Bayesian Statistics

Lecture notes of a course by
Botond Szabó and Aad van der Vaart
(Università Bocconi and TU Delft)
version 20-7-2023

(Warning: may look convincing at first reading.)

Contents

Notation

$X \sim Y$: the random variables X and Y possess the same distribution.
$X \sim P$: the random variable has distribution P.
$X_i \overset{\text{iid}}{\sim} P$: the random variables $X_1, X_2, \ldots$, are independent and $X_i \sim P$.
$X_i \overset{\text{ind}}{\sim} P_i$: the random variables $X_1, X_2, \ldots$, are independent and $X_i \sim P_i$.
$X \sim Y|\, Z$: same as above but conditionally on Z.
$X_i|\, Z \overset{\text{iid}}{\sim} P$: same as above but conditionally on Z.
$X_i|\, Z_i \overset{\text{ind}}{\sim} P_i$: same as above but conditionally on Z.
$U \perp\!\!\!\perp V$: the random variables U and V are independent.
$U \perp\!\!\!\perp V|\, W$: the random variables U and V are conditionally independent given W.
Pf: $\int f\, dP = \mathrm{E} f(X)$, if $X \sim P$.
P^n: product measure of n copies of P.
$P^n f$: $\int f\, dP^n = \mathrm{E} f(X_1, \ldots, X_n)$, if $X_1, \ldots, X_n \overset{\text{iid}}{\sim} P$.
$\mathbb{P}_n$: *empirical measure* of $X_1, \ldots, X_n$, the random discrete probability measure that puts mass $1/n$ at every X_i.
$\mathbb{P}_n f$: $n^{-1} \sum_{i=1}^n f(X_i)$.
$\mathbb{S}_n$: n-dimensional unit simplex, all probability vectors $(w_1, \ldots, w_n)$ of length n.
$\mathbb{S}_\infty$: all probability vectors of countably infinite length.
$\|x\|_r$: $(\sum |x_i|^r)^{1/r}$, for $r \geq 1$ and $x \in \mathbb{R}^d$, or $x \in \mathbb{R}^\infty$.
$\|f\|_r$: $(\int |f|^2\, d\nu)^{1/r}$, for a function $f: \mathfrak{X} \to \mathbb{R}$.
h: Hellinger distance.
d_{TV}: total variation distance.
$K(P; Q)$, $K(p; q)$: Kullback–Leibler (KL) divergence.
$V(P; Q)$, $V(p; q)$: Kullback–Leibler (KL) variation, see (5.4).
$K_2(P; Q)$, $K_2(p; q)$: maximum of K and V.
$D(\epsilon, S, d)$: ϵ-packing number of B with respect to distance d.
$N(\epsilon, S, d)$: ϵ-covering number of B with respect to distance d.
$N_{[]}(\epsilon, S, d)$: ϵ-bracketing number of B with respect to distance d.
$\mathfrak{M}(\mathfrak{X})$: space of probability measures on a sample space $\mathfrak{X}$
$\mathfrak{M}_\infty(\mathfrak{X})$: space of (positive) measures on a sample space $\mathfrak{X}$
$\mathfrak{C}(\mathfrak{X})$: space of continuous functions on $\mathfrak{X}$
$\mathfrak{C}_b(\mathfrak{X})$: space of bounded continuous functions on $\mathfrak{X}$
$\mathfrak{C}^\alpha(\mathfrak{X})$: Hölder space of functions of smoothness α on $\mathfrak{X}$
$\mathfrak{UC}(\mathfrak{X})$: space of uniformly continuous functions on $\mathfrak{X}$
$\mathfrak{W}^\alpha(\mathfrak{X})$: Sobolev space of functions of smoothness α on $\mathfrak{X}$
$\xrightarrow{P}$: convergence in probability

$\rightsquigarrow$: convergence in law/distribution, weak convergence
$\lesssim$: smaller than up to a multiplicative constant that is fixed in the context.
$\mathrm{Nor}(\mu, \sigma^2)$: normal distribution with mean μ and variance σ^2.
$\mathrm{Nor}_k(\mu, \Sigma)$: k-variate normal distribution with mean vector μ and dispersion matrix Σ.
$\phi_{\mu,\Sigma}$: normal density with mean (vector) μ and dispersion (matrix) Σ.
$\mathrm{Bin}(n, p)$: binomial distribution with parameters n and p.
$\mathrm{Exp}(\lambda)$: exponential distribution with mean $1/\lambda$.
$\mathrm{Ga}(a, b)$: gamma distribution with shape a and scale b.
$\mathrm{Be}(a, b)$: beta distribution with parameter a and b.
$\mathrm{MN}_k(n; p_1, \ldots, p_k)$: k-variate multinomial distribution with n trials.
$\mathrm{Dir}(k; \alpha_1, \ldots, \alpha_k)$: k-dimensional Dirichlet distribution.

Preface

These notes are a mathematically rigorous introduction to Bayesian statistical inference, with an emphasis on nonparametric methods.

Knowledge of measure theory and probability are assumed throughout. Each chapter ends with a section that reviews other useful background.

1

The Bayesian Paradigm

In statistics *data* x is modelled as a realization of a random variable X. The *statistical model* is the set of possible probability distributions P_θ of X, indexed by a *parameter* θ running through a *parameter set* Θ.

The Bayesian statistical approach adds the assumption, or working hypothesis, that the parameter θ itself is also a realization of a random variable, say ϑ. The distribution P_θ is then considered the *conditional distribution* of X given $\vartheta = \theta$. In this setup the pair (X, ϑ) then has a joint probability distribution, given by

$$\Pr(X \in A, \vartheta \in B) = \int_B P_\theta(A)\, d\Pi(\theta). \tag{1.1}$$

Here Π is the *marginal distribution* of ϑ, which is called the *prior distribution* in this context, meaning the distribution of the parameter before the data x was collected. Once the data is available the Bayesian statistician will "update" this prior distribution to the conditional distribution of ϑ given $X = x$, called the *posterior distribution*: the distribution on Θ given by

$$\Pi(B|\, x) = \Pr(\vartheta \in B|\, X = x).$$

This is considered to contain all the relevant information; all further inference will be based on it.

This updating is perfectly natural to anybody with a basic knowledge of probability. However, the Bayesian paradigm is a little controversial in its insistence that the parameter is a random variable. As an assumption on the world this is untenable to most scientists. As a working hypothesis or an "expression of uncertainty" it is much easier to accept, with still the objection that the final result of the analysis will depend on the prior distribution Π, and this may be chosen differently by different scientists.

De Finetti's theorem is often brought forward in favor of the random variable assumption. According to this theorem a sequence $X_1, X_2, \ldots$ is *exchangeable*, that is has a joint distribution that is invariant under permutation of its elements, if and only if there exists a variable ϑ such that given ϑ the variables $X_1, X_2, \ldots$ are i.i.d. Since exchangeability is a reasonable assumption in many situations, the prior variable ϑ arises naturally.

The debate may also zoom in on the question whether there is a "true parameter". For a subjectivist Bayesian (i.e. a true Bayesian) there is none; statistical inference is only an expression of uncertainty. For most scientists there is a true parameter, often denoted by θ_0; it is assumed that the data x was generated according to the measure P_{θ_0}.

In this course we shall not dwell too much on philosophical questions. In fact, we shall be

both Bayesian and non-Bayesian, or *frequentist* as the opposite of Bayesian is often called, for not too clear reasons. We shall adopt the Bayesian approach to obtain posterior distributions, and next study these posterior distributions (in their dependence on x) under the assumption that x was generated from a "true distribution". We can then pose and answer the question whether the Bayesian methods "works", relate it to other statistical methods, and compare different priors. Some priors turn out to work better than others.

1.1 Measure-theoretic definitions

Random variables and their (conditional) distributions are mathematically defined in measure theory. We assume the relevant definitions known, but in this section review some facts on conditional distributions. These are complicated objects, in particular in the nonparametric setup that we shall be concerned with later on.

The random variables X and ϑ are formally defined as measurable maps on a probability space, with values in measurable spaces $(\mathfrak{X}, \mathscr{X})$ and $(\Theta, \mathscr{B})$. Here $\mathscr{X}$ and $\mathscr{B}$ are σ-fields of subsets of the (arbitrary) sets $\mathfrak{X}$ and Θ, called the *sample space* and the *parameter space*. We are usually not concerned with the way that X and Θ are defined on the underlying probability space, but only with their induced laws on $(\mathfrak{X}, \mathscr{X})$ and $(\Theta, \mathscr{B})$, or their joint law on the product of these spaces. To work with these laws the following definition is crucial.

Definition 1.1 (Markov kernel)**.** A *Markov kernel* from a measurable space $(\mathfrak{X}, \mathscr{X})$ into another measurable space $(\mathfrak{Y}, \mathscr{Y})$ is a map $Q\colon \mathfrak{X} \times \mathscr{Y} \to [0, 1]$ such that:

(i) The map $B \mapsto Q(x, B)$ is a probability measure for every $x \in \mathfrak{X}$.
(ii) The map $x \mapsto Q(x, B)$ is measurable for every $B \in \mathscr{Y}$.

Informally the Markov kernel $Q(x, \cdot)$ can be viewed as the distribution of a variable Y with values in $\mathfrak{Y}$ given that a variable X with values in $\mathfrak{X}$ takes the value x. Corresponding to this a Markov kernel is also called a *regular conditional distribution*, and we shall typically write $Q(B|\, x)$ instead of $Q(x, B)$. This is the probability that the (imagined) variable Y falls in the set B given the knowledge that the (imagined) variable X has been realized as the value x.

For the basic definitions of Bayesian statistics we need two types of Markov kernels. As starting point we assume that the statistical model $(P_\theta\colon \theta \in \Theta)$ is given by a Markov kernel from $(\Theta, \mathscr{B})$ into $(\mathfrak{X}, \mathscr{X})$:

(i) The map $A \mapsto P_\theta(A)$ is a probability measure for every $\theta \in \Theta$.
(ii) The map $\theta \mapsto P_\theta(A)$ is measurable for every $A \in \mathscr{X}$.

This formalizes viewing P_θ as the conditional distribution of X given $\vartheta = \theta$.

The assumed measurability (ii) ensures that the integral on the right side of (1.1) is well defined, for every measurable sets $A \in \mathscr{X}$ and $B \in \mathscr{B}$. The integral defines a number in $[0, 1]$ for every set $A \times B$ in the product σ-field $\mathscr{X} \times \mathscr{B}$. By the usual measure-theoretic construction this can be extended to exactly one probability measure on this product σ-field. This measure formalizes the joint distribution of the pair (X, ϑ) mentioned in the introduction.

Starting from the statistical model as a Markov kernel we have constructed the "law of the pair" (X, ϑ), but not the pair itself. The latter has little relevance, as all later developments

will concern the law and not the pair of variables. However, viewing the right hand side of (1.1) as the distribution of some pair (X, ϑ) yields very convenient notation. A trivial way to define also the random variables is to take the underlying probability space equal to $\Omega = \mathfrak{X} \times \Theta$ with the product σ-field $\mathscr{X} \times \mathscr{B}$, and define $(X, \vartheta)\colon \Omega \to \mathfrak{X} \times \Theta$ as the identity map: $X(\omega) = x$ and $\vartheta(\omega) = \theta$ if $\omega = (x, \theta) \in \Omega$. Then $\{\omega\colon (X(\omega), \vartheta(\omega)) \in A \times B\} = A \times B$ and hence (1.1) holds.

From (1.1) the marginal law of ϑ is obtained by setting $A = \mathfrak{X}$. For this choice the equation reduces to $\Pr(\vartheta \in B) = \Pi(B)$, whence ϑ has law Π.

The marginal distribution of X is similarly obtained by setting $B = \Theta$ in (1.1), and is given by

$$P(A) := \Pr(X \in A) = \int P_\theta(A)\, d\Pi(\theta). \tag{1.2}$$

This is a *mixture* of the distributions P_θ. This *Bayesian marginal law* should not be confused with the "frequentist law" of the observation X, which is the distribution P_{θ_0} of X under the "true" parameter θ_0. Instead of marginal law, Bayesians also say *prior predictive distribution*.

Next we turn to the posterior distribution, which should be the conditional distribution of ϑ given $X = x$. Now conditioning on an event of probability zero is problematic. Since the events $\{X = x\}$ are often zero events, we *define* the posterior distribution through integration, as follows.

Definition 1.2 (Posterior distribution)**.** Given a model $(P_\theta\colon \theta \in \Theta)$, given as a Markov kernel, and a prior probability distribution Π, a *posterior distribution* is a Markov kernel $(x, B) \mapsto \Pi(B|\, x)$ from $(\mathfrak{X}, \mathscr{X})$ into $(\Theta, \mathscr{B})$ such that, for every $A \in \mathscr{X}$ and $B \in \mathscr{B}$,

$$\Pr(X \in A, \vartheta \in B) = \int_A \Pi(B|\, x)\, dP(x). \tag{1.3}$$

Here (X, ϑ) is the random variable satisfying (1.1) constructed in the preceding, and P is the marginal law of X defined in (1.2).

Equation (1.3) is analogous to (1.1), but on the right side the roles of X and ϑ are swapped. One says that both equations give *disintegrations* of the joint law of (X, ϑ), but in different orders.

Another Markov kernel $(x, B) \mapsto \Pi_1(B|\, x)$ such that $\Pi_1(B|\, x) = \Pi(B|\, x)$, for almost every x under the marginal law P, will also satisfy (1.3), for every A. Hence a posterior distribution is uniquely defined only up to a null set of x-values. Since we think of these as negligible, we shall often speak of *the* posterior distribution, despite the possible lack of uniqueness. A potential embarrassment is that the null sets are relative to the marginal law P, which is the relevant law for Bayesians, but not for a statistician who believes the data were in reality generated from a true distribution P_{θ_0}. Usually, but not always, a null set for P is also a null set for every P_θ; if not, then additional criteria are needed to work with a posterior distribution.

It is not true in full generality that a posterior distribution *exists*. By Kolmogorov's definition of conditional expectation from measure theory, for every $B \in \mathscr{B}$ there always exists a measurable map $x \mapsto \Pi(B|\, x)$ that satisfies (1.3) for every $A \in \mathscr{X}$ (see proof of next proposition). The map $(x, B) \mapsto \Pi(B|\, x)$ then satisfies requirement (ii) of the definition of a Markov

kernel. However, requirement (i) of the definition of a Markov kernel imposes relations between the maps $x \mapsto \Pi(B|\, x)$ for different B, and it is not true that the maps can always be chosen so that these are satisfied and $(x, B) \mapsto \Pi(B|\, x)$ is a Markov kernel. In the following proposition we see that usually they can. The key condition is that there are "not too many" sets B, a condition that is expressed in topological terms.

Proposition 1.3 (Existence of posterior distribution)**.** *Suppose that there exists a metric d on Θ under which Θ is separable and complete, and assume that $\mathscr{B}$ is the σ-field generated by the open sets for this metric. Then there exists a Markov kernel $(x, B) \mapsto \Pi(B|\, x)$ from $(\mathfrak{X}, \mathscr{X})$ into $(\Theta, \mathscr{B})$ that satisfies (1.3) for every $A \in \mathscr{X}$ and $B \in \mathscr{B}$. Furthermore, any other such Markov kernel $(x, B) \mapsto \Pi_1(B|\, x)$ satisfies $\Pi_1(B|\, x) = \Pi(B|\, x)$ for every $B \in \mathscr{B}$, for all x except possibly in a set N with $\Pr(X \in N) = 0$, where the set N can be chosen independent of B.*

Proof This result can be found in any good text on measure-theoretic probability. E.g. Bauer (1981), Theorem 10.3.5 (about three or four pages of proof). The main work is to put versions of the conditional probabilities $\Pi(B|\, x)$ together to a probability measure $B \mapsto \Pi(B|\, x)$. The existence of the conditional probabilities itself is a consequence of the Radon-Nikodym theorem, as follows. For given B we consider the measure $Q(A) = \Pr(X \in A, \vartheta \in B)$. This is absolutely continuous with respect to P: if $\Pr(X \in A) = 0$, then also $Q(A) = 0$. Thus there exists a measurable map $x \mapsto f_B(x)$ on $\mathfrak{X}$ such that $Q(A) = \int_A f_B(x)\, dP(x)$, for all $A \in \mathscr{X}$. This is a version of the map $\Pi(B|\, x)$. □

The σ-field generated by the open sets in a topological space is called the *Borel σ-field*. A topological space whose topology is generated by a metric under which the space is separable and complete is called a *Polish topological space*. Thus the preceding proposition can be paraphrased as saying that a posterior distribution exists as soon as $(\Theta, \mathscr{B})$ is a Polish space with its Borel σ-field. This will be the case in all examples of interest.

Example 1.4 (Euclidean space)**.** The space $\mathbb{R}^d$ is complete and separable under the usual metric and hence Polish. The usual σ-field, which is also generated by the cells $(a, b]$, is the Borel σ-field.

Example 1.5 (Unit interval)**.** The open unit interval $(0, 1)$ is not complete under the usual metric, but it is still Polish. For instance, the topology is also generated by the metric $d(x, y) = \left|\Phi^{-1}(x) - \Phi^{-1}(y)\right|$, for Φ the standard normal distribution function, which makes $(0, 1)$ into a complete space. (Sequences $\{x_n\} \subset (0, 1)$ that approach 0 or 1 in the usual topology are not Cauchy for d. This solves the problem that $(0, 1)$ is not closed at its end points.) The usual σ-field, which is also generated by the cells $(a, b]$, is the Borel σ-field for this metric.

Example 1.6 (Sequence space)**.** The space $\mathbb{R}^\infty$ is complete and separable under the metric $d(x, y) = \sum_{i=1}^\infty 2^{-i}(|x_i - y_i| \wedge 1)$, if $x = (x_1, x_2, \ldots)$ and $y = (y_1, y_2, \ldots)$, and hence Polish. The usual σ-field, generated by the cylindrical sets, can be shown to be the Borel σ-field for d.

The second and third examples are generalized in the following lemma, which shows that Polish spaces abound. We shall encounter other examples of interest later on, including spaces of measures and spaces of functions.

Lemma 1.7. (i) *Every closed subset of a Polish space is again a Polish space.*
(ii) *Every open subset of a Polish space is again a Polish space.*
(iii) *The product of finitely many or countably many Polish spaces is a Polish space.*

Proof Assertion (i) is trivial, as a closed subset of a separable, complete space is separable and complete. Assertion (iii) follows by constructing a metric on the product in the manner of Example 1.6, where the Euclidean distances $|x_i - y_i|$ are replaced by the (complete) distances of the individual spaces in the product. For the proof of (ii), let G be an open subset of the Polish space E. Since $F = G^c$ is closed, the map $\psi\colon \mathbb{R} \times E \to \mathbb{R}$ given by $\psi(r, x) = r\, d(x, F)$ is continuous. Hence the inverse image $G_0 := \psi^{-1}(\{1\}) = \{(r, x)\colon r\, d(x, F) = 1\}$ is closed in $\mathbb{R} \times E$. By (iii) the latter product is Polish and hence by (i) this inverse image is Polish. Now the map $\phi\colon G_0 \to E$ given by $(r, x) \mapsto x$ can be seen to be a homeomorphism of G_0 onto G. This implies that G inherits its Polishness from G_0. □

1.2 Bayes's rule

Nowadays Bayes's rule is taught as a probability rule, but Thomas Bayes (1701–1761), was in fact a statistician (the first Bayesian statistician). We present a version of his rule as a concrete formula for the posterior distribution, available in a special but common situation.

This situation is that the measures P_θ in the statistical model permit jointly measurable densities relative to a given σ-finite measure μ on the sample space $(\mathfrak{X}, \mathscr{X})$. We assume that there exists a map $(x, \theta) \mapsto p_\theta(x)$ that is measurable for the product σ-field $\mathscr{X} \times \mathscr{B}$ such that $P_\theta(A) = \int_A p_\theta(x)\, d\mu(x)$, for every $A \in \mathscr{X}$. In this case a version of the posterior distribution is given by *Bayes's formula*:

$$\Pi(B|\, x) = \frac{\int_B p_\theta(x)\, d\Pi(\theta)}{\int p_\theta(x)\, d\Pi(\theta)}. \tag{1.4}$$

Of course, this expression is defined only if the denominator $\int p_\theta(x)\, d\Pi(\theta)$ is nonzero. For x such that the denominator is zero the definition is not important, but for definiteness we define the quotient in that case to be equal to $Q(B)$, for an arbitrary probability measure Q on $(\Theta, \mathscr{B})$.

Proposition 1.8 (Bayes's formula)**.** *If there exists a σ-finite measure μ on the sample space $(\mathfrak{X}, \mathscr{X})$ and jointly measurable maps $(x, \theta) \mapsto p_\theta(x)$ such that $P_\theta(A) = \int_A p_\theta(x)\, d\mu(x)$, for every $A \in \mathscr{X}$, then formula (1.4) gives an expression for the posterior distribution.*

Proof It is one of the assertions of Fubini's theorem that the integrals with respect to one argument $\int_B p_\theta(x)\, d\Pi(\theta)$ and $\int p_\theta(x)\, d\Pi(\theta)$ of the jointly measurable function $(x, \theta) \mapsto p_\theta(x)$ are measurable functions of the remaining argument x. Their quotient is then also measurable, and hence $\Pi(B|\, x)$ defined by (1.4) satisfies requirement (ii) of a Markov kernel.

It is clear from the definition that $\Pi(\varnothing|\, x) = 0$ and $\Pi(\Theta|\, x) = 1$. Furthermore, the map $B \mapsto \Pi(B|\, x)$ is additive over disjoint sets, by the properties of integrals. Applying the monotone convergence theorem to the indicators of a sequence $\cup_{i=1}^n B_i$ of unions of disjoint sets, we see that is also countably additive. Thus requirement (i) is satisfied as well, so that $(x, B) \mapsto \Pi(B|\, x)$ is a Markov kernel.

By (1.1) and the representation of $P_\theta(A)$ as an integral, followed by Fubini's theorem,

$$\Pr(X \in A, \vartheta \in B) = \int_B \int_A p_\theta(x)\, d\mu(x)\, d\Pi(\theta) = \int_A \int_B p_\theta(x)\, d\Pi(\theta)\, d\mu(x).$$

Setting $B = \Theta$ we see that $p(x) := \int p_\theta(x)\, d\Pi(\theta)$ is a density of P relative to μ. Then the set $N = \{x\colon p(x) = 0\}$ is a null set for P, and the preceding display does not change if we intersect A with N^c. On $A \cap N^c$, we can rewrite the inner integral in the expression on the far right as $\Pi(B|\, x)p(x)$, by (1.4), and hence we find that the left side is equal to $\int_{A\cap N^c} \Pi(B|\, x)p(x)\, d\mu(x)$. This does not change if we replace $A \cap N^c$ again by A. Thus we have verified that $\Pi(B|\, x)$ satisfies the disintegration (1.3). □

Bayes's formula can be best memorized as a reweighting rule. A priori the value θ of the parameter has weight $d\Pi(\theta)$. The data x is observed with "probability proportional to" the likelihood $p_\theta(x)$. A posteriori the weight of θ is proportional to $p_\theta(x)\, d\Pi(\theta)$. With the symbol $\propto$ meaning "is proportional to as a function of θ", we can express this by the formula

$$d\Pi(\theta|\, x) \propto p_\theta(x)\, d\Pi(\theta).$$

Bayes's rule is the integrated form of this formula. The denominator $p(x) = \int p_\theta(x)\, d\Pi(\theta)$ in (1.4) acts as the norming constant for the density, which makes the quotient (1.4) equal to 1 for $B = \Theta$. Note that p is precisely the marginal density of X. Bayesians call it also the *evidence* or *prior predictive density*.

Example 1.9 (Binomial distribution)**.** Thomas Bayes derived his rule for a special model and prior, which we nowadays can describe as a binomial observation with uniform prior on the success probability. Thus

$$p_\theta(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, \qquad d\Pi(\theta) = d\theta, \qquad x = 0, 1, \ldots, n,\ \theta \in (0,1).$$

In this case the posterior distribution takes the form

$$d\Pi(\theta|\, x) \propto \theta^x(1-\theta)^{n-x}d\theta, \qquad \theta \in (0,1).$$

This distribution is known as the *beta distribution* with parameters $x + 1$ and $n - x + 1$. Figure 1.1 shows the prior density and the posterior density for $n = 20$ and $x = 7$.

The denominator in Bayes's rule ensures that the total mass of the measure is equal to 1. This is true even if Π itself is not a probability measure. A true prior is a probability measure, and it is not useful to replace it by a finite measure of total mass unequal to 1, but it can be useful to employ an infinite measure. Provided the integral in the denominator is finite, the right side of formula (1.1) still defines a random probability measure, which we could call a posterior measure. A "prior" of infinite mass is called *improper*.

Example 1.10 (Normal distribution)**.** For an observation X from a normal distribution with unknown mean θ and variance 1 and Π Lebesgue measure, Bayes's rule gives

$$d\Pi(\theta|\, x) = \frac{\phi(x-\theta)\, d\theta}{\int \phi(x-\theta)\, d\theta} = \phi(x-\theta)\, d\theta.$$

Thus the posterior distribution is Gaussian with mean x and variance 1.

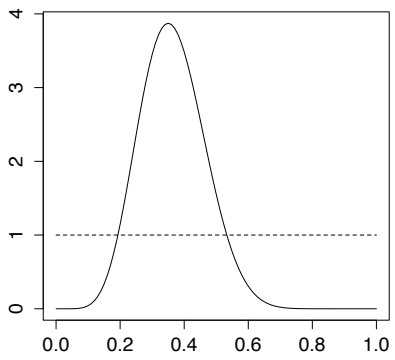

Figure 1.1 Uniform prior density (dashed) and posterior density for observation $x = 7$ from a binomial distribution with parameters $n = 20$ and success probability θ.

1.3 Hierarchical modelling

The Bayesian model can be portrayed as a two-step algorithm to generate the pair (x, θ) of a data point and a parameter:

- θ is generated from Π.
- x is generated from P_θ.

The two steps concern the parameter θ and the data x, respectively. There is no distinction between these quantities other than that the first is not observed, whereas the second is. Bayesian modelling can be viewed more generally as describing the mechanism by which data and unobservables have been generated. Here the "unobservables" may well consist of other variables besides the parameter that indexes the model for the observables, and the *generative model* may well consist of a hierarchy of more than two steps. Every unobservable receives a distribution in this hierarchy, and after observing the data these distributions are updated to their conditional distributions given the data. Depending on their nature, unobservables may be called *latent variables* or parameters, but the conditional distribution of any unobservable given the data is called a "posterior distribution".

Some examples make this clearer.

Example 1.11 (Linear mixed model)**.** In a *linear mixed model* (also called *random effects model*) the observation is a vector Y with values in $\mathbb{R}^n$ satisfying the regression equation

$$Y = X\beta + Z\gamma + e,$$

where X and Z are known matrices (of dimensions $(n \times p)$ and $(n \times q)$ say), $\beta \in \mathbb{R}^p$ is an unknown parameter vector, and γ and e are independent random variables, in the simplest case with zero-mean normal distributions of dimensions q and n with unknown covariance

matrices D and $\sigma^2 I$, say. The "random effects" γ are not observed, but used to model dependencies between the coordinates of Y. The "error vector" e is also not observed, as usual in a regression model.

The distribution of the observable Y is normal with mean $X\beta$ and covariance matrix $ZDZ^T + \sigma^2 I$.

As a parameter vector we can take $\theta = (\beta, D, \sigma^2)$. In its most basic form the Bayesian hierarchy would be to generate θ from a prior (on $\mathbb{R}^p \times \mathfrak{D} \times (0, \infty)$, for $\mathfrak{D}$ the set of positive-definite matrices), and next Y from the $N_n(X\beta, ZDZ^T + \sigma^2 I)$-distribution.

However, it is attractive to split the generative model in more than two steps:

- (β, D, σ^2) are generated from a prior.
- γ is generated from $N_q(0, \sigma^2 D)$.
- e is generated from $N_n(0, \sigma^2 I)$.
- $Y = X\beta + Z\gamma + e$.

In this hierarchy the unobservables γ receive a similar treatment as the unobservables β, which are part of the parameter vector θ. This is even more so if the prior for β would also be chosen Gaussian, as is customary.

One concrete example is a *longitudinal study*, in which individuals are followed over time and measured at multiple occasions. The observational vector Y would then consist of blocks, and could be indexed as $(Y_{1,1}, \ldots, Y_{1,T}, Y_{2,1}, \ldots, Y_{2,T}, \ldots, Y_{S,1}, \ldots, Y_{S,T})^T$, where $Y_{s,1}, \ldots, Y_{s,T}$ are the consecutive measurements on individual s. A simple linear model would be

$$Y_{s,t} = \beta_0 + \beta_1 t + \gamma_{s,0} + \gamma_{s,1} t + e_{s,t},$$

where $e_{s,t}$ are i.i.d. univariate normal variables. The idea of this model is that every individual s follows a linear model in time t, but the intercept and slopes vary over the individuals: for individual s these are $\beta_0 + \gamma_{s,0}$ and $\beta_1 + \gamma_{s,1}$. The parameters β_0 and β_1 are the "population intercept and slope", while the parameters $\gamma_{s,0}$ and $\gamma_{s,1}$ are the deviations of individual s from the average. If the data consists of measurements on individuals sampled from some population, then it makes sense to think of the pairs $(\gamma_{s,0}, \gamma_{s,1})$ as an i.i.d. sample from a distribution. The vector $\gamma = (\gamma_{1,0}, \gamma_{1,1}, \ldots, \gamma_{S,0}, \gamma_{S,1})^T$ is then also random (with covariance matrix $\sigma^2 D$ having block structure, as individuals are independent).

One advantage of the Bayesian model is that one can speak naturally of the conditional distribution of the latent variable γ given the data Y. This will reveal the hidden structure behind the data. In the concrete example the posterior distribution of γ would show the variability of the intercepts and slopes in the population.

A possible prior for (β, D, σ^2) is given by the following scheme:

- $1/\sigma^2$ is generated from a $\Gamma(a, b)$-distribution.
- β is generated from a $N(0, \sigma^2 C)$-distribution.
- D is generated from a Wishart distribution with parameters $\sigma^2 \Delta$.

These three lines could replace the first line of the preceding hierarchy. The parameters a, b, C, Δ could be chosen fixed constants (such as $a = b = 0.001$ to obtain a widely spread prior), or one could add a further hierarchical step by generating these in term from a prior.

The subsequent steps in this hierarchical description may use values from previous steps.

The distributions from which the variables are generated are then understood to be conditional distributions given those variables, something that is often not made explicit. Furthermore and more confusing, if variables are not mentioned, then this usually is meant to imply stochastic independence.

Example 1.12 (Probit regression)**.** Suppose that the data consists of independent random variables $Y_1, \ldots, Y_n$ taking values in two-point set $\{0, 1\}$ and of fixed constants $x_1, \ldots, x_n$ such that

$$\Pr(Y_i = 1) = \Phi(\beta_0 + \beta_1 x_i) = 1 - \Pr(Y_i = 0).$$

The only parameters are β_0 and β_1, and hence a Bayesian analysis proceeds from a prior on $\beta = (\beta_0, \beta_1)$.

One could consider that the normal distribution function Φ appears only to map the linear regressions $\beta_0 + \beta_1 x$ into the interval $(0, 1)$, which is necessary to model the probabilities $\Pr(Y_i = 1)$. One could also give a more structural motivation for the form of the model through the following hierarchical Bayesian model:

- Generate (β_0, β_1) from a prior.
- Generate i.i.d. $\gamma_1, \ldots, \gamma_n$ from a standard normal distribution.
- Form $Z_i = \beta_0 + \beta_1 x_i + \gamma_i$.
- Set $Y_i = 1_{Z_i > 0}$.

One can check that this gives variables $Y_1, \ldots, Y_n$ following the model as before. The variables Z_i are not observable, but they explain the observables: Z_i can be viewed as a measure of fitness, or "propensity", of the ith unit. If the fitness is above a threshold (taken zero 0 here), then the observed value is 1; otherwise it is 0.

1.4 Empirical Bayes

Even in a multistep hierarchical model some prior specifications may be difficult. One may then opt for *objective priors* (see Section 2.2) or (other) *vague priors*, defined as priors that spread their mass evenly over the parameter space, in some way. A Gamma distribution with parameters $(0.001, 0.001)$ (so that the mean is 1 and the variance 1000) is a typical example of a vague prior on $(0, \infty)$.

An alternative is to "estimate the prior from the data". The standard *empirical Bayes* approach uses maximum likelihood for this purpose. Given a collection of priors Π_α that depends on some parameter α, this parameter is estimated by

$$\hat{\alpha} := \operatorname*{argmax}_\alpha \int p_\theta(X)\, d\Pi_\alpha(\theta).$$

The reasoning here is that $x \mapsto \int p_\theta(x)\, d\Pi_\alpha(\theta)$ is the density of the data X if the Bayesian model with prior Π_α would indeed be valid: it is the marginal density of X if θ is generated from the prior Π_α and next X generated from p_θ. This marginal density depends on α and hence it makes sense to apply the maximum likelihood principle to estimate it.

The prior $\Pi_{\hat{\alpha}}$ is next used to form a posterior as if $\hat{\alpha}$ were fixed from the beginning.

We shall later use this method to determine hyperparameters of nonparametric priors,

which are essential to good performance. The empirical Bayes method is also important in many other settings, as illustrated in the following intriguing example.

Example 1.13 (Shrinkage estimation)**.** Consider estimating a vector-valued parameter $\theta \in \mathbb{R}^d$ based on an observation X with the $N_d(\theta, \sigma^2 I)$-distribution, i.e. the coordinates $X_1, \ldots, X_d$ of X are independent random variables with $X_i \sim N(\theta_i, \sigma^2)$. There are no known relationships between the coordinates $\theta_1, \ldots, \theta_d$ of θ, and we consider σ^2 to be known.

The maximum likelihood estimator of θ is the vector X itself and has mean square error $\mathrm{E}_\theta\|X - \theta\|^2 = d\sigma^2$. Somewhat surprisingly, for $d \geq 3$ there exist estimators with a smaller mean square error, and even much smaller if d is big.

One way of deriving such an improvement is by a combination of Bayesian and empirical Bayesian thinking. As a working hypothesis assume that the coordinates $\theta_1, \ldots, \theta_d$ are realizations of i.i.d. variables $\vartheta_1, \ldots, \vartheta_d$ following a $N(0, A)$-prior distribution, for some $A > 0$. Then $(X_1, \vartheta_1)^T, \ldots, (X_d, \vartheta_d)^T$ are i.i.d. random vectors following a bivariate normal distribution

$$\begin{pmatrix} X_i \\ \vartheta_i \end{pmatrix} \overset{\text{iid}}{\sim} N_2\left(0, \begin{pmatrix} \sigma^2 + A & A \\ A & A \end{pmatrix}\right).$$

Under this model

$$\mathrm{E}(\vartheta_i | X_1, \ldots X_d) = \mathrm{E}(\vartheta_i | X_i) = \frac{A}{\sigma^2 + A} X_i = \Big(1 - \frac{\sigma^2}{\sigma^2 + A}\Big) X_i.$$

If we really believe the prior model, then these would be natural estimators for the parameters. Even then, we would probably not want to pretend to know the value of A.

We may estimate this prior parameter from the data. Under the Bayesian model the variables $X_1, \ldots, X_d$ are i.i.d. with Bayesian marginal distribution $N(0, \sigma^2 + A)$. The maximum likelihood estimator for $\sigma^2 + A$ in this model is given by $d^{-1}\|X\|^2$ and hence we might estimate $1/(\sigma^2 + A)$ by $d/\|X\|^2$. The variable $\|X\|^2/(\sigma^2 + A)$ possesses a chisquared distribution with d degrees of freedom. A calculation based on this will show that $\mathrm{E}(d-2)/\|X\|^2 = 1/(\sigma^2 + A)$. This motivates to replace d^{-1} in $d^{-1}\|X\|^2$ by $(d-2)^{-1}$ and use as estimators

$$\hat{\theta}_i = \Big(1 - \frac{\sigma^2(d-2)}{\|X\|^2}\Big) X_i.$$

It can be shown that $\mathrm{E}_\theta\|\hat{\theta} - \theta\|^2 < d\sigma^2$, for every $\theta \in \mathbb{R}^d$, and hence this estimator improves the standard estimator X for every parameter value.

The estimator $\hat{\theta}$ is known as the *Stein shrinkage estimator*. Because the multiplicative factor is strictly smaller than 1, and typically will be nonnegative, it "shrinks" the vector X to zero. That shrinkage is a good idea is also suggested by the fact that $\mathrm{E}_\theta\|X\|^2 = \|\theta\|^2 + d\sigma^2$: the vector X is "too big" for estimating the vector θ.

1.5 Bayesian inference

The posterior distribution, or the conditional distribution of some unobservable given the data, is the end-point of the Bayesian analysis. It expresses our knowledge on the value of the unobservable in the form of probabilities. However, we may wish to summarize the information in this full probability distribution by simpler quantities.

1.5.1 Location

The "location" of the posterior distribution is a natural point estimator of the parameter.

If the parameter set is a linear space that allows integration, then the *posterior mean* $\int \theta \, d\Pi(\theta|\, X)$ is the most popular measure of location.

If the parameter set is a subset of Euclidean space and the prior distribution has a density π relative to Lebesgue measure, and the model is given by densities p_θ, then the *posterior mode* is defined as

$$\operatorname*{argmax}_\theta \log \pi(\theta|\, X) = \operatorname*{argmax}_\theta \Big[\log p_\theta(X) + \log \pi(\theta)\Big].$$

The second way of writing the posterior shows that the posterior mode is a *penalized maximum likelihood estimator*, with penalty the log prior density. If the prior density were constant, then the posterior mode would be the maximum likelihood estimator.

The posterior mode means to give a point of highest posterior probability, but the latter concept is usually ill defined, as single parameter values typically have posterior probability zero. An alternative is to define for a given $\delta > 0$ a *δ-posterior mode* as

$$\operatorname*{argmax}_\theta \Pi(B(\theta, \delta)|\, X),$$

where $B(\theta, \delta)$ is the ball of radius δ around θ, for some metric on Θ. If the limit of the δ-modes as $\delta \downarrow 0$ exists, then this would be a good candidate for a definition of a posterior mode in general.

Another possible point estimator is the *spatial short*, which is defined as the center of the smallest ball containing posterior mass at least $3/4$ (or some other prescribed number).

1.5.2 Spread and credible sets

The spread is the second important characteristic of a distribution. The spread of the posterior distribution can be interpreted as measuring our uncertainty about the parameter after seeing the data.

The *posterior standard deviation* is a natural quantitative measure of spread.

A *credible set* is a map $x \mapsto C_x$ from $\mathfrak{X}$ to the measurable subsets $\mathscr{B}$ of the parameter space such that, for some prescribed level $1 - \alpha$,

$$\Pi(C_x|\, x) \geq 1 - \alpha.$$

The concept is reminiscent of a confidence set in non-Bayesian statistics, which is a set C_x such that $P_\theta(C_X \ni \theta) \geq 1 - \alpha$, for every $\theta \in \Theta$. However, the two concepts are different. The only link in general is the equality

$$\int \Pi(C_x|\, x) \, dP(x) = \Pr(\vartheta \in C_X) = \int P_\theta(C_X \ni \theta) \, d\Pi(\theta),$$

of the *integrals* of the two defining quantities. Confidence sets have a somewhat complicated interpretation, the level referring to the probability that data has been obtained so that the set covers the parameter. Credible sets express a probability concerning the parameter, and seem to have a more natural interpretation.

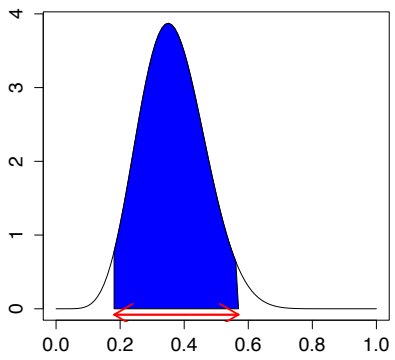

Figure 1.2 Credible set for observation $x = 7$ in a a binomial experiment with parameter $n = 20$ and uniform prior on the success probability θ. The shaded area corresponds to 95% posterior mass; the credible set is its projection on the horizontal axis.

Example 1.14 (Normal mean)**.** The posterior distribution for $\theta \in \mathbb{R}$ based on a sample of size n from the $N(\theta, 1)$-distribution relative to a $N(0, \lambda)$ prior distribution is the normal distribution with mean $(1 + \lambda^{-1}/n)^{-1}\bar{X}_n$ and variance $(n + \lambda)^{-1}$. A central interval of posterior probability $1-\alpha$ is $(1+\lambda^{-1}/n)^{-1}\bar{X}_n \pm (n+\lambda^{-1})^{-1/2}\xi_{1-\alpha/2}$, for ξ_α the standard normal α-quantile. The interval depends on the prior distribution through λ. For $\lambda \to \infty$ the interval tends to the "usual" interval $\bar{X}_n \pm n^{-1/2}\xi_{1-\alpha/2}$. We shall see later on that for $n \to \infty$ the interval also approaches the usual interval, for any λ.

1.5.3 Prediction

Within the Bayesian framework a *prediction* of a variable Y given data X is based on the conditional distribution Y of X. If desired this could be summarized by the location (such as $\mathrm{E}(Y|\, X)$) or a credible set (called predictive set in this context) corresponding to this conditional distribution.

If Y and X are conditionally independent given a parameter (or latent variable) ϑ and $\Pi(\cdot|\, x)$ gives the posterior distribution of ϑ given $X = x$, then the predictive distribution can be expressed as

$$\Pr(Y \in C|\, X = x) = \int \Pr(Y \in C|\, \vartheta = \theta)\, d\Pi(\theta|\, x).$$

Therefore, if Y and X are conditionally independent given $\vartheta = \theta$ and possess conditional densities q_θ and p_θ, then the predictive distribution has density

$$y \mapsto \int q_\theta(y)\, d\Pi(\theta|\, x).$$

This is known as the *posterior predictive density*. In particular, if $X_1, X_2, \ldots,$ are i.i.d. given

ϑ from a density p_θ, then the posterior predictive density of X_{n+1} given data $X_1, \ldots, X_n$ is

$$x_{n+1} \mapsto \int p_\theta(x_{n+1})\, d\Pi(\theta|\, x_1, \ldots, x_n).$$

This may be compared to the *prior predictive density*, which is obtained if $n = 0$, with $\Pi(\cdot|\, x_1, \ldots, x_n)$ interpreted as the prior distribution.

1.5.4 Bayes factors and testing

Statistical testing is concerned with choosing between competing hypotheses on the parameter. The usual setup, with a significance level and power function, does not fit well into the Bayesian approach. Instead Bayesians advocate to calculate the posterior probabilities of the various hypotheses.

Suppose that the hypotheses correspond to disjoint subsets Θ_i of the parameter set. Every set Θ_i receives both a prior weight λ_i (with $\lambda_i \geq 0$ and $\sum_i \lambda_i = 1$) and a prior distribution Π_i, that spreads its mass over Θ_i. This gives an overall prior $\Pi = \sum_i \lambda_i \Pi_i$ over Θ. If the data X has a density p_θ with respect to some dominating measure, then Bayes's rule gives the posterior distribution of θ as

$$d\Pi(\theta|\, x) \propto p_\theta(x)\, d\Pi(\theta) = \sum_i \lambda_i\, p_\theta(x)\, d\Pi_i(\theta).$$

The posterior probability of the set Θ_i is

$$\Pi(\Theta_i|\, x) = \frac{\lambda_i \int_{\Theta_i} p_\theta(x)\, d\Pi_i(\theta)}{\sum_j \lambda_j \int p_\theta(x)\, d\Pi_j(\theta)}.$$

If we had to choose for one of the hypotheses Θ_i, then it would be natural to choose the one with the highest posterior probability.

The comparison between a pair of models i and j is often summarized by the *Bayes factor*

$$\mathrm{BF}(\Theta_i, \Theta_j)(x) = \frac{\int_{\Theta_i} p_\theta(x)\, d\Pi_i(\theta)}{\int_{\Theta_j} p_\theta(x)\, d\Pi_j(\theta)}.$$

The Bayes factor is precisely the quotient of the Bayesian marginal densities of the data given models i and j, respectively. It is the likelihood ratio statistic for testing the two hypotheses in the Bayesian setup; it gives the most powerful test according to the Neyman-Pearson theory. A large value of $\mathrm{BF}(\Theta_i, \Theta_j)$ is an indication that model i gives a better fit to the data. This explains that the Bayesian marginal density is also called "evidence".

A proper comparison of the models should also take account of their prior probabilities λ_i. Indeed, the quotient of the posterior probabilities of the models is

$$\frac{\Pi(\Theta_i|\, x)}{\Pi(\Theta_j|\, x)} = \frac{\lambda_i}{\lambda_j}\, \mathrm{BF}(\Theta_i, \Theta_j)(x).$$

In words this reads: *posterior odds is equal to prior odds times Bayes factor*.

Bayesian model choice seems straightforward, but unfortunately there are many problems with its implementation. The following examples illustrate that one should be careful with models of different complexities, and that one should not use improper priors.

Example 1.15 (Lindley's paradox)**.** Consider data consisting of a random sample $X_1, \ldots, X_n$ from a normal distribution with mean θ and variance 1. We wish to decide between the hypotheses $H_0\colon \theta = 0$ and $H_1\colon \theta \neq 0$. The only possible prior under H_0 is the Dirac measure at 0. As a prior under H_1 we choose a normal distribution with mean 0 and variance τ^2, still to be determined. The Bayes factor between the hypotheses is then given by

$$\frac{\int \prod_{i=1}^n \phi(X_i - \theta)\, \phi(\theta/\tau)/\tau \, d\theta}{\prod_{i=1}^n \phi(X_i)} = \frac{1}{\sqrt{1+n\tau^2}}\, e^{\frac{1}{2} n\bar{X}^2/(1+(n\tau^2)^{-1})}.$$

For $n\tau^2 \gg 1$ and a fixed observation this expression has the same order of magnitude as $\tau^{-1} n^{-1/2} \exp(n\bar{X}^2/2)$.

A first observation is that the prior standard deviation τ plays a crucial role. A large value of τ leads to a small Bayes factor and hence evidence in favor of the null hypothesis. Large τ corresponds to giving more weight to bigger values of $|\theta|$ in the alternative hypothesis, and apparently makes the alternative hypothesis less likely. As $\tau \to \infty$ the Bayes factor even tends to zero, for any fixed values of the observations. This is usually not viewed as an advantage, as larger values of τ specify greater uncertainty on the value of θ, yielding *uninformative priors*.

For the Bayes factor to be larger than a threshold $c \geq 1$ it is necessary that

$$\sqrt{n}|\bar{X}| > \sqrt{2\log(c\tau) + \log n}.$$

Thus the Bayes factor gives evidence for H_1 only if $\sqrt{n}|X_n|$ exceeds $\sqrt{\log n}$ (if $\tau > 1$). On the other hand, the usual test rejects H_0 if $\sqrt{n}|\bar{X}_n| > \xi_{1-\alpha/2}$. For values between $\xi_{1-\alpha/2}$ and $\sqrt{\log n}$ the frequentist and Bayesian conclusions will be different. The p-value $2(1-\Phi(\sqrt{n}|\bar{X}_n|))$ can be very small even though the Bayes factor strongly favors H_0.

If the observations are distributed according to a true parameter value $\theta_{1,n}$, then $\sqrt{n}\bar{X}$ is $N(\sqrt{n}\theta_{1,n}, 1)$-distributed and hence takes its values in the interval $(\sqrt{n}\theta_{1,n} - M_n,\ \sqrt{n}\theta_{1,n} + M_n)$ with probability tending to one, for any $M_n \to \infty$. Then for values $\theta_{1,n}$ such that $\sqrt{n}|\theta_{1,n}| \to \infty$ slowly enough that $\sqrt{n}|\theta_{1,n}| \ll \sqrt{\log n}$, the frequentist and Bayesian procedures make different decisions with probability tending to one: the frequentist procedure rejects the null hypothesis (correctly) as $\sqrt{n}|\bar{X}_n| \geq \sqrt{n}|\theta_{1,n}| - M_n \to \infty$ and hence exceeds $\xi_{1-\alpha/2}$ with probability tending to one, whereas the Bayesian procedure does not reject the null hypothesis (incorrectly) as $\sqrt{n}|\bar{X}_n| \leq \sqrt{n}|\theta_{1,n}| + M_n \ll \sqrt{\log n}$ with probability tending to 1. (For sequences with $\sqrt{\log\log n} \ll \sqrt{n}|\theta_{1,n}| \ll \sqrt{\log n}$, the argument can be refined to an almost sure statement, as $\sqrt{n}\bar{X}$ will be in the interval as above almost surely as $n \to \infty$ for M_n of the order $\sqrt{\log\log n}$, by the law of the iterated logarithm.)

This phenomenon was apparently first noted by Jeffreys and later called a *paradox* by Lindley. Since then it has been subject of much debate. That the Bayes factor may lead to the opposite conclusion as simple frequentist statistical reasoning seems not necessarily paradoxical, but it is certainly embarrassing to the Bayesian. One conclusion could be that we should be careful in giving equal prior weights to models of different complexities, such as the zero-dimensional null model and the one-dimensional alternative in the present case. The data may not in all cases correct the prior bias to the more precise smaller model.

Example 1.16 (Improper priors)**.** For an improper prior there is usually no natural norming constant, in that $c\Pi$ is just as reasonable as Π, for any $c > 0$. For the posterior distribution

this makes no difference as the multiplicative constant cancels from Bayes's formula. However, in the posterior probability $\Pi(\Theta_i|\, x)$ when there are multiple models Θ_i, or the Bayes factor between two models, the norming constants of the various models cancel only if they are all identical. If the models or their priors are not comparable, then there may not be good arguments for the relative sizes of the norming constants. It is therefore a common opinion that improper priors should not be used in the case of multiple models, or at best are admissible for parameters that are common to the models and a priori independent of the other parameters, so that the norming constants of their priors cancel.

1.6 Decision theory

Suppose that one is required to choose an *action* or *decision* t from a set $\mathfrak{D}$ leading to a *loss* $\ell(\theta, t)$ if $\theta \in \Theta$ is the true state of the world, for a given function $\ell\colon \Theta \times \mathfrak{D} \to [0, \infty)$. One is provided with data x that is generated from the measure P_θ, and is allowed to randomize the action by using a number u that is independently generated from the uniform distribution on $[0, 1]$, and choose an action $T(x, u)$, where $T\colon \mathfrak{X} \times [0, 1] \to \mathfrak{D}$ is a given map.[1] The *risk function* of the *procedure* (or *randomized decision function*) T is defined as

$$\theta \mapsto \mathrm{E}_\theta \ell(\theta, T) = \int \int_0^1 \ell(\theta, T(x, u))\, du\, dP_\theta(x),$$

and the *Bayes risk* of T for prior Π is the average risk

$$\int \mathrm{E}_\theta \ell(\theta, T)\, d\Pi(\theta) = \mathrm{E}\ell(\vartheta, T(X, U)).$$

To make these expressions well defined, we assume given a σ-field on $\mathfrak{D}$, and require that ℓ and T are measurable maps.

The purpose is to find a procedure with small Bayes risk.

Example 1.17 (Estimation with quadratic loss)**.** A possible loss function for $\Theta \subset \mathbb{R}^d$ and $\mathfrak{D} = \mathbb{R}^d$ is *square loss* $\ell(\theta, t) = \|\theta - t\|^2$. The risk function of the randomized estimator T is then its *mean square error* $\mathrm{E}_\theta \|T - \theta\|^2$, as a function of the parameter θ.

Given a procedure T, the procedure $\int_0^1 T(x, u)\, du$ can be seen to have smaller risk, by Jensen's inequality, and hence randomization is not useful in this case. This is true more generally for convex loss functions.

The risk function will typically be minimized by different procedures for different θ, and then there will not be a uniformly optimal procedure. However, the Bayes risk reduces the risk function to a number and there will usually be a procedure that minimizes the Bayes risk. It suffices that the infimum of the Bayes risks over all procedures is assumed. The minimizer is called the *Bayes procedure*.

[1] The map $(x, B) \mapsto \Pr(T(x, U) \in B)$ is a Markov kernel. Instead of maps T, one could also consider Markov kernels from the observational space into the decision space as the possible "randomised decisions". For a Polish decision space this is no more general in the sense that any Markov kernel $(x, B) \mapsto Q(B|\, x)$ arises from some T as considered here. We prefer the more concrete T notation, as it hides the randomisation and simplifies formulas. See Lemma 1.25.

The following theorem shows that the Bayes procedure can be found by minimizing the *posterior risk*, given by, for $\Pi(\cdot|\, x)$ the posterior distribution of θ,

$$\mathrm{E}(\ell(\vartheta, T)|\, X = x) = \int\int_0^1 \ell(\theta, T(x,u))\, du\, d\Pi(\theta|\, x).$$

Theorem 1.18. *Assume that there exists a procedure T_0 such that $T_0(x, U)$ minimizes the posterior risk at x over all procedures T, for every x. Then T_0 is a Bayes procedure.*

Proof The Bayes risk can be written in the form $\mathrm{E}\ell(\vartheta, T) = \int L(x,T)\, dP(x)$, for $L(x,T)$ the posterior risk, and P the marginal distribution of X. Minimization of the integral on the right is achieved by minimizing the integrand $L(x,T)$, for every x separately. □

Example 1.19 (Estimation with quadratic loss)**.** For the quadratic loss function $\ell(\theta, t) = \|\theta - t\|^2$ on $\Theta \subseteq \mathbb{R}^d$, the posterior risk of a nonrandomized procedure is

$$\int \big\|T(x) - \theta\big\|^2\, d\Pi(\theta|\, x).$$

Minimization over all procedures T leads to the posterior mean $T(x) = \int \theta\, d\Pi(\theta|\, x)$. This follows because $\mu \mapsto \mathrm{E}\|Y - \mu\|^2$ is minimized by $\mu = \mathrm{E}Y$, for any random vector Y; we apply this to Y distributed according to $\Pi(\cdot|\, x)$.

Example 1.20 (Binomial distribution)**.** Suppose the observation X is $\mathrm{bin}(n, \theta)$-distributed, with n known and $0 \le \theta \le 1$ the success probability. A convenient class of prior distributions on $[0, 1]$ is the class of *Beta-distributions*, given by the densities, for $\alpha > 0$ and $\beta > 0$,

$$\pi(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

The (omitted) norming constant for this density is the *beta-function* $B(\alpha,\beta) = \int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}\, d\theta$. By Bayes's rule the posterior density is given by

$$\pi(\theta|\, x) \propto \binom{n}{x}\theta^x(1-\theta)^{n-x}\pi(\theta) \propto \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}.$$

In other words: the a-posteriori distribution of θ is a Beta-distribution with parameters $x + \alpha$ and $n - x + \beta$. The posterior mean is

$$T_{\alpha,\beta}(x) = \int_0^1 \theta\, \pi(\theta|\, x)\, d\theta = \frac{B(x+\alpha+1, n-x+\beta)}{B(x+\alpha, n-x+\beta)} = \frac{x+\alpha}{n+\alpha+\beta}.$$

Here we use that $B(\alpha,\beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$ so that $B(\alpha+1,\beta) = \alpha/(\alpha+\beta)\, B(\alpha,\beta)$. For every parameter (α,β) we find a different estimator. The "natural estimator" X/n is not a Bayes estimator, but is obtained as a limit as $(\alpha,\beta) \to (0,0)$.

Which of the estimators is best? None of the Bayes estimators is "wrong", because each one is optimal in terms of its own Bayes risk as criterion. The mean square errors of the estimators can be computed as

$$\begin{aligned}\mathrm{E}_\theta(T-\theta)^2 &= \mathrm{E}_\theta\Big(\frac{X+\alpha}{n+\alpha+\beta} - \theta\Big)^2 = \frac{\mathrm{var}_\theta\, X}{(n+\alpha+\beta)^2} + \Big(\frac{\mathrm{E}_\theta X + \alpha}{n+\alpha+\beta} - \theta\Big)^2 \\ &= \frac{\theta^2((\alpha+\beta)^2 - n) + \theta(n - 2\alpha(\alpha+\beta)) + \alpha^2}{(n+\alpha+\beta)^2}.\end{aligned}$$

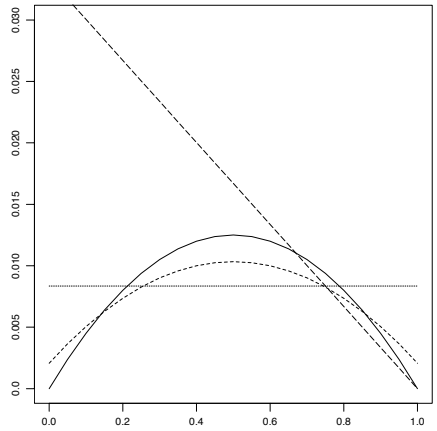

Figure 1.3 Mean square error of the Bayes estimators $T_{\alpha,\beta}$ with $\alpha = \beta = \frac{1}{2}\sqrt{n}$ (constant), $\alpha = \beta = 0$ (curved, solid), $\alpha = \sqrt{n}, \beta = 0$ (dashed, linear), $\alpha = \beta = 1$ (dashed).

Figure 1.3 shows the graphs of some of these functions. Every Bayes estimator is best for some region of θ-values, but worse in other regions, and there is no absolutely best estimator.

Minimizing the Bayes risk entails minimizing a weighted average of the risk function, with the weights determined by the prior distribution. Different prior distributions will give different solutions. The *complete class theorem* shows that all procedures of interest are obtained in this manner. The *minimax theorem* shows that the minimax risk is the maximum over all Bayes risks. Unfortunately, neither of the two theorems is true in exactly this simple form, but both require significant regularity conditions and/or a closure operation. The following are versions of these results that are compromises between generality and simplicity.

A randomized estimator T is called *admissible* if there is no randomized estimator T_1 with $\mathrm{E}_\theta \ell(\theta, T_1) \leq \mathrm{E}_\theta \ell(\theta, T)$, for every $\theta \in \Theta$ with strict inequality for at least one value of θ. A loss function $\ell\colon \Theta \times \mathfrak{D}$ is called *subcompact* if for every θ and $c \in \mathbb{R}$ the set $\{t\colon \ell(\theta, t) \leq c\}$ is either compact or the full space $\mathfrak{D}$.

Theorem 1.21 (Complete class theorem)**.** *Suppose that the measures P_θ permit densities relative to a σ-finite measure on $(\mathfrak{X}, \mathscr{X})$, assume that the decision space $(\mathfrak{D}, \mathscr{D})$ is Polish with its Borel σ-field, and let the loss function ℓ be subcompact.*

(i) *If Θ is finite, then every admissible procedure is Bayes for some prior on Θ.*

(ii) *If $(\Theta, \mathscr{B})$ is a compact metric space with its Borel σ-field and every risk function*

$\mathrm{E}_\theta \ell(\theta, T)$ *is continuous, then every admissible procedure is Bayes with respect to some prior on* $(\Theta, \mathscr{B})$.

(iii) *For any* Θ *the risk function of any admissible procedure is the pointwise limit (on* Θ*) of risk functions of a net of Bayes procedures for finitely discrete priors on* Θ.

Theorem 1.22 (Minimax theorem)**.** *Under the conditions of the preceding theorem*

$$\sup_{\Pi} \inf_{T} \int \mathrm{E}_\theta \ell(\theta, T)\, d\Pi(\theta) = \inf_{T} \sup_{\theta \in \Theta} \mathrm{E}_\theta \ell(\theta, T),$$

where the first supremum is taken over all finitely discrete measures Π *on* Θ *and the infima are taken over all randomized estimators* T.

Proofs of these results can be found in abstract form in Le Cam (1986) or Strasser (1985), and in different forms in other books.

1.7 Complements

Theorem 1.23 (De Finetti's theorem)**.** *If* $X_1, X_2, \ldots$ *are random variables with values in a Polish measurable space* $(\mathfrak{X}, \mathscr{X})$ *such that* $(X_{\sigma_1}, \ldots, X_{\sigma_n})$ *and* $(X_1, \ldots, X_n)$ *are equal in distribution for every permutation* $(\sigma_1, \ldots, \sigma_n)$ *of* $\{1, \ldots, n\}$ *and every* n*, then there exists a Markov kernel from* $[0, 1]$ *into* $(\mathfrak{X}, \mathscr{X})$ *and a probability measure* Π *on* $[0, 1]$ *such that* $\Pr(X_1 \in A_1, X_2 \in A_2, \ldots)) = \int_0^1 \prod_{i=1}^\infty Q(A_i|\, u)\, d\Pi(u)$*, for every* $A_i \in \mathscr{X}$.

Lemma 1.24. *If* $(x, B) \mapsto Q(B|\, x)$ *is a Markov kernel from a measurable space* $(\mathfrak{X}, \mathscr{X})$ *into a measurable space* $(\mathfrak{Y}, \mathscr{Y})$ *and* P *is a probability measure on* $(\mathfrak{X}, \mathscr{X})$*, then there exists a unique probability measure* R *on* $(\mathfrak{X} \times \mathfrak{Y}, \mathscr{X} \times \mathscr{Y})$ *such that* $R(A \times B) = \int_A Q(B|\, x)\, dP(x)$.

Lemma 1.25. *For any Markov kernel* $(x, B) \mapsto Q(B|\, x)$ *from a measurable space* $(\mathfrak{X}, \mathscr{X})$ *into a Polish space* $\mathfrak{D}$ *there exists a measurable map* $T\colon \mathfrak{X} \times [0, 1] \to \mathfrak{D}$ *such that* $Q(B|\, x) = \Pr(T(x, U) \in B)$*, for every* $x \in \mathfrak{X}$*, where* U *is a uniform variable.*

Exercises

1.1 Consider the decision problem with decision space $\{0, 1\}$ and loss function $\ell(\theta, i) = c_i 1_{\theta \notin \Theta_i}$, for given constants $c_i > 0$ and a given partition $\Theta = \Theta_0 \cup \Theta_1$. Determine the Bayes procedure relative to a given prior Π.

1.2 Suppose that $V = V(X)$ is a sufficient statistic for θ. Show that the posterior distribution of θ given X is the same as the posterior distribution of θ given V. [As a definition of sufficiency you may use that there exists a regular version of the conditional distribution of X given V that does not depend on θ.]

1.3 Suppose the vector X is normally distributed with unknown mean $\theta \in \mathbb{R}^k$ and known covariance matrix Σ. Show that the posterior distribution for θ relative to the prior $\Pi = N_k(0, \Lambda)$ is the normal distribution with mean $(\Sigma^{-1} + \Lambda^{-1})^{-1}\Sigma^{-1}X$ and covariance matrix $(\Sigma^{-1} + \Lambda^{-1})^{-1}$.

1.4 Suppose the data is a sample $X_1, \ldots, X_n$ from the Poisson distribution with mean θ, and choose the prior to have density $\pi(\theta) = e^{-\theta}$. Find the posterior distribution. Find the posterior mean and the posterior variance.

1.5 Given an improper prior Π let $\Theta_1 \subset \Theta_2 \subset \cdots$ be a sequence of subsets of the parameter set with $\Theta = \cup_i \Theta_i$ and such that $\Pi(\Theta_i) < \infty$, for every i. Assume that $\int p_\theta(x)\, d\Pi(\theta) < \infty$, for a given statistical model with densities p_θ. Show that the posterior distributions $\Pi_i(\cdot|\, x)$ for the

proper priors $\Pi(\cdot \cap \Theta_i)/\Pi(\Theta_i)$ converge to the posterior distribution for the improper prior Π in the sense of Kullback-Leibler divergence: $\int \log(\pi_i(\theta|\, x)/\pi(\theta|\, x))\, d\Pi_i(\theta|\, x) \to 0$, as $i \to \infty$.

2

Parametric Models

A statistical model is called *parametric* if it is finite-dimensional, as opposed to the "non-parametric" infinite-dimensional models in a later chapter. This definition is empty without a further specification of "dimension", but there is no need to be very specific. The examples in this chapter concern models with parameter set Θ a subset of $\mathbb{R}^d$ and such that the dependence $\theta \mapsto P_\theta$ is continuous in some way.

Many of the well-known statistical models are parametric. In this chapter we give examples of standard prior specifications. We also present the Bernstein-von Mises theorem on the asymptotic behavior of posterior distributions for *differentiable* parametric models, and discuss its consequences.

2.1 Conjugate priors

Despite its simplicity Bayes's formula (1.4) may be hard to apply, because numerical evaluations, such as computing the posterior mean or a credible set, require integration over the parameter space. One solution are simulation schemes, as discussed in Chapter 3. Another way out is to use priors that make the computations easy.

Definition 2.1 (Conjugacy)**.** A parametrized family $(\Pi_\alpha : \alpha \in A)$ of priors is called *conjugate* with respect to a statistical model if the posterior distribution relative to a member of the family is again a member of the family.

Example 2.2 (Dirichlet-Multinomial)**.** In Example 1.20 the family of Beta distributions was seen to be conjugate with respect to the binomial likelihood. A generalization to higher dimensions is the family of Dirichlet distributions relative to the multinomial distribution.

A random vector $\vartheta = (\vartheta_1, \ldots, \vartheta_k)$ is said to possess a *Dirichlet distribution* with parameters $k \in \mathbb{N}$ and $\alpha_1, \ldots, \alpha_k > 0$ if $\vartheta_1 + \cdots + \vartheta_k = 1$ and the vector $(\vartheta_1, \ldots, \vartheta_{k-1})$ of its first $k-1$ coordinates has Lebesgue density, for $\theta_i > 0$ and $\sum_{i=1}^{k-1} \theta_i < 1$,

$$(\theta_1, \ldots, \theta_{k-1}) \mapsto \frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)}\, \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \cdots \theta_{k-1}^{\alpha_{k-1} - 1} (1 - \theta_1 - \cdots - \theta_{k-1})^{\alpha_k - 1}.$$

The full vector $(\vartheta_1, \ldots, \vartheta_k)$ is restricted to the $(k-1)$-dimensional unit simplex in $\mathbb{R}^k$ and has no density, but it is convenient to think of it as having a density proportional to $\theta_1^{\alpha_1 - 1} \cdots \theta_k^{\alpha_k - 1}$ with respect to the "Lebesgue measure on the unit simplex".

For $k = 2$ the vector $(\vartheta_1, \vartheta_2)$ is completely described by a single coordinate, where $\vartheta_1 \sim \mathrm{Be}(\alpha_1, \alpha_2)$ and $\vartheta_2 = 1 - \vartheta_1 \sim \mathrm{Be}(\alpha_2, \alpha_1)$. Thus the Dirichlet distribution is a multivariate generalization of the Beta distribution. The $\mathrm{Dir}(k; 1, \ldots, 1)$-distribution is the uniform

distribution on the unit simplex. See Section 2.8.1 for more information on the Dirichlet distribution.

The Dirichlet family is conjugate for data X following a multinomial distribution with success parameter θ. Indeed if $X = (X_1, \dots, X_k)$ has density

$$p_\theta(x) = \binom{n}{x} \theta_1^{x_1} \theta_2^{x_2} \cdots \theta_k^{x_k}, \qquad x_i \in \mathbb{Z}^+, \sum_i x_i = n,$$

and the prior is chosen to be $\mathrm{Dir}(k; \alpha)$, then the posterior density satisfies

$$\pi(\theta|\, x) \propto \theta_1^{\alpha_1 + x_1 - 1} \theta_2^{\alpha_2 + x_2 - 1} \cdots \theta_k^{\alpha_k + x_k - 1}.$$

Thus the posterior distribution is $\mathrm{Dir}(k; \alpha + x)$.

From the properties of the Dirichlet distribution it follows that the posterior mean is equal to $(\alpha + X)/(n + |\alpha|)$, for $|\alpha| = \sum_i \alpha_i$. For $\alpha \to 0$ this tends to the maximum likelihood estimator X/n.

Example 2.3 (Normal/Gamma-Gaussian)**.** The likelihood of an i.i.d. sample $X_1, \dots, X_n$ from the $N(\mu, \sigma^2)$-distribution is proportional to

$$(\mu, \sigma) \mapsto \frac{1}{\sigma^n} e^{-\sum_{i=1}^n (X_i - \mu)^2/(2\sigma^2)} = \frac{1}{\sigma^n} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n X_i^2 + \frac{\mu}{\sigma^2} \sum_{i=1}^n X_i - \frac{n\mu^2}{2\sigma^2}}.$$

It follows that a conjugate prior density is of the form, for given constants $\gamma_1, \gamma_2, \gamma_3, \gamma_4$,

$$\pi(\mu, \sigma) \propto \frac{1}{\sigma^{\gamma_1}} e^{-\frac{\gamma_2}{\sigma^2} + \frac{\gamma_3 \mu}{\sigma^2} - \frac{\gamma_4 \mu^2}{\sigma^2}}.$$

This depends on μ through a parabola in the exponent, in which we recognize a Gaussian density with scale σ and mean and variance determined by the pair (γ_3, γ_4). For μ given σ distributed according to $N(\nu, \sigma^2 \lambda)$, we have

$$\pi(\mu|\, \sigma) = \frac{1}{\sigma \sqrt{2\pi\lambda}} e^{-(\mu - \nu)^2/(2\sigma^2 \lambda)}.$$

If we choose (ν, λ) so that the parabola in the exponent matches the parabola in the exponent of $\pi(\mu, \sigma)$, then we see that, for some γ_5, γ_6,

$$\pi(\sigma) = \frac{\pi(\mu, \sigma)}{\pi(\mu|\, \sigma)} \propto \frac{1}{\sigma^{\gamma_5}} e^{-\frac{\gamma_6}{\sigma^2}}.$$

It can be verified that this is equivalent to $1/\sigma^2$ following a Gamma distribution with certain parameters (α, β).

Thus the conjugate family for a normal sample with unknown mean and variance can be conveniently described by the hierarchy $1/\sigma^2 \sim \Gamma(\alpha, \beta)$ and $\mu|\, \sigma^2 \sim N(\nu, \sigma^2 \lambda)$, for hyper parameters $\alpha > 0, \beta > 0, \nu \in \mathbb{R}, \lambda > 0$. The update formula for the posterior can be computed to be, for $x = (x_1, \dots, x_n)$,

$$\frac{1}{\sigma^2}|\, x \sim \Gamma\Big(\alpha + \frac{n}{2}, \beta + \frac{1}{2} \sum_{i=1}^n (x_i - \bar{x})^2 + \frac{n(\bar{x} - \nu)^2}{2n\lambda + 2}\Big),$$

$$\mu|\, x, \sigma \sim N\Big(\frac{n\lambda(\bar{x} - \nu)}{n\lambda + 1} + \nu, \frac{\sigma^2 \lambda}{n\lambda + 1}\Big).$$

Choosing the prior for μ proportional to σ seems not easily defendable from a subjectivist Bayesian point of view, but is often considered reasonable as σ measures the noise level of the data. The posterior means are given by

$$\begin{aligned}\mathrm{E}\Big(\frac{1}{\sigma^2}\,|\,x\Big) &= \frac{\alpha + n/2}{\beta + \sum_{i=1}^n (x_i - \bar{x})^2/2 + n(\bar{x}-\nu)^2/(2n\lambda + 2)},\\ \mathrm{E}(\mu|\,x) &= \frac{n\lambda(\bar{x}-\nu)}{n\lambda+1} + \nu.\end{aligned}$$

For $\nu = 0$ and $\lambda \to \infty$ this reduces to the maximum likelihood estimators of $1/\sigma^2$ and μ.

Example 2.4 (Regression)**.** The likelihood for observing a variable Y with the $N_n(X\beta, \sigma^2 I)$-distribution, where X is a known and fixed $(n \times p)$-matrix, β a vector in $\mathbb{R}^p$ and $\sigma > 0$, is

$$p_{\beta,\sigma}(y) = \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\|y - X\beta\|^2/(2\sigma^2)} = \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\frac{\|y\|^2}{2\sigma^2} + \frac{\beta^T}{\sigma^2} X^T y - \frac{\beta^T X^T X \beta}{2\sigma^2}}.$$

This is similar in structure as in Example 2.3, except that β is a vector and enters linearly and through a quadratic form in the exponent. Similar reasoning shows that a conjugate prior is given by $1/\sigma^2 \sim \Gamma(a, b)$ and $\beta|\,\sigma \sim N_p(\nu, \sigma^2\Lambda)$, for hyper parameters $a > 0, b > 0$, $\nu \in \mathbb{R}^p$ and a positive-definite matrix Λ. The posterior distribution can be shown to take the form

$$\begin{aligned}\frac{1}{\sigma^2}\,|\,Y &\sim \Gamma\Big(a + \frac{n}{2}, b + \tfrac{1}{2}\|Y - X\nu\|^2 - \tfrac{1}{2}(Y - X\nu)^T X (X^T X + \Lambda^{-1})^{-1} X^T (Y - X\nu)\Big),\\ \beta|\,Y, \sigma &\sim N\Big((X^T X + \Lambda^{-1})^{-1} X^T (Y - X\nu) + \nu, \sigma^2 (X^T X + \Lambda^{-1})^{-1}\Big).\end{aligned}$$

For $a, b \to 0$, $\nu = 0$, and $\Lambda^{-1} \to 0$, the posterior means of β and $1/\sigma^2$ tend to the maximum likelihood (or least squares) estimators $\hat{\beta} = (X^T X)^{-1} X^T Y$ and n/SS_{res}, where $SS_{res} = \|Y - X\hat{\beta}\|^2$ is the residual sum of squares. (Use the decomposition $\|Y\|^2 = \|Y - X\hat{\beta}\|^2 + \|X\hat{\beta}\|^2$, which follows from the fact that $X\hat{\beta}$ is the orthogonal projection of Y onto the column space of X.)

Since the regression matrix X is considered fixed, it can be used to construct the prior. The *Zellner g-prior*, or just short, g-prior, uses $\beta|\,\sigma, g \sim N(\nu, \sigma^2 g (X^T X)^{-1})$, for a hyperparameter $g > 0$. It is usually combined with the improper prior density $\pi(\sigma) \propto 1/\sigma$, for σ, which can be viewed as the limit of $1/\sigma^2 \sim \Gamma(a, b)$ as $(a, b) \to 0$. The posterior is then given by

$$\begin{aligned}\frac{1}{\sigma^2}\,|\,Y, g &\sim \Gamma\Big(\frac{n}{2}, \frac{SS_{res}}{2} + \frac{(\hat{\beta} - \nu) X^T X (\hat{\beta} - \nu)}{2(g+1)}\Big),\\ \beta|\,Y, \sigma, g &\sim N_p\Big(\frac{\nu + \hat{\beta}}{g+1}, \sigma^2 \frac{g}{g+1} (X^T X)^{-1}\Big).\end{aligned}$$

The hyper parameter g is fixed in these calculations, as giving it a prior destroys conjugacy. It may be estimated by the empirical Bayes method, by maximizing the marginal likelihood. Since the marginal likelihood has an interpretation only for a proper prior, we may compute

this first for the situation that $1/\sigma^2 \sim \Gamma(a,b)$ as, setting $\nu = 0$ for simplicity,

$$p(Y|\, g) = \int_0^\infty \int \frac{u^{n/2} e^{-\frac{1}{2}u\|y-X\beta\|^2}}{(2\pi)^{n/2}} \frac{u^{a-1} b^a e^{-bu}}{\Gamma(a)} \frac{u^{p/2} e^{-\frac{1}{2}u\beta^T g^{-1} X^T X\beta}}{(2\pi)^{p/2} (\det g(X^T X)^{-1})^{1/2}}\, d\beta\, du$$
$$= \frac{b^a \Gamma(n/2+a)}{\pi^{n/2}\Gamma(a)} \frac{(1+g)^{n/2-p/2-a}}{(b(1+g) + g\|Y - X\hat\beta\|^2 + \|Y\|^2)^{n/2-a}}.$$

As $b \to 0$ this tends to a limit, but as $a \to 0$ the factor $\Gamma(a)$ tends to infinity and the expression approaches 0. This is usually solved by dropping the factor $b^a/\Gamma(a)$, and setting the other appearances of a and b to zero.

The main motivation for Zellner's prior appears to be that it leads to simple formulas, which are reminiscent of least squares theory. However, the same computations are possible for a general covariance matrix Λ.

The regression matrix X usually contains a column of 1s, to model a general intercept, and then the other columns are taken orthogonal to this column. A variation is to model the intercept coefficient by the improper Lebesgue prior, and equip only the other coefficients with a Zellner g-prior. This leads to slightly different formulas.

See Liang et al. (2008) for further discussion.

Example 2.5 (Exponential families)**.** A statistical model follows an *exponential family* if it is given by densities relative to some σ-finite dominating measure of the form, for given functions $c, Q_1, \ldots, Q_k$ on Θ and $h, \tau_1, \ldots, \tau_k$ on $\mathfrak{X}$ and given $k \in \mathbb{N}$,

$$p_\theta(x) = c(\theta) h(x) e^{\sum_{j=1}^k Q_j(\theta)\tau_j(x)}.$$

Then a conjugate family of priors is given by the densities

$$\pi_{\alpha,\beta}(\theta) \propto c(\theta)^\alpha e^{\sum_{j=1}^k Q_j(\theta)\beta_j},$$

indexed by parameters α and $\beta = (\beta_1, \ldots, \beta_k)$ such that the functions in the display are integrable with respect to the dominating measure. The posterior is obtained by updating the parameters from (α, β) to $(\alpha + 1, \beta + \tau(x))$ for $\tau(x) = (\tau_1(x), \ldots, \tau_k(x))$.

The distributions of a random sample $X_1, \ldots, X_n$ from p_θ form an exponential family with the same functions Q_i, but with the statistics $\sum_i \tau(x_i)$ replacing the statistics $\tau(x)$. The conjugacy is then retained with the update being from (α, β) to $(\alpha + n, \beta + \sum_i \tau(x_i))$.

Example 2.6 (Multivariate normal)**.** The likelihood for a sample $X_1, \ldots, X_n$ from the multivariate-normal distribution $N_d(\mu, \Sigma)$ is given by

$$p_{\mu,\Sigma}(x_1, \ldots, x_n) = \prod_{i=1}^n \frac{1}{(2\pi)^{d/2}(\det\Sigma)^{1/2}} e^{-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)}$$
$$\propto \frac{1}{(\det\Sigma)^{n/2}} e^{-\frac{1}{2}\mathrm{tr}(S_x\Sigma^{-1}) + \mu^T\Sigma^{-1}n\bar x - \frac{1}{2}n\mu^T\Sigma^{-1}\mu},$$

for $S_x = \sum_{i=1}^n x_i x_i^T$. A conjugate prior for (μ, Σ) is given by

$$\pi(\Sigma) \propto \frac{1}{(\det\Sigma)^{(a+d+1)/2}} e^{-\frac{1}{2}\mathrm{tr}(B\Sigma^{-1})},$$
$$\pi(\mu|\,\Sigma) = \frac{1}{\tau^d(\det\Sigma)^{1/2}} e^{-\frac{1}{2}(\mu-\nu)^T\Sigma^{-1}(\mu-\nu)/\tau^2}.$$

The first is the density of an *inverse Wishart* distribution with a degrees of freedom and precision matrix B. The second corresponds to a multivariate normal density.

2.2 Default priors

There may be compelling reasons to choose a particular prior, but often there are none. For the latter cases it is desirable to have a *default prior*, or *objective prior*, or *non-informative prior*, terms that roughly mean the same, but do not have an accepted rigorous definition.

One might think that a "uniform prior" or "flat prior" (also called "vague prior" if the parameter space is large), one that spreads its mass evenly over the parameter space, could serve this purpose. However, such priors are often criticized for lacking invariance. One argues that a parametrization $\theta \mapsto P_\theta$ is only a way to describe a statistical model $(P_\theta\colon \theta \in \Theta)$, which is a *set* of possible laws for the data, whence the parameter itself has no intrinsic meaning. As a consequence, "objectivity" cannot be independent of the statistical model.

To descirbe this more precisely, consider a bijection $\phi\colon H \to \Theta$ from some set H onto Θ. Then the model is equally well described as the set of laws $(P_\theta\colon \theta \in \Theta)$ or the set of laws $(P_{\phi(\eta)}\colon \eta \in H)$. Should "vagueness" then refer to a uniform prior for θ or for η? For a nonlinear reparametrization only one of the two priors can be uniform. Indeed, if θ has a density $\theta \mapsto \pi(\theta)$, then for a given diffeomorphism ϕ the variable $\eta = \phi^{-1}(\theta)$ will have density π_1 given by

$$\pi_1(\eta) = \pi(\phi(\eta))\,|\det \phi'(\eta)|.$$

Unless the Jacobian $|\det \phi'(\eta)|$ is constant, at most one of the two densities π or π_1 can be uniform.

This argument supposedly rules out the uniform measure as a prior for a fraction or success probability θ in a binomial experiment on parameter space $[0, 1]$, as for instance the odds ratio $\theta/(1-\theta)$ or its logarithm also give natural parametrizations.

2.2.1 Group invariance

Some statistical models are *invariant* under the action of a group: to every element g of a given group correspond measurable bijections $\bar{g}\colon \mathfrak{X} \to \mathfrak{X}$ and $\tilde{g}\colon \Theta \to \Theta$ on the sample space and the parameter space such that $\bar{g}(X)$ has law $P_{\tilde{g}(\theta)}$ if X has law P_θ.[1] It is reasonable that, given data $\bar{g}(X)$, a statistician would form a posterior law for $\tilde{g}(\theta)$ that is the transformation under the map $\tilde{g}$ of the posterior law for of θ given data X. In other words, it is reasonable that the prior is chosen such that

$$\Pr(\tilde{g}(\vartheta) \in B |\, \bar{g}(X)) = \Pr(\vartheta \in B |\, X).$$

In words: the conditional law of $\tilde{g}(\vartheta)$ given $\bar{g}(X)$ is the same as the conditional law of ϑ given X. We also have that $\bar{g}(X) |\, \tilde{g}(\vartheta) = \tilde{g}(\theta)$ is the same as $\bar{g}(X) |\, \vartheta = \theta \sim X |\, \vartheta = \tilde{g}(\theta)$, by the invariance assumption, whence $\bar{g}(X) |\, \tilde{g}(\vartheta) \sim X |\, \vartheta$. Thus it follows that $(\tilde{g}(\vartheta), \bar{g}(X))$ and (ϑ, X)

[1] So the pair $(\bar{g}, \tilde{g})$ refers to the element g of the group; a preciser notation would add the group element g as a subscript to $\bar{g}$ and $\tilde{g}$, and perhaps use different letters to denote the latter two maps, but the sloppy notation seems the clearer one. Even sloppier would be to use the same letter g for the group element and the two maps.

must be equal in distribution, by Lemma 2.36. Consequently, the marginal distributions of these pairs must be the same, and in particular $\tilde{g}(\vartheta)$ must be distributed as ϑ. We conclude that the prior must be invariant under the group action.

The invariance principle is known to lead to reasonable solutions, but also known to rule out many statistical procedures of interest. The preceding argument shows that it clearly rules out many priors.

Another point of criticism is that many statistical models are not invariant. The following two examples are important exceptions.

Example 2.7 (Location family)**.** The statistical model corresponding to observing a random sample from a density of the form $p_\theta(x) = f(x-\theta)$, for $\theta \in \mathbb{R}^d$ and a fixed probability density f on $\mathbb{R}^d$, is invariant under the translation group on $\mathbb{R}^d$. The group actions are given by the shift maps $(x_1, \ldots, x_n) \mapsto (x_1 + g, \ldots, x_n + g)$ and $\theta \mapsto \theta + g$. The only measures on the parameter set $\mathbb{R}^d$ that are invariant for this group action are multiples of Lebesgue measure, which are improper as priors.

Example 2.8 (Scale family)**.** The statistical model corresponding to observing a random sample from a density of the form $p_\theta(x) = f(x/\theta)/\theta$, for $\theta \in (0, \infty)$ and a fixed probability density f on $\mathbb{R}$, is invariant under the multiplicative group on $\mathbb{R}^+$. The group actions are given by the maps $(x_1, \ldots, x_n) \mapsto (x_1/g, \ldots, x_n/g)$ and $\theta \mapsto \theta/g$, for $g > 0$. The only measures on the parameter set $(0, \infty)$ that are invariant for this group action are the measures with Lebesgue measure $\pi(\theta) \propto 1/\theta$, which are improper as priors.

The groups in the two preceding examples are both locally compact topological groups. The invariant measures on such a group are called *Haar measure*. Haar measure is finite and gives rise to a proper prior (only) in the case that the group is compact. Although "invariant" under the group action, Haar priors are not invariant under nonlinear reparametrizations, as discussed previously. Real Bayesians seem not to like them.

Example 2.9 (Orthogonal group)**.** The statistical model corresponding to observing a random sample from a density of the form $p_\theta(x) = f(\theta x)$, for $\theta\colon \mathbb{R}^d \to \mathbb{R}^d$ an orthogonal linear map and a fixed probability density f on $\mathbb{R}^d$, is invariant under the orthogonal group on $\mathbb{R}^d$. The group actions are given by the maps $(x_1, \ldots, x_n) \mapsto (gx_1, \ldots, gx_n)$ and $\theta \mapsto g\theta$, for g an orthogonal matrix. The Haar measure on the orthogonal group is the only invariant prior and it can be normalised to a probability measure.

2.2.2 *Jeffreys priors*

Recall that for a statistical model given by densities p_θ that are smoothly parametrized by $\theta \in \mathbb{R}^d$ the gradient of the map $\theta \mapsto \log p_\theta(x)$ is known as the *score function* of the model, and its covariance matrix is the *Fisher information*:

$$I_\theta = \operatorname{Cov}_\theta(\dot{\ell}_\theta(X_1)), \qquad \dot{\ell}_\theta(x) = \frac{\partial}{\partial\theta} \log p_\theta(x).$$

Definition 2.10 (Jeffreys prior)**.** If the model $(P_\theta\colon \theta \in \Theta)$ with parameter set $\Theta \subset \mathbb{R}^d$ permits finite Fisher information I_θ, the *Jeffreys prior* is defined by the Lebesgue density

$$\pi(\theta) \propto \sqrt{\det I_\theta}.$$

Even if the Fisher information is finite, it may be that the function $\theta \mapsto \sqrt{\det I_\theta}$ is not integrable over the parameter set Θ. In that case, which is common, the Jeffreys prior is improper.

In the Cramér-Rao theorem, or asymptotic estimation theory (e.g. van der Vaart (1998), Chapter 8), the inverse of the Fisher information arises as the minimal variance or mean square error of (unbiased) estimators of θ from data generated from p_θ. If I_θ is large, then the parameter is easy to estimate. Thus the Jeffreys prior weighs the parameters in accordance to the intrinsic difficulty by which they can be estimated, putting more mass on parameter values that are easy.

To give intuition for why the Jeffreys prior is "non-informative" this is usually rephrased as follows. Because a large value of I_θ makes the data informative about θ, the Jeffreys prior prefers parameters that make the data informative. If the data is informative, then it will have strong influence when forming the posterior distribution and can easily overcome the prior. Thus the role of the prior is diminished: the prior has less influence and hence is uninformative.

Not every statistician is convinced by this reasoning.

Real Bayesians find another reason to dislike Jeffreys priors: they involve an expectation (in the definition of the Fisher information). Because this is an average over the sample space, this means that the Jeffreys prior "involves samples that were never observed", and does not satisfy the likelihood principle.[2] See Exercise 2.5 for a concrete illustration.

Because the Fisher information is additive over independent observations, the Jeffreys prior for a random sample from p_θ is the same as for a single observation. Another desirable property is invariance.

Lemma 2.11. *The Jeffreys prior is equivariant under reparametrization: the Jeffreys prior of $\eta = \phi^{-1}(\theta)$ for a smooth reparametrization is given by $\eta \mapsto \pi(\phi(\eta))|\det \phi'(\eta)|$ if π is the Jeffreys prior of θ.*

Proof By the chain rule the derivative of $\eta \mapsto \log p_{\phi(\eta)}$ is equal to $\dot\ell^T_{\phi(\eta)}\phi'(\eta)$, where $\phi'(\eta)$ is the derivative matrix of ϕ. Thus the score function of the densities $\eta \mapsto p_{\phi(\eta)}$ is equal to $\phi'(\eta)^T\dot\ell_{\phi(\eta)}$ and the Fisher information matrix of this model is equal to $\phi'(\eta)^T I_{\phi(\eta)}\phi'(\eta)$. (Note that the score function is defined as a $(d \times 1)$-vector, while the derivative matrix of the map $\theta \mapsto \log p_\theta$ from $\mathbb{R}^d$ into $\mathbb{R}$, which arises in the chain rule, is a $(1 \times d)$-matrix.) Thus the Jeffreys prior for η is proportional to the root of the determinant of this matrix, which is $(\det I_{\phi(\eta)})^{1/2}\,|\det \phi'(\eta)|$. This reduces to $\pi(\phi(\eta))\,|\det\phi'(\eta)|$ if π is the Jeffreys prior for θ. □

Example 2.12 (Binomial)**.** The score function and Fisher information for the binomial density $p_\theta(x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$ are equal to

$$\dot\ell_\theta(x) = \frac{x - n\theta}{\theta(1-\theta)}, \qquad I_\theta = \frac{n}{\theta(1-\theta)}.$$

Thus the Jeffreys prior has density proportional to $\theta^{-1/2}(1-\theta)^{-1/2}$. On the natural parameter

[2] The *likelihood principle* says that if the observed likelihoods $\theta \mapsto p_\theta(x)$ resulting from two possibly different models are proportional, then the statistical inferences should be the same. In particular, only the observed value of x of the data matters, not which other values could have been observed.

space $[0,1]$, this is the Beta distribution with parameters $(1/2, 1/2)$. Values of θ close to 0 or 1 receive much more prior weight than under the uniform measure.

Example 2.13 (Location-scale)**.** The score function and Fisher information matrix of the location-scale family $p_{\mu,\sigma}(x) = f((x-\mu)/\sigma)/\sigma$, for a fixed density f on $\mathbb{R}$, are equal to

$$\dot{\ell}_{\mu,\sigma}(x) = \begin{pmatrix} -(f'/f)((x-\mu)/\sigma)/\sigma \\ -1/\sigma - (f'/f)((x-\mu)/\sigma)(x-\mu)/\sigma^2 \end{pmatrix},$$
$$I_{\mu,\sigma} = \frac{1}{\sigma^2}\begin{pmatrix} \int (f'/f)^2(x)\, f(x)\, dx & \int x(f'/f)^2(x)\, f(x)\, dx \\ \int x(f'/f)^2(x)\, f(x)\, dx & \int (1 + x(f'/f)(x))^2\, f(x)\, dx \end{pmatrix}.$$

Thus the Jeffreys prior for (μ, σ) satisfies $\pi(\mu,\sigma) \propto 1/\sigma^2$. For the natural parameter space $\mathbb{R} \times (0,\infty)$, this prior is improper.

If we consider the scale parameter σ fixed, then the Jeffreys prior for μ is proportional to $1/\sigma$, which is constant in μ and hence still improper for the natural parameter space $\mathbb{R}$.

If we consider the location parameter μ fixed, then the Jeffreys prior for σ is also proportional to $1/\sigma$, and hence is still improper for the natural parameter space $(0,\infty)$ (at both ends of the parameter space). It is often thought to be embarrassing that this distribution is not the "marginal" of the Jeffreys prior for (μ,σ) jointly, which is proportional to $1/\sigma^2$. The "marginal Jeffreys prior" is often considered more natural.

Example 2.14 (Regression)**.** The score function and Fisher information matrix for the parameter (β,σ) of the linear regression model with one observation y from a $N_n(X\beta, \sigma^2 I)$-distribution are

$$\dot{\ell}_{\beta,\sigma}(y) = \begin{pmatrix} X^T(y - X\beta)/\sigma^2 \\ -n/\sigma + \|y - X\beta\|^2/\sigma^3 \end{pmatrix}, \qquad I_{\beta,\sigma} = \frac{1}{\sigma^2}\begin{pmatrix} X^T X & 0 \\ 0 & 2n \end{pmatrix}.$$

Therefore the joint Jeffreys prior is given by $\pi(\beta,\sigma) \propto 1/\sigma^2$.

*2.2.3 *Reference priors*

A reference prior is meant to be a prior such that the data is maximally informative, and hence the prior is uninformative. It can be viewed as an attempt to abstract the intuition behind the Jeffreys prior. Unfortunately, it is involved.[3]

Assume that the statistical model is given by densities p_θ relative to some σ-finite measure μ, so that the posterior distribution is given by Bayes's rule, and hence has a density relative to the prior. For ease of notation we write $\pi(\theta|\, x)$ and $\pi(\theta)$ for the posterior and prior densities of θ relative to some dominating measure ν. Recall that in the Bayesian model the marginal density of the observation X is given by $p_\pi(x) := \int p_\theta(x)\, \pi(\theta)\, d\theta$. In the following definition it is helpful to write this as p_π (and not as p as before) to stress its dependence on the prior.

The following definition concerns a prior to be used with data X (which will often be a vector), but the definition involves a random sample $X^{(n)} = (X_1, \ldots, X_n)$ of independent

[3] The name appears to originate from the advice to perform a data analysis multiple times, with different priors, including an uninformative prior as a "reference" to assess the influence of the other priors on the final conclusion (Bernardo (1979)).

replications X_i of X, where n will tend to infinity. For a given prior π, define

$$I_n(\pi) = \int\int \pi(\theta|\, x^{(n)}) \log \frac{\pi(\theta|\, x^{(n)})}{\pi(\theta)}\, d\nu(\theta)\, p_\pi(x^{(n)})\, d\mu^n(x^{(n)}).$$

The inner integral is the *Kullback-Leibler divergence* between the posterior density $\pi(\cdot|\, x^{(n)})$ and the prior density π given the sample $x^{(n)} = (x_1, \ldots, x_n)$ of observations from p_θ. This expression is always nonnegative, and zero if and only if these two densities are equal, and may be viewed a measure of "distance" between the two densities. (See Complements. It is called "divergence", because it is not a mathematical distance.) The outer integral computes the expectation of this divergence under the Bayesian marginal distribution of $x^{(n)}$.

For $K \subset \Theta$, let Π_K be the renormalized restriction of the prior to K, given by $\Pi(\cdot \cap K)/\Pi(K)$, and let π_K its density.

Definition 2.15 (Reference prior)**.** A density π is a *reference prior* for the statistical model with densities p_θ if

$$\lim_{n\to\infty} (I_n(\pi_K) - I_n(\tilde{\pi}_K)) \geq 0,$$

for every compact $K \subset \Theta$ and every other prior density $\tilde{\pi}$ on Θ.

For multivariate parameters the preceding definition is sometimes replaced by a definition that sequentially considers single coordinates conditionally on other coordinates. We omit a discussion. The definition may also allow improper priors, provided the corresponding posteriors are defined and the restrictions Π_K are proper. It is then also required that π and $\tilde{\pi}$ are *permissible*: the Kullback-Leibler divergence between the posteriors based on Π_K and Π must tend to zero as the sets K increase to Θ, and the same for $\tilde{\Pi}_K$ and $\tilde{\Pi}$ (see Berger et al. (2009)).

So a reference prior asymptotically maximizes the "distance" $I_n(\pi)$ between posterior and prior. The reasoning is that if the posterior is far from the prior, then the prior cannot have been "informative" and hence was "objective". That the distance is maximized "on the average under p_π" and in an asymptotic sense is important, but does not make the intuition easier to grasp.

For $(\theta, x^{(n)}) \mapsto \gamma_\pi(\theta, x^{(n)}) = \pi(\theta|\, x^{(n)}) p_\pi(x^{(n)})$ the joint density of $(\vartheta, X_1, \ldots, X_n)$ in the Bayesian setup, the expected Kullback-Leibler divergence can also be rewritten in the form

$$I_n(\pi) = \int\int \gamma_\pi(\theta, x^{(n)}) \log \frac{\gamma_\pi(\theta, x^{(n)})}{\pi(\theta) p_\pi(x^{(n)})}\, d\nu(\theta)\, d\mu^n(x^{(n)}).$$

This exhibits $I_n(\pi)$ as the Kullback-Leibler divergence between the joint distribution of $(\vartheta, X_1, \ldots, X_n)$ and the product of the marginal distributions of ϑ and $(X_1, \ldots, X_n)$. In information theory this is known as the *mutual information*, and interpreted as the "amount of information shared by ϑ and $(X_1, \ldots, X_n)$". The mutual information is a measure for dependence between ϑ and $(X_1, \ldots, X_n)$, large values meaning stronger dependence. Thus the reference prior makes these variables maximally dependent, asymptotically. This does not seem to yield a better intuition for the definition.

Still another motivation is that the reference prior is least favorable in a decision problem. Suppose that the statistician is to choose a probability density q that is closest to p_θ with

respect to the (random) loss $\ell(\theta, q) = \log(p_\theta/q)(x)$. The risk function for this problem is $\theta \mapsto \int p_\theta(x) \log(p_\theta/q)(x)\, d\mu(x)$, and the Bayes risk for the prior Π is

$$\int\int p_\theta(x) \log \frac{p_\theta(x)}{q(x)}\, d\mu(x)\, d\Pi(\theta) = -\int p_\pi(x) \log q(x)\, d\mu(x) + \int\int p_\theta(x) \log p_\theta(x)\, d\mu(x)\, d\Pi(\theta).$$

The second term on the right does not depend on q, so that minimization with respect to q entails minimizing the first term. The non-negativeness of the Kullback-Leibler divergence implies that the minimizer is the marginal density $q = p_\pi$. The minimal risk attained by this procedure is equal to

$$\int\int p_\theta(x) \log \frac{p_\theta(x)}{p_\pi(x)}\, d\mu(x)\, d\Pi(\theta) = \int\int \pi(\theta|\, x) \log \frac{\pi(\theta|\, x)}{\pi(\theta)}\, d\nu(\theta)\, p_\pi(x)\, d\mu(x),$$

in view of Bayes's rule. When applied to the product density of n copies of p_θ instead of p_θ, the right side is equal to $I_n(\pi)$, the quantity that the reference prior is defined to maximize, up to truncation to compact sets and taking the limit as $n \to \infty$. The reference prior can thus be seen as a distribution that makes the minimal Bayes risk of the decision problem as large as possible: it is nature's choice of prior that is *least favorable* to the statistician.

It turns out that for smoothly parameterized models the *Jeffreys prior is a reference prior*. We only give a heuristic derivation of this result, using an expansion of the likelihood from the next section. A precise statement, with regularity conditions, can be found in Clarke and Barron (1994), together with a proof that for finite parameter sets the uniform prior is a reference prior. The paper Berger et al. (2009) discusses still other examples, of non-smooth models.

Write $p_\theta^{(n)}$ and $p_\pi^{(n)}$ for the joint density of $(X_1, \ldots, X_n)$. Using the expansion (2.1), we find that

$$\begin{aligned} \frac{p_\pi^{(n)}}{p_\theta^{(n)}}(X^{(n)}) &= \int \prod_{i=1}^n \frac{p_{\theta+h/\sqrt{n}}}{p_\theta}(X_i)\, \pi\Big(\theta + \frac{h}{\sqrt{n}}\Big)\, \frac{dh}{n^{d/2}} \\ &\approx \int e^{h^T \Delta_{n,\theta} - \frac{1}{2} h^T I_\theta h} \pi(\theta)\, \frac{dh}{n^{d/2}} = \frac{\pi(\theta)}{\sqrt{\det I_\theta}} \frac{(2\pi)^{d/2}}{n^{d/2}} e^{\frac{1}{2}\Delta_{n,\theta}^T I_\theta^{-1} \Delta_{n,\theta}}. \end{aligned}$$

See Theorem 2.22 below for precise conditions for this approximation. It follows that

$$\mathrm{E}_\theta \log \frac{p_\theta^{(n)}}{p_\pi^{(n)}}(X^{(n)}) \approx \log \frac{\sqrt{\det I_\theta}}{\pi(\theta)} + \frac{d}{2} \log \frac{n}{2\pi} - \tfrac{1}{2} \mathrm{E}_\theta \Delta_{n,\theta}^T I_\theta^{-1} \Delta_{n,\theta}.$$

Because the variables $\Delta_{n,\theta}$ are under θ asymptotically normally distributed with mean zero and covariance matrix I_θ, the quadratic form $\Delta_{n,\theta}^T I_\theta^{-1} \Delta_{n,\theta}$ is asymptotically chisquare distributed with d degrees of freedom, and hence its expectation ought to be close to d. Replacing the last term of the preceding display by $-d/2$ and taking the integral with respect to Π, yields

$$I_n(\pi) \approx \int \log \frac{\sqrt{\det I_\theta}}{\pi(\theta)}\, d\Pi(\theta) + \frac{d}{2} \log \frac{n}{2\pi e}.$$

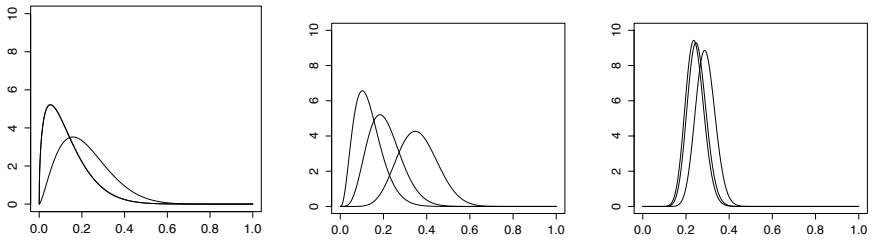

Figure 2.1 Posterior distribution for data from the binomial distribution with $n = 10$, $n = 25$, and $n = 100$ (left to right) relative to a Beta prior with parameters 0.5 and 1. Each panel shows three realizations of the posterior density, based on data generated according to success probability $1/3$.

Non-negativeness of the Kullback-Leibler divergence shows that this is maximized for $\pi(\theta) \propto \sqrt{\det I_\theta}$.

If the model were reparametrized as $\theta = \phi(\eta)$ for a measurable bijection $\phi: H \to \Theta$, then we would form the risk measure, say $I_{n,1}(\pi_1)$, based on priors Π_1 on H and likelihood $p_{\phi(\eta)}(x)$. Then $I_{n,1}(\pi_1) = I_n(\pi)$ if $\Pi = \Pi_1 \circ \phi^{-1}$ is the law of $\phi(\eta)$ if $\eta \sim \Pi_1$. Maximizing $I_{n,1}$ is equivalent to maximizing I_n, where the maximizers, the reference priors for η and θ, are related by the relation $\Pi = \Pi_1 \circ \phi^{-1}$. This means that the reference prior is invariant, at least under transformations that map compact sets to compact sets.

2.3 Bernstein-von Mises theorem

The Bernstein-von Mises theorem is the main theoretical result on Bayesian methods for parametric models. It asserts that the posterior distribution can be approximated by a normal distribution as the sample size increases. This normal distribution has a fixed covariance, but a random location, reflecting that a posterior distribution is a random probability distribution, depending on the data. Figure 2.1 illustrates the theorem in the case of the binomial distribution. Each panel shows multiple realizations of the posterior density, where the sample size increases from left to right. The shape of the density becomes more symmetric and normal when the sample size increases, while the scale decreases in size. For fixed sample size (in a single panel) the locations vary among the realizations, but the scale is roughly constant. We shall see below that the scale decreases to zero (at the speed $1/\sqrt{n}$), so that in the limit as $n \to \infty$ the posterior distributions contract to a Dirac measure; the normal approximation works after rescaling the parameter. This is reminiscent of the behavior of averages of samples of random variables, which tend to a constant (the mean) by the law of large numbers, and tend to a normal distribution after rescaling by the central limit theorem. By analogy the Bernstein-von Mises theorem is sometimes called the "Bayesian central limit theorem".

We first give a heuristic derivation of the theorem. Suppose that the data are a random sample $X_1, \ldots, X_n$ from a density p_θ that depends smoothly on the parameter $\theta \in \Theta$ and

allows a Taylor expansion of the form, as $h \to 0$,

$$\log \frac{p_{\theta+h}}{p_\theta}(x) = h^T \dot{\ell}_\theta(x) + \tfrac{1}{2} h^T \ddot{\ell}_\theta(x) h + \cdots .$$

Here $\dot{\ell}_\theta(x)$ is the gradient of the log likelihood $\theta \mapsto \log p_\theta(x)$, and is known as the *score function* of the model. Replacing h by $h/\sqrt{n}$ and summing over the observations, we find

$$\log \prod_{i=1}^{n} \frac{p_{\theta+h/\sqrt{n}}}{p_\theta}(X_i) = h^T \Delta_{n,\theta} - \tfrac{1}{2} h^T I_{n,\theta} h + \cdots , \tag{2.1}$$

where

$$\Delta_{n,\theta} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{\ell}_\theta(X_i), \qquad I_{n,\theta} = -\frac{1}{n} \sum_{i=1}^{n} \ddot{\ell}_\theta(X_i). \tag{2.2}$$

It can be shown that the score function has mean zero (see Complements), so that the sequence $\Delta_{n,\theta}$ will tend (if θ is the true parameter) in distribution to a Gaussian variable with mean zero and covariance the Fisher information $I_\theta = \mathrm{Cov}_\theta(\dot{\ell}_\theta(X_1))$, by the central limit theorem. Furthermore, by the law of large numbers the sequence $I_{n,\theta}$ will converge in probability to its expectation. It can be shown (see Complements) that this expectation is equal to the Fisher information I_θ, which explains the notation $I_{n,\theta}$. The remainder term in the expansion (the dots) is a sum over n terms of the order $(h/\sqrt{n})^3$ and should therefore be negligible.

We compare this to a statistical model with a single observation X from the $N(h, I_\theta^{-1})$-distribution, where h is an unknown parameter and the precision matrix I_θ is considered fixed. If $dN(\mu, \Sigma)$ denotes the density of the normal distribution with mean μ and covariance Σ, then by direct computation

$$\log \frac{dN(h, I_\theta^{-1})}{dN(0, I_\theta^{-1})}(X) = h^T I_\theta X - \tfrac{1}{2} h^T I_\theta h.$$

The right side is similar in form to the right side of (2.1), where $\Delta_\theta := I_\theta X$ has taken the place of $\Delta_{n,\theta}$, and I_θ the place of $I_{n,\theta}$. In the model for $X_1, \ldots, X_n$ the parameter $h = 0$ corresponds to the parameter θ, and $\Delta_{n,\theta}$ was seen to be approximately $N(0, I_\theta)$-distributed, which is also the distribution of Δ_θ under true parameter $h = 0$. Thus apart from having similar form the two expansions also agree in a distributional sense. One therefore says that the *sequence of statistical models* $(P^n_{\theta+h/\sqrt{n}} : h \in \mathbb{R}^d)$ *converges to the models* $(N(h, I_\theta^{-1}) : h \in \mathbb{R}^d)$, as $n \to \infty$.

We do not develop the theory of convergence of models here, but use the correspondence for a heuristic derivation of the Bernstein-von Mises theorem. The posterior density of ϑ given $X_1, \ldots, X_n$ relative to a prior density π is given by

$$\pi(\theta \,|\, X_1, \ldots, X_n) = \frac{\prod_{i=1}^n p_\theta(X_i)\, \pi(\theta)}{\int \prod_{i=1}^n p_\theta(X_i)\, \pi(\theta)\, d\theta}.$$

Typically, the distribution corresponding to this measure will shrink to the true parameter value θ_0, as $n \to \infty$. To obtain a more interesting limit we rescale the parameter, and study the sequence of posterior distributions of $\sqrt{n}(\vartheta - \theta_0)$, whose densities are given by

$$\pi_n(h \,|\, X_1, \ldots, X_n) = \frac{\prod_{i=1}^n p_{\theta_0+h/\sqrt{n}}(X_i)\, \pi(\theta_0 + h/\sqrt{n})}{\int \prod_{i=1}^n p_{\theta_0+h/\sqrt{n}}(X_i)\, \pi(\theta_0 + h/\sqrt{n})\, dh}.$$

If the prior density π is continuous, then $\pi(\theta_0 + h/\sqrt{n}) \approx \pi(\theta_0)$, for large n, and π will cancel from the right side. We can also divide both numerator and denominator by the "constant" $\prod_{i=1}^n p_{\theta_0}(X_i)$ to find that the posterior density is approximately given by

$$\frac{\prod_{i=1}^n p_{\theta_0+h/\sqrt{n}}/p_{\theta_0}(X_i)}{\int \prod_{i=1}^n p_{\theta_0+h/\sqrt{n}}/p_{\theta_0}(X_i)\, dh}.$$

As we noted the likelihood ratio processes of the statistical models $(P^n_{\theta_0+h/\sqrt{n}}: h \in \mathbb{R}^d)$ are asymptotically similar to the likelihood ratio process of the normal statistical experiment $(N(h, I_{\theta_0}^{-1}): h \in \mathbb{R}^d)$. Then we may expect the preceding display to be "asymptotically equivalent in distribution" to

$$\frac{dN(h, I_{\theta_0}^{-1})/dN(0, I_{\theta_0}^{-1})(X)}{\int dN(h, I_{\theta_0}^{-1})/dN(0, I_{\theta_0}^{-1})(X)\, dh} = \frac{dN(h, I_{\theta_0}^{-1})(X)}{\int dN(h, I_{\theta_0}^{-1})(X)\, dh} = dN(X, I_{\theta_0}^{-1})(h).$$

Thus the sequence of rescaled posterior distributions should converge to the (random) normal distribution $N(X, I_{\theta_0}^{-1})$, as $n \to \infty$. The expression in the middle of the display is precisely the posterior density for h based on an observation X from the $N(h, I_{\theta_0}^{-1})$-distribution relative to the (improper) Lebesgue prior. Thus we can also say heuristically that the posterior distribution of $\sqrt{n}(\vartheta - \theta_0)$ tends to a posterior distribution in the "limit experiment" $(N(h, I_{\theta_0}^{-1}): h \in \mathbb{R}^d)$, the one relative to the Lebesgue prior dh. Interestingly, when taking the limit, the prior π in the original experiment has disappeared; in the asymptotic experiment the prior is always Lebesgue measure.

We now give a rigorous formulation of the Bernstein-von Mises theorem. The smoothness of the map $\theta \mapsto p_\theta$ is essential, but it turns out that we can do with existence of a first derivative in an appropriate sense, rather than assume three derivatives as in the preceding heuristic discussion. We assume that for every θ in the interior of $\Theta \subset \mathbb{R}^d$, there exists a vector-valued measurable function $\dot\ell_\theta: \mathfrak{X} \to \mathbb{R}^d$ such that, as $h \to 0$,

$$\int \left[p_{\theta+h}^{1/2} - p_\theta^{1/2} - \tfrac{1}{2} h^T \dot\ell_\theta\, p_\theta^{1/2}\right]^2 d\mu = o(\|h\|^2). \tag{2.3}$$

The function $\dot\ell_\theta$ in (2.3) agrees with the score function in the heuristic discussion, because, by the chain rule,

$$\frac{\partial}{\partial\theta}\, p_\theta^{1/2} = \frac{1}{2}\Big(\frac{\partial}{\partial\theta}\, \log p_\theta\Big)\, p_\theta^{1/2}.$$

We define the Fisher information from the function in (2.3) as $I_\theta = \mathrm{Cov}_\theta(\dot\ell_\theta(X_1))$, and assume that this matrix is nonsingular for every θ and that the map $\theta \mapsto I_\theta$ is continuous. Finally, we assume that the parameter is *identifiable*, i.e. the map $\theta \mapsto P_\theta$ is one-to-one.

In the case that the parameter set Θ is not bounded we assume in addition that there exists a sequence of uniformly consistent tests for testing the null hypothesis $H_0: \theta = \theta_0$ against the alternative $H_1: \theta \notin \Theta_0$, for some compact neighborhood $\Theta_0 \subset \Theta$ of θ_0. Recall that a *test* is a measurable function of the observations taking values in the interval $[0, 1]$; in the present context this means a measurable function $\phi_n: \mathfrak{X}^n \to [0, 1]$. The interpretation is that H_0 is rejected with probability $\phi_n(x)$ and hence the expectations $P^n_{\theta_0}\phi_n$ and $P^n_\theta(1 - \phi_n)$ appearing in the following theorem are the probabilities of errors of the first and second kinds of the tests. It can be shown that for a compact parameter set Θ the required tests exist automatically

under (2.3) and identifiability (see van der Vaart (1998), Chapters 6 and 7). The condition in the theorem serves to extend to non-compact parameter sets.

The *total variation norm* between two probability measures P and Q is given by

$$\|P - Q\|_{TV} = 2 \sup_B \bigl|P(B) - Q(B)\bigr| = \int |p - q|\, d\nu,$$

where the supremum is taken over all measurable sets B and p and q are densities of P and Q relative to some measure ν (see Lemma 2.28).

Theorem 2.16 (Bernstein-von Mises)**.** *Suppose that for some compact neighborhood $\Theta_0 \subset \Theta$ of θ_0, there exists a sequence of tests ϕ_n such that*

$$P_{\theta_0}^n \phi_n \to 0, \qquad \sup_{\theta \notin \Theta_0} P_\theta^n (1 - \phi_n) \to 0. \tag{2.4}$$

Furthermore, assume that (2.3) holds at every θ in the interior of Θ with nonsingular Fisher information, and that the map $\theta \mapsto P_\theta$ is one-to-one. If θ_0 is an inner point of Θ and the prior measure is absolutely continuous with a bounded density that is continuous and positive in a neighborhood of θ_0, then the corresponding posterior distributions satisfy

$$\Bigl\| \Pi\bigl(\sqrt{n}(\vartheta - \theta_0) \in \cdot \,|\, X_1, \ldots, X_n\bigr) - N_d\bigl(I_{\theta_0}^{-1} \Delta_{n,\theta_0}, I_{\theta_0}^{-1}\bigr) \Bigr\|_{TV} \overset{P_{\theta_0}^n}{\to} 0.$$

Here the sequence of variables Δ_{n,θ_0} is defined in (2.2) and converges under θ_0 in distribution to a $N_d(0, I_{\theta_0})$-distribution.

Proof For a given sequence $M_n \to \infty$, to be determined later in the proof, we consider the posterior distribution of $\sqrt{n}(\vartheta - \theta_0)$ conditioned on ϑ to belong to $\Theta_n = \{\theta \colon \|\theta - \theta_0\| < M_n/\sqrt{n}\}$, given by

$$\Pi_n\bigl(B \,|\, X^{(n)}\bigr) = \frac{\int_{\Theta_n \cap (\theta_0 + B/\sqrt{n})} \prod_{i=1}^n (p_\theta / p_{\theta_0})(X_i)\, d\Pi(\theta)}{\int_{\Theta_n} \prod_{i=1}^n (p_\theta / p_{\theta_0})(X_i)\, d\Pi(\theta)}. \tag{2.5}$$

The total variation distance between a probability distribution Q and its renormalized restriction $Q(\cdot \,|\, \Theta_n) = Q(\cdot \cap \Theta_n)/Q(\Theta_n)$ to a set Θ_n is bounded above by $2Q(\Theta_n^c)$. Therefore it suffices to show that $\Pi(\sqrt{n}(\vartheta - \theta_0) \in \Theta_n^c \,|\, X_1, \ldots, X_n) \to 0$ in probability, and that the measures in (2.5) approximate the Gaussian distribution in the theorem.

Let A_n be the event that $\int \prod_{i=1}^n (p_\theta / p_{\theta_0})(X_i)\, d\Pi(\theta) \geq n^{-d/2} \epsilon_n$, for a given sequence $\epsilon_n \to 0$. We prove below that the probability of the events A_n tends to one for any $\epsilon_n \to 0$. By Bayes's formula,

$$\Pi\bigl(\sqrt{n}(\vartheta - \theta_0) \in \Theta_n^c \,|\, X_1, \ldots, X_n\bigr) 1_{A_n} (1 - \phi_n) \leq \frac{\int_{\Theta_n^c} \prod_{i=1}^n (p_\theta / p_{\theta_0})(X_i)(1 - \phi_n)\, d\Pi(\theta)}{n^{-d/2} \epsilon_n},$$

where we leave off the argument $(X_1, \ldots, X_n)$ of ϕ_n for ease of notation. The expectation of the integrand on the right is equal to $\int_{\prod p_{\theta_0}(x_i) > 0} \prod_{i=1}^n p_\theta(x_i)(1 - \phi_n) \otimes d\mu(x_i) \leq \mathrm{P}_\theta^n (1 - \phi_n)$. By Lemma 2.35 we can assume without loss of generality that the tests ϕ_n satisfy $P_\theta^n (1 - \phi_n) \leq e^{-cn(\|\theta - \theta_0\|^2 \wedge 1)}$, for every $\theta \in \Theta_n^c$. Then, by Fubini's theorem, the expectation of the preceding

display is bounded above by

$$\frac{\int_{\|\theta-\theta_0\|>M_n/\sqrt{n}} e^{-cn(\|\theta-\theta\|^2\wedge 1)}\, d\Pi(\theta)}{n^{-d/2}\epsilon_n}$$
$$\leq \epsilon_n^{-1} \int_{M_n<\|h\|<\sqrt{n}} e^{-c\|h\|^2}\, \|\pi\|_\infty\, dh + n^{d/2}\epsilon_n^{-1} \int_{\|\theta-\theta_0\|\geq 1} e^{-cn}\, d\Pi(\theta).$$

This tends to zero if ϵ_n tends to zero sufficiently slowly (relative to M_n).

Trivially, we also have $\mathrm{E}_{\theta_0}\Pi(\sqrt{n}(\vartheta-\theta_0)\in\Theta_n^c|\, X_1,\ldots,X_n)\phi_n \leq P_{\theta_0}^n\phi_n \to 0$. We are left to prove that $P_{\theta_0}^n(A_n)\to 1$ and that the measures in (2.5) approximate the Gaussian distribution in the theorem.

Under condition (2.3) the log likelihood can be expanded as in (2.1) with a remainder term that tends to zero (see Lemma 2.30). By replacing $\Delta_{n,\theta}$ as given in (2.2), if necessary, by "truncated versions", it is possible to ensure that (2.1) holds with random vectors $\Delta_{n,\theta}$ that tend under θ to a $N_d(0, I_\theta^{-1})$-distribution not only in the ordinary sense of distributional convergence (which is always valid for the vectors in (2.2) by the central limit theorem), but also in the sense of convergence of Laplace transforms: for every M:

$$\sup_{\|h\|<M} \left|\mathrm{E}_\theta e^{h^T\Delta_{n,\theta}-\frac{1}{2}h^TI_\theta h} - 1\right| \to 0.$$

See Lemma 2.32 in the Complements. Note that $\mathrm{E}e^{h^T\Delta_\theta - h^TI_\theta h/2} = 1$, for every h, when Δ_θ is a $N_d(0, I_\theta^{-1})$-vector. By an extension of Scheffé's lemma (see Lemma 2.38), it follows that the sequence $e^{h^T\Delta_{n,\theta}-h^TI_\theta h/2}$ is uniformly integrable. Similarly, by (2.1) the sequence of likelihood ratios $L_n := \prod_{i=1}^n (p_{\theta+h/\sqrt{n}}/p_\theta)(X_i)$ tends in distribution to $e^{h^T\Delta_\theta - h^TI_\theta h/2}$, and, by Fubini's theorem,

$$\mathrm{E}_\theta L_n = \int_{\prod p_\theta(x_i)>0} \prod_{i=1}^n p_{\theta+h/\sqrt{n}}(x_i)\otimes d\mu(x_i) = (1-P_{\theta+h/\sqrt{n}}(p_\theta=0))^n \to 1,$$

as $P_{\theta+h}(p_\theta=0) = o(\|h\|^2)$, as $h\to 0$, by (2.3) with the integral restricted to the set $\{p_\theta=0\}$. Hence by Lemma 2.38 the sequence L_n is also uniformly integrable, whence (2.1) can be strengthened to convergence in mean: for every h:

$$\mathrm{E}_\theta \left| \prod_{i=1}^n \frac{p_{\theta+h/\sqrt{n}}}{p_\theta}(X_i) - e^{h^T\Delta_{n,\theta}-\frac{1}{2}h^TI_\theta h}\right| \to 0.$$

By dominated convergence the integral with respect to h over the set $\{h\in B\colon \|h\|<M\}$ also tends to zero, uniformly in B, for every fixed M. Next, by Jensen's inequality and Fubini's theorem, we obtain that

$$\mathrm{E}_\theta \sup_B \left| \int_{h\in B:\|h\|<M} \prod_{i=1}^n \frac{p_{\theta+h/\sqrt{n}}}{p_\theta}(X_i)\, dh - \int_{h\in B:\|h\|<M} e^{h^T\Delta_{n,\theta}-\frac{1}{2}h^TI_\theta h}\, dh\right| \to 0.$$

Since this is true for every fixed M, there also exists $M=M_n\to\infty$ such that the expression at $M=M_n$ also tends to zero.

By changing coordinates $\sqrt{n}(\theta-\theta_0)\hookrightarrow h$ in the two integrals in the right side of (2.5) we

can write this quotient in the form $Y_n(B)/Y_n(\mathbb{R}^d)$, for

$$Y_n(B) = \int_{h\in B:\|h\|<M_n} \prod_{i=1}^n \frac{p_{\theta_0+h/\sqrt{n}}}{p_{\theta_0}}(X_i)\,\frac{\pi(\theta_0+h/\sqrt{n})}{\pi(\theta_0)}\,dh.$$

The quotient $\pi(\theta_0 + h/\sqrt{n})/\pi(\theta_0)$ tends to 1, uniformly in $\|h\| \le M_n$, provided $M_n/\sqrt{n} \to 0$. We compare $Y_n(B)$ to $Z_n(B)$, given by

$$Z_n(B) = \int_{h\in B:\|h\|<M_n} e^{h^T\Delta_{n,\theta_0}-\frac{1}{2}h^T I_{\theta_0}h}\,dh.$$

For $B = \mathbb{R}^d$ and $M_n = \infty$ the integral on the right can be computed explicitly, using the expression for a multivariate normal density. For finite M_n we have

$$Z_n(\mathbb{R}^d) = \frac{(2\pi)^{d/2}}{\sqrt{\det I_{\theta_0}}}e^{\frac{1}{2}\Delta_{n,\theta_0}^T I_{\theta_0}^{-1}\Delta_{n,\theta_0}}(1-c_n), \qquad c_n = \Phi_{I_{\theta_0}^{-1}\Delta_{n,\theta_0},I_{\theta_0}^{-1}}(h\colon \|h\| \ge M_n),$$

for $\Phi_{\mu,\Sigma}$ the normal distribution with mean μ and covariance matrix Σ. For $M_n \to \infty$ the sequence c_n tends in mean to 0, whence $Z_n(\mathbb{R}^d)$ is bounded away from zero and tends in distribution to a positive random variable. Then $0 \le Z_n(B) \le Z_n(\mathbb{R}^d)$ is bounded, uniformly in B, and we can conclude that $\sup_B\big|Y_n(B) - Z_n(B)\big| \to 0$, in probability, by the preceding paragraph, and also that $\sup_B |Y_n(B)/Y_n(\mathbb{R}^d) - Z_n(B)/Z_n(\mathbb{R}^d)|$ tends to zero, uniformly in B. The quotient $Z_n(B)(1-c_n)/Z_n(\mathbb{R}^d)$ is equal to $N_d(I_{\theta_0}^{-1}\Delta_{n,\theta_0}, I_{\theta_0}^{-1})(B \cap \{h\colon \|h\| < M_n\})$, which converges to $N_d(I_{\theta_0}^{-1}\Delta_{n,\theta_0}, I_{\theta_0}^{-1})(B)$, uniformly in B. This completes the proof that the measures (2.5) approximate the Gaussian distribution in the theorem in the total variation norm.

Finally the events A_n contain the events that $Y_n(\mathbb{R}^d) \ge c\epsilon_n$, if $c \le \pi(\theta_0 + h/\sqrt{n})/\pi(\theta_0)$ for every $\|h\| < M_n$. If $\epsilon_n \to 0$ their probability tends to one, as $Y_n(\mathbb{R}^d)$ tends to a strictly positive random variable. □

Under conditions somewhat stronger than imposed in the preceding theorem (see e.g. van der Vaart (1998), Theorem 5.39), the maximum likelihood estimators $\hat\theta_n$ of θ satisfy

$$\sqrt{n}(\hat\theta_n - \theta_0) - I_{\theta_0}^{-1}\Delta_{n,\theta_0} \overset{P_{\theta_0}^n}{\to} 0. \tag{2.6}$$

This motivates to restate the theorem in a different form.

Corollary 2.17. *Under the conditions of the preceding theorem, for any estimators $\hat\theta_n$ that satisfy (2.6),*

$$\Big\|\Pi(\vartheta \in \cdot\,|\,X_1,\ldots,X_n) - N_d\Big(\hat\theta_n, \frac{1}{n}I_{\theta_0}^{-1}\Big)\Big\|_{TV} \overset{P_{\theta_0}^n}{\to} 0.$$

Under regularity conditions this is true for the maximum likelihood estimators $\hat\theta_n$.

Proof We can first replace the mean $I_{\theta_0}^{-1}\Delta_{n,\theta_0}$ of the normal approximation in Theorem 2.16 by $\sqrt{n}(\hat\theta_n - \theta_0)$, using the fact that the total variation distance between two normal distributions $N(\mu,\Sigma)$ and $N(\nu,\Sigma)$ with means μ and ν and equal nonsingular covariance matrix Σ is bounded above by a multiple of $\|\mu-\nu\|$ (with constant depending on Σ). Because it is the supremum over all Borel sets, the total variation distance is invariant under shifting and scaling: if $P_{\mu,\sigma}$ and $Q_{\mu,\sigma}$ are the laws of $\sigma Z + \mu$ if some variable Z has laws P and Q, then $\|P_{\mu,\sigma} - Q_{\mu,\sigma}\| = \|P - Q\|$. We apply this to replace $\sqrt{n}(\vartheta - \theta_0)$ by ϑ in

$\Pi(\sqrt{n}(\vartheta - \theta_0) \in \cdot \,|\, X_1, \ldots, X_n)$, changing the location and scale of the approximating normal distribution $N_d(\sqrt{n}(\hat{\theta}_n - \theta_0), I_{\theta_0}^{-1})$ simultaneously. □

2.4 Credible regions

An important consequence of the Bernstein-von Mises theorem is that credible sets are asymptotic confidence sets: *for parametric models Bayesian and frequentist inference are asymptotically equivalent*.

We consider the case of setting a confidence interval for a linear combination $g^T\theta$ of the parameters, for given g, for instance an individual coordinate θ_i. Then a natural Bayesian credible interval would range between two quantiles of the posterior distribution of $g^T\vartheta$ given $X_1, \ldots, X_n$. Denote the distribution function and quantile function of this posterior distribution by

$$F_{g^T\vartheta}(y \,|\, x_1, \ldots, x_n) = \Pi(\theta \colon g^T\theta \le y \,|\, x_1, \ldots, x_n),$$
$$F_{g^T\vartheta}^{-1}(\alpha \,|\, x_1, \ldots, x_n) = \inf\bigl\{y \colon F_{g^T\vartheta}(y \,|\, x_1, \ldots, x_n) \ge \alpha\bigr\}.$$

The natural Bayesian credible interval of level $1 - 2\alpha$ for $g^T\theta$ is

$$\hat{C}_n = \Bigl[F_{g^T\vartheta}^{-1}\bigl(\alpha \,|\, X_1, \ldots, X_n\bigr),\ F_{g^T\vartheta}^{-1}\bigl(1 - \alpha \,|\, x_1, \ldots, x_n\bigr)\Bigr]. \tag{2.7}$$

We compare this to the usual *Wald asymptotic confidence interval* for $g^T\theta$. If $\hat{\theta}_n$ is a sequence of estimators satisfying (2.6), then the sequence $\sqrt{n}(g^T\hat{\theta}_n - g^T\theta_0)/(g^T I_{\theta_0}^{-1} g)^{1/2}$ converges under θ_0 in distribution to a standard normal distribution. Therefore this variable is an asymptotic pivot (i.e. its distribution is free of the parameter) and gives rise to the asymptotic $1 - 2\alpha$-confidence interval

$$\left[g^T\hat{\theta}_n + \Phi^{-1}(\alpha)\sqrt{\frac{g^T I_{\hat{\theta}_n}^{-1} g}{n}},\quad g^T\hat{\theta}_n + \Phi^{-1}(1-\alpha)\sqrt{\frac{g^T I_{\hat{\theta}_n}^{-1} g}{n}}\right].$$

The following theorem shows that the credible and Wald intervals are asymptotically equivalent.

Theorem 2.18. *Under the conditions of Corollary 2.17, for any $\hat{\theta}_n$ satisfying (2.6), and $0 < \alpha < 1$,*

$$F_{g^T\vartheta}^{-1}(\alpha \,|\, X_1, \ldots, X_n) - g^T\hat{\theta}_n - \Phi^{-1}(\alpha)\sqrt{\frac{g^T I_{\hat{\theta}_n}^{-1} g}{n}} = o_{P_\theta^n}\Bigl(\frac{1}{\sqrt{n}}\Bigr).$$

Consequently, the frequentist coverage $\Pr_{\theta_0}(g^T\theta_0 \in \hat{C}_n)$ of the credible interval (2.7) tends to $1 - 2\alpha$, as $n \to \infty$.

Proof Because the total variation norm is a supremum over all Borel sets, the Bernstein-von Mises theorem implies that

$$\hat{\epsilon}_n \colon = \sup_y \bigl|F_{g^T\vartheta}(y \,|\, X_1, \ldots, X_n) - \Phi\bigl((g^T I_{\theta_0}^{-1} g/n)^{-1/2}(y - g^T\hat{\theta}_n)\bigr)\bigr| \xrightarrow{P_{\theta_0}^n} 0.$$

By the definition of a quantile function F^{-1} and right-continuity of a distribution function,

we have $F(F^{-1}(\alpha) - 1/n) \le \alpha \le F(F^{-1}(\alpha))$. Therefore, applying the preceding display with $\hat{\xi}_n = F^{-1}_{g^T\hat{\vartheta}}(\alpha | X_1, \ldots, X_n)$ and this same value $-1/n$, we conclude that

$$\Phi((g^T I_{\theta_0}^{-1} g/n)^{-1/2}(\hat{\xi}_n - 1/n - g^T\hat{\theta}_n)) \le \alpha + \hat{\epsilon}_n,$$
$$\alpha - \hat{\epsilon}_n \le \Phi((g^T I_{\theta_0}^{-1} g/n)^{-1/2}(\hat{\xi}_n - g^T\hat{\theta}_n)).$$

We apply Φ^{-1} across these inequalities to obtain the first assertion of the theorem.

For the proof of the final assertion we note that θ_0 falls to the left of $\hat{C}_n$ if and only if $g^T\theta_0 < F^{-1}_{g^T\hat{\vartheta}}(\alpha | X_1, \ldots, X_n)$. In view of the first assertion this is equivalent to

$$\sqrt{n}(g^T\hat{\theta}_n - g^T\theta_0) > -\Phi^{-1}(\alpha)\sqrt{g^T I^{-1}_{\hat{\theta}_n} g} + o_P(1).$$

The probability of this event tends to α by assumption (2.6). We combine this with a similar argument for the right tail. □

Warning 2.19. It is tempting to think that under the conditions of the Bernstein-von Mises theorem any data-based $\hat{C}_n$ that satisfies $\Pi(\hat{C}_n | X_1, \ldots, X_n) = 1-\alpha$ almost surely, also satisfies $P^n_\theta(\theta \in \hat{C}_n) \to 1 - \alpha$, for every θ. In other words, that every credible set is automatically a confidence set of asymptotically the same level. However, this is false (see Problem 2.11). The sets in the preceding theorem are special in that they have a pivotal property.

2.5 Bayes estimators

The Bernstein-von Mises theorem shows that the posterior laws converge in distribution to a Gaussian posterior law in total variation norm. As a consequence, by the continuous mapping theorem for convergence in distribution any "location functional" that is continuous relative to the total variation norm applied to the sequence of posterior laws will converge to the same location functional applied to the limiting Gaussian posterior distribution. For most choices this means to X, i.e. a $N(0, I_{\theta_0}^{-1})$-distribution. For instance, this argument may be applied to show that the posterior median is asymptotically normal (Problem 2.8).

Not every location functional is continuous relative to the total variation norm. For instance, to show that the posterior mean is asymptotically normally distributed we need to do extra work. Consider general Bayes estimators relative to a loss function $\ell: \mathbb{R}^d \to [0, \infty)$: for fixed $X_1, \ldots, X_n$ let T_n minimize the posterior risk

$$t \mapsto \frac{\int \ell(\sqrt{n}(t-\theta)) \prod_{i=1}^n p_\theta(X_i)\, d\Pi(\theta)}{\int \prod_{i=1}^n p_\theta(X_i)\, d\Pi(\theta)}.$$

We assume that the integrals in the preceding display exist, for almost every sequence of observations, and that T_n can be selected as a measurable function of the observations.

We restrict ourselves to loss functions with the property, for every $M > 0$,

$$\sup_{\|h\| \le M} \ell(h) \le \inf_{\|h\| \ge 2M} \ell(h),$$

with strict inequality for at least one M[4]. This is true, for instance, for loss functions of the form $\ell(h) = \ell_0(\|h\|)$ for a nondecreasing function $\ell_0: [0, \infty) \to [0, \infty)$ that is not constant

[4] The 2 is for convenience, any other number would do.

on $(0, \infty)$. Furthermore, we suppose that ℓ grows at most polynomially: for some constant $p \geq 0$,

$$\ell(h) \leq 1 + \|h\|^p.$$

Theorem 2.20. *Let the conditions of Theorem 2.16 hold, and let ℓ satisfy the conditions as listed, for a p such that $\int \|\theta\|^p \, d\Pi(\theta) < \infty$. Then the sequence $\sqrt{n}(T_n - \theta_0)$ converges under θ_0 in distribution to the minimizer of $t \mapsto \int \ell(t-h)\, dN(X, I_{\theta_0}^{-1})(h)$, for X possessing the $N(0, I_{\theta_0}^{-1})$-distribution, provided that any two minimizers of this process coincide almost surely. In particular, for a loss function of the form $\ell(h) = \ell_0(\|h\|)$ for a nonzero, nondecreasing function ℓ_0 the sequence $\sqrt{n}(T_n - \theta_0)$ converges under θ_0 to the $N(0, I_{\theta_0}^{-1})$-distribution.*

Proof See van der Vaart (1998), Theorem 10.8. The final assertion follows because $t \mapsto \int \ell_0(\tau - X - h\|)\, dN(0, \Sigma)(h)$ is minimized at $t - X = 0$, by the symmetry of the normal distribution. □

Example 2.21 (Posterior mean)**.** The Bayes estimator for $\ell(x) = \|x\|^2$ is the posterior mean

$$T_n = \frac{\int \theta \prod_{i=1}^n p_\theta(X_i)\, \pi(\theta)\, d\theta}{\int \prod_{i=1}^n p_\theta(X_i)\, \pi(\theta)\, d\theta}.$$

The sequence $\sqrt{n}(T_n - \theta_0)$ converges in distribution to a $N_d(0, I_{\theta_0}^{-1})$-distribution provided the prior has a finite second moment (and the conditions of Theorem 2.16 hold).

By extending the argument, it can also be shown that the difference $\sqrt{n}(T_n - \hat{\theta}_n)$ between a Bayes estimator and the centering $\hat{\theta}_n$ in the Bernstein-von Mises theorem tends to zero in probability. As the latter can typically be taken equal to the maximum likelihood estimator, it follows that Bayes and maximum likelihood estimators are often asymptotically equivalent.

2.6 Bayes factors and BIC

A Bayes factor between two models is defined as the quotient of the evidences of the models, and can be used to select the model that best fits the data. For parametric models Bayes factors turn out to be equivalent, in the large sample limit, to the maximum of the likelihood penalized by a multiple of the dimension of the model. This connection leads to the *Bayesian information criterion* for model selection.

We start with an expansion of the evidence for a given prior Π on a smooth d-dimensional parametric model $(P_\theta \colon \theta \in \Theta)$.

Theorem 2.22 (Evidence)**.** *Under the conditions of Theorem 2.16,*

$$\log \int \prod_{i=1}^n \frac{p_\theta}{p_{\theta_0}}(X_i)\, d\Pi(\theta) = -\frac{d}{2} \log n + \log \frac{(2\pi)^{d/2} \pi(\theta_0)}{\sqrt{\det I_{\theta_0}}} + \frac{1}{2} \Delta_{n,\theta_0}^T I_{\theta_0}^{-1} \Delta_{n,\theta_0} + o_{P_{\theta_0}^n}(1).$$

Proof For a given sequence $M_n \to \infty$, which will be determined later in the proof, we partition the domain $\mathbb{R}^d$ of the integral into the sets $\Theta_{n,1}$, $\Theta_{n,2}$ and $\Theta_{n,3}$, consisting of the vectors θ such that $\|\theta - \theta_0\|$ is contained in $[0, M_n/\sqrt{n})$, or $[M_n/\sqrt{n}, 1)$ or $[1, \infty)$. We write

the left side of the theorem accordingly as $\log(B_{n,1} + B_{n,2} + B_{n,3})$, for

$$B_{n,i} = \int_{\Theta_{n,i}} \prod_{i=1}^{n} \frac{p_\theta}{p_{\theta_0}}(X_i)\, d\Pi(\theta).$$

It suffices to show that there exists $M_n \to \infty$ such that $\log B_{n,1}$ has the expansion as given in the right side of the theorem, while $B_{n,i}/B_{n,1} \to 0$ in probability, for both $i = 2, 3$.

By Lemma 2.35 there exist tests ϕ_n with $P^n_{\theta_0}\phi_n \to 0$ and $P^n_\theta(1-\phi_n) \le e^{-cn(\|\theta-\theta_0\|^2 \wedge 1)}$, for every $\theta \in \Theta_{n,2} \cup \Theta_{n,3}$. Then $P^n_{\theta_0} 1_{B_{n,i}/B_{n,1} > \epsilon}\phi_n \to 0$, for every $\epsilon > 0$, and

$$P^n_{\theta_0} B_{n,2}(1-\phi_n) \le \int_{\Theta_{n,2}} e^{-cn\|\theta-\theta_0\|^2}\, d\Pi(\theta) \le \|\pi\|_\infty \int_{M \le \|h\| < \sqrt{n}} e^{-c\|h\|^2} \frac{dh}{n^{d/2}},$$

$$P^n_{\theta_0} B_{n,3}(1-\phi_n) \le \int_{\Theta_{n,3}} e^{-cn}\, d\Pi(\theta) \le e^{-cn}.$$

The second term tends to zero at exponential speed and hence is $o(n^{-d/2})$, while the first term is $o(n^{-d/2})$ for any $M_n \to \infty$. We show below that the sequence $B_{n,1}n^{d/2}$ tends in probability to a positive constant for a particular $M_n \to \infty$, and can then conclude that $P^n_{\theta_0} 1_{B_{n,i}/B_{n,1} > \epsilon}(1 - \phi_n) \le P^n_{\theta_0} 1_{B_{n,i} > \epsilon C n^{d/2}}(1-\phi_n) + o(1) \to 0$, for every $\epsilon > 0$ and $i = 2, 3$, by Markov's inequality.

Under condition (2.3) the log likelihood can be expanded as in (2.1) with a remainder term that tends to zero (see Lemma 2.30). By replacing $\Delta_{n,\theta}$ as given in (2.2), if necessary, by a "truncated version", it is possible to ensure that (2.1) holds with random vectors $\Delta_{n,\theta}$ that tend under θ to a $N_d(0, I_\theta^{-1})$-distribution not only in the ordinary sense of distribution convergence (which is always valid for the vectors in (2.2) by the central limit theorem), but also in the sense of convergence of Laplace transforms: for every M:

$$\sup_{\|h\|<M} \left| \mathrm{E}_\theta e^{h^T\Delta_{n,\theta} - \frac{1}{2}h^T I_\theta h} - 1 \right| \to 0.$$

See Lemma 2.32 in the Complements. (Note that $\mathrm{E}e^{h^T\Delta_\theta - h^T I_\theta h/2} = 1$, for every h, when Δ_θ is a $N_d(0, I_\theta^{-1})$-vector.) By an extension of Scheffé's lemma (see Lemma 2.38), it follows that the sequence $e^{h^T\Delta_{n,\theta} - h^T I_\theta h/2}$ is uniformly integrable. Similarly, by (2.1) the sequence of likelihood ratios $L_n := \prod_{i=1}^n (p_{\theta+h/\sqrt{n}}/p_\theta)(X_i)$ tends in distribution to $e^{h^T\Delta_\theta - h^T I_\theta h/2}$, and, by Fubini's theorem,

$$\mathrm{E}_\theta L_n = \int_{\prod p_\theta(x_i)>0} \prod_{i=1}^{n} p_{\theta+h/\sqrt{n}}(x_i) \otimes d\mu(x_i) = \left(1 - P_{\theta+h/\sqrt{n}}(p_\theta = 0)\right)^n \to 1,$$

as $P_{\theta+h}(p_\theta = 0) = o(\|h\|^2)$, as $h \to 0$, by (2.3) with the integral restricted to the set $\{p_\theta = 0\}$. Hence by Lemma 2.38 the sequence L_n is also uniformly integrable, whence (2.1) can be strengthened to convergence in mean: for every h:

$$\mathrm{E}_\theta \left| \prod_{i=1}^{n} \frac{p_{\theta+h/\sqrt{n}}}{p_\theta}(X_i) - e^{h^T\Delta_{n,\theta} - \frac{1}{2}h^T I_\theta h} \right| \to 0.$$

By dominated convergence the integral with respect to h over the set $\{h: \|h\| < M\}$ also tends to zero. Next, by Jensen's inequality and Fubini's theorem, we obtain that

$$\mathrm{E}_\theta \left| \int_{\|h\|<M} \prod_{i=1}^{n} \frac{p_{\theta+h/\sqrt{n}}}{p_\theta}(X_i)\, dh - \int_{\|h\|<M} e^{h^T\Delta_{n,\theta} - \frac{1}{2}h^T I_\theta h}\, dh \right| \to 0.$$

The second, Gaussian integral in this display can be explicitly evaluated using the formula for a density of the multivariate distribution when $M = \infty$, and hence

$$\int_{\|h\|<M} \prod_{i=1}^{n} \frac{p_{\theta+h/\sqrt{n}}}{p_\theta}(X_i)\, dh = \frac{(2\pi)^{d/2}}{\sqrt{\det I_\theta}} e^{\frac{1}{2}\Delta_{n,\theta}^T I_\theta^{-1}\Delta_{n,\theta}} \Big(1 - \Phi_{I_\theta^{-1}\Delta_{n,\theta}, I_\theta^{-1}}(B_{0,M}^c)\Big) + r_{n,M},$$

for $\Phi_{\mu,\Sigma}$ the normal distribution with mean μ and covariance matrix Σ and $B_{0,M}^c$ the complement of a ball of radius M around $0 \in \mathbb{R}^d$, and $r_{n,M}$ the discrepancy in the preceding display. As $n \to \infty$ followed by $M \to \infty$, the term within brackets tends to 1 in mean and $\mathrm{E}_\theta r_{n,M} \to 0$. Then there also exists $M_n \to \infty$ such that evaluated at $M = M_n$ this remains true when $n \to \infty$.

The final step is to note that by a change of variables

$$B_{n,1} = \int_{\|h\|<M_n} \prod_{i=1}^{n} \frac{p_{\theta_0+h/\sqrt{n}}}{p_{\theta_0}}(X_i)\, \pi(\theta_0 + h/\sqrt{n})\, \frac{dh}{n^{d/2}}.$$

Here $\pi(\theta_0 + h/\sqrt{n}) = \pi(\theta_0)(1 + o(1))$, uniformly in $\|h\| < M_n$, provided $M_n \ll \sqrt{n}$, which we can assume without loss of generality. We can then equate the expression to $\pi(\theta_0)/n^{d/2}$ times the left side of the preceding display evaluated at $\theta = \theta_0$. □

It can be shown, under stronger conditions than in the preceding theorem, that the *log likelihood ratio statistics* for testing the null hypothesis $H_0: \theta = \theta_0$ are under the null hypothesis asymptotic to the random part of the expansion in the preceding theorem (e.g. van der Vaart (1998), Chapter 16):

$$\sup_{\theta\in\Theta} \log \prod_{i=1}^{n} \frac{p_\theta}{p_{\theta_0}}(X_i) = \frac{1}{2}\Delta_{n,\theta_0}^T I_{\theta_0}^{-1}\Delta_{n,\theta_0} + o_{P_{\theta_0}^n}(1).$$

This expansion is the justification for the usual chisquare approximation to the null distribution of twice the log likelihood ratio statistics. More interesting in the present context is that combined the two approximations give

$$\log \int \prod_{i=1}^{n} p_\theta(X_i)\, d\Pi(\theta) = \sup_{\theta\in\Theta} \log \prod_{i=1}^{n} p_\theta(X_i) - \frac{d}{2}\log n + \log \frac{(2\pi)^{d/2}\pi(\theta_0)}{\sqrt{\det I_{\theta_0}}} + o_{P_{\theta_0}^n}(1).$$

This *Laplace expansion* can also be obtained by expanding the integrand in the integral on the left around its point of maximum $\prod_{i=1}^n p_{\hat\theta}(X_i)$, where $\hat\theta$ is the maximum likelihood estimator. (For large n the integrand is determined by a small neighborhood of its point of maximum, and can be approximated by replacing the logarithm of its integrand by a quadratic Taylor expansion. The latter calculation is similar to the one in the preceding proof, which uses an expansion around the true value of the parameter.)

The preceding display shows that the log evidence is equal to the maximum of the log likelihood minus a remainder term. For $n \to \infty$, the factor $-(d/2)\log n$ is the dominant part of the remainder, the other part being fixed, and hence the remainder is negative and can be viewed as a *penalty* that pulls down the maximum of the likelihood. Twice this penalty is known as the *BIC penalty*, and the BIC criterion is defined by

$$\mathrm{BIC} = -2\Big[\sup_{\theta\in\Theta} \log \prod_{i=1}^{n} p_\theta(X_i) - \frac{d}{2}\log n\Big].$$

Models of higher dimension d are penalized more, which is meant to compensate for the bigger maximum over a bigger model. The maximum value of the penalized likelihood is often used to compare the quality of competing models of different dimensions. The preceding approximation shows that the model with the largest penalized likelihood (and hence smallest BIC-value) is, up to a constant, the same as the model with the largest evidence. [5]

The real Bayesian will of course not need BIC, but use the evidence of a model directly, through the comparison of models by Bayes factors. The following theorem shows that the criterion is *consistent* in that given a finite set of models, it chooses the smallest model that fits the data.

Suppose given a *finite* set of models $(P_{\theta,k} : \theta \in \Theta_k)$ indexed by parameter sets $\Theta_k \subset \mathbb{R}^{d_k}$ of dimensions d_k, for k running through some finite index set. Each model comes with a prior Π_k, and hence models k and l can be compared by their Bayes factor

$$\mathrm{BF}_n(\Theta_k, \Theta_l) = \frac{\int \prod_{i=1}^n p_{\theta,k}(X_i)\, d\Pi_k(\theta)}{\int \prod_{i=1}^n p_{\theta,l}(X_i)\, d\Pi_l(\theta)}.$$

Theorem 2.23. *Assume that the conditions of Theorem 2.16 hold for every of the finitely many models $(P_{\theta,k} : \in \Theta_k)$. If the observations are distributed according to a parameter $\theta_0 \in \Theta_{k_0}$, then $\mathrm{BF}_n(\Theta_k, \Theta_{k_0}) \to 0$ in $P^n_{\theta_0,k_0}$-probability, for any k satisfying one of the two following conditions:*

(i) *$d_k > d_{k_0}$ and $p_{\theta_0,k_0} = p_{\theta_k,k}$, for some $\theta_k \in \Theta_k$.*
(ii) *there exists tests ϕ_n with $P^n_{\theta_0,k_0}\phi_n \to 0$ and $\sup_{\theta\in\Theta_k} P^n_{\theta,k}(1-\phi_n) \to 0$.*

Proof To prove assertion (i) we apply Theorem 2.22 to find that the logarithm of the Bayes factors $\mathrm{BF}_n(\Theta_k, \Theta_{k_0}) \xrightarrow{P} 0$ behave asymptotically as $-(d_k/2 - d_{k_0}/2)\log n$ up to a term that is bounded in probability. If $d_k > d_{k_0}$, then this tends to $-\infty$.

For the proof of (ii) we note first that by Theorem 2.22 the probability of the events A_n such that $\int \prod_{i=1}^n (p_{\theta,k_0}/p_{\theta_0,k_0})(X_i)\, d\Pi_{k_0}(\theta) \geq n^{-d_{k_0}/2}\epsilon_n$, tends to one, for an arbitrary given $\epsilon_n \to 0$. On the events A_n the Bayes factors are bounded above by

$$\frac{n^{d_{k_0}/2}}{\epsilon_n} \int \prod_{i=1}^n \frac{p_{\theta,k}}{p_{\theta_0,k_0}}(X_i)\, d\Pi_k(\theta),$$

and hence it suffices to show that this sequence tends to zero in probability. By Lemma 2.33, the tests ϕ_n in (ii) can be improved, if necessary, to have error probabilities satisfying $P^n_{\theta,k}(1-\phi_n) \leq e^{-cn}$, for every $\theta \in \Theta_k$, and some $c > 0$. The event $K_n = \{\phi_n > 1/2\}$ has probability $P^n_{\theta_0,k_0}(K_n) \leq 2P^n_{\theta_0,k_0}\phi_n \to 0$, while, by Fubini's theorem,

$$P^n_{\theta_0,k_0}\, \mathrm{BF}_n(\Theta_k, \Theta_{k_0}) 1_{K_n^c \cap A_n} \leq \frac{n^{d_{k_0}/2}}{\epsilon_n} \int P^n_{\theta,k}(K_n^c)\, d\Pi_k(\theta) \leq 2\frac{n^{d_{k_0}/2}}{\epsilon_n} e^{-cn} \to 0,$$

since $1_{K_n^c} \leq 2(1-\phi_n)$. Thus on the events $K_n^c \cap A_n$ the Bayes factors tend to zero in mean. An application of Markov's inequality completes the proof. □

[5] The "B" of BIC is for "Bayesian", but it can also be viewed as the next letter following the "A" of the competing AIC penalty (Akaike's information criterion), which uses the penalty d instead of $(d/2)\log n$. For large n the BIC penalty pulls the likelihood of bigger models down more than the AIC penalty, with as a result that BIC tends to choose smaller models than AIC.

Part (i) of the theorem shows that the Bayes factors (or equivalently the BIC criterion) will asymptotically choose the model with the smaller dimension if both models contain the true distribution of the data. This is comforting; it implies for instance that no unnecessary explanatory variables will be included in a regression model. Part (ii) is applicable in the situation that the alternative model does not contain the true distribution of the data. In that case the Bayes factors will still choose the correct model provided that there is some procedure (the test ϕ_n) that is able to make the discrimination between the true parameter and the alternative. The condition will typically be satisfied if the true distribution P_{θ_0,k_0} is at a positive distance from the model $(P_{\theta,k}\colon \theta \in \Theta_k)$, as is the case for most of the usual models. The testing criterion rules out situations, where the true distribution P_{θ_0,k_0} can be approximated by distributions $P_{\theta,k}$ for θ approaching the boundary of the parameter set. We discuss the construction of tests in Section 5.2.

*Uniform consistency

These positive assertions concern *pointwise* consistency, in that the probability of choosing the correct model tends to one under a fixed true distribution. This convergence is not necessarily uniform in the underlying law P_{θ_0,k_0}, so that the size of the dataset needed for a correct choice may be much larger for some laws than for other laws. The lack of uniformity was illustrated in Example 1.15, where the Bayes factor made the wrong choice for parameter sequences in the alternative model approaching the null model at the rate $1/\sqrt{n}$. This is a general phenomenon. Probability measures in a smooth parametric model $(P_\theta\colon \theta \in \mathbb{R}^k)$ at parameter values separated by a distance of the order $1/\sqrt{n}$ are *contiguous*: for any events A_n and any $h \in \mathbb{R}^k$:

$$P^n_{\theta+h/\sqrt{n}}(A_n) \to 1, \qquad \text{if and only if} \qquad P^n_\theta(A_n) \to 1.$$

(See e.g. van der Vaart (1998), Chapters 6 and 7.) This implies that the Bayes factor (or any other model selection method) will choose the *same* model when the true parameter is $\theta+h/\sqrt{n}$ or θ if it consistently chooses some model under one of these parameters. However, the parameters $\theta + h/\sqrt{n}$ may not belong to the same model as θ, as the perturbation $h/\sqrt{n}$ may point outside the model for θ, depending on h. In the latter case no method can choose the correct model both when $\theta + h/\sqrt{n}$ or θ is the true parameter, as any method will choose the same model in both cases. Thus there cannot be uniformity in the parameter for choosing the correct model, not even locally, as θ and $\theta + h/\sqrt{n}$ are close.

2.7 DIC

For data X, a statistical model given by densities p_θ, and a given prior, the *deviance information criterion* (DIC) is defined as, for $\bar\theta = \bar\theta(X) = \mathrm{E}(\vartheta|\, X)$ the posterior mean,

$$\mathrm{DIC} = -2\Big[\log p_{\bar\theta}(X) - 2\,\mathrm{E}\Big(\log \frac{p_{\bar\theta}}{p_\vartheta}(X)\big|\, X\Big)\Big].$$

One prefers the model that scores the smallest value on this criterion. The conditional expectation on the right side refers to the variable ϑ given X, and hence is relative to the posterior distribution.

The first term of the criterion may be compared to the maximum of the likelihood, to which it would reduce if the Bayes estimator $\bar{\theta}$ were replaced by the maximum likelihood estimator. The second term is a penalty, which should prevent fitting a too large model. Both $\bar{\theta}$ and the penalty are posterior means and hence may be approximated by an average over a large sample of values generated from the posterior distribution. As explained in the next chapter, generating such samples is a popular way of computing a posterior distribution. This explains that the DIC has become standard output of computer packages for Bayesian computation, and hence has gained some popularity.

We shall now show that DIC is closely related to the AIC criterion for model selection, and hence possesses good properties.[6]

In the case that $X = (X_1, \ldots, X_n)$ is a sample from a smoothly parametrized model $\theta \mapsto p_\theta$, the difference between the Bayes estimator $\bar{\theta}$ and the maximum likelihood estimator $\hat{\theta}$ will be negligible, in view of the Bernstein-von Mises theorem, and hence the leading term in the DIC criterion will essentially be equal to the maximum of the likelihood. Furthermore, the linear term in an expansion of the log likelihood ratio $\log p_\theta$ around $\bar{\theta}$ will approximately vanish (as $\bar{\theta} \approx \hat{\theta}$), and it will certainly vanish after taking a conditional expectation given the data, and hence in the penalty term we can make the approximation

$$\log \frac{p_{\bar{\theta}}}{p_\vartheta}(X) \approx \tfrac{1}{2} n(\vartheta - \bar{\theta})^T I_\theta (\vartheta - \bar{\theta}),$$

where I_θ is the Fisher information in a single observation. By the Bernstein-von Mises theorem ϑ is conditionally given X approximately normally distributed with mean $\bar{\theta}$ (of course!) and covariance $(nI_\theta)^{-1}$, whence $n(\vartheta - \bar{\theta})^T I_\theta (\vartheta - \bar{\theta})$ is approximately chisquare distributed with d degrees for freedom, for d the dimension of the parameter, still given X. This suggests that the posterior expectation of the right side of the preceding display is approximately equal to $d/2$. Thus the DIC penalty is close to d, and the DIC criterion is close to the AIC criterion, which is given by

$$\text{AIC} = -2\Big[\sup_{\theta \in \Theta} \log p_\theta(X) - d\Big].$$

Relative to BIC this replaces the penalty $d \log n/2$ by d and hence puts a smaller penalty on higher dimensions (when $\log n > 2$), and favors bigger models. AIC is typically not consistent for model selection, but can be shown to yield a "best fitting model" in terms of Kullback-Leibler distance.

*Misspecified model

If the true distribution of the data does not belong to the model, as can be expected for a model that is too small, then the preceding derivation should be adapted, although the conclusion will be the same. The Bayesian and maximum likelihood estimators will still be close, but in the expansion of the log likelihood the Fisher information matrix should be replaced by the expectation $J = -P\ddot{\ell}_{\theta^*}$ of the second derivative matrix of the likelihood under the true distribution P, evaluated at a parameter θ^* in the model for which $K(p; p_\theta)$ is minimal, which is known to be the asymptotic value of $\bar{\theta}$ (see Kleijn and van der Vaart

[6] We have also seen it applied to choose between different priors. That appears not to be justifiable.

(2012)). The matrix J does not necessarily reduce to the Fisher information. Furthermore, the Bernstein-von Mises theorem should be adapted to the misspecified model as well, and will now show that the posterior distribution is normal with mean $\bar{\theta}$ (of course!) and covariance matrix $n^{-1}J$. The same argument as before will show that the DIC penalty will again be close to the dimension d of the model. This is not necessarily good news, as a different penalty than AIC may be preferable for smaller models. [7]

2.8 Complements

2.8.1 Finite-dimensional Dirichlet distribution

Definition 2.24 (Dirichlet distribution)**.** The *Dirichlet distribution* $\mathrm{Dir}(k;\alpha)$ with parameters $k \in \mathbb{N} - \{1\}$ and $\alpha = (\alpha_1, \ldots, \alpha_k) > 0$ is the distribution of a vector $(X_1, \ldots, X_k)$ such that $\sum_{i=1}^k X_i = 1$ and such that $(X_1, \ldots, X_{k-1})$ has density

$$\frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)} x_1^{\alpha_1 - 1} x_2^{\alpha_2 - 1} \cdots x_{k-1}^{\alpha_{k-1} - 1} (1 - x_1 - \cdots - x_{k-1})^{\alpha_k - 1}, \qquad x_i > 0, \sum_i x_i < 1. \tag{2.8}$$

The Dirichlet distribution with parameters k and $\alpha \geq 0$, where $\alpha_i = 0$ for $i \in I \subsetneq \{1, \ldots, k\}$, is the distribution of the vector $(X_1, \ldots, X_k)$ such that $X_i = 0$ for $i \in I$ and such that $(X_i : i \notin I)$ possesses a lower-dimensional Dirichlet distribution, given by a density of the form (2.8).

Proposition 2.25 (Gamma representation)**.** *If* $Y_i \overset{ind}{\sim} \mathrm{Ga}(\alpha_i, 1)$ *for* $i = 1, \ldots, k$ *and* $Y := \sum_{i=1}^k Y_i$*, then* $(Y_1/Y, \ldots, Y_k/Y) \sim \mathrm{Dir}(k; \alpha_1, \ldots, \alpha_k)$*, and is independent of* Y*.*

Proof We may assume that all α_i are positive. The Jacobian of the inverse of the transformation $(y_1, \ldots, y_k) \mapsto (y_1/y, \ldots, y_{k-1}/y, y) =: (x_1, \ldots, x_{k-1}, y)$ is given by y^{k-1}. The density of the $\mathrm{Ga}(\alpha_i, 1)$-distribution is proportional to $e^{-y_i} y_i^{\alpha_i - 1}$. Therefore the joint density of $(Y_1/Y, \ldots, Y_{k-1}/Y, Y)$ is, proportional to,

$$e^{-y} y^{|\alpha| - 1} x_1^{\alpha_1 - 1} \cdots x_{k-1}^{\alpha_{k-1} - 1} (1 - x_1 - \cdots - x_{k-1})^{\alpha_k - 1}.$$

This factorizes into a Dirichlet density dimension $k - 1$ and the $\mathrm{Ga}(|\alpha|, 1)$-density of Y. □

Proposition 2.26 (Aggregation)**.** *If* $X \sim \mathrm{Dir}(k; \alpha_1, \ldots, \alpha_k)$ *and* $Z_j = \sum_{i \in I_j} X_i$ *for a given partition* $I_1, \ldots, I_m$ *of* $\{1, \ldots, k\}$*, then*

(i) $(Z_1, \ldots, Z_m) \sim \mathrm{Dir}(m; \beta_1, \ldots, \beta_m)$*, where* $\beta_j = \sum_{i \in I_j} \alpha_i$*, for* $j = 1, \ldots, m$*.*
(ii) $(X_i/Z_j : i \in I_j) \overset{ind}{\sim} \mathrm{Dir}(\#I_j; \alpha_i, i \in I_j)$*, for* $j = 1, \ldots, m$*.*
(iii) $(Z_1, \ldots, Z_m)$ *and* $(X_i/Z_j : i \in I_j, j = 1, \ldots, m)$ *are independent.*

Conversely, if X *is a random vector such that (i)–(iii) hold, for a given partition* $I_1, \ldots, I_m$ *and* $Z_j = \sum_{i \in I_j} X_i$*, then* $X \sim \mathrm{Dir}(k; \alpha_1, \ldots, \alpha_k)$*.*

Proof In terms of the Gamma representation $X_i = Y_i/Y$ of Proposition 2.25 we have

$$Z_j = \frac{\sum_{i \in I_j} Y_i}{Y}, \qquad \text{and} \qquad \frac{X_i}{Z_j} = \frac{Y_i}{\sum_{i \in I_j} Y_i}.$$

[7] Under misspecification the posterior mean and the maximum likelihood estimator will be approximately normal with mean θ^* and covariance matrix $n^{-1} J K^{-1} J$, for $K = P(\dot{\ell}_{\theta^*} \dot{\ell}_{\theta^*}^T)$. This "sandwich" covariance matrix causes trouble for the usual interpretation of the AIC penalty, which is sometimes solved by replacing the penalty d by an estimate of $\mathrm{tr}(KJ^{-1})$, the so-called Takeuchi correction, TIC. See Claeskens and Hjort (2008), Sections 2.5 and 3.5. In the discussion following formula (3.16) they appear to suggest that in the misspecified case the DIC criterion is asymptotic to TIC, not AIC, but this is wrong?

Because $W_j := \sum_{i\in I_j} Y_i \overset{\text{ind}}{\sim} \text{Ga}(\beta_j, 1)$ for $j = 1, \ldots, m$, and $\sum_j W_j = Y$, the Dirichlet distributions in (i) and (ii) are immediate from Proposition 2.25. The independence in (ii) is immediate from the independence of the groups $(Y_i : i \in I_j)$, for $j = 1, \ldots, m$. By Proposition 2.25 W_j is independent of $(Y_i/W_j : i \in I_j)$, for every j, whence by the independence of the groups the variables $W_j, (Y_i/W_j : i \in I_j)$, for $j = 1, \ldots, m$, are jointly independent. Then (iii) follows, because $(X_i/Z_j : i \in I_j, j = 1, \ldots, m)$ is a function of $(Y_i/W_j : i \in I_j, j = 1, \ldots, m)$ and $(Z_1, \ldots, Z_m)$ is a function of $(W_j : j = 1, \ldots, m)$.

The converse also follows from the Gamma representation. □

Proposition 2.27 (Moments)**.** *If* $X \sim \text{Dir}(k; \alpha_1, \ldots, \alpha_k)$*, then* $X_i \sim \text{Be}(\alpha_i, |\alpha| - \alpha_i)$*, where* $|\alpha| = \sum_{i=1}^k \alpha_i$*. In particular,*

$$\text{E}(X_i) = \frac{\alpha_i}{|\alpha|}, \qquad \text{var}(X_i) = \frac{\alpha_i(|\alpha| - \alpha_i)}{(|\alpha|^2(|\alpha| + 1))}.$$

Furthermore, $\text{cov}(X_i, X_j) = -\alpha_i\alpha_j/(|\alpha|^2(|\alpha| + 1))$ *and, with* $r = r_1 + \cdots + r_k$*,*

$$\text{E}(X_1^{r_1} \cdots X_k^{r_k}) = \frac{\Gamma(\alpha_1 + r_1) \cdots \Gamma(\alpha_k + r_k)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_k)} \times \frac{\Gamma(|\alpha|)}{\Gamma(|\alpha| + r)}. \tag{2.9}$$

In particular, if $r_1, \ldots, r_k \in \mathbb{N}$*, then the expression in* (2.9) *is equal to* $\alpha_1^{[r_1]} \cdots \alpha_k^{[r_k]}/|\alpha|^{[r]}$*, where* $x^{[m]} = x(x + 1) \cdots (x + m - 1)$*,* $m \in \mathbb{N}$*, stands for the ascending factorial.*

Proof The first assertion follows from Proposition 2.26 by taking $m = 2$, $I_i = \{i\}$, $I_2 = I \setminus \{i\}$, for $I = \{1, \ldots, k\}$. Next the expressions for expectation and variance follow by the properties of the beta distribution.

For the second assertion, we take $m = 2$, $I_1 = \{i, j\}$ and $I_2 = I \setminus I_1$ in Proposition 2.26 to see that $X_i + X_j \sim \text{Be}(\alpha_i + \alpha_j, |\alpha| - \alpha_i - \alpha_j)$. This gives $\text{var}(X_i + X_j) = (\alpha_i + \alpha_j)(|\alpha| - \alpha_i - \alpha_j)/(|\alpha|^2(|\alpha| + 1))$, and allows to obtain the expression for the covariance from the identity $2\,\text{cov}(X_i, X_j) = \text{var}(X_i + X_j) - \text{var}(X_i) - \text{var}(X_j)$.

For the derivation of (2.9), observe that the mixed moment is the ratio of two Dirichlet forms with parameters $(\alpha_1 + r_1, \ldots, \alpha_k + r_k)$ and $(\alpha_1, \ldots, \alpha_k)$. □

2.8.2 Distances

Let P and Q be probability measures on the measurable space $(\mathfrak{X}, \mathscr{X})$, having densities p and q with respect to some σ-finite measure μ. Three measures of discrepancy between P and Q are

$$\begin{aligned} K(p; q) &= \int \log \frac{p}{q}\, dP, && \textit{Kullback-Leibler divergence},\\ \|P - Q\|_{TV} &= \sup_{B\in\mathscr{X}} \big|P(B) - Q(B)\big|, && \textit{total variation distance},\\ h(p, q) &= \Big(\int (\sqrt{p} - \sqrt{q})^2\, d\mu\Big)^{1/2}, && \textit{Hellinger distance}. \end{aligned}$$

In the definition of $K(p; q)$ the logarithm $\log(p/q)$ is understood to be ∞ if $p > 0 = q$; thus $K(p; q) = \infty$ if $P(q = 0) > 0$.

Lemma 2.28. *For any probability measures P and Q:*

(i) $h^2(p, q) \le K(p; q)$.
(ii) $\|P - Q\|_{TV} = 2\int |p - q|\, d\mu$.
(iii) $\|P - Q\|_{TV}^2 \le 16K(p; q)$.
(iv) $\|P - Q\|_{TV} \le 4h(p; q) \le 2\sqrt{2}\|P - Q\|_{TV}^{1/2}$.

Proof (i). Because $\log x \le 2(\sqrt{x}-1)$, for every $x \ge 0$, we have that $\int \log(q/p)\,dP \le 2\int(\sqrt{q/p}-1)\,dP = 2\int \sqrt{pq} - 2 = -h^2(p,q)$.

(ii). For any measurable set B we have $P(B) - Q(B) = \int_B (p-q)\,d\mu \le \int_{p>q}(p-q)\,d\mu$. Furthermore, since $\int (p-q)\,d\mu = 0$, we have $\int_{p>q}(p-q)\,d\mu = \int_{p<q}(q-p)\,d\mu$, while the sum of these integral is $\int |p-q|\,d\mu$.

(iv). The first inequality follows from (ii) and the Cauchy-Schwarz inequality applied to $\int |p-q|\,d\mu = \int |\sqrt{p}-\sqrt{q}|(\sqrt{p}+\sqrt{q})\,d\mu$. The second inequality follows from $(\sqrt{x}-\sqrt{y})^2 \le |x-y|$, for every $x, y \ge 0$.

(iii). This follows by combining (iv) and (i). (A preciser proof can improve on the constant 16.) □

2.8.3 Local asymptotic normality

Lemma 2.29. *If $(p_\theta\colon \theta \in \Theta \subset \mathbb{R}^d)$ are probability densities that satisfy (2.3), then $\int \dot\ell_\theta p_\theta\,d\mu = 0$ and $\int \|\dot\ell_\theta\|^2 p_\theta\,d\mu < \infty$.*

Proof Equation (2.3) entails that the difference $(p_{\theta+h}^{1/2} - p_\theta^{1/2})/h$ tends to $\frac{1}{2}\dot\ell_\theta p_\theta^{1/2}$ in $L_2(\mu)$, as $h \to 0$. This implies that $p_{\theta+h}^{1/2} - p_\theta^{1/2} \to 0$ and hence $p_{\theta+h}^{1/2} + p_\theta^{1/2} \to 2p_\theta^{1/2}$. Then

$$0 = \int \frac{p_{\theta+h}-p_\theta}{h}\,d\mu = \int \frac{p_{\theta+h}^{1/2}-p_\theta^{1/2}}{h}(p_{\theta+h}^{1/2}+p_\theta^{1/2})\,d\mu \to \int \tfrac{1}{2}\dot\ell_\theta p_\theta^{1/2}\,2p_\theta^{1/2}\,d\mu = \int \dot\ell_\theta p_\theta\,d\mu.$$

The second assertion is a consequence of the fact that $(p_{\theta+h}^{1/2} - p_\theta^{1/2})/h$ is in $L_2(\mu)$ for every h, which implies that also its limit $\frac{1}{2}\dot\ell_\theta p_\theta^{1/2}$ is contained in this space. □

Lemma 2.30. *If $(p_\theta\colon \theta \in \Theta \subset \mathbb{R}^d)$ are probability densities that satisfy (2.3), then, for $I_\theta = P_\theta(\dot\ell_\theta \dot\ell_\theta^T)$,*

$$\log \prod_{i=1}^n \frac{p_{\theta+h/\sqrt{n}}}{p_\theta}(X_i) = \frac{1}{\sqrt{n}}\sum_{i=1}^n h^T\dot\ell_\theta(X_i) - \tfrac{1}{2}h^T I_\theta h + o_{P_\theta^n}(1).$$

Proof See van der Vaart (1998), Theorem 7.2. □

Lemma 2.31. *If $\theta \mapsto p_\theta(x)$ is twice continuously differentiable, for every x, with derivatives $\dot\ell_\theta(x)$ and $\ddot\ell_\theta(x)$ whose norms are for every θ bounded by functions $x \mapsto L_1(x)$ and $x \mapsto L_2(x)$ with $P_\theta L_1^2 < \infty$ and $P_\theta L_2 < \infty$, then (2.3) holds and $-\int \ddot\ell_\theta p_\theta\,d\mu = P_\theta(\dot\ell_\theta\dot\ell_\theta^T)$.*

Proof For a proof (under weaker conditions) that (2.3) holds, see Lemma 7.6 in van der Vaart (1998). The equality can be proved by differentiating the identity $\int \dot\ell_\theta p_\theta\,d\mu = 0$, which itself follows from differentiating the identity $\int p_\theta\,d\mu = 1$, or from Lemma 2.29. □

Lemma 2.32. *If Δ_n are random vectors with $\Delta_n \rightsquigarrow N_d(0,J)$, then there exists $M_n \to \infty$ such that $\Delta_n 1_{\|\Delta_n\| \le M_n} - \Delta_n \to 0$ and, for every $C > 0$,*

$$\sup_{\|h\|<C} \left|\mathrm{E} e^{h^T\Delta_n 1_{\|\Delta_n\|\le M_n} - h^T J h/2} - 1\right| \to 0.$$

Proof The first property is automatic for any $M_n \to \infty$, since $\Pr(\|\Delta_n\| > M_n) \le \Pr(\|\Delta_n\| > M) \to \Pr(\|\Delta\| > M)$, for a Gaussian variable Δ, for any M. Because the limit becomes arbitrarily small if M is large enough, we can conclude that the probability that Δ_n and its truncation coincide tends to 1, for any $M_n \to \infty$.

To construct M_n with the second property define $a(M,C,n)$ to be the expression in the display with M_n replaced by M. Since the vectors $\Delta_n 1_{\|\Delta_n\|\le M}$ tend in distribution to the truncated Gaussian vector $\Delta 1_{\|\Delta\|\le M}$, for any fixed M as $n \to \infty$, the exponential functions $x \mapsto \exp(h^T x)$ are

uniformly bounded and uniformly equicontinuous on compact sets, it follows that $a(M, C, n) \to \sup_{\|h\|<C} |\mathrm{E} \exp(h^T \Delta 1_{\|\Delta\| \le M} - h^T J h/2) - 1|$, as $n \to \infty$ by the portmanteau lemma for weak convergence, for every fixed (M, C). Next, as $M \to \infty$ this tends to zero, for every fixed C. One can then show that there exist $M_n \to \infty$ and $C_n \to \infty$ such that $a(M_n, C_n, n) \to 0$. □

2.8.4 Tests

Lemma 2.33. *If there exist tests ϕ_n such that $\sup_{\theta\in\Theta_0} P_\theta^n \phi_n \to 0$ and $\sup_{\theta\in\Theta_1} P_\theta^n(1-\phi_n) \to 0$, for given fixed sets Θ_0 and Θ_1 and a given statistical model, then there exist tests ψ_n and $c > 0$ such that $\sup_{\theta\in\Theta_0} P_\theta^n \phi_n \le e^{-cn}$ and $\sup_{\theta\in\Theta_1} P_\theta^n(1-\phi_n) \le e^{-cn}$.*

Proof Fix k large enough such that $P_{\theta_0}^k \phi_k$ and $P_{\theta_1}^k(1-\phi_k)$ are smaller than $1/4$ for every $\theta_0 \in \Theta_0$ and $\theta_1 \in \Theta_1$. Let $n = mk + r$ for $0 \le r < k$, and define $Y_{n,1}, \ldots, Y_{n,m}$ as ϕ_k applied in turn to $X_1, \ldots, X_k$, to $X_{k+1}, \ldots, X_{2k}$, etcetera. Let $\bar{Y}_m$ be their average and then define $\psi_n = 1\{\bar{Y}_m \ge 1/2\}$. Since $\mathrm{E}_\theta Y_j \ge 3/4$ for every $\theta \in \Theta_1$ and every j, Hoeffding's inequality, Lemma 2.34, implies that

$$P_\theta^n(1-\psi_n) = \Pr_\theta(\bar{Y}_m \le 1/2) \le e^{-2m(\frac{1}{2}-\frac{3}{4})^2} \le e^{-m/8}.$$

Since m is proportional to n, this gives the desired exponential decay. Since $\mathrm{E}_\theta Y_j \le 1/4$, for every $\theta \in \Theta_0$, the expectations $P_\theta^n \psi_n$ are similarly bounded for $\theta \in \Theta_0$. □

Lemma 2.34 (Hoeffding)**.** *For any independent random variables $Y_1, \ldots, Y_n$ such that $a \le Y_i \le b$ for every i, and any $y > 0$,*

$$\mathrm{P}(\bar{Y}_n - \mathrm{E}\bar{Y}_n \ge y) \le e^{-2ny^2/(b-a)^2}.$$

Proof By Markov's inequality applied to the variable $e^{hn(\bar{Y}_n - \mathrm{E}\bar{Y}_n)}$, for $h > 0$ to be chosen later, we obtain

$$\mathrm{P}\Big(\sum_{i=1}^n (Y_i - \mathrm{E}Y_i) \ge y\Big) \le e^{-hny} \mathrm{E} \prod_{i=1}^n e^{h(Y_i - \mathrm{E}Y_i)}.$$

By independence of the Y_i the order of expectation and product on the right side can be swapped. By convexity of the exponential function $e^{hY} \le ((b-Y)e^{ha} + (Y-a)e^{hb})/(b-a)$ whenever $a \le Y \le b$, whence, by taking expectation,

$$\mathrm{E}e^{hY} \le e^{ha} \frac{b - \mathrm{E}Y}{b-a} + e^{hb} \frac{\mathrm{E}Y - a}{b-a} = e^{g(\xi)},$$

where $g(\xi) = \log(1 - p + pe^{\xi}) - p\xi$, for $\xi = (b-a)h$ and $p = (\mathrm{E}Y - a)/(b-a)$.

Now $g(0) = 0$, $g'(0) = 0$ and $g''(\xi) = (1-p)pe^{\xi}/(1-p+pe^{\xi})^2 \le \frac{1}{4}$ for all ξ, so that a second order Taylor's expansion gives $g(\xi) \le \xi^2/8$. Combining this with the preceding displays, we obtain, for any $h > 0$,

$$\mathrm{P}\Big(\sum_{i=1}^n (Y_i - \mathrm{E}Y_i) \ge y\Big) \le \exp(-hny + h^2 n(b-a)^2).$$

The result follows upon choosing $h = 4y/(b-a)^2$. □

Lemma 2.35. *Under the conditions of Theorem 2.16, there exists for every $M_n \to \infty$ a sequence of tests ϕ_n and a constant $c > 0$ such that, for every sufficiently large n and every $\|\theta - \theta_0\| \ge M_n/\sqrt{n}$,*

$$P_{\theta_0}^n \phi_n \to 0, \qquad P_\theta^n(1-\phi_n) \le e^{-cn\left(\|\theta-\theta_0\|^2 \wedge 1\right)}.$$

Proof We construct two sequences of tests, which "work" for the ranges $M_n/\sqrt{n} \le \|\theta-\theta_0\| \le \epsilon$ and $\|\theta-\theta_0\| > \epsilon$, respectively, and a given $\epsilon > 0$. Then the ϕ_n of the lemma can be defined as the maximum of the two sequences.

First consider the range $M_n/\sqrt{n} \le \|\theta-\theta_0\| \le \epsilon$. Let $\dot\ell^L_{\theta_0}$ be the score function truncated to the interval $[-L, L]$. By the dominated convergence theorem, $P_{\theta_0}\dot\ell^L_{\theta_0}\dot\ell^T_{\theta_0} \to I_{\theta_0}$ as $L \to \infty$. Hence, there exists $L > 0$ such that the matrix $P_{\theta_0}\dot\ell^L_{\theta_0}\dot\ell^T_{\theta_0}$ is nonsingular. Fix such an L and define

$$\omega_n = 1\{\big\|(\mathbb{P}_n - P_{\theta_0})\dot\ell^L_{\theta_0}\big\| \ge \sqrt{M_n/n}\}.$$

By the central limit theorem, $P^n_{\theta_0}\omega_n \to 0$, so that ω_n satisfies the first requirement. By the triangle inequality,

$$\big\|(\mathbb{P}_n - P_\theta)\dot\ell^L_{\theta_0}\big\| \ge \big\|(P_{\theta_0} - P_\theta)\dot\ell^L_{\theta_0}\big\| - \big\|(\mathbb{P}_n - P_{\theta_0})\dot\ell^L_{\theta_0}\big\|.$$

Since $P_\theta\dot\ell^L_{\theta_0} - P_{\theta_0}\dot\ell^L_{\theta_0} = (P_{\theta_0}\dot\ell^L_{\theta_0}\dot\ell^T_{\theta_0} + o(1))(\theta-\theta_0)$, by the differentiability (2.3) of the model, the first term on the right is bounded below by $c\|\theta-\theta_0\|$ for some $c > 0$, for every θ that is sufficiently close to θ_0, say for $\|\theta-\theta_0\| < \epsilon$ and a sufficiently small ϵ. If $\omega_n = 0$, then the second term is bounded below by $\sqrt{M_n/n}$. Consequently, for every $c\|\theta-\theta_0\| \ge 2\sqrt{M_n/n}$, and hence for every $\|\theta-\theta_0\| \ge M_n/\sqrt{n}$ and every sufficiently large n,

$$P^n_\theta(1-\omega_n) \le \Pr_\theta\Big(\big\|(\mathbb{P}_n - P_\theta)\dot\ell^L_{\theta_0}\big\| \ge \tfrac{1}{2}c\|\theta-\theta_0\|\Big) \le e^{-Cn\|\theta-\theta_0\|^2},$$

by Hoeffding's inequality, Lemma 2.34, for a sufficiently small constant C.

Next, consider the range $\|\theta-\theta_0\| > \epsilon$ for a fixed $\epsilon > 0$. By assumption (2.4) there exist uniformly consistent tests ϕ_n of θ_0 versus the complement of some compact neighborhood Θ_0 of θ_0. It can be shown that under assumption (2.3) and identifiability of the model, there automatically exists such a test for *any* neighborhood Θ_0 of θ_0, in particular for $\Theta_0 = \{\theta\colon \|\theta-\theta_0\| \le \epsilon\}$ (see van der Vaart (1998), pages 144-145.) By Lemma 2.33 applied with $\Theta_0 = \{\theta_0\}$ and $\Theta_1 = \{\theta\colon \|\theta-\theta_0\| > \epsilon\}$ there also exists tests with exponential error probabilities. □

2.8.5 Miscellaneous results

Lemma 2.36. *If (X_i, Y_i) are random vectors ($i = 1, 2$) such that the conditional distributions $X_i|\,Y_i = y$ and $Y_i|\,X_i = x$ possess regular versions that do not depend on i and are given by densities relative to σ-finite measures, then (X_1, Y_1) and (X_2, Y_2) are identically distributed.*

Proof If $p_i(x|\,y)$ and $p_i(y|\,x)$ are densities of the conditional distributions and $p_i(x)$ and $p_i(y)$ are marginal densities (with the usual abuse of notation that the arguments reveal which density is which), then by Bayes's formula $p_i(x|\,y)/p_i(y|\,x) = p_i(x)/p_i(y)$. By assumption the left side is the same for $i = 1, 2$. We conclude that $p_1(x)p_2(y) = p_2(x)p_1(y)$. Integrate with respect to y to see that $p_1(x) = p_2(x)$, or: $X_1 \sim X_2$. Together with the equality of the conditional distributions this gives equality of the joint distributions. □

A sequence of random variables Z_n is said to be *uniformly integrable* if $\sup_n \mathrm{E}|Z_n|1_{|Z_n|\ge M} \to 0$, as $M \to \infty$. This is weaker than *domination* of the variables ($|Z_n| \le Z$ for all n) by an integrable variable (and also does not require the variables to be defined on the same probability space), but is exactly what is needed for convergence in mean (i.e. $\mathrm{E}|Z_n - Z| \to 0$).

Lemma 2.37. *A sequence of integrable random variables Z_n converges in mean to a random variable Z if and only if the sequence Z_n is uniformly integrable and converges in probability to Z.*

Proof See e.g. Bauer (1981), Theorem 2.12.4. Alternatively, the lemma may be proved using the dominated convergence theorem. □

Lemma 2.38 (Scheffé's theorem)**.** *Let Z_n be a sequence of integrable random variables such that* $\limsup \mathrm{E}|Z_n| \le \mathrm{E}|Z|$ *for an integrable random variable Z.*

(i) *If all variables are defined on the same probability space and $Z_n \xrightarrow{P} Z$, then* $\mathrm{E}|Z_n - Z| \to 0$.
(ii) *If $Z_n \ge 0$ and $Z_n \rightsquigarrow Z$, then the sequence Z_n is uniformly integrable.*

Proof (i). Since $0 \le |Z_n|+|Z|-|Z_n-Z| \to 2|Z|$ in probability, Fatou's lemma gives that $\liminf \mathrm{E}(|Z_n|+|Z| - |Z_n - Z|) \ge 2\mathrm{E}|Z|$. The liminf is also bounded above by $2\mathrm{E}|Z| - \limsup \mathrm{E}|Z_n - Z|$, by assumption. By combining these inequalities we conclude that $\limsup \mathrm{E}|Z_n - Z| \le 0$.

(ii). Uniform integrability and the assumption $\limsup \mathrm{E}|Z_n| \le \mathrm{E}|Z|$ depend only on the marginal distributions of the Z_n and Z. Thus it suffices to prove the assertion for some sequence Z_n and Z with the same distributions. We may choose $Z_n = F_n^{-1}(U)$ and $Z = F^{-1}(U)$, for F_n and F the distribution functions of the original Z_n and Z, and U a uniform variable. This sequence satisfies the assumption of (i) and hence converges in mean. Then it is uniformly integrable by Lemma 2.37. □

Exercises

2.1 Find a conjugate prior family for $X_1, \ldots, X_n$ a sample from the Poisson distribution with parameter θ.

2.2 Derive the updating formula for the posterior distribution for the Gaussian location-scale family given in Example 2.3.

2.3 Show that the marginal density for Y in the regression model $Y|\beta, \sigma \sim N_n(X\beta, \sigma^2 I)$ with prior $1/\sigma^2 \sim \Gamma(a, b)$ and $\beta|\,\sigma \sim N_p(0, \sigma^2\Lambda)$ is given by

$$\frac{b^a\Gamma(n/2+a)}{(2\pi)^{n/2}\Gamma(a)} \frac{(b + \frac{1}{2}\|Y\|^2 - \frac{1}{2}Y^T X(X^T X + \Lambda^{-1})^{-1}X^T Y)^{-n/2+a}}{(\det(X^T X + \Lambda^{-1})\Lambda)^{1/2}}.$$

2.4 Specialize the formula in the preceding exercise to the special cases of

- the g-prior $\Lambda = g(X^T X)^{-1}$;
- the adapted g-prior in which the first column of X is the vector with entries only 1, the remaining columns are orthogonal to this column, and $\Lambda_{1,1} \to \infty$, $\Lambda_{>1,>1} = g(X_{>1}^T X_{>1})^{-1}$, and the remaining $\Lambda_{i,j}$ are zero.

2.5 Find the Jeffreys prior for the success parameter in a negative-binomial distribution with fixed number of required successes. [The likelihoods of the negative-binomial and binomial distributions for the same observation x are proportional, but the Jeffreys priors are different. Bayesians consider this a violation of the *likelihood principle* according to which proportional likelihoods should give the same inference.]

2.6 Compute the total variation norm between two univariate normal distributions with different means and unit variance.

2.7 Let $X_1, \ldots, X_n$ be the k-dimensional normal distribution with mean θ and covariance matrix the identity, and choose as prior $\Pi = N(0, \Lambda)$ for some nonsingular matrix Λ. Show by direct computation that the Bernstein-von Mises theorem is true in this case.

2.8 Suppose that θ is one-dimensional. Show, using Theorem 2.16, that the median M_n of the posterior distribution satisfies that $\sqrt{n}(M_n - \theta_0)$ converges in distribution to a $N(0, I_{\theta_0}^{-1})$-distribution.

2.9 Suppose that, in the case of a one-dimensional parameter, we use the loss function $\ell(h) = 1_{(-1,2)}(h)$. Find the limit distribution of the corresponding Bayes point estimator, assuming that the conditions of the Bernstein-von Mises theorem hold.

2.10 A credible ball around the posterior mean $\hat\theta_n$ takes the form $\{\theta\colon \|\theta - \hat\theta_n\| \le \hat r_n\}$, for $\hat r_n$ defined by $\Pi(\theta\colon \|\theta - \hat\theta_n\| \le \hat r_n) = 1 - \alpha$. Show that under the conditions of Theorem 2.16 the confidence

level of a credible ball tends to $1 - \alpha$, as $n \to \infty$. Assume that the prior has a finite second moment.

2.11 Suppose that the Bernstein-von Mises theorem holds for the one-dimensional parameter θ, with Fisher information equal to 1. Define the sets $\hat{C}_n$ as $(-\infty, \hat{\theta}_n + \xi_\alpha/\sqrt{n})$ if $\hat{\theta}_n < 0$ and as $(\hat{\theta}_n - \xi_\alpha/\sqrt{n}, \infty)$ if $\hat{\theta}_n \geq 0$. Show that:

(a) $\Pi(\hat{C}_n | X_1, \ldots, X_n) \to 1 - \alpha$.
(b) $P_0^n(0 \in \hat{C}_n) \to 1 - 2\alpha$.

Conclude that not *every* credible set is a confidence set of the same asymptotic level.

2.12 Derive the formulas for $(1/\sigma^2)| Y$ and $\beta| Y, \sigma$ in Example 2.4.

3

Bayesian Computation

Given an observation x of a variable with density p_θ and a prior density π, the posterior density $\pi(\theta|\, x)$ is proportional to

$$p_\theta(x)\pi(\theta).$$

In most cases it is easy to compute this number, because it is directly related to the specifications of model and prior. However, to compute a posterior mean or a posterior credible region we must also evaluate the proportionality constant: the integral of the preceding display relative to θ, for fixed x. Except in special cases, for instance a prior that is conjugate with respect to the statistical model, this could be cumbersome. The lack of analytical expressions for the posterior distribution in general cases has hampered the popularity of Bayes estimation for many years. It is simply not attractive to have to select a prior density on the basis that the computations are easy.

Stochastic approximation methods allow to overcome this difficulty. We focus on evaluating integrals of the type $\int_\Theta f(\theta)\, d\Pi(\theta|\, x)$, for some given measurable function $f\colon \Theta \mapsto \mathbb{R}$. The idea is to generate a sample of values $\theta_1, \theta_2, ..., \theta_n$ with the property

$$\frac{1}{n}\sum_{i=1}^{n} f(\theta_i) \overset{a.s.}{\to} \int_\Theta f(\theta)\, d\Pi(\theta|\, x).$$

By the law of large numbers this is certainly the case if the values $\theta_1, \theta_2, ...,$ are i.i.d. from the posterior distribution, but it is rarely practical to simulate such variables. Instead one simulates *dependent* values $\theta_1, \theta_2, ...,$, forming a Markov chain. *Markov Chain Monte Carlo* (MCMC) was developed in image analysis and statistical physics since the 1960s, and consists of simulating values from a Markov chain with equilibrium distribution approximately equal to the posterior distribution.

In this chapter we introduce the most popular MCMC algorithms, and investigate their approximation properties. We start with a review of Markov chains and their ergodic properties. Next we discuss a general MCMC algorithm, the Metropolis Hastings (MH) algorithm, with special cases the independent MH and the random walk MH. This algorithm is very generally applicable, but may have a slow rate of convergence, in particular to approximate complex distributions. More efficient methods can be constructed by exploiting specific properties of the target density. We discuss the slice sampler, the Gibbs sampler and Hamiltonian MCMC as examples. We also discuss the variational Bayes method, which cuts down on computational cost, by targetting an approximation to the posterior distribution. Simulating from mixture distributions, as arising in model selection problems, requires special

techniques, in particular if the components possess different dimensions. This is explained in the section on reversible jump MCMC.

In all cases our goal will be to simulate from a given target density or distribution. In the Bayesian setup this would be the posterior distribution, for fixed data, but in this chapter we use the notation π and Π for the target density and distribution from which we wish to simulate, not showing the dependence on the data. *Thus, in this chapter* Π *denotes the posterior, not the prior.* We also denote the underlying sample space by $\mathfrak{Y}$, with elements y rather than θ.

For more details on Markov chains we refer to Meyn and Tweedie (2012), and for a longer introduction to MCMC techniques, and other sampling algorithms, to, for instance, Robert and Casella (2004).

3.1 Brief introduction to Markov chains

A Markov chain is a sequence of random variables $Y_1, Y_2, \ldots$ with values in a measurable space $(\mathfrak{Y}, \mathscr{Y})$ with the *Markov property*: for every measurable set B and $y_1, y_2, \ldots, y_n \in \mathfrak{Y}$:

$$\Pr(Y_{n+1} \in B | Y_n = y_n, Y_{n-1} = y_{n-1}, \ldots, Y_1 = y_1) = \Pr(Y_{n+1} \in B | Y_n = y_n).$$

The right side of this equation is a Markov kernel in the sense of Definition 1.1, and is also called *transition kernel* in this context. The measurable space $(\mathfrak{Y}, \mathscr{Y})$ is the *state space* of the chain, which we shall assume to be a Polish space with Borel σ-field to be sure that conditional distributions are well defined.

In this chapter we consider only *time homogeneous Markov chains*, for which the transition kernel is independent of n. The evolution of the chain is then completely described by the starting value and the transition kernel Q, defined by, for $y \in \mathfrak{Y}$, $n \in \mathbb{N}$, and $B \in \mathscr{Y}$,

$$Q(y, B) = \Pr(Y_{n+1} \in B | Y_n = y).$$

This kernel may be given by a *transition density* q relative to a σ-finite measure μ on $(\mathfrak{Y}, \mathscr{Y})$ in that

$$Q(y, B) = \int_B q(y, z)\, d\mu(z), \qquad B \in \mathscr{Y}.$$

The function q is assumed to be jointly measurable in its two arguments, and $z \mapsto q(y, z)$ is a probability density, for every $\in \mathfrak{Y}$.[1]

Example 3.1. The AR(1) model $Y_n = \theta Y_{n-1} + \varepsilon_n$, with $\varepsilon_n \overset{iid}{\sim} N(0, \sigma^2)$ independent of a given starting value Y_0, is a time-homogeneous Markov chain with transition kernel $Q(y, B) = \Pr(Z + \theta y \in B)$, with $Z \sim N(0, \sigma^2)$.

A Markov chain with transition kernel Q is called *ψ-irreducible*, for a given nondegenerate measure ψ on the state space of the chain, if for every measurable set B with $\psi(B) > 0$ and every $y \in \mathfrak{Y}$ there exists a time point n such that $Q^n(y, B) > 0$. This means that from every possible starting point $y \in \mathfrak{Y}$, every set B of positive ψ-measure can be reached with

[1] The notation $q(z|y)$ would show that the latter function is a conditional density, but in this chapter we use the notation $q(y, z)$, which better expresses the order in a "transition" from y to z.

positive probability. The arbitrariness of the starting point means that the chain does not partition into separate pieces. The irreducibility measure ψ is not unique, but it can be shown (see Meyn and Tweedie (2012), Proposition 4.2.2) that there is a maximal one, in the sense that it dominates any other such measure in the sense of absolute continuity. In the following ψ will always refer to this maximal measure. [2]

Just having a positive probability of visiting a set is a weak sense of connection. For good long term behavior the chain must also visit sets often enough. We measure this by the count of the total number of visits to a measurable set B:

$$\eta_B = \sum_{n=1}^{\infty} 1_{Y_n \in B}.$$

The Markov chain (Y_n) is called *recurrent* if it is ψ-irreducible and $\mathrm{E}(\eta_B | Y_0 = y) = \infty$, for every set B with $\psi(B) > 0$ and every $y \in B$. The chain is called *Harris recurrent* if, moreover, $P(\eta_B = \infty | Y_0 = y) = 1$, for every set B with $\psi(B) > 0$ and every $y \in B$. Thus the chain is recurrent or Harris recurrent if the expected number of returns to a set B is infinite, or the number of returns is infinite almost surely, for every set B of positive ψ-measure, whenever the chain is started somewhere in B. Harris recurrence is of course (much) stronger than recurrence.

A probability distribution Π is called a *stationary distribution* for Q if, for every measurable set B,

$$\int_{\mathcal{Y}} Q(y, B)\, d\Pi(y) = \Pi(B).$$

This implies that, if $Y_1 \sim \Pi$, then $Y_2 \sim \Pi$, and so on. If Q allows a transition density q and Π a density π relative to some σ-finite measure μ, then this is equivalent to, for μ-almost every $z \in \mathcal{Y}$),

$$\int_{\mathcal{Y}} q(y, z)\, \pi(y)\, d\mu(y) = \pi(z).$$

A stationary, ψ-irreducible Markov chain is called *positive Harris recurrent* if it is Harris recurrent and its transition kernel possesses a stationary distribution.

We shall construct and generate Markov chains with a given stationary distribution, in order to estimate the latter distribution. Typically it will be impossible to generate the starting value Y_1 from the stationary distribution and the chain will not be stationary, but we still use the empirical distribution of $Y_1, \ldots, Y_n$ as an estimate of the stationary distribution Π. This is justified by the following version of the law of large numbers.

Theorem 3.2 (Ergodic theorem)**.** *If the Markov chain $Y_1, Y_2, \ldots$ is Harris recurrent and its transition kernel has stationary probability distribution Π, then for every Π-integrable f and $n \to \infty$,*

$$\frac{1}{n} \sum_{i=1}^{n} f(Y_i) \to \int f\, d\Pi, \qquad a.s..$$

[2] For this maximal irreducibility measure ψ, a set B with $\psi(B) = 0$ has the property that the set of y such that $Q^n(y, B) > 0$ for some n has ψ measure zero. Thus a set B with $\psi(B) = 0$ can be reached only from a null set of starting points. We note that in Meyn and Tweedie (2012) the notation ψ always refers to a maximal irreducibility measure, while the notation ϕ is used for a general measure.

Proof See Meyn and Tweedie (2012), Theorem 12.1.7. We note that "positive Harris recurrent" in this reference means "Harris recurrent and admits a stationary probability measure". □

A sequence of random variables $Y_1, Y_2, \ldots$ that satisfies the law of large numbers in the theorem (for every bounded measurable function f and some distribution Π) is called *ergodic*. The practical use of the theorem is that an average over $Y_1, Y_2, \ldots, Y_n$, for a large n, can be used to approximate a (posterior) mean. The theorem generalizes the ordinary strong law of large numbers, in that an i.i.d. sequence of variables $Y_1, Y_2, \ldots$ is a Harris recurrent Markov chain with its marginal law as stationary distribution. It can also be proved that the averages $n^{-1}\sum_{i=1}^{n} f(Y_i)$ of any strictly stationary sequence $Y_1, Y_2, \ldots$ of variables with $\mathrm{E}|f(Y_1)| < \infty$ tends almost surely to a limit (Birkhoff-Khinchin ergodic theorem), but in general the limit will be a random variable. In the special case that this variable is degenerate (for every f), the sequence is called ergodic. So a Harris recurrent Markov chain which has a stationary (probability) distribution is ergodic.

The averages may converge even if the individual variables Y_n are far from distributed according to the stationary distribution. However, even when the starting value Y_1 is not drawn from the stationary distribution, we may hope that the marginal distributions of the Y_n approach stationarity for large n. This "ergodicity of the Markov chain" is often true, but only under the additional assumption of aperiodicity, which prevents cycling of the chain between certain subsets of the state space in a periodic way. This somewhat involved concept is defined in terms of "small sets", as follows.

With $Q^1 = Q$ the kernel determining one transition, the n-transition kernel is given by the recursion formula

$$Q^n(y, B) = \int_{\mathfrak{Y}} Q^{n-1}(z, B)\, Q(y, dz).$$

By the Markov property it is easily derived that $Y_n | Y_0 = y \sim Q^n(y, \cdot)$, for every n. We wish to find conditions so that for large n the distributions $Q^n(y, \cdot)$ closely resemble the stationary distribution, for any starting point y of the chain. The Markov chain with transition kernel $(y, B) \mapsto Q(y, B)$ is called *ergodic* if

$$\lim_{n\to\infty} \left\| Q^n(y, \cdot) - \Pi(\cdot) \right\|_{TV} = 0, \quad \text{for all } y \in \mathfrak{Y}. \tag{3.1}$$

This property does not show how many iterations are needed to come close to the target distribution. In general this depends on the particular chain and the starting point, and is a complicated matter. Exponentially fast convergence is desirable. The Markov chain with transition kernel $(y, B) \mapsto Q(y, B)$ is called *geometrically ergodic* if there exists a constant $r > 1$ such that

$$\sum_{n=1}^{\infty} r^n \| Q^n(y, \cdot) - \Pi(\cdot) \|_{TV} < \infty, \quad \text{for all } y \in \mathfrak{Y}.$$

This implies that $\| Q^n(y, \cdot) - \Pi(\cdot) \|_{TV}$ tends to zero exponentially fast for every y, but the speed need not be uniform in y. The Markov chain is called *uniformly ergodic* if

$$\lim_{n\to\infty} \sup_{y\in\mathfrak{Y}} \left\| Q^n(y, \cdot) - \Pi(\cdot) \right\|_{TV} = 0.$$

This definition does not explicitly require a rate of convergence, but the theorem below shows that the speed is automatically exponentially fast, so that uniform ergodicity is stronger than geometric ergodicity.

An *atom* of the Markov chain (Y_n) is a measurable set $C \subset \mathcal{Y}$ such that there exists a measure ν with $Q(y, B) = \nu(B)$, for every $y \in C$ and $B \in \mathscr{Y}$. Thus from any starting point $y \in C$ the chain moves in an identical manner, to a point picked from the measure ν. An atom C in a ψ-irreducible chain is called *accessible* if $\psi(C) > 0$. Accessible atoms can be thought of as counterparts in a general Markov chain of the states/atoms of a Markov chain with countable state space. They are helpful to analyze long term behavior, as every time the chain enters an atom, it starts "from the beginning".

However, many continuous state space Markov chains do not have useful atoms, necessitating the following relaxation. A measurable set C is called a *small set* if there exist $n \in \mathbb{N}$, $\delta > 0$ and a probability measure ν such that

$$Q^n(y, B) \geq \delta \nu(B), \qquad \text{for all } y \in C, \quad B \in \mathscr{Y}. \tag{3.2}$$

For a ψ-irreducible Markov chain it can be shown (see Proposition 5.2.4 of Meyn and Tweedie (2012)) that it is no loss of generality to require that n, δ and ν in (3.2) be chosen such that $\nu(C) > 0$. Then the equation implies that $Q^n(y, C) \geq \delta \nu(C) > 0$, for every $y \in C$, so that the chain will return to the set C in n steps with positive probability. Next the way the chain leaves C may depend on the exact state in C, but with probability δ it moves independently from this state, according to ν. This turns out to be almost as good as an atom.[3]

The Markov chain with transition kernel Q is said to have *period* d if there exists a small set C with associated integer n and probability distribution ν with $\nu(C) > 0$ such that d is the greatest common divisor of the set

$$\left\{k \in \mathbb{N}\colon Q^k(y, B) \geq \delta_k \nu(B), \text{ for all } y \in C \text{ and } B \in \mathscr{Y}, \text{ for some } \delta_k > 0\right\}.$$

The latter greatest common divisor can be shown not to depend on the choice of the small set C and to correspond to a partition of the sample space such that the chain visits the partitioning sets in a given fixed order (Meyn and Tweedie (2012), Theorem 5.44). The chain is called *aperiodic* if its period is equal to one.

Theorem 3.3. *Any ψ-irreducible, Harris recurrent and aperiodic Markov chain is ergodic in the sense of* (3.1)*. A Markov chain is uniformly ergodic if and only if it is ψ-irreducible, aperiodic, and the full state space $\mathcal{Y}$ is a small set. In this case there exist numbers $R > 0$ and $\rho \in (0, 1)$ such that*

$$\left\|Q^n(y, \cdot) - \Pi(\cdot)\right\|_{TV} \leq R\rho^n, \quad \textit{for all } y \in \mathcal{Y}.$$

Proof For the first assertion see Theorem 13.0.1 in Meyn and Tweedie (2012). For the second, see Theorem 6.59 of Robert and Casella (2004). Or see Theorems 16.2.1 and 16.2.2 in Meyn and Tweedie (2012), together with Theorem 5.5.7. □

[3] Small sets abound. In fact, it can be shown (see again Proposition 5.2.4 of Meyn and Tweedie (2012)) that for a ψ-irreducible chain there always exists a countable collection of small sets C_i that covers the state space: $\mathcal{Y} = \bigcup_i C_i$. This collection of small sets presents an analogue to the single points in a countable state space.

Example 3.4. Suppose that the transition kernel $(y, B) \mapsto Q(y, B)$ has a density q such that the function

$$h(z) = \inf_{y \in \mathcal{Y}} q(y, z)$$

is measurable and strictly positive. Intuitively, this means that the chain may move from anywhere to anywhere in a single step. The assumption of positivity of the infimum is unnecessary strong, in particular for unbounded spaces, but the example is useful for illustration.

Then $Q(y, B) = \int_B q(y, z)\, d\mu(z) \geq \int_B h\, d\mu$, for every $B \in \mathscr{Y}$ and $y \in \mathcal{Y}$, which is positive whenever $\mu(B) > 0$. Thus the chain is ψ-irreducible for $\psi = \mu$. (This is a maximal irreducibility measure, because any irreducibility measure must be absolutely continuous with respect to $Q(y, \cdot)$, for any y.)

Also $Q(y, B) \geq \delta \nu(B)$, for every $y \in \mathcal{Y}$ and $B \in \mathscr{Y}$, where $\delta = \int h\, d\mu > 0$ and ν is the probability measure with density h/δ. It follows that the state space $\mathcal{Y}$ is a small set, with corresponding period 1.

By Theorem 3.3 the Markov chain is uniformly ergodic.

Although the ergodic theorem shows that averages over $Y_1, \ldots, Y_n$ converge to the target value, as $n \to \infty$, it is customary to throw away the earlier variables $Y_1, Y_2, \ldots, Y_m$ and use only $Y_{m+1}, Y_{m+2}, \ldots$ for a sufficiently large m. This is called the "burn-in" of the "MCMC sampler", and is justified if the chain is ergodic, since the laws $Q^m(Y_0, \cdot)$ will typically be close(r) to the target for larger m.

Detailed balance and reversibility

A transition density q is said to satisfy the *detailed balance* relationship relative to a density π if for every $y, z \in \mathcal{Y}$:

$$\pi(y) q(y, z) = \pi(z) q(z, y). \tag{3.3}$$

The following lemma shows that the density π is necessarily a stationary density. In the next section we shall see how the detailed balance relation may be used to construct a transition density q that corresponds to a given stationary density π.

Lemma 3.5 (Detailed balance)**.** *If the probability density π and transition density q satisfy the detailed balance relationship (3.3), then π is a stationary density for the transition kernel with density q.*

Proof This is immediate from integrating (3.3) with respect to y and using the identity $\int q(z, y)\, d\mu(y) = 1$, on the right side. □

The detailed balance relationship expresses that starting the chain at y and then moving to z is "equally likely" as starting in z and then moving to y. Intuitively, the dynamics of the chain remain the same if the direction is reversed. Because $(y, z) \mapsto \pi(y) q(y, z)$ is the joint density of two consecutive variables of the (stationary) chain, the detailed balance relationship is equivalent to $(Y_n, Y_{n+1}) \sim (Y_{n+1}, Y_n)$, for every n. A Markov chain with this property is called *reversible*. Thus the stationary Markov chains that satisfy the detailed balance condition are precisely the reversible Markov chains.[4]

[4] In general a Markov chain (Y_n) remains a Markov chain if run in the opposite time direction: the sequence of

Equation (3.3) is the key to constructing a transition density q for which a given target density π is a stationary density. To verify that a given (reversible) chain has a given stationary density or distribution, the following (very) trivial observation is useful.

Lemma 3.6. *For a given measure* Π *and transition kernel* Q*, let* $Y_n \sim \Pi$ *and* $Y_{n+1}|\, Y_n = y \sim Q(y, \cdot)$*. If the probability* $\Pr(Y_n \in A, Y_{n+1} \in B)$ *is invariant under swapping* A *and* B*, for every pair of measurable sets* A *and* B*, then* Π *is a stationary measure of* Q*.*

Proof The assumption on the probabilities is equivalent to $(Y_n, Y_{n+1}) \sim (Y_{n+1}, Y_n)$. By marginalization to the first marginal, we see that $Y_n \sim Y_{n+1}$, and this common distribution is Π by the assumption $Y_n \sim \Pi$. □

3.2 Metropolis-Hastings

Let q be a transition density such that it is easy to draw a sample from the density $z \mapsto q(y, z)$, for given y. We further require that we know q up to a multiplicative constant, or that it is symmetric, i.e. $q(y, z) = q(z, y)$, for every y and z. We define the *Metropolis-Hastings acceptance probability* as

$$\alpha(y, z) = \frac{\pi(z)q(z, y)}{\pi(y)q(y, z)} \wedge 1.$$

To calculate $\alpha(y, z)$ it suffices to know π and q up to multiplicative constants (in case of a symmetric q even this is not necessary). This is essential for Bayesian computation, as the norming constant of the posterior distribution poses the main challenge.

Given a starting value Y_0 we proceed recursively, for $n = 0, 1, 2, \ldots$, as follows:

Given Y_n.
Generate Z_{n+1} from the distribution with density $q(Y_n, \cdot)$.
Generate U_{n+1} from the uniform distribution on $[0, 1]$.
If $U_{n+1} < \alpha(Y_n, Z_{n+1})$, set $Y_{n+1} := Z_{n+1}$,
else set $Y_{n+1} := Y_n$.

The acceptance probability $\alpha(y, z)$ is defined only if $\pi(y) > 0$, but if the starting value Y_0 is chosen to satisfy $\pi(Y_0) > 0$, then $\pi(Y_n) > 0$, for every n, since moves to states z with $\pi(z) = 0$ are rejected, as $\alpha(y, z) = 0$ if $\pi(z) = 0$. Similarly the chain will with probability one never reach states such that $q(Y_n, Z_{n+1}) = 0$.

By construction (Y_n) is a Markov chain. Its transition kernel is given in the next theorem. It is not absolutely continuous, but consists of two parts, which are typically orthogonal: either the test involving U_{n+1} is rejected and the chain does not move, or the test is accepted and the proposal Z_{n+1} becomes the new state of the chain. Staying put is the same as a move

variables (Z_n) defined by $Z_n = Y_{-n}$ is a Markov chain. This follows for instance from the fact that the Markov property can also be formulated in the symmetric form: "past and future are independent given the present". The chain is reversible if and only if the transition distributions in the two directions are identical. Indeed $Y_{n+1}|\, Y_n = y \sim Y_n|\, Y_{n+1} = y$, for every y, implies that $(Y_n, Y_{n+1}) \sim (Y_{n+1}, Y_n)$, by Lemma 2.36, applied to the special case that (X_1, Y_1) is (Y_n, Y_{n+1}) and (X_2, Y_2) is (Y_{n+1}, Y_n).

from y to y and intuitively this is clearly reversible. The other part moves according to the density $\alpha(y,z)q(y,z)$. The acceptance probability $\alpha,(y,z)$ is chosen such that

$$\pi(y)\alpha(y,z)q(y,z) = \pi(z)\alpha(z,y)q(z,y). \tag{3.4}$$

(To see this, split out in the two cases that $\pi(z)q(z,y)$ is bigger or smaller than $\pi(y)q(y,z)$, respectively, when $\alpha(y,z)$ is equal to 1 or smaller than 1, respectively, and conversely for $\alpha(z,y)$.) This is a detailed balance condition relative to π and together with Lemma 3.5 suggests that π is indeed a stationary distribution.

The following theorem makes this precise, under the assumption that the kernel is positive (which could be relaxed to the kernel $z \mapsto q(y,z)$ to contain the support of π).

Theorem 3.7. *The transition kernel P of the sequence (Y_n) produced by the Metropolis-Hastings algorithm with proposal density $z \mapsto q(y,z)$ is given by, for δ_y the Dirac measure at y,*

$$P(y,B) = \int_B \alpha(y,z)q(y,z)\,d\mu(z) + \int (1-\alpha(y,z))q(y,z)\,d\mu(z)\,\delta_y(B). \tag{3.5}$$

If $q(y,z) > 0$ for every $(y,z) \in \mathfrak{Y}\times\mathfrak{Y}$, then π is a stationary density of the chain, and for any π-integrable function f

$$\frac{1}{n}\sum_{i=1}^{n} f(Y_n) \xrightarrow{P} \int f\pi\,d\mu.$$

If (Y_n) is also aperiodic, then also

$$\left\|P^n(y,\cdot) - \Pi(\cdot)\right\|_{TV} \to 0, \qquad \forall y \in \mathfrak{Y}.$$

Proof By the tower rule for conditional expectation, for every measurable set B and y,

$$\begin{aligned}\Pr(Y_{n+1}\in B|\,Y_n = y) &= \mathrm{E}[\Pr(Y_{n+1}\in B|\,Y_n = y, Z_{n+1})|\,Y_n = y] \\ &= \mathrm{E}[1_B(Z_{n+1})\alpha(Y_n,Z_{n+1}) + 1_B(Y_n)(1-\alpha(Y_n,Z_{n+1}))|\,Y_n = y].\end{aligned}$$

Since $Z_{n+1}|\,Y_n = y \sim q(y,\cdot)$, this reduces to to the expression given in (3.5). The number $t(y) = \int(1-\alpha(y,z))q(y,z)\,d\mu(z)$ is the probability that the chain does not move.

To prove that π is a stationary density, note that

$$\begin{aligned}\int P(y,B)\,d\Pi(y) &= \int\int_B \alpha(y,z)q(y,z)\pi(y)\,d\mu(z)\,d\mu(y) + \int_B\int (1-\alpha(y,z))q(y,z)\,d\mu(z)\,d\Pi(y) \\ &= \int\int_B \alpha(z,y)q(z,y)\pi(z)\,d\mu(z)\,d\mu(y) + \int_B\int (1-\alpha(z,y))q(z,y)\,d\mu(y)\,d\Pi(z) \\ &= \int\int_B q(z,y)\pi(z)\,d\mu(z)\,d\mu(y) = \int_B \pi(z)\,d\mu(z).\end{aligned}$$

In the second equality we use the detailed balanced condition (3.4) in the first integral, and simply swap the notations y and z in the second integral.

Alternatively, and perhaps easier, we can express a joint probability of a chain started

from $Y_n \sim \pi$, by

$$\Pr(Y_n \in A, Y_{n+1} \in B) = \int\int 1_A(y)\pi(y)q(y,z)\alpha(y,z)1_B(z)\,d\mu(y)\,d\mu(z) + \int\int 1_A(y)\pi(y)q(y,z)(1-\alpha(y,z))1_B(y)\,d\mu(y)\,d\mu(z).$$

The second integral depends on (A, B) only through $A \cap B$ and hence is symmetric in A and B. The first integral is invariant under swapping A and B if the integrand $\pi(y)q(y,z)\alpha(y,z)$ is symmetric in y and z. This is the detailed balance condition (3.4), which is ensured by the definition of α.

The convergence follows by combining Theorem 6.51 and Lemma 7.3 of Robert and Casella (2004). □

The performance of the Metropolis-Hastings algorithm depends heavily on the proposal kernel q. Choosing a good kernel is somewhat of an art. The general principle is to choose a kernel that proposes variables Z_{n+1} throughout the support of π (is "sufficiently mixing" or "sufficiently explores the space"; possibly only when taking multiple steps together), and at the same time does not run into the "else" step too often, for efficiency reasons.

The two most popular choices are the independent- and the random walk MH algorithms.

3.2.1 Independent Metropolis-Hastings

In the *independent Metropolis-Hastings* algorithm the proposals are independent of the current states. For a given probability density g on $\mathcal{Y}$, the transition density is chosen equal to $q(y,z) = g(z)$, giving the acceptance probability

$$\alpha(y,z) = \frac{\pi(z)g(y)}{\pi(y)g(z)} \wedge 1$$

The algorithm reduces to (only the first step changes):

```
Given Y_n.
Generate Z_{n+1} from g.
Generate U_{n+1} from the uniform distribution on [0,1].
If U_{n+1} < α(Y_n, Z_{n+1}), set Y_{n+1} := Z_{n+1},
else  set Y_{n+1} := Y_n.
```

Example 3.8. Consider estimating the expected value of the variable $f(Y) = \sqrt{|Y|}$, where the random variable Y follows the density π satisfying $\pi(y) \propto e^{-y^4}$. Based on the shape of π the standard normal distribution seems reasonable as a transition kernel. Given starting value $y_0 = 0$ we iterate the following steps:

- independently draw $z_{n+1} \sim N(0,1)$ and $u_{n+1} \sim U(0,1)$,
- set $y_{n+1} := z_{n+1}$ if $u_{n+1} < e^{y_n^4 + z_{n+1}^2/2 - z_{n+1}^4 - y_n^2/2}$, or else set $y_{n+1} := y_n$.

After n iterations we compute the estimator $n^{-1}\sum_{i=1}^n \sqrt{|y_i|}$ of $\mathrm{E}\sqrt{|Y|}$.

For appropriate choice of g one can prove uniform ergodicity for the Markov chain (Y_n).

Theorem 3.9. *If there exists a constant M with $\pi(y) \leq Mg(y)$, for every $y \in \mathcal{Y}$, then the Markov chain (Y_n) generated by the independent Metropolis-Hastings algorithm is uniformly ergodic. More specifically*

$$\|P^n(y,\cdot) - \Pi(\cdot)\|_{TV} \leq (1 - 1/M)^n, \qquad \text{for all } y \in \mathcal{Y}.$$

Proof See Theorem 2.1 of Mengersen and Tweedie (1996). □

The speed of convergence of the IMH-algorithm depends on the choice of the proposal density g. If this resembles the target density π closely, then the proposals Z_{n+1} are nearly from the target distribution, and the acceptance probabilities $\alpha(Y_n, Z_{n+1})$ will be large. The extreme case would be to choose g equal to π, when the acceptance probability becomes 1. A more feasible approach for finding a suitable g is to start with a family of proposal densities, parametrized by some hyper-parameter τ, and next choose τ that maximizes the acceptance rate. This rate can be empirically evaluated by running the chain for a given number of iterations with a given value of τ.

The IMH-sampler works well if the proposal density resembles the target density globally, but this may be difficult to achieve especially in high-dimensional state spaces. Global resemblance is necessary, because the proposals are generated independently of the past values of the chain, but must explore the target density. In the following section we consider the alternative of drawing proposals more locally, from neighborhoods of the current state. The long run of the chain may then take care of the exploration of the state space.

3.2.2 Random walk Metropolis-Hastings

In the *random walk Metropolis-Hastings* algorithm the transition kernel takes the form $q(y,z) = f(z-y)$, for a probability density f on the state space (which is assumed to be a vector space). The name comes from the representation $Z_{n+1} = Y_n + \varepsilon_{n+1}$, where $\varepsilon_{n+1} \sim f$ is the proposed step. For state space the real line, common choices for the density f are the normal, student or symmetric uniform distribution.

When f is symmetric around zero, the acceptance probability $\alpha(y,z)$ reduces to $\pi(z)/\pi(y)$, simplifying the algorithm, which becomes:

```
Given Y_n.
Generate Z_{n+1} from f(· − Y_n).
Generate U_{n+1} from the uniform distribution on [0, 1].
If U_{n+1} < π(Z_{n+1})/π(Y_n), set Y_{n+1} := Z_{n+1},
else  set Y_{n+1} := Y_n.
```

Example 3.10. Consider simulating from the probability density satisfying

$$\pi(y) \propto y^2 e^{-(y-1)^2} + e^{-|y|}.$$

We may use a random walk Metropolis-Hastings algorithm with proposal density f the normal density with mean zero and variance τ^2. The scale parameter τ can be used to tune

the step sizes of the random walk. Starting from some value y_0, in each iteration step $n \geq 0$ we draw $z_{n+1} \sim N(y_n, \tau^2)$ and $u_{n+1} \sim U(0,1)$, and next set

$$y_{n+1} := \begin{cases} z_{n+1}, & \text{if } u_{n+1} \leq v_{n+1}, \\ y_n, & \text{if } u_{n+1} > v_{n+1}, \end{cases} \qquad v_{n+1} = \frac{z_{n+1}^2 e^{-(z_{n+1}-1)^2} + e^{-|z_{n+1}|}}{y_n^2 e^{-(y_n-1)^2} + e^{-|y_n|}}.$$

The chain $y_m, y_{m+1}, \ldots$ remaining after a sufficiently long burn-in can be used as an approximate sample from π.

Theorem 3.11. *If the target density π is log-concave, or more generally $\log \pi(x) - \log \pi(y) \geq \alpha|y - x|$, for every $M \leq |x| \leq |y|$ and a sufficiently large M, then the Markov chain produced by the random walk Metropolis-Hastings algorithm with kernel f is geometrically ergodic whenever f is positive, symmetric, and has exponentially small tails (in case π is non-symmetric).*

The choice of the random walk kernel strongly influences the speed of convergence. In contrast to the independent Metropolis-Hastings algorithm, it is not optimal to maximize the acceptance rate. Accepting too many moves might mean that the steps of the random walk are too small and the state space is not explored appropriately, resulting in slow convergence of the Markov chain. As always the acceptance rate should not be too small either, as rejections mean wasted simulations.

Thus the step size of the random walk should neither be too small, nor too large. As this is relative to π, again choice of a good proposal distribution is a bit of an art.

3.2.3 Metropolis-adjusted Langevin

The *Langevin diffusion* with drift function $\nabla \log \pi$ is the stochastic process $(X_t : t \geq 0)$ on $\mathfrak{Y} = \mathbb{R}^d$ following the stochastic differential equation, for W_t a d-dimensional Brownian motion,

$$dX_t = \nabla \log \pi(X_t)\, dt + \sqrt{2}\, dW_t.$$

It is known that, as $t \to \infty$, the distribution of X_t will approach the distribution with density π whatever the initial value X_0, and if $X_0 \sim \pi$, then X_t will have density π for every $t > 0$. (In other words π is a stationary density for the diffusion.) Thus we could approximately draw from π by running the Langevin process for a sufficiently large time.

In practice, the stochastic differential equation cannot be solved in continuous time, and its solution is approximated by taking discrete time steps. The *Euler-Maruyama scheme* simply computes, for a given time step ϵ, the sequence of variables $X_0, X_\epsilon, X_{2\epsilon}, \ldots$ by the recursions

$$X_{i\epsilon+\epsilon} = X_{i\epsilon} + \epsilon \nabla \log \pi(X_{i\epsilon}) + \sqrt{2}(W_{i\epsilon+\epsilon} - W_{i\epsilon}).$$

Here the increments $W_{i\epsilon+\epsilon} - W_{i\epsilon}$ of Brownian motion are normal vectors with mean zero and covariance matrix ϵ times the identity matrix I. As $\epsilon \downarrow 0$, the variables will resemble the sampled Langevin diffusion more and more.

The *Metropolis-adjusted Langevin algorithm* or *MALA* turns this idea into a sampler by combining it with a Metropolis-Hastings step. It uses the discrete time steps to generate proposals and next accepts or rejects these by ordinary Metropolis-Hastings, thus correcting

for the discretization error. For a given time step ϵ, the proposal distribution is taken normal with mean $y + \epsilon\nabla \log \pi(y)$ and covariance matrix $2\epsilon I$, leading to the acceptance probability

$$\alpha(y, z) = \frac{\pi(z) \exp(-\|y - z - \epsilon\nabla \log \pi(z)\|^2/(4\epsilon))}{\pi(y) \exp(-\|z - y - \epsilon\nabla \log \pi(y)\|^2/(4\epsilon))} \wedge 1.$$

A potential advantage is that the proposals involve the gradient of π, and not just the values of the function π itself. They should be better guided by the target density, leading to fewer rejections even if the moves are larger. The value of ϵ must of course be chosen appropriately.

The algorithm becomes as follows.

Given Y_n.
Generate Z_{n+1} from the normal distribution with mean $Y_n + \epsilon\nabla \log \pi(Y_n)$ and covariance matrix $2\epsilon I$.
Generate U_{n+1} from the uniform distribution on $[0, 1]$.
If $U_{n+1} < \alpha(Y_n, Z_{n+1})$, set $Y_{n+1} := Z_{n+1}$,
else set $Y_{n+1} := Y_n$.

For results on ergodicity of the Langevin chain, see Roberts and Tweedie (1996).

3.3 Slice sampling

The slice sampling algorithm does not use a proposal density, but is defined directly in terms of the target density. It relies on the fact that making a draw from a density π on $\mathbb{R}^d$ is equivalent to drawing a point from the uniform distribution on the *subgraph of* π, defined by

$$\mathcal{F}(\pi) = \{(y, u) \in \mathbb{R}^d \times \mathbb{R}^+ \colon 0 \le u \le \pi(y)\}.$$

Indeed, if (Y, U) is distributed according to the uniform distribution on $\mathcal{F}(\pi)$, then the marginal density of Y is given by π:

$$\Pr(Y \in B) = \frac{\lambda((y, u) \colon y \in B, 0 \le u \le \pi(y))}{\lambda((y, u) \colon 0 \le u \le \pi(y))} \propto \int_B \int_0^{\pi(y)} du\, d\lambda(y) = \int_B \pi(y)\, d\lambda(y).$$

Here λ is the Lebesgue measure. We may thus construct a (nearly) uniform Markov chain (Y_n, U_n) on the subgraph, next throw away the uniform variables, and obtain a sample (Y_n) from π. The same strategy works using the subgraph $\mathcal{F}(c\pi)$ of a fixed multiple $c\pi$ of π, and hence can also be used if π is only known up to a constant.

The *slice sampler* constructs a Markov chain on $\mathcal{F}(c\pi)$ by repeated consecutive, uniformly distributed moves along the y- and z-axis:

Given Y_n.
Generate U_{n+1} from $U(0, c\pi(Y_n))$.
Generate Y_{n+1} from $U(B_{n+1})$, where $B_{n+1} = \{y \colon c\pi(y) \ge U_{n+1}\}$.

We note that the set B_{n+1} is never empty, since Y_n is inside. An advantage of the algorithm is that it does not have an accept/reject step, all draws are accepted.

Theorem 3.12. *The sequence (Y_n, U_n) produced by the slice sampler is a Markov chain with stationary distribution equal to the uniform distribution on the subgraph $\mathcal{F}(\pi^*)$ and hence the sequence (Y_n) is a Markov chain with stationary density π. Furthermore, if π is bounded and has bounded support in $\mathbb{R}$, then the slice sampler is uniformly ergodic.*

Proof From the description of the algorithm it is seen that the conditional density of (Y_{n+1}, U_{n+1}) given $(Y_n = y, U_n)$ is given by

$$(y_{n+1}, u_{n+1}) \mapsto \frac{I_{\pi^*(y_{n+1}) \geq u_{n+1}}}{\lambda(y \colon \pi^*(y) \geq u_{n+1})} \frac{I_{0 \leq u_{n+1} \leq \pi^*(y)}}{\pi^*(y)}.$$

Note that this does not depend on the value of U_n. If (Y_n, U_n) is uniformly distributed over $\mathcal{F}(\pi^*)$, then Y_n has marginal density π. In that case the density of (Y_{n+1}, U_{n+1}) is given by

$$\int \frac{I_{\pi^*(y_{n+1}) \geq u_{n+1}}}{\lambda(y \colon \pi^*(y) \geq u_{n+1})} \frac{I_{0 \leq u_{n+1} \leq \pi^*(y)}}{\pi^*(y)} \pi(y)\, dy = \frac{1}{c} I_{\pi^*(y_{n+1}) \geq u_{n+1}},$$

for c the proportionality constant so that $\pi^* = c\pi$. This shows that (Y_{n+1}, U_{n+1}) is uniformly distributed over $\mathcal{F}(\pi^*)$.

Since the conditional distribution of (Y_{n+1}, U_{n+1}) given (Y_n, U_n) depends only on Y_n, the Markov property of the joint chain $((Y_n, U_n))$ gives that

$$\Pr((Y_{n+1}, U_{n+1}) \in B | (Y_n, U_n), \ldots, (Y_1, U_1)) = \Pr((Y_{n+1}, U_{n+1}) \in B | Y_n).$$

Taking the conditional expectation given $Y_n, \ldots, Y_1$ across this identity, shows that (Y_n) is a Markov chain. It was already noted that the first marginal density of the uniform density on $\mathcal{F}(\pi^*)$ is π. This proves the second assertion.

For the final assertion of the theorem, see Lemma 8.5 of Robert and Casella (2004). □

Example 3.13. Consider the density $\pi(y) \propto e^{-y^4}$. From an arbitrary starting value, for instance, $y_0 = 0$, the slice sampler repeats the steps $u_{n+1} \sim U(0, e^{-y_n^4})$ and $y_{n+1} \sim U(-(-\log u_{n+1})^{1/4}, (-\log u_{n+1})^{1/4})$, for $n = 0, 1, 2, \ldots$. The resulting values $y_1, y_2, \ldots$ will, after a burn-in period, be approximate draws from π.

The difficulty of implementing the slice sampling algorithm is to determine the sets B_{n+1}. These sets may have a complicated form, especially for multi-modal distributions. This problem can be alleviated by introducing multiple slices.

Assume that the density π can be factorized into nonnegative functions $\pi_1, \ldots, \pi_k$ as

$$\pi(y) \propto \prod_{i=1}^{k} \pi_i(y).$$

The quantities $\pi_i(y)$ can be considered "new dimensions", which can be consecutively explored by a random walk, one direction at a time. The *generalized slice sampling* algorithm takes the form:

Given Y_n,
generate $U_{n+1}^{(1)}$ from $U(0, \pi_1(Y_n))$,
generate $U_{n+1}^{(2)}$ from $U(0, \pi_2(Y_n))$,
$\vdots$

generate $U_{n+1}^{(k)}$ from $U(0, \pi_k(Y_n))$,
generate Y_{n+1} from $U(B_{n+1})$, where $B_{n+1} = \bigcap_{i=1}^{k} \{y : \pi_i(y) \geq U_{n+1}^{(i)}\}$.

The decomposition, although seemingly slowing down the algorithm, may make it easier to compute the sets B_{n+1}, as the intersection of the sets $B_{n+1}^{(i)} = \{y : \pi_i(y) \geq U_{n+1}^{(i)}\}$.

Example 3.14. Consider sampling from the density $\pi(y) \propto \cos(y)^2 y^{-2} e^{-2|y|}$. One factorization of the function is as the product of the three function $\pi_1(y) = \cos(y)^2$, $\pi_2(y) = y^{-2}$, and $\pi_3(y) = e^{-2|y|}$. Given a starting point, for instance $y_0 = 0$, we iterate the general slice sampling algorithm for a large number of times. by drawing $u_{n+1,1} \sim U(0, \cos^2 y_n)$, $u_{n+1,2} \sim U(0, y_n^2)$, and $u_{n+1,3} \sim U(0, e^{-2|y_n|})$, and finally sample y_{n+1} uniformly from the set

$$\{y : |y| \leq \sqrt{u_{n+1,1}}\} \cap \{y : \arccos(\sqrt{u_{n+1,2}}) \leq y - k\pi \leq \arccos(-\sqrt{u_{n+1,2}}), k \in \mathbb{Z}\} \cap \{y : |y| \leq -\log(u_{n+1,3})/2\}.$$

3.4 Gibbs sampler

The Gibbs sampler reduces the problem of simulating a vector $(y_1, \ldots, y_m)$ in a product space $\mathcal{Y} = \mathcal{Y}_1 \times \cdots \times \mathcal{Y}_m$ to simulating from lower-dimensional distributions. Suppose that π is a density relative to a product measure $\mu_1 \times \cdots \times \mu_m$ on a product space, and suppose that we can generate variables from each of the conditional densities,

$$\pi_i(y_i | y_1, \ldots, y_{i-1}, y_{i+1}, \ldots y_m) = \frac{\pi(y)}{\int \pi(y)\, d\mu_i(y_i)}.$$

The corresponding distributions are known as the *full conditional distributions*, and it is of course helpful if these can be characterized analytically.

The *Gibbs sampling* algorithm proceeds, from a given initial value $Y_0 = (Y_{0,1}, \ldots, Y_{0,m})$, recursively for $n = 0, 1, 2, \ldots$ by the following steps:

Given $Y_n = (Y_{n,1}, \ldots, Y_{n,m})$.
Generate $Y_{n+1,1}$ from $\pi_1(\cdot | Y_{n,2}, \ldots, Y_{n,m})$.
Generate $Y_{n+1,2}$ from $\pi_2(\cdot | Y_{n+1,1}, Y_{n,3} \ldots, Y_{n,m})$.
$\vdots$
Generate $Y_{n+1,m}$ from $\pi_m(\cdot | Y_{n+1,1}, \ldots, Y_{n+1,m-1})$.

Thus we update each of the coordinates in turn, every time conditioning on the last available value of the other coordinates.

A variation is to update not in the preceding deterministic order, but to choose the coordinate to be updated randomly with respect to some distribution.

Example 3.15. The Ising model describes the distribution of a $(d \times d)$-matrix σ, whose coordinates σ_i, for $i \in \{1, \ldots, d\}^2$, are random variables with values in $\{-1, 1\}$. The variables represent the magnetic dipole moments of atomic spins in a geometric configuration given by the matrix. The energy function of a configuration $\sigma \in \{-1, 1\}^{d \times d}$ is given as

$$H(\sigma) = -J \sum_{i \sim j} \sigma_i \sigma_j - \mu \sum_i \sigma_i,$$

where $i \sim j$ denotes that the variables σ_i and σ_j are neighbours in the matrix configuration. The probability mass function of the random matrix σ is given as

$$\pi(\sigma) \propto e^{-H(\sigma)}, \qquad \sigma \in \{-1, 1\}^{d \times d}.$$

The conditional probability mass function of the variable σ_i given all the other variables is

$$\pi(\sigma_i | \, \sigma_j \colon j \neq i) = \frac{e^{-J \sum_{j:j\sim i} \sigma_i \sigma_j - \mu \sigma_i}}{e^{-J \sum_{j:j\sim i} \sigma_j - \mu} + e^{J \sum_{j:j\sim i} \sigma_j + \mu}}, \qquad \sigma_i \in \{-1, 1\}.$$

A Gibbs sampler for the Ising-model can be implemented by simulating iteratively from these two-point distributions.

Example 3.16. The joint distribution of the pair of variables (θ, τ^2), where $\theta | \, \tau \sim N(\mu, \tau^2)$ and $1/\tau^2 \sim \Gamma(a, b)$, has density

$$\pi(\theta, \tau^2) \propto \tau^{-2a+1} e^{-\frac{(\theta-\mu)^2+2b}{2\tau^2}}, \quad \theta \in \mathbb{R}, \tau^2 > 0.$$

The conditional density of τ^2 given θ is proportional to this function as a function of τ^2. It follows that the second conditional distribution satisfies $1/\tau^2 | \, \theta \sim \Gamma(a + 1/2, (\theta - \mu)^2/2 + b)$. The first conditional is already given in the definition of the model. Therefore the Gibbs sampler to simulate from the distribution of (θ, τ^2) proceeds by the sequence of updates

$$\theta_{n+1} \sim N(\mu, \tau_n^2), \qquad \frac{1}{\tau_{n+1}^2} \sim \Gamma(a + \tfrac{1}{2}, \tfrac{1}{2}(\theta_{n+1} - \mu)^2 + b).$$

Alternatively (and preferably?), one might simulate from this distribution by repeatedly drawing a value τ^2 from its marginal distribution, and then θ from its conditional distribution given τ. This would give a sequence of independent variables with exactly the correct distribution.

Next we discuss the stationarity and ergodicity properties of the two stage ($m = 2$) Gibbs sampler. We say that a density π satisfies the *positivity condition* if the positivity of the marginal densities implies the positivity of the joint density: if $\pi_i(y_i) > 0$ for $i = 1, ..., m$, then $\pi(y_1, ..., y_m) > 0$, for every $(y_1, \ldots, y_m)$.

Theorem 3.17. *The density π is a stationary distribution of the Markov chain (Y_n) generated by the Gibbs sampling algorithm. If the density π satisfies the positivity condition, then the chain $(Y_{n,1}, Y_{n,2})$ resulting from the two stage Gibbs sampler is ergodic.*

Proof If Y_n is distributed according to π, then $(Y_{n,2}, \ldots, Y_{n,m})$ is distributed according to marginal of π on the last $m-1$ coordinates. The first step of the Gibbs sampler is to generate $Y_{n+1,1}$ from the conditional distribution of the first coordinate given the last $m-1$ coordinates under π, evaluated at $(Y_{n,2}, \ldots, Y_{n,m})$. Then $(Y_{n+1,1}, Y_{n,2}, \ldots, Y_{n,m})$ is distributed according to π. (If $X_2 \sim \Pi_2$ and $X_1 | \, X_2 \sim \Pi_{1|2}(\cdot | \, X_2)$, for Π_2 and $\Pi_{1|2}$ the second marginal and first conditional of a distribution Π, then $(X_1, X_2) \sim \Pi$.) This shows that the stationary distribution is preserved under the first step of the Gibbs sampler. By the same argument it is preserved under every of the m steps of an iteration.

For a proof of the second assertion, see ? □

3.5 Missing data and data augmentation

Suppose that we wish to generate a sample from a density π, but this density is only known as the marginal density $\pi(y) = \int \bar{\pi}(y, z)\, d\mu(z)$ of a joint density $(y, z) \mapsto \bar{\pi}(y, z)$. Any of the sampling schemes would suffer in efficiency if at every step we would have to approximate the integral numerically. An easy way around this is to generate a sample $(Y_1, Z_1), \ldots, (Y_n, Z_n)$ from $\bar{\pi}$, and then discarding the Zs. Marginalizing an empirical sample just means throwing away the other variables!

This situation arises commonly in Bayesian computation with "missing data". Suppose that instead of the "full data" (X, Y) we only observe X. If $(x, y) \mapsto p_\theta(x, y)$ is a density of (X, Y), then $\int p_\theta(x, y)\, d\mu(y)$ is the density of the observation X. Given a prior density π the posterior density of interest is proportional to

$$\theta \mapsto \int p_\theta(x, y)\, d\mu(y)\, \pi(\theta).$$

If the integration cannot be performed analytically, implementing an MCMC scheme will be awkward.

The posterior distribution of ϑ given the observation X is the first marginal distribution of the conditional distribution of (ϑ, Y) given X. If $(\vartheta_1, Y_1), \ldots, (\vartheta_n, Y_n)$ are a sample from the latter conditional distribution, then $\vartheta_1, \ldots, \vartheta_n$ are a sample from the distribution of interest. This leads to the following strategy. We apply an MCMC scheme to simulate from the density that is proportional to $(\theta, y) \mapsto p_\theta(x, y)\pi(\theta)$, with x fixed to the observed value. Next we throw away the Y-values.

In this scheme the data y may be naturally missing, for instance due to the way the data is ascertained. However, the preceding strategy is valid for any y added to the data x, as long as there is joint distribution of (x, y, ϑ). When y is added for the convenience of sampling, one speaks of *data augmentation*. The augmented data could be part of any Bayesian hierarchy of distributional assumptions, provided the pair (x, ϑ) is contained in the hierarchy and its marginal distribution is the same as in the problem at hand.

3.6 *Hamiltonian MCMC

The Hamiltonian Markov chain Monte Carlo method is inspired by the laws of motion from physics. The negative log target density $-\log \pi$ is viewed as the potential energy of a particle, which is located at the value of the target variable y. The movement of the particle is controlled by the *Hamiltonian*

$$H(y, v) = -\log \pi(y) + \tfrac{1}{2} v^T M^{-1} v,$$

where M is a positive-definite matrix, usually diagonal, often a scalar multiple of the identity matrix.[5] The variable v is interpreted as the velocity of the particle and the term $\frac{1}{2} v^T M v$ as its kinetic energy. In the MCMC scheme it just serves as an augmented variable, which facilitates the MCMC steps. The function

$$(y, v) \mapsto e^{-H(y,v)} = \pi(y) e^{-v^T M^{-1} v/2}$$

[5] Choosing M^{-1} approximately equal to the covariance matrix of π typically leads to a more stable algorithm.

is proportional to the joint density of the target variable $Y \sim \pi$ taken together with an independent multivariate-normal variable $V \sim N(0, M)$. The algorithm generates a Markov chain of variables (Y_n, V_n) and then discards the second coordinates to obtain a sample from π.

Hamiltonian MCMC was introduced in the 1980s, but is only recently (and rapidly) gaining popularity, because it can be more efficient, in particular for irregularly shaped target densities π, for instance densities that spread very differently in different directions.

The algorithm consists of two steps. The first step changes only the momentum v, simply by generating a new value from the $N(0, M)$-distribution. By the independence of Y and V in the target density, this step clearly retains the stationary density. The second step is more interesting and changes both variables. In the idealized version of the algorithm, this step is deterministic. It replaces (y, v) by the value $(y(T), v(T))$ at time T of the curve (or *integral flow*) started at $(y(0), v(0)) = (y, v)$ and evolving according to the system of differential equations

$$y'(t) = \frac{\partial H}{\partial v}(y(t), v(t)),$$
$$v'(t) = -\frac{\partial H}{\partial y}(y(t), v(t)).$$

This is the so-called *Hamiltonian flow*, which describes the joint location and momentum of a particle as determined by the energy. The time step T is a parameter of the algorithm and must be chosen carefully for efficiency. In practice, it is typically impossible to solve the system exactly and $(y(T), v(T))$ is replaced by a numerical approximation, which we shall describe in a moment. The discrete approximation can be justified without studying the continuous flow (see Theorem 3.19), but for intuition it is instructive to explain why Hamiltonian MCMC works in the idealized setup.

Most deterministic algorithms do not retain a stationary density. The Hamiltonian flow is special in that (see next lemma) it retains the value of the Hamiltonian (i.e. $H(y(T), v(T)) = H(y(0), v(0))$), it preserves volume (the map $(y(0), v(0)) \mapsto (y(T), v(T))$ has Jacobian equal to 1), and it is reversible, up to sign changes. The first ("preservation of energy") means that the value of the target density $e^{-H(y,v)}$ does not change. Combination with the second (and the transformation rule for densities) shows that the new point $(y(T), v(T))$ is distributed according to this target if this is true for the initial point $(y(0), v(0))$. The third property seems not to be needed to justify the idealized Hamiltonian algorithm, but it will be seen to be important for discretization.

We record the three properties in a lemma, where we restrict to densities on $\mathcal{Y} = \mathbb{R}^d$.[6] Let $\phi: \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d \times \mathbb{R}^d$ give the map that sends the initial value $(y(0), v(0)) = (y, v)$ into $\phi(y, v, t) := (y(t), -v(t))$. The sign of the second coordinate is flipped to make the map exactly reversible (see (iii) below). This does not affect preservation of Hamiltonian or volume, as these are symmetric in momentum.

Lemma 3.18. *The Hamiltonian flow satisfies*

(i) The map $t \mapsto H(\phi(y, v, t))$ is constant, for every fixed (y, v).
(ii) The map $(y, v) \mapsto \phi(y, v, t)$ has Jacobian $\det(\frac{\partial}{\partial(y,v)}\phi(y, v, t))$ equal to 1, for every fixed t.
(iii) The inverse of the map $(y, v) \mapsto \phi(y, v, t)$ is the map itself, for every fixed t.

[6] Hamiltonian flow can be defined on general differential manifolds.

Proof Abbreviate the partial derivatives (gradients) of H with respect to y and v by H_y and H_v, respectively, and use similar notation for other functions. For (i) and (ii) we may ignore the sign flip of the second coordinate and prove the assertions for the map ϕ without the minus sign, which we call ψ. Thus $\psi\colon \mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d \times \mathbb{R}^d$ is the map that sends the initial value $(y(0), v(0)) = (y, v)$ into $\psi(y, v, t) = (y(t), v(t))$, for $(y(t), v(t))$ defined as the solutions of the system

$$\begin{aligned} \frac{\partial}{\partial t}\psi(y, v, t)_1 &= y'(t) = H_v(y(t), v(t)) = H_v(\psi(y, v, t)), \\ \frac{\partial}{\partial t}\psi(y, v, t)_2 &= v'(t) = -H_y(y(t), v(t)) = -H_y(\psi(y, v, t)). \end{aligned}$$

If we abbreviate $x = (y, v)$, then we have $\psi(x, 0) = x$ and we can write the pair of equations as the single equation $\psi_t(x, t) = F(\psi(x, t))$, for $F(x) = (H_v(x), -H_y(x))$.

The map in (i) with ϕ replaced by ψ can be written as $t \mapsto H(y(t), v(t))$, and has derivative $H_y(y(t), v(t))y'(t) + H_v(y(t), v(t))v'(t)$, which vanishes, in view of the system of differential equations.

Integrating the vector equation $\psi_t(x, t) = F(\psi(x, t))$ with respect to t, we obtain $\psi(x, t) = x + \int_0^t F(\psi(x, s))\, ds$. Differentiating this with respect to x, we obtain that the derivative matrix of $x \mapsto \psi(x, t)$ satisfies $\psi_x(x, t) = I + \int_0^t F'(\psi(x, s))\psi_x(x, s)\, ds$, for $F'(x)$ the derivative matrix of F at x, given by

$$F'(x) = \begin{pmatrix} H_{vy}(x) & H_{vv}(x) \\ -H_{yy}(x) & -H_{yv}(x) \end{pmatrix}.$$

Since the derivative of the map $t \mapsto \det A_t$ is equal to $\det A_t \operatorname{tr}(A_t^{-1}\dot{A}_t)$, we see that the derivative of $t \mapsto \det \psi_x(x, t)$ is given by

$$\det \psi_x(x, t) \operatorname{tr}\bigl(\psi_x(x, t)^{-1} F'(\psi(x, s))\psi_x(x, s)\bigr) = \det \psi_x(x, t) \operatorname{tr}(F'(\psi(x, s))),$$

since $\operatorname{tr}(AB) = \operatorname{tr}(BA)$, for any matrices A and B. From the expression for $F'(x)$, we see that $\operatorname{tr}(F'(x)) = H_{vy}(x) - H_{yv}(x) = 0$.[7] Thus the time derivative in the display is zero and hence the map $t \mapsto \det \psi_x(x, t)$ is constant and equal to its value 1 at $t = 0$.

The inverse of the map $x \mapsto \psi(x, t)$ is the flow obtained by negating the derivatives, i.e. by replacing F by $-F$: the map $x \mapsto \bar{\psi}(x, t)$ satisfying $\bar{\psi}(x, 0) = x$ and $\partial/\partial t\, \bar{\psi}(x, t) = -F(\bar{\psi}(x, t))$. We claim that this inverse is given by first flipping the sign of the momentum, next following the original flow, and finally flipping the sign of the momentum again, i.e. $\bar{\psi}(y, v, t)_1 = \psi(y, -v, t)_1$ and $\bar{\psi}(y, v, t)_2 = -\psi(y, -v, t)_2$. Indeed $\psi(y, -v, t)_1 = H_v(\psi(y, -v, t)) = -H_v(\psi(y, -v, t)_1, -\psi(y, -v, t)_2)$, since $H_v(y, v) = -H_v(y, -v)$ by symmetry of the Hamiltonian in v; and $-\psi(y, -v, t)_2 = H_y(\psi(y, -v, t)) = H_y(\bar{\psi}(y, v, t))$, because $H_y(y, v)$ depends on the first coordinate y only, which is the identical for $\psi(y, -v, t)$ and $\bar{\psi}(y, v, t)$), by definition.

Assertion (iii) is just a relabelling of the statement on the form of the inverse of $\bar{\psi}$, putting one of the two flips in the definition of ψ. □

The preservation of the Hamiltonian under the flow suggests why the algorithm may be efficient. The easy part of the (augmented) target distribution, the Gaussian factor, is updated

[7] The trace of the Jacobian matrix is the *divergence* $\operatorname{tr}(F') = \operatorname{div} F$ and describes the flow of the vector field. Under Hamiltonian flow this is zero.

efficiently by exact and independent sampling. Next the pair (y, v) is updated moving along the level sets of the Hamiltonian, thus preventing a drift to parts of the parameter space with little mass under the target. The algorithm achieves this by incorporating the gradient of the target density π next to the density itself, through the update of the momentum using the gradient of H.

Which time step T to use? The Hamiltonian flow may eventually lead a particle back to a position near its starting point. (Some flows are even periodic.) A too large value of T may therefore lead to inefficiency. A too small value is not desirable either, as this will correspond to small moves. We shall come back to this in the context of the discretized algorithm.

For most target densities π it will be impossible to derive a usable, analytic formula for the Hamiltonian flow. Instead we may solve the differential equation numerically. Euler's method would approximate the solution $(y(T), v(T))$ by recursively computing approximations $(y(\epsilon), v(\epsilon))$, $(y(2\epsilon), v(2\epsilon)), \ldots, (y(L\epsilon), v(L\epsilon))$, for a given (small) step size $\epsilon > 0$ and $T = L\epsilon$, using the difference equations

$$\begin{aligned} y(i\epsilon + \epsilon) &= y(i\epsilon) + \epsilon \frac{\partial H}{\partial v}\big(y(i\epsilon), v(i\epsilon)\big), \\ v(i\epsilon + \epsilon) &= v(i\epsilon) - \epsilon \frac{\partial H}{\partial y}\big(y(i\epsilon), v(i\epsilon)\big). \end{aligned}$$

It turns out that this does not work so well. First the final value $(y(T), v(T))$ will be far from the continuous solution, unless the step size ϵ is very small. Second (and related), the discretization loses the three good properties of the continuous flow. A simple fix to retain volume preservation and reversibility is to update the two coordinates y and v separately, using the last values of the coordinates as input in every update. Even better is to split one of the two updates in two "half updates", leading to the *leap-frog algorithm*

$$\begin{aligned} v(i\epsilon + \epsilon/2) &= v(i\epsilon) - \frac{\epsilon}{2} \frac{\partial H}{\partial y}\big(y(i\epsilon, v(i\epsilon)\big), \\ y(i\epsilon + \epsilon) &= y(i\epsilon) + \epsilon \frac{\partial H}{\partial v}\big(y(i\epsilon), v(i\epsilon + \epsilon/2)\big), \\ v(i\epsilon + \epsilon) &= v(i\epsilon + \epsilon/2) - \frac{\epsilon}{2} \frac{\partial H}{\partial y}\big(y(i\epsilon + \epsilon), v(i\epsilon + \epsilon/2)\big). \end{aligned}$$

This algorithm preserves volume and is reversible, and is also numerically more stable than Euler's algorithm (see lemma below). To compensate for the fact that it still does not reproduce the exact continuous flow (and does not exactly retain the Hamiltonian), Hamiltonian MCMC introduces a Metropolis-Hastings step, using the acceptance probability

$$\alpha\big(y(0), v(0), y(L\epsilon), v(L\epsilon)\big) = \frac{e^{-H\left(y(L\epsilon), v(L\epsilon)\right)}}{e^{-H\left(y(0), v(0)\right)}} \wedge 1.$$

Because the leap-frog algorithm is deterministic, the acceptance probability is actually only a function of the starting value $(y(0), v(0))$, unlike in the general Metropolis-Hastings al-

gorithm, but we included the endpoint as an argument of α for consistency with earlier notation.[8] We now define *Hamiltonian MCMC* as follows:[9]

Given (Y_n, V_n).
Generate $W \sim N(0, M)$.
Compute (Z_{n+1}, W_{n+1}) as the endpoint of L leap-frog iterations starting from (Y_n, W).
Generate U_{n+1} from the uniform distribution on $[0, 1]$.
If $U_{n+1} < \alpha(Y_n, W, Z_{n+1}, W_{n+1})$, set $(Y_{n+1}, V_{n+1}) := (Z_{n+1}, -W_{n+1})$,
else set $(Y_{n+1}, V_{n+1}) := (Y_n, V_n)$.

We may now discard the V_n and are left with a sample of Y_n that will be approximately from π, after burn-in. The swapping of the sign of W_{n+1} is included, because it makes the map reversible, but this step can be left out, as the variables V_n are not used and regenerated independently from their Gaussian distribution, at the beginning of every iteration.

The step size ϵ and number of leap-frog iterations L are parameters of the algorithm, which are chosen in advance. They are crucial to the efficiency of the algorithm, but not easy to tune in general. A too small step size ϵ will unnecessarily increase the number of computations, while a too large step size will lead to low acceptance probabilities, as the Hamiltonian will not be preserved. If the total path length $L\epsilon$ is too small, then the sampler will move too little, and if it is too long, the Hamiltonian dynamics may actually bring back the particle to (almost) its initial state. In practice one determines good values by trial and error. It can also be useful to choose a new, random value for the parameters at the beginning of every update. The so-called no U-turn criterion is meant to automatically tune the parameters. It is based on running the Hamiltonian flow both forward and backward until the difference of the forward and backward locations possesses a negative inner product with the momentum, which would indicate that the particle may be starting a return.

Theorem 3.19. *The density proportional to* $(y, v) \mapsto e^{-H(y,v)}$ *is stationary for the leap-frog Hamiltonian MCMC algorithm.*

Proof For $H(y, v) = U(y) + K(v)$, the basic leap-frog updates $(y, v) \mapsto (y, v - \epsilon/2U_y(y))$ and $(y, v) \mapsto (y + \epsilon K_v(v), v)$ possess Jacobian matrices

$$\begin{pmatrix} 1 & 0 \\ -\epsilon/2U_{yy}(y) & 1 \end{pmatrix} \qquad \begin{pmatrix} 1 & \epsilon K_{vv}(v) \\ 0 & 1 \end{pmatrix}.$$

[8] We could think of it as being of ordinary Metropolis-Hastings form, with the transition density $q(y, z)$ in the latter algorithm presently replaced by a Dirac function $\delta_{\phi(y)}(z) = 1\{z = \phi(y)\}$, giving the certain transition from y to $\phi(y)$. In this interpretation, the same transition kernel, but with the arguments switched, hence $\delta_{\phi(z)}(y) = 1\{\phi(z) = y\}$, should appear in the numerator of the acceptance probability. This is different, and hence the two terms would not "cancel from the quotient", unless $\phi = \phi^{-1}$. The latter is true for the Hamiltonian flow, and this may explain that no transition density enters in the present definition of α. In the last paragraph of the proof of next theorem it is shown in general, that a deterministic move retains a stationary distribution if it is strongly reversible in the latter sense.

[9] The half step at the end of a leap-frog update is often combined with the half-step at the beginning of the next leap-frog update into a single update. In other words, the "endpoint" in the algorithm is typically computed by starting with a half-step update of the momentum, next alternating L full-step updates of the location and $L - 1$ full-step updates of momentum and finally a half-step update of the momentum.

The Jacobians of these transformations are 1 and hence both updates preserve volume.

The inverses of these maps are given by $(y, v) \mapsto (y, v + \epsilon/2U_y(y))$ and $(y, v) \mapsto (y - \epsilon K_v(v), v)$. From the fact that $K_v(-v) = -K_v(v)$, we see that the inverses can also be obtained by first flipping the sign of v, then applying the original map, and finally flipping the sign of the new v: i.e. for ψ the original map, the inverse is $\bar{\psi}(y, v) = (\psi(y, -v)_1, -\psi(y, -v)_2)$. This means that the map ϕ that first flips the sign of v, then performs the leap-frog transformations and finally flips the sign of the new momentum, is its own inverse.

We can now conclude the proof by proving in general that a deterministic, volume-preserving transformation which is its own inverse ($\phi = \phi^{-1}$) followed by a Metropolis-Hastings step with acceptance probability $\alpha(x) = (\pi \circ \phi(x)/\pi(x)) \wedge 1$ preserves a stationary density π.

If $X_0 \sim \pi$ and $X_1 = \phi(X_0)$ with probability $\alpha(X_0)$ and is equal to X_0 otherwise, then $\Pr(X_1 \in B)$ is equal to

$$\int \pi(x)\big[\alpha(x)1_B(\phi(x)) + (1 - \alpha(x))1_B(x)\big]\,dx = \Pi(B) + \int \pi \circ \phi(x) \wedge \pi(x)(1_{\phi(x)\in B} - 1_{x\in B})\,dx.$$

Applying the substitution $\phi(x) = y$ to the integral on the right, we see that this integral is equal to $\int \pi(y) \wedge \pi \circ \phi^{-1}(y)(1_{y\in B} - 1_{\phi^{-1}(y)\in B})\,dy$, if the Jacobian is unity. If also $\phi^{-1} = \phi$, then the expression is identical to minus the original integral. Thus the integral vanishes. □

See Neal (2011) for an extended discussion of Hamiltonian MCMC and Livingstone et al. (2019) for results on ergodicity of the Hamiltonian chain.

3.7 *Reversible jump chains

In some situations the state space $\mathfrak{Y}$ carrying the target distribution Π is the disjoint union of spaces of different dimensions. This is true for instance for a posterior distribution in Bayesian model selection, when priors on statistical models of different dimensions are combined in an overall prior on the union of the models by mixing over the model index. In such a situation a Markov chain with target Π must 'jump' from one model to another model from time to time. Suitable chains can be constructed using the Metropolis-Hastings sampling scheme, but the jumps require special care. To make them reversible, as with the ordinary Metropolis-Hastings sampler, the different subspaces must be linked in some way.[10]

Consider a state space $\mathfrak{Y} = \cup_k \{k\} \times \mathfrak{Y}_k$ that is the union of disjoint spaces $\{k\} \times \mathfrak{Y}_k$, for an index k that ranges over a finite or countable set $\mathfrak{K}$. (The added notation $\{k\}\times$ helps to stress that the spaces $\{k\} \times \mathfrak{Y}_k$ are disjoint, and makes it possible to denote points in $\mathfrak{Y}$ by pairs (k, y_k).) The target distribution Π is a probability distribution on $\mathfrak{Y}$, and can be decomposed over the disjoint union. This is easiest described in terms of a random point (K, Y_K) in $\mathfrak{Y}$ drawn according to Π. Then K is a random variable with values in $\mathfrak{K}$ with a distribution (ρ_k), and given $K = k$ the variable Y_K is chosen from $\mathfrak{Y}_k$ according to some distribution Π_k:

$$\rho_k := \Pr(K = k), \qquad \Pi_k(A) := \Pr(Y_K \in A \,|\, K = k).$$

A general subset of $\mathfrak{Y}$ can be written as $A = \cup_k \{k\} \times A_k$, for measurable sets $A_k \subset \mathfrak{Y}_k$, and then $\Pi(A) = \sum_k \rho_k \Pi_k(A_k)$.

[10] Alternatively, non-reversible Markov chains have been developed for certain applications.

We wish to construct a Markov chain (Y_n) with state space $\mathfrak{Y}$ and stationary distribution Π. Given the current state $Y_n = (k, y_k)$ the chain will move to a state $Y_{n+1} = (l, y_l)$ (which as usual is allowed to be identical to (k, y_k), so that the chain stays put). The idea is first to decide on l using a transition distribution $l \mapsto r(k, l)$ from $\mathfrak{K}$ to $\mathfrak{K}$, and next to decide on the value $y_l \in \mathfrak{Y}_l$. If $l = k$, then the move is within $\mathfrak{Y}_k$, and we can just follow any of the schemes discussed before. On the other hand, if $l \neq k$, the move may be between spaces of a different character. This could be complicated, lacking an easy common dominating measure on the two spaces, that could be used to describe transition proposals $q(y_k, y_l)$ and $q(y_l, y_k)$ and corresponding acceptance probabilities, making the moves reversible. We concentrate on moves of the second type, and more precisely on moves between spaces of different dimensions.

For $k \in \mathfrak{K}$ assume that $\mathfrak{Y}_k \subset \mathbb{R}^{d_k}$ is an open subset of Euclidean space of dimension d_k. For distinct indices $k, l \in \mathfrak{K}$ such that $d_k < d_l$, let $g_{k,l}\colon \mathfrak{Y}_k \times \mathfrak{U}_{k,l} \to \mathfrak{Y}_l$ be a diffeomorphism, for a given open subset $\mathfrak{U}_{k,l} \subset \mathbb{R}^{d_l - d_k}$. The latter space will be used to generate an auxiliary variable that makes up for the difference in dimensions between the two spaces. Let $q_{k,l}$ be a probability density on $\mathfrak{U}_{k,l}$ relative to Lebesgue measure.

Let $g'_{k,l}$ be the derivative of $g_{k,l}$, and let $|g'_{k,l}(y_k, u_k)|$ be the Jacobian at (y_k, u_k), the absolute value of the determinant of the $d_l \times d_l$ matrix $g'_{k,l}(y_k, u_k)$.

Assume that Π_k has a density π_k with respect to Lebesgue measure on $\mathfrak{Y}_k$.

As mentioned, the transition of the Markov chain starts by proposing a new index l:

```
Given Y_n = (k, y_k).
Generate l from r(k, ·),
```

The next step is to generate a value $y_l \in \mathfrak{Y}_l$. If $d_k = d_l$ this is done by any of the methods discussed before, after suitably identifying the spaces $\mathfrak{Y}_k$ and $\mathfrak{Y}_l$, if necessary. We concentrate on the other case, where the move depends on whether $d_k < d_l$ or $d_k > d_l$.

```
Generate U_{n+1} from the uniform distribution on [0, 1].
If d_k < d_l generate u_{k,l} from q_{k,l}(·), and set z_l = g_{k,l}(y_k, u_{k,l}),
    if U_{n+1} < α_{k,l}(y_k, u_{k,l}, z_l), set Y_{n+1} := (l, z_l),
    else  set Y_{n+1} := Y_n,
else  if d_k > d_l set (z_l, u_{l,k}) = g^{-1}_{l,k}(y_k),
    if U_{n+1} < β_{k,l}(y_k, z_l, u_{l,k}), set Y_{n+1} := (l, z_l),
    else  set Y_{n+1} := Y_n,
```

Here the acceptance probabilities $\alpha_{k,l}\colon \mathfrak{Y}_k \times \mathfrak{U}_{k,l} \times \mathfrak{Y}_l \to [0, 1]$ and $\beta_{k,l}\colon \mathfrak{Y}_k \times \mathfrak{Y}_l \times \mathfrak{U}_{l,k} \to [0, 1]$ are given by[11]

$$\alpha_{k,l}(y, u, z) = \frac{\rho_l \pi_l(z)\, r(l, k) |g'_{k,l}(y, u)|}{\rho_k \pi_k(y)\, r(k, l) q_{k,l}(u)} \wedge 1, \qquad \beta_{k,l}(y, z, u) = \frac{\rho_l \pi_l(z)\, r(l, k) q_{l,k}(u)}{\rho_k \pi_k(y)\, r(k, l) |g'_{l,k}(y, u)|} \wedge 1.$$

The Jacobian $|g'_{k,l}(y, u)|$ arises in these formulas, because the densities are understood relative

[11] These are acceptance probabilities for moves from $\mathfrak{Y}_k$ to $\mathfrak{Y}_l$ in the cases $d_k < d_l$ and $d_k > d_l$, respectively. For a reverse move from $\mathfrak{Y}_l$ to $\mathfrak{Y}_k$ the indices k and l must be switched!

to Lebesgue measure. To see this better it could be written $|\partial z/\partial(u, y)|$, which suggests to read $\alpha_{k,l}(y, u, z)$ as

$$\frac{\rho_l \pi_l(z)\, r(l,k)\, |dz|}{\rho_k \pi_k(y)\, r(k,l) q_{k,l}(u)\, |dy\, du|} \wedge 1.$$

One finds the formula in this way in the literature, but this may not be helpful for clarity.

Regrettably the norming constants of the densities π_k, π_l and $q_{k,l}, q_{l,k}$ depend on k and l, and do not cancel from the quotients. Unlike in the ordinary Metropolis-Hastings, these densities must be known completely, except possibly for a single constant common to them all. This can be inconvenient. For instance, to compute a posterior distribution in the Bayesian setup the density π_k would be proportional to the function $\theta \mapsto p_{\theta,k}(x)\pi_k(\theta)$, for $p_{\theta,k}$ the likelihood and π_k the prior density of the parameter of the kth model, and the norming constant would be the inverse of $\int p_{\theta,k}(x)\pi_k(\theta)\, d\theta$. The sampling scheme would suffer in efficiency if this (multivariate) integral would not be analytically computable. Choosing the prior π_k conjugate to the family $p_{\theta,k}$ will help, provided the norming constant of the latter family is explicit.

Theorem 3.20. *Assume that the moves between spaces of equal dimension are reversible with respect to* Π. *Then the sequence* (Y_n) *produced by the reversible jump scheme is a Markov chain on* $\mathfrak{Y}$ *with stationary measure* Π.

Proof It suffices to show that the chain is reversible: if Y_n is sampled from Π and Y_{n+1} is produced as described, then $\Pr(Y_n \in A, Y_{n+1} \in B) = \Pr(Y_n \in B, Y_{n+1} \in A)$, for every pair of measurable sets $A, B \subset \mathfrak{Y}$.

If $A = \{k\} \times A_k$ for $A_k \subset \mathfrak{Y}_k$ and $B = \{l\} \times B_l$ for $B_l \subset \mathfrak{Y}_l$ and $d_k < d_l$, then a jump must take place for both events to occur, and hence

$$\begin{aligned}
\Pr(Y_n \in A, Y_{n+1} \in B) &= \int\!\!\int 1_{A_k}(y)\rho_k\pi_k(y)r(k,l)q_{k,l}(u)\alpha_{k,l}(y,u,g_{k,l}(y,u))1_{B_l}(g_{k,l}(y,u))\, dy\, du,\\
\Pr(Y_n \in B, Y_{n+1} \in A) &= \int 1_{B_l}(z)\rho_l\pi_l(z)r(l,k)\beta_{l,k}(z, g_{k,l}^{-1}(z))1_{A_k}(g_{k,l}^{-1}(z)_1)\, dz\\
&= \int\!\!\int 1_{B_l}(g_{k,l}(y,u))\rho_l\pi_l(g_{k,l}(y,u))r(l,k)\beta_{l,k}(g_{k,l}(y,u),y,u)1_{A_k}(y)\, |g'_{k,l}(y,u)|\, dy\, du,
\end{aligned}$$

where in the second line $g_{k,l}^{-1}(z)_1$ is the first coordinate of the pair $g_{k,l}^{-1}(z) =: (y, u)$, and the last line is obtained through the substitution $z = g_{k,l}(y, u)$, with $dz = |g'_{k,l}(y, u)|\, dy\, du$. The two expressions are equal if

$$\rho_k\pi_k(y)r(k,l)q_{k,l}(u)\alpha_{k,l}(y,u,g_{k,l}(y,u)) = \rho_l\pi_l(g_{k,l}(y,u))r(l,k)\beta_{l,k}(g_{k,l}(y,u),y,u)|g'_{k,l}(y,u)|.$$

This is ensured by the definitions of $\alpha_{k,l}$ and $\beta_{l,k}$, where we simplify notation by setting $z = g_{k,l}(y, u)$, and the expression for $\beta_{l,k}$ is derived from the formula for $\beta_{k,l}$ given preceding the theorem by switching k and l, and y and z in the notation.

For the same type of sets, but with $d_k > d_l$, the equality entails just swapping A and B, and hence follows by symmetry.

If $d_k = d_l$ and $k = l$, either the new index generated from $r(k, \cdot)$ is equal to $k = l$ and the chain moves by a sampler that is reversible by assumption, or the new index proposed

is different from $k = l$, but next the proposed move is rejected and the chain stays put. The latter is clearly reversible.

For the same type of sets, but with $d_k = d_l$ and $k \neq l$, the equality is true by assumption that the sampler used is reversible.

For general sets $A, B \subset \mathfrak{Y}$, the probability $\Pr(Y_n \in A, Y_{n+1} \in B)$ can be decomposed as $\sum_{k,l} \Pr(Y_n \in \{k\} \times A_k, Y_{n+1} \in \{l\} \times B_l)$, and the reversal follows by symmetry of all terms. □

The sampler moves differently when jumping up or down in dimension, which is intrinsic to the problem. However, the resulting asymmetry in the formulas is sometimes removed by also generating an auxiliary variable when moving down in dimension. Then for every $k, l \in \mathfrak{K}$ a diffeomorphism $g_{k,l}\colon \mathfrak{Y}_k \times \mathfrak{U}_{k,l} \to \mathfrak{Y}_l \times \mathfrak{U}_{l,k}$ between two augmented spaces (which must have the same dimensions) is employed, where $g_{k,l} = g_{l,k}^{-1}$, and a jump between arbitrary spaces $\mathfrak{Y}_k$ and $\mathfrak{Y}_l$ is described by

```
Generate U_{n+1} from the uniform distribution on [0,1].
Generate u_{k,l} from q_{k,l}(·), and set (z_l, u_{l,k}) = g_{k,l}(y_k, u_{k,l}).
If U_{n+1} < α_{k,l}(y_k, u_{k,l}, z_l, u_{l,k}), set Y_{n+1} := (l, z_l),
else  set Y_{n+1} := Y_n.
```

This works whatever the dimensions d_k and d_l, and there is a single acceptance probability $\alpha_{k,l}\colon \mathfrak{Y}_k \times \mathfrak{U}_{k,l} \times \mathfrak{Y}_l \times \mathfrak{U}_{l,k} \to [0, 1]$, given by

$$\alpha_{k,l}(y, u, z, v) = \frac{\rho_l \pi_l(z)\, r(l, k) q_{l,k}(v) |g'_{k,l}(y, u)|}{\rho_k \pi_k(y)\, r(k, l) q_{k,l}(u)} \wedge 1.$$

The values (y, u, z, v) appearing in this context satisfy $(z, v) = g_{k,l}(y, u)$, and then $g'_{k,l}(y, u) = \partial(z, v)/\partial(y, u)$, and the right side can be further symmetrized (at least in notation) as

$$\frac{\rho_l \pi_l(z)\, r(l, k) q_{l,k}(v)}{\rho_k \pi_k(y)\, r(k, l) q_{k,l}(u)} \frac{|\partial(z, v)|}{|\partial(y, u)|} \wedge 1.$$

Theorem 3.21. *The sequence (Y_n) produced by the symmetrized reversible jump scheme is a Markov chain on $\mathfrak{Y}$ with stationary measure Π.*

Proof As in the proof of the preceding theorem it suffices to show that the chain is reversible, and this can be reduced to showing that $\Pr(Y_n \in A, Y_{n+1} \in B) = \Pr(Y_n \in B, Y_{n+1} \in A)$, for every pair of measurable sets of the form $A = \{k\} \times A_k$ for $A_k \subset \mathfrak{Y}_k$ and $B = \{l\} \times B_l$ for $B_l \subset \mathfrak{Y}_l$ if Y_n is sampled from Π and Y_{n+1} is produced as described.

Presently, if $k \neq l$,

$$\Pr(Y_n \in A, Y_{n+1} \in B) = \int\!\!\int 1_{A_k}(y) \rho_k \pi_k(y) r(k, l) q_{k,l}(u) \alpha_{k,l}(y, u, g_{k,l}(y, u)) 1_{B_l}(g_{k,l}(y, u)_1)\, dy\, du,$$

$$\Pr(Y_n \in B, Y_{n+1} \in A) = \int\!\!\int 1_{B_l}(z) \rho_l \pi_l(z) r(l, k) q_{l,k}(v) \alpha_{l,k}(z, v, g_{l,k}(z, v)) 1_{A_k}(g_{l,k}(z, v)_1)\, dz\, dv.$$

The first formula is as before, except that now the proposed value in $\mathfrak{Y}_l$ is only the first coordinate $g_{k,l}(y, u)_1$ of $g_{k,l}(y, u)$, the second coordinate being used for computing the acceptance probability only. We rewrite the second integral using the substitution $(z, v) = g_{k,l}(y, u)$,

which by assumption is equivalent to $(y, u) = g_{l,k}(z, v)$, giving

$$\int\int 1_{B_l}(g_{k,l}(y,u)_1)\rho_l\pi_l(g_{k,l}(y,u)_1)r(l,k)q_{l,k}(g_{k,l}(y,u)_2)\alpha_{l,k}(g_{k,l}(y,u),y,u)1_{A_k}(y)\,|g'_{k,l}(y,u)|\,dy\,du.$$

This shows that the two probabilities are equal if

$$\begin{aligned}&\rho_k\pi_k(y)r(k,l)q_{k,l}(u)\alpha_{k,l}(y,u,g_{k,l}(y,u))\\&\quad=\rho_l\pi_l(g_{k,l}(y,u)_1)r(l,k)q_{l,k}(g_{k,l}(y,u)_2)\alpha_{l,k}(g_{k,l}(y,u),y,u)\,|g'_{k,l}(y,u)|.\end{aligned}$$

We simplify this by substituting $g_{k,l}(y, u) = (z, v)$, and then see that this is ensured by the definition of $\alpha_{k,l}$, and the implied definition of $\alpha_{l,k}$, where we note that $g'_{l,k}(z, v)$ is the inverse of $g'_{k,l}(y, u)$.

If $k = l$, then the probabilities as computed must be enlarged with the probabilities that the chain does not move and the value $Y_n = Y_{n+1}$ belongs to both A_k and B_l. In both cases this probability is

$$\int\int 1_{A_k\cap B_l}(y)\rho_k\pi_k(y)r(k,k)q_{k,k}(u)(1-\alpha_{k,k}(y,u,g_{k,k}(y,u))\,dy\,du.$$

This is clearly invariant under swapping A_k and B_l. □

The reversible jump scheme generates a chain of joint values (k, y_k). Sometimes it is possible and advantageous to separate the index k and value y_k. Given k, sampling y_k from π_k is a problem of fixed dimension and can be solved without jumps. Sampling k involves only the marginal distribution (ρ_k) on the countable set $\aleph$ and hence also does not need jumps. Failing more efficient methods, we could implement a Metropolis-Hastings sampler with proposal distribution $r(k, \cdot)$ and acceptance probabilities

$$\frac{\rho_l\, r(l,k)}{\rho_k\, r(k,l)} \wedge 1.$$

Unfortunately, although in our abstract notation this looks easy, in practice it may not, as the ρ_k may not be explicit. In particular, in the context of Bayesian model selection the ρ_k would involve the likelihood and be proportional to $\int p_{\theta,k}(x)\,\rho_k\pi_k(\theta)\,d\theta$, which could be a high-dimensional integral of the type we wanted to avoid evaluating in the first place.

An intermediate approach, known as *partial analytic structure*, separates out the part of y_k that is not common to the value y_l in the intended jump space, as follows. Given k and l, assume that $y_k = (y_{k,l}, y_{k,-l})$ and $y_l = (y_{l,k}, y_{l,-k})$, where $y_{k,l} = y_{l,k}$. Then when jumping from k to l, the common part $y_{k,l}$ could be retained, and we could conditionally on $y_{k,l}$ apply the idea of the preceding paragraph to separate generating the new index l from updating the second part of y_k, as follows. Let

$$\begin{aligned}\rho(k|\,y_{k,l}) &= \frac{\int \rho_k\pi_k(y_{k,l},y_{k,-l})\,dy_{k,-l}}{\sum_m \int \rho_m\pi_m(y_{k,l},y_{m,-l})\,dy_{m,-l}},\\ \pi_k(\cdot|\,y_{k,l}) &= \frac{\pi_k(y_{k,l},\cdot)}{\int \pi_k(y_{k,l},y_{k,-l})\,dy_{k,-l}}.\end{aligned}$$

The first gives the probabilities $\Pr(K = k|\,Y_{K,l} = y_{k,l})$, while the second is the conditional density of $Y_{k,-l}$ given $K = k, Y_{k,l} = y_{k,l}$. (The sum over m in the first is restricted to the values of m for which $y_{k,l}$ is possible as first coordinate of $y_m = (y_{m,l}, y_{m,-l})$ and $y_{m,l} = y_{k,l}$.)

Given $Y_n = (k, y_k)$,
Generate l from $r(k, \cdot)$.
Split $y_k = (y_{k,l}, y_{k,-l})$ and set $y_{l,k} = y_{k,l}$.
Generate U_{n+1} from the uniform distribution on $[0, 1]$.
If $U_{n+1} < \alpha_{k,l}(y_{k,l})$, generate $z_{l,-k}$ from $\pi_l(\cdot|\, y_{l,k})$ and set $Y_{n+1} := (l, y_{l,k}, z_{l,-k})$,
else generate $z_{k,-l}$ from $\pi_k(\cdot|\, y_{k,l})$ and set $Y_{n+1} := (k, y_{k,l}, z_{k,-l})$.

Here the acceptance probability is given by

$$\alpha_{k,l}(y) = \frac{\rho(l|\, y)\, r(l,k)}{\rho(k|\, y)\, r(k,l)} \wedge 1.$$

3.8 *Exchange algorithm

It was noted that the norming constant in the target density π cancels from the acceptance probability of a Metropolis-Hastings chain, so that it suffices to know π up to a proportionality factor. This is useful, but not always enough. In a Bayesian setup the target density is proportional to $\theta \mapsto \pi_1(\theta)p_\theta(x)$, for π_1 the prior density, and hence we must still know the norming constant to the data density p_θ, as this is a function of θ. In a model selection setup the target of the kth model is proportional to $(k, \theta) \mapsto \rho_k\pi_k(\theta)p_{k,\theta}(x)$, and we must know the norming constant of the posterior density $\theta \mapsto \pi_k(\theta)p_{k,\theta}(x)$ of the kth model. Such difficulties can be overcome with an extension of the Metropolis-Hastings algorithm.

Suppose that the target density is $\pi(y) = c(y)\bar{\pi}(y)$, where $\bar{\pi}(y)$ can be computed for every y, but $c(y)$ cannot. For instance, the target density is $\theta \mapsto \pi_1(\theta)c(\theta)\bar{p}_\theta(x)$, where c is the norming constant to $\bar{p}_\theta$. Consider the following algorithm, with arbitrary proposal densities q and r.

Given a starting value Y_0 we proceed recursively, for $n = 0, 1, 2, \ldots$, as follows.

Given Y_n.
Generate Z_{n+1} from the distribution with density $q(Y_n, \cdot)$.
Generate V_{n+1} from the distribution with density $r(Z_{n+1}, \cdot)$.
Generate U_{n+1} from the uniform distribution on $[0, 1]$.
If $U_{n+1} < \alpha(Y_n, Z_{n+1}, V_{n+1})$, set $Y_{n+1} := Z_{n+1}$,
else set $Y_{n+1} := Y_n$.

Just as in the ordinary Metropolis-Hastings algorithm, the chain accepts a move to the proposed value z with probability $\alpha(y, z, v)$, but this now also depends on the extra variable v. By integrating out this variable, we obtain an ordinary Metropolis-Hastings chain with acceptance probability $\bar{\alpha}(y, z) = \int \alpha(y, z, v)r(z, v)\, d\mu(v)$. If we set this as in the ordinary chain, then π will be a stationary density. Thus we assume that

$$\frac{\int \alpha(y, z, v)r(z, v)\, d\mu(v)}{\int \alpha(z, y, v)r(y, v)\, d\mu(v)} = \frac{\pi(z)q(z, y)}{\pi(y)q(y, z)}.$$

Theorem 3.22. *The Markov chain (Y_n) has stationary density π.*

Proof If $Y_n \sim \pi$, then $\Pr(Y_n \in A, Y_{n+1} \in B)$ is given by

$$\int\int\int 1_A(y)\pi(y)\alpha(y,z,v)1_B(z)q(y,z)r(z,v)\,d\mu(v)\,d\mu(z)\,d\mu(y)$$
$$+\int\int\int 1_A(y)\pi(y)(1-\alpha(y,z,v))1_B(y)q(y,z)r(z,v)\,d\mu(v)\,d\mu(z)\,d\mu(y)$$

The second integral depends on (A, B) only through $A \cap B$ and hence is symmetric in A and B. The first integral is symmetric in A and B if the integral $\int \pi(y)\alpha(y,z,v)q(y,z)r(z,v)\,d\mu(v)$ is invariant under exchanging y and z, for every (y, z). This is ensured by the definition of α, which is equivalent to symmetry of the latter integral. □

As an example consider the *exchange algorithm* meant to approximate the posterior distribution in the case that the constant in the likelihood is not explicit. The target density is proportional to $\theta \mapsto \pi_1(\theta)p_\theta(x)$, for fixed x, where $p_\theta(x) = c(\theta)\bar{p}_\theta(x)$, for a computable value $\bar{p}_\theta(x)$ and non-explicit norming constant $c(\theta)^{-1} = \int \bar{p}_\theta(x)\,d\mu(x)$. The algorithm assumes that we can simulate variables from p_θ.

The acceptance probability is set equal to

$$\alpha(\theta,\theta',v) = \frac{\pi_1(\theta')\bar{p}_{\theta'}(x)q(\theta',\theta)\bar{p}_\theta(v)}{\pi_1(\theta)\bar{p}_\theta(x)q(\theta,\theta')\bar{p}_{\theta'}(v)} \wedge 1.$$

Given $Y_n = \theta$.
Generate θ' from the distribution with density $q(\theta,\cdot)$.
Generate v from the distribution with density $p_{\theta'}$.
Generate U_{n+1} from the uniform distribution on $[0,1]$.
If $U_{n+1} < \alpha(\theta,\theta',v)$, set $Y_{n+1} := \theta'$,
else set $Y_{n+1} := \theta$.

To see that this works, it suffices to verify that $\int \alpha(\theta,\theta',v)p_{\theta'}(v)\,d\mu(v)\,\pi_1(\theta)p_\theta(x)q(\theta,\theta')$ is symmetric in θ and θ', which readily follows from the definition of α.

3.9 *Connections between the samplers

The slice sampler is actually a special Gibbs sampler, and the coordinate updates of the Gibbs sampler are special Metropolis-Hastings steps.

Lemma 3.23. *The slice sampler is the special case of the Gibbs sampler applied to generating bivariate samples (Y_n, U_n) from the uniform distribution on $\mathcal{F}(\pi) = \{(y,u)\colon 0 \le u \le \pi^*(y)\}$.*

Proof The density of the uniform distribution satisfies $\pi(y,u) \propto 1_{0\le u\le\pi^*(y)}$. The full conditional densities $\pi(y|\,u)$ and $\pi(u|\,y)$ correspond to the uniform distributions $U(B(u))$ and $U(0,\pi^*(y))$, for $B(u) = \{y\colon u \le \pi^*(y)\}$. These correspond to the two updating steps of the slice sampler. □

For the next lemma we extend the Metropolis-Hastings sampler to proposal distributions that may not be dominated and given by a transition density. In the case that $y = (y_1, y_2)$

and only y_2 is updated and y_1 is left unchanged, we require that the transition proposal for y_2 is given by a density and read the quotient $q(z, y)/q(y, z)$ in the acceptance probability accordingly as referring to the transition density for y_2 only.

Lemma 3.24. *Every coordinate update of the Gibbs sampler is a Metropolis-Hastings sampler with acceptance probability equal to one.*

Proof The transition kernel of the ith update of the Gibbs sampler is given by, for $y_{-i} = (y_1, ..., y_{i-1}, y_{i+1}, ..., y_m)$,

$$q_i(y, y') = \delta_{y_{-i}}(y'_{-i})\pi_i(y'_i|\, y_{-i}).$$

If this is taken to be the proposal distribution of a Metropolis-Hastings algorithm, then the acceptance probability is equal to

$$\frac{\pi(y')q_i(y', y)}{\pi(y)q_i(y, y')} = \frac{\pi(y')\pi_i(y_i|\, y_{-i})}{\pi(y)\pi_i(y'_i|\, y_{-i})} = 1,$$

since $y_{-i} = y'_{-i}$ by construction, and $\pi(y')/\pi_i(y'_i|\, y_{-i}) = \pi_{-i}(y_{-i})$, for π_{-i} the marginal density of y_{-i}, and similarly $\pi(y)/\pi_i(y_i|\, y_{-i}) = \pi_{-i}(y_{-i})$. □

The lemma does not mean that every Gibbs sampler is a Metropolis-Hastings algorithm. In fact, there are elementary examples (see for instance VanDerwerken (2016)) of Gibbs samplers that do not satisfy the detailed balance property, which automatically holds for every Metropolis-Hastings algorithm, as seen in the proof of Theorem 3.7.

Notwithstanding the close relationship between the Gibbs sampler and the Metropolis-Hastings algorithm, certain features of the Gibbs sampler are worth noting separately. The Gibbs sampler accepts all proposed moves, whereas the acceptance rate is a delicate parameter of the general Metropolis-Hastings algorithm (see the discussion in Sections 3.2.1 and 3.2.2). The primary use of the Gibbs sampler is to reduce the dimension of the variables to be sampled, although sometimes artificial variables are added to break up sampling in multiple steps (the slice sampler is an example). The Gibbs sampler is less useful when the conditional distributions are not analytically available. Incorrect parametrization of the model can substantially increase the convergence time of the Gibbs sampler, or even result in the algorithm getting "trapped" in certain states (see for instance Hills and Smith (1992)). Finally, the Gibbs sampler does not apply to a union of models of different dimensions.

There are various ways to combine the Gibbs sampler and the Metropolis-Hastings algorithm. The *hybrid MCMC algorithm* simultaneously utilizes Gibbs and MH steps. For instance, every rth iteration (or every iteration with probability $1/r$) of a Gibbs sampler is replaced by a Metropolis-Hastings step. In case one of the full conditional distributions is not analytically tractable, one might sample from this conditional by a general Metropolis-Hastings algorithm. This is called *Metropolis-Hastings embedded in a Gibbs sampler*.

3.10 Variational Bayes

MCMC methods can be slow, or even computationally infeasible, when the target distribution is complex. In the Bayesian framework this may happen when the data set is large or the parameter set is high dimensional. To speed up the computations we may choose to compute

an attractive approximation, rather than the exact target distribution. A popular approach is the *variational method*, which casts inference as finding a best approximation of the target distribution Π in a given class of distributions $\mathcal{Q}$. When applied to compute a posterior distribution, this leads to the *variational Bayes method*.[12]

The standard variational method uses the *Kullback-Leibler divergence* to measure the discrepancy between the target and its approximation. For two probability distributions P and Q with densities p and q relative to a measure μ on a given measurable space this is defined as

$$K(Q;P) = \int \log \frac{q}{p}\, dQ = \int \Big(\log \frac{q}{p}\Big) q\, d\mu.$$

This number can be shown to be independent of the dominating measure μ, and nonnegative, but can be ∞. (The quotient q/p is defined to be infinite if $q > p = 0$, with logarithm also infinite; the integral over the set where $q = 0$ is understood to be 0.) Note that the divergence is asymmetric in P and Q. We shall write $K(O;P)$ and $K(q;p)$ interchangeably.

Now given a target probability distribution Π and a given class $\mathcal{Q}$ of probability distributions Q, the variational approximation is defined by

$$Q^* = \operatorname*{argmin}_{Q\in\mathcal{Q}} K(Q;\Pi)$$

When applied with Π equal to a posterior distribution, this gives the variational Bayes posterior. If Π and every distribution in $\mathcal{Q}$ possesses a density, then the minimization problem can be stated in terms of densities, and leads to a density q^*.

The choice of the variational class is crucial and somewhat a form of art. Larger, more flexible classes provide better approximations, but may be computationally less attractive, and vice versa. Without an efficient and accurate algorithm to find the minimum, the method is empty. By the asymmetry of the Kullback-Leibler divergence the positioning of Q and Π in the preceding display is important: the integral is taken with respect to Q, not Π. This choice seems to be motivated by computational efficiency.

3.10.1 Evidence lower bound

In the Bayesian framework the target is the posterior distribution $\Pi(\cdot|\,X)$ given the data X. If the likelihood is given by a density $x \mapsto p_\theta(x)$ and the prior by a density π, then Bayes's rule gives the posterior density as $\pi(\theta|\,X) = p_\theta(X)\pi(\theta)/p(X)$, for $x \mapsto p(x) = \int p_\theta(x)\, d\Pi(\theta)$ the Bayesian marginal density of X (or evidence), and the Kullback-Leibler divergence between $\Pi(\cdot|\,X)$ and a probability measure Q with density q is

$$K(q;\pi(\cdot|\,X)) = \int \Big[\log q - \log(\pi(\theta)p_\theta(X))\Big]\, dQ + \log p(X).$$

The term $\log p(X)$ on the far right does not depend on q and hence can be ignored when searching a minimum over q. Thus minimizing the variational criterion is equivalent to maximizing the quantity

$$\mathrm{ELBO}(q) = -K(q;\pi(\cdot|\,X)) + \log p(X) = \mathrm{E}_q \log(p_\vartheta(X)\pi(\vartheta)) - \mathrm{E}_q \log q(\vartheta).$$

[12] For a more detailed introduction of the topic, we recommend Chapter 10 of Bishop (2006) or the review article Blei et al. (2017).

Here E_q refers to the expectation relative to $\vartheta \sim q$, for fixed X. The nonnegativity of the Kullback-Leibler divergence and the definition immediately that $\mathrm{ELBO}(q) \le \log p(X)$. This explains the name *evidence lower bound* and the acronym "ELBO".

The variational class $\mathcal{Q}$ is typically chosen independent of the data X, but the *variational posterior distribution*, the maximizer of the ELBO for fixed X, of course depends on the given data.

3.10.2 Mean-field variational family

The most popular variational class $\mathcal{Q}$ is the set of all product measures in case the state space is a product space $\mathfrak{Y} = \mathfrak{Y}_1 \times \cdots \times \mathfrak{Y}_m$. As in the case of the Gibbs sampler, the coordinates y_i of $y = (y_1, \ldots, y_m) \in \mathfrak{Y}$ may arise naturally, or consist of blocks that are formed for computational convenience. Suppose that the target distribution has density π relative to a product measure $\mu_1 \times \cdots \times \mu_m$, and consider the class $\mathcal{Q}$ of all distributions with a density q that can be decomposed as

$$q(y_1, \ldots, y_m) = \prod_{i=1}^{m} q_i(y_i),$$

for some densities $q_1, ..., q_m$. Often the densities are left unspecified, and the variational criterion is minimized over all densities. Alternatively, the densities $q_1, \ldots, q_m$ may be restricted to some model.

The Kullback-Leibler divergence between the variational approximation to a target density π can be written, with $y = (y_1, \ldots, y_m)$ and $\mu = \prod_{i=1}^m \mu_i$,

$$\begin{aligned}\int \Bigl(\log \frac{\prod_{i=1}^m q_i(y_i)}{\pi(y)}\Bigr) \prod_{i=1}^{n} q_i(y_i)\, d\mu(y_1, \ldots, y_m) \\ = \sum_{j=1}^{m} \int (\log q_j)\, q_j \, d\mu_j - \int \log \pi(y) \prod_{i=1}^{m} q_i(y_i)\, d\mu(y_1, \ldots, y_m).\end{aligned}$$

Finding densities $q_1, \ldots, q_m$ that minimize this Kullback-Leibler divergence is not necessarily an easy problem. However, minimization with respect to a single density q_i given fixed values of the other densities q_j, for $j \ne i$, is often feasible. We rewrite the preceding display as

$$\begin{aligned}\sum_{j=1}^{m} \int (\log q_j)\, q_j \, d\mu_j - \int \Bigl[\int \log \pi(y) \prod_{j \ne i} q_j(y_j)\, d\mu_{-i}(y_{-i}) \Bigr] q_i(y_i)\, d\mu_i(y_i) \\ = \sum_{j \ne i} \int (\log q_j)\, q_j \, d\mu_j + \int \left[\log \frac{q_i(y_i)}{\exp \int \log \pi(y) \prod_{j \ne i} q_j(y_j)\, d\mu_{-i}(y_{-i})} \right] q_i(y_i)\, d\mu_i(y_i).\end{aligned}$$

Here $\mu_{-i} = \prod_{j \ne i} \mu_j$ and $y_{-i} = (y_j)_{j \ne i}$ is the vector y with its ith coordinate left out. The function in the denominator of the fraction within the logarithm in the second integral is independent of q_i and hence fixed given q_j, for $j \ne i$. It is not a probability density, but its norming constant is also independent of q_i and hence the expression is up to an additive constant equal to the Kullback-Leibler divergence between q_i and the normalized function. The nonnegativity of the Kullback-Leibler divergence shows that the expression is minimized

with respect to q_I ranging over all densities, by taking q_i equal to the normalized function, i.e. the minimizer is

$$q_i^*(y_i) \propto \exp\Big(\int \log \pi(y) \prod_{j\neq i} q_j(y_j)\, d\mu_{-i}(y_{-i})\Big). \tag{3.6}$$

This depends of course on q_j, for $j \neq i$, and requires evaluation of the integral. In many examples the latter is feasible. An iterative procedure can then cycle through updating the q_i in turn, each time fixing the other densities at their current values, a procedure known as *coordinate ascent variational inference*, or *CAVI*. The iterations are repeated "until convergence". It is apparently unknown whether the algorithm converges, and, if it does, whether the limit achieves the global optimum.

Equation (3.6) corresponds to unrestrained optimization of q_i. The same coordinate ascent algorithm also applies to constrained optimization. When using a parametric model for q_i, the coordinate ascents can be given in terms of updates of the parameters of the model. In some "conjugate" examples, the unconstrained optimizer may also remain in a given parametric model, and be given by a parameter update at each step.

In the Bayesian context the target is the posterior density $\pi(\theta|\, X) \propto p_\theta(X)\pi(\theta)$, for a partitioned parameter $\theta = (\theta_1, \ldots, \theta_m)$, and the coordinate ascent (3.6) is given by

$$q_i^*(\theta_i) \propto \exp\Big(\mathrm{E}_{-i} \log(p_\vartheta(X)\pi(\vartheta))\Big), \tag{3.7}$$

where E_{-i} refers to the expectation relative to $\vartheta_{-i} = (\vartheta_j)_{j\neq i}$ under the distribution $\vartheta_j \overset{\text{ind}}{\sim} q_j$, for $j \neq i$.

Example 3.25 (Normal location-scale)**.** Suppose that the data $X = (X_1, \ldots, X_n)$ are a random sample $X_i \overset{\text{iid}}{\sim} N(\mu, 1/\tau)$. In Example 2.3 it is seen that the prior for the parameter $\theta = (\mu, \tau)$ given by $\tau \sim \Gamma(\alpha, \beta)$ and $\mu|\, \tau \sim N(\nu, \tau^{-1})$ is conjugate, and an explicit expression for the posterior is obtained. For illustration consider the mean-field variational approximation of the posterior in which the parameters μ and τ are independent, i.e. we assume that the variational class $\mathcal{Q}$ consists of all densities q of the form

$$q(\mu, \tau) = q_\mu(\mu) q_\tau(\tau),$$

for univariate densities q_μ and q_τ without any additional restriction on the form of the latter densities.

The joint density takes the form

$$p_{\mu,\tau}(X)\pi(\mu, \tau) \propto \exp\Big(\Big(\frac{n}{2} + \frac{1}{2} + \alpha - 1\Big)\log\tau - \frac{\tau}{2}\sum_{i=1}^n (X_i - \mu)^2 - \frac{\tau}{2}(\mu - \nu)^2 - \beta\tau\Big).$$

In view of (3.7), the variational Bayes solutions satisfy

$$q_\mu^*(\mu) \propto \exp\Big(-\frac{\mathrm{E}_\tau\tau}{2}\Big(\sum_{j=1}^n (X_j - \mu)^2 + (\mu - \nu)^2\Big)\Big) = \exp\Big(-\frac{(1+n)\mathrm{E}_\tau\tau}{2}\Big(\mu - \frac{\nu + \sum_{j=1}^n X_j}{n+1}\Big)^2\Big),$$

$$q_\tau^*(\tau) \propto \exp\Big(\Big(\frac{n}{2} + \frac{1}{2} + \alpha - 1\Big)\log\tau - \frac{\tau}{2}\, E_\mu\Big(\sum_{j=1}^n (X_j - \mu)^2 + (\mu - \nu)^2\Big) - \beta\tau\Big).$$

Here E_τ and E_μ denote the expectations with respect to the variational densities q_τ and q_μ.

We conclude that q_μ^* is the density of the $N((\nu + n\bar{X}_n)/(n+1), 1/(\mathrm{E}_\tau\tau(1+n)))$-distribution, and q_τ^* is the density of the $\Gamma(\alpha + n/2, \beta + \mathrm{E}_\mu(\sum_{j=1}^n (X_j - \mu)^2 + (\mu - \nu)^2)/2)$-distribution.

In practice, the formulas are used recursively to update the two distributions, which boils down to updating the parameters of the normal and Gamma distributions. The update for q_μ^* depends on $\mathrm{E}_\tau\tau$, while the update of q_τ^* depends on the two moments $\mathrm{E}_\mu\mu$ and $\mathrm{E}_\mu\mu^2$.

Example 3.26 (Multivariate normal)**.** Suppose that the target distribution Π is the $N_m(\mu, \Omega^{-1})$-distribution, for given mean vector μ and precision matrix Ω. Unless Ω is diagonal, the coordinates of Π are not independent and hence a mean field variational approximation cannot be exact. In this example we derive that it is equal to the product of the univariate normal distributions $N(\mu_i, 1/\Omega_{i,i})$. Thus the marginals of the mean field approximation are normal, like the marginals of Π, with the same means μ_i. Unfortunately, the marginals of $N(\mu, \Omega^{-1})$ have variances $(\Omega^{-1})_{i,i}$, where

$$(\Omega^{-1})_{i,i} \geq \frac{1}{\Omega_{i,i}}.$$

Thus the marginals of the mean field approximation have smaller variance than the marginals of the true distribution.

The example of a multivariate normal distribution is relevant in Bayesian inference, as the posterior distribution of a multi-dimensional parameter $\theta = (\theta_1, \ldots, \theta_m)$ will typically be approximately a normal distribution $N_m(\hat{\theta}_n, I_\theta^{-1})$, for I_θ the Fisher information, in view of the Bernstein-von Mises theorem, Theorem 2.16. The marginals of the variational Bayes approximation to the posterior distribution then have a smaller variance than the true posterior marginals. The difference can be substantial if the Fisher information matrix is far from diagonal. The difference between the two variances is equal to the difference between the true Cramér-Rao bounds $(I_\theta^{-1})_{i,i}$ and $1/(I_\theta)_{i,i}$ for estimating θ_i when the other coordinates of θ are unknown or known, respectively. The second, the variational posterior variance is smaller.

To derive the variational approximation, we start from the log likelihood of the $N_m(\mu, \Omega^{-1})$ distribution

$$-\log \pi(\theta) = \tfrac{1}{2}(\theta - \mu)^T \Omega(\theta - \mu) - \tfrac{1}{2}\log\det\Omega + \tfrac{1}{2}m\log(2\pi).$$

We drop the last two terms and minimize over a product density q the expression

$$\begin{aligned}\int \log q(\theta)\, q(\theta)\, d\theta + \int \tfrac{1}{2}(\theta - \mu)^T \Omega(\theta - \mu)\, q(\theta)\, d\theta \\ = \sum_{i=1}^m \int \log q_i(\theta_i)\, q_i(\theta_i)\, d\theta_i + \tfrac{1}{2}\mathrm{tr}(\Omega\Sigma_q) + \tfrac{1}{2}(\mu_q - \mu)^T\Omega(\mu_q - \mu),\end{aligned}$$

where the q_i are the marginals of q, and μ_q and Σ_q the mean vector and covariance matrix of q. The first two terms on the right are invariant under shifting the q_i, while the third becomes zero if the q_i are centered at the μ_i. Furthermore, for diagonal $\Sigma_q = \mathrm{diag}(\sigma_{i,q}^2)$, the second term is equal to $\sum_{i=1}^m \Omega_{i,i}\sigma_{i,q}^2$. It follows that after shifting to mean zero, the problem becomes to minimize

$$\sum_{i=1}^m \Big[\int \log q_i(\theta_i)\, q_i(\theta_i)\, d\theta_i + \tfrac{1}{2}\Omega_{i,i}\sigma_{i,q}^2\Big],$$

where q_i is centered and has second moment $\sigma_{i,q}^2$. This splits in m separate minimization

problems. By Lemma 3.27 the normal distribution has maximal entropy $-\int \log q_i(\theta_i)\, q_i(\theta_i)\, d\theta_i$ among distributions with fixed variance, which can be computed to be $\frac{1}{2}\log(2\pi e\sigma_{i,q}^2)$. Hence we take q_i a centered normal density and need only minimize over its variance $\sigma_{i,q}^2$. This leads to minimization of $-\frac{1}{2}\log(2\pi e\sigma_{i,q}^2) + \frac{1}{2}\Omega_{i,i}\sigma_{i,q}^2$, and gives $\sigma_{i,q}^2 = 1/\Omega_{i,i}$. Thus the variational Bayes approximation to the $N_m(\mu, \Omega^{-1})$ distribution is the product of $N(\mu_i, 1/\Omega_{i,i})$-distributions.

3.11 Complements

Lemma 3.27. *The entropy $-\int(\log f(x))\, f(x)\, dx$ of a probability density f on $\mathbb{R}$ with variance 1 is maximal for the standard normal density.*

Proof For any centered density f with second moment 1, we have

$$0 \le \int (\log f/\phi) f\, d\lambda = \int (\log f) f\, d\lambda + \tfrac{1}{2}\log(2\pi e),$$

with equality if and only if $f = \phi$. Rearrange to see $\int(\log f) f\, d\lambda \ge -\frac{1}{2}\log(2\pi e) = \int(\log\phi)\phi\, d\lambda$. □

Exercises

3.1 Prove that the sequence (Y_n) produced by the multivariate slice sampler is a Markov chain with stationary distribution π.

3.2 Show that $(Y_n)_{n\ge1}$ is a Markov chain if and only if $(Y_1, \ldots, Y_{n-1})$ and $(Y_{n+1}, Y_{n+2}, \ldots)$ are conditionally independent given Y_n, for every n.

3.3 Implement a random-walk Metropolis-Hastings algorithm to compute the posterior distribution for θ based on a sample of size 50 from the $N(\theta, 1)$-distribution, relative to a standard Cauchy prior. Use $N(0, \tau^2)$ steps for the proposal distribution. Experiment with different values of τ. How does τ influence the acceptance rate? What seems a reasonable order of magnitude of τ? Give an estimate of the posterior density through a histogram of the MCMC output. Use data generated according to the model with $\theta_0 = 0$.

3.4 Implement the algorithm described in Example 3.8. Experiment with setting the burn-in period and by choosing different proposal distributions.

3.5 Implement the algorithm described in Example 3.10. Experiment with the choice of the scaling parameter τ. Using a test data set, try to find the scaling parameter which results in acceptance rate around 25%.

3.6 As data generate a sample $X_1, \ldots, X_n$ with of size $n = 50$ from the standard normal distribution. Suppose we want to fit a $N(\theta, 1)$ model to this data, using a Bayesian approach with a prior on θ defined hierarchically in two steps. The (hyper)parameter $1/\tau^2$ is Gamma distributed with parameters $(4/10, 2/10)$; given τ, the parameter θ is $N(0, \tau^2)$ distributed.

- Analytically derive the full conditionals of $\theta|\,\tau, X_1, \ldots, X_n$ and $1/\tau^2|\,\theta, X_1, \ldots, X_n$.
- Implement the Gibbs sampler to draw an approximate sample from the posterior distribution of (θ, τ).
- Use this sampler to estimate the marginal posterior mean of both θ and τ^2, and to form a 95% credible interval for θ.

3.7 Let $Y = f(X)$ be a measurable transformation of some random vector X with distribution Π and let $Q(y, B) = \Pr(X \in B|\, Y = y)$ be a version of the conditional distribution of X given $Y = f(X)$.

If we generate X from Π, coarsen X to $Y = f(X)$ and next generate Z from $Q(Y, \cdot)$, then Z possesses law Π. Show this.

3.8 Implement a MCMC scheme to estimate the parameter β in a linear regression model $Y_i = \beta X_i + e_i$, where $X_1, \ldots, X_n$ and $e_1, \ldots, e_n$ are assumed to be independently generated from a normal distribution, and where we assume that we observe only $Y_1, \ldots, Y_n$ and $X_{m+1}, \ldots, X_n$ for some $m \geq 2$. (E.g. $n = 10$, $m = 5$.) Use a Gaussian prior on β and set the error variance equal to 1.

3.9 Show that the Gibbs sampler is reversible only in the trivial case that the coordinates of $Y = (Y_1, \ldots, Y_m) | \pi$ are independent. (A proof for $m = 2$ suffices.)

3.10 Implement the variational Bayes algorithm described in Example 3.25. Compare (empirically) the resulting variational Bayes posterior with the true posterior.

3.11 Show that the leap-frog updates of the Hamiltonian MCMC algorithm for the Hamiltonian function $H(y, v) = y^2/(2\sigma^2) + v^2/2$ are given by the linear map with matrix

$$\begin{pmatrix} 1 - \epsilon^2/(2\sigma^2) & \epsilon \\ -\epsilon/\sigma^2 + \epsilon^3/(4\sigma^2) & 1 - \epsilon^2/(2\sigma^2) \end{pmatrix}.$$

Show that the updates are stable if $\epsilon < 2\sigma$. Implement Hamiltonian MCMC algorithms to sample from the normal distribution ($\sigma = 1$) with step sizes $\epsilon = 1$, $\epsilon = 3/2$ and $\epsilon = 3$, keeping the path length $L\epsilon$ constant. What are the acceptance rates? Conclusion?

4

Dirichlet Process

The Dirichlet process is the "normal distribution of Bayesian nonparametrics". It is the default prior on spaces of probability measures, and a building block for priors on other structures.

Throughout this chapter let the sample space $(\mathfrak{X}, \mathscr{X})$ be a Polish space with its Borel σ-field, and let $\mathfrak{M}$ be the collection of all probability measures on $(\mathfrak{X}, \mathscr{X})$. The Dirichlet process prior will be defined as a probability measure on $\mathfrak{M}$.

4.1 Random measures

In this section we explain the measure-theoretic details of putting a prior on a probability measure.

A prior on the set of probability measures $\mathfrak{M}$ is a probability measure on a σ-field $\mathscr{M}$ of subsets of $\mathfrak{M}$. Alternatively, it can be viewed as the law of a *random measure*: a measurable map from some probability space into $(\mathfrak{M}, \mathscr{M})$. As usual we denote the prior by Π; we denote the random measure by P. Thus Π is a probability measure on $(\mathfrak{M}, \mathscr{M})$ and P is a random element with values in $\mathfrak{M}$, and the correspondence is that $P \sim \Pi$.

It is natural to choose the σ-field $\mathscr{M}$ such that at least every $P(A)$, for $A \in \mathscr{X}$, is a random variable, i.e. the concatenation of the map P from the probability space into $\mathfrak{M}$ and the map $M \mapsto M(A)$ is a measurable map from $\mathfrak{M}$ into $[0, 1]$, for every $A \in \mathscr{X}$. We define the σ-field $\mathscr{M}$ as the minimal one to make this true.

Definition 4.1. Given the set $\mathfrak{M}$ of all Borel probability measures on a given Polish sample space $(\mathfrak{X}, \mathscr{X})$, the σ-field $\mathscr{M}$ is defined as the smallest σ-field on $\mathfrak{M}$ such that all maps $M \mapsto M(A)$ from $\mathfrak{M}$ to $\mathbb{R}$ are measurable, for $A \in \mathscr{X}$. A measurable map P from a probability space into $(\mathfrak{M}, \mathscr{M})$ is called a *random measure*.

With the preceding definition a map P from a probability space into $\mathfrak{M}$ is a random measure if and only if $P(A)$ is a random variable, for every $A \in \mathscr{X}$. (See Exercise 4.1.) Equivalently, $(P(A): A \in \mathscr{X})$ is a *stochastic process* on the underlying probability space. (A stochastic process is by definition just a collection of random variables defined on a common probability space. See Definition 7.1.)

Although other measurability structures on $\mathfrak{M}$ are possible, the σ-field $\mathscr{M}$ is attractive also because it is identical to the Borel σ-field for the weak topology on $\mathfrak{M}$. We state this fact in the following proposition, which also gives other useful characterizations. (See Section 4.9.1 for a review on the weak topology, also known as the topology of "convergence in distribution".)

Proposition 4.2. *If $(\mathfrak{X}, \mathscr{X})$ is a Polish space with its Borel σ-field, then the σ-field $\mathscr{M}$ of Definition 4.1 is also:*

(i) *the Borel σ-field for the weak topology on $\mathfrak{M}$;*
(ii) *the smallest σ-field on $\mathfrak{M}$ making all maps $M \mapsto M(A)$ measurable, for A in a generator $\mathscr{X}_0$ for $\mathscr{X}$;*
(iii) *the smallest σ-field on $\mathfrak{M}$ making all maps $M \mapsto \int \psi \, dM$ measurable, for bounded continuous functions $\psi\colon \mathfrak{X} \to \mathbb{R}$.*

A finite measure on $(\mathfrak{M}, \mathscr{M})$ is completely determined by the set of distributions induced under the maps $M \mapsto (M(A_1), \ldots, M(A_k))$, for $A_1, \ldots, A_k \in \mathscr{X}_0$ and $k \in \mathbb{N}$; and also under the maps $M \mapsto \int \psi \, dM$, for bounded continuous functions $\psi\colon \mathfrak{X} \to \mathbb{R}$.

For a proof see Proposition A.5 in Ghoshal and van der Vaart (2017).

As $\mathfrak{M}$ is Polish under the weak topology (see Proposition 4.21), the proposition implies that $(\mathfrak{M}, \mathscr{M})$ is a Polish space with its Borel σ-field. Thus taking the parameter θ from Section 1.1 that indexes the statistical model $(P_\theta\colon \theta \in \Theta)$ equal to the distribution P itself, and $\mathfrak{M}$ as the parameter set, we obtain a Polish parameter set. As is noted in Section 1.1, this is desirable for the definition of posterior distributions. We also note that, with respect to the σ-field $\mathscr{M}$ on the parameter set $\mathfrak{M}$, the data distributions are trivially "regular conditional probabilities": the map $(\theta, A) \mapsto P_\theta(A)$, which becomes $(P, A) \mapsto P(A)$, satisfies:

(i) $A \mapsto P(A)$ is a probability measure for every $P \in \mathfrak{M}$.
(ii) $P \mapsto P(A)$ is $\mathscr{M}$-measurable for every $A \in \mathscr{X}$,

This verifies our assumption in Section 1.1, that the measures P_θ in the statistical model are Markov kernels, and justifies speaking of "drawing a measure P from the prior Π and next sampling observations X from P".

Random measures as a stochastic process

By Kolmogorov's consistency theorem (see Proposition 4.18) a stochastic process $(P(A)\colon A \in \mathscr{X})$ can be constructed simply from a specification of the distributions of all "marginal" vectors $(P(A_1), \ldots, P(A_k))$, for all finite collections $A_1, \ldots, A_k$ of Borel sets in $\mathfrak{X}$. Indeed, Kolmogorov's theorem asserts that given any *consistent* specification of these distributions, there exists a suitable probability space $(\Omega, \mathscr{U}, \Pr)$ and a stochastic process $(P(A)\colon A \in \mathscr{X})$ defined on it with the given finite-dimensional distributions.

This falls short of constructing a random measure P, as the properties of measures are much richer than can be described by the finite-dimensional distributions. While a stochastic process is a measurable map from a probability space into $\mathbb{R}^{\mathscr{X}}$, a random measure is a measurable map in $\mathfrak{M}$. Through the identification $M \leftrightarrow (M(A)\colon A \in \mathscr{X})$, the space $\mathfrak{M}$ can be identified with a subset of $\mathbb{R}^{\mathscr{X}}$; the σ-field $\mathscr{M}$ is then the trace of the product σ-field on $\mathbb{R}^{\mathscr{X}}$, by Proposition 4.2. In the next theorem we give sufficient conditions that a stochastic process can indeed be realized as a random measure, i.e. that there exists a version of the stochastic process that takes its values in the smaller space $\mathfrak{M}$.

If the marginal distributions of the stochastic process $(P(A)\colon A \in \mathscr{X})$ correspond to those of a random measure, then it will be true that

(i) $P(\varnothing) = 0$, $P(\mathfrak{X}) = 1$, a.s.

(ii) $P(A_1 \cup A_2) = P(A_1) + P(A_2)$, a.s., for any disjoint A_1, A_2.

Assertion (i) follows, because the distributions of $P(\emptyset)$ and $P(\mathfrak{X})$ will be specified to be degenerate at 0 and 1, respectively, while (ii) can be read off from the degeneracy of the joint distribution of the three variables $P(A_1)$, $P(A_2)$ and $P(A_1 \cup A_2)$. Thus the process $(P(A): A \in \mathscr{X})$ will automatically define a *finitely-additive* measure on $(\mathfrak{X}, \mathscr{X})$.

A problem is that the exceptional null sets in (ii) might depend on the pair (A_1, A_2). If restricted to a countable sub-collection $\mathscr{X}_0 \subset \mathscr{X}$ there would only be countably many pairs and the null sets could be gathered in a single null set. This would give a version of P that is an additive measure on $\mathfrak{X}_0$ almost surely. Then still when extending (ii) to σ-additivity, which is typically possible by similar distributional arguments, there would be uncountably many sequences of sets. This problem can be overcome through existence of a *mean measure*

$$\mu(A) = \mathrm{E}P(A).$$

For a valid random measure P, this necessarily defines a Borel measure on $\mathfrak{X}$. Existence of a mean measure is also sufficient for existence of version of $(P(A): A \in \mathscr{X})$ that is a Borel measure on $\mathfrak{X}$.

Theorem 4.3. *Suppose that $(P(A): A \in \mathscr{X})$ is a stochastic process defined on a probability space $(\Omega, \mathscr{U}, \Pr)$ that satisfies (i) and (ii) and whose mean $A \mapsto \mathrm{E}P(A)$ is a Borel measure on $\mathfrak{X}$. Then there exists a version of P that is a random Borel measure on $\mathfrak{X}$. More precisely, there exists a measurable map $\tilde{P}: (\Omega, \mathscr{U}, \Pr) \to (\mathfrak{M}, \mathscr{M})$ such that $P(A) = \tilde{P}(A)$ almost surely, for every $A \in \mathscr{X}$.*

Proof Let $\mathscr{X}_0$ be a countable field that generates the Borel σ-field $\mathscr{X}$, enumerated arbitrarily as $A_1, A_2, \ldots$. Because the mean measure $\mu(A) := \mathrm{E}P(A)$ is regular, there exists for every $i, m \in \mathbb{N}$ a compact set $K_{i,m} \subset A_i$ with $\mu(A_i - K_{i,m}) < 2^{-2i-2m}$. By Markov's inequality

$$\Pr(P(A_i - K_{i,m}) > 2^{-i-m}) \leq 2^{i+m}\mathrm{E}P(A_i - K_{i,m}) \leq 2^{-i-m}.$$

Consequently, the event $\Omega_m = \cap_i \{P(A_i - K_{i,m}) \leq 2^{-i-m}\}$ possesses probability at least $1 - 2^{-m}$, and $\liminf \Omega_m$ possesses probability 1, by the Borel-Cantelli lemma.

Because $\mathscr{X}_0$ is countable, the null sets involved in (i)-(ii) with $A_1, A_2 \in \mathscr{X}_0$ can be aggregated into a single null set N. For every $\omega \notin N$ the process P is a finitely additive measure on $\mathscr{X}_0$, with the resulting usual properties of monotonicity and sub-additivity. By increasing N if necessary we can also ensure that it is sub-additive on all finite unions of sets $A_i - K_{i,m}$.

Let $A_{i_1} \supset A_{i_2} \supset \cdots$ be an arbitrary decreasing sequence of sets in $\mathscr{X}_0$ with empty intersection. Then, for every fixed m, the corresponding compacts $K_{i_j,m}$ possess empty intersection also, whence there exists a finite J_m such that $\cap_{j \leq J_m} K_{i_j,m} = \emptyset$. This implies that

$$A_{i_{J_m}} = \cap_{j=1}^{J_m} A_{i_j} - \cap_{j=1}^{J_m} K_{i_j,m} \subset \cup_{j=1}^{J_m} (A_{i_j} - K_{i_j,m}).$$

Consequently, on the event $\Omega_m \setminus N$,

$$\limsup_j P(A_{i_j}) \leq P(A_{i_{J_m}}) \leq \sum_{j=1}^{J_m} P(A_{i_j} - K_{i_j,m}) \leq 2^{-m}.$$

Thus on the event $\Omega_0 = \liminf \Omega_m \setminus N$ the limit is zero. We conclude that for every $\omega \in \Omega_0$,

the restriction of $A \mapsto P(A)$ to $\mathscr{X}_0$ is countably additive. By Carathéodory's theorem it extends to a measure $\tilde{P}$ on $\mathscr{X}$.

By construction $\tilde{P}(A) = P(A)$ almost surely, for every $A \in \mathscr{X}_0$. In particular $\mathrm{E}\tilde{P}(A) = \mathrm{E}P(A) = \mu(A)$, for every A in the field $\mathscr{X}_0$, whence by uniqueness of extension the mean measure of $\tilde{P}$ coincides with the original mean measure μ on $\mathscr{X}$. For every $A \in \mathscr{X}$, there exists a sequence $\{A_m\} \subset \mathscr{X}_0$ such that $\mu(A \,\Delta\, A_m) \to 0$. Then both $P(A_m \,\Delta\, A)$ and $\tilde{P}(A_m \,\Delta\, A)$ tend to zero in mean. Finite-additivity of P gives that $\left|P(A_m) - P(A)\right| \leq P(A_m \,\Delta\, A)$, almost surely, and by σ-additivity the same is true for $\tilde{P}$. This shows that $\tilde{P}(A) = P(A)$ almost surely, for every $A \in \mathscr{X}$.

This also proves that $\tilde{P}(A)$ is a random variable for every $A \in \mathscr{X}$, whence $\tilde{P}$ is a measurable map in $(\mathfrak{M}, \mathscr{M})$. □

Discrete random measures

Given random variables $W_1, W_2, \ldots$ with $0 \leq W_j \leq 1$ and $\sum_{j=1}^{\infty} W_j = 1$, and random variables $\theta_1, \theta_2, \ldots$ with values in $(\mathfrak{X}, \mathscr{X})$, we can define a random probability measure by

$$P = \sum_{j=1}^{\infty} W_j \delta_{\theta_j}. \tag{4.1}$$

The realizations of this measure are discrete with countably many support points, which may be different for each realization. The vectors of weights $(W_1, W_2, \ldots)$ and "locations" $(\theta_1, \theta_2, \ldots)$ are often chosen independent.

The *support* of a Borel probability measure on a Polish space is defined as the smallest closed set with probability one. Equivalently, it is the set of all points such that every open neighborhood of the point receives positive probability. The *support* of a random variable with values in a Polish space is understood to be the support of its induced distribution. The following lemma shows that the random probability measure (4.1) has "full support" (i.e. support the whole space) in the set of measures $(\mathfrak{M}, \mathscr{M})$, under the mild conditions that the weight vector and the location variables have full support. For the weight vector this refers to the infinite-dimensional unit simplex $\mathbb{S}_\infty$: the set of all probability vectors $(w_1, w_2, \ldots)$.

Lemma 4.4. *If the weight vector $(W_1, W_2, \ldots)$ has support $\mathbb{S}_\infty$ and is independent of the vector $(\theta_1, \theta_2, \ldots)$ of locations, and the vector $(\theta_1, \ldots, \theta_k)$ has support $\mathfrak{X}^k$, for every k, then the random measure P defined in (4.1) has weak support $\mathfrak{M}$, i.e. $\Pr(P \in B) > 0$ for every weakly open subset $B \subset \mathfrak{M}$.*

Proof Every probability measure Q on a Polish space $(\mathfrak{X}, \mathscr{X})$ is the limit in distribution (or "weak limit") of a sequence of finitely discrete distributions (see Proposition 4.22). Therefore, it suffices that to show that $\Pr(P \in B) > 0$ for every weak neighborhood B of a distribution $P^* = \sum_{j=1}^{k} w_j^* \delta_{\theta_j^*}$ with finite support. All distributions $P' = \sum_{j=1}^{\infty} w_j \delta_{\theta_j}$ with $\sum_{j>k} w_j$ sufficiently small, for some given k, and $(w_1, \ldots, w_k)$ and $(\theta_1, \ldots, \theta_k)$ sufficiently close to $(w_1^*, \ldots, w_k^*)$ and $(\theta_1^*, \ldots, \theta_k^*)$ are in such a neighborhood B. Thus the assertion follows if for every $\epsilon > 0$ the intersection of the events $\{\sum_{j>k} W_j < \epsilon, \max_{j \leq k} |W_j - w_j^*| < \epsilon\}$ and $\{\max_{j \leq k} d(\theta_j, \theta_j^*) < \epsilon\}$ has positive probability. Now both events have positive probability, as they refer to open subsets of $\mathbb{S}_\infty$ and $\mathfrak{X}^k$, respectively, and hence so does their intersection, as the events are independent. □

An important special case is obtained by choosing the locations $\theta_1, \theta_2, \ldots$ an i.i.d. sequence in $\mathfrak{X}$, and by choosing the weights $W_1, W_2, \ldots$ by the following *stick-breaking algorithm*. Given a sequence of random variables $Y_1, Y_2, \ldots$ with values between 0 and 1, we set

$$W_j = \Big(\prod_{l=1}^{j-1}(1 - Y_l)\Big)Y_j. \tag{4.2}$$

The idea is to divide the total weight (a "stick" of length 1) first in parts Y_1 and $1 - Y_1$; the first weight W_1 is set equal to Y_1. Next the remaining weight $1 - Y_1$ is divided in $(1 - Y_1)Y_2$ and $(1 - Y_1)(1 - Y_2)$; the second weight W_2 is set equal to $(1 - Y_1)Y_2$. The remaining weight is divided in $(1 - Y_1)(1 - Y_2)Y_3$ and $(1 - Y_1)(1 - Y_2)(1 - Y_3)$, etc.

Under mild conditions on the sequence $Y_1, Y_2, \ldots$, the probabilities W_j will sum to one, and the distribution will have full support.

Lemma 4.5. *For any random variables Y_l with $0 \le Y_l \le 1$, the random vector $(W_1, W_2, \ldots)$ defined by (4.2) has nonnegative coordinates with $\sum_j W_j \le 1$. We have $\sum_j W_j = 1$ almost surely iff $\mathrm{E}[\prod_{l=1}^{j}(1 - Y_l)] \to 0$ as $j \to \infty$. For independent variables $Y_1, Y_2, \ldots$ this condition is equivalent to $\sum_{l=1}^{\infty} \log \mathrm{E}(1 - Y_l) = -\infty$. In particular, for i.i.d. variables $Y_1, Y_2, \ldots$ it suffices that $\mathrm{P}(Y_1 > 0) > 0$. Finally, for any independent variables $Y_1, Y_2, \ldots$ with support $[0, 1]$, the vector $(W_1, W_2, \ldots)$ has support $\mathbb{S}_\infty$.*

Proof The first assertion is immediate from the construction. By induction, it easily follows that the leftover mass at stage j is equal to $1 - \sum_{l=1}^{j} W_l = \prod_{l=1}^{j}(1 - Y_l)$. Hence $\sum_j W_j = 1$ a.s. iff $\prod_{l=1}^{j}(1 - Y_l) \to 0$ a.s.. Since the leftover sequence is decreasing, nonnegative and bounded by 1, the almost sure convergence is equivalent to convergence in mean. If the Y_j's are independent, then this condition becomes $\prod_{l=1}^{j}(1 - \mathrm{E}Y_l) \to 0$ as $j \to \infty$, which is equivalent to the condition $\sum_{l=1}^{\infty} \log \mathrm{E}(1 - Y_l) = -\infty$.

Because the map $y \mapsto w$ from stick lengths to weights is continuous, it follows that the event $\{(W_1, W_2, \ldots) \in B\}$ for an open set $B \subset \mathbb{S}_\infty$ translates into an event $\{(Y_1, Y_2, \ldots) \in C\}$ for an open set $C \subset [0, 1]^\infty$. Hence the sequence of weights has full support as soon as the sequence of stick lengths has full support. Now the cylindrical sets, of the form $C = \{(y_1, y_2, \ldots) : (y_1, \ldots, y_k) \in C_k\}$, for $C_k \subset \mathbb{R}^k$ an open interval, form a basis of the open sets of $[0, 1]^\infty$, and the probability that $(Y_1, \ldots, Y_k) \in C_k$ is positive, for every such C_k. Thus the vector $(Y_1, Y_2, \ldots)$ has support $[0, 1]^\infty$. □

4.2 Definition and existence

Definition 4.6 (Dirichlet process)**.** A random measure P on $(\mathfrak{X}, \mathscr{X})$ is said to possess a *Dirichlet process* distribution $\mathrm{DP}(\alpha)$ with *base measure* α, if for every finite measurable partition $A_1, \ldots, A_k$ of $\mathfrak{X}$,

$$(P(A_1), \ldots, P(A_k)) \sim \mathrm{Dir}(k; \alpha(A_1), \ldots, \alpha(A_k)). \tag{4.3}$$

In this definition α is a given finite positive Borel measure on $(\mathfrak{X}, \mathscr{X})$. We write $|\alpha| = \alpha(\mathfrak{X})$ for its total mass and $\bar{\alpha} = \alpha/|\alpha|$ for the probability measure obtained by normalizing α, respectively, and use the notations $P \sim \mathrm{DP}(\alpha)$ and $P \sim \mathrm{DP}(|\alpha|, \bar{\alpha})$ interchangeably to say that P has a Dirichlet process distribution with base measure α.

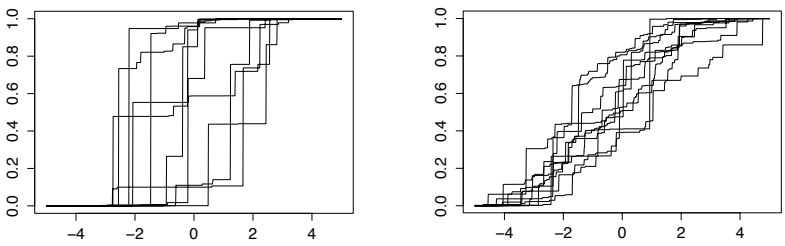

Figure 4.1 Cumulative distribution functions of 10 draws from the Dirichlet process with base measures $N(0, 2)$ (left) and $10N(0, 2)$ (right). (Computations based on stick-breaking representation truncated to 1000 terms.)

The existence of the Dirichlet process is not obvious, but can be proved using Theorem 4.3.

Theorem 4.7. *The Dirichlet process* $\mathrm{DP}(\alpha)$ *exists: for any finite Borel measure* α *on the Polish space* $\mathfrak{X}$ *there exists a random measure* P *satisfying* (4.3) *for every finite measurable partition* $A_1, \dots, A_k$ *of* $\mathfrak{X}$.

Proof Definition 4.6 specifies the joint distribution of the vector $(P(A_1), \dots, P(A_k))$, for any measurable partition $\{A_1, \dots, A_k\}$ of the sample space. In particular, it specifies the distribution of $P(A)$, for every measurable set A, and hence the mean measure $A \mapsto \mathrm{E}[P(A)]$. By Proposition 2.27,

$$\mathrm{E}P(A) = \bar{\alpha}(A).$$

Thus the mean measure is the normalized base measure $\bar{\alpha}$, which is a valid Borel measure by assumption. Therefore Theorem 4.3 implies existence of the Dirichlet process $\mathrm{DP}(\alpha)$ provided the specification of distributions can be consistently extended to any vector of the type $(P(A_1), \dots, P(A_k))$, for arbitrary measurable sets and not just partitions, in such a way that it gives a finitely-additive measure.

An arbitrary collection $A_1, \dots, A_k$ of measurable sets defines a collection of 2^k atoms of the form $A_1^* \cap A_2^* \cap \cdots \cap A_k^*$, where A^* stands for A or A^c. These atoms $\{B_j \colon j = 1, \dots, 2^k\}$ (some of which may be empty) form a partition of the sample space, and hence the joint distribution of $(P(B_j) \colon j = 1, \dots, 2^k)$ is defined by Definition 4.6. Every A_i can be written as a union of atoms, and $P(A_i)$ can be defined accordingly as the sum of the corresponding $P(B_j)$'s. This defines the distribution of the vector $(P(A_1), \dots, P(A_k))$.

To prove the existence of a stochastic process $(P(A) \colon A \in \mathscr{X})$ that possesses these marginal distributions, it suffices to verify that this collection of marginal distributions is consistent in the sense of Kolmogorov's extension theorem. Consider the distribution of the vector $(P(A_1), \dots, P(A_{k-1}))$. This has been defined using the coarser partitioning in the 2^{k-1} sets of the form $A_1^* \cap A_2^* \cap \cdots \cap A_{k-1}^*$. Every set in this coarser partition is a union of two sets in the finer partition used previously to define the distribution of $(P(A_1), \dots, P(A_k))$. Therefore, consistency pertains if the distributions specified by Definition 4.6 for two partitions, where one is finer than the other, are consistent.

Let $\{A_1, \dots, A_k\}$ be a measurable partition and let $\{A_{i1}, A_{i2}\}$ be a further measurable parti-

tion of A_i, for $i = 1, \ldots, k$. Then Definition 4.6 specifies that

$$(P(A_{11}), P(A_{12}), P(A_{21}), \ldots, P(A_{k1}), P(A_{k2})) \\ \sim \mathrm{Dir}(2k; \alpha(A_{11}), \alpha(A_{12}), \alpha(A_{21}), \ldots, \alpha(A_{k1}), \alpha(A_{k2})).$$

In view of the group additivity of finite dimensional Dirichlet distributions given by Proposition 2.26, this implies

$$\Big(\sum_{j=1}^{2} P(A_{1j}), \ldots, \sum_{j=1}^{2} P(A_{kj})\Big) \sim \mathrm{Dir}\Big(k; \sum_{j=1}^{2} \alpha(A_{1j}), \ldots, \sum_{j=1}^{2} \alpha(A_{kj})\Big).$$

Consistency follows as the right side is $\mathrm{Dir}(k; \alpha(A_1), \ldots, \alpha(A_k))$, since α is a measure.

That $P(\varnothing) = 0$ and $P(\mathfrak{X}) = 1$ almost surely follow from the fact that $\{\varnothing, \mathfrak{X}\}$ is an eligible partition in Definition 4.6, whence $(P(\varnothing), P(\mathfrak{X})) \sim \mathrm{Dir}(2; 0, |\alpha|)$ by (4.3). That $P(A_1 \cup A_2) = P(A_1) + P(A_2)$ almost surely for every disjoint pair of measurable sets A_1, A_2, follows similarly from consideration of the distributions of the vectors $(P(A_1), P(A_2), P(A_1^c \cap A_2^c))$ and $(P(A_1 \cup A_2), P(A_1^c \cap A_2^c))$, whose three and two components both add up to 1. □

4.3 Stick-breaking representation

The Dirichlet process can be represented as a random discrete measure of the type discussed in Section 4.1. This representation gives an easy method to simulate a Dirichlet process, at least approximately. It also proves the remarkable fact that realizations from the Dirichlet measure are discrete measures, with probability one.

The weights are given by the stick-breaking algorithm with relative stick lengths from the Beta-distribution. The random support points are generated from the normalized base measure.

Theorem 4.8 (Stick-breaking). *If $\theta_1, \theta_2, \ldots \overset{iid}{\sim} \bar{\alpha}$ and $Y_1, Y_2, \ldots \overset{iid}{\sim} \mathrm{Be}(1, M)$ are independent random variables and $W_j = Y_j \prod_{l=1}^{j-1}(1 - Y_l)$, then $\sum_{j=1}^{\infty} W_j \delta_{\theta_j} \sim \mathrm{DP}(M\bar{\alpha})$.*

Proof Because $\mathrm{E}(\prod_{l=1}^{j}(1 - Y_l)) = (M/(M+1))^j \to 0$, the stick-breaking weights W_j form a probability vector a.s. (c.f. Lemma 4.5), so that P is a probability measure a.s..

For $j \geq 2$ define $W'_{j-1} = Y_j \prod_{l=2}^{j-1}(1 - Y_l)$ and $\theta'_j = \theta_{j+1}$. Then $W_j = (1 - Y_1)W'_{j-1}$ for every $j \geq 1$ and hence

$$P := W_1 \delta_{\theta_1} + \sum_{j=2}^{\infty} W_j \delta_{\theta_j} = Y_1 \delta_{\theta_1} + (1 - Y_1) \sum_{j=1}^{\infty} W'_j \delta_{\theta'_j}.$$

The random measure $P' := \sum_{j=1}^{\infty} W'_j \delta_{\theta'_j}$ has exactly the same structure as P, and hence possesses the same distribution. Furthermore, it is independent of (Y_1, θ_1).

We conclude that P satisfies the distributional equation (4.4) given below, and the theorem follows from Lemma 4.9. □

For independent random variables $Y \sim \mathrm{Be}(1, |\alpha|)$ and $\theta \sim \bar{\alpha}$, consider the equation

$$P =_d Y\delta_\theta + (1 - Y)P. \tag{4.4}$$

We say that a random measure P that is independent of (Y, θ) is a solution to equation (4.4) if

for every measurable partition $\{A_1, \dots, A_k\}$ of the sample space the random vectors obtained by evaluating the random measures on its left and right sides are equal in distribution in $\mathbb{R}^k$.

Lemma 4.9. *For given independent $\theta \sim \bar{\alpha}$ and $Y \sim \mathrm{Be}(1, |\alpha|)$, the Dirichlet process $\mathrm{DP}(\alpha)$ is the unique solution of the distributional equation (4.4).*

Proof For a given measurable partition $\{A_1, \dots, A_k\}$, the equation requires that the random vector $Q := (P(A_1), \dots, P(A_k))$ has the same distribution as the vector $YN + (1 - Y)Q$, for $N \sim \mathrm{MN}(1; \bar{\alpha}(A_1), \dots, \bar{\alpha}(A_k))$ and (Y, N) independent of Q.

We first show that the solution is unique in distribution. Let (Y_n, N_n) be a sequence of i.i.d. copies of (Y, N), and for two solutions Q and Q' that are independent of this sequence and suitably defined on the same probability space, set $Q_0 = Q$, $Q_0' = Q'$, and recursively define $Q_n = Y_n N_n + (1 - Y_n)Q_{n-1}$, $Q_n' = Y_n N_n + (1 - Y_n)Q_{n-1}'$, for $n \in \mathbb{N}$. Then every Q_n is distributed as Q and every Q_n' is distributed as Q', because each of them satisfies the distributional equation. Also

$$\|Q_n - Q_n'\| = |1 - Y_n| \, \|Q_{n-1} - Q_{n-1}'\| = \prod_{i=1}^{n} |1 - Y_i| \, \|Q - Q'\| \to 0$$

with probability 1, since the Y_i are i.i.d. and are in $(0, 1)$ with probability one. This forces the distributions of Q and Q' to agree.

To prove that the Dirichlet process is a solution (4.4), let $W_0, W_1, \dots, W_k \overset{\text{ind}}{\sim} \mathrm{Ga}(\alpha_i, 1)$, for $i = 0, 1, \dots, k$, where $\alpha_0 = 1$. Then by Proposition 2.26 the vector (W_0, W), for $W = \sum_{i=1}^{k} W_i$, is independent of the vector $Q := (W_1/W, \dots, W_k/W) \sim \mathrm{Dir}(k, \alpha_1, \dots, \alpha_k)$. Furthermore, $Y := W_0/(W_0 + W) \sim \mathrm{Be}(1, |\alpha|)$ and $(Y, (1 - Y)Q) \sim \mathrm{Dir}(k + 1; 1, \alpha_1, \dots, \alpha_k)$. Thus for any $i = 1, \dots, k$, merging the 0th cell with the ith, we obtain from Proposition 2.26 that, with e_i the ith unit vector,

$$Ye_i + (1 - Y)Q \sim \mathrm{Dir}(k; \alpha + e_i), \quad i = 1, \dots, k. \tag{4.5}$$

This gives the conditional distribution of the vector $YN + (1 - Y)Q$ given $N = e_i$. It follows that $YN + (1 - Y)Q$ given N possesses a $\mathrm{Dir}(k; \alpha + N)$-distribution, just as θ given X in Example 2.2. Because also the marginal distributions of N and X in the two cases are the same, so must be the marginal distributions of $YN + (1 - Y)Q$ and θ, where the latter is $\theta \sim \mathrm{Dir}(k; \alpha)$. □

4.4 Tail-freeness

Consider a sequence $\mathcal{T}_0 = \{\mathfrak{X}\}$, $\mathcal{T}_1 = \{A_0, A_1\}$, $\mathcal{T}_2 = \{A_{00}, A_{01}, A_{10}, A_{11}\}$, and so on, of measurable partitions of the sample space $\mathfrak{X}$, obtained by splitting every set in the preceding partition into two new sets. See Figure 4.2.

With $\mathcal{E} = \{0, 1\}$ and $\mathcal{E}^* = \cup_{m=0}^{\infty} \mathcal{E}^m$, the set of all finite strings $\varepsilon_1 \cdots \varepsilon_m$ of 0's and 1's, we can index the 2^m sets in the mth partition $\mathcal{T}_m$ by $\varepsilon \in \mathcal{E}^m$, in such a way that $A_\varepsilon = A_{\varepsilon 0} \cup A_{\varepsilon 1}$ for every $\varepsilon \in \mathcal{E}^*$. Here $\varepsilon 0$ and $\varepsilon 1$ are the extensions of the string ε with a single symbol 0 or 1; the empty string indexes $\mathcal{T}_0$. Let $|\varepsilon|$ stand for the length of a string ε, and let $\varepsilon\delta$ be the concatenation of two strings $\varepsilon, \delta \in \mathcal{E}^*$.

Because the probability of any A_ε must be distributed to its "offspring" $A_{\varepsilon 0}$ and $A_{\varepsilon 1}$, a

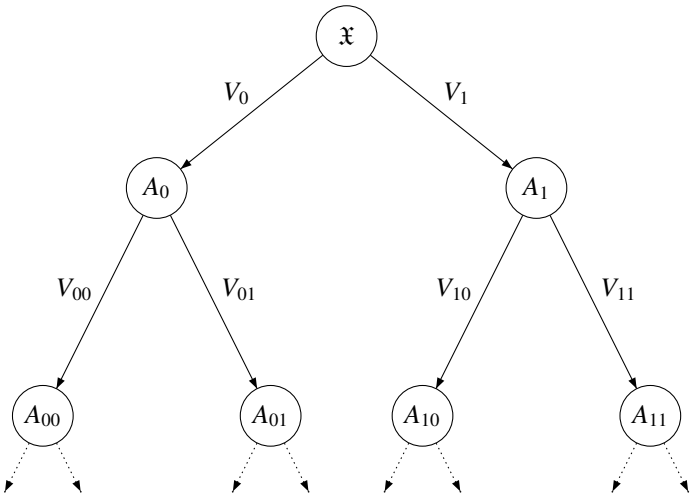

Figure 4.2 Tree diagram showing the distribution of mass over the first two partitions $\mathfrak{X} = A_0 \cup A_1 = (A_{00} \cup A_{01}) \cup (A_{10} \cup A_{11})$ of the sample space. Mass at a given node is distributed to its two children proportionally to the weights on the arrows. Every pair of V's on arrows originating from the same node add to 1.

probability measure P must satisfy the *tree additivity* requirement $P(A_\varepsilon) = P(A_{\varepsilon 0}) + P(A_{\varepsilon 1})$. The relative weights of the offspring sets are the conditional probabilities

$$V_{\varepsilon 0} = P(A_{\varepsilon 0} | A_\varepsilon), \qquad \text{and} \qquad V_{\varepsilon 1} = P(A_{\varepsilon 1} | A_\varepsilon). \tag{4.6}$$

With V_0 and V_1 interpreted as the unconditional probabilities $P(A_0)$ and $P(A_1)$, it follows that

$$P(A_{\varepsilon_1 \cdots \varepsilon_m}) = V_{\varepsilon_1} V_{\varepsilon_1 \varepsilon_2} \cdots V_{\varepsilon_1 \cdots \varepsilon_m}, \qquad \varepsilon = \varepsilon_1 \cdots \varepsilon_m \in \mathcal{E}^m. \tag{4.7}$$

In Figure 4.2 this corresponds with multiplying the weights on the arrows leading to a node at the bottom level.

Write $U \perp\!\!\!\perp V$ to denote that random variables U and V are independent, and $U \perp\!\!\!\perp V | Z$ to say that U and V are conditionally independent given a random variable Z.

Definition 4.10 (Tail-free)**.** The random measure P is a *tail-free process* with respect to the sequence of partitions $\mathcal{T}_m$ if $\{V_0\} \perp\!\!\!\perp \{V_{00}, V_{10}\} \perp\!\!\!\perp \cdots \perp\!\!\!\perp \{V_{\varepsilon 0} \colon \varepsilon \in \mathcal{E}^m\} \perp\!\!\!\perp \cdots$.

A degenerate prior is certainly tail-free according to this definition (with respect to any sequence of partitions), since all its V-variables are degenerate at appropriate values. The Dirichlet process is a more interesting example.

Theorem 4.11. *The* DP(α) *prior is tail-free. All splitting variables* $V_{\varepsilon 0}$ *are independent and* $V_{\varepsilon 0} \sim \text{Be}(\alpha(A_{\varepsilon 0}), \alpha(A_{\varepsilon 1}))$.

Proof We must show that the vectors $(V_{\varepsilon 0}\colon \varepsilon \in \mathcal{E}^m)$ defined in (4.6) are mutually independent across levels m. It suffices to show sequentially for every m that this vector is independent of the vectors corresponding to lower levels. Because the vectors $(V_\varepsilon\colon \varepsilon \in \cup_{k\le m}\mathcal{E}^k)$, and $(P(A_\varepsilon)\colon \varepsilon \in \mathcal{E}^m)$ generate the same σ-field, it suffices to show that $(V_{\varepsilon 0}\colon \varepsilon \in \mathcal{E}^m)$ is independent of $(P(A_\varepsilon)\colon \varepsilon \in \mathcal{E}^m)$, for every fixed m.

This follows by an application of Proposition 2.26 to the vectors $(P(A_{\varepsilon\delta})\colon \varepsilon \in \mathcal{E}^m, \delta \in \mathcal{E})$ with the aggregation of the pairs $(P(A_{\varepsilon 0}), P(A_{\varepsilon 1}))$ into the sums $P(A_\varepsilon) = P(A_{\varepsilon 0}) + P(A_{\varepsilon 1})$.

The beta distributions also follow by Proposition 2.26 (and the fact that the first marginal of a $\mathrm{Dir}(2;\alpha,\beta)$ is a $\mathrm{Be}(\alpha,\beta)$-distribution). □

The mass $P(A_\varepsilon)$ of a partitioning set at level m can be expressed in the V-variables up to level m (see (4.7)), while, by their definition (4.6), the V-variables at higher levels control conditional probabilities. Therefore, tail-freeness makes the distribution of mass *within* every partitioning set in $\mathcal{T}_m$ independent of the distribution of the total mass one *among* the sets in $\mathcal{T}_m$. Definition 4.10 refers only to masses of partitioning sets, but under the assumption that the partitions generate the Borel sets, the independence extends to all Borel sets.

Lemma 4.12. *If P is a random measure that is tail-free relative to a sequence of partitions $\mathcal{T}_m = \{A_\varepsilon\colon \varepsilon \in \mathcal{E}^m\}$ that generates the Borel sets $\mathcal{X}$ in $\mathfrak{X}$, then for every $m \in \mathbb{N}$ the process $(P(A|A_\varepsilon)\colon A \in \mathcal{X}, \varepsilon \in \mathcal{E}^m)$ is independent of the random vector $(P(A_\varepsilon)\colon \varepsilon \in \mathcal{E}^m)$.*

Proof Because P is a random measure, its mean measure $\mu(A) = \mathrm{E}[P(A)]$ is a well defined Borel probability measure. As $\mathcal{T} := \cup_m \mathcal{T}_m$ is a field, which generates the Borel σ-field by assumption, there exists for every $A \in \mathcal{X}$ a sequence A_n in $\mathcal{T}$ such that $\mu(A_n \,\Delta\, A) \to 0$. Because P is a random measure $P(A_n|A_\varepsilon) \to P(A|A_\varepsilon)$ in mean and hence a.s. along a subsequence. It follows that the random variable $P(A|A_\varepsilon)$ is measurable relative to the completion $\mathcal{U}_0$ of the σ-field generated by the variables $P(C|A_\varepsilon)$, for $C \in \mathcal{T}$. Every of the latter variables is a finite sum of probabilities of the form $P(A_{\varepsilon\delta}|A_\varepsilon) = V_{\varepsilon\delta_1}\cdots V_{\varepsilon\delta_1\cdots\delta_k}$, for $\varepsilon \in \mathcal{E}^m$, $\delta = \delta_1\cdots\delta_k \in \mathcal{E}^k$ and $k \in \mathbb{N}$. Therefore, by tail-freeness the σ-field $\mathcal{U}_0$ is independent of the σ-field generated by the variables $P(A_\varepsilon) = V_{\varepsilon_1}\cdots V_{\varepsilon_1\cdots\varepsilon_m}$, for $\varepsilon = \varepsilon_1\cdots\varepsilon_m \in \mathcal{E}^m$. □

Relative to the σ-field $\mathcal{M}$ the process $(P(A|A_\varepsilon)\colon A \in \mathcal{X})$ contains all information about the conditional random measure $P(\cdot|A_\varepsilon)$. Thus the preceding lemma truly expresses that the "conditional measure within partitioning sets is independent of the distribution of mass among them".

Suppose that the data consists of an i.i.d. sample $X_1, \ldots, X_n$ from a distribution P, which is a-priori modelled as a tail-free process. For each $\varepsilon \in \mathcal{E}^*$, denote the number of observations falling in A_ε by

$$N_\varepsilon := \#\{1 \le i \le n\colon X_i \in A_\varepsilon\}. \tag{4.8}$$

For each m the vector $(N_\varepsilon\colon \varepsilon \in \mathcal{E}^m)$ collects the counts of all partitioning sets at level m. The following theorem shows that this vector contains all information (in the Bayesian sense) about the probabilities $(P(A_\varepsilon)\colon \varepsilon \in \mathcal{E}^m)$ of these sets: the additional information about the precise positions of the X_i within the partitioning sets is irrelevant.

Theorem 4.13. *If a random measure P is tail-free relative to a given sequence of partitions $\mathcal{T}_m = \{A_\varepsilon\colon \varepsilon \in \mathcal{E}^m\}$ that generates the Borel sets, then for every m and n the posterior*

distribution of $(P(A_\varepsilon)\colon \varepsilon \in \mathcal{E}^m)$ *given an i.i.d. sample* $X_1, \ldots, X_n$ *from* P *is the same as the posterior distribution of this vector given* $(N_\varepsilon\colon \varepsilon \in \mathcal{E}^m)$ *defined in* (4.8)*, a.s.*.

Proof In view of Theorem 4.11 and Lemma 4.12, we may generate the variables $P, X_1, \ldots, X_n$ in four steps:

(a) Generate the vector $\theta := (P(A_\varepsilon)\colon \varepsilon \in \mathcal{E}^m)$ from its prior.
(b) Given θ generate a multinomial vector $N = (N_\varepsilon\colon \varepsilon \in \mathcal{E}^m)$ with parameters n and θ.
(c) Generate the process $\eta := (P(A|\, A_\varepsilon)\colon A \in \mathscr{X}, \varepsilon \in \mathcal{E}^m)$.
(d) Given (N, η) generate for every $\varepsilon \in \mathcal{E}^m$ a random sample of size N_ε from the measure $P(\cdot|\, A_\varepsilon)$, independently across $\varepsilon \in \mathcal{E}^m$, and let $X_1, \ldots, X_n$ be the n values so obtained in a random order.

For a tail-free measure step (c) is independent of step (a). Together with the fact that step (b) uses θ but not η this implies that $(N, \theta) \perp\!\!\!\perp \eta$. Finally, that step (d) does not use θ can be expressed as $X \perp\!\!\!\perp \theta|\,(N, \eta)$. Together these (conditional) independencies imply that $\theta \perp\!\!\!\perp X|\, N$, which is equivalent to the statement of the theorem. (Note that $\mathrm{E}(f(\theta)g(X)|\, N, \eta) = \mathrm{E}(f(\theta)|\, N, \eta)\mathrm{E}(g(X)|\, N, \eta)$, for every bounded measurable functions f and g. In the first conditional expectation η can be deleted in the conditioning. Next we can take the conditional expectation relative to N across. See Exercise 4.14 for more formal manipulation of these (in)dependencies.)

Thus the theorem is proved for this special representation of prior and data. Because the assertion depends on the joint distribution of (P, X, N) only, it is true in general. □

4.5 Posterior distribution

Consider observations $X_1, X_2, \ldots, X_n$ sampled independently from a distribution P that was drawn from a Dirichlet prior distribution, i.e.

$$P \sim \mathrm{DP}(\alpha), \qquad X_1, X_2, \ldots |\, P \overset{\text{iid}}{\sim} P. \tag{4.9}$$

By an abuse of language, which we shall follow, such observations are often termed a *sample from the Dirichlet process*.

One of the most remarkable properties of the Dirichlet process prior is that the posterior distribution is again Dirichlet.

Theorem 4.14 (Conjugacy)**.** *The posterior distribution of* P *given an i.i.d. sample* $X_1, \ldots, X_n$ *from a* $\mathrm{DP}(\alpha)$ *process is* $\mathrm{DP}(\alpha + \sum_{i=1}^n \delta_{X_i})$.

Proof Because the Dirichlet process is tail-free for any sequence of partitions by Theorem 4.13, and a given measurable partition $\{A_1, \ldots, A_k\}$ of $\mathfrak{X}$ can be viewed as part of a sequence of successive binary partitions, the posterior distribution of the random vector $(P(A_1), \ldots, P(A_k))$ given $X_1, \ldots, X_n$ is the same as the posterior distribution of this vector given the vector $N = (N_1, \ldots, N_k)$ of cell counts, defined by $N_j = \#(1 \le i \le n\colon X_i \in A_j)$. Given P the vector N possesses a multinomial distribution with parameter $(P(A_1), \ldots, P(A_k))$, which has a $\mathrm{Dir}(k; \alpha(A_1), \ldots, \alpha(A_k))$ prior distribution. The posterior distribution can be obtained using Bayes' rule applied to these finite-dimensional vectors, as in Example 2.2. □

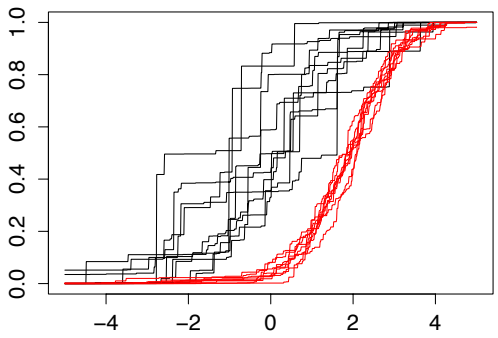

Figure 4.3 Cumulative distribution functions of 10 draws (black) from the Dirichlet process with base measure $5N(0, 2)$, and of 10 draws (red) from the realization of the posterior distribution based on a sample of size 100 from a $N(2, 1)$ distribution.

Theorem 4.14 can be remembered as the updating rule $\alpha \mapsto \alpha + \sum_{i=1}^{n} \delta_{X_i}$ for the base measure of the Dirichlet distribution. In terms of the parameterization $\alpha \leftrightarrow (M = |\alpha|, \bar{\alpha})$ of the base measure, this rule takes the form

$$M \mapsto M + n \quad \text{and} \quad \bar{\alpha} \mapsto \frac{M}{M+n}\bar{\alpha} + \frac{n}{M+n}\mathbb{P}_n, \tag{4.10}$$

where $\mathbb{P}_n = n^{-1} \sum_{i=1}^{n} \delta_{X_i}$ is the *empirical distribution* of $X_1, \ldots, X_n$. Because the mean measure of a Dirichlet process is the normalized base measure, we see that

$$\mathrm{E}(P(A) | X_1, \ldots, X_n) = \frac{|\alpha|}{|\alpha| + n}\bar{\alpha}(A) + \frac{n}{|\alpha| + n}\mathbb{P}_n(A). \tag{4.11}$$

Thus the posterior mean (the "Bayes estimator" of P) is a convex combination of the prior mean $\bar{\alpha}$ and the empirical distribution, with weights $M/(M+n)$ and $n/(M+n)$, respectively. For a given sample it is close to the prior mean if M is large, and close to the empirical distribution (which is based only on the data) if M is small. Thus M determines the extent to which the prior controls the posterior mean — a Dirichlet process prior with precision M contributes information equivalent to a sample of size M (although M is not restricted to integer values). This invites to view M as the *prior sample size*, or the "number of pre-experiment observations". In this interpretation the sum $M+n$ is the "posterior sample size".

As the number of observations n tends to infinity, the prior sample size M, if fixed, is negligible. The posterior mean (4.11) is then dominated by the empirical part. The difference with the empirical distribution is of order $1/n$, and hence the posterior mean tends almost surely to $P(A)$, just as the empirical distribution $\mathbb{P}_n(A)$. The posterior variance of $P(A)$ for a given set A satisfies, for $\tilde{\mathbb{P}}_n = (\alpha + n\mathbb{P}_n)/(|\alpha| + n)$ the posterior mean,

$$\mathrm{var}(P(A) | X_1, \ldots, X_n) = \frac{\tilde{\mathbb{P}}_n(A)\tilde{\mathbb{P}}_n(A^c)}{1 + |\alpha| + n} \leq \frac{1}{4(1 + |\alpha| + n)}.$$

This is also of the order $1/n$, whence the posterior distribution contracts to the Dirac measure at $P(A)$. We say that the posterior distribution is *consistent*.

Closer inspection shows that the posterior distribution converges at the rate $1/\sqrt{n}$ and satisfies a Bernstein-von Mises theorem.

Theorem 4.15 (Bernstein–von Mises theorem for Dirichlet process)**.** *For any measurable set A the process $\sqrt{n}(P-\mathbb{P}_n)(A)$ with $P \sim \mathrm{DP}(\alpha+n\mathbb{P}_n)$ converges conditionally in distribution given $X_1, X_2, \ldots$ to a normal distribution with mean zero and variance $P_0(A)(1-P_0)(A)$, a.s., $[P_0^\infty]$, as $n \to \infty$. The same conclusion is valid if the centering $\mathbb{P}_n$ is replaced by the posterior mean $\bar{\mathbb{P}}_n = (\alpha + n\mathbb{P}_n)/(|\alpha| + n)$.*

Proof The conditional distribution of $P(A)$ is beta with parameters $((\alpha + n\mathbb{P}_n)(A), (\alpha + n\mathbb{P}_n)(A^c))$. By Proposition 2.25 this can be represented as the quotient $V_n/(V_n + W_n)$, for independent Gamma variables $V_n \sim \Gamma((\alpha + n\mathbb{P}_n)(A), 1)$ and $W_n \sim \Gamma((\alpha + n\mathbb{P}_n)(A^c), 1)$. The sequences $\sqrt{n}(V_n/n-\mathbb{P}_n)(A))$ and $\sqrt{n}(W_n/n-\mathbb{P}_n)(A^c))$ are asymptotically normal by the central limit theorem and Slutzky's lemma. For instance, write V_n as the sum of n i.i.d. Gamma variables with parameters $(\mathbb{P}_n(A), 1)$ and an independent Gamma variable with parameters $(\alpha(A), 1)$. The last variable is negligible in the sum. The distribution of the first n variables depends on n, but asymptotic normality follows by a version of the central limit theorem that allows this, such as the Lindeberg-Feller theorem.

Next the asymptotic normality of the sequence $\sqrt{n}(V_n/(V_n + W_n) - \mathbb{P}_n(A))$ follows by the delta-method applied with the function $(v, w) \mapsto v/(v + w)$. □

4.6 Predictive distribution

The joint distribution of a sequence $X_1, X_2, \ldots$ generated from a Dirichlet process, as in (4.9), has a complicated structure, but can be conveniently described by its sequence of *predictive distributions*: the laws of X_1, $X_2|\, X_1$, $X_3|\, X_1, X_2$, etc.

Because $\Pr(X_1 \in A) = \mathrm{E}\Pr(X_1 \in A|\, P) = \mathrm{E}P(A) = \bar{\alpha}(A)$, the marginal distribution of X_1 is $\bar{\alpha}$.

Because $X_2|\,(P, X_1) \sim P$ and $P|\, X_1 \sim \mathrm{DP}(\alpha + \delta_{X_1})$, we can apply the same reasoning again, but now conditionally given X_1, to see that $X_2|\, X_1$ follows the normalization of $\alpha + \delta_{X_1}$. This is a mixture of α and δ_{X_1} with weights $|\alpha|/(|\alpha| + 1)$ and $1/(|\alpha| + 1)$, respectively.

Repeating this argument, using that $P|\, X_1, \ldots, X_{i-1} \sim \mathrm{DP}(\alpha + \sum_{j=1}^{i-1} \delta_{X_j})$, we find that

$$X_i|\, X_1, \ldots, X_{i-1} \sim \begin{cases} \delta_{X_1}, & \text{with probability } \frac{1}{|\alpha|+i-1}, \\ \vdots & \vdots \\ \delta_{X_{i-1}}, & \text{with probability } \frac{1}{|\alpha|+i-1}, \\ \bar{\alpha}, & \text{with probability } \frac{|\alpha|}{|\alpha|+i-1}. \end{cases} \tag{4.12}$$

Being a mixture of a product of identical distributions, the joint distribution of $X_1, X_2, \ldots$ is exchangeable. This means that the ordering of the variables in (4.12) is actually not important; the formulas are valid for any permutation of the indices of $X_1, X_2, \ldots$.

The recipe (4.12) is called the *generalized Polya urn scheme*, and can be viewed as a continuous analog of the familiar Polya urn scheme. Consider balls which can carry a continuum $\mathfrak{X}$ of "colors". Initially the "number of balls" is $M = |\alpha|$, which may be any positive number, and the colors are distributed according to $\bar{\alpha}$. We draw a ball from the collection, observe its color X_1, and return it to the urn along with an additional ball of the same color. The total number of balls is now $M+1$, and the colors are distributed according to $(M\bar{\alpha}+\delta_{X_1})/(M+1)$. We draw a ball from this updated urn, observe its color X_2, and return it to the urn along with

an additional ball of the same color. The probability of picking up the ball that was added after the first draw is $1/(M+1)$, in which case $X_2 = X_1$; otherwise, with probability $M/(M+1)$, we make a fresh draw from the original urn. This process continues indefinitely, leading to the conditional distributions in (4.12).

4.7 Number of distinct values

It is clear from the predictive distribution that a realization of $(X_1, \dots, X_n)$ will have ties (equal values) with positive probability. For instance, with probability at least

$$\frac{1}{M+1}\frac{2}{M+2}\cdots\frac{n-1}{M+n-1}$$

all X_i will even be identical to the first drawn value. For simplicity assume that the base measure α is non-atomic. Then the ith value X_i in the Polya scheme (4.12) is different from the previous $X_1, \dots, X_{i-1}$ if it is drawn from $\bar{\alpha}$. The vector $(X_1, \dots, X_n)$ induces a random partition $\{\mathcal{P}_1, \dots, \mathcal{P}_{K_n}\}$ of the set of indices $\{1, 2, \dots, n\}$ corresponding to the ties (i.e. the equivalence classes under the relation $i \equiv j$ iff $X_i = X_j$). It is intuitively clear that given this partition the K_n distinct values are an i.i.d. sample from $\bar{\alpha}$.

The number of distinct values is remarkably small.

Proposition 4.16. *If the base measure α is nonatomic and of strength $|\alpha| = M$, then, as $n \to \infty$,*

(i) $\mathrm{E}K_n \asymp M \log n \asymp \operatorname{var} K_n$.
(ii) $K_n / \log n \to M$, *a.s.*
(iii) $(K_n - \mathrm{E}K_n)/\operatorname{sd}(K_n) \rightsquigarrow \mathrm{Nor}(0, 1)$.

Proof For $i \in \mathbb{N}$ define $D_i = 1$ if the ith observation X_i is a "new value", i.e. if $X_i \notin \{X_1, \dots, X_{i-1}\}$, and set $D_i = 0$ otherwise. Then $K_n = \sum_{i=1}^n D_i$ is the number of distinct values among the first n observations. Given $X_1, \dots, X_{i-1}$ the variable X_i is "new" if and only if it is drawn from $\bar{\alpha}$, which happens with probability $M/(M+i-1)$. It follows that the variables $D_1, D_2, \dots$ are independent Bernoulli variables with success probabilities $\Pr(D_i = 1) = M/(M+i-1)$.

Assertion (i) can be derived from the exact formulas

$$\mathrm{E}K_n = \sum_{i=1}^n \frac{M}{M+i-1}, \qquad \operatorname{var} K_n = \sum_{i=1}^n \frac{M(i-1)}{(M+i-1)^2}.$$

Next, assertion (ii) follows from Kolmogorov's strong law of large numbers for independent variables, since

$$\sum_{i=1}^\infty \frac{\operatorname{var} D_i}{(\log i)^2} = \sum_{i=1}^\infty \frac{M(i-1)}{(M+i-1)^2(\log i)^2} < \infty.$$

Finally (iii) is a consequence of the Lindeberg central limit theorem. □

Thus the number of distinct values in a (large) sample from a distribution taken from a fixed Dirichlet prior is logarithmic in the sample size, and the fluctuations of this number around its mean are of the order $\sqrt{\log n}$.

The following proposition gives the distribution of the partition $\{\mathcal{P}_1, \ldots, \mathcal{P}_{K_n}\}$ induced by $(X_1, \ldots, X_n)$.

Proposition 4.17. *A random sample $X_1, \ldots, X_n$ from a Dirichlet process with nonatomic base measure of strength $|\alpha| = M$ induces a given partition of $\{1, 2, \ldots, n\}$ into k sets of sizes $n_1, \ldots, n_k$ with probability equal to*

$$\frac{M^k \Gamma(M) \prod_{j=1}^k \Gamma(n_j)}{\Gamma(M+n)}. \tag{4.13}$$

Proof By exchangeability the probability depends on the sizes of the partitioning sets only. The probability that the partitioning set of size n_1 consists of the first n_1 variables, the one of size n_2 of the next n_2 variables, etc. can be obtained by multiplying the appropriate conditional probabilities for the consecutive draws in the Polya urn scheme in their natural order of occurrence. For $r_j = \sum_{l=1}^j n_l$, it is given by

$$\begin{aligned}\frac{M}{M}\frac{1}{M+1}\frac{2}{M+2}\cdots\frac{n_1-1}{M+n_1-1}\frac{M}{M+n_1}\frac{1}{M+n_1+1}\times\cdots\\ \cdots\times\frac{M}{M+r_{k-1}}\frac{1}{M+r_{k-1}+1}\cdots\frac{n_k-1}{M+r_{k-1}+n_k-1}.\end{aligned}$$

This can be rewritten as in the proposition. □

4.8 *Mixtures of Dirichlet processes

Application of the Dirichlet prior requires a choice of a base measure α. It is often reasonable to choose the center measure $\bar{\alpha}$ from a specific family such as the normal family, but then the parameters of the family must still be specified. It is natural to give these a further prior. Similarly, one may put a prior on the precision parameter $|\alpha|$.

For a base measure α_ξ that depends on a parameter ξ the Bayesian model then consists of the hierarchy

$$X_1, \ldots, X_n | P, \xi \overset{\text{iid}}{\sim} P, \qquad P|\xi \sim \mathrm{DP}(\alpha_\xi), \qquad \xi \sim \pi. \tag{4.14}$$

We denote the induced (marginal) prior on P by $\mathrm{MDP}(\alpha_\xi, \xi \sim \pi)$. Many properties of this *mixture Dirichlet prior* follow immediately from those of a Dirichlet process. For instance, any P following an MDP is almost surely discrete. However, unlike a Dirichlet process, an MDP is not tail-free.

Given ξ we can use the posterior updating rule for the ordinary Dirichlet process, and obtain that

$$P|\xi, X_1, \ldots, X_n \sim \mathrm{DP}(\alpha_\xi + n\mathbb{P}_n).$$

To obtain the posterior distribution of P given $X_1, \ldots, X_n$, we need to mix this over ξ relative to its posterior distribution given $X_1, \ldots, X_n$. By Bayes's theorem the latter has density proportional to

$$\xi \mapsto \pi(\xi)\, p(X_1, \ldots, X_n|\xi). \tag{4.15}$$

Here the marginal density of $X_1, \ldots, X_n$ given ξ (the second factor) is described by the generalized Polya urn scheme (4.12) with α_ξ instead of α. In general, this has a somewhat

complicated structure due to the ties between the observations. However, for a posterior calculation we condition on the observed data $X_1, \ldots, X_n$, and know the partition that they generate. Given this information the density takes a simple form. For instance, if the observations are distinct (which happens with probability one if the observations actually follow a continuous distribution), then the Polya urn scheme must have simply generated a random sample from the normalized base measure $\bar{\alpha}_\xi$, in which case the preceding display becomes

$$\pi(\xi) \prod_{i=1}^{n} d\alpha_\xi(X_i) \prod_{i=1}^{n} \frac{1}{|\alpha_\xi| + i - 1},$$

for $d\alpha_\xi$ a density of α_ξ. Further calculations depend on the specific family and its parameterization.

Typically the precision parameter M and center measure G in $\alpha = MG$ will be modelled as independent under the prior. The posterior calculation then factorizes in these two parameters. To see this, consider the following scheme to generate the parameters and observations:

(i) Generate M from its prior.
(ii) Given M generate a random partition $\mathcal{P} = \{\mathcal{P}_1, \ldots, \mathcal{P}_{K_n}\}$ according to the distribution given in Proposition 4.17.
(iii) Generate G from its prior, independently of $(M, \mathcal{P})$.
(iv) Given $(\mathcal{P}, G)$ generate a random sample of size K_n from G, independently of M, and set X_i with $i \in \mathcal{P}_j$ equal to the jth value in this sample.

By the description of the Polya urn scheme this indeed gives a sample $X_1, \ldots, X_n$ from the mixture of Dirichlet processes $\mathrm{MDP}(MG, M \sim \pi, G \sim \pi)$. We may now formally write the density of $(M, \mathcal{P}, G, X_1, \ldots, X_n)$ in the form, with π abusively denoting prior densities for both M and G and p conditional densities of observed quantities,

$$\pi(M)\, p(\mathcal{P}|\, M)\, \pi(G)\, p(X_1, \ldots, X_n|\, G, \mathcal{P}).$$

Since this factorizes in terms involving M and G, these parameters are also independent under the posterior distribution, and the computation of their posterior distributions can be separated.

The term involving M depends on the data through K_n only (the latter variable is *sufficient* for M). Indeed, by Proposition 4.17 it is proportional to,

$$M \mapsto \pi(M) \frac{M^{K_n}\Gamma(M)}{\Gamma(M+n)} \propto \pi(M) M^{K_n} \int_0^1 \eta^{M-1}(1-\eta)^{n-1}\, d\eta.$$

Rather than by (numerically) integrating this expression, the posterior density is typically computed by simulation. Suppose that $M \sim \mathrm{Ga}(a, b)$ a priori, and consider a fictitious random vector (M, η) with $0 \le \eta \le 1$ and joint (Lebesgue) density proportional to

$$\pi(M) M^{K_n} \eta^{M-1}(1-\eta)^{n-1} \propto M^{a+K_n-1} e^{-M(b-\log\eta)} \eta^{-1}(1-\eta)^{n-1}.$$

Then by the preceding display the marginal density of M is equal to its posterior density (given K_n, which is fixed for the calculation). Thus simulating from the distribution of (M, η)

and dropping η simulates M from its posterior distribution. The conditional distributions are given by

$$M|\,\eta, K_n \sim \mathrm{Ga}(a + K_n, b - \log\eta), \qquad \eta|\,M, K_n \sim \mathrm{Be}(M, n). \tag{4.16}$$

We can use these in a *Gibbs sampling scheme*: given an arbitrary starting value η_0 we generate a sequence $M_1, \eta_1, M_2, \eta_2, M_3, \ldots$, by repeatedly generating M from its conditional distribution given (η, K_n) and η from its conditional distribution given (M, K_n), each time setting the conditioning variable (η or M) equal to its last value. After an initial *burn-in* the values $M_k, M_{k+1}, \ldots$ will be approximately from the posterior distribution of M given K_n.

4.9 Complements

Proposition 4.18 (Kolmogorov extension theorem)**.** *For every finite subset S of an arbitrary set T let P_S be a probability distribution on $\mathbb{R}^S$. Then there exists a probability space $(\Omega, \mathscr{U}, \Pr)$ and measurable maps $X_t\colon \Omega \to \mathbb{R}$ such that $(X_t\colon t \in S) \sim P_S$ for every finite set S if and only if for every pair $S' \subset S$ of finite subsets the measure $P_{S'}$ is the marginal distribution of P_S on $\mathbb{R}^{S'}$.*

For a proof see a book on measure-theoretic probability or stochastic processes. The sets S in this formulation are unordered; if ordered sets were used, then of course there must also be consistency between marginals P_S and $P_{S'}$ for every pair (S, S') with S' a reordering of S.

4.9.1 Weak topology on measures

Definition 4.19. The *weak topology* on the set $\mathfrak{M}$ of all probability measures on a given Polish space $(\mathfrak{X}, \mathscr{X})$ with its Borel σ-field, is the weakest topology such that all maps $P \mapsto \int \psi\, dP$, for a bounded continuous function $\psi\colon \mathfrak{X} \to \mathbb{R}$, are continuous. Equivalently, it is the topology such that $P_n \to P$ if and only if $\int \psi\, dP_n \to \int \psi\, dP$, for every bounded continuous function $\psi\colon \mathfrak{X} \to \mathbb{R}$.

The weak topology is also called the topology of *convergence in distribution*. In particular, one says that a sequence of random variables X_n with values in a Polish space converges in distribution to a random variable X if the sequence of induced laws converges weakly to the law of X, i.e. $\mathrm{E}\psi(X_n) \to \mathrm{E}\psi(X)$, for every bounded continuous $\psi\colon \mathfrak{X} \to \mathbb{R}$. For random vectors with values in Euclidean space, this is equivalent to convergence of the corresponding cumulative distribution functions at continuity points of the limit.

Proposition 4.20 (Portmanteau)**.** *A sequence of probability measures P_n on $\mathbb{R}^d$ converges weakly to a probability measure P if and only if $F_n(x) := P_n((-\infty, x]) \to F(x) := P((-\infty, x])$, for every $x \in \mathbb{R}^d$ such that F is continuous at x.*

For a proof see many books on probability, or van der Vaart (1998), Lemma 2.2.

Proposition 4.21. *The weak topology $\mathscr{W}$ on the set $\mathfrak{M}$ of Borel measures on a Polish space $\mathfrak{X}$ is Polish.*

For a proof see e.g. Dudley (2002), Corollary 11.5.5. In particular, the proposition implies that convergence in distribution can be understood as convergence $d_w(P_n, P) \to 0$ for a metric d_w on $\mathfrak{M}$. There are several possibilities for this metric, perhaps the simplest being the *bounded Lipschitz metric*

$$d_w(P, Q) = \sup_{\psi} \Big| \int \psi\, dP - \int \psi\, dQ \Big|,$$

where the supremum is taken over all functions $\psi\colon \mathfrak{X} \to [0, 1]$ such that $|\psi(x) - \psi(y)| \le d(x, y)$, for

every $x, y \in \mathfrak{X}$. (This metric can also be defined on the set of all signed measures on $(\mathfrak{X}, \mathscr{X})$, the linear span of probability measures, and be derived from a norm on this space. The dual of the resulting normed space is the space of bounded continuous functions and the weak topology can be understood in terms of the functional analytic concept of the same name.)

Proposition 4.22. *The set of discrete probability measures with finitely many support points is weakly dense in* $\mathfrak{M}$.

Proof It suffices to show that any probability measure P is the weak limit of a sequence of finitely discrete measures. To see this, let $X_1, X_2, \ldots$ be a sequence of independent random variables with distribution P. By the law of large numbers $n^{-1} \sum_{i=1}^n \psi(X_i) \to \int \psi \, dP$, almost surely, for every integrable function $\psi: \mathfrak{X} \to \mathbb{R}$, in particular for every bounded continuous function ψ. This is then true also simultaneously almost surely for every ψ in a countable set of bounded continuous functions. Thus there certainly exists a sequence of realizations $x_1, x_2, \ldots$ with $\int \psi \, d\mathbb{P}_n \to \int \psi \, dP$, for $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{x_i}$. It may be shown that there exists a countable set of bounded continuous functions such that the weak topology is generated by the maps $P \mapsto \int \psi \, dP$. □

Exercises

4.1 For each A in an arbitrary index set $\mathscr{A}$ let $f_A: \mathfrak{M} \to \mathbb{R}$ be an arbitrary map.

(a) Show that there exists a smallest σ-field $\mathscr{M}$ such that every map f_A is measurable.

(b) Show that a map $T: (\Omega, \mathscr{U}) \to (\mathfrak{M}, \mathscr{M})$ is measurable if and only if $f_A \circ T: \Omega \to \mathbb{R}$ is measurable for every $A \in \mathscr{A}$.

4.2 Let $K_1, K_2, \ldots$ be compact sets in a topological space such that $\cap_i K_i = \varnothing$. Show that $\cap_{i=1}^m K_i = \varnothing$ for some m.

4.3 Show that for any Borel set $A \subset \mathbb{R}^d$ and finite measure μ on the Borel sets, and every $\epsilon > 0$, there exists a compact set K with $K \subset A$ and $\mu(A - K) < \epsilon$. [Let $\mathscr{X}_0$ be the set of all Borel sets A such that there exists for every $\epsilon > 0$ a closed set F and open set G with $F \subset A \subset G$ and $\mu(G - F) < \epsilon$. Show that $\mathfrak{X}$ is a σ-field. Show that it is the Borel σ-field. Show that the sets F can be taken compact without loss of generality.]

4.4 Show that the discrete probability measures with finitely many support points are dense in the set of all Borel probability measures on a Polish space (or $\mathbb{R}^d$) relative to the weak topology.

4.5 Show that the support of the Dirichlet process prior with base measure α such that $\alpha(G) > 0$ for every open subset $G \subset \mathfrak{X}$ is equal to the set $\mathfrak{X}$ of all Borel probability measures on $(\mathfrak{X}, \mathscr{X})$, in the sense that every weakly open subset of $\mathfrak{X}$ has positive probability under $\mathrm{DP}(\alpha)$. [Hint: one possibility is to use the preceding exercise and the series representation of the Dirichlet process.]

4.6 Consider a stick-breaking scheme with independent variables Y_k with $1 - \Pr(Y_k = 0) = 1/k^2 = \Pr(Y_k = 1 - e^{-k})$. Show that the "stick is not finished": $\sum_k p_k < 1$ almost surely.

4.7 Show that if $P \sim \mathrm{DP}(\alpha)$ and $\psi: \mathfrak{X} \to \mathfrak{Y}$ is a measurable mapping, then $P \circ \psi^{-1} \sim \mathrm{DP}(\beta)$, for $\beta = \alpha \circ \psi^{-1}$.

4.8 Show that if $P \sim \mathrm{DP}(\alpha)$, then $\mathrm{E} \int \psi \, dP = \int \psi \, d\bar{\alpha}$, and $\mathrm{var} \int \psi \, dP = \int (\psi - \int \psi \, d\bar{\alpha})^2 \, d\bar{\alpha}/(1 + |\alpha|)$, for any measurable function ψ for which the integrals make sense. [Hint: prove this first for $\psi = 1_A$.]

4.9 Let $0 = T_0 < T_1 < T_2 < \cdots$ be the events of a standard Poisson process and let $\theta_1, \theta_2, \ldots \overset{\text{iid}}{\sim} \bar{\alpha}$ and independent of $(T_1, T_2, \ldots)$. Show that

$$P = \sum_{k=1}^{\infty} (e^{-T_{k-1}} - e^{-T_k}) \delta_{\theta_k}$$

follows a Dirichlet process DP($\bar{\alpha}$). How can we change the prior precision to $M \neq 1$?

4.10 Let $F \sim \mathrm{DP}(MG)$ be a Dirichlet process on $\mathfrak{X} = \mathbb{R}$, for a constant $M > 0$ and probability distribution G, identified by its cumulative distribution function $x \mapsto G(x)$. So F can be viewed as a random cumulative distribution function. Define its median as any value m_F such that $F(m_F-) \leq 1/2 \leq F(m_F)$. Show that

$$\Pr(m_F \leq x) = \int_{1/2}^{1} \beta(u, MG(x), M(1-G(x)))\, du,$$

where $\beta(\cdot, \alpha, \beta)$ is the density of the Beta-distribution.

4.11 Simulate and plot the cumulative distribution function of the Dirichlet processes $F \sim \mathrm{DP}(\Phi)$, $F \sim \mathrm{DP}(0.1\Phi)$, and $F \sim \mathrm{DP}(10\Phi)$. Do the same with the Cauchy base measure. [Suggestion use Sethuraman's presentation. Cut the series at an appropriate point.]

4.12 Let G be a given continuous probability measure on $\mathbb{R}$, identified by its cumulative distribution function. Let the partition at level m consist of the sets $(G^{-1}((i-1)2^{-m}), G^{-1}(i2^{-m})]$, for $i = 1, 2 \ldots, 2^m$. Let the variables $V_{\varepsilon 0}$ be independent with mean $1/2$ and define a random probability measure by (4.7). Find $\mathrm{E}P(A)$, for a given measurable set.

4.13 Suppose that $N \sim \mathrm{MN}(1, p_1, \ldots, p_k)$ and given $N = e_j$ let X be drawn from a given probability measure P_j. Show that $X \sim \sum_j p_j P_j$. What is this latter measure if $p_j = P(A_j)$ and $P_j = P(\cdot | A_j)$ for a given measure P and measurable partition $\mathfrak{X} = \cup_j A_j$?

4.14 Let θ, η, N, X be random elements defined on a common probability space with values in Polish spaces. Show that

(a) $\eta \perp\!\!\!\perp \theta$ if and only if $\Pr(\eta \in A | \theta) = \Pr(\eta \in A)$ almost surely, for all measurable sets A.
(b) $\theta \perp\!\!\!\perp X | N$ if and only if $\Pr(\theta \in A | N) = \Pr(\theta \in A | X, N)$ almost surely, for all measurable sets A.
(c) $\eta \perp\!\!\!\perp (\theta, N)$ if and only if ($\eta \perp\!\!\!\perp N | \theta$ and $\eta \perp\!\!\!\perp \theta$).
(d) if $X \perp\!\!\!\perp \theta | (N, \eta)$ and $\theta \perp\!\!\!\perp \eta | N$, then $\theta \perp\!\!\!\perp X | N$.

Conclude that if $\eta \perp\!\!\!\perp (N, \theta)$ and $X \perp\!\!\!\perp \theta | (N, \eta)$, then $\theta \perp\!\!\!\perp X | N$. [The Polish assumptions guarantee that conditional distributions are well defined. Conditional independence of X and Y given Z means that $\Pr(X \in A, Y \in B | Z) = \Pr(X \in A | Z) \Pr(Y \in B | Z)$ almost surely, for every measurable sets A, B. The conditional expectation $\Pr(X \in A | Z)$ is a measurable function of Z such that $\mathrm{E} \Pr(X \in A | Z) 1_C(Z) = \Pr(X \in A, Z \in C)$ for every measurable sets A, C.]

4.15 Let ψ be a given bounded measurable function. Show that if $P \sim \mathrm{DP}(\alpha)$ and $X_1, \ldots, X_n | P \overset{\mathrm{iid}}{\sim} P$, then the posterior distribution of $\int \psi\, dP$ given $X_1, \ldots, X_n$ tends in distribution to a Dirac measure at $\int \psi\, dP_0$ for a.e. sequence $X_1, X_2, \ldots$ generated iid from P_0.

4.16 In the model (4.14) assume that the total mass $|\alpha_\xi|$ is bounded uniformly in ξ. Show that the posterior distribution of $P(A)$ is consistent.

4.17 Simulate and plot the cumulative distribution functions of realizations of some posterior Dirichlet processes. First use several fixed prior strengths. Second put a Gamma prior on the prior strength.

5

Posterior Contraction

This chapter investigates the question whether nonparametric Bayesian procedures "work". We adopt the "frequentist" point of view that the observations are generated according to some "true parameter" and study whether the posterior distribution is able to recover this parameter if the number of observations (or, more generally, their "informativeness") increases indefinitely. We start with posterior consistency, which is the basic form of recovery, and next turn to the rate of posterior contraction, which is a much more informative refinement of consistency.

5.1 Consistency

For every $n \in \mathbb{N}$ let $X^{(n)}$ be an observation in a sample space $(\mathfrak{X}^{(n)}, \mathscr{X}^{(n)})$ with distribution $P_\theta^{(n)}$ indexed by a parameter θ belonging to a semi-metric space (Θ, d). For instance $X^{(n)}$ may be sample of size n from a given distribution P_θ, and $(\mathfrak{X}^{(n)}, \mathscr{X}^{(n)}, P_\theta^{(n)})$ the corresponding product probability space. Given a prior Π on the Borel sets of Θ, let $\Pi_n(\cdot | X^{(n)})$ be a version of the posterior distribution.

Definition 5.1 (Posterior consistency)**.** The posterior distributions $\Pi_n(\cdot | X^{(n)})$ are said to be *(weakly) consistent* at $\theta_0 \in \Theta$ if $\Pi_n(\theta\colon d(\theta, \theta_0) > \epsilon | X^{(n)}) \to 0$ in $P_{\theta_0}^{(n)}$-probability, as $n \to \infty$, for every $\epsilon > 0$. The posterior distributions are said to be *strongly consistent* at $\theta_0 \in \Theta$ if the convergence is in the almost sure sense.

Both forms of consistency are of interest. Naturally, strong consistency is more appealing as it is stronger. (To be well defined it presumes that the observations $X^{(n)}$ are defined on a common underlying probability space, which may or may not be natural.)

Consistency entails that the full posterior distribution *contracts* to within arbitrarily small distance ϵ to the true parameter θ_0. It can also be summarized as saying that the posterior distributions converge weakly to a Dirac measure at θ_0, in probability or almost surely.

Naturally an appropriate summary of the "location" of the posterior distribution should provide a point estimator that is consistent in the usual sense of consistency of estimators. The following proposition gives a summary that works without further conditions. (The value $1/2$ could be replaced by any other number between 0 and 1.)

Proposition 5.2 (Point estimator)**.** *Suppose that the posterior distributions $\Pi_n(\cdot | X^{(n)})$ are consistent (or strongly consistent) at θ_0 relative to the metric d on Θ. Then $\hat{\theta}_n$ defined as the center of a (nearly) smallest ball that contains posterior mass at least $1/2$ satisfies $d(\hat{\theta}_n, \theta_0) \to 0$ in $P_{\theta_0}^{(n)}$-probability (or almost surely $[P_{\theta_0}^{(\infty)}]$, respectively).*

Proof For $B(\theta, r) = \{s \in \Theta: d(s, \theta) \leq r\}$ the closed ball of radius r around $\theta \in \Theta$, let $\hat{r}_n(\theta) = \inf\{r: \Pi_n(B(\theta, r)|\, X^{(n)}) \geq 1/2\}$, where the infimum over the empty set is ∞. Taking the balls closed ensures that $\Pi_n(B(\theta, \hat{r}_n(\theta))|\, X^{(n)}) \geq 1/2$, for every θ. Let $\hat{\theta}_n$ be a near minimizer of $\theta \mapsto \hat{r}_n(\theta)$ in the sense that $\hat{r}_n(\hat{\theta}_n) \leq \inf_\theta \hat{r}_n(\theta) + 1/n$.

By consistency $\Pi_n(B(\theta_0, \epsilon)|\, X^{(n)}) \to 1$ in probability or almost surely, for every $\epsilon > 0$. As a first consequence, $\hat{r}_n(\theta_0) \leq \epsilon$ with probability tending to one, or eventually almost surely, and hence $\hat{r}_n(\hat{\theta}_n) \leq \hat{r}_n(\theta_0) + 1/n$ is bounded by $\epsilon + 1/n$ with probability tending to one, or eventually almost surely. As a second consequence the balls $B(\theta_0, \epsilon)$ and $B(\hat{\theta}_n, \hat{r}_n(\hat{\theta}_n))$ cannot be disjoint, as their union would contain mass nearly $1 + 1/2$. This shows that $d(\theta_0, \hat{\theta}_n) \leq \epsilon + \hat{r}_n(\hat{\theta}_n)$ with probability tending to one, or eventually almost surely, which is further bounded by $2\epsilon + 1/n$. As this is true for every $\epsilon > 0$, the result follows. □

An alternative point estimator is the *posterior mean* $\int \theta\, d\Pi_n(\theta|\, X^{(n)})$ (available when Θ has a vector space structure). This is attractive for computational reasons, as it can be approximated by the average of the output of a simulation run. Usually the posterior mean is also consistent, but in general this requires additional assumptions, just as weak convergence to a Dirac measure on a Euclidean space does not imply convergence of moments. Consistency and boundedness of $\int \|\theta\|^p\, d\Pi_n(\theta|\, X^{(n)})$ in probability or almost surely for some $p > 1$ would be sufficient for consistency of the posterior mean.

Example 5.3 (Dirichlet process)**.** If the observations are a random sample $X_1, \ldots, X_n$ from a distribution that is equipped with a Dirichlet process prior $\mathrm{DP}(\alpha)$, then the posterior distribution is a Dirichlet process $\mathrm{DP}(\alpha + n\mathbb{P}_n)$, for $\mathbb{P}_n$ the empirical distribution of the observations (see Theorem 4.14). For a fixed measurable set A, the posterior distribution of $P(A)$ is beta distributed with parameters $(\alpha(A) + n\mathbb{P}_n(A), \alpha(A^c) + n\mathbb{P}_n(A^c))$, whence

$$\tilde{\mathbb{P}}_n(A) := \mathrm{E}(P(A)|\, X_1, \ldots, X_n) = \frac{|\alpha|}{|\alpha| + n}\bar{\alpha}(A) + \frac{n}{|\alpha| + n}\mathbb{P}_n(A),$$

$$\mathrm{var}(P(A)|\, X_1, \ldots, X_n) = \frac{\tilde{\mathbb{P}}_n(A)\tilde{\mathbb{P}}_n(A^c)}{1 + |\alpha| + n} \leq \frac{1}{4(1 + |\alpha| + n)}.$$

The first equation and the law of large numbers applied to $\mathbb{P}_n(A)$ shows that the posterior mean tends almost surely to $P_0(A)$ if $X_1, X_2, \ldots$ are sampled from "true" distribution P_0. The second equation shows that the posterior variance tends almost surely to zero. An application of Markov's inequality gives that, for every $\epsilon > 0$,

$$\Pi_n(P: |P(A) - P_0(A)| > \epsilon|\, X_1, \ldots, X_n) \leq \frac{1}{\epsilon^2}\Big[|\tilde{\mathbb{P}}_n(A) - P_0(A)|^2 + \mathrm{var}(P(A)|\, X_1, \ldots, X_n)\Big] \overset{as}{\to} 0.$$

It follows that the posterior distribution is strongly consistent at P_0 for $d(P, P_0) = |P(A) - P_0(A)|$. This is true for any base measure α and any P_0. (We should note that the posterior distribution is unique only up to null sets under the marginal distribution of the data. The claim is valid for the particular choice $\mathrm{DP}(\alpha + n\mathbb{P}_n)$ of posterior distribution.)

The present d is a semi-metric only (in that $d(P_1, P_2) = 0$ does not imply $P_1 = P_2$), but as the result is valid for any measurable set A, it can be inferred that the consistency is also true relative to the weak topology on the set of probability measures. Stronger metrics, such as $d(P_0, P) = \sup_{A \in \mathscr{A}} |P(A) - P_0(A)|$ for a collection of set $\mathscr{A}$, could also be used with the help of the Glivenki-Cantelli theorem.

5.1.1 Doob's theorem

Doob's theorem basically says that for any fixed prior, the posterior distribution is consistent at every θ except those in a set that is "small" when seen from the prior point of view. We first present the theorem, for i.i.d. observations only, and next argue that the message is not as positive as it may seem. For a proof of the theorem, see e.g. van der Vaart (1998), Chapter 10.

Theorem 5.4 (Doob)**.** *Let* $(\mathfrak{X}, \mathscr{X}, P_\theta : \theta \in \Theta)$ *be experiments with* $(\mathfrak{X}, \mathscr{X})$ *a Polish space with Borel* σ*-field and* Θ *a Borel subset of a Polish space such that* $\theta \mapsto P_\theta(A)$ *is Borel measurable for every* $A \in \mathscr{X}$ *and the map* $\theta \mapsto P_\theta$ *is one-to-one. Then for any prior* Π *on the Borel sets of* Θ *the posterior* $\Pi_n(\cdot \mid X_1, \ldots, X_n)$ *in the model* $X_1, \ldots, X_n \mid \theta \overset{iid}{\sim} p_\theta$ *and* $\theta \sim \Pi$ *is strongly consistent at* θ*, for* Π*-almost every* θ*.*

Doob's theorem is remarkable in that virtually no condition is imposed on the model or the parameter space. A Bayesian will "almost always" have consistency, as long as she is certain of her prior.

However, in practice no one can be certain of the prior, and troublesome values of the parameter may really obtain. In fact, the Π-null set could be very large if *not* judged from the point of view of the prior. An extreme example is a prior that assigns all its mass to some fixed point θ_0. The posterior then also assigns mass one to θ_0 and hence is *in*consistent at every $\theta \neq \theta_0$. Doob's theorem is still true, of course; the point is that the set $\{\theta : \theta \neq \theta_0\}$ is a null set under the present prior.

Thus Doob's theorem should not create a false sense of satisfaction about Bayesian procedures in general. It is important to know, for a given "reasonable" prior, at which parameter values consistency holds.

An exception is the case that the parameter set Θ is countable. Then Doob's theorem shows that consistency holds at θ as long as Π assigns positive mass to it. More generally, consistency holds at any atom of a prior. However, even in these cases the theorem is of "asymptopia" type only, in that at best it gives convergence, without quantification of the approximation error, or uniformity in the parameter.

5.1.2 Schwartz's theorem

In this section we take the parameter equal to a probability density, relative to a given dominating measure ν on the sample space $(\mathfrak{X}, \mathscr{X})$. We denote this parameter by p rather than θ, and the corresponding parameter set by $\mathcal{P}$. We consider estimating p based on a random sample $X_1, \ldots, X_n$ of observations, with true density p_0. As notational convention we denote a density by a lower case letter p and the measure induced by it by the uppercase letter P. The parameter set is equipped with a semi-metric d, which is left unspecified for the moment.

A key condition for posterior consistency is that the prior assigns positive probability to any Kullback-Leibler (or KL) neighborhood of the true density. The *Kullback-Leibler divergence* between two densities p_0 and p is defined as

$$K(p_0; p) = \int p_0 \log \frac{p_0}{p} \, d\nu.$$

Note that it is asymmetric in its arguments. We write $K(p_0; \mathcal{P}_0) = \inf_{p \in \mathcal{P}_0} K(p_0; p)$ for the minimal divergence of p_0 to a set $\mathcal{P}_0$ of densities.

Definition 5.5 (KL-property)**.** A density p_0 is said to possess the *Kullback-Leibler property* relative to a prior Π if $\Pi(p\colon K(p_0; p) < \epsilon) > 0$ for every $\epsilon > 0$. This is denoted $p_0 \in \mathrm{KL}(\Pi)$. Alternatively, we say that p_0 belongs to the *Kullback-Leibler support* of Π.[1]

Schwartz's theorem is the basic result on posterior consistency for dominated models. It has two conditions: the true density p_0 should be in the KL-support of the prior, and the hypothesis $p = p_0$ should be testable against complements of neighborhoods of p_0. The first is clearly a Bayesian condition, but the second may be considered a condition to enable recovery of p_0 by any statistical method. Although in its original form the theorem has limited applicability, extensions go far deeper, and lead to a rich theory of posterior consistency.

In the present context *tests* ϕ_n are understood to refer both to measurable mappings $\phi_n\colon \mathfrak{X}^n \to [0, 1]$, and to the corresponding statistics $\phi_n(X_1, \ldots, X_n)$. The interpretation of a test ϕ_n is that a null hypothesis is rejected with probability ϕ_n, whence $P^n\phi_n$ is the probability of rejection if the data are sampled from P. It follows that $P_0^n\phi_n$ is the probability of a type I error for testing $H_0\colon P = P_0$, and $P^n(1 - \phi_n) = 1 - P^n\phi_n$ is the probability of a type II error if $P \neq P_0$.

Theorem 5.6 (Schwartz)**.** *If $p_0 \in \mathrm{KL}(\Pi)$ and for every neighborhood $\mathcal{U}$ of p_0 there exist tests ϕ_n such that $P_0^n\phi_n \to 0$ and $\sup_{p \in \mathcal{U}^c} P^n(1 - \phi_n) \to 0$, then the posterior distribution $\Pi_n(\cdot \mid X_1, \ldots, X_n)$ in the model $X_1 \ldots, X_n \mid p \overset{iid}{\sim} p$ and $p \sim \Pi$ is strongly consistent at p_0.*

Proof By Lemma 2.33 it is not a loss of generality to assume that the tests ϕ_n as in the theorem have exponentially small error probabilities, in the sense that, for some positive constant C,

$$P_0^n\phi_n \leq e^{-Cn}, \qquad \sup_{p \in \mathcal{U}^c} P^n(1 - \phi_n) \leq e^{-Cn}.$$

Then the theorem follows from an application of Theorem 5.9 below, with $\mathcal{P}_n = \mathcal{P}$ for every n. □

The weak topology on the set of probability measures P can also be viewed as a topology on the corresponding densities $p \in \mathcal{P}$. For this topology, consistent tests as in Schwartz's theorem, Theorem 5.6, always exist, and hence the posterior distribution is consistent at every density that possesses the KL-property.

Corollary 5.7 (Consistency for weak topology)**.** *The posterior distribution is consistent for the weak topology at any density p_0 that has the Kullback-Leibler property for the prior.*

Proof It suffices to construct tests with vanishing error probabilities for complements of weak neighbourhoods of p_0. The weak topology is by definition the weakest topology such that the maps $p \mapsto \int \psi \, dP$ are continuous, for bounded continuous functions $\psi\colon \mathfrak{X} \to \mathbb{R}$. (See Section 4.9.1.) Thus all sets of the form $\mathcal{U} = \{p\colon P\psi < P_0\psi + \epsilon\}$, for a continuous

[1] The Kullback-Leibler divergence is typically measurable in its second argument, and then Kullback-Leibler neighbourhoods are measurable in the space of densities. If not, then we interpret the KL-property in the sense of inner probability: it suffices that there exists measurable sets $\mathcal{B} \subset \{p\colon K(p_0; p) < \epsilon\}$ with $\Pi(\mathcal{B}) > 0$.

function $\psi\colon \mathfrak{X} \to [0,1]$ and $\epsilon > 0$, are weakly open, and all such sets form a sub-base for the weak topology. This means that any weakly open set contains the intersection of finitely many of sets of this form (and hence is the union of all such finite intersections). Thus it suffices to construct consistent tests for every finite intersection. Now given a test for any neighborhood $\mathcal{U}$ of the given type, we can form a test for a finite intersection $\cap_i \mathcal{U}_i$ by rejecting P_0 as soon as P_0 is rejected for one of the finitely many neighborhoods $\mathcal{U}_i$. The resulting error probabilities are bounded by the sum of the error probabilities of the finitely many tests, and hence will tend to zero if each test is consistent. Thus it suffices to construct consistent tests for a single neighborhood $\mathcal{U}$.

For a given function $\psi\colon \mathfrak{X} \to [0,1]$, consider the test

$$\phi_n = \mathbb{1}\Big\{\frac{1}{n}\sum_{i=1}^{n}\psi(X_i) > P_0\psi + \frac{\epsilon}{2}\Big\}.$$

By Hoeffding's inequality, Lemma 2.34, this test has type I error satisfying $P_0^n\phi_n \le e^{-n\epsilon^2/2}$. Furthermore, since $P_0\psi - P\psi < -\epsilon$ whenever $P \in \mathcal{U}^c$, we have $P^n(1-\phi_n) \le P^n(n^{-1}\sum_{i=1}^n(\psi(X_i) - P\psi) < -\epsilon/2)$ for $P \in \mathcal{U}^c$ and this is bounded by $e^{-n\epsilon^2/2}$, by a second application of Hoeffding's inequality. □

Example 5.8 (Finite-dimensional models)**.** If the model is smoothly parameterized by a finite-dimensional parameter that varies over a bounded set, then consistent tests as required in Schwartz's theorem, Theorem 5.6, typically exist under mere regularity conditions on the model. For unbounded Euclidean sets some conditions may be needed.

One may show this by direct arguments, or alternatively we can derive this from Corollary 5.7 under the condition that the map $P_\theta \mapsto \theta$ is well defined (i.e. the model is identifiable) and continuous for the weak topology on P_θ. See Problem 5.9.

In its original form Schwartz's theorem requires that the complement of every neighborhood of p_0 can be "tested away". For strong metrics, such as the L_1-distance, such tests may not exist, even though the posterior distribution may be consistent. The following extension of the theorem is useful for these situations. The idea is that sets of very small prior mass need not be tested.

Theorem 5.9 (Extension of Schwartz's theorem)**.** *If $p_0 \in \mathrm{KL}(\Pi)$ and for every neighborhood $\mathcal{U}$ of p_0 there exist a constant $C > 0$, measurable sets $\mathcal{P}_n \subset \mathcal{P}$ and tests ϕ_n such that*

$$\Pi(\mathcal{P} \setminus \mathcal{P}_n) < e^{-Cn}, \qquad P_0^n\phi_n \le e^{-Cn}, \qquad \sup_{p \in \mathcal{P}_n \cap \mathcal{U}^c} P^n(1-\phi_n) \le e^{-Cn},$$

then the posterior distribution $\Pi_n(\cdot|X_1,\ldots,X_n)$ in the model $X_1\ldots,X_n|\,p \overset{iid}{\sim} p$ and $p \sim \Pi$ is strongly consistent at p_0.

Proof We first show that for any $\epsilon > 0$ eventually a.s. $[P_0^\infty]$:

$$\int \prod_{i=1}^{n} \frac{p}{p_0}(X_i)\, d\Pi(p) \ge e^{-n\epsilon}. \tag{5.1}$$

Equivalently, the liminf of $e^{n\epsilon}$ times the integral in the left side is bounded below by 1, almost surely. We shall show that in fact this liminf is ∞, almost surely.

For any set $\mathcal{P}_0 \subset \mathcal{P}$ with positive prior mass the integral in (5.1) is bounded below by $\Pi(\mathcal{P}_0) \int \prod_{i=1}^n (p/p_0)(X_i)\, d\Pi_0(p)$, for Π_0 the renormalized restriction $\Pi(\cdot \cap \mathcal{P}_0)/\Pi(\mathcal{P}_0)$ of Π to $\mathcal{P}_0$. Therefore the logarithm of the integral is bounded below by

$$\log \Pi(\mathcal{P}_0) + \log \int \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi_0(p) \geq \log \Pi(\mathcal{P}_0) + \int \log \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi_0(p),$$

by Jensen's inequality applied to the logarithm (which is concave). The first term times n^{-1} tends to zero, while n^{-1} times the second term is the average

$$\frac{1}{n} \sum_{i=1}^n \int \log \frac{p}{p_0}(X_i)\, d\Pi_0(p) \to P_0 \int \log \frac{p}{p_0}\, d\Pi_0(p), \qquad a.s.$$

by the strong law of large numbers. The right side is $-\int K(p_0; p)\, d\Pi_0(p)$, and is strictly bigger than $-\epsilon$ for $\mathcal{P}_0 = \{p\colon K(p_0; p) < \epsilon\}$. We infer that n^{-1} times the logarithm of $e^{n\epsilon}$ times the left side of (5.1) is strictly positive, eventually, almost surely. Hence the liminf of this expression without the n^{-1} is ∞. This concludes the proof of (5.1).

Next fix a neighborhood $\mathcal{U}$ of p_0, and let C, $\mathcal{P}_n$ and the tests ϕ_n be as in the statement of the theorem. We shall show separately that $\Pi_n(\mathcal{P}_n \cap \mathcal{U}^c | X_1, \ldots, X_n) \to 0$ and that $\Pi_n(\mathcal{P}_n^c | X_1, \ldots, X_n) \to 0$, almost surely.

In view of Bayes's rule (1.4),

$$\Pi_n(\mathcal{P}_n \cap \mathcal{U}^c | X_1, \ldots, X_n) \leq \phi_n + \frac{(1-\phi_n) \int_{\mathcal{P}_n \cap \mathcal{U}^c} \prod_{i=1}^n (p/p_0)(X_i)\, d\Pi(p)}{\int \prod_{i=1}^n (p/p_0)(X_i)\, d\Pi(p)}.$$

The expectation of the first term is bounded by e^{-Cn} by assumption, whence $\sum_n P_0^n(\phi_n > \delta) < \sum_n \delta^{-1} e^{-Cn} < \infty$, by Markov's inequality, for every $\delta > 0$. This implies that $\phi_n \to 0$ almost surely, by the Borel-Cantelli lemma.

By (5.1) and the fact that p_0 is in the Kullback-Leibler support of Π the denominator of the second term is bounded below by a constant times $e^{-n\epsilon}$ eventually a.s., for every given ϵ. Thus this term of the display tends to zero if $e^{n\epsilon}$ times the numerator tends to zero. By Fubini's theorem,

$$\begin{aligned} P_0^n\Big((1-\phi_n) \int_{\mathcal{P}_n \cap \mathcal{U}^c} \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi(p)\Big) &= \int_{\mathcal{P}_n \cap \mathcal{U}^c} P_0^n\Big[(1-\phi_n) \prod_{i=1}^n \frac{p}{p_0}(X_i)\Big]\, d\Pi(p) \\ &\leq \int_{\mathcal{P}_n \cap \mathcal{U}^c} P^n(1-\phi_n)\, d\Pi(p) \leq e^{-Cn}. \end{aligned}$$

Since $\sum_n e^{n\epsilon} e^{-Cn} < \infty$ if $\epsilon < C$, the desired convergence of $e^{n\epsilon}$ times the numerator follows by Markov's inequality.

Finally we apply the argument of the preceding paragraph with $\mathcal{P}_n \cap \mathcal{U}^c$ replaced by $\mathcal{P}_n^c$ and the tests $\phi_n = 0$ instead of the given tests. The "power" $P^n(1-\phi_n)$ of this test is equal to one, but the final term of the preceding display can be bounded by $\Pi(\mathcal{P}_n^c)$, which is also of the order e^{-Cn}, by assumption. This shows that $\Pi_n(\mathcal{P}_n^c | X_1, \ldots, X_n) \to 0$, almost surely. □

5.2 Tests

In Schwartz's theorem, and also the theorem on contraction rates later on, the existence of tests ensures that the statistical model is not too complex. In this section we derive tests in the important case of i.i.d. observations and distances bounded above by the Hellinger distance.

5.2.1 Minimax theorem

The *minimax risk for testing* a probability P versus a set $\mathcal{Q}$ of probability measures is defined by

$$\pi(P, \mathcal{Q}) = \inf_{\phi} \Big(P\phi + \sup_{Q \in \mathcal{Q}} Q(1 - \phi) \Big), \tag{5.2}$$

where the infimum is taken over all *tests*, i.e. measurable functions $\phi \colon \mathfrak{X} \to [0, 1]$. The problem is to give a manageable bound on this risk, or equivalently on its two components, the probabilities of errors of the first kind $P\phi$ and of the second kind $Q(1 - \phi)$. We assume throughout that P and $\mathcal{Q}$ are dominated by a σ-finite measure μ, and denote by p and q the densities of the measures P and Q. Let $\operatorname{conv}(\mathcal{Q})$ denote the *convex hull* of $\mathcal{Q}$: the set of all finite convex combinations $\sum_{i=1}^{k} \lambda_i Q_i$ of elements $Q_i \in \mathcal{Q}$, where $(\lambda_1, \ldots, \lambda_k)$ is a probability vector.

The *Hellinger affinity* of two densities p and q is defined as

$$\rho_{1/2}(p, q) = \int \sqrt{p}\, \sqrt{q}\, d\mu.$$

It is related to the *Hellinger distance* $h(p, q)$ between p and q, whose square is defined by

$$h^2(p, q) = \int (\sqrt{p} - \sqrt{q})^2\, d\mu = 2 - 2\rho_{1/2}(p, q). \tag{5.3}$$

Proposition 5.10 (Minimax theorem for testing)**.** *For a probability measure P and dominated set of probability measures $\mathcal{Q}$,*

$$\pi(P, \mathcal{Q}) = 1 - \tfrac{1}{2} \|P - \operatorname{conv}(\mathcal{Q})\|_{TV} \leq \sup_{Q \in \operatorname{conv}(\mathcal{Q})} \rho_{1/2}(p, q).$$

Proof Replacing the set $\mathcal{Q}$ in the definition (5.2) of the minimax risk for testing by its convex hull certainly does not make this quantity smaller. Since the minimax risk takes the supremum over $\mathcal{Q}$, the replacement makes it no bigger either. Therefore the minimax risk for testing can be written in the form

$$\pi(P, \mathcal{Q}) = \inf_{\phi \in \Phi} \sup_{Q \in \operatorname{conv}(\mathcal{Q})} (P\phi + Q(1 - \phi)).$$

The next step is to interchange the order of inf and sup in this expression. Since the domains for both ϕ and Q are convex and the function $(\phi, Q) \mapsto P\phi + Q(1 - \phi)$ is linear in both arguments, this is permitted by the minimax theorem, Theorem 5.21, as soon as this function is continuous in ϕ relative to a topology that makes Φ a compact subset of a topological vector space. Now the set Φ of test-functions ϕ can be identified with the nonnegative functions in the unit ball of $L_\infty(\mathfrak{X}, \mathscr{X}, \mu)$, which is dual to $L_1(\mathfrak{X}, \mathscr{X}, \mu)$, since μ is σ-finite. The unit ball is compact and Hausdorff with respect to the weak*-topology, by the Banach-Alaoglu

theorem (cf. Theorem 3.15 of Rudin (1973)), and the nonnegative functions in this unit ball form a weak-* closed subset. Therefore the minimax theorem applies and the minimax risk for testing is equal to

$$\sup_{Q \in \mathrm{conv}(\mathcal{Q})} \inf_{\phi \in \Phi} (P\phi + Q(1-\phi)).$$

We finish the proof by explicitly determining the infimum in this expression, for fixed Q.

For fixed p, q the expression $P\phi + Q(1-\phi) = 1 + \int \phi(p-q)\,d\mu$ is minimized over all test functions by choosing ϕ equal to the minimal permitted value 0 if $p - q > 0$ and equal to the maximal permitted value 1 if $p - q < 0$. In other words, the infimum in the right side is attained for $\phi = \mathbb{1}\{p < q\}$, and the minimal value is equal to $1 + \int_{p<q}(p-q)\,d\mu = 1 - \frac{1}{2}\|p-q\|_1$, since $0 = \int (p-q)\,d\mu = \int_{p>q}(p-q)\,d\mu - \int_{p<q}(q-p)\,d\mu$. This proves the equality in the assertion of the theorem. For the inequality we write

$$P(p < q) + 1 - Q(p < q) = \int_{p<q} p\,d\mu + \int_{p \geq q} q\,d\mu,$$

and bound p in the first integral and q in the second by $\sqrt{p}\sqrt{q}$. □

The proposition shows the importance of the convex hull of $\mathcal{Q}$. Not the separation of $\mathcal{Q}$ from the null hypothesis, but the separation of its convex hull drives the error probabilities.

5.2.2 Product measures

We shall be interested in tests based on n i.i.d. observations, and therefore wish to apply Proposition 5.10 with the general P and Q replaced by product measures P^n and Q^n. Because the total variation distance between product measures is difficult to handle, the further bound by the Hellinger affinity is useful. By Fubini's theorem this is multiplicative in product measures:

$$\rho_{1/2}(p_1 \times p_2, q_1 \times q_2) = \rho_{1/2}(p_1, q_1)\rho_{1/2}(p_2, q_2).$$

When we take the supremum over convex hulls of sets of densities, then this multiplicativity is lost, but the following lemma shows that the Hellinger affinity is still "sub-multiplicative".

For $i = 1, \ldots, n$ let P_i and $\mathcal{Q}_i$ be a probability measure and a set of probability measures on an arbitrary measurable space $(\mathfrak{X}_i, \mathscr{X}_i)$, and consider testing the product $\otimes_i P_i$ versus the set $\otimes_i \mathcal{Q}_i$ of products $\otimes_i Q_i$ with Q_i ranging over $\mathcal{Q}_i$. For simplicity write $\rho_{1/2}(P, \mathcal{Q})$ for $\sup_{Q \in \mathcal{Q}} \rho_{1/2}(P, Q)$.

Lemma 5.11. *For any probability measures P_i and classes $\mathcal{Q}_i$ of probability measures (for $i = 1, \ldots, n$),*

$$\rho_{1/2}(\otimes_i P_i, \mathrm{conv}(\otimes_i \mathcal{Q}_i)) \leq \prod_i \rho_{1/2}(P_i, \mathrm{conv}(\mathcal{Q}_i)).$$

Proof It suffices to give the proof for $n = 2$; the general case follows by induction, as $\mathrm{conv}(\otimes_{i=1}^k \mathcal{Q}_i) \subset \mathrm{conv}(\mathrm{conv}(\otimes_{i=1}^{k-1} \mathcal{Q}_i) \times \mathcal{Q}_k)$. Any measure $Q \in \mathrm{conv}(\mathcal{Q}_1 \otimes \mathcal{Q}_2)$ can be represented by a density of the form $q(x, y) = \sum_j \kappa_j q_{1j}(x) q_{2j}(y)$, for nonnegative constants κ_j with $\sum_j \kappa_j =$

1, and q_{ij} densities of measures belong to $\mathcal{Q}_i$. Then $\rho_{1/2}(p_1 \times p_2, q)$ can be written in the form

$$\int p_1(x)^{1/2}\Big(\sum_j \kappa_j q_{1j}(x)\Big)^{1/2}\Big[\int p_2(y)^{1/2}\Big(\frac{\sum_j \kappa_j q_{1j}(x) q_{2j}(y)}{\sum_j \kappa_j q_{1j}(x)}\Big)^{1/2} d\mu_2(y)\Big]\, d\mu_1(x).$$

(If $\sum_j \kappa_j q_{1j}(x) = 0$, the quotient in the inner integral is interpreted as 0.) The inner integral is bounded by $\rho_{1/2}(P_2, \mathrm{conv}(\mathcal{Q}_2))$ for every fixed $x \in \mathfrak{X}$, since the function of y within the brackets is for every fixed x a convex combination of the densities q_{2j} (with weights proportional to $\kappa_j q_{1j}(x)$). After substitution of this upper bound the remaining integral is bounded by $\rho_{1/2}(P_1, \mathrm{conv}(\mathcal{Q}_1))$. □

Combining the preceding lemma with Proposition 5.10, we see that, for every *convex* set $\mathcal{Q}$ of measures:

$$\pi(P^n, \mathcal{Q}^n) \le \rho_{1/2}(P^n, \mathrm{conv}(\mathcal{Q}^n)) \le \rho_{1/2}(P, \mathcal{Q})^n.$$

Thus any convex set $\mathcal{Q}$ with Hellinger affinity to P smaller than 1 can be tested with error probabilities that decrease exponentially in the number of observations.

Proposition 5.12. *For any probability measure P and convex, dominated set of probability measures $\mathcal{Q}$ with $h(p, q) > \epsilon$ for every $q \in \mathcal{Q}$ and any $n \in \mathbb{N}$, there exists a test ϕ such that*

$$P^n\phi \le e^{-n\epsilon^2/2}, \qquad \sup_{Q \in \mathcal{Q}} Q^n(1-\phi) \le e^{-n\epsilon^2/2}.$$

Proof By (5.3) we have $\rho_{1/2}(P, \mathcal{Q}) = 1 - h^2(P, \mathcal{Q})/2$, which is bounded above by $1 - \epsilon^2/2$ by assumption. Combined with the display preceding the proposition we see that $\pi(P^n, \mathcal{Q}^n) \le (1 - \epsilon^2/2)^n \le e^{-n\epsilon^2/2}$, since $1 - x \le e^{-x}$, for every x. □

5.2.3 Entropy

The preceding theorem applies to convex alternatives. To handle alternatives that are not convex, such as the complement of a ball, we cover these with convex sets, and combine the corresponding tests into a single overall test. The power will then depend on the number of sets needed in a cover.

Definition 5.13 (Covering number and entropy)**.** Given a semi-metric d on a set $\mathcal{Q}$ and $\epsilon > 0$, the *covering number* $N(\epsilon, \mathcal{Q}, d)$ is defined as the minimal number of balls of radius ϵ needed to cover $\mathcal{Q}$. The logarithm of the covering number is called (metric) *entropy*. A set of points $Q_1, \ldots, Q_N$ in $\mathcal{Q}$ such that the balls $\{Q\colon d(Q, Q_i) < \epsilon\}$ cover $\mathcal{Q}$ is called an *ϵ-net*.

The covering number increases as ϵ decreases to zero. Except in trivial cases, they increase to infinity as $\epsilon \downarrow 0$. The rate of increase is a measure of the size of $\mathcal{Q}$. For instance, the covering numbers of a compact subset of Euclidean space of dimension d increase at rate $(1/\epsilon)^d$, as $\delta \downarrow 0$; the covering numbers of a space of functions with bounded derivatives are much bigger, their entropy is polynomial in $(1/\epsilon)$. There is a rich literature on covering numbers. See Section 5.6.1 for some examples.

Proposition 5.14. *Let d be a metric whose balls are convex and which is bounded above by the Hellinger distance h. Then for every $\epsilon > 0$ and n there exists a test ϕ such that*

$$P^n\phi \le N(\epsilon/4, \mathcal{Q}, d)\, e^{-n\epsilon^2/8} \qquad \sup_{Q\in\mathcal{Q}: d(P,Q)>\epsilon} Q^n(1-\phi) \le e^{-n\epsilon^2/8}.$$

Proof Choose a maximal set of points $Q_1, \ldots, Q_N$ in the set $\mathcal{Q}' := \{Q \in \mathcal{Q}: d(P, Q) > \epsilon\}$ such that $d(Q_k, Q_l) \ge \epsilon/2$, for every $k \ne l$. Because every ball in a cover of $\mathcal{Q}$ by balls of radius $\epsilon/4$ can contain at most one Q_l, it follows that $N(\epsilon/4, \mathcal{Q}, d) \ge N$. Furthermore, the N balls $B_l = \{Q': d(Q', Q_l) < \epsilon/2\}$ of radius $\epsilon/2$ around the Q_l cover $\mathcal{Q}'$, as otherwise the set $Q_1, \ldots, Q_N$ would not be maximal. Since $Q_l \in \mathcal{Q}'$, the distance of Q_l to P is at least ϵ and hence $h(B_l, P) \ge d(B_l, P) > \epsilon/2$ for every ball B_l. The balls B_l are convex by assumption. Therefore, by Proposition 5.12 there exists a test ϕ_l of P versus B_l with error probabilities bounded above by $e^{-n\epsilon^2/8}$. Let $\phi = \max_l \phi_l$ be the maximum of all the tests ϕ_l obtained in this way, for $l = 1, 2, \ldots, N$. Then

$$P^n\phi \le \sum_{l=1}^{N} P^n\phi_l \le N(\epsilon/4, \mathcal{Q}, d)e^{-n\epsilon^2/8}, \qquad \text{and} \qquad \sup_{Q\in\mathcal{Q}'} Q^n(1-\phi) \le e^{-n\epsilon^2/8}.$$

The first inequality follows, because the maximum is smaller than the sum; the second, because by construction for every $Q \in \mathcal{Q}'$ there exists a ball with $Q \in B_l$, and $1 - \phi \le 1 - \phi_l$, by construction. □

5.3 Consistency under an entropy bound

We combine the results on testing with the extended Schwartz theorem, Theorem 5.9, to obtain sufficient conditions for consistency in terms of entropy. We could use any distance that is bounded above by a multiple of the Hellinger distance, but for simplicity choose the $\mathbb{L}_1$-distance (which is twice the total variation distance). By the Cauchy-Schwarz inequality and the inequality $(\sqrt{p} + \sqrt{q})^2 \le 2(p + q)$, for any probability distances p and q,

$$\int |p - q|\, d\nu \le \Big(\int (\sqrt{p} - \sqrt{q})^2\, d\nu \int (\sqrt{p} + \sqrt{q})^2\, d\nu\Big)^{1/2} \le 2h(p, q).$$

Theorem 5.15 (Consistency in total variation)**.** *The posterior distribution is strongly consistent relative to the L_1-distance at every $p_0 \in \mathrm{KL}(\Pi)$ if for every $\epsilon > 0$ there exist measurable sets $\mathcal{P}_n \subset \mathcal{P}$ (which may depend on ϵ) such that, for constants $C > 0$, $\xi < 1/2$, and sufficiently large n,*

$$\Pi(\mathcal{P} \setminus \mathcal{P}_n) \le e^{-Cn}, \qquad \log N(\epsilon, \mathcal{P}_n, \|\cdot\|_1) \le \xi n\epsilon^2.$$

Proof Because the L_1-distance is bounded above by twice the Hellinger distance, we can apply Theorem 5.14 with $d = \|\cdot\|_1/2$. Then $N(\epsilon/2, \mathcal{P}_n, d) \le N(\epsilon, \mathcal{P}_n, \|\cdot\|_1) \le e^{\xi n\epsilon^2}$. Therefore, by Proposition 5.14 applied with 2ϵ instead of ϵ, there exists a test ϕ_n with

$$P_0^n\phi_n \le e^{\xi n\epsilon^2} e^{-n\epsilon^2/2}, \qquad \sup_{p\in\mathcal{P}_n: \|p-p_0\|_1>4\epsilon} P^n(1-\phi_n) \le e^{-n\epsilon^2/2}.$$

Thus the conditions of Theorem 5.9 are satisfied, where we take $\mathcal{U}$ the ball of radius 5ϵ around p_0 and set the constant C equal to the minimum of $(1/2 - \xi)\epsilon^2/2$ and the present constant C. □

Example 5.16 (Totally bounded model)**.** As a very crude example consider a model $\mathcal{P}$ such that $N(\epsilon, \mathcal{P}, \|\cdot\|_1) < \infty$, for every $\epsilon > 0$. Clearly for every fixed $\epsilon > 0$ the inequality $\log N(\epsilon, \mathcal{P}, \|\cdot\|_1) < n\epsilon^2/4$, is satisfied for sufficiently large n (which depends on ϵ). Thus the entropy condition is satisfied for $\mathcal{P}_n = \mathcal{P}$, and then also $\Pi(\mathcal{P} \setminus \mathcal{P}_n) = 0$ is trivially satisfied. Thus the theorem gives consistency at every probability density p_0 that possesses the Kullback-Leibler property.

A concrete example is the set of all densities $p: [0,1] \to [0, K]$ that are uniformly Lipschitz: $|p(x) - p(y)| \le K|x - y|^\beta$, for every $x, y \in [0,1]$ and some given constants $K, \beta > 0$ (see Proposition 5.24). This example shows that the entropy condition in the preceding theorem is not overly strong, as a Lipschitz condition is relatively weak. That the condition is quantitative through the constant K is less pleasant. It would force us to use a prior that is dependent on K, and which K would we use?

We can use the sieves $\mathcal{P}_n$ to resolve such questions and prove consistency for many more priors. However, consistency is a very weak property, and it is not overly satisfying to apply the preceding theorem to more general situations. In the next section we turn to stronger results.

5.4 Rate of contraction

For every $n \in \mathbb{N}$ let $X^{(n)}$ be an observation in a sample space $(\mathfrak{X}^{(n)}, \mathscr{X}^{(n)})$ with distribution $P_\theta^{(n)}$ indexed by a parameter θ belonging to a metric space Θ. Given a prior Π_n on the Borel sets of Θ, let $\Pi_n(\cdot | X^{(n)})$ be a version of the posterior distribution.

Definition 5.17 (Posterior rate of contraction)**.** The posterior distribution $\Pi_n(\cdot | X^{(n)})$ is said to *contract at rate* $\epsilon_n \to 0$ at $\theta_0 \in \Theta$ if $\Pi_n(\theta: d(\theta, \theta_0) > M_n\epsilon_n | X^{(n)}) \to 0$ in $P_{\theta_0}^{(n)}$-probability, for every $M_n \to \infty$ as $n \to \infty$.

A rough interpretation of the rate ϵ_n is that the posterior distribution concentrates on balls of radius "of the order ϵ_n" around θ_0. The construction using the additional sequence M_n expresses the "of the order" part of this assertion. For "every $M_n \to \infty$" must be read as "whenever $M_n \to \infty$, no matter how slowly". Actually, in most nonparametric applications the fixed sequence $M_n = M$ for a large constant M also works. (In many parametric applications, the posterior distribution tends after scaling to a distribution that is supported on the full space, and letting M_n tend to infinity is important.)

We may view the rate of contraction as the natural refinement of consistency. Consistency requires that the posterior distribution contracts to within arbitrarily small distance ϵ to the true parameter θ_0; the rate as defined here quantifies "arbitrarily small". Typically contraction rates are much more informative about the quality of a Bayesian procedure than is revealed by mere consistency.

If ϵ_n is a rate of contraction, then every sequence that tends to zero at a slower rate is also a contraction rate, according to the definition. Saying that the contraction rate is *at least* ϵ_n would be appropriate. Naturally we are interested in the fastest contraction rate, but we are typically already satisfied with knowing some rate that is valid for every θ_0 in a given class of true parameters.

An appropriate summary of the location of the posterior distribution inherits its rate of

contraction.‘ The same summary as used in Proposition 5.2 also works for rates. The proof is very similar too and omitted.

Proposition 5.18 (Point estimator)**.** *Suppose that the posterior distribution $\Pi_n(\cdot|X^{(n)})$ contracts at rate ϵ_n at θ_0 relative to the metric d on Θ. Then $\hat\theta_n$ defined as the center of a (nearly) smallest ball that contains posterior mass at least $1/2$ satisfies $d(\hat\theta_n,\theta_0) = O_P(\epsilon_n)$ under $P_{\theta_0}^{(n)}$.*

In particular, the posterior distribution cannot contract faster than the best point estimator. This makes it possible to connect the theory of posterior contraction rates to the theory of "optimal" rates of estimation, which are typically defined by the minimax criterion. For a given set Θ of possible parameters the *minimax rate* is the fastest rate so that, for some sequence of estimators and every $M_n \to \infty$,

$$\sup_{\theta\in\Theta} P_\theta^{(n)}(d(\hat\theta_n,\theta) > M_n\epsilon_n) \to 0.$$

We would like priors that give the same rate of contraction ϵ_n, preferably uniformly in $\theta \in \Theta$.

5.5 Contraction under an entropy bound

Let the observations be a random sample $X_1,\dots,X_n$ from a density p that belongs to a set of densities $\mathcal{P}$, relative to a given σ-finite measure ν. Let Π_n be a prior on $\mathcal{P}$, and let p_0 denote the true density of the observations.

Let d be a distance on $\mathcal{P}$ that is bounded above by the Hellinger distance h, and set

$$K(p_0;p) = P_0\log\frac{p_0}{p}, \qquad V(p_0;p) = P_0\Big(\log\frac{p_0}{p}\Big)^2. \tag{5.4}$$

The first is the Kullback-Leibler divergence, the second a corresponding second moment. For simplicity of notation also define K_2 as the maximum of these quantities:

$$K_2(p_0;p) = \max\Big(K(p_0;p), V(p_0;p)\Big). \tag{5.5}$$

Theorem 5.19. *Let $d \le h$ be a metric whose balls are convex. The posterior distribution contracts at rate ϵ_n at p_0 for any ϵ_n such that there exists $\underline\epsilon_n \le \epsilon_n$ such that $n\underline\epsilon_n^2 \to \infty$ and such that, for positive constants c_1, c_2 and sets $\mathcal{P}_n \subset \mathcal{P}$,*

$$\log N(\epsilon_n,\mathcal{P}_n,d) \le c_1 n\epsilon_n^2, \tag{5.6}$$

$$\Pi_n(p\colon K_2(p_0;p) < \underline\epsilon_n^2) \ge e^{-c_2 n\underline\epsilon_n^2}, \tag{5.7}$$

$$\Pi_n(\mathcal{P}\setminus\mathcal{P}_n) \le e^{-(c_2+3)n\underline\epsilon_n^2}. \tag{5.8}$$

Proof For every $\epsilon \ge 4\epsilon_n$ we have $\log N(\epsilon/4,\mathcal{P}_n,d) \le \log N(\epsilon_n,\mathcal{P}_n,d) \le c_1 n\epsilon_n^2$, by assumption (5.6). Therefore, by Proposition 5.14 applied with $\epsilon = M\epsilon_n$, where $M \ge 4$ is a large constant to be chosen later, there exist tests ϕ_n with errors

$$P_0^n\phi_n \le e^{c_1 n\epsilon_n^2}e^{-nM^2\epsilon_n^2/8}, \qquad \sup_{p\in\mathcal{P}_n:d(p,p_0)>M\epsilon_n} P^n(1-\phi_n) \le e^{-nM^2\epsilon_n^2/8}.$$

For $M^2/8 > c_1$ the first expression tends to zero. For A_n the event $\{\int\prod_{i=1}^n(p/p_0)(X_i)\,d\Pi_n(p) \ge$

$e^{-(2+c_2)n\underline{\epsilon}_n^2}\}$ we can bound $\Pi_n(p\colon d(p,p_0) > M\epsilon_n | X_1,\ldots,X_n)$ by

$$\phi_n + \mathbb{1}\{A_n^c\} + e^{(2+c_2)n\underline{\epsilon}_n^2}(1-\phi_n)\int_{d(p,p_0)>M\epsilon_n}\prod_{i=1}^n \frac{p}{p_0}(X_i)\,d\Pi_n(p).$$

The expected value under P_0^n of the first term tends to zero, by choice of M and the preceding display. The same is true for the second term, by (5.7) and Lemma 5.20 (below). We split the integral in the third term over the intersection with the domain $\mathcal{P}_n$ and its complement.

The first is

$$P_0^n\Big[(1-\phi_n)\int_{p\in\mathcal{P}_n:d(p,p_0)>M\epsilon_n}\prod_{i=1}^n \frac{p}{p_0}(X_i)\,d\Pi_n(p)\Big] \le \int_{p\in\mathcal{P}_n:d(p,p_0)>M\epsilon_n} P^n(1-\phi_n)\,d\Pi_n(p),$$

which is bounded by $e^{-nM^2\epsilon_n^2/8}$, by the construction of the tests. The second is bounded by

$$P_0^n\int_{\mathcal{P}\setminus\mathcal{P}_n}\prod_{i=1}^n \frac{p}{p_0}(X_i)\,d\Pi_n(p) \le \Pi_n(\mathcal{P}\setminus\mathcal{P}_n).$$

This is bounded above by $e^{-(c_2+3)n\underline{\epsilon}_n^2}$ by (5.8). For $M^2/8 > 2 + c_2$ all terms tend to zero. □

The sequence $\underline{\epsilon}_n$ may be chosen equal to the rate ϵ_n, but allowing a different rate gives some flexibility that is important to treat some examples.

The condition $n\underline{\epsilon}_n^2 \to \infty$ excludes the parametric rate $\epsilon_n = n^{-1/2}$, and merely says that we are considering the nonparametric situation, where slower rates obtain. The main conditions of the theorem are (5.6)-(5.8).

Condition (5.8) is trivially satisfied by choosing $\mathcal{P}_n = \mathcal{P}$ for every n. Similar as in the consistency theorems, this condition expresses that a subset $\mathcal{P}\setminus\mathcal{P}_n$ of the model $\mathcal{P}_n$ that receives very little prior mass does not play a role in the rate of contraction.

The remaining pair (5.6)-(5.7) of conditions is more structural. For given $\mathcal{P}_n$ and c_1, c_2 each of the two conditions on its own determines a minimal value of ϵ_n (as their left sides decrease and their right sides increase if ϵ_n is replaced by a bigger value). The rate of contraction is the slowest one defined by the two inequalities. Condition (5.7) involves the prior, whereas condition (5.6) does not. The latter condition bounds the complexity of the model, and should be viewed as characterizing the best rate of estimation by any method.

Lemma 5.20 (Evidence bound)**.** *For any probability measure* Π *on* $\mathcal{P}$*, and positive constant* ϵ*, with* P_0^n*-probability at least* $1-(n\epsilon^2)^{-1}$*,*

$$\int\prod_{i=1}^n \frac{p}{p_0}(X_i)\,d\Pi(p) \ge \Pi(p\colon K_2(p_0;p) < \epsilon^2)\,e^{-2n\epsilon^2}.$$

Proof The integral becomes smaller by restricting it to the set $B := \{p\colon K_2(p_0;p) < \epsilon_n^2\}$. By next dividing the two sides of the inequality by $\Pi(B)$, we can rewrite the inequality in terms of the prior Π_0 obtained by restricting Π to B and renormalizing it to a probability measure. By Jensen's inequality applied to the logarithm,

$$\log\int\prod_{i=1}^n \frac{p}{p_0}(X_i)\,d\Pi_0(P) \ge \sum_{i=1}^n\int\log\frac{p}{p_0}(X_i)\,d\Pi_0(P) =: Z.$$

The right side has mean $\mathrm{E}Z = -n\int K(p_0;p)\,d\Pi_0(p) > -n\epsilon^2$, because Π_0 concentrates on B, and variance satisfying

$$\operatorname{var} Z \le nP_0\Big(\int \log\frac{p_0}{p}\,d\Pi_0(p)\Big)^2 \le nP_0\int\Big(\log\frac{p_0}{p}\Big)^2 d\Pi_0(p) \le n\epsilon^2,$$

by Jensen's inequality, Fubini's theorem, and again the fact that $\Pi_0(B) = 1$. It follows that

$$P_0^n(Z < -2n\epsilon^2) \le P_0^n(Z - \mathrm{E}Z < -n\epsilon^2) \le \frac{n\epsilon^2}{(n\epsilon^2)^2},$$

by Chebyshev's inequality. □

5.6 Complements

Theorem 5.21 (Minimax theorem)**.** *Let T be a compact, convex set of a locally convex topological vector space and S a convex subset of a linear space. Let $f\colon T\times S\to\mathbb{R}$ be a function such that*

(i) *$t\mapsto f(t,s)$ is continuous and concave for all $s\in S$;*
(ii) *$s\mapsto f(t,s)$ is convex for all $s\in S$.*

Then

$$\inf_{s\in S}\sup_{t\in T} f(t,s) = \sup_{t\in T}\inf_{s\in S} f(t,s). \tag{5.9}$$

For a proof, see Strasser (1985), pages 239–241.

The following proposition is a sharper version of Proposition 5.14. Instead of counting the cover number of the full alternative, it partitions the alternative in rings $\{Q\in\mathcal{Q}\colon j\epsilon < d(Q,P) < 2j\epsilon\}$ of increasing radius, given by $j = 1,2,\ldots$ and uses the supremum of the $j\epsilon/2$-cover numbers of these rings. This makes a difference mostly for finite-dimensional models $\mathcal{Q}$, as for big models the orders of the cover numbers for different j are similar.

Proposition 5.22. *Let d be a metric whose balls are convex and which is bounded above by the Hellinger distance h. Then for every $\epsilon > 0$ and n there exists a test ϕ such that, for all $j\in\mathbb{N}$,*

$$P^n\phi \le \sup_{j\in\mathbb{N}} N(\epsilon j/4, \{Q\in\mathcal{Q}\colon j\epsilon < d(Q,P) < 2j\epsilon\}, d)\frac{e^{-n\epsilon^2/8}}{1-e^{-n\epsilon^2/8}}, \qquad \sup_{Q\in\mathcal{Q}:d(P,Q)>j\epsilon} Q^n(1-\phi) \le e^{-n\epsilon^2 j^2/8}.$$

Proof For a given $j\in\mathbb{N}$, choose a maximal set of points $Q_{j,1},\ldots,Q_{j,N_j}$ in the set $\mathcal{Q}_j := \{Q\in\mathcal{Q}\colon j\epsilon < d(P,Q) < 2j\epsilon\}$ such that $d(Q_{j,k},Q_{j,l}) \ge j\epsilon/2$ for every $k\ne l$. Because every ball in a cover of $\mathcal{Q}_j$ by balls of radius $j\epsilon/4$ then contains at most one $Q_{j,l}$, it follows that $N_j \le N(j\epsilon/4,\mathcal{Q}_j,d)$. Furthermore, the N_j balls $B_{j,l}$ of radius $j\epsilon/2$ around the $Q_{j,l}$ cover $\mathcal{Q}_j$, as otherwise this set was not maximal. Since $Q_{j,l}\in\mathcal{Q}_j$, the distance of $Q_{j,l}$ to P is at least $j\epsilon$ and hence $h(P,B_{j,l}) \ge d(P,B_{j,l}) > j\epsilon/2$ for every ball $B_{j,l}$. By Proposition 5.12 there exists a test $\phi_{j,l}$ of P versus $B_{j,l}$ with error probabilities bounded above by $e^{-nj^2\epsilon^2/8}$. Let ϕ be the supremum of all the tests $\phi_{j,l}$ obtained in this way, for $j = 1,2,\ldots$, and $l = 1,2,\ldots,N_j$. Then

$$P^n\phi \le \sum_{j=1}^{\infty}\sum_{l=1}^{N_j} e^{-nj^2\epsilon^2/8} \le \sum_{j=1}^{\infty} N(j\epsilon/4,\mathcal{Q}_j,d)e^{-nj^2\epsilon^2/8} \le N(\epsilon/4,\mathcal{Q},d)\frac{e^{-n\epsilon^2/8}}{1-e^{-n\epsilon^2/8}}$$

and, for every $j\in\mathbb{N}$,

$$\sup_{Q\in\cup_{l>j}\mathcal{Q}_l} Q^n(1-\phi) \le \sup_{l>j} e^{-nl^2\epsilon^2/8} \le e^{-nj^2\epsilon^2/8},$$

since for every $Q\in\mathcal{Q}_j$ there exists a test $\phi_{j,l}$ with $1-\phi \le 1-\phi_{j,l}$, by construction. □

5.6.1 Examples of entropy

Lemma 5.23 (Euclidean ball). *For $\|x\|_p = (\sum_i |x_i|^p)^{1/p}$ and $p \geq 1$, for any M and $\epsilon \in (0, M)$,*

$$N(\epsilon, \{x \in \mathbb{R}^d : \|x\|_p \leq M\}, \|\cdot\|_p) \leq \left(\frac{3M}{\epsilon}\right)^d. \tag{5.10}$$

Proof We can reduce to the case $M = 1$ by scaling. Let $x_1, \ldots, x_N$ be a maximal set of points in the closed unit ball $\{x : \|x\|_p \leq 1\}$ such that $\|x_i - x_j\|_p \geq \epsilon$, for $i \neq j$, where "maximal" means that it is impossible to add a further point while keeping the inequality $\|x_i - x_j\|_p \geq \epsilon$. Then any point of the unit ball is within distance ϵ of at least one $x_1, \ldots, x_N$ and hence $N(\epsilon, \{x : \|x\|_p \leq 1\}, \|\cdot\|_p) \leq N$. The open balls $B(x_i, \epsilon/2)$ around the points $x_1, \ldots, x_N$ are disjoint and their union is contained in $B(0, 1 + \epsilon/2)$. Comparing the volume of the union to the volume of the latter ball gives the inequality $N(\epsilon/2)^d \leq (1 + \epsilon/2)^d$. Hence $N \leq ((2 + \epsilon)/\epsilon)^d \leq (3/\epsilon)^d$, for $\epsilon < 1$. □

The preceding bounds show that entropy numbers of sets in Euclidean spaces grow logarithmically. For infinite-dimensional spaces the growth is much faster, as is illustrated by the following example.

The *Hölder norm* of order β of a continuous function $f : \mathfrak{X} \to \mathbb{R}$ on a bounded, convex subset $\mathfrak{X} \subset \mathbb{R}^d$ is defined as

$$\|f\|_\beta = \max_{k:|k| \leq \underline{\beta}} \sup_{x \in \mathfrak{X}} |D^k f(x)| + \max_{k:|k| = \underline{\beta}} \sup_{x,y \in \mathfrak{X}: x \neq y} \frac{|D^k f(x) - D^k f(y)|}{\|x - y\|^{\beta - \underline{\beta}}}. \tag{5.11}$$

Here $\underline{\beta}$ is the biggest integer strictly smaller than β, and for a vector $k = (k_1, \ldots, k_n)$ of integers, D^k is the partial differential operator

$$D^k = \frac{\partial^{|k|}}{\partial x_1^{k_1} \cdots \partial x_d^{k_d}}.$$

Lemma 5.24 (Hölder ball). *There exists a constant K depending only on $\mathfrak{X}$, d and β such that,*

$$\log N(\epsilon, \{f : \|f\|_\beta \leq M\}, \|\cdot\|_\infty) \leq K \left(\frac{M}{\epsilon}\right)^{d/\beta}.$$

Definition 5.25 (Sobolev space). The *Sobolev space* $\mathfrak{W}^\alpha(\mathfrak{X})$ of order $\alpha \in \mathbb{N}$ for an interval $\mathfrak{X} \subset \mathbb{R}$ is the class of all functions $f \in \mathbb{L}_2(\mathfrak{X})$ that possess an absolutely continuous $(\alpha-1)$th derivative whose (weak) derivative $f^{(\alpha)}$ is contained in $\mathbb{L}_2(\mathfrak{X})$; the space is equipped with the norm $\|f\|_{2,2,\alpha} = \|f\|_2 + \|f^{(\alpha)}\|_2$. The Sobolev space $\mathfrak{W}^\alpha(\mathbb{R}^m)$ of order $\alpha > 0$ is the class of all functions $f \in \mathbb{L}_2(\mathbb{R}^m)$ with Fourier transform $\hat{f}$ satisfying $v_\alpha(f)^2 := \int |\lambda|^{2\alpha} |\hat{f}(\lambda)|^2 \, d\lambda < \infty$; the space is equipped with the norm $\|f\|_{2,2,\alpha} = \|f\|_2 + v_\alpha(f)$.

The preceding definition is restricted to functions of "integral" smoothness $\alpha \in \mathbb{N}$ with domain an interval in the real line or functions of general smoothness $\alpha > 0$ with domain a full Euclidean space. The relation between the two cases is that for $\alpha \in \mathbb{N}$ the function $\lambda \mapsto (i\lambda)^\alpha \hat{f}(\lambda)$ is the Fourier transform of the (weak) αth derivative $f^{(\alpha)}$ of a function $f : \mathbb{R} \to \mathbb{R}$ and hence $v_\alpha(f) = \|f^{(\alpha)}\|_2$. The Fourier transform is not entirely natural for functions not defined on the full Euclidean space, which makes definitions of Sobolev spaces for general domains and smoothness a technical matter. One possibility suggests itself by the fact that in both cases the Sobolev space is known to be equivalent to the Besov space $\mathfrak{B}^\alpha_{2,2}(\mathfrak{X})$. Hence we may define a Sobolev space $\mathfrak{W}^\alpha(\mathfrak{X})$ in general as the corresponding Besov space. This identification suggested the notation $\|\cdot\|_{2,2,\alpha}$ for the norm. Another possibility arise for *periodic* functions, which we briefly indicate below.

The Sobolev spaces of functions on a compact domain (identified with the corresponding Besov space) are compact in $\mathbb{L}_r(\mathbb{R})$ if the smoothness level is high enough: $\alpha > m(1/2 - 1/r)_+$. The entropy is not bigger than the entropy of the smaller Hölder spaces and hence the following proposition generalizes and improves Lemma 5.24.

Proposition 5.26 (Sobolev space)**.** *For $\alpha > m(1/2 - 1/r)_+$ and $r \in (0, \infty]$ there exists a constant K that depends only on α and r such that*

$$\log N(\epsilon, \{f \in \mathfrak{W}^\alpha[0,1]^m \colon \|f\|_{2,2,\alpha} \leq M\}, \mathbb{L}_r([0,1]^m) \leq K\Big(\frac{M}{\epsilon}\Big)^{m/\alpha}.$$

Proposition 5.27 (Analytic functions)**.** *The class $\mathfrak{A}_A[0,1]^m$ of all functions $f \colon [0,1]^m \to \mathbb{R}$ that can be extended to an analytic function on the set $G = \{z \in \mathbb{C}^m \colon \|z - [0,1]^m\|_\infty < A\}$ with $\sup_{z \in G}\big|f(z)\big| \leq 1$ satisfies, for $\epsilon < 1/2$ and a constant c that depends on m only,*

$$\log N(\epsilon, \mathfrak{A}_A[0,1]^m, \|\cdot\|_\infty) \leq c\Big(\frac{1}{A}\Big)^m\Big(\log\frac{1}{\epsilon}\Big)^{1+m}.$$

A function $f \in \mathbb{L}_2[0, 2\pi]$ may be identified with its sequence of Fourier coefficients $(f_j) \in \ell_2$. The αth derivative of f has Fourier coefficients $((ij)^\alpha f_j)$. This suggests to think of the ℓ_2-norm of the sequence $(j^\alpha f_j)$ as a *Sobolev norm* of order α. The following proposition gives the entropy of the corresponding Sobolev sequence space relative to the ℓ_2-norm.

Proposition 5.28 (Sobolev sequence)**.** *For $\|\vartheta\|_2 = \big(\sum_{i=1}^\infty \theta_i^2\big)^{1/2}$ the norm of ℓ_2 and $\alpha > 0$, for all $\epsilon > 0$,*

$$\log D\Big(\epsilon, \{\vartheta \in \ell_2 \colon \sum_{i=1}^\infty i^{2\alpha}\theta_i^2 \leq B^2\}, \|\cdot\|_2\Big) \leq \log(4(2e)^{2\alpha})\Big(\frac{3B}{\epsilon}\Big)^{1/\alpha}.$$

Exercises

5.1 Suppose that the posterior distribution is strongly consistent. Show that the sequence of posterior distributions converges almost surely to the Dirac measure relative to the weak topology.

5.2 Extend the preceding exercise to weak consistency.

5.3 Suppose that $\Theta = \mathbb{R}^d$, and the posterior distribution is strongly consistent and $\int \|\theta\|^p \, d\Pi_n(\theta \,|\, X^{(n)})$ is bounded in probability or almost surely for some $p > 1$. Show that the posterior mean $\int \theta \, d\Pi_n(\theta \,|\, X^{(n)})$ is consistent as a point estimator.

5.4 Suppose that the posterior distribution $\Pi(\cdot \,|\, X^{(n)})$ of a probability density is consistent relative to the L_1-distance on the parameter set of densities. Show that the posterior mean density $x \mapsto \int p(x) \, d\Pi_n(p \,|\, X^{(n)})$ is consistent in L_1, as a point estimator for a density.

5.5 Consider the set $\mathcal{F}$ of functions $f \colon [0,1] \to [0,1]$ such that $\big|f(x) - f(y)\big| \leq |x - y|$, for every $x, y \in [0,1]$. Show that there exists a constant K such that $\log N(\epsilon, \mathcal{F}, \|\cdot\|_\infty) \leq K(1/\epsilon)$, for $\epsilon < 1$. [This is a special case of Lemma 5.24. Give a direct proof. Use balls around piecewise constant (or linear) functions.]

5.6 Suppose d_1 and d_2 are metrics with $d_1 \leq d_2$. Show that $N(\epsilon, Q, d_1) \leq N(\epsilon, Q, d_2)$, for every $\epsilon > 0$.

5.7 Prove Proposition 5.18.

5.8 Suppose that $X_1, \ldots, X_n$ are a random sample from the uniform distribution on $[0, \theta]$ and θ is equipped with a gamma prior with parameters (r, s). Use Schwartz's theorem to show that the posterior distribution of θ is consistent at θ_0. [Hint: the Kullback-Leibler divergence $K(U_{[0,\theta_0]}; U_{[0,\theta]})$ is infinite if $\theta < \theta_0$, but continuous in θ for $\theta \geq \theta_0$. Devise uniformly consistent tests for $H_0 \colon \theta = \theta_0$ versus $H_1 \colon |\theta - \theta_0| > \epsilon$.]

5.9 If a model is parameterized by a parameter θ and $g \colon \Theta \to H$ is a map that is continuous at θ_0, then consistency of the posterior distribution of θ at θ_0 implies the consistency of the posterior distribution of the parameter $\eta = g(\theta)$ at $g(\theta_0)$. Show this.

6

Dirichlet Process Mixtures

Because the Dirichlet process is discrete, it is useless as a prior for a distribution of which we wish to estimate the density. This can be remedied by convolving it with a kernel. For each θ in a parameter set Θ, let $x \mapsto \psi(x, \theta)$ be a probability density function relative to some σ-finite dominating measure μ, measurable in its two arguments. For a measure F on Θ define a *mixture density* by

$$p_F(x) = \int \psi(x, \theta)\, dF(\theta).$$

By equipping F with a prior, we obtain a prior on densities. Densities p_F with F following a Dirichlet process prior are known as *Dirichlet mixtures*.

As the Dirichlet process prior is defined for any Polish space Θ, the parameter θ in this mixture representation can be high-dimensional, but in most applications it is Euclidean of low dimension. It is often considered convenient to make the kernel depend on an additional parameter $\sigma \in \mathcal{S}$, which is given its own prior, giving mixtures of the form

$$p_{F,\sigma}(x) = \int \psi(x, \theta, \sigma)\, dF(\theta). \tag{6.1}$$

An example is to set $\psi(\cdot, \theta, \sigma)$ equal to the density of the $N(\theta, \sigma^2)$-distribution. Equipping this with a prior on σ and a Dirichlet process prior on F, results in a "mixture of Dirichlet mixtures".

Throughout the chapter the observations $X_1, \ldots, X_n$ will be a random sample from a density p_0 (relative to μ) on the sample space. When discussing posterior contraction, this density is not necessarily assumed to be one of the mixtures $p_{F,\sigma}$. The idea is that for a well chosen kernel *any* density can be approximated by mixture densities with appropriately chosen (F, σ), so that contraction to p_0 may pertain even if p_0 is not a mixture itself. On the other hand, when discussing the computation of the posterior distribution in the next section, we adopt the Bayesian model, which assumes that the observations are a sample from $p_{F,\sigma}$ given the pair (F, σ) generated from the prior.

6.1 Computation

In this section we discuss an MCMC algorithm to compute the posterior distribution resulting from a Dirichlet mixture.

For $x \mapsto \psi(x;\theta,\sigma)$ a probability density function, consider the Bayesian model

$$X_i|\, F,\sigma \overset{\text{iid}}{\sim} p_{F,\sigma}(x) = \int \psi(x;\theta,\sigma)\, dF(\theta), \qquad i = 1,\ldots,n. \tag{6.2}$$

We equip F and σ with independent priors $F \sim \mathrm{DP}(\alpha)$ and $\sigma \sim \pi$. The resulting model can be equivalently written in terms of *n latent variables* $\theta_1,\ldots,\theta_n$ as

$$X_i|\, \theta_i, F,\sigma \overset{\text{ind}}{\sim} \psi(\cdot;\theta_i,\sigma), \qquad \theta_i|\, F,\sigma \overset{\text{iid}}{\sim} F, \qquad F \sim \mathrm{DP}(\alpha), \qquad \sigma \sim \pi. \tag{6.3}$$

The posterior distribution of any object of interest can be described in terms of the posterior distribution of (F,σ) given $X_1,\ldots,X_n$. The latent variables $\theta_1,\ldots,\theta_n$ help to make the description simpler. Indeed,

- $F|\, \theta_1,\ldots,\theta_n \sim \mathrm{DP}(\alpha + \sum_{i=1}^n \delta_{\theta_i})$, by Theorem 4.14.
- Given $\sigma,\theta_1,\ldots,\theta_n$, the observations $X_1,\ldots,X_n$ and F are independent.

Hence the conditional distribution of F given $\sigma,\theta_1,\ldots,\theta_n,X_1,\ldots,X_n$ is free of the observations $X_1,\ldots,X_n$, and equal to the Dirichlet process $\mathrm{DP}(\alpha + \sum_{i=1}^n \delta_{\theta_i})$. We thus obtain

$$\begin{aligned}\Pr(F \in B|\, X_1,\ldots,X_n) &= \int \Pr(F \in B|\, \sigma,\theta_1,\ldots,\theta_n,X_1,\ldots,X_n)\, dP^{\sigma,\theta_1,\ldots,\theta_n|X_1,\ldots,X_n}(\sigma,\theta_1,\ldots,\theta_n)\\ &= \int \mathrm{DP}(\alpha + \sum_{i=1}^n \delta_{\theta_i})(B)\, dP^{\sigma,\theta_1,\ldots,\theta_n|X_1,\ldots,X_n}(\sigma,\theta_1,\ldots,\theta_n).\end{aligned}$$

The integrand is the probability of the set B under the Dirichlet process $\mathrm{DP}(\alpha + \sum_{i=1}^n \delta_{\theta_i})$-distribution. The set B is a general set of measures, and hence the probability is somewhat abstract. However, for given $\theta_1,\ldots,\theta_n$ we can replace the Dirichlet process by its series approximation, or we can consider sets B of the form $\{F\colon (F(A_1),\ldots,F(A_k)) \in B_k\}$ for partitions $\mathfrak{X} = \cup_i A_i$ of the sample space and use the definition of the Dirichlet process. The integrating measure is the posterior distribution of $\sigma,\theta_1,\ldots,\theta_n$. There are analytical formulas for this distribution, but these are too unwieldy for practical use. Computation is typically done by simulating samples $\sigma,\theta_1,\ldots,\theta_n$ from their posterior distribution. The theorem below describes a Gibbs sampling scheme for this purpose.

The decomposition can also be read as a recipe for simulating F from its posterior distribution: first simulate $\sigma,\theta_1,\ldots,\theta_n$ from its posterior distribution (e.g. by the Gibbs sampler below), next given these values simulate F from the $\mathrm{DP}(\alpha + \sum_{i=1}^n \delta_{\theta_i})$-distribution.

For the expectation of a linear functional of a Dirichlet process there is an explicit formula (see Exercise 4.8). This makes one of the two simulations unnecessary. Applied to the preceding, the formula gives, for any measurable function ψ,

$$\mathrm{E}\Big(\int \psi\, dF|\, \sigma,\theta_1,\ldots,\theta_n,X_1,\ldots,X_n\Big) = \frac{1}{|\alpha| + n}\Big[\int \psi\, d\alpha + \sum_{j=1}^n \psi(\theta_j)\Big]. \tag{6.4}$$

The advantage of this representation is that the infinite-dimensional parameter F has been eliminated. To compute the posterior expectation of $\int \psi\, dF$ it now suffices to take the conditional expectation with respect to $X_1,\ldots,X_n$ across the display. In other words, we average out the right hand side of (6.4) with respect to the posterior distribution of $(\theta_1,\ldots,\theta_n)$.

Example 6.1 (Density estimation)**.** The choice $\psi(\theta) = \psi(x, \theta, \sigma)$ in (6.4), for a given x, gives the integral $\int \psi\, dF = \int \psi(x, \theta, \sigma)\, dF(\theta)$, which is the mixture density $p_{F,\sigma}(x)$. Thus the posterior mean density satisfies

$$\mathrm{E}(p_{F,\sigma}(x)|\, \sigma, X_1, \ldots, X_n) = \frac{1}{|\alpha| + n}\Big[\int \psi(x; \theta, \sigma)\, d\alpha(\theta) + \mathrm{E}\Big(\sum_{j=1}^{n} \psi(x; \theta_j, \sigma)|\, X_1, \ldots, X_n\Big)\Big].$$

The next theorem explains a *Gibbs sampling scheme* to simulate from the posterior distribution of $(\theta_1, \ldots, \theta_n)$, based on a weighted generalized Polya urn scheme. Inclusion of a possible parameter σ and other hyperparameters is tackled in the next section. This Gibbs scheme can be used to generate samples $(\theta_1^{(b)}, \ldots, \theta_n^{(b)})$, for $b = 1, \ldots, B$, that are (approximately, after a burn-in) samples from the posterior distribution of $(\theta_1, \ldots, \theta_n)$ given $(X_1, \ldots, X_n, \sigma)$. The conditional expectation $\mathrm{E}[\sum_{j=1}^{n} \psi(\theta_j)|\, X_1, \ldots, X_n]$ of the second term in (6.4) can then be approximated by the average

$$\frac{1}{B} \sum_{b=1}^{B} \sum_{j=1}^{n} \psi(\theta_j^{(b)}).$$

Thus we obtain a Monte Carlo estimate of the posterior mean of a linear functional $\int \psi\, dF$. We can also generate samples from the posterior distribution, by generating F from the $\mathrm{DP}(\alpha + \sum_{i=1}^{n} \delta_{\theta_i})$-distribution, for instance by using the series representation of the Dirichlet process, after first generating $(\theta_1, \ldots, \theta_n)$ from its posterior distribution given $(X_1, \ldots, X_n)$.

We use the subscript $-i$ to denote every index $j \neq i$, and $\theta_{-i} = (\theta_j\colon j \neq i)$.

Theorem 6.2 (Gibbs sampler)**.** *The conditional posterior distribution of θ_i is given by:*

$$\theta_i|\, \theta_{-i}, \sigma, X_1, \ldots, X_n \sim \sum_{j \neq i} q_{i,j} \delta_{\theta_j} + q_{i,i} G_{b,i}, \tag{6.5}$$

where $(q_{i,1}, \ldots, q_{i,n})$ is the probability vector satisfying

$$q_{i,j} \propto \begin{cases} \psi(X_i; \theta_j, \sigma), & j \neq i, j \geq 1, \\ \int \psi(X_i; \theta, \sigma)\, d\alpha(\theta), & j = i, \end{cases} \tag{6.6}$$

and $G_{b,i}$ is the "baseline posterior measure" given by

$$dG_{b,i}(\theta|\, \sigma, X_i) \propto \psi(X_i; \theta, \sigma)\, d\alpha(\theta). \tag{6.7}$$

Proof Since the parameter σ is fixed throughout, we suppress it from the notation. For measurable sets A and B,

$$\mathrm{E}(\mathbb{1}_A(X_i)\mathbb{1}_B(\theta_i)|\, \theta_{-i}, X_{-i}) = \mathrm{E}\Big(\mathrm{E}(\mathbb{1}_A(X_i)\mathbb{1}_B(\theta_i)|\, F, \theta_{-i}, X_{-i})|\, \theta_{-i}, X_{-i}\Big).$$

Because (θ_i, X_i) is conditionally independent of (θ_{-i}, X_{-i}) given F, the inner conditional expectation is equal to $\mathrm{E}(\mathbb{1}_A(X_i)\mathbb{1}_B(\theta_i)|\, F) = \int\int \mathbb{1}_A(x)\mathbb{1}_B(\theta)\psi(x; \theta)\, d\mu(x)\, dF(\theta)$. In the outer layer of conditioning the variables X_{-i} are superfluous, by the conditional independence of F and X_{-i} given θ_{-i}. Therefore, by Exercise 4.8 the preceding display is equal to

$$\frac{1}{|\alpha| + n - 1} \int\int \mathbb{1}_A(x)\mathbb{1}_B(\theta)\psi(x; \theta)\, d\mu(x)\, d\Big(\alpha + \sum_{j \neq i} \delta_{\theta_j}\Big)(\theta).$$

This determines the joint conditional distribution of (X_i, θ_i) given (θ_{-i}, X_{-i}). By Bayes's rule (applied to this joint law conditionally given (θ_{-i}, X_{-i})) we infer that

$$\Pr(\theta_i \in B | X_i, \theta_{-i}, X_{-i}) = \frac{\int_B \psi(X_i;\theta)\, d(\alpha + \sum_{j\neq i}\delta_{\theta_j})(\theta)}{\int \psi(X_i;\theta)\, d(\alpha + \sum_{j\neq i}\delta_{\theta_j})(\theta)}.$$

This in turn is equivalent to the assertion of the theorem. □

6.1.1 *MCMC method

Theorem 6.2 describes the full conditionals of the parameters θ_i in a Gibbs sampler. The parameter σ is added by separate update steps. In practice one often also equips the base measure with hyper parameters, giving the following augmented model:

$$X_i | \theta_i, \sigma, M, \xi, F \overset{\text{ind}}{\sim} \psi(\cdot;\theta_i,\sigma), \quad \theta_i | F, \sigma, M, \xi \overset{\text{iid}}{\sim} F, \quad F | M, \xi \sim \text{DP}(M, G_\xi),$$

where σ, M and ξ are independently generated hyperparameters. A basic algorithm uses the Gibbs sampling scheme of Theorem 6.2 to generate $\theta_1, \ldots, \theta_n$ given $X_1, \ldots, X_n$ in combination with the Gibbs sampler for the posterior distribution of M given in Section 4.8, and/or additional Gibbs steps. The prior densities of the hyperparameters are denoted by a generic π.

Algorithm Generate samples by sequentially executing steps (i)–(iv) below:

(i) Given the observations and σ, M and ξ, update each θ_i sequentially using (6.5) inside a loop $i = 1, \ldots, n$.
(ii) Update $\sigma \sim p(\sigma | \theta_1, \ldots, \theta_n, X_1, \ldots, X_n) \propto \pi(\sigma) \prod_{i=1}^n \psi(X_i;\theta_i,\sigma)$.
(iii) Update $\xi \sim p(\xi | \theta_1, \ldots, \theta_n) \propto \pi(\xi) p(\theta_1, \ldots, \theta_n | \xi)$, where the marginal distribution of $(\theta_1, \ldots, \theta_n)$ is as in the Polya scheme (4.12).
(iv) Update M and next the auxiliary variable η using (4.16), for K_n the number of distinct values in $\{\theta_1, \ldots, \theta_n\}$.

In practice one chooses conjugate priors in steps (ii)-(iii), so that updating σ and ξ is easy. Also the efficiency of the algorithm can be increased by making use of the fact that the number of distinct values among $\theta_1, \ldots, \theta_n$ can be small.

An efficient implementation in R is given in the package `DPpackage`.

6.2 Consistency

For a given family of probability densities $x \mapsto \psi(x;\theta,\sigma)$, form the mixtures $p_{F,\sigma}$ as in (6.1), and construct a prior distribution on densities by equipping F with the Dirichlet prior and independently σ with another prior Π_σ. Form the posterior distribution based on a random sample of observations $X_1, \ldots, X_n$ from $p_{F,\sigma}$. This can be viewed as a distribution on the pair (F, σ), but in this section we are interested in the induced posterior distribution on the densities $p_{F,\sigma}$, which we consider as a method of reconstructing the density of the observations.

We are interested in the success of this procedure in the (frequentist) situation that the observations are a random sample from a given density p_0, which is not necessarily of mixture

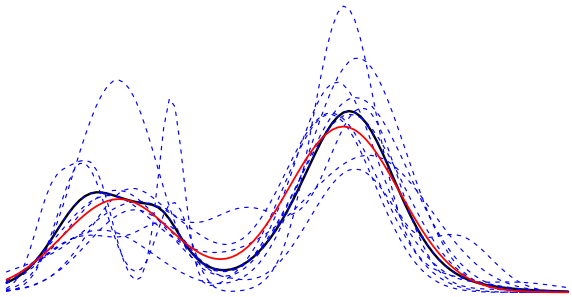

Figure 6.1 Posterior mean (solid black) and ten draws of the posterior distribution of a Dirichlet location-scale mixture of a Gaussian density for data consisting of a sample of size 50 from a mixture of two normals (shown in red). Computations in R using the DPpackage version 1.1-6 (Jara et al. (2015)).

form. We ask whether the posterior distribution of $p_{F,\sigma}$ is consistent at p_0 in the sense of Definition 5.1, where we use the $\mathbb{L}_1$-norm as the metric. We shall answer this question with the help of Theorem 5.15.

Obviously consistency is impossible if the distance of p_0 to the set of all mixtures $p_{F,\sigma}$ is positive: p_0 must be in the closure of the set of all mixtures. Theorem 5.15 imposes the slightly stronger condition that p_0 possesses the Kullback-Leibler property relative to the prior. Whether this is true depends on the combination of p_0 and the kernels $\psi(\cdot;\theta,\sigma)$. In the following section we consider in detail the case of the normal location-scale kernel, and shall see that almost every continuous density p_0 then qualifies. In general, one might decompose the Kullback-Leibler divergence as, for given $(F_\epsilon,\sigma_\epsilon)$,

$$K(p_0;p_{F,\sigma}) = K(p_0;p_{F_\epsilon,\sigma_\epsilon}) + P_0 \log \frac{p_{F_\epsilon,\sigma_\epsilon}}{p_{F,\sigma}}.$$

If for every $\epsilon > 0$ there exists $(F_\epsilon,\sigma_\epsilon)$ such that $K(p_0;p_{F_\epsilon,\sigma_\epsilon}) < \epsilon$, and the set of parameters (F,σ) such that the second term is also smaller than ϵ has positive prior probability, then p_0 possesses the Kullback-Leibler property. The following lemma gives a sufficient condition for the latter, for the Dirichlet prior on F.

Lemma 6.3 (KL-property)**.** *Suppose that for every $\epsilon > 0$ there exists $(F_\epsilon,\sigma_\epsilon)$ such that $K(p_0;p_{F_\epsilon,\sigma_\epsilon}) < \epsilon$, and such that the map $(F,\sigma) \mapsto P_0 \log(p_{F_\epsilon,\sigma_\epsilon}/p_{F,\sigma})$ is continuous at $(F_\epsilon,\sigma_\epsilon)$ relative to the product of the weak topology on F and a given topology on σ. If $F \sim \mathrm{DP}(\alpha)$ for a base measure α whose support contains the support of every F_ϵ and $\sigma \sim \Pi_\sigma$ for a prior Π_σ whose support contains every σ_ϵ, then p_0 is in the Kullback-Leibler support of the prior on $p_{F,\sigma}$ induced by independent priors $F \sim \mathrm{DP}(\alpha)$ and $\sigma \sim \Pi_\sigma$.*

Proof By Lemma 4.4 and Theorem 4.8 the Dirichlet prior gives positive mass to every open neighborhood A of F_ϵ for the weak topology. By assumption $\Pi_\sigma(B) > 0$ for every open neighborhood B of σ_ϵ. By independence every product open neighborhood $A \times B$ of $(F_\epsilon,\sigma_\epsilon)$ possesses positive prior mass, and hence so does *every* open neighborhood for

the product topology. By the assumed continuity the set $\{(F,\sigma)\colon P_0 \log(p_{F_\epsilon,\sigma_\epsilon}/p_{F,\sigma}) < \epsilon\}$ contains an open neighborhood of $(F_\epsilon,\sigma_\epsilon)$, and hence has positive prior mass. We finish with the decomposition argument given preceding the statement of the lemma. □

The second condition for posterior consistency in Theorem 5.15 concerns the entropy of the support of the prior. The following theorem shows that this is often satisfied, and gives sufficient conditions for consistency given the Kullback-Leibler property.

The theorem assumes that the parameters θ and σ of the kernel range over subsets Θ and $\mathcal{S}$ of Euclidean spaces and that the kernel depends smoothly on these parameters.

Theorem 6.4 (Consistency Dirichlet mixtures)**.** *Suppose that for any given $\epsilon > 0$ and n, there exist convex subsets $\Theta_n \subset \Theta \subset \mathbb{R}^k$ and $\mathcal{S}_n \subset \mathcal{S} \subset \mathbb{R}^l$ and constants $a_n, A_n, b_n, B_n > 1$ such that*

(i) $\big\|\psi(\cdot;\theta,\sigma) - \psi(\cdot;\theta',\sigma')\big\|_1 \le a_n\|\theta-\theta'\| + b_n\|\sigma-\sigma'\|$, *for all $\theta,\theta' \in \Theta_n$ and $\sigma,\sigma' \in \mathcal{S}_n$,*
(ii) $\operatorname{diam}(\Theta_n) \le A_n$ *and* $\operatorname{diam}(\mathcal{S}_n) \le B_n$,
(iii) $\log(a_nA_n) \le C\log n$, *for some $C > 0$, and* $\log(b_nB_n) \le n\epsilon^2/(8l)$,
(iv) $\max(\alpha(\Theta_n^c), \Pi_\sigma(\mathcal{S}_n^c)) \le e^{-Cn}$, *for some $C > 0$.*

Then the posterior distribution $\Pi_n(\cdot|\,X_1,\ldots,X_n)$ for $p_{F,\sigma}$ in the model $X_1,\ldots,X_n|\,(F,\sigma) \overset{iid}{\sim} p_{F,\sigma}$, for $(F,\sigma) \sim \mathrm{DP}(\alpha)\times\Pi_\sigma$, is strongly consistent relative to the $\mathbb{L}_1$-norm at every p_0 in the Kullback-Leibler support of the prior of $p_{F,\sigma}$.

Proof We apply Theorem 5.15 with, for given $\epsilon > 0$ the set $\mathcal{P}_n$ defined by, for $N_n \sim \eta n/\log n$ and small $\eta > 0$ to be determined at the end of the proof,

$$\mathcal{P}_n = \Big\{\sum_{j=1}^{\infty} w_j\psi(\cdot;\theta_j,\sigma)\colon (w_j)\in\mathbb{S}_\infty, \sum_{j>N_n} w_j < \frac{\epsilon}{8},\ \theta_1,\ldots,\theta_{N_n}\in\Theta_n, \sigma\in\mathcal{S}_n\Big\}.$$

By the series representation $F = \sum_j W_j\delta_{\theta_j}$ of the Dirichlet process, given in Theorem 4.8, the prior density $p_{F,\sigma}$ is contained in $\mathcal{P}_n$, unless the weights of the representation satisfy $\sum_{j>N_n} W_j \ge \epsilon/8$, or at least one of $\theta_1,\ldots,\theta_{N_n} \overset{\text{iid}}{\sim} \bar\alpha$ falls outside Θ_n, or $\sigma\notin\mathcal{S}_n$. It follows that

$$\Pi(p_{F,\sigma}\notin\mathcal{P}_n) \le \Pr\Big(\sum_{j>N_n} W_j \ge \frac{\epsilon}{8}\Big) + N_n\bar\alpha(\Theta_n^c) + \Pi_\sigma(\mathcal{S}_n^c).$$

The last two terms are exponentially small by assumption (iv) and the choice of N_n. The stick-breaking weights in the first term satisfy $W_j = V_j\prod_{l=1}^{j-1}(1-V_l)$, for $V_l \overset{\text{iid}}{\sim} \mathrm{Be}(1,|\alpha|)$ and $\sum_{j>N} W_j = \prod_{j=1}^{N}(1-V_j)$. It can be checked that $-\log(1-V_j)$ is exponentially distributed with parameter $|\alpha|$, so that $R_n := -\log\sum_{j>N_n} W_j$ possesses a gamma distribution with parameters N_n and $|\alpha|$. Therefore the first term is bounded above by

$$\Pr\Big(R_n < \log\frac{8}{\epsilon}\Big) = \int_0^{\log(8/\epsilon)} \frac{1}{(N_n-1)!} r^{N_n-1}|\alpha|^{N_n}e^{-|\alpha|r}\,dr \le \frac{(\log(8/\epsilon)|\alpha|)^{N_n}}{N_n!} \le \Big(\frac{e\log(8/\epsilon)|\alpha|}{N_n}\Big)^{N_n}.$$

The last inequality follows, since $x^N/N! \le e^x$, for every $x > 0$, including $x = N$. For large n the term within brackets is (easily, because $\log n \ll \sqrt{n}$) bounded above by $1/\sqrt{n}$ and hence the right side is bounded by $e^{-(N_n/2)\log n}$, which is also exponentially small, by the choice of N_n, for every $\eta > 0$.

Given any infinite probability vector $w = (w_1, w_2, \ldots)$, the probability vector $\tilde w$ obtained

by setting the coordinates with $j > N_n$ to zero and renormalizing the remaining coordinates to sum to 1 (hence $\tilde{w}_j = 0$ for $j > N_n$ and $\tilde{w}_j = w_j / \sum_{j \le N_n} w_j$, otherwise) satisfies $\|w - \tilde{w}\|_1 = 2 \sum_{j > N_n} w_j$. By the triangle inequality and the fact that $\|\psi(\cdot; \theta, \sigma)\|_1 = 1$, it follows that

$$\Big\| \sum_{j=1}^{\infty} w_j \psi(\cdot; \theta_j, \sigma) - \sum_{j \le N_n} \tilde{w}_j \psi(\cdot; \theta_j, \sigma) \Big\|_1 \le \|w - \tilde{w}\|_1.$$

Therefore, the finite mixtures $\sum_{j=1}^{N_n} w_j \psi(\cdot; \theta_j, \sigma)$, with $(w_1, \ldots, w_{N_n})$ an arbitrary probability vector, form an $\epsilon/4$-net over $\mathcal{P}_n$ for the $\mathbb{L}_1$-norm. Consequently, a $3\epsilon/4$-net over these finite mixtures is an ϵ-net over $\mathcal{P}_n$. To construct such a net we restrict $(w_1, \ldots, w_{N_n})$ to an $\epsilon/4$-net over all probability vectors of length N_n, restrict $\theta_1, \ldots, \theta_{N_n}$ to an $\epsilon/(4a_n)$-net over Θ_n, and restrict σ to an $\epsilon/(4b_n)$-net over $\mathcal{S}_n$. By the triangle inequality

$$\Big\| \sum_{j \le N_n} w_j \psi(\cdot; \theta_j, \sigma) - \sum_{j \le N_n} \tilde{w}'_j \psi(\cdot; \theta'_j, \sigma') \Big\|_1 \le \|w - w'\|_1 + \max_{j \le N_n} \Big\| \psi(\cdot; \theta_j, \sigma) - \psi(\cdot; \theta'_j, \sigma') \Big\|_1.$$

In view of (i) it follows that for every w, θ, σ there exist w', θ', σ' in the restricted set, so that the right side is smaller than $3\epsilon/4$. By Lemma 5.23 and (ii) the cardinality of this $3\epsilon/4$-net is bounded above by

$$\Big(\frac{24}{\epsilon}\Big)^{N_n} \times \Big(\frac{12 A_n a_n}{\epsilon}\Big)^{k N_n} \times \Big(\frac{12 B_n b_n}{\epsilon}\Big)^{l}.$$

Together with (iii) and the definition $N_n \sim n\eta / \log n$, this shows that $\log N(\epsilon, \mathcal{P}_n, \|\cdot\|_1) \le n\epsilon^2/4$, for sufficiently large n, provided η is chosen small enough. □

Example 6.5 (Location-scale family)**.** The density $x \mapsto \psi(x; \theta, \sigma) = \psi((x - \theta)/\sigma)/\sigma$ of the location scale family of a smooth density ψ will typically satisfy, for some $D > 0$,

$$\Big\| \psi(\cdot; \theta, \sigma) - \psi(\cdot; \theta', \sigma') \Big\|_1 \le \frac{D}{\sigma \wedge \sigma'} [|\theta - \theta'| + |\sigma - \sigma'|].$$

For instance, this is true for the normal density. In this case we may set $\Theta_n = [-n^a, n^a] \subset \Theta = \mathbb{R}$ and $\mathcal{S}_n = [n^{-b}, e^{cn}] \subset \mathcal{S} = (0, \infty)$, and apply the preceding theorem with constants $a_n = D n^b = b_n$, $A_n = 2n^a$, and $B_n = e^{cn}$. This gives $\log(a_n A_n) \sim (a + b) \log n$ and $\log(b_n B_n) \sim cn$, so that (iii) of the theorem is satisfied provided c is chosen small enough (depending on ϵ). Condition (iv) of the theorem becomes

$$\Pi_\sigma\Big(\frac{1}{\sigma} < e^{-cn} \text{ or } \frac{1}{\sigma} > n^b\Big) \le e^{-Cn}, \qquad \alpha(z : |z| > n^a) < e^{-Cn}, \qquad \text{for some } C > 0.$$

This is true for instance if $(1/\sigma)^{1/b} \sim \Gamma(r, s)$, for some $r > 1$ and $s > 0$, and α is a multiple of the normal distribution and $a = 1/2$.

6.3 Rate of contraction

In this section we specialize to Dirichlet mixtures of a Gaussian location family. For ϕ the standard normal density, set

$$p_{F,\sigma}(x) = \int \frac{1}{\sigma} \phi\Big(\frac{x - \theta}{\sigma}\Big) \, dF(\theta).$$

Such Gaussian mixtures can approximate any density p_0. In fact, the mixture $p_{F,\sigma}$ is the convolution of F and a Gaussian density with variance σ^2 and hence $p_{F,\sigma}$ is the density of the sum $\Theta + \sigma\epsilon$, for independent variables $\Theta \sim F$ and $\epsilon \sim N(0,1)$. It is immediate that $\Theta + \sigma\epsilon \rightsquigarrow \Theta$ as $\sigma \to 0$, and it is not surprising that this can be extended to convergence of $p_{F,\sigma}$ to the density of F, if this exists and is sufficiently regular.

This suggests that Dirichlet mixtures of the normal family may provide consistent posterior approximation for a density of general form. It is remarkable that in combination with an inverse gamma prior on the scale σ of the normal density, Dirichlet mixtures even give a near optimal rate of reconstruction of any smooth density.

Here "optimality" may be understood in the sense of *minimax rate* of estimation. There exist density estimators $\hat{p}_n(\cdot) = \hat{p}_n(\cdot; X_1, \ldots, X_n)$ such that, for a random sample $X_1, \ldots, X_n$ from a density p_0 on a compact interval of $\mathbb{R}$ that is β times continuously differentiable,

$$P_0^n(d(\hat{p}_n, p_0) \geq Cn^{-\beta/(2\beta+1)}) \to 0.$$

Thus the rate of estimation for the distance d is $n^{-\beta/(2\beta+1)}$. The distance can be taken an L_r-distance, given by $d^r(p,q) = \int |p-q|^r(x)\,dx$, or also the Hellinger distance, and the preceding display can be valid uniformly in sets of densities p_0 whose derivatives are uniformly bounded (and bounded away from zero). It can be shown that no estimators can attain a faster rate uniformly in these sets of densities, making $n^{-\beta/(2\beta+1)}$ the *minimax* rate.

The same rate applies to smoothness levels β that are not integer values. For general $\beta > 0$ one says that p_0 is *smooth of order* β if p_0 is $\underline{\beta}$ times continuously differentiable, for $\underline{\beta}$ the largest integer strictly smaller than β, with bth derivative satisfying, for some $c > 0$,

$$|p_0^{(b)}(x) - p_0^{(b)}(y)| \leq c|x-y|^{\beta-\underline{\beta}}, \qquad \text{for every } x, y.$$

Furthermore, the density need not be limited to a compact interval, but it should have small tails, as otherwise the rate of estimation will be determined by the difficulty of estimating the tails, rather than estimating the bulk of the density.

The minimax rate of estimation is faster if β is bigger, and approaches $n^{-1/2}$ if $\beta \to \infty$, which shows that estimation can be more accurate if p_0 is known to be smoother. The rate is attained for instance by kernel estimators, or truncated series estimators, such as those based on a wavelet expansion for the underlying density. These estimators can also be tuned in such a way that a single estimator $\hat{p}_n$ attains the rate $n^{-\beta/(2\beta+1)}$ whenever p_0 is β-smooth, whatever the value of $\beta > 0$. One says that such estimators *adapt* to the underlying smoothness.

The following theorem shows that the posterior distribution arising from Dirichlet mixtures of normal densities also adapts to the smoothness level, and attains the optimal rate up to a logarithmic factor. It assumes that $p_0 \colon \mathbb{R} \to [0, \infty)$ is smooth of order β in the sense that it is $\underline{\beta}$ times continuously differentiable with derivatives satisfying,

$$P_0\left(\frac{|p_0^{(k)}|}{p_0}\right)^{2\beta/k} < \infty, \qquad k = 1, \ldots, \underline{\beta}, \tag{6.8}$$

$$\left|p_0^{(\underline{\beta})}(x+h) - p_0^{(\underline{\beta})}(x)\right| \leq |h|^{\beta-\underline{\beta}} e^{b_0 h^2} L(x), \tag{6.9}$$

for some $b_0 > 0$ and a function $L \colon \mathbb{R} \to \mathbb{R}$ such that $P_0(L/p_0)^2 < \infty$.

Theorem 6.6 (Contraction Dirichlet mixture)**.** *Suppose that p_0 is β-smooth in the sense of*

(6.9) *and satisfies* $p_0(x) \leq ae^{-b_0|x|^{d_0}}$ *for some constants* $a_0, b_0, d_0 > 0$. *Then the posterior distribution of* $p_{F,\sigma}$, *for the prior induced by choosing* $F \sim \mathrm{DP}(\alpha)$ *independently of* $1/\sigma \sim \Gamma(r, s)$, *for* $r, s > 0$ *and a base measure* α *with a positive continuous density satisfying* $\alpha(z\colon |z| > M) \leq a_1 e^{-b_1 M^{d_1}}$ *for some constants* $a_1, b_1, d_1 > 0$, *has contraction rate relative to the Hellinger distance at least* $n^{-\beta/(2\beta+1)}(\log n)^{(2\beta+\beta/d_0+1)/(2\beta+1)}$.

The theorem can be proved by verifying the conditions of Theorem 5.19. Here the set $\mathcal{P}_n$ is defined similarly to the set used in Theorem 6.4, and its entropy is computed in a similar manner. Computing a lower bound on the prior mass, as in (5.7), is much more involved. This is achieved by first approximating the density p_0 by a density of the form p_{F_n,σ_n}, for a finitely-discrete mixture distribution F_n and a suitable "bandwidth" σ_n, and next analyzing probabilities concerning F_n under its prior using the characterization of the finite-dimensional distributions of the Dirichlet process, as given in Definition 4.6.

The proof takes many steps, and occupies the next section.

6.3.1 *Proof of Theorem 6.6

Let ϵ_n be the contraction rate as claimed, and define $N_n = Cn\epsilon_n^2/\log n$ and $\tau_n = (Cn\epsilon_n^2)^{-1}$, for a constant C to be chosen. We verify the conditions of Theorem 5.19 with

$$\mathcal{P}_n = \Big\{p_{F,\sigma}\colon F = \sum_{j=1}^{\infty} w_j\delta_{z_j}, \sum_{j>N_n} w_j < \epsilon_n^2, \max_{1\leq j\leq N_n}|z_j| \leq (Cn\epsilon_n^2)^{1/d_1}, 1 \leq \frac{\sigma}{\tau_n} \leq (1+\epsilon_n^2)^{Cn}\Big\}.$$

This set has the same form as in the proof of Theorem 6.4. By the arguments given there, based on the series representation of the Dirichlet process given in Theorem 4.8,

$$\begin{aligned}\Pi(\mathcal{P}_n^c) &\leq \Pi\Big(\sum_{j>N_n} W_j \geq \epsilon_n^2\Big) + \Pi\Big(\max_{1\leq j\leq N_n}|\theta_j| > (Cn\epsilon_n^2)^{1/d_1}\Big) + \Pi\Big(\frac{\sigma}{\tau_n} < 1\Big) + \Pi\Big(\frac{\sigma}{\tau_n} > (1+\epsilon_n^2)^{Cn}\Big)\\ &\lesssim \Big(\frac{e|\alpha|(\log\epsilon_n^{-2}}{N_n}\Big)^{N_n} + N_n e^{-b_1 Cn\epsilon_n^2} + \frac{e^{-s/\tau_n}}{\tau_n^r} + \frac{(1+\epsilon_n^2)^{-rCn}}{\tau_n^r}.\end{aligned}$$

All terms can be seen to be bounded be $e^{-b_2 Cn\epsilon_n^2}$, for a constant b_2, so that (5.8) is satisfied, for an arbitrary c_2 if C is chosen large enough.

To bound the entropy of $\mathcal{P}_n$ we consider the set $\mathcal{P}_n'$ of all elements of $\mathcal{P}_n$ such that also:

- $w_j = 0$, for $j > N_n$,
- $(w_1, \ldots, w_{N_n})$ belongs to a maximal ϵ_n^2-net over the set $\mathbb{S}_{N_n}$ of all probability vectors of length N_n,
- every z_j belongs to the grid $0, \pm\tau_n\epsilon_n^2, \pm 2\tau_n\epsilon_n^2, \ldots$ intersecting $\{z\colon |z| \leq (Cn\epsilon_n^2)^{1/d_1}\}$.
- σ belongs to the grid $\tau_n, \tau_n(1+\epsilon_n^2), \tau_n(1+\epsilon_n^2)^2, \ldots, \tau_n(1+\epsilon_n^2)^{Cn-1}$.

By Lemma 5.23 the cardinality of this set satisfies

$$\log \#\mathcal{P}_n' \leq \log\Big[\Big(\frac{3}{\epsilon_n^2}\Big)^{N_n}\Big(\frac{3(Cn\epsilon_n^2)^{1/d_1}}{\tau_n\epsilon_n^2}\Big)^{N_n} Cn\Big] \leq C_2 n\epsilon_n^2,$$

for a constant C_2 that depends on C and d_1 only. If we can show that $\mathcal{P}_n'$ is an $C_3\epsilon_n$-net over $\mathcal{P}_n$ for the Hellinger distance, for some C_3, then It follows that (5.6) is satisfied for a multiple of ϵ_n. Now a given $p_{F,\sigma}$ in $\mathcal{P}_n$ can be identified with the triple (w, z, σ) of weights w

and locations z of F and scale σ, and can be linked with the density $p_{F',\sigma'} \in \mathcal{P}'_n$ given by the triple (w', z', σ') defined by:

- w' equal to a closest point in the net over $\mathbb{S}_{N_n}$ to the renormalized restriction of w to $\mathbb{S}_{N_n}$; so that $\|w - w'\|_1 \le 3\epsilon_n^2$.
- z'_j equal to the projection of z_j on the grid; so that $\|z - z'\|_\infty < \tau_n \epsilon_n^2$.
- $\sigma' = \tau_n(1+\epsilon_n^2)^k$ if $\tau_n(1+\epsilon_n^2)^k \le \sigma \le \tau_n(1+\epsilon_n^2)^{k+1}$; so that $0 \le \sigma/\sigma' - 1 = (\sigma - \sigma')/\sigma' \le \epsilon_n^2$.

By Lemmas 6.10 and 6.9, we have $h(p_{F,\sigma}, p_{F,\sigma'}) \le \epsilon_n^2$. By the triangle inequality and again the second lemma $\|p_{F,\sigma'} - p_{F',\sigma'}\|_1 \le \|w - w'\|_1 + 2\|z - z'\|_\infty/\sigma' \le 3\epsilon_n^2 + 2\epsilon_n^2$, since $\sigma' \ge \tau_n$. By Lemma 2.28 and another application of the triangle inequality, we see that $h(p_{F,\sigma}, p_{F',\sigma'}) \le 4\epsilon_n$.

It remains to verify (5.7). Fix small $\epsilon, \tau > 0$ and set $A_\tau = A(\log \tau^{-1})^{1/d_0}$, for a sufficiently large constant A. By Lemma 6.7 there exists a discrete distribution $F_{\epsilon,\tau}$ with no more than $N \asymp \tau^{-1} A_\tau (\log \epsilon^{-1})$ support points contained in the interval $[-A_\tau, A_\tau]$ such that $h(p_0, p_{F,\tau}) \lesssim \epsilon + \tau^\beta$. The support points can be chosen $\epsilon\tau$-separated without loss of generality, so that there exist intervals $U_1, \ldots, U_N$ of lengths $\epsilon^2\tau$, each containing one support point. We combine this with Lemmas 6.11 and 6.10 to see that

$$\Big\{(F,\sigma)\colon \sum_{j=1}^N |F(U_j) - F_{\epsilon,\tau}(U_j)| < \epsilon^2, 1 \le \frac{\sigma}{\tau} \le 1 + \epsilon\Big\} \subset \Big\{(F,\sigma)\colon h(p_{F,\sigma}, p_{F,\tau}) \lesssim \epsilon + \tau^\beta\Big\}.$$

We can extend $U_1, \ldots, U_N$ to a partition of $[-A_\tau, A_\tau]$ in intervals of length at most τ and next into a partition $U_1, \ldots, U_K$ of $\mathbb{R}$ such that the total number K of sets is bounded above by a multiple of N and such that $\tau\epsilon^2 \le \alpha(U_j) \le 1$ for every j. Set $w_j = 0$ for $N < j \le K$. By Lemma 6.12

$$\Pi\Big(\sum_{i=1}^K |F(U_j) - w_j| \le 2\epsilon^2, \min_{1\le j\le K} F(U_j) > \frac{\epsilon^4}{2}\Big) \ge Ce^{-cK\log\epsilon^{-1}}.$$

Combining the preceding with the inverse gamma distribution of $1/\sigma$ we see that

$$\Pi\Big((F,\sigma)\colon h(p_{F,\sigma}, p_{F,\tau}) \lesssim \epsilon + \tau^\beta, \min_{1\le j\le K} F(U_j) > \frac{\epsilon^4}{2}, \tau \le \sigma \le 2\tau\Big) \ge Ce^{-c_3 N\log\epsilon^{-1}} e^{-s/\tau}\tau^{-r}\epsilon.$$

Next we relate the Hellinger distance to the Kulback-Leibler divergence and variation, with the help of Lemma 6.13. Every $x \in [-A_\tau, A_\tau]$ is contained in a set U_j of diameter at most τ, say $U_j(x)$. Hence

$$p_{F,\sigma}(x) \ge \begin{cases} \int_{U_i(x)} \frac{e^{-(x-z)^2/(2\sigma^2)}}{\sigma\sqrt{2\pi}}\, dF(z) \gtrsim \frac{e^{-\tau^2/(2\sigma^2}}{\sigma\sqrt{2\pi}} F(U_i(x)), & \text{if } |x| \le A_\tau, \\ \int_{-A}^{A} \frac{e^{-(x^2+z^2)^2/\sigma^2}}{\sigma\sqrt{2\pi}}\, dF(z) \ge \frac{e^{-2x^2/\sigma^2}}{\sigma\sqrt{2\pi}} F[-A_\tau, A_\tau], & \text{if } |x| > A_\tau. \end{cases}$$

For F and σ contained in the event in the second last display, it follows that $\log(p_0/p_{F,\sigma})(x) \lesssim \log(\tau/\epsilon^4)\mathbb{1}\{|x| \le A_\tau\} + x^2/\tau^2\mathbb{1}\{|x| > A_\tau\}$, and hence, for sufficiently small $C_4 > 0$,

$$P_0\Big(\log\frac{p_0}{p_{F,\sigma}}\Big)^2 \mathbb{1}\Big\{\frac{p_0}{p_{F,\sigma}} < \frac{c_4\tau}{\epsilon^4}\Big\} \le 2\int_{A_\tau}^\infty \frac{x^4}{\tau^4}\, p_0(x)\, dx \le \tau^k,$$

by the assumed tail condition on p_0, for any k, if the constant A in A_τ is chosen large

enough. Lemma 6.13 now gives that both $K_2(p_0; p_{F,\sigma})$ is bounded above by a multiple of $h^2(p_0; p_{F,\sigma})(\log(\tau/\epsilon^4))^2 + \tau^k \lesssim (\epsilon + \tau^\beta)^2(\log(\tau/\epsilon^4))^2$, provided $\epsilon^4/(c_4\tau) < 0.4$.

Taking everything together we see that that the left side of (5.7) is bounded below by $e^{-n\epsilon_n^2}$ for ϵ_n such that there exist ϵ and τ such that, with $N \asymp \tau^{-1}A_\tau(\log 1/\tau)^{1/d_0}$,

$$c_5(\epsilon + \tau^\beta)(\log(\tau/\epsilon^4)) \le \epsilon_n, \qquad \epsilon^4 \le 0.4c_4\tau, \qquad e^{-c_3 N \log \epsilon^{-1}} e^{-s/\tau}\tau^{-r}\epsilon \ge e^{-n\epsilon_n^2}.$$

We can choose $\epsilon \ll \tau$ to satisfy the second requirement. In the third requirement the terms $e^{-s/\tau}$ and τ^{-r} are much bigger than the first exponential, and in view of the bound on N the inequality can be reduced to $n\epsilon_n^2 \ge c_6\tau^{-1}(\log n)^{2+1/d_0}$, if τ and ϵ are chosen so that both $\log\tau^{-1}$ and $\log\epsilon^{-1}$ are equivalent to a constant times $\log n$. We finish by choosing $\tau^{2\beta+1} = n^{-1}(\log n)^{1/d_0}$ and next ϵ a big multiple of $\tau^\beta(\log n)$. For τ as given, the latter ϵ has the same order as ϵ_n in the statement of the theorem.

Supporting lemmas

Lemma 6.7. *For every sufficiently small $\epsilon, \sigma > 0$ there exists a discrete probability measure F with no more that $A(\sigma^{-1}(\log\sigma^{-1})^{1/d_0}(\log\epsilon^{-1})$ support points contained in the interval $[-A(\log\sigma^{-1})^{1/d_0}, A(\log\sigma^{-1})^{1/d_0}]$ such that $h(p_0, p_{F,\sigma}) \le \epsilon + D\sigma^\beta$. Here A and D are constants that depend on p_0 only.*

Proof Let $k = \underline{\beta}$. The first step is to show that there exist constants $c_0 = 1, c_1 = 0, c_2, \ldots, c_k$ and C_0 such that

$$\left| p_0(x) - \phi_\sigma * \sum_{l=0}^{k} \sigma^l c_l p_0^{(l)}(x) \right| \le C_0\sigma^\beta L(x). \tag{6.10}$$

Since p_0 satisfies (6.9), a Taylor expansion gives

$$p_0(x) - p_0(x-y) = \sum_{l=1}^{k} \frac{(-y)^l}{l!} p_0^{(l)}(x) + R(x,y), \qquad |R(x,y)| \le |y|^\beta L(x)e^{cy^2}.$$

Multiplying this equation by $\phi_\sigma(y)$ and integrating with respect to y, we find, with m_l the lth moment of the standard normal distribution,

$$p_0(x) - \psi_\sigma * p_0(x) = \sum_{l=1}^{k} \frac{(-\sigma)^l m_l}{l!} p_0^{(l)}(x) + \bar{R}(x), \qquad |R(x)| \le C_1 L(x).$$

For $k = 0$ the sum on the right is absent, and for $k = 1$ it also vanishes, as $m_1 = 0$, and the proof is complete. We proceed to the case $k > 1$ by induction on k. Suppose that (6.10) is valid with α instead of β for every function p_0 that satisfies (6.9) with α instead of β (but the same L and c), for every $\alpha < \beta$ with the same fractional part as β. This class of functions includes the functions $p_0^{(l)}$ with $2 \le l \le k$, with $\alpha = \beta - l$. By the induction hypothesis these function possess an approximation of the form $\phi_\sigma * \sum_{j=0}^{k-l} \sigma^j c_{l,j} p_0^{(l+j)}$. We substitute these approximations in the preceding display and rearrange the resulting double sum by powers of σ to arrive at the desired approximation (6.10).

Since $\int p_0(x)\,dx = 1$ and $\int p_0^{(l)}(x)\,dx = 0$, for $l > 0$, the function $p_\sigma = \sum_{l=0}^{k} \sigma^l c_l p_0^{(l)}$ integrates to one. However, it may assume negative values. To remedy this we consider its absolute value $|p_\sigma|$. On the set where $p_\sigma < 0$, where this differs from p_σ, we have that

$|p_\sigma| - p_\sigma = -2p_\sigma \le 2(p_0 - p_\sigma)$. On the set $E_\sigma = \cap_{2\le l\le k}\{\sigma^l|p^{(l)}| < \eta p_0\}$, we have that $|p_\sigma - p_0| \lesssim \eta p_0$, and hence $p_\sigma \ge 0$, for sufficiently small η. It follows that

$$\int (|p_\sigma| - p_\sigma)\,d\lambda \le 2\int_{E_\sigma^c}(p_0 - p_\sigma)\,d\lambda \le \sum_{2\le l\le k}|c_l|\sigma^l \int_{E_\sigma^c}|p_0^{(l)}|\,d\lambda$$
$$\le \sum_{2\le l\le k}|c_l|\sigma^l P_0\Big(\frac{|p_0^{(l)}|}{p_0}\Big)^{2\beta/l} d\lambda\Big(\frac{\sigma^l}{\eta}\Big)^{2\beta/l-1} \lesssim \sigma^{2\beta}.$$

Next

$$h^2(p_0, \phi_\sigma * |p_\sigma|) \le \int \frac{(p_0 - \phi_\sigma * |p_\sigma|)^2}{p_0 + \phi_\sigma * |p_\sigma|}\,d\lambda$$
$$\le 2\int \frac{(p_0 - \phi_\sigma * p_\sigma)^2}{p_0}\,d\lambda + 2\int \frac{(\phi_\sigma * (|p_\sigma| - p_\sigma))^2}{\phi_\sigma * |p_\sigma|}\,d\lambda.$$

The first term on the right is bounded above by $2C_0P_0(L/p_0)^2\sigma^{2\beta}$, by the construction of p_σ. In the second term we can bound $\phi_\sigma * (|p_\sigma| - p_\sigma)$ by $\phi_\sigma * |p_\sigma|$ and hence the whole term by $\int \phi_\sigma * (|p_\sigma| - p_\sigma)\,d\lambda = \int |p_\sigma|\,d\lambda - 1$, which is of the order $\sigma^{2\beta}$.

Next we restrict and renormalize p_σ to the given interval, and consider

$$f_\sigma(x) = \frac{|p_\sigma|(x)\mathbb{1}\{|x| \le A_\sigma\}}{\int |p_\sigma|(x)\mathbb{1}\{|x| \le A_\sigma\}\,dx}, \qquad A_\sigma = A(\log \sigma^{-1})^{1/d_0}.$$

Since $p_0(x) \lesssim e^{-b_0|x|^{d_0}}$, we have that $P_0(x\colon |x| \ge A_\sigma) \lesssim \sigma^k$ if $b_0A^{d_0} > k$. Thus $P_0(x\colon |x| \ge A_\sigma)$ is smaller than any power of k if A is small enough. Using the fact that $P_0(|p_0^{(l)}/p_0|^r) < \infty$ is assumed finite for some $r > 1$, it can be seen that the integrals of the derivatives $\int_{x:|x|>A_\sigma}|p_0^{(l)}|(x)\,dx$ similarly can be made smaller than any power of σ, by choosing a sufficiently large A (use Hölder's inequality). Then $\int_{x:|x|>A_\sigma}|p_\sigma^{(l)}|(x)\,dx$ is equally small, by its definition and the triangle inequality, and hence

$$h^2(\phi_\sigma * |p_\sigma|, \phi_\sigma * f_\sigma) = \int |p_\sigma|(x)\,dx + 1 - 2\sqrt{\int_{-A_\sigma}^{A_\sigma}|p_\sigma|(x)\,dx}$$

The first term on the right is $1 + O(\sigma^{2\beta})$, and the same is true for the term under the root. It follows that the expression is bounded above by a multiple of $\sigma^{2\beta}$.

The distribution F_σ corresponding to f_σ satisfies the requirements for F, except that it is not discrete. We can replace it by a discrete distribution in view Lemma 6.8, at the cost of increasing the distance with ϵ. □

Lemma 6.8 (Finite approximation)**.** *For every probability measure F on $[-A, A]$, $\sigma \in (0, A)$ and sufficiently small $\epsilon > 0$, there exists a discrete probability measure F^* on $[-A, A]$ with no more than $16eA\sigma^{-1}\log\epsilon^{-1}$ support points such that $h(p_{F,\sigma}, p_{F^*,\sigma}) \le \epsilon$. Without loss of generality the support points can be chosen in the set $0, \pm\sigma\epsilon, \pm 2\sigma\epsilon, \ldots$.*

Proof Because the Hellinger distance is invariant under a change of scale and $p_{F,\sigma}(x) = \det\sigma^{-1}p_{F_\sigma,1}(\sigma^{-1}x)$, for F_σ the distribution of Z/σ if $Z \sim F$, the distance in the lemma is equal to $h(p_{F_\sigma,1}, p_{F_\sigma^*,1})$. The measures F_σ concentrate on the interval $[-A/\sigma, A/\sigma]$. Thus the problem can be reduced to mixtures of the standard normal kernel relative to mixing

distributions on the latter interval. This interval can be partitioned into fewer than $2a/\sigma$ intervals of length at most 1. If $F = \sum_i F(I_i)F_i$, where each F_i is a probability measure on a partitioning interval I_i, then $p_{F,1} = \sum_i F(I_i)p_{F_i,1}$. If F_i^* is a discrete distribution on I_i with at most $D\log(1/\epsilon)$ support points such that $h(p_{F_i^*,1}, p_{F,1}) \le \epsilon$, then $F^* = \sum_i F(I_i)F_i^*$ will be the appropriate approximation to F, by convexity of the square of the Hellinger distance. Because the distance is invariant under shifting, we can shift the intervals I_i to the origin, and hence it is no loss of generality to construct the approximation only for $I_i = [0, 1]$. Thus for the remainder of the proof assume that $\sigma = 1$ and that F is concentrated on the unit interval.

A Taylor expansion of the exponential function gives

$$\phi(x-z) = \frac{1}{\sqrt{2\pi}}\Big[\sum_{l=0}^{k-1} \frac{[-(x-z)^2/2]^l}{l!} + R(x-z)\Big], \qquad \big|R(x)\big| \le \frac{(x^2/2)^k}{k!}.$$

Let F^* be a probability measure on $[0, 1]$ such that $\int z^l\, dF^*(z) = \int z^l\, dF(z)$, for $l = 1, \ldots, 2k-2$. Then integrating the last display with respect to F^*-F, which gives $(p_{F^*,1}-p_{F,1})(x)$, leaves only the integral over R. For $|x| \le T$ and $|z| \le 1$ and $T \ge 2$, we have $|x - T| \le 2T$ and hence $|R(x-z)| \le (2eT^2/k)^k$, in view of the inequality $k! \ge k^k e^{-k}$. If $|x| > T \ge 2$ and $|z| \le 1$, then $|x - z| \ge x/2$ and hence $p_{F,1}(x) \le e^{-x^2/8}$, for any probability measure F on $[0, 1]$. Since $\int_{|x|\ge T} p_{F,1}(x)\, dx \le e^{-T^2/8}$, we have

$$\|p_{F^*,1} - p_{F,1}\|_1 \le 2T\Big(\frac{2eT^2}{k}\Big)^k + e^{-T^2/8}.$$

We choose $T = 4(\log \epsilon^{-1})^{1/2}$ to reduce the second term to ϵ^2. Next we choose $k \ge 4eT^2$ and such that $2T2^{-k} \le \epsilon^2$ to reduce also the first term to ϵ^2. The choice $k \sim 16e\log(1/\epsilon)$ suffices for both. a suitable multiple of $(\log \epsilon^{-1})^{1/2}$ to reduce the right side to ϵ^2. The Hellinger distance is bounded by the root of the L_1-distance and hence is bounded by $\sqrt{2}\epsilon$.

By the geometric form of Jensen's inequality, the vector $(\int z\, dF(z), \ldots, \int z^{2k-1}\, dF(z))$, is contained in the convex hull of the curve $\{(z, z^2, \ldots, z^{2k-1}): 0 \le z \le 1\} \subset \mathbb{R}^{2k-1}$. By Carathéodory's theorem any point of the convex hull can be written as the convex combination of $(2k-1)+1 = 2k$ points from the curve. This convex combination corresponds to a probability distribution F^* on $[01,]$ with at most $2k$ support points that matches the moments of F up to order $2k-1$.

The final assertion of the lemma follows from the fact that moving the support points of F^* to a closest point in the given lattice increases the Hellinger distance by at most a multiple of ϵ, by Lemmas 6.10 and 6.9. □

Lemma 6.9. *Let $\phi_{\mu,\sigma}$ the density of the normal distribution with parameters (μ, σ^2). Then $h(\phi_{\mu,\sigma}, \phi_{\nu,\tau}) \le |\mu-\nu|/(\sigma+\tau) + 2|\sigma-\tau|/(\sigma+\tau)$, for any $\mu, \nu \in \mathbb{R}$ and any $\sigma, \tau > 0$.*

Proof The square Hellinger distance between normal densities with standard deviations σ and τ can be calculated explicitly as

$$h^2(\phi_{\mu,\sigma}, \phi_{\nu,\tau}) = 2 - 2\int \sqrt{\phi_{\mu,\sigma}(x)}\,\sqrt{\phi_{\nu,\tau}(x)}\, dx = 2 - 2\sqrt{1-(\sigma-\tau)^2/(\sigma^2+\tau^2)}\, e^{-|\mu-\nu|^2/(4\sigma^2+4\tau^2)}.$$

Next we use the inequalities $|1 - \sqrt{1-s}| \le s$, for $s \in [0, 1]$, and $1 - e^{-t} \le t$, for $t \ge 0$. □

Lemma 6.10. *For any probability distribution F on $\mathbb{R}$ we have $h(p_{F,\sigma}, p_{F,\tau}) \leq h(\phi_\sigma, \phi_\tau)$.*

Proof The density $p_{F,\sigma}$ is the convolution $\phi_\sigma * F$, and convolution decreases the Hellinger distance $h(p_{F,\sigma}, p_{F,\tau}) \leq h(\phi_\sigma, \phi_\tau)$. Indeed, by Jensen's inequality applied to the convex function $(u, v) \mapsto (\sqrt{u} - \sqrt{v})^2$ and the expectations $u = \mathrm{E}\phi_\sigma(x-Z) = p_{F,\sigma}(x)$ and $v = \mathrm{E}\phi_\tau(x-Z) = p_{F,\tau}(x)$, for $Z \sim F$, we see that, for every x,

$$|\sqrt{p_{F,\sigma}(x)} - \sqrt{p_{F,\tau}(x)}|^2 \leq \mathrm{E}|\sqrt{\phi_\sigma(x-Z)} - \sqrt{\phi_\tau(x-Z)}|^2.$$

We integrate this with respect to x, and apply Fubini's theorem to the right side, to see that the integral is bounded above by $\int |\sqrt{\phi_\sigma(x)} - \sqrt{\phi_\tau(x)}|^2\, dx$. □

Lemma 6.11. *If $U_1, \ldots, U_N$ are arbitrary disjoint intervals of length bounded by $\epsilon^2\sigma$, then $\sum_{j=1}^N |F(U_j) - G(U_j)| < \epsilon^2$ implies that $h(p_{F,\sigma}, p_{G,\sigma}) \lesssim \epsilon$, for any probability measures F on $\mathbb{R}$, and G on $\cup_{j=1}^N U_j$, and $\sigma > 0$.*

Proof For arbitrary points $z_1, \ldots, z_N$ contained in the intervals $U_1, \ldots, U_N$ and a point $z_0 \in U_0 := \mathbb{R} \setminus \cup_{j=1}^N U_j$, let $F' = \sum_{j=1}^N F(U_j)\delta_{z_j} + F(V)\delta_z$. Then we can decompose $p_{F,\sigma}(x) - p_{F',\sigma}(x)$ as $\sum_{j=0}^N \int_{U_j} (\phi_\sigma(x-z) - \phi_\sigma(x-z_j))\, dF(z)$ and hence by the triangle inequality and Fubini's theorem,

$$\left\| p_{F,\sigma} - p_{F',\sigma} \right\|_1 \leq 2F(U_0) + \sum_{j=1}^N \frac{\mathrm{diam}(U_j)}{\sigma} F(U_j) \lesssim 2F(U_0) + \epsilon^2,$$

since the distance between two normal densities with the same scale σ is is bounded by the difference between their locations divided by σ and the set U_j have length bounded by $\epsilon^2\sigma$ by assumption.

Define G' analogously and use the triangle inequality to see that $\|p_{F,\sigma} - p_{G,\sigma}\|_1$ is bounded above by $2F(U_0) + 2G(U_0) + 2\epsilon^2 + \|p_{F',\sigma} - p_{G',\sigma}\|_1$. Since $p_{F',\sigma}(x) - p_{G',\sigma}(x) = \sum_{j=0}^N \phi_\sigma(x - z_j)(F(U_j) - G(U_j))$, the last term is bounded above by $\sum_{j=0}^N \left|F(U_j) - G(U_j)\right|$. It follows that $\|p_{F,\sigma} - p_{G,\sigma}\|_1$ is bounded above by $3F(U_0) + 3G(U_0) + 3\epsilon^2$. Because G is concentrated on $\cup_j U_j$ by assumption, we have that $G(U_0) = 0$, while $F(U_0) = F(U_0) - G(U_0) = \sum_{j=1}^N (G(U_j) - F(U_j)) \leq \epsilon^2$. Conclude that $\|p_{F,\sigma} - p_{G,\sigma}\|_1$ is bounded above by $6\epsilon^2$. The lemma follows as $h^2 \leq \|\cdot\|_1$. □

Lemma 6.12 (Prior mass Dirichlet)**.** *If $(X_1, \ldots, X_N) \sim \mathrm{Dir}(N; \alpha_1, \ldots, \alpha_N)$, where $A\epsilon^b \leq \alpha_j \leq M$, $MN\epsilon \leq 1$, and $\sum_{j=1}^N \alpha_j = m$ for some constants A, ϵ, b, M and $M \geq m$, then there exist positive constants c and C depending only on A, M, m and b such that for any point $(w_1, \ldots, w_N)$ in the N-simplex $\mathbb{S}_N$,*

$$\Pr\Big(\sum_{i=1}^N |X_i - w_i| \leq 2\epsilon, \min_{1 \leq i \leq N} X_i > \frac{\epsilon^2}{2}\Big) \geq Ce^{-cN\log_- \epsilon}.$$

Proof First assume that $M = 1$, so that $\epsilon < N^{-1}$. There is at least one index i with $w_i \geq N^{-1}$; by relabeling, we can assume that $i = N$, and then $\sum_{i=1}^{N-1} w_i = 1 - w_N \leq (N-1)/N$. If $(y_1, \ldots, y_N)$ is contained in the N-simplex and $|y_i - w_i| \leq \epsilon^2$ for $i = 1, \ldots, N-1$, then

$$\sum_{i=1}^{N-1} y_i \leq \sum_{i=1}^{N-1} w_i + (N-1)\epsilon^2 \leq (N-1)(N^{-1} + \epsilon^2) \leq 1 - \epsilon^2 < 1.$$

Furthermore, $\sum_{i=1}^N |y_i - w_i| \le 2\sum_{i=1}^{N-1} |y_i - w_i| \le 2\epsilon^2(N-1) \le 2\epsilon$ and $y_N > \epsilon^2 > \epsilon^2/2$ in view of the preceding display. Therefore the probability on the left side of the lemma is bounded below by

$$\Pr\Big(\max_{1\le i\le N-1} |X_i - w_i| \le \epsilon^2\Big) \ge \frac{\Gamma(m)}{\prod_{i=1}^N \Gamma(\alpha_i)} \prod_{i=1}^{N-1} \int_{\max((w_i-\epsilon^2),0)}^{\min((w_i+\epsilon^2),1)} 1\, dy_i,$$

because $\prod_{i=1}^N y_i^{\alpha_i - 1} \ge 1$ for every $(y_1, \ldots, y_N) \in \mathbb{S}_N$, as $\alpha_i - 1 \le 0$ by assumption. Since each interval of integration contains an interval of at least length ϵ^2, and $\alpha\Gamma(\alpha) = \Gamma(\alpha+1) \le 1$ for $0 < \alpha \le 1$, the last display is bounded from below by

$$\Gamma(m)\epsilon^{2(N-1)} \prod_{i=1}^N \alpha_i \ge \Gamma(m)\epsilon^{2(N-1)}(A\epsilon^b)^N \ge Ce^{-cN\log_-\epsilon}.$$

This concludes the proof in the case that $M = 1$.

We may assume without loss of generality that a "general" M is an integer, and represent the Dirichlet vector as the aggregation $\sum_{m=1}^M (X_{1,m}, \ldots, X_{N,M})$ of a Dirichlet vector $(X_{j,m}\colon j = 1, \ldots, N, m = 1, \ldots, M)$ with parameters $(\alpha_{j,m}\colon j = 1, \ldots, N, m = 1, \ldots, M)$, where $\alpha_{j,m} = \alpha_j/M$. The event on the left side of the lemma contains the event

$$\Big\{\sum_{j=1}^N \sum_{m=1}^M \Big|X_{j,m} - \frac{w_j}{M}\Big| \le 2\epsilon, \min_{1\le j\le N-1, 1\le m\le M} X_{j,m} > \epsilon^2/2\Big\}.$$

The result now follows from the special case. □

6.4 Complements

Lemma 6.13. *For every pair of probability densities p and q,*

$$K(p;q) \lesssim 2h^2(p,q)\Big(1 + \log\Big\|\frac{p}{q}\Big\|_\infty\Big),$$
$$V(p;q) \lesssim h^2(p,q)\Big(1 + \log\Big\|\frac{p}{q}\Big\|_\infty\Big)^2.$$

Furthermore, for every pair of probability densities p and q and every $r \in (0, 0.4)$,

$$K_2(p;q) \le h^2(p,q)\Big(4 + 2\log\frac{1}{r}\Big)^2 + 8P\Big[\Big(\log\frac{p}{q}\Big)^2 \mathbb{1}\Big\{\frac{p}{q} > \frac{1}{r}\Big\}\Big].$$

Exercises

6.1 Simulate vectors $(\theta_1, \theta_2, \ldots, \theta_n)$ from the posterior distribution of the Dirichlet process mixture using the Gibbs sampler. For simplicity take the kernel $x \mapsto \psi(x,\theta)$ to be the density of the $N(\theta, 1)$-distribution, the base measure α equal to $N(0,1)$, and use data $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(0,1)$, with $n = 20$. Start by analytically calculating the (unnormalized) weights $q_{i,i}$ and the measures $G_{b,i}$ (which will be normal).

6.2 For ϕ the standard normal density, let $p_F(x) = \int \phi(x-\theta)\, dF(\theta)$. We take the domain of F to be the set of all probability measures on the interval $[-1, 1]$.

(a) Show that the map $F \mapsto p_F(x)$ is continuous relative to the weak topology, for every x.

(b) Show that $p_F(x) \le 1$ and $p_F(x) \ge e^{-x^2}/(e\sqrt{2\pi})$, for every x.

(c) Show that $F \mapsto P_{F_0} \log(p_{F_0}/p_F)$ is continuous at any probability measure F_0 on $[-1, 1]$.
(d) Show that P_{F_0} possesses the Kullback-Leibler property relative to the prior induced on p_F by $F \sim \mathrm{DP}(\alpha)$, for α a distribution with positive density on $[-1, 1]$.
(e) Show that the posterior distribution of p_F relative to this prior based on a random sample of size n from P_F is $\mathbb{L}_1$-consistent at p_{F_0}. [Hint: apply Theorem 5.15 with the entropy estimated as in Example 6.5.]

7

Gaussian Process Priors

Gaussian processes are widely used as prior distributions for functional parameters. They are stochastic processes whose finite-dimensional distributions are multivariate Gaussian vectors, and can take many different forms. In this chapter we define Gaussian processes, study basic properties and examples, and derive rates of posterior contraction.

7.1 Stochastic processes

We already encountered stochastic processes when discussing random measures. The general definition, for an arbitrary set T, is as follows.

Definition 7.1 (Stochastic process)**.** A *stochastic process* indexed by T is a collection $W = (W_t : t \in T)$ of random variables W_t defined on a common probability space $(\Omega, \mathscr{U}, \Pr)$, i.e. a collection of measurable maps $W_t : \Omega \to \mathbb{R}$.

If T is a subset of $\mathbb{R}$, then we may think of the index t as "time". Many stochastic processes model the evolution of a process in time, but in Bayesian nonparametric applications T often has a different meaning.

Instead of real-valued random variables one may also consider stochastic processes with values in a more general state space, but the space $\mathbb{R}$ will suffice for our purposes.

Because the variables W_t are defined on the same probability space, the map $t \mapsto W_t(\omega)$ is well defined, for every $\omega \in \Omega$. These maps are called the *trajectories*, or *sample paths* of the process. They are functions from T to $\mathbb{R}$, i.e. elements of $\mathbb{R}^T$. Correspondingly, a stochastic process W can be considered a random function, and its induced law is a prior probability measure on functions. This "law" is always defined on the product σ-field on $\mathbb{R}^T$. In fact $W : \Omega \to \mathbb{R}^T$ is a stochastic process if and only if it is measurable relative to this product σ-field (see Proposition 7.47).

Often we prefer a prior process W that has "nice" sample paths, for instance continuous, integrable, or differentiable. The process W can then be viewed as a measurable map $W : (\Omega, \mathscr{U}, \Pr) \to (\Theta, \mathscr{B})$ in a subset $\Theta \subset \mathbb{R}^T$, equipped with some natural σ-field $\mathscr{B}$. For instance, if $T = [0, 1]$ and the sample paths of W are continuous, then we may consider W as a map in $\mathfrak{C}[0, 1]$.

We have insisted on defining a stochastic process as a map on an underlying probability space, but shall mostly be interested in its "law", and not in the exact construction. The law of the process on $\mathbb{R}^T$ is determined by the collection of all distributions of vectors of the the form $(W_{t_1}, \ldots, W_{t_n})$, for $n \in \mathbb{N}$ and $t_1, \ldots, t_n \in T$. These vectors are called the *finite-dimensional marginals* of the process W, and their distribution are its *finite-dimensional*

distributions (or *fdds*). Kolmogorov's extension theorem, Proposition 4.18, allows to start with a *consistent* set of finite-dimensional distributions, and then yields a construction of a stochastic process with these fdds: a probability space $(\Omega, \mathscr{U}, \Pr)$ and a measurable map $W\colon \Omega \to \mathbb{R}^T$ whose finite-dimensional marginals possess the given finite-dimensional distributions.

A next step could be to verify that there also exists a *version* of this map W whose sample paths have a desired property, such as continuity. A process $\tilde{W}$ is called a *version* or *modification* of a process W if it is defined on the same probability space and $\tilde{W}_t = W_t$, almost surely, for every t. Clearly this implies that $\tilde{W}$ and W possess the same marginal distributions. However, because T is typically uncountable, it may well be that a process and a version satisfy $\Pr(\tilde{W}_t = W_t, \forall t \in T) < 1$, so that their sample paths do not agree. This makes sample path properties a subtle matter, and it often requires work to construct a nice version of a stochastic process, such as a continuous one, or even prove that such a version exists. (Actually even the definition of the preceding probability requires care, as the event in question may not be measurable!)

7.2 Gaussian processes

Definition 7.2. A stochastic process is called *Gaussian* if all its finite-dimensional marginals are multivariate-normally distributed.

A multivariate-normal distribution is determined by a mean vector and a covariance matrix. Correspondingly, the set of finite-dimensional distributions of a Gaussian process W is determined by the two functions $m\colon T \to \mathbb{R}$ and $r\colon T \times T \to \mathbb{R}$ given by

$$\begin{aligned} m(t) &= \mathrm{E} W_t, \\ r(s,t) &= \operatorname{cov}(W_s, W_t). \end{aligned}$$

The function m is called the *mean function* and the function r the *covariance function* of the process. Any two Gaussian processes with the same mean function and covariance function are versions of each other. A process with mean function equal to zero is called *centered*.

The mean function of a Gaussian process can be an arbitrary function, but the covariance function must be symmetric and positive-definite. The latter means that, for all $a_1, \dots, a_n \in \mathbb{R}$ and $t_1, \dots, t_n \in T$,

$$\sum_{i=1}^{n} \sum_{i=1}^{n} a_i a_j r(t_i, t_j) \geq 0.$$

In fact, this expression can be seen to be $\operatorname{var} \sum_i a_i W_{t_i}$. That *every* symmetric and positive-definite function is the covariance function of some Gaussian process can be shown with the help of Kolmogorov's extension theorem, Proposition 4.18.

The question whether the sample paths of a stochastic process have a certain regularity is not well posed, but we can give conditions under which a version with a certain regularity exists. For a Gaussian process these conditions will only depend on the mean and covariance function. Because a process minus its mean function is centered, it is not a great loss of generality to consider only centered processes. For the case that the index set T is a subset of Euclidean space, there are the following basic results, which essentially show that the

regularity of the covariance function carries over to the sample paths (almost) of a suitable version.

Proposition 7.3 (Modulus)**.** *If $(W_t: t \in T)$ is a centered Gaussian process indexed by a compact set $T \subset \mathbb{R}^d$ with $\mathrm{E}|W_s - W_t|^2 \le \|s-t\|^{2\alpha}$, for all $s, t \in T$ and some $\alpha \in (0, 1]$, then W possesses a version with continuous sample paths such that $|W_s - W_t| = O(\|s-t\|^\alpha \log(1/\|s-t\|))$, uniformly in (s, t) with $\|s-t\| \to 0$, almost surely.*

Since $\log(1/\|s-t\|) \to \infty$ as $\|s-t\| \to 0$, the preceding proposition shows that the sample paths of W are nearly Lipschitz continuous of order α; they are Lipschitz continuous of any order $a < \alpha$.

For a multi-index $j = (j_1, \ldots, j_d)$ of nonnegative integers , let D^j denote the mixed partial derivative operator $\partial^{|j|}/\partial t_1^{j_1} \cdots \partial t_d^{j_d}$. Furthermore, let $D_s^i D_t^j r(s, t)$ denote the function obtained by differentiating the covariance function i times with respect to s and j times with respect to t.

Proposition 7.4 (Differentiability)**.** *If the partial derivatives $(s, t) \mapsto D_s^j D_t^j r(s, t)$ of order j of the covariance function r of the centered Gaussian process $(W_t: t \in T)$, for T an interval in $\mathbb{R}^d$, exist and are Lipschitz continuous of order $\alpha > 0$, then W possesses a version whose sample paths are partially differentiable up to order j with jth order derivative that is Lipschitz of order a, for any $a < \alpha$.*

For proofs, see e.g. Appendix I of Ghoshal and van der Vaart (2017). For one-dimensional processes a slightly weaker result than Proposition 7.3 can be obtained from Kolmogorov's classical criterion (see Proposition 7.48).

7.3 Gaussian regression

Although Gaussian process priors are commonly used for many statistical problems, they are particularly attractive in the regression problem. This is a rare case of a nonparametric problem with a *conjugate* family of priors.

Suppose we have observations $Y = (Y_1, \ldots, Y_n)^T$ satisfying the regression relation

$$Y_i = \theta(t_i) + \epsilon_i, \qquad i = 1, 2 \ldots, n,$$

where the t_i are fixed, known elements of a set T, the ϵ_i are independent standard normal variables, and the unknown regression function $\theta: T \to \mathbb{R}$ is the object of interest. We wish to make Bayesian inference about the function θ using a Gaussian process prior with mean function m and covariance function r.

Although a prior draw gives a full function $(\theta(t): t \in T)$, the likelihood of the data Y actually depends only on the vector $\vec{\theta} = (\theta(t_1), \ldots, \theta(t_n))^T$. This implies that in the Bayesian setup the process $(\theta(t): t \in T)$ is conditionally independent of Y given $\vec{\theta}$. Consequently, the posterior distribution of θ can be decomposed as

$$\Pi(\theta \in B | Y) = \int \Pi(\theta \in B | \vec{\theta}) \, \Pi(d\vec{\theta} | Y).$$

The integrating measure $\Pi(\vec{\theta} \in \cdot | Y)$ of the integral is the posterior distribution of $\vec{\theta}$. The integrand concerns the conditional law of θ given $\vec{\theta}$, i.e. given its values at the design points

$t_1, \ldots, t_n$. This is free of the observations, and hence completely determined by the prior. Properties of the multivariate-normal distribution (see Lemma 7.56) show that it is the law of a Gaussian process, with mean equal to a linear function of $\vec{\theta}$ and covariance function independent of $\vec{\theta}$. (The corresponding process is called the *kriging* of θ. Its values at the design points $t_1, \ldots, t_n$ are degenerate at the values $\theta(t_i)$: the sample paths of the kriging process interpolate the observed values.) Thus it has a representation

$$\theta | \vec{\theta} \sim b_0 + \sum_{i=1}^{n} \theta(t_i) b_i + W,$$

for deterministic functions $t \mapsto b_i(t)$ (for $i = 0, 1, \ldots, n$) and a centered Gaussian process $W = (W_t : t \in T)$ that is independent of $(\vec{\theta}, Y)$. By the conjugacy of the normal distribution as a prior for the multivariate-normal distribution (see Example 2.4), the posterior distribution $\Pi(\vec{\theta} \in \cdot \,| Y)$ is multivariate-normal. It can be represented as

$$\vec{\theta} | Y \sim (I + R_n^{-1})^{-1}(Y - \mu_n) + \mu_n + V,$$

for $\mu_n = (m(t_i))$ and $R_n = (r(t_i, t_j))$ the prior mean vector and covariance matrix of $\vec{\theta}$, and V a mean-zero, multivariate-normally distributed vector with covariance matrix $(I + R_n^{-1})^{-1}$. Combining the preceding two displays, we conclude that under the posterior distribution the process θ is distributed as

$$\theta | Y \sim b_0 + \Big((I + R_n^{-1})^{-1}(Y - \mu_n) + \mu_n + V \Big)^T b + W,$$

for independent Gaussian variables V and W, where $b = (b_1, \ldots, b_n)^T$. This is the sum of two independent Gaussian processes, and hence is itself a Gaussian process.

Thus a Gaussian process is a conjugate prior in the nonparametric regression model. The updating formula can be cast in updating formulas for the mean function and covariance function, which can be obtained along the preceding lines.

Figure 7.3 shows a simulation example, with $T = [0, 1]$ and integrated Brownian as a prior: this has mean function $m = 0$ and covariance function $r(s, t) = s^2 t/2 - t^3/16$, for $s < t$. We simulated $n = 200$ observations, with $t_i = i/n$. The black curve depicts the true regression function used in the simulations, the black dots are the simulated noisy data points. The left panel shows 10 draws from the prior; the right panel 10 draws from the corresponding posterior.

In more realistic situations the errors ϵ_i have an unknown variance σ^2, which is then endowed with a prior distribution as well. The resulting posterior distribution for θ is then not Gaussian any more. However, one might use an MCMC method to generate (approximate) draws from the posterior. This is particularly simple for an inverse Gamma prior of σ^2, in which case the posterior of $1/\sigma^2$ is Gamma, while the conditional of θ given (σ^2, Y) is Gaussian, in view of the preceding (see Example 2.4).

7.4 RKHS and concentration

The *first chaos* $\overline{\text{lin}}(W)$ of a Gaussian process $W = (W_t : t \in T)$ defined on some probability space $(\Omega, \mathscr{U}, \Pr)$ is the closure in $L_2(\Pr)$ of the collection of all linear combinations $\sum_{i=1}^{n} a_i W_{t_i}$ of variables W_t, for $n \in \mathbb{N}$, $t_1, \ldots, t_n \in T$, and $a_1, \ldots, a_n \in \mathbb{R}$.

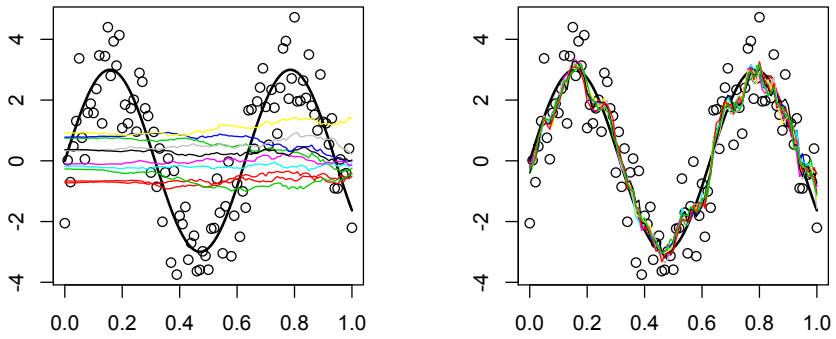

Figure 7.1 Left: 10 draws from the integrated Brownian motion prior (gray), the true regression function and the data. Right: 10 draws from the posterior (gray) and the true regression function and the data.

Definition 7.5 (RKHS)**.** The *reproducing kernel Hilbert space* (or *RKHS*) $\mathbb{H}$ associated to the centered Gaussian process W is the set of functions

$$t \mapsto h_L(t) := \mathrm{E}[W_t L], \qquad L \in \overline{\mathrm{lin}}(W)$$

equipped with the inner product and norm given by

$$\langle h_{L_1}, h_{L_2} \rangle_{\mathbb{H}} = \mathrm{E}[L_1 L_2], \qquad \|h_L\|_{\mathbb{H}} = \sqrt{\mathrm{E}[L^2]}.$$

To verify that the given formula indeed defines an inner product, one should first check that the variable $L \in \overline{\mathrm{lin}}(W)$ is uniquely determined by the function h_L. Next one readily sees that the map $L \mapsto h_L$ defines a Hilbert space isometry between $\overline{\mathrm{lin}}(W)$ and $\mathbb{H}$, whence the RKHS inherits its Hilbert space structure from the first chaos space.

Property (ii) of the following lemma is called the *reproducing property*, which gives the RKHS its name. It shows that a function in $\mathbb{H}$ can be evaluated at the point t by taking its inner product with the function $r(\cdot, t)$, derived from the covariance function r, which is called also *reproducing kernel* in this context. The final assertion shows that for a process with uniformly continuous sample paths, the functions in the RKHS are also uniformly continuous, and that the embedding of the RKHS in the space of uniformly continuous functions is continuous with norm bounded by $\sigma(W)$.

Lemma 7.6. *For a centered Gaussian process $W = (W_t : t \in T)$ with covariance function r:*

- *the function $t \mapsto r(s,t)$ is contained in $\mathbb{H}$, for every $s \in T$, and represented by W_s.*
- *$h(t) = \langle h, r(\cdot, t) \rangle_{\mathbb{H}}$, for every $h \in \mathbb{H}$ and $t \in T$.*
- *$\langle r(\cdot, s), r(\cdot, t) \rangle_{\mathbb{H}} = r(s,t)$, for every $s, t \in T$.*

Furthermore, if W has uniformly continuous sample paths relatively to some metric ρ on T, then so does every element $h \in \mathbb{H}$, and $\sup_{t\in T} |h(t)| \leq \sigma(W)\,\|h\|_{\mathbb{H}}$, for every $h \in \mathbb{H}$, where $\sigma^2(W) = \sup_{t\in T} \mathrm{E}W_t^2$.

Proof For (i) it suffices to note that $h_{W_s}(t) = \mathrm{E}W_tW_s$, by applying the definition of h_L to $L = W_s$, and $\mathrm{E}W_tW_s = r(s,t)$. Next for (ii) we have have $\langle h_L, r(\cdot,t)\rangle_{\mathbb{H}} = \langle h_L, h_{W_t}\rangle_{\mathbb{H}} = \mathrm{E}LW_t = h_L(t)$, for every $L \in \overline{\mathrm{lin}}(W)$. Property (iii) is immediate upon choosing $h = r(\cdot, s)$ in (ii).

We have $|h_L(s) - h_L(t)| = \mathrm{E}[(W_s - W_t)L]$, which is bounded above by the square root of $\mathrm{E}(W_s - W_t)^2\mathrm{E}L^2$, by the Cauchy-Schwarz inequality. If the sample paths of W are uniformly continuous, then the right side tends to zero as $\rho(s,t) \to 0$ (Exercise 7.10). It follows that every function in the RKHS is uniformly continuous. The inequality $|h_L(t)| \leq \sigma(W)\,\|h_L\|_{\mathbb{H}}$, for every $t \in T$, follows similarly. □

Example 7.7 (Euclidean space)**.** To gain insight in the RKHS it helps to consider a Gaussian random vector $W \sim \mathrm{Nor}_k(0,\Sigma)$ in $\mathbb{R}^k$. This can be identified with the stochastic process $W = (W_i\colon i = 1,\ldots,k)$ on the time set $T = \{1,2,\ldots,k\}$ and is of course also a random element in Euclidean space $\mathbb{R}^k$, which is trivially identical to $\mathfrak{UC}(T,\rho)$, for ρ any metric on T that generates the discrete topology (e.g. $\rho(s,t) = 1_{s\neq t}$). The covariance kernel is $K(i,j) = \Sigma_{i,j}$ and the RKHS is the space of functions $z_\alpha\colon \{1,\ldots,k\} \to \mathbb{R}$ given by $z_\alpha(i) = \mathrm{E}\left[W_i(\alpha^T W)\right] = (\Sigma\alpha)_i$ indexed by the (coefficients of the) linear combinations $\alpha^\top W \in \mathrm{lin}(W_1,\ldots,W_k)$, with inner product $\langle z_\alpha, z_\beta\rangle_{\mathbb{H}} = \mathrm{E}\left[(\alpha^\top W)(\beta^\top W)\right] = \alpha^\top\Sigma\beta$. We can identify z_α with the vector $\Sigma\alpha$, and the inner product then satisfies $\langle\Sigma\alpha,\Sigma\beta\rangle_{\mathbb{H}} = \alpha^\top\Sigma\beta$.

In other words, the RKHS is the range of the covariance matrix, with inner product given through the (generalized) inverse of the covariance matrix. If the covariance matrix is non-singular, then the RKHS is $\mathbb{R}^k$, but equipped with the inner product generated by the inverse covariance matrix Σ^{-1}.

The RKHS reflects the familiar ellipsoid contours of the density of the multivariate normal distribution.

Lemma 7.8. *If V and W are independent centered Gaussian processes defined on the same probability space, then the RKHS of the process $V + W$ is equal to the direct sum of the RKHS of V and W: all functions $g + h$ with g and h ranging over the RKHS of V and W, respectively, with norm $(\|g\|_{\mathbb{H}}^2 + \|h\|_{\mathbb{H}}^2)^{1/2}$.*

Proof By the independence the first chaos of $V + W$ is the set of variables $K + L$, for $K \in \overline{\mathrm{lin}}(V)$ and $L \in \overline{\mathrm{lin}}(W)$. The variables K and L are independent, and also independent of W and V, respectively. Therefore $\mathrm{E}(V + W)_t(K + L) = \mathrm{E}V_tK + \mathrm{E}W_tL$, and $\mathrm{E}(K + L)^2 = \mathrm{E}K^2 + \mathrm{E}L^2$. □

Three properties of Gaussian processes relating to their RKHS are important for the analysis of Gaussian priors. In the following propositions and lemma we assume that T is a totally bounded metric space with metric ρ, and that W has uniformly continuous sample paths relative to ρ. For instance, let T be the unit square $[0,1]^d$ with the usual metric. The process W is then automatically a Borel measurable map in the Banach space $\mathfrak{UC}(T,\rho)$ of uniformly continuous functions $z\colon T \to \mathbb{R}$ equipped with the supremum norm

$$\|z\| = \sup_{t\in T} |z(t)|.$$

See Lemma 7.43. The results extend to more general spaces (although the definition of RKHS may have to be adapted). This motivates to formulate the results for a general Banach space $\mathbb{B}$ with norm $\|\cdot\|$.

The first property of the RKHS relates to the shape of the Gaussian distribution, and is described in *Borell's inequality*. Let $\mathbb{B}_1$ and $\mathbb{H}_1$ stand for the unit balls of the spaces $\mathbb{B}$ and $\mathbb{H}$, respectively, under their norms, and let Φ be the cumulative distribution function of the standard normal distribution. Furthermore, let $\varphi_0(\epsilon)$ be the *small ball exponent*, defined through, for $\epsilon > 0$,

$$\Pr(\|W\| < \epsilon) = e^{-\varphi_0(\epsilon)}. \tag{7.1}$$

Proposition 7.9 (Borell's inequality)**.** *For any centered Gaussian random element W in a separable Banach space and every $\epsilon, M > 0$,*

$$\Pr(W \in \epsilon\,\mathbb{B}_1 + M\mathbb{H}_1) \geq \Phi\Big(\Phi^{-1}(e^{-\varphi_0(\epsilon)}) + M\Big).$$

For $M = 0$ the inequality in the proposition is an equality, and just reduces to the definition of the small ball exponent. For $M \to \infty$ and fixed $\epsilon > 0$, the right side tends to 1 like the tail of the normal distribution. This shows that the bulk of the distribution of W is contained in an ϵ-shell of a big multiple of the unit ball of the RKHS. We should keep in mind here the ellipsoid shapes found in Example 7.7, which is coded in general in the shape of $\mathbb{H}_1$ within the Banach space $\mathbb{B}$. Not only does the Gaussian distribution concentrate most mass close to zero (it has small tails), but also it distributes the mass unevenly in the infinitely many possible directions, as determined by the shape of the RKHS.

The addition of the small ball $\epsilon\mathbb{B}_1$ creates an ϵ-cushion around the multiple $M\mathbb{H}_1$. This is necessary to capture the mass of W, because the RKHS itself may have probability zero. Since $M\mathbb{H}_1 \uparrow \mathbb{H}$ as $M \uparrow \infty$, we have the equality $\Pr(W \in \epsilon\mathbb{B}_1 + \mathbb{H}) = 1$, for any $\epsilon > 0$, and hence W is supported on the closure $\bar{\mathbb{H}}$ of the RKHS in $\mathbb{B}$. Thus to use a process W as a prior for estimating a function, we must at least ensure that the closure of its RKHS contains the true function.

The second property of a Gaussian variable relates to the change in measure under changes of location. Zero-mean Gaussian variables put most mass near 0; in the Euclidean case this results from the mode of the density being at the mean, while *Anderson's lemma* expresses this in general. The decrease of mass in a ball of a given radius if the center of this ball is moved away from 0 may be studied quantitatively using Radon-Nikodym derivatives.

The distributions of the two Gaussian variables W and $W + h$ can be shown to be either mutually absolutely continuous or orthogonal, depending on whether the shift h is contained in the RKHS or not. In the first case, when $h \in \mathbb{H}$, a density of the law of $W + h$ relative to the law of W can be shown to take the form

$$\frac{dP_{W+h}}{dP_W}(W) = e^{Uh - \frac{1}{2}\|h\|_{\mathbb{H}}^2},$$

where $U\colon \mathbb{H} \to \overline{\text{lin}}(W)$ is the linear isometry defined by $U(h_L) = L$, for $L \in \overline{\text{lin}}(W)$. The formula allows us to compute the small ball probability $\Pr(\|W + h\| < \epsilon)$ of the shifted variable (a *decentered small ball probability*) in terms of the distribution of W, and leads to the following lemma.

Lemma 7.10 (Decentered small ball)**.** *For any $h \in \mathbb{H}$ and every $\epsilon > 0$ we have*

$$\Pr(\|W + h\| < \epsilon) \geq e^{-\frac{1}{2}\|h\|_{\mathbb{H}}^2} \Pr(\|W\| < \epsilon).$$

Proof Since W and $-W$ have the same distribution, $\Pr(\|W + h\| < \epsilon) = \Pr(\|W - h\| < \epsilon)$. By the Cameron-Martin formula,

$$\Pr(\|W + h\| < \epsilon) = \int 1_{\|z\|<\epsilon} \, dP_{W+h}(z) = \mathrm{E} e^{Uh - \frac{1}{2}\|h\|_{\mathbb{H}}^2} 1_{\|W\|<\epsilon}.$$

This is true with $-h$ in the place of h as well. Combining these two facts we get

$$\begin{aligned}\Pr(\|W - h\| < \epsilon) &= \frac{1}{2}\mathrm{E} e^{Uh - \frac{1}{2}\|h\|_{\mathbb{H}}^2} 1_{\|W\|<\epsilon} + \frac{1}{2}\mathrm{E} e^{U(-h) - \frac{1}{2}\|-h\|_{\mathbb{H}}^2} 1_{\|W\|<\epsilon} \\ &= e^{-\frac{1}{2}\|h\|_{\mathbb{H}}^2} \mathrm{E} \cosh(Uh) 1_{\|W\|<\epsilon},\end{aligned}$$

where $\cosh(x) = (\exp(x) + \exp(-x))/2 \geq 1$, for every x, so that the expectation is at least $\Pr(\|W\| < \epsilon)$. □

The lemma only applies to shifts within the reproducing kernel Hilbert space, but it can be extended to general shifts by approximation. Here we restrict to shifts w inside the closure of the RKHS, as otherwise a small enough ball around w will have probability zero. Define the *concentration function* of W at w by

$$\varphi_w(\epsilon) = \inf_{h \in \mathbb{H} : \|h - w\| \leq \epsilon} \tfrac{1}{2}\|h\|_{\mathbb{H}}^2 - \log \Pr(\|W\| < \epsilon). \tag{7.2}$$

For $w = 0$ this reduces to the small ball exponent $\varphi_0(\epsilon)$ defined in (7.1) (the infimum is achieved for $h = 0$), and it measures concentration of W at 0. The extra term if $w \neq 0$, which will be referred to as the *decentering function*, measures the decrease in mass when shifting from the origin to w. Indeed, up to constants, the concentration function is the exponent of the small ball around w, for every w in the support of W.

Proposition 7.11 (Small ball exponent)**.** *For any centered Gaussian random element in a separable Banach space, and any w in the closure of its RKHS, and any $\epsilon > 0$,*

$$\varphi_w(\epsilon) \leq -\log \Pr(\|W - w\| < \epsilon) \leq \varphi_w(\epsilon/2).$$

The asymptotic behavior for $\epsilon \to 0$ of the centered small ball probability $\Pr(\|W\| < \epsilon)$ on the logarithmic scale turns out to be closely related to the "size" of the RKHS unit ball $\mathbb{H}_1$. More precisely, it is determined by the asymptotic behavior for $\epsilon \to 0$ of the metric entropy $\log N(\epsilon, \mathbb{H}_1, \|\cdot\|)$. The following special cases suffice for our purposes.

Proposition 7.12. *For any centered Gaussian random element in a separable Banach space, and any $\alpha, \gamma > 0$, as $\epsilon \to 0$,*

$$\log N(\epsilon, \mathbb{H}_1, \|\cdot\|) \asymp \epsilon^{-2\alpha/(2+\alpha)} \iff -\log \Pr(\|W\| < \epsilon) \asymp \epsilon^{-\alpha}, \tag{7.3}$$

$$\log N(\epsilon, \mathbb{H}_1, \|\cdot\|) \asymp \log^{\gamma} \frac{1}{\epsilon} \iff -\log \Pr(\|W\| < \epsilon) \asymp \log^{\gamma} \frac{1}{\epsilon}. \tag{7.4}$$

7.5 Posterior contraction rates

In this section we first give a generic result on Gaussian priors, which characterizes rates such that the prior has sufficient mass near a given "true function" w_0 and almost all its mass in a set of bounded complexity. The formulation of the result reminds of Theorem 5.19 on posterior contraction rates. However, the result is purely in terms of the norm of the Banach space in which the prior process lives. Next we apply the generic result to obtain rates of posterior contraction for Gaussian process priors in standard statistical settings by relating the statistically relevant norms and discrepancies to the Banach space norm.

Theorem 7.13 (Gaussian contraction rate)**.** *Let W be a centered Gaussian random element in a separable Banach space $\mathbb{B}$ with RKHS $\mathbb{H}$ and let $w_0 \in \bar{\mathbb{H}}$. If $\epsilon_n > 0$ is such that*

$$\varphi_{w_0}(\epsilon_n) \le n\epsilon_n^2, \tag{7.5}$$

then

$$\Pr\left(\|W - w_0\| < 2\epsilon_n\right) \ge e^{-n\epsilon_n^2}, \tag{7.6}$$

and for any $C > 1$ such that $Cn\epsilon_n^2 > \log 2$ there exists a measurable set $B_n \subset \mathbb{B}$ such that

$$\log N(3\epsilon_n, B_n, \|\cdot\|) \le 9Cn\epsilon_n^2, \tag{7.7}$$

$$\Pr(W \notin B_n) \le e^{-Cn\epsilon_n^2},. \tag{7.8}$$

Proof Inequality (7.6) is an immediate consequence of (7.5) and Proposition 7.11. We need to prove existence of the set B_n such that (7.7) and (7.8) hold.

Set $B_n = \epsilon_n\mathbb{B}_1 + M_n\mathbb{H}_1$, where $\mathbb{B}_1$ and $\mathbb{H}_1$ are the unit balls of $\mathbb{B}$ and $\mathbb{H}$, respectively, and M_n is a positive constant. Then $\Pr(W \notin B_n) \le 1 - \Phi(\alpha_n + M_n)$ for α_n given by $\Phi(\alpha_n) = \Pr(W \in \epsilon_n\mathbb{B}_1) = e^{-\varphi_0(\epsilon_n)}$, by Borell's inequality. Since $\varphi_0(\epsilon_n) \le \varphi_{w_0}(\epsilon_n) \le n\epsilon_n^2$ and $C > 1$, we have that $\alpha_n \ge -M_n/2$ if $M_n = -2\Phi^{-1}(e^{-Cn\epsilon_n^2})$. It follows that for this choice $\Pr(W \notin B_n) \le 1 - \Phi(M_n/2) = e^{-Cn\epsilon_n^2}$.

It remains to verify the complexity estimate (7.7). If $h_1, \ldots, h_N \in M_n\mathbb{H}_1$ are $2\epsilon_n$-separated in terms of the Banach space norm $\|\cdot\|$, then the ϵ_n-balls $h_1 + \epsilon_n\mathbb{B}_1, \ldots, h_N + \epsilon_n\mathbb{B}_1$ are disjoint and hence, by Lemma 7.10,

$$1 \ge \sum_{j=1}^{N} \Pr(W \in h_j + \epsilon_n\mathbb{B}_1) \ge \sum_{j=1}^{N} e^{-\|h_j\|_{\mathbb{H}}^2/2} \Pr(W \in \epsilon_n\mathbb{B}_1) \ge Ne^{-M_n^2/2}e^{-\varphi_0(\epsilon_n)}.$$

For a maximal $2\epsilon_n$-separated set $h_1, \ldots, h_N$, the balls around $h_1, \ldots, h_N$ of radius $2\epsilon_n$ cover the set $M_n\mathbb{H}_1$ and hence we obtain the estimates $N(2\epsilon_n, M_n\mathbb{H}_1, \|\cdot\|) \le N \le e^{M_n^2/2}e^{\varphi_0(\epsilon_n)}$. Since any point of B_n is within ϵ_n of an element of $M_n\mathbb{H}_1$, this is also a bound on $N(3\epsilon_n, B_n, \|\cdot\|)$. To complete the proof, observe that $M_n^2/2 \le 8Cn\epsilon_n^2$ if $e^{-Cn\epsilon_n^2/2} < 1$, by the definition of M_n, because $\Phi^{-1}(y) \ge -2\sqrt{\log(1/y)}$ and is negative for every $y \in (0, 1/2)$. □

Ignoring multiplicative constants in the rate, we can rewrite relation (7.5) as the pair of inequalities

$$-\log\Pr\left(\|W\| < \epsilon_n\right) \le n\epsilon_n^2, \qquad \inf_{h\in\mathbb{H}:\|h-w_0\|\le\epsilon_n} \|h\|_{\mathbb{H}}^2 \le n\epsilon_n^2.$$

Both inequalities have a minimal solution ϵ_n, and the final rate ϵ_n satisfying (7.5) is the maximum of the two minimal solutions, up to a constant. The first inequality in the preceding

display concerns the small ball probability at 0. It depends on the prior, but not on the true parameter w_0: priors that put little mass near 0 will give slow rates ϵ_n, whatever the true parameter w_0. The second inequality measures the decrease of prior mass around the true parameter w_0 relative to the zero parameter (on a logarithmic scale). A prior that puts much mass around 0 may still give bad performance for a nonzero w_0, depending on its position relative to the RKHS. The most favorable situation is that the true parameter is contained in the RKHS: then the choice $h = w_0$ is eligible in the infimum, whence the infimum is bounded by $\|w_0\|_{\mathbb{H}}^2$, and the condition merely says that ϵ_n must not be smaller than a multiple of the "parametric rate" $n^{-1/2}$. However, the RKHS can be a very small space, and hence this favorable situation is rare.

By Proposition 7.11 the concentration function $\varphi_{w_0}(\epsilon)$ measures the prior mass around w_0. Thus the theorem shows that for Gaussian priors the rate of contraction is driven by the prior mass condition only. The existence of sieves of a prescribed complexity, required in (5.6) and (5.7) of the general rate theorems, is implied by the prior mass condition. The preceding theorem establishes this only for entropy, neighborhoods and metric of convergence all described in terms of the Banach space norm, but in the sequel we show that for the standard inference problems this can be translated into the Hellinger distance and Kullback-Leibler discrepancies used in Theorem 5.19.

7.5.1 Density estimation

Consider estimating a probability density p on $[0,1]^d$ based on a sample of observations $X_1, \ldots, X_n | p \overset{\text{iid}}{\sim} p$. Because a Gaussian process can take on negative values, it would be an unnatural prior for a density function. Instead we use an exponential link function, and construct a prior Π for p as the exponential, normalized transform of a Gaussian process $W = (W_x : x \in [0,1]^d)$, given by

$$p(x) = \frac{e^{W_x}}{\int e^{W_y}\, dy}.$$

Theorem 7.14. *Let $W = (W_t : t \in [0,1]^d)$ be a centered Gaussian process with uniformly continuous sample paths. If $w_0 = \log p_0$ and ϵ_n satisfy the rate equation $\varphi_{w_0}(\epsilon_n) \le n\epsilon_n^2$, then $\Pi_n(p : h(p, p_0) > M\epsilon_n | X_1, \ldots, X_n) \to 0$ in P_0^n-probability, for some sufficiently large constant M.*

Proof The proof is based on Theorem 5.19, which is combined with Lemma 7.15 (below), and Theorem 7.13.

We choose the set $\mathcal{P}_n$ in Theorem 5.19 equal to the set $\mathcal{P}_n = \{p_w : w \in B_n\}$, for B_n the measurable set as in Theorem 7.13, with C a large constant. In view of Lemma 7.15(i) for sufficiently large n the $4\epsilon_n$-entropy of $\mathcal{P}_n$ relative to the Hellinger distance is bounded above by the $3\epsilon_n$-entropy of the set B_n relative to the uniform distance, which is bounded by $6Cn\epsilon_n^2$ by Theorem 7.13. The prior probability $\Pi(\mathcal{P}_n^c)$ outside the set $\mathcal{P}_n$ as in (5.8) is bounded by the probability of the event $\{W \notin B_n\}$, which is bounded by $e^{-Cn\epsilon_n^2}$ by Theorem 7.13. Finally, by Lemma 7.15(ii)-(iii) the prior probability as in (5.7), but with ϵ_n replaced by a multiple of ϵ_n, is bounded above by the probability of the event $\{\|W - w_0\|_\infty < 2\epsilon_n\}$, which is bounded below by $e^{-n\epsilon_n^2}$ by Theorem 7.13. □

Lemma 7.15. *For any measurable functions $f, g: \mathfrak{X} \to \mathbb{R}$, and p_f the probability densities defined by $p_f(x) = e^{f(x)-c((f)}$,*

(i) $h(p_f, p_g) \le \|f-g\|_\infty e^{\|f-g\|_\infty/2}$,
(ii) $K(p_f; p_g) \lesssim \|f-g\|_\infty^2 e^{\|f-g\|_\infty}(1+\|f-g\|_\infty)$,
(iii) $V(p_f; p_g) \lesssim \|f-g\|_\infty^2 e^{\|f-g\|_\infty}(1+\|f-g\|_\infty)^2$.

Proof The triangle inequality and simple algebra give

$$h(p_f, p_g) = \left\| \frac{e^{f/2}}{\|e^{f/2}\|_2} - \frac{e^{g/2}}{\|e^{g/2}\|_2} \right\|_2 \le 2\frac{\|e^{f/2}-e^{g/2}\|_2}{\|e^{g/2}\|_2}.$$

Because $|e^{f/2}-e^{g/2}| \le e^{g/2}e^{|f-g|/2}|f-g|/2$ for any $f, g \in \mathbb{R}$, the square of the right side is bounded by

$$\frac{\int e^g e^{|f-g|}|f-g|^2\,d\nu}{\int e^g\,d\nu} \le e^{\|f-g\|_\infty}\|f-g\|_\infty^2.$$

This proves assertion (i) of the lemma.

Next assertions (ii) and (iii) follow from (i) and the equivalence of K, V and the squared Hellinger distance if the quotient of the densities is uniformly bounded (see Lemma 6.13). From $g-\|f-g\|_\infty \le f \le g+\|f-g\|_\infty$ it follows that $c(g)-\|f-g\|_\infty \le c(f) \le c(g)+\|f-g\|_\infty$. Therefore $\left\|\log(p_f/p_g)\right\|_\infty = \left\|f-g-c(f)+c(g)\right\|_\infty \le 2\|f-g\|_\infty$. Assertions (ii) and (iii) now follow by Lemma 6.13. □

7.5.2 Nonparametric logistic regression

Consider estimating a binary regression function $p(x) = \Pr(Y=1|X=x)$ based on an i.i.d. sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ from the distribution of (X, Y), where $Y \in \{0, 1\}$, and X follows a distribution G on some totally bounded metric space $\mathfrak{X}$. Because a Gaussian process ranges over the full set of reals $\mathbb{R}$, whereas p is restricted to $[0, 1]$, it is natural to employ a link function. For $\Psi(x) = (1+e^{-x})^{-1}$ the logistic function construct a prior for p through a Gaussian process $W = (W_x: x \in \mathfrak{X})$ by

$$p(x) = \Psi(W_x).$$

The likelihood for (X, Y) factorizes as $p(x)^y(1-p(x))^{1-y}\,dG(x)$. Because the contribution of G cancels out from the posterior distribution for p, we can consider G known and need not specify a prior for it. Assume that p_0 is never zero and let $w_0 = \Psi^{-1}(p_0)$, for p_0 the true value of p.

Theorem 7.16. *Let $W = (W_x: x \in \mathfrak{X})$ be a centered Gaussian process with uniformly continuous sample paths. If $w_0 = \Psi^{-1}(p_0)$ and ϵ_n satisfies the rate equation $\varphi_{w_0}(\epsilon_n) \le n\epsilon_n^2$, then $\Pi_n(\|p-p_0\|_{2,G} > M\epsilon_n|X_1, Y_1, \ldots, X_n, Y_n) \to 0$ in P_0^n-probability, for some $M > 0$.*

Proof This follows from combining Lemma 7.17 (below), Theorem 5.19, and Theorem 7.13. □

Lemma 7.17. *For any measurable functions $f, g: \mathfrak{X} \to \mathbb{R}$, and $p_f(y|x) = \Psi(f(x))^y(1-\Psi(f(x)))^{1-y}$,*

(i) $h(p_f, p_g) \lesssim \|f - g\|_{2,G}$.
(ii) $K(p_f; p_g) \lesssim \|f - g\|_{2,G}$.
(iii $V(p_f; p_g) \lesssim \|f - g\|_{2,G}$.

Proof The square of the Hellinger distance in (i) can be written

$$\int \left| \sqrt{\Psi \circ f} - \sqrt{\Psi \circ g} \right|^2 + \left| \sqrt{1 - \Psi \circ f} - \sqrt{1 - \Psi \circ g} \right|^2 dG.$$

The functions $\sqrt{\Psi}$ and $\sqrt{1-\Psi}$ possess derivatives $\frac{1}{2}\psi/\sqrt{\Psi}$ and $\frac{1}{2}\psi/\sqrt{1-\Psi}$, which are both uniformly bounded. This readily gives (i).

To derive (ii) we express the Kullback-Leibler divergence as $K(p_f; p_g) = \int \phi_f(g)\, dG$, where

$$\phi_u(v) = \Psi(u) \log \frac{\Psi(u)}{\Psi(v)} + (1 - \Psi(u)) \log \frac{1 - \Psi(u)}{1 - \Psi(v)}.$$

The function ϕ_u possesses derivative $\phi_u'(v) = \Psi(v) - \Psi(u)$, since the function $\psi/(\Psi(1-\Psi))$ is identically equal to 1. Because the function ϕ_u vanishes at u, the mean value theorem gives that $|\phi_f(g)| \le |f - g|$, and assertion (iii) follows.

For the proof of (iii) we write $V(p_f, p_g) = \int \phi_f(g)\, dG$, where the function ϕ_f is as ϕ_f, but with the logarithmic factors squared. The result follows because both $\log(\Psi(g)/\Psi(f))$ and $\log[(1 - \Psi(g))/(1 - \Psi(f)]$ are bounded by 1. □

7.5.3 Nonparametric normal regression

Consider estimating a regression function f based on observations $Y_1, \ldots, Y_n$ in the normal regression model with fixed covariates $Y_i = f(x_i) + \varepsilon_i$, where $\varepsilon_i \overset{\text{iid}}{\sim} \text{Nor}(0, \sigma_0^2)$ and the covariates $x_1, \ldots, x_n$ are fixed elements from a set $\mathfrak{X}$.

A prior on f is induced by setting $f(x) = W_x$, for a Gaussian process $(W_x : x \in \mathfrak{X})$. If σ is unknown, then we also put a prior on σ, which we assume to be supported on an interval $[a, b] \subset (0, \infty)$ with a density π that is bounded away from zero.

Let $\|\cdot\|_n$ be the $\mathbb{L}_2(\mathbb{P}_n^x)$-norm for the empirical measure $\mathbb{P}_n^x$ of the design points $x_1, \ldots, x_n$, and let $\varphi_{w_0,n}$ be the concentration function of W viewed as a map in $\mathbb{L}_2(\mathbb{P}_n^x)$. The inconvenience that this depends on n can be removed by bounding the empirical norm by the uniform norm, which gives a corresponding bound on the concentration function.

Theorem 7.18. *Let W be a centered Gaussian random element in $\mathbb{L}_2(\mathbb{P}_n^x)$, for every n, and suppose that the true values f_0 of f and σ_0 of σ belong to the supports of W and π. If ϵ_n satisfies the rate equation $\varphi_{w_0,n}(\epsilon_n) \le n\epsilon_n^2$, then $\Pi_n((w, \sigma) : \|w - w_0\|_n + |\sigma - \sigma_0| > M\epsilon_n | Y_1, \ldots, Y_n) \to 0$ in P_0^n-probability, for some $M > 0$.*

Proof Because the observations are not identically distributed, we need an extension of Theorem 5.19 to non-i.i.d. observations. Apart from this, the theorem can be derived from Theorem 7.13. □

7.6 Examples

Brownian motion

Brownian motion is a fundamental Gaussian process and can be used as a building block to construct many more examples.

Definition 7.19. The stochastic process $B = (B_t : t \geq 0)$ is called a (standard) *Brownian motion*, or *Wiener process*, if

(i) $B_0 = 0$ a.s.,
(ii) $B_t - B_s$ is independent of $(B_u : u \leq s)$ for all $s \leq t$,
(iii) $B_t - B_s$ has a $N(0, t-s)$-distribution for all $s \leq t$,
(iv) almost all sample paths of B are continuous.

A process with property (ii) is called a process with *independent increments*. Property (iii) implies that the distribution of the increment $B_t - B_s$ only depends on $t - s$. This is called the *stationarity of the increments*. Property (ii) implies that $\operatorname{cov}(B_t, B_s) = \operatorname{var} B_s$, for $s \leq t$, and together with property (iii) then gives the covariance function

$$r(s,t) = E[B_s B_t] = s \wedge t.$$

The mean function of B is zero in view of (i) and (iii).

It is not immediately clear that Brownian motion exists. With the help of Proposition 4.18 it is not too difficult to show that there exists a stochastic process B that satisfies properties (i)–(iii) (see Exercise 7.2). Next, since $\mathrm{E}|B_s - B_t|^2 = |s-t|$ by (iii), Proposition 7.3 gives that there exists a version whose sample paths satisfy $|B_s - B_t| = O(|s-t|^\alpha)$, for every $\alpha < 1/2$. Thus this version is not only continuous, but also Hölder continuous of every order strictly less than $1/2$. [1]

It can be proved that no version of Brownian motion has sample paths that are Hölder of order exactly equal to $1/2$. (In particular, they are non-differentiable functions.) However, although then strictly speaking it is inaccurate (in the Hölder sense), we think of the sample paths of Brownian motion as having "regularity $1/2$". Figure 7.6 shows an example of a typical Brownian sample path.

Brownian motion B is a good starting point for modeling functions on $[0,1]$. The RKHS and small ball probabilities of Brownian motion are classical; the RKHS is also called the *Cameron-Martin space*. Let $\mathfrak{W}^k[0,1]$ be the *Sobolev space* of all functions $f \in \mathbb{L}_2[0,1]$ that are $k-1$ times differentiable with $f^{(k-1)}$ absolutely continuous with derivative $f^{(k)}$ belonging to $\mathbb{L}_2[0,1]$.

Lemma 7.20 (RKHS)**.** *The RKHS of Brownian motion is equal to* $\{f \in \mathfrak{W}^1[0,1] : f(0) = 0\}$ *with the inner product* $\langle f, g\rangle_{\mathbb{H}} = \int_0^1 f'(t)g'(t)\,dt$.

Lemma 7.21 (Small ball)**.** *The small ball exponent of Brownian motion viewed as a map into* $\mathfrak{C}[0,1]$ *or* $\mathbb{L}_r[0,1]$, *for some* $r \geq 1$, *satisfies, as* $\epsilon \downarrow 0$,

$$\varphi_0(\epsilon) \asymp \epsilon^{-2}.$$

[1] Recall that a function $f : T \subset \mathbb{R}$ on an interval $T \subset \mathbb{R}^d$ is called (uniformly) *Hölder continuous* of order $\alpha \in (0,1]$ if there exists a constant $C > 0$ such that $|f(t) - f(s)| \leq C\|t - s\|^\alpha$ for all $s, t \in T$.

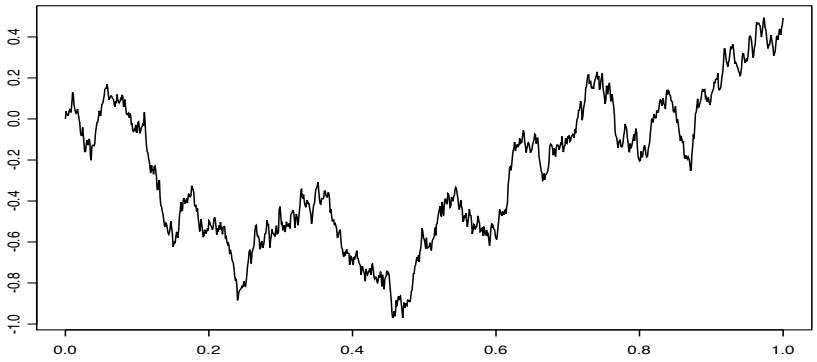

Figure 7.2 A sample path of one-dimensional Brownian motion

Proof The second lemma may be derived from the first using Proposition 7.12 and the characterization of the entropy of a Sobolev space (see Proposition 5.26), or alternatively by a probabilistic proof. For the first lemma we provide two proofs.

Since $r(s,t) = s \wedge t$, the RKHS is equal to the completion of the linear span of the functions $\{t \mapsto s \wedge t \colon s \in [0,1]\}$ under the inner product determined by

$$\langle s_1 \wedge \cdot, s_2 \wedge \cdot \rangle_{\mathbb{H}} = s_1 \wedge s_2 = \int_0^1 (s_1 \wedge t)'(s_2 \wedge t)' \, dt.$$

Here $t \mapsto (s \wedge t)' = \mathbb{1}\{[0,s]\}(t)$ is the derivative of the function $s \wedge \cdot$. The equation shows that the RKHS inner product is indeed the inner product of $\mathfrak{W}^1[0,1]$. It suffices to show that the linear span of all functions of this type is dense in $\{f \in \mathfrak{W}^1[0,1] \colon f(0) = 0\}$. Now the linear span contains every function that is 0 at 0, continuous, and piecewise linear on a partition $0 = s_0 < s_1 \cdots < s_N = 1$, since such a function with slopes α_j on the intervals (s_{j-1}, s_j), for $j = 1, \ldots, N$, can be constructed as a linear combination of the functions $s \wedge \cdot$ by first determining the coefficient of $(s_N \wedge \cdot)$ to have the correct slope on (s_{N-1}, s_N), next determining the coefficient of $(s_{N-1} \wedge \cdot)$ to have the correct slope on (s_{N-2}, s_{N-1}), etc. The derivatives of these piecewise linear functions are piecewise constant, and the set of piecewise constant functions is dense in $\mathbb{L}_2[0,1]$. Thus the completion of the linear span is as claimed.

Another proof can be based on the characterization of the first chaos $\overline{\text{lin}}(B)$ of Brownian motion as the collection of all *Wiener integrals* $\int_0^1 f(u)\, dW_u$, where $f \in L_2[0,1]$ (see Section 7.8.3 and Exercise 7.9). The isometry (7.16) implies that

$$\mathrm{E}[W_t L] = \int_0^t f(u)\, du, \quad \mathrm{E}L^2 = \int_0^1 f^2(u)\, du, \qquad \text{for } L = \int_0^1 f(u)\, dW_u.$$

Thus the RKHS consists of all integrals $\int_0^t f(u)\, du$, which have square RKHS-norms given by the L_2-norm $\|f\|_2$ of f. □

Standard Brownian motion is zero at zero, and hence for use as a prior it is preferable to

"release it at zero," by adding a variable that gives a prior for the unknown function at zero. Adding an independent Gaussian variable $Z \sim \mathrm{Nor}(0,1)$ yields another Gaussian process, of the form $t \mapsto Z + B_t$, which we call "Brownian motion released at zero".

Lemma 7.22 (RKHS)**.** *The RKHS of Brownian motion released at zero is equal to* $\mathfrak{W}^1[0,1]$ *with the inner product* $\langle f, g\rangle_{\mathbb{H}} = f(0)g(0) + \int_0^1 f'(t)g'(t)\,dt$.

Lemma 7.23 (Small ball)**.** *The small ball exponent of Brownian motion released at zero viewed as a map into* $\mathfrak{C}[0,1]$ *or* $\mathbb{L}_r[0,1]$*, for some* $r \geq 1$*, satisfies, as* $\epsilon \downarrow 0$*,*

$$\varphi_0(\epsilon) \asymp \epsilon^{-2}.$$

Proof The RKHS of the constant process $t \mapsto Z$ consists of the constant functions, and the RKHS of a sum of two independent processes is the direct sum of the RKHS of the two processes, by Lemma 7.8. This readily shows that the RKHS of $Z + B$ is the direct sum of the constant functions and the RKHS of B, which was derived in Lemma 7.20.

The addition of the single variable Z makes the small ball probability at 0 smaller, but it does not change the rate ϵ^{-2} obtained Lemma 7.21, as $-\log \Pr(|Z| < \epsilon) \asymp \log_- \epsilon \ll \epsilon^{-2}$. □

The concentration function $\varphi_{w_0}(\epsilon)$ of $Z + B$ depends on the position of the true parameter w_0 relative to the RKHS. It can be computed using a kernel smoother $w_0 * \psi_\sigma$, which is contained in the RKHS if ψ is a smooth kernel.

Lemma 7.24 (Decentering)**.** *If* $w_0 \in \mathfrak{C}^\beta[0,1]$ *for some* $\beta \in (0,1]$*, then the decentering of the concentration function of Brownian motion released at zero viewed as map in* $\mathfrak{C}[0,1]$ *satisfies, for* $\epsilon \downarrow 0$*,*

$$\inf_{h:\|h-w_0\|_\infty<\epsilon} \|h'\|_2^2 \lesssim \epsilon^{-(2-2\beta)/\beta}.$$

Proof If ψ_σ is the density of the $\mathrm{Nor}(0,\sigma^2)$-distribution, then $\|w_0 * \psi_\sigma - w_0\| \lesssim \sigma^\beta$, as $\sigma \to 0$, and the squared RKHS-norm of $w_0 * \psi_\sigma$ is given by $(w_0 * \psi_\sigma)(0)^2 + \|(w_0 * \psi_\sigma)'\|_2^2 \asymp \sigma^{-(2-2\beta)}$. Choosing $\sigma \asymp \epsilon^{1/\beta}$, we obtain the assertion. □

Combining the preceding we see that the concentration function of released Brownian motion satisfies, if $w_0 \in \mathfrak{C}^\beta[0,1]$,

$$\varphi_{w_0}(\epsilon) \lesssim \epsilon^{-(2-2\beta)/\beta} + \epsilon^{-2}.$$

For $\beta \geq 1/2$, the second term ϵ^{-2} dominates, and hence the minimal solution to the rate equation $\varphi_{w_0}(\epsilon_n) \leq n\epsilon_n^2$ satisfies $\epsilon_n^{-2} \asymp n\epsilon_n^2$, or $\epsilon_n \asymp n^{-1/4}$. For $\beta \in (0, 1/2)$, the first term dominates, leading to $\epsilon_n^{-(2-2\beta)/\beta} \lesssim n\epsilon_n^2$, with minimal solution $\epsilon_n \asymp n^{-\beta/2}$.

The resulting contraction rate can be summarized as

$$n^{-(\beta\wedge 1/2)/2}.$$

It is equal to the minimax rate of estimation $n^{-\beta/(2\beta+1)}$ for functions in a Hölder space of order β if and only if $\beta = 1/2$. This is intuitively understandable, as the sample paths of a Brownian motion are regular of that order: only matching of prior and true smoothness yields optimal results. For any positive $\beta \neq 1/2$ the posterior distribution is consistent, but the performance of the Brownian motion prior is suboptimal. The discrepancy is most felt for smooth w_0: the contraction rate is $n^{-1/4}$, no matter how smooth w_0 is. This is caused by

the rough paths of Brownian motion, which result in a tiny small ball probability. Even the zero function, which may be viewed the smoothest function of all, receives probability only of the order $\exp(-C\epsilon^{-2})$ in a ball of radius ϵ. No amount of data will fully wash out this prior preference for non-smooth functions.

Integrated Brownian motion

This shortcoming of Brownian motion as a prior for smooth functions can be remedied by integrating its sample paths. For a given function f, denote by $I_{0+}f$ its primitive function $t \mapsto \int_0^t f(s)\,ds$, and define by induction $I_{0+}^k = I_{0+}^{k-1} I_{0+}$. Since taking a primitive function increases the smoothness by 1, the k-fold integrated Brownian motion $I_{0+}^k B$ is smooth of order (nearly) $k + 1/2$. Its k vanishing derivatives at zero can be released by adding a random polynomial, yielding the process

$$W_t = \sum_{i=0}^{k} Z_i \frac{t^i}{i!} + (I_{0+}^k B)_t, \qquad Z_0, \ldots, Z_k \overset{\text{iid}}{\sim} \mathrm{Nor}(0,1) \perp\!\!\!\perp B. \tag{7.9}$$

Lemma 7.25 (RKHS)**.** *The RKHS of the process W given in (7.9), for B a Brownian motion independent of the variables $Z_1, \ldots, Z_k \overset{iid}{\sim} \mathrm{Nor}(0,1)$, is the Sobolev space $\mathfrak{W}^{k+1}[0,1]$, with inner product*

$$\langle f, g\rangle_{\mathbb{H}} = \sum_{i=0}^{k} f^{(i)}(0) g^{(i)}(0) + \int_0^1 f^{(k+1)}(t) g^{(k+1)}(t)\,dt.$$

Lemma 7.26 (Small ball)**.** *The small ball exponents of k-fold integrated Brownian motion $I_{0+}^k B$ and the process W given in (7.9), for B a Brownian motion independent of $Z_1, \ldots, Z_k \overset{iid}{\sim} \mathrm{Nor}(0,1)$, viewed as maps into $\mathfrak{C}[0,1]$ satisfy, as $\epsilon \downarrow 0$,*

$$\varphi_0(\epsilon) \asymp \epsilon^{-2/(2k+1)}.$$

Lemma 7.27 (Decentering)**.** *If $w_0 \in \mathfrak{C}^\beta[0,1]$ for $\beta \le k+1$, then the decentering function of the process in (7.9), for B a Brownian motion independent of $Z_1, \ldots, Z_k \overset{iid}{\sim} \mathrm{Nor}(0,1)$, viewed as map in $\mathfrak{C}[0,1]$ satisfies, as $\epsilon \downarrow 0$,*

$$\inf_{h:: \|h - w_0\|_\infty < \epsilon} \|h\|_{\mathbb{H}}^2 \lesssim \epsilon^{-(2k-2\beta+2)/\beta}.$$

Proofs The first lemma can be obtained from the corresponding results for Brownian motion: the RKHS of integrated Brownian motion is essentially the integral of the RKHS of Brownian motion.

The second lemma may be derived from the first using Proposition 7.12 and the characterization of the entropy of a Sobolev space (see Proposition 5.26), or alternatively by a probabilistic proof.

For a proof of the third consider the convolution $w_0 * \psi_\sigma$, of w_0 with a scaled version ψ_σ of a smooth kth order kernel (an integrable function ψ satisfying $\int \psi(t)\,dt = 1$ and $\int t^r \psi(t)\,dt = 0$ for $r = 1, \ldots, k$, and $\int |t|^{k+1} \psi(t)\,dt < \infty$). By a Taylor expansion argument this can be seen to satisfy $\|w_0 * \psi_\sigma - w_0\|_\infty \lesssim \sigma^\beta$. Furthermore, the function $w_0 * \psi_\sigma$ belongs to the RKHS and satisfies $(w_0 * \psi_\sigma)^{(l)} = w_0^{(b)} * \psi_\sigma^{(l-b)}$, for b the largest integer strictly smaller than β. Hence $\|(w_0 * \psi_\sigma)^{(l)}\|_\infty \lesssim \sigma^{-(l-\beta)}$ if $w_0 \in \mathfrak{C}^\beta[0,1]$ and $l \ge \beta$. Consequently $\int (w_0 * \psi_\sigma)^{(k+1)}(t)^2\,dt \lesssim \sigma^{-(2k-2\beta+2)}$ and the square derivatives at 0 up to order k are bounded or of smaller order in

$1/\sigma$. It follows that $\|w_0 * \psi_\sigma\|_{\mathbb{H}} \lesssim \sigma^{-(k-\beta+1)}$, if $w_0 \in \mathfrak{C}^\beta[0,1]$. The choice $\sigma \asymp \epsilon^{1/\beta}$ leads to $\|w_0 * \psi_\sigma - w_0\|_\infty \lesssim \epsilon$ and $\|w_0 * \psi_\sigma\|_{\mathbb{H}}^2 \lesssim \epsilon^{-(2k-2\beta+2)/\beta}$. □

It follows that the concentration function of "released integrated Brownian motion" (7.9) takes the form

$$\varphi_{w_0}(\epsilon) \lesssim \epsilon^{-(2k-2\beta+2)/\beta} + \epsilon^{-2/(2k+1)}.$$

For $\beta \geq k + 1/2$ the second term dominates, and the rate inequality becomes $\epsilon_n^{-2/(2k+1)} \leq n\epsilon_n^2$, with as minimal solution $\epsilon_n \asymp n^{-(2k+1)/(4k+4)}$. For $\beta \leq k+1/2$ the first term in the concentration function dominates, and the rate inequality $\epsilon_n^{-(2k-2\beta+2)/\beta} \lesssim n\epsilon_n^2$ has the minimal solution $\epsilon_n \asymp n^{-\beta/(2k+2)}$. The posterior contraction rate can be summarized as the maximum

$$n^{-(\beta\wedge k+1/2)/(2k+2)}$$

of the two rates. This is the minimax rate for β-regular functions if and only if $\beta = k + 1/2$.

* *Riemann-Liouville processes*

Brownian motion and its repeated integrals give Gaussian processes of regularities $1/2, 3/2, \ldots$. We can interpolate between these values by performing fractional integrations. By Fubini's theorem and integration by parts (Proposition 7.50),

$$\int_0^t \int_0^{t_{n-1}} \cdots \int_0^{t_1} W_{t_0}\, dt_0 dt_1 \cdots t_{n-1} = \frac{1}{n!} \int_0^t (t-s)^n\, dW_s, \qquad a.s. \tag{7.10}$$

(See Exercise 7.4). Now the right-hand side of this equation is well defined not just for $n \in \mathbb{N}$, but for every $n \in \mathbb{R}$ such that $s \mapsto (t-s)^n$ belongs to $L_2[0,t]$, i.e. for every $n > -1/2$. This leads to the definition of the following process, which can be viewed as the $(\alpha - 1/2)$-fold iterated integral of Brownian motion.

Definition 7.28. For $\alpha > 0$ and W a Brownian motion, the *Riemann-Liouville process* with parameter $\alpha > 0$ is defined by

$$R_t^\alpha = \frac{1}{\Gamma(\alpha + 1/2)} \int_0^t (t-s)^{\alpha-1/2}\, dW_s, \qquad t \geq 0.$$

It can be shown that, if used as a prior, the Riemann-Liouville process indeed gives the optimal contraction rate when the true function is α-regular, for any $\alpha > 0$.

Changing the length scale

The Gaussian process priors in the preceding example possess sample paths of a given smoothness. The generic picture is that they lead to optimal or near-optimal posterior contraction rates if this smoothness matches that of the target function, but to suboptimal rates otherwise, although some rate and hence consistency is always attained. One way of overcoming this limitation is to rescale the sample paths and use a process of the form $W^a = (W_{at} : t \in \mathbb{R}^d)$, for a given Gaussian process W. This is called changing the *length scale* of the prior, and is common in practice.

If the parameter a is chosen a fixed constant, then nothing much changes in the asymptotic setup. However, even though the scaling does not change the qualitative properties of the sample paths, a scaling that tends to zero or infinity with the number of observations may

change the contraction rates dramatically. For $a < 1$ this entails shrinking a process on a bigger time set to the time set $[0,1]^d$, whereas $a > 1$ corresponds to stretching. Intuitively shrinking makes the sample paths more variable, as the randomness on a bigger time set is packed inside $[0,1]^d$, whereas stretching creates a smoother process.

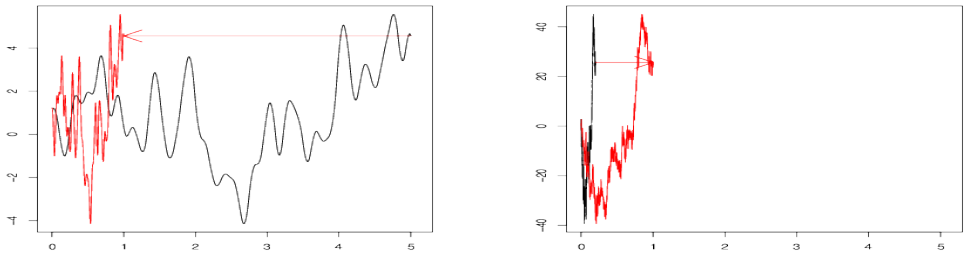

Figure 7.3 Changing the length scale of a process

The RKHS of a rescaled process is a rescaling of the RKHS of the original process.

Lemma 7.29 (RKHS)**.** *If the map $w \mapsto (t \mapsto w(at))$ is a continuous, linear map from the Banach spaces $\mathbb{B}_a$ into $\mathbb{B}$, and W is a random element in $\mathbb{B}_a$, then the RKHS of the process W^a given by $W^a_t = W_{at}$ consists of the functions $t \mapsto h(at)$ for h in the RKHS of W, with identical norms.*

The small ball probability is equally easily to obtain from the original one for self-similar processes. A stochastic process W is *self-similar* of order α if the processes $(W_{at}: 0 \leq t \leq 1)$ and $(a^\alpha W_t: 0 \leq t \leq 1)$ are equal in distribution. Thus the rescaling of the time-axis is equivalent to a rescaling of the vertical axis. This observation makes the following two lemmas evident.

Lemma 7.30 (RKHS)**.** *The RKHS $\mathbb{H}^a$ of the rescaled process W^a corresponding to a self-similar process W of order α is the RKHS $\mathbb{H}$ of W, but equipped with the norm $\|h\|_{\mathbb{H}^a} = a^{-\alpha}\|h\|_{\mathbb{H}}$.*

Lemma 7.31 (Small ball)**.** *The small ball exponent φ^a_0 of the rescaled process W^a corresponding to a self-similar process W of order α satisfies $\varphi^a_0(\epsilon) = \varphi_0(a^{-\alpha}\epsilon)$, for φ_0 the small ball exponent of W.*

Proofs The function $t \mapsto \mathrm{E}[W_{at}Z] = a^\alpha \mathrm{E}[W_t Z]$ is contained in both $\mathbb{H}^a$ and $\mathbb{H}$ and has square norm $\mathrm{E}[Z^2]$ in $\mathbb{H}^a$ and $a^{2\alpha}\mathrm{E}[Z^2]$ in $\mathbb{H}$. The second lemma is immediate from the fact that the events $\{\|W^a\| < \epsilon\}$ and $\{\|a^\alpha W\| < \epsilon\}$ are equal in probability. □

It follows that the concentration function $\varphi^a_{w_0}$ of the rescaled, self-similar process W^a can be written as

$$\varphi^a_{w_0}(\epsilon) = \varphi_0(a^{-\alpha}\epsilon) + a^{-2\alpha} \inf_{\|h - w_0\| \leq \epsilon} \|h\|^2_{\mathbb{H}}.$$

Increasing the scaling factor a makes the first term on the right side bigger, but decreases

the second term. We may now choose $a = a_n$ such that the solution ϵ_n of the rate equation $\varphi^{a_n}_{w_0}(\epsilon_n) \asymp n\epsilon_n^2$ is smallest. In the case that $\varphi_0(\epsilon) \asymp \epsilon^{-r}$ and $\inf\{\|h\|^2_{\mathbb{H}} \colon \|h - w_0\| \le \epsilon\} \asymp \epsilon^{-s}$, for some $r, s > 0$, the optimal scaling value and resulting contraction rate can be seen to be

$$a_n = n^{(s-r)/(4\alpha+4r\alpha+rs\alpha)}, \qquad \epsilon_n = n^{-(2+r)/(4+4r+rs)}.$$

It may be noted that the exponent of self-similarity appears in the scaling factor, but not in the contraction rate.

Rescaled integrated Brownian motion

The k-fold integrated Brownian motion is self-similar of order $k + 1/2$. Its small ball probability is of the order $\epsilon^{-1/(k+1/2)}$, and its decentering function for a function w_0 belonging to the Hölder space of order $\beta \le k + 1$ and having vanishing derivatives at 0 is of the order $\epsilon^{-(2k-2\beta+2)/\beta}$. Substitution of $\alpha = k + 1/2$, $r = 1/(k + 1/2)$ and $s = (2k - 2\beta + 2)/\beta$ in the preceding display yields the rescaling rate $a_n = n^{(k+1/2-\beta)/((k+1/2)(2\beta+1))}$ and the contraction rate $n^{-\beta/(2\beta+1)}$, for $\beta \le k + 1$. Thus the minimax rate is obtained for all $\beta \le k + 1$. For $\beta < k + 1/2$ the integrated Brownian motion is shrunk, and every possible smoothness level is attained. For $\beta > k + 1/2$ the process is stretched, but this is successful only up to smoothness level $k + 1$.

In order to drop the restriction that w_0 has vanishing derivatives at 0 we added a random polynomial of order k to the integrated Brownian motion. Because this polynomial process is not self-similar, the preceding argument does not apply to this extension. Actually rescaling the polynomial part in the same way as the integrated Brownian motion may also not be natural. Now it may be checked that the decentering function of the process $a_n^{k+1/2} I^k_{0+} B + b_n \sum_{i=1}^k Z_i t^i$ satisfies, for $w_0 \in \mathfrak{C}^\beta[0, 1]$ and $\beta \le k + 1$,

$$\inf_{\|h-w\|_\infty \le \epsilon} \|h\|^2_{\mathbb{H}_{a,b}} \lesssim a_n^{-(2k+1)} \epsilon^{-(2k-2\beta+2)/\beta} + b_n^{-2}[\epsilon^{-(2k-2\beta)/\beta} \vee 1].$$

This is dominated by the first term (arising from the integrated Brownian motion) if $b_n \ge a_n^{k+1/2} \epsilon_n^{((k+1-\beta)\wedge 1)/\beta}$. Since the small ball exponent of the process is hardly determined by the polynomial part, under the latter condition the preceding derivation goes through without essential changes for any w_0 in the Hölder space of order $\beta \le k + 1$.

Stationary Gaussian processes

Definition 7.32. A centered process $W = (W_t \colon t \in \mathbb{R}^d)$ with finite second moments is called (wide sense) *stationary* if its covariance function satisfies $\mathrm{E} W_s W_t = r(t-s)$ for some function $r \colon \mathbb{R}^d \to \mathbb{R}$.

For a centered Gaussian process stationarity is equivalent to the invariance of the distribution of the process $(W_t \colon t \in \mathbb{R}^d)$ under "time shifts": the finite-dimensional distributions of the processes $(W_{t+h} \colon t \in \mathbb{R}^d)$ are the same, for every $h \in \mathbb{R}^d$. This seems to be a reasonable property of a prior process.

Lemma 7.33. *The process W is a mean-square continuous stationary process indexed by $\mathbb{R}^d$ if and only if there exists a finite, symmetric Borel measure μ on $\mathbb{R}^d$ such that, for all $s, t \in \mathbb{R}^d$,*

$$\mathrm{E} W_s W_t = \int e^{i\lambda^T (s-t)} \, d\mu(\lambda). \tag{7.11}$$

Proof A process W with covariance function of the form (7.11) is clearly stationary. Moreover,

$$\mathrm{E}(W_t - W_s)^2 = \int |e^{i\lambda^T t} - e^{i\lambda^T s}|^2 \, \mu(d\lambda) \to 0,$$

as $t \to s$, by dominated convergence. Hence, the process is mean-square continuous.

Conversely, if the process W is stationary, and $\mathrm{E}W_s W_t = r(t-s)$, for all $s, t \in \mathbb{R}^d$, then, by the Cauchy-Schwarz inequality,

$$\big|r(t) - r(s)\big|^2 = \big|\mathrm{E}(W_t - W_s)W_0\big|^2 \le \mathrm{E}(W_t - W_s)^2 \mathrm{E}W_0^2.$$

Thus mean-square continuity of W implies that r is a continuous function. Since it is also positive-definite, existence of μ follows from Bochner's theorem (Theorem 7.51). □

Definition 7.34. The measure μ in (7.11) is called the *spectral measure* of the process W. If it admits a Lebesgue density, then this is called the *spectral density*.

The two sides of the spectral representation (7.11) are both inner products: on the left between the variables W_s and W_t in $L_2(\Pr)$, and on the right between the (complex-valued) functions e_s and e_t, defined by $e_t(\lambda) = \exp(i\lambda t)$, in $L_2(\mu)$. The continuous, linear extension of the association $W_t \leftrightarrow e_t$ provides an isometry between the first chaos $\overline{\mathrm{lin}}(W)$ of the process W and the closure of the linear span of the functions e_t in $L_2(\mu)$. This *spectral isometry* can often be used to translate probabilistic problems concerning the process W into analytical problems regarding the functions e_t in $L_2(\mu)$.

The weight of the tails of the spectral distribution determine the regularity of a stationary Gaussian process W. Heavier tails means "more high frequencies", and give sample paths that are less regular.

Proposition 7.35 (Regularity of stationary processes)**.** *Suppose that $(W_t : t \in \mathbb{R}^d)$ is a centered stationary Gaussian process with spectral measure μ.*

(i) *If $\int \|\lambda\|^{2\alpha} \, d\mu(\lambda) < \infty$, then W has a version whose sample paths are partially differentiable up to order the biggest integer k strictly smaller than α with partial derivatives of order k that are Lipschitz of order $a - k$, for any $a < \alpha$.*

(ii) *If $\int e^{c\|\lambda\|} \, d\mu(\lambda) < \infty$ for some $c > 0$, then W has a version with analytic sample paths.*

Proof By the dominated convergence theorem, for any multi-indices j and h with $\sum_l (j_l + h_l) \le 2k$,

$$D_s^j D_t^h r(s,t) = \int (i\lambda)^j (-i\lambda)^h e^{i\lambda^T(s-t)} \, d\mu(\lambda).$$

Furthermore, the right-hand side is Lipschitz in (s,t) of order $\alpha - k$. Thus (i) follows from Proposition 7.4.

For the proof of (ii) we note that the function $r(s,t) = \int e^{i\lambda^T(s-\bar{t})} \, d\mu(\lambda)$ is now also well defined for complex-valued s, t with absolute imaginary parts smaller than $c/2$ (and $\bar{t}$ the complex conjugate of t). The function is conjugate-symmetric and nonnegative-definite and hence defines a covariance function of a (complex-valued) stochastic process $(W_t : t \in T)$, indexed by $T = \{t \in \mathbb{C} : |\operatorname{Im} t| < c/2\}$. By an extension of Proposition 7.4 this can be seen to have sample paths with continuous partial derivatives, which satisfy the Cauchy-Riemann equations, and hence the process is differentiable on its complex domain. □

The RKHS of a stationary process can be described in spectral terms. As before, we define $e_t\colon \mathbb{R}^d \to \mathbb{C}$ by $e_t(\lambda) = \exp(i\lambda^T t)$.

Lemma 7.36. *The RKHS of a centered, continuous, stationary Gaussian process with $W = (W_t\colon t \in [0,1]^d)$ spectral measure μ is the set of (real parts of) the functions (from $[0,1]^d$ to $\mathbb{C}$)*

$$t \mapsto \int e^{i\lambda^T t}\psi(\lambda)\,\mu(d\lambda),$$

where ψ runs through the complex Hilbert space $L_2(\mu)$. The RKHS-norm of the displayed function equals the norm in $L_2(\mu)$ of the projection of ψ onto the closed linear span of the set of functions $\{e_t\colon t \in [0,1]^d\}$ (or, equivalently, the infimum of $\|\psi\|_{L_2(\mu)}$ over all functions ψ giving the same function in the preceding display).

Proof By the spectral isometry (7.11), the first chaos $\overline{\mathrm{lin}}(W)$ of W is isometric to the space of functions $\mathscr{L}' \subset L_2(\mu)$ defined as the closure in $L_2(\mu)$ of the linear span of the functions $\{e_t\colon t \in [0,1]^d\}$. Since every element of $\mathbb{H}$ is of the form $t \mapsto \mathrm{E}W_t L$ for $L \in \mathscr{L}$, using the spectral isometry again shows that every element of $\mathbb{H}$ is of the form $t \mapsto \langle e_t, \psi\rangle_{2,\mu}$, for $\psi \in \mathscr{L}'$, and the RKHS-norm of such a function is given by the $L_2(\mu)$-norm of ψ.

Now let $P\colon L_2(\mu) \to L_2(\mu)$ be the orthogonal projection onto the closed subspace $\mathscr{L}'$. Then for every $\psi \in L_2(\mu)$, $\langle e_t, \psi\rangle_{2,\mu} = \langle Pe_t, \psi\rangle_{2,\mu} = \langle e_t, P\psi\rangle_{2,\mu}$. Hence $t \mapsto \langle e_t, \psi\rangle_{2,\mu}$ belongs to $\mathbb{H}$ and its RKHS norm is given by the $L_2(\mu)$-norm of the projection $P\psi$. □

Matérn process

For $d \in \mathbb{N}$ and $\alpha > 0$, the *Matérn process* on $\mathbb{R}^d$ with parameter α is defined as the centered stationary process with spectral density

$$\lambda \mapsto \frac{1}{(1 + \|\lambda\|^2)^{\alpha + d/2}}.$$

The parameter α describes the regularity of the process. For k the largest integer strictly smaller than α, the spectral measure of the Matérn process W has a finite moment of order $2k$, hence it is k times differentiable in mean-square sense.

Example 7.37. The *Ornstein-Uhlenbeck process* is the special instance of the Matérn process, corresponding to the choices $d = 1$ and $\alpha = 1/2$. Its covariance function can be explicitly computed as

$$\mathrm{E}W_s W_t = \frac{\sigma^2}{2\theta} e^{-\theta|t-s|}.$$

This process also has a representation as an integral relative to Brownian motion B, of the form $W_t = \sigma \int_{-\infty}^t e^{-\theta(t-s)}\,dB_s$. This shows that its sample paths have the same regularity as those of Brownian motion.

Lemma 7.38 (Small ball)**.** *The small ball exponent of the Matérn process W viewed as a map in $\mathfrak{C}[0,1]$ satisfies, as $\epsilon \downarrow 0$,*

$$\varphi_0(\epsilon) \lesssim \epsilon^{-d/\alpha}.$$

Proof The Fourier transform of $H\psi$ is, up to a constant, the function $\phi = \psi m$, if $d\mu(\lambda) = m\,d\lambda$. For ψ the choice of minimal norm in the definition of $H\psi$, this function satisfies

$$\int \bigl|\phi(\lambda)\bigr|^2 (1 + \|\lambda\|^2)^{\alpha+d/2}\,d\lambda = \|H\psi\|_{\mathbb{H}}^2.$$

In other words, the unit ball $\mathbb{H}_1$ of the RKHS is contained in a Sobolev ball of order $\alpha + d/2$. The metric entropy relative to the uniform norm of such a Sobolev ball is bounded by a constant times $\epsilon^{-d/(\alpha+d/2)}$, by Proposition 5.26. The lemma next follows from Lemma 7.12, which characterizes the small ball probability in terms of the entropy of the RKHS-unit ball. □

To estimate the infimum in the definition of the concentration function φ_{w_0} for a nonzero response function w_0, we approximate w_0 by elements of the RKHS. The idea is to write w_0 in terms of its Fourier inverse $\hat{w}_0$ as, with $d\mu(\lambda) = m(\lambda)\,d\lambda$,

$$w_0(x) = \int e^{i\lambda^T x}\hat{w}_0(\lambda)\,d\lambda = \int e^{i\lambda^T x}\frac{\hat{w}_0}{m}(\lambda)\,d\mu(\lambda). \tag{7.12}$$

If $\hat{w}_0/m$ were contained in $\mathbb{L}_2(\mu)$, then w_0 would be contained in the RKHS, with RKHS-norm bounded by the $\mathbb{L}_2(\mu)$-norm of $\hat{w}_0/m$, i.e. the square root of $\int(|\hat{w}_0|^2/m)(\lambda)\,d\lambda$. In general this integral may be infinite, but we can remedy this by truncating the tails of $\hat{w}_0/m$.

A natural a priori condition on the true response function $w_0\colon [0,1]^d \to \mathbb{R}$ is that this function is contained in a Sobolev space of order β. The *Sobolev space* $\mathfrak{W}^\alpha[0,1]^d$ is here defined as the set of functions $w\colon [0,1]^d \to \mathbb{R}$ that are restrictions of a function $w\colon \mathbb{R}^d \to \mathbb{R}$ with Fourier transform $\hat{w}(\lambda) = (2\pi)^{-d}\int e^{i\lambda^T t}w(t)\,dt$ such that

$$\|w\|_{2,2,\alpha}^2 := \int (1 + \|\lambda\|^2)^\alpha \bigl|\hat{w}(\lambda)\bigr|^2\,d\lambda < \infty.$$

Roughly speaking, for integer α, a function belongs to $\mathfrak{W}^\alpha([0,1]^d)$ if it has partial derivatives up to order α that are all square integrable. This follows, because the αth derivative of a function w has Fourier transform $\lambda \mapsto (i\lambda)^\alpha\hat{w}(\lambda)$,

Lemma 7.39 (Decentering)**.** *If $w_0 \in \mathfrak{C}^\beta([0,1]^d) \cap \mathfrak{W}^\beta([0,1]^d)$ for $\beta \le \alpha$, then the decentering function of the Matérn process satisfies, for $\epsilon < 1$,*

$$\inf_{h:\|h-w_0\|_\infty<\epsilon} \|h\|_{\mathbb{H}}^2 \lesssim \epsilon^{-(2\alpha+d-2\beta)/\beta}.$$

Proof Let $\kappa\colon \mathbb{R} \to \mathbb{R}$ be a function with a real, symmetric Fourier transform $\hat{\kappa}$, which equals $1/(2\pi)$ in a neighborhood of 0 and which has compact support. From $\hat{\kappa}(\lambda) = (2\pi)^{-1}\int e^{i\lambda t}\kappa(t)\,dt$ it then follows that $\int \kappa(t)\,dt = 1$ and $\int (it)^k\kappa(t)\,dt = 0$ for $k \ge 1$. For $t = (t_1, \ldots, t_d)$, define $\phi(t) = \kappa(t_1)\cdots\kappa(t_d)$. Then ϕ integrates to 1, has finite absolute moments of all orders, and vanishing moments of all orders bigger than 0.

For $\sigma > 0$ set $\phi_\sigma(x) = \sigma^{-d}\phi(x/\sigma)$ and $h = \phi_\sigma * w_0$. Because ϕ is a higher order kernel, standard arguments from the theory of kernel estimation show that $\|w_0 - \phi_\sigma * w_0\|_\infty \lesssim \sigma^\beta$.

The Fourier transform of h is the function $\lambda \mapsto \hat{h}(\lambda) = \hat{\phi}(\sigma\lambda)\hat{w}_0(\lambda)$, and therefore, by

(7.12),

$$\|h\|_{\mathbb{H}}^2 \lesssim \int \left|\hat{\phi}(\sigma\lambda)\hat{w}_0(\lambda)\right|^2 \frac{1}{m(\lambda)}\, d\lambda \lesssim \sup_\lambda \left[(1+\|\lambda\|^2)^{\alpha+d/2-\beta}\left|\hat{\phi}(\sigma\lambda)\right|^2\right] \|w_0\|_{2,2,\beta}^2$$
$$\lesssim C(\sigma) \sup_\lambda \left[(1+\|\lambda\|^2)^{\alpha+d/2-\beta}\left|\hat{\phi}(\lambda)\right|^2\right] \|w_0\|_{2,2,\beta}^2,$$

for

$$C(\sigma) = \sup_\lambda \Big(\frac{1+\|\lambda\|^2}{1+\|\sigma\lambda\|^2}\Big)^{\alpha+d/2-\beta} \lesssim \Big(\frac{1}{\sigma}\Big)^{2\alpha+d-2\beta},$$

if $\sigma \le 1$. The assertion of the lemma follows upon choosing $\sigma \sim \epsilon^{1/\beta}$. □

It follows that the rate equation $\varphi_{w_0}(\epsilon_n) \le n\epsilon_n^2$ has the minimal solution $\epsilon_n \asymp n^{-\beta/(2\alpha+d)}$, for $\beta \le \alpha$. Thus again the rate is minimax if and only if prior and smoothness match.

Squared exponential process

The *squared exponential process* is the zero-mean Gaussian process with covariance function

$$\mathrm{E} W_s W_t = e^{-\|s-t\|^2}, \qquad s,t \in \mathbb{R}^d.$$

Its spectral measure can be found using the fact that the Fourier transform of the Gaussian density is Gaussian. Specifically, the spectral measure has density

$$\lambda \mapsto \frac{1}{2^d \pi^{d/2}} e^{-\|\lambda\|^2/4}.$$

The sample paths of the squared exponential process are analytic. This makes the centered small ball probability of this process much larger than that of the processes considered so far: it is nearly "parametric."

Lemma 7.40 (Small ball)**.** *The small ball exponent of the square exponential process viewed as a map in* $\mathfrak{C}([0,1]^d)$ *satisfies, for a constant* C *depending only on* d*, as* $\epsilon \downarrow 0$*,*

$$\varphi_0(\epsilon) \le C(\log_- \epsilon)^{1+d/2}.$$

Lemma 7.41 (Decentering)**.** *If* $w_0 \in \mathfrak{W}^\beta([0,1]^d)$ *for* $\beta > d/2$*, then the concentration function of the square exponential process satisfies, for* $\epsilon < 1$*, a constant* C *that depends only on* w_0*,*

$$\inf_{h:\|h-w_0\|_\infty<\epsilon} \|h\|_{\mathbb{H}}^2 \lesssim \exp(C\epsilon^{-2/(\beta-d/2)}).$$

Proofs Every function $H\psi$ in the (complex) RKHS as described in Lemma 7.36 can be extended to an entire function $H\psi\colon \mathbb{C} \to \mathbb{C}$ defined by $H\psi(z) = \int e^{\langle \lambda, z\rangle}\psi(\lambda)\, d\mu(\lambda)$. If the function $H\psi$ is contained in the unit ball of the RKHS, then $\|\psi\|_{2,\mu} \le 1$, and an application of the Cauchy-Schwarz inequality gives that $|H\psi(z)|^2 \le \int e^{2\|\lambda\|\,|z|}\, d\mu(\lambda) \le e^{2C|z|^2}$, for some universal constant C. In particular, the functions $H\psi$ can be extended to analytic functions on the strip $\{z \in \mathbb{C}\colon \|z\|_\infty \le A\}$ that are uniformly bounded by e^{CA^2}, for any A. It follows by Proposition 5.27 that $N(\epsilon, \mathbb{H}_1, \|\cdot\|_\infty) \le A^{-d}\log(e^{CA^2}/\epsilon)^{1+d}$. Choosing A of the order $\log_-\epsilon$ leads to the bound $N(\epsilon, \mathbb{H}_1, \|\cdot\|_\infty) \le (\log_-\epsilon)^{1+d/2}$.

The first lemma now follows from the characterization of the small ball exponent by the entropy of the RKHS unit ball, Lemma 7.12.

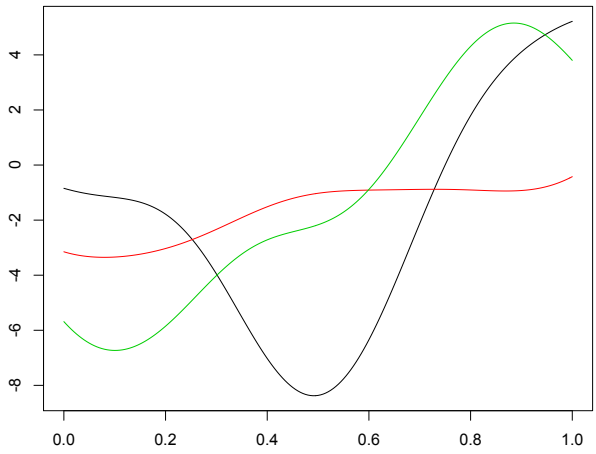

Figure 7.4 Three realizations of the squared exponential process.

For given $K > 0$ let $\psi(\lambda) = (\hat{w}_0/m)(\lambda)\mathbb{1}_{\|\lambda\| \le K}$, for m the density in (7.6). The function $H\psi$ satisfies

$$\begin{aligned}\|H\psi - w_0\|_\infty &\le \int_{\|\lambda\|>K} |\hat{w}_0(\lambda)|\, d\lambda \le \|w_0\| \Big[\int_{\|\lambda\|>K} (1+\|\lambda\|^2)^{-\beta}\, d\lambda\Big]^{1/2} \\ &\lesssim \|w_0\|_{2,2,\beta} K^{-(\beta-d/2)}.\end{aligned}$$

Furthermore, the squared RKHS-norm of $H\psi$ is given by

$$\|H\psi\|_{\mathbb{H}}^2 = \int_{\|\lambda\|\le K} \frac{|\hat{w}_0|^2}{m}(\lambda)\, d\lambda \le \sup_{\|\lambda\|\le K}\Big[m(\lambda)^{-1}(1+\|\lambda\|^2)^{-\beta}\Big] \|w_0\|_{2,2,\beta}^2 \lesssim e^{K^2/4}\|w_0\|_{2,2,\beta}^2.$$

We conclude the proof by choosing $K \asymp \epsilon^{-1/(\beta-d/2)}$. □

Combining the preceding we see that the concentration function of the square exponential process satisfies

$$\varphi_{w_0}(\epsilon) \lesssim \exp(C\epsilon^{-2/(\beta-d/2)}) + (\log_- \epsilon)^{1+d/2}.$$

The first term (decentering function) dominates the second (centered small ball exponent) for any $\beta > 0$, and the contraction rate for a β-smooth function satisfies $\epsilon_n \asymp (\log n)^{-(\beta/2-d/4)}$. This extremely slow rate is the result of the discrepancy between the infinite smoothness of the prior and the finite smoothness of the true parameter. A remedy for this mismatch is to rescale the sample paths and is discussed in Section 7.7.

7.7 Mixtures of Gaussian processes

In the preceding it was seen that changing the length scale of a Gaussian prior may adapt this prior better to a true target function of a given regularity. Because the latter true regularity is usually not known, it is attractive to change the length scale in a way that can adapt to the data. In the Bayesian setup the most attractive method is to choose the length scale itself from a prior. The resulting prior will not be Gaussian, but a mixture of Gaussians.

Consider a stationary Gaussian process $W = (W_t : t \in \mathbb{R}^d)$ with spectral density with exponentially small tails, so that the process has analytic sample paths. Rather than scaling the process deterministically, consider the process $W^A = (W_{At} : t \in [0,1]^d)$, for A a random variable independent of W. The resulting prior is a mixture of Gaussian processes. Specifically we study the case that the variable A^d follows a gamma distribution. The parameters of this gamma distribution are inessential, and the gamma distribution can be replaced by another distribution with tails of the same weights, but the power d in A^d appears important.

The following theorem gives the contraction rate for the prior process W^A in the abstract setting of a mixed Gaussian prior, and may be compared with Theorem 7.13. The theorem can be translated to contraction rates in concrete settings, such as density estimation, classification, and regression, in the same way as Theorems 7.14–7.16 are derived from Theorem 7.13. The theorem gives the existence of sets B_n and rates ϵ_n and $\bar{\epsilon}_n$ such that

$$\log N(\epsilon_n, B_n, \|\cdot\|) \le n\epsilon_n^2, \tag{7.13}$$

$$\Pr(W^A \notin B_n) \le e^{-4n\bar{\epsilon}_n^2}, \tag{7.14}$$

$$\Pr(\|W^A - w_0\| \le \bar{\epsilon}_n) \ge e^{-n\bar{\epsilon}_n^2}. \tag{7.15}$$

The rates are specified for two situations: the ordinary smooth case, where the true function w_0 belongs to a Hölder class, and the super-smooth case, where w_0 is analytic. The following theorem shows that choosing the length scale according to dth root of a gamma variable, leads to near optimal rate of contraction in both worlds.

Theorem 7.42 (Mixed Gaussian contraction rate)**.** *If $(W_t : t \in \mathbb{R}^d)$ is a stationary Gaussian process with spectral density m such that $a \mapsto m(a\lambda)$ is decreasing on $(0,\infty)$, for every $\lambda \in \mathbb{R}^d$, and $\int e^{\delta\|\lambda\|} m(\lambda)\, d\lambda < \infty$, for some $\delta > 0$, and A^d is an independent gamma variable, then there exist measurable sets $B_n \subset \mathfrak{C}([0,1]^d)$ such that (7.13)–(7.15) hold for the process $W_t^A = W_{At}$ and ϵ_n and $\bar{\epsilon}_n$ defined as follows.*

(i) *If $w_0 \in \mathfrak{C}^\beta([0,1]^d)$, then $\bar{\epsilon}_n \asymp n^{-\beta/(2\beta+d)}(\log n)^{(1+d)\beta/(2\beta+d)}$ and $\epsilon_n \asymp \bar{\epsilon}_n(\log n)^{(1+d)/2}$.*

(ii) *If w_0 is the restriction of a function in $\mathcal{A}^{\gamma,r}(\mathbb{R}^d)$ to $[0,1]^d$ and $m(\lambda) \ge C_3 \exp(-D_3\|\lambda\|^\gamma)$ for constants $C_3, D_3, \gamma > 0$, then $\bar{\epsilon}_n \asymp n^{-1/2}(\log n)^{(d+1)/2}$ and $\epsilon_n \asymp \bar{\epsilon}_n(\log n)^\kappa$, where $\kappa = 0$ for $r \ge \gamma$, and $\kappa = d/(2r)$ for $r < \gamma$.*

7.8 Complements

7.8.1 Random elements in the space of uniformly continuous functions

Let T be a totally bounded metric space, with metric denoted by ρ, and assume that W is a Gaussian process with uniformly continuous sample paths.

The set $\mathfrak{UC}(T,\rho)$ of uniformly continuous functions $z : T \to \mathbb{R}$ is a separable Banach space under

the uniform norm $\|\cdot\|$. A stochastic process $(W_t\colon t \in T)$ with uniformly continuous sample paths can be viewed as a map $W\colon \Omega \to \mathfrak{UC}(T,\rho)$. By the next lemma this map is automatically measurable relative to the Borel σ-field on $\mathfrak{UC}(T,\rho)$.

Lemma 7.43. *Let (T,ρ) be a totally bounded metric space and W a stochastic process whose sample paths $t \mapsto W_t(\omega)$ belong to $\mathfrak{UC}(T,\rho)$. Then the map $W\colon (\Omega,\mathscr{U}) \to \mathfrak{UC}(T,\rho)$ is measurable relative to the Borel σ-field relative to the uniform norm on $\mathfrak{UC}(T,\rho)$.*

Proof Because the space $\mathfrak{UC}(T,\rho)$ is separable under the uniform norm, every open set is a countable union of open, and then also of closed, balls. (Take the balls centered at the points of a countable dense set with rational radius.) Therefore it suffices to show that the inverse image under W of an arbitrary closed ball is measurable. For a countable, dense subset $D \subset T$, continuity of the sample paths of W implies that

$$\bigl\{\omega\colon \bigl\|W(\omega)\bigr\| \le a\bigr\} = \bigcap_{t\in D}\bigl\{\omega\colon |W_t(\omega)| \le a\bigr\}.$$

Every set appearing on the right is in $\mathscr{U}$, since every W_t is a random variable. Since the intersection is countable, it follows that the left-hand side belongs to $\mathscr{U}$ as well. □

The coordinates W_t of a Borel measurable map $W\colon \Omega \to \mathfrak{UC}(T,\rho)$ are clearly random variables, by the continuity and hence measurability of the projections $z \mapsto z(t)$. Thus the converse of the preceding lemma is true as well, so that stochastic processes with uniformly continuous sample paths and random elements in $\mathfrak{UC}(T,\rho)$ are identical objects.

7.8.2 Regular versions of stochastic processes

A minimal regularity property of a stochastic process is *separability*. Roughly speaking, the behavior of a separable process over the whole index set is determined by its behavior on a countable subset.

Definition 7.44. Let $(X_t\colon t \in T)$ be a stochastic process indexed by a topological space T, with state space $(E,\mathcal{E})$, where E is a topological space and $\mathcal{E}$ is its Borel σ-field. The process is called *separable* if there exists an event N with $\Pr(N) = 0$ and a countable set $S \subset T$ such that for all open $U \subset T$ and closed $F \subset E$, the subsets $\bigcap_{t\in U}\{X_t \in F\}$ and $\bigcap_{t\in S\cap U}\{X_t \in F\}$ of the underlying outcome space differ by at most a subset of N. Any countable set S with the stated property is called a *separability set*.

The definition immediately implies that for a separable process X indexed by T and defined on a complete probability space, the set $\bigcap_{t\in U}\{X_t \in F\}$ is a measurable event for every open $U \subset T$ and closed $F \subset E$. If the process is real-valued we have in particular that for every $b \in \mathbb{R}$,

$$\Bigl\{\sup_{t\in T} X_t \le b\Bigr\} = \bigcap_{t\in T}\{X_t \le b\}$$

is measurable, and hence $\sup_{t\in T} X_t$ is a well-defined random variable. Moreover, it is a.s. equal to $\sup_{t\in T\cap S} X_t$. Similarly, the variables $\inf X_t$, $\sup|X_t|$ and $\inf|X_t|$ are measurable as well.

Another useful consequence of separability is that to show that a real-valued separable process X vanishes identically with probability one, it suffices to show that $X_t = 0$ a.s. for every t in a separability set S. Indeed, suppose that $X_t = 0$ a.s., for all $t \in S$. Then $\bigcap_{t\in S}\{X_t = 0\}$ has probability one. By separability the event $\bigcap_{t\in T}\{X_t = 0\}$ is measurable and has probability one as well, i.e. $\Pr(X_t = 0 \text{ for all } t \in T) = 1$.

In addition to separability it is useful to know whether a stochastic process X is measurable as function of the pair (ω,t). By Fubini's theorem this implies for instance that the sample paths of the process are measurable functions.

Definition 7.45. Let X be a real-valued process indexed by a metric space (T, d), defined on $(\Omega, \mathscr{U}, \Pr)$. Let $\mathbb{B}(T)$ be the Borel σ-field of T. The process X is called *(Borel) measurable*, if the map from $(\Omega \times T, \mathscr{U} \times \mathbb{B}(T))$ to $(\mathbb{R}, \mathbb{B}(\mathbb{R}))$ given by $(\omega, t) \mapsto X_t(\omega)$ is measurable.

A process X indexed by a metric space (T, d) is called *continuous in probability* if for all $s \in T$, $X_t \to X_s$ in probability if $d(t, s) \to 0$. The following theorem says that a process admits a measurable and separable modification if it is continuous in probability. Note that this property only depends on the fdds of the process X.

Theorem 7.46. *Let X be a real-valued process indexed by a separable metric space (T, d). If the process is continuous in probability, it admits a Borel measurable, separable version, which may take the value ∞. Any countable, dense subset of (T, d) is a separability set.*

The *product σ-field* on $\mathbb{R}^T$ is the smallest σ-field containing all sets of the form $\{x \in \mathbb{R}^T : x_t \in B\}$, for a Borel set $B \subset \mathbb{R}$ and $t \in T$.

Proposition 7.47. *A map $W : (\Omega, \mathscr{U}) \to \mathbb{R}^T$ is measurable relative to the product σ-field if and only if all coordinate maps $W_t : (\Omega, \mathscr{U}) \to \mathbb{R}$ are measurable.*

Proposition 7.48 (Kolmogorov's continuity criterion)**.** *If X is a real-valued process indexed by a compact interval $T \subset \mathbb{R}$ such that there exist constants $p, q, C > 0$ with $\mathrm{E}|X_t - X_s|^p \le C|t - s|^{1+q}$, for every $s, t \in T$, then X admits a continuous modification with sample paths that are almost surely Hölder continuous of order α for every $\alpha < q/p$.*

7.8.3 Wiener integrals

One way to construct new processes from a Brownian motion W is to integrate functions relative to W, that is, to consider integrals of the form $\int f\,dW$. However, since the sample paths of W are very rough, these integrals can not be defined path-wise in the ordinary Lebesgue-Stieltjes sense. The way out is to define them via a Hilbert space isometry. In general this leads to integrals that are not defined path-wise, but only in an L_2-sense.

To have more flexibility we define integrals with respect to a *two-sided* Brownian motion. Let W^1 and W^2 be two independent Brownian motions. Construct a two-sided Brownian motion $W = (W_t : t \in \mathbb{R})$, emanating from 0, by setting

$$W_t = \begin{cases} W^1_t & \text{if } t \ge 0, \\ W^2_{-t} & \text{if } t < 0. \end{cases}$$

For real numbers $t_0 < \cdots < t_n$ and $a_1, \ldots, a_n$, consider the simple function $f = \sum a_k 1_{(t_{k-1}, t_k]}$. We define the "integral" of f relative to W in the obvious way by setting

$$\int f\,dW = \sum a_k (W_{t_k} - W_{t_{k-1}}).$$

Using the basic properties of the Brownian motion it is straightforward to verify that for two simple functions f, g, we have

$$\mathrm{E}\Big(\int f\,dW\Big)\Big(\int g\,dW\Big) = \int_{\mathbb{R}} f(x)g(x)\,dx. \tag{7.16}$$

In other words, the linear map $f \mapsto \int f\,dW$ is an isometry from the collection of simple functions in $L_2(\mathbb{R})$ into $L_2(\Pr)$. Since the simple functions are dense in $L_2(\mathbb{R})$, the map can be extended to the whole space $L_2(\mathbb{R})$. This defines $\int f\,dW$ for all $f \in L_2(\mathbb{R})$.

Note that by construction the integral is almost surely unique. It is a centered Gaussian random variable and the isometry relation (7.16) holds for all $f, g \in L_2(\mathbb{R})$.

Definition 7.49. We call $\int f\,dW$ the *Wiener integral* of f relative to W. If $f \in L_2(\mathbb{R})$ and $t \geq s$, we write $\int_s^t f(u)\,dW_u$ for $\int 1_{(s,t]} f\,dW$.

Under appropriate conditions, some of the usual calculus rules still hold for the Wiener integral, in particular a version of Fubini's theorem and the integration by parts formula. Recall that we say that $f\colon [s,t] \to \mathbb{R}$ is of bounded variation if

$$\operatorname{var} f = \sup \sum |f(t_k) - f(t_{k-1})|$$

is finite, where the supremum is over all finite partitions of $[s,t]$. Note that such a function is necessarily square integrable on $[s,t]$.

Proposition 7.50. *(i) (Fubini for Wiener integrals) Let (S, Σ, μ) be a finite measure space and $f \in L_2(\mathrm{Leb} \times \mu)$. Then it almost surely holds that*

$$\int \Big(\int f(u,v)\,dW_u \Big) \mu(dv) = \int \Big(\int f(u,v)\,\mu(dv) \Big) dW_u.$$

(ii) (Integration by parts) If $t \geq s$ and $f\colon [s,t] \to \mathbb{R}$ is of bounded variation, then

$$\int_s^t f(u)\,dW_u = W_t f(t) - W_s f(s) - \int_s^t W_u\,df(u)$$

almost surely.

Proof (i). If f is a simple function of the form $f = \sum a_i 1_{I_i \times E_i}$, for real numbers a_i, intervals I_i and $E_i \in \Sigma$, the statement is trivially true. For a general $f \in L_2(\mathrm{Leb} \times \mu)$, there exists a sequence of simple f_n of the form just described such that $f_n \to f$ in $L_2(\mathrm{Leb} \times \mu)$. Then by Jensen's inequality,

$$\int \Big(\int f_n(u,v)\,\mu(dv) - \int f(u,v)\,\mu(dv) \Big)^2 du \leq \|f_n - f\|_{L_2}^2 \to 0.$$

Hence, by definition of the Wiener integral,

$$\int \Big(\int f_n(u,v)\,\mu(dv) \Big) dW_u \to \int \Big(\int f(u,v)\,\mu(dv) \Big) dW_u$$

in $L_2(\mathrm{Pr})$. On the other hand, the convergence $\|f_n - f\|_{L_2}^2 \to 0$ implies that there exists a subsequence n' and a set $S' \subset S$ of full μ-measure such that

$$\int (f_{n'}(u,v) - f(u,v))^2\,du \to 0$$

for all $v \in S'$. Again by definition of the Wiener integral it follows that for $v \in S'$,

$$\int f_{n'}(u,v)\,dW_u \to \int f(u,v)\,dW_u$$

in $L_2(\mathrm{Pr})$. First, this implies that there is a further subsequence along the convergence takes place almost surely. Hence, since the left-hand side is a measurable function of v, so is the right-hand side. Second, by Jensen and the ordinary Fubini theorem we have

$$\mathrm{E}\Big(\int \Big(\int f_{n'}(u,v)\,dW_u \Big) \mu(dv) - \int \Big(\int f(u,v)\,dW_u \Big) \mu(dv) \Big)^2 \to 0.$$

(ii). The function f can be written as the difference of two non-decreasing, cadlag functions on $[s,t]$.

Hence it suffices to prove the statement under the assumption that f itself is such a non-decreasing function, so that df is an ordinary Lebesgue-Stieltjes measure. By (i), we then have, a.s.,

$$\begin{aligned}\int_s^t W_u\, df(u) &= \int_s^t \Big(\int_0^s dW_v \Big) df(u) + \int_s^t \Big(\int_s^u dW_v \Big) df(u) \\ &= (f(t) - f(s))W_s + \int_s^t (f(t) - f(u))\, dW_u.\end{aligned}$$

Rearranging gives the equality we have to prove. □

7.8.4 Bochner's theorem

A function $f\colon \mathbb{R}^d \to \mathbb{R}$ is called *positive definite* if for all $a_1, \dots, a_n \in \mathbb{R}$ and $t_1, \dots, t_n \in \mathbb{R}^d$,

$$\sum \sum a_i a_j f(t_i - t_j) \geq 0.$$

Bochner's theorem asserts that among the continuous functions on $\mathbb{R}^d$, the positive definite ones are precisely the Fourier transforms of finite measures.

Theorem 7.51 (Bochner)**.** *A continuous function f on $\mathbb{R}^d$ is positive definite if and only there exists a finite Borel measure μ such that*

$$f(t) = \int e^{i\lambda^T t}\, \mu(d\lambda), \quad t \in \mathbb{R}^d.$$

A collection of functions $\mathscr{F} \subset \mathfrak{C}[0,1]^d$ is called *uniformly bounded* if $\sup_{f\in\mathscr{F}} \|f\|_\infty < \infty$. It is called *uniformly equicontinuous* is for all $\epsilon > 0$, there exists a $\delta > 0$ such that $\|s - t\| < \delta$ implies that $|f(s) - f(t)| \leq \epsilon$ for all $s, t \in [0,1]^d$ and $f \in \mathscr{F}$.

Theorem 7.52 (Arzelà-Ascoli)**.** *The set $\mathscr{F} \subset \mathfrak{C}[0,1]^d$ is pre-compact if and only if it is uniformly bounded and uniformly equicontinuous.*

Theorem 7.53. *(Hahn-Banach) Let W be a linear subspace of a normed linear space V. If every bounded linear functional on V that vanishes on W, vanishes on the whole space V, then W is dense in V.*

Lemma 7.54 (Cameron-Martin)**.** *The measures P^W and P^{W+h} are equivalent Borel measures if and only if $h \in \mathbb{H}$. In this case a Radon-Nikodym density given by*

$$\frac{dP^{W+h}}{dP^W}(W) = e^{Uh - \frac{1}{2}\|h\|_{\mathbb{H}}^2}.$$

Sketch of proof The process W can be written as $W = \sum Z_i h_i$, for $Z_i = Uh_i$ independent, standard normal variables and the h_i an orthonormal basis of the RKHS $\mathbb{H}$. The convergence of the series takes place in $\mathfrak{C}[0,1]^d$ almost surely. The function $h \in \mathbb{H}$ admits an expansion $h = \sum c_i h_i$ for some $c \in \ell^2$. The series converges in $\mathbb{H}$, but it converges in $\mathfrak{C}[0,1]^d$ as well, as the RKHS-norm is stronger. We can thus write $W + h = \sum (Z_i + c_i)h_i$, convergence taking place in $\mathfrak{C}[0,1]^d$.

It can be proved that W and $W + h$ are measurable functions of the sequences Z and $Z + c$, respectively. This implies that to prove the equivalence of the laws P^W and P^{W+h} it suffices to show that the laws of the sequences Z and $Z + c$ are equivalent measures on the sequence space $\mathbb{R}^\infty$. Now for a fixed i, the squared Hellinger distance (see (5.3)) between the laws of Z_i and $Z_i + c_i$ equals

$$1 - e^{-\frac{1}{8}c_i^2} \leq \tfrac{1}{8}c_i^2.$$

Since $c \in \ell^2$, Theorem 7.55 yields the equivalence of the laws.

The ratio of the densities of Z_i and $Z_i + c_i$ at the point z_i is given by $\exp((z_i c_i - c_i^2/2))$. Therefore, the Radon-Nikodym derivative of the law of $Z + c$ relative to the law of Z at the point $Z = (Z_1, Z_2, \ldots)$ is given by

$$\prod_{i=1}^{\infty} e^{c_i Z_i - \frac{1}{2}c_i^2} = e^{\sum c_i Z_i - \frac{1}{2}\sum c_i^2}.$$

This completes the proof, since $Uh = \sum c_i Z_i$ and $\sum c_i^2 = \|h\|_{\mathbb{H}}^2$. □

In the following theorems, h denotes the Hellinger distance between densities (see (5.3)).

Theorem 7.55 (Kakutani)**.** *Let $X = (X_1, X_2, \ldots)$ and $Y = (Y_1, Y_2, \ldots)$ be two sequences of independent random variables. Assume X_i has a positive density f_i with respect to a dominating measure μ, and Y_i has a positive density g_i with respect to μ. Then the laws of the sequences X and Y are equivalent probability measures on $\mathbb{R}^\infty$ if and only if*

$$\sum_{i=1}^{\infty} h^2(f_i, g_i) < \infty.$$

If the laws are not equivalent, they are mutually singular.

7.8.5 Miscellaneous results

Lemma 7.56. *If the vector $(X, Y)^T$ possesses a multivariate normal distribution*

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \\ \nu \end{pmatrix}, \begin{pmatrix} R & C \\ C^T & S \end{pmatrix}\right),$$

with R nonsingular, then $Y|X \sim N(\nu + A(X - \mu), S - ARA^T)$, for $A = C^T R^{-1}$.

Proof We can reduce to the case $\mu = \nu = 0$ by considering the conditional distribution of $Y - \nu$ given $X - \mu$. Because $\mathrm{E}(Y - AX)X^T = C^T - AR = 0$, by the definition of A, the vector $(X, Y - AX)$ is multivariate normal with mean zero and $\mathrm{cov}(Y - AX, X) = 0$. Properties of the multivariate normal distribution imply that $Y - AX$ and X are independent. It follows that $Y = (Y - AX) + AX$ is conditionally given $X = x$ distributed as $(Y - AX) + Ax$, where $Y - AX$ follows its unconditional distribution, which is normal with covariance matrix $S - ARA^T$. □

Exercises

7.1 Derive the updating formulas for the mean and covariance function in the context of Section 7.3.

7.2 Prove that there exists a stochastic process W that satisfies conditions (i)–(iii) of Definition 7.19.

7.3 Prove that there exists a stochastic process W that satisfies conditions (i)–(iv) of Definition 7.19.

7.4 Verify (7.10).

7.5 Show that the Riemann-Liouville process R^α with parameter $\alpha > 0$ is *self-similar*: for $c > 0$, the process $(c^{-\alpha} R_{ct} : t \geq 0)$ is again a Riemann-Liouville process with parameter α.

7.6 Prove that the Ornstein-Uhlenbeck process admits a version that is locally Hölder continuous of every order strictly less than $1/2$.

7.7 Show that the Ornstein-Uhlenbeck process satisfies the integral equation

$$X_t - X_0 = -\theta \int_0^t X_s \, ds + \sigma W_t, \quad t \geq 0,$$

almost surely.

7.8 Verify the posterior computations in Section 7.3.

7.9 Prove that the first chaos of the Brownian motion $W = (W_t : t \in [0, 1])$ can be identified with the collection of Wiener integrals $\{\int_0^1 f(u)\, dW_u : f \in L_2[0, 1]\}$.

7.10 Prove that a Gaussian process with continuous sample paths is mean-square continuous.

7.11 Prove that the RKHS is a separable Hilbert space.

7.12 Determine the support of the Brownian motion indexed by $[0, 1]$.

7.13 Determine the RKHS of integrated Brownian motion.

7.14 Determine the RKHS and the RKHS-norm of the Brownian motion with standard normal initial distribution.

References

Bauer, H. (1981). *Probability theory and elements of measure theory*. Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], London-New York. Second edition of the translation by R. B. Burckel from the third German edition, Probability and Mathematical Statistics.

Berger, J. O., J. M. Bernardo, and D. Sun (2009). The formal definition of reference priors. *Ann. Statist. 37*(2), 905–938.

Bernardo, J.-M. (1979). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. Ser. B 41*(2), 113–147. With discussion.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Blei, D. M., A. Kucukelbir, and J. D. McAuliffe (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association 112*(518), 859–877.

Claeskens, G. and N. L. Hjort (2008). *Model selection and model averaging*, Volume 27 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

Clarke, B. S. and A. R. Barron (1994). Jeffreys' prior is asymptotically least favorable under entropy risk. *J. Statist. Plann. Inference 41*(1), 37–60.

Dudley, R. (2002). *Real Analysis and Probability*, Volume 74 of *Cambridge Studies in Advanced Mathematics*. Cambridge: Cambridge University Press. Revised reprint of the 1989 original.

Ghoshal, S. and A. van der Vaart (2017). *Fundamentals of nonparametric Bayesian inference*. Cambridge University Press.

Hills, S. E. and A. F. Smith (1992). Parameterization issues in bayesian inference. *Bayesian statistics 4*, 227–246.

Jara, A., T. Hanson, F. Quintana, P. Mueller, and G. Rosner (2015). Package dppackage.

Kleijn, B. J. K. and A. W. van der Vaart (2012). The Bernstein-Von-Mises theorem under misspecification. *Electron. J. Stat. 6*, 354–381.

Le Cam, L. (1986). *Asymptotic methods in statistical decision theory*. Springer Series in Statistics. Springer-Verlag, New York.

Liang, F., R. Paulo, G. Molina, M. A. Clyde, and J. O. Berger (2008). Mixtures of g priors for Bayesian variable selection. *J. Amer. Statist. Assoc. 103*(481), 410–423.

Livingstone, S., M. Betancourt, S. Byrne, and M. Girolami (2019). On the geometric ergodicity of Hamiltonian Monte Carlo. *Bernoulli 25*(4A), 3109–3138.

Mengersen, K. L. and R. L. Tweedie (1996, 02). Rates of convergence of the hastings and metropolis algorithms. *Ann. Statist. 24*(1), 101–121.

Meyn, S. P. and R. L. Tweedie (2012). *Markov chains and stochastic stability*. Springer Science & Business Media.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In *Handbook of Markov chain Monte Carlo*, Chapman & Hall/CRC Handb. Mod. Stat. Methods, pp. 113–162. CRC Press, Boca Raton, FL.

Robert, C. P. and G. Casella (2004). *Monte Carlo statistical methods* (Second ed.). Springer Texts in Statistics. Springer-Verlag, New York.

Roberts, G. O. and R. L. Tweedie (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli 2*(4), 341–363.

Rudin, W. (1973). *Functional Analysis*. New York: McGraw-Hill Book Co. McGraw-Hill Series in Higher Mathematics.

Strasser, H. (1985). *Mathematical theory of statistics*, Volume 7 of *De Gruyter Studies in Mathematics*. Walter de Gruyter & Co., Berlin. Statistical experiments and asymptotic decision theory.

van der Vaart, A. W. (1998). *Asymptotic statistics*, Volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

VanDerwerken, D. (2016). Not every gibbs sampler is a special case of the metropolis-hastings algorithm. *Communications in Statistics - Theory and Methods. To appear*.

Subject index