\alpha)$, $\alpha \in I$. Then there exists a probability measure Π on $(\mathscr{Y}, \mathscr{B})$ such that $\Pi_\alpha = \Pi \circ \varphi_\alpha^{-1}$ for all $\alpha \in I$.*

Choksi (1958) specializes to compact Hausdorff spaces and does not rely on the notion of sequential maximality. Other approaches based on notions of inner regularity and compactness are considered in Metivier (1963) [194], Schwartz (1973) [227], Bourbaki (2010) [49] and discussed comprehensively in Rao (1981) [216] and Bogachev (2007) [42]. Definitions of measure-theoretic inverse limit constructions that (mostly) by-pass topological notions and result in probability measures on the inverse limit space of finitely additive set-functions (rather than its subspace $\mathscr{M}^1(\mathscr{X})$ of probability measures) are presented in Mallory and Sion (1971) [189], Rao (1971) [215], Pintér (2010) [210], and Beznea and Cimpean (2014) [26].

Limits of systems of random histograms in the Bayesian non-parametric literature start with the seminal work of Freedman (1963,1965) [97, 98] and Fabius (1964) [92] and the notion of tailfreeness introduced there. The work of Kingman (1967,1975) [148, 149] on completely random measures forms a leading example in this context and is continued with Ferguson (1973,1974) [94, 95] who defines the Dirichlet process as a *normalized completely random measure*, with an existence proof based on Kolmogorov's theorem and the argument put forth in lemma 8.2.2. Alternative methods to construct the Dirichlet, Pólya-tree and an array of other concrete families of priors on the full model (*e.g. Pólya urn*, *Gibbs type*, *stick-breaking*, *Chinese restaurant* and *Indian buffet* processes) have been studied extensively and imply their own, specific proofs of existence (for overviews, see [209, 65, 110]). Ghosh and Ramamoorthi (2003) [111] construct priors with Kraft-type conditions on an inverse system for distribution functions on $\mathbb{R}$ and prove existence. Based on Harris's theorem [123], Orbanz (2011) [205] requires that there exists a Borel measure G with histogram projections G_α such that,

$$G_\alpha(A) = \int P_\alpha(A)\, d\Pi_\alpha(P_\alpha), \tag{8.17}$$

for all partitions α and all $A \in \alpha$. Ghosal and van der Vaart (2017) [110] formulate the related condition that the set-function $G : \mathscr{B} \to [0,1]$,

$$G(A) = \int P(A)\, d\Pi(P),$$

defines a Borel measure. The Ghosal-van der Vaart condition is certainly a necessary condition (see lemma 2.1.5), but, strictly speaking, tautological when it comes to the existence of Π. It can only be justified by realizing that in some examples (*e.g.* the Dirichlet process) the set-function G is fixed through the definition of the histogram distributions. (Or, in other examples, the random histograms fix G on a ring that generates $\mathscr{B}$; see theorems 3.2 and 3.9 in [110], which require an involved form of inner regularity and do not lead to a sharp existence condition for the Pólya tree distribution, *c.f.* remark 8.6.23). Note that the more generally accepted [109] Bayesian condition that the random P satisfy,

$$\Pi\big(0 \leq P \leq 1 \text{ is } \sigma\text{-additive}\big) = 1, \tag{8.18}$$

also refers to Π before proving it exists as a measure, which does not tally with the fact that the set of countably additive P does not have a well-defined meaning in terms of the histogram distributions (even if we take a Daniell-Kolmogorov-type existence assertion like that of theorem 8.2.1 for granted).

8.5.2 The Bourbaki-Prokhorov-Schwartz theorem

The conditions we derive in subsequent sections are based on the equivalence referred to as Prokhorov's theorem in Bourbaki (2010) [49], which is based on a form of inner regularity that holds for all measures in the inverse system simultaneously. This leads to characterization of those inverse systems $(\Pi_\alpha : \alpha \in \mathscr{A})$ that consistently define Radon probability measures Π on $M^1(\mathscr{X})$ with various topologies.

The relevant theorem is first stated in [227] and re-stated more generally in [49], Ch. IX, § 4, No. 2. It is formulated involving general inverse systems of measures, does not refer to maximal sequentiality and results in a Radon measure as the limit. Before we give the theorem, we state relevant definitions and after, we indicate how we apply it in the setting of inverse systems of random histograms.

Definition 8.5.4. Let $(\mathscr{Y}_\alpha, \varphi_{\alpha\beta})$ be an inverse system of topological spaces and let T be a topological space. A family of functions $\varphi_\alpha : T \to \mathscr{Y}_\alpha$, $\alpha \in I$, is said to be *coherent* if, for all α, β, $\alpha \leq \beta$, $\varphi_\alpha = \varphi_{\alpha\beta} \circ \varphi_\beta$, and it is said to be *separating* if, for all $x, y \in T$, $x \neq y$, there exists an $\alpha \in I$ such that $\varphi_\alpha(x) \neq \varphi_\alpha(y)$.

Theorem 8.5.5. *(Bourbaki-Prokhorov-Schwartz)*
Let $(\mathscr{Y}_\alpha, \varphi_{\alpha\beta})$ be an inverse system of Hausdorff topological spaces, indexed by $\mathscr{A}$, T a Hausdorff topological space and $\varphi_\alpha : T \to \mathscr{Y}_\alpha$ a coherent and separating family of continuous mappings. Let $(\mu_\alpha, \varphi_{*\alpha\beta})$ be a coherent inverse system of positive measures on $(\mathscr{Y}_\alpha, \varphi_{\alpha\beta})$. There exists a bounded positive Radon measure μ on T projecting to μ_α for all $\alpha \in I$, if and only if the following property is satisfied:

(P) for every $\varepsilon > 0$, there is a compact $H \subset T$ such that for all $\alpha \in I$,

$$\mu_\alpha(\mathscr{Y}_\alpha \setminus \varphi_\alpha(H)) \leq \varepsilon.$$

When (P) holds, the measure μ is uniquely determined and $\mu(L) = \inf\{\mu_\alpha(\varphi_\alpha(L)) : \alpha \in I\}$ for every compact set L in T.

Let us prepare the subsequent discussion with some specifications pertaining to the situation where $I = \mathscr{A}$, $T = M^1(\mathscr{X})$ and $\mathscr{Y}_\alpha = M^1(\mathscr{X}_\alpha)$. Because the spaces $\mathscr{X}_\alpha$ are discrete, *any* $g : \mathscr{X}_\alpha \to \mathbb{R}$ lies in the (finite-dimensional) Banach space $C(\mathscr{X}_\alpha)$ of bounded, continuous, real-valued maps with the uniform norm. The continuous dual $M(\mathscr{X}_\alpha)$ of $C(\mathscr{X}_\alpha)$ is a (finite-dimensional) Banach space when equipped with the total-variational norm. For $\mu \in M(\mathscr{X}_\alpha)$ and $g \in C(\mathscr{X}_\alpha)$, $\int g\, d\mu$ is denoted in bi-linear form $\langle \mu, g \rangle_\alpha$.

Let $\alpha, \beta \in \mathscr{A}$ with $\alpha \leq \beta$ be given. For any $g \in C(\mathscr{X}_\alpha)$, the map $g \circ \varphi_{\alpha\beta} : \mathscr{X}_\beta \to \mathbb{R}$ is an element of $C(\mathscr{X}_\beta)$. Because $\varphi_{\alpha\beta}$ is surjective, the induced map $\varphi^*_{\alpha\beta} : C(\mathscr{X}_\alpha) \to C(\mathscr{X}_\beta)$ is a bounded linear operator with norm equal to one. The transpose map $\varphi_{*\alpha\beta} : M(\mathscr{X}_\beta) \to M(\mathscr{X}_\alpha)$ is defined by,

$$\langle \varphi_{*\alpha\beta}(\mu), g \rangle_\alpha = \langle \mu, \varphi^*_{\alpha\beta}(g) \rangle_\beta = \langle \mu, g \circ \varphi_{\alpha\beta} \rangle_\beta$$

for all $\mu \in M(\mathscr{X}_\beta)$ and $g \in C(\mathscr{X}_\alpha)$. The linear map $\varphi_{*\alpha\beta}$ is bounded with norm less than or equal to one. Note that if we express $\mu \in M(\mathscr{X}_\beta)$ as a vector $(\mu_1, \ldots, \mu_{N(\beta)})$ in $\mathbb{R}^{N(\beta)}$,

$$\begin{aligned} \langle \mu, g \circ \varphi_{\alpha\beta} \rangle_\beta &= \sum_{j \in I(\beta)} \mu_j\, g(\varphi_{\alpha\beta}(e_j)) \\ &= \sum_{j \in I(\beta)} \mu_j\, g(e_{i_{\alpha\beta}(j)}) = \sum_{i \in I(\alpha)} \Big(\sum_{j \in J_{\alpha\beta}(i)} \mu_j \Big) g(e_i), \end{aligned}$$

from which one determines that coherence of the inverse system of measures amounts to,

$$\varphi_{*\alpha\beta}(\mu)_i = \sum_{j \in J_{\alpha\beta}(i)} \mu_j, \tag{8.19}$$

for all $1 \leq i \leq N(\alpha)$, *c.f.* (8.1). Note that for any $\alpha, \beta, \gamma \in \mathscr{A}$, $\alpha \leq \beta \leq \gamma$, $\varphi_{*\alpha\gamma} = \varphi_{*\alpha\beta} \circ \varphi_{*\beta\gamma}$ and that $\varphi_{*\alpha\alpha}$ is the identity. The following proposition formalizes this construction as an inverse limit of spaces of probability measures on $\mathscr{X}_\alpha$ (that is, α-histograms).

Proposition 8.5.6. *Let P_α, $(\alpha \in \mathscr{A})$ be a coherent inverse system of probability measures on the inverse system $(\mathscr{X}_\alpha, \varphi_{\alpha\beta})$. Then,*

$$P_\alpha = \varphi_{*\alpha\beta}(P_\beta) = P_\beta \circ \varphi^{-1}_{\alpha\beta}, \tag{8.20}$$

and $(M^1(\mathscr{X}_\alpha), \varphi_{*\alpha\beta})$ forms an inverse system of non-empty, compact topological spaces, with non-empty, compact inverse limit N that can be identified with the set of all finitely additive probability set-functions on a domain equal to the union $\{A \subset \mathscr{X} : A \in \alpha, \alpha \in \mathscr{A}\}$.

As it turns out the space of all probability measures on $\mathscr{X}$, $M^1(\mathscr{X})$, can be mapped into N injectively. Note however, that the space T in theorem 8.5.5 is not identified

with a subset of N directly; instead, the essential relationship is formed by requiring *continuity* of the maps $\varphi_\alpha : T \to \mathscr{X}_\alpha$.

8.6 Inverse limit measures in Prokhorov's weak topology

In this section we consider theorem 8.5.5 for $T = M^1(\mathscr{X})$ with Prokhorov's weak topology. In order to satisfy the continuity condition for the maps $\varphi_\alpha : T \to \mathscr{X}_\alpha$, we are forced to consider a zero-dimensional version $\hat{\mathscr{X}}$ of the sample space $\mathscr{X}$, which has a space of bounded Borel measures $M(\hat{\mathscr{X}})$ that is mapped continuously but not surjectively into $M(\mathscr{X})$.

8.6.1 Inverse limit sample spaces

We require $\mathscr{X}$ to be a *completely regular space* or, more specifically, a Polish space (with *basis* $\mathscr{U}$). Given that the central condition of the Bourbaki-Prokhorov-Schwartz theorem requires a form of inner regularity *for all* $\alpha \in \mathscr{A}$ *simultaneously*, smaller collections of partitions are preferable. A minimal requirement is the following.

Definition 8.6.1. A collection of partitions $\mathscr{A}$ of a topological space $\mathscr{X}$ is said to is said to be *separating*, if for any pair $x, y \in \mathscr{X}$ such that $x \neq y$, there exists an $\alpha \in \mathscr{A}$ such that $\varphi'_\alpha(x) \neq \varphi'_\alpha(y)$. We say that $\mathscr{A}$ *resolve* $\mathscr{X}$ if, for every open $U \subset \mathscr{X}$, there exists a $\alpha \in \mathscr{A}$ and an element $A \in \alpha$, such that $A \subset U$.

If $\mathscr{X}$ is Hausdorff, any $\mathscr{A}$ that resolves $\mathscr{X}$ also separates $\mathscr{X}$. A collection of partitions $\mathscr{A}$ that resolves $\mathscr{X}$ makes it possible for every $x \in \mathscr{X}$, to find a filter $\mathscr{F}$ that converges to x and consists of only sets that contain some $A \in \alpha$, for some $\alpha \in \mathscr{A}$.

Definition 8.6.2. Let $\mathscr{U}$ be a *topological basis* for $\mathscr{X}$. We say that a partition α is *generated by the basis* $\mathscr{U}$, if any set in α is the union of a finite number of subsets obtained through a finite number of intersections of $\mathscr{X}$ with U or $\mathscr{X} \setminus U$, $U \in \mathscr{U}$. We say that a collection of partitions$\mathscr{A}$ is *generated by the basis* $\mathscr{U}$, if $\mathscr{A}$ consists only of partitions generated by $\mathscr{U}$.

Example 8.6.3. In case $\mathscr{X}$ is *second countable* (*e.g.* when $\mathscr{X}$ is separable and metrizable), we consider the following sequence of partitions $\mathscr{A} = \{\alpha_n : n \geq 1\}$: let $\{U_n : n \geq 1\}$ enumerate $\mathscr{U}$, set $\alpha_1 = \{\mathscr{X}\}$ and define recursively,

$$\alpha_{n+1} = \{A \cup U_n, A \setminus U_n : A \in \alpha_n\},$$

for all $n \geq 1$. Because $\mathscr{U}$ is a basis, $\mathscr{A}$ resolves (and if $\mathscr{X}$ is Hausdorff $\mathscr{A}$ also separates) $\mathscr{X}$. Any partition α generated by the basis $\mathscr{U}$ is a coarsening of a partition α_n, for some $n \geq 1$. Recall that if $\mathscr{X}$ is metrizable with countable dense subset Q

(and a countable set $R \subset (0,\infty)$ such that $R \cap U \neq \varnothing$ for any neighbourhood of 0 in $[0,\infty)$), the (countable) collection $\mathscr{U}$ of *open balls* $\{(q-r, q+r) : q \in Q, r \in R\}$ forms a countable basis.

Example 8.6.4. A different, more organised choice in $\mathbb{R}$ for the partitions $\{\alpha_n : n \geq 1\}$ is described as follows: choose two sequences of integers (m_n), (k_n) with $m_n \geq 0$ and $k_n \geq 1$, define $q_1 = 1$, $\delta_1 = 1$ to start the recursion,

$$q_{n+1} - q_n = m_n \delta_n, \quad \delta_{n+1} = \frac{\delta_n}{k_n},$$

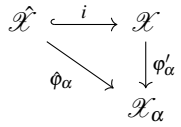
and consider the partition α_n with elements,

$$A_{0,n} = (-\infty, -q_n) \cup (q_n, \infty), \quad A_{1,n} = [-q_n, -q_n + \delta_n],$$
$$A_{l,n} = (-q_n + (l-1)\delta_n, -q_n + l\delta_n],$$

with $l = 1, \ldots, N_n$, and $N_1 = 3$, $N_{n+1} = ((N_n - 1) + 2m_n)k_n + 1$. A moments thought shows that these partitions of $\mathbb{R}$ are generated by a basis, which, for every $n \geq 1$, covers $[-q_n, q_n]$ with half-overlapping open sets of width δ_n and then coarsens the resulting partition to obtain α_n. The resulting α_n are partitions of $\mathbb{R}$ generated by $\mathscr{U}$, which resolve $\mathbb{R}$, if and only if $\limsup_n m_n > 1$ and $\limsup_n k_n > 1$. If $\lim_n m_n = 0$ but $\limsup_n k_n > 2$, the interval $[-q, q]$ (with $q = \lim_n q_n = \sum_n m_n < \infty$) is resolved by the 'subset partitions' $\alpha'_n = \{A_{l,n} : 1 \leq l \leq N_n\}$ that cover $[-q, q]$ for large enough n. (If $\lim_n k_n = 1$, the resulting sequence of partitions does not resolve $\mathbb{R}$ (nor $[-q, q]$).) The choices $m_n = 0$ and $k_n = 2$ for all $n \geq 1$ can be mapped to the dyadic partitions of $[0,1]$ that underpin the Pólya trees of example 8.3.1.

Note that the spaces $\mathscr{X}_\alpha$, $(\alpha \in \mathscr{A})$, have finite cardinals and can be viewed as *discrete topological spaces*. Then the maps $\varphi_{\alpha\beta}$ are trivially continuous and organize the $\mathscr{X}_\alpha$ into an inverse system of topological spaces, $(\mathscr{X}_\alpha, \varphi_{\alpha\beta})$ (see definition C.5.2). Since each of the $\mathscr{X}_\alpha$ is non-empty and compact, the inverse limit Y is non-empty and compact. For any $\alpha \in \mathscr{A}$, denote by φ_α the continuous projection that takes Y into $\mathscr{X}_\alpha$. The maps φ_α form a *coherent family of maps*, but it is not the only one: for any $\alpha = (A_1, \ldots, A_{N(\alpha)})$, $\mathscr{X}$ is mapped surjectively to $\mathscr{X}_\alpha$ by the φ'_α defined in (8.16) and for all $\alpha, \beta \in \mathscr{A}$ with $\alpha \leq \beta$, $\varphi_{\alpha\beta} \circ \varphi'_\beta = \varphi'_\alpha$. For every $x \in \mathscr{X}$, $(\varphi'_\alpha(x) : \alpha \in \mathscr{A}) \in \prod_{\alpha\in\mathscr{A}} \varphi'_\alpha(\mathscr{X}) = \prod_{\alpha\in\mathscr{A}} \mathscr{X}_\alpha$, is a point in the inverse limit space Y. This defines a mapping $\varphi' : \mathscr{X} \to Y$ that is neither injective nor surjective in general.

Proposition 8.6.5. *Let $\mathscr{X}$ be a Hausdorff space with a directed set of finite partitions $\mathscr{A}$ that resolves $\mathscr{X}$ and let $(\mathscr{X}_\alpha, \varphi_{\alpha\beta})$ denote the associated inverse system. The inverse limit Y is a zero-dimensional compact Hausdorff space that contains a subspace $\hat{\mathscr{X}}$ with a continuous bijection $i : \hat{\mathscr{X}} \to \mathscr{X}$ as well as a coherent family of continuous surjective maps $\hat{\varphi}_\alpha : \hat{\mathscr{X}} \to \mathscr{X}_\alpha$ such that for all $\alpha \in \mathscr{A}$,*



(8.21)

is a commutative diagram.

Proof. As said, the inverse limit Y is non-empty and compact. Since $\mathscr{A}$ is a directed set, a basis for the topology on Y is given by the collection of all sets $\varphi_\alpha^{-1}(A)$, where $\alpha \in \mathscr{A}$ and A any subset of $\mathscr{X}_\alpha$. Note that if $x, y \in \mathscr{X}$, $x \neq y$, the Hausdorff property of $\mathscr{X}$ and the assumption that $\mathscr{A}$ resolves $\mathscr{X}$, imply that there exists an $\alpha \in \mathscr{A}$ with distinct $A_x, A_y \in \alpha$ such that $x \in A_x$ and $y \in A_y$. Conclude that since the spaces $\mathscr{X}_\alpha$ are Hausdorff, Y is Hausdorff. Since any $A \subset \mathscr{X}_\alpha$ is *clopen* in $\mathscr{X}_\alpha$ and $\varphi_\alpha : Y \to \mathscr{X}_\alpha$ is continuous, $\varphi_\alpha^{-1}(A)$ is clopen in Y for any α and A. With a clopen neighbourhood basis for any $y \in Y$, Y is a *zero-dimensional space*. Again because for $x, y \in \mathscr{X}$, $x \neq y$, there exists an $\alpha \in \mathscr{A}$ with distinct $A_x, A_y \in \alpha$ such that $x \in A_x$ and $y \in A_y$, the map $\varphi' : \mathscr{X} \to Y$ is injective, and has an inverse defined on $\hat{\mathscr{X}} = \varphi'(\mathscr{X})$ which we denote by $i : \hat{\mathscr{X}} \to \mathscr{X}$. Note that the maps $\hat{\varphi}_\alpha : \hat{\mathscr{X}} \to \mathscr{X}_\alpha$, defined as the restrictions to $\hat{\mathscr{X}}$ of the projection maps $\varphi_\alpha : Y \to \mathscr{X}_\alpha$, are continuous. Because $i^{-1}(x) = \varphi'(x) = (\varphi'_\alpha(x) : \alpha \in \mathscr{A}) \in \hat{\mathscr{X}} \subset \prod_\alpha \mathscr{X}_\alpha$ and $\hat{\varphi}_\alpha$ is the restriction of $\mathrm{pr}_\alpha : \prod_\alpha \mathscr{X}_\alpha \to \mathscr{X}_\alpha$ to $\hat{\mathscr{X}}$, the relation $\hat{\varphi}_\alpha \circ i^{-1} = \varphi'_\alpha$ is immediate. Composition with i shows that diagram (8.21) is commutative. Surjectivity of i and φ'_α implies surjectivity of $\hat{\varphi}_\alpha$. As for the continuity of i, let $(x_j : j \in \mathscr{J})$ denote a net in $\hat{\mathscr{X}}$, converging to $y \in \hat{\mathscr{X}}$. Denote $z = i(y) \in X$ and U an open neighbourhood of z. By assumption, there exists a $\beta \in \mathscr{A}$ and an element $A \in \beta$, such that $A \subset U$. Since $x_j \to y$, $x_{j\alpha} \to y_\alpha$ in $\mathscr{X}_\alpha$ for all $\alpha \in \mathscr{A}$, so in particular, $x_{j\beta} \to y_\beta$. Because $\mathscr{X}_\beta$ is discrete, there exists a $J \in \mathscr{J}$ such that $x_{j\beta} = y_\beta = (0, \ldots, 0, 1, 0, \ldots, 0)$ for all $j \geq J$ (with the one corresponding to the element A in β). Therefore, $i(x_j) \in U$ for all $j \geq J$, which proves that i is continuous.

As a subspace of a compact space, $\hat{\mathscr{X}}$ is completely regular (see [47], Ch. IX, § 1, No. 5, proposition 3). If we assume $\mathscr{X}$ is not itself zero-dimensional, it is not possible for the space $\hat{\mathscr{X}}$ to be closed as a subset of Y. (For, if we assume that $\hat{\mathscr{X}}$ is closed, $\hat{\mathscr{X}}$ is compact and Hausdorff, so i has a continuous inverse. However, $\hat{\mathscr{X}}$ is zero-dimensional while $\mathscr{X}$ is not.) Therefore it is possible that there exist subsets that are compact in $\mathscr{X}$ and *not compact* in $\hat{\mathscr{X}}$. (To appreciate the prevalence of this phenomenon, see example 8.6.8.)

The essential difference between the maps $\hat{\varphi}_\alpha$ and φ'_α in diagram (8.21) is that the former are continuous. The space $\hat{\mathscr{X}}$ has a topology that is a minimal, zero-dimensional refinement of $\mathscr{X}$ that enables continuity of the coherent system associated with the maps φ_α defined by (8.16). It is illustrative to consider a space $\mathscr{X}$ which is itself zero-dimensional: in that case it is possible to let $\mathscr{A}$ consist of finite partitions of *clopen* sets $A \subset \mathscr{X}$. Consequently, the maps φ'_α form a coherent system of *continuous* maps, which means they can play the role of the maps $\hat{\varphi}_\alpha$ in subsequent steps.

To find a concrete representation of $\hat{\mathscr{X}}$, consider the following: let $\mathscr{X}$ be a (for now) completely regular space with topology $\mathscr{T}$ and basis $\mathscr{U}$. We refine $\mathscr{T}$ on the set $\mathscr{X}$ to defining a zero-dimensional space that can play the role of $\hat{\mathscr{X}}$ above: with the same underlying set $\mathscr{X}$, define a topological *subbasis*,

$$\mathscr{S} = \{U, \mathscr{X} \setminus U : U \in \mathscr{U}\}, \tag{8.22}$$

for a topology $\hat{\mathscr{T}}$ on the set $\mathscr{X}$ in which each basis element $U \in \mathscr{U}$ is *clopen*; denote the resulting topological space by $\hat{\mathscr{X}}$. In the case of a *second countable* space $\mathscr{X}$, enumerate $\mathscr{U} = \{U_n : n \geq 1\}$ and define $\mathscr{X}_0 = \mathscr{X}$ and $\mathscr{X}_n$, $(n \geq 1)$ to be the topological sum of U_n and $\mathscr{X}_{n-1} \setminus U_n$. Note that there is a corresponding sequence of partitions $\mathscr{A} = \{\alpha_n : n \geq 1\}$ (as in example 8.6.3) that are generated by the basis $\mathscr{U}$, and that the $\mathscr{X}_n$ form an inverse system of topological spaces $(\mathscr{X}_n, \varphi_{mn})$ (where the set-theoretic identity on the set $\mathscr{X}$ acts as continuous projection map φ_{mn} for all $m \leq n$) with inverse limit homeomorphic to the space $\hat{\mathscr{X}}$.

Proposition 8.6.6. *The space $\hat{\mathscr{X}}$ is zero-dimensional and the identity map $i : \hat{\mathscr{X}} \to \mathscr{X}$ is continuous. If $\mathscr{X}$ is Polish, then $\hat{\mathscr{X}}$ is also Polish.*

Proof. The subbasis $\mathscr{S}$ gives rise to a basis consisting of clopen sets, so $\hat{\mathscr{X}}$ is zero-dimensional. Furthermore the identity i is continuous because $\hat{\mathscr{T}}$ refines $\mathscr{T}$. Assuming $\mathscr{X}$ is Polish, the countable product space $\mathscr{X}^{\mathbb{N}} = \prod_{n\geq 1} \mathscr{X}$ is Polish (see [143], prop. 3.3) and has a diagonal $\Delta = \{(x, x, \ldots) \in \prod_{n\geq 1} \mathscr{X} : x \in \mathscr{X}\}$ that is a closed subspace, homeomorphic to $\mathscr{X}$. The spaces $\mathscr{X}_n$ are all Polish, since both U_n and $\mathscr{X} \setminus U_n$ are Polish and countable topological sums of Polish spaces are Polish spaces. There is a canonical set-theoretic identification $i_n : \mathscr{X}_n \to \mathscr{X}$, which is continuous: $\mathscr{X}_n$ is the refinement of $\mathscr{X}$ with $U_1, \ldots, U_n$ made clopen. The product space $\prod_{n\geq 1} \mathscr{X}_n$ is Polish and the map $j : \prod_{n\geq 1} \mathscr{X}_n \to \mathscr{X}^{\mathbb{N}}$ is continuous. Then $j^{-1}(\Delta)$ is homeomorphic to $\hat{\mathscr{X}}$ and (closed in Polish $\mathscr{X}^{\mathbb{N}}$, therefore) Polish.

Alternatively a zero-dimensional space $\hat{\mathscr{X}}$ can be constructed as the *Stone space* of all ultra-filters of the Boolean algebra $\mathbb{B}(\mathscr{U})$ generated by the basis $\mathscr{U}$. Both representations can play the role of the space $\hat{\mathscr{X}}$ in the context of theorem 8.5.5.

Proposition 8.6.7. *Let $\mathscr{X}$ be a Hausdorff completely regular space. The Borel sets on $\mathscr{X}$ and $\hat{\mathscr{X}}$ are equal and any set function μ that is a (bounded/positive/probability) Borel measure on $\mathscr{X}$ if and only if μ is a (bounded/positive/probability) Borel measure on $\hat{\mathscr{X}}$.*

Proof. Note that the Borel σ-algebra on $\mathscr{X}$ generated by the basis $\mathscr{U}$ is identical to the σ-algebra generated by $\mathscr{U}$ and its complements, which form the subbasis for $\hat{\mathscr{X}}$. Conclude that $\mathscr{X}$ and $\hat{\mathscr{X}}$ have the same Borel sets. Boundedness, positivity, being a probability measure and countable additivity are then identical as properties of set functions μ.

As was noted, inner-regularity of the Borel measure μ on $\hat{\mathscr{X}}$ implies regularity on $\mathscr{X}$, but the converse may fail because, in general, $\mathscr{X}$ has compact sets that are not compact in $\hat{\mathscr{X}}$: consider a typical compact subset of $\mathbb{R}$.

Example 8.6.8. Take $\mathscr{X} = \mathbb{R}$ in its usual topology, let $(0,1)$ be an element in the basis used in the construction of subbasis (8.22). Consider the interval $[0,1]$ (which is compact in $\mathscr{X}$). In the zero-dimensional space $\hat{\mathscr{X}}$, the interval $[0,1]$ has an infinite open cover that cannot be reduced to finite: namely, take,

$$A_0 = (-\infty, 0] \cup [1, \infty), \quad A_k = \left(\tfrac{1}{k}, 1 - \tfrac{1}{k}\right),$$

where $k \geq 1$. All of these subsets A_k, $(k \geq 0)$ belong to the basis of $\hat{\mathscr{X}}$ so they are open, and any finite sub-collection does not cover $[0,1]$. Note also that a subset A of $(0,1)$ that is closed in $\mathscr{X}$ (possibly unified with any or both of the points $\{0\}, \{1\}$) *is* compact in $\hat{\mathscr{X}}$.

Zero-dimensional compact subsets K of Polish spaces are characterized as by *Brouwer's theorem*: K can be written as a union of a subspace homeomorphic to the *Cantor space* with a finite collection of isolated points. Above we have seen that a Borel measure on a Polish space $\mathscr{X}$ is also a Borel measure on the Polish space $\hat{\mathscr{X}}$; due to the *Radon property* of Polish spaces, on both, μ is a Radon measure and comparing *inner regularity* of μ in these two roles, leads to the conclusion that there exist approximating zero-dimensional subspaces that are compact in both $\hat{\mathscr{X}}$ and $\mathscr{X}$, and that, in addition, there are approximating subspaces that are compact only in $\mathscr{X}$. In fact, every approximating K that is compact in $\mathscr{X}$ has an approximating zero-dimensional subspace $\hat{K}$ that is compact in $\hat{\mathscr{X}}$.

Lemma 8.6.9. *Let $\mathscr{X}$ be a Polish space with countable basis $\mathscr{U}$ and let $\hat{\mathscr{X}}$ be the zero-dimensional Polish space with subbasis (8.22). Let μ be a bounded, positive Borel measure on $\mathscr{X}$ and $\hat{\mathscr{X}}$. For every $\varepsilon > 0$, there exists a K compact in $\mathscr{X}$ and a $\hat{K} \subset K$ that is compact in $\hat{\mathscr{X}}$ such that,*

$$\mu(\mathscr{X} \setminus K) < \tfrac{1}{2}\varepsilon, \quad \mu(\hat{\mathscr{X}} \setminus \hat{K}) < \varepsilon.$$

Proof. Let $\varepsilon > 0$ be given; by the Radon property of Polish spaces and *inner regularity* of Radon measures, there exists a non-empty, compact K in $\mathscr{X}$ such that $\mu(\mathscr{X} \setminus K) < \varepsilon/2$. To find $\hat{K}$, we construct a suitable decreasing sequence of non-empty, compact sets in the spaces $\mathscr{X}_n$, $(n \geq 0)$, by induction: $K_0 = K$ is non-empty, compact in $\mathscr{X}_0$ and $\mu(\mathscr{X}_0 \setminus K_0) < \varepsilon/2$. Now, suppose that K_n is non-empty, compact in $\mathscr{X}_n$ with,

$$\mu(\mathscr{X}_n \setminus K_n) < \frac{2^{n+1} - 1}{2^{n+1}}\varepsilon.$$

Since K_n is compact in $\mathscr{X}_n$, any union $K_{n+1} = (K_n \setminus U_{n+1}) \cup A_n$, where A_n is closed in $K_n \cap U_{n+1}$, is compact in $\mathscr{X}_{n+1}$. Again by inner regularity of μ, there exists a sequence of compact sets in U_{n+1} that increases to U_{n+1} in μ-measure, so there is a choice for A_n such that $\mu(A_n) > \mu(K_n \cap U_{n+1}) - 2^{-(n+2)}\varepsilon$. Then,

$$\begin{aligned}\mu(K_{n+1}) &= \mu(K_n \setminus U_{n+1}) + \mu(A_n) > \mu(K_n) - 2^{-(n+2)}\varepsilon \\ &> \mu(\mathscr{X}_n) - \sum_{i=1}^{n+1} 2^{-(i+1)}\varepsilon = \mu(\mathscr{X}_n) - \frac{2^{n+2} - 1}{2^{n+2}}\varepsilon,\end{aligned}$$

completing the inductive step. Define $\hat{K}$ to be the inverse limit of the inverse system $(K_n, \varphi_{\alpha_m \alpha_n})$. The space $\hat{K}$ is non-empty, compact (as it is the inverse limit of non-empty, compact spaces) and $\mu(\hat{\mathscr{X}} \setminus \hat{K}) = \sup_{n \geq 0} \mu(\mathscr{X}_n \setminus K_n) = \varepsilon$ by monotone convergence, theorem B.3.5.

The relation with Brouwer's compact zero-dimensional *Cantor spaces* is intuitively clear: at every step of the induction in the proof of lemma 8.6.9, we split the previous compact into two closed components, so each point x in $\hat{K}$ can de identified with an infinite binary sequence that tells us at each step in which of the two closed components x lies.

To rephrase the relationship between Radon measures on $\mathscr{X}$ and $\hat{\mathscr{X}}$ somewhat, consider the associated vector spaces of Radon measures: any bounded, continuous $f : \mathscr{X} \to \mathbb{R}$ is also bounded, continuous when viewed as $f : \hat{\mathscr{X}} \to \mathbb{R}$, so there exists a linear, injective map $j : C^b(\mathscr{X}) \to C^b(\hat{\mathscr{X}})$ of norm one, and transpose to that, a bounded linear $j^t : M^b(\hat{\mathscr{X}}) \to M^b(\mathscr{X})$ of norm one (see [50], Ch. II, § 6, No. 4, proposition 5 and [50], Ch. IV, § 1, No. 3, proposition 8).

Remark 8.6.10. It is tempting to appeal to the locally convex version of the open mapping theorem (see, *e.g.*, Rudin (1991) [221], theorem 2.11) to conclude that j^t has a continuous inverse, thus rendering $M^b(\hat{\mathscr{X}})$ and $M^b(\mathscr{X})$ isomorphic as locally convex spaces. However, the space $M^b(\hat{\mathscr{X}})$ is not necessarily complete, even though $M^b_+(\hat{\mathscr{X}})$ is complete (see [49], Ch. IX, § 5, No. 4, proposition 10), and hence $M^b(\hat{\mathscr{X}})$ is not necessarily a Fréchet space. To appreciate the problem in a more concrete form, we consider $\mathscr{X} = [0,1]$ and $\hat{\mathscr{X}}$ of example 8.6.8: the sequence of probability measures (P_n) with Lebesgue probability densities $p_n : [0,1] \to \mathbb{R}$,

$$p_n(x) = n\, 1_{[0,n^{-1}]}(x),$$

converges to the *Dirac measure* δ_0 in $M^b(\mathscr{X})$, so $H = \{P_n : n \geq 1\} \cup \{\delta_0\}$ is a compact subset of $M^b(\mathscr{X})$ in Prokhorov's weak topology. However, the function $f : [0,1] \to \mathbb{R}$,

$$f(x) = 1_{\{0\}}(x),$$

is *continuous* on $\hat{\mathscr{X}}$, while $\int f\, dP_n = 0$, for all $n \geq 1$ and $\int f\, d\delta_0 = f(0) = 1$. So (P_n) does not converge to δ_0 and H is not compact in $M^b(\hat{\mathscr{X}})$ in Prokhorov's weak topology.

The proposition below notes that the projection maps for finite, clopen partitions α are continuous (as required in theorem 8.5.5), if we equip $M^b(\hat{\mathscr{X}})$ with Prokhorov's weak topology or stronger. Finally, we come back to the inverse limits associated with partitions of the sample space. Recall the compact inverse limit N associated with the inverse system $(M^1(\mathscr{X}_\alpha), \varphi_{*\alpha\beta})$ of proposition 8.5.6. Here too, all $\mu \in M^1(\hat{\mathscr{X}})$ map to points in N, but not all points in N correspond to (Radon) probability measures on $\hat{\mathscr{X}}$. To relate N and $M^1(\hat{\mathscr{X}})$ directly, consider for $\alpha \in \mathscr{A}$ with $\alpha = (A_1, \ldots, A_{N(\alpha)})$, the mapping $\hat{\varphi}_{*\alpha} : M^b(\hat{\mathscr{X}}) \to M(\mathscr{X}_\alpha)$,

$$\hat{\varphi}_{*\alpha}(\mu) = \big(\mu(A_1), \ldots, \mu(A_{N(\alpha)})\big), \tag{8.23}$$

that takes any bounded measure on $\hat{\mathscr{X}}$ into its α-histogram.

Proposition 8.6.11. *For all $\alpha \in \mathscr{A}$, assume that all $A \in \alpha$ are clopen in $\hat{\mathscr{X}}$. Then the map $\hat{\varphi}_{*\alpha} : M^b(\hat{\mathscr{X}}) \to M(\mathscr{X}_\alpha)$ that takes any bounded measure μ into its α-histogram is continuous for Prokhorov's weak topology.*

Proof. For any $\alpha \in \mathscr{A}$ and $A \in \alpha$, the indicator function $1_A : \hat{\mathscr{X}} \to [0,1]$ satisfies $1_A^{-1}(U) = A$ if U is an open neighbourhood of 1 that does not contain 0, and $1_A^{-1}(V) = \hat{\mathscr{X}} \setminus A$ if V is a open neighbourhood of 0 that does not contain 1, and (because A is clopen) both A and $\hat{\mathscr{X}} \setminus A$ are open, so 1_A is a bounded, continuous function on $\hat{\mathscr{X}}$. Therefore, $M^b(\hat{\mathscr{X}}) \to \mathbb{R} : \mu \mapsto \mu(A)$ is continuous with respect to Prokhorov's weak topology and so is $\hat{\varphi}_{*\alpha}$.

8.6.2 The double Prokhorov condition

The goal of this section, is to use theorem 8.5.5 to prove the existence of inverse limit distributions Π on $M^1(\mathscr{X})$, based on coherent inverse systems of probability measures on the (countable) inverse system $(M^1(\mathscr{X}_\alpha), \varphi_{*\alpha\beta})$. We construct Π as a Borel probability measure on $M^1(\hat{\mathscr{X}})$ (which represents the same set of Borel measures, but with a Prokhorov's weak topology that is finer) and restricting the domain to the Borel sets of $M^1(\mathscr{X})$.

To prepare theorem 8.6.12 we make the following definitions. For every compact subset $K \subset \mathscr{X}$ ($\hat{K} \subset \hat{\mathscr{X}}$) and every partition $\alpha \in \mathscr{A}$ of $\mathscr{X}$ ($\hat{\mathscr{X}}$), define,

$$K_\alpha = \cup\{A : A \in \alpha, A \cap K \neq \varnothing\}.$$

(and analogously for $\hat{K}_\alpha$). Note that $P(K_\alpha) = P_\alpha(\varphi_\alpha(K))$ (and $P(\hat{K}_\alpha) = P_\alpha(\hat{\varphi}_\alpha(\hat{K}))$). Because the spaces $\mathscr{X}_\alpha$ are finite, the spaces $M^1(\mathscr{X}_\alpha)$ are compact and homeomorphic to simplices. Consider the spaces $C(M^1(\mathscr{X}_\alpha))$ of continuous, real-valued functions on $M^1(\mathscr{X}_\alpha)$ and the continuous maps $\varphi_{*\alpha\beta} : M^1(\mathscr{X}_\beta) \to M^1(\mathscr{X}_\alpha)$ of (8.19), which induce continuous, linear maps ${\varphi^*}_{*\alpha\beta} : C(M^1(\mathscr{X}_\alpha)) \to C(M^1(\mathscr{X}_\beta))$ through ${\varphi^*}_{*\alpha\beta}(f) = f \circ \varphi_{*\alpha\beta}$, with transposes $\varphi_{**\beta\alpha} : M^b(M^1(\mathscr{X}_\beta)) \to M^b(M^1(\mathscr{X}_\alpha))$ for all $\alpha, \beta \in \mathscr{A}$, $\alpha \leq \beta$. Like before, the system $(M^1(M^1(\mathscr{X}_\alpha)), \varphi_{**\alpha\beta})$ forms an coherent inverse system with inverse limit that we denote N', projections $\varphi_{**\alpha} : N' \to M^1(M^1(\mathscr{X}_\alpha))$ and an injective embedding $M^1(M^1(\hat{\mathscr{X}})) \to \hat{M} \subset N'$ (with restrictions of the projection maps that we denote $\hat{\varphi}_{**\alpha} : \hat{M} \to M^1(M^1(\mathscr{X}_\alpha))$).

Theorem 8.6.12. *Let $\mathscr{X}$ be Polish and endow $M^1(\mathscr{X})$ with Prokhorov's weak topology. Let $\mathscr{A}$ be a countable collection of partitions generated by a countable basis that resolves $\mathscr{X}$. Let $(\mathscr{X}_\alpha, \varphi_{\alpha\beta})$ be the corresponding inverse system and let $(\Pi_\alpha, \varphi_{**\alpha\beta})$ be a coherent inverse system of probability measures on $(M^1(\mathscr{X}_\alpha), \varphi_{*\alpha\beta})$. There exists a unique Borel probability measure Π on $M^1(\mathscr{X})$ such that $\hat{\varphi}_{**\alpha}(\Pi) = \Pi_\alpha$, for all $\alpha \in \mathscr{A}$, if and only if:*

(P1) for every $\delta, \varepsilon > 0$ there exists a compact K in $\mathscr{X}$ such that for every $\alpha \in \mathscr{A}$,

$$\Pi_\alpha\big(P_\alpha(\varphi_\alpha(K)) < 1-\delta\big) < \varepsilon.$$

When property (P1) holds, $\Pi(L) = \inf\big\{\Pi_\alpha(\hat{\varphi}_{*\alpha}(L)) : \alpha \in \mathscr{A}\big\}$, for every compact set L in $M^1(\mathscr{X})$.

Remark 8.6.13. Since $M^1(\mathscr{X})$ is a closed subspace of $M_+^b(\mathscr{X})$, it is Polish (and hence *Souslin*) as a result of theorem C.8.9, so theorem C.8.11 says that $M^1(\mathscr{X})$ is a *Radon space*. That means that, in principle, the assertion of theorem 8.5.5 is stronger, in that Π is a *Radon* probability measure, rather than just Borel, but the distinction is redundant.

Proof. Let $\hat{\mathscr{X}}$ denote the zero-dimensional Polish space associated with the basis that generates $\mathscr{A}$, *c.f.* proposition 8.6.5. Proposition 8.5.6 establishes that $(\hat{\varphi}_{*\alpha}, \varphi_{*\alpha\beta})$ is a coherent system of functions and separates $M^1(\hat{\mathscr{X}})$ if $\{A \subset \mathscr{X} : A \in \alpha, \alpha \in \mathscr{A}\}$ generates the Borel σ-algebra for $\hat{\mathscr{X}}$, *c.f.* the *Carathéodory extension*. This is the case, since $\mathscr{A}$ is generated by the basis and resolves $\mathscr{X}$, so that all sets in the basis can be written as (unions of) sets in the partitions α. According to proposition 8.6.11, the histogram projections $\hat{\varphi}_{*n}$ are continuous, so the Hausdorff space $M^1(\hat{\mathscr{X}})$ may play the role of T in theorem 8.5.5. Next, consider *property (P)* of theorem 8.5.5. Let $\varepsilon > 0$ be given and define $\varepsilon_n = 2^{-n}\varepsilon$ for $n \geq 1$. Given some sequence (δ_n) such that $\delta_n > 0$, $\delta_n \downarrow 0$, let K_n, $(n \geq 1)$, be compact subsets of $\mathscr{X}$ such that,

$$\Pi_\alpha\big(P_\alpha(\varphi_\alpha(K_n)) < 1 - \tfrac{1}{2}\delta_n\big) < \tfrac{1}{2}\varepsilon_n,$$

for every $\alpha \in \mathscr{A}$. According to lemma 8.6.15, there exist compact subsets $\hat{K}_n$, $(n \geq 1)$, in $\hat{\mathscr{X}}$ such that,

$$\Pi_\alpha(P_\alpha(\hat{\varphi}_\alpha(\hat{K}_n)) < 1 - \delta_n) < \varepsilon_n,$$

for every $\alpha \in \mathscr{A}$. Define the set,

$$H = \bigcap_{n\geq 1}\{P \in M^1(\hat{\mathscr{X}}) : P(\hat{K}_n) \geq 1 - \delta_n\}.$$

The set H is relatively compact in $M^1(\hat{\mathscr{X}})$; namely, let $\delta > 0$ be given and choose $n \geq 1$ such that $\delta_n < \delta$. Then $\inf\{P(\hat{K}_n) : P \in H\} \geq 1 - \delta$, showing that the (bounded) set H is *uniformly tight* and theorem C.7.15 establishes that the closure $\overline{H}$ of H is compact in $M^1(\hat{\mathscr{X}})$. For any $m \geq 1$, we have,

$$\begin{aligned}\Pi_\alpha\big(M^1(\mathscr{X}_\alpha) \setminus \hat{\varphi}_{*\alpha}(\overline{H})\big) &\leq \Pi_\alpha\left(\bigcup_{n=1}^{\infty}\{P_\alpha \in M^1(\mathscr{X}_\alpha) : P_\alpha(\hat{\varphi}_\alpha(\hat{K}_n)) < 1 - \delta_n\}\right)\\ &\leq \sum_{n=1}^{\infty}\Pi_\alpha\big(P_\alpha(\hat{\varphi}_\alpha(\hat{K}_n)) < 1 - \delta_n\big) < \varepsilon.\end{aligned}$$

which shows that condition *(P)* of theorem 8.5.5 is satisfied for $\overline{H}$. Conclude that there exists a Radon (which is equivalent to Borel, *c.f.* theorem C.8.11) probability

measure $\hat{\Pi}$ on $M^1(\hat{\mathscr{X}})$ such that $\hat{\varphi}_{**\alpha}(\hat{\Pi}) = \Pi_\alpha$ for all $\alpha \in \mathscr{A}$. Since the restriction of $j^t : M^b(\hat{\mathscr{X}}) \to M^b(\mathscr{X})$ to $M^1(\hat{\mathscr{X}})$ is continuous, $\hat{\Pi}$ induces a Borel probability measure $\Pi = \hat{\Pi} \circ (j^t)^{-1}$ on $M^1(\mathscr{X})$ with the same distributions Π_α for its α-histograms.

Conversely, let Π be a Borel measure on $M^1(\mathscr{X})$ such that $\hat{\varphi}_{**\alpha}(\Pi) = \Pi_\alpha$, for all $\alpha \in \mathscr{A}$. Since $M^1(\mathscr{X})$ is a Radon space (as per remark 8.6.13), Π is a Radon measure and inner regularity implies that for every $\varepsilon > 0$, there exists a compact H in $M^1(\mathscr{X})$ such that,

$$\Pi\big(M^1(\mathscr{X}) \setminus H\big) < \varepsilon.$$

By theorem C.7.16, H is uniformly tight, which means that for any $\delta > 0$, there exists a compact K in $\mathscr{X}$ such that for any $P \in H$, $P(\mathscr{X} \setminus K) < \delta$. Combining these two points, for any $\varepsilon, \delta > 0$, there exists a compact K in $\mathscr{X}$ such that,

$$\Pi\big(P(K) < 1 - \delta\big) < \varepsilon.$$

Since $K \subset K_\alpha$ and $P(K_\alpha) = P_\alpha(\varphi_\alpha(K))$ for all $\alpha \in \mathscr{A}$, we have,

$$\Pi_\alpha\big(P_\alpha(\varphi_\alpha(K)) < 1 - \delta\big) = \Pi\big(P(K_\alpha) < 1 - \delta\big) \leq \Pi\big(P(K) < 1 - \delta\big) < \varepsilon,$$

proving *property (P1)*.

Remark 8.6.14. Although trivial, it is important to make the following observation explicit: once a unique distribution Π on $M^1(\mathscr{X})$ is fixed by any (*e.g.* countable) collection of random histograms, *random histograms for Borel measurable partitions are all fixed*: given any Borel measurable partition $\alpha = \{A_1, \ldots, A_N\}$, Π fixes the distribution Π_α for P_α as in (8.4) and coherence of the resulting inverse system of measures $(\Pi_\alpha, \varphi_{**\alpha\beta})$ is implied by theorem 8.1.3.

Below we give the lemma that provides the zero-dimensional version of *property (P1)* needed in the proof of theorem 8.6.12.

Lemma 8.6.15. *Let $\mathscr{X}$ be a Polish space with countable basis $\mathscr{U}$ and let $\hat{\mathscr{X}}$ be the zero-dimensional Polish space with subbasis (8.22). Let $\mathscr{A}$ be a countable collection of partitions generated by $\mathscr{U}$ that resolves $\mathscr{X}$. For given $\delta, \varepsilon > 0$, let K be a compact subset of $\mathscr{X}$ such that,*

$$\Pi_\alpha\big(P_\alpha(\varphi_\alpha(K)) < 1 - \tfrac{1}{2}\delta\big) < \tfrac{1}{2}\varepsilon,$$

for all $\alpha \in \mathscr{A}$. Then there exists a compact subset $\hat{K}$ of $\hat{\mathscr{X}}$ such that,

$$\Pi_\alpha\big(P_\alpha(\hat{\varphi}_\alpha(\hat{K})) < 1 - \delta\big) < \varepsilon,$$

for all $\alpha \in \mathscr{A}$.

Proof. Let $\delta, \varepsilon > 0$ and K be as assumed. We assume that the partitions $\alpha \in \mathscr{A}$ form the sequence (α_n) as defined in example 8.6.3, since any partition α generated by the basis $\mathscr{U}$ is a coarsening of a partition α_n for some $n \geq 1$. To find $\hat{K}$, we

construct a suitable decreasing sequence of non-empty, compact sets in the spaces $\mathscr{X}_n$, $(n \geq 0)$, by induction: the assumption asserts that for $\mathscr{X}_0 = \mathscr{X}$, $K_0 = K$ is a compact set such that,

$$\Pi_{\alpha_0}\big(P_{\alpha_0}(\hat{\varphi}_{\alpha_0}(K)) < 1 - \tfrac{1}{2}\delta\big) < \tfrac{1}{2}\varepsilon,$$

Now, suppose that K_n is non-empty, compact in $\mathscr{X}_n$ with,

$$\Pi_{\alpha_m}\Big(P_{\alpha_m}(\hat{\varphi}_{\alpha_m}(K_n)) < 1 - \frac{2^{n+1}-1}{2^{n+1}}\delta\Big) < \frac{2^{n+1}-1}{2^{n+1}}\varepsilon,$$

for all $0 \leq m \leq n$. Since K_n is compact in $\mathscr{X}_n$, any union $K_{n+1} = (K_n \setminus U_{n+1}) \cup A_{n+1}$, where A_{n+1} is closed in $K_n \cap U_{n+1}$, is compact in $\mathscr{X}_{n+1}$. Letting A_{n+1} grow to U_{n+1}, $P_{\alpha_{n+1}}(\hat{\varphi}_{\alpha_{n+1}}(A_{n+1}))$ grows to $P_{\alpha_{n+1}}(\hat{\varphi}_{\alpha_{n+1}}(U_{n+1}))$, for every $P_{\alpha_{n+1}} \in M^1(\mathscr{X}_{n+1})$, by inner regularity. Define M_{n+1} to be the set of $P_{\alpha_{n+1}} \in M^1(\mathscr{X}_{n+1})$ for which,

$$\varphi_{*\,\alpha_m\alpha_{n+1}}(P_{\alpha_{n+1}})\big(\hat{\varphi}_{\alpha_m}(K_n)\big) \geq 1 - \frac{2^{n+1}-1}{2^{n+1}}\delta,$$

for all $0 \leq m \leq n$. The set M'_{n+1} of $P_{\alpha_{n+1}} \in M_{n+1}$ such that, for all $0 \leq m \leq n+1$,

$$\begin{aligned}\varphi_{*\,\alpha_m\alpha_{n+1}}(P_{\alpha_{n+1}})\big(\hat{\varphi}_{\alpha_m}(K_{n+1})\big) &\geq P_{\alpha_{n+1}}\big(\hat{\varphi}_{\alpha_{n+1}}(K_n \setminus U_{n+1})\big) + P_{\alpha_{n+1}}\big(\hat{\varphi}_{\alpha_{n+1}}(A_{n+1})\big)\\ &> P_{\alpha_{n+1}}\big(\hat{\varphi}_{\alpha_{n+1}}(K_n)\big) - 2^{-(n+2)}\delta = 1 - \frac{2^{n+2}-1}{2^{n+2}}\delta,\end{aligned}$$

grows to M_{n+1} as A_{n+1} grows to U_{n+1}. By assumption,

$$\Pi_{\alpha_m}\big(\varphi_{*\,\alpha_m\alpha_{n+1}}(M_{n+1})\big) \geq 1 - \frac{2^{n+1}-1}{2^{n+1}}\varepsilon.$$

for all $0 \leq m \leq n+1$. Because the measures Π_{α_m}, $(0 \leq m \leq n+1)$, are countably additive and theorem B.2.7, letting A_{n+1} grow until $\Pi_{\alpha_m}(\varphi_{\alpha_m\alpha_{n+1}}(M'_{n+1}))$ lies only $2^{-(n+2)}\varepsilon$ below the above lower bound, for all $0 \leq m \leq n+1$, leads to,

$$\Pi_{\alpha_m}\Big(P_{\alpha_m}(\hat{\varphi}_{\alpha_m}(K_{n+1})) < 1 - \frac{2^{n+2}-1}{2^{n+2}}\delta\Big) < \frac{2^{n+2}-1}{2^{n+2}}\varepsilon,$$

for all $0 \leq m \leq n+1$, completing the inductive step. Define $\hat{K}$ to be the inverse limit of the inverse system $(K_n, \varphi_{\alpha_m\alpha_n})$. The space $\hat{K}$ is non-empty, compact (as it is the inverse limit of non-empty, compact spaces) and

$$\Pi_{\alpha_n}\big(P_{\alpha_n}(\hat{\varphi}_{\alpha_n}(\hat{K})) < 1 - \delta\big) < \varepsilon,$$

for all $n \geq 0$.

8.6.3 Existence of inverse limits on compact spaces

Below we illustrate the intuition behind *property (P1)* with a counterexample in which probability mass of the histogram distributions P_α vanishes to infinity with non-zero probability in the limit. This eventuality is impossible if the underlying space $\mathscr{X}$ is compact, and we formulate two corollaries to accommodate the exceptional cases of *property (P1)* with a different existence result.

Example 8.6.16. Here we refer to the sequence (α_n) of partitions for $\mathbb{R}$ of example 8.6.4, each of which contains a element of the form,

$$A_{0,n} = (-\infty, -q_n) \cup (q_n, \infty),$$

with $q_n \to \infty$. We give a coherent definition for histograms distributions $\Pi_n = \Pi_{\alpha_n}$, $n \geq 1$, that does not satisfy *property (P1)*. For all $n \geq 1$, consider degenerate marginals Π_{α_n} which assign P_{α_n}-mass one to the set $A_{0,n}$, with Π_{α_n}-probability one:

$$\Pi_{\alpha_n}\big(P_{\alpha_n}(\{(1,0,\ldots,0)\}) = 1\big) = 1.$$

The resulting $(\Pi_{\alpha_n}, \varphi_{\alpha_m \alpha_n})$ forms a coherent inverse system of measures on the inverse system $(M^1(\mathscr{X}_{\alpha_n}), \varphi_{*\,\alpha_m \alpha_n})$. For any compact K in $\mathbb{R}$, $K \cap A_{n,0} = \varnothing$ for large enough n, which invalidates *property (P1)* for $\varepsilon < 1$. Explained more intuitively, the problem occurs because the marginals P_{α_n} on which the Π_{α_n} fixate, shift mass towards $\pm\infty$ without limitation, which would mean that compact sets K receive mass zero from any presumed limit measure. There is no such limit, because *any* Borel measure on $\mathbb{R}$ is inner regular. Counterexamples need not be this extreme: it is enough that for some $\delta > 0$,

$$\lim_{n\to\infty} \Pi_{\alpha_n}\big(P_{\alpha_n}(\{(1,0,\ldots,0)\}) > \delta\big) > 0,$$

to invalidate *property (P1)*.

To interpret and accommodate the above counterexample in a more concrete fashion, we formulate the following straightforward corollaries of theorem 8.6.12.

Corollary 8.6.17. *Let $\mathscr{X}$ be compact metrizable and endow $M^1(\mathscr{X})$ with Prokhorov's weak topology. Let $\mathscr{A}$ be a collection of partitions generated by a countable basis. Let $(\mathscr{X}_\alpha, \varphi_{\alpha\beta})$ be the corresponding inverse system and let $(\Pi_\alpha, \varphi_{**\,\alpha\beta})$ be a coherent inverse system of probability measures on $(M^1(\mathscr{X}_\alpha), \varphi_{*\,\alpha\beta})$. Then there exists a unique Borel probability measure Π on $M^1(\mathscr{X})$ such that $\hat{\varphi}_{**\,\alpha}(\Pi) = \Pi_\alpha$ for all $\alpha \in \mathscr{A}$.*

Proof. The space $\mathscr{X}$ is Polish and can play the role of all sets K in *property (P1)*. Alternatively, note that in $M^b(\mathscr{X})$, $M^1(\mathscr{X})$ is a closed, bounded, uniformly tight subset. According to theorem C.7.15, $M^1(\mathscr{X})$ is compact and can play the role of all sets H in property *(P)*.

Corollary 8.6.18. *Let $\mathscr{X}$ be a Polish space and let $\mathscr{A}$ be a collection of partitions generated by a countable basis. Let $(\mathscr{X}_\alpha, \varphi_{\alpha\beta})$ be the corresponding inverse system and let $(\Pi_\alpha, \varphi_{**\alpha\beta})$ be a coherent inverse system of probability measures on $(M^1(\mathscr{X}_\alpha), \varphi_{*\alpha\beta})$. Then there exists a compactification $\mathscr{Y}$ of $\mathscr{X}$ and a (possibly non-unique) Borel probability measure Π on $M^1(\mathscr{Y})$ with Prokhorov's weak topology, such that $\hat{\varphi}_{**\alpha}(\Pi) = \Pi_\alpha$ for all $\alpha \in \mathscr{A}$. If $\mathscr{X}$ is locally compact and σ-compact, $\mathscr{Y}$ can be chosen equal to the one-point-compactification and Π is then unique.*

Proof. Every Polish space is homeomorphic to a *G_δ-subspace* A of the *Hilbert cube* and $\mathscr{Y}$ can be chosen equal to its closure $\bar{A}$ (see theorem C.4.9 and the remark preceding it). The subset $\mathscr{Y} \setminus A$ is Borel measurable and we construct the zero-dimensional space $\hat{\mathscr{Y}}$ as the topological sum of $\mathscr{Y} \setminus A$ and A, followed by refinement by the sub-basis (8.22). The set $\mathscr{Y} \setminus A$ can also be added to each of the partitions in $\mathscr{A}$ to form a collection of partitions of $\hat{\mathscr{Y}}$ (although these partitions do not necessarily separate $M^1(\hat{\mathscr{Y}})$ and the ring they generate may not generate the full Borel σ-algebra on $\hat{\mathscr{Y}}$). Using the generalization of the *Bourbaki-Prokhorov-Schwartz theorem* 8.5.5 that does not require separation of T by the maps φ_α (see the remark following theorem 1 of [49], Ch. IX, § 4, No. 2), we see that corollary 8.6.17 continues to be valid, however, without fixing the Borel probability measure Π on $M^1(\mathscr{Y})$ uniquely: the mass $P \sim \Pi$ assigns to $\mathscr{Y} \setminus A$ is fixed as in example 8.6.16, but exactly how it is spread out among the Borel subsets of $\mathscr{Y} \setminus A$ is left undetermined by the inverse system on $\mathscr{X}$ and implies non-uniqueness of the inverse limit. If $\mathscr{X}$ is locally compact and σ-compact, there exists a Polish *one-point-compactification* $\mathscr{Y} = A \cup \{\omega\}$ of $\mathscr{X}$ (see [46], Ch. I, § 9, No. 8, theorem 4 and [46], Ch. I, § 9, No. 9, corollary 2) and the only possible inverse limit Π is a unique convex combination of a Borel measure concentrated on $\mathscr{X}$ with the Dirac measure δ_ω.

So in cases where *property (P1)* is violated, a weaker existence result holds, requiring the concession that the space $\mathscr{X}$ be extended to accommodate the probability mass that vanishes to infinity in example 8.6.16.

8.6.4 Dirichlet process distributions and other examples

To illustrate the uses of theorem 8.6.12, we finally prove the existence of the Dirichlet process distributions that were introduced in section 8.2 and the Pólya tree distributions of section 8.3.

Note that in the formulation of existence of the Dirichlet process distribution, we use the collection of *all measurable partitions* in the formulation of random histograms. Although this is not necessary (a collection of partitions generated by a basis that resolves $\mathscr{X}$ is sufficient for the application of theorem 8.6.12, see also remark 8.6.14), in the case of the Dirichlet process, it is the most natural perspective.

Theorem 8.6.19. *(Existence of Dirichlet process distributions)*
Let $\mathscr{X}$ be a Polish space, endow $M^1(\mathscr{X})$ with Prokhorov's weak topology and let

μ be a bounded, positive Borel measure on $\mathscr{X}$. There exists a Borel probability measure $D_\mu \in M^1(M^1(\mathscr{X}))$ that is the unique inverse limit measure with histogram distributions (8.5).

Proof. Let $\mathscr{U}$ be a countable basis for $\mathscr{X}$ and let $\mathscr{A}$ be a countable collection of partitions generated by $\mathscr{U}$ that resolves $\mathscr{X}$. By assumption there exist distributions $D_{\mu,\alpha}$ for the corresponding random histograms $P_\alpha \in M^1(\mathscr{X}_\alpha)$, $(\alpha \in \mathscr{A})$. Coherence of the inverse system $(D_{\mu,\alpha}, \varphi_{**\alpha\beta})$ has been proved in theorem 8.2.1.

Let $\varepsilon > 0, \delta > 0$ be given. Since $\mathscr{X}$ is Polish, the Borel measure μ is a Radon measure and inner regularity implies that there exists a compact subset K in $\mathscr{X}$ such that,

$$\mu(\mathscr{X} \setminus K) < \delta\,\varepsilon\,\mu(\mathscr{X}) < \infty.$$

Let α be given. By Markov's inequality and the fact that under Π_α, $P_\alpha(A') \sim \text{Beta}(\mu(A), \mu(\mathscr{X} \setminus A))$ for any $A' \subset \mathscr{X}_\alpha$ with $A = \varphi_\alpha^{-1}(A')$ (see example 3.6.2),

$$\begin{aligned} D_{\mu,\alpha}\big(P_\alpha(\mathscr{X}_\alpha \setminus \varphi_\alpha(K)) > \delta\big) &\leq \frac{1}{\delta}\int P_\alpha(\mathscr{X}_\alpha \setminus \varphi_\alpha(K))\, dD_{\mu,\alpha}(P_\alpha) \\ &= \frac{1}{\delta}\frac{\mu(\mathscr{X} \setminus K_\alpha)}{\mu(\mathscr{X})} \leq \frac{1}{\delta}\frac{\mu(\mathscr{X} \setminus K)}{\mu(\mathscr{X})} < \varepsilon, \end{aligned}$$

since $K \subset K_\alpha$.

Existence of Pólya tree distributions as Borel probability distributions on $M^1([0,1])$ with Prokhorov's weak topology follows directly from corollary 8.6.17. The existence on $M^1(\mathbb{R})$ (or $M^1((0,1))$ is analysed with theorem 8.6.12. The result is best understood from the perspective sketched in example 8.6.16: if the sets $A_{00\ldots}$ and $A_{11\ldots}$ do not retain any probability mass in the limit that $n \to \infty$ with high Π_{α_m}-probability, then the Pólya tree distribution Π is well-defined.

Theorem 8.6.20. *Consider $M^1(\mathbb{R})$ with Prokhorov's weak topology. Let $\mathscr{A}$ be a dyadic tree of partitions of $\mathbb{R}$ as in example 8.3.2 and for every $\varepsilon \in \mathscr{E}$, let $\beta_{\varepsilon 0}, \beta_{\varepsilon 1} > 0$ be the parameters for splitting variables $V_{\varepsilon 0} \sim \text{Beta}(\beta_{\varepsilon 0}, \beta_{\varepsilon 1})$, leading to the inverse system (8.10) of measures Π_{α_m}, $m \geq 0$. Then there exist a unique Borel probability measure Π on $M^1(\mathbb{R})$ such that $\hat{\varphi}_{**\alpha_m}(\Pi) = \Pi_{\alpha_m}$, for all $m \geq 0$, if and only if,*

$$\prod_{m\geq 0} \frac{\beta_{\varepsilon_m 0}}{\beta_{\varepsilon_m 0} + \beta_{\varepsilon_m 1}} = 0, \quad \prod_{m\geq 0} \frac{\beta_{\varepsilon'_m 1}}{\beta_{\varepsilon'_m 0} + \beta_{\varepsilon'_m 1}} = 0, \tag{8.24}$$

with $\varepsilon_m = 0\ldots0 \in \mathscr{E}_m$ and $\varepsilon'_m = 1\ldots1 \in \mathscr{E}_m$, for all $m \geq 0$.

Proof. A dyadic trees of partitions as in example 8.3.2 is generated by a basis for $\mathbb{R}$ and resolves $\mathbb{R}$. Moreover, for every $m \geq 1$, $K_m = \mathbb{R} \setminus (A_{0\ldots0} \cup A_{1\ldots1}) = [-a_m, a_m]$ forms an increasing sequence of compacta that are unions of sets in α_m, and the sets K_m, $(m \geq 1)$, cover $\mathbb{R}$. Because every compact K in $\mathbb{R}$ is contained in K_m for m large enough, and because $P_{\alpha_n}(\varphi_{\alpha_n}(K_m))$ has the same distribution under Π_{α_n} as $P_{\alpha_m}(\varphi_{\alpha_m}(K_m))$ under Π_{α_m}, for all $n \geq m$, we can re-write *property (P1)* in the following form: for every $\delta, \varepsilon > 0$ there exists an $m \geq 1$, such that,

$$\Pi_{\alpha_m}\big(P_{\alpha_m}(\varphi_{\alpha_m}(K_m)) < 1-\delta\big) = \Pi_{\alpha_m}\big(P_{\alpha_m}(A_{0\ldots0}) + P_{\alpha_m}(A_{1\ldots1}) > \delta\big) < \varepsilon.$$

This condition is more easily interpretable when expressed in terms of convergence in Π_{α_m}-probability: there exist a unique Borel probability measure Π on $M^1(\mathbb{R})$ such that $\hat{\varphi}_{**\,\alpha_m}(\Pi) = \Pi_{\alpha_m}$ for all $m \geq 0$, if and only if,

$$P_{\alpha_m}(A_{0\ldots0}) \xrightarrow{\Pi_{\alpha_m}} 0, \quad P_{\alpha_m}(A_{1\ldots1}) \xrightarrow{\Pi_{\alpha_m}} 0. \tag{8.25}$$

Because $P_{\alpha_m}(A_{0\ldots0})$ and $P_{\alpha_m}(A_{1\ldots1})$ are bounded, condition (8.25) holds, if and only if,

$$\int P_{\alpha_m}(A_{0\ldots0})\, d\Pi_{\alpha_m}(P_{\alpha_m}) \to 0, \quad \int P_{\alpha_m}(A_{1\ldots1})\, d\Pi_{\alpha_m}(P_{\alpha_m}) \to 0 \tag{8.26}$$

which amounts to condition (8.24), *c.f.* equation (8.11).

Remark 8.6.21. Existence of Pólya tree distributions as Borel probability distributions on $M^1((0,1))$ with Prokhorov's weak topology is analogous: again we require that the dyadic tree of partitions consists of a finite collection of intervals, as in example 8.3.1. Based on example 8.6.16 and the proof of theorem 8.6.20, it is then clear that the probabilities of intervals of the forms $A_{0\ldots0} = (0,u)$ and $A_{1\ldots1} = (l,1)$ have to satisfy (8.26) as well, leading to the same condition (8.24).

Note that existence of the Pólya tree distribution does not depend on the absolute sizes of the β parameters, only on their ratios: for any $B_m > 0$, $(m \geq 0)$, simultaneous re-scalings of the form $(\beta_{\varepsilon_m 0}, \beta_{\varepsilon_m 1}) \mapsto (B_m\beta_{\varepsilon_m 0}, B_m\beta_{\varepsilon_m 1})$ leave condition (8.24) invariant.

Example 8.6.22. To show an example of a coherent system of Pólya tree random histograms that does not give rise to a well-defined inverse limit distribution, consider example 8.3.3 *(iv)*: take $\mathscr{X} = \mathbb{R}$ with a *dyadic tree* of partitions as defined in example 8.3.2, and, for all $m \geq 0$, $\varepsilon \in \mathscr{E}_m$,

$$\beta_{\varepsilon 0} = \cos\big(\tfrac{1}{2}\pi x(\varepsilon)\big), \quad \beta_{\varepsilon 1} = \sin\big(\tfrac{1}{2}\pi x(\varepsilon)\big),$$

(see example C.4.8 for the definition of the *Cantor mid-point function* $x : \mathscr{E} \to [0,1]$). Note that if we let $\varepsilon_m = 0\ldots0 \in \mathscr{E}_m$ for all $m \geq 1$,

$$\begin{aligned}\prod_{m\geq 0}\frac{\beta_{\varepsilon_m 0}}{\beta_{\varepsilon_m 0}+\beta_{\varepsilon_m 1}} &= \prod_{m\geq 0}\frac{\cos\big(\frac{1}{2}\pi x(\varepsilon_m)\big)}{\cos\big(\frac{1}{2}\pi x(\varepsilon_m)\big)+\sin\big(\frac{1}{2}\pi x(\varepsilon_m)\big)} \\ &= \prod_{m\geq 0}\Big(1+\tan\big(\tfrac{1}{2}\pi x(\varepsilon_m)\big)\Big)^{-1} = \exp\Big(-\sum_{m\geq 0}\log\Big(1+\tan\big(\tfrac{1}{2}\pi x(\varepsilon_m)\big)\Big)\Big).\end{aligned}$$

It is noted that $x(\varepsilon_m) = 1/2(1/3)^m$ and,

$$\sum_{m\geq 0}\log\Big(1+\tan\big(\tfrac{1}{2}\pi x(\varepsilon_m)\big)\Big)\approx\sum_{m\geq 0}\tan\big(\tfrac{1}{2}\pi x(\varepsilon_m)\big)$$
$$\approx\frac{\pi}{2}\sum_{m\geq 0}x(\varepsilon_m)=\frac{\pi}{4}\sum_{m\geq 0}\Big(\frac{1}{3}\Big)^m=\frac{3\pi}{8}<\infty,$$

Similarly, if we let $\varepsilon'_m=1\ldots 1\in\mathscr{E}_m$ for all $m\geq 0$,

$$\prod_{m\geq 0}\frac{\beta_{\varepsilon'_m 1}}{\beta_{\varepsilon'_m 0}+\beta_{\varepsilon'_m 1}}=\prod_{m\geq 0}\frac{\sin\big(\frac{1}{2}\pi x(\varepsilon'_m)\big)}{\cos\big(\frac{1}{2}\pi x(\varepsilon'_m)\big)+\sin\big(\frac{1}{2}\pi x(\varepsilon'_m)\big)}$$
$$=\prod_{m\geq 0}\Big(1+1/\tan\big(\tfrac{1}{2}\pi x(\varepsilon'_m)\big)\Big)^{-1}=\exp\Big(-\sum_{m\geq 0}\log\Big(1+1/\tan\big(\tfrac{1}{2}\pi x(\varepsilon'_m)\big)\Big)\Big)$$

Since $x(\varepsilon'_m)=1-x(\varepsilon_m)$ for all $m\geq 0$,

$$1/\tan\big(\tfrac{1}{2}\pi x(\varepsilon'_m)\big)=1/\tan\big(\tfrac{1}{2}\pi(1-x(\varepsilon_m))\big)=\tan\big(\tfrac{1}{2}\pi x(\varepsilon_m)\big).$$

Conclude that,

$$\prod_{m\geq 0}\frac{\beta_{\varepsilon_m 0}}{\beta_{\varepsilon_m 0}+\beta_{\varepsilon_m 1}}=\prod_{m\geq 0}\frac{\beta_{\varepsilon'_m 1}}{\beta_{\varepsilon'_m 0}+\beta_{\varepsilon'_m 1}}>0,$$

which implies that the Pólya tree random histograms defined in example 8.3.3 *(iv)* form a coherent system that *does not* lead to a Borel probability measure on $M^1(\mathbb{R})$ with Prokhorov's weak topology. A similar construction can be based on $\mathscr{X}=(0,1)$ with a dyadic tree of partitions as in example 8.3.1, *c.f.* remark 8.6.21. In both cases, a compactification of $\mathscr{X}$ is needed to accommodate the inverse system, as in corollary 8.6.18.

Remark 8.6.23. Condition (8.24) is weaker than condition *(ii)* on p. 49 of section 3.7 of Ghosal and van der Vaart (2017) [110] (which is based on sufficiency as in theorem 3.9 on p. 38 therein).

8.7 Inverse limit measures in the Le Cam-Schwartz topology

Consider the case where $T=M^1(\mathscr{X})$ with topology $\mathscr{T}_1$, *i.e.* the subspace topology that $M^1(\mathscr{X})$ inherits from $M(\mathscr{X})$ when placed in duality with the space of all bounded Borel-measurable functions f on $\mathscr{X}$. (It is noted that $M(\mathscr{X})$ and $M(\hat{\mathscr{X}})$ are isomorphic as locally convex spaces in this duality and we identify them throughout below.) Let $\mathscr{A}$ denote the directed set of all finite measurable partitions of $\mathscr{X}$. Then the maps $\hat{\varphi}_{*\alpha}$ that project $M(\mathscr{X})$ onto the $M(\mathscr{X}_\alpha)$ are continuous, separating and they form a coherent system. Theorem 8.5.5 then takes the following form.

Theorem 8.7.1. *Let $\mathscr{X}$ be a completely regular space and consider $M^1(\mathscr{X})$ with the $\mathscr{T}_1$ topology. Let $\mathscr{A}$ be a refining collection of Borel measurable partitions that*

resolves $\mathscr{X}$. *Let* $(\mathscr{X}_n, \varphi_{nm})$ *be the corresponding inverse system and assume that* $(\Pi_\alpha, \varphi_{**\alpha\beta})$ *form a coherent system of probability measures on the inverse system* $(M^1(\mathscr{X}_\alpha), \varphi_{*\alpha\beta})$. *There exists a unique Radon probability measure* Π *on* $M^1(\mathscr{X})$ *such that* $\varphi_{**\alpha}(\Pi) = \Pi_\alpha$ *for all* $\alpha \in \mathscr{A}$, *if and only if,*

(P3) for every $\varepsilon, \delta > 0$, *there is finite, positive Borel measure* $Q \in M_+(\mathscr{X})$ *and a constant* $L > 0$ *such that,*

$$\Pi_\alpha(\{P_\alpha \in M^1(\mathscr{X}_\alpha) : \|P_\alpha - P_\alpha \wedge LQ_\alpha\| > \delta\}) < \varepsilon,$$

for all $\alpha \in \mathscr{A}$.

When (P3) holds, $\Pi(M) = \inf\{\Pi_\alpha(\hat{\varphi}_{*\alpha}(M)) : \alpha \in \mathscr{A}\}$, *for every compact set* M *in* $M^1(\mathscr{X})$.

Proof. For all $\alpha \in \mathscr{A}$, $\varphi_{*\alpha} : M^1(\hat{\mathscr{X}}) \to M^1(\mathscr{X}_\alpha)$ is continuous with respect to the Le Cam-Schwartz topology. Moreover, $(\varphi_{*\alpha}, \varphi_{\alpha\beta})$ is a coherent and separating family. The assertion now follows from theorem 8.5.5 if we can show that property *(P)* of theorem 8.5.5 holds. To that end, let $\varepsilon > 0$ be given and define $\varepsilon_n = 2^{-n}\varepsilon$. Given some sequence (δ_n) such that $\delta_n > 0$, $\delta_n \downarrow 0$, let Q_n be finite, positive Borel measures on $\mathscr{X}$ and L_n positive constants such that,

$$\Pi_\alpha(\|P_\alpha - P_\alpha \wedge L_n Q_{n,\alpha}\| > \delta_n\}) < \varepsilon_n,$$

for every $\alpha \in \mathscr{A}$. Define the set,

$$H = \bigcap_{n\geq 1} \{P \in M^1(\mathscr{X}) : \|P - P \wedge L_n Q_n\| > \delta_n\},$$

Let $\delta > 0$ be given and choose $n \geq 1$ such that $\delta_n < \delta$ and define $Q = Q_n$, $L = L_n$. Then, $\sup\{\|P - P \wedge LQ\| : P \in H\} < \delta$, so H is relatively compact with respect to the Le Cam-Schwartz topology, according to the Dunford-Pettis-Grothendieck theorem. For any $\alpha \in \mathscr{A}$, we have,

$$\begin{aligned}\Pi_\alpha\Big(\bigcup_{n=1}^{\infty} \{P_\alpha \in M^1(\mathscr{X}_\alpha) : \|P_\alpha - P_\alpha \wedge L_n Q_{n,\alpha}\| > \delta_n\}\Big)& \\ \leq \sum_{n=1}^{\infty} \Pi_\alpha\big(\{P_\alpha \in M^1(\mathscr{X}_\alpha) : \|P_\alpha - P_\alpha \wedge L_n Q_{n,\alpha}\| > \delta_n\}\big) < \varepsilon.&\end{aligned}$$

which shows that condition *(P)* of theorem 8.5.5 is satisfied for the compact closure $\overline{H}$ of H.

Given the inaccessible nature of property *(P3)*, is it really necessary to prove the Radon property for Π? One might be content with the assertion that Π is a Borel measure on $M^1(\mathscr{X})$. Aside from the Radon-Nykodym theorem and the existence of a support, we argue that the Radon property is indispensable due to the intended applications, *e.g.* in Bayesian statistics and machine-learning, where *conditioning* is

required to turn an (inverse limit) prior distribution into a posterior distribution. In [228], conditions are formulated for existence of an appropriate *disintegration*: to summarize the application in the case of the Bayesian posterior, it is required that the product measure Π' defined by $\Pi'(A \times B) = \int_B P(A)\, d\Pi(P)$ for Borel sets A in $\mathscr{X}$ and B in $M^1(\mathscr{X})$, is Radon on $\mathscr{X} \times M^1(\mathscr{X})$, and there exists a sequence K_n of metrizable compacta in $\mathscr{X} \times M^1(\mathscr{X})$ such that $\Pi(\cup_n K_n) = 1$ (see, *e.g.*, theorems 43, 44 in [228]). We do not analyse the existence of posteriors completely but restrict attention to the Radon property of Π on $M^1(\mathscr{X})$, which is also important for the following.

8.7.1 Existence of Pólya tree priors

The more interesting question concerns the existence of Pólya tree priors on $M^1(\mathbb{R})$ with the Le Cam-Schwartz topology, which we consider in the following theorem.

Theorem 8.7.2. *Let $\mathscr{A} = \{\alpha_m : m \geq 1\}$, $\alpha_m = \{A_\varepsilon : \varepsilon \in \mathscr{E}_m\}$ denote a Pólya tree of refining partitions of $\mathbb{R}$ and let $V_{\varepsilon 0} \sim \text{Beta}(\beta_{\varepsilon 0}, \beta_{\varepsilon 1})$, $\varepsilon \in \mathscr{E}$ denote the random fractions that divide probability mass over dyadic splits. If,*

$$\textit{Something with } \beta_{\varepsilon 0}, \beta_{\varepsilon 1},$$

then the corresponding Pólya tree process defines a Radon probability measure on $M^1(\mathbb{R})$ with the Le Cam-Schwartz topology.

Proof. Consider the requirement of $\mathscr{T}_1$-compactness: for every $\delta, \varepsilon > 0$ there exists an $L > 0$ such that for all $m \geq 1$,

$$\Pi_m\Big(\big\{P_m \in M^1(\mathscr{X}_m) : \|P_m - P_m \wedge LG\| > \delta\big\}\Big) < \varepsilon.$$

Also consider the following curious example.

Example 8.7.3. In example 8.3.1, we may consider the (non-random) choices $V_{\varepsilon 0} = 1/2$ to see quickly that the resulting (non-random) measure P is Lebesgue measure on $[0,1]$ (or $(0,1)$). To approximate this situation randomly, we may think of equal parameters $\beta_{\varepsilon 0} = \beta_{\varepsilon 1} = \beta_m > 0$ for all $\varepsilon \in \mathscr{E}_m$. Then $G(A_\varepsilon) = 2^{-m}$ for every $\varepsilon \in \mathscr{E}_m$, corresponding to the histogram projections G_m of Lebesgue measure G on $[0,1]$ (or $(0,1)$). We are interested to see how closely the random measures $P_m \sim \Pi_m$ resemble the G_m (on the (finite) σ-algebras generated by the α_m), for all $m \geq 1$. To quantify differences, we specify some degree of approximation for the total-variational distance, *i.e.* a sequence ε_m such that $\|P_m - G_m\|_{TV,m} \leq \varepsilon_m$, and require this approximation to hold with Π_m-probability growing to one. Note first that the variance is given by,

$$\int \Big(\frac{P_m(A_\varepsilon)}{G_m(A_\varepsilon)} - 1\Big)^2 d\Pi_m(P_m) = \Big(\frac{2\beta_m + 2}{2\beta_m + 1}\Big)^m - 1,$$

so if we choose $\beta_m = (2-(1+\delta_m))^{1/m})/(2(1+\delta_m)^{1/m}-2)$ for some $0<\delta_m$, then the variance of $P_m(A_\varepsilon)/G_m(A_\varepsilon)$ is less than or equal to δ_m. With the help of the union bound and Chebyshev's inequality, we see that for all $m \geq 1$,

$$\begin{aligned}\Pi_m\big(\|P_m - G_m\|_{TV,m} > \varepsilon_m\big) &= \Pi_m\left(\sum_{\varepsilon\in\mathscr{E}_m}\left|\frac{P_m(A_\varepsilon)}{G_m(A_\varepsilon)}-1\right|G_m(A_\varepsilon) > \varepsilon_m\right)\\ &\leq \Pi_m\left(\max\left\{\left|\frac{P_m(A_\varepsilon)}{G_m(A_\varepsilon)}-1\right| : \varepsilon\in\mathscr{E}_m\right\} > \varepsilon_m\right) \leq \frac{2^m\delta_m}{\varepsilon_m^2} = o(1),\end{aligned}$$

if the $delta_m$ are of order $o(2^{-m}\varepsilon_m^2)$. This example demonstrates that with appropriate choices for the parameters β ($\beta_m \approx 2^m m\varepsilon_m^{-2}$ in the above), random histrograms distributed according to a Pólya tree construction are concentrated around histogram-projections of the expected measure G with degrees of approximation decreasing arbitrarily quickly with m (for more of this nature, see [191, 167, 168]).

8.8 Supports of inverse limit distributions

To understand the nature of Dirichlet process distributions better, it helps to formulate with precision on which subsets of the model these measures concentrate their mass. In this subsection, we first view the Dirichlet process prior as a Borel measure with respect to Prokhorov's weak model topology $\mathscr{T}_C$ and characterize the $\mathscr{T}_C$-support in terms of the support of the base measure. We then show that, if one assumes existence of the Dirichlet process prior as a Borel measure with respect to the more refined topology $\mathscr{T}_1$ (see definition A.0.5), an analogous derivation characterizes the $\mathscr{T}_1$-support to consist of those measure that are absolutely continuous with respect to the base measure. However, this contradicts a discreteness property that all Dirichlet random measures share and we conclude that the Dirichlet process prior *does not exist* as a probability measure on the model with respect to the $\mathscr{T}_1$-topology (or refinements thereof like the total-variational topology).

8.8.1 Support in Prokhorov's weak topology

Theorem 8.6.12 is used for an straightforward, accessible proof of the fact that the support of D_β in $M^1(\mathscr{X})$ consists of all distributions with a support that lies inside that of the base-measure. (For an alternative proof that does not use *property (P1)*, see theorem 4.15 in [110]).

Proposition 8.8.1. *Consider $M^1(\mathscr{X})$ with Prokhorov's weak topology. Let β be a finite measure on $\mathscr{X}$. The support of D_β is given by,*

$$\operatorname{supp}_{\mathscr{T}_C}(D_\beta) = \big\{\mu\in M^1(\mathscr{X}) : \operatorname{supp}(\mu)\subset\operatorname{supp}(\beta)\big\}.$$

Proof. If P is such that $\mathrm{supp}(P) \not\subset \mathrm{supp}(\beta)$, there exist an $x \in \mathrm{supp}(P) \setminus \mathrm{supp}(\beta)$ and a continuous $f : \mathbb{R} \to [0,1]$ with $f = 0$ on $\mathrm{supp}(\beta)$ and $f(x) = 1$, by complete regularity (see definition C.2.3). This implies that $Pf > 0$ and $\beta f = 0$. So, if Q lies in the $\mathscr{T}_C$-neighbourhood $\{Q : |(P-Q)f| < \varepsilon\}$ (for small enough $\varepsilon > 0$), the expectation values Q for f are lower bounded by,

$$\inf\{Qf : Q \in M(\mathscr{X}),\ |(P-Q)f| < \varepsilon\} > \varepsilon.$$

However, by Markov's inequality,

$$D_\beta\big(Qf > \varepsilon\big) \leq \frac{1}{\varepsilon}\int Qf\, dD_\beta(Q) = \frac{1}{\varepsilon}\frac{\beta f}{\beta(\mathscr{X})} = 0$$

Conclude that P has a $\mathscr{T}_C$-neighbourhood of D_β-mass zero, which means that $P \notin \mathrm{supp}_{\mathscr{T}_C}(D_\beta)$.

Conversely, suppose that P is such that $\mathrm{supp}(P) \subset \mathrm{supp}(\beta)$. Let a continuous $f : \mathscr{X} \to [0,1]$ and $\varepsilon > 0$ be given. Since $\mathscr{X}$ is Polish and P is a probability measure, there exists a compact $K' \subset \mathscr{X}$ such that $P(K') > 1 - \frac{1}{6}\varepsilon$. As a (countable) basis for $\mathscr{X}$, choose the collection of all rational-radius balls centred on a dense countable subset and let $0 < \delta < 1$ be given; according to theorem 8.6.12, there exits a compact $K'' \subset \mathscr{X}$ such that for any partition $(A_0, A_1, \ldots, A_n)$ generated by the basis,

$$D_\beta\Big(\sum\{Q(A_i) : 0 \leq i \leq n, A_i \cap K'' = \varnothing\} > \tfrac{1}{6}\varepsilon\Big) < \delta, \tag{8.27}$$

which remains true if we replace K'' by the union $K = K'' \cup K'$. On K the continuous function f is uniformly continuous, so there exist an integer $n \geq 1$, a rational $\eta > 0$, and rational points $x_1, \ldots, x_n \in K$, such that for all $x \in \mathbb{R}$, $|x - x_i| < \eta$ implies that $|f(x) - f(x_i)| < \varepsilon$. We define a partition $A_0, \ldots, A_n$ as follows: $A_1 = \{x \in \mathbb{R} : |x - x_1| < \eta\}$, $A_i = \{x \in \mathbb{R} \setminus A_{i-1} : |x - x_i| < \eta\}$ for $2 \leq i \leq n$. The union is denoted $A = A_1 \cup \ldots \cup A_n$ and $A_0 = \mathbb{R} \setminus A$ is its complement. (The constructed partition $A_0, A_1, \ldots, A_n$ is a partition generated by the basis.) Note that $K \cap A_i \neq \varnothing$ iff $1 \leq i \leq n$, so (8.27) reduces to $D_\beta(Q : Q(A_0) > \frac{\varepsilon}{6}) < \delta$. There exist $f_1, \ldots, f_n \in [0,1]$ such that,

$$\sup_{x \in A}\Big| f(x) - \sum_{i=1}^{n} f_i\, 1_{A_i}(x)\Big| < \tfrac{1}{3}\varepsilon.$$

Hence, in case $Q(A_0) \leq \frac{\varepsilon}{6}$,

$$\begin{aligned}
\big|(P-Q)f\big| &\leq \big|(P-Q)f\,1_A\big| + P(A_0) + Q(A_0) \\
&\leq \Big|(P-Q)\Big(f\,1_A - \sum_{i=1}^{n} f_i\, 1_{A_i}\Big)\Big| + \Big|\sum_{i=1}^{n} f_i\big(P(A_i) - Q(A_i)\big)\Big| + \tfrac{1}{3}\varepsilon \\
&\leq \sum_{i=1}^{n}\big|P(A_i) - Q(A_i)\big| + \tfrac{2}{3}\varepsilon
\end{aligned}$$

Therefore,

$$D_\beta\Big(\big|(P-Q)f\big|<\varepsilon\Big)\geq D_\beta\Big(\big|(P-Q)f\big|<\varepsilon\;\Big|\;Q(A_0)\leq\tfrac{1}{6}\varepsilon\Big)D_\beta\Big(Q(A_0)\leq\tfrac{1}{6}\varepsilon\Big)$$
$$\geq(1-\delta)D_\beta\Big(\sum_{i=1}^{n}\big|P(A_i)-Q(A_i)\big|<\tfrac{1}{3}\varepsilon\;\Big|\;Q(A_0)\leq\tfrac{1}{6}\varepsilon\Big)$$

Since the support of P lies inside the support of β, $P(A_i)>0$ implies $\beta(A_i)>0$, which implies that $D_\beta(Q(A_i)\in U)>0$ for any open U in $[0,1]$, in particular, $D_\beta(|Q(A_i)-P(A_i)|<\frac{1}{3n}\varepsilon)>0$ and this remains true, if we condition on $Q(A_0)\leq\frac{1}{6}\varepsilon$. Conclude that $D_\beta(|(P-Q)f|<\varepsilon)>0$.

This proposition shows that Dirichlet priors spread their mass over the full non-parametric model for data in $\mathbb{R}$, if we choose the base measure to have full support.

8.8.2 Support in the Le Cam-Schwartz topology

Definition 8.8.2. Consider $M^1(\mathscr{X})$ with Prokhorov's weak topology $\mathscr{T}_1$ (or finer) and a Borel probability measure Π. The *expected measure* G under Π is defined pointwise, for every Borel set A in $\mathscr{X}$,

$$G(A)=\int P(A)\,d\Pi(P).$$

The expected measure under Π is a Borel probability measure on $\mathscr{X}$, because σ-additivity of G follows from positivity of P and monotone convergence. Note the following direct consequence.

Lemma 8.8.3. *Consider $M^1(\mathscr{X})$ with the topology $\mathscr{T}_1$ and a Borel probability measure Π. For any Borel set A in $\mathscr{X}$, $G(A)=0$ implies that $\Pi(\{P\in M^1(\mathscr{X}):P(A)>0\})=0$.*

Proof. Let a Borel set A in $\mathscr{X}$ be given. If the Borel set $B=\{P\in M^1(\mathscr{X}):P(A)>0\}$ in $M^1(\mathscr{X})$ has probability $\Pi(B)>0$, then for some $\varepsilon>0$, the Borel set $B'=\{P\in M^1(\mathscr{X}):P(A)>\varepsilon\}$ has probability $\Pi(B')>0$, which implies that $G(A)\geq\int_{B'}P(A)\,d\Pi(P)\geq\varepsilon\Pi(B')>0$.

This domination property of the expected measure plays a role in the following proposition concerning the support of $\mathscr{T}_1$-Radon measures on $M^1(\mathscr{X})$.

Proposition 8.8.4. *Consider $M^1(\mathscr{X})$ with the topology $\mathscr{T}_1$ and a Radon probability distribution Π. Let G be the expected measure under Π. Then $\{P\in M^1(\mathscr{X}):P\ll G\}$ is closed in $M^1(\mathscr{X})$ and,*

$$\operatorname{supp}_{\mathscr{T}_1}(\Pi)\subset\{P\in M^1(\mathscr{X}):P\ll G\}.$$

If, in addition, for every $P \in M^1(\mathscr{X})$ with $P \ll G$, every $\delta > 0$ and every finite partition α of $\mathscr{X}$ into non-empty, measurable $A_1, \ldots, A_{N(\alpha)}$,

$$\Pi_\alpha\big(\{Q \in M^1(\mathscr{X}_\alpha) : |Q(A_i) - P(A_i)| < \delta, 1 \le i \le N(\alpha)\}\big) > 0, \tag{8.28}$$

then also,

$$\mathrm{supp}_{\mathscr{T}_1}(\Pi) \supset \{P \in M^1(\mathscr{X}) : P \ll G\}.$$

Proof. If $P \in M^1(\mathscr{X})$ is not dominated by G, then there exists a Borel set A such that $P(A) > 0 = G(A)$. Consequently for small enough $\varepsilon' > 0$, the $\mathscr{T}_1$-open neighbourhood $U = \{Q \in M^1(\mathscr{X}) : |Q(A) - P(A)| < \varepsilon'\}$ does not meet $\{Q \in M^1(\mathscr{X}) : Q \ll G\}$, so $\{Q \in M^1(\mathscr{X}) : Q \ll G\}$ is closed. In addition, $\Pi(U) \le \Pi(\{Q \in M^1(\mathscr{X}) : Q(A) > 0\}) = 0$, so U receives Π-mass zero, implying that $P \notin \mathrm{supp}_{\mathscr{T}_1}(\Pi)$. Conversely, let $P \in M^1(\mathscr{X})$, $\varepsilon > 0$, $k \ge 1$ and $\mathscr{B}$-measurable $\phi_l : \mathbb{R} \to [0,1]$ $(1 \le l \le k)$ be given. There exists an $n \ge 1$ such that for every l, there is a $\mathscr{B}$-measurable partition $\{A_{1,l}, \ldots, A_{n,l}\}$ of $\mathbb{R}$ and constants $0 \le f_{m,l} \le 1$, $1 \le m \le n$, $1 \le l \le k$, such that the simple functions $f_l(x) = \sum_{m=1}^n f_{m,l}\, 1\{x \in A_{m,l}\}$ approximate ϕ_l uniformly: $\sup_{x \in \mathbb{R}} |\phi_l(x) - f_l(x)| < \varepsilon/4$. Re-labelling the $A_{m,l}$ as $A_1, \ldots, A_I$, we write,

$$\begin{aligned} \big\{Q \in M(\mathscr{X}) : &|Q(A_i) - P(A_i)| < \varepsilon/m, 1 \le i \le I\big\} \\ &\subset \big\{Q \in M(\mathscr{X}) : |\langle Q, \phi_l\rangle - \langle P, \phi_l\rangle| < \varepsilon, 1 \le l \le k\big\}. \end{aligned} \tag{8.29}$$

Conclude that for every $P \in M^1(\mathscr{X})$ such that $P \ll G\}$ and every open neighbourhood U of P, there exists a set V of the form on the left-hand-side of inclusion (8.30) such that $V \subset U$, proving the second assertion.

(In (8.28) and in the proof, we have abused notation slightly: for $Q \in M^1(\mathscr{X}_\alpha)$, the domain is $\mathscr{X}_\alpha$ rather than α, so that the argument should be $\varphi'_\alpha(A_i)$ rather than A_i. The form given is more intuitive, however.) One concludes from this, that domination by the expected measure G is among minimal sufficient conditions for existence (in concrete applications).

Generally, weak compactness of the type needed in property *(P3)* is the domain of the Dunford-Pettis-Grothendieck theorem [117] and is best analysed in the context of L- and M-spaces [179]. Given the previous conclusion, however, we shall have to analyse this requirement only in constructions where all relevant $P \in M^1(\mathscr{X})$ are *dominated* by the expected measure G. We do this in the context of certain Pólya-tree distributions in section **??**. In that case, weak compactness is characterized by uniform integrability and the Dunford-Pettis theorem [76]. Property *(P3)* can then be re-formulated as follows,

(P3') *for every $\varepsilon > 0$, there exists a $K \subset M^1(\mathscr{X})$ such that,*

$$\forall_{\delta>0}\, \exists_{L>0}\, \forall_{P \in K} : \int_{\{dP/dG > L\}} \frac{dP}{dG}\, dG < \delta,$$

and for all α, $\Pi_\alpha\big(\varphi_{*\alpha}(K)\big) > 1 - \varepsilon$.

(Note that the K in the first part is relatively compact, while taking its closure increases $\Pi_\alpha(\varphi_{*\alpha}(K))$, so we do not have to insist on a *closed* K in this formulation.)

To conclude we note that there is an equivalent formulation of the Dunford-Pettis theorem [76], that implies another re-formulation of condition *(P3)*:

(P3'') *for every* $\varepsilon > 0$*, there exists a* $K \subset M^1(\mathscr{X})$ *such that,*

$$\forall_{\delta>0}\exists_{\eta>0}\forall_{P\in K}\forall_{A\in\mathscr{B}} : G(A) < \eta \Rightarrow P(A) < \delta,$$

and for all α, $\Pi_\alpha\big(\varphi_{*\alpha}(K)\big) > 1-\varepsilon$.

Let Π be a Borel probability measure on $M^1(\mathscr{X})$ with the $\mathscr{T}_1$-topology, with expected measure G. For every $\eta > 0$ and every $\alpha \in \mathscr{A}$, let $\mathscr{C}_{G,\alpha}(\eta)$ denote the collection of all Borel sets B in $\mathscr{X}$ that can be approximated by elements of the σ-algebra σ_α to within G-measure η: $\inf_{C\in\sigma_\alpha} G(B\triangle C) < \eta$. (Note that for any $\eta > 0$ and any B, there exists an α such that $B \in \mathscr{C}_{G,\alpha}(\eta)$, see *e.g.* theorem 4.4 in [147].)

Proposition 8.8.5. *If* Π *is a Radon probability measure on* $M^1(\mathscr{X})$ *with the* $\mathscr{T}_1$*-topology, with expected measure* G*, then for every* $\delta, \varepsilon > 0$*, there exists a partition* α *and an* $\eta > 0$*, such that for all* $B \in \mathscr{C}_{G,\alpha}(\eta)$*,*

$$\Pi\left(\left\{P \in M^1(\mathscr{X}) : \inf_{C\in\sigma_\alpha} P(B\triangle C) > \delta\right\}\right) < \varepsilon.$$

Proof. Let $\varepsilon > 0$ be given. There exists a compact K in $M^1(\mathscr{X})$ such that $\Pi(K) > 1-\varepsilon$. For every $\delta >$, there exists an $\eta > 0$, such that for all Borel sets A in $\mathscr{X}$, $G(A) < \eta$ implies that for all $P \in K$: $P(A) < \delta$. In particular, if $B \in \mathscr{C}_{G,\alpha}(\eta)$, then for some $C \in \sigma_\alpha$, $G(B\triangle C) < \eta$, so that for all $P \in K$, $P(B\triangle C) < \delta$.

This fact is important from the numerical perspective: the practitioner chooses a partition α to perform computations and would like to be able to control accuracy of his approximations for the P in terms of their restrictions to σ_α. He has control over the finite-dimensional marginals, commonly leading to a tractable expression for the expected measure G, and approximations in G-measure by α-measurable coarsening can be verified readily for all Borel sets in $\mathscr{X}$. The compactness condition implied by the Radon property ensures that the approximation in G-measure carries over to approximation in P-measure, uniformly in P, with arbitrarily high Π-probability, depending on the degree of approximation in the level α that is chosen for actual computations. Such a guarantee concerning degrees of approximation for Borel sets is not automatic if Π is a Borel measure only for Prokhorov's weak topology: there may be Borel sets B in $\mathscr{X}$, for which $G(B\triangle C)$ is small, while $P(B\triangle C)$ is large with non-negligible Π-probability.

And finally, the Radon property is important for the existence of conditional distributions like the Bayesian posterior: we do not explore this difficult subject further here, but note that in [228], it is shown that conditional distributions (or *disintegrations*, as in [48, 49]) exist if the unconditioned measure is Radon and permits a so-called Π*-concassage* (a partition $M^1(\mathscr{X}) = \cup_n K_n \cup N$ into countably many

compact K_n, and some N such that $\Pi(N) = 0$) with *metrizable* compacta K_n (see theorem 41 in [228]). The most straightforward construction involves a perspective where one is interested in a one-to-one, continuous parameter $\theta : M^1(\mathscr{X}) \to D$, where D is a metric space (*c.f.* the construction in [173]) and θ^{-1} is continuous on compact subsets of the image $\theta(M^1(\mathscr{X}))$.

But this perspective changes dramatically if we refine the topology slightly, and consider the support with respect to $\mathscr{T}_1$. Note that for the next proposition, we *assume* that the Dirichlet process prior exists as a Borel probability measure with respect to $\mathscr{T}_1$, an assumption that will turn out to be untenable.

Proposition 8.8.6. *($\mathscr{T}_1$-support of the Dirichlet process prior)*
Consider $M(\mathbb{R})$ with the weak topology $\mathscr{T}_1$ and assume that the Dirichlet process prior exists as a Borel probability measure $M(\mathbb{R})$. Let α be a bounded, positive measure on $(\mathbb{R}, \mathscr{B})$. The $\mathscr{T}_1$-support of D_α is given by,

$$\operatorname{supp}_{\mathscr{T}_1}(D_\alpha) = \{P \in M(\mathbb{R}) : P \ll \alpha\}.$$

Proof. If P is not dominated by α, then there exists an $A \in \mathscr{B}$ such that $P(A) > 0 = \alpha(A)$. Consequently for small enough $\varepsilon' > 0$, the $\mathscr{T}_1$-open neighbourhood $U = \{Q \in M(\mathbb{R}) : |Q(A) - P(A)| < \varepsilon'\}$ does not meet $\{Q \in M(\mathbb{R}) : Q \ll \alpha\}$, so $\{Q \in M(\mathbb{R}) : Q \ll \alpha\}$ is closed. In addition, *c.f.* lemma 3.6.6, $D_\alpha(Q(A) > 0) = 0$, so U receives D_α-mass zero, and we conclude that P does not lie in the support of D_α. Conversely, let $P \in M(\mathbb{R})$, $\varepsilon > 0$, $k \geq 1$ and $\mathscr{B}$-measurable $\phi_l : \mathbb{R} \to [0,1]$ ($1 \leq l \leq k$) be given. There exists an $n \geq 1$ such that for every l, there is a *$\mathscr{B}$-measurable partition* $\{A_{1,l}, \ldots, A_{n,l}\}$ of $\mathbb{R}$ and constants $0 \leq f_{m,l} \leq 1$, $1 \leq m \leq n$, $1 \leq l \leq k$, such that the simple functions $f_l(x) = \sum_{m=1}^n f_{m,l} 1\{x \in A_{m,l}\}$ approximate ϕ_l uniformly,

$$\sup_{x \in \mathbb{R}} |\phi_l(x) - f_l(x)| < \varepsilon/4.$$

Re-labelling the $A_{m,l}$ as $A_1, \ldots, A_I$, we write,

$$\begin{aligned} \big\{Q \in M(\mathbb{R}) : &|Q(A_i) - P(A_i)| < \varepsilon/m, 1 \leq i \leq I\big\} \\ &\subset \big\{Q \in M(\mathbb{R}) : |Q\phi_l - P\phi_l| < \varepsilon, 1 \leq l \leq k\big\}. \end{aligned} \tag{8.30}$$

Conclude that for every P in $\{Q \in M(\mathbb{R}) : Q \ll \alpha\}$ and every open neighbourhood U of P, there exists a set V of the form on the left-hand-side of inclusion (8.30) such that $V \subset U$. One sees from definition 3.6.1 (for a detailed proof, see [111], theorem 3.2.4) that for any $P \ll \alpha$, such sets V receive non-zero probability under D_α, so that $P \in \operatorname{supp}_{\mathscr{T}_1}(D_\alpha)$.

If we maintain the (erroneous) assumption that the Dirichlet process prior exists as a Borel probability measure with respect to $\mathscr{T}_1$, the conclusion is rather disappointing: contrary to what proposition 8.8.1 suggests, any Dirichlet distribution is actually concentrated on a space of measures that can be represented as an L^1-type family of densities (with respect to the base-measure α). However, that conclusion contradicts the well-known *discreteness property* of the Dirichlet process prior. Considering the

conclusions of propositions 8.8.6 and 8.3.11, we note that any non-discrete choice for the base measure β leads to a contradiction: for example, if we choose a base measure that is absolutely continuous with respect to Lebesgue measure,

$$D(\mathbb{R}) \cap \{P \in M(\mathbb{R}) : P \ll \beta\} = \varnothing,$$

while both are assigned mass equal to one. Conclude as follows.

Corollary 8.8.7. *(Non-Existence of Dirichlet process distributions)*
Let β be a bounded, positive measure on a Polish space $\mathscr{X}$ with inverse system of measures (8.5). Unless β is a discrete measure, the Dirichlet process prior D_β is not a Radon probability measure on $M^1(\mathscr{X})$ with respect to the topology $\mathscr{T}_1$ (or refinements thereof, like the total-variational topology).

Corollary 8.8.7 has consequences when it comes to measurability. Hellinger/total-variational metrics and Kullback-Leibler divergences are not measurable for the domain of D_β; appropriate are the Lévy-Prokhorov, Kantorovitch-Rubinstein and Wasserstein metrics. For Bayesian and frequentist density estimation, Dirichlet process distributions are used in an in-direct fashion, as priors for the parameter P in a model of (Lebesgue) so-called *mixture densities*:

$$P \mapsto p_P(x) = \int \varphi(x-y)\, dP(y),$$

where $\mathbb{R} \to [0,\infty) : z \mapsto \varphi(z)$ denotes some bounded, continuous Lebesgue probability density on $\mathbb{R}$. Here, P is called the *mixing distribution* and the map $P \mapsto p_P$ is one-to-one as well as completely continuous. Hence any subset K of mixing distributions in $M^1(\mathscr{X})$ that is relatively weakly compact, is mapped to a Hellinger relatively compact subset of densities.

8.9 Inverse limit measures in the total-variational topology

Proposition 8.9.1. *, A model $\mathscr{P}$ is TV-separable, if and only if, $\mathscr{P}$ is dominated and $\mathscr{B}$ is generated countably.*

8.10 Exercises

8.10.1. Find which line contains the mistake in the proof of proposition 8.2.2. Speculate on possible solutions, to conclude that there is no easy way out. *[Hint: Recall that conditional probabilities are defined almost-surely for every A separately, c.f. definition (B.9), but not automatically also almost-surely for all A simultaneously, as a regular conditional distribution. The present problem is similar.]*

8.10.2. In the proof of proposition 8.2.2, it is claimed that if a positive sequence $(x_n)_{n\geq 1}$ converges to zero, then there exists a subsequence $(x_{n_j})_{j\geq 1}$ with finite sum. Prove the following, slightly stronger fact: for every positive sequence $(x_n)_{n\geq 1}$ and every $s>0$, there exists a subsequence $(x_{n_j})_{j\geq 1}$ such that $\sigma_{j\geq 1}x_{n_j} < s$.

8.10.3. Complete the proof of theorem 8.2.1, by verifying the Kolmogorov consistency conditions (K1) and (K2) of theorem B.6.3 explicitly, based on the histogram marginals.

8.10.4. Comparing definition C.1.18 with the assertion of proposition 8.2.2 one might expect the formulation of the latter to read "the support of D_μ is $M^1(\mathbb{R})$". Why does proposition 8.2.2 not mention "the support" of D_μ? *A fortiori*, point out why not just the topological nature of the subset $M^1(\mathbb{R})$ but even measurability forms a problem.

8.10.5. As is mentioned after the statement of theorem 8.2.3, the proof contains a point that is, at best, correct but passed too quickly. Find this point. *[Hint: the proof is correct only because Dirichlet distributions are tailfree.]*

8.10.6. Let $\mathscr{A} = \{\alpha_n : n \geq 1\}$ denote a dyadic tree of refining partitions for a Pólya tree prior as in section 8.3. Assume that for every $n,m \geq 1$, $n \neq m$, the vectors of splitting variables $(V_\varepsilon : \varepsilon \in \mathscr{E}_n)$ and $(V_\varepsilon : \varepsilon \in \mathscr{E}_m)$ are independent. Show that the resulting inverse system of measures Π_α is tailfree.

8.10.7. Show that the open interval $(0,1)$ with the subspace topology it inherits from $\mathbb{R}$ (with the usual topology), is a Polish space (even though it is not complete for the usual metric $d(x,y) = |x-y|$).

8.10.8. Show that if $\mathscr{X}$ is a Hausdorff topological space and $\mathscr{P} = \{\delta_x : x \in \mathscr{X}\}$ is the space of all *Dirac measures* on $\mathscr{X}$, and we equip $\mathscr{P}$ with Prokhorov's weak topology, then the map $x \mapsto \delta_x$ is a continuous bijection. Also show that if $\mathscr{X}$ is completely regular, the inverse mapping $\delta_x \mapsto x$ is continuous, so that $\mathscr{X}$ and $\mathscr{P}$ are homeomorphic. Next, prove that that the convex hull of $\mathscr{P}$ is $\mathscr{T}_C$-dense in the space $\mathscr{M}^1(\mathscr{X},\mathscr{B})$ of all Borel probability measures on $(\mathscr{X},\mathscr{B})$, but not $\mathscr{T}_\infty$-dense unless $\mathscr{X}$ is countable.

Chapter 9
Consistent tests and model selection

The question, "Which pairs of model subsets can be told apart asymptotically and which cannot?", is not just of direct practical importance (*e.g.* for model selection with large amounts of data) and of essential value in the development of theory. It is also a fundamental matter at the heart of statistics: which model questions have a truly statistical nature (that is, questions answerable from the data), and which do not? Of course, there are two versions of this question, one that requires only a proof of the *existence* of tests, and another that asks for the actual *construction* of such tests. In this chapter, the existence question is answered first and the constructive question is considered as in [157], promoting the existence result to a guarantee that posteriors achieve the correct conclusion, also for the frequentist.

9.1 Asymptotic testability

To make the issue precise, consider a situation where we observe *i.i.d.* data $X^n \sim P^n$, $(n \geq 1)$, with a model $\mathscr{P}$ such that $P \in \mathscr{P}$. Suppose that, for disjoint $B, V \subset \mathscr{P}$, we are interested whether,

$$H_0 : P \in B, \quad \text{or} \quad H_1 : P \in V.$$

In an asymptotic, symmetric testing procedure, one requires a sequence of test functions (ϕ_n) with type-I and type-II error probabilities (resp. $P^n\phi_n$ for $P \in B$ and $P^n(1-\phi_n)$ for $P \in V$) that go to zero. Equivalently (see [203, 91] and proposition 9.4.2) one requires existence of some testing procedure with the following property,

> A testing procedure that chooses for B or V based on X^n for every $n \geq 1$, has property (D) if it is wrong only a finite number of times with P^∞-probability one.

Property (D) is referred to as "discernibility" in [69, 203, 91] and it is also the basis for the tests in many other publications, for example [60, 70].

To do justice to the level of generality that the title promises, a real answer requires various things: ideally, there should be *no restrictions* on the model at all; furthermore, the answer should *characterise* the pairs B,V for which test sequences exist, as well as the pairs for which this is not the case, preferably in the form of an equivalence: given any model $\mathscr{P}$, whether P belongs to B or to V can be tested asymptotically, if and only if, *etcetera*. Answers depend crucially on the formal/philosophical framework: a Bayesian who gives his answer based on posterior odds or a Bayes factor, and who disregards potential prior null-sets of exceptions, answers this question differently from a frequentist, who formulates his answer in terms of test functions and insists on asymptotic consistency for all points in the model (or even uniformly). These distinctions lead to differing answers to the question "What is asymptotically testable and what is not?", and hence, to differing notions of answerable and non-answerable statistical questions.

In section 9.2, we consider three forms of asymptotic testability: uniform testability, pointwise testability and Bayesian testability. In subsequent sections we prove for each form an equivalence characterising pairs B,V for which consistent tests exist: in section 9.3 equivalent formulations of uniform testability are given; in section 9.4 we characterise hypotheses that are pointwise testable; and in section 9.5, it is shown that Bayesian tests exist for a very wide variety of hypotheses.

As stressed already, we do not restrict attention to *subclasses* of models, the model choice is left completely free. (We make one exception: in theorem 9.4.16 we require a *dominated model*, see the discussion in section 9.7.) Characterisations of testability are formulated in terms of conditions on the sets B,V only. Otherwise, we would not characterise testability itself but how it manifests in subclasses. Of course, it is possible that a hypothesis is not testable versus its complement in a large model $\mathscr{P}$, while becoming testable when $\mathscr{P}$ is restricted. The consequences of topologically suitable, general restrictions (like completeness, metrizability, metric totally-boundedness, or weak-relative-compactness) are accommodated in corollaries. Such restrictions form connections with previous work and motivate examples.

9.1.1 Some examples and unexpected answers

Intuition regarding the existence problem of asymptotic tests is greatly helped by some examples that typify the nature of possible answers: distinctions between smoothness classes for a regression function $f : X \to \mathbb{R}$:

$$H_0 : f \in C^1(X \to \mathbb{R}), \quad H_1 : f \in C^2(X \to \mathbb{R}), \tag{9.1}$$

cannot be tested consistently according to the frequentist. However, to the Bayesian using the posterior, smoothness classes are asymptotically testable without any reservations, for prior-almost-all points in the model. To mention another instance, the frequentist cannot test to distinguish asymptotically between classes of densities p on $\mathbb{R}$ with or without a second moment:

$$H_0 : \int x^2\, p(x)\, dx < \infty, \quad H_1 : \int x^2\, p(x)\, dx = \infty.$$

That simple fact implies no statistician dealing with unbounded data can ever use the central limit theorem with asymptotic certainty that it applies for the true distribution P, unless square-integrability is established externally, before observation of the data.

Bayesians can make the distinction between the hypotheses H_0 and H_1 asymptotically (but, again, only prior-almost-surely). Similarly, Bayesians can test consistently whether a distribution on $\mathbb{R}$ is compactly supported, whether its Lebesgue density is square-integrable, *etcetera*, distinctions that are not testable for frequentists [70].

The intricacy of the question is emphasized further by the unexpected answer to *Cover's rational mean problem* [60]: for an *i.i.d.* sequence of coin-flips $X_1, X_2, \ldots$ (with all X_i distributed marginally Bernoulli-p with $p \in [0,1]$), consider the hypotheses:

$$H_0 : p \in [0,1] \cap \mathbb{Q}, \quad H_1 : p \in [0,1] \setminus \mathbb{Q}. \tag{9.2}$$

Rather surprisingly, Cover shows that there exists a test sequence $\phi_n(X_1, \ldots, X_n)$ that goes to one if $p \in [0,1] \cap \mathbb{Q}$, and to zero for *Lebesgue-almost-all* $p \in [0,1] \setminus \mathbb{Q}$. It is not possible to find a test sequence for Cover's problem without such an exceptional null-set (see corollary 9.4.2). However, it is possible to restrict the model to enable testability: Dembo and Peres [69] show that there exist asymptotically consistent tests for,

$$H_0 : p \in [0,1] \cap \mathbb{Q}, \quad H_1 : p \in [0,1] \cap \sqrt{2} + \mathbb{Q}, \tag{9.3}$$

without measure-theoretic exceptions. But one does not have to restrict to countable hypotheses to find testability for apparently deeply intertwined hypotheses: example 9.4.25 shows it is possible to test whether p lies in the *Cantor subset* C or not,

$$H_0 : p \in C, \quad H_1 : p \in [0,1] \setminus C, \tag{9.4}$$

It is noted that C is zero-dimensional and nowhere-dense, while both C and its complement are uncountable. And although C has Lebesgue measure zero, there are Cantor subspaces of $[0,1]$ that have non-zero Lebesgue measure, for which testability also holds. So if testability is ruined by certain forms of denseness but not for others, and maintained for self-similar sets like C, what does the distinction depend on (preferably characterized in topological terms)?

9.1.2 Testability over the decades

Of course the question has a long history: the first attempts to answer general questions on testability appear already in the 1950's: Hoeffding and Wolfowitz [126] give sufficient conditions that are also necessary in some cases (see also, [173]). Kraft [161] studied consistent tests for families of general, dependent data distributions

and gives a separation condition in terms of the separation of convex hulls of finite-dimensional projections, much like the Hahn-Banach theorem and its specializations suggest [50]. Berger [17] gives necessary and sufficient conditions for the existence of uniformly consistent tests, and subsequent work [18, 173] extends the approach to pointwise consistent estimation problems. It is noted that the present work is inspired first and foremost on the Le Cam-Schwartz theorem [173], which provides necessary and sufficient conditions for the existence of uniform and pointwise tests, in terms of a particular uniformity we denote $\mathscr{U}_\infty$ with associated topology $\mathscr{T}_\infty$. (Appendix **??** offers a comparison of $\mathscr{T}_\infty$ with other, better-known model topologies.) In the form applicable to pointwise testing, the Le Cam-Schwartz theorem states that,

Theorem 9.1.1. (Le Cam-Schwartz, 1960) *Let $\mathscr{P}$ be a model for i.i.d. data X^n with disjoint subsets B,V. There exist (uniformly) consistent tests for B versus V, if and only if, there exists a sequence of $\mathscr{U}_\infty$-uniformly continuous functions $\psi_n : \mathscr{P} \to [0,1]$ such that,*

$$\psi_n(P) \to 1_V(P), \tag{9.5}$$

(uniformly) over all $P \in \mathscr{P}$.

So the Le Cam-Schwartz theorem provides the definitive answer to our question. However, its formulation is in terms of a uniformity $\mathscr{U}_\infty$ that is "rather inaccessible" [179] (see [69, 203, 91] for more detailed comments), and it is perhaps this inaccessibility that explains why the entire body of subsequent work on the subject mentions the Le Cam-Schwartz theorem but does not relate to it at any formal level. Most sensitive to the argument put forth by Le Cam and Schwartz appears to be the insightful work of Ermakov [91], which departs from necessary conditions for the existence of pointwise consistent tests in terms of uniformly consistent tests. However, a weakly compact, dominated model is required for Ermakov's results and Prokhorov's weak topology rather than Le Cam-Schwartz's uniformity $\mathscr{U}_\infty$ is used to formulate testability conditions.

A separate but related historical line of research originates from Cover's rational mean problem [60], and answers Cover's specific (but prototypical, see theorem 9.4.15) question from the probabilistic point of view (see also, [208]). As a second inspirational reference for this work, we mention Dembo and Peres [69], who show that the limited version of Cover's problem in (9.3) has a solution and subsequently prove the following theorem.

Theorem 9.1.2. (Dembo and Peres, 1995) *Let $\mathscr{P}$ be a model dominated by Lebesgue measure μ for i.i.d. data X^n. Model subsets B,V that are contained in disjoint countable unions of closed sets for Prokhorov's weak topology have tests with property (D). If there exists an $\alpha > 1$ such that $\int (dP/d\mu)^\alpha \, d\mu < \infty$ for all $P \in \mathscr{P}$, then the converse is also true.*

Note the recurrence of weakly compact, dominated models with Prokhorov's weak topology. Kulkarni and Zeitouni [164] accept Cover's exceptional null-set and consider the question when such tests (which we call Bayesian, see definition 9.2.3) exist in more general setting. Nobel [203] notices that the approach of Dembo and Peres can be extended from *i.i.d.* setting to a framework where the data is dependent,

e.g. to test between disjoint families of (uniformly) ergodic processes. Ermakov's work also appears to be inspired by the results of Dembo and Peres.

If one departs from the strictly constructive, statistical perspective (*e.g.* in non-parametric density estimation), first of all, there are many solutions that are specific to model and hypotheses presented, often involving specific test-statistics and critical regions, roughly following the classical approach of [170]. If we restrict attention to (a non-exhaustive list of) references that aim to answer the more general question, it is worth mentioning Donoho (1988) [77], who discusses non-parametric confidence sets and testing of hypotheses for aspects of the density of the data in dominated, non-parametric models. Similar in intention, and a third major inspiration for this chapter, is Devroye and Lugosi (2003) [70] who construct solutions in many diverse and practical examples of non-parametric testing problems for densities, based largely on contemporary methods of kernel estimation.

9.1.3 The forms that answers take

Given the rather intricate examples of subsection 9.1.1, one wonders which expectations one should have regarding the *forms* in which answers to the testability question are formulated. Based on the examples of pointwise testing in subsection 9.1.1, it is clear that model topology plays a central role in characterising which disjoint pairs B,V are testable and which are not. Exactly *which* topology we deploy here, is prescribed by the necessary and sufficient conditions that the Le Cam-Schwartz theorem formulates: we are obliged to view the model as a uniform space with the uniformity $\mathscr{U}_\infty$. (In the examples of subsection 9.1.1, the topology $\mathscr{T}_\infty$ coincides with the usual topology of $[0,1]$.) This rather technical starting point is not a choice but an imperative (if we insist on total freedom of model choice); only by setting model conditions (*e.g.* like uniform integrability, as in [69, 203, 91] and corollary 9.4.23) can this be avoided.

But having decided which topology is relevant, we also need to determine what type of topological condition we expect to determine testability of disjoint pairs B,V. For uniform testability of disjoint B,V, it is necessary and sufficient (see theorem 9.3.3) that B and V are $\mathscr{U}_\infty$*-uniformly separated*: there exists an entourage U that does not meet $B\times V\cup V\times B$. Regarding pointwise testability one expects countable unions of weakly closed sets to be important, based on [69, 203, 91]; as we shall see, disjoint B,V that are pointwise testable can be characterised as sets that are both countable unions of closed sets and countable intersections of open sets in $B\cup V$. This condition holds for the *Cantor set* C and its complement in $[0,1]$ and for the countable sets $[0,1]\cap\mathbb{Q}$ and $[0,1]\cap\sqrt{2}+\mathbb{Q}$, but not for $[0,1]\cap\mathbb{Q}$ and its complement in $[0,1]$.

Bayesian testability is different and forms the constructive contribution. Bayesian testability does not fit the formulation of the Le Cam-Schwartz theorem and, as such, escapes its topological imperatives. The existence of a Bayesian test sequence is equivalent to the consistency of posterior odds or Bayes factors (see theorem 9.5.1),

at least, if one is willing to permit prior null-sets of exceptions. This presents the opportunity to promote mere *existence* proofs for (Bayesian) testability, to *constructive* proofs in the sense that they imply an actual way to perform the test based on the data, using the posterior. This resolves the matter of testability for the Bayesian, but for the frequentist there remains the rather unwelcome possibility of exceptional null-sets [98]. To bridge the discrepancy, two things are required [157]: a Bayesian test sequence with known testing power and a prior that induces a property termed *remote contiguity* for local prior predictive distributions. Then the Bayesian conclusion that the posterior provides a consistent test sequence remains valid for the frequentist, that is, without exceptional null-sets. Using a generality concerning the testing power of uniform test sequences (see proposition 9.3.1), and remote contiguity as it applies for Kullback-Leibler priors, we indicate a practical way to perform consistent, frequentist *model selection* with posteriors, and demonstrate how to use it in two model selection problems, selection of the number of clusters in a clustering problem, and selection of a directed acyclical graph in a graphical model.

9.2 Existence of test sequences

Let the model $\mathscr{P}$ be a collection of distributions P on a measurable space $(\mathscr{X},\mathscr{B})$, to model *i.i.d.* samples $X^n = (X_1, X_2, \ldots, X_n) \in \mathscr{X}^n$, $X^n \sim P^n$. The relevant model topology is the (subspace) topology $\mathscr{T}_\infty$ defined in definition **??**. We consider two disjoint model subsets B, V and wonder whether there is a way to tell whether the true distribution of the data lies in B or in V with asymptotic certainty. More particularly, we wonder whether there exists a sequence of test functions $\phi_n : \mathscr{X}^n \to [0,1]$ that converge to one or to zero, depending on $P \in B$ or $P \in V$ (in probability/expectation or almost-surely, see proposition 9.4.2). Given a topological space X, we say that the testing problem *has a (uniform) representation on* X, if there exists a $\mathscr{T}_\infty$-(uniformly-)continuous surjective map $f : B \cup V \to X$ such that $f(B) \cap f(V) = \varnothing$. Given a Hausdorff topological space Θ, we say that the model is *parametrized by* Θ, if there exists a $\mathscr{T}_\infty$-continuous bijection $P_\cdot : \Theta \to \mathscr{P}$ (*i.e.* for every $m \geq 1$ and measurable $f : \mathscr{X}^m \to [0,1]$, the map $\theta \mapsto P_\theta^m f$ is continuous). This condition is satisfied quite easily, for example, it is weaker than continuity with respect to the total-variational topology (see proposition C.7.17). It does not imply that Θ and $\mathscr{P}$ are homeomorphic, unless Θ is compact. If Θ is compact and $P_\cdot$ is $\mathscr{T}_1$-continuous (see proposition C.7.13), then $\mathscr{P}$ is parametrized by Θ and $P_\cdot$ is a homeomorphism. When considering a represented testing problem on a parametrized model with $X = \Theta$, Θ and the model are homeomorphic.

Test sequences come in various kinds, *e.g.* uniform or pointwise, or Bayesian in nature.

Definition 9.2.1. We say that (ϕ_n) is a *uniform test sequence* for B versus V, if,

$$\sup_{P\in B} P^n\phi_n \to 0, \quad \sup_{Q\in V} Q^n(1-\phi_n) \to 0. \tag{9.6}$$

Existence of a uniform test sequence for B versus V implies the existence of a uniform test sequence *of exponential power* (see proposition 9.3.1), *i.e.* (9.6) implies there exist a test sequence ψ_n whose sum of type-I and type-II errors goes to zero exponentially fast,

$$\sup_{P\in B} P^n\phi_n + \sup_{Q\in V} Q^n(1-\phi_n) \leq e^{-nD},$$

for some constant $D > 0$. This fact is exploited in section 9.7, where it is used together with a suitable prior to do demonstrate model-selection with Bayes factors, in a *constructive* way, while satisfying *frequentist* consistency criteria.

Definition 9.2.2. The (ϕ_n) are a *pointwise test sequence* for B versus V, if,

$$\phi_n(X^n) \xrightarrow{P} 0, \quad \phi_n(X^n) \xrightarrow{Q} 1, \tag{9.7}$$

for *all* $P \in B$ and $Q \in V$.

Existence of a pointwise test sequence for B versus V is equivalent to the existence of test sequence with property (D), (see proposition 9.4.2).

Aside from these two frequentist notions of testability, we also consider a version of the pointwise test that is strictly Bayesian, because it leaves room for a prior-null-set of exceptions [60, 164, 157].

Definition 9.2.3. Let $(\mathscr{P}, \mathscr{G})$ be a measurable space with prior Π and assume $B, V \in \mathscr{G}$. We say that (ϕ_n) is a *Bayesian test sequence* for B versus V (under Π), if,

$$\phi_n \xrightarrow{P} 0, \quad \phi_n \xrightarrow{Q} 1, \tag{9.8}$$

for Π*-almost-all* $P \in B$ and $Q \in V$.

The goal of this chapter is to characterize the existence of the test sequences with as much precision and in as much detail as possible, for the three definitions 9.2.1–9.2.3. We require an "accessible" form, that is, firstly we insist on easy illustration with a wide variety of examples and counterexamples, and secondly, that we elevate results of *existence* to *constructive* results (by applying the methods of [157]).

9.2.1 The Le Cam-Schwartz theorem

The basis for sections 9.3 and 9.4 is the Le Cam-Schwartz theorem. The following theorem is the Le Cam-Schwartz theorem, restated in test-specific form. Below $\mathscr{F}$ denotes a *increasingly directed* collection of model subsets (for any finite subset $\{F_1, \ldots, F_m\} \subset \mathscr{F}$, there exists an $F \in \mathscr{F}$ such that $F_1 \cup \ldots \cup F_m \subset F$. Examples: $\mathscr{F} = \{\mathscr{P}\}$; $\mathscr{F}$ consists of all *finite* subsets of $\mathscr{P}$; $\mathscr{F}$ consists of all *compact* subsets of $\mathscr{P}$, *etcetera*).

Theorem 9.2.4. (Le Cam-Schwartz, 1960) *Let $\mathscr{P}$ with hypotheses B, V $B \cap V = \varnothing$ be given. There exists an $\mathscr{F}$-uniform test sequence for B versus V, if and only if,*

there exists a sequence (ψ_n) *of* $\mathscr{U}_\infty$*-uniformly continuous functions* $\psi_n : B \cup V \to [0,1]$ *such that,*

$$\sup_{P \in F} \left| \psi_n(P) - 1_V(P) \right| \to 0, \tag{9.9}$$

for every $F \in \mathscr{F}$.

Proof. (See [173] and also section 17.5 of [179].) If there exists an $\mathscr{F}$-uniform test sequence (ϕ_n) for B versus V, then the functions $\psi_n : B \cup V \to [0,1]$,

$$\psi_n(P) = P^n \phi_n(X^n),$$

are $\mathscr{U}_\infty$-uniformly continuous and converge $\mathscr{F}$-uniformly to 1_V on $B \cup V$. Conversely, suppose first that $\psi_n = \psi$ for some $\mathscr{U}_\infty$-uniformly continuous $\psi : B \cup V \to [0,1]$. Let $\varepsilon > 0$ be given. There exist $m, J \geq 1$, $\delta > 0$ and $f_j : \mathscr{X}^m \to [0,1]$ $(1 \leq j \leq J)$, such that for all $P, Q \in B \cup V$,

$$\rho(P,Q) := \max_{1 \leq j \leq J} \left| P^m f_j - Q^m f_j \right| < \delta,$$

implies that $|\psi(P) - \psi(Q)| < \varepsilon$. Define M to be the smallest integer greater than $9J\varepsilon^{-1}\delta^{-2}$ and $\mathbb{P}_M f_j = M^{-1} \sum_{i=1}^M f_j(X_i)$, for random $X_1, \ldots, X_M \in \mathscr{X}^m$. Because any probability model is pre-compact for the uniform structure $\mathscr{U}_\infty$, there exist $L \geq 1$ and $\{Q_1, \ldots, Q_L\}$ such that, for all $P \in B \cup V$,

$$\min_{1 \leq l \leq L} \rho(P, Q_l) < \tfrac{1}{3}\delta.$$

Let $\hat{Q}_M$ denote the minimizer of $Q \mapsto \rho(\mathbb{P}_M, Q)$ over $\{Q_1, \ldots, Q_L\}$. For any P and l such that $\rho(P, Q_l) < \frac{1}{3}\delta$, we have,

$$\rho(\hat{Q}_M, P) \leq \rho(\mathbb{P}_M, P) + \rho(\mathbb{P}_M, \hat{Q}_M) \leq \rho(\mathbb{P}_M, P) + \rho(\mathbb{P}_M, Q_l) \leq 2\rho(\mathbb{P}_M, P) + \tfrac{1}{3}\delta.$$

For any $P \in B \cup V$, Chebyshev's inequality gives,

$$P^{Mm}(\rho(\mathbb{P}_M, P) \geq \tfrac{1}{3}\delta) \leq \sum_{j=1}^J P^{Mm}\left(|\mathbb{P}_M f_j - P^m f_j| \geq \tfrac{1}{3}\delta\right) \leq \sum_{j=1}^J \frac{9}{\delta^2} \mathrm{Var}(\mathbb{P}_M f_j) \leq \frac{9J}{\delta^2 M} < \varepsilon.$$

Conclude that for all $P \in B \cup V$, $P^{Mm}(|\psi(\hat{Q}_M) - \psi(P)| < \varepsilon) \geq 1 - \varepsilon$. This proves the following intermediate result: *for every* $\varepsilon > 0$ *and uniformly continuous* ψ*, the exists a sequence of estimators* $(\hat{Q}_n)$ *which satisfy,*

$$\sup_{P \in B \cup V} P^n |\psi(\hat{Q}_n) - \psi(P)| < \varepsilon, \tag{9.10}$$

for large enough n. Generalizing to the sequential case of uniformly continuous (ψ_k) satisfying (9.9), let a sequence (ε_k) be given that decreases to zero. For every $k \geq 1$, let $(\hat{Q}_{k,n})$ denote the estimator sequence that satisfies (9.10) for ψ_k and ε_k. By traversing the sequences labelled with k slowly enough with increasing n, we can

guarantee that,

$$\sup_{P\in B\cup V} P^n|\psi(\hat{Q}_{k(n),n}) - \psi_{k(n)}(P)| < \varepsilon_{k(n)}.$$

Combined with assumption (9.9), we see that,

$$\sup_{P\in F} P^n|\psi(\hat{Q}_{k(n),n}) - 1_V(P)| \to 0,$$

completing the proof.

This theorem solves the question for a topological characterization of an $\mathscr{F}$-uniform test sequence for B versus V elegantly and completely, at least, from a strictly mathematical perspective. However, the necessary and sufficient condition stated is not only technical and difficult to handle, it is very hard to interpret.

A possible interpretation runs as follows: model subsets can be told apart by *some* procedure involving *i.i.d.* data, if and only if, their distinctions can be expressed in terms of one specific model topology (or rather, uniformity), the topology $\mathscr{T}_\infty$ (uniformity $\mathscr{U}_\infty$), using uniform continuity of a sequence of functions on the model to specify possible distinctions exactly. The topology $\mathscr{T}_\infty$ (see definition **??**) is of the weak type and non-metrizable in all but the simplest cases.

Moreover $\mathscr{T}_\infty$ does not display any close relation to the samplespace topology, like Prohorov's weak topology. Le Cam qualifies the uniformity $\mathscr{U}_\infty$ as "not very easily accessible" [173], and more than 25 years later, as "rather inaccessible" [179]. It is therefore warranted to look for other equivalent formulations or accessible sufficient conditions. We start with the existence of uniform test sequences.

9.3 Uniform testability

A very natural question concerns conditions under uniform test sequences exists [17, 208]. Let us first establish the following useful equivalence [226, 177, 10, 68].

Proposition 9.3.1. *Let $\mathscr{P}$ be a model with hypotheses B and V, $B\cap V = \varnothing$. The following are equivalent:*

1. *there exists a uniform test sequence (ϕ_n) such that,*

$$\sup_{P\in B} P^n\phi_n \to 0, \quad \sup_{Q\in V} Q^n(1-\phi_n) \to 0,$$

2. *there exists a test sequence (ϕ_n) and a constant $D > 0$ such that,*

$$\sup_{P\in B} P^n\phi_n \leq e^{-nD}, \quad \sup_{Q\in V} Q^n(1-\phi_n) \leq e^{-nD}.$$

This fact can be exploited, for example, in Bayesian model selection, and consequently, in frequentist model selection with posteriors as well if a Kullback-Leibler prior is used: *c.f.* proposition 9.3.1, existence of a uniform test implies existence of

an exponentially powerful uniform test, which is enough to compensate for prior-mass lower-bounds for Kullback-Leibler neighbourhoods, *e.g.* through remote contiguity. According to theorem 4.8 of [157] posterior odds or Bayes factors then select the correct model consistently.

Here, we derive necessary and sufficient conditions for the existence of a uniform test sequence, in terms of the uniformity $\mathscr{U}_\infty$.

Definition 9.3.2. Let $(\mathscr{P},\mathscr{U})$ be a model with uniformity. We say that two model subsets B and V are *uniformly separated by* $\mathscr{U}$ if there exists an entourage $U \in \mathscr{U}$ such that for every $P,Q \in \mathscr{P}$, $(P,Q) \in U$ implies that either $P,Q \in B$, or $P,Q \in V$.

Theorem 9.3.3. *Let $\mathscr{P}$ be a model and let B,V be model subsets. The following are equivalent:*

(i.) there exists a uniform test sequence (ϕ_n) for B versus V, c.f. (9.6);
(ii.) the indicator $1_V : B \cup V \to \{0,1\}$ is $\mathscr{U}_\infty$-uniformly continuous;
(iii.) the subsets B and V are uniformly separated by $\mathscr{U}_\infty$.

Proof. Assume that there exists a uniform test sequence (ϕ_n) for B versus V. Define the $\mathscr{U}_\infty$-uniformly continuous functions $\psi_n : \mathscr{P} \to [0,1]$, $\psi_n(P) = P^n\phi_n$ and note that the difference $|\psi_n(P) - 1_V(P)|$ goes to zero uniformly over $B \cup V$. So, for every $\varepsilon > 0$ there exist an $N \geq 1$ such that for all $n \geq N$, $\sup_P |\psi_n - 1_V|(P) < \varepsilon/3$ and an entourage $W \in \mathscr{U}_\infty$ such that for all $(P,Q) \in W$, $|\psi_N(P) - \psi_N(Q)| < \varepsilon/3$. Therefore,

$$|1_V(P) - 1_V(Q)| \leq |1_V(P) - \psi_N(P)| + |\psi_N(P) - \psi_N(Q)| + |\psi_N(Q) - 1_V(Q)| < \varepsilon,$$

for all $(P,Q) \in V$. To show that *(ii.)* implies *(i.)*, choose $\psi_n = 1_V$. The equivalence of *(ii.)* to *(iii.)* follows directly from the definition of $\mathscr{U}_\infty$-uniform continuity of $1_V : \mathscr{P}' \to \{0,1\}$.

To expand on formulation *(iii.)*, B and V are uniformly separated by $\mathscr{U}_\infty$, if and only if, there exist $J,m \geq 1$, $\varepsilon > 0$ and bounded, measurable functions $f_1,\ldots,f_J : \mathscr{X}^m \to [0,1]$ such that,

$$\max_{1 \leq j \leq J} \left| P^m f_j - Q^m f_j \right| < \varepsilon,$$

implies that either $P,Q \in B$, or $P,Q \in V$. If the model is $\mathscr{T}_\infty$-compact, $m = 1$ suffices.

Proposition 9.3.4. *Let $\mathscr{P}$ be a model and let B,V be model subsets with $\mathscr{T}_\infty$-closures $\overline{B}$ and $\overline{V}$. If B and V are uniformly separated by $\mathscr{U}_\infty$, then $\overline{B} \cap \overline{V} = \varnothing$. If $\mathscr{P}$ is relatively $\mathscr{T}_\infty$-compact, the converse is also true.*

Proof. Suppose that there exists a $P \in \overline{B} \cap \overline{V}$. Let $W \in \mathscr{U}_\infty$ be any entourage. There exists an entourage $W' \in \mathscr{U}_\infty$ such that $W' \circ W' \subset W$. (Recall that $W' \circ W'$ denotes the collection of all pairs $(P,Q) \in \mathscr{P} \times \mathscr{P}$ for which there exists an $R \in \mathscr{P}$ such that $(P,R) \in W'$ and $(R,Q) \in W'$; more generally, see [46].) The sets $U_1 = \{P' \in \mathscr{P} : (P,P') \in W'\}$ and $U_2 = \{P' \in \mathscr{P} : (P',P) \in W'\}$ are neighbourhoods of the point P, so $U_1 \cap B \neq \varnothing$ and $U_2 \cap V \neq \varnothing$. Pick $P_1 \in U_1 \cap B$ and $P_2 \in U_2 \cap V$. Then $(P_1,P_2) \in W' \circ W' \subset W$ so that $W \cap B \times V \neq \varnothing$, *i.e.* W does not separate B from V

uniformly. Conversely, assume that the $\mathscr{T}_\infty$-closure $\overline{\mathscr{P}}$ of $\mathscr{P}$ is $\mathscr{T}_\infty$-compact and that B and V are not uniformly separated by $\mathscr{U}_\infty$, that is, for every $W \in \mathscr{U}_\infty$, there exists a pair $(P_W, Q_W) \in B \times V \cap W$. The collection $\{(P_W, Q_W) : W \in \mathscr{U}_\infty\}$ forms a net in $B \times V$. By the compactness of $\overline{\mathscr{P}}$, there exists a convergent subnet $(P_{W'}, Q_{W'})$ with a limit (P,P) on the diagonal of $\overline{\mathscr{P}} \times \overline{\mathscr{P}}$. So there exists a $P \in \overline{\mathscr{P}}$ and nets $(P_{W'}) \subset B$ and $(Q_{W'}) \subset V$ such that $P_{W'} \to P$ and $Q_{W'} \to P$. This shows that $P \in \overline{B} \cap \overline{V}$.

If $\mathscr{P}$ is not compact in $\mathscr{T}_\infty$, it is possible for two closed subsets to have no points in common, yet fail to be uniformly separated. (For comparison: two closed subsets of (the non-compact sets) $\mathbb{R}^k$ for $k \geq 1$ can have empty intersections but be at distance zero, for example $\{(x,y) \in (0,\infty) \times \mathbb{R} : y \geq 1/x\}$ and $\{(x,y) \in (0,\infty) \times \mathbb{R} : y \leq -1/x\}$ in $\mathbb{R}^2$. Note that this is not possible if we replace $\mathbb{R}^k$ by some compact subset.)

Example 9.3.5. Hellinger tests (from the minimax theorem)

It is shown in [177] that the centre of a total-variational ball cannot be tested with uniform testing power against the interior of said ball and this negative result stays true if we change from an L_1- to L_p-metrics for $p > 1$.

9.4 Pointwise testability

This section focusses on pointwise testability. Between Bayesian testability (which requires only Borel measurability) and uniform testability (which requires no less than uniform separation), pointwise testability is an interesting test-case where the question of testability may find its most natural or balanced answer.

In the first subsection, we consider pointwise testability according to definition 9.2.2 in models that are not dominated. Subsequent subsections focus on hypotheses that are asymptotically indistinguishable by statistical testing, and on a characterization of testable hypotheses in dominated models.

To start with a situation for which most have more intuition, consider the case of a model in which *consistent estimators* $\hat{P}_n : \mathscr{X}^n \to \mathscr{P}$ exists, $n \geq 1$. Here $\mathscr{P}$ is a set of single-observation distributions P, assumed Hausdorff in some topology $\mathscr{T}$. Consistency says that for every $P \in \mathscr{P}$ and neighbourhood U of P, we have $P^n(\hat{P}_n \in U) \to 1$. Given two *open* hypotheses $B, V \subset \mathscr{P}$ with $B \cap V = \varnothing$, define $\phi_n(X^n) = 1\{\hat{P}_n \in V\}$ and note that for any $P \in B$, B is a neighbourhood of P so $P^n\phi_n = P^n(\hat{P}_n \in V) \leq P^n(\hat{P}_n \notin B) \to 0$, and for any $Q \in V$, $Q^n(1-\phi_n) \to 0$. So (ϕ_n) is a pointwise test sequence for B versus V. If we restrict attention to $\mathscr{P}' = B \cup V$, B and V are complementary, so that B and V are both *clopen* sets, *i.e.* B, V both lie in the first ambiguous class $\Delta_1^0(\mathscr{P}')$. We summarize with the following proposition.

Proposition 9.4.1. *If $P \in \mathscr{P}$ can be estimated consistently and B is clopen, there exist pointwise tests for B versus its complement.*

So clopenness is *sufficient* if we can estimate, but is it also necessary? And which topologies on $\mathscr{P}$ are strong enough? Below we shall see that the topology $\mathscr{T}_\infty$ imposes itself and that, in fact, the requirement that B is both an F_σ and a G_δ (*i.e.*

that B lie in the second ambiguous class $\Delta_2^0(\mathscr{P})$) is necessary for the existence of pointwise tests, and in complete models, also sufficient.

9.4.1 Pointwise testability in non-dominated models

As a general, introductory remark, let us first prove that there is no difference between pointwise testability in its "almost-sure", "in-probability" and "in-expectation" versions.

Proposition 9.4.2. *Let $\mathscr{P}$ be a model with hypotheses B and V, $B \cap V = \varnothing$. The following are equivalent:*

1. *there exists a test sequence (ϕ_n) such that for all $P \in B$ and $Q \in V$,*

$$P^n \phi_n \to 0, \quad Q^n(1-\phi_n) \to 0,$$

2. *there exists a test sequence (ϕ_n) such that for all $P \in B$ and $Q \in V$,*

$$\phi_n(X^n) \xrightarrow{P} 0, \quad (1-\phi_n(X^n)) \xrightarrow{Q} 0,$$

3. *there exists a test sequence (ϕ_n) such that for all $P \in B$ and $Q \in V$,*

$$\phi_n(X^n) \xrightarrow{P\text{-a.s.}} 0, \quad (1-\phi_n(X^n)) \xrightarrow{Q\text{-a.s.}} 0.$$

This equivalence has a very useful and immediate implication: one is often interested in testing procedures that have property (D), which is a way to formulate the almost-sure version of testing in proposition 9.4.2. The *construction* of almost-sure test sequences is often difficult (see, however, [70]), but their *existence* can be inferred from the much-easier-to-prove in-probability pointwise testability of B versus V. This fact can be exploited, for example, in Bayesian model selection, and consequently [157], in frequentist model selection with posteriors, if a suitable prior is used.

Some testing problems do not require analysis at the level of the Le Cam-Schwartz theorem because a test sequence can readily be constructed.

Example 9.4.3. For fixed, measurable $D \subset \mathscr{X}$, can we test whether $\mathrm{supp}(P) \subset D$? *(I)*
For a measurable $D \subset X$, it is possible to test,

$$H_0 : P(D) = 1, \quad H_1 : P(D) < 1.$$

Namely, take test functions $\phi_n : \mathscr{X}^n \to [0,1]$, defined by,

$$\phi_n(X_1, \ldots, X_n) = 1 - \prod_{i=1}^{n} 1\{X_i \in D\},$$

Then, for all P satisfying H_0, $\phi_n(X^n) \xrightarrow{P\text{-a.s.}} 0$ and for all Q satisfying H_1, $\phi_n(X^n) \xrightarrow{Q\text{-a.s.}} 1$, which implies testability according to definition 9.2.2.

The above may seem trivial but it has many applications, for example the following.

Example 9.4.4. For any random $X \in [-\infty,\infty]$, is tightness of X testable?
Suppose we have a measurable map $X : \mathscr{X} \to [-\infty,\infty]$ and hypotheses,

$$H_0 : X \text{ is tight}, \quad H_1 : X \text{ is not tight}.$$

Define $C = \{\omega \in \mathscr{X} : |X(\omega)| = \infty\}$ and $D = \mathscr{X} \setminus C$ and apply example 9.4.3 to conclude that tightness is testable.

Another example concerns the presence of point-masses in the data-distribution.

Example 9.4.5. Can we test whether a distribution contains any point-masses?
In other words, we require a test for the hypotheses,

$$H_0 : \forall x \in \mathscr{X}, P(\{x\}) = 0, \quad H_1 : \exists x \in \mathscr{X}, P(\{x\}) > 0.$$

A suitable test sequence is constructed from ties in the sample,

$$\phi_n(X_1,\ldots,X_n) = 1 - \prod_{i,j=1, i\neq j}^{n} 1\{X_i \neq X_j\}.$$

If there exists an $x \in \mathscr{X}$ such that $P(\{x\}) = p > 0$, then the probabilities of seeing no ties decrease like $(1-p)^n$, so $P^n\phi_n \to 1$; while if H_0 holds, probabilities for ties are zero, $P(X_i = X_j) = 0$ if $i \neq j$, so $P^n\phi_n = 0$.

But in more complicated cases one needs the Le Cam-Schwartz theorem. Without formulating requirements on the model, we focus solely on the testability question itself: a pointwise test sequence (ϕ_n) for B versus V exists, if and only if, there exists a sequence of $\mathscr{U}_\infty$-uniformly continuous $\psi_n : B \cup V \to [0,1]$ such that $\psi_n(P) \to 1_V(P)$ for all $P \in B \cup V$. Let us first look at example 9.4.3 through the Le Cam-Schwartz equivalence.

Example 9.4.6. For fixed, measurable $D \subset \mathscr{X}$, can we test whether $\mathrm{supp}(P) \subset D$? *(II)*
Take the hypotheses of example 9.4.3. Let $B = \{P \in \mathscr{P} : P(D) = 1\}$ and $V = \{P \in \mathscr{P} : P(D) < 1\}$. Define the function $f : \mathscr{P} \to [0,1]$ by $f(P) = P(D)$ and the sequence $\psi_n = 1 - f^n$. Then the ψ_n are $\mathscr{U}_\infty$-uniformly continuous and $\psi_n(P) \to 0$ for all $P \in B$, while $\psi_n(P) \to 1$ for all $P \in V$.

Example 9.4.7. Can we test independence of two events A and B?
Let A and B be two measurable subsets of the sample space $\mathscr{X}$ for a single-observation. We test the hypotheses,

$$H_0 : P(A \cap B) = P(A)P(B), \quad H_1 : P(A \cap B) \neq P(A)P(B).$$

Consider the three $\mathscr{U}_1$-uniformly continuous functions $f_i : \mathscr{P} \to [0,1]$, $(i = 1,2,3)$,

$$f_1(P) = P(A \cap B), \quad f_2(P) = P(A), \quad f_3(P) = P(B),$$

and the uniformly continuous function $g : [0,1]^3 \to [1,-1]$, $g(x_1,x_2,x_3) = x_1 - x_2x_3$. The composition $h : \mathscr{P} \to [-1,1]$, $h = g \circ (f_1,f_2,f_3)$ and $|h|$ are $\mathscr{U}_1$-uniformly continuous, and so are the functions $\psi_n = |h|^{1/n}$. Note that $\psi_n(P) = 0$ for all $n \geq 1$ if $h(P) = 0$, and $\psi_n(P) \to 1$ in all other cases. So independence of events A and B is asymptotically testable.

9.4.2 Pointwise non-testability

As is pointed out in [91], conditions and examples for *non*-testability of hypotheses have been largely lacking for a long time (but see [70] for a notable exception). To better understand potential problems obstructing testability, we focus on necessary conditions for testability and hypotheses that are impossible to test.

Suppose that there exists a pointwise consistent test sequence (ϕ_n) for B versus V. Defining $\psi_n : \mathscr{P} \to [0,1]$,

$$\psi_n(P) = P^n\phi_n,$$

it is immediate that the ψ_n are all $\mathscr{U}_\infty$-uniformly continuous and that $\psi_n(P) \to 1_V(P)$ for every P. This implies that the nature of pointwise testable pairs of hypotheses B,V can be described quite precisely.

Proposition 9.4.8. *Suppose that there exists a pointwise consistent test sequence (ϕ_n) for B versus V. Then both B and V are both G_δ- and F_σ-sets with respect to $\mathscr{T}_\infty$ in the subspace $B \cup V$.*

Proof. Let $\varepsilon < 1/2$ be given, consider the closed sets $B_n = \{P \in B \cup V : \psi_n(P) \leq \varepsilon\}$ and $V_n = \{P \in B \cup V : \psi_n(P) \geq 1-\varepsilon\}$. For every $P \in B$ there exists an $N \geq 1$ such that P lies in the closed set $\cap_{n\geq N}^{\infty} B_n$. So P lies in the F_σ-set $\cup_{N\geq 1} \cap_{n\geq N}^{\infty} B_n$. Conversely, if P lies in $\cup_{N\geq 1} \cap_{n\geq N}^{\infty} B_n$, then $\psi_n(P) \leq \varepsilon$ for large enough n, which implies that $\psi_n(P) \to 0$ because we assume testability, so $P \in B$. Conclude that B is an F_σ-set. Since the same holds for V by symmetry, the complement of B in $B \cup V$ is also F_σ, that is, B is also G_δ.

In the language of descriptive set theory, hypotheses that are testable versus their complements in the model belong to the class of ambiguous sets $\Delta_2^0(\mathscr{P})$. Because $\mathscr{P}$ with the $\mathscr{T}_\infty$-topology is not necessarily metrizable, there is no guarantee that $\mathscr{T}_\infty$-open (or -closed) subsets are F_σ (or G_δ) in general. However, in *Polish* models (for examples, see remark 9.4.22 and corollary 9.4.23) testability implies completeness.

Corollary 9.4.9. *Suppose that $\mathscr{P} = B \cup V$ is Polish in the $\mathscr{T}_\infty$-topology and that B is pointwise testable versus V. Then the hypotheses B and V are complete (and hence Polish) subspaces of $\mathscr{P}$.*

Proof. Hypotheses B and V are both F_σ in $B\cup V$, if and only if, they are both G_δ in $B\cup V$, if and only if, they are both Polish in $B\cup V$. Metrizability and separability (in metrizable spaces) are subspace properties but completeness is not.

Testability of hypotheses in Polish spaces implies that the hypothesized sets B and V themselves are complete. That suggests that pointwise non-testability of a hypothesis can be shown based on the properties of *Baire spaces*.

Corollary 9.4.10. *Suppose $\mathscr{P}$ is a Baire space in a topology that is $\mathscr{T}_\infty$ or finer. If B and $\mathscr{P}\setminus B$ are dense in $\mathscr{P}$, then B is not pointwise testable versus $\mathscr{P}\setminus B$.*

Proof. We prove this by contradiction: assume that B is testable versus its complement in $\mathscr{P}$. Then B and its complement $C := \mathscr{P}\setminus B$ are both G_δ-sets, so there exist sequences of open sets (B_n) and (C_n) such that $B = \cap_{n=1}^\infty B_n$ and $C = \cap_{n=1}^\infty C_n$. Because both B and C are dense in the Baire space $\mathscr{P}$, the intersection $D = \cap\{B_n \cap C_n : n \geq 1\}$ is a countable intersection of dense open subsets, so D is dense. However, B and C are disjoint so,

$$\Big(\bigcap_{n=1}^{\infty} B_n\Big) \cap \Big(\bigcap_{n=1}^{\infty} C_n\Big) = B \cap C = \varnothing,$$

so the intersection D cannot be dense.

Remark 9.4.11. The condition that $\mathscr{P}$ be a Baire space is not as stringent as it looks: if $C \subset B$ and $W \subset V$, then non-testability of C versus W implies non-testability for B versus V. So the above corollary could have been formulated slightly more generally as follows: *if $B, V \subset \mathscr{P}$, $B\cap V = \varnothing$ are given and there exists a Baire subspace D of $\mathscr{P}$ in which both $D\cap B$ and $D\cap V$ are dense, then B is not testable versus V*. Aside from the remark that the Polish spaces we have discussed are Baire spaces, the Baire property is often applicable in dominated (sub-)models, under the condition of uniform integrability. Namely, because (locally) compact Hausdorff spaces are Baire spaces, and (relative) $\mathscr{T}_\infty$-compactness is often an easily accessible property (see the argument leading up to corollary 9.4.23), finding a Baire sub-problem D in examples is perhaps less demanding than it appears.

Example 9.4.12. Is Cover's rational means problem testable?
The above proves that Cover's rational mean problem has a negative answer. To prove this, first note that $\mathscr{P}$ is dominated and the Dunford-Pettis theorem shows that $\mathscr{P}$ is $\mathscr{T}_\infty$-compact (so that $\mathscr{T}_\infty = \mathscr{T}_1$). There is an injective parametrization $P_\cdot : [0,1] \to \mathscr{P}$, with $P_p(\{1\}) = 1 - P_p(\{0\}) = p$: any $p, q \in [0,1]$, $P_p \neq P_q$ and, given $f : \{0,1\} \to [0,1]$,

$$\big|(P_p - P_q)f\big| = \big|(p-q)(f(1) - f(0))\big| \leq |p-q|,$$

so that $P_\cdot$ is a $\mathscr{T}_\infty$-continuous injection and therefore a homeomorphism. Since $[0,1]$ is a complete metric space, $\mathscr{P}$ is a Baire space for the $\mathscr{T}_\infty$-topology. Because both $[0,1]\cap\mathbb{Q}$ and $[0,1]\setminus\mathbb{Q}$ are dense in $[0,1]$, the images $\mathscr{P}_0 := \{P_p : p \in [0,1]\cap\mathbb{Q}\}$ and

$\mathscr{P}_1 := \{P_p : p \in [0,1] \setminus \mathbb{Q}\}$ are $\mathscr{T}_\infty$-dense in $\mathscr{P}$. So there does not exist a pointwise test for $p \in [0,1] \cap \mathbb{Q}$ versus $p \in [0,1] \setminus \mathbb{Q}$.

Unfortunately, many common statistical assumptions are of this type.

Example 9.4.13. Is integrability of a random variable X, $P|X| < \infty$, testable?
Let X be a real-valued random variable with some distribution in the space $\mathscr{P}$ of all probability distributions on $\mathbb{R}$. When equipped with the total-variation norm or the Hellinger metric, $(\mathscr{P}, \mathscr{T}_d)$ refines $\mathscr{T}_\infty$, is a Polish space and therefore has the Baire property. Define the dichotomy $\mathscr{P}_0 = \{P \in \mathscr{P} : P|X| < \infty\}$, $\mathscr{P}_1 = \{P \in \mathscr{P} : P|X| = \infty\}$. The sets $\mathscr{P}_0$ and $\mathscr{P}_1$ are non-empty, so let $P \in \mathscr{P}_0$ and $Q \in \mathscr{P}_1$ be given. For any $0 < \varepsilon < 1$, $P' = (1-\varepsilon)P + \varepsilon Q$ satisfies $\|P' - P\| = \varepsilon\|(P+Q)\| \leq 2\varepsilon$, but $P' \in \mathscr{P}_1$. Conclude that $\mathscr{P}_1$ lies $\mathscr{T}_d$-dense in $\mathscr{P}$. Conversely, tightness of Q implies that for every $\varepsilon > 0$, there exists a constant $M > 0$ such that $|Q(A) - Q(A||X| \leq M)| < \varepsilon$ for all measurable $A \subset \mathbb{R}$. Since $Q(\cdot||X| \leq M) \in \mathscr{P}_0$, we also see that $\mathscr{P}_0$ lies $\mathscr{T}_d$-dense in $\mathscr{P}$. So $\mathscr{P}_0$ cannot be tested versus $\mathscr{P}_1$.

Since we cannot test for integrability of X, there is no asymptotic, statistical way of finding out whether use of the Law of Large Numbers is justified. In fact, integrability with regard to any unbounded random variable on $\mathbb{R}$ (*e.g.* $P|f(X)| < \infty$ for some $f : \mathbb{R} \to \mathbb{R}$) cannot be tested: in particular, square-integrability of X cannot be tested, so use of the Central Limit Theorem cannot be justified with tail-probability one either, based on an *i.i.d.* sample.

Example 9.4.14. Can we test whether a random variable X is compactly supported?

$$H_0 : \exists_K : P(X \in K) = 1, \quad H_1 : \forall_K : P(X \in K) < 1$$

Let all $P \in B$ be such that $P(X \in K) = 1$ for some compact K and all $Q \in V$ such that there is no such K. Then for all $0 < \varepsilon < 1$, $P' = (1-\varepsilon)P + \varepsilon Q \in V$ while $\|P - P'\| \leq 2\varepsilon$, so V lies $\mathscr{T}_d$-dense in $B \cup V$. Since $\mathbb{R}$ is a Radon space, for $\varepsilon > 0$ and any $Q \in V$ there exists a compact K such that $|Q(A) - Q(A|K)| < \varepsilon$ for all A. Therefore, also B lies $\mathscr{T}_d$-dense in $B \cup V$. Since the collection of all probability measures on $\mathbb{R}$ is completely metrizable in the $\mathscr{T}_d$ topology, $B \cup V$ is a Baire space and we conclude that there does not exist a pointwise test sequence for H_0 versus H_1.

Cover's rational mean problem can be called prototypical for non-testability of hypotheses, at least, if we are willing to restrict the issue to models that are Polish for $\mathscr{T}_\infty$ (for examples, see remark 9.4.22 and corollary 9.4.23). In Polish models we consider the potential testability of hypotheses that correspond to *analytic subsets*. (A subset A is *analytic* if it is the continuous image of a Polish space ([143], sections 7.F, 21.F.). The class of all analytic sets is very large; it contains all Borel subsets of $\mathscr{P}$.) To demonstrate how Cover's problem makes an appearance when non-testability is in play, we consider analytic subsets B and V of a Polish $\mathscr{P} = B \cup V$ that are not both F_σ-sets. Clearly B is not asymptotically pointwise testable versus V. Hurewicz's theorem [143] provides crucial insight.

Theorem 9.4.15. *(Hurewicz) Let $\mathscr{P}$ be a Polish space and let A be analytic in $\mathscr{P}$. If A is not F_σ, then there exists a Cantor subset C such that $C \setminus A$ is countably dense in C, homeomorphic to $[0,1] \cap \mathbb{Q}$, while $C \cap A$ is closed in A, and homeomorphic to $[0,1] \setminus \mathbb{Q}$.*

On the basis of the (more general) *Kechris-Louveau-Woodin theorem* [142, 143] we may say the following: suppose that $\mathscr{P}$ is Polish in the $\mathscr{T}_\infty$-topology. If we have two disjoint analytic model subsets B and V that are not both F_σ in $\mathscr{T}_\infty$ (which is necessary for testability), then there exists a sub-testing-problem, in the form of a *(Cantor) subset* $C \subset B \cup V$ with a representation on $[0,1]$ in which the testing of $B \cap C$ versus $V \cap C$ is represented on $[0,1]$ as Cover's rational mean problem. So if, for example, we are in a model that has finite total-variational entropy numbers (so that on the separable completion, total-variational and $\mathscr{T}_\infty$ topologies coincide), non-F_σ-ness of any analytic hypotheses can always be reduced to a non-testable Cover sub-problem.

9.4.3 Pointwise testability in dominated models

For the following theorem, recall that the testing problem *has a (uniform) representation on X*, if there exists a $\mathscr{T}_\infty$-(uniformly-)continuous surjective map $f : B \cup V \to X$ such that $f(B) \cap f(V) = \varnothing$.

Theorem 9.4.16. *Let a dominated model $\mathscr{P}$ with hypotheses B,V $B \cap V = \varnothing$ be given. The following are equivalent,*

i. there exists a pointwise test sequence for B versus V;
ii. the testing problem has a representation $f : B \cup V \to X$ on a normal space X and there exist disjoint F_σ-sets $B', V' \subset X$ such that $f(B) \subset B'$, $f(V) \subset V'$;
iii. the testing problem has a uniform representation $\psi : B \cup V \to X$ on a separable, metrizable space X with $\psi(B), \psi(V) \in \Delta_2^0(X)$.

The proof of this theorem requires some vector-space reasoning. Given the model $\mathscr{P}$ in the $\mathscr{T}_\infty$-topology, we define the linear space E of all bounded, continuous $f : \mathscr{P} \to \mathbb{R}$ and the linear space F that is the linear span of the collection of all degenerate (Borel) measures δ_P on $\mathscr{P}$: for any $\lambda \in F$, there exist $m \geq 1$, $\lambda_1, \ldots, \lambda_m \in \mathbb{R} \setminus \{0\}$ and distinct $P_1, \ldots, P_m \in \mathscr{P}$ such that λ can be written (uniquely) as:

$$\lambda = \sum_{i=1}^{m} \lambda_i \, \delta_{P_i}. \tag{9.11}$$

Definition of the bi-linear form $\langle \cdot, \cdot \rangle : E \times F \to \mathbb{R}$,

$$\langle f, \lambda \rangle = \int_{\mathscr{P}} f(P) \, d\lambda(P) = \sum_{i=1}^{m} \lambda_i \, f(P_i),$$

puts E and F in duality. This duality is separating for both E and F. (Because $\mathscr{P}$ is a uniform space (hence regular), point-sets are separated by continuous functions.) With the corresponding weak topologies $\sigma(E,F)$ and $\sigma(F,E)$, the spaces E and F form a dual pair of Hausdorff locally convex spaces (see, [50], Ch. 4). Note that the topology of pointwise convergence for bounded, continuous functions on $\mathscr{P}$ coincides with $\sigma(E,F)$. Within E, define,

$$H = \{f \in E : 0 \le f \le 1, f \ \mathscr{U}_\infty\text{-unif. cont.}\}.$$

The bi-polar theorem guarantees that for H, the closure $\overline{H}$ equals the bi-polar $H^{\circ\circ}$ which enables the following result.

Lemma 9.4.17. *Every $\mathscr{T}_\infty$-continuous $f : \mathscr{P} \to [0,1]$ lies in the $\sigma(E,F)$-closure $\overline{H}$ of H.*

Proof. According to the bi-polar theorem (see theorem 1 of [50], Ch. II, § 6, No. 3), the polar $H^{\circ\circ} \subset E$ of the polar $H^\circ \subset F$ is equal to the closed convex envelope of $H \cup \{0\}$. Since H is convex and contains $0 \in E$, $H^{\circ\circ} = \overline{H}$. For given $\lambda \in F$ there exists an $m \ge 1$, $\lambda_1, \ldots, \lambda_m \in \mathbb{R} \setminus \{0\}$ and distinct $P_1, \ldots, P_m \in \mathscr{P}$ such that λ is written uniquely as $\lambda = \Sigma_i \lambda_i \delta_{P_i}$. Fix some $m \ge 1$ and distinct $P_1, \ldots, P_m \in \mathscr{P}$ and consider the finite-dimensional subspace of F we obtain when we vary $w = (\lambda_1, \ldots, \lambda_m) \in \mathbb{R}^m$. Any $f \in H$ is represented on this subspace only through the values $v = (f(P_1), \ldots, f(P_m)) \in [0,1]^m$ and any λ supported on $\{P_1, \ldots, P_m\}$ lies in H° whenever the inner product $\langle v, w \rangle$ in $\mathbb{R}^m$ is greater than or equal to -1. Because the cube $[0,1]^m$ is the convex hull of its corner points, we see that if the coefficients $\lambda_1, \ldots, \lambda_m$ are such that,

$$\sum_{i \in M} \lambda_i \ge -1, \tag{9.12}$$

for any finite subset M of $\{1, \ldots, m\}$, then $\lambda \in H^\circ$. Conclude that if we define L to be the subset of all $\lambda \in F$ that satisfy (9.12) when decomposed according to (9.11), then $L \subset H^\circ$. Conversely, let $\lambda \in H^\circ$ be given (again represented in the form $\lambda = \Sigma_i \lambda_i \delta_{P_i}$). For every $1 \le i < j \le m$, define the $\mathscr{B}_n$-measurable maps $x^n \mapsto \phi_{ij,n}(x^n)$ to be likelihood ratio tests (with $\mu = P_i + P_j$ and $p_i = dP_i/d\mu$, $p_j = dP_j/d\mu$):

$$\phi_{ij,n}(X^n) = 1\{p_i^n(X^n) < p_j^n(X^n)\}.$$

Then, because the Hellinger distance $H(P_i, P_j)$ between P_i and P_j is strictly positive,

$$\begin{aligned}
&P_i^n \phi_{ij,n} + P_j^n(1 - \phi_{ij,n}) \\
&\quad = \int \left(p_i^n(x^n) 1\{p_i^n(x^n) < p_j^n(x^n)\} + p_j(x^n) 1\{p_i^n(x^n) \ge p_j^n(x^n)\}\right) d\mu^n(x^n) \\
&\quad \le \int \sqrt{p_i^n(x^n) p_j^n(x^n)}\, d\mu^n(x^n) = 1 - \frac{1}{2} \int \left(\sqrt{p_i^n(x^n)} - \sqrt{p_j^n(x^n)}\right)^2 d\mu^n(x^n) \\
&\quad = 1 - H^2(P_i^n, P_j^n) \le e^{-nH^2(P_i,P_j)} \to 0.
\end{aligned}$$

(For the last inequality, see, for example, lemma 2.17 in Strasser [236].) Choose some $0 < \varepsilon < 1/2$ and N large enough such that $P_i^N \phi_{ij,N} < \varepsilon$ and $P_j^N \phi_{ij,N} > 1 - \varepsilon$.

Then the function $f_{ij} : \mathscr{P} \to [0,1]$,

$$f_{ij}(P) = \Big(\frac{P^N \phi_{ij,N} - P_i^N \phi_{ij,N}}{P_j^N \phi_{ij,N} - P_i^N \phi_{ij,N}} \Big) \vee 0 \wedge 1,$$

is uniformly continuous with respect to $\mathscr{U}_N$ (which coarsens $\mathscr{U}_\infty$) and satisfies $f_{ij}(P_i) = 0$, $f_{ij}(P_j) = 1$, so f_{ij} lies in H and separates P_i and P_j. For every $M \subset \{1,\ldots,m\}$, we can construct an $f \in H$ from the collection $\{f_{ij} : 1 \leq i < j \leq m\}$, such that,

$$\langle f, \lambda \rangle = \sum_{i \in M} \lambda_i,$$

so that $H^\circ \subset L$ as well. Conclude that $H^\circ = L$. Next, consider the polar $H^{\circ\circ} \subset E$ of H°: let $\lambda \in H^\circ$ and $f : \mathscr{P} \to [0,1]$ in E be given. Reasoning like before we see that $\langle f, \lambda \rangle$ can be replaced by an inner product $\langle v, w \rangle$ in $\mathbb{R}^m$ and that $\langle f, \lambda \rangle \geq -1$ because the coefficients $\lambda_1, \ldots, \lambda_m$ satisfy (9.12). Conclude that $f \in H^{\circ\circ}$.

So any $\mathscr{T}_\infty$-continuous $f : \mathscr{P} \to [0,1]$ is approximated arbitrarily closely by $\mathscr{U}_\infty$-uniformly continuous functions, with respect to any of the semi-norms that define $\sigma(E,F)$. Although, for every continuous f, this implies the existence of *nets* of uniform functions that converge to f, nothing is implied regarding the existence of a convergent *sequence* of uniform functions. For that step, the conclusion of the next lemma is sufficient, however.

Lemma 9.4.18. *If $\mathscr{P}$ is dominated, $\overline{H}$ is separable and metrizable with respect to $\sigma(E,F)$.*

Proof. Because we assume that the σ-algebra $\mathscr{B}$ is countably generated, Strasser's lemma 4.1 [236] says that $\mathscr{P}$ is separable with respect to the total-variational topology. This implies that $\mathscr{P}$ is also separable in the $\mathscr{T}_\infty$-topology (because $\mathscr{T}_d$ refines $\mathscr{T}_\infty$, but see also theorems 4.4 and 21.3 in [236]). As the linear span of a set with countable dense subset, F (has a total set and) is separable with respect to $\sigma(F,E)$. And as a consequence of that, E is first-countable at zero with respect to $\sigma(E,F)$. The total-variational norm $\|\cdot\|_{TV}$ makes F a normed space, with continuous dual F', and F' can be equipped with the (weak-star) topology $\sigma(F',F)$. If we define, for every bounded $\mathscr{T}_\infty$-continuous $f : \mathscr{P} \to \mathbb{R}$, the linear map $g_f : F \to \mathbb{R}$,

$$g_f(\lambda) = \int_{\mathscr{P}} f(P)\, d\lambda(P) = \langle f, \lambda \rangle, \tag{9.13}$$

then, with $0 \leq |f| \leq \|f\| = \sup_{P \in \mathscr{P}} |f(P)|$,

$$\big| g_f(\lambda) \big| \leq \|f\|\, \|\lambda\|_{TV},$$

so g_f lies in F', for every $f \in E$. This map is one-to-one and a $\sigma(E,F)$-to-$\sigma(F',F)$ homeomorphism between E and $\underline{E} = \{g_f : f \in E\} \subset F'$, and we conclude that $\underline{E}$ is first-countable at zero. Also, every norm-bounded set in E (and in particular the set $\overline{H}$) is mapped to a norm-bounded subset of $\underline{E}$ (denoted G in the case of $\overline{H}$)

by (9.13), with norm-bounded closure $\overline{G}$ with respect to the norm-topology on F'. Then, according to cor. 2 of [50], Ch. III, § 3, No. 4, $\overline{G}$ is a compact, metrizable space for $\sigma(F',F)$, which implies that G is separable and metrizable with respect to $\sigma(F',F)$, which is equivalent to separability and metrizability of $\overline{H}$ for $\sigma(E,F)$.

Now a diagonalization argument suffices to draw the conclusion that if B and V are separated by a sequence of continuous functions (*i.e.* there exist continuous $\psi_n : B \cup V \to [0,1]$ such that $\psi_n \to 1_V$), then B is pointwise testable versus V. The representation in terms of F_σ-sets on a normal space guarantees the existence of the ψ_n through *Urysohn's lemma*.

Proof. (of theorem 9.4.16)
Assume condition *(ii)*. The disjoint sets B' and V' can be written as countable unions of closed sets and because X is a normal space there exists a sequence of continuous $g_n : X \to [0,1]$, $(n \geq 1)$ such that for each $x \in B'$ (resp. $y \in V'$), there is an $N \geq 1$ such that $g_n(x) = 0$ (resp. $g_n(y) = 1$) for all $n \geq N$. Composition with $f : B \cup V \to X$ gives rise to a sequence of $\mathscr{T}_\infty$-continuous $\psi_n = g_n \circ f : B \cup V \to [0,1]$ such that $\psi_n(P) \to 1_V(P)$ for all $P \in B \cup V$. For each $n \geq 1$, lemma 9.4.17 asserts that ψ_n lies in $\overline{H}$ and, according to lemma 9.4.18, $\overline{H}$ is metrizable for $\sigma(E,F)$, which implies the existence of a sequence $\{\psi_{n,m} : m \geq 1\} \subset H$ such that $\psi_{n,m} \to \psi_n$ with respect to $\sigma(E,F)$ as $m \to \infty$. Letting $m(n)$ increase with n slowly enough, a 'diagonal' sequence $\{\psi_{n,m(n)} : n \geq 1\}$ is constructed such that $\psi_{n,m(n)}(P) \to 1_V(P)$ for all $P \in B \cup V$. According to the Le Cam-Schwartz theorem, that implies the existence of a consistent pointwise test for B versus V, *i.e.* condition *(i)* follows from condition *(ii)*.

Next, assume condition *(i)*, that (ϕ_n) is a pointwise test sequence for B versus V. Define the $\mathscr{U}_\infty$-uniformly continuous maps $\psi_n : B \cup V \to [0,1]$, $\psi_n(P) = P^n\phi_n$, and the mapping $\psi : B \cup V \to \prod_n [0,1]$, $\psi(P) = (\psi_n(P) : n \geq 1)$. The map ψ is $\mathscr{U}_\infty$-uniformly continuous and the image $X = \psi(B \cup V)$ in the (separable, metrizable) product space $\prod_n [0,1]$ is separable and metrizable. Next, we reason similar to proposition 9.4.8: let $0 < \varepsilon < 1/2$ be given and consider the closed product sets $c_n, w_n \subset \prod_n [0,1]$,

$$c_n = [0,1] \times \ldots \times [0,1] \times [0,\varepsilon] \times [0,1] \times \ldots,$$
$$w_n = [0,1] \times \ldots \times [0,1] \times [\varepsilon,1] \times [0,1] \times \ldots,$$

(with the ε-dependent intervals as the n-th factors) and the sets $b_N = \cap_{n \geq N}(c_n \cap X)$, $v_N = \cap_{n \geq N}(w_n \cap X)$ for all $N \geq 1$ which are closed in the subspace X. Note that for any $P \in B$ (resp. any $Q \in V$), there exists an $N \geq 1$ such that $\psi(P) \in b_N$ (resp. $\psi(Q) \in v_N$), so $\psi(B)$ is a subset of the F_σ-set $\cup_N b_N$ in X (resp. $\psi(V)$ is a subset of the F_σ-set $\cup_N v_N$ in X). Conversely, if $x \in \cup_N b_N$ (resp. $y \in \cup_N v_N$), there exists a $P \in B \cup V$ such that $x = \psi(P)$ (resp. $y = \psi(P)$) and $\lim_n \psi_n(P) < 1/2$ (resp. $\lim_n \psi_n(P) > 1/2$), which means that $P \in B$ (resp. $P \in V$), *i.e.* $\cup_N b_N \subset \psi(B)$ (resp. $\cup_N v_N \subset \psi(V)$). So $\psi(B) = X \setminus \psi(V) \in \Delta_2^0(X)$. Condition *(iii)* follows from condition *(i)*. Condition *(ii)* follows from condition *(iii)* because metrizable spaces are normal spaces.

Corollary 9.4.19. *Suppose that $\mathscr{P}$ is dominated and there exist disjoint F_σ-sets B',V' in the completion $\hat{\mathscr{P}}$ (for $\mathscr{U}_\infty$) such that $B \subset B'$, $V \subset V'$. Then B is pointwise testable versus V.*

Proof. Since $\mathscr{P}$ is pre-compact for $\mathscr{U}_\infty$, the completion $\hat{\mathscr{P}}$ is compact (and hence normal) and the canonical embedding $\mathscr{P} \to \hat{\mathscr{P}}$ is continuous. Formulation *ii.* of theorem 9.4.16 is then satisfied, and we conclude formulation *i.*.

Remark 9.4.20. Based on corollary 9.4.19, it is tempting to conclude that testability must be equivalent to the existence of disjoint F_σ-sets B'',V'' such that $B \subset B''$ and $V \subset V''$ in the *original model* $\mathscr{P}$. But corollary 9.4.19 requires more: the existence of *disjoint* F_σ-sets B',V' in $\hat{\mathscr{P}}$ cannot be guaranteed from the existence of disjoint B'',V'' that are F_σ in $\mathscr{P}$. For the same reason, the converse of corollary 9.4.19 does not follow from corollary 9.4.8. This observation does allow for the following re-formulation, however: *suppose that $\mathscr{P}$ is dominated and complete for $\mathscr{U}_\infty$ with disjoint subsets B,V. Then B is pointwise testable versus V, if and only if, there exist disjoint F_σ-sets $B',V' \subset \mathscr{P}$ such that $B \subset B'$, $V \subset V'$.*

Although perhaps pleasantly succinct from a mathematical perspective, corollary 9.4.19 is not practical unless the model can easily be shown to be complete for $\mathscr{U}_\infty$. More common, for example, are models that describe a (possibly non-parametric) family of Lebesgue densities as a metric space, where the metric is related in some way (*e.g.* through inequalities) to the Hellinger or total-variational metrics. For the following corollary, we think of the model $\mathscr{P}$ as a metric space with a metric d that (is equal to or) refines the total-variational metric (*e.g.* for all $P,Q \in \mathscr{P}$, $\|P-q\| \leq f(d(P,Q))$ for some strictly increasing $f : [0,\infty) \to [0,\infty)$). The argument below gives an explanation for the ubiquity in the mathematical statistics literature of the assumption that the model has finite metric entropy numbers (*e.g.* for all $\varepsilon > 0$, the covering number $N(\varepsilon,\mathscr{P},d) < \infty$).

Corollary 9.4.21. *Suppose that $\mathscr{P}$ is dominated and totally bounded with respect to the total-variational metric. Then $B,V \subset \mathscr{P}$, $B \cap V = \varnothing$ are pointwise testable, if and only if, B,V are F_σ-sets for the total-variational topology in $B \cup V$.*

Proof. The closure $\overline{\mathscr{P}}$ of $\mathscr{P}$ in $\mathscr{M}(\mathscr{X},\mathscr{B})$ with respect to $\mathscr{T}_{TV}$ is compact. Because $\mathscr{T}_{TV}$ refines $\mathscr{T}_\infty$, the identity $i : \overline{\mathscr{P}} \to \overline{\mathscr{P}}$ is a $\mathscr{T}_{TV}$-to-$\mathscr{T}_\infty$ homeomorphism. The inverse i^{-1} is $\mathscr{T}_\infty$-to-$\mathscr{T}_{TV}$ continuous (and so is its restriction to $B \cup V$). Since the subspace $B \cup V$ remains metrizable for $\mathscr{T}_{TV}$, $i^{-1} : B \cup V \to B \cup V$ is a representation of the testing problem on a normal space.

Remark 9.4.22. To appreciate the role that metric entropy numbers play here, consider a model $\mathscr{P}$ of Lebesgue densities $p : [0,1] \to \mathbb{R}$ and equip it with the L^∞-norm, $d(P,Q) = \|p-q\|_\infty$, then certainly, $\|P-Q\|_{TV} \leq d(P,Q)$. Hypotheses that involve d, like testing for $\|\cdot\|_\infty$-neighbourhoods of a fixed $Q \in \mathscr{P}$,

$$H_0 : \|p-q\|_\infty < \delta, \quad H_1 : \|p-q\|_\infty \geq \delta, \tag{9.14}$$

are, in principle, not testable because these hypotheses are F_σ-sets in a topology that is strictly stronger than $\mathscr{T}_\infty$. But if the model is totally bounded with respect to d, the proof of corollary 9.4.21 extends because d refines total variation. That renders the hypotheses of (9.14) testable, because now they correspond to F_σ sets in the total-variational and $\mathscr{T}_\infty$ topologies.

Because $\mathscr{T}_\infty$ refines Prohorov's weak topology $\mathscr{T}_C$, we may also weaken the model topology to $\mathscr{T}_C$ by imposing $\mathscr{T}_\infty$-compactness. (Relative $\mathscr{T}_\infty$-compactness is a weaker requirement than relative compactness in total-variation.) The results of Ermakov (2017) [91] are formulated under this assumption and the main result of Dembo and Peres (1994) [69] relates to ours through the same construction.

Corollary 9.4.23. *Suppose that $\mathscr{P}$ is dominated by a probability measure, with a uniformly integrable family of densities. Then $B,V \subset \mathscr{P}$, $B \cap V = \varnothing$ are pointwise testable, if and only if, B,V are F_σ-sets for Prokhorov's weak topology in $B \cup V$.*

Proof. Consider the model $\mathscr{P}$ as a subspace of $\mathscr{M}(\mathscr{X},\mathscr{B})$ equipped with the $\mathscr{U}_1$-uniformity, and denote by $\overline{\mathscr{P}}$ the $\mathscr{T}_1$-closure of $\mathscr{P}$. Because $\mathscr{P}$ is dominated by a probability measure Q, $\overline{\mathscr{P}}$ is dominated by Q (see, *e.g.*, lemma 4.3 and theorem 4.8 in [236]). Clearly, the resulting family $\overline{\mathscr{P}}_Q$ of Q-densities in $L^1(Q)$ is the weak closure of $\mathscr{P}_Q$. In fact, embedding $L^1(Q)$ in $\mathscr{M}(\mathscr{X},\mathscr{B})$ canonically, $\overline{\mathscr{P}}_Q$ with the weak topology and $\overline{\mathscr{P}}$ with $\mathscr{T}_1$ are homeomorphic. By assumption, $\mathscr{P}_Q$ is relatively weakly compact, so $\overline{\mathscr{P}}_Q$ is weakly compact and $\overline{\mathscr{P}}$ is $\mathscr{T}_1$-compact. It is shown in the proof of lemma 3 of section 17.5 of Le Cam (1986) [179] (in the somewhat broader context of theorem 6 of appendix 8 in [179]) that weak convergence of a net $f_\alpha \to f$ in $L^1(Q)$ implies weak convergence of product densities $f_\alpha^n \to f^n$ weakly in $L^1(Q^n)$, as a result of the Dunford-Pettis theorem (see also lemma 3.8 in [236]). Suppose that f_α is an arbitrary net in $\overline{\mathscr{P}}_Q$, then there exists a convergent subnet $f_\beta \to f \in \overline{\mathscr{P}}_Q$ which implies that $f_\beta^n \to f^n$ weakly in $L^1(Q^n)$. That means that the set $\{f^n \in L^1(Q^n) : f \in \overline{\mathscr{P}}_Q\}$ is weakly compact and $\overline{\mathscr{P}}$ is compact with respect to $\mathscr{T}_n$, for all $n \geq 1$. Because compact spaces have unique uniformities compatible with their topologies, $\mathscr{U}_n = \mathscr{U}_1$ for all $n \geq 1$ and consequently $\mathscr{U}_\infty = \mathscr{U}_1 = \mathscr{U}_C$ on $\overline{\mathscr{P}}$. Therefore, the identity $i : \overline{\mathscr{P}} \to \overline{\mathscr{P}}$ is a $\mathscr{U}_\infty$-to-$\mathscr{U}_C$ homeomorphism of uniform spaces. Since $\overline{\mathscr{P}}$ is metrizable for $\mathscr{T}_C$, so is the subspace $B \cup V$, which means that $i : B \cup V \to B \cup V$ is a (uniform) representation of the testing problem on a normal space. Theorem 9.4.16 then asserts the existence of a pointwise test of B versus V.

Corollary 9.4.23 is related to theorem 2 in Dembo and Peres (1994) [69], which says that in a dominated model, a test sequence for B versus V exists if B and V are contained in disjoint F_σ-sets *for the $\mathscr{T}_C$-topology*, while the converse holds true whenever $\int (dP/dQ)^p dQ < \infty$ for some $p > 1$ and all $P \in \mathscr{P}$ (which implies uniform integrability). Ermakov formulates the following strengthening of the Dembo-Peres result.

Corollary 9.4.24. *(Ermarkov (2014), theorem 3.2)*
Suppose that $\mathscr{P}$ is dominated by a probability measure, with a uniformly integrable family of densities. Then $B,V \subset \mathscr{P}$, $B \cap V = \varnothing$ are pointwise testable, if and only if,

there exist sequences (B_m), (V_m) *with* $\cup_m B_m = B$ *and* $\cup_m V_m = V$, *and uniform test sequences* $(\phi_{m,n} : \mathscr{X}^n \to [0,1] : n \geq 1)$, $(m \geq 1)$, *such that,*

$$\sup_{P \in B_m} P^n \phi_{m,n} + \sup_{Q \in V_m} Q^n (1 - \phi_{m,n}) \to 0,$$

for all $m \geq 1$.

Proof. Assume that B is pointwise testable versus V; according to proposition 9.4.8, there exist sequences (B_m), (V_m), $(m \geq 1)$ of sets that are closed in the subspace $B \cup V$. Since $B \cup V \subset \mathscr{P}$ is relatively $\mathscr{T}_\infty$-compact, B_m and V_m are uniformly separated by $\mathscr{T}_\infty$ *c.f.* proposition 9.3.4 and theorem 9.3.3 implies existence of a uniform test sequence $(\phi_{m,n})$ for B_m versus V_m. Conversely, given tests $(\phi_{m,n})$, we choose $m(n)$ to traverse the sequence in $m = 1, 2, \ldots$ slowly enough to guarantee that $\phi_{m(n),n} : \mathscr{X}^n \to [0,1]$ is a pointwise test sequence for B versus V.

The above proof of Ermakov's theorem relies in a crucial way on (relative) compactness with respect to $\mathscr{T}_\infty$, because the all-important *existence* assertion in the proof follows from proposition 9.3.4.

Example 9.4.25. Are Cantor subsets of the right topological type to be testable? Cover's rational mean problem of hypotheses (9.2), concerning the parameter $p \in [0,1]$ of and *i.i.d.* sequence $X_1, X_2, \ldots$ of coin-flips, may also be posed with other hypotheses such as those of (9.4): can we test whether p lies in the Cantor subset $B = C \subset [0,1]$, or in its complement $V = [0,1] \setminus C$? As we have seen in example 9.4.12, the map $[0,1] \to \mathscr{P} : p \mapsto P_p$ is a $\mathscr{T}_\infty$-homeomorphism. In particular, this implies $\mathscr{P}$ is compact and metrizable for $\mathscr{T}_\infty$. Because B is closed, B is F_σ in $[0,1]$ and so is its image in $\mathscr{P}$. Because open sets in metrizable spaces are F_σ, V and its image in $\mathscr{P}$ are F_σ. We may now use theorem 9.4.16 or corollary 9.4.19 to conclude that C is testable versus its complement. More broadly, any (non-empty) topological space is homeomorphic to C, if and only if, it is perfect, compact, totally disconnected and metrizable. So the above concrete example represents a whole class of testing problems, those in which one of the hypotheses satisfies said characteristic topological properties as a subspace of a model $\mathscr{P} = B \cup V$ that is metrizable with respect to $\mathscr{T}_\infty$.

9.5 Bayesian test sequences

First of all, the existence of a Bayesian test sequence is linked directly to behaviour of the posterior itself. In the following, $\mathscr{P}$ is a model for *i.i.d.* data X^n taking values in a measurable space $(\mathscr{X}^n, \mathscr{B}^n)$. Assume that $X^n \sim P^n$, for some $P \in \mathscr{P}$ and all $n \geq 1$. Assume also that $\mathscr{P}$ has a σ-algebra $\mathscr{G}$ such that $P \mapsto P^n(A)$ is measurable for all $n \geq 1$ and $A \in \mathscr{B}^n$, and a prior $\Pi : \mathscr{G} \to [0,1]$.

Theorem 9.5.1. *Let a model* $(\mathscr{P}, \mathscr{G}, \Pi)$ *with hypotheses* $B, V \in \mathscr{G}$ *be given, with* $\Pi(B) > 0, \Pi(V) > 0$. *The following are equivalent,*

i. there exists a Bayesian test sequence for B versus V,

ii. there are test functions $\phi_n : \mathcal{X}^n \to [0,1]$ such that for Π-almost-all $P \in B, Q \in V$,

$$P^n\phi_n \to 0, \quad Q^n(1-\phi_n) \to 0,$$

iii. there are test functions $\phi_n : \mathcal{X}^n \to [0,1]$ such that,

$$\int_B P^n\phi_n \, d\Pi(P) + \int_V Q^n(1-\phi_n)\, d\Pi(Q) \to 0,$$

iv. for Π-almost-all $P \in B$, $Q \in V$,

$$\Pi(V|X^n) \xrightarrow{P} 0, \qquad \Pi(B|X^n) \xrightarrow{Q} 0.$$

Proof. Because $0 \le \phi_n \le 1$, *i.* implies *ii.*; by dominated convergence, *ii.* implies *iii.*; *iii.* leads to *iv.* through Martingale convergence and the inequality (see lemma 2.2 in [157]),

$$\int_B P^n \Pi(V|X^n)\, d\Pi(P) \le \int_B P^n\phi_n \, d\Pi(P) + \int_V Q^n(1-\phi_n)\, d\Pi(Q),$$

which holds for all $n \ge 1$ and any test sequence (ϕ_n); *iv.* gives *i.* when we set $\phi_n = \Pi(V|X^n)$. For details, see theorem 2.4 in [157].

An almost-sure version of definition 9.8 is also equivalent, through direct application of proposition 9.4.2, pointwise in a set of prior mass one. The interpretation of this theorem is gratifying to supporters of the likelihood principle and pure Bayesians: distinctions between model subsets are Bayesian testable, if and only if, they are picked up by the posterior asymptotically and the posterior itself can be viewed as the test function.

Breiman, Le Cam and Schwartz (1964) [51] provide a careful measurability argument to explain the essence of Doob's consistency theorem. The astonishing generality of Doob's theorem comes from the measure-theoretical (rather than topological) answer to questions related to posterior convergence. (Although the original reference for these notions is [51], a more complete exposé is found in Le Cam (1986) [179].)

Definition 9.5.2. Let $\mathscr{P}$ be a model with prior Π. An event $B \in \mathscr{B}^{(\infty)}$ is called a *Π-zero-one set*, if $P^\infty(B) = P^\infty(B)^2$, for Π-almost-all $P \in \mathscr{P}$. A model subset $G \in \mathscr{G}$ is called a *Π-one set* if there exists a Π-zero-one set B such that $G = \{P \in \mathscr{P} : P^\infty(B) = 1\}$.

The collection of all Π-one sets forms a sub-σ-algebra of $\mathscr{G}$, which we denote by $\mathscr{G}_1$. Let $\mathscr{G}_0$ denote the initial σ-algebra for the collection $\{P \mapsto P(A) : A \in \mathscr{B}\}$ (which coincides with the initial σ-algebra for the collection $\{P \mapsto P^\infty(A) : A \in \mathscr{B}^\infty\}$, see lemma 3.10 in [236]). Then $\mathscr{G}_0$ is contained in the Borel σ-algebra for $\mathscr{T}_1$. In order to make the next argument, we assume that the domain of the prior contains $\mathscr{G}_0$, for example if the prior is Borel for $\mathscr{T}_1$, $\mathscr{T}_\infty$ or total-variation. Asymptotic posterior convergence is then fully specified by the following observation.

Proposition 9.5.3. *Let $\mathscr{P}$ be a model with prior Π that contains $\mathscr{G}_0$. Let V be a Π-one set. Then,*

$$\Pi(V|X^n) \xrightarrow{P\text{-a.s.}} 1_V(P), \tag{9.15}$$

for Π-almost-all $P \in \mathscr{P}$.

Proof. Define the products $\Omega_n = \mathscr{P} \times \mathscr{X}^n$ with full product σ-algebras $\sigma(\mathscr{G} \times \mathscr{B}^n)$ and sub-σ-algebras $\mathscr{F}_n = \{\varnothing, \mathscr{P}\} \times \mathscr{B}^n$. The product $\Omega = \mathscr{P} \times \mathscr{X}^\infty$ with full product σ-algebra $\mathscr{F} = \sigma(\mathscr{G} \times \mathscr{B}^\infty)$ has a sub-σ-algebra $\mathscr{F}_\infty = \{\varnothing, \mathscr{P}\} \times \mathscr{B}^\infty$, and the filtration $\{\mathscr{F}_n : n \geq 1\}$ has limit $\mathscr{F}_\infty$. Given a prior Π on $(\mathscr{P}, \mathscr{G})$, a joint distribution $S : \mathscr{F} \mapsto [0,1]$ on Ω is fixed by defining $S(A \times B) = \int_A P^\infty(B)\, d\Pi(P)$ for $A \in \mathscr{G}$ and $B \in \mathscr{B}^\infty$ (which requires measurability of $P \mapsto P^\infty(B)$ for $B \in \mathscr{B}^\infty$). For any $\mathscr{F}$-measurable $g : \Omega \to [0,1]$ the conditional expectations $\{E_S[g|\mathscr{F}_n] : n \geq 1\}$ form a martingale. If, with slight abuse of notation, we maintain 1_V for the function $\Omega \to [0,1] : (P, x_\infty) \mapsto 1_V(P)$, we observe that the posteriors $\Pi(V|X^n) = E[1_V|\mathscr{F}_n]$ form a martingale relative to S. According to Doob's martingale convergence theorem there exists an $\mathscr{F}_\infty$-measurable random variable f_V such that $\Pi(V|X^n) \to f_V$, S-almost-surely. Since V is a Π-one set, there exists an event $B \in \mathscr{B}^\infty$ such that $1_V(P) = 1_B(x_\infty)$, S-almost-surely. Hence,

$$\Pi(V|X^n) = E_S\big[1_V \,\big|\, \mathscr{F}_n\big] = E_S\big[1_B \,\big|\, \mathscr{F}_n\big] \to E_S\big[1_B \,\big|\, \mathscr{F}_\infty\big] = 1_B = 1_V,$$

S-almost-surely, which amounts to, P-almost-surely for Π-almost-all P (by Fubini's theorem).

The remaining question, then, is whether the σ-algebra of Π-one sets is large enough to be interesting. The answer is given in proposition 2 of section 17.7 in Le Cam (1986) [179]: if the model is a Hausdorff space with Radon prior Π and the σ-field $\mathscr{B}$ on $\mathscr{X}$ is countably generated, then $\mathscr{G} = \mathscr{G}_1$. This implies Doob's consistency theorem (since any prior on a Polish space is Radon) and more, *c.f.* the corollary to proposition 2 of section 17.7 in [179] (beware of some typos and omissions in the proofs). We summarize and conclude as follows.

Theorem 9.5.4. *Let $(\mathscr{P}, \mathscr{G})$ be a measurable model, with a prior Π that is a Radon measure, and hypotheses B, V. There is a Bayesian test sequence for B versus V, if and only if, B, V are $\mathscr{G}$-measurable.*

Proof. In order for the definition of Bayesian testing to make sense, it is necessary that B and V are measurable. Conversely, if B is measurable and $V \subset \mathscr{P} \setminus B$, then $\phi_n(X^n) = \Pi(\mathscr{P} \setminus B|X^n)$ is a Bayesian test sequence for B versus V.

Example 9.5.5. Is Cover's rational mean problem Bayesian testable?
Let's revisit Cover's rational mean problem to illustrate the Bayesian answer: consider priors Π_B and Π_V for $B = [0,1] \cap \mathbb{Q}$ and $V = [0,1] \setminus \mathbb{Q}$ such that $\Pi_B(B) = 1$ and $\Pi_V(V) = 1$, (for example, enumerate $[0,1] \cap \mathbb{Q} = \{q_i : i \geq 1\}$ and define, for every measurable $F \subset [0,1]$,

$$\Pi_B(F) = \sum_{i \geq 1} 2^{-i}\, 1_F(q_i). \tag{9.16}$$

For Π_V we may simply choose Lebesgue measure. Set $\Pi = \frac{1}{2}\Pi_B + \frac{1}{2}\Pi_V$ on $[0,1]$. As we have seen in example 9.4.12, the compact space $[0,1]$ is homeomorphic to the model $\mathscr{P} = \{P_p : p \in [0,1]\}$ with the $\mathscr{T}_\infty$-topology through $p \mapsto P_p$. Therefore, $\mathscr{P}$ is Polish for $\mathscr{T}_\infty$, which implies that Π is Radon. Since B is Borel measurable in $[0,1]$, the corresponding subset of $\mathscr{P}$ is Borel measurable for $\mathscr{T}_\infty$, implying Bayesian testablility of B versus V. Proposition 9.5.3 even strengthens that to,

$$\Pi(p \in V|X^n) \xrightarrow{P_q\text{-a.s.}} 0, \quad \Pi(p \in V|X^n) \xrightarrow{P_r\text{-a.s.}} 1,$$

for $q \in [0,1] \cap \mathbb{Q}$ and $r \in [0,1] \setminus \mathbb{Q}$. So the tests $\phi(X^n) = \Pi(p \in V|X^n)$ (or the indicators for posterior odds of proposition 9.6.2) have property (D), albeit with a Π-null-set of exceptions. Indeed corollary 9.4.10 and example 9.4.12 establish that this Π-null-set is non-empty. So Cover's rational mean problem *does* have a Bayesian type solution. (It appears [60] that D. Blackwell made Cover aware of a Bayesian approach leading to a solution of the rational mean problem but failed to convince him fully of the validity of his alternative.)

To conclude this section, we provide an unexpected frequentist consequence of the Bayesian considerations of this section.

Theorem 9.5.6. *Let $\mathscr{P}$ be a model that is countable. Any $B, V \subset \mathscr{P}$ with $B \cap V = \varnothing$ are pointwise testable.*

Proof. For any two $P, Q \in \mathscr{P}$ there exists a measurable $0 \leq f \leq 1$ such that $Pf \neq Qf$, so $\mathscr{T}_1$ is the discrete topology on $\mathscr{P}$ (and so is $\mathscr{T}_\infty$). Any countable discrete space is Polish and the corresponding Borel σ-algebra is the power set of $\mathscr{P}$. Pick any (Borel) prior Π on $\mathscr{P}$ such that $\Pi(\{P\}) > 0$ for all $P \in \mathscr{P}$. Any V is Bayesian testable versus any disjoint B under Π and the test functions $\phi_n(X_1, \ldots, X_n) = \Pi(V|X_1, \ldots, X_n)$ form a Bayesian test sequence for B versus V. Because the only null-set of the prior is $\varnothing$, Bayesian test sequences under Π are also pointwise test sequences.

That means that the example of hypotheses (9.3) has full validity as a frequentist procedure.

Corollary 9.5.7. *(Dembo and Peres (1994))*
Regarding the parameter $p \in [0,1]$ for i.i.d.-Bernoulli-p distributed $X_1, X_2, \ldots$, there exists a pointwise test sequence that distinguishes,

$$H_0 : p \in [0,1] \cap \mathbb{Q}, \quad H_1 : p \in [0,1] \cap \sqrt{2} + \mathbb{Q},$$

asymptotically.

9.6 Bayesian testing power and model selection for frequentists

Proposition 9.5.3 settles the Bayesian question, but with Bayesian tests, more is possible. In the frequentist, constructive answer to the testability question, we shall

require control over the power of the tests. The following proposition from [157] formulates a general upper bound based on barycentres. (Denote the density for the local prior predictive distribution $P_n^{\Pi|B}$ with respect to $\mu_n = P_n^{\Pi|B} + P_n^{\Pi|V}$ by $p_{B,n}$, and similar for $P_n^{\Pi|V}$.)

Proposition 9.6.1. *Let $(\mathscr{P},\mathscr{G})$ be a model with priors (Π_n) and two measurable model subsets B,V with $\Pi(B),\Pi(V) > 0$. For every $n \geq 1$, there exists a $\phi_n : \mathscr{X}^n \to [0,1]$ such that,*

$$\int_B P^n \phi_n \, d\Pi(P) + \int_V Q^n (1-\phi_n) \, d\Pi(Q) \leq \int \Big(\Pi(B)\, p_{B,n}(x) \Big)^{\alpha} \Big(\Pi(V)\, p_{V,n}(x) \Big)^{1-\alpha} d\mu_n(x), \tag{9.17}$$

for any $0 \leq \alpha \leq 1$.

Proof. See proposition 2.6 in [157].

The following demonstrates that a sequence of tests based on posterior odds (or Bayes factors) is optimal, and thus obeys any upper bound for Bayesian testing power, including that of proposition 9.6.1 and the exponential bounds that follow from uniformly testable hypotheses and proposition 9.3.1.

Proposition 9.6.2. *Let $(\mathscr{P},\mathscr{G})$ be a model with priors (Π_n) and measurable model subsets B,V. For every $n \geq 1$, the test $\phi_n(X^n) = 1\{X^n : \Pi(V|X^n) \geq \Pi(B|X^n)\}$ based on posterior odds has optimal Bayesian testing power.*

Proof. Consider the decision-theoretic problem of setting the optimal $\phi \in [0,1]$ for picking B or V based on the data, with loss $\ell : \mathscr{P} \times [0,1] \to [0,1]$,

$$\ell(P,\phi) = \begin{cases} 0, & \text{if } P \notin B \cup V, \\ \big|\phi - 1_V(P)\big|, & \text{if } P \in B \cup V. \end{cases}$$

Data-driven decisions $\phi_n(X^n)$ for all $n \geq 1$ are test functions. The Bayesian risk functions,

$$r_n(\phi_n,\Pi) = \int_{\mathscr{P}} P^n \ell(P,\phi_n) \, d\Pi(P),$$

equal the Bayesian testing power,

$$\begin{aligned} r_n(\phi_n,\Pi) &= \int_B P^n |\phi_n - 1_V(P)| \, d\Pi(P) + \int_V Q^n |\phi_n - 1_V(Q)| \, d\Pi(Q) \\ &= \int_B P^n \phi_n \, d\Pi(P) + \int_V Q^n (1-\phi_n) \, d\Pi(Q), \end{aligned}$$

for all $n \geq 1$. Bayes's rule implies that if, for all $n \geq 1$ and P_n^{Π}-almost-all $x^n \in \mathscr{X}^n$, $\phi_n(x^n)$ is the minimizer,

$$\int_{\mathscr{P}} \ell(P,\phi_n(x^n)) \, d\Pi(P|X^n = x^n) = \inf_{\psi \in [0,1]} \int_{\mathscr{P}} \ell(P,\psi) \, d\Pi(P|X^n = x^n),$$

then $\phi_n(x^n)$ optimizes Bayesian testing power:

$$r_n(\phi_n, \Pi) = \inf_{\psi_n} r_n(\psi_n, \Pi),$$

(where the infimum runs over all possible choices $\psi_n : \mathscr{X}^n \to [0,1]$ for the n-th test function). To conclude, note that,

$$\begin{aligned}
&\int_{\mathscr{P}} \ell(P, \psi_n(x^n))\, d\Pi(P|X^n = x^n) \\
&\quad = \int_B \psi_n(x^n)\, d\Pi(P|X^n = x^n) + \int_V (1 - \psi_n(x^n))\, d\Pi(Q|X^n = x^n) \\
&\quad = \psi_n(x^n)\Pi(B|X^n = x^n) + (1 - \psi_n(x^n))\Pi(V|X^n = x^n),
\end{aligned}$$

is minimal if we choose $\psi_n(x^n) = 1\{x^n : \Pi(V|x^n) \geq \Pi(B|x^n)\}$.

We appeal to a theorem from [157] to make the final step in the proof that the existence of sufficiently powerful Bayesian tests, in combination with the requirement of remote contiguity (see definition 3.1 in [157]) of the local prior predictive distributions $P_n^{\Pi|B}$ with respect to the true distribution of the data P^n, implies that posteriors select the correct underlying hypothesis with probability growing to one.

Theorem 9.6.3. *For all $n \geq 1$, let the model be a measurable space $(\mathscr{P}, \mathscr{G})$ with priors $\Pi_n : \mathscr{G} \to [0,1]$. Consider disjoint, measurable $B, V \subset \Theta$ with $\Pi_n(B), \Pi_n(V) > 0$ such that,*

i. There exist Bayesian tests for B versus V of power $a_n \downarrow 0$,

$$\int_B P^n \phi_n\, d\Pi_n(P) + \int_V Q^n (1 - \phi_n)\, d\Pi_n(Q) = o(a_n),$$

ii. For every $P \in B$, $P^n \lhd a_n^{-1} P_n^{\Pi_n|B}$, and for every $Q \in V$, $Q^n \lhd a_n^{-1} P_n^{\Pi_n|V}$.

Then the indicators $\phi_n(X^n) = 1\{X^n : \Pi(V|X^n) \geq \Pi(B|X^n)\}$ for posterior odds form a pointwise test sequence for B versus V.

So if we can find a sequence of priors for which, *(a.)* we can prove the *existence* of a suitably powerful Bayesian test sequence, and which, *(b.)* induces remote contiguity at the right rate, the resulting posterior forms a *constructive* means (through posterior odds) to achieve consistent frequentist model selection.

9.7 Conclusions and discussion

Theorem 9.4.16 does not leave the model choice completely free, a *dominated* model is required. This restriction, however annoying, is of an essential nature because it is ultimately due to the sequential nature of the *i.i.d.* experiment. The shortest way to explain this is as follows. Using the Le Cam-Schwartz theorem to prove

the existence of tests, the most peculiar aspect of the conditions is the fact that a *sequence* of uniformly continuous functions is required. In the context of weak topologies, which are not first-countable in general, requiring existence of a convergent net or filter is natural but requiring existence of convergent *sequences* poses a considerable extra burden. If the weak topology in question happens to be metrizable, like Prokhorov's weak topology and the $\sigma(E,F)$-topology on the norm-bounded subset $\overline{H}$ of lemma 9.4.18, first-countability is recovered and sequential convergence coincides with net convergence, but $\mathscr{T}_\infty$ is not first-countable in general. Earlier work [69, 203, 91] solves this problem by requiring domination and uniform integrability, using $\mathscr{T}_\infty$-compactness *c.f.* the Dunford-Pettis theorem to equate $\mathscr{T}_\infty$ to Prokhorov's weak topology. In our argument, the problem is lifted by the (admittedly only sufficient) condition that the model is dominated. Given that the reason for this restriction (see the proof of theorem 9.2.4) is the sequential nature of *i.i.d.* experiments, it seems unlikely that there is a formulation of theorem 9.4.16 for non-dominated models of the same or very similar form.

9.7.1 Model assumptions

"There are statistical questions that I shouldn't even be thinking about... I can't afford to waste my time like that."

9.7.2 Model selection

Let $\mathscr{P}$ be a model for *i.i.d.* data $X^n \sim P^n$, $(n \geq 1)$, consisting of $M \geq 1$ disjoint sub-models, $\mathscr{P} = \mathscr{P}_1 \cup \ldots \cup \mathscr{P}_M$. Assume that $P \in \mathscr{P}$. The simplest form that the model-selection question takes, is to require asymptotically consistent selection of the sub-model $\mathscr{P}_i$ such that $P \in \mathscr{P}_i$.

9.8 Exercises

9.8.1. how that a topological space $\mathscr{X}$ is a *Baire space*, if and only if every *residual* subset A of $\mathscr{X}$ is *dense*, if and only if, every non-empty open subset of $\mathscr{X}$ is of the second categoryBaire category!second in $\mathscr{X}$.

Chapter 10
Application: non-parametric errors-in-variables regression

To demonstrate in a typical way how the methods presented in chapter 6 are applied in practice, we consider the asymptotic behaviour of the posterior distribution for the errors-in-variables model. The model describes measurements consisting of paired observations (X,Y) that are represented in terms of an unobserved Z. The random variable Z is related to X directly and to Y through a regression function f, both perturbed by Gaussian errors. We assume that Z falls into a (known) bounded subset of the real line with probability one, but otherwise leave its distribution unconstrained. In the semi-parametric literature, the regression function comes from a parametric (see Taupin (2001) [240]), or even linear (see, *e.g.* Anderson (1984) [6]) class of functions. In the following, we broaden that assumption to non-parametric regression classes, discussing the errors-in-variables problem also for Lipschitz and smooth functions.

Hence, the formulation we use involves two non-parametric components, the distribution of Z and the regression function f. We give Hellinger rates of convergence for the posterior distribution of the errors-in-variables density in non-parametric and parametric regression classes, using the posterior rate-of-convergence theorem 6.4.3 (or rather, a version based on the Hellinger metric entropy, *c.f.* Ghosal *et al.* (2000) [106]). Conditions that bound the rate of convergence can be decomposed into two terms, one for each of the non-parametric components of the model. The rate is then determined by the term that dominates the bound.

10.1 Errors-in-variables regression

When data is observed in pairs $(X,Y) \in \mathbb{R}^2$ and there is reason to assume that there is some unknown functional relation $f : \mathbb{R} \to \mathbb{R}$ between X and Y, observed with an additive *regression error* $e \in \mathbb{R}$, the most straightforward model is,

$$Y = f(X) + e. \tag{10.1}$$

Estimation then occurs in a family of possible *regression functions* f based on a sample (X_i, Y_i), $i \geq 1$, usually assuming that the e_i form an *i.i.d.* sample from some *error distribution*, typically with expectation equal to 0 (and commonly a known normal distribution). However popular, this model suffers from a serious shortcoming: it assumes that X has been observed with infinite precision, while very often there is some uncertainty in the observation of X. This uncertainty causes what is known as *attenuation bias*: because the observed X are only a noisy reflection of some unobserved quantity that determines the value of Y through f, the calculation is "blurred" horizontally. In the common case of an unknown linear $f(x) = ax + b$, estimation of a in the model (10.1) is biased towards 0, attenuating the regression function and making it resemble a constant function more appears reasonable graphically.

To prevent this, the errors-in-variables model studies a version of the regression model that takes the error in observed X into account explicitly: pairs (X, Y) are assumed to be distributed as,

$$\begin{aligned} X &= Z + e_1, \\ Y &= f(Z) + e_2, \end{aligned} \tag{10.2}$$

where (e_1, e_2) and Z are independent and $f : \mathbb{R} \to \mathbb{R}$ belongs to a family of regression functions. Usually, the distribution of the errors (e_1, e_2) is assumed to be known up to a (finite-dimensional) parameter σ whereas the distribution F of Z is completely unknown in the most general case. The primary interest lies in estimation of the regression function f from a *i.i.d.* sample of pairs $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ in the presence of the nuisance parameter F. Applications include all situations in which a functional dependence between measurements with errors is to be established.

The primary difference between errors-in-variables and ordinary regression using a set of design points $x_1, \ldots, x_n$, is the stochastic nature of the variable X. Regarding X, the variable e_1 is referred to as the "random error", whereas the variability of Z is said to be the "systematic error" (Anderson (1984) [6]). Kendall and Stuart (1979) [144] distinghuish between the "functional" errors-in-variables problem, in which Z is non-stochastic, taking on the values of 'design points' $z_1, \ldots, z_n$, and the "structural" errors-in-variables problem, in which Z is stochastic. Best known is linear errors-in-variables regression, in which f is assumed to depend linearly on z (see, *e.g.* [6] for an extensive overview of the literature). Efficient estimators for the parameters of f have been constructed by Bickel and Ritov (1987) [28], Bickel *et al.* (1998) [29] and Van der Vaart (1988, 1996) [245, 246]. Errors-in-variables regression involving a parametric family of non-linear regression functions has been analysed by Taupin and others (see Taupin (2001) [240] and references therein). In Fan and Troung (1993) [93], the rate of convergence (in a weighted L_2-sense) of Nadaraya-Watson-type kernel estimators for the conditional expectation of Y given Z (and hence for the regression function) are considered using deconvolution methods.

In this chapter we analyse the structural errors-in-variables problem for non-parametric families of regression functions in a Bayesian setting; we consider the

behaviour of posterior distributions for the parameter (σ, f, F) in the asymptotic limit. It is stressed that in this formulation, the errors-in-variables problem has two non-parametric components, one being the distribution of the underlying variable Z and the other the regression function. The emphasis lies on the interplay between these two non-parametric aspects of the model, as illustrated by their respective contributions to the rate of convergence (see, *e.g.* theorems 10.3.7 and 10.4.2).

10.1.1 Definition of the EIV model

We assume throughout this chapter that there is some known constant $A > 0$ such that $Z \in [-A, A]$ with probability one. Furthermore, we assume (unless indicated otherwise) that the errors e_1 and e_2 are independent and distributed according to the same normal distribution Φ_σ on $\mathbb{R}$ with mean zero and variance σ^2 (*i.e.* a special case of *restricted Gaussian errors* in the terminology of [28]). Writing φ_σ for the normal density of both e_1 and e_2, the model consists of a family of distributions for the observations (X, Y), parametrized by $(\sigma, f, F) \in I \times \mathscr{F} \times D$, where it is assumed that:

(a) I is a closed interval in the positive reals, bounded away from zero and infinity, *i.e.* $I = [\underline{\sigma}, \overline{\sigma}] \subset (0, \infty)$.
(b) D is the collection of all probability distributions on the compact symmetric interval $[-A, A]$, parametrized by all corresponding Stieltjes functions F.
(c) $\mathscr{F} \subset C_B[-A, A] \subset C[-A, A]$ is a bounded family of continuous regression functions $f : [-A, A] \to [-B, B]$. We shall distinguish various cases by further requirements, including equicontinuity, Lipschitz- and smoothness-bounds. Also considered is the parametric case, in which $\mathscr{F}$ is parametrized by a subset of $\mathbb{R}^k$.

For all $(\sigma, f, F) \in I \times \mathscr{F} \times D$, we define the following convoluted density for the distribution of observated pair (X, Y):

$$p_{\sigma,f,F}(x, y) = \int_{\mathbb{R}} \varphi_\sigma\big(x - z\big)\, \varphi_\sigma\big(y - f(z)\big)\, dF(z), \tag{10.3}$$

for all $(x, y) \in \mathbb{R}^2$.

It is stressed that when we speak of the errors-in-variables *model* $\mathscr{P}$, we refer to the collection of probability measures $P_{\sigma,f,F}$ on $\mathbb{R}^2$ defined by the Lebesgue-densities parametrized in the above display (rather than the parameter space $I \times \mathscr{F} \times D$). In many cases we regard $\mathscr{P}$ as a metric space, using either the Hellinger metric or $L_1(\mu)$-norm. As far as the parameter space is concerned, the model may not be identifiable: if, for given $F \in D$, two regression functions $f, g \in \mathscr{F}$ differ only on a set of F-measure zero, the corresponding densities $p_{\sigma,f,F}$ and $p_{\sigma,g,F}$ are equal on all of $\mathbb{R}^2$ (for all $\sigma \in I$). Determination of the true regression function f_0 based on an *i.i.d.* P_0-distributed sample can therefore be done only F_0-almost-everywhere (where $P_0 = P_{\sigma_0,f_0,F_0}$). Ultimately, the results we give are based on the Hellinger distance,

which, in the present circumstances, gives rise to a semi-metric on the parameter space $I \times \mathscr{F} \times D$ for the same reason. The 'well-known' identifiability problems in the linear errors-in-variables model (see *e.g.* Reiersøl (1950) [217]) arising due to interchangability of Gaussian components of the distribution of Z with the error-distribution (see also [6] and [28]) do not occur in our considerations, because we assume the distribution of Z to be compactly supported.

10.1.2 Posterior concentration theorem

Conditions for the theorem on Bayesian rates of convergence are again formulated in terms of the specific Kullback-Leibler neighbourhoods (6.13) of $P_0 \in \mathscr{P}$. Recall the Ghosh-Ghosal-van der Vaart theorem, which we write here with the help of entropy condition (6.21), where $N(\varepsilon, \mathscr{P}, H)$ denote the *covering numbers* with respect to the Hellinger metric on $\mathscr{P}$, *i.e.* the minimal number of Hellinger balls of radius $\varepsilon > 0$ needed to cover $\mathscr{P}$.

Theorem 10.1.1. *Let $\mathscr{P}$ be a model and assume that the sample $U_1, U_2, \ldots$ is i.i.d. P_0-distributed for some $P_0 \in \mathscr{P}$. For a given prior Π, suppose that there exists a sequence of strictly positive numbers ε_n with $\varepsilon_n \to 0$ and $n\varepsilon_n^2 \to \infty$ and constants $R_1, R_2 > 0$, such that:*

$$\Pi\big(B(\varepsilon_n; P_0)\big) \geq e^{-R_1 n \varepsilon_n^2}, \tag{10.4}$$

$$\log N\big(\varepsilon_n, \mathscr{P}, H\big) \leq R_2 n \varepsilon_n^2, \tag{10.5}$$

for all large enough n. Then, for every sufficiently large constant M, the posterior distribution satisfies:

$$\Pi_n\big(P \in \mathscr{P} : H(P, P_0) \geq M\varepsilon_n \,\big|\, U_1, \ldots, U_n\big) \to 0, \tag{10.6}$$

as $n \to \infty$, in P_0-expectation.

The assumption that the model is well-specified, *i.e.* $P_0 \in \mathscr{P}$, can be relaxed. In Kleijn and Van der Vaart (2006) [151], the above theorem is given in the case of a misspecified model. We do not give misspecified versions of the results, although we believe that the conditions of the necessary theorems in [151] are met in the model we consider.

10.2 Rates of posterior convergence in function spaces

10.2.1 Lipschitz and smoothness classes

We consider regression classes $\mathscr{F}$ contained within the class $C_B[-A,A]$ of all continuous functions $f : [A,A] - \to \mathbb{R}$ bounded by a (known) constant $B > 0$. At the very least, we also require equicontinuity of $\mathscr{F}$, which guarantees compactness in the topology of the uniform norm $\|\cdot\|$ according to the *Arzelá-Ascoli theorem*. Consequently, covering numbers $N(\varepsilon, \mathscr{F}, \|\cdot\|)$ are finite and an important part of the argument rests on bounds on these covering numbers we establish later. We distinguish several non-parametric and parametric examples of such classes below, but remark that other regression classes for which bounds on covering numbers exist, can also be used.

(i) $\mathrm{Lip}_M(\alpha)$ (for some $M > 0$ and $0 < \alpha \leq 1$), the class of all Lipschitz functions $f \in C_B[-A,A]$ with constant M and exponent α, *i.e.*

$$\left|f(z) - f(z')\right| \leq M|z - z'|^{\alpha}, \tag{10.7}$$

for all $z, z' \in [-A,A]$.

(ii) $D_{\alpha,M}(q)$ (for some $0 < \alpha \leq 1$, $M > 0$ and an integer $q \geq 1$), the class of all q times differentiable functions $f \in C_B[-A,A]$ for which the q-th derivative $f^{(q)}$ belongs to $\mathrm{Lip}_M(\alpha)$.

(iii) $\mathscr{F}_\Theta$, a parametric class of regression functions which forms a subset of $\mathrm{Lip}_M(\alpha)$ for some $\alpha \in (0,1]$ and $M > 0$. We assume that there exists a bounded, open subset $\Theta \subset \mathbb{R}^k$ for some $k \geq 1$ such that $\mathscr{F}_\Theta = \{f_\theta : \theta \in \Theta\}$. Furthermore, we assume that the map $\theta \mapsto f_\theta$ is Lipschitz-continuous, *i.e.* there exist constants $L > 0$ and $\rho \in (0,1]$ such that for all $\theta_1, \theta_2 \in \Theta$:

$$\|f_{\theta_1} - f_{\theta_2}\| \leq L\|\theta_1 - \theta_2\|^{\rho}_{\mathbb{R}^k}. \tag{10.8}$$

Often, it is more convenient to unify cases (i) and (ii) above, by considering the family of classes $C_{\beta,L}[-A,A]$ defined as follows. For given $\beta > 0$ and $L > 0$, we define β to be the greatest integer such that $\beta < \beta$ and we consider, for suitable functions $f : [-A,A] \to \mathbb{R}$, the norm:

$$\|f\|_\beta = \max_{k \leq \beta} \|f^{(k)}\| + \sup_{z_1, z_2} \frac{\left|f^{(\beta)}(z_1) - f^{(\beta)}(z_2)\right|}{|z_1 - z_2|^{\beta - \beta}},$$

where the supremum is taken over all pairs $(z_1, z_2) \in [-A,A]^2$ such that $z_1 \neq z_2$. The class $C_{\beta,L}[-A,A]$ is then taken to be the collection of all continuous $f : [-A,A] \to \mathbb{R}$ for which $\|f\|_\beta \leq L$. Note that for $0 < \beta \leq 1$, $\beta = 0$ and $C_{\beta,L}[-A,A]$ is a Lipschitz class bounded by L; if $\beta > 1$, differentiability of a certain order is implied, as well as boundedness of all derivatives and a Lipschitz property for the highest derivative.

10.2.2 Competing entropy bounds

As indicated in subsection 10.1.2, the Hellinger rate of convergence ε_n is bounded by two conditions, one related to the small-ε behaviour of the (Hellinger) entropy of the model, the other by the small-ε behaviour of the prior mass in Kullback-Leibler neighbourhoods of the form (6.13). The first condition is considered in section 10.3: theorem 10.3.7 says that the Hellinger covering number of the errors-in-variables model has an upper bound that consists of two terms, one resulting from the (σ, F)-part of the model and the other from minimal covering of the regression class:

$$\log N(\varepsilon, \mathscr{P}, H) \leq L_0 \left(\log \frac{1}{\varepsilon} \right)^3 + \log N(L\varepsilon, \mathscr{F}, \|.\|), \tag{10.9}$$

for small $\varepsilon > 0$ and some constants $L, L_0 > 0$. If the regression class $\mathscr{F}$ is 'small' enough, in the sense that the first term in the entropy bound displayed above dominates in the limit $\varepsilon \to 0$, the candidate rates of convergence ε_n are parametric up to a logarithmic factor.

Lemma 10.2.1. *If there exists a constant $L_1 > 0$ such that:*

$$\log N(\varepsilon, \mathscr{F}, \|.\|) \leq L_1 \left(\log \frac{1}{\varepsilon} \right)^3, \tag{10.10}$$

for small enough $\varepsilon > 0$, then the entropy condition (10.5) is satisfied by the sequence:

$$\varepsilon_n = n^{-1/2} (\log n)^{3/2}, \tag{10.11}$$

for large enough n.

Proof. Under the above assumption, $\log N(\varepsilon, \mathscr{P}, H)$ is upper bounded by the first term in (10.9) with a larger choice for the constant. Note that the sequence ε_n as defined in (10.11) satisfies $\varepsilon_n \downarrow 0$ and $n\varepsilon_n^2 \to \infty$. Also note that $\varepsilon_n \geq 1/n$ for large enough n, so that for some $L > 0$,

$$\log N(\varepsilon_n, \mathscr{F}, \|.\|) \leq \log N(1/n, \mathscr{F}, \|.\|) \leq L(\log n)^3,$$

and $n\varepsilon_n^2 = (\log n)^3$, which proves that ε_n satisfies (10.5).

It is also possible that the small-ε behaviour of the errors-in-variables entropy is dominated by the covering numbers of the regression class. In that case the *r.h.s.* of (10.9) is replaced by a single term proportional to $\log N(L\varepsilon, \mathscr{F}, \|.\|)$ for small enough ε. If the regression functions constitute a Lipschitz or smoothness class, lemma 10.5.1 gives the appropriate upper bound for the entropy, leading to the following candidate rates of convergence.

Lemma 10.2.2. *For an errors-in-variables model $\mathscr{P}$ based on a regression class $C_{\beta,M}[-A, A]$, the entropy condition (10.5) is satisfied by the sequence:*

$$\varepsilon_n = n^{-\frac{\beta}{2\beta+1}}, \tag{10.12}$$

for large enough n.

Proof. As argued above, the Hellinger entropy of the errors-in-variables model is upper-bounded as follows:

$$\log N(\varepsilon, \mathscr{P}, H)) \leq \frac{K}{\varepsilon^{1/\beta}},$$

for some constant $K > 0$ and small enough ε. The sequence ε_n satisfies $\varepsilon_n \downarrow 0$ and $n\varepsilon_n^2 \to \infty$. Furthermore, note that:

$$\log N(\varepsilon_n, \mathscr{P}, H) \leq K n^{1/(2\beta+1)} = Kn \cdot n^{-\frac{2\beta}{2\beta+1}} = Kn\varepsilon_n^2,$$

for large enough n.

10.2.3 Competing lower bounds on prior mass

Similar reasoning applies to condition (10.4) for the small-ε behaviour of the prior mass of Kullback-Leibler neighbourhoods of the form (6.13). Section 10.4 discusses the necessary lemmas in detail. We define priors Π_I, $\Pi_{\mathscr{F}}$ and Π_D on the parametrizing spaces I, $\mathscr{F}$ and D respectively and choose the prior Π on the model $\mathscr{P}$ as induced by their product under the map $(\sigma, f, F) \mapsto P_{\sigma,f,F}$ (which is measurable, as shown in lemma 10.4.1). The prior Π_I is chosen as a probability measure on I with continuous and strictly positive density with respect to the Lebesgue measure on I. Priors for the various regression classes discussed in the beginning of this section are discussed in subsection 10.5.2. The prior Π_D on D is based on a Dirichlet process with base measure α which has a continuous and strictly positive density on all of $[-A, A]$.

As with the covering numbers discussed above, we find (see theorem 10.4.2) that (the logarithm of) the prior mass of Kullback-Leibler neighbourhoods is lower bounded by two terms, one originating from the prior on the regression class and the other from the priors on the remaining parameters in the model:

$$\log \Pi\Big(B\big(K\delta\log(1/\delta); P_0\big)\Big) \geq -c\Big(\log\frac{1}{\delta}\Big)^3 + \log \Pi_{\mathscr{F}}\big(f \in \mathscr{F} : \|f - f_0\| \leq \delta\big), \tag{10.13}$$

for some constants $K, c > 0$ and small enough $\delta > 0$. If the prior mass in $\mathscr{F}$ around the true regression function f_0 does not decrease too quickly with decreasing δ, the bound that dominates (10.13) is proportional to the first term on the *r.h.s.*, which leads to near-parametric candidate rates of convergence.

Lemma 10.2.3. *If there exists a constant $c' > 0$ such that:*

$$\log \Pi_{\mathscr{F}}\big(f \in \mathscr{F} : \|f - f_0\| \leq \varepsilon\big) \geq -c'\Big(\log\frac{1}{\varepsilon}\Big)^3, \tag{10.14}$$

for small enough $\varepsilon > 0$, then the prior-mass condition (10.4) is satisfied by the sequence (10.11) for large enough n.

Proof. Condition (10.14) implies that (10.13) holds with the lower bound on the *r.h.s.* replaced by only its first term with a larger choice for the constant c. The substitution $\varepsilon = K\delta\log(1/\delta)$ leads to a constant and a $\log\log(1/\delta)$ correction, both of which are dominated by $\log(1/\delta)$ for small enough δ. (See the proof of lemma 10.2.4, where a similar step is made.) It follows that:

$$\log\Pi\big(B(P_0,\varepsilon)\big) \geq -c''\left(\log\frac{1}{\varepsilon}\right)^3,$$

for some constant $c'' > 0$ and small enough ε. The remainder of the proof is identical to that of lemma 10.2.1.

However, it is also possible that the prior mass around f_0 in the regression class decreases more quickly than (10.14). In that case the lower bound on the *r.h.s.* of (10.13) is determined by the prior on $\mathscr{F}$. The following lemma assumes a so-called *net-prior* on the regression class $\mathscr{F}$, a construction that is explained in subsection 10.5.2.

Lemma 10.2.4. *For an errors-in-variables model $\mathscr{P}$ based on a regression class $C_{\beta,M}[-A,A]$ with a net-prior Π, the prior-mass condition (10.4) is satisfied by the sequence:*

$$\varepsilon_n = n^{-\frac{\beta}{2\beta+1}}(\log n)^{\frac{1}{2\beta}}, \tag{10.15}$$

for large enough n.

Proof. Given β, the prior mass in neighbourhoods of the true regression function f_0 for a net prior Π is lower bounded by the expression on the *r.h.s.* in (10.39). Since this term dominates in the *r.h.s.* of (10.13) for small δ, the prior mass of Kullback-Leibler neighbourhoods of P_0 in $\mathscr{P}$ satisfies the following lower bound:

$$\log\Pi\Big(B\big(K\delta\log(1/\delta);P_0\big)\Big) \geq -L\frac{1}{\delta^{1/\beta}},$$

for some constants $K, L > 0$ and small enough δ. Define $\varepsilon = K\delta\log(1/\delta)$ and note that, for small enough δ:

$$\begin{aligned}
\frac{1}{\varepsilon^{1/\beta}}\left(\log\frac{1}{\varepsilon}\right)^{1/\beta} &= K^{-1/\beta}\frac{1}{\delta^{1/\beta}}\left(\log\frac{1}{\delta}\right)^{-1/\beta}\left(\log\frac{1}{\delta} - \log K - \log\log\frac{1}{\delta}\right)^{1/\beta} \\
&\geq K^{-1/\beta}\frac{1}{\delta^{1/\beta}}\left(\log\frac{1}{\delta}\right)^{-1/\beta}\left(\tfrac{1}{2}\log\frac{1}{\delta}\right)^{1/\beta} \\
&\geq \left(\tfrac{1}{2}\right)^{1/\beta}K^{-1/\beta}\frac{1}{\delta^{1/\beta}}.
\end{aligned}$$

For the first inequality in the above display, we have used that $\log K \leq \log\log\frac{1}{\delta} \leq \frac{1}{4}\log\frac{1}{\delta}$ (for small enough δ). We see that there exists a constant $L' > 0$, such that,

for small enough $\varepsilon > 0$:

$$\log \Pi\big(B(P_0, \varepsilon)\big) \geq -L' \frac{1}{\varepsilon^{1/\beta}} \Big(\log \frac{1}{\varepsilon}\Big)^{1/\beta}.$$

The sequence ε_n satisfies $\varepsilon_n \downarrow 0$ and $n\varepsilon_n^2 \to \infty$. Define the sequence $a_n = n^{-\beta/(2\beta+1)}$ and note that $\varepsilon_n \geq a_n$ (for large enough n) so that for some constant $R > 0$:

$$\log \Pi\big(B(P_0, \varepsilon_n)\big) \geq \log \Pi\big(B(a_n; P_0)\big) \geq -L' \frac{1}{a_n^{1/\beta}} \Big(\log \frac{1}{a_n}\Big)^{1/\beta}$$

$$= -Rn^{\frac{1}{2\beta+1}} (\log n)^{\frac{1}{\beta}} = -Rn\varepsilon_n^2,$$

for large enough n.

10.2.4 Various rates of posterior convergence

In the case of a parametric regression class ($\mathscr{F}_\Theta$ as defined under case *(iii)* in the beginning of this section) and a prior on Θ with strictly positive and continuous density, the conditions of lemmas 10.2.1 and 10.2.3 are satisfied. From lemma 10.5.3, we know that in the case of a parametric class of regression functions, covering numbers satisfy (10.10). Furthermore, from lemma 10.5.5, we know that for a parametric class, the prior mass in neighbourhoods of f_0 satisfies (10.14). The resulting conclusion for the rate of convergence in parametric regression classes is given in the theorem below.

We summarize the main results in the following theorem by stating the rates of convergence for the classes defined in the beginning of this section. The proof consists of combination of the preceding lemmas.

Theorem 10.2.5. *For the specified regression classes, the assertion of theorem 10.1.1 holds with the following rates of convergence.*

(i) If $\mathscr{F} = \mathrm{Lip}_M(\alpha)$ (for some $\alpha \in (0, 1]$ and $M > 0$) with a net prior, the prior-mass condition for neighbourhoods of f_0 in the regression class determines the rate, given by the sequence ε_n defined in lemma 10.2.4 with $\beta = \alpha$:

$$\varepsilon_n = n^{-\frac{\alpha}{2\alpha+1}} (\log n)^{\frac{1}{2\alpha}}.$$

(ii) If $\mathscr{F} = D_{\alpha,M}(q)$ (for some $M > 0$ and integer $q \geq 1$) with a net prior, the prior-mass condition for neighbourhoods of f_0 again determines the rate, given by the sequence ε_n defined in lemma 10.2.4 with $\beta = q + \alpha$:

$$\varepsilon_n = n^{-\frac{q+\alpha}{2q+2\alpha+1}} (\log n)^{\frac{1}{2q+2\alpha}}.$$

(iii) If $\mathscr{F} = \mathscr{F}_\Theta$ is a parametric class with a prior that has a continuous and strictly positive density throughout Θ, the rate is determined by the posterior convergence with regard to the parameter F and is given by:

$$\varepsilon_n = n^{-1/2}(\log n)^{3/2}.$$

Concerning the parametric rate of convergence, it is stressed that this rate applies to the full, non-parametric problem and can not be compared with semi-parametric rates for estimation of the parameter θ in the presence of the nuisance parameter F. With regard to the logarithmic corrections to the powers of n in the expressions for the rate of convergence in Lipschitz- and smoothness-classes, we note that they originate from (the proof of) lemma 10.2.4: the logarithm is introduced by the transition from δ to ε, which compensates for the logarithmic correction in the extent of the Kullback-Leibler neighbourhoods $B(K\delta\log(1/\delta);P_0)$. When considering near-parametric rates (as in lemmas 10.2.1 and 10.2.3), logarithmic corrections of this kind do not influence the calculation, but they do play a role in non-parametric regression. It is possible that these logarithmic corrections to the rate can be omitted, the proof depending on a version of theorem 10.1.1 along the lines of theorem 2.4 of Ghosal *et al.* (2000) [106], in which the prior-mass condition is replaced by a more complicated, but less demanding bound on a ratio of prior masses. Note that the rate (10.15) approaches that given in (10.12) for large values of β, *i.e.* for regression classes with a high degree of differentiability.

Regarding classes with a high degree of differentiabilty, one might expect that suitably restricted classes of analytic regression functions would allow for convergence at the rate (10.15) in the limit $\beta \to \infty$, *i.e.* $1/\sqrt{n}$. However, in that case (10.9) and (10.13) are dominated by the contribution from the parameter $F \in D$, so the expected result would be the parametric rate of convergence given above, *i.e.* $1/\sqrt{n}$ with logarithmic correction of the order $(\log n)^{3/2}$.

10.3 Model entropy

One of the two primary conditions in theorems on non-parametric Bayesian rates of convergence (see, *e.g.* theorem 10.1.1), is an upper-bound on the covering numbers with respect to a metric on the model, in our case the Hellinger metric. In this section, we relate the Hellinger metric entropy of the model to entropy numbers of the three parametrizing spaces, *i.e.* I, $\mathscr{F}$ and D. Due to technical reasons (see subsection 10.3.3, which contains the proofs of all lemmas in this section), we can and shall express most results in terms of the $L_1(\mu)$-norm rather than the Hellinger metric, demonstrating in the (proof of) theorem 10.3.7 that this does not influence the entropy calculation.

10.3.1 Nets in parametrizing spaces

We start the discussion by considering the $L_1(\mu)$-distance between densities in the model that differ only in one of the three parameters (σ, f, F), the goal being the definition of an ε-net over $\mathscr{P}$ from ε-nets over the spaces I, $\mathscr{F}$ and D separately.

With the following lemma, we indicate the possibility of generalizing the discussion that follows to situations in which less is known about the error distribution, by a bound on the $L_1(\mu)$-difference under variation of the parameter for the error distribution. For the next lemma only, we define $\{\psi_\sigma : \sigma \in \Sigma\}$ to be a family of Lebesgue densities of probability distributions on $\mathbb{R}^2$, parametrized by σ in some (parametric or non-parametric) set Σ. The densities $p_{\sigma,f,F}$ are still given by a convolution *c.f.* (10.3) (because we maintain the assumption of independence of Z and (e, f)).

Lemma 10.3.1. *For every $f \in \mathscr{F}$ and $F \in D$,*

$$\|p_{\sigma,f,F} - p_{\tau,f,F}\|_{1,\mu} \leq \|\psi_\sigma - \psi_\tau\|_{1,\mu},$$

for all $\sigma, \tau \in \Sigma$.

Specializing back to the situation of interest, we find the following lemma.

Lemma 10.3.2. *In the case of equally distributed, independent normal errors (e_1, e_2) with mean zero and equal but unknown variance in the interval $[\underline{\sigma}, \overline{\sigma}]$:*

$$\|\psi_\sigma - \psi_\tau\|_{1,\mu} \leq 4\overline{\sigma}\underline{\sigma}^{-2}|\sigma - \tau|.$$

Similar inequalities can be derived for other parametric families of kernels, for instance the Laplace kernel. In the case of a non-parametric family of error distributions, it may be necessary to derive a (sharper) bound, based on the Hellinger distance between $p_{\sigma,f,F}$ and $p_{\tau,f,F}$. This generalized approach is not pursued here and the rest of this chapter relies on the assumption that the errors (e_1, e_2) are as in the above lemma.

Next we consider the dependence of densities in the model on the regression function f.

Lemma 10.3.3. *There exists a constant $K > 0$ such that for all $\sigma \in I$ and all $F \in D[-A, A]$:*

$$\|p_{\sigma,f,F} - p_{\sigma,g,F}\|_{1,\mu} \leq K\|f - g\|_{1,F}, \tag{10.16}$$

for all $f, g \in \mathscr{F}$.

The bound depends on the distribution F for the underlying random variable Z and proves the claim we made earlier, concerning identifiability of the regression function only up to null-sets of the distribution F. To derive a bound that is independent of F, we note that for all $F \in D$ and all $f, g \in C[-A, A]$:

$$\|f - g\|_{1,F} \leq \sup\big\{|f - g|(z) : z \in [-A, A]\big\} = \|f - g\|, \tag{10.17}$$

the right side being finite as a result of continuity of f and g and compactness of the interval $[-A, A]$. Note that we cannot simply equate the uniform norm $\|\,.\,\|$ in

(10.17) to the L_∞-norm because the Lebesgue measure on $[-A,A]$ does not dominate all $F \in D$.

The bound $H^2(P,Q) \leq \|p-q\|_{1,\mu}$ suggests that metric entropy numbers for the Hellinger distance can safely be upper-bounded by those for the $L_1(\mu)$-norm. In cases where the class of regression functions is non-parametric and in fact large enough to dominate the metric entropy of the model, this line of reasoning is insufficient for optimal rates of convergence in the Hellinger distance. The reason is the fact that it is the *squared* Hellinger distance that is dominated by the $L_1(\mu)$-distance and not the Hellinger distance itself. As long as $L_1(\mu)$ entropy numbers are logarithmic, transition from $L_1(\mu)$- to Hellinger coverings leads only to a larger constant. However, if the small-ε behaviour of $L_1(\mu)$ entropy numbers is dominated by terms of the form (10.30)), the replacement of ε by ε^2 influences the calculation. Therefore, we also provide the following lemma.

Lemma 10.3.4. *For all $\sigma \in I$, $f,g \in \mathscr{F}$ and $F \in D$:*

$$H(P_{\sigma,f,F}, P_{\sigma,g,F}) \leq \frac{1}{2\sigma}\Big(\int_{[-A,A]} \big(f(z)-g(z)\big)^2 \, dF(z)\Big)^{1/2}.$$

Although useful, the above bound depends on the particular values of σ, F, which is undesirable in situations below. The lower bound for the interval I and the uniform bound on $|f-g|(z)$ serve to prove a bound on the Hellinger distance proportional to the uniform norm (as opposed to its square-root) of the difference between regression parameters.

Corollary 10.3.5. *There exists a constant $L > 0$ such that for all $\sigma \in I$, $f,g \in \mathscr{F}$ and $F \in D$:*

$$H(P_{\sigma,f,F}, P_{\sigma,g,F}) \leq L\|f-g\|. \tag{10.18}$$

The above two lemmas and the fact that approximation in the uniform norm of subclasses of bounded continuous functions on closed intervals is well-understood, strongly suggests that the class of regression functions is to be endowed with the uniform norm to find nets. We do this in subsection 10.5.1 for the regression classes mentioned earlier.

To bound the contribution of the parameter F to the covering numbers of the model, we approximate F by a discrete distribution F' with a number of support points that is bounded by the approximation error in $L_1(\mu)$. Note that the number of support points needed depends on a power of $\log(1/\varepsilon)$, so that a sharper bound in terms of the Hellinger distance is not necessary (see above).

Lemma 10.3.6. *There exist constants $C, C' > 0$ such that for all $(\sigma, f) \in I \times \mathscr{F}$ and $F \in D$, there is a discrete F' on $[-A,A]$ with less than $C(\log(1/\varepsilon))^2$ support points such that*

$$\|p_{\sigma,f,F} - p_{\sigma,f,F'}\|_{1,\mu} \leq C'\varepsilon.$$

We stress that the particular choice F' depends on the regression function f. The above lemma implies that the set D_ε of all discrete $F \in D$ with less than $C(\log(1/\varepsilon))^2$ support points parametrizes an ε-net over $\mathscr{P}$. For any fixed pair

$(\sigma, f) \in I \times \mathscr{F}$, the ε-net parametrized by D_ε is a 2ε-net over the submodel $\mathscr{P}_{\sigma,f} = \{p_{\sigma,f,F} \in \mathscr{P} : F \in D\}$ so that

$$N\big(\varepsilon, \mathscr{P}_{\sigma,f}, \|\,.\,\|_{1,\mu}\big) \le N\big(2\varepsilon, \{p_{\sigma,f,F} \in \mathscr{P} : F \in D_\varepsilon\}, \|\,.\,\|_{1,\mu}\big).$$

The direct nature of the above approximation (as opposed to the procedure for the parameters σ and f, where we first bound by a norm on the parametrizing variable and then calculate the entropy in the parametrizing space) circumvents the notoriously difficult dependence of mixture densities on their mixing distribution, responsible for the (logarithmically) slow rate of convergence in deconvolution problems. Indeed, problems of this nature plague the method of Fan and Truong (1993) [93], which is based on a kernel-estimate for F and leads to a Nadaraya-Watson-type of estimator for the regression function. Here we are only interested in covering the model $\mathscr{P}$, which allows us to by-pass the deconvolution problem by means of the above lemma.

10.3.2 Metric entropy of the errors-in-variables model

This subsection is devoted entirely to the following theorem, which uses the lemmas of the previous subsection to calculate the Hellinger entropy of the errors-in-variables model $\mathscr{P}$.

Theorem 10.3.7. *Suppose that the regression family $\mathscr{F}$ is one of those specified in the beginning of section 10.2). Then there exist constants $L, L' > 0$ such that the Hellinger covering numbers of the model $\mathscr{P}$ satisfy:*

$$\log N(\varepsilon, \mathscr{P}, H) \le L'\Big(\log\frac{1}{\varepsilon}\Big)^3 + \log N(L\varepsilon, \mathscr{F}, \|\,.\,\|), \tag{10.19}$$

for small enough ε.

Proof. If the class of regression functions $\mathscr{F}$ is a Lipschitz-class with exponent in $(0,1)$, we set α equal to that exponent. In other cases we set $\alpha = 1$.

Let $\varepsilon > 0$ be given, fix some $\sigma \in I$, $f \in \mathscr{F}$. According to lemma (10.3.6) the collection $\mathscr{P}^\varepsilon_{\sigma,f}$ of all $p_{\sigma,f,F'}$ where F' is a discrete distribution in D with at most $N_\varepsilon = \alpha^2 C\big(\log(1/\varepsilon)\big)^2$ support points, forms an ε^α-net over $\mathscr{P}_{\sigma,f}$ with respect to the $L_1(\mu)$-norm. Therefore any ε^α-net $\mathscr{Q}^\varepsilon_{\sigma,f}$ over $\mathscr{P}^\varepsilon_{\sigma,f}$ is a $2\varepsilon^\alpha$-net over $\mathscr{P}_{\sigma,f}$. Let $\mathscr{S}_\varepsilon$ be a minimal ε^α-net for the simplex with ℓ_1-norm in $\mathbb{R}^{N_\varepsilon}$. As is shown by lemma A.4 in Ghosal and Van der Vaart (2001) [107], the order of $\mathscr{S}_\varepsilon$ does not exceed $(5/\varepsilon^\alpha)^{N_\varepsilon}$. Next we define the grid $G_\varepsilon = \{0, \pm\varepsilon, \pm 2\varepsilon, \ldots\} \subset [-A, A]$ and $\mathscr{Q}^\varepsilon_{\sigma,f}$ as the collection of all distributions on $[-A,A]$ obtained by distributing the weights in a vector from $\mathscr{S}_\varepsilon$ over the points in G_ε. We project an arbitrary $p_{\sigma,f,F'}$ in $\mathscr{P}^\varepsilon_{\sigma,f}$ onto $\mathscr{Q}^\varepsilon_{\sigma,f}$ in two steps: given that $F' = \sum_{i=1}^{N_\varepsilon} \lambda_i \delta_{z_i}$, for some set of N_ε points $z_i \in [-A,A]$ and non-negative weights such that $\sum_i \lambda_i = 1$, we first project the vector

λ onto a vector in $\mathscr{S}_\varepsilon$ and second, shift the resulting masses to the closest point in G_ε. One easily sees that the first step leads to a new distribution F'' such that:

$$\|p_{\sigma,f,F'} - p_{\sigma,f,F''}\|_{1,\mu} \leq \varepsilon.$$

As for the second step, in which $F'' = \sum_{i=1}^{N_\varepsilon} \lambda_i' \delta_{z_i}$ is 'shifted' to a new distribution $F''' = \sum_{i=1}^{N_\varepsilon} \lambda_i' \delta_{z_i'}$ such that $|z_i - z_i'| \leq \varepsilon$, we note that:

$$\begin{aligned}
\big|p_{\sigma,f,F''} - p_{\sigma,f,F'''}\big|(x,y) &\leq \sum_{i=1}^{N_\varepsilon} \lambda_i' \big|\varphi_\sigma(x-z_i)\varphi_\sigma(y-f(z_i)) - \varphi_\sigma(x-z_i')\varphi_\sigma(y-f(z_i'))\big| \\
&\leq \sum_{i=1}^{N_\varepsilon} \lambda_i' \Big(\big|\varphi_\sigma(x-z_i) - \varphi_\sigma(x-z_i')\big|\varphi_\sigma(y-f(z_i)) \\
&\qquad + \big|\varphi_\sigma(y-f(z_i)) - \varphi_\sigma(y-f(z_i'))\big|\varphi_\sigma(x-z_i')\Big),
\end{aligned}$$

which implies that the $L_1(\mu)$-difference satisfies:

$$\begin{aligned}
&\|p_{\sigma,f,F''} - p_{\sigma,f,F'''}\|_{1,\mu} \\
&\leq \sum_{i=1}^{N_\varepsilon} \lambda_i' \Big(\int \big|\varphi_\sigma(x-z_i) - \varphi_\sigma(x-z_i')\big|\,dx + \int \big|\varphi_\sigma(y-f(z_i)) - \varphi_\sigma(y-f(z_i'))\big|\,dy \Big).
\end{aligned}$$

By assumption, the family of regression functions satisfies (10.7), which is used to establish that there exists a constant $K > 0$ such that

$$\|p_{\sigma,f,F''} - p_{\sigma,f,F'''}\|_{1,\mu}, \leq K\varepsilon^\alpha,$$

(for small enough ε), along the same lines as the proof of lemma 10.3.3. Summarizing, we assert that for some constant $K_3 > 0$, $\mathscr{Q}^\varepsilon_{\sigma,f}$ is a $K_3^2\varepsilon^\alpha$-net over $\mathscr{P}_{\sigma,f}$. There exist an ε^α-net I_ε over I (with norm equal to absolute differences) and an $\varepsilon^{\alpha/2}$-net $\mathscr{F}_\varepsilon$ over $\mathscr{F}$ in the uniform norm. (The order of $\mathscr{F}_\varepsilon$ is bounded in lemmas 10.5.1 and 10.5.3.) By virtue of the triangle inequality and with the help of lemma 10.3.1 and corollary 10.3.5, we find that constants $K_1, K_2 > 0$ exist such that:

$$\begin{aligned}
H(P_{\sigma,f,F}, P_{\tau,g,F'}) &\leq H(P_{\sigma,f,F}, P_{\tau,f,F}) + H(P_{\tau,f,F}, P_{\tau,g,F}) + H(P_{\tau,g,F}, P_{\tau,g,F'}) \\
&\leq \|p_{\sigma,f,F} - p_{\tau,f,F})\|_{1,\mu}^{1/2} + K\|f-g\| + \|p_{\tau,g,F} - p_{\tau,g,F'}\|_{1,\mu}^{1/2} \\
&\leq K_1|\sigma-\tau|^{1/2} + K_2\|f-g\| + \|p_{\tau,g,F} - p_{\tau,g,F'}\|_{1,\mu}^{1/2},
\end{aligned}$$

for all $\sigma \in I$, $\tau \in I_\varepsilon$, $f \in \mathscr{F}$, $g \in \mathscr{F}_\varepsilon$ and $F, F' \in D$. For every fixed pair $(\tau, g) \in I_\varepsilon \times \mathscr{F}_\varepsilon$, we define the $K_3^2\varepsilon^\alpha$-net $\mathscr{Q}^\varepsilon_{\tau,g}$ like above and choose F' in the above display so that $p_{\tau,g,F'}$ lies in $\mathscr{Q}^\varepsilon_{\tau,g}$ and approximates $p_{\tau,g,F}$ to within $L_1(\mu)$-distance proportional to ε^α. This shows that the set:

$$\mathscr{Q}_\varepsilon = \bigcup \{ \mathscr{Q}^\varepsilon_{\tau,g} : \tau \in I_\varepsilon, g \in \mathscr{F}_\varepsilon \},$$

forms a $K\varepsilon^{\alpha/2}$-net over $\mathscr{P}$ with respect to the Hellinger distance, where $K = K_1 + K_2 + K_3$. The order of this net can be calculated and forms an upper bound for the Hellinger covering number of the model.

$$\log N(K\varepsilon^{\alpha/2}, \mathscr{P}, H) \leq \log N(\varepsilon^\alpha, I, |.|) + \log N(\varepsilon^{\alpha/2}, \mathscr{F}, \|.\|) + \log N(\mathscr{Q}^\varepsilon_{\tau,g}),$$

where $N(\mathscr{Q}^\varepsilon_{\tau,g})$ denotes the uniform bound on the number of points in the nets $\mathscr{Q}^\varepsilon_{\tau,g}$, given by:

$$\log N(\mathscr{Q}^\varepsilon_{\tau,g}) = L'' \left(\log \frac{1}{\varepsilon} \right)^3,$$

for some constant $L'' > 0$ as is easily checked from the above. Moreover, the covering numbers for the finite-dimensional, bounded space I satisfy, for some constant $L''' > 0$:

$$\log N(\varepsilon^\alpha, I, |.|) \leq L''' \log \frac{1}{\varepsilon}.$$

(Note that in the two displays above, any exponent for ε (*e.g.* $\alpha/2$) is absorbed in the constants L' and L''). Note that for small enough ε, the contribution from the mixing parameter F dominates that of the parameter σ. Eventually, we find the bound:

$$\log N(\varepsilon, \mathscr{P}, H) \leq L' \left(\log \frac{1}{\varepsilon} \right)^3 + \log N(L\varepsilon, \mathscr{F}, \|.\|),$$

for small enough $\varepsilon > 0$ and some $L, L' > 0$.

10.3.3 Proofs of several lemmas

Proof. Proof of lemma 10.3.1 Fix $f \in \mathscr{F}$ and $F \in D$, let $\sigma, \tau \in \Sigma$ be given. Consider the $L_1(\mu)$ difference:

$$\|p_{\sigma,f,F} - p_{\tau,f,F}\|_{1,\mu} \leq \int_{\mathbb{R}} \int_{\mathbb{R}^2} \left| \psi_\sigma\big(x-z, y-f(z)\big) - \psi_\tau\big(x-z, y-f(z)\big) \right| d\mu(x,y)\, dF(z),$$

by Fubini's theorem. Translation invariance of the Lebesgue measure and the domain of integration $\mathbb{R}^2$ make it possible to translate over $(z, f(z))$ to render the inner integral independent of z and integrate with respect to F with the following result:

$$\|p_{\sigma,f,F} - p_{\tau,f,F}\|_{1,\mu} \leq \int_{\mathbb{R}^2} \left| \psi_\sigma(x,y) - \psi_\tau(x,y) \right| d\mu(x,y),$$

thus leading to an upper bound that is independent of both f and F.

Proof. Proof of lemma 10.3.2 The $L_1(\mu)$-difference of the densities ψ_σ and ψ_τ equals the total-variational difference between the corresponding distributions Ψ_σ and Ψ_τ and can be expressed in terms of the event $\{\psi_\sigma > \psi_\tau\}$ as follows:

$$\|\psi_\sigma - \psi_\tau\|_{1,\mu} = 2\left(\Psi_\sigma(\psi_\sigma > \psi_\tau) - \Psi_\tau(\psi_\sigma > \psi_\tau)\right).$$

In the case of normally and equally distributed, independent errors (e_1, e_2) the kernel is $\psi_\sigma(x,y) = \varphi_\sigma(x)\varphi_\sigma(y)$, with $\sigma \in I$. Assuming that $\sigma < \tau$, the event in question is a ball in $\mathbb{R}^2$ of radius r_0 centred at the origin (and its complement if $\sigma > \tau$), where $r_0^2 = (2\sigma^2\tau^2/(\tau^2 - \sigma^2))\log(\tau^2/\sigma^2)$. Integrating the normal kernels over this ball, we find:

$$\|\psi_\sigma - \psi_\tau\|_{1,\mu} = 2\left|e^{-\frac{1}{2}(r_0/\sigma)^2} - e^{-\frac{1}{2}(r_0/\tau)^2}\right| = 2e^{-\frac{1}{2}(r_0/\sigma)^2}\left|1 - \frac{\sigma^2}{\tau^2}\right| \leq \frac{4\overline{\sigma}}{\underline{\sigma}^2}|\sigma - \tau|,$$

where we have used the upper and lower bounds for the interval I.

Proof. Proof of lemma 10.3.3 Let $\sigma \in I$, $F \in D[-A,A]$ and $f, g \in \mathscr{F}$ be given. Since the x-dependence of the densities $p_{\sigma,f,F}$ and $p_{\sigma,g,F}$ is identical and can be integrated out, the $L_1(\mu)$-difference can be upper-bounded as follows:

$$\|p_{\sigma,f,F} - p_{\sigma,g,F}\|_{1,\mu} \leq \int_{\mathbb{R}}\int_{\mathbb{R}} \left|\varphi_\sigma\big(y - f(z)\big) - \varphi_\sigma\big(y - g(z)\big)\right| dy\, dF(z).$$

Fix a $y \in \mathbb{R}$ and $z \in [-A,A]$. We note:

$$\left|\varphi_\sigma(y - f(z)) - \varphi_\sigma(y - g(z))\right| \leq \left|\int_{y-f(z)}^{y-g(z)} \varphi'_\sigma(u)\, du\right| \leq \sup_{u \in J}|\varphi'_\sigma(u)|\,\big|f(z) - g(z)\big|,$$

where $J = [y - f(z) \vee g(z), y - f(z) \wedge g(z)]$. The uniform bound on the functions in the regression class $\mathscr{F}$ guarantees that $J \subset J' = [y - B, y + B]$. If $y \geq 2B$, then $y - B \geq \frac{1}{2}y \geq B > 0$, so if, in addition, $\frac{1}{2}y \geq \overline{\sigma}$, we see that for all $u \in J'$, $u \geq \frac{1}{2}y \geq \sigma$, thus restricting u to the region in which the derivative of the normal density decreases monotonously:

$$\left|\varphi'_\sigma(u)\right| \leq \left|\varphi'_\sigma(\tfrac{1}{2}y)\right|.$$

Symmetry of the normal density allows us to draw the same conclusion if y lies below $-2B$ and $-2\overline{\sigma}$. Using the explicit form of the normal density and the constant $T = 2(B \vee \overline{\sigma})$, we derive the following upper bound on the supremum:

$$\sup\left\{\left|\varphi'_\sigma(u)\right| : u \in J\right\} \leq Ks(y),$$

where the function s is given by:

$$s(y) = \begin{cases} |y|\,\varphi_{2\overline{\sigma}}(y), & \text{if } |y| \geq T, \\ \|\varphi'_{\underline{\sigma}}\|_\infty, & \text{if } |y| < T. \end{cases}$$

Note that s does not depend on the values of the parameters. Therefore:

$$\|p_{\sigma,f,F} - p_{\sigma,g,F}\|_{1,\mu} \leq \int_{\mathbb{R}}\int_{\mathbb{R}} Ks(y)\big|f(z) - g(z)\big|\, dy\, dF(z).$$

Since the integral over $s(y)$ is finite, the asserted bound follows.

Proof. (Proof of lemma 10.3.4) Consider a binary experiment $E_1 = (\mathbb{R}^3, \mathscr{B}^{(3)}, \{P,Q\})$, giving two possible distributions P,Q for the triplet (X,Y,Z) that describes the errors-in-variables model (*c.f.* (10.2)). The map T that projects by $T(X,Y,Z) = (X,Y)$ leads to another binary experiment $E_2 = (\mathbb{R}^2, \mathscr{B}^{(2)}, \{P^T, Q^T\})$ which is less informative than E_1. (The phrase "less informative" is defined in the sense of Le Cam, *i.e.* for every test function ϕ_2 in E_2, there exists a test function ϕ_1 in E_1 such that $P\phi_1 \leq P^T\phi_2$ and $Q\phi_1 \geq Q^T\phi_2$ (see, for instance, Strasser (1985) [236], definition 15.1).) This property follows from the fact that $\sigma(X,Y) \subset \mathscr{B}^{(2)}$ is such that $T^{-1}\big(\sigma(X,Y)\big) \subset \sigma(X,Y,Z) \subset \mathscr{B}^{(3)}$, which makes it possible to identify every test function in E_2 with a test function in E_1, while there may exist test functions on $\mathbb{R}^3$ that are not measurable with respect to $T^{-1}\big(\sigma(X,Y)\big)$. Corollary 17.3 in Strasser (1985) [236] asserts that the Hellinger distance decreases when we make the transition from a binary experiment to a less informative binary experiment, so we see that:

$$H(P^T, Q^T) \leq H(P,Q). \tag{10.20}$$

In the case at hand, we choose $P^T = P_{\sigma,f,F}$ and $Q^T = P_{\sigma,g,F}$. From the definition of the errors-in-variables model (10.2), we obtain the conditional laws:

$$\begin{aligned}
\mathscr{L}_P\big(X,Y \,\big|\, Z\big) &= N(Z,\sigma^2) \times N(f(Z),\sigma^2),\\
\mathscr{L}_Q\big(X,Y \,\big|\, Z\big) &= N(Z,\sigma^2) \times N(g(Z),\sigma^2),
\end{aligned}$$

and, of course, $\mathscr{L}_P(Z) = \mathscr{L}_Q(Z) = F$. It follows that:

$$\begin{aligned}
H^2(P,Q) &= \int_{\mathbb{R}^3} \big(dP^{1/2} - dQ^{1/2}\big)^2\\
&= \int_{\mathbb{R}^3} \varphi_\sigma(x-z)\Big(\varphi_\sigma\big(y-f(z)\big)^{1/2} - \varphi_\sigma\big(y-g(z)\big)^{1/2}\Big)^2 dF(z)\,dx\,dy\\
&= \int_{[-A,A]} H^2\big(N(f(z),\sigma^2), (N(g(z),\sigma^2)\big)\, dF(z),
\end{aligned}$$

by Fubini's theorem. A straightforward calculation shows that:

$$H^2\big(N(f(z),\sigma^2), (N(g(z),\sigma^2)\big) = 2\Big(1 - e^{-\frac{1}{2}\left(f(z)-g(z)\right)^2/(2\sigma)^2}\Big) \leq \frac{1}{4\sigma^2}\big(f(z)-g(z)\big)^2,$$

where we use that $1 - e^{-x} \leq x$ for all $x \geq 0$. Upon combination of the above two displays and (10.20), we obtain:

$$H^2(P_{\sigma,f,F}, P_{\sigma,g,F}) \leq \frac{1}{4\sigma^2}\int_{[-A,A]} \big(f(z)-g(z)\big)^2 dF(z),$$

which proves the assertion.

Proof. Proof of lemma 10.3.6 Let $\varepsilon > 0$, $\sigma \in I$, $f \in \mathscr{F}$ be given, fix $M \geq 2A \vee 2B$ and $k \geq 1$. A Taylor-expansion up to order $k-1$ of the exponential in the normal

density demonstrates that:

$$\Big|\varphi_\sigma(x-z)-\frac{1}{\sigma\sqrt{2\pi}}\sum_{j=0}^{k-1}\frac{1}{j!}(-\tfrac{1}{2})^j\Big(\frac{x-z}{\sigma}\Big)^{2j}\Big|\le\frac{1}{\sigma\sqrt{2\pi}}\frac{1}{k!}(\tfrac{1}{2})^k\Big(\frac{x-z}{\sigma}\Big)^{2k}$$
$$\le\frac{1}{\sigma\sqrt{2\pi}}\Big(\frac{e}{2k}\Big)^k\Big(\frac{x-z}{\sigma}\Big)^{2k},$$

where we have used that $k!\ge k^k e^{-k}$. Similarly, we obtain:

$$\Big|\varphi_\sigma\big(y-f(z)\big)-\frac{1}{\sigma\sqrt{2\pi}}\sum_{j=0}^{k-1}\frac{1}{j!}(-\tfrac{1}{2})^j\Big(\frac{y-f(z)}{\sigma}\Big)^{2j}\Big|\le\frac{1}{\sigma\sqrt{2\pi}}\Big(\frac{e}{2k}\Big)^k\Big(\frac{y-f(z)}{\sigma}\Big)^{2k}.$$

Considering $|x|,|y|\le M$ and using that $\sigma\ge\overline{\sigma}>0$, we see that there exists a constant $C_1>0$ (independent of σ and f) such that both residuals of the last two displays are bounded above by $(C_1M^2/k)^k$. So for all x,y like above,

$$\begin{aligned}&\big|p_{\sigma,f,F}-p_{\sigma,f,F'}\big|(x,y)\\&\quad\le\frac{1}{2\pi\sigma^2}\Big|\int\sum_{i,j=0}^{k-1}\frac{1}{i!j!}(-\tfrac{1}{2})^{i+j}\Big(\frac{x-z}{\sigma}\Big)^{2i}\Big(\frac{y-f(z)}{\sigma}\Big)^{2j}d(F-F')(z)\Big|\\&\qquad+4\Big(\frac{C_1M^2}{k}\Big)^k+\Big(\frac{C_1M^2}{k}\Big)^{2k}.\end{aligned}\tag{10.21}$$

Lemma A.1 in Ghosal and Van der Vaart (2001) [107] asserts that there exists a discrete distribution F' on $[-A,A]$ with at most (k^2+1) support points such that for all functions $\psi_{f,ij}(z)=z^{2i}f^{2j}(z)$ the F- and F'-expectations coincide, *i.e.*:

$$\int_{[-A,A]}\psi_{f,ij}\,dF=\int_{[-A,A]}\psi_{f,ij}\,dF'.$$

Thus choosing F', the first term in (10.21) vanishes and we see that (for large enough k):

$$\sup_{|x|\vee|y|\le M}\big|p_{\sigma,f,F}-p_{\sigma,f,F'}\big|(x,y)\le 5\Big(\frac{C_1M^2}{k}\Big)^k.\tag{10.22}$$

For points (x,y) outside $[-M,M]\times[-M,M]$, we note that there exists a constant $C_2>0$ such that for all $|x|\ge 2A$, $|y|\ge 2B$:

$$\varphi_\sigma(x-z)\le\varphi_\sigma\big(\tfrac{x}{2}\big)\le C_2\varphi_{\overline{\sigma}}\big(\tfrac{x}{2}\big),$$
$$\varphi_\sigma\big(y-f(z)\big)\le\varphi_\sigma\big(\tfrac{y}{2}\big)\le C_2\varphi_{\overline{\sigma}}\big(\tfrac{y}{2}\big),$$

($C_2=\|\varphi_{\underline{\sigma}}\|_\infty/\|\varphi_{\overline{\sigma}}\|_\infty$ will do). Since $M\ge 2A\vee 2B$, there exists a constants $C_3,C_4>0$ such that:

$$\begin{aligned}\int_{|x|\vee|y|>M} p_{\sigma,f,F}(x,y)\,d\mu(x,y) &\le C_2 \int_{|x|>M} \varphi_{\overline{\sigma}}\left(\tfrac{x}{2}\right) dx \int\int \varphi_\sigma\big(y-f(z)\big)\,dF(z)\,dy \\ &\quad + C_2 \int_{|y|>M} \varphi_{\overline{\sigma}}\left(\tfrac{y}{2}\right) dy \int\int \varphi_\sigma(x-z)\,dF(z)\,dx \\ &= 4C_2 \int_{x>M} \varphi_{\overline{\sigma}}\left(\tfrac{x}{2}\right) dx \le 4C_2 \int_{x>M} \frac{x}{M}\varphi_{\overline{\sigma}}\left(\tfrac{x}{2}\right) dx \\ &\le C_3 e^{-C_4 M^2},\end{aligned} \tag{10.23}$$

where we have used Fubini's theorem and translation invariance of Lebesgue measure in the second step and the fact that $\varphi'_\sigma(x) = -(x/\sigma^2)\varphi_\sigma(x)$ in the last. Now, let $\varepsilon > 0$ be given. We decompose the domain of integration for the $L_1(\mu)$-difference between $p_{\sigma,f,F}$ and $p_{\sigma,f,F'}$ into the region where $|x| \vee |y| \le M$ and its complement. Using the uniform bound (10.22) on the region bounded by M and (10.23) for the tails, we find that there is a constant D_1 such that:

$$\|p_{\sigma,f,F} - p_{\sigma,f,F'}\|_{1,\mu} \le D_1\left(M^2\left(\frac{C_1 M^2}{k}\right)^k + e^{-C_4 M^2}\right). \tag{10.24}$$

In order to bound the *r.h.s.* by ε we fix M in terms of ε:

$$M = \sqrt{\tfrac{1}{C_4}\log\tfrac{1}{\varepsilon}},$$

and note that the lower bound $M \ge 2A \vee 2B$ is satisfied for small enough ε. Upon substitution, the first term in (10.24) leads to $(D_1/C_4)D_2^k e^{(k+1)\log\log\frac{1}{\varepsilon}} e^{-k\log k}$ (where $D_2 = C_1/C_4$), so that the choice:

$$k \ge D_3 \log\tfrac{1}{\varepsilon},$$

(for some large $D_3 > D_2$) suffices to upper bound the $L_1(\mu)$-difference appropriately. The smallest integer k above the indicated bound serves as the minimal number of support points needed.

Note that the f-dependence of the functions $\psi_{f,ij}$ carries over to the choice for F', which is therefore f-dependent as well.

10.4 Model prior

Assume that the model is well-specified and denote by $P_0 \in \mathscr{P}$ (corresponding to some, not necessarily unique, $\sigma_0 \in I$, $f_0 \in \mathscr{F}$ and $F_0 \in D$) the true distribution underlying the *i.i.d.* sample. We define a prior Π on $\mathscr{P}$ by defining priors on the parameter spaces I, $\mathscr{F}$ and D and taking Π equal to the probability measure induced by the map $(\sigma,f,F) \mapsto P_{\sigma,f,F}$ from $I \times \mathscr{F} \times D$ with product-measure to $\mathscr{P}$. The prior on I is denoted Π_I and is assumed to have a density π_I, continuous and strictly

positive at σ_0. The prior $\Pi_{\mathscr{F}}$ on $\mathscr{F}$ is specified differently for each of the classes defined in the beginning of section 10.2, but all have as their domain the Borel σ-algebra generated by the norm topology on $C[-A,A]$. The definition of these priors is postponed to subsection 10.5.2. The prior Π_D on D is based on a Dirichlet process with base measure α which has a continuous and strictly positive density on all of $[-A,A]$. The domain of Π_D is the Borel σ-algebra generated by the topology of weak convergence.

The fact that these priors are defined on the product of the parameter spaces rather than the errors-in-variables model $\mathscr{P}$ itself, necessitates a lemma asserting appropriate measurability. So before we discuss the properties of priors, we show that the map $\hat{p}$ that takes parameters (σ, f, F) into densities $p_{\sigma,f,F}$ (*c.f.* (10.3)) is measurable.

Lemma 10.4.1. *Endow I and $\mathscr{F}$ with their norm topology and D with the topology of weak convergence. Then the map $\hat{p} : I \times \mathscr{F} \times D \to L_1(\mu)$ is continuous in the product topology.*

Proof. The space D with the topology of weak convergence is metric, so the product topology on $I \times \mathscr{F} \times D$ is a metric topology as well. Let (σ_n, f_n, F_n) be a sequence, converging to some point (σ, f, F) in $I \times \mathscr{F} \times D$ as $n \to \infty$. As a result of the triangle inequality and lemmas 10.3.1–10.3.3, the $L_1(\mu)$-distance satisfies:

$$\left\| p_{\sigma_n,f_n,F_n} - p_{\sigma,f,F} \right\|_{1,\mu} \leq K_1 |\sigma_n - \sigma| + K_2 \|f_n - f\| + \left\| p_{\sigma,f,F_n} - p_{\sigma,f,F} \right\|_{1,\mu}, \quad (10.25)$$

for some constants $K_1, K_2 > 0$. Since F_n converges to F weakly, the continuity of the regression function f, combined with the continuity and boundedness of the Gaussian kernel and the portmanteau lemma guarantee that

$$\int_{[-A,A]} \varphi_\sigma(x-z)\,\varphi_\sigma\big(y-f(z)\big)\,dF_n(z) \to \int_{[-A,A]} \varphi_\sigma(x-z)\,\varphi_\sigma\big(y-f(z)\big)\,dF(z),$$

as $n \to \infty$ for all $(x,y) \in \mathbb{R}^2$. Using the (μ-integrable) upper-envelope for the model $\mathscr{P}$ and dominated convergence, we see that

$$\left\| p_{\sigma,f,F_n} - p_{\sigma,f,F} \right\|_{1,\mu} \to 0,$$

and hence the *r.h.s.* of (10.25) goes to zero. We conclude that $\hat{p}$ is continuous in the product topology.

Note that the $L_1(\mu)$- and Hellinger topologies on the model $\mathscr{P}$ are equivalent, so that the above lemma implies continuity of $\hat{p}$ in the Hellinger topology. Hence $\hat{p}^{-1}$ is a well-defined map between the Borel σ-algebras of the model with the Hellinger topology and the product $I \times \mathscr{F} \times D$.

The following lemma establishes that the prior-mass condition (10.4) can be analysed for the regression class and the parameter space for (σ, F) separately. Lower bounds for the prior mass in appropriate neighbourhoods of the point (σ_0, F_0) are incorporated immediately.

Theorem 10.4.2. *Suppose that the regression family $\mathscr{F}$ is one of those specified in the beginning of section 10.2. Assume that the prior Π on $\mathscr{P}$ is of the product form indicated above. Then there exist constants $K, c, C > 0$ such that:*

$$\Pi\Big(B\big(K\delta\log(1/\delta);P_0\big)\Big) \geq C\exp\Big(-c\big(\log(1/\delta)\big)^3\Big)\,\Pi_{\mathscr{F}}\big(f\in\mathscr{F} : \|f-f_0\|\leq\delta\big),$$

for small enough δ.

Proof. If the class of regression functions $\mathscr{F}$ is a Lipschitz-class with exponent in $(0,1)$, we set α equal to that exponent. In other cases we set $\alpha = 1$.

Let $\varepsilon > 0$ be given. By lemma 10.3.6 there exists a discrete F_0' in D with at most $N_\varepsilon = C(\log(1/\varepsilon))^2$ support points $z_1,\ldots,z_{N_\varepsilon}$ of the form $F_0' = \sum_{i=1}^{N_\varepsilon} p_i\delta_{z_i}$ with $\sum_{i=1}^{N_\varepsilon} p_i = 1$, such that:

$$\|p_{\sigma_0,f_0,F_0'} - p_{\sigma_0,f_0,F_0}\|_{1,\mu} \leq C'\varepsilon^\alpha,$$

for some constant $C' > 0$. Although the assertion of lemma 10.3.6 is stronger, we include the power of α because we assume (without loss of generality) that the set of support points for F_0' is 2ε-separated. If this is not the case, take a maximal 2ε-separated subset and shift the masses of other support points of F_0' to points in the chosen subset within distance 2ε, to obtain a new discrete distribution F_0''. Arguing as in the proof of theorem 10.3.7, we see that the corresponding change in $L_1(\mu)$-distance between $p_{\sigma_0,f_0,F_0'}$ and $p_{\sigma_0,f_0,F_0''}$ is upper-bounded by a multiple of ε^α, since the family of regression functions satisfies (10.7) by assumption. The distribution function F_0'' so obtained may then replace F_0'. By lemma 10.4.3, there exists a constant $K_3 > 0$ such that for all $F \in D$:

$$\|p_{\sigma_0,f_0,F} - p_{\sigma_0,f_0,F'}\|_{1,\mu} \leq K_3\Big(\varepsilon^\alpha + \sum_{i=1}^{N_\varepsilon}\big|F[z_i-\varepsilon,z_i+\varepsilon] - p_i\big|\Big).$$

Let (σ,f,F) be a point in the parameter space of the model. The Hellinger distance between $p_{\sigma,f,F}$ and p_{σ_0,f_0,F_0} is upper-bounded as follows (for constants $K_1, K_2 > 0$):

$$\begin{aligned}
H(P_{\sigma,f,F},P_{\sigma_0,f_0,F_0}) &\leq H(P_{\sigma,f,F},P_{\sigma_0,f,F}) + H(P_{\sigma_0,f,F},P_{\sigma_0,f_0,F}) + H(P_{\sigma_0,f_0,F},P_{\sigma_0,f_0,F_0}) \\
&\leq \big\|p_{\sigma,f,F} - p_{\sigma_0,f,F}\big\|_{1,\mu}^{1/2} + H(P_{\sigma_0,f,F},P_{\sigma_0,f_0,F}) + \big\|p_{\sigma_0,f_0,F} - p_{\sigma_0,f_0,F_0}\big\|_{1,\mu}^{1/2} \\
&\leq K_1\big|\sigma-\sigma_0\big|^{1/2} + K_2\big\|f-f_0\big\| \\
&\quad + \Big(\big\|p_{\sigma_0,f_0,F} - p_{\sigma_0,f_0,F_0'}\big\|_{1,\mu} + \big\|p_{\sigma_0,f_0,F_0'} - p_{\sigma_0,f_0,F_0}\big\|_{1,\mu}\Big)^{1/2},
\end{aligned} \tag{10.26}$$

where we have used lemmas 10.3.1, 10.3.2 and corollary 10.3.5. Moreover, we see that there exists a constant $K_4 > 0$ such that for small enough $\eta > 0$ and $P \in \mathscr{P}$ such

that $H(P,P_0) \leq \eta$:

$$-P_0 \log \frac{p}{p_0} \vee P_0 \Big(\log \frac{p}{p_0} \Big)^2 \leq K_4^2 \eta^2 \Big(\log \frac{1}{\eta} \Big)^2,$$

as a result of lemma 10.4.4. Combining the last two displays and using definition (6.13), we find that, for some constants $K_5, K_6 > 0$, the following inclusions hold:

$$\begin{aligned}
&\Big\{ (\sigma, f, F) \in I \times \mathscr{F} \times D : \\
&\qquad |\sigma - \sigma_0|^{1/2} \leq \varepsilon^\alpha, \, \|f - f_0\| \leq \varepsilon^{\alpha/2}, \, \sum_{j=1}^{N_\varepsilon} \big| F[z_j - \varepsilon, z_j + \varepsilon] - p_j \big| \leq \varepsilon^\alpha \Big\} \\
&\qquad \subset \Big\{ (\sigma, f, F) \in I \times \mathscr{F} \times D : H(P_{\sigma,f,F}, P_0) \leq K_5 \varepsilon^{\alpha/2} \Big\} \\
&\qquad \Big\{ P \in \mathscr{P} : H(P, P_0) \leq K_5 \varepsilon^{\alpha/2} \Big\} \subset B\big(K_6 \varepsilon^{\alpha/2} \log(1/\varepsilon); P_0 \big),
\end{aligned} \tag{10.27}$$

for small enough ε and with the notation p_0 for the density of P_0 ($p_0 = p_{\sigma_0, f_0, F_0}$). Using the fact that the prior measure of the rectangle set on the *l.h.s.* of the first inclusion above factorizes, we find that:

$$\begin{aligned}
\Pi\Big(B\big(K_6 \varepsilon^{\alpha/2} \log(1/\varepsilon); P_0 \big) \Big) &\geq \Pi_I\big(\sigma \in I : |\sigma - \sigma_0|^{1/2} \leq \varepsilon^\alpha \big) \, \Pi_{\mathscr{F}}\big(f \in \mathscr{F} : \|f - f_0\| \leq \varepsilon^{\alpha/2} \big) \\
&\quad \times \Pi_D\Big(F \in D : \sum_{j=1}^{N_\varepsilon} \big| F[z_j - \varepsilon, z_j + \varepsilon] - p_j \big| \leq \varepsilon^\alpha \Big).
\end{aligned}$$

Note that $\varepsilon^\alpha \geq \varepsilon$ for small enough ε, so that

$$\Pi_D\Big(\sum_{j=1}^{N_\varepsilon} \big| F[z_j - \varepsilon, z_j + \varepsilon] - p_j \big| \leq \varepsilon^\alpha \Big) \geq \Pi_D\Big(\sum_{j=1}^{N_\varepsilon} \big| F[z_j - \varepsilon, z_j + \varepsilon] - p_j \big| \leq \varepsilon \Big).$$

According to lemma 6.1 in Ghosal *et al.* (2000) [106] (also given as lemma A.2 in Ghosal and Van der Vaart (2001) [107]), there are constants $C', c' > 0$ such that

$$\Pi_D\Big(\sum_{j=1}^{N_\varepsilon} \big| F[z_j - \varepsilon, z_j + \varepsilon] - p_j \big| \leq \varepsilon \Big) \geq C' \exp\big(-c' N_\varepsilon \log(1/\varepsilon) \big) \geq C' \exp\big(-c' C \big(\log(1/\varepsilon) \big)^3 \big).$$

Furthermore, continuity and strict positivity of the density of the prior Π_I imply that (see the proof of lemma 10.5.5):

$$\Pi_I(\sigma \in I : |\sigma - \sigma_0| \leq \varepsilon^\alpha) \geq \pi_1 \, \varepsilon^\alpha = \pi_1 \exp\big(-\alpha \log(1/\varepsilon) \big),$$

for some constant $\pi_1 > 0$. Note that the exponent on the *r.h.s.* falls above all multiples of $-(\log(1/\varepsilon))^3$ for small enough ε. Substitution of $\delta = \varepsilon^{\alpha/2}$ leads to the conclusion that there exist constants $K, c, C > 0$ such that:

$$\Pi\Big(B\big(K\delta\log(1/\delta);P_0\big)\Big) \geq C\exp\Big(-c\big(\log(1/\delta)\big)^3\Big)\,\Pi_{\mathscr{F}}\big(f\in\mathscr{F}:\|f-f_0\|\leq\delta\big),$$

for small enough δ.

If the model is not identifiable in the parameter space $I\times\mathscr{F}\times D$, the above conditions are more stringent than necessary. The point (σ_0,f_0,F_0) may not be the only one that is mapped to P_0, so the first inclusion in (10.27) may discount parts of the parameter space that also contribute to the Kullback-Leibler neighbourhoods $B(P_0,\varepsilon)$. However, the methods we use to lower-bound the prior mass rely on uniformity in the sense that neighbourhoods of *every* point in the parameter space receive a certain minimal fraction of the total prior mass. Therefore, identifiability issues do not affect the argument.

10.4.1 Lemmas

In the following lemma, it is assumed that the regression class $\mathscr{F}$ is one of those specified in the beginning of section 10.2. If the class of regression functions is a Lipschitz-class with exponent in $(0,1)$, we set α equal to that exponent. In other cases we set $\alpha=1$.

Lemma 10.4.3. *Let $\varepsilon>0$ be given and let $F'=\sum_{i=1}^N p_i\delta_{z_i}$ be a convex combination of point-masses, where the set $\{z_i:i=1,\ldots,N\}$ is 2ε-separated. Then there exists a constant $K>0$ such that for all $\sigma\in I$, $f\in\mathscr{F}$ and all $F\in D$:*

$$\|p_{\sigma,f,F}-p_{\sigma,f,F'}\|_{1,\mu}\leq K\Big(\varepsilon^\alpha+\sum_{i=1}^N\big|F[z_i-\varepsilon,z_i+\varepsilon]-p_i\big|\Big),$$

for small enough ε.

Proof. Let F be given. We partition the real line by $\mathbb{R}=\cup_i A_i\cup B$, with $B=\big(\cap_i B_i\big)$, where

$$A_i=\big\{z:|z-z_i|\leq\varepsilon\big\},\quad B_i=\big\{z:|z-z_i|>\varepsilon\big\},$$

and decompose the absolute difference between $p_{\sigma,f,F}$ and $p_{\sigma,f,F'}$ accordingly:

$$\begin{aligned}\big|p_{\sigma,f,F}-p_{\sigma,f,F'}\big|(x,y)&=\Big|\int_{\mathbb{R}}\varphi_\sigma(x-z)\varphi_\sigma\big(y-f(z)\big)\,d(F-F')(z)\Big|\\&=\Big|\sum_{i=1}^N\int_{A_i}\varphi_\sigma(x-z)\varphi_\sigma(y-f(z))\,d(F-F')(z)+\int_B\varphi_\sigma(x-z)\varphi_\sigma(y-f(z))\,dF(z)\Big|,\end{aligned}$$

for all $(x,y)\in\mathbb{R}^2$. Integrating this expression over $\mathbb{R}^2$, we find that the $L_1(\mu)$-difference is bounded as follows:

$$\begin{aligned}\big\|p_{\sigma,f,F}-p_{\sigma,f,F'}\big\|_{1,\mu} &\le \sum_{i=1}^{N}\big|F[z_i-\varepsilon,z_i+\varepsilon]-p_i\big|+F\Big(\bigcap_{i=1}^{N}B_i\Big)\\ &\quad+\sum_{i=1}^{N}\int_{A_i}\int_{\mathbb{R}^2}\big|\varphi_\sigma(x-z)\varphi_\sigma(y-f(z))-\varphi_\sigma(x-z_i)\varphi_\sigma(y-f(z_i))\big|\,d\mu(x,y)\,dF(z),\end{aligned}$$

by Fubini's theorem and the triangle inequality. To upper-bound the last term on the *r.h.s.* in the above display, we use that for all $x,y\in\mathbb{R}$ and $z\in[-A,A]$:

$$\begin{aligned}&\big|\varphi_\sigma(x-z)\varphi_\sigma(y-f(z))-\varphi_\sigma(x-z_i)\varphi_\sigma(y-f(z_i))\big|\\ &\quad\le\big|\varphi_\sigma(x-z)-\varphi_\sigma(x-z_i)\big|\varphi_\sigma(y-f(z))+\big|\varphi_\sigma(y-f(z))-\varphi_\sigma(y-f(z_i))\big|\varphi_\sigma(x-z_i),\end{aligned}$$

and argue as in the proof of lemma 10.3.3, to see that the integrand is bounded by a multiple of $|z-z_i|^\alpha$ for small enough ε. Noting that the intervals $[z_i-\varepsilon,z_i+\varepsilon]$ are disjoint due to 2ε-separation of the set $\{z_i : i=1,\dots,N\}$, we see that there exists a constant $L'>0$ such that

$$\big\|p_{\sigma,f,F}-p_{\sigma,f,F'}\big\|_{1,\mu}\le L'\varepsilon^\alpha+\sum_{i=1}^{N}\big|F[z_i-\varepsilon,z_i+\varepsilon]-p_i\big|+F\Big(\bigcap_{i=1}^{N}B_i\Big).$$

Furthermore, by De Morgan's law and the disjointness of the intervals $[z_i-\varepsilon,z_i+\varepsilon]$:

$$\begin{aligned}F\Big(\bigcap_{i=1}^{N}\big\{z:|z-z_i|>\varepsilon\big\}\Big)&=1-F\Big(\bigcup_{i=1}^{N}\big\{z:|z-z_i|\le\varepsilon\big\}\Big)\\ &=\sum_{i=1}^{N}p_i-\sum_{i=1}^{N}F[z_i-\varepsilon,z_i+\varepsilon]\le\sum_{i=1}^{N}\big|F[z_i-\varepsilon,z_i+\varepsilon]-p_i\big|,\end{aligned}$$

which proves the assertion.

Lemma 10.4.4. *Let $P,Q\in\mathscr{P}$ be given. There exists a constant $K>0$ such that for small enough $H(P,Q)$:*

$$\begin{aligned}\int p\log\frac{p}{q}\,d\mu&\le K^2H^2(P,Q)\Big(\log\frac{1}{H(P,Q)}\Big)^2,\\ \int p\Big(\log\frac{p}{q}\Big)^2d\mu&\le K^2H^2(P,Q)\Big(\log\frac{1}{H(P,Q)}\Big)^2.\end{aligned}\tag{10.28}$$

The constant K does not depend on P,Q.

Proof. Fix $\delta\in(0,1]$ and consider the integral:

$$M_\delta^2=\int p\Big(\frac{p}{q}\Big)^\delta d\mu.$$

We shall prove that for a suitable choice of δ, $M_\delta^2<\infty$. Since all densities involved are bounded away from zero and infinity on compacta, we consider only the domain

$O = \mathbb{R}^2 \setminus [-C,C] \times [-C,C]$, for some large constant $C \geq A \vee B$. Note that:

$$\int_O p\Big(\frac{p}{q}\Big)^\delta d\mu \leq \int_O U\Big(\frac{U}{L}\Big)^\delta d\mu,$$

where (L,U) forms an envelope for the model. This envelope follows from the fact that the regression densities (10.3) fall in the class of mixture densities obtained by mixing the normal kernel $\varphi_\sigma(x)\varphi_\sigma(y)$ on $\mathbb{R}^2$ by means of a two-dimensional distribution that places all its mass in the rectangle $[-A,A] \times [-B,B]$. There exists a lower bound for this envelope which factorizes into x- and y-envelopes (L_X,U_X) and (L_Y,U_Y) that are constant on sets that include $[-A,A]$ and $[-B,B]$ respectively and have Gaussian tails. The domain O can therefore be partitioned into four subdomains in which either x or y is bounded and four subdomains in which both coordinates are unbounded. Reflection-symmetries of the envelope functions suffice to demonstrate that integrals of $U(U/L)^\delta$ can be expressed as products of trivially finite factors and integrals of the form:

$$\int_L^\infty U_X(x)\Big(\frac{U_X}{L_X}\Big)^\delta(x)\,d\mu(x), \quad \int_L^\infty U_Y(y)\Big(\frac{U_Y}{L_Y}\Big)^\delta(y)\,d\mu(y).$$

For large enough C, the envelope functions $L_X(x)$ and $U_X(x))$ are equal to multiples of $\varphi_{\underline{\sigma}}(x+A)$ and $\varphi_{\overline{\sigma}}(x-A))$ on the domain (C,∞) and hence, for some constants $c,K > 0$:

$$\int_L^\infty U_X(x)\Big(\frac{U_X}{L_X}\Big)^\delta(x)\,d\mu(x) \leq K\int_L^\infty e^{c\delta x^2}\,\varphi_{\overline{\sigma}}(x-A)\,dx,$$

which is finite for small enough $\delta > 0$. Similarly, one can prove finiteness of the integrals over y. This proves that the condition for theorem 5 in Wong and Shen (1995) [**?**] is satisfied. Note that the choice for δ is independent of p,q. Furthermore, the value of M_δ can be upper-bounded independent of p,q, as is apparent from the above. Hence, for small enough $\eta > 0$, (10.28) holds.

10.5 Regression classes

Theorems 10.3.7 and 10.4.2 demonstrate that both the entropy and prior-mass conditions in theorem 10.1.1 can be decomposed in a term that pertains to the regression function f and a term pertaining to the parameters (σ,F). This makes it possible to consider entropy and prior-mass restricted to the regression class separately.

In the first subsection, we state a bound on the metric entropy of the classes $C_{\beta,M}[-A,A]$ due to Kolmogorov, who derived it shortly after his introduction of the concept of covering numbers. This bound is used in the second subsection to demonstrate that so-called *net priors* can be used for non-parametric regression classes in this situation. Also discussed is an alternative approach, that uses (adapted versions of) *Jackson's approximation theorem*. Up to a logarithmic correction, the second approach reproduces Kolmogorov's bound for the metric entropy, but upon application

in the form of so-called *sieve priors*, the resulting lower bounds for the prior mass in neighbourhoods of the true regression function are sub-optimal in a more grave manner. Nevertheless, we indulge in an explanation of the second approach, because it provides a good example of the methods and subtleties of Bayesian procedures in non-parametric problems. We also give the necessary bounds on the entropy and prior mass of parametric regression classes.

10.5.1 Covering numbers of regression classes

The usefulness of bounds (10.17) and (10.18) indicates that the class of regression functions parametrizing the model is best chosen within the (Banach-)space $C[-A,A]$ of continuous functions on the closed interval $[-A,A]$ with the uniform norm $\|.\|$. According to the *Weierstrass approximation*, polynomials are dense in $C[-A,A]$; bounded families of polynomials can therefore be used to approximate regression families $\mathscr{F}$ as characterised in point *(c)* at the beginning of subsection 10.1.1. The Ascoli-Arzelà theorem asserts that if, in addition, $\mathscr{F}$ is equi-continuous, it is relatively compact. Hence bounded, equi-continuous families $\mathscr{F}$ are totally bounded in the norm-topology, rendering covering numbers finite,

$$N(\varepsilon,\mathscr{F},\|.\|)<\infty, \tag{10.29}$$

for all $\varepsilon>0$. However, since we are interested in rates of convergence, finiteness of covering numbers is not enough and a more detailed analysis of the behaviour of $N(\varepsilon,\mathscr{F},\|.\|)$ for small ε is needed. We reproduce here a result due to Kolmogorov and Tikhomirov (1961) [159] (in a version as presented in Van der Vaart and Wellner (1996) [247]), that gives the required bound:

Lemma 10.5.1. *Let $\beta>0$, $M>0$ be given. There exists a constant K depending only on β and A, such that:*

$$\log N\big(\varepsilon,C_{\beta,M}[-A,A],\|\cdot\|\big)\leq K\Big(\frac{1}{\varepsilon}\Big)^{1/\beta}, \tag{10.30}$$

for all $\varepsilon>0$.

The proof of this lemma is a special version of the proof of theorem 2.7.1 in [247], which consists of a fairly technical approximation by polynomials. To improve our understanding of the above result, we briefly digress on an approach that is based on Jackson's approximation theorem.

Fix an $n\geq 1$; *Jackson's approximation theorem* (see Jackson (1930) [**?**]) says that if $f\in \mathrm{Lip}_M(\alpha)$, there exists an n-th order polynomial p_n such that:

$$\|f-p_n\|\leq\frac{K}{n^\alpha}, \tag{10.31}$$

where $K > 0$ is a constant that depends only on A and M. Moreover, if $f \in D_{\alpha,M}(q)$, there exists a polynomial p_n of degree n such that:

$$\|f - p_n\| \leq \frac{K'}{n^{q+\alpha}}, \tag{10.32}$$

where $K' > 0$ is a constant that depends on A, q, α and M. Indeed, in its most general formulation, Jackson's theorem applies to arbitrary continuous functions f, relating the degree of approximation to the modulus of continuity. As such, it provides a more precise version of Weierstrass' theorem.

The class of *all* n-th degree polynomials is larger than needed for the purpose of defining nets over the bounded regression classes we are interested in. Let $B > 0$ denote the constant that bounds all functions in $\mathscr{F}$. With given $\gamma > 0$, define $P_n' = \{p \in P_n : \|p\| \leq (1+\gamma)B\}$. By virtue of the triangle inequality, any polynomial used to approximate f as in (10.31) or (10.32) satisfies a bound slightly above and arbitrarily close to B with increasing n. Hence, for large enough n, P_n' is a L/n^{β}-net over $C_{\beta,M}[-A,A]$, where $L > 0$ is a constant that depends only on the constants defining the regression class. For these finite-dimensional, bounded subsets of $C[-A,A]$, the order of suitable nets can be calculated. The upper-bound for the metric entropy of Lipschitz and smoothness classes based on Jackson's theorem takes the following form.

Lemma 10.5.2. *Let $\beta > 0$ and $M > 0$ be given. There exists a constant $K' > 0$ such that:*

$$\log N(\varepsilon, C_{\beta,M}[-A,A], \|\,.\,\|) \leq K'\Big(\frac{1}{\varepsilon}\Big)^{1/\beta} \log\frac{1}{\varepsilon},$$

for small enough $\varepsilon > 0$.

Proof. Let $\varepsilon > 0$ be given and choose n to be the smallest integer satisfying $n^{\beta} \geq 1/\varepsilon$. Define $P_n'' = \{p \in P_n : \|p\| \leq L\}$ for some $L > B$. As argued after (10.32), there is a uniformly bounded set P_n' of polynomials of degree n that forms an ε-net over $C_{\beta,M}[-A,A]$. If n is chosen large enough, P_n' is a proper subset of P_n''. To calculate an upper bound for the covering number of P_n', let $\delta > 0$ be given and let $p_1, \ldots, p_D$ be a (maximal) set of δ-separated polynomials in P_n', where D is the packing number $D(\delta, P_n', \|\,.\,\|)$. Note that the balls $B_i = \{p \in P_n' : \|p - p_i\| < \frac{1}{2}\delta\}$, $(i = 1, \ldots, D)$, do not intersect. If δ is chosen small enough, $B_i \subset P_n''$. The linear map $\hat{p} : \mathbb{R}^{n+1} \to P_n$ that takes a vector $(a_0, \ldots, a_n)$ into the polynomial $\sum_{m=0}^{n} a_m z^m$ is Borel measurable and is used to define the sets $C_i = \hat{p}^{-1}(B_i)$. Note that the sets C_i are obtained from $C = \hat{p}^{-1}(P_n'')$ by rescaling and translation for all i. By the same argument as used in the proof of lemma 10.33, we conclude that there is a constant L such that the packing number satisfies:

$$D(\delta, P_n', \|\,.\,\|) \leq \Big(\frac{L}{\delta}\Big)^{n+1},$$

for small enough $\delta > 0$, which serves as an upper bound for the covering number as well. Choosing δ equal to a suitable multiple of $n^{-\beta}$ for large enough n, we find

a constant $K' > 0$ and a net over $C_{\beta,M}[-A,A]$ in P_n of order bounded by $(K'n^\beta)^{n+1}$. The triangle inequality then guarantees the existence of a slightly less dense net over $C_{\beta,M}[-A,A]$ inside $C_{\beta,M}[-A,A]$ of the same order. We conclude that there exists a constant $K'' > 0$ such that:

$$\log N(\varepsilon, C_{\beta,M}[-A,A], \|\,.\,\|) \leq K'' n \log n^\beta,$$

for large enough n, which leads to the stated bound upon substitution of the relation between ε and n.

The power of ε in the bound asserted by the above lemma is that of lemma 10.5.1. The logarithmic correction can be traced back to the n-dependence of the radius of the covering balls B_i, *i.e.* the necessity of using finer and finer nets over P_n' to match the n-dependence in the degree of approximation. Therefore, there is no obvious way of adapting the above proof to eliminate the $\log(1/\varepsilon)$-factor and Kolmogorov's approach gives a strictly smaller bound on the entropy. However, the above illustrates the origin of the β-dependence in the power of ε more clearly.

For parametric classes (as given under *(iii)* in the beginning of section 10.2), the entropy is bounded in the following lemma.

Lemma 10.5.3. *For a parametric class $\mathscr{F}_\Theta$, there exists a constant $K > 0$ such that the metric entropy is bounded as follows:*

$$\log N(\varepsilon, \mathscr{F}_\Theta, \|\,.\,\|) \leq K \log \frac{1}{\varepsilon}, \tag{10.33}$$

for small enough $\varepsilon > 0$.

Proof. Since, by assumption, $\Theta \subset \mathbb{R}^k$ is bounded by some constant $M' > 0$, the covering numbers of Θ are upper-bounded by the covering numbers of the ball $B(0,M') \subset \mathbb{R}^k$ of radius M' centred on 0. Let $\delta > 0$ be given. Since covering numbers are bounded by packing numbers, we see that:

$$N(\delta, \Theta, \|\,.\,\|_{\mathbb{R}^k}) \leq D(\delta, B(0,M'), \|\,.\,\|_{\mathbb{R}^k}).$$

Let $\theta_1, \ldots, \theta_D$ (with $D = D(\delta, B(0,M'), \|\,.\,\|_{\mathbb{R}^k})$) be a maximal δ-separated subset of $B(0,M')$. The balls $B_i = B(\theta_i, \frac{1}{2}\delta)$ do not intersect and are all contained in the ball $B(0,M'+\frac{1}{2}\delta)$ by virtue of the triangle inequality. Therefore, the sum of the volumes of the balls B_i (which are all equal and proportional to $(\frac{1}{2}\delta)^k$, due to translation invariance and scaling behaviour of the Lebesgue measure) lies below the volume of the ball $B(0,M'+\frac{1}{2}\delta)$. We conclude that:

$$D(\delta, B(0,M'), \|\,.\,\|_{\mathbb{R}^k})(\tfrac{1}{2}\delta)^k \leq (M'+\tfrac{1}{2}\delta)^k.$$

Assuming that $\delta < 2M'$, we see that:

$$D(\delta, B(0,M'), \|\,.\,\|_{\mathbb{R}^k}) \leq \left(\frac{4M'}{\delta}\right)^k. \tag{10.34}$$

Next, note that due to (10.8), any δ-net over Θ leads to a $L\delta^\rho$-net over the regression class $\mathscr{F}_\Theta$, whence we see that:

$$N(L\delta^\rho, \mathscr{F}_\Theta, \|.\|) \leq N(\delta, \Theta, \|.\|_{\mathbb{R}^k}). \tag{10.35}$$

Let $\varepsilon > 0$ be given and choose $\delta = (\varepsilon/L)^{1/\rho}$. Combining (10.34) and (10.35), we find that there exists a constant $K > 0$ such that:

$$\log N(\varepsilon, \mathscr{F}_\Theta, \|.\|) \leq K \log \frac{1}{\varepsilon},$$

for small enough ε.

These bounds on the small-ε behaviour of the entropy are incorporated in the calculation of bounds for the entropy of the errors-in-variables model through theorem 10.3.7.

10.5.2 Priors on regression classes

This subsection is devoted to the definition of a suitable prior $\Pi_{\mathscr{F}}$ on the regression class $\mathscr{F}$. The challenge is to show that $\Pi_{\mathscr{F}}$ places 'enough' mass in small neighbourhoods of any point in the regression class. More specifically, a lower bound is needed for the prior mass of neighbourhoods of the (unknown) regression function $f_0 \in \mathscr{F}$:

$$\Pi_{\mathscr{F}}\big(f \in \mathscr{F} : \|f - f_0\| \leq \delta\big), \tag{10.36}$$

for small enough $\delta > 0$ (refer to theorem 10.4.2).

Jackson's theorem suggests that a natural definition of a prior on $\mathscr{F}$ entails the placement of prior mass on all (finite-dimensional) linear spaces of n-th degree polynomials P_n on $[-A, A]$, since their union is dense in $C[-A, A]$ and therefore also in $\mathscr{F}$. Fix the regression class $\mathscr{F}$. For all $n \geq 1$ we define:

$$\mathscr{F}_n = \mathscr{F} \cap P_n,$$

i.e. the subsets of n-th degree polynomials in the regression class. Note that $\mathscr{F}_n \subset \mathscr{F}_{n+1}$ for all n, and that $\mathscr{F}$ lies in the closure of their union. The linear map $\hat{p} : \mathbb{R}^{n+1} \to P_n$ that takes a vector $(a_0, ..., a_n)$ into the polynomial $\sum_{m=0}^n a_m z^m$ can be used to define a subset $\hat{p}^{-1}(\mathscr{F}_n) \subset \mathbb{R}^k$ with Lebesgue measure strictly above zero. Normalizing the Lebesgue measure to 1 on $\hat{p}^{-1}(\mathscr{F}_n)$, the inverse map $\hat{p}^{-1}$ serves to define a probability measure Π_n on $\mathscr{F}_n$. Any sequence $(b_n)_{n\geq 0}$ such that $b_n \geq 0$ and $\sum_{n=0}^\infty b_n = 1$, may be used to define a prior $\Pi_{\mathscr{F}}$ by the infinite convex combination:

$$\Pi_{\mathscr{F}}(A) = \sum_{n=0}^{\infty} b_n \Pi_n(A) = \sum_{n=0}^{\infty} b_n \Pi_n(A \cap \mathscr{F}_n), \tag{10.37}$$

for all A in the Borel σ-algebra generated by the norm topology on $\mathscr{F}$. Following Huang [129], we refer to priors obtained in this manner as *sieve priors*.

With a sieve prior, a proof of (10.4) amounts to showing that neighbourhoods of f_0 have intersections with the sets $\mathscr{F}_n$ and that the sum of the masses of these intersections is large enough. Obviously, Jackson's approximation provides a useful way to assert that balls centred on f_0 intersect with all P'_n from a certain minimal n onward. However, as is apparent from (10.36), this is not sufficient, because the relevant neighbourhoods are restricted to the regression class $\mathscr{F}$. One would have to show that these *restricted* neighbourhoods intersect with the sets $\mathscr{F}_n$.

Jackson's theorem does not assert anything concerning Lipschitz-bounds of the approximating polynomial or derivatives thereof. The assertion that p_n approximates f in uniform norm leaves room for very sharp fluctuations of p_n on small scales, even though it stays within a bracket of the form $[f - K/n^\beta, f + K/n^\beta]$. It is therefore possible that p_n lies far outside $\mathscr{F}_n$, rendering neighbourhoods of p_n in P_n unfit for the purpose. Although it is possible to adapt Jackson's theorem in such a way that the approximating polynomials satisfy a Lipschitz condition that is arbitrarily close to that of the regression class, this adaptation comes at a price with regard to the degree of approximation. As it turns out, this price leads to substantial corrections for the rate of convergence and ultimately to sub-optimality (with respect to the *power* of ε rather than logarithmically). That is not to say that sieve priors are in any sense sub-optimal. (Indeed, sieve priors have been used with considerable success in certain situations; for an interesting example, see the developments in adaptive Bayesian estimation, for instance in Huang [129].) The calculation underlying the claims made above merely shows that the construction via adapted versions of Jackson's theorem does not lead to optimal results, leaving the possibility that a sieve prior satisfies (10.4) open. What it does show, however, is that this may be very hard to demonstrate.

Therefore, we define the prior on the regression class in a different fashion (first proposed in Le Cam (197X) [178], based on ideas from Le Cam (1973) [177]), based on the upper bounds for covering numbers obtained in the previous subsection. Let the regression class $\mathscr{F}$ be a bounded, equi-continuous subset of $C[-A,A]$, so that the covering numbers $N(\varepsilon, \mathscr{F}, \|\,.\,\|)$ are finite for all $\varepsilon > 0$. Let $(a_m)_{m\geq 1}$ be a monotonically decreasing sequence, satisfying $a_m > 0$ (for all $m \geq 1$), and $a_m \downarrow 0$. For every $m \geq 1$, there exists an a_m-net $\{f_i \in \mathscr{F} : i = 1, \ldots, N_m\}$ over $\mathscr{F}$, where $N_m = N(a_m, \mathscr{F}, \|\,.\,\|)$. We define, for every $m \geq 1$, a discrete probability measure Π_m that distributes its mass uniformly over the set $\{f_i : i = 1, \ldots, N_m\}$:

$$\Pi_m = \sum_{i=1}^{N_m} \frac{1}{N_m} \delta_{f_i}.$$

Any sequence $(b_n)_{n\geq 0}$ such that $b_n \geq 0$ and $\sum_{n=0}^{\infty} b_n = 1$, may be used to define a prior $\Pi_{\mathscr{F}}$ on $\mathscr{F}$ by the infinite convex combination:

$$\Pi_{\mathscr{F}}(A) = \sum_{m=0}^{\infty} b_m \Pi_m(A), \tag{10.38}$$

for all A in the Borel σ-algebra generated by the norm topology on $\mathscr{F}$. Priors defined in this manner are referred to as a *net priors* and resemble those defined in Ghosal, Ghosh and Ramamoorthi (1997) [104], (see also, Ghosal *et al.* (2000) [106]).

Note that for all $m \geq 1$ and every $f \in \mathscr{F}$, there is an f_i satisfying $\|f - f_i\| \leq a_m$. So for every $f_0 \in \mathscr{F}$ and all $\delta > 0$, we have:

$$\Pi_m\big(\|f - f_0\| \leq \delta\big) \geq \frac{1}{N_m},$$

if $a_m \leq \delta$, *i.e.* for all m large enough. This means that the priors Π_m satisfy lower bounds for the mass in neighbourhoods of points in the regression class, that are inversely related to upper bounds satisfied by the covering numbers. As is demonstrated below, choices for the sequences a_m and b_m exist such that this property carries over to a prior of the form (10.38).

Lemma 10.5.4. *Let $\beta > 0$ and $M > 0$ be given and define $\mathscr{F}$ to be the class $C_{\beta,M}[-A,A]$. There exists a net prior $\Pi_{\mathscr{F}}$ and a constant $K > 0$ such that*

$$\log \Pi_{\mathscr{F}}\big(f \in \mathscr{F} : \|f - f_0\| \leq \delta\big) \geq -K\frac{1}{\delta^{1/\beta}}, \tag{10.39}$$

for small enough δ.

Proof. Define, for all $m \geq 1$, $a_m = m^{-\beta}$. Then the covering number N_m satisfies, for some constant $K' > 0$:

$$\log N_m = \log N(a_m, \mathscr{F}, \|\,.\,\|) \leq K' a_m^{-1/\beta} = K' m,$$

according to lemma 10.5.1. Let $\delta > 0$ be given and choose the sequence $b_m = (1/2)^m$. Let M be an integer such that:

$$\frac{1}{\delta^{1/\beta}} \leq M \leq \frac{1}{\delta^{1/\beta}} + 1.$$

Then for all $m \geq M$, $a_m \leq \delta$ and, due to the inequality (10.38), the net prior $\Pi_{\mathscr{F}}$ satisfies:

$$\begin{aligned}\Pi_{\mathscr{F}}\big(f \in \mathscr{F} : \|f - f_0\| \leq \delta\big) &\geq \sum_{m \geq M} b_m \Pi_m\big(\|f - f_0\| \leq \delta\big) \geq \sum_{m \geq M}\Big(\frac{e^{-K'}}{2}\Big)^m \\ &\geq \tfrac{1}{2}e^{-K'M} \geq \tfrac{1}{2}e^{-K'(\delta^{-1/\beta}+1)} \geq \tfrac{1}{2}e^{-2K'\delta^{-1/\beta}},\end{aligned} \tag{10.40}$$

for small enough δ.

For parametric classes, the prior mass in neighbourhoods of f_0 is lower-bounded in the following lemma.

Lemma 10.5.5. *Assume that the regression class $\mathscr{F}$ is parametric: $\mathscr{F} = \mathscr{F}_\Theta$. Any prior Π_Θ on Θ induces a prior $\Pi_{\mathscr{F}}$ with the Borel σ-algebra generated by the*

topology of the norm $\|.\|$ as its domain. Furthermore, if Π_Θ is dominated by the Lebesgue measure and has a density that is strictly positive at θ_0, then there exists a constant $R > 0$ such that the prior mass in neighbourhoods of f_0 is bounded as follows:

$$\log \Pi_{\mathscr{F}}\big(f \in \mathscr{F} : \|f - f_0\| \leq \varepsilon\big) \geq -R \log \frac{1}{\varepsilon}, \tag{10.41}$$

for small enough $\varepsilon > 0$.

Proof. The Lipschitz condition (10.8) ensures that the map $\hat{f} : \Theta \to \mathscr{F}_\Theta : \theta \mapsto f_\theta$ is continuous, implying measurability with respect to the corresponding Borel σ-algebras. So composition of Π_Θ with $\hat{f}^{-1}$ induces a suitable prior on $\mathscr{F}_\Theta$. As for the second assertion, let $\delta > 0$ be given. Since Π_Θ has a continuous Lebesgue density $\pi : \Theta \to \mathbb{R}$ that satisfies $\pi(\theta_0) > 0$ by assumption and since θ_0 is internal to Θ, there exists an open neighbourhood $U \subset \Theta$ of θ_0 and a constant $\pi_1 > 0$ such that $\pi(\theta) \geq \pi_1$ for all $\theta \in U$. Therefore, for all balls $B(\delta, \theta_0) \subset U$ (*i.e.* for small enough $\delta > 0$), we have:

$$\Pi_\Theta\big(B(\delta, \theta_0)\big) = \int_{B(\delta,\theta_0)} \pi(\theta)\, d\theta \geq V_k \pi_1 \delta^k,$$

where V_k is the Lebesgue measure of the unit ball in $\mathbb{R}^k$. Note that due to property (10.8),

$$\big\{\theta \in \Theta : \|\theta - \theta_0\| \leq \delta\big\} \subset \big\{\theta \in \Theta : \|f_\theta - f_0\| \leq L\delta^\rho\big\},$$

so that, for given $\varepsilon > 0$ and the choice $\delta = (\varepsilon/L)^{1/\rho}$:

$$\begin{aligned}\log \Pi_{\mathscr{F}}\big(f \in \mathscr{F} : \|f - f_0\| \leq \varepsilon\big) &\geq \log \Pi_\Theta\big(\theta \in \Theta : \|\theta - \theta_0\| \leq (\varepsilon/L)^{1/\rho}\big) \\ &\geq \log\big(V_k \pi_1 (\varepsilon/L)^{k/\rho}\big) \geq -R \log \frac{1}{\varepsilon},\end{aligned}$$

for some constant $R > 0$ and small enough ε.

The bounds on the small-ε behaviour of prior mass presented in this subsection are incorporated in the calculation of bounds for the prior mass of Kullback-Leibler neighbourhoods $B(P_0, \varepsilon)$ through theorem 10.4.2.

10.6 Asymptotic uncertainty quantification with Hellinger balls

Theorem 10.2.5 says the posterior converges at rate ε_n.

Theorem 10.6.1. *Let $\hat{B}_n = B_n(\hat{\theta}_n, \hat{r}_n)$ be level-$1 - \varepsilon$ credible balls of minimal radii and $C_n = B_n(\hat{\theta}_n, \hat{r}_n + \varepsilon_n) \subset B_n(\hat{\theta}_n, 2(1 + o(1))\varepsilon_n)$.*

$\mathscr{F} = \mathrm{Lip}_M(\alpha)$ with a net prior, the sets $C(X^n)$ have asymptotic coverage, and shrink like,

$$\varepsilon_n = n^{-\frac{\alpha}{2\alpha+1}} (\log n)^{\frac{1}{2\alpha}}.$$

$\mathscr{F} = D_{\alpha,M}(q)$ with a net prior, the sets $C(X^n)$ have asymptotic coverage, and shrink like,

$$\varepsilon_n = n^{-\frac{q+\alpha}{2q+2\alpha+1}} (\log n)^{\frac{1}{2q+2\alpha}}.$$

$\mathscr{F} = \mathscr{F}_\Theta$ with Lebesgue prior with continuous and strictly positive density, the sets $C(X^n)$ have asymptotic coverage, and shrink like,

$$\varepsilon_n = n^{-1/2} (\log n)^{3/2}.$$

Note that the assertion of theorem 10.2.5, based on the Ghosal-Ghosh-van der Vaart theorem, theorem 10.1.1, is *not strong enough* to actually construct such Hellinger confidence balls. The unknown constant M weakens the conclusion to, *there exists a constant $M > 0$* such that the $M\varepsilon_n$-enlargements (where ε_n represents the rates of theorem 10.2.5) of credible balls $\hat{B}_n$ are frequentist asymptotic confidence sets. This is not the case with posterior convergence as formulated in theorem 7.5.1 (provided the sets V_n are complements of balls of *known* radii, of course).

10.7 Exercises [EMPTY]

Chapter 11
Application: community detection in the planted bi-section model

To demonstrate how the methods presented in chapter 7 are applied in practice, we consider a sparse stochastic block model, focussing on the questions of community recovery, detection and uncertainty quantification.

11.1 Communities in random graphs

The stochastic block model is a generalization of the Erdős-Rényi random graph model [90] where one studies a version X^n of the complete graph between n vertices under percolation, with edge probability $p_n \in [0,1]$. Stochastic block models [127] are similar but concern random graphs with vertices that belong to one of several classes and edge probabilities that depend on those classes. If we think of the graph X^n as data and the class assignments of the vertices as unobserved, an interesting statistical challenge presents itself regarding estimation of (and other forms of inference on) the vertices' class assignments, a task referred to as *community detection* [112]. The stochastic block model and its generalizations have applications in physics, biology, sociology, image processing, genetics, medicine, logistics, *etcetera* and are widely employed as canonical models for the study of clustering and community detection [96, 1]; [261] even state that, *"Community detection for the stochastic block model is probably the most studied topic in network analysis."*

In an asymptotic sense one may wonder under which conditions on edge probabilities, community detection can be done in a 'statistically consistent' way as the number of vertices n grows; particularly, whether it is possible to estimate the true class assignments correctly (*exact recovery*), or correctly for a fraction of the vertices that goes to one (*almost-exact recovery*), with high probability (see definitions 11.2.1 and 11.2.2 for details). Note that the Erdős-Rényi graph already displays very rich asymptotic behaviour: edge probabilities $p_n \geq An^{-1}\log n$ lead to a connected graph with high probability, if and only if $A > 1$; a *giant component* occurs with high probability, if and only if $p_n \geq C/n$ with $C > 1$; below the $1/n$-threshold, the graph X^n fragments into many disconnected sub-graphs of order no larger than

$O(\log n)$ with high probability. At the boundaries $1/n$ and $n^{-1}\log n$, the Erdős-Rényi graph is said to undergo *phase transitions* [43], from the *fragmented phase* to the sparse *Kesten-Stigum phase* and then to the less sparse *Chernoff-Hellinger phase*.

Here and in [2, 190, 199], the community detection problem is studied in the context of the so-called *planted bi-section model*, which is a stochastic block model with two classes, each of n vertices and edge probabilities p_n (within-class) and q_n (between-class). A famous sufficient condition for exact recovery of the class assignment in the planted bi-section model comes from [84]: if there exists a constant $A > 0$ such that, $p_n - q_n \geq An^{-1}\log n$, then community detection by minimization of the number of edges between estimated classes achieves exact recovery. In in [66, 67], it was conjectured that almost-exact recovery is possible in block models, if $n(p_n - q_n)^2 > 2(p_n + q_n)$. [199] prove a definitive assertion: *almost-exact recovery* is possible (by any estimator or algorithm), if and only if,

$$\frac{n(p_n - q_n)^2}{p_n + q_n} \to \infty. \tag{11.1}$$

An analogous claim in the Chernoff-Hellinger phase was first considered more rigorously in [190] and later confirmed, both from a probabilistic/statistical perspective in [197, 199], and independently from an information theoretic perspective in [2]. Defining a_n and b_n by $np_n = a_n \log n$ and $nq_n = b_n \log n$ and assuming that $C^{-1} \leq a_n, b_n \leq C$ for all but finitely many $n \geq 1$, the class assignment in the planted bi-section model can be *recovered exactly*, if and only if,

$$(a_n + b_n - 2\sqrt{a_n b_n} - 1)\log n + \tfrac{1}{2}\log\log n \to \infty, \tag{11.2}$$

(see [199]). Conditions (11.1) and (11.2) not only lower-bound the degree of edge-sparsity, but also guarantee sufficient distance [7] from the Erdős-Rényi graph ($p_n = q_n$), in which communities are not identifiable.

Estimation methods used for the community detection problem include spectral clustering (see [163] and many others), maximization of the likelihood and other modularities [112, 30, 57, 5], semi-definite programming [119, 118], and penalized ML detection of communities with minimax optimal misclassification ratio [261, 102]. More generally, we refer to [1] and the very informative introduction of [102] for extensive bibliographies and a more comprehensive discussion. Bayesian methods have been popular throughout, *e.g.* the original work [**?**], the work of [66, 67] and more recently, [238], based on an empirical prior choice, and [198]. The machine learners' interest in the stochastic block model has generated a wealth of algorithms that estimate the class assignment. We mention only maximization of the likelihood or other modularities [112, 30] and refer to the discussions in [261, 102].

In this paper, the first goal is to explore the limits of what is possible from the statistical point of view, similar to what Mossel *et al.* do from the probabilistic point of view and Abbe *et al.* from the information theoretic point of view. So first of all, in section 11.3 it is shown that posteriors recover underlying class assignments exactly and almost-exactly, under conditions on (p_n) and (q_n) that are sharp. To be more precise: condition (11.1) is found to be sufficient for almost-exact recovery with

posteriors with uniform priors. This implies that if there *exists* an estimator or algorithm that recovers the class assignment almost-exactly, then posteriors *also recover* the class assignment *almost exactly*. Similarly, (a slight variation on) the necessary condition 11.2 is shown to be sufficient for posteriors to recover the community assignment exactly.

The second goal concerns a far more important advantage posteriors offer over other estimation methods: in section 11.5, *credible sets for community assignment* are shown to be (or can be enlarged to form) *asymptotic confidence sets*. Since sampling distributions of other estimators are mostly prohibitively hard to analyse, obtaining (asymptotic) confidence sets for community assignment in other ways may prove very hard. To the best of the authors' knowledge, frequentist uncertainty quantification with *confidence sets for class assignment* has not been addressed in the literature. We conclude that, in the context of the planted bi-section model and also much wider, the relatively high computational cost of simulating a posterior is quite justifiable if one is interested in uncertainty quantification. Section 11.4 provides a sharp calculation of testing power for likelihood ratio tests; section **??** applies remote contiguity to convert credible sets to confidence sets as in chapter 7.

11.2 The planted bi-section model

In a stochastic block model, each vertex is assigned to one of $K \geq 2$ classes through an unobserved *class assignment vector* θ'. Each vertex belongs to a class and any edge occurs (independently of others) with a probability depending on the classes of the vertices that it connects. In the *planted bi-section model*, there are only two classes ($K = 2$) and, at the n-th iteration ($n \geq 1$), there are $2n$ vertices (labelled with indices $1 \leq i \leq 2n$), n in each class, with class assignment vector $\theta' \in \Theta'_n$ (with components $\theta'_1, \ldots, \theta'_{2n} \in \{0,1\}$), where Θ'_n is the subset of $\{0,1\}^{2n}$ of all finite binary sequences that contain as many ones as zeroes. Denote that space in which the random graph X^n takes its values by $\mathscr{X}_n$ (*e.g.* represented by its adjacency matrix with entries $\{X_{ij} : 1 \leq i, j \leq 2n\}$). The ($n$-dependent) probability of an edge occuring ($X_{ij} = 1$) between vertices $1 \leq i, j \leq 2n$ *within the same class* is denoted $p_n \in (0,1)$; the probability of an edge *between classes* is denoted $q_n \in (0,1)$,

$$Q_{ij}(\theta') := P_{\theta,n}(X_{ij} = 1) = \begin{cases} p_n, & \text{if } \theta'_{n,i} = \theta'_{n,j}, \\ q_n, & \text{if } \theta'_{n,i} \neq \theta'_{n,j}, \end{cases} \tag{11.3}$$

Note that if $p_n = q_n$, X^n is the Erdős-Rényi graph $G(2n, p_n)$ and the class assignment $\theta_n \in \Theta'_n$ is not identifiable. Another identifiability issue that arises is that the model is invariant under interchange of class labels 0 and 1. This is expressed in the parameter spaces Θ'_n through equivalence relations: $\theta'_1 \sim_n \theta'_2$, if $\theta'_{2,n} = \neg\theta'_{1,n}$ (by componentwise negation). To prevent non-identifiability, we parametrize the model for X^n in terms of a parameter θ_n in a quotient space $\Theta_n = \Theta'_n / \sim_n$, for every $n \geq 1$. For $\theta'_n \in \Theta'_n$ we denote the equivalence class $\{\theta'_n, \neg\theta'_n\}$ by θ_n. Note that the set Θ_n

can be identified with the set of partitions of $\{1,\ldots,2n\}$ consisting of exactly two sets with n elements, via the identification

$$\theta_n \longleftrightarrow \left\{\left\{i : \theta'_{n,i} = 0\right\}, \left\{i : \theta'_{n,i} = 1\right\}\right\},$$

and note that this is independent of the choice of the representation.

The probability measure for the graph X^n corresponding to parameter θ is denoted $P_{\theta,n}$. The likelihood is given by,

$$p_{\theta,n}(X^n) = \prod_{i<j} Q_{i,j}(\theta)^{X_{ij}}(1 - Q_{i,j}(\theta))^{1-X_{ij}}.$$

For the *sparse versions* of the planted bi-section model, we also define edge probabilities that vanish with growing n: take (a_n) and (b_n) such that $a_n \log n = np_n$ and $b_n \log n = nq_n$ for the Chernoff-Hellinger phase; take (c_n) and (d_n) such that $c_n = np_n$ and $d_n = nq_n$ for the Kesten-Stigum phase. The fact that we do not allow loops (edges that connect vertices with themselves) leaves room for $2 \cdot \frac{1}{2}n(n-1) + n^2 = 2n^2 - n = \frac{1}{2} \cdot (2n)(2n-1)$ possible edges in the random graph X^n observed at iteration n.

The statistical question of interest in this model is to reconstruct the unobserved class assignment vectors θ_n *consistently*, that is, (close to) correctly with probability growing to one as $n \to \infty$. Consistency can be stated in various ways, as defined below.

Definition 11.2.1. Let $\theta_{0,n} \in \Theta_n$ be given. An estimator sequence $\hat{\theta}_n : \mathscr{X}_n \to \Theta_n$ is said to *recover the class assignment* $\theta_{0,n}$ *exactly* if,

$$P_{\theta_{0,n}}\left(\hat{\theta}_n(X^n) = \theta_{0,n}\right) \to 1,$$

that is, if $\hat{\theta}_n$ indicates the correct communities with high probability.

We also relax this consistency requirement somewhat in the form of the following definition, *c.f.* [199] and others: for $n \geq 1$ and two class assignments $\theta_{0,n}, \theta_n \in \Theta_n$, let $k(\theta_n, \theta_{0,n})$ denote the *minimal number of pair exchanges needed to transform* θ_n *into* $\theta_{0,n}$ (for further details, see the definition of k, just before eq. (11.6) below).

Definition 11.2.2. Let $\theta_{0,n} \in \Theta_n$ be given. An estimator sequence $\hat{\theta}_n : \mathscr{X}_n \to \Theta_n$ is said to *recover* $\theta_{0,n}$ *almost-exactly*, if $k(\hat{\theta}_n, \theta_{0,n})$ is of order $o_P(n)$ under $P_{\theta_{0,n}}$. If, for some sequence $l_n = o(n)$,

$$P_{\theta_{0,n}}\left(k(\hat{\theta}_n, \theta_{0,n}) \leq l_n\right) \to 1.$$

we say that $\hat{\theta}_n$ *recovers* $\theta_{0,n}$ *with error rate* l_n.

Below, we specialize to the Bayesian approach: we choose prior distributions π_n for all Θ_n, $(n \geq 1)$ and calculate the posterior: denoting the likelihood by $p_{\theta,n}(X^n)$, the posterior for the parameter θ_n is written as a fraction of sums, for all $A \subset \Theta_n$,

$$\Pi(A|X^n) = \sum_{\theta_n \in A} p_{\theta,n}(X^n)\,\pi_n(\theta_n) \Big/ \sum_{\theta_n \in \Theta_n} p_{\theta,n}(X^n)\,\pi_n(\theta_n),$$

where $\pi_n : \Theta_n \to [0,1]$ is the probability mass function for the prior Π_n. Here, we only consider *uniform priors* (Π_n) for $\theta_n \in \Theta_n$, so for all $n \geq 1$ and $\theta_n \in \Theta_n$, $\pi(\theta_n) = \pi_n := (|\Theta_n|)^{-1}$.

11.3 Exact and almost-exact recovery with posteriors

Consider the sequence of experiments in which we observe random graphs $X^n \in \mathscr{X}_n$ generated by the planted bi-section model of definition (11.3). Assuming for every $n \geq 1$, that the prior is uniform, we have $\pi_n = (\frac{1}{2}\binom{2n}{n})^{-1}$.

Given true parameters $\theta_{0,n} \in \Theta_n$ ($n \geq 1$), choose representations $\theta'_{0,n} \in \Theta'_n$ and define $Z_n(\theta'_0) \subset \{1,\ldots,2n\}$ to be *class zero* (the set of all those i such that $\theta'_{0,i} = 0$) and call the complement $Z_n^c(\theta'_0)$ *class one*. For the questions concerning exact recovery and detection, we are interested in the sets $V'_{n,k} \subset \Theta'_n$, defined to contain all those θ'_n that differ from $\theta'_{0,n}$ by exactly k exchanges of pairs: for $\theta'_n \in \Theta'_n$ we have $\theta'_n \in V'_{n,k}$, if the set of vertices in class zero *c.f.* $\theta'_{0,n}$, $Z(\theta'_{0,n}) = \{1 \leq i \leq 2n : \theta'_{0,n,i} = 0\}$, from which we leave out the set of vertices in class zero *c.f.* θ'_n, $Z(\theta'_n) = \{1 \leq i \leq 2n : \theta'_{n,i} = 0\}$, has k elements. Conversely, for any $\theta'_{1,n}$ and $\theta'_{2,n}$ in Θ'_n, we denote the minimal number of pair-exchanges necessary to take $\theta'_{1,n}$ into $\theta'_{2,n}$ by $k'(\theta'_{1,n},\theta'_{2,n})$. Note that $k'(\theta'_{1,n},\neg\theta'_{2,n}) = n - k'(\theta'_{1,n},\theta'_{2,n})$, which leads to the metric between two representation classes,

$$k(\theta_{1,n},\theta_{2,n}) = k'(\theta'_{1,n},\theta'_{2,n}) \wedge k'(\theta'_{1,n},\neg\theta'_{2,n}) \tag{11.4}$$

and note that this is independent of choice of the representations and that this function k takes values in $\{0,\ldots,\lfloor n/2 \rfloor\}$. Now define,

$$V_{n,k} = V_{n,k}(\theta_{0,n}) = \{\theta_n : k(\theta_n,\theta_{0,n}) = k\} = \left\{\theta_n : \theta'_n \in V'_{n,k}\right\}, \tag{11.5}$$

for $k \in \{1,\ldots,\lfloor n/2 \rfloor\}$. Given some sequence (k_n) of positive integers we then define V_n as the disjoint union,

$$V_n = \bigcup_{k=k_n}^{\lfloor n/2 \rfloor} V_{n,k} \tag{11.6}$$

Since we can choose two subsets of k elements from two sets of size n in $\binom{n}{k}^2$ ways, the cardinal of $V_{n,k}$ is $\binom{n}{k}^2$, when $k < n/2$ and $\frac{1}{2}\binom{n}{n/2}^2$ when n is even and $k = n/2$. In both cases the number of elements in $V_{n,k}$ is therefore bounded by $\binom{n}{k}^2$.

According to lemma 2.2 in [157] (with $B_n = \{\theta_{0,n}\}$), for any test sequences $\phi_{k,n} : \mathscr{X}_n \to [0,1]$ ($k \geq 1, n \geq 1$), we have,

$$P_{\theta_0,n}\Pi(V_n|X^n) = \sum_{k=k_n}^{\lfloor n/2\rfloor} P_{\theta_0,n}\Pi(V_{n,k}|X^n)$$
$$\leq \sum_{k=k_n}^{\lfloor n/2\rfloor}\Big(P_{\theta_0,n}\phi_{k,n}(X^n) + \sum_{\theta_n\in V_{n,k}} P_{\theta_n,n}(1-\phi_{k,n}(X^n))\Big)$$

for every $n \geq 1$. Suppose that for any $k \geq 1$ there exists a sequence $(a_{n,k})_{n\geq 1}$, $a_{n,k} \downarrow 0$ and, for any $\theta_n \in V_{n,k}$, a test function $\phi_{\theta_n,n}$ that distinguishes $\theta_{0,n}$ from θ_n as follows,

$$P_{\theta_0,n}\phi_{\theta_n,n}(X^n) + P_{\theta_n,n}(1-\phi_{\theta_n,n}(X^n)) \leq a_{n,k}, \tag{11.7}$$

for all $n \geq 1$. Then using test functions $\phi_{k,n}(X^n) = \max\{\phi_{\theta_n,n}(X^n) : \theta_n \in V_{n,k}\}$, as well as the fact that,

$$P_{\theta_0,n}\phi_{k,n}(X^n) \leq \sum_{\theta_n\in V_{n,k}} P_{\theta_0,n}\phi_{\theta_n,n}(X^n),$$

we see that,

$$P_{\theta_0,n}\Pi(V_n|X^n) \leq \sum_{k=k_n}^{\lfloor n/2\rfloor}\sum_{\theta_n\in V_{n,k}}\Big(P_{\theta_{0,n},n}\phi_{\theta_n,n}(X^n) + P_{\theta_n,n}(1-\phi_{\theta_n,n}(X^n))\Big)$$
$$\leq \sum_{k=k_n}^{\lfloor n/2\rfloor}\binom{n}{k}^2 a_{k,n}.$$

This inequality forms the basis for the results in the next two subsections, on exact recovery and almost-exact recovery.

11.3.1 Posterior consistency: exact recovery

We are interested in the expected posterior masses of subsets of Θ_n of the form:

$$V_n = \{\theta_n \in \Theta_n : \theta_n \neq \theta_{0,n}\} = \bigcup_{k=1}^{\lfloor n/2\rfloor} V_{n,k}.$$

The theorem states a sufficient condition for (p_n) and (q_n), which is related to requirement (11.2) in the Chernoff-Hellinger phase.

Theorem 11.3.1. *For some $\theta_{0,n} \in \Theta_n$, assume that $X^n \sim P_{\theta_0,n}$, for every $n \geq 1$. If we equip every Θ_n with its uniform prior and (p_n) and (q_n) are such that,*

$$\Big(1 + \big(1 - p_n - q_n + 2p_nq_n + 2\sqrt{p_n(1-p_n)q_n(1-q_n)}\big)^{n/2}\Big)^{2n} \to 1, \tag{11.8}$$

as $n \to \infty$, then,

$$\Pi\big(\,\theta_n = \theta_{0,n}\,\big|\,X^n\,\big) \xrightarrow{P_{\theta_{0,n}}} 1, \tag{11.9}$$

as $n \to \infty$, i.e. the posterior recovers the community assignment exactly.

Proof. According to lemma 11.4.1, for every $n \geq 1$, $k \geq 1$ and given, $\theta_{0,n}$, there exists a test sequence satisfying (11.7) with $a_{n,k} = (1-\mu_n)^{2k(n-k)}$ and $\mu_n = p_n + q_n - 2p_n q_n - 2(p_n(1-p_n)q_n(1-q_n))^{1/2} \in [0,1]$. Therefore, with $z_n = (1-\mu_n)^{n/2}$,

$$\begin{aligned} P_{\theta_{0,n}}\Pi(V_n|X^n) &\leq \sum_{k=1}^{\lfloor n/2\rfloor}\binom{n}{k}^2(1-\mu_n)^{2k(n-k)} \leq \sum_{k=1}^{\lfloor n/2\rfloor}\binom{n}{k}^2(1-\mu_n)^{nk} \\ &\leq \sum_{k=1}^{\lfloor n/2\rfloor}\binom{2n}{2k}(1-\mu_n)^{nk} \leq \sum_{l=1}^{2n}\binom{2n}{l}z_n^l = (1+z_n)^{2n}-1 \end{aligned}$$

The right-hand side goes to zero if (11.8) is satisfied.

Example 11.3.2. Consider (11.8) in the sparse Chernoff-Hellinger phase, where $np_n = a_n \log n$, $nq_n = b_n \log n$ with $a_n, b_n = O(1)$. In that case,

$$\begin{aligned} &\Big(1+\big(1-p_n-q_n+2p_nq_n+2\sqrt{p_n(1-p_n)q_n(1-q_n)}\big)^{n/2}\Big)^{2n} \\ &\quad= \left(1+\left(1-\left(a_n+b_n-2\sqrt{a_nb_n}+o(n^{-1}\log n)\right)\frac{\log n}{n}\right)^{n/2}\right)^{2n} \\ &\quad\approx \left(1+n^{-\frac{1}{2}(a_n+b_n-2\sqrt{a_nb_n}))}\right)^{2n} = \left(1+\frac{1}{n}n^{-\frac{1}{2}(a_n+b_n-2\sqrt{a_nb_n}-2)}\right)^{2n} \\ &\quad\approx \exp\big(2e^{-\frac{1}{2}(a_n+b_n-2\sqrt{a_nb_n}-2)\log n}\big) \end{aligned} \tag{11.10}$$

Accordingly, in the Chernoff-Hellinger phase (11.8) amounts to the sufficient condition,

$$(a_n+b_n-2\sqrt{a_nb_n}-2)\log n \to \infty, \tag{11.11}$$

which closely resembles (but is not exactly equal to) (11.2), the requirement of [199]. In fact there is a trade-off: (11.2) is slightly weaker than (11.11) but applies only if there exists a $C>0$ such that $C^{-1} \leq a_n, b_n \leq C$ for large enough n [199, 261]. This bound excludes some interesting examples in which one of the sequences (a_n) and (b_n) may fade away with growing n or equal zero outright. For instance, if $b_n = 0$ and $\liminf_n a_n > 2$, edges between classes are completely absent but, separately, the Erdős-Rényi graphs spanned by vertices in $Z_n(\theta_0')$ and $Z_n^c(\theta_0')$ respectively are connected with high probability. Similarly, if $a_n = 0$ and $\liminf_n b_n > 2$, the posterior succeeds in exact recovery: possibly, with b_n above 2, edges between classes are abundant enough to guarantee the existence of a path in X^n that visits all vertices at least once, with high probability. It is tempting to state the following, well-known [2, 199] sufficient condition for the sequences $a_n > 0$ and $b_n > 0$:

$$(\sqrt{a_n}-\sqrt{b_n})^2 > c, \textit{ for some } c > 2 \textit{ and } n \textit{ large enough,} \tag{11.12}$$

(even though it ignores the logarithm in (11.11)).

Corollary 11.3.3. *Under the conditions of (11.3.1), the MAP-/ML-estimator recovers $\theta_{0,n}$ exactly.*

Proof. Due to the uniformity of the prior, for every $n \geq 1$, maximization of the posterior density (with respect to the counting measure) on Θ_n, is the same as maximization of the likelihood. Due to (11.9), the posterior density in the points $\theta_{0,n}$ in Θ_n converges to one in $P_{\theta_{0,n}}$-probability. Accordingly, the point of maximization is $\theta_{0,n}$ with high probability.

11.3.2 Posterior consistency: almost-exact recovery

For the case of almost-exact recovery, the requirement of convergence is less stringent: as said, [199, proposition 2.9] states that condition (11.1) is *necessary and sufficient* for almost-exact recovery. Below we show that posteriors with uniform priors *recover* the true class assignment *almost exactly* if (11.1) holds.

We are interested in the expected posterior masses of subsets of Θ_n of the form:

$$W_n = \bigcup_{k=k_n}^{\lfloor n/2 \rfloor} V_{n,k},$$

for a sequence k_n of order $o(n)$ or $O(n)$: the posterior concentrates on class assignments θ_n that differ from $\theta_{0,n}$ by no more than k_n pair exchanges.

Theorem 11.3.4. *For some $\theta_{0,n} \in \Theta_n$, let $X^n \sim P_{\theta_{0,n}}$ for every $n \geq 1$. If we equip all Θ_n with uniform priors and edge-probabilities (p_n), (q_n) and error rates (k_n) are such that,*

$$\frac{n}{k_n}\left(1 - p_n - q_n + 2p_n q_n + 2\sqrt{(p_n(1-p_n)q_n(1-q_n))}\right)^{n/2} \to 0, \tag{11.13}$$

as $n \to \infty$, then,

$$\Pi(W_n | X^n) \xrightarrow{P_0} 0, \tag{11.14}$$

as $n \to \infty$, i.e. the posterior recovers $\theta_{0,n}$ with error rate k_n.

Proof. According to lemma 11.4.1, for every $n \geq 1$, $k \geq 1$ and given $\theta_{0,n}$, there exists a test sequence satisfying (11.7) with $a_{n,k} = (1-\mu_n)^{2k(n-k)}$. Therefore, using the inequalities $\binom{2n}{k} \leq \frac{(2n)^k}{k!}$ and $(n+m)! \geq n!m!$, the Stirling lower bound formula, and finally our assumption $n(1-\mu_n)^{n/2}/k_n \to 0$, we see that for big enough n,

$$P_{\theta_0,n}\Pi(W_n|X^n) \leq \sum_{k=k_n}^{\lfloor n/2\rfloor} \binom{n}{k}^2 (1-\mu_n)^{2k(n-k)}$$

$$\leq \sum_{k=2k_n}^{n} \binom{2n}{k} (1-\mu_n)^{k(n-k/2)} \leq \sum_{k=2k_n}^{\infty} \frac{1}{k!}(2n)^k (1-\mu_n)^{kn/2}$$

$$\leq \frac{\left(2n(1-\mu_n)^{n/2}\right)^{2k_n}}{(2k_n)!} e^{2n(1-\mu_n)^{n/2}}.$$

We then see that,

$$P_{\theta_0,n}\Pi(W_n|X^n) \leq \frac{1}{\sqrt{4\pi k_n}} \left(\frac{n(1-\mu_n)^{n/2}}{k_n}\right)^{2k_n} e^{2k_n+2n(1-\mu_n)^{n/2}}$$

$$\leq \frac{1}{\sqrt{4\pi k_n}} \left(\frac{n(1-\mu_n)^{n/2}}{k_n} e^{1+n(1-\mu_n)^{n/2}/k_n}\right)^{2k_n}$$

$$\leq \frac{n(1-\mu_n)^{n/2}}{k_n} e^{1+n(1-\mu_n)^{n/2}/k_n}$$

which converges to zero as $n \to \infty$.

Example 11.3.5. Note that as $p_n, q_n = O(n^{-1}) = o(1)$, we may expand,

$$\sqrt{p_n} - \sqrt{q_n} = \frac{1}{2\sqrt{\frac{1}{2}(p_n+q_n)}}(p_n - q_n) + O(|p_n - q_n|^2).$$

which means that,

$$\mu_n = (\sqrt{p_n} - \sqrt{q_n})^2 + O(n^{-2}) = \frac{(p_n-q_n)^2}{2(p_n+q_n)} + O(n^{-2})$$

Assuming only that $n(p_n - q_n)^2 > 2(p_n + q_n)$, as in [66, 67], we would arrive at the conclusion that $n\mu_n > 1 + O(n^{-1})$, which is insufficient in the proof of theorem 11.3.4. Note that a non-divergent choice $k_n = O(1)$ forces us back into the Chernoff-Hellinger phase where exact recovery is possible.

Corollary 11.3.6. *Under the conditions of theorem 11.3.4 with (p_n) and (q_n) such that,*

$$n\big(p_n + q_n - 2p_n q_n - 2\sqrt{(p_n(1-p_n)q_n(1-q_n))}\big) \to \infty, \tag{11.15}$$

as $n \to \infty$, posteriors recover $\theta_{0,n}$ partially,

$$\Pi\big(k(\theta_n, \theta_{0,n}) \geq \beta n \;\big|\; X^n\big) \xrightarrow{P_0} 0,$$

for any *fraction $\beta \in (0, \frac{1}{2})$, which implies that the posterior recovers $\theta_{0,n}$ almost-exactly.*

Proof. Let $\beta \in (0, \frac{1}{2})$ be given. Follow the proof of theorem 11.3.4 with $k_n = \beta n$ and note that,

$$P_{\theta_0,n}\Pi\big(k(\theta_n, \theta_{0,n}) \geq \beta n \,\big|\, X^n\big) \leq \frac{1}{\beta}(1-\mu_n)^{n/2} e^{1+\beta^{-1}(1-\mu_n)^{n/2}}.$$

Due to eq. (11.15),

$$(1-\mu_n)^{n/2} = \big(1 - p_n - q_n + 2p_n q_n + 2\sqrt{(p_n(1-p_n)q_n(1-q_n))}\big)^{n/2} \to 0,$$

so $P_{\theta_0,n}\Pi(k(\theta_n, \theta_{0,n}) \geq \beta n \,|\, X^n) \to 0$. For almost-exact recovery, let $\beta_m \downarrow 0$ be given; if we let $m(n)$ go to infinity slowly enough, posterior convergence continues to hold with β equal to $\beta_{m(n)}$.

The condition that $n\mu_n \to \infty$ is *not just sufficient* for almost-exact posterior recovery; as said [199, proposition 2.10], it is *also necessary* for any form of almost-exact recovery. A somewhat provocative way of re-phrasing this, is as follows.

Corollary 11.3.7. *If there* exist any estimators $\hat{\theta}_n : \mathscr{X}_n \to \Theta_n$ *that* recover *the class assignment* almost exactly*, then posteriors with uniform priors* also recover *the class assignment* almost-exactly.

The latter result is encouraging to the Bayesian and to the frequentist who use Bayesian methods in this model and in models like it, *e.g.* the stochastic block model.

11.4 Existence of suitable tests

Given $n \geq 1$ and two class assignment vectors $\theta_{0,n}, \theta_n \in \Theta_n$, we are interested in calculation of the likelihood ratio $dP_{\theta,n}/dP_{\theta_0,n}$, because it determines testing power as well as the various forms of remote contiguity that play a role.

Choose representations θ_0' of θ_0 and θ' of θ so that $k'(\theta_0', \theta') = k(\theta_0, \theta)$, where k and k' are as in section 11.3. Recall that, $Z_n(\theta_0') \subset \{1, \ldots, 2n\}$ is class zero and the complement $Z_n^c(\theta_0')$ class one. For the sake of presentation (in figure 11.1 below), re-label the vertices such that $Z(\theta_0') = \{1, \ldots, n\}$ and $Z^c(\theta_0') = \{n+1, \ldots, 2n\}$. In the case $n = 4$, figure 11.1 shows edge probabilities in the familiar block arrangement.

Recall that the likelihood under θ_0 is given by,

$$p_{\theta_0,n}(X^n) = \prod_{i<j} Q_{i,j}(\theta_0)^{X_{ij}}(1 - Q_{i,j}(\theta_0))^{1-X_{ij}}.$$

If we assume that $\theta_{0,n}'$ and θ_n' differ by k pair-exchanges among respective members of the zero- and one-classes, then a look at figure 11.1 reveals that the likelihood-ratio depends only on the edges for which exactly one of its end-points changes class. Define,

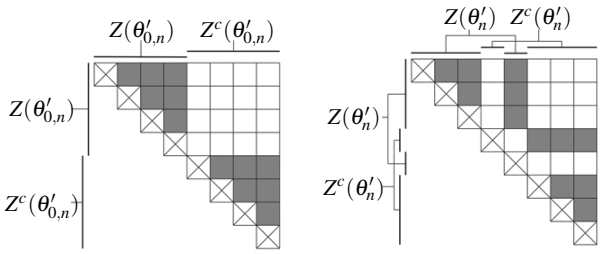


Fig. 11.1 Class assignments and edge probabilitites according to $\theta'_{0,n}$ and to θ'_n for $n = 4$ and $k = 1$. Vertex sets $Z(\cdot)$ and $Z^c(\cdot)$ correspond to zero- and one-classes for the given class assignment. Dark squares correspond to edges that occur with (within-class) probability p_n, and light squares to edges that occur with (between-class) probability q_n.

$$A_n = \{(i,j) \in \{1,\ldots,2n\} : i < j,\, \theta'_{0,n,i} = \theta'_{0,n,j},\, \theta'_{n,i} \neq \theta'_{n,j}\},$$
$$B_n = \{(i,j) \in \{1,\ldots,2n\} : i < j,\, \theta'_{0,n,i} \neq \theta'_{0,n,j},\, \theta'_{n,i} = \theta'_{n,j}\}.$$

Also define,

$$(S_n, T_n) := \Big(\sum\{X_{ij} : (i,j) \in A_n\}, \sum\{X_{ij} : (i,j) \in B_n\}\Big),$$

and note that the likelihood ratio can be written as,

$$\frac{p_{\theta,n}}{p_{\theta_0,n}}(X^n) = \left(\frac{1-p_n}{p_n}\,\frac{q_n}{1-q_n}\right)^{S_n - T_n} \tag{11.16}$$

where,

$$(S_n, T_n) \sim \begin{cases} \mathrm{Bin}(2k(n-k), p_n) \times \mathrm{Bin}(2k(n-k), q_n), & \text{if } X^n \sim P_{\theta_0,n}, \\ \mathrm{Bin}(2k(n-k), q_n) \times \mathrm{Bin}(2k(n-k), p_n), & \text{if } X^n \sim P_{\theta,n}. \end{cases} \tag{11.17}$$

Based on that, we derive the following lemma.

Lemma 11.4.1. *Let $n \geq 1$, $\theta_{0,n}, \theta_n \in \Theta_n$ be given. Assume that $\theta_{0,n}$ and θ_n differ by k pair-exchanges. Then there exists a test function $\phi_n : \mathscr{X}_n \to [0,1]$ such that,*

$$P_{\theta_0,n}\phi_n(X^n) + P_{\theta,n}(1 - \phi_n(X^n)) \leq a_{n,k},$$

with testing power,

$$a_{n,k} = \big(1 - p_n - q_n + 2p_n q_n + 2\sqrt{p_n(1-p_n)}\sqrt{q_n(1-q_n)})\big)^{2k(n-k)}.$$

Proof. The likelihood ratio test $\phi_n(X^n)$ has testing power bounded by the so-called Hellinger transform,

$$P_{\theta_0,n}\phi_n(X^n) + P_{\theta,n}(1-\phi_n(X^n)) \le \inf_{0\le\alpha\le 1} P_{\theta_0,n}\Big(\frac{p_{\theta,n}}{p_{\theta_0,n}}(X^n)\Big)^{\alpha},$$

(see, *e.g.* [179] and proposition 2.6 in [157]). Using $\alpha = 1/2$ (which is the minimum), we find that,

$$P_{\theta_0,n}\left(\frac{p_{\theta,n}}{p_{\theta_0,n}}(X^n)\right)^{\frac{1}{2}} = P_{\theta_0,n}\left(\frac{p_n}{1-p_n}\frac{1-q_n}{q_n}\right)^{\frac{1}{2}(T_n-S_n)} = Pe^{\frac{1}{2}\lambda_n S_n}\, Pe^{-\frac{1}{2}\lambda_n T_n}$$

where $\lambda_n := \log(1-p_n) - \log(p_n) + \log(q_n) - \log(1-q_n)$ and (S_n, T_n) are distributed binomially, as in the first case of (11.17). Using the moment-generating function of the binomial distribution, we conclude that,

$$\begin{aligned}
&P_{\theta_0,n}\left(\frac{p_{\theta,n}}{p_{\theta_0,n}}(X^n)\right)^{1/2} \\
&\quad= \left(\left(1-p_n+p_n\Big(\frac{1-p_n}{p_n}\frac{q_n}{1-q_n}\Big)^{1/2}\right)\right. \\
&\qquad\qquad \left.\times\left(1-q_n+q_n\Big(\frac{p_n}{1-p_n}\frac{1-q_n}{q_n}\Big)^{1/2}\right)\right)^{2k(n-k)} \\
&\quad= \left(\left((1-p_n)+p_n^{1/2}q_n^{1/2}\Big(\frac{1-p_n}{1-q_n}\Big)^{1/2}\right)\right. \\
&\qquad\qquad \left.\times\left((1-q_n)+p_n^{1/2}q_n^{1/2}\Big(\frac{1-q_n}{1-p_n}\Big)^{1/2}\right)\right)^{2k(n-k)} \\
&\quad= \Big((1-p_n)(1-q_n)+2\big(p_nq_n(1-p_n)(1-q_n)\big)^{1/2}+p_nq_n\Big)^{2k(n-k)}
\end{aligned}$$

which proves the assertion.

11.5 Uncertainty quantification

The most immediate results on uncertainty quantification are obtained with the help of the results in the previous section: if we know that the sequences (p_n) and (q_n) satisfy requirements like (11.8) or (11.13), so that exact or almost-exact recovery is guaranteed, then a consistent sequence of confidence sets is easily constructed from credible sets, as shown in subsection 11.5.1 and the sizes of these credible sets as well as the sizes of associated confidence sets are controlled. If the sequences (p_n) and (q_n) are unknown, or if we require explicit confidence levels, confidence sets can still be constructed from credible sets under conditions requiring that credible levels grow to one quickly enough. Enlargement of credible sets may be used to mitigate this condition, whenever we are close to the Erdős-Rényi submodel, as discussed in subsection 11.5.2.

Regarding the sizes of credible sets, the most natural way to compile a minimal-order credible set $E_n(X^n)$ in a discrete space like Θ_n, is to calculate the posterior weights $\Pi(\{\theta_n\}|X^n)$ of all $\theta_n \in \Theta_n$, order Θ_n by decreasing posterior weight into a finite sequence $\{\theta_{n,1}, \theta_{n,2}, \ldots, \theta_{n,|\Theta_n|}\}$ and define $E_n(X^n) = \{\theta_{n,1}, \ldots, \theta_{n,m}\}$, for the smallest $m \geq 1$ such that $\Pi(E_n(X^n)|X^n)$ is greater than or equal to the required credible level. To provide guarantees regarding the sizes of credible sets, one would like to show that these $E_n(X^n)$ are of an order that is upper bounded with high probability. (Although it is not so clear what the upper bound should be, ideally.)

Here we shall follow a different path based on the smallest number $k(\theta_n, \eta_n)$ of pair-exchanges between two representations θ'_n and η'_n in Θ'_n of θ_n and η_n respectively, see (11.4). The map $k : \Theta_n \times \Theta_n \to \{0, 1, \ldots, \lfloor n/2 \rfloor\}$ is interpreted in a role similar to that of a metric on larger parameter spaces: the *diameter* $\mathrm{diam}_n(C)$ of a subset $C \subset \Theta_n$ is,

$$\mathrm{diam}_n(C) = \max\big\{k(\theta_n, \eta_n) : \theta_n, \eta_n \in C\big\}.$$

by definition.

11.5.1 Posterior recovery and confidence sets

If the posteriors concentrate amounts of mass on $\{\theta_{0,n}\}$ arbitrarily close to one with growing n, then a sequence of credible sets of a certain, fixed level contains $\theta_{0,n}$ for large enough n. If such posterior concentration occurs with high $P_{\theta_{0,n}}$-probability, then the sequence of credible sets is also an asymptotically consistent sequence of confidence sets.

Theorem 11.5.1. *Let $c_n \in [0,1]$ be given, with $c_n > \varepsilon > 0$ for large enough n. Suppose that the posterior recovers the communities exactly,*

$$\Pi\big(\,\theta = \theta_{0,n} \,\big|\, X^n\,\big) \xrightarrow{P_{\theta_{0,n}}} 1. \tag{11.18}$$

Then any sequence (D_n) of (P_n^Π-almost-sure) credible sets of levels c_n satisfies,

$$P_{\theta_{0,n}}\big(\,\theta_{0,n} \in D_n(X^n)\,\big) \to 1,$$

i.e. (D_n) is a consistent sequence of confidence sets. Credible sets of minimal order/diameter equal $\{\theta_0\}$ with high $P_{\theta_{0,n}}$-probability.

Proof. Note that with uniform priors Π_n, $P_{\theta_{0,n}} \ll P_n^\Pi$ for all $n \geq 1$, so that P_n^Π-almost-surely defined credible sets D_n of credible level at least ε, also satisfy,

$$P_{\theta_{0,n}}\big(\,\Pi(D_n(X^n)|X^n) \geq \varepsilon\,\big) = 1.$$

So if, in addition,

$$P_{\theta_{0,n}}\big(\,\Pi(\{\theta_{0,n}\}|X^n) > 1 - \varepsilon\,\big) \to 1,$$

then $\theta_{0,n} \in D_n(X^n)$ with high $P_{\theta_{0,n}}$-probability. Since all posterior mass is concentrated at $\theta_{0,n}$ with high probability, the $\{\theta_{0,n}\}$ form a sequence of unique credible sets of minimal order (or minimal diameter $k_n = 0$) with confidence levels greater than $\varepsilon > 0$ for large enough n.

In the Kesten-Stigum phase, enlargement of credible sets is sufficient to obtain confidence sets. Recall the definition of the $V_{n,k}(\theta_n)$ in (11.5) (with $\theta_{0,n}$ replaced by θ_n). Given some fixed underlying $\theta_{0,n} \in \Theta_n$, we write $V_{n,k}$ for $V_{n,k}(\theta_{0,n})$. Making a certain choice for the upper bounds $k_n \geq 1$, we arrive at,

$$B_n(\theta_n) = \bigcup_{k=0}^{k_n} V_{n,k}(\theta_n), \tag{11.19}$$

for every $n \geq 1$ and $\theta_n \in \Theta_n$. Similar as for $V_{n,k}$ we write B_n for $B_n(\theta_{0,n})$. Given a subset D_n of Θ_n, the set $C_n \subset \Theta_n$ associated with D_n under $B_n(\theta_n)$ (see definition 7.7.3) then is the set of $\theta_n \in \Theta_n$ whose k-distance from some element of D_n is at most k_n,

$$C_n = \{\theta_n \in \Theta_n : \exists_{\eta_n \in D_n}, k(\eta_n, \theta_n) \leq k_n\},$$

the k_n-enlargement of D_n. If we know that the sequences (p_n) and (q_n) satisfy requirement (11.13), posterior concentration occurs around $\{\theta_{0,n}\}$ in 'balls' of diameters $2k_n$ with growing n, and there exist credible sets D'_n of levels greater than $1/2$ and of diameters $2k_n$ centred on $\theta_{0,n}$. The credible sets D_n of *minimal diameters* of any level greater than $1/2$ must intersect D_n. Then the k_n-enlargements C_n of the D_n contain $\theta_{0,n}$.

Theorem 11.5.2. *Suppose that the posterior recovers communities with error rate* (k_n),

$$\Pi\big(k(\theta_n, \theta_{0,n}) \leq k_n \,\big|\, X^n\big) \xrightarrow{P_{\theta_{0,n}}} 1.$$

Let $c_n \in [0,1]$ *be given, with* $c_n > \varepsilon > 0$ *for large enough n and let* (D_n) *denote a sequence of* (P_n^Π*-almost-sure) credible sets of levels* c_n*. Then the* k_n*-enlargements* $C_n(X^n)$ *of the* $D_n(X^n)$ *satisfy,*

$$P_{\theta_{0,n}}\big(\theta_{0,n} \in C_n(X^n)\big) \to 1,$$

i.e. the k_n*-enlargements* (C_n) *form a consistent sequence of confidence sets. If the sets* D_n *have* minimal diameters*, then* $\mathrm{diam}_n(D_n(X^n)) \leq 2k_n$ *and* $\mathrm{diam}_n(C_n(X^n)) \leq 4k_n$ *with high* $P_{\theta_{0,n}}$*-probability.*

Proof. As in the proof of theorem 11.5.1, P_n^Π-almost-surely defined credible sets D_n of credible level at least c_n also satisfy,

$$P_{\theta_{0,n}}\big(\Pi(D_n(X^n)|X^n) \geq c_n\big) = 1.$$

Convergence of the posterior implies that with growing n, the balls $B_n(\theta_{0,n})$ of radii k_n centred on $\theta_{0,n}$ contain an arbitrarily large fraction of the total posterior mass, so

assuming that n is large enough, $c_n > \varepsilon > 0$ and $\Pi(B_n(\theta_{0,n})|X^n) > 1-\varepsilon$ with high $P_{\theta_{0,n}}$-probability. Conclude that,

$$B_n(\theta_{0,n}) \cap D_n(X^n) \neq \varnothing,$$

with high $P_{\theta_{0,n}}$-probability, which amounts to asymptotic coverage of $\theta_{0,n}$ for the k_n-enlargement $C_n(X^n)$ of $D_n(X^n)$. Now fix $n \geq 1$. For every $\theta_n \in \Theta_n$ and every $x^n \in \mathscr{X}_n$, let $k_n(\theta_n, x^n)$ denote the minimal radius of balls B in Θ_n centred on θ_n of posterior mass $\Pi(B|x^n) \geq c_n$. Let $\hat{\theta}_n(x^n) \in \Theta_n$ be such that,

$$k_n(\hat{\theta}_n(x^n)) = \min\{k_n(\theta_n, x^n) : \theta_n \in \Theta_n\},$$

i.e. the centre point of a *smallest* level-c_n credible ball in Θ_n. To conclude, note $k_n(\hat{\theta}_n(X^n)) \leq k_n$ with high $P_{\theta_{0,n}}$-probability and if the $D_n(X^n)$ are of minimal diameters, then they are contained in $k_n(\hat{\theta}_n(X^n))$-balls centred on some $\hat{\theta}_n(X^n)$.

11.5.2 Confidence sets directly from credible sets

To use theorems 11.5.1 or 11.5.2, the statistician needs to know that the sequences (p_n) and (q_n) satisfy (11.8) or (11.13), basically to satisfy the testing condition (11.7). Particularly, condition (11.15) is *not* strong enough to apply theorem 11.5.2. But even if that knowledge is not available and testing cannot serve as a condition, the use of credible sets as confidence sets remains valid, as long as credible levels grow to one fast enough. The following proposition also provides lower bounds for confidence levels of credible sets. (Write $b_n = |\Theta_n|^{-1} = (\frac{1}{2}\binom{2n}{n})^{-1}$.)

Proposition 11.5.3. *Let $\theta_{0,n}$ in Θ_n with uniform priors Π_n, $n \geq 1$, be given and let D_n be a sequence of credible sets, such that,*

$$\Pi(D_n(X^n)|X^n) \geq 1 - a_n,$$

for some sequence (a_n) with $a_n = o(b_n)$. Then,

$$P_{\theta_{0,n}}\big(\theta_0 \in D_n(X^n)\big) \geq 1 - b_n^{-1} a_n.$$

Proof. If $\theta_{0,n} \notin D_n(X^n)$ then $\Pi(\{\theta_{0,n}\}|X^n) \leq a_n$, P_n^{Π}-almost-surely. Then,

$$\begin{aligned} P_{\theta_{0,n}}\big(\theta_0 \in \Theta \setminus D_n(X^n)\big) &= P_n^{\Pi|\{\theta_0\}}\big(\theta_0 \in \Theta \setminus D_n(X^n)\big) \\ &= b_n^{-1} \int_{\{\theta_{0,n}\}} P_{\theta,n}\big(\theta_0 \in \Theta \setminus D_n(X^n)\big)\, d\Pi_n(\theta) \\ &= b_n^{-1} P_n^{\Pi}\big(1\{\theta_0 \in \Theta_n \setminus D_n(X^n)\}\, \Pi(\{\theta_{0,n}\}|X^n)\big) \leq b_n^{-1} a_n, \end{aligned}$$

by Bayes's Rule (A.4).

Note that only the b_n is specific to the planted bi-section model; the proposition as stated holds with any discrete (Θ_n) with a uniform prior.

But as we have seen in chapter 7, remote contiguity enables conversion of sequences of credible sets to asymptotic confidence sets more generally, as in theorem 7.7.4. Until this point, theorem 7.7.4's sets B_n are simply chosen as singletons,

$$B_n(\theta_n) = \{\theta_n\},$$

for every $n \geq 1$ and every $\theta_n \in \Theta_n$, so that the confidence sets C_n associated with any credible sets $D_n \subset \Theta_n$ under B_n are simply *equal to* D_n. In that case, $P_{\theta_{0,n}} \lhd c_n^{-1} P_n^{\Pi|B(\theta_0)}$ for any rate (c_n), $c_n \downarrow 0$, so all sequences $a_n = o(b_n)$ are permitted. Since the prior mass in $B_n(\theta_{0,n})$ is fixed, theorem 7.7.4 says that, if we have a sequence of credible sets $D_n(X^n) \subset \Theta_n$ of high enough credible levels $1 - a_n$, then these $D_n(X^n)$ are also asymptotically consistent confidence sets (see proposition 11.5.3).

Next we consider bigger sets $B_n = B_n(\theta_{0,n})$ like in (11.19); there are two competing influences when enlarging: on the one hand, the prior masses $b_n = \Pi_n(B_n(\theta_{0,n}))$ become larger, relaxing the lower bounds for credible levels. On the other hand, enlargement leads to likelihood ratios with random fluctuations that take them further away from one, thus interfering with notions like contiguity and remote contiguity. Whether proposition 11.5.3 is useful and whether enlargement of credible sets helps, depends on the sequences (p_n) and (q_n). We shall consider the 'statistical phase' where distinctions between within-class and between-class edges become less-and-less pronounced:

$$p_n - q_n = o\big(n^{-1}\big), \tag{11.20}$$

while satisfying also the condition that,

$$p_n^{1/2}(1-p_n)^{1/2} + q_n^{1/2}(1-q_n)^{1/2} = o\big(n|p_n - q_n|\big). \tag{11.21}$$

In this regime, $p_n, q_n \to 0$ or $p_n, q_n \to 1$. If $p_n, q_n \to 0$ as in the sparse phases, (11.21) amounts to,

$$n(p_n^{1/2} - q_n^{1/2}) \to \infty, \tag{11.22}$$

so differences between p_n and q_n may not converge to zero too fast. (Note however that extreme sparsity levels of order $p_n, q_n \propto n^{-\gamma}$ with $1 < \gamma < 2$ are allowed.) For the following lemma we define,

$$\rho_n = \min\left\{ \left(\frac{1-p_n}{p_n}\frac{q_n}{1-q_n}\right), \left(\frac{p_n}{1-p_n}\frac{1-q_n}{q_n}\right) \right\} = e^{-|\lambda_n|}.$$

where $\lambda_n := \log(1-p_n) - \log(p_n) + \log(q_n) - \log(1-q_n)$, and,

$$\alpha_n = \int 2k(\theta_{0,n}, \theta_n)(n - k(\theta_{0,n}, \theta_n))\, d\Pi_n(\theta_n|B_n) = \frac{1}{|B_n|}\sum_{k=0}^{k_n} \binom{n}{k}^2 2k(n-k)$$

with the following rate for remote contiguity (see definition 7.2.1):

$$d_n = \rho_n^{C\alpha_n|p_n - q_n|}, \tag{11.23}$$

for some $C > 1$.

Lemma 11.5.4. *Let (k_n) be given and assume that (11.21) holds. Then for any $\theta_{0,n} \in \Theta_n$,*

$$P_{\theta_{0,n}} \lhd d_n^{-1} P_n^{\Pi|B},$$

with $B_n = B_n(\theta_{0,n})$ like in (11.19).

Proof. Let (k_n) and $\theta_{0,n} \in \Theta_n$ be given. We denote $P_n = P_n^{\Pi|B}$, $Q_n = P_{\theta_{0,n}}$ and apply Jensen's inequality to obtain,

$$\begin{aligned}\frac{dP_n}{dQ_n}(X^n) &= \frac{1}{|B_n|}\sum_{\theta_n \in B_n}\left(\frac{1-p_n}{p_n}\,\frac{q_n}{1-q_n}\right)^{S_n(\theta_n)-T_n(\theta_n)} \\ &\geq \exp\Big(\frac{\lambda_n}{|B_n|}\sum_{\theta_n \in B_n}\big(S_n(\theta_n) - T_n(\theta_n)\big)\Big)\end{aligned}$$

where $(S_n(\theta_n), T_n(\theta_n))$ is distributed as in (11.17). By invariance of the sum under permutations of the vertices, we re-sum as follows for any $k \geq 1$,

$$\frac{1}{|V_{n,k}|}\sum_{\theta_n \in V_{n,k}} S_n(\theta_n) = \frac{2k(n-k)}{n(n-1)} S_n,\ \frac{1}{|V_{n,k}|}\sum_{\theta_n \in V_{n,k}} T_n(\theta_n) = \frac{2k(n-k)}{n^2} T_n,$$

where, with the notation $Z_n = Z(\theta'_{0,n}) \subset \{1,\dots,2n\}$, for a certain representation $\theta'_{0,n}$ of $\theta_{0,n}$, for the zero elements of $\theta'_{0,n}$,

$$\begin{aligned}S_n &= \sum_{i,j \in Z_n} X_{ij} + \sum_{i,j \in Z_n^c} X_{ij} \sim \mathrm{Bin}(n(n-1), p_n), \\ T_n &= \sum_{i \in Z_n,\, j \in Z^c} X_{ij} + \sum_{i \in Z_n^c,\, j \in Z} X_{ij} \sim \mathrm{Bin}(n^2, q_n)\end{aligned}$$

which gives us the upper bound,

$$\frac{dP_n}{dQ_n}(X^n) \geq \rho_n^{\sum_{k=0}^{k_n} 2k(n-k)\frac{|V_{n,k}|}{|B_n|}|\bar{S}_n - \bar{T}_n|} = \rho_n^{\alpha_n|\bar{S}_n - \bar{T}_n|},$$

where $\bar{S}_n = S_n/(n(n-1))$ and $\bar{T}_n = T_n/n^2$. By the central limit theorem,

$$\left(\frac{n(\bar{S}_n - p_n)}{p_n^{1/2}(1-p_n)^{1/2}}, \frac{n(\bar{T}_n - q_n)}{q_n^{1/2}(1-q_n)^{1/2}}\right) \xrightarrow{Q_n\text{-w.}} N(0,1) \times N(0,1),$$

which implies that for every $\varepsilon > 0$ there exists an $M > 0$ such that,

$$\sup_{n \geq 1} Q_n\left(\frac{n(\bar{S}_n - p_n)}{p_n^{1/2}(1-p_n)^{1/2}} \vee \frac{n(\bar{T}_n - q_n)}{q_n^{1/2}(1-q_n)^{1/2}} > M\right) < \varepsilon$$

Conclude that,

$$\sup_{n\geq 1} Q_n\left(\left(\frac{dP_n}{dQ_n}\right)^{-1} \leq \rho_n^{-\alpha_n\left(\frac{M}{n}(p_n^{1/2}(1-p_n)^{1/2}+q_n^{1/2}(1-q_n)^{1/2})+|p_n-q_n|\right)}\right) \geq 1-\varepsilon.$$

Note that the term in the exponent proportional to M is dominated by $|p_n - q_n|$ by (11.21). Hence for every $C > 1$ and every $\varepsilon > 0$,

$$Q_n\left(\left(\frac{dP_n}{dQ_n}(X^n)\right)^{-1} \leq \rho_n^{-C\alpha_n|p_n-q_n|}\right) \geq 1-\varepsilon,$$

for large enough n. Based on Prokhorov's theorem, conclude that $P_{\theta_{0,n}} \lhd d_n^{-1} P_n^{\Pi|B}$.

This amounts to a proof for the following theorem (immediate from theorem 7.7.4).

Theorem 11.5.5. *Let (k_n) be given and assume that (p_n) and (q_n) satisfy (11.20) and (11.21). Let $\theta_{0,n}$ in Θ_n with uniform priors Π_n be given and let D_n be a sequence of credible sets of credible levels $1-a_n$, for some sequence (a_n) such that $b_n^{-1}a_n = o(d_n)$. Then the sets C_n, associated with D_n under B_n as in (11.19) satisfy,*

$$P_{\theta_{0,n}}\big(\,\theta_0 \in C_n(X^n)\,\big) \to 1,$$

i.e. the C_n are asymptotic confidence sets.

Consider the possible choices for (a_n) if we assume $k_n = \beta n$ for some fixed $\beta \in (0, \frac{1}{2})$ (as in the proof of corollary 11.3.6), which leads to the type of exponential correction factor in the prior mass sequence b_n that is required to move the restriction on the credible levels $1-a_n$ substantially. First of all, Stirling's approximation gives rise to the following approximate lower bound on the factor between prior mass and prior mass without enlargement:

$$\frac{\Pi_n(B_n)}{\Pi_n(\{\theta_{0,n}\}))} = \sum_{k=0}^{k_n}\binom{n}{k}^2 \geq \binom{n}{k_n}^2 \geq \frac{1}{2\pi n}\frac{1}{\beta(1-\beta)}f(\beta)^n,$$

where $f : (0, \frac{1}{2}) \to (1,4)$ is given by,

$$f(\beta) = (1-\beta)^{-2(1-\beta)}\beta^{-2\beta}.$$

Approximating $\alpha_n \approx 2k_n(n-k_n)$ for large n and using (11.20), we also have,

$$d_n = \rho_n^{C\alpha_n|p_n-q_n|} \approx \rho_n^{2Cn^2\beta(1-\beta)|p_n-q_n|} = e^{-|\lambda_n|o(n)}.$$

So if we assume that $\lambda_n = O(1)$, d_n is sub-exponential and does not play a role for the improvement factor.

Conclude as follows: (let $a_n = o(|\Theta_n|^{-1}) \approx o(4^{-n})$ denote the rates appropriate in proposition 11.5.3 and assume $\lambda_n = O(1)$) if we have credible sets $D_n(X^n)$ of

credible levels $1 - a_n f(\beta)^{n(1+o(1))}$, then the sequence of enlarged confidence sets $(C_n(X^n))$, associated with $D_n(X^n)$ through B_n with $k_n = \beta n$, covers the true value of the class assignment parameter with high probability. Credible levels that had to be of order $1 - a_n \approx 1 - o(4^{-n})$ previously, can be of approximate order $1 - o(c^{-n})$ for any $1 < c < 4$ by enlargement by B_n if conditions (11.20) and (11.21) hold; the closer $0 < \beta < \frac{1}{2}$ is to $\frac{1}{2}$, the closer c is to 1.

11.6 Exercises [EMPTY]

Appendix A
Notation, definitions and conventions

Because we take the perspective of a frequentist using Bayesian methods, we are obliged to demonstrate that Bayesian definitions continue to make sense under the assumptions that the data X is distributed according to a true, underlying P_0.

Remark A.0.1. We assume given for every $n \geq 1$, a measurable (sample) space $(\mathscr{X}_n, \mathscr{B}_n)$ and random sample $X^n \in \mathscr{X}_n$, with a model $\mathscr{P}_n$ of probability distributions $P_n : \mathscr{B}_n \to [0,1]$. It is also assumed that there exists an n-independent parameter space Θ with a Hausdorff, completely regular topology $\mathscr{T}$ and associated Borel σ-algebra $\mathscr{G}$, and, for every $n \geq 1$, a bijective model parametrization $\Theta \to \mathscr{P}_n : \theta \mapsto P_{\theta,n}$ such that for every $n \geq 1$ and every $A \in \mathscr{B}_n$, the map $\Theta \to [0,1] : \theta \mapsto P_{\theta,n}(A)$ is measurable. Any prior Π on Θ is assumed to be a Borel probability measure $\Pi : \mathscr{G} \to [0,1]$ and can vary with the sample-size n. (Note: in *i.i.d.* setting, the parameter space Θ is $\mathscr{P}_1$, θ is the single-observation distribution P and $\theta \mapsto P_{\theta,n}$ is $P \mapsto P^n$.) As frequentists, we assume that there exists a 'true, underlying distribution for the data; in this case, that means that for every $n \geq 1$, there exists a distribution $P_{0,n}$ from which the n-th sample X^n is drawn.

Often one assumes that the model is *well-specified*: that there exists a $\theta_0 \in \Theta$ such that $P_{0,n} = P_{\theta_0,n}$ for all $n \geq 1$. We think of Θ as a topological space because we want to discuss estimation as a procedure of sequential, stochastic approximation of and convergence to such a 'true parameter value θ_0. In theorem 9.5.1 and definition 6.1.1 we assume, in addition, that the observations X^n are *coupled*, *i.e.* there exists a probability space $(\Omega, \mathscr{F}, P_0)$ and random variables $X^n : \Omega \to \mathscr{X}_n$ such that $P_0((X^n)^{-1}(A)) = P_{0,n}(X^n \in A)$ for all $n \geq 1$ and $A \in \mathscr{B}_n$.

Definition A.0.2. Given $n, m \geq 1$ and a prior probability measure $\Pi_n : \mathscr{G} \to [0,1]$, define the *n-th prior predictive distribution on* $\mathscr{X}_m$ as follows:

$$P_m^{\Pi_n}(A) = \int_\Theta P_{\theta,m}(A)\, d\Pi_n(\theta), \tag{A.1}$$

for all $A \in \mathscr{B}_m$. If the prior is replaced by the posterior, the above defines the *n-th posterior predictive distribution on* $\mathscr{X}_m$,

$$P_m^{\Pi_n|X^n}(A) = \int_\Theta P_{\theta,m}(A)\, d\Pi(\theta|X^n), \tag{A.2}$$

for all $A \in \mathscr{B}_m$. For any $B_n \in \mathscr{G}$ with $\Pi_n(B_n) > 0$, define also the *n-th local prior predictive distribution on* $\mathscr{X}_m$,

$$P_m^{\Pi_n|B_n}(A) = \frac{1}{\Pi_n(B_n)} \int_{B_n} P_{\theta,m}(A)\, d\Pi_n(\theta), \tag{A.3}$$

as the predictive distribution on $\mathscr{X}_m$ that results from the prior Π_n when conditioned on B_n. If m is not mentioned explicitly, it is assumed equal to n.

The prior predictive distribution $P_n^{\Pi_n}$ is the marginal distribution for X^n in the Bayesian perspective that considers parameter and sample jointly $(\theta, X^n) \in \Theta \times \mathscr{X}_n$ as the random quantity of interest.

Definition A.0.3. Given $n \geq 1$, a *(version of) the posterior* is any map $\Pi(\,\cdot\,|X^n = \cdot\,) : \mathscr{G} \times \mathscr{X}_n \to [0,1]$ such that,

(i) for $B \in \mathscr{G}$, the map $\mathscr{X}_n \to [0,1] : x_n \mapsto \Pi(B|X^n = x_n)$ is $\mathscr{B}_n$-measurable,
(ii) for all $A \in \mathscr{B}_n$ and $V \in \mathscr{G}$,

$$\int_A \Pi(V|X^n)\, dP_n^{\Pi_n} = \int_V P_{\theta,n}(A)\, d\Pi_n(\theta). \tag{A.4}$$

Bayes's Rule is expressed through equality (A.4) and is sometimes referred to as a 'disintegration' (of the joint distribution of (θ, X^n)). If the posterior is a Markov kernel, it is a $P_n^{\Pi_n}$-almost-surely well-defined probability measure on $(\Theta, \mathscr{G})$. But it does not follow from the definition above that a version of the posterior actually exists as a regular conditional probability measure. Under mild extra conditions, regularity of the posterior can be guaranteed: for example, if sample space and parameter space are Polish, the posterior is regular; if the model $\mathscr{P}_n$ is dominated (denote the density of $P_{\theta,n}$ by $p_{\theta,n}$), the fraction of integrated likelihoods,

$$\Pi(V|X^n) = \int_V p_{\theta,n}(X^n)\, d\Pi_n(\theta) \Big/ \int_\Theta p_{\theta,n}(X^n)\, d\Pi_n(\theta), \tag{A.5}$$

for $V \in \mathscr{G}$, $n \geq 1$ defines a regular version of the posterior distribution. (Note also that there is no room in definition (A.4) for X^n-dependence of the prior, so 'empirical Bayes' methods must be based on data Y^n independent of X^n, *i.e.* sample-splitting.)

Remark A.0.4. As a consequence of the frequentist assumption that $X^n \sim P_{0,n}$ for all $n \geq 1$, the $P_n^{\Pi_n}$-almost-sure definition (A.4) of the posterior $\Pi(V|X^n)$ does not make sense automatically [97, **?**]: null-sets of $P_n^{\Pi_n}$ on which the definition of $\Pi(\,\cdot\,|X^n)$ is ill-determined, may not be null-sets of $P_{0,n}$. To prevent this, we impose the domination condition,

$$P_{0,n} \ll P_n^{\Pi_n}, \tag{A.6}$$

for every $n \geq 1$.

To understand the reason for (A.6) in a perhaps more familiar way, consider a dominated model and assume that for certain n, (A.6) is *not* satisfied. Then, using (A.1), we find,

$$P_{0,n}\Big(\int p_{\theta,n}(X^n)\,d\Pi_n(\theta)=0\Big)>0,$$

so the denominator in (A.5) evaluates to zero with non-zero $P_{0,n}$-probability. A sufficient condition for (A.6) is obtained with the help of the topologies $\mathscr{T}_n$ (see also remark 3.6 (2) in Strasser (1985) [236]).

Definition A.0.5. For all $n\geq 1$, let F_n denote the class of all bounded, $\mathscr{B}_n$-measurable $f:\mathscr{X}_n\to\mathbb{R}$. The topology $\mathscr{T}_n$ is the initial topology on $\mathscr{P}_n$ for the functions $\{P\mapsto Pf : f\in F_n\}$.

If we model single-observation distributions $P\in\mathscr{P}$ for an *i.i.d.* sample, the topology $\mathscr{T}_n$ on $\mathscr{P}_n=\mathscr{P}^n$ induces a topology on $\mathscr{P}$ (which we also denote by $\mathscr{T}_n$) for each $n\geq 1$. The union $\mathscr{T}_\infty=\cup_n\mathscr{T}_n$ is an inverse-limit topology that allows formulation of conditions for the existence of consistent estimates that are not only sufficient, but also necessary [173], offering a precise perspective on what is estimable *and what is not* in *i.i.d.* context. The associated strong topology is that generated by total variation (or, equivalently, the Hellinger metric).

For more on these topologies, the reader is referred to Strasser (1985) [236] and to Le Cam (1986) [179]. We note explicitly the following fact, which is a direct consequence of Hoeffding's inequality.

Proposition A.0.6. *(Uniform $\mathscr{T}_n$-tests)*
Consider a model $\mathscr{P}$ of single-observation distributions P for i.i.d. samples $(X_1,X_2,\ldots,X_n)\sim P^n$, $(n\geq 1)$. Let $m\geq 1$, $\varepsilon>0$, $P_0\in\mathscr{P}$ and a measurable $f:\mathscr{X}^m\to[0,1]$ be given. Define $B=\{P\in\mathscr{P} : |(P^m-P_0^m)f|<\varepsilon\}$, and $V=\{P\in\mathscr{P} : |(P^m-P_0^m)f|\geq 2\varepsilon\}$. There exist a uniform test sequence (ϕ_n) such that,

$$\sup_{P\in B}P^n\phi_n\leq e^{-nD},\quad \sup_{Q\in V}Q^n(1-\phi_n)\leq e^{-nD},$$

for some $D>0$.

Proof. The proof is an application of Hoeffding's inequality for the sum $\sum_{i=1}^n f(X_i)$ and is left to the reader.

The topologies $\mathscr{T}_n$ also play a role for condition (A.6).

Proposition A.0.7. *Let (Π_n) be Borel priors on the Hausdorff uniform spaces $(\mathscr{P}_n,\mathscr{T}_n)$. For any $n\geq 1$, if $P_{0,n}$ lies in the $\mathscr{T}_n$-support of Π_n, then $P_{0,n}\ll P_n^{\Pi_n}$.*

Proof. Let $n\geq 1$ be given. For any $A\in\mathscr{B}_n$ and any $U'\subset\Theta$ such that $\Pi_n(U')>0$,

$$P_{0,n}(A)\leq\int P_{\theta,n}(A)\,d\Pi_n(\theta|U')+\sup_{\theta\in U'}|P_{\theta,n}(A)-P_{0,n}(A)|.$$

Let $A\in\mathscr{B}_n$ be a null-set of $P_n^{\Pi_n}$; since $\Pi_n(U')>0$, $\int P_{\theta,n}(A)\,d\Pi_n(\theta|U')=0$. For some $\varepsilon>0$, take U' equal to the $\mathscr{T}_n$-basis element $\{\theta\in\Theta : |P_{\theta,n}(A)-P_{\theta_0,n}(A)|<\varepsilon\}$ to conclude that $P_{\theta_0,n}(A)<\varepsilon$ for all $\varepsilon>0$.

In many situations, priors are Borel for the Hellinger topology, so it is useful to observe that the Hellinger support of Π_n in $\mathscr{P}_n$ is always contained in the $\mathscr{T}_n$-support.

Notation and conventions

l.h.s. and *r.h.s.* refer to left- and right-hand sides respectively. For given probability measures P, Q on a measurable space $(\Omega, \mathscr{F})$, we define the Radon-Nikodym derivative $dP/dQ : \Omega \to [0,\infty)$, P-almost-surely, referring *only* to the Q-dominated component of P, following [179]. We also *define* $(dP/dQ)^{-1} : \Omega \to (0,\infty] : \omega \mapsto 1/(dP/dQ(\omega))$, Q-almost-surely. Given a σ-finite measure μ that dominates both P and Q (*e.g.* $\mu = P+Q$), denote $dP/d\mu = q$ and $dQ/d\mu = p$. Then the measurable map $p/q\,1\{q>0\} : \Omega \to [0,\infty)$ is a μ-almost-everywhere version of dP/dQ, and $q/p\,1\{q>0\} : \Omega \to [0,\infty]$ of $(dP/dQ)^{-1}$. Define total-variational and Hellinger distances by $\|P-Q\| = \sup_A |P(A)-Q(A)|$ and $H(P,Q)^2 = 1/2 \int (p^{1/2}-q^{1/2})^2\, d\mu$, respectively. Given random variables $Z_n \sim P_n$, weak convergence to a random variable Z is denoted by $Z_n \xrightarrow{P_n\text{-w.}} Z$, convergence in probability by $Z_n \xrightarrow{P_n} Z$ and almost-sure convergence (with coupling P^∞) by $Z_n \xrightarrow{P^\infty\text{-a.s.}} Z$. The integral of a real-valued, integrable random variable X with respect to a probability measure P is denoted PX, while integrals over the model with respect to priors and posteriors are always written out in Leibniz's notation. For any subset B of a topological space, $\bar{B}$ denotes the closure, $\mathring{B}$ the interior and ∂B the boundary. Given $\varepsilon > 0$ and a metric space (Θ, d), the covering number $N(\varepsilon, \Theta, d) \in \mathbb{N} \cup \{\infty\}$ is the minimal cardinal of a cover of Θ by d-balls of radius ε. Given real-valued random variables $X_1, \ldots, X_n$, the first order statistic is $X_{(1)} = \min_{1 \le i \le n} X_i$. The Hellinger diameter of a model subset C is denoted $\mathrm{diam}_H(C)$ and the Euclidean norm of a vector $\theta \in \mathbb{R}^n$ is denoted $\|\theta\|_{2,n}$. The cardinal of a set B is denoted $N(B)$. The space of all bounded, real-valued, continuous maps defined on a Hausdorff completely regular space $\mathscr{X}$ is denoted $C^b(\mathscr{X})$.

Appendix B
Measure theory

In this appendix we collect some important notions from measure theory. The goal is not a self-contained presentation but rather to establish the basic definitions and theorems from the theory for reference in the main text. As such, the presentation omits certain existence theorems and many of the proofs of other theorems (although references are given). The focus is strongly on bounded (*e.g.* probability-)measures, in places at the expense of generality. Some background in elementary set-theory and analysis is required. As comprehensive references and sources for all proofs we note Kingman and Taylor (1966) [147], Dudley (1989) [81] and Billingsley (1986) [31], among many others.

B.1 Sets and sigma-algebras

It is assumed that the reader is familiar with the following notions in set theory: *set*, *subset*, *empty set*, *union*, *intersection*, *complement*, *symmetric difference* and *disjointness*. Let Ω be a set. The *powerset* 2^Ω is the collection of all subsets of Ω. A *partition* of Ω is an $\mathscr{A} \subset s^\Omega$ such that $\Omega = \cup_{A\in\mathscr{A}} A$ and $A \cap A' = \varnothing$ for any $A, A' \in \mathscr{A}$ such that $A \neq A'$. Let (A_n) be a sequence of subsets of Ω. We say that (A_n) is *monotone* decreasing (resp. monotone increasing) if $A_{n+1} \subset A_n$ (resp. $A_n \subset A_{n+1}$) for all $n \geq 1$. A monotone decreasing (resp. increasing) sequence (A_n) has a *set-theoretic limit* $\lim A_n$ defined as $\cap_{n\geq 1} A_n$ (resp. $\cup_{n\geq 1} A_n$). For any sequence of subsets (A_n), the sequence $(\cup_{m\geq n} A_m)_{n\geq 1}$ (resp. $(\cap_{m\geq n} A_m)_{n\geq 1}$) is monotone decreasing (resp. increasing) and, accordingly, for any sequence (A_n) we define

$$\limsup A_n = \cap_{n\geq 1} \cup_{m\geq n} A_m, \quad \liminf A_n = \cup_{n\geq 1} \cap_{m\geq n} A_m.$$

The sequence (A_n) is said to converge, if $\limsup A_n = \liminf A_n$.

Definition B.1.1. Let Ω be a non-empty set. A non-empty collection $\mathscr{R}$ of subsets of Ω is called a *ring*, if $\mathscr{R}$ has the following properties.

1. If $A, B \in \mathscr{R}$, then $A \cup B \in \mathscr{R}$,

2. If $A, B \in \mathscr{R}$, then $A \setminus B \in \mathscr{R}$.

Note that any ring $\mathscr{R}$ contains $\varnothing$, and for any $A, B \in \mathscr{R}$, $A \cap B = A \setminus (A \setminus B)$, so rings are also closed under (finite) intersections.

Definition B.1.2. Let Ω be a non-empty set. A non-empty collection $\mathscr{F}$ of subsets of Ω is called a σ-*algebra*, if $\mathscr{F}$ has the following properties.

1. $\varnothing \in \mathscr{F}$,
2. If $A \in \mathscr{F}$, then $\Omega \setminus A \in \mathscr{F}$,
3. If $(A_n) \subset \mathscr{F}$, then $\cup_{n \geq 1} A_n \in \mathscr{F}$.

Example B.1.3. Let X be a topological space. The collection $\mathscr{R}$ of all open sets in X is a ring and so is the collection of all closed sets in X. Neither of these collections is a σ-algebra. Namely, finite intersections and unions of open (resp. closed) sets are open (resp. closed), but countable intersections of open sets are not necessarily open, nor are countable unions of closed sets always closed.

Lemma B.1.4. *If I is any set and $\mathscr{F}_i$, $(i \in I)$ are σ-algebra's of subsets of a non-empty set Ω, then $\cap_{i \in I} \mathscr{F}_i$ is also a σ-algebra.*

Definition B.1.5. A *measurable space* $(\Omega, \mathscr{F})$ consists of a non-empty set Ω and a σ-algebra $\mathscr{F}$ of subsets of Ω.

A subset A of a measurable space $(\Omega, \mathscr{F})$ is called *measurable* if $A \in \mathscr{F}$. It can be shown that a σ-algebra is a *monotone class* which means that if $(A_n) \subset \mathscr{F}$ is a *monotone sequence*, then $\lim A_n \in \mathscr{F}$.

Definition B.1.6. Let Ω be a non-empty set and let $\mathscr{C}$ be a collection of subsets of Ω. The σ-*algebra generated by* $\mathscr{C}$, denoted $\sigma(\mathscr{C})$ is the smallest σ-algebra that contains $\mathscr{C}$. Then,

$$\sigma(\mathscr{C}) = \bigcap \big\{ \Sigma \subset 2^{\Omega} : \mathscr{C} \subset \Sigma,\ \Sigma \text{ is a } \sigma\text{-algebra} \big\}.$$

Definition B.1.7. Let $\mathscr{X}$ be a topological space with topology $\mathscr{T}$. The *Borel* σ-*algebra* is the σ-algebra $\sigma(\mathscr{T})$ generated by the *open* (or closed) sets. The Borel σ-algebra on $\mathscr{X}$ is denoted $\mathscr{B}(\mathscr{X})$ (or simply $\mathscr{B}$ if it is clear what the underlying space $\mathscr{X}$ is).

Example B.1.8. Let $\mathscr{X} = \mathbb{R}$ and let $\mathscr{R}$ be the ring consisting of $\varnothing$ and all finite unions of half-open intervals $(a, b]$ with $a, b \in \mathbb{R}$, $a < b$. Then $\mathscr{R}$ generates the Borel σ-algebra $\mathscr{B}$. Indeed, the same holds if we restrict to half-open intervals $(a, b]$ with rational end-points $a, b \in \mathbb{Q}$. In that case the ring $\mathscr{R}$ has a countable number of elements, and we say that $\mathscr{B}$ is *countably generated*.

B.2 Measures

From here on, let $(\Omega, \mathscr{F})$ denote a measurable space. A *set-function* ν is any mapping $\mathscr{F} \to \mathbb{R}$.

Definition B.2.1. A set-function $\nu : \mathscr{F} \to \mathbb{R}$ is said to be *(finitely) additive* if, for any $k \geq 1$ and any $\mathscr{A} = \{A_1, \ldots, A_k\} \subset \mathscr{F}$ such that $A_i \cap A_j = \varnothing$ for all $1 \leq i < j \leq k$,

$$\nu\Big(\bigcup_{i=1}^{k} A_i\Big) = \sum_{i=1}^{k} \nu(A_i). \tag{B.1}$$

A set-function ν is said to be *countably additive* (or σ*-additive*) if, for any *countable* $\mathscr{A} = \{A_n : n \geq 1\} \subset \mathscr{F}$ such that $A_i \cap A_j = \varnothing$ for all $i, j \geq 1$, $i \neq j$:

$$\nu\Big(\bigcup_{i\geq 1} A_i\Big) = \sum_{i\geq 1} \nu(A_i). \tag{B.2}$$

Definition B.2.2. Given a measurable space $(\Omega, \mathscr{F})$, a set-function $\mu : \mathscr{F} \to \mathbb{R}$ is a *signed measure* if μ is countably additive and μ is a *(positive) measure* if μ is countably additive and $\mu \geq 0$. A (signed) measure with a Borel σ-algebra for a domain is called a *(signed) Borel measure*. If μ is a measure, $(\Omega, \mathscr{F}, \mu)$ is called a *measure space*.

Whenever we refer to a measure, it is implied that the measure is positive; if a measure is signed, this is mentioned explicitly. For the construction of measures, the following theorem is instrumental.

Theorem B.2.3. *(Carathéodory extension)*
Let Ω be a non-empty set and let $\mathscr{R}$ be a ring of of subsets of Ω. Denote by $\mathscr{F}$ the σ-algebra generated by $\mathscr{R}$. If $\hat{\mu} : \mathscr{R} \to \mathbb{R}$ is a measure on $\mathscr{R}$ (that is, if countable additivity holds for any sequence of disjoint sets in $\mathscr{R}$ whose union lies in $\mathscr{R}$), then there exists a measure $\mu : \mathscr{F} \to \mathbb{R}$ that extends $\hat{\mu}$ from $\mathscr{R}$ to $\mathscr{F}$. If $\hat{\mu}$ is σ-finite, then the extension μ is σ-finite and unique.

The Carathéodory extension is used to define a measure on a σ-algebra, by constructing it only on generating ring (where countable additivity is relatively easy to verify) and then to infer existence of its (unique) extension to the full sigma-algebra.

Example B.2.4. Consider the real line $\mathbb{R}$ and the collection of all half-open intervals of the form $(a, b]$, for some $a, b \in \mathbb{R}$ with $a < b$. Taking all finite unions and complements, we generate a ring $\mathscr{R}$ of subsets of $\mathbb{R}$. It is easily seen that the σ-algebra generated by $\mathscr{R}$ coincides with the *Borel* σ*-algebra* $\mathscr{B}$ for $\mathbb{R}$ (with its usual (norm) topology). On $\mathscr{R}$ we may define, $\hat{\mu}(\varnothing) = 0$,

$$\hat{\mu}\big((a, b]\big) = b - a,$$

extend $\hat{\mu}$ by finite additivity to all finite unions, and by $\hat{\mu}(C) = \infty$ for any complements C thereof. One verifies easily that the resulting set function $\hat{\mu} : \mathscr{R} \to [0,\infty]$ is positive, countably additive and σ-finite, and the Carathéodory extension guarantees existence of a unique positive, σ-finite Borel measure $\mu : \mathscr{B} \to [0,\infty]$ that is usually called *Lebesgue measure* on $\mathbb{R}$.

The extension theorem holds only for positive measures, but signed measures can be defined like this as well, if we decompose them into positive and negative parts, as per the following definition.

Definition B.2.5. Let $(\mathscr{Y},\mathscr{F})$ be a measurable space. Given a signed measure $\nu : \mathscr{F} \to \mathbb{R}$, the *total-variation norm* of ν is defined as follows. We decompose ν into positive and negative parts $\nu = \nu_+ - \nu_-$ uniquely, where ν_+, ν_- are positive measures (the so-called *Hahn-Jordan decomposition*). Then we define the *total variation measure* as the positive measure $|\nu| = \nu_+ + \nu_-$ and assign $\|\nu\| = |\nu|(\mathscr{Y})$. The signed measure ν is said to be *bounded* if its total variation is finite. A signed measure $\nu : \mathscr{F} \to \mathbb{R}$ is said to be σ*-finite* if there exists a measurable countable partition (A_n) of Ω such that $|\nu|(A_n) < \infty$ for all $n \geq 1$. A positive measure ν such that $\|\nu\| = \nu(\mathscr{Y}) = 1$ is a *probability measure*. Then the triple $(\mathscr{Y},\mathscr{B},\nu)$ is called a *probability space*.

The Hann-Jordan decomposition adds that for any signed measure ν, there exist $A_+, A_- \in \mathscr{B}$ such that $A_+ \cap A_- = \varnothing$, and $\nu_+(B) = \nu(B \cap A_+)$, $\nu_-(B) = -\nu(B \cap A_-)$, for any $B \in \mathscr{B}$. The map $\nu \mapsto \|\nu\|$ defines a *norm* on the linear space of all bounded, signed measure and for any two probability measures P, Q, the *total-variational distance* (or *total-variational metric*) is defined as,

$$\|P - Q\|_{TV} = \sup_{A \in \mathscr{F}} |P(A) - Q(A)| = \tfrac{1}{2}\|P - Q\|.$$

(Note the norm-like notation for this metric and the relative factor 2 which can lead to confusion. A note of caution: for bounded, signed measures μ, ν, $\|\mu - \nu\|$ does *not equal* $2\sup_{A \in \mathscr{F}} |\mu(A) - \nu(A)|$ in general.) A *null-set* of a measure μ on $(\mathscr{Y},\mathscr{F})$ is an $A \in \mathscr{F}$ such that $\mu(A) = 0$. If a property holds for all points in $\mathscr{Y}$, except in a null-set $A \subset \mathscr{Y}$ of a measure μ, we say that the property holds *(μ-)almost-everywhere* (notation: μ-a.e.) or, if μ is a probability measure, *(μ-)almost-surely*) (notation: μ-a.s.). For any two positive measures μ and ν on $(\mathscr{Y},\mathscr{F})$, we say that μ *dominates* ν (notation: $\nu \ll \mu$), if $\mu(A) = 0$ implies $\nu(A) = 0$ for all $A \in \mathscr{F}$. We say that μ and ν are *(mutually) singular* (notation: $\mu \perp \nu$), if there exists a measurable partition $\{A, B\}$ of $\mathscr{Y}$ such that $\mu(C) = 0$ for any measurable $C \subset B$ and $\nu(D) = 0$ for any measurable $D \subset A$.

Proposition B.2.6. *Let $(\Omega,\mathscr{F})$ be a measurable space. The collection of all bounded signed measures on $\mathscr{F}$ forms a linear space $\mathscr{M}(\Omega,\mathscr{F})$ which is a Banach space for the total variation norm. The linear subspace of all bounded positive measures on $\mathscr{F}$ is denoted $\mathscr{M}_+(\Omega,\mathscr{F})$, and the space of all probability measures by $\mathscr{M}^1(\Omega,\mathscr{F})$.*

(See definition C.7.2 for the notion of a Banach space.) Observe the notational difference with the space of definition C.8.1. As a result of countable additivity, measures display a form of continuity expressed by the following theorem.

Theorem B.2.7. *Let $(\Omega, \mathscr{F})$ be a measurable space with measure $\mu : \mathscr{F} \to [0,\infty]$. Then,*

(i) for any monotone decreasing sequence (F_n) in $\mathscr{F}$ such that $\mu(F_n) < \infty$ for some n,

$$\lim_{n\to\infty} \mu(F_n) = \mu\Big(\bigcap_{n=1}^{\infty} F_n\Big), \tag{B.3}$$

(ii) for any monotone increasing sequence (G_n) in $\mathscr{F}$,

$$\lim_{n\to\infty} \mu(G_n) = \mu\Big(\bigcup_{n=1}^{\infty} G_n\Big), \tag{B.4}$$

Theorem B.2.7 is sometimes referred to as the *continuity theorem for measures*, because if we view $\cap_n F_n$ as the monotone limit $\lim F_n$, (B.3) can be read as $\lim_n \mu(F_n) = \mu(\lim_n F_n)$, expressing continuity from below. Similarly, (B.4) expresses continuity from above. Note that theorem B.2.7 does *not* guarantee continuity for arbitrary sequences in $\mathscr{F}$. It should also be noted that theorem B.2.7 is presented here in simplified form: the full theorem states that continuity from below is equivalent to countable additivity of μ (for a more comprehensive formulation and a proof of theorem B.2.7, see [147], theorem 3.2).

Example B.2.8. Let Ω be a discrete set and let $\mathscr{F}$ be the *powerset* 2^Ω of Ω, *i.e.* $\mathscr{F}$ is the collection of *all* subsets of Ω. The *counting measure* $n : \mathscr{F} \to [0,\infty]$ on $(\Omega, \mathscr{F})$ is defined simply to count the number $n(F)$ of points in $F \subset \Omega$. If Ω contains a finite number of points, n is a bounded measure; if Ω contains a countably infinite number of points, n is σ-finite. The counting measure is countably additive.

Example B.2.9. We consider $\mathbb{R}$ with any σ-algebra $\mathscr{F}$ (for example the power-set $2^{\mathbb{R}}$), let $x \in \mathbb{R}$ be given and define the measure $\delta_x : \mathscr{F} \to [0,1]$ by,

$$\delta_x(A) = 1\{x \in A\},$$

for any $A \in \mathscr{F}$. The probability measure δ_x is called the *Dirac measure* (or *delta measure*, or *atomic measure*) degenerate at x and it concentrates all its mass in the point x. Clearly, δ_x is bounded and countably additive. Convex combinations of Dirac measures, *i.e.* measures of the form

$$P = \sum_{j=1}^{m} p_j \delta_{x_j}, \tag{B.5}$$

for some $m \geq 1$ (where $m = \infty$ is permitted) with $(p_1, \ldots, p_m) \in S_m$ (see (1.4)) and any $x_1, \ldots, x_m \in \mathbb{R}$, can be used as a statistical model for an observation X that take values in a discrete (but unknown) subset $\{x_1, \ldots, x_m\}$ of $\mathbb{R}$. The resulting model is not dominated. For later reference, we introduce the set of all *discrete probability measures* (or *purely atomic probability measures*) $D(\mathbb{R}) = \{P : P = \sum_{j=1}^{\infty} p_j \delta_{x_j}\}$ for sequences $(x_j) \subset \mathbb{R}$ and $(p_j) \subset [0,1]$ such that $\sum_{j=1}^{\infty} p_j = 1$.

Example B.2.10. In the context of *i.i.d.* samples of data $X_1, \dots, X_n \in \mathscr{X}$ distributed according to the product distribution P_0^n, an obvious estimator for the single-observation distribution $P_0 \in \mathscr{M}^1(\mathscr{X}, \mathscr{B})$ is the so-called *empirical distribution*,

$$\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i},$$

which is of the form (B.5). By the *law of large numbers*, for any P_0-integrable function $f : \mathscr{X} \to \mathbb{R}$,

$$\mathbb{P}_n f(X) = \frac{1}{n} \sum_{i=1}^{n} f(X_i) \xrightarrow{P_0\text{-a.s.}} P_0 f(X).$$

So $\mathbb{P}_n$ converges in the Le Cam-Schwartz topology (see definition C.9.5) to the true distribution of a single-observation form the sample almost-surely. To study this convergence more closely, we consider the *central limit theorem*, which guarantees that, for every P_0-square-integrable $f : \mathscr{X} \to \mathbb{R}$

$$\sqrt{n}\big(\mathbb{P}_n f(X) - P_0 f(X)\big) \xrightarrow{P_0\text{-w.}} N(0, \sigma^2(f)),$$

where $\sigma^2(f) = P_0(f(X) - P_0 f(X))^2$.

Often, one has a sequence of events (A_n) and one is interested in the probability of a limiting event A, for example the event that A_n occurs infinitely often. The following lemmas pertain to this situation.

Lemma B.2.11. *(First Borel-Cantelli lemma)*
Let $(\Omega, \mathscr{F}, P)$ be a probability space with a sequence $(A_n) \subset \mathscr{F}$ and denote $A = \limsup A_n$. If $\sum_{n\geq 1} P(A_n) < \infty$, then $P(A) = 0$.

In the above lemma, the sequence (A_n) is general. To draw the converse conclusion, the sequence needs to exist of *independent* events: $A, B \in \mathscr{F}$ are said to be *independent* under P if $P(A \cap B) = P(A)\,P(B)$.

Lemma B.2.12. *(Second Borel-Cantelli lemma)*
Let $(\Omega, \mathscr{F}, P)$ be a probability space and let $(A_n) \subset \mathscr{F}$ be independent and denote $A = \limsup A_n$. If

$$\sum_{n\geq 1} P(A_n) = \infty,$$

then $P(A) = 1$.

Together, the Borel-Cantelli lemmas assert that for a sequence of independent events (A_n), $P(A)$ equals zero or one, according as $\sum_n P(A_n)$ converges or diverges. As such, this corollary is known as a *zero-one law*.

To conclude this section, we consider a property of random vectors called exchangeability.

Definition B.2.13. A random vector $(X_1, \dots, X_n) \in \mathbb{R}^n$ with distribution P_n is said to be *exchangeable*, if, for any permutation π of $\{1, \dots, n\}$, the random vector $(X_{\pi(1)}, \dots, X_{\pi(n)})$ also has distribution P_n.

This property is a generalization of *i.i.d.*-ness: note that if $(X_1,\dots,X_n) \sim P_0^n$ then $(X_1,\dots,X_n)$ is exchangeable. The converse does not hold but exchangeable distributions can be characterized in terms of *i.i.d.* distributions, as the following result demonstrates. In the following theorem, the space $\mathscr{M}^1(\mathbb{R},\mathscr{B})$ is endowed with the Borel σ-algebra corresponding to Prokhorov's weak topology (see definition C.8.8), which makes all functions $P \mapsto P(A)$, $(A \in \mathscr{B})$, measurable.

Theorem B.2.14. *(De Finetti's theorem) The random vector $(X_1,\dots,X_n) \in \mathbb{R}^n$ distributed according to a probability measure P_n is exchangeable* if and only if *there exists a unique Borel probability measure Π on the collection $\mathscr{M}^1(\mathbb{R},\mathscr{B})$ of all Borel probability measures on $\mathbb{R}$ such that,*

$$P_n(A_1 \times \ldots \times A_n) = \int_{\mathscr{M}(\mathbb{R})} \prod_{i=1}^{n} P(A_i)\, d\Pi(P),$$

for all $A_1,\dots,A_n \in \mathscr{B}(\mathbb{R})$.

B.3 Measurability, random variables and integration

In this section we consider random variables and their expectation values. Throughout this section, let $(\Omega,\mathscr{F},P)$ denote a probability space.

Definition B.3.1. Given a map $X : A \to B$ and a subset $C \subset B$, the *pre-image* of C under X, is defined as,

$$X^{-1}(C) = \{a \in A : X(a) \in C\} \subset A.$$

Given two measurable spaces $(\Omega,\mathscr{F})$ and $(\mathscr{X},\mathscr{B})$, a map $X : \Omega \to \mathscr{X}$ is called *measurable* if, for all $B \in \mathscr{B}$, $X^{-1}(B) \in \mathscr{F}$. These subsets form a sub-σ-algebra $\sigma(X) = \{X^{-1}(B) : B \in \mathscr{B}\}$ called the *σ-algebra generated by X*.

Essentially, measurability makes it possible to speak of "the probability that X lies in B":

$$P(X \in B) = P\big(\{\omega \in \Omega : X(\omega) \in B\}\big),$$

is well-defined only if $X^{-1}(B)$ belongs to the domain of P. Specializing to real-valued measurable maps, it follows from elementary manipulation of set-limits that suprema of sequences of measurable maps are again measurable. This statement can be framed in the following central theorem in measure theory.

Theorem B.3.2. *(Monotone class theorem) For every $n \geq 1$, let $f_n : \Omega \to \overline{\mathbb{R}}$ be measurable and assume that $f_{n+1}(\omega) \geq f_n(\omega)$ for all $n \geq 1$ and $\omega \in \Omega$. Then $f(\omega) = \lim_{n\to\infty} f_n(\omega)$ defines a measurable map $f : \Omega \to \overline{\mathbb{R}}$.*

This means that the set of all measurable $f : \Omega \to \overline{\mathbb{R}}$ forms what is know as a *monotone class*, an partially ordered set that is closed for limits over monotone sequences. Although measurability is preserved under linear combinations, the space

of all measurable $f : \Omega \to \overline{\mathbb{R}}$ is *not* a linear space because if, for some $\omega \in \Omega$, $f(\omega) = \infty$ and $g(\omega) = -\infty$, then $(f+g)(\omega) = \infty - \infty$ is ill-defined. No such problems arise when we restrict to the set of all measurable $f \geq 0$, which form a cone. Restriction to measurable $f : \Omega \to \mathbb{R}$, on the other hand, invalidates the monotone class theorem.

Definition B.3.3. Let $(\Omega, \mathscr{F}, P)$ be a probability space. A *random variable* is a measurable map $X : \Omega \to \overline{\mathbb{R}}$ with the property that $P(|X| = \infty) = 0$. Therefore, every random variable can be represented by a real-valued $X' : \Omega \to \mathbb{R}$, up to null-sets of P, *i.e.* $P(X = X') = 1$.

Note that random variables do not form a monotone class (take $f_n = n$), but they do form a linear space. To define expectations (integrals with respect to P), we extend by monotone limit starting from the following definition.

Definition B.3.4. A measurable map $f : \Omega \mapsto \mathbb{R}$ is called *simple* if there exists a $k \geq 1$, a k-set partition $A_1, \ldots, A_k$ of Ω and $a_1, \ldots, a_k \in \mathbb{R}$ such that,

$$f(\omega) = \sum_{i=1}^{k} a_i \, 1_{A_i}(\omega).$$

The *integral* of a simple f with respect to P is defined as,

$$\int f \, dP = \sum_{i=1}^{k} a_i \, P(A_i).$$

A straightforward construction shows that for every measurable $f \geq 0$, there exists an increasing sequence (f_n) of non-negative, simple functions such that $f_n(\omega) \uparrow f(\omega)$ for all $\omega \in \Omega$. By the monotony of (f_n), this defines an integral for every non-negative, measurable f,

$$\int f \, dP = \lim_{n \to \infty} \int f_n \, dP,$$

(after one demonstrates that the *l.h.s.* does not depend on the particular (f_n) we choose to approximate f). Extension to real-valued measurable functions that take on negative values as well is done by treating negative f_- and non-negative f_+ parts of f separately. Extension to $\mathbb{R}^d$ with $d > 1$ proceeds component-wise. The most important result in integration theory is the following elementary theorem.

Theorem B.3.5. *(Monotone convergence) Let (f_n) be a monotone sequence of measurable maps $\Omega \to \mathbb{R}$. Then $\lim_n \int f_n \, dP = \int (\lim_n f_n) \, dP$.*

Before we can state Fatou's lemma and the dominated convergence theorem, we define integrability of measurable maps.

Definition B.3.6. Let $(\Omega, \mathscr{F}, P)$ be a probability space. A real-valued measurable function $f : \Omega \to \mathbb{R}$ is said to be *integrable* with respect to P if

$$\int_{\Omega} |f| \, dP < \infty. \tag{B.6}$$

It follows immediately from the definition that an integrable f is a random variable. Note that *any* sequence of measurable f_n is dominated by the sequence $(\sup_{m\geq n} f_m)$. By the monotone class theorem the suprema are measurable and the resulting sequence of maps is monotone decreasing.

Lemma B.3.7. *(Fatou's lemma) Let $f_n : \Omega \to \overline{\mathbb{R}}$ be a sequence of measurable maps such that $f_n \leq g$, P-almost-surely for all $n \geq 1$, for some P-integrable $g : \Omega \to \mathbb{R}$. Then,*

$$\limsup_{n\to\infty} \int f_n \, dP \leq \int (\limsup_{n\to\infty} f_n) \, dP.$$

An obvious extension provides an inequality for the limes inferior. When combined, the lim sup and lim inf versions of Fatou's lemma imply the following result, known as Lebesgue's (dominated convergence) theorem.

Theorem B.3.8. *(Dominated convergence) Let $f_n : \Omega \to \overline{\mathbb{R}}$ be a sequence of measurable maps such that $\lim_n f_n : \Omega \to \overline{\mathbb{R}}$ exists and $|f_n| \leq g$, P-almost-surely for all $n \geq 1$, for some P-integrable $g : \Omega \to \mathbb{R}$. Then,*

$$\lim_{n\to\infty} \int f_n \, dP = \int (\lim_{n\to\infty} f_n) \, dP.$$

For any two probability spaces $(\Omega_1, \mathscr{F}_1, P_1)$ and $(\Omega_2, \mathscr{F}_2, P_2)$, the set $\Omega_1 \times \Omega_2$ can be endowed with the σ-algebra generated by products of the form $A_1 \times A_2$ where $A_1 \in \mathscr{F}_1$, $A_2 \in \mathscr{F}_2$, which is called the product σ-algebra, denoted $\mathscr{F} = \sigma(\mathscr{F}_1 \times \mathscr{F}_2)$ and a product measure $P = P_1 \times P_2$, to arrive at a probability space $(\Omega, \mathscr{F}, P)$, for which the following elementary theorem on the interchangability of integrals applies.

Theorem B.3.9. *(Fubini's theorem) Let $(\Omega_1, \mathscr{F}_1, P_1)$ and $(\Omega_2, \mathscr{F}_2, P_2)$ be probability spaces and denote their product by $(\Omega, \mathscr{F}, P)$. For any non-negative, $\mathscr{F}$-measurable $f : \Omega \to \overline{\mathbb{R}}$ and any $\omega_1 \in \Omega_1$, $f(\omega_1, \cdot) : \Omega_2 \to \bar{\mathbb{R}}$ is $\mathscr{F}_2$-measurable. Furthermore, for any $A_1 \in \mathscr{F}_1$ and $A_2 \in \mathscr{F}_2$,*

$$\begin{aligned}\int_{A_1\times A_2} f(\omega_1,\omega_2)\, dP(\omega) &= \int_{A_1} \Big(\int_{A_2} f(\omega_1,\omega_2)\, dP_2(\omega_2) \Big) dP_1(\omega_1) \\ &= \int_{A_2} \Big(\int_{A_1} f(\omega_1,\omega_2)\, dP_1(\omega_1) \Big) dP_2(\omega_2).\end{aligned}$$

Another central result from integration theory forms the foundation for the *probability density* we associate with many distributions.

Theorem B.3.10. *(Radon-Nikodym theorem) Let $(\Omega, \mathscr{F})$ be a measurable space and let $\mu, \nu : \mathscr{F} \to [0, \infty]$ be two σ-finite measures on $(\Omega, \mathscr{F})$. There exists a unique decomposition*

$$\mu = \mu_{\|} + \mu_{\perp},$$

such that $\mu_{\|} \ll \nu$ and $\mu_{\perp}$ and ν are mutually singular. Furthermore, there exists a $\mathscr{F}$-measurable function $f : \Omega \to \mathbb{R}$ such that for all $F \in \mathscr{F}$,

$$\mu_{\|}(F) = \int_F f \, d\nu. \tag{B.7}$$

The function f is ν-almost-everywhere unique.

The function $f : \Omega \to \mathbb{R}$ in the above theorem is called the *Radon-Nikodym derivative* of μ with respect to ν. If μ is a probability distribution, then f is called the *(probability) density* for μ with respect to ν. The Radon-Nikodym derivative is sometimes denoted $d\mu/d\nu$. The assertion that f is "ν-almost-everywhere unique" means that if there exists a measurable function $g : \Omega \to \mathbb{R}$ such that (B.7) holds with g replacing f, then $f = g$, (ν-a.e.), *i.e.* f and g may differ only on a set of ν-measure equal to zero. Through a construction involving increasing sequences of simple functions, we see that the Radon-Nikodym theorem has the following implication.

Corollary B.3.11. *Assume that the conditions for the Radon-Nikodym theorem are satisfied. Let $X : \Omega \to [0,\infty]$ be measurable and μ-integrable. Then the product Xf is ν-integrable and*

$$\int X \, d\mu = \int X f \, d\nu.$$

Remark B.3.12. Integrability is not a necessary condition here, but the statement of the corollary becomes rather less transparent if generalized.

B.4 Conditional distributions

In this section, we consider conditioning of probability measures. In first instance, we consider straightforward conditioning on events and illustrate Bayes's rule, but we also cover conditioning on σ-algebras and random variables, to arrive at the posterior distribution and Bayes's rule for densities.

Definition B.4.1. Let $(\Omega, \mathscr{F}, P)$ be a probability space and let $B \in \mathscr{F}$ be such that $P(B) > 0$. For any $A \in \mathscr{F}$, the *conditional probability* of the event A given event B is defined:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{B.8}$$

Conditional probability given B describes a set-function on $\mathscr{F}$ and one easily checks that this set-function is a probability measure that assigns probability one to B. The conditional probability measure $P(\cdot\,|B) : \mathscr{F} \to [0,1]$ can be viewed as the restriction of P to $\mathscr{F}$-measurable subsets of B, normalized to be a probability measure. Definition B.4.1 gives rise to a relation between $P(A|B)$ and $P(B|A)$ (in case both $P(A) > 0$ and $P(B) > 0$, of course), which is called Bayes's Rule.

Proposition B.4.2. (Bayes's Rule)
Let $(\Omega, \mathscr{F}, P)$ be a probability space and let $A, B \in \mathscr{F}$ be such that $P(A) > 0$, $P(B) > 0$. Then

$$P(A|B)\,P(B) = P(B|A)\,P(A).$$

However, only being able to condition on events B of non-zero probability is too restrictive. Moreover, B above is a definite event; it is desirable also to be able to discuss probabilities conditional on events that have not been measured yet, *i.e.* to condition on a whole σ-algebra of events like B above.

Definition B.4.3. Let $(\Omega, \mathscr{F}, P)$ be a probability space, let $\mathscr{C}$ be a sub-σ-algebra of $\mathscr{F}$ and let X be a real-valued P-integrable random variable. The *conditional expectation* of X given $\mathscr{C}$ is any $\mathscr{C}$-measurable random variable $E[X|\mathscr{C}] : \Omega \to \mathbb{R}$ such that,

$$\int_C X\, dP = \int_C E[X|\mathscr{C}]\, dP,$$

for all $C \in \mathscr{C}$,

If we consider some $B \in \mathscr{F}$ with $P(B) > 0$ and the σ-algebra $\sigma_B = \{\varnothing, B, \Omega \setminus B, \Omega\}$, and we consider definition B.4.3 for $X = 1_A$, we recover,

$$E[1_A|\sigma_B] = P(A|B)1_B + P(A|\Omega \setminus B)1_{\Omega \setminus B}.$$

The condition that X be P-integrable is sufficient for existence and uniqueness of $E[X|\mathscr{C}]$ P-almost-surely, the proof being an application of the Radon-Nikodym theorem (see theorem 10.1.1 in Dudley (1989)). So conditional expectations are not unique but if we have two different random variables e_1 and e_2 satisfying the defining conditions for $E[X|\mathscr{C}]$, then $e_1 = e_2$, P-almost-surely. Often, the σ-algebra $\mathscr{C}$ is the σ-algebra $\sigma(Z)$ generated by another random variable Z. In that case we denote the conditional expectation by $E[X|Z]$ and realizations are denoted $E[X|Z = z]$.

Definition B.4.4. Let $(\Omega, \mathscr{F}, P)$ be a probability space and let $\mathscr{C}$ be a sub-σ-algebra of $\mathscr{F}$. Furthermore, let $(\mathscr{Y}, \mathscr{B})$ be a measurable space and let $Y : \Omega \to \mathscr{Y}$ be a random variable. For each $A \in \mathscr{F}$, the *conditional distribution* of Y given $\mathscr{C}$ is defined as follows:

$$P_{Y|\mathscr{C}}(A, \omega) = E[1\{Y \in A\}|\mathscr{C}](\omega), \tag{B.9}$$

P-almost-surely.

Although seemingly innocuous, the fact that conditional expectations are defined only P-almost-surely poses a rather subtle problem: for every $A \in \mathscr{B}$ there exists an A-dependent null-set on which $P_{Y|\mathscr{C}}(A, \cdot)$ is not defined. This is not a problem if we are interested only in A (or in a countable number of sets). But usually, we wish to view $P_{Y|\mathscr{C}}$ as a probability measure, that is to say, it must be well-defined as a *map* on the σ-algebra $\mathscr{B}$ almost-surely. Since most σ-algebras are uncountable, there is no guarantee that the corresponding union of exceptional null-sets has measure zero as well. This means that definition B.4.4 only defines $P_{Y|\mathscr{C}}(A, \cdot)$ *per individual* A, and *not as a map* $A \mapsto P_{Y|\mathscr{C}}(A, \omega)$ for P-almost-all $\omega \in \Omega$. The extra property that the conditional distribution is well-defined P-almost-surely as a map is called *regularity* of the conditional distribution.

Definition B.4.5. If $\pi : \mathscr{B} \times \Omega \to [0, 1]$ is such that,

1. for every $B \in \mathscr{B}$, $\omega \mapsto \pi(B, \omega)$ is $\mathscr{C}$-measurable,

2. there is an $E \in \mathscr{C}$ with $P(E) = 0$ such that for all $\omega \in \Omega \setminus E$, $B \mapsto \pi(B, \omega)$ is a probability measure,
3. for all $C \in \mathscr{C}$,
$$\int_C \pi(B, \omega)\, dP(\omega) = P(B \cap C),$$

then π is said to be a *regular conditional distribution*

The existence of a regular conditional probability cannot be guaranteed without further condtions on the underlying probability space.

Definition B.4.6. A topological space $(S, \mathscr{T})$ is said to be a *Polish space* if $\mathscr{T}$ is metrizable, complete and separable. Any topological space that is the continuous image of a Polish space is called a *Souslin space*; any topological space that is the one-to-one continuous image of a Polish space is called a *Lusin space*.

Polish spaces appear in many subjects in measure theory: the existence of a countable, dense subset in a metric setting allows constructions based *countable* covers by metric balls. In this manner Polish spaces allow countable formulations for properties that would involve uncountable collections of subsets otherwise, in correspondence with countability restrictions arising from measure theory. Such a construction occurs in a theorem that guarantees the existence of regular conditional distributions.

Theorem B.4.7. *Let $(\Omega, \mathscr{F}, P)$ be a probability space and let Θ be a Polish space with Borel σ-algebra $\mathscr{G}$. If $\vartheta : \Omega \to \Theta$ is a Borel measurable random variable taking values in Θ and $\mathscr{C}$ is any sub-σ-algebra of $\mathscr{F}$, there exists a P-almost-surely unique regular conditional distribution $P_{Y|\mathscr{C}}(A, \omega) : \mathscr{G} \times \Omega \to [0, 1]$.*

Proof. For a proof of this theorem, the reader is referred to Dudley (1989) [81], theorem 10.2.2).

So for Bayesian purposes, where $\Omega = \mathscr{Y} \times \Theta$ and we condition on $Y \in \mathscr{Y}$ (by choosing $\mathscr{C} = \sigma(Y)$), Polishness of the parameter space Θ is enough to guarantee the existence of a regular version of the conditional probability for ϑ given Y, the posterior. In cases where the model topology is not Polish, posteriors are not guaranteed to exist as Borel probability measures on the model, unless we impose more.

B.5 Martingale convergence [EMPTY]

B.6 Existence of stochastic processes

A stochastic process has the following broad definition, intended to enable discussion of 'random functions'.

Definition B.6.1. Let $(\Omega, \mathscr{F}, P)$ be a probability space, let T be an arbitrary set. A collection X of $\mathscr{F}$-measurable random variables $X = \{X_t : \Omega \to \mathbb{R} : t \in T\}$ is called a *stochastic process* (or *coupling* of the random variables X_t) indexed by T.

The perspective we assume here, is that one starts with a collection of random quantities with possible dependency, without the certainty that there exists such a *coupling*; the Daniell-Kolmogorov theorem formulates conditions for existence [160].

Indeed, the Daniell-Kolmogorov theorem provides an explicit construction of $(\Omega, \mathscr{F}, P)$. Clearly, if the X_t take their values in a space $\mathscr{X}$, the obvious choice for Ω is the product $\mathscr{X}^T$ in which the process takes its values. The question remains how to characterize P and the σ-algebra $\mathscr{F}$ that forms its domain. Kolmogorov's point of departure is the assumption that for *any finite subset* $S = \{t_1, \ldots, t_k\} \subset T$, the distribution $P_{t_1 \ldots t_k}$ of the k-dimensional stochastic vector $(X_{t_1}, \ldots, X_{t_k})$ in $\mathscr{X}^k$ is given.

Example B.6.2. Choose $T = [0,1]$ and define random quantities $f(t)$ for each $t \in T$, by considering multivariate normal distributions P_S with the properties that $f(0) = 0$, that the expectation $P(f(s) - f(t)) = 0$ for any $s, t \in S$, that $f(s) - f(t)$ is independent of $f(u) - f(v)$ if $s < t \leq u < v$, and that the variance of $f(s) - f(t)$ is proportional to $|s - t|$. If a coupling exists, the resulting stochastic process $(f(t) : t \in [0,1])$ describes random *functions* $f : [0,1] \to \mathbb{R}$. (Such a coupling exists and gives rise to Brownian motion without drift).

Since the distributions $P_{t_1 \ldots t_k}$ are as yet unrelated and given for *all* finite subsets of T, consistency requirements are implicit if they are to serve as marginals to the probability distribution P: if two finite subsets $S_1, S_2 \subset T$ satisfy $S_1 \subset S_2$, then the distribution of $\{X_t : t \in S_1\}$ should be marginal to that of $\{X_t : t \in S_2\}$. Similarly, permutation of the components of the stochastic vector in the above display should be reflected in the respective distributions as well. The requirements for consistency are formulated in two requirements called Daniell-Kolmogorov *consistency conditions*:

(K1) Let $k \geq 1$ and $\{t_1, \ldots, t_{k+1}\} \subset T$ be given. For any $C \in \sigma(\mathscr{B}^k)$,

$$P_{t_1 \ldots t_k}(C) = P_{t_1 \ldots t_{k+1}}(C \times \mathscr{X}),$$

(K2) Let$k \geq 1$, $\{t_1, \ldots, t_k\} \subset T$ and a permutation π of k elements be given. For any $A_1, \ldots, A_k \in \mathscr{B}$,

$$P_{t_{\pi(1)} \ldots t_{\pi(k)}}(A_1 \times \ldots \times A_k) = P_{t_1 \ldots t_k}(A_{\pi^{-1}(1)} \times \ldots \times A_{\pi^{-1}(k)}).$$

Theorem B.6.3. *(Daniell-Kolmogorov existence theorem)*
Let T be any set and let X_t, $t \in T$ be random variables taking values in a Polish space $\mathscr{X}$, with finite-dimensional marginal distributions,

$$(X_{t_1}, \ldots, X_{t_k}) \sim P_{t_1 \ldots t_k}, \tag{B.10}$$

for any $k \geq 1$ and all $t_1, \ldots, t_k \in T$. Suppose that the marginals $P_{t_1 \ldots t_k}$ satisfy conditions (K1) and (K2). Then there exists a probability space $(\Omega, \mathscr{F}, P)$ and Borel-

measurable $X_t : \Omega \to \mathscr{X}$, $t \in T$ such that all distributions of the form (B.10) are marginal to P.

Proof. The proof of this theorem can be found in many places, *e.g.* [160].

Kolmogorov's approach to the definition and characterization of stochastic processes in terms of finite-dimensional marginals is of great practical value: the infinite-dimensional objects in $\mathscr{X}^T$ are somewhat elusive and their distributions are hard to characterize in principle, whereas the finite-dimensional marginals are concrete, explicit objects. The power of the Daniell-Kolmogorov theorem is that it reduces the inaccessible infinite-dimensional issue to a much simpler issue involving an infinity of finite-dimensional building blocks. This makes the analysis accessible and permits calculations regarding the infinite-dimensional objects as (limits of) calculations that are explicit in terms of the finite-dimensional marginal distributions.

The drawback of the construction becomes apparent only upon closer inspection of the domain of P: $\mathscr{F}$ is the σ-algebra generated by the so-called *cylinder sets* that are involved in the finite-dimensional marginals: a cylinder set is of the form,

$$\{X \in \mathscr{X}^T : X_{t_1} \in B_1, \ldots, X_{t_k} \in B_k\},$$

for some $k \geq 1$, some $t_1, \ldots, t_k \in T$ and some Borel sets $B_1, \ldots, B_k$ in $\mathscr{X}$. This implies that $\mathscr{F}$-measurable subsets of $\Omega = \mathscr{X}^T$ restrict at most a countable number of X_t's simultaneously and properties of the random function that involve uncountable subsets of T do not necessarily correspond to $\mathscr{F}$-measurable subsets. In practice, this often leads to topological restrictions on the set T: if T is a second-countable topological space, or compact, then many properties of random functions $f : T \to \mathscr{X}$ (like continuity, differentiability, *etcetera*) can be formulated equivalently as properties that hold only on a dense, countable subset of T. The latter may then be formulated in terms of $\mathscr{F}$-measurable subsets of $\mathscr{X}^T$ while this would not be possible otherwise. Kolmogorov's existence theorem requires consistency but has no other conditions and, hence, it always works; but this general applicability has the downside that it does not give rise to a 'comfortably large' domain for the resulting probability measure P.

Appendix C
Topology

In this appendix we collect some results from topology: in the first sections there is a brief review of elementary point-set topology, followed by a more detailed discussion of inverse limits of topological and uniform spaces, with a review of locally convex spaces. We conclude with a review of the topologies that are used for spaces of signed, positive and probability measures.

C.1 Topological basics

In the first section we concentrate on precise definitions of elements of general point-set topology and give some of their most important properties in the main text, without proofs or examples. For those, the reader is referred to several other sources: a very readable introduction is the first part of Munkres (2000), [200]. A more comprehensive treatise is provided in Bourbaki (1998, 1989), [46, 47].

Definition C.1.1. A *topological space* is a non-empty set $\mathscr{X}$ with a collection $\mathscr{T}$ of subsets of $\mathscr{X}$ such that:

(i) $\varnothing$ and $\mathscr{X}$ are in $\mathscr{T}$,
(ii) the union of any subcollection of $\mathscr{T}$ is in $\mathscr{T}$,
(iii) the intersection of any *finite* subcollection of $\mathscr{T}$ is in $\mathscr{T}$.

The collection $\mathscr{T}$ is called a *topology* on $\mathscr{X}$ The subsets in $\mathscr{T}$ are called *open subsets* and their complements are called *closed subsets*. A subset A of $\mathscr{X}$ that is both open and closed is called a *clopen subset*. In the language of descriptive set theory [143], a clopen set A is said to be of the *first ambiguous class*. A countable intersection of open sets is called a G_δ*-set*; a countable union of closed sets is called an F_σ*-set*; a set that is both G_δ and F_σ is said to be of the *second ambiguous class*. If $x \in W \subset \mathscr{X}$ and there is an open U in $\mathscr{X}$ such that $x \in U \subset W$, then W is called a *neighbourhood* of x. The *interior* $\mathring{A}$ of a subset A of $\mathscr{X}$ is the union of all open sets contained in A. The *closure* $\bar{A}$ of A is the intersection of all closed sets that contain A. The *boundary* ∂A of A is $\bar{A} \setminus \mathring{A}$.

A topology for a set describes, in a very abstract manner, what it means for one point of $\mathscr{X}$ to be 'close' to another. Thus topology forms the abstract foundation for any form of approximation within sets of functions, measures, or other mathematical objects. In the notation of topological spaces, we often omit explicit mention of the topology and write simply $\mathscr{X}$ when it is clear which topology is intended.

Definition C.1.2. Given a set $\mathscr{X}$, a *topological basis* is a collection $\mathscr{U}$ of subsets of $\mathscr{X}$ such that:

(i) for every $x \in \mathscr{X}$, there is at least one $B \in \mathscr{U}$ such that $x \in B$,
(ii) for all $B_1, B_2 \in \mathscr{U}$ and all $x \in B_1 \cap B_2$, there is a $B_3 \in \mathscr{U}$ such that $x \in B_3 \subset B_1 \cap B_2$.

The *topology generated by the basis* $\mathscr{U}$ consists of all unions of sets in the basis $\mathscr{U}$. Any collection of subsets $\mathscr{S}$ that covers $\mathscr{X}$ is a *topological subbasis*, the *basis generated by the subbasis* $\mathscr{S}$ generated by $\mathscr{S}$ consists of all finite intersections of subsets in $\mathscr{S}$ and the *topology generated by the subbasis* $\mathscr{S}$ is the collection of all unions of finite intersections of subsets from $\mathscr{S}$.

Given a topological space $(\mathscr{X}, \mathscr{T})$, any collection $\mathscr{C} \subset \mathscr{T}$ such that for every $x \in \mathscr{X}$ and any open U that contains x, there is a $B \in \mathscr{C}$ such that $x \in B \subset U$, then $\mathscr{C}$ is a basis for the topology $\mathscr{T}$.

Definition C.1.3. If we have a topological space $\mathscr{X}$, a directed set I and, for every $\alpha \in I$, a point $x_\alpha \in \mathscr{X}$, then the subset $\{x_\alpha : \alpha \in I\}$ is called a *net* in $\mathscr{X}$ (denoted (x_α)). If $I = \{1, 2, \ldots\}$, (x_α) is called a *sequence* and usually denoted with (x_n). A net (x_α) is said to *converge* to a point $x \in \mathscr{X}$, if for every neighbourhood U of x, there is an index α such that $\beta \geq \alpha$ implies $x_\beta \in U$.

A more general and sophisticated notion of convergence is provided through the following set-theoretic definition.

Definition C.1.4. Given a set $\mathscr{X}$, a *filter* $\mathscr{F}$ is a collection of subsets of $\mathscr{X}$ with the following properties:

(i) The empty set $\varnothing$ does not belong to $\mathscr{F}$,
(ii) Finite intersections of sets from $\mathscr{F}$ belong to $\mathscr{F}$,
(iii) If $A \in \mathscr{F}$ and $A \subset B$, then $B \in \mathscr{F}$.

A collection $\mathscr{U}$ of subsets of $\mathscr{X}$ form a *filter basis* for $\mathscr{F}$, if $\mathscr{U} \subset \mathscr{F}$ and every $A \in \mathscr{F}$ contains some $B \in \mathscr{U}$.

A collection $\mathscr{U}$ of subsets of a set $\mathscr{X}$ forms the basis of a filter if $\varnothing \notin \mathscr{U}$ and for any $B_1, B_2 \in \mathscr{U}$, there is a $B_3 \in \mathscr{U}$ such that $B_3 \subset B_1 \cap B_2$. For example, given a set $\mathscr{X}$ with a net (x_α) (resp. a sequence (x_n)), the collection of all *tails* $\{x_\beta : \beta \geq \alpha\}$ for some $\alpha \in I$ (resp. $\{x_N, x_{N+1}, \ldots\}$ for some $N \geq 1$), of (x_α) (resp. of (x_n)) form a filter basis and the induced filter is the collection of all subsets of $\mathscr{X}$ that contain some tail.

Definition C.1.5. Given a set $\mathscr{X}$, a collection of *neighbourhood filters* is a collection $\mathscr{F}(x)$ of subsets of $\mathscr{X}$ for each point $x \in \mathscr{X}$, such that, for each $x \in \mathscr{X}$,

(i) if $U \in \mathscr{F}(x)$, then $x \in U$,
(ii) if $U \in \mathscr{F}(x)$ then there is a $V \in \mathscr{F}(x)$ such that for all $y \in V$, $U \in \mathscr{F}(y)$.

Condition *(ii)* above looks difficult but simply expresses the following intuitive fact: a subset of points that is 'close' to x (the neighbourhood U), is also 'close' to a point y (also a neighbourhood of y), if y lies 'close enough' to x (*i.e.* anywhere in the neighbourhood V). A collection of neighbourhood filters on a set $\mathscr{X}$ induces a topology: a subset of $\mathscr{X}$ is open, if it is a neighbourhood of each of the points it contains.

Definition C.1.6. A filter $\mathscr{F}$ on a space $\mathscr{X}$ with neighbourhood filters $\mathscr{F}(x)$ *converges* to a point $x \in \mathscr{X}$ if $\mathscr{F}$ is finer than $\mathscr{F}(x)$. A point $x \in \mathscr{X}$ is called an *accumulation point* of a filter $\mathscr{F}$, if x belongs to the closure of every subset in $\mathscr{F}$.

One may verify that a filter of tails of a net converges in the sense of definition C.1.6, if and only if, the net converges in the sense of definition C.1.3. If a filter $\mathscr{F}$ converges to a point x, then x is an accumulation point of $\mathscr{F}$. A point x is an accumulation point of a filter $\mathscr{F}$, if and only if, there exists a filter $\mathscr{F}'$ which is finer than both $\mathscr{F}$ and the neighbourhood filter $\mathscr{F}(x)$ (*i.e.* $\mathscr{F} \subset \mathscr{F}'$ and $\mathscr{F}'$ converges to x).

Definition C.1.7. Given a set $\mathscr{X}$ with two topologies $\mathscr{T}_1, \mathscr{T}_2$ such that $\mathscr{T}_1 \subset \mathscr{T}_2$, $\mathscr{T}_1$ is said to be *coarser* than $\mathscr{T}_2$, and $\mathscr{T}_2$ is said to be *finer* than $\mathscr{T}_1$. Given a set $\mathscr{X}$ with two filters $\mathscr{F}_1, \mathscr{F}_2$ such that $\mathscr{F}_1 \subset \mathscr{F}_2$, $\mathscr{F}_1$ is said to be *coarser* than $\mathscr{F}_2$, and $\mathscr{F}_2$ is said to be *finer* than $\mathscr{F}_1$.

(The above terminology does not imply strictness: any filter is finer and coarser than itself.) Note that any topology on a set $\mathscr{X}$ is finer than the topology $\{\varnothing, \mathscr{X}\}$ (which is called the *trivial topology*) and coarser than the topology formed by the powerset (which is called the *discrete topology*). If we define two topologies $\mathscr{T}_1, \mathscr{T}_2$ on $\mathscr{X}$ through neighbourhood filters $\mathscr{F}_1(x), \mathscr{F}_2(x)$, $(x \in \mathscr{X})$, and $\mathscr{F}_1(x)$ is coarser than $\mathscr{F}_2(x)$ for all $x \in \mathscr{X}$, then $\mathscr{T}_1$ is coarser than $\mathscr{T}_2$.

Definition C.1.8. A filter $\mathscr{F}$ on a set $\mathscr{X}$ is an *ultrafilter* if every filter that is finer than $\mathscr{F}$ is equal to $\mathscr{F}$.

For any filter $\mathscr{F}$, there is at least one ultrafilter that is finer than $\mathscr{F}$, by Zorn's lemma. If an ultrafilter $\mathscr{F}$ has an accumulation point x, then $\mathscr{F}$ converges to x. Given a set $\mathscr{X}$ with an ultrafilter $\mathscr{F}$ and $A, B \subset \mathscr{X}$ such that $A \cup B \in \mathscr{F}$, then either $A \in \mathscr{F}$ or $B \in \mathscr{F}$ (and not both). In particular, for any $A \subset \mathscr{X}$, either A or $\mathscr{X} \setminus A$ belongs to $\mathscr{F}$.

Definition C.1.9. If S is a subset of $\mathscr{X}$ with topology $\mathscr{T}$, S is a topological space called a *subspace* of $(\mathscr{X}, \mathscr{T})$ when it is given the *subspace topology* $\mathscr{T}_S = \{U \cap S : U \in \mathscr{T}\}$.

Definition C.1.10. A space $\mathscr{X}$ with basis $\mathscr{U}$ has a *countable basis at a point* $x \in \mathscr{X}$, if there exists a countable collection of neighbourhoods of x such that each contains at least one element of the basis $\mathscr{U}$. A space that has countable bases for all its points, is called *first countable*. A space that has a countable basis is called *second countable*.

Any subspace of a first (resp. second) countable space is first (resp. second) countable. If a space $\mathscr{X}$ is first countable and $A \subset \mathscr{X}$, then for every $x \in \bar{A}$ there exists a sequence $(x_n) \subset A$ such that $x_n \to x$.

Definition C.1.11. A collection of subsets of a set $\mathscr{X}$ is called a *cover* if the union of those subsets equals $\mathscr{X}$. The cover is called open (or closed, *etcetera*), if those subsets are open (or closed, *etcetera*).

Any open cover of a second countable space $\mathscr{X}$ contains a countable subcover (and $\mathscr{X}$ is called a *Lindelöf space*).

Definition C.1.12. A subset A of a topological space $\mathscr{X}$ is *dense* if every open set U in $\mathscr{X}$ satisfies $U \cap A \neq \varnothing$. The space $\mathscr{X}$ is called *separable* if it has a dense subset that is countable.

A second-countable space is separable.

Definition C.1.13. Given a collection $\{(\mathscr{X}_\alpha, \mathscr{T}_\alpha) : \alpha \in I\}$ of topological spaces, the *product space* $\prod_\alpha \mathscr{X}_\alpha$ is the Cartesian product with elements of the I-tuple form $(x_\alpha : \alpha \in I)$ with the topology generated by the basis of sets of the form,

$$\prod\{U_\beta : \beta \in J\} \times \prod\{\mathscr{X}_\alpha : \alpha \in I \setminus J\},$$

where J is any *finite* subset of I and the sets U_β are open (or even, basis-sets) in $\mathscr{X}_\beta$. The spaces $\mathscr{X}_\alpha$ are called the *factors* of $\prod_\alpha \mathscr{X}_\alpha$. For any $\beta \in I$, the *projection map* $\mathrm{pr}_\beta : \prod_\alpha \mathscr{X}_\alpha \to \mathscr{X}_\beta$ is the map that takes the I-tuple $(x_\alpha : \alpha \in I)$ into its β-component x_β. The product of a countable number of copies of a topological space $\mathscr{X}$ is denoted $\mathscr{X}^{\mathbb{N}}$, or $\mathscr{X}^\infty$.

Given a product space as in definition C.1.13 and subsets $A_\alpha \subset \mathscr{X}_\alpha$ for all $\alpha \in I$, the closure of $\prod_\alpha A_\alpha$ is the product of the closures $\bar{A}_\alpha$. Any product space with countable many first (resp. second) countable factors is first (resp. second) countable.

Definition C.1.14. Given a collection $\{(\mathscr{X}_\alpha, \mathscr{T}_\alpha) : \alpha \in I\}$ of topological spaces, the *topological sum* $\mathscr{X}$ is the set-theoretic disjoint union,

$$\mathscr{X} = \bigcup_{\alpha \in I} \mathscr{X}_\alpha,$$

endowed with the final topology for the collection of injection maps $i_\alpha : \mathscr{X}_\alpha \to \mathscr{X} : x \mapsto (x, \alpha)$, $(\alpha \in I)$.

So a subset U of a topological sum $\mathscr{X}$ is open if each of its disjoint components $U \cap \mathscr{X}_\alpha$, $(\alpha \in I)$, has an open pre-image $i_\alpha^{-1}(U \cap \mathscr{X}_\alpha)$ in $\mathscr{X}_\alpha$.

Definition C.1.15. Given two topological spaces $(\mathscr{X}_1, \mathscr{T}_1)$ and $(\mathscr{X}_2, \mathscr{T}_2)$, a map $f : \mathscr{X}_1 \to \mathscr{X}_2$ is said to be *continuous* if $f^{-1}(V) \in \mathscr{T}_1$ for any $V \in \mathscr{T}_2$. A bijective map f is said to be a *homeomorphism*, if both f and f^{-1} are continuous. If there exists a homeomorphic map between two topological spaces $\mathscr{X}_1, \mathscr{X}_2$, then these spaces are called *homeomorphic*.

For any $x \in \mathscr{X}_1$, f is *continuous in* x if for any $x \in \mathscr{X}_1$ and any neighbourhood W_2 of $f(x)$, there is a neighbourhood W_1 of x such that $f(W_1) \subset W_2$. A map is continuous, if it is continuous in all $x \in \mathscr{X}_1$. A map is continuous, if and only if every converging filter $\mathscr{F}$ in $\mathscr{X}_1$ is mapped to a filter $f(\mathscr{F})$ that converges. Given a subspace S of a topological space $(\mathscr{X}, \mathscr{T})$, the *inclusion map* $j : (S, \mathscr{T}_S) \mapsto (\mathscr{X}, \mathscr{T}) : x \mapsto x$ is continuous; if a set $\mathscr{X}$ has two topologies $\mathscr{T}_1, \mathscr{T}_2$, then the *identity map* $i : (\mathscr{X}, \mathscr{T}_1) \to (\mathscr{X}, \mathscr{T}_2) : x \mapsto x$ is continuous, if and only if $\mathscr{T}_2$ is finer than $\mathscr{T}_1$. Given a topological space $\mathscr{X}$ and a product space as in definition C.1.13, a map $f : \mathscr{X} \to \prod_\alpha \mathscr{X}_\alpha$ is continuous, if and only if $\mathrm{pr}_\alpha \circ f : \mathscr{X} \to \mathscr{X}_\alpha$ is continuous for all $\alpha \in I$.

Definition C.1.16. Let $(\mathscr{Y}_\alpha, \mathscr{T}_\alpha)$, $(\alpha \in I)$, be a collection of topological spaces and $\mathscr{X}$ a set. Given maps $f_\alpha : \mathscr{X} \to \mathscr{Y}_\alpha$, $(\alpha \in I)$, the coarsest topology on $\mathscr{X}$ for which all f_α are continuous, is called the *initial topology* for the collection of maps f_α, $(\alpha \in I)$. Given maps $f_\alpha : \mathscr{Y}_\alpha \to \mathscr{X}$, $(\alpha \in I)$, the finest topology on $\mathscr{X}$ for which all f_α are continuous, is called the *final topology* for the collection of maps f_α, $(\alpha \in I)$.

Given a product space as in definition C.1.13, the product topology is the initial topology for the collection of all projection maps.

The following definitions are specific to functions defined on topological spaces, taking values on the (extended) real line.

Definition C.1.17. A function $f : \mathscr{X} \to [-\infty, \infty]$ is *upper (lower) semi-continuous* at $x \in \mathscr{X}$ if, for every $y > f(x)$ $(y < f(x))$, there exists a neighbourhood U of x such that $f(z) < y$ $(f(z) > y)$ for all $z \in U$.

Definition C.1.18. Given a topological space $\mathscr{X}$, the *support* of a function $f : \mathscr{X} \to [-\infty, \infty]$ is the closure of $\{x \in \mathscr{X} : f(x) \neq 0\}$.

(For the support of a Borel measure, see definition C.1.18; for the support of a Radon measure (see definition C.8.1), see proposition 2.1.16).)

C.2 Separation axioms and compactness

Topologies and filters are very general notions and in most cases where one conceptualizes what it means that one point in a set is 'close' to another, the resulting definitions have certain immediate properties that have become known as *separation axioms*. These are linked intimately with the concept of compactness, which plays a central role in any topological argument.

Definition C.2.1. A topological space $(\mathscr{X}, \mathscr{T})$ is said to be *Hausdorff*, if for every $x, y \in \mathscr{X}$, $x \neq y$, there exist neighbourhoods V, W of x, y respectively, such that $V \cap W = \varnothing$.

A subspace of a Hausdorff space is Hausdorff; if each factor $\mathscr{X}_\alpha$ of a product space is Hausdorff, then the product space is Hausdorff.

Definition C.2.2. A topological space is is said to be *regular*, if for every closed $A \subset \mathscr{X}$ and every point $x \in \mathscr{X} \setminus A$, there exist disjoint open U, V such that $A \subset U$ and $x \in V$.

A subspace of a regular space is regular; if each factor $\mathscr{X}_\alpha$ of a product space is regular, then the product space is regular. A topological space $\mathscr{X}$ is regular, if and only if, for every $x \in \mathscr{X}$ and every neighbourhood U of x, there exists a neighbourhood V of x such that $\bar{V} \subset U$. Every regular space is a Hausdorff space.

Definition C.2.3. A topological space $(\mathscr{X}, \mathscr{T})$ is said to be *completely regular*, if $\mathscr{X}$ is Hausdorff and for every closed subset A of $\mathscr{X}$ any point $x \in \mathscr{X} \setminus A$, there exists a continuous function $f : \mathscr{X} \to [0,1]$ such that $f = 0$ on A and $f(x) = 1$.

A subspace of a completely regular space is completely regular; if each factor $\mathscr{X}_\alpha$ of a product space is completely regular, then the product space is completely regular. Every completely regular space is a regular space but the opposite does not hold.

Definition C.2.4. A topological space $\mathscr{X}$ is said to be *normal*, if for every pair of disjoint, closed subsets A, B of $\mathscr{X}$, there exist disjoint, open U, V such that $A \subset U$ and $B \subset V$.

A topological space $\mathscr{X}$ is normal, if and only if, for every closed $A \subset \mathscr{X}$ and open U such that $A \subset U$, there exists an open V such that $A \subset \bar{V} \subset U$. Every normal space is a completely regular space but the opposite does not hold. Metric spaces are normal spaces. If $\mathscr{X}$ is a normal space and A, B are disjoint, closed subsets of $\mathscr{X}$, then there exists a continuous function $f : \mathscr{X} \to [0,1]$ such that $f = 0$ on A and $f = 1$ on B (a result sometimes referred to as *Urysohn's lemma*). If A is a closed subspace of a normal space $\mathscr{X}$ and $f : A \to [0,1]$ (resp. $f : A \to \mathbb{R}$) is continuous, then there exists a continuous extension $g : \mathscr{X} \to [0,1]$ (resp. $g : \mathscr{X} \to \mathbb{R}$) of f to all of $\mathscr{X}$.

Definition C.2.5. A topological space $\mathscr{X}$ is said to be *connected*, if it cannot be written as the union of two open subsets that are disjoint.

Definition C.2.6. A Hausdorff topological space is a *zero-dimensional space* if its topology has a basis of *clopen subsets*.

In a zero-dimensional space there is no subspace that is connected. A subspace of a zero-dimensional space is zero-dimensional; a product of zero-dimensional spaces is zero-dimensional; a topological sum of zero-dimensional spaces is zero-dimensional.

Definition C.2.7. A topological space $\mathscr{X}$ is called *compact* if every open cover of $\mathscr{X}$ contains a finite sub-collection that also covers $\mathscr{X}$. A topological space is *locally compact* if every point x has a compact neighbourhood. A subset A of $\mathscr{X}$ is *relatively compact* is its closure $\bar{A}$ is a compact subspace of $\mathscr{X}$. A topological space $\mathscr{X}$ is σ-*compact*, if $\mathscr{X}$ equals a countable union of compact subsets.

If $(\mathscr{X},\mathscr{T})$ is compact and $\mathscr{T}'$ coarsens $\mathscr{T}$, then $\mathscr{T}=\mathscr{T}'$. Every closed subspace of a compact space is compact and every compact subspace of a Hausdorff space is closed. If A is a compact subspace of a Hausdorff space $\mathscr{X}$ and $x\in\mathscr{X}\setminus A$, then there exist disjoint open U,V such that $A\subset U$ and $x\in V$. Every compact Hausdorff space is a normal space. If $\mathscr{X}$ is a Hausdorff space, then $\mathscr{X}$ is compact, if and only if, every filter on $\mathscr{X}$ has at least one accumulation point, if and only if, every ultrafilter on $\mathscr{X}$ converges, if and only if, every collection of closed subset with an empty intersection has a finite sub-collection with empty intersection. If $\mathscr{X}$ is compact, $\mathscr{Y}$ is Hausdorff and $f:\mathscr{X}\to\mathscr{Y}$ is continuous, then $f(\mathscr{X})$ is a compact subspace of $\mathscr{Y}$; if, in addition, f is injective, then $\mathscr{X}$ and the subspace $f(\mathscr{X})$ of $\mathscr{Y}$ are homeomorphic. *Tychonov's theorem* says that if $\mathscr{X}$ is a product space $\prod_\alpha\mathscr{X}_\alpha$ with factors $\mathscr{X}_\alpha$ that are all compact, then $\mathscr{X}$ is compact; if $\mathscr{X}$ is a product space $\prod_\alpha\mathscr{X}_\alpha$ with locally compact factors $\mathscr{X}_\alpha$, then $\mathscr{X}$ is locally compact, if and only if all but finitely many factors are compact. A space is completely regular, if and only if it is homeomorphic to a Hausdorff subspace of a compact space. A locally compact Hausdorff space is completely regular.

Definition C.2.8. A *compactification* of a topological space $\mathscr{X}$ is a dense subspace A of a compact topological space $\mathscr{Y}$ that is that is homeomorphic with $\mathscr{X}$. A *one-point-compactification* of $\mathscr{X}$ is a compactification such that $\mathscr{Y}\setminus A$ consists of a single point $\omega\in\mathscr{Y}$.

A locally compact space $\mathscr{Y}$ has a one-point-compactification that is unique up to homeomorphisms, and the point ω has a countable basis of neighbourhoods if and only if $\mathscr{Y}$ is σ-compact.

C.3 Uniform spaces and complete spaces

Whereas topological spaces give an abstract notion of 'closeness' of one point in the space to another, there is no notion of 'relative closeness', that is, no way to compare what is close to one point and what is close to another point in the same topological space. Of course, relative closeness is well defined in metric spaces (see subsection C.4), where we can compare the metric distance between points x_1 and x_2, with the metric distance between two other points y_1 and y_2. To define 'relative closeness' more generally, we introduce uniform spaces below, as a natural abstraction from the metric spaces introduced before. This enables definition of the important concepts of Cauchy nets and completeness of a uniform space.

In the definition below we use the following notation: if $\mathscr{X}$ is a set and $U,V\subset X\times X$, then the *composite* $U\circ V$ denotes the set $\{(x,z)\in\mathscr{X}\times\mathscr{X}:\exists_{y\in\mathscr{X}}(x,y)\in U,(y,z)\in V\}$.

Definition C.3.1. A *uniform space* is a non-empty set $\mathscr{X}$ with a filter $\mathscr{W}$ of subsets of $\mathscr{X}\times\mathscr{X}$ such that:

(i) every W contains the diagonal $\Delta=\{(x,x):x\in\mathscr{X}\}$,

(ii) if $W \in \mathscr{W}$, then $W^{-1} = \{(y,x) : (x,y) \in W\} \in \mathscr{W}$,
(iii) for every $V \in \mathscr{W}$, there exists a $W \in \mathscr{W}$ such that $W \circ W \subset V$.

The collection $\mathscr{W}$ is called a *uniformity* on $\mathscr{X}$. The sets in $\mathscr{W}$ are called *entourages*. A *fundamental system of entourages* $\mathscr{F}$ for $\mathscr{W}$ is any collection of entourages such that any $W \in \mathscr{W}$ contains an entourage from $\mathscr{F}$.

Property *(ii)* says that entourages reflected in the diagonal remain entourages (symmetry of $\mathscr{W}$); property *(iii)* can be interpreted as an abstraction of the triangle inequality that holds in metric spaces (see definition C.4.1). Given a set $\mathscr{X}$, a collection $\mathscr{F}$ of subsets of $\mathscr{X} \times \mathscr{X}$ is a fundamental system of entourages for a uniformity, if and only if,

(i) every $W \in \mathscr{F}$ contains the diagonal $\Delta = \{(x,x) : x \in \mathscr{X}\}$,
(ii) for any $W, W' \in \mathscr{F}$, there is a $W'' \in \mathscr{F}$ such that $W'' \subset W \cap W'$,
(iii) for any $W \in \mathscr{F}$, there is a $W' \in \mathscr{F}$ such that $W' \subset W^{-1}$,
(iv) for any $W \in \mathscr{F}$, there exists a $V \in \mathscr{F}$ such that $V \circ V \subset W$.

Given a uniform structure $\mathscr{W}$ on a set $\mathscr{X}$ and a point $x \in \mathscr{X}$, the sets $\{y \in \mathscr{X} : (x,y) \in W\}$ form a basis of neighbourhoods for the point x and there is a unique topology $\mathscr{T}$ on $\mathscr{X}$ for which these sets form a collection of neighbourhoods for x, called the *topology induced by* $\mathscr{W}$. The topology $\mathscr{T}$ is Hausdorff if and only if the intersection of all entourages in $\mathscr{W}$ is the diagonal δ; every Hausdorff uniform space is completely regular and every completely regular space $(\mathscr{X}, \mathscr{T})$ has a uniformity $\mathscr{W}$ that induces the topology $\mathscr{T}$ (although $\mathscr{W}$ is not unique unless $\mathscr{X}$ is compact). A Hausdorff space $\mathscr{X}$ is a uniform space, if and only if every lower semi-continuous $f : \mathscr{X} \to \mathbb{R}$ satisfies $f(x) = \sup\{g(x) : g \text{ continuous}, g \leq f\}$.

Definition C.3.2. Given two uniform spaces $(\mathscr{X}_1, \mathscr{W}_1)$ and $(\mathscr{X}_2, \mathscr{W}_2)$, a map $f : \mathscr{X}_1 \to \mathscr{X}_2$ is said to be *uniformly continuous* if for any entourage $V \in \mathscr{W}_2$, there exists an entourage $W \in \mathscr{W}_1$ such that $(x,y) \in V$ implies $(f(x), f(y)) \in V$. A bijective map f is said to be a *uniform homeomorphism* , if both f and f^{-1} are uniformly continuous. If there exists a uniformly homeomorphic map between two uniform spaces $\mathscr{X}_1, \mathscr{X}_2$, then these spaces are called *uniformly homeomorphic*.

Any uniformly continuous map $f : \mathscr{X}_1 \to \mathscr{X}_2$ is continuous for the induced topologies $\mathscr{T}_1$ and $\mathscr{T}_2$. If $\mathscr{X}_1$ is compact and $\mathscr{X}_2$ is a uniform space, any continuous map $f : \mathscr{X}_1 \to \mathscr{X}_2$ is uniformly continuous. Other definitions made earlier in this section for topological spaces (*e.g.* subspaces, initial topologies, products, *etcetera*) have obvious generalizations to uniform spaces.

Definition C.3.3. Let $(\mathscr{X}, \mathscr{W})$ be a uniform space. A *Cauchy net* in $\mathscr{X}$ is any net (x_α), $(\alpha \in I)$, such that, for any $W \in \mathscr{W}$, there exists a $\alpha \in I$ such that for all $\beta, \gamma \geq \alpha$, $(x_\beta, x_\gamma) \in W$. A Cauchy filter $\mathscr{F}$ is a filter $\mathscr{F}$ such that for every entourage $W \in \mathscr{W}$, there exists a $U \in \mathscr{F}$ such that $U \times U \subset W$.

If a net (x_α) in a uniform space $\mathscr{X}$ converges, then (x_α) is a Cauchy net; if a filter $\mathscr{F}$ in a uniform space $\mathscr{X}$ converges, then $\mathscr{F}$ is a Cauchy filter. The opposites of these two facts do not hold in general, which motivates the following definition.

Definition C.3.4. A uniform space $\mathscr{X}$ is *complete*, if every Cauchy net in $\mathscr{X}$ converges to a point in $\mathscr{X}$. Equivalently, a uniform space is complete if every Cauchy filter with an accumulation point x, converges to x.

Every closed subspace of a complete uniform space is a complete uniform space and every complete subspace of a Hausdorff uniform space is closed. If $(\mathscr{X},\mathscr{T})$ is compact, the (unique) uniformity associated with $\mathscr{T}$ is complete. If A is a subset of a uniform space $\mathscr{X}$ and $f : A \to \mathscr{Y}$ maps A to a complete Hausdorff uniform space $\mathscr{Y}$ and f is uniformly continuous, then f can be extended to a uniformly continuous function $g : \mathscr{X} \to \mathscr{Y}$.

Definition C.3.5. For any uniform space $\mathscr{X}$, there exists a *complete* Hausdorff uniform space $\mathscr{Y}$ and a uniformly continuous map $i : \mathscr{X} \to \mathscr{Y}$ such that $i(\mathscr{X})$ is dense in $\mathscr{Y}$; $\mathscr{Y}$ is called the *(Hausdorff) completion* of $\mathscr{X}$ and $\mathscr{Y}$ is unique up to uniform homeomorphisms. A uniform space is *pre-compact* if its Hausdorff completion is compact.

A subset A of a Hausdorff uniform space is pre-compact, if and only if, the closure of $i(A)$ is compact in the Hausdorff completion, if and only if, for every entourage $W \in \mathscr{W}$, there exists a *finite* cover $\{A_1,\ldots,A_n\}$ of A such that $A_i \times A_i \subset W$, for every $1 \leq i \leq n$. A Hausdorff uniform space is compact, if and only if it is complete and pre-compact. Given a set $\mathscr{X}$ and a collection of pre-compact spaces $\{\mathscr{X}_\alpha : \alpha \in I\}$ with maps $\{f_\alpha : \mathscr{X} \to \mathscr{X}_\alpha : \alpha \in I\}$, the smallest uniformity for which all f_α are uniformly continuous makes $\mathscr{X}$ a pre-compact space.

C.4 Metric spaces and Polish spaces

Metric spaces are ubiquitous and because many topological spaces used in this book are complete (and often also separable) for a metrizable topology, we discuss metric spaces with specific attention for Baire and Polish spaces. Much more on these subjects can be found in [47], chapter IX, § 5 and § 6, and comprehensively, in Kechris (1994), [143]. With its focus on descriptive set theory, the latter book goes much further. Most of its material is not used in this book but certainly warrants attention from readers interested in what mathematics lies beyond the realm of Borel sets.

Definition C.4.1. Let $\mathscr{X}$ be a set with a function $d : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ such that,

(i) for all $x,y \in \mathscr{X}$, $d(x,y) \geq 0$ and $d(x,y) = 0$ if and only if $x = y$,
(ii) for all $x,y \in \mathscr{X}$, $d(x,y) = d(y,x)$,
(iii) for all $x,y,z \in \mathscr{X}$, $d(x,z) \leq d(x,y) + d(y,z)$.

Then d is called a *metric* and $(\mathscr{X},d)$ is called a *metric space*. An *open (d-)ball* $B_d(x,r)$ (or $B(x,r)$, if the subscript d can be omitted unambiguously) centred on $x \in \mathscr{X}$ of radius $r \geq 0$ is the set $\{y \in \mathscr{X} : d(x,y) < r\}$. The collection of all d-balls in $\mathscr{X}$ forms a basis for a topology $\mathscr{T}_d$ on $\mathscr{X}$ called the *metric topology* associated with the metric d. The sets $\{(x,y) \in \mathscr{X} \times \mathscr{X} : d(x,y) < \varepsilon\}$, $(\varepsilon > 0)$, form a fundamental

system of entourages for a corresponding *metric uniformity* $\mathscr{W}_d$ on $\mathscr{X}$. An *isometry* is a bijection $f : \mathscr{X}_1 \to \mathscr{X}_2$ between two metric spaces $(\mathscr{X}_1, d_1)$ and $(\mathscr{X}_2, d_2)$, such that $d_2(f(x), f(y)) = d_1(x, y)$ for all $x, y \in \mathscr{X}_1$. A subset $\mathscr{Y}$ of a metric space is *bounded* if $\sup\{d(x, y) : x, y \in \mathscr{Y}\} < \infty$. A metric space $(\mathscr{X}, d)$ is bounded, if $\mathscr{X}$ is bounded as a subset, in which case the metric d is referred to as a *bounded metric*.

Property *(iii)* above is referred to as the *triangle inequality* associated with the metric d. Every isometry is a (uniform) homeomorphism. If properties *(i)–(iii)* hold, except that $d(x, y) = 0$ does not imply $x = y$, then d is called a *semi-metric*. Metric spaces are normal spaces, but a topology induced by a semi-metric that is not a metric, is not even Hausdorff. A compact subset of a metric space is bounded.

Definition C.4.2. A topological space $(\mathscr{X}, \mathscr{T})$ is *metrizable* if there exists a *(topologically) compatible metric* for $\mathscr{T}$, that is, if there exists a metric d with $\mathscr{T}_d = \mathscr{T}$. Similarly, a uniform space $(\mathscr{X}, \mathscr{W})$ is *metrizable*, if there exists a *(uniformly) compatible metric* for $\mathscr{W}$, that is, if there exists a metric d with $\mathscr{W}_d = \mathscr{W}$. A topological space $(\mathscr{X}, \mathscr{T})$ is *completely metrizable*, if there exists a compatible metric for $\mathscr{T}$ with a metric uniformity for which $\mathscr{X}$ is complete.

For every metrizable space $(\mathscr{X}, \mathscr{T})$, there exists a bounded metric d such that $\mathscr{T} = \mathscr{T}_d$. A subspace of a metrizable space is metrizable; a countable product of metrizable spaces is metrizable. A metrizable topological space $\mathscr{X}$ is first countable and $\mathscr{X}$ is second countable, if and only if, $\mathscr{X}$ is separable, if and only if, $\mathscr{X}$ is a Lindelöf space.

Theorem C.4.3. *(Urysohn metrization)*
Every regular space that is second countable is metrizable.

A uniformity $\mathscr{W}$ on a space $\mathscr{X}$ is metrizable, if and only if there exists a *countable* fundamental system of entourages for $\mathscr{W}$. Every closed subset of a metrizable space $\mathscr{X}$ is a G_δ*-set*. If $\mathscr{X}$ is a topological space and $\mathscr{Y}$ is a metrizable space, then the points of continuity of a map $f : \mathscr{X} \to \mathscr{Y}$ form a G_δ-set in $\mathscr{X}$. If $\mathscr{X}$ is metrizable and $\mathscr{Y} \subset \mathscr{X}$ is completely metrizable, then $\mathscr{Y}$ is a G_δ-set in $\mathscr{X}$; if $\mathscr{X}$ is completely metrizable and $\mathscr{Y} \subset \mathscr{X}$ is a G_δ-set, then $\mathscr{Y}$ is completely metrizable. A subspace $\mathscr{Y}$ of a metrizable space is compact, if and only if, every sequence (x_n) in $\mathscr{Y}$ has a convergent sub-sequence.

Completely metrizable spaces play a central role in this book: the set of all probability measure on the real line (and all other realistic sample spaces) is completely metrizable in the two most common topologies, Prokhorov's weak topology and the total-variational topology (see section C.5). Completely metrizable spaces have one topological property that stands out and which they share with locally compact Hausdorff spaces.

Definition C.4.4. A topological space $\mathscr{X}$ is a *Baire space*, if any countable intersection of open, dense subsets of $\mathscr{X}$, is again dense in $\mathscr{X}$.

Equivalently $\mathscr{X}$ is a Baire space if any countable union of closed sets with empty interiors, again has empty interior. Any open subspace of a Baire space is a Baire

space. A subset A of a topological space $\mathscr{X}$ is *nowhere dense* if the closure $\bar{A}$ has empty interior. A subset A of $\mathscr{X}$ is *meager* (or *of first (Baire) category* in $\mathscr{X}$), if A is a countable union of nowhere dense subsets; the complement $\mathscr{X} \setminus A$ of a meager set A is called *residual*; any subset of $\mathscr{X}$ that is not meager is said to be *of second (Baire) category* in $\mathscr{X}$. An example of a meager subset in $\mathbb{R}$ is $\mathbb{Q}$, and the set of all irrational real numbers is an example of a residual set. A topological space $\mathscr{X}$ is a Baire space, if and only, if every meager subset A is nowhere dense (see exercise **??**), if and only if, every point $x \in \mathscr{X}$ has a neighbourhood U that is a Baire space.

Theorem C.4.5. *(Baire)*
Locally compact Hausdorff spaces and completely metrizable spaces are Baire spaces.

Definition C.4.6. A subset A of a topological space $\mathscr{X}$ has the *Baire property* if there exists an open subset U of $\mathscr{X}$ such that the symmetric difference $(A \setminus U) \cup (U \setminus A)$ between A and U is meager in $\mathscr{X}$.

The subsets of a topological space with the Baire property form a σ-algebra, the smallest σ-algebra that contains all open and all meager sets. In particular, all open, closed, G_δ- and F_σ-sets have the Baire property. In fact, a subset A of $\mathscr{X}$ has the Baire property, if and only if, A is the disjoint union of a G_δ-set and a meager set, if and only if, A is contained in an F_σ-set F and $F \setminus A$ is meager.

Now we are in a position to consider Polish spaces.

Definition C.4.7. A topological space $\mathscr{X}$ is a *Polish space*, if $\mathscr{X}$ is completely metrizable and separable. Given a Polish space $\mathscr{X}$, a metrizable space $\mathscr{Y}$, and a continuous $f : \mathscr{X} \to \mathscr{Y}$, the subspace $f(\mathscr{X})$ is called a *Souslin space*; if, in addition, f is injective, $f(\mathscr{X})$ is called a *Lusin space*.

Countable products and countable topological sums of Polish spaces are Polish spaces. Some of the most important examples of Polish spaces are countable discrete spaces.

Example C.4.8. The space $\{0,1\}^{\mathbb{N}}$ is called the *Cantor space* and $\mathbb{N}^{\mathbb{N}}$ is (confusingly) called *the Baire space*. The Cantor space is compact and homeomorphic to the fractal subspace $\mathscr{C}$ of $[0,1]$ that is obtained by deleting an open interval, then deleting open intervals from the two remaining closed intervals, and repeating ad infinitum (*e.g.* first delete $(1/3,2/3)$ from $[0,1]$, then $(1/9,2/9)$ from $[0,1/3]$ and $(7/9,8/9)$ from $[2/3,1]$, *etcetera*). In this context, we also define the function x that parametrizes the set of all mid-points of deleted intervals in terms of finite binary sequences, which we shall refer to as the *Cantor mid-point function*. Like in section 8.3, we define, for every $m \geq 0$ the set $\mathscr{E}_m$ as the set of all binary sequences ε of length m (including the $m = 0$ case of the empty binary sequence $\varepsilon_\varnothing$), and the set $\mathscr{E} = \cup_{m \geq 0} \mathscr{E}_m$ of all *finite binary sequences*. The function $x : \mathscr{E} \to [0,1]$ maps $\varepsilon \in \mathscr{E}_m$ to the midpoint of the interval that is deleted in the m-th transition in the construction of the set $\mathscr{C}$: for example, $x(\varepsilon_\varnothing) = 1/2$ in $\mathscr{E}_0$, $x(0) = 1/6$, $x(1) = 5/6$ in $\mathscr{E}_1$, $x(00) = 1/18$, $x(01) = 5/18$, $x(10) = 13/18$, $x(11) = 17/18$ in $\mathscr{E}_2$, *etcetera*. In particular, for any $m \geq 1$, $x(\varepsilon) = 1/2(1/3)^m$ and $x(\varepsilon') = 1 - 1/2(1/3)^m$ if $\varepsilon = 0\ldots0 \in \mathscr{E}_m$ and $\varepsilon' = 1\ldots1 \in \mathscr{E}_m$.

Other examples of Polish spaces are $[0,1]^{\mathbb{N}}$ (usually referred to as the *Hilbert cube*), and of course separable Banach and Hilbert spaces, including the Euclidean spaces $\mathbb{R}^d$, ($d \geq 1$). The Hilbert cube is compact and every separable metrizable space is homeomorphic to a subspace of the Hilbert cube.

Theorem C.4.9. *A subspace $\mathscr{Y}$ of a Polish space $\mathscr{X}$ is a Polish space, if and only if $\mathscr{Y}$ is a G_δ-set in $\mathscr{X}$.*

Every non-empty compact metrizable space is a continuous image of the Cantor space. Any compact, metrizable space is Polish. If $\mathscr{X}$ is Hausdorff and locally compact, $\mathscr{X}$ is *second countable*, if and only if, $\mathscr{X}$ is metrizable and σ-*compact*, if and only if, $\mathscr{X}$ is Polish, if and only if, $\mathscr{X}$ is homeomorphic to an open subset of a compact metrizable space.

Lusin and Souslin spaces are important because of their relation to *Borel measurability*. A subset A of a Lusin space is a Lusin space, if and only if A is a Borel set; every Lusin subspace of a metrizable space is a Borel set.

Theorem C.4.10. *If $\mathscr{X},\mathscr{Y}$ are Souslin spaces, then $f : \mathscr{X} \to \mathscr{Y}$ is Borel measurable, if and only if the graph $\{(x, f(x)) \in \mathscr{X} \times \mathscr{Y} : x \in \mathscr{X}\}$ of f is a Souslin subspace of $\mathscr{X} \times \mathscr{Y}$.*

If $\mathscr{X},\mathscr{Y}$ are Souslin spaces and $f : \mathscr{X} \to \mathscr{Y}$ is a Borel measurable injection, then its inverse on $f(\mathscr{X})$ is also Borel measurable. For any two metrizable Lusin spaces of the same cardinal, there exists a Borel measurable bijection.

Zero-dimensionality plays a role for the characterization of Polish spaces. A separable metrizable space $\mathscr{X}$ is zero-dimensional, if and only if, for every closed subset A of $\mathscr{X}$ there exists a continuous surjection $f : \mathscr{X} \to A$ such that $f(x) = x$ for all $x \in A$. A metrizable space $\mathscr{Y}$ is a Lusin space, if and only if there exists a zero-dimensional Polish space $\mathscr{X}$ and a continuous bijection $f : \mathscr{X} \to \mathscr{Y}$.

Theorem C.4.11. *(Alexandrov-Urysohn)*
Up to homeomorphisms, the Baire space $\mathbb{N}^{\mathbb{N}}$ is the only non-empty, zero-dimensional Polish space in which all compact subspaces have empty interior.

Theorem C.4.12. *(Brouwer)*
Up to homeomorphisms, the only non-empty, compact, zero-dimensional space without isolated points is the Cantor space.

Every zero-dimensional Polish space is homeomorphic to a closed subspace of the Baire space and to a G_δ-set in the Cantor space.

C.5 Inverse limit spaces

Bourbaki introduces the so-called *inverse limit space* (known also as *projective limit space* [223]) as a construction that can be interpreted at many levels of detail. The set-theoretic definition ([45], Ch. III, § 7, No. 1; [45], Ch. III, § 7, No. 2; [45], Ch. R,

§ 6, No. 2) is described as follows. Recall that a *directed set* is a set I with a *partial ordering* relation $\leq$ (that is: $\alpha \leq \alpha$; $\alpha \leq \beta$ and $\beta \leq \alpha \Rightarrow \alpha = \beta$; $\alpha \leq \beta$ and $\beta \leq \gamma \Rightarrow \alpha \leq \gamma$), such that for all pairs $\alpha, \beta \in I$ there exits a $\gamma \in I$ with $\alpha \leq \gamma$ and $\beta \leq \gamma$.

Definition C.5.1. Let I be a directed set and let $(\mathscr{X}_\alpha)_{\alpha \in I}$ be sets with maps $f_{\alpha\beta} : \mathscr{X}_\beta \to \mathscr{X}_\alpha$ such that,

(i) for all $\alpha \leq \beta \leq \gamma$, we have $f_{\alpha\gamma} = f_{\alpha\beta} \circ f_{\beta\gamma}$,
(ii) for all $\alpha \in I$, $f_{\alpha\alpha} = i_\alpha$, the identity mapping on $\mathscr{X}_\alpha$.

The *inverse limit* of the *inverse limit system* $(\mathscr{X}_\alpha, f_{\alpha\beta})$ is then defined as the set $\mathscr{X}$ of all x in the (set-theoretic) product $\prod_{\alpha \in I} \mathscr{X}_\alpha$ that satisfy,

$$\mathrm{pr}_\alpha(x) = f_{\alpha\beta}(\mathrm{pr}_\beta(x)),$$

for all $\alpha \leq \beta$.

Conceptually, the maps $f_{\alpha\beta}$ can be thought of as a system of projections between the spaces $\mathscr{X}_\alpha$. The restriction of pr_α to $\mathscr{X}$ is denoted $f_\alpha : \mathscr{X} \to \mathscr{X}_\alpha$ and called the *canonical mapping* of $\mathscr{X}$ onto $\mathscr{X}_\alpha$; these mappings form a so-called *coherent family*, *i.e.* they satisfy $f_\alpha = f_{\alpha\beta} \circ f_\beta$ for all $\alpha \leq \beta$. An immediate point of caution concerns the still-open possibility that $\mathscr{X} = \varnothing$: not every inverse system has a well-defined inverse limit. Non-emptiness is most conveniently demonstrated by injection of another space, *e.g.* as in proposition 8.6.5. For an inverse system of topological spaces $(\mathscr{X}_\alpha, \mathscr{T}_\alpha)$ more can be said (see [46], Ch. I, § 4, No. 4).

Definition C.5.2. In the above setup, assume that the $\mathscr{X}_\alpha$ are topological spaces and that the maps $f_{\alpha\beta}$ are continuous for all $\alpha \leq \beta$. The *topological inverse limit* $\mathscr{X}$ is the set-theoretic inverse limit $\mathscr{X}$ with the initial topology for the canonical mappings, that is, the coarsest topology that makes all $f_\alpha : \mathscr{X} \to \mathscr{X}_\alpha$ continuous.

Example C.5.3. Let I be a directed set and denote by $\mathscr{S}$ the collection of all finite $S \subset I$. Given a family of topological spaces $(\mathscr{X}_\alpha)$, the *product space* $\prod_\alpha \mathscr{X}_\alpha$ of definition C.1.13 is defined equivalently as the inverse limit of the finite topological products,

$$\mathscr{X}_S = \prod_{\alpha \in S} \mathscr{X}_\alpha$$

where $\mathscr{S}$ is directed by inclusion. The maps $f_{ST} : \mathscr{X}_T \to \mathscr{X}_S$ for $S \subset T$ are projections between finite product spaces. The canonical mappings $f_S : X \to \mathscr{X}_S$ are the usual projections pr_S. The corresponding inverse limit topology is therefore the coarsest topology on the set-theoretic product that makes all projections continuous, *c.f.* the usual definition of the topological product.

To characterize convergence and continuity in inverse limit spaces, we have the following specification of [46], Ch. I, § 2, No. 3, Prop. 4.

Proposition C.5.4. *Let the topological space $(X, \mathscr{T})$ be the inverse limit of the inverse system of topological spaces $(\mathscr{X}_\alpha, \mathscr{T}_\alpha)$. Then the collection of all finite intersections of sets of the form $f_\alpha^{-1}(U)$ ($\alpha \in I$, $U \in \mathscr{T}_\alpha$), forms a basis for $\mathscr{T}$. Furthermore, given a topological space $\mathscr{Y}$, a map $h : \mathscr{Y} \mapsto X$ is continuous, if and only if $f_\alpha \circ h : \mathscr{Y} \to \mathscr{X}_\alpha$ is continuous for all $\alpha \in I$.*

One may wonder which topological properties lift from the spaces $\mathscr{X}_\alpha$ to the inverse limit space $\mathscr{X}$. For one, an inverse limit is compact and non-empty whenever the spaces $\mathscr{X}_\alpha$ are compact and non-empty ([46], Ch. I, § 9, No. 6, prop. 8); a so-called *inverse limit of uniform spaces* requires the $f_{\alpha\beta}$ to be *uniformly continuous* and leads to a uniformity on the inverse limit space, the coarsest that makes all f_α uniformly continuous [46], Ch. II, § 2, No. 7, prop. 8. In case the uniform spaces $\mathscr{X}_\alpha$ are complete, the inverse limit $\mathscr{X}$ is also complete ([46], Ch. II, § 3, No. 5, Prop. 10 and Cor.). Note the following criterion for the *Cauchy property* (which is a specific version of [46], Ch. II, § 3, No. 1, Prop. 4).

Proposition C.5.5. *Let $(\mathscr{X},\mathscr{W})$ be the inverse limit of uniform spaces $(\mathscr{X}_\alpha,\mathscr{W}_\alpha)$. Then the collection of all finite intersections of subsets of the form $(f_\alpha,f_\alpha)^{-1}(V)$, where $\alpha\in I$ and V is an entourage from $\mathscr{W}_\alpha$ forms a fundamental system of entourages for $\mathscr{W}$. Furthermore, given a uniform space $\mathscr{Y}$, a map $h:\mathscr{Y}\to\mathscr{X}$ is uniformly continuous, if and only if the maps $f_\alpha\circ h:\mathscr{Y}\to\mathscr{X}_\alpha$ are uniformly continuous for all $\alpha\in I$. Moreover, a filter base $\mathscr{C}$ on $\mathscr{X}$ is Cauchy, if and only if $f_\alpha(\mathscr{C})$ is Cauchy for all $\alpha\in I$.*

Statistical models for the distributions of *i.i.d.* samples $X=(X_1,X_2,\ldots)$ from a topological space $\mathscr{X}$, carry a natural uniformity that arises as an inverse limit. Let $(\mathscr{X},\mathscr{B})$ denote the Borel space in which each of the sample points X_i, $(i\geq 1)$ takes its values, so X lies in the countable product space $\mathscr{X}^{\mathbb{N}}$, with σ-algebra $\mathscr{B}^{\mathbb{N}}$ generated by the cylinder sets. The sample space $\mathscr{X}^n$ for the first n sample points is denoted $\mathscr{X}_n$ with product σ-algebra $\mathscr{B}_n$. Let $\mathscr{P}$ denote a collection of Borel probability measures on $(\mathscr{X},\mathscr{B})$. Note that the model $\mathscr{P}$ is mapped one-to-one to a collection of infinite product measures with domain $\mathscr{B}^{\mathbb{N}}$: $\mathscr{P}^\infty=\{P^\infty:\mathscr{B}^{\mathbb{N}}\to[0,1]:P\in\mathscr{P}\}$. As in example C.5.3, the inverse limit of the spaces $(\mathscr{X}_n,\mathrm{pr}_{nm})$ is $\mathscr{X}^{\mathbb{N}}$. On the space $\mathscr{P}_\infty$, we define uniformities $\mathscr{W}_n$ with an inverse limit $\mathscr{W}_\infty$ as follows.

Definition C.5.6. For each $n\geq 1$, consider the linear space $\mathscr{F}_n$ of all bounded $\mathscr{B}_n$-to-Borel-measurable $f:\mathscr{X}_n\to\mathbb{R}$ and consider the fundamental system of entourages on $\mathscr{P}^\infty$, obtained by choosing $k\geq 1$, and $f_1,\ldots,f_k\in\mathscr{F}_n$ to define,

$$W_{n;f_1,\ldots,f_k}=\Big\{(P,Q)\in\mathscr{P}^\infty\times\mathscr{P}^\infty:\big|(P-Q)f_i\big|<1,1\leq i\leq k\Big\}.$$

For every $n\geq 1$, these subsets of $\mathscr{P}^\infty\times\mathscr{P}^\infty$ form a fundamental system of entourages for the so-called *n-th Le Cam-Schwartz uniformity* $\mathscr{W}_n$ on $\mathscr{P}^\infty$ (and by extension with slight abuse of notation, also on $\mathscr{P}$, which is in bijective correspondence with the diagonal in $\mathscr{P}^\infty$ and inherits the subspace uniformity). The associated *n-th Le Cam-Schwartz topology* (on $\mathscr{P}^\infty$ and, again by extension, also on $\mathscr{P}$) is denoted $\mathscr{T}_n$. Identifying $n\geq 1$ as the index α, we call the inverse limit $(\mathscr{P}^\infty,\mathscr{W}_\infty)$ of the inverse system of uniform spaces $((\mathscr{P}^\infty,\mathscr{W}_n),f_{nm})$ the *Le Cam-Schwartz inverse limit uniformity*. Again with slight abuse of notation, we also denote by $\mathscr{W}_\infty$ the subspace uniformity on $\mathscr{P}$. The associated *Le Cam-Schwartz inverse limit topologies* on $\mathscr{P}^\infty$ and $\mathscr{P}$ are both denoted $\mathscr{T}_\infty$. The topology $\mathscr{T}_1$ on $\mathscr{P}$ plays a central role in many arguments in this book, and if no confusion can arise, will be called simply the *Le Cam-Schwartz topology*.

Another, much more common topology on a model for *i.i.d.* observations on a topological space $\mathscr{X}$ is defined as follows: we consider the collection $\mathscr{C}^b(\mathscr{X})$ of all bounded *continuous* maps $f : \mathscr{X} \mapsto \mathbb{R}$ to define a fundamental systems of entourages.

Definition C.5.7. Let $\mathscr{X}$ be a *completely regular space*. Consider the space $C^b(\mathscr{X})$ of all bounded, continuous $f : \mathscr{X} \to \mathbb{R}$ and consider the fundamental system of entourages on a subspace $\mathscr{P}$ of the space of all Borel probability measures on $\mathscr{X}$, obtained by choosing $k \geq 1$, and $f_1, \ldots, f_k \in C^b(\mathscr{X})$ to define,

$$W^C_{f_1,\ldots,f_k} = \left\{(P,Q) \in \mathscr{P} \times \mathscr{P} : \left|(P-Q)f_i\right| < 1, 1 \leq i \leq k\right\}.$$

These subsets of $\mathscr{P} \times \mathscr{P}$ form a fundamental system of entourages for the so-called *Prokhorov uniformity* $\mathscr{W}^C$ on $\mathscr{P}$. We call the associated topology *Prokhorov's weak topology* and denote it $\mathscr{T}_C$.

The topology $\mathscr{T}_C$ is referred to as the *tight topology* in [49], chapter XI. Completeness of the space of bounded, positive Borel measures on a Polish space $\mathscr{X}$ for Prokhorov's weak uniformity is the subject of theorem C.8.9.

Definition C.5.8. Let $\mathscr{X}$ be a *locally compact space*. Consider the space $\mathscr{K}(\mathscr{X})$ of all continuous $f : \mathscr{X} \to \mathbb{R}$ with compact support and consider the fundamental system of entourages on a subspace $\mathscr{P}$ of the space of all Borel probability measures on $\mathscr{X}$, obtained by choosing $k \geq 1$, and $f_1, \ldots, f_k \in \mathscr{K}(\mathscr{X})$ to define,

$$W^K_{f_1,\ldots,f_k} = \left\{(P,Q) \in \mathscr{P} \times \mathscr{P} : \left|(P-Q)f_i\right| < 1, 1 \leq i \leq k\right\}.$$

These subsets of $\mathscr{P} \times \mathscr{P}$ form a fundamental system of entourages for the so-called *vague uniformity* $\mathscr{W}^K$ on $\mathscr{P}$. We call the associated topology the *vague topology* and denote it $\mathscr{T}_K$.

On any locally compact space $\mathscr{X}$, the vague topology is coarser than Prokhorov's weak topology, unless $\mathscr{X}$ is compact. The vague topology on spaces of Borel measures on locally compact spaces is of central importance in appendix C.8. To conclude the present perspective, we focus on the differences between the topologies $\mathscr{T}_1$ and $\mathscr{T}_C$, as illustrated in the following example.

Example C.5.9. Suppose that we consider a topological space $\mathscr{X}$ with its Borel σ-algebra and we take the (deterministic) collection of all atomic measures $\mathscr{P} = \{\delta_x : x \in \mathscr{X}\}$ as our model. We identify $\mathscr{X}$ and $\mathscr{P}$ through the bijection, $\mathscr{X} \to \mathscr{P} : x \mapsto \delta_x$. Note that for every $x \in \mathscr{X}$, there exists a (measurable but, in general, discontinuous) f such that $f(x) = 1$ and $f(y) = 0$ for all $y \in [0,1]$, $y \neq x$. Conclude that $\mathscr{W}_1$ is the *discrete uniformity* on $\mathscr{P}$, and hence, so is $\mathscr{W}_\infty$. That means that any function $g : \mathscr{P} \to [0,1]$ is $\mathscr{W}_\infty$-uniformly-continuous. According to the *Le Cam-Schwartz theorem* this fact renders any pair of disjoint model subsets B, V (uniformly) testable, which is appropriate in deterministic setting. Note that the map $x \mapsto \delta_x$ is *not continuous* unless we equip $\mathscr{X}$ also with the discrete topology (and hence, *not* a parametrization *c.f.* the definition at the beginning of section 9.2).

By contrast, the map $x \mapsto \delta_x$ *is* continuous if we equip $\mathscr{P}$ with the $\mathscr{T}_C$ topology; in fact, if $\mathscr{X}$ is *completely regular*, the inverse mapping $\delta_x \mapsto x$ is also continuous and $\mathscr{X}$ and $\mathscr{P}$ are homeomorphic (see exercise 8.10.8). To extend this point, note that the convex hull of $\mathscr{P}$ is $\mathscr{T}_C$-dense in the space $\mathscr{M}^1(\mathscr{X},\mathscr{B})$ of all Borel probability measures on $(\mathscr{X},\mathscr{B})$, but not $\mathscr{T}_\infty$-dense unless $\mathscr{X}$ is countable.

C.6 Function spaces

We consider spaces of functions from the general perspective of Bourbaki (1989) [47], Ch. X, with special attention for the notions of completeness and compactness.

Let $\mathscr{X}$ be a topological space and let $(\mathscr{Y},\mathscr{W})$ be a Hausdorff uniform space. Consider the set $\mathscr{F}(\mathscr{X},\mathscr{Y})$ of all maps $f : \mathscr{X} \to \mathscr{Y}$. Denote the subspace of all continuous such f by $C(\mathscr{X},\mathscr{Y})$.

Definition C.6.1. Let $\mathscr{F}$ be a subspace of $\mathscr{F}(\mathscr{X},\mathscr{Y})$. Let Σ be a collection of subsets of $\mathscr{X}$; for every $S \in \Sigma$ and every entourage $W \in \mathscr{W}$, define,

$$V(S,W) = \big\{(f,g) \in \mathscr{F} \times \mathscr{F} : \forall_{x\in S}\,(f(x)),g(x)) \in V\big\}.$$

Finite intersections of the sets $V(S,W)$, $(S \in \Sigma, W \in \mathscr{W}\})$ form a fundamental system of entourages for a uniformity $\mathscr{V}_\Sigma$ on $\mathscr{F}$ that is called the *uniformity of Σ-convergence*. The uniform (or topological) space $(\mathscr{F},\mathscr{V}_\Sigma)$ is denoted $\mathscr{F}_\Sigma$.

If $\Sigma = \{\mathscr{X}\}$, the uniformity/topology is referred to as the *uniformity/topology of uniform convergence*; if $\Sigma = \{\{x\} : x \in \mathscr{X}\}$, the uniformity/topology is referred to as the *uniformity/topology of pointwise convergence*; if Σ consists of all compact subsets of $\mathscr{X}$, the uniformity/topology is referred to as the *uniformity/topology of compact convergence*, *etcetera*. If $\mathscr{Y}$ is Hausdorff and the sets of Σ cover $\mathscr{X}$, then $\mathscr{F}_\Sigma(\mathscr{X},\mathscr{Y})$ is Hausdorff. If $\mathscr{Y}_1,\mathscr{Y}_2$ are two uniform spaces and $h : \mathscr{Y}_1 \to \mathscr{Y}_2$ is uniformly continuous, then the map $\mathscr{F}_\Sigma(\mathscr{X},\mathscr{Y}_1) \to \mathscr{F}_\Sigma(\mathscr{X},\mathscr{Y}_2) : f \mapsto h \circ f$ is uniformly continuous. If $\mathscr{X}_1,\mathscr{X}_2$ are two topological spaces, with Σ_1 (resp. Σ_2) a collection of subsets of $\mathscr{X}_1$ (resp. of $\mathscr{X}_2$) and $g : \mathscr{X}_1 \to \mathscr{X}_2$ is such that, for any $S \in \Sigma_1$, $g(S)$ is contained in a finite union of sets from Σ_2, then the map $\mathscr{F}_\Sigma(\mathscr{X},\mathscr{Y}_1) \to \mathscr{F}_\Sigma(\mathscr{X},\mathscr{Y}_2) : f \mapsto f \circ g$ is uniformly continuous. Pointwise convergence plays an ultimate role when it comes to completeness in functions space: if $\mathscr{G}$ is a filter in $\mathscr{F}_\Sigma(\mathscr{X},\mathscr{Y})$, then $\mathscr{G}$ converges to f, if and only if, $\mathscr{G}$ is Cauchy for the Σ-uniformity and converges *pointwise* to f.

Theorem C.6.2. *(Completeneness of functions spaces)*
A subspace H of $\mathscr{F}_\Sigma(\mathscr{X},\mathscr{Y})$ is complete, if and only if, every filter in H that is Cauchy for the Σ-uniformity converges pointwise *to an $f \in H$ (for all points in $\cup\{S : S \in \Sigma\}$).*

If $S \subset \mathscr{F}_{\Sigma_1}(\mathscr{X},\mathscr{Y})$ is complete and $\Sigma_1 \subset \Sigma_2$, then S is complete also in $\mathscr{F}_{\Sigma_2}(\mathscr{X},\mathscr{Y})$.

Theorem C.6.3. *(Pointwise completeneness)*
Let H be a subset of $\mathscr{F}_\Sigma(\mathscr{X},\mathscr{Y})$. If, for all $x \in \cup\{S : S \in \Sigma\}$,

$$H(x) = \{f(x) : f \in H\},$$

is complete in $\mathscr{Y}$, then the closure $\bar{H}$ in $\mathscr{F}_\Sigma(\mathscr{X},\mathscr{Y})$ is complete.

If $\mathscr{Y}$ is complete, then $\mathscr{F}_\Sigma(\mathscr{X},\mathscr{Y})$ is complete.

Next, we restrict attention to *continuous* maps, *i.e.* we apply the preceding generalities to the subspaces $C_\Sigma(\mathscr{X},\mathscr{Y})$. Since pointwise limits of continuous functions may have points of discontinuity, completeness is no longer a straightforward property, unless we impose *uniform* convergence: if $\mathscr{Y}$ is complete, the space $C(\mathscr{X},\mathscr{Y})$ is complete for the uniformity of uniform convergence. (It is noted that on $C(\mathscr{X},\mathscr{Y})$ the uniformity of uniform convergence on a subset B is the same for all dense subsets B of $\mathscr{X}$.) By extension, the space $\tilde{C}_\Sigma(\mathscr{X},\mathscr{Y})$ of all functions $f : \mathscr{X} \to \mathscr{Y}$ with continuous restrictions $f\,\big|_S$ for all $S \in \Sigma$, is complete for the Σ-uniformity, if $\mathscr{Y}$ is complete. If $\mathscr{X}$ can be covered by open subsets U, each of which is contained in some $S \in \Sigma$, then $\tilde{C}_\Sigma(\mathscr{X},\mathscr{Y}) = C(\mathscr{X},\mathscr{Y})$ and $C_\Sigma(\mathscr{X},\mathscr{Y})$ is complete, if $\mathscr{Y}$ is complete. For example: if $\mathscr{X}$ is locally compact and $\mathscr{Y}$ is complete, then $C(\mathscr{X},\mathscr{Y})$ is complete for the uniformity of compact convergence.

Compactness in function spaces revolves around the notion of equi-continuity.

Definition C.6.4. Let $\mathscr{X}$ be a topological space and let $(\mathscr{Y},\mathscr{W})$ be a uniform space. A subset H of $\mathscr{F}(\mathscr{X},\mathscr{Y})$ is said to be *equi-continuous* at a point $x \in \mathscr{X}$, if for every $W \in \mathscr{W}$ there is a neighbourhood U of x, such that $(f(x),f(y)) \in W$ for all $y \in U$ and every $f \in H$. The subset H is said to be *equi-continuous*, if H is equi-continuous in every point of $\mathscr{X}$. If $(\mathscr{X},\mathscr{U})$ is a uniform space, a subset H of $\mathscr{F}(\mathscr{X},\mathscr{Y})$ is said to be *uniformly equi-continuous*, if for every $W \in \mathscr{W}$ there is a $U \in \mathscr{U}$, such that $(f(x),f(y)) \in W$ for all $(x,y) \in U$ and every $f \in H$.

Uniform equi-continuity of a subset H implies equi-continuity of H; (uniform) equi-continuity of H implies (uniform) continuity of each $f \in H$. If $\mathscr{X}$ is a compact space, every equi-continuous subset is uniformly equi-continuous. Subsets and finite unions of (uniformly) equi-continuous subsets are again (uniformly) equi-continuous. A subset H is equi-continuous at x, if and only if the closure of H in $\mathscr{F}(\mathscr{X},\mathscr{Y})$ with the topology of pointwise convergence is equi-continuous.

Proposition C.6.5. *Let $\mathscr{X}$ be a topological space, $\mathscr{Y}$ a uniform space and H an equi-continuous subset of $C(\mathscr{X},\mathscr{Y})$. If H is endowed with the topology of pointwise convergence, then the map $H \times \mathscr{X} \to \mathscr{Y} : (h,x) \mapsto h(x)$ is continuous.*

The above is reflected in more intuitive form as follows: if $h \to g$ in an equi-continuous subset and $x \to y$ in $\mathscr{X}$, then also $h(x) \to g(y)$. Equi-continuity also makes composition of functions continuous for the topology of pointwise convergence.

Proposition C.6.6. *Let $\mathscr{X}$ be a topological space, $\mathscr{Y},\mathscr{Z}$ uniform spaces and H an equi-continuous subset of $C(\mathscr{Y},\mathscr{Z})$. If we endow H, $C(\mathscr{X},\mathscr{Y})$ and $C(\mathscr{X},\mathscr{Z})$*

with the topology of pointwise convergence, the map $C(\mathscr{X},\mathscr{Y})\times H \to C(\mathscr{X},\mathscr{Z}) : (f,h)\mapsto h\circ f$ *is continuous.*

If $\mathscr{X}$ is a topological space, $\mathscr{Y}$ a uniform space and H is an equi-continuous subset of $\mathscr{F}(\mathscr{X},\mathscr{Y})$, the topologies of compact convergence, pointwise convergence and pointwise convergence on a dense subset are identical. Compact sets of continuous mappings are characterized as follows.

Theorem C.6.7. *(Ascoli-Arzelà)*
Let $\mathscr{X}$ *be a topological (resp. uniform) space and let* Σ *be a covering of* $\mathscr{X}$ *by compact (resp. pre-compact) subsets. Let* $\mathscr{Y}$ *be a uniform space and let* H *be a subset of* $\mathscr{F}(\mathscr{X},\mathscr{Y})$ *such that, for every* $S\in\Sigma$ *and every* $h\in H$*, the restriction* $h\big|_S$ *is continuous (resp. uniformly continuous). Then, for* H *to be pre-compact with respect to the uniformity of* Σ*-convergence, it is necessary and sufficient that,*

(i) for each $S\in\Sigma$*, the set of restrictions* $\{h\big|_S : h\in H\}$ *is equi-continuous (resp. uniformly equi-continuous) in* $\mathscr{F}(S,\mathscr{Y})$*;*
(ii) for each $x\in\mathscr{X}$*, the set* $H(x)$ *is pre-compact in* $\mathscr{Y}$*.*

If $\Sigma_1\subset\Sigma_2$ and the sets of Σ_1 cover $\mathscr{X}$, then compactness of $S\subset\mathscr{F}(\mathscr{X},\mathscr{Y})$ for Σ_1-convergence and Σ_1-convergence are equivalent, and on S, the Σ_1- and Σ_2-topologies are the same. Two straightforward corollaries are as follows.

Corollary C.6.8. *Let* $\mathscr{X}$ *be a topological space and let* $\mathscr{Y}$ *be a Hausdorff uniform space. Any equi-continuous* H *in* $C(\mathscr{X},\mathscr{Y})$ *such that* $H(x)$ *is relatively compact in* $\mathscr{Y}$ *for all* $x\in\mathscr{X}$*, is relatively compact for the topology of compact convergence.*

Corollary C.6.9. *Let* $\mathscr{X}$ *be a locally compact space, let* $\mathscr{Y}$ *be a Hausdorff uniform space and let* H *be a subset of* $C(\mathscr{X},\mathscr{Y})$*. Then* H *is relatively compact for the topology of compact convergence, if and only if* H *is equi-continuous and* $H(x)$ *is relatively compact in* $\mathscr{Y}$ *for all* $x\in\mathscr{X}$*.*

For example, if $\mathscr{X}$ is $\mathbb{R}^k$ and $\mathscr{Y}=\mathbb{R}^l$ for some $k,l\geq 1$, then any bounded, equi-continuous subset H of $C(\mathscr{X},\mathscr{Y})$ is relatively compact.

To conclude this section, we consider some cases specified by extra properties for the spaces $\mathscr{X}$ and $\mathscr{Y}$. If $\mathscr{Y}$ is a *metrizable* uniform space, the uniformity of uniform convergence is metrizable; if, in addition, $\mathscr{X}$ is σ*-compact*, then the uniformity of compact convergence is metrizable. If $\mathscr{X}$ is compact metrizable and $\mathscr{Y}$ is Polish, then $C(\mathscr{X},\mathscr{Y})$ is Polish. For a metric space $(\mathscr{Y},d)$, a mapping $f:\mathscr{X}\to\mathscr{Y}$ is said to be *bounded*, if $\sup\{d(f(x),f(y)) : x,y\in\mathscr{X}\}<\infty$. The space of all bounded (continuous) mappings is *clopen* in $\mathscr{F}(\mathscr{X},\mathscr{Y})$ (in $C(\mathscr{X},\mathscr{Y})$) for the topology of uniform convergence, and it is complete if $\mathscr{Y}$ is complete. If $\mathscr{X}$ is *locally compact*, $\mathscr{Y}$ is a metrizable uniform space and both $\mathscr{X}$ and $\mathscr{Y}$ are *second countable*, then the space $C(\mathscr{X},\mathscr{Y})$ is metrizable and second countable for the topology of compact convergence. It is noted that the space $C^b(\mathbb{R},\mathbb{R})$ of all bounded, continuous, real-valued functions on $\mathbb{R}$ is not second countable for the topology of uniform convergence. The subspace of all bounded continuous functions on $\mathbb{R}$ with a limit at infinity, however, *is* second countable.

C.7 Vector spaces and locally convex spaces

The fundamental concepts in the theory of functional analysis are usually introduced using complete normed and inner-product spaces, see for example Megginson (1998) [192]. While this approach exposes the extent of the theory on function spaces quite fully, the resulting perspective leaves something to be desired in terms of the generality that is required for other applications, most notably the theory of generalized functions (see *e.g.* [243]) and many aspects of the theory of (Borel and Radon) measures. The more general version of functional analysis that underpins the statistical mathematics of Le Cam (1986) [179] revolves around so-called topological vector spaces and locally convex spaces. Bourbaki (1987, 2004, 2010) [50, 48, 49] covers these subjects at a very formal level, while more accessible versions can be found in Schaefer (1999) [223], Rudin (1991) [221] and in Trèves (2006) [243]. We start this summary with a quick look at function spaces and then give the basic elements of vector spaces and locally convex spaces. For the sake of brevity and because they used only in passing, the more evolved theory of Riesz spaces (with a central role in [179]) is not discussed (see, however, Luxemburg and Zaanen (1971) [188] and Zaanen and Luxemburg (1983) [260]).

[...] A *linear space* (also, *vector space*) E is a space closed under the usual linear operations; a *topological vector space* is a vector space with a topology, in which the linear operations are continuous.

Definition C.7.1. A *semi-norm* on a (real or complex) linear space E is a map $p : E \to [0,\infty)$ such that,

(i) p is *sub-additive*: for all $x,y \in E$, $p(x+y) \leq p(x)+p(y)$,
(ii) p is *positively homogeneous*: for any (real or complex) λ and any $x \in E$, $p(\lambda x) = |\lambda| p(x)$.

A *norm* $\|\cdot\| : E \to [0,\infty)$ is a semi-norm with the additional property that $\|x\| = 0$ implies $x = 0$.

For a linear space E with a norm $\|\cdot\|$, it follows directly that $d(x,y) = \|x-y\|$ is a *metric* on E. Correspondingly the collection of all *norm-balls*, $B(x,r) = \{y \in E : \|y-x\| < r\}$, $(x \in E, r > 0)$, forms the basis for a metric topology on E called the *norm topology*.

Definition C.7.2. A *normed space* $(E, \|\cdot\|)$ is a linear space E with a norm $\|\cdot\| : E \to [0,\infty)$, equipped with the metric topology generated by the norm balls. A *Banach space* is a normed space that is complete for the norm topology.

A linear form on a vector space E is any linear map $f : E \to \mathbb{R}$.

Definition C.7.3. If E is a topological vector space and E^* denotes so-called *algebraic dual* of E, which is the vector space of all linear forms on E, then the linear subspace of those $f \in E^*$ that are continuous for the topology on E is called the *continuous dual* of E, denoted E'.

[...]

Two vector spaces E and F are *placed in dual correspondence* by a (real) bi-linear form $B : E \times F \to \mathbb{R}$.

Definition C.7.4. The weak topology $\sigma(E,F)$ is a topology on E, generated by a basis of open neighbourhoods of the form,

$$U_{e,f_1\dots f_k} = \left\{ e' \in E \,:\, \max_{1\le i\le k} |B(e-e',f)| < 1 \right\}$$

where $e \in E$, and $f_1,\dots,f_k \in F$. Similarly, the sets,

$$V_{f,e_1\dots e_k} = \left\{ f' \in F \,:\, \max_{1\le i\le k} |B(e,f-f')| < 1 \right\}$$

where $f \in F$, and $e_1,\dots,e_k \in E$, form a basis for the weak topology $\sigma(F,E)$ on F.

The continuous dual of a space E in dual correspondence with another space F, is $E' = F$. If a locally convex space E has the topology $\sigma(E,F)$ for some dual space F, then a subset A of E is called *bounded*, if $\{|B(e,f)| : e \in A\}$ is bounded in $\mathbb{R}$, for every f in F.

Proposition C.7.5. *A bounded subset of a weak space is pre-compact. A complete bounded subset of a weak space is compact.*

[...]

Definition C.7.6. Let E and F be two vector spaces in dual correspondence. For every subset M of E, the *polar (set)* is defined,

$$M^\circ = \left\{ y \in F : B(x,y) \ge -1, x \in M \right\}.$$

The *bi-polar (set)* of a subset M of E is defined,

$$M^{\circ\circ} = \left\{ x \in E : B(x,y) \ge -1\, y \in M^c irc \right\}.$$

Theorem C.7.7. *(Bi-polar theorem)theorem!bi-polar*
Let E and F be two vector spaces in dual correspondence. For every subset M of E, the bi-polar $M^{\circ\circ}$ equals the closed (for $\sigma(E,F)$) convex hull of $M \cup \{0\}$.

[...]

Definition C.7.8. A real topological vector space is a *locally convex space* if there exists a fundamental system of neighbourhoods of 0 consisting of convex sets.

A topological vector space E is a locally convex space, if and only if the topology on E is defined by a collection of semi-norms.

[...]

Theorem C.7.9. *(Hahn-Banach theorem (analytic))*
Let p be a semi-norm on a vector space E. Let V be a vector subspace of E and f a linear form on V such that, for all $y \in V$, $f(y) \le p(y)$. Then there exists a linear form h on E that extends f, such that $f(x) \le p(x)$ for all $x \in E$.

(See [50], Ch. II, § 3, No. 2, Theorem 1.)

Corollary C.7.10. *If a linear functional f on a subspace V of a normed space $(E, \|\cdot\|)$ is continuous, then there exists a continuous linear functional h that extends f to the whole space.*

The Hahn-Banach theorem can also be formulated in an equivalent, strictly geometric form.

Theorem C.7.11. *(Hahn-Banach theorem (geometric))*
Let A be an open convex non-empty subset of a topological vector space and let M be a non-empty linear variety that does not intersect A. Then there exists a closed hyperplane that contains M and does not intersect A.

(See [50], Ch. II, § 5, No. 1, Theorem 1.) In a locally convex space, any closed convex set is the intersection of all closed halfspaces that contain it.

[...]

Note that $\mathscr{P}^\infty$ may be replaced by its linear span $\mathscr{L}$, with straightforward extension of the definition of the entourages $W_{n,f_1,\ldots,f_k}$, without changing the conclusions above. In that case, also define the linear space $\mathscr{F}$ that consists of all maps $f : \mathscr{X}^\infty \to \mathbb{R}$ that are in the union of the images of the spaces $\mathscr{F}_n$ under the canonical embeddings of $\mathscr{F}_n$ in the space of all maps $X^\infty \to \mathbb{R}$. To define $\mathscr{F}$ topologically, we view $\{\mathscr{F}_n : n \geq 1\}$ as a system of locally convex spaces (by means of the collection semi-norms $p_\mu : \mathscr{F}_n \to \mathbb{R} : f \mapsto |\mu f|$ for $\mu \in \mathscr{L}$). The space $\mathscr{F}$ is the *direct* (or *inductive limit*) of the system $(\mathscr{F}_n, f_{nm})$, with canonical injections for all $n \leq m$, $f_{nm} : \mathscr{F}_n \to \mathscr{F}_m$ that are trivially continuous (see [50], Ch. II, § 4, No. 5, Example II). The locally convex spaces $\mathscr{L}$ and $\mathscr{F}$ are then placed in dual correspondence, via the bilinear form $B(\mu, f) = \mu f$. Particularly, the topology $\mathscr{T}_\infty$ on $\mathscr{L}$ associated with $\mathscr{W}_\infty$ coincides with the weak topology $\sigma(\mathscr{L}, \mathscr{F})$; the topology on the direct limit $\mathscr{F}$ is $\sigma(\mathscr{F}, \mathscr{L})$.

[...]

For the next theorem only, assume that $\mathscr{P}$ is norm-bounded collection of bounded, positive measures, dominated by a probability measure Q and represented as a family $\mathscr{P}_Q = \{dP/dQ : P \in \mathscr{P}\}$ in $L^1(Q)$. The continuous dual of $L^1(Q)$ is $L^\infty(Q)$ and the model $\mathscr{P}$ with the $\mathscr{T}_\infty$-topology is homeomorphic with $\mathscr{P}_Q$ as a subspace of $L^1(Q)$ with the weak topology. (see [76].)

Theorem C.7.12. (Dunford-Pettis) *Assume $\mathscr{P}_Q$ is a norm-bounded subset of $L^1(Q)$; $\mathscr{P}_Q$ is relatively weakly compact,* if and only if, *for every $\varepsilon > 0$ there is an $M > 0$ such that,*

$$\sup_{P \in \mathscr{P}} \int_{\{dP/dQ > M\}} \frac{dP}{dQ} dQ < \varepsilon,$$

that is, $\mathscr{P}_Q$ is uniformly Q-integrable.

It is shown in the proof of lemma 3 of section 17.5 of Le Cam (1986) [179] (in the somewhat broader context of theorem 6 of appendix 8 in [179]) that weak convergence of a net $f_\alpha \to f$ in $L^1(Q)$ implies weak convergence of product densities $f_\alpha^n \to f^n$ weakly in $L^1(Q^n)$, as a result of the Dunford-Pettis theorem (see also

lemma 3.8 in [236]). Consequently, a net in $\mathscr{P}$ that has a $\mathscr{T}_1$-convergent subnet, also has a $\mathscr{T}_n$-convergent subnet, so $\mathscr{T}_1$-compactness implies $\mathscr{T}_n$-compactness for all $n \geq 1$, which implies $\mathscr{T}_\infty$-compactness.

Proposition C.7.13. *Let $\mathscr{P}$ be a model for i.i.d. data X^n; $\mathscr{P}$ is $\mathscr{T}_1$-compact, if and only if, $\mathscr{P}$ is $\mathscr{T}_\infty$-compact.*

[...]

(Relative) compactness of the model $\mathscr{P}$ for $\mathscr{T}_C$ is the realm of Prokhorov's theorem (see Prokhorov (1956) [211]). Here it is assumed that $\mathscr{X}$ is a Hausdorff topological space with Borel σ-algebra. In [47], Ch. IX, § 5, No. 5, the following is referred to as *Prokhorov's property*.

Definition C.7.14. Let H be a subset of $M(\mathscr{X})$; H is said to be *uniformly tight* if,

(i) $\sup\{\|\mu\| : \mu \in H\} < \infty$,
(ii) for every $\varepsilon > 0$, there is a compact K in $\mathscr{X}$ such that,

$$\sup\{|\mu(\mathscr{X} \setminus K)| : \mu \in H\} \leq \varepsilon.$$

For probability models $\mathscr{P}$ the uniform bound in norm is always satisfied and only the second condition plays a role when one verifies relative compactness for $\mathscr{T}_C$.

Theorem C.7.15. *Assume that $\mathscr{X}$ is completely regular. A subset H of $M(\mathscr{X})$ that is uniformly tight, is H relatively compact for $\mathscr{T}_C$.*

Proof. For a proof, see [47], Ch. IX, § 5, No. 5, theorem 1.

In locally compact or Polish spaces, uniform tightness is equivalent to $\mathscr{T}_C$-relative compactness.

Theorem C.7.16. *(Prokhorov)*
Assume that $\mathscr{X}$ is locally compact or Polish. A subset H of $M(\mathscr{X})$ that is relatively compact for $\mathscr{T}_C$, is uniformly tight.

Proof. For a proof, see [47], Ch. IX, § 5, No. 5, theorem 2.

Note that also regarding matters of compactness, the $\mathscr{T}_1$ and $\mathscr{T}_C$ topologies are different in that the $\mathscr{T}_C$ compactness criterion refers to a topological feature of the sample space (the compact subset K of $\mathscr{X}$), while the $\mathscr{T}_1$ compactness criterion does not and is formulated as a property that derives from $\mathscr{X}$ as a measurable space (uniform integrability). The associated strong topologies also maintain a distinction of the type.

Proposition C.7.17. *The strong topologies associated with $\mathscr{T}_1$ and $\mathscr{T}_\infty$ are equal to the total-variational topology. The strong topology associated with $\mathscr{T}_C$ is $\mathscr{T}_C$ itself.*

To conclude with an example, we consider a sequence (P_n) of probability measures that converges in $\mathscr{T}_C$ but not $\mathscr{T}_1$. The example also shows how $\mathscr{T}_C$-compact sets can be non-compact for $\mathscr{T}_1$.

Example C.7.18. Consider $\mathscr{X} = [0,1]$ with the Borel σ-algebra with distributions P_n defined by their Lebesgue measures p_n for all $n \geq 1$, $p_n(x) = n\,1\{0 \leq x \leq 1/n\}$. For any continuous $g : [0,1] \to \mathbb{R}$,

$$\inf_{0 \leq x \leq 1/n} g(x) \leq \int_0^1 g(x)\, dP_n(x) \leq \sup_{0 \leq x \leq 1/n} g(x),$$

and both bounds go to $g(0)$ as $n \to \infty$, so $P_n \to \delta_0$ in $\mathscr{T}_C$. However, the collection $\{P_n : n \geq 1\}$ does not satisfy the condition of theorem C.7.12, so (P_n) does not converge for $\mathscr{T}_1$.

C.8 Radon measures

Radon measures are best viewed as continuous linear forms on spaces of bounded, continuous functions on a topological space. They can be identified as Borel measures as in appendix B, with extra properties. Although abstract measure theory is sufficient for most applications in probability theory and statistics, certain important aspects of abstract Borel measures, like the support of a Borel measure (see definition C.1.18), remain nebulous and are insufficient from the perspective of (functional) analysis. Excellent references for the theory of Radon measures is Schwartz (1973) [227] and Bourbaki (1998,1989) [48, 49]. (Below, we do not consider complex measures, all measures are real-valued, signed measures.)

The following definition starts from the perspective that a Radon measure is an abstract measure defined on a Borel σ-algebra with additional properties related to compactness and the support of a measure.

Definition C.8.1. Given a Hausdorff topological space $\mathscr{X}$, a *Radon measure* Π is a Borel measure that is:

(i) *locally bounded*: any point in $\mathscr{X}$ has a neighbourhood U such that $\Pi(U) < \infty$;
(ii) *inner regular*: for any open subset $U \subset \mathscr{X}$ and any $\varepsilon > 0$, there exists a compact $K \subset U$ such that $\mu(U \setminus K) < \varepsilon$;
(iii) *outer regular*: for any Borel $B \subset \mathscr{X}$ and any $\varepsilon > 0$, there exists an open $V \subset \mathscr{X}$ such that $\mu(V \setminus B) < \varepsilon$.

This definition does no do justice to the real intention, however. To appreciate the concept more appropriately, we first observe the following well-known way to represent elements of the dual of the space $C(K)$ of continuous functions on a fixed compact domain K (see, for example, Dunford and Schwartz (1988) [82]).

Theorem C.8.2. *(Riesz representation)*
Let K be a compact space and I a continuous linear form on the normed space $C(K)$. Then there exists a bounded Borel measure μ such that $I(f) = \int f\, d\mu$ for all $f \in C(K)$.

Note that the measure μ is, in fact, a Radon measure (see exercise 4.4.10). It is this functional-analytic representation that we intend to use as a starting point. But compactness is, of course, a rather restrictive condition on sample spaces and even more on the (often infinite-dimensional) parameter spaces that play a role in Bayesian statistics, so there exists a clear need to generalize. We start the generalized construction with the assumption that $\mathscr{X}$ is locally compact and the specification to Radon measures becomes relevant (see Bourbaki (1998) [48]).

Let $C(\mathscr{X})$ denote the vector space of all continuous real-valued functions on a *locally compact space* $\mathscr{X}$ and let $\mathscr{K}(\mathscr{X})$ denote the linear subspace of continuous real-valued functions on $\mathscr{X}$ with compact supports; for compact $K \subset \mathscr{X}$, denote by $\mathscr{K}(\mathscr{X},K)$ the linear subspace of continuous real-valued functions on $\mathscr{X}$ with compact support contained in K. Note that on $\mathscr{K}(\mathscr{X},K)$, the *uniform norm* $f \mapsto \|f\|_K = \sup\{|f(x)| : x \in K\}$ is well-defined. Note also that if $K,K' \subset \mathscr{X}$ are compact and $K \subset K'$, then there is a natural norm-to-norm continuous embedding of $\mathscr{K}(\mathscr{X},K)$ into $\mathscr{K}(\mathscr{X},K')$, with corresponding *direct limit* space that is identified with $\mathscr{K}(\mathscr{X})$ and each $\mathscr{K}(\mathscr{X},K)$ corresponds to a closed linear subspace with the topology generated by the norm $\|\cdot\|_K$ (see [48], Ch. III, § 1, No. 1, proposition 1). (Note, however, that in general the direct-limit topology on $\mathscr{K}(\mathscr{X})$ is weaker than the topology generated by the uniform norm $\|f\| = \sup\{|f(x)| : x \in \mathscr{X}\}$.) The space $\mathscr{K}(\mathscr{X})$ is *barrelled* (since a direct limit of barrelled spaces is again a barrelled space, see [50], Ch. III, §), No. 4, 1corollary 5 of proposition 3). Compactness of a subset H of $\mathscr{K}(\mathscr{X})$ is characterized by the Ascoli-Arzelà theorem ([47], Ch. X, § 2, No. 5, corollary 3 of theorem 2): H is compact, if and only if H is closed and equi-continuous. This gives rise to the following alternative definition of a Radon measure (which can be shown to be equivalent to definition C.8.1).

Theorem C.8.3. *Given a locally compact Hausdorff space $\mathscr{X}$, a continuous linear form I on $\mathscr{K}(\mathscr{X})$ is a (real, signed) Radon measure. If $I(f) \geq 0$ for all $f \in \mathscr{K}(\mathscr{X})$ such that $f \geq 0$, then I is a positive Radon measure.*

Every signed Radon measure I can be written as the difference of two positive Radon measures I_+, I_-: $I = I_+ - I_-$; the *absolute value* of I is the positive measure $|I| = I_+ + I_-$. By the characterization of continuous linear forms on direct limit spaces (see [50], Ch. II, § 4, No. 4, proposition 5), a linear form I on $\mathscr{K}(\mathscr{X})$ is a Radon measure, if and only if for every compact K there exists an $M_K > 0$ such that for all $f \in \mathscr{K}(\mathscr{X},K)$.:

$$|I(f)| \leq M_K \|f\|_K.$$

(Indeed, this property has to be shown only for a collection of compacta whose interiors cover $\mathscr{X}$.) The vector space of all Radon measures on $\mathscr{X}$ is therefore identified as the dual of $\mathscr{K}(\mathscr{X})$. To make clear notational distinction between spaces of Borel measures and the spaces of abstract measures of appendix B, we denote this space by $M(\mathscr{X})$. Positive measures in $M(\mathscr{X})$ form a subset of $M(\mathscr{X})$ that we denote by $M_+(\mathscr{X})$.

Definition C.8.4. Given a locally compact Hausdorff space $\mathscr{X}$, a linear form I on $\mathscr{K}(\mathscr{X})$ that is continuous for the uniform norm $\|\cdot\|$ is a *bounded Radon measure*. We denote the linear space of all bounded Radon measures on $\mathscr{X}$ by $M^b(\mathscr{X})$.

A linear form on is a *bounded* Radon measures on $\mathscr{X}$, if and only if there exists an $M > 0$ such that for all compact $f \in \mathscr{K}(\mathscr{X})$,

$$|I(f)| \leq M\|f\|.$$

Clearly, if $\mathscr{X}$ is compact, every Radon measure is bounded. If P is a positive, bounded Radon measure such that the smallest M like above is $M = 1$, then we say that P is a *Radon probability measure*. We denote the subset of all probability measures with $M^1(\mathscr{X})$.

Mostly we shall consider $\mathscr{K}(\mathscr{X})$ and $M(\mathscr{X})$ in *duality*.

Definition C.8.5. Consider the fundamental system of entourages obtained by choosing $k \geq 1$, and $f_1, \ldots, f_k \in \mathscr{K}(\mathscr{X})$ to define,

$$W_{f_1,\ldots,f_k} = \left\{(\mu,\nu) \in M(\mathscr{X}) \times M(\mathscr{X}) : \left|(\mu-\nu)f_i\right| < 1, 1 \leq i \leq k\right\}.$$

These subsets of $\mathscr{P}^\infty \times \mathscr{P}^\infty$ form a fundamental system of entourages for the *vague uniformity* $\mathscr{W}_K$ on $M(\mathscr{X})$, with corresponding *vague topology* $\mathscr{T}_K$.

It is noted that $M(\mathscr{X})$ is Hausdorff and that every closed, bounded subset of $M(\mathscr{X})$ is complete (see [48], Ch. III, § 1, No. 3, proposition 7).

In the case of arbitrary (that is, non-locally-compact) Hausdorff spaces, the identification of Radon measures and continuous linear forms is slightly more involved. In the definitions, we replace the direct limit space $\mathscr{K}(\mathscr{X})$ in the above, by the space of all bounded, continuous functions.

Definition C.8.6. For a Hausdorff topological space $\mathscr{X}$, $C^b(\mathscr{X})$ denotes the linear space of all *bounded, continuous* $f : \mathscr{X} \to \mathbb{R}$, equipped with the uniform norm.

On completely regular spaces, bounded Radon measures are identified with elements of a dual (now of $C^b(\mathscr{X})$ with the uniform norm $\|\cdot\|$), while the compactness requirement is additional (see [49], Ch. IX, § 5, No. 2, proposition 5).

Theorem C.8.7. *(Riesz-Markov-Kakutani)*
Let $\mathscr{X}$ be a completely regular space and I a continuous linear form on the normed space $C^b(\mathscr{X})$. In order that there exist a bounded Radon measure μ such that $I(f) = \int f\,d\mu$, for all $f \in C^b(\mathscr{X})$, it is necessary and sufficient that,

(R) for every $\varepsilon > 0$, there is a compact K in $\mathscr{X}$, such that $\|g\| \leq 1$, $g\,\big|_K = 0$ imply $|I(g)| \leq \varepsilon$.

Consider the linear space of bounded Radon measures $M^b(\mathscr{X})$ and the linear space of bounded, continuous, real-valued maps $C^b(\mathscr{X})$ as a dual pair with bi-linear form $\langle \mu, g\rangle = \int g\,d\mu$, for all $\mu \in M(\mathscr{X})$ and $g \in C^b(\mathscr{X})$. Re-phrased in the terminology of uniform spaces, we define this as follows.

Definition C.8.8. Let $\mathscr{X}$ be a *completely regular space* and let $M^b(\mathscr{X})$ denote the topological vector space of Radon measures on $\mathscr{X}$. Consider the fundamental system of entourages obtained by choosing $k \geq 1$, and $f_1, \ldots, f_k \in C^b(\mathscr{X})$ to define,

$$W_{f_1,\ldots,f_k} = \left\{(\mu,\nu) \in M^b(\mathscr{X}) \times M^b(\mathscr{X}) : \left|(\mu-\nu)f_i\right| < 1, 1 \le i \le k\right\}.$$

These subsets of $\mathscr{P}^\infty \times \mathscr{P}^\infty$ form a fundamental system of entourages for the *tight uniformity* $\mathscr{W}_c$ on $M^b(\mathscr{X})$, with corresponding *tight topology* $\mathscr{T}_C$, which is equal to the topology that we have called Prokhorov's weak topology so far.

The space $M^b(\mathscr{X})$ is a *completely regular space*.

Theorem C.8.9. *If $\mathscr{X}$ is a Polish space, $M^b_+(\mathscr{X})$ is a Polish space.*

(see Proposition 10 of [49], Ch. IX, § 5, No. 4.) It is noted that the above theorem does not imply that $M^b(\mathscr{X})$ is a Polish space.

Definition C.8.10. A topological space $\mathscr{X}$ has the *Radon property* (also, $\mathscr{X}$ is said to be a *Radon space*) if every Borel measure on $\mathscr{X}$ is a Radon measure.

Theorem C.8.11. *All Souslin spaces are Radon spaces.*

(see [49], Ch. IX, § 3, No. 1, proposition 3) or [227], Ch. 2, theorem 10). Polish spaces are Souslin spaces, so all Polish spaces have the Radon property.

Definition C.8.12. Lusin-measurablility

If $\mathscr{X},\mathscr{Y}$ are Souslin spaces, μ is a Radon measure on $\mathscr{Y}$ and $f : \mathscr{X} \to \mathscr{Y}$ is surjective and continuous, then there exists a Lusin μ-measurable $g : \mathscr{Y} \to \mathscr{X}$ such that $f \circ g$ is the identity on $\mathscr{Y}$.

C.9 Convergence in spaces of probability measures

Let $M^1(\mathbb{R})$ denote the space of all Borel probability measures on $\mathbb{R}$.

Definition C.9.1. *(vague topology)*
Let (Q_n) and Q in $M^1(\mathbb{R})$ be given. We say that Q_n converges vaguely to Q if for every continuous $f : \mathbb{R} \to \mathbb{R}$ with compact support, $Q_n f \to Qf$.

Definition C.9.2. *(Prokhorov's weak topology)*
Let (Q_n) and Q in $M^1(\mathbb{R})$ be given. We say that Q_n converges weakly to Q if for every bounded, continuous $f : \mathbb{R} \to \mathbb{R}$, $Q_n f \to Qf$.

Trivially, if Q_n converges to Q and $f : \mathbb{R} \to \mathbb{R}$ is continuous, then $Q_n \circ f^{-1}$ converges to $Q \circ f^{-1}$, for Prokhorov's weak topology; for sequences of real-valued random variables $X_n \sim P_n$ converging to $X \sim Q$, this amounts to $f(X_n)$ converging to $f(X)$, a result known as the *continuous mapping theorem*. Relative compactness in Prokhorov's weak topology is characterized as uniform tightness (for every $\varepsilon > 0$, there is an $M > 0$ such that for all $n \ge 1$, we have $P_n(|X_n| > M) < \varepsilon$), which we also indicate with the *stochastic order symbol* $X_n = O_{P_n}(1)$. The other stochastic order symbol signifies convergence to zero in probability (for all $\delta, \varepsilon > 0$, there is

an $n \geq 1$ such that for all $m \geq n$, $P_n(|X_n| > \delta) < \varepsilon$), and is indicated $X_n = o_{P_n}(1)$. When two sequences of random variables $(X_n), (Y_n)$ are involved, $X_n = O_{P_n}(Y_n)$ (or $X_n = o_{P_n}(Y_n)$) simply means that $Y_n^{-1} X_n = O_{P_n}(1)$ (or $Y_n^{-1} X_n = o_{P_n}(1)$).

Convergence in Prokhorov's weak topology has several equivalent formulations.

Lemma C.9.3. *Portmanteau lemma*
Let (Q_n) and Q in $M^1(\mathbb{R})$ be given. The following are equivalent:

(i) Q_n converges weakly to Q.
(ii) For all $t \in C$, $Q_n(-\infty, t] \to Q(-\infty, t]$, where C denotes the set of continuity points of $\mathbb{R} \to [0,1] : t \mapsto Q(-\infty, t]$.
(iii) For every bounded, Lipschitz $g : \mathbb{R} \to \mathbb{R}$, $Q_n g \to Qg$.
(iv) For all non-negative, continuous $h : \mathbb{R} \to \mathbb{R}$, $\liminf_{n\to\infty} Q_n f \geq Qf$.
(v) For every open set $F \subset \mathbb{R}$, $\liminf_{n\to\infty} Q_n(F) \geq Q(F)$.
(vi) For every closed set $G \subset \mathbb{R}$, $\limsup_{n\to\infty} Q_n(G) \leq Q(G)$.
(vii) For every Borel set B such that $Q(\partial B) = 0$, $Q_n(B) \to Q(B)$.

In *(vii)* above, ∂B denotes the boundary of B, which is defined as the closure of B minus the interior of B.

Proposition C.9.4. *Let $\mathscr{Y}$ be a Polish space and let $\mathscr{B}$ be its Borel σ-algebra. With the topology $\mathscr{T}_C$ of weak convergence, the space $M_+^b(\mathbb{R})$ is Polish. Since the space of probability measures $M^1(\mathbb{R})$ is a closed subset of $M_+^b(\mathbb{R})$, $M^1(\mathbb{R})$ is also Polish.*

Proof. See theorem (17.23) in [143].

Definition C.9.5. *(Le Cam-Schwartz topology)*
Let Q in $M^1(\mathbb{R})$ and a net (Q_α) in $M^1(\mathbb{R})$ be given. We say that Q_α converges to Q in the Le Cam-Schwartz topology, if for every bounded, measurable $f : \mathbb{R} \to \mathbb{R}$, $Q_\alpha f \to Qf$.

See also definition C.9.5.

Definition C.9.6. *(topology of pointwise convergence)*
Let (Q_n) and Q in $M^1(\mathbb{R})$ be given. We say that Q_n converges pointwise to Q if, for all $B \in \mathscr{B}$, $Q_n(B) \to Q(B)$.

Definition C.9.7. *(topology of total variation)*
Let (Q_n) and Q in $M^1(\mathbb{R})$ be given. We say that Q_n converges in total variation to Q if,

$$\|Q_n(B) - Q(B)\| = \sup_{B \in \mathscr{B}} \left| Q_n(B) - Q(B) \right| \to 0.$$

In exercise **??**, it is shown that this distance can also be calculated as the L_1-difference between densities for Q_n and Q.

Lemma C.9.8. *(Scheffé's lemma)*
Let $(\Omega, \mathscr{F}, \mu)$ be a measure space. Given a sequence (f_n) of integrable functions and a measurable function f such that $f_n(\omega) \to f(\omega)$ for μ-almost-all ω, then $\int |f_n - f| \, d\mu \to 0$ if and only if $\int |f_n| \, d\mu \to \int |f| \, d\mu$.

Corollary C.9.9. *Let $(\Omega, \mathscr{F}, \mu)$ be a measure space. Let p, p_n $(n \geq 1)$ be probability densities with respect to μ. If $p_n(\omega) \to p(\omega)$ for μ-almost-all $\omega \in \Omega$), then $\|P_n - P\| \to 0$.*

C.10 Contiguity

First, let us recall the definition of contiguity [174] (see [179] for alternatives, *e.g.* in terms of limiting domination in a sequence of binary experiments).

Definition C.10.1. Given measurable spaces $(\mathscr{X}_n, \mathscr{B}_n)$, $n \geq 1$ with two sequences (P_n) and (Q_n) of probability measures, we say that Q_n is *contiguous with respect to* P_n, notation $Q_n \lhd P_n$, if,

$$P_n \phi_n(X^n) = o(1) \quad \Rightarrow \quad Q_n \phi_n(X^n) = o(1), \tag{C.1}$$

for every sequence of $\mathscr{B}_n$-measurable $\phi_n : \mathscr{X}_n \to [0,1]$.

The value of the notion of contiguity does not just reside with the usefulness of the property itself, but also with the multitude of accessible characterizations listed in Le Cam's famous First Lemma (see, *e.g.*, Hajék and Šidák (1967) [120]). (One of the formulations requires that we define the so-called Hellinger transform $\psi(P, Q; \alpha) = \int p^\alpha q^{1-\alpha} d\mu$, where p and q denote densities for P and Q with respect to a σ-finite measure that dominates both P and Q.)

Lemma C.10.2. *(Le Cam's First Lemma)*
Given measurable spaces $((\mathscr{X}_n, \mathscr{B}_n) : n \geq 1)$ with two sequences (P_n) and (Q_n) of probability measures, the following are equivalent:

(i) $Q_n \lhd P_n$,
(ii) for any measurable $T_n : \mathscr{X}_n \to \mathbb{R}$, *if* $T_n \xrightarrow{P_n} 0$, *then* $T_n \xrightarrow{Q_n} 0$,
(iii) given $\varepsilon > 0$, *there is a* $b > 0$ *such that* $Q_n(dQ_n/dP_n > b) < \varepsilon$, *for large enough* n,
(iv) given $\varepsilon > 0$, *there is a* $c > 0$ *such that* $\|Q_n - Q_n \wedge c P_n\| < \varepsilon$, *for large enough* n,
(v) if $dP_n/dQ_n \xrightarrow{Q_n\text{-w.}} f$ *along a subsequence, then* $P(f > 0) = 1$,
(vi) if $dQ_n/dP_n \xrightarrow{P_n\text{-w.}} g$ *along a subsequence, then* $Eg = 1$,
(vii) Hellinger transforms satisfy, $\liminf_n \lim_{\alpha \uparrow 1} \psi(P_n, Q_n; \alpha) = 1$.

A proof of this form of the First Lemma can be found in [179], section 6.3. Note the relation to testing: for two sequences (P_n), (Q_n) that are mutually contiguous ($P_n \lhd Q_n$ and $Q_n \lhd P_n$), there exists no test sequence that separates (P_n) from (Q_n) asymptotically. Loosely said, (P_n) and (Q_n) are indistinguishable statistically regardless of the amount of data available. Much more can be said about contiguity (to begin with, see, Roussas (1972) [220] and Greenwood and Shiryaev (1985) [116]), for instance in relation to Le Cam's convergence of experiments, but also, specific

relations that exist in the locally asymptotically normal case (*e.g.* Le Cam's Third lemma [120], which relates the laws of a statistic under P_n and Q_n in such context).

Appendix D
Inverse limit measures

D.1 Inverse limits of positive measures

In this section, we assume that $\mathscr{X}$ is a a Hausdorff completely regular space and denote by $\mathscr{A}$ a collection of finite partitions of $\mathscr{X}$ (into non-empty sets).

Definition D.1.1. A collection $\mathscr{A}$ is said to be *separating* if, for every $x, y \in \mathscr{X}$, $x \neq y$, there exist $\alpha \in \mathscr{A}$ and $A \in \alpha$ such that $x \in A$ and $y \notin A$.

For any $\alpha \in \mathscr{A}$, let $N(\alpha) \in \mathbb{N}$ denote its cardinal, let $I(\alpha)$ denote the set $\{1, \ldots, N(\alpha)\}$ and write $\alpha = (A_1, \ldots, A_{N(\alpha)})$ for non-empty A_i ($1 \leq i \leq N(\alpha)$). For every $\alpha \in \mathscr{A}$, let $\mathscr{X}_\alpha$ denote the discrete space of unit vectors in $\mathbb{R}^N(\alpha)$, $\mathscr{X}_\alpha = \{e_1, \ldots, e_{N(\alpha)}\} \subset \mathbb{R}^{N(\alpha)}$ (where e_i denotes the i-th unit vector in $\mathbb{R}^{N(\alpha)}$) and,

$$\varphi'_\alpha : \mathscr{X} \to \mathscr{X}_\alpha : x \mapsto (1_{A_1}(x), \ldots, 1_{A_{N(\alpha)}}(x)).$$

Definition D.1.2. Given a space $\mathscr{X}$ and a collection of spaces $\mathscr{X}_\alpha$, $\alpha \in \mathscr{A}$, a collection of functions $\varphi_\alpha : \mathscr{X} \to \mathscr{X}_\alpha$ is said to be *separating*, if for all $x, y \in \mathscr{X}$, $x \neq y$, there exist an $\alpha \in \mathscr{A}$ and such that $\varphi_\alpha(x) \neq \varphi_\alpha(y)$.

We assume that $\mathscr{A}$ is a *directed set* with respect to the natural pre-order: for $\alpha, \beta \in \mathscr{A}$, $\alpha \leq \beta$ whenever β refines α. Let $\alpha, \beta \in \mathscr{A}$ be such that $\alpha \leq \beta$. Denote $\alpha = \{A_1, \ldots, A_{N(\alpha)}\}$ and $\beta = \{B_1, \ldots, B_{N(\beta)}\}$ and for every $1 \leq i \leq N(\alpha)$, let $J_{\alpha\beta}(i) \subset \{1, \ldots, N(\beta)\}$ be such that $A_i = \cup_{j \in J_{\alpha\beta}(i)} B_j$. Equivalently, this can be expressed through a mapping $i_{\alpha\beta} : I(\beta) \to I(\alpha)$ such that $i_{\alpha\beta}(j) = i$ whenever $j \in J_{\alpha\beta}(i)$. Based on that, define $\varphi_{\alpha\beta} : \mathscr{X}_\beta \to \mathscr{X}_\alpha$ to be the map that takes e_j into e_i whenever $j \in J_{\alpha\beta}(i)$. Note that $\varphi_{\alpha\beta}(e_j) = e_{i_{\alpha\beta}(j)}$ for all $1 \leq j \leq N(\beta)$. The maps $\varphi_{\alpha\beta}$ are coherent (in the sense that for $\alpha \leq \beta \leq \gamma$, $\varphi_{\alpha\beta} \circ \varphi_{\beta\gamma} = \varphi_{\alpha\gamma}$) and (since the spaces $\mathscr{X}_\alpha$ are discrete) continuous. The projection maps φ'_α form a coherent system, *i.e.* for all $\alpha, \beta \in \mathscr{A}$, if $\alpha \leq \beta$

$$\varphi_{\alpha\beta} \circ \varphi'_\beta = \varphi'_\alpha.$$

Next note that any $g : \mathscr{X}_\alpha \to \mathbb{R}$ is continuous: such g lie in the (finite-dimensional) Banach space $C(\mathscr{X}_\alpha)$ of all continuous, real-valued maps with the uniform norm, with continuous dual we denote by $M(\mathscr{X}_\alpha)$, the finite-dimensional space of all finite, signed measures on $\mathscr{X}_\alpha$, which is a Banach space with the total-variational norm. For $\mu \in M(\mathscr{X}_\alpha)$ and $g \in C(\mathscr{X}_\alpha)$, $\int g\, d\mu$ is denoted in bi-linear form $\langle \mu, g\rangle_\alpha$.

Let $\alpha, \beta \in \mathscr{A}$ such that $\alpha \leq \beta$ be given. For any $g \in C(\mathscr{X}_\alpha)$, the map $g \circ \varphi_{\alpha\beta} : \mathscr{X}_\beta \to \mathbb{R}$ is an element of $C(\mathscr{X}_\beta)$. Because $\varphi_{\alpha\beta}$ is surjective, the induced map $\varphi^*_{\alpha\beta} : C(\mathscr{X}_\alpha) \to C(\mathscr{X}_\beta)$ is a bounded linear operator with norm equal to one. The transpose map $\varphi_{*\,\alpha\beta} : M(\mathscr{X}_\beta) \to M(\mathscr{X}_\alpha)$ is defined by,

$$\langle \varphi_{*\,\alpha\beta}(\mu), g\rangle_\alpha = \langle \mu, \varphi^*_{\alpha\beta}(g)\rangle_\beta = \langle \mu, g \circ \varphi_{\alpha\beta}\rangle_\beta$$

for all $\mu \in M(\mathscr{X}_\beta)$ and $g \in C(\mathscr{X}_\alpha)$. The linear map $\varphi_{*\,\alpha\beta}$ is bounded with norm less than or equal to one. Note that if we express $\mu \in M(\mathscr{X}_\alpha)$ as a vector $(\mu_1, \ldots, \mu_{N(\alpha)})$ in $\mathbb{R}^{N(\alpha)}$,

$$\begin{aligned}\langle \mu, g \circ \varphi_{\alpha\beta}\rangle_\beta &= \sum_{j \in I(\beta)} \mu_j\, g(\varphi_{\alpha\beta}(e_j)) \\ &= \sum_{j \in I(\beta)} \mu_j\, g(\varphi_{\alpha\beta}(e_{i_{\alpha\beta}(j)})) = \sum_{i \in I(\alpha)} \Big(\sum_{j \in J(i)} \mu_j \Big) g(e_i),\end{aligned}$$

from which one determines that,

$$\varphi_{*\,\alpha\beta}(\mu)_i = \sum_{j \in J_{\alpha\beta}(i)} \mu_j, \tag{D.1}$$

for all $1 \leq i \leq N(\alpha)$.

Definition D.1.3. Any collection of Borel measures $\mu_\alpha \in M(\mathscr{X}_\alpha)$ on a coherent system of topological spaces $(\mathscr{X}_\alpha, \varphi_{\alpha\beta})$, such that for all $\alpha, \beta \in \mathscr{A}$ with $\alpha \leq \beta$,

$$\mu_\alpha = \varphi_{*\,\alpha\beta} \circ \mu_\beta, \tag{D.2}$$

is called a (coherent) *inverse system of measures* $(\mu_\alpha, \varphi_{*\,\alpha\beta})$ on the inverse system of spaces $(\mathscr{X}_\alpha, \varphi_{\alpha\beta})$.

The relations (D.2) express coherency, in the same role as the Kolmogorov consistency relations that play a role in the proof of theorem 8.2.1. But here, we consider (D.2) from the dual point of view: for all $\alpha, \beta, \gamma \in \mathscr{A}$ such that $\alpha \leq \beta \leq \gamma$, $\varphi_{*\,\alpha\beta} \circ \varphi_{*\,\beta\gamma} = \varphi_{*\,\alpha\gamma}$ and $\varphi_{*\,\alpha\alpha}$ is the identity on $M(\mathscr{X}_\alpha)$, so $(M(\mathscr{X}_\alpha), \varphi_{*\,\alpha\beta})$ is an inverse system of (non-empty, compact) topological spaces, with (non-empty, compact) inverse limit N. And, again, all $\mu \in M(\hat{\mathscr{X}})$ map to points in N, but not all points in N correspond to bounded Radon measures on $\hat{\mathscr{X}}$. To relate N and $M(\hat{\mathscr{X}})$ directly, consider for $\alpha \in \mathscr{A}$, with $\alpha = (A_1, \ldots, A_{N(\alpha)})$, the mapping $\hat{\varphi}_{*\,\alpha} : M(\hat{\mathscr{X}}) \to M(\mathscr{X}_\alpha)$,

$$\hat{\varphi}_{*\,\alpha}(\mu) = \big(\mu(A_1), \ldots, \mu(A_{N(\alpha)})\big), \tag{D.3}$$

which form a coherent system. We are now in a position to formulate the following theorem (see [49], Ch. IX, § 4, No. 2).

Theorem D.1.4. *(Bourbaki-Prokhorov) Let $(\mathscr{X}_\alpha, \varphi_{\alpha\beta})$ be an inverse system of topological spaces, indexed by $\mathscr{A}$, T a topological space and $\hat{\varphi}_\alpha : T \to \mathscr{X}_\alpha$ a coherent and separating family of continuous mappings. Finally, let $(\mu_\alpha, \varphi_{*\alpha\beta})$ be an inverse system of positive measures on $(\mathscr{X}_\alpha, \varphi_{\alpha\beta})$. For there to exist a bounded Radon measure μ on T such that $\hat{\varphi}_{*\alpha}(\mu) = \mu_\alpha$ for all $\alpha \in \mathscr{A}$, it is necessary and sufficient that the following condition is satisfied:*

(P) for every $\varepsilon > 0$, there is a compact $K \subset T$ such that $\mu_\alpha(\mathscr{X}_\alpha \setminus \hat{\varphi}_\alpha(K)) \leq \varepsilon$ for all $\alpha \in \mathscr{A}$.

When this is so, the measure μ is uniquely determined and

$$\mu(L) = \inf\{\mu_\alpha(\hat{\varphi}_\alpha(L)) : \alpha \in \mathscr{A}\},$$

for every compact set L in T.

The most direct application of this theorem to the inverse systems at hand is through the measure-theoretic approach.

Example D.1.5. Consider the case where $T = M(\mathscr{X})$ with topology $\mathscr{T}_1$, that is, $\mu \to \nu$ if $\mu f \to \nu f$ for every bounded, measurable $f : \mathscr{X} \to \mathbb{R}$. Let $\mathscr{A}$ denote the directed set of all finite measurable partitions of $\mathscr{X}$. Then the maps $\hat{\varphi}_{*\alpha}$ that project $M(\mathscr{X})$ onto the $M(\mathscr{X}_\alpha)$ are *continuous* (and separating) and they form a coherent system. If we require specified marginals Π_α to form a coherent system of probability measures, we can say the following.

Proposition D.1.6. *There exists a bounded Radon measure Π on $M(\mathscr{X})$ with marginals Π_α for all α,* if and only if*,*

for every $\varepsilon > 0$, there is a $\mathscr{T}_1$-compact K in $M(\mathscr{X})$ such that $\mu_\alpha(\mathscr{X}_\alpha \setminus \hat{\varphi}_\alpha(K)) \leq \varepsilon$ for all $\alpha \in \mathscr{A}$.

But now the problem becomes apparent: this type of weak compactness is the domain of the Dunford-Pettis theorem, which is formulated in the context of *dominated* models only, *or* must be extended to the context of the *L*- and *M*-spaces that feature centrally in Le Cam's perspective [179]. The former option appears difficult to formulate from the point of view of an inverse system of priors, unless one considers reasonable the assumption that the measures ultimately described in the model are all dominated by a single probability measure Q and even then, that road leads to the type of tautology we have seen earlier: the requirement becomes hard to formulate but would look something like:

for every $\varepsilon, \delta > 0$, there is an $M > 0$ such that,

$$\Pi_\alpha\Big(\hat{\varphi}_\alpha(\mu \in M(\mathscr{X}) : \int_{\{d\mu/dQ > M\}} \frac{d\mu}{dQ}\, dQ < \delta\Big) \leq \varepsilon,$$

for all $\alpha \in \mathscr{A}$.

which is difficult to verify since in practical situations we have only an inverse system of marginal measures available. Somewhat more generally, we may assume that there exists a probability measure Q and bounded linear operators $T_\alpha : L^1(Q) \to \mathscr{P}$ with compact $K_\alpha \subset L^1(Q)$ for all $\alpha \in \mathscr{A}$ such that,

$$\Pi_\alpha(\hat{\varphi}_\alpha(M(\mathscr{X}) \setminus T_\alpha(K_\alpha))) \geq \delta)) \to 0,$$

but things are not improving, the above amounts basically to the same unverifiable condition from the point of view that only an inverse system of marginal measures is available. Not only do we see that the inclusion of a *large* $\mathscr{A}$ (like that of all finite, measurable partitions) implies that the compactness condition is to be verified for a *large* set of projections, compacta do not organise with the projections in a harmonious way, as expressed by the explicit appearance of the rather difficult map $\hat{\varphi}_\alpha$ above.

In what follows, we apply theorem 8.5.5 with Prokhorov's weak topology and a zero-dimensional version $\hat{\mathscr{X}}$ of the sample-space $T = M(\hat{\mathscr{X}})$, which leads to a more controllable condition.

D.2 Inverse limit priors

The goal, here, is to use theorem 8.5.5 again to prove the existence of inverse limit priors Π. Like before, the spaces $M(\mathscr{X}_\alpha)$ are all compact and we consider the spaces $C(M(\mathscr{X}_\alpha))$ of continuous functions $M(\mathscr{X}_\alpha) \to \mathbb{R}$ and the continuous maps $\varphi_{*\alpha\beta} : M(\mathscr{X}_\beta) \to M(\mathscr{X}_\alpha)$ of (8.19), which induce $\varphi^*_{*\alpha\beta} : C(M(\mathscr{X}_\alpha)) \to C(M(\mathscr{X}_\beta))$ through $\varphi^*_{*\alpha\beta}(f) = f \circ \varphi_{*\alpha\beta}$, with transpose $\varphi_{**\alpha\beta} : M(M(\mathscr{X}_\beta)) \to M(M(\mathscr{X}_\alpha))$, for all $\alpha, \beta \in \mathscr{A}$, $\alpha \leq \beta$. Like before, $(M(M(\mathscr{X}_\alpha)), \varphi_{**\alpha\beta})$ form an inverse system with inverse limit N', projections $\varphi_{**\alpha} : N' \to M(M(\mathscr{X}_\alpha))$ and injective embedding $\hat{M} \subset N'$ of $M(M(\hat{\mathscr{X}}))$ (with restrictions $\hat{\varphi}_{**\alpha} : \hat{M} \to M(M(\mathscr{X}_\alpha))$).

If we assume continuity of the projections, $M(\mathscr{X})$ (identified with $\hat{M}$) may play the role of T in theorem 8.5.5: take an inverse system of measure $(\Pi_\alpha, \varphi_{**\alpha\beta})$ on the inverse system $(M(\mathscr{X}_\alpha), \varphi_{*\alpha\beta})$ and $T = M(\hat{\mathscr{X}})$, for a space $\mathscr{X}$ that is completely regular. The bounded Radon measure Π for which existence is proved, lies in $M(M(\hat{\mathscr{X}}))$, where $\hat{\mathscr{X}}$ is the space obtained in proposition 8.6.6.

Theorem D.2.1. *Let $\mathscr{X}$ be Hausdorff completely regular with basis $\mathscr{U}$; let $\mathscr{A}$ consist of partitions generated by $\mathscr{U}$ and resolve $\mathscr{X}$. Denote the corresponding inverse system by $(\mathscr{X}_\alpha, \varphi_{\alpha\beta})$ and by $(\Pi_\alpha, \varphi_{*\alpha\beta})$ an inverse system of positive measures on $(M(\mathscr{X}_\alpha), \varphi_{\alpha\beta})$. Endow $M(\hat{\mathscr{X}})$ with Prokhorov's weak topology. For there to exist a bounded Radon measure Π on $M(\hat{\mathscr{X}})$ such that $\hat{\varphi}_{**\alpha}(\Pi) = \Pi_\alpha$ for all $\alpha \in \mathscr{A}$, it is necessary and sufficient that the following condition be satisfied:*

(P') for every $\varepsilon > 0$, there is a compact $H \subset M(\hat{\mathscr{X}})$ such that,

$$\Pi_\alpha\big(M(\mathscr{X}_\alpha) \setminus \hat{\varphi}_{*\alpha}(H)\big) < \varepsilon,$$

for all $\alpha \in \mathscr{A}$.

When this is so, the measure Π *is uniquely determined and*

$$\Pi(L) = \inf\{\Pi_\alpha(\hat{\varphi}_{*\alpha}(L)) : \alpha \in \mathscr{A}\},$$

for every compact set L *in* $M(\hat{\mathscr{X}})$.

Proof. For every $A \in \alpha$ in every $\alpha \in \mathscr{A}$, $i^{-1}(A) \subset \hat{\mathscr{X}}$ is a clopen element of the subbasis that defines the topology for $\hat{\mathscr{X}}$. So for all $\alpha \in \mathscr{A}$, $\hat{\varphi}_{*\alpha} : M(\hat{\mathscr{X}}) \to M(\mathscr{X}_\alpha)$ is continuous with respect to Prokhorov's weak topology, by lemma 8.6.11. Moreover, $(\hat{\varphi}_{*\alpha}, \varphi_{\alpha\beta})$ is a coherent and separating family. The assertion now follows directly from theorem 8.5.5.

This leads to the following double version of Prokhorov's condition for the most common types of sample spaces.

Corollary D.2.2. *Let* $\mathscr{X}$ *be Polish in theorem 8.6.12. For there to exist a bounded measure* Π *on* $M(\mathscr{X})$ *such that* $\hat{\varphi}_{**\alpha}(\Pi) = \Pi_\alpha$ *for all* $\alpha \in \mathscr{A}$*, it is necessary and sufficient that the following condition be satisfied:*

(P") *for every* $\varepsilon > 0$*, there exist* $\beta \in \mathscr{A}$*,* $R > 0$ *such that* $\Pi_\beta(\mu_\beta(\mathscr{X}_\beta) \le R) > 1 - \varepsilon$ *and, for every* $\delta > 0$*, there is a compact* $K \subset \mathscr{X}$ *such that,*

$$\Pi_\alpha\big(\{\mu_\alpha \in M(\mathscr{X}_\alpha) : \mu_\alpha(\mathscr{X}_\alpha \setminus \hat{\varphi}_\alpha(K)) > \delta\}\big) < \varepsilon,$$

for all $\alpha \in \mathscr{A}$.

When this is so, the measure Π *is uniquely determined and*

$$\Pi(L) = \inf\{\Pi_\alpha(\hat{\varphi}_{*\alpha}(L)) : \alpha \in \mathscr{A}\},$$

for every compact set L *in* $M(\mathscr{X})$.

Proof. Since $\mathscr{X}$ is Polish, $M(\mathscr{X})$ is identified with $M(\hat{\mathscr{X}})$ according to proposition **??**, and *c.f.* theorems C.7.15 and C.7.16, $H \subset M(\mathscr{X})$ is compact iff H is closed,

$$\sup\{\|\mu\| : \mu \in H\} < \infty,$$

and for every $\delta > 0$, there exists a compact K in $\mathscr{X}$ such that for all $\mu \in H$,

$$|\mu(\mathscr{X} \setminus K)| \le \delta.$$

The set H_1,

$$H_1 = \{\mu \in M(\mathscr{X}) : \hat{\varphi}_{*\alpha}(\mu)(\mathscr{X}_\alpha \setminus \hat{\varphi}_\alpha(K)) \le \delta\},$$

is closed in $M(\mathscr{X})$, because of proposition **??** and the fact that $\mathscr{X} \setminus \hat{\varphi}_\alpha(K)$ is clopen, *c.f.* lemma 8.6.11. Because $\mathscr{X}_\beta$ is clopen, the set $H_2 = \{\mu \in M(\mathscr{X}) : \mu_\beta(\mathscr{X}_\beta) \le L\}$ is closed. Since $\mu_\alpha(\mathscr{X}_\alpha) = \mu_\beta(\mathscr{X}_\beta)$ for all α, *c.f.* (D.2), the intersection $H = H_1 \cap H_2$ is compact in $M(\mathscr{X})$ and satisfies property *(P')*.

It is noted that property *(P")* is also *sufficient* in the situation where $\mathscr{X}$ is completely regular and we require that,

(P"') for every $\varepsilon, \delta > 0$, there is a compact $K \subset \hat{\mathscr{X}}$ such that,

$$\Pi_\alpha\big(\{\mu_\alpha \in M(\hat{\mathscr{X}}_\alpha) : \mu_\alpha(\mathscr{X}_\alpha \setminus \hat{\varphi}_\alpha(K)) > \delta\}\big) < \varepsilon,$$

for all $\alpha \in \mathscr{A}$.

Note also that corollary D.2.2 can be read with the following consequence.

Corollary D.2.3. *Let $\mathscr{X}$ be Polish in theorem 8.6.12 with $\mathscr{A}$ the partitions generated by a countable basis $\mathscr{U}$. For any bounded measure Π on $M(\mathscr{X})$ and any $\varepsilon, \delta > 0$, there is a compact $K \subset \hat{\mathscr{X}}$ such that,*

$$\Pi_\alpha\big(\{\mu_\alpha \in M(\mathscr{X}_\alpha) : \mu_\alpha(\mathscr{X}_\alpha \setminus \hat{\varphi}_\alpha(K)) > \delta\}\big) < \varepsilon,$$

for all $\alpha \in \mathscr{A}$.

With reference to Brouwer's theorem C.4.12, the point of this complication is that *any random probability measure on a Polish space places arbitrarily mass arbitrarily close to one on fixed zero-dimensional compact sets*. The suggestion is that there exists a relation with the *discreteness* of Π-almost-all realizations of the Dirichlet random probability measure *c.f.* lemma **??**, which is perhaps somewhat surprising without the zero-dimensional perspective on existence of inverse limit measures.

References

References

1. E. ABBE, *Community Detection and Stochastic Block Models: Recent Developments,* Journal of Machine Learning Research **18.177** (2018), 1–86.
2. E. ABBE, A. BANDEIRA, and G. HALL, *Exact Recovery in the Stochastic Block Model,* IEEE Transactions on Information Theory **62.1** (2016).
3. M. ALPERT and H. RAIFFA, *A progress report on the training of probability assessors,* In *Judgement under uncertainty: heuristics and biases*, eds. D. Kahneman, P. Slovic and A. Tversky, Cambridge University Press, Cambridge (1982).
4. S. AMARI, *Differential-geometrical methods in statistics,* Lecture Notes in Statistics No. 28, Springer Verlag, Berlin (1990).
5. A. AMINI, et al. *Pseudo-likelihood methods for community detection in large sparse networks,* Ann. Statist. **41.4** (2013), 2097-–2122.
6. T. ANDERSON, *Estimating linear statistical relationships,* Ann. Statist. **12** (1984), 1–45.
7. D. BANERJEE, *Contiguity and non-reconstruction results for planted partition models: the dense case,* Electron. J. Probab. 23 (2018).
8. A. BARRON, Discussion on *Diaconis and Freedman: the consistency of Bayes estimates,* Ann. Statist. **14** (1986), 26–30.
9. A. BARRON, *The exponential convergence of posterior probabilities with implications for Bayes estimators of density functions,* Technical Report **7** (1988), Dept. Statistics, Univ. Illinois.
10. A. BARRON, *Uniformly powerful goodness-of-fit tests,* Ann. Statist. **17** (1989), 107—124.
11. A. BARRON, M. SCHERVISH and L. WASSERMAN, *The consistency of distributions in nonparametric problems,* Ann. Statist. **27** (1999), 536—561.
12. M. BAYARRI and J. BERGER, *The interplay of Bayesian and frequentist analysis,* Preprint (2004).
13. T. BAYES, *An essay towards solving a problem in the doctrine of chances,* Phil. Trans. Roy. Soc. **53** (1763), 370–418.
14. A. Barron, L. Birgé and P. Massart, *Risk bounds for model selection via penalization*, Probability Theory and Related Fields **113** (1999), pp. 301–413.
15. S. BERNSTEIN, *Theory of probability,* (in Russian), Moskow (1917).
16. A. BARRON, M. SCHERVISH and L. WASSERMAN, *The consistency of posterior distributions in nonparametric problems,* Ann. Statist. **27** (1999), 536–561.
17. A. BERGER, *On uniformly consistent tests,* Ann. Math. Statist. **18** (1951), 289–293.
18. A. BERGER, *On orthogonal probability measures,* Proc. Amer. Math. Soc. **4** (1953), 800–806.
19. J. BERGER, *Statistical decision theory and Bayesian analysis,* Springer, New York (1985).

20. J. BERGER and J. BERNARDO, *On the development of reference priors,* Bayesian Statistics **4** (1992), 35–60.
21. R. BERK, *Limiting behaviour of posterior distributions when the model is incorrect,* Ann. Math. Statist. **37** (1966), 51–58.
22. R. BERK, *Consistency of a posteriori,* Ann. Math. Statist. **41** (1970), 894–906.
23. R. BERK and I. SAVAGE, *Dirichlet processes produce discrete measures: an elementary proof,* Contributions to statistics, Reidel, Dordrecht (1979), 25–31.
24. J. BERNARDO, *Reference posterior distributions for Bayesian inference,* J. Roy. Statist. Soc. **B41** (1979), 113–147.
25. J. BERNARDO and A. Smith, *Bayesian theory,* John Wiley & Sons, Chichester (1993).
26. L. BEZNEA and I. CIMPEAN, *On Bochner-Kolmogorov Theorem,* in Donati-Martin, *et al.* (eds.), *Séminaire de Probabilités XLVI,* Lecture Notes in Mathematics **2123**, Springer (2014).
27. P. BICKEL and J. YAHAV, *Some contributions to the asymptotic theory of Bayes solutions,* Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete **11** (1969), 257–276.
28. P. BICKEL and Y. Ritov, *Efficient estimation in the errors in variables model.* Ann. Statist. **15** (1987), 513–540.
29. P. BICKEL, Y. RITOV, C. KLAASSEN and J. WELLNER, *Efficient and adaptive estimation for semiparametric models (2nd edition),* Springer, New York (1998).
30. P. BICKEL, and A. CHEN, *A nonparametric view of network models and Newman-Girvan and other modularities,* Proceedings of the National Academy of Sciences **106.50** (2009), 21068–21073.
31. P. BILLINGSLEY, *Probability and Measure, 2nd edition,* John Wiley & Sons, Chichester (1986).
32. L. BIRGÉ, *Approximation dans les espaces métriques et théorie de l'estimation,* Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete **65** (1983), 181–238.
33. L. BIRGÉ, *Sur un théorème de minimax et son application aux tests,* Probability and Mathematical Statistics **3** (1984), 259–282.
34. L. BIRGÉ and P. MASSART, *From model selection to adaptive estimation,* Festschrift for Lucien Le Cam, Springer, New York (1997), 55–87.
35. L. BIRGÉ and P. MASSART, *Gaussian model selection*, J. Eur. Math. Soc. **3** (2001), 203–268.
36. C. BISHOP, *Pattern Recognition and Machine Learning,* Springer, New York (2006).
37. D. BLACKWELL and L. DUBINS, *Merging of opinions with increasing information,* Ann. Math. Statist. **33** (1962), 882–886.
38. D. BLACKWELL, *Discreteness of Ferguson Selections,* Ann. Statist. **1.2** (1973), 356–358.
39. D. BLACKWELL and J. MACQUEEN, *Ferguson Distributions Via Polya Urn Schemes,* Ann. Statist. **1.2** (1973), 353–355.
40. D. BLEI, A. NG and M. Jordan, Latenet Dirichlet Allocation, J. Mach. Learn. Res. **3** (2003), 993—1022.
41. S. BOCHNER, *Harmonic Analysis and the Theory of Probability,* University of California Press, Los Angeles (1955).
42. BOGACHEV, *Measure theory (vol. I and II),* Springer, New York (2007).
43. B. BOLLOBÁS, S. JANSON, and O. RIORDAN, *The phase transition in inhomogeneous random graphs,* Random Structures and Algorithms **31.1** (2007), 3–122.
44. L. BOLTZMANN, *Vorlesungen über Gastheorie,* (2 Volumes), Leipzig (1895, 1898).
45. N. BOURBAKI, *Theory of Sets,* Springer, New York (2004).
46. N. BOURBAKI, *General Topology: Chapters 1-4,* Springer, New York (1998).
47. N. BOURBAKI, *General Topology: Chapters 5-10,* Springer, New York (1989).
48. N. BOURBAKI, *Integration I: Chapters 1-6,* Springer, New York (2004).
49. N. BOURBAKI, *Integration II: Chapters 7-9,* Springer, New York (2010).
50. N. BOURBAKI, *Topological vector spaces: Chapters 1-5,* Springer, New York (1987).
51. L. BREIMAN, L. LE CAM and L. SCHWARTZ, *Consistent Estimates and Zero-One Sets,* Ann. Math. Statist. **35** (1964), 157–161.
52. P. BÜHLMAN and S. VAN DE GEER, *Statistics for High-Dimensional Data,* Springer Verlag, New York (2011).

53. O. BUNKE and X. MILHAUD, *Asymptotic behavior of Bayes estimates under possibly incorrect models,* Ann. Statist. **26** (1998), 617–644.
54. T. CAI, M. LOW and Y. XIA, *Adaptive confidence interval for regression functions under shape constraints,* Ann. Statist. **41** (2013), 722–750.
55. C. CARVALHO, N. POLSON, and J. SCOTT, *The horseshoe estimator for sparse signals,* Biometrika **97** (2010), 465–480.
56. I. CASTILLO and A. VAN DER VAART, *Needles and straw in a haystack: posterior concentration for possibly sparse sequences,* Ann. Statist. **40** (2012), 2069–2101.
57. D. CHOI, P. WOLFE, and E. AIROLDI, *Stochastic blockmodels with a growing number of classes,* Biometrika **99.2** (2012), 273–284.
58. J. CHOKSI, *Inverse limits of measure spaces,* Proceedings of the London Mathematical Society **8.3** (1958), 321–342.
59. L. COMMINGES, and A. DALALYAN, *Minimax testing of a composite null hypothesis defined via a quadratic functional in the model of regression.* Electr. J. Statist. **7** (2013), 146–190.
60. T. COVER, *On determining the irrationality of the mean of a random variable,* Ann. Statist. **1** (1973), 862–871.
61. D. COX, *An analysis of Bayesian inference for non-parametric regression,* Ann. Statist. **21** (1993), 903–924.
62. H. CRAMÉR, *Mathematical methods of statistics,* Princeton University Press, Princeton (1946).
63. A. DAWID, *On the limiting normality of posterior distribution,* Proc. Canad. Phil. Soc. **B67** (1970), 625—633.
64. P. DE BLASI, A. LIJOI, and I. PRÜNSTER, *An asymptotic analysis of a class of discrete nonparametric priors,* Statist. Sinica **23** (2013), 1299–1322.
65. P. DE BLASI, S. FAVARO, A. LIJOI, R.H. MENA, I. PRUENSTER and M. RUGGIERO, *Are Gibbs-type priors the most natural generalization of the Dirichlet process?* IEEE Transactions Pattern Analysis and Machine Intelligence **37.2** (2015), 212–229.
66. A. DECELLE, et al. *Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications,* Phys. Rev. **E84.6** (2011), p. 066106.
67. A. DECELLE, et al. *Inference and Phase Transitions in the Detection of Modules in Sparse Networks,* Phys. Rev. Lett. **107.6**, p. 065701.
68. A. DEMBO and O. ZEITOUNI, *Large Deviations Techniques and Applications,* Jones and Bartlett, Boston (1993).
69. A. DEMBO and Y. PERES, *A topological criterion for hypothesis testing,* Ann. Statist. **22** (1994), 106–117.
70. L. DEVROYE and G. LUGOSI, *Almost sure classification of densities,* J. Nonparametric Statist. **14** (2002), 675-–698.
71. P. DIACONIS and D. FREEDMAN, *On the consistency of Bayes estimates,* Ann. Statist. **14** (1986), 1–26.
72. P. DIACONIS and D. FREEDMAN, *On inconsistent Bayes estimates of location,* Ann. Statist. **14** (1986), 68–87.
73. P. DIACONIS and D. FREEDMAN, *Nonparametric Binary Regression: A Bayesian Approach,* Ann. Statist. **21** (1993), 2108–2137.
74. P. DIACONIS and D. FREEDMAN, *Consistency of Bayes estimates for nonparametric regression: Normal theory,* Bernoulli, **4** (1998), 411-–444.
75. J. DIESTEL, *A survey or results related to the Dunford–Pettis property,* Contemp. Math. **2**, Amer. Math. Soc. (1980) 15—60.
76. J. DIESTEL, *Uniform integrability: an introduction,* Lectures presented at the *School of Measure theory and Real Analysis*, Grado, Italy (1991).
77. D. DONOHO, *One-sided inference about functionals of a density,* Ann. Statist. **16** (1988), 1390—1420.
78. D. DONOHO, I. JOHNSTONE, J. HOCH and A. STERN, *Maximum entropy and the nearly black object,* J. Roy. Statist. Soc. **B54** (1992), 41–81.
79. D. DONOHO, and I. JOHNSTONE, *Minimax risk over ℓ_p-balls for ℓ_q-error,* Probab. Theory Related Fields, **99** (1994), 277-303.

80. J. DOOB, *Applications of the theory of martingales,* Le calcul des Probabilités et ses Applications, Colloques Internationales du CNRS, Paris (1948), 22–28.
81. R. DUDLEY, *Real analysis and probability,* Wadsworth & Brooks-Cole, Belmont (1989).
82. N. DUNFORD and J. SCHWARTZ, *Linear Operators, Part 1: General Theory (Vol. 1),* John Wiley and Sons, Hoboken, NJ (1988).
83. A. DVORETZKY, J. KIEFER, and J. WOLFOWITZ, *Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator,* Ann. Math. Statist. **27** (1956), 642–669.
84. M. DYER, and **A. Frieze,** *The solution of some random NP-hard problems in polynomial expected time,* Journal of Algorithms **10.4** (1989), 451–489.
85. P. EICHELSBACHER and A. GANESH, *Bayesian inference for Markov chains,* J. Appl. Probab. **39** (2002), 91–99.
86. B. EFRON and C. MORRIS, *Stein's estimation rule and its competitors - an empirical Bayes approach,* Journal of the American Statistical Association **68** (1973), 117–130.
87. B. EFRON, *Defining curvature on a statistical model,* Ann. Statist. **3** (1975), 1189–1242.
88. B. EFRON and R. Tibshirani, An introduction to the Bootstrap, Chapman and Hall, London (1993).
89. M. ESCOBAR and M. WEST, *Bayesian density estimation and inference with mixtures,* Journal of the American Statistical Association **90** (1995), 577–588.
90. P. ERDÖS, A. RÉNYI, *On Random Graphs I,* PUBLICATIONES MATHEMATICAE **6** (1959).
91. M. ERMAKOV, *On consistent hypothesis testing,* J. MATH. SC. **225** (2017), 751-–769. *(Translated from Zapiski Nauchnykh Seminarov POMI* **442** *(2015), 48-–74.)*
92. J. FABIUS, *Asymptotic behavior of Bayes' estimates,* ANN. MATH. STATIST. **35** (1964), 846–856.
93. J. FAN AND Y. TRUONG, *Nonparametric regression with errors in variables,* ANN. STATIST. **21** (1993), 1900–1925.
94. T. FERGUSON, *A Bayesian analysis of some non-parametric problems,* ANN. STATIST. **1** (1973), 209–230.
95. T. FERGUSON, *Prior distribution on the spaces of probability measures,* ANN. STATIST. **2** (1974), 615–629.
96. S. FORTUNATO, *Community detection in graphs,* PHYSICS REPORTS **486.3** (2010), 75–174.
97. D. FREEDMAN, *On the asymptotic behavior of Bayes estimates in the discrete case I,* ANN. MATH. STATIST. **34** (1963), 1386–1403.
98. D. FREEDMAN, *On the asymptotic behavior of Bayes estimates in the discrete case II,* ANN. MATH. STATIST. **36** (1965), 454–456.
99. D. FREEDMAN AND P. DIACONIS, *On inconsistent M-estimators,* ANN. STATIST. **10.2** (1982), 454–461.
100. D. FREEDMAN, AND P. DIACONIS, *On Inconsistent Bayes Estimates in the Discrete Case,* ANN. STATIST. **11** (1983), 1109–1118.
101. D. FREEDMAN, *On the Bernstein-von Mises theorem with infinite dimensional parameters,* ANN. STATIST. **27.4** (1999), 1119–1140.
102. C. GAO, ET AL. *Achieving Optimal Misclassification Proportion in Stochastic Block Models,* JOURNAL OF MACHINE LEARNING RESEARCH **18.60** (2017), 1–45.
103. S. VAN DE GEER, *Empirical Processes in M-Estimation,* CAMBRIDGE UNIVERSITY PRESS, CAMBRIDGE (2000).
104. S. GHOSAL, J. GHOSH AND R. RAMAMOORTHI, *Non-informative priors via sieves and packing numbers,* ADVANCES IN STATISTICAL DECISION THEORY AND APPLICATIONS (EDS. S. PANCHAPAKESHAN, N. BALAKRISHNAN), BIRKHÄUSER, BOSTON (1997).
105. S. GHOSAL, J. GHOSH, AND R. RAMAMOORTHI, *Consistency issues in Bayesian nonparametrics,* IN *Asymptotics, Nonparametrics and Time Series: A Tribute to Madan Lal Puri (Subir Ghosh, ed.)* DEKKER, NEW YORK (1999), 639—667.
106. S. GHOSAL, J. GHOSH AND A. VAN DER VAART, *Convergence rates of posterior distributions,* ANN. STATIST. **28** (2000), 500–531.

107. S. GHOSAL AND A. VAN DER VAART, *Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities,* ANN. STATIST. **29** (2001), 1233 – 1263.
108. S. GHOSAL AND Y.-Q. TANG *Bayesian Consistency for Markov Processes,* SANKHYA **68** (2006), 227–239.
109. S. GHOSAL, *Dirichlet process, related priors and posterior asymptotics,* IN N. L. HJORT ET AL. (EDS.), *Bayesian Nonparametrics*, CAMBRIDGE UNIVERSITY PRESS, CAMBRIDGE (2010).
110. S. GHOSAL AND A. VAN DER VAART, *Fundamentals of nonparametric Bayesian statistics,* CAMBRIDGE UNIVERSITY PRESS, CAMBRIDGE (2017).
111. J. GHOSH AND R. RAMAMOORTHI, *Bayesian Nonparametrics,* SPRINGER VERLAG, BERLIN (2003).
112. M. GIRVAN, AND *M. Newman, Community structure in social and biological networks,* PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES **99.12** (2002), 7821–7826.
113. P. GLYNN and D. ORMONEIT, *Hoeffding's inequality for uniformly ergodic Markov chains,* Statist. Probab. Lett. **56** (2002), 143–146.
114. E. GOBET, *LAN property for ergodic diffusions with discrete observations,* Ann. I. H. Poincaré PR **38** (200), 711–737.
115. P. GREEN, *Reversible jump Markov chain Monte Carlo computation and Bayesian model determination*, Biometrika **82** (1995), 711–732.
116. P. GREENWOOD and A. SHIRYAEV, *Contiguity and the statistical invariance principle,* Gordon and Breach, New York (1985).
117. A. GROTHENDIECK, *Sur les applications lineaires faiblement compactes d'espaces du type C(K),* Canadian Journal of Mathematics **5** (1953), 129—173.
118. O. GUÉDON, and R. VERSHYNIN, *Community detection in sparse networks via Grothendieck's inequality,* Probability Theory and Related Fields **165.3** (2016), 1025–1049.
119. B. HAJEK, Y. WU, and J. XU, *Achieving Exact Cluster Recovery Threshold via Semidefinite Programming,* IEEE Trans. Inf. Theor. **62.5** (2016), 2788–2797.
120. J. HAJÉK and Z. Šidák, *Theory of rank tests,* Academic Press, New York (1967).
121. J. HÁJEK, *A characterization of limiting distributions of regular estimates*, Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete **14** (1970), 323–330.
122. J. HÁJEK, *Local asymptotic minimax and admissibility in estimation*, Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability **1**, 175–194. University of California Press, Berkeley (1972).
123. T. HARRIS, *Counting measures, monotone random set functions,* Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete **10.2** (1968), 102–119.
124. K. HAYASHI, T. KONISHI and T. KAWAMOTO, *A tractable fully Bayesian method for the stochastic block model,* `arxiv:1602.02256 [cs.LG]`
125. N. HENGARTNER, P. Stark, *Finite-sample confidence envelopes for shape-restricted densities,* Ann. Statist. **23** (1995), 525–550.
126. W. HOEFFDING, and J. WOLFOWITZ, *Distinguishability of sets of distributions,* Ann. Math. Statist. **29** (1958), 700–718.
127. P. HOLLAND, K. LASKEY, and S. LEINHARDT, *Stochastic block models: First steps,* Social Networks **5.2** (1983), 109–137.
128. R. HÖPFNER, J. JACOD and L. LADELLI, *Local asymptotic normality and mixed normality for Markov statistical models,* Probab. Th. Rel. Fields (1990) 86: 105.
129. T.-M. HUANG, *Convergence rates for posterior distributions and adaptive estimation,* Carnegie Mellon University, preprint (accepted for publication in Ann. Statist.).
130. P. HUBER, *The behavior of maximum likelihood estimates under nonstandard conditions,* Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability **1**, 221–233. University of California Press, Berkeley (1967).
131. I. IBRAGIMOV and R. HAS'MINSKII, *Statistical estimation: asymptotic theory,* Springer, New York (1981).
132. INGSTER, I. SUSLINA, *Nonparametric Goodness-of-fit Testing under Gaussian Models,* Lecture Notes in Statistics **169** (2002), Springer N.Y.

133. W. JAMES and C. STEIN, *Estimation with quadratic loss,* Proc. Fourth Berkeley Symp. Math. Statist. Prob. **1** (1961), 361—379.
134. L. JAMES, *A simple proof of the almost sure discreteness of a class of random measures,* Statistics and Probability Letters **65.4** (2003), 363–368.
135. L. JAMES, A. LIJOI and I. PRÜNSTER, *Posterior analysis for normalized random measures with independent increments,* Scandinavian journal of statistics **36.1** (2009), 76–97.
136. H. JEFFREYS, *An invariant form for the prior probability in estimation problems,* Proc. Roy. Soc. London **A186** (1946), 453–461.
137. H. JEFFREYS, *Theory of probability (3rd edition),* Oxford University Press, Oxford (1961).
138. I. JOHNSTONE and B. SILVERMAN, *Needles and straw in haystacks: empirical Bayes setimates of possibly sparse sequences,* Ann. Statist. **32** (2004), 1594–1649.
139. S. KAKUTANI, *Concrete representation of abstract (L)-spaces and the mean ergodic theorem,* Annals of Mathematics **42** (1941), 523–537.
140. R. KASS and A. RAFTERY, *Bayes factors,* Journal of the American Statistical Association **90** (1995), 773–795.
141. R. KASS and L. WASSERMAN, *A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion,* Journal of the American Statistical Association **90** (1995), 928–934.
142. A. KECHRIS, A. LOUVEAU and W. WOODIN, *The structure of σ-ideals of compact spaces,* Trans. Amer. Math. Soc. **301** (1987), 747–758.
143. A. KECHRIS, *Classical descriptive set theory,* Springer, New York (1994).
144. M. KENDALL and A. STUART, *The advanced theory of statistics, Vol. 2, (4th edition),* Griffin, London (1979).
145. YONGDAI KIM and JAEYONG LEE, *The Bernstein-von Mises theorem of survival models,* (accepted for publication in Ann. Statist.)
146. YONGDAI KIM and JAEYONG LEE, *The Bernstein-von Mises theorem of semiparametric Bayesian models for survival data,* (accepted for publication in Ann. Statist.)
147. J. KINGMAN and S. TAYLOR, *Introduction to measure and probability,* Cambridge University Press, Cambridge (1966).
148. J. KINGMAN, *Completely random measures,* Pacific J. Math. **21.1** (1967).
149. J. KINGMAN, *Random discrete distributions (with discussion),* J. Roy. Statist. Soc. **B37** (1975), 1–22.
150. B. KLEIJN, *Bayesian asymptotics under misspecification.* PhD. Thesis, Free University Amsterdam (2004).
151. B. KLEIJN and A. VAN DER VAART, *Misspecification in Infinite-Dimensional Bayesian Statistics,* Ann. Statist. **34** (2006), 837–877.
152. B. KLEIJN and A. VAN DER VAART, *The Bernstein-Von-Mises theorem under misspecification,* Electron. J. Statist. **6** (2012), 354–381.
153. B. KLEIJN, *A Bayesian analysis of errors-in-variables regression,* (2004, unpublished).
154. B. KLEIJN, *Criteria for posterior consistency,* (2013) arxiv:1308.1263 [MATH.ST].
155. B. KLEIJN and J. VAN WAAIJ, *Recovery, detection and confidence sets of communities in a sparse stochastic block model,* (2018) `arxiv:1810.09533 [math.ST]`
156. B. KLEIJN and Y. Y. ZHAO, *Criteria for posterior consistency and convergence at a rate,* Electron. J. Statist. (13.2) (2019), 4709–4742.
157. B. KLEIJN, *Frequentist validity of Bayesian limits,* Ann. Statist. **49.1** (2021), 182–202.
158. B. KLEIJN and J. VAN WAAIJ, *Confidence sets in a sparse stochastic block model with two communities of unknown sizes,* (2021) `arxiv2108.07078 [math.ST]`
159. A. KOLMOGOROV and V. TIKHOMIROV, *Epsilon-entropy and epsilon-capacity of sets in function spaces,* American Mathematical Society Translations (series 2), **17** (1961), 277–364.
160. A. KOLMOGOROV, *Grundbegriffe der Wahrscheinlichkeitsrechnung,* Springer Verlag, Berlin (1933).
161. C. KRAFT, *Some conditions for consistency and uniform consistency of statistical procedures,* Univ. Californ. Publ. Stat. **2** (1955), 125–142.
162. C. KRAFT, *A class of distribution function processes which have derivatives,* Journal of Applied Probability **2** (1964), 385–388.

163. F. KRZAKALA, et al. *Spectral redemption in clustering sparse networks,* Proceedings of the National Academy of Sciences **110.52** (2013), 20935–20940.
164. S. KULKARNI and O. ZEITOUNI, *A general classification rule for probability measures,* Ann. Statist. **23** (1995), 1393-–1407.
165. P. LAPLACE, *Mémoire sur la probabilité des causes par les évenements,* Mem. Acad. R. Sci. Presentés par Divers Savans **6** (1774), 621–656. (Translated in Statist. Sci. **1**, 359–378.)
166. P. LAPLACE, *Théorie Analytique des Probabilités (3rd edition),* Courcier, Paris (1820).
167. M. LAVINE, *Some aspects of Pólya tree distributions for statistical modelling,* Ann. Statist. **20.3** (1992), 1222–1235.
168. M. LAVINE, *More aspects of Pólya tree distributions for statistical modelling,* Ann. Statist. **22.3** (1994), 1161–1176.
169. E. LEHMANN and G. CASELLA, *Theory of point-estimation, (2nd ed.)* Springer, New York (1998).
170. E. LEHMANN and J. ROMANO, *Testing statistical hypothesis,* Springer, New York (2005).
171. L. LE CAM, *On some asymptotic properties of maximum-likelihood estimates and related Bayes estimates,* University of California Publications in Statistics, **1** (1953), 277–330.
172. L. LE CAM, *An extension of Wald's theory of statistical decision functions,* Ann. Math. Statist. **26.1** (1955), 69–81.
173. L. LE CAM and L. SCHWARTZ, *A necessary and sufficient condition for the existence of consistent estimates,* Ann. Math. Statist. **31** (1960), 140–150.
174. L. LE CAM, *Locally asymptotically normal families of distributions,* University of California Publications in Statistics, **3** (1953), 37–98.
175. L. LE CAM, *Limits of experiments,* Proc. Sixth Berkeley Symp. on Math. Statist. and Prob. **1** (1972), 245–261.
176. L. LE CAM, *On the assumptions used to prove asymptotic normality of maximum likelihood estimators,* Ann. Math. Statist. **41** (1970), 802–828.
177. L. LE CAM, *Convergence of estimates under dimensionality restrictions,* Ann. Statist. **1.1** (1973), 38–53.
178. L. LE CAM, *An inequality concerning Bayes estimates,* University of California, Berkeley (197X), unpublished.
179. L. LE CAM, *Asymptotic methods in statistical decision theory,* Springer, New York (1986).
180. L. LE CAM, Comment on *Consistency of Bayes estimates,* by D. A. Freedman and P. Diaconis, Ann. Statist. **14** (1986), 59–60.
181. L. LE CAM and G. YANG, *On the preservation of local asymptotic normality under information loss,* Ann. Statist. **16** (1988), 483–520.
182. L. LE CAM, *Maximum likelihood; an introduction,* International Statistical Review **58** (1990), 153–171.
183. L. LE CAM and G. YANG, *Asymptotics in Statistics: some basic concepts,* Springer Verlag, New York (1990).
184. A. LIJOI, I. PRÜNSTER and S. WALKER, *Extending Doob's consistency theorem to nonparametric densities,* Bernoulli **10** (2004), 651–663.
185. D. LINDLEY, *A measure of the information provided by an experiment,* Ann. Math. Statist. **27** (1956), 986–1005.
186. D. LINDLEY and A. SMITH, *Bayes estimates for the linear model,* J. Roy. Statist. Soc. **B43** (1972), 1–41.
187. M. LOW, *On nonparametric confidence intervals,* Ann. Statist. **25** (1997), 2547–2554.
188. W. LUXEMBURG, and A. ZAANEN, *Riesz spaces, volume I*, North-Holland, Amsterdam (1971).
189. D. MALLORY and M. SION, *Limits of inverse systems of measures,* Ann. Inst. Fourier **21** (1971), 25–57.
190. L. MASSOULIÉ, *Community detection thresholds and the weak Ramanujan property,* Proceedings of STOC 2014: 46th Annual Symposium on the Theory of Computing, New York (2014), 1–10.
191. R. MAULDIN, W. SUDDERTH AND S. WILLIAMS, *Pólya trees and random distributions,* Ann. Statist. **20.3** (1992), 1203–1221.

192. R. MEGGINSON, *An introduction to Banach Space Theory,* Springer, New York (1998).
193. S. MEYN and R. TWEEDIE, *Markov Chains and Stochastic Stability,* Cambridge University Press, New York (2009).
194. M. METIVIER, *Limites projectives de measures, martingales, applications,* Annali di Matematica, **63** (1963), 225–352.
195. R. VON MISES, *Wahrscheinlichkeitsrechnung,* Springer Verlag, Berlin (1931).
196. T. MITCHELL, and J. BEAUCHAMP, *Bayesian Variable Selection in Linear Regression,* J. Amer. Statist. Assoc. **83** (1988), 1023–1032.
197. E. MOSSEL, J. NEEMAN, and A. SLY, *Reconstruction and estimation in the planted partition model,* Probability Theory and Related Fields **162.3** (2015), 431–461.
198. E. MOSSEL, J. NEEMAN, and A. SLY, *Belief propagation, robust reconstruction and optimal recovery of block models,* Ann. Appl. Probab. **26.4** (2016), 2211–2256.
199. E. MOSSEL, J. NEEMAN, and A. SLY, *Consistency thresholds for the planted bisection model,* Electron. J. Probab. **21** (2016).
200. J. MUNKRES, *Topology (2nd edition),* Prentice Hall, Upper Saddle River (2000).
201. S. MURPHY and A. VAN DER VAART, *On Profile Likelihood,* Journal of the American Statistical Association, **95.450** (2000), 449–465.
202. R. NEAL, *Probabilistic Inference Using Markov Chain Monte Carlo Methods,* Technical Report CRG-TR-93-1, Dept. of Computer Science, University of Toronto, Toronto, (1993).
203. A. NOBEL, *Hypothesis testing for families of ergodic processes,* Bernoulli **12** (2006), 251—269.
204. K. NOWICKI, and T. SNIJDERS, *Estimation and Prediction for Stochastic Blockstructures,* Journal of the American Statistical Association **96.455** (2001), 1077–1087.
205. P. ORBANZ, *Projective limit random probabilities on Polish spaces,* Electron. J. Statist. **5** (2011), 1354–1373.
206. F. PAPANGELOU, *Large-deviations and the Bayesian estimation of higher-order Markov chains,* J. Appl. Probab. **33** (1996), 18–27.
207. S. PETRONE, M. GUINDANI, and A. GELFAND, *Hybrid Dirichlet mixture models for functional data,* J. Roy. Statist. Soc. **B71** (2009), 755–782.
208. J. PFANZAGL, *On the existence of consistent estimates and tests,* Z. Wahrsch. verw. Gebiete **10** (1968), 43—62.
209. E. PHADIA, *Prior processes and their applications: nonparametric Bayesian statistics,* Springer, New York (2013).
210. M. PINTER, *The existence of an inverse limit of an inverse system of measure spaces – a purely measurable case,* Acta Math. Hungar. **126.1-2** (2010), 65–77.
211. YU. PROKHOROV, *Convergence of random processes and limit theorems in probability theory,* Theory Probab. Appl. **1.2** (1956), 157—214.
212. H. RAIFFA, and R. SCHLAIFER, *Decision analysis: introductory lectures on choices under uncertainty,* Addison-Wesley, Reading (1961).
213. J. RAMSAY, and B. SILVERMAN, *Functional Data Analysis (2nd edn.),* Springer, New York (2005).
214. C. RAO, *Information and the accuracy attainable in the estimation of statistical parameters,* Bull. Calcutta Math. Soc. **37** (1945), 81–91.
215. M. RAO, *Projective limits of probability spaces,* Journal of Multivariate Analysis **1.1** (1971), 28–57.
216. M. RAO, *Foundations of stochastic analysis,* Academic press, New York (1981).
217. O. REIERSØL, *Identifiability of a linear relation between variables which are subject to error,* Econometrica **18** (1950), 375–389.
218. C. ROBERT, *The Bayesian choice: from decision-theoretic foundations to computational implementation,* Springer, New York (2001).
219. B. RIPLEY, *Pattern recognition and neural networks*, Cambridge University Press, Cambridge (1996).
220. G. ROUSSAS, *Contiguity of probability measures: some applications in statistics,* Cambridge Tracts in Mathematics and Mathematical Physics **63** (1972), Cambridge University Press, London-New York.

221. W. RUDIN, *Functional Analysis (Second edition),* McGraw-Hill, New York (1991).
222. L. SAVAGE, *The subjective basis of statistical practice,* Technical report, Dept. Statistics, University of Michigan (1961).
223. H. SCHAEFER, *Topological vector spaces,* Springer, New York (1999).
224. M. SCHERVISH, *Theory of statistics,* Springer, New York (1995).
225. LORRAINE SCHWARTZ, *Consistency of Bayes procedures,* PhD. thesis, Dept. of Statistics, University of California, Berkeley (1961).
226. LORRAINE SCHWARTZ, *On Bayes procedures,* Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete **4** (1965), 10–26.
227. LAURENT SCHWARTZ, *Radon measures on arbitrary topological spaces and cylindrical measures,* Tata Institute of Fundamental Research, Oxford University Press, Oxford, (1973).
228. LAURENT SCHWARTZ, *Lectures on disintegration of measures,* Tata Institute of Fundamental Research, Oxford University Press, Oxford, (1975).
229. G. Schwarz, *Estimating the dimension of a model*, Ann. Statist. **6** (1978), 461–464.
230. C. SHANNON, *A Mathematical Theory of Communication,* Bell System Technical Journal **27** (1948), 379–423, 623–656.
231. X. SHEN and L. WASSERMAN, *Rates of convergence of posterior distributions*, Ann. Statist. **29** (2001), 687–714.
232. X. SHEN, *Asymptotic normality of semiparametric and nonparametric posterior distributions,* Journal of the American Statistical Association **97** (2002), 222–235.
233. M. SION, *On general minimax theorems,* Pacific J. Math **8** (1958), 171–176.
234. C. STEIN, *Inadmissibility of the usual estimator for the mean of a multivariate distribution,* Proc. Third Berkeley Symp. Math. Statist. Prob. **1** (1956), 197—206.
235. F. STEUTEL and K. VAN HARN, *Infinite Divisibility of Probability Distributions on the Real Line,* Marcel Dekker, New York (2004).
236. H. STRASSER, *Mathematical Theory of Statistics,* de Gruyter, Amsterdam (1985).
237. C.STRELIOFF, J. CRUTCHFIELD and A. HUBLER, *Inferring Markov chains: Bayesian estimation, model comparison, entropy rate, and out-of-class modeling,* arXiv:math/0703715 [math.ST] (2007).
238. S. SUWAN, et al. *Empirical Bayes estimation for the stochastic block- model,* Electron. J. Statist. **10.1** (2016), 761–782.
239. B. SZABÓ, A. VAN DER VAART and J. VAN ZANTEN, Frequentist coverage of adaptive nonparametric Bayesian credible sets, Ann. Statist. **43** (2015), 1391–1428.
240. M. TAUPIN, *Semi-parametric estimation in the nonlinear structural errors-in-variables model,* Ann. Statist. **29** (2001), 66–93.
241. J. TAYLOR and R. TIBSHIRANI, Statistical learning and selective inference, Proc. Natl. Acad. Sc. **112** (2016), 7629–7634.
242. E. TORGESEN, *Comparison of statistical experiments,* Cambridge University Press, Cambridge (1991).
243. F. TRÈVES, *Topological Vector Spaces, Distributions and Kernels,* Dover Publications, Mineola (NY) (2006).
244. J. TUKEY, *Exploratory data analysis,* Addison-Wesley, Reading (1977).
245. A. VAN DER VAART, *Estimating a real parameter in a class of semiparametric models,* Ann. Statist. **16** (1988), 1450–1474.
246. A. VAN DER VAART, *Efficient maximum-likelihood estimation in semiparametric mixture models,* Ann. Statist. **24** (1996), 862–878.
247. A. VAN DER VAART and J. WELLNER, *Weak Convergence and Empirical Processes,* Springer, New York (1996).
248. A. VAN DER VAART, *Asymptotic Statistics,* Cambridge University Press, Cambridge (1998).
249. A. WALD, *An essentially complete class of admissible decision functions,* Ann. Math. Statist. **18.4** (1947), 549–555.
250. A. WALD, *Statistical Decision Functions,* J. Wiley, New York (1950).
251. A. WALKER, *On the asymptotic behaviour of posterior distributions,* J. Roy. Statist. Soc. **B31** (1969), 80–88.

252. S. WALKER, New approaches to Bayesian consistency, Ann. Statist. **32** (2004), 2028–2043.
253. S. WALKER, A. LIJOI and I. PRÜNSTER, *Contributions to the understanding of Bayesian consistency,* ICER Applied Mathematics Working Paper Series **13** (2004).
254. S. WALKER, A. LIJOI and I. PRÜNSTER, Data tracking and the understanding of Bayesian consistency, Biometrika **92** (2005), 765–778.
255. S. WALKER, A. LIJOI and I. PRÜNSTER, On rates of convergence for posterior distributions in infinite-dimensional models, Ann. Statist. **35** (2007), 738–746.
256. L. WASSERMAN, *Bayesian model selection and model averaging*, J. Math. Psych. **44** (2000), 92–107.
257. L. WASSERMAN, *Bayesian Model Selection and Model Averaging,* J. Math. Psychology **44** (2000), 92-–107.
258. G. YANG, A conversation with Lucien Le Cam, Statist. Sc. **14** (1999), 223–241.
259. Y. YANG and A. BARRON, *An asymptotic property of model selection criteria*, IEEE Transactions on Information Theory **44** (1998), 95–116.
260. A. ZAANEN, and W. LUXEMBURG, *Riesz spaces, volume II*, North-Holland, Amsterdam (1983).
261. A. ZHANG, and H. ZHOU, *Minimax rates of community detection in stochastic block models,* Ann. Statist. **44.5** (2016), 2252–2280.
262. C.-H. ZHANG and J. HUANG, *The sparsity and bias of the LASSO selection in high-dimensional linear regression,* Ann. Statist. **36** (2008), 1567–1594.

Index