8

Contraction Rates: General Theory

A posterior contraction rate quantifies the speed at which a posterior distribution approaches the true parameter of the distribution of the data. Characterising a contraction rate can be seen as a significant refinement of establishing posterior consistency. It links nonparametric Bayesian analysis to the minimax theory of general statistical estimation. In this chapter we present rates of posterior contraction for dominated experiments. We begin with the general notion of a contraction rate, and its implications for Bayesian point estimation. We then present a basic rate theorem for i.i.d. observations, several refinements and a number of examples of priors, which illustrate optimal rates. Next we deal with general observations, with emphasis on independent, non-identically observations and Markov processes, and present a theory for misspecified models, illustrated by nonparametric regression under a misspecified error distribution. Finally we show that the α-posterior possesses a contraction rate that is solely determined by the prior concentration rate (for $\alpha < 1$), and conclude with an information-theoretic analysis.

8.1 Introduction

The contraction rate of a posterior distribution may be viewed as elaborating on its consistency. Both are asymptotic properties, and therefore we consider the same setup as in Chapter 6, of a sequence of statistical experiments $(\mathfrak{X}^{(n)}, \mathscr{X}^{(n)}, P_\theta^{(n)}: \theta \in \Theta_n)$ with parameter spaces Θ_n and observations $X^{(n)}$ indexed by a parameter $n \to \infty$. To measure a speed of contraction, the parameter spaces Θ_n must carry semimetrics d_n[1], unlike in Chapter 6, where a neighborhood system was enough for characterizing consistency. The parameter space and semimetric may depend on n, but we shall make this explicit only when needed, as in Section 8.3.

As in Chapter 6 we fix particular versions of the posterior distributions $\Pi_n(\cdot|\, X^{(n)})$ relative to a prior Π_n, which we refer to as “the” posterior distribution.

Definition 8.1 (Contraction rate) A sequence ϵ_n is a *posterior contraction rate* at the parameter θ_0 with respect to the semimetric d if $\Pi_n(\theta: d(\theta, \theta_0) \geq M_n\epsilon_n|\, X^{(n)}) \to 0$ in

[1] A semimetric satisfies the axiom of symmetry and the triangle inequality like a metric, but two distinct elements need not be at a positive distance. A semimetric space can always be converted to a metric space by identifying any pair of elements with zero distance.

$P_{\theta_0}^{(n)}$-probability, for every $M_n \to \infty$. If all experiments share the same probability space and the convergence to zero takes place almost surely $[P_{\theta_0}^{(\infty)}]$, then ϵ_n is said to be a *posterior contraction rate in the strong sense.*

We defined "*a*" rather than *the* rate of contraction, and hence logically any rate slower than a contraction rate is also a contraction rate. Naturally we are interested in a fastest decreasing sequence ϵ_n, but in general this may not exist or may be hard to establish. Thus our rate is an upper bound for a targeted rate, and generally we are happy if our rate is equal to or close to an "optimal" rate. With an abuse of terminology we often make statements like "ϵ_n is *the* rate of contraction."

In most infinite-dimensional models, the constants M_n in the definition of a contraction rate can be fixed to a single large constant M without changing ϵ_n, which gives a contraction rate in a slightly stronger sense. However, for typical finite-dimensional models, the sequence M_n must be allowed to grow indefinitely to obtain the "usual" contraction rate (such as $n^{-1/2}$ in a "smooth" model). The requirement for "every $M_n \to \infty$" should be understood as "for any arbitrarily slow $M_n \to \infty$."

We begin with some examples where the rate of contraction can be directly calculated, often using the following simple observation.

Lemma 8.2 *If $\Theta \subset \mathbb{R}$ and $\mathrm{E}(\theta|\, X^{(n)}) = \theta_0 + O_p(\epsilon_n)$ and $\mathrm{var}(\theta|\, X^{(n)}) = O_p(\epsilon_n^2)$, with respect to the distribution generated by the true parameter θ_0, then ϵ_n is a rate of contraction at θ_0 with respect to the Euclidean distance.*

Proof By Chebyshev's inequality, the probability $\Pi_n(\theta\colon |\theta - \mathrm{E}(\theta|\, X^{(n)})| > M_n\epsilon_n|\, X^{(n)})$ is bounded above by $\mathrm{var}(\theta|\, X^{(n)})/(M_n\epsilon_n)^2$ and hence tends to zero for any $M_n \to \infty$. Furthermore, the variable $\Pi_n(\theta\colon |\mathrm{E}(\theta|\, X^{(n)}) - \theta_0| > M_n\epsilon_n|\, X^{(n)})$ is the indicator of the event where $|\mathrm{E}(\theta|\, X^{(n)}) - \theta_0| > M_n\epsilon_n$, which has probability tending to zero. □

Example 8.3 (Bernoulli) If $X_1, \ldots, X_n|\, \theta \overset{\text{iid}}{\sim} \mathrm{Bin}(1, \theta)$ and $\theta \sim \mathrm{Be}(a, b)$, then by elementary calculation the posterior is $\theta|\, X_1, \ldots, X_n \sim \mathrm{Be}(a + n\bar{X}_n, b + n - n\bar{X}_n)$. Thus

$$\mathrm{E}(\theta|\, X_1, \ldots, X_n) = \frac{a + n\bar{X}_n}{a + b + n} = \theta_0 + O_p\Big(\frac{1}{\sqrt{n}}\Big),$$

$$\mathrm{var}\,(\theta|\, X_1, \ldots, X_n) = \frac{(a + n\bar{X}_n)(b + n - n\bar{X}_n)}{(a + b + n)^2(a + b + n + 1)} \le \frac{1}{n}.$$

Hence the rate of contraction is $\epsilon_n = n^{-1/2}$.

Example 8.4 (Uniform) If $X_1, \ldots, X_n|\, \theta \overset{\text{iid}}{\sim} \mathrm{Unif}(0, \theta)$ and θ has the improper density proportional to θ^{-a}, then the posterior density is proportional to $\theta^{-(n+a)}\mathbb{1}\{\theta > X_{(n)}\}$, where $X_{(n)}$ is the maximum of the observations. Thus

$$\mathrm{E}(\theta|\, X_1, \ldots, X_n) = \frac{n + a - 1}{n + a - 2} X_{(n)} = \theta_0 + O_p\Big(\frac{1}{n}\Big),$$

$$\mathrm{var}\,(\theta|\, X_1, \ldots, X_n) = \Big(\frac{n+a-1}{n+a-3} - \Big(\frac{n+a-1}{n+a-2}\Big)^2\Big) X_{(n)}^2 = O_p\Big(\frac{1}{n^2}\Big).$$

Hence the rate of contraction is $\epsilon_n = n^{-1}$.

In the preceding examples the prior is conjugate to the model, enabling explicit expressions for the posterior mean and variance. However, the conclusions go through for all priors that possess a continuous and positive density at θ_0. In fact, for most parametric models the posterior distribution is asymptotically free of the prior, provided that it satisfies this weak restriction. For regular models, such as Example 8.3, this follows from the Bernstein–von Mises theorem (see e.g. Theorem 10.1 of van der Vaart 1998), which shows that the posterior distribution of θ is asymptotically a normal distribution with mean centered at an efficient estimator (e.g. the maximum likelihood estimator) and with variance proportional to n^{-1}, giving an $n^{-1/2}$ contraction rate. More generally, the rate of contraction can often be identified with a *local scaling rate* r_n that ensures that the local likelihood ratio process $Z_n(u) = \prod_{i=1}^n (p_{\theta_0+r_n u}/p_{\theta_0})(X_i)$ converges in (marginal) distribution to a limit (see Chapter I of Ibragimov and Has'minskiĭ 1981, and the proof of Proposition 1 of Ghosal et al. 1995).

Example 8.5 (Dirichlet process) The posterior distribution in the model $X_1, \ldots, X_n|\, P \overset{\text{iid}}{\sim} P$ and $P \sim \mathrm{DP}(\alpha)$ is the $\mathrm{DP}(\alpha + n\mathbb{P}_n)$-distribution, for $\mathbb{P}_n$ the empirical distribution of the observations. The posterior mean and variance of $P(A)$ are given by (4.11) and (4.12), and are $P_0(A) + O_P(n^{-1/2})$ and $O_P(n^{-1})$, respectively. Thus the posterior rate of contraction relative to the semimetric $d(P_1, P_2) = |P_1(A) - P_2(A)|$ is $n^{-1/2}$. (This result can also be obtained from Example 8.3.)

As an example of a global metric d, consider sample space $\mathbb{R}$ and the $\mathbb{L}_2(\nu)$-distance between the distribution functions of the measures, for a σ-finite positive measure ν on $\mathbb{R}$: $d^2(F_1, F_2) = \int (F_1(t) - F_2(t))^2\, d\nu(t)$. By Markov's inequality followed by Fubini's theorem

$$\Pi_n(F\colon d(F, F_0) \geq \epsilon|\, X_1, \ldots, X_n) \leq \frac{1}{\epsilon^2} \int \mathrm{E}\Big((F(t) - F_0(t))^2|\, X_1, \ldots, X_n\Big)\, d\nu(t).$$

The integrand can be further expanded as the sum of $\{\mathrm{E}(F(t)|\, X_1, \ldots, X_n) - F_0(t)\}^2$ and $\mathrm{var}\,(F(t)|\, X_1, \ldots, X_n)$, which are given by (4.11) and (4.12). Thus the preceding display is bounded by, for $\alpha = MG$,

$$\frac{2M^2 \int (G - F_0)^2\, d\nu}{(M+n)^2} + \frac{2n^2 \int (\mathbb{F}_n - F_0)^2\, d\nu}{(M+n)^2} + \frac{Mn \int F_0(1 - F_0)\, d\nu}{(M+n)^2(M+n+1)}.$$

For a finite measure ν all terms are of the order $O_P(n^{-1})$, whence the rate of contraction is $n^{-1/2}$. Even without finiteness of ν the rate is $n^{-1/2}$ provided the integrals $\int G^2(1-G)^2\, d\nu$ and $\int F_0(1 - F_0)\, d\nu$ are finite. This follows since

$$\begin{aligned}\int (G - F_0)^2 d\nu &\leq 2\int_{-\infty}^0 G^2 d\nu + 2\int_{\infty}^0 F_0^2 + \int_0^\infty (1-G)^2 d\nu + 2\int_0^\infty (1 - F_0)^2 d\nu \\ &\lesssim \int G^2(1-G)^2 d\nu + \int F_0^2(1 - F_0)^2 d\nu\end{aligned}$$

and $F_0^2(1-F_0)^2 \leq F_0(1-F_0)$. In particular, this applies to the Lebesgue measure, if both G and F_0 have finite integrals.

Example 8.6 (White noise model) Suppose we observe an infinite-dimensional random vector $X^{(n)} = (X_{n,1}, X_{n,2}, \ldots)$, where $X_{n,i} \,|\, \theta \overset{\text{ind}}{\sim} \text{Nor}(\theta_i, n^{-1})$ and the parameter $\theta = (\theta_1, \theta_2, \ldots)$ belongs to ℓ_2, with square norm $\|\theta\|^2 = \sum_{i=1}^{\infty} \theta_i^2$. This model is equivalent to the white noise model considered in Section 8.3.4, and hence can be interpreted as concerned with estimating a function with Fourier coefficients (θ_i).

The priors $\theta_i \overset{\text{ind}}{\sim} \text{Nor}(0, i^{-2\alpha-1})$, for some fixed $\alpha > 0$, are conjugate in this model. The posterior distribution can be computed by considering the countably many coordinates separately, and is explicitly given by

$$\theta_i \,|\, X^{(n)} \overset{\text{ind}}{\sim} \text{Nor}\Big(\frac{nX_{n,i}}{n+i^{2\alpha+1}}, \frac{1}{n+i^{2\alpha+1}}\Big). \tag{8.1}$$

We claim that the rate of posterior contraction is $\epsilon_n = n^{-\min(\alpha,\beta)/(1+2\alpha)}$ at true parameters θ_0 with finite square Sobolev norm $\|\theta_0\|_{2,2,\beta}^2 = \sum_{i=1}^{\infty} i^{2\beta}\theta_{i,0}^2$.

By Chebyshev's inequality, $\Pi_n(\|\theta-\theta_0\| \geq \epsilon \,|\, X^{(n)})$ is bounded above by ϵ^{-2} times

$$\int \|\theta-\theta_0\|^2 \, d\Pi_n(\theta \,|\, X^{(n)}) = \sum_{i=1}^{\infty} (\text{E}(\theta_i \,|\, X^{(n)}) - \theta_{0,i})^2 + \sum_{i=1}^{\infty} \text{var}(\theta_i \,|\, X^{(n)}).$$

The expectations of the terms of the first series on the right are the mean square errors of the univariate estimators $\text{E}(\theta_i \,|\, X^{(n)}) = nX_{n,i}/(n+i^{2\alpha+1})$ of the coordinates $\theta_{i,0}$, and can be easily evaluated as the sum of a square bias and variance, while the second series on the right is deterministic and can be directly read off from (8.1). We conclude that the expectation of the right side of the preceding display is given by

$$\sum_{i=1}^{\infty} \frac{\theta_{i,0}^2 i^{4\alpha+2}}{(n+i^{2\alpha+1})^2} + \sum_{i=1}^{\infty} \frac{n}{(n+i^{2\alpha+1})^2} + \sum_{i=1}^{\infty} \frac{1}{n+i^{2\alpha+1}}.$$

By Lemma K.7, the second and third series are of the order $n^{-2\alpha/(2\alpha+1)}$; and the first is of the order $n^{-2\min(\alpha,\beta)/(2\alpha+1)}$ if the true parameter θ_0 satisfies $\sum_{i=1}^{\infty} i^{2\beta}\theta_{i,0}^2 < \infty$. Lemma K.7 also shows that the latter bound is best possible if it is to be uniform in θ_0 such that $\sum_{i=1}^{\infty} i^{2\beta}\theta_{i,0}^2 \leq 1$. We may interpret the preceding as saying that $n^{-\min(\alpha,\beta)/(2\alpha+1)}$ is the (best possible) contraction rate if the only information on θ_0 is that it has a finite Sobolev norm of order β.

For given β, the prior corresponding to the choice $\alpha = \beta$ gives the best contraction rate for these parameters θ_0. Interestingly, under this prior $\sum_{i=1}^{\infty} i^{2\beta}\theta_i^2 = \infty$ a.s., which might be interpreted as indicating that the prior gives probability zero to the parameters it seems to target. This seems embarrassing to the Bayesian method. However we may argue that the support of any of the priors under consideration is the whole of ℓ_2, and approximation of a true parameter by elements of the support is of greater importance for a posterior contraction rate than belonging to a set of prior probability 1. We revisit this situation in Section 12.4.

Contraction of the posterior distribution at a rate implies the existence of point estimators that converge at the same rate. The same construction as in Proposition 6.7 applies.

Theorem 8.7 *If the posterior contraction rate at θ_0 is ϵ_n, then the center $\hat{\theta}_n$ of the smallest ball that contains posterior mass at least $1/2$ satisfies $d(\hat{\theta}_n, \theta_0) = O_P(\epsilon_n)$ in $P_{\theta_0}^{(n)}$-probability (or almost surely if the posterior contraction rate is in the strong sense).*

Proof The proof of Proposition 6.7 with ϵ replaced by $M_n\epsilon_n$ shows that $d(\hat{\theta}_n, \theta_0) \leq 2M_n\epsilon_n + 1/n$, where the $1/n$ term can be replaced by an arbitrarily small, positive quantity. Since this is true for every $M_n \to \infty$, the sequence $d(\hat{\theta}_n, \theta_0)/\epsilon_n$ is tight for every $M_n \to \infty$. This implies the claim. □

For the posterior mean, if this is defined, the argument of Theorem 6.8 gives the following theorem.

Theorem 8.8 *If d is bounded and $\theta \mapsto d^s(\theta, \theta_0)$ is convex for some $s \geq 1$, then the posterior mean $\hat{\theta}_n = \int \theta\, d\Pi_n(\theta|\, X^{(n)})$ satisfies $d(\hat{\theta}_n, \theta_0) \leq M_n\epsilon_n + \|d\|_\infty^{1/s}\, \Pi_n\, (\theta\colon d(\theta, \theta_0) > M_n\epsilon_n|\, X^{(n)})^{1/s}$.*

Thus the rate of posterior contraction is transferred to the posterior mean provided the posterior probability of the complement of (a multiple) of an ϵ_n-ball around θ_0 converges to zero sufficiently fast. For instance, for the Hellinger distance we can choose $s = 2$, and $\Pi_n(\theta\colon d(\theta, \theta_0) > M_n\epsilon_n|\, X^{(n)})$ should be bounded by a multiple of ϵ_n^2. This is usually the case; in fact, the contraction of the posterior is typically exponentially fast.

The applicability of Lemma 8.2 is typically limited to the situation where the posterior distribution can be computed explicitly. In general, closed-form expressions are not available, and the posterior needs to be analyzed through more abstract arguments. In the subsequent sections we develop a theory of posterior contraction rates for dominated models. We describe the rates in terms of a prior concentration rate and the size of the model, which may be viewed as quantitative analogs of Schwartz's conditions for posterior consistency.

8.2 Independent Identically Distributed Observations

In this section we take the parameter equal to a probability density p, belonging to a class $\mathcal{P}$ of probability densities relative to a given dominating measure ν on a sample space $(\mathfrak{X}, \mathscr{X})$. We consider the posterior distribution $\Pi_n(\cdot|\, X_1, \ldots X_n)$ based on a sequence of priors Π_n on $\mathcal{P}$ and observations satisfying $X_1, \ldots, X_n|\, p \overset{\text{iid}}{\sim} p$, where $p \sim \Pi_n$ in the Bayesian setup and the posterior distribution is studied under the assumption that $p = p_0$. As usual we denote a density and the corresponding probability measure by lower- and uppercase letters (e.g. p and P).

We derive posterior contraction rates relative to a metric d on the parameter set $\mathcal{P}$ that satisfies the requirement: for every $n \in \mathbb{N}$ and $\epsilon > 0$ and p_1 with $d(p_1, p_0) > \epsilon$ there exists a test ϕ_n with, for some universal constants $\xi, K > 0$,

$$P_0^n\phi_n \leq e^{-Kn\epsilon^2}, \qquad \sup_{d(p,p_1)<\xi\epsilon} P^n(1-\phi_n) \leq e^{-Kn\epsilon^2}. \tag{8.2}$$

Thus the "null hypothesis" P_0 can be separated from every d-ball at some distance from it, by a test with exponentially small error probabilities, smaller if the alternative is farther from the null hypothesis. It is shown in Proposition D.8 that this requirement is fulfilled (with $\xi = 1/2$ and $K = 1/8$) by any semimetric d that generates convex balls and satisfies $d(p_0, p) \le d_H(p_0, p)$ for every density p, where d_H is the Hellinger metric. Examples are the Hellinger metric itself, but also the total variation metric, and for a bounded set of densities a multiple of the $\mathbb{L}_2(\nu)$-metric (namely this metric divided by twice the upper bound $\sup_{p\in\mathcal{P}} \|\sqrt{p}\|_\infty$ on the densities; see Lemma B.1).

The following theorem describes the posterior contraction rate in terms of a prior concentration rate and the entropy of the model. The concentration rate is measured in terms of a combination of the Kullback-Leibler divergence $K(p_0; p) = P_0 \log(p_0/p)$ and the kth Kullback-Leibler variation $V_{k,0}(p_0; p) = P_0|\log(p_0/p) - K(p_0; p)|^k$. For every $\epsilon > 0$ define neighborhoods of p_0 by

$$\begin{aligned} B_0(p_0, \epsilon) &= \{p\colon K(p_0; p) < \epsilon^2\}, \\ B_k(p_0, \epsilon) &= \{p\colon K(p_0; p) < \epsilon^2, V_{k,0}(p_0; p) < \epsilon^k\}, \qquad (k > 0). \end{aligned} \tag{8.3}$$

Theorem 8.9 (Basic contraction rate) *Given a distance d for which* (8.2) *is satisfied, suppose that there exist partitions $\mathcal{P} = \mathcal{P}_{n,1} \cup \mathcal{P}_{n,2}$ and a constant $C > 0$, such that, for constants $\bar\epsilon_n \le \epsilon_n$ with $n\bar\epsilon_n^2 \to \infty$,*

(i) $\Pi_n(B_2(p_0, \bar\epsilon_n)) \ge e^{-Cn\bar\epsilon_n^2}$. (8.4)

(ii) $\log N(\xi\epsilon_n, \mathcal{P}_{n,1}, d) \le n\epsilon_n^2$. (8.5)

(iii) $\Pi_n(\mathcal{P}_{n,2}) \le e^{-(C+4)n\bar\epsilon_n^2}$. (8.6)

Then the posterior rate of contraction at p_0 is ϵ_n. If $\epsilon_n \gtrsim n^{-\alpha}$ for some $\alpha \in (0, 1/2)$ and (i) holds with $B_2(p_0, \bar\epsilon_n)$ replaced by $B_k(p_0, \bar\epsilon_n)$ for some k such that $k(1 - 2\alpha) > 2$, then the contraction rate is also in the almost sure sense $[P_0^\infty]$.

Proof Assumption (8.2) implies assumption (D.6) with K in the latter condition equal to the present Kn and $c = 1$. Hence by assumption (ii) and Theorem D.5 applied with $\epsilon = M\epsilon_n$ and $j = 1$ in its assertion, there exist tests ϕ_n with error probabilities

$$P_0^n\phi_n \le e^{n\epsilon_n^2}\frac{e^{-KnM^2\epsilon_n^2}}{1 - e^{-KnM^2\epsilon_n^2}}, \qquad \sup_{p\in\mathcal{P}_{n,1}: d(p,p_0)>M\epsilon_n} P^n(1 - \phi_n) \le e^{-KnM^2\epsilon_n^2}.$$

For $KM^2 > 1$ both error probabilities tend to zero. In view of Bayes's formula, for A_n the event that $\int \prod_{i=1}^n (p/p_0)(X_i)\, d\Pi_n(p) \ge e^{-(2+C)n\bar\epsilon_n^2}$, we can bound the posterior mass of the set $\{p\colon d(p, p_0) > M\epsilon_n\}$ by

$$\phi_n + \mathbb{1}\{A_n^c\} + e^{(2+C)n\bar\epsilon_n^2}\int_{d(p,p_0)>M\epsilon_n} \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi_n(p)(1 - \phi_n). \tag{8.7}$$

The expected value under P_0^n of the first term tends to zero by construction of the tests ϕ_n. The same is true for the second term, by assumption (i) and Lemma 8.10 (below, with the constant D of the lemma taken equal to 1). The expected value of the third term is bounded above by

$$e^{(2+C)n\bar\epsilon_n^2}\int_{d(p,p_0)>M\epsilon_n} P^n(1-\phi_n)\,d\Pi_n(p) \le e^{(2+C)n\bar\epsilon_n^2}\Bigl(e^{-KnM^2\epsilon_n^2}+\Pi_n(\mathcal{P}_{n,2})\Bigr).$$

The last inequality follows by splitting the integral in parts over $\mathcal{P}_{n,1}$ and $\mathcal{P}_{n,2}$, and next using the bounds on the tests in the first part and the trivial bound $P^n(1-\phi_n)\le 1$ in the second part. For $KM^2>2+C$ the right side tends to zero by assumption (iii).

For the final assertion on almost sure contraction we note that for a general $k\ge 2$ Lemma 8.10 gives the bound $\mathrm{P}(A_n^c)\lesssim(\sqrt{n}\epsilon_n)^{-k}$, so that $\sum_n \mathrm{P}(A_n^c)<\infty$ under the stated conditions on ϵ_n and k, whence $P_0^\infty(\limsup A_n)=0$, by the Borel-Cantelli lemma, which means that $\mathbb{1}\{A_n^c\}\to 0$ almost surely. The other terms in (8.7) converge to zero exponentially fast, and can be treated by the same reasoning. □

From inspection of the proof, it is clear that if the constant C can be chosen the same for a collection of values of the true density p_0, then the contraction rate sequence ϵ_n can be chosen to be the same for all p_0 in the collection. Hence it makes sense to compare the posterior contraction rate with the corresponding minimax rate for the collection of true p_0.

The theorem copies the form of Theorem 6.23 on posterior consistency. The Kullback-Leibler property of p_0 has been quantified in condition (i), and the fixed ϵ is replaced by the sequences $\bar\epsilon_n$ and ϵ_n. Conditions (i) and (ii) (for given $\mathcal{P}_{n,1}$) are monotone in $\bar\epsilon_n$ or ϵ_n in the sense that they are satisfied for every larger value of these rates as soon at they are true for given values. They can thus be seen as defining minimal possible values of $\bar\epsilon_n$ and ϵ_n. The rate of posterior contraction is the slowest of the minimal values satisfying (ii) and the pair (i), (iii), respectively.

The *prior mass condition* (i) requires that the priors put a sufficient amount of mass near the true density p_0. In the basic theorem “near” is measured through the Kullback-Leibler divergence and variation $K(p_0;\,p)$ and $V_{2,0}(p_0;\,p)$, which define the neighborhoods $B_2(p_0;\,\epsilon)$. Later results will show that the variation may be superfluous and the neighborhoods enlarged to the Kullback-Leibler balls $B_0(p_0;\,\epsilon)$.

The *entropy condition* (ii) is the other main determinant of the contraction rate. It bounds the complexity of the model $\mathcal{P}_{n,1}$, which is the bulk of the “support” of the prior, in view of condition (iii). It is known that a rate ϵ_n satisfying (ii) for d the Hellinger metric gives the *minimax optimal rate* of contraction for estimators of p relative to the Hellinger metric, given the model $\mathcal{P}_{n,1}$ (see Birgé 1983a, Yang and Barron 1999). In the present situation the contraction rate can be seen to be uniform in p_0 that satisfy the conditions for given constants, so that the contraction is uniform and meets the minimax criterion. Technically the entropy condition ensures the existence of good tests of p_0 versus alternatives at some distance. For more flexibility it can be replaced by a testing condition; see Theorem 8.12.

The conditions (i) and (ii) are related. For the sake of the argument, assume that the square distance d^2 and the discrepancies K and $V_{2,0}$ are equivalent (which is the case if $d=d_H$ and the likelihood ratios p_0/p are uniformly bounded; see (8.8) below). Let ϵ_n be the minimal value that satisfies (ii), i.e. the optimal rate of contraction for the model $\mathcal{P}_{n,1}$. Thus, a minimal ϵ_n-cover of $\mathcal{P}_{n,1}$ consists of $e^{n\epsilon_n^2}$ balls. If the prior Π_n would spread its mass uniformly over $\mathcal{P}_{n,1}$, then every ball would receive prior mass approximately 1 over the number of balls in the cover, say $e^{-Cn\epsilon_n^2}$; the constant C may express the use of multiple distances and possible overlap between the balls in the cover. On the other hand, if Π_n is not evenly spread, then we should expect (i) to fail for some $p_0 \in \mathcal{P}_{n,1}$.

"Uniform priors" do not exist in infinite-dimensional models, and actually condition (i) is stronger than needed and will be improved ahead in Theorem 8.11. Nevertheless, a rough implication of this argument is that Π_n should be "not very unevenly spread" in order for the posterior distribution to attain the optimal rate of contraction at all elements p_0 in the parameter space.

Condition (iii), relative to (ii), can be interpreted as saying that a part of the model that barely receives prior mass need not have bounded complexity. It is trivially satisfied for the partition with $\mathcal{P}_{n,1} = \mathcal{P}$; we can make this choice if the entropy of the full model is bounded as in (ii).

The prior mass and entropy conditions of Theorem 8.9 use the Kullback-Leibler divergence and the metric d, respectively. Mixing two types of neighborhoods is not pretty, but seems to be unavoidable in general. An advantage is that the metric d reappears in the assertion of the theorem, giving a stronger sense of contraction when using a stronger metric (at the cost of a more restrictive complexity condition). It is of interest to note that for sufficiently regular models, the square Hellinger distance and Kullback-Leibler discrepancies are equivalent. In particular, by Lemmas B.2 and B.3, for any $k \geq 2$,

$$\Bigl\{p\colon d_H(p, p_0)\Bigl\|\frac{p_0}{p}\Bigr\|_\infty < \sqrt{2k!}\,\epsilon\Bigr\} \subset B_k(p_0, \epsilon) \subset \{p\colon d_H(p, p_0) < \epsilon\}. \tag{8.8}$$

Thus, given bounded likelihood ratios, a theorem can be stated solely in terms of the Hellinger distance (or, equivalently the Kullback-Leibler divergence). The same lemmas also apply under just moment conditions on the ratios p_0/p, but only up to factors $\log_- \epsilon$, giving equivalent rates of contraction up to logarithmic factors.

The testing condition (8.2) links the metric d to the likelihood. In the case of i.i.d. observations the Hellinger distance d_H seems to be a canonical choice. For a metric d that is stronger than the Hellinger distance the condition may not hold. Posterior contraction may still take place, but may involve properties of prior and model additional to the prior mass and entropy considered in the basic theorem. One possible approach to handling a different metric is to show, by additional arguments, that the posterior distribution gives mass tending to one to sets $\mathcal{S}_n \subset \mathcal{P}$ on which the new metric is comparable to the Hellinger distance. Specifically, if $d(p_0, p) \leq m_n d_H(p_0, p)$ for all $p \in \mathcal{S}_n$ and some sequence of numbers m_n, and $\Pi_n(\mathcal{S}_n^c | X_1, \ldots, X_n) \to 0$ in probability, then a posterior contraction rate ϵ_n relative to the Hellinger distance implies a contraction rate $m_n \epsilon_n$ with respect to d. One method to show that the posterior distributions concentrate on $\mathcal{S}_n$ is to prove that $\Pi_n(\mathcal{S}_n^c)$ is exponentially small, and apply Theorem 8.20 (below, which gives a generalization of (iii) in the basic theorem above). This simple device has been used to deal with the supremum distance and certain inverse problems.

The following lemma was used in the proof of Theorem 8.9. It gives a lower bound on the denominator of the posterior measure (the "evidence"), similar to Lemma 6.26. Employment of the stronger neighborhoods $B_k(p_0; \epsilon)$, rather than Kullback-Leibler balls $B_0(p_0; \epsilon)$ as in the latter lemma, gives better control of the probability of failure of the lower bound. Further variations, which give exponential bounds on these probabilities, are given in Problems 8.6 and 8.8.

Lemma 8.10 (Evidence lower bound) *For every $k \geq 2$ there exists a constant $d_k > 0$ (with $d_2 = 1$) such that for any probability measure Π on $\mathcal{P}$, and any positive constants ϵ, D, with P_0^n-probability at least $1 - d_k(D\sqrt{n}\epsilon)^{-k}$,*

$$\int \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi(p) \geq \Pi(B_k(p_0, \epsilon))e^{-(1+D)n\epsilon^2}.$$

Proof The integral becomes smaller by restricting it to the set $B := B_k(p_0, \epsilon)$. By next dividing the two sides of the inequality by $\Pi(B)$, we can rewrite the inequality in terms of the prior Π restricted and renormalized to a probability measure on B. Thus we may without loss of generality assume that $\Pi(B) = 1$. By Jensen's inequality applied to the logarithm,

$$\log \int \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi(p) \geq \sum_{i=1}^n \int \log \frac{p}{p_0}(X_i)\, d\Pi(p) =: Z.$$

The right side has mean $\mathrm{E}Z = -n\int K(p_0; p)\, d\Pi(p) > -n\epsilon^2$, by the definition of B. Furthermore, by the Marcinkiewicz-Zygmund inequality (Lemma K.4), followed by Jensen's inequality, Fubini's theorem, and again the definition of B,

$$\mathrm{E}\left|\frac{Z - \mathrm{E}Z}{\sqrt{n}}\right|^k \lesssim P_0 \left|\int \left(\log \frac{p}{p_0} - P_0 \log \frac{p}{p_0}\right) d\Pi(p)\right|^k \leq \int V_{k,0}(p_0; p)\, d\Pi(p) \leq \epsilon^k.$$

The constant d_k in the lemma is the proportionality constant in the first inequality (e.g. $d_2 = 1$). Finally $P_0^n(Z < -(1 + D)n\epsilon^2) \leq P_0^n(Z - \mathrm{E}Z < -Dn\epsilon^2)$ is bounded by $\mathrm{E}|Z - \mathrm{E}Z|^k/(Dn\epsilon^2)^k$, by Markov's inequality, which can be further bounded as in the lemma. □

One deficit of Theorem 8.9 is that it does not satisfactorily cover finite-dimensional models. For models of fixed dimension it yields the rate $n^{-1/2}$ times a logarithmic factor rather than the correct $n^{-1/2}$, and when applied to sieves of finite, increasing dimensions it incurs a similar unnecessary logarithmic term. To improve this situation, both the prior mass (i) and entropy (ii) conditions must be refined.

The following theorem is also more precise in employing the neighborhoods $B_0(p_0; \epsilon)$, which are solely in terms of the Kullback-Leibler divergence and not the variation.

Theorem 8.11 (Refined contraction rate) *Given a distance d for which (8.2) is satisfied, suppose that there exist partitions $\mathcal{P} = \mathcal{P}_{n,1} \cup \mathcal{P}_{n,2}$, such that, for constants $\epsilon_n, \bar\epsilon_n \geq n^{-1/2}$, and every sufficiently large j,*

$$\text{(i)}\quad \frac{\Pi_n(P{:}\, j\epsilon_n < d(p, p_0) \leq 2j\epsilon_n)}{\Pi_n(B_0(p_0, \epsilon_n))} \leq e^{Kn\epsilon_n^2 j^2/2}, \tag{8.9}$$

$$\text{(ii)}\quad \sup_{\epsilon \geq \epsilon_n} \log N\Big(\xi\epsilon, \{p \in \mathcal{P}_{n,1}{:}\, d(p, p_0) \leq 2\epsilon\}, d\Big) \leq n\epsilon_n^2, \tag{8.10}$$

$$\text{(iii)}\quad \frac{\Pi_n(\mathcal{P}_{n,2})}{\Pi_n(B_0(p_0, \bar\epsilon_n))} = o(e^{-D_n n\bar\epsilon_n^2}), \qquad \textit{for some } D_n \to \infty. \tag{8.11}$$

Then the posterior rate of contraction at p_0 is ϵ_n. This remains true if (iii) is replaced by

$$\text{(iii}'\text{)}\quad \frac{\Pi_n(\mathcal{P}_{n,2})}{\Pi_n(B_2(p_0, \bar\epsilon_n))} = o(e^{-2n\bar\epsilon_n^2}). \tag{8.12}$$

Proof The proof is similar to the proof of Theorem 8.9, but we use Lemma 6.26 next to Lemma 8.10, more refined tests, and split the sieves in "shells." We prove first that the posterior probability of the set $\{p \in \mathcal{P}_{n,1}: d(p, p_0) > M\epsilon_n\}$ tends to zero for large M or $M \to \infty$ (under (i) and (ii)), and next separately the same for the sets $\mathcal{P}_{n,2}$ (under either (iii) or (iii$'$)).

By (8.10) and Theorem D.5, applied with $\epsilon = M\epsilon_n$, a given $M > 1$, and N the constant function $N(\epsilon) = n\epsilon_n^2$, there exist tests ϕ_n such that, for every $j \in \mathbb{N}$,

$$P_0^n \phi_n \le e^{n\epsilon_n^2} \frac{e^{-KnM^2\epsilon_n^2}}{1 - e^{-KnM^2\epsilon_n^2}}, \qquad \sup_{p \in \mathcal{P}_{n,1}: d(p,p_0) > Mj\epsilon_n} P^n(1 - \phi_n) \le e^{-KnM^2\epsilon_n^2 j^2}.$$

The set $\{p \in \mathcal{P}_{n,1}: d(p, p_0) > M\epsilon_n\}$ can be partitioned into the countably many "shells" $\mathcal{S}_{n,j} = \{p \in \mathcal{P}_{n,1}: M\epsilon_n j < d(p, p_0) \le M\epsilon_n(j + 1)\}$, for $j \in \mathbb{Z}$. For every $j \in \mathbb{N}$, by Fubini's theorem and the preceding display,

$$P_0^n \Big[\int_{\mathcal{S}_{n,j}} \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi_n(p)(1 - \phi_n)\Big] \le e^{-KnM^2\epsilon_n^2 j^2} \Pi_n(\mathcal{S}_{n,j}).$$

For A_n the event that $\int \prod_{i=1}^n (p/p_0)(X_i)\, d\Pi_n(p) \ge e^{-2Dn\epsilon_n^2} \Pi_n(B_0(p_0, \epsilon_n))$, we now replace the upper bound (8.7) on the posterior probability of the set $\{p \in \mathcal{P}_{n,1}: d(p, p_0) > M\epsilon_n\}$ by

$$\phi_n + \mathbb{1}\{A_n^c\} + \frac{1}{e^{-2Dn\epsilon_n^2} \Pi_n(B_0(p_0, \epsilon_n))} \int_{p \in \mathcal{P}_{n,1}: d(p,p_0) > M\epsilon_n} \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi_n(p)(1 - \phi_n).$$

By construction of the tests, the expected value of the first term tends to zero if $n\epsilon_n^2 \to \infty$ and $KM^2 > 1$, and can be made arbitrarily small by choosing M big if $n\epsilon_n^2$ is only known to be bounded away from zero. The expected value of the second term is bounded above by $1/D$ by Lemma 6.26, and can be made arbitrarily small by choice of D. The domain of the integral in the third term is contained in $\cup_{j \ge 1} \mathcal{S}_{n,j}$, and hence the expected value of this term is bounded above by

$$\sum_{j \ge 1} \frac{e^{-KnM^2\epsilon_n^2 j^2} \Pi_n(\mathcal{S}_{n,j})}{e^{-2Dn\epsilon_n^2} \Pi_n(B_0(p_0, \epsilon_n))} \le \sum_{j \ge 1} e^{-n\epsilon_n^2(KM^2 j^2 - 2D - \frac{1}{2}KM^2 j^2)},$$

by (8.9). This converges to zero as $n\epsilon_n^2 \to \infty$ and $KM^2 > 4D$, and can be made arbitrarily small by choosing large M if $n\epsilon_n^2$ is only known to be bounded away from zero.

Next we consider the posterior probability of the sets $\mathcal{P}_{n,2}$. Given validity of (iii), we define the events A_n as previously, but with ϵ_n replaced by $\bar\epsilon_n$. Reasoning as before we obtain that the posterior probability of $\mathcal{P}_{n,2}$ is bounded above by

$$\mathbb{1}\{A_n^c\} + \frac{\Pi_n(\mathcal{P}_{n,2})}{e^{-2Dn\bar\epsilon_n^2} \Pi_n(B_0(p_0, \bar\epsilon_n))}.$$

The expected value of the first term is bounded by $1/D$ and hence can be made arbitrarily small by choosing a large D. The second term tends to zero under assumption (iii), for any D. If not (iii) but (iii$'$) is assumed, then we redefine the A_n as the events that $\int \prod_{i=1}^n (p/p_0)(X_i)\, d\Pi_n(p) \ge e^{-2Dn\bar\epsilon_n^2} \Pi_n(B_2(p_0, \bar\epsilon_n))$. The posterior probability of $\mathcal{P}_{n,2}$

is then bounded by the preceding display, but with B_0 replaced by B_2. By Lemma 8.10 the expectation of the first term $\mathbb{1}\{A_n^c\}$ is bounded above by $(2D-1)^{-2}(n\bar\epsilon_n^2)^{-1}$. To finish the proof, we may assume that either $n\bar\epsilon_n^2 \to \infty$ or $n\bar\epsilon_n^2$ is bounded; otherwise we argue along subsequences. If $n\bar\epsilon_n^2 \to \infty$, then we choose $D = 1$ and the expectation of $\mathbb{1}\{A_n^c\}$ tends to zero, as does the second term in the preceding display, by assumption (iii′). If $n\bar\epsilon_n^2$ remains bounded then the factor $e^{-2Dn\bar\epsilon_n^2}$ is bounded away from zero and infinity for every fixed D, since $n\bar\epsilon_n^2 \geq 1$, by assumption. By (iii′) the second term of the preceding display then tends to zero for every fixed $D > 0$, and the expectation of the first can be made arbitrarily small by choosing a large D. □

The left side of (8.10) is called the *local entropy* or *Le Cam dimension* of the set $\mathcal{P}_{n,1}$ at p_0. It measures the number of balls of radius $\xi\epsilon$ needed to cover a ball (or a shell) of radius 2ϵ. That the quotient of the latter two radii is independent of ϵ makes the Le Cam dimension smaller for “finite-dimensional” models, where it indeed behaves as a dimension (see Appendix D.1). For genuinely infinite-dimensional models the local entropy in the left side of (8.10) is typically comparable in magnitude to $\log N(\xi\epsilon_n, \mathcal{P}_{n,1}, d)$. As the latter quantity is always an upper bound on the left side of (8.10), this condition is implied by (8.5).

The numerator of (8.9) is trivially bounded above by 1, and hence this condition is implied by (8.4) (where B_2 may even be replaced by the larger set B_0). Similarly, given (8.4), condition (8.12) relaxes the corresponding condition (8.6) of Theorem 8.9.

8.2.1 Further Refinements

The entropy conditions appearing in the preceding theorems are used to construct tests against (nonconvex) alternatives of the form $\{p \in \mathcal{P}_{n,1}: \epsilon < d(p, p_0) \leq 2\epsilon\}$. Tests are more fundamental than entropies for contraction rates, and in some applications can be constructed by direct methods. Thus the testing conditions in the following theorem give more flexibility and maneuverability. The proof is essentially contained in that of Theorem 8.11 with the choice $M = 1$.

Theorem 8.12 (Rates by testing) *If there exist partitions $\mathcal{P} = \mathcal{P}_{n,1} \cup \mathcal{P}_{n,2}$ and a constant $C > 0$, such that* (8.9) *and* (8.12) *hold for constants $\bar\epsilon_n \leq \epsilon_n$ with $n\bar\epsilon_n^2 \geq 1$ and every sufficiently large j, and in addition there exists a sequence of tests ϕ_n such that for some constant $K > 0$ and for every sufficiently large j*

$$P_0^n\phi_n \to 0, \qquad \sup_{p\in\mathcal{P}_{n,1}: j\epsilon_n < d(p,p_0) \leq 2j\epsilon_n} P^n(1-\phi_n) \leq e^{-Kn\epsilon_n^2 j^2}, \tag{8.13}$$

then the posterior rate of contraction at p_0 is ϵ_n.

Like for consistency, testing conditions for a posterior contraction rate result are almost necessary. Given prior concentration, posterior contraction at an exponential speed implies the existence of tests and a sieve, as hypothesized in Theorem 8.12.

Theorem 8.13 (Necessity of testing) *If $P_0^n \Pi_n(p\colon d(p_0, p) \geq M\epsilon_n | X_1, \ldots, X_n) \leq e^{-Cn\epsilon_n^2}$, for given positive constants C and M, then there exists a partition $\mathcal{P} = \mathcal{P}_{1,n} \cup \mathcal{P}_{2,n}$ and a sequence of tests ϕ_n such that $P_0^n \phi_n \to 0$, $\sup\{P^n(1-\phi_n)\colon p \in \mathcal{P}_{1,n}\} \leq e^{-Cn\epsilon_n^2/4}$ and $\Pi_n(\mathcal{P}_{2,n}) \leq e^{-Cn\epsilon_n^2/4}$.*

Proof Define an event $S_n = \{\Pi_n(p\colon d(p_0, p) \geq M\epsilon_n | X_1, \ldots, X_n) > e^{-Cn\epsilon_n^2/2}\}$ and proceed as in the proof of Theorem 6.22. The test $\phi_n = \mathbb{1}_{S_n}$ has type I error probability $P_0^n \phi_n = P_0^n(S_n) \leq e^{-Cn\epsilon_n^2/2} \to 0$ and $P^n(1-\phi_n) = P^n(S_n^c) \leq e^{-Cn\epsilon_n^2/2}$ for $p \in \mathcal{P}_{n,1} = \mathcal{P} \setminus \mathcal{P}_{n,2}$, where $\mathcal{P}_{n,2} = \{p\colon P^n(S_n^c) \geq e^{-Cn\epsilon_n^2/2}\}$. Finally $\Pi(\mathcal{P}_{n,2}) \leq e^{-Cn\epsilon_n^2/4}$ by arguments as before. □

In Theorems 8.9 and 8.11, the sieves $\mathcal{P}_{n,1}$ are controlled by their entropy, without taking account of the prior mass they carry, whereas their complements $\mathcal{P}_{n,2}$ are measured solely by their prior probability, regardless of their complexity. These are two extreme sides of what matters for posterior concentration. In the following theorem, which parallels Theorems 8.9, the parameter space is divided into more than two pieces to make a smoother transition from one method to another. Theorem 8.11 can be similarly refined.

Theorem 8.14 (Rate by partition entropy) *Given a distance d for which* (8.2) *is satisfied suppose that there exists disjoint subsets $\mathcal{P}_{n,j}$ of $\mathcal{P}$ such that, for constants ϵ_n with $n\epsilon_n^2 \to \infty$,*

$$\sum_{j=1}^{\infty} \sqrt{N(\xi\epsilon_n, \mathcal{P}_{n,j}, d)} \sqrt{\Pi_n(\mathcal{P}_{n,j})} \leq e^{n\epsilon_n^2}. \tag{8.14}$$

If (8.4) *holds for some constant $C > 0$ and $\bar{\epsilon}_n = \epsilon_n$, then $\Pi_n(p \in \cup_{j=1}^{\infty} \mathcal{P}_{n,j}\colon d(p, p_0) > M\epsilon_n | X_1, \ldots, X_n) \to 0$ in P_0^n-probability, for every sufficiently large M.*

Proof Without loss of generality, we may assume that the priors charge only $\cup_{j \geq 1} \mathcal{P}_{n,j}$. On the event $A_n = \{\int \prod_{i=1}^n (p/p_0)(X_i)\, d\Pi_n(p) \geq e^{-(C+2)n\epsilon_n^2}\}$ the posterior mass of the set $\{p\colon d(p, p_0) > M\epsilon_n\}$ can be bounded above, for arbitrary tests $\phi_{n,j}$, by

$$\sum_{j=1}^{\infty} \phi_{n,j} + \mathbb{1}\{A_n^c\} + e^{(C+2)n\epsilon_n^2} \sum_{j=1}^{\infty} \int_{p \in \mathcal{P}_{n,j}: d(p,p_0) > M\epsilon_n} \prod_{i=1}^{n} \frac{p}{p_0}(X_i)\, d\Pi_n(p)^n (1-\phi_{n,j}).$$

The event A_n has probability tending to one, by Lemma 8.10 and assumption (8.4). By Theorem D.5 applied for every $j \in \mathbb{Z}$ separately with $c \leftarrow c_j$ and $\mathcal{Q} = \{p \in \mathcal{P}_{n,j}\colon d(p_0, p) \geq M\epsilon_n\}$, there exist tests $\phi_{n,j}$ so that the expected value under P_0^n of the preceding display is bounded above by

$$\sum_{j=1}^{\infty} c_j\, N(\xi\epsilon_n, \mathcal{P}_{n,j}, d) \frac{e^{-KnM^2\epsilon_n^2}}{1 - e^{-KnM^2\epsilon_n^2}} + P_0^n(A_n^c) + e^{(C+2)n\epsilon_n^2} \sum_{j=1}^{\infty} \frac{1}{c_j} e^{-KnM^2\epsilon_n^2} \Pi_n(\mathcal{P}_{n,j}).$$

Now choose $c_j^2 = \Pi_n(\mathcal{P}_{n,j})/N(\xi\epsilon_n, \mathcal{P}_{n,j}, d)$ and $KM^2 > C+2$, and sum over j to obtain that this tends to zero, in view of (8.14). □

Theorem 8.14 contains Theorem 8.9 as the special case of a partition in a single set $\mathcal{P}_{n,1}$: if $\mathcal{P}_{n,j} = \varnothing$ for $j \geq 2$, then (8.14) reduces to the bound $\log N(\xi\epsilon_n, \mathcal{P}_{n,1}, d) \leq n\epsilon_n^2$. (The present set $\mathcal{P}_{n,0}$ plays the role of $\mathcal{P}_{n,2}$ in Theorem 8.9, and is presently not addressed.) At the other extreme, if the model is decomposed fine enough so that each $\mathcal{P}_{n,j}$ has diameter smaller than $\xi\epsilon_n$, then the covering numbers appearing in (8.14) are all 1, and hence (8.14) reduces to

$$\sum_{j=1}^{\infty} \sqrt{\Pi_n(\mathcal{P}_{n,j})} \leq e^{n\epsilon_n^2}. \tag{8.15}$$

This condition is a quantitative analog of the summability condition for posterior consistency in Example 6.24 (with $\alpha = 1/2$). (A contraction rate theorem based on (8.15) may also be derived using the martingale approach discussed in Section 6.8.4; see Problem 8.11.)

Theorem 8.14 can be written more neatly in terms of a new measure of size, called *Hausdorff entropy*, which takes into account both number and prior probabilities of ϵ-coverings; see Problem 8.13.

Surprisingly, Theorem 8.14 can also be proved as a corollary to Theorem 8.9 by a clever definition of the sieves; see Problem 8.13.

8.2.2 Priors Based on Finite Approximating Sets

The rate of posterior contraction is determined by the complexity of a model and the prior concentration. Although in the Bayesian setup the “model” can be viewed as being implicitly determined by the prior, complexity is predominantly a non-Bayesian quality. In fact, under general conditions a rate ϵ_n satisfying the entropy inequality $\log N(\epsilon_n, \mathcal{P}, d_H) \leq n\epsilon_n^2$ is *minimax optimal* for estimating a density in the model $\mathcal{P}$, relative to the Hellinger distance, for any method of point estimation. In this section we show that this rate can always be obtained (perhaps up to a logarithmic factor) also as a rate of posterior contraction, for some prior. This prior is constructed as a convex combination of discrete uniform measures on a sequence of ϵ_n-nets over the model.

We start by assuming the slightly stronger entropy inequality $\log N_{]}(\epsilon_n, \mathcal{P}, d_H) \leq n\epsilon_n^2$, defined in terms of upper brackets, rather than covers by balls. The ϵ-*upper bracketing number* $N_{]}(\epsilon, \mathcal{P}, d_H)$ is defined as the minimal number of functions $u_{1,n}, \ldots, u_{N_n,n}$ such that for every $p \in \mathcal{P}$ there exists j with $p \leq u_{j,n}$ and $d_H(p, u_{j,n}) \leq \epsilon$. Given a minimal set of $N_n = N_{]}(\epsilon_n, \mathcal{P}, d_H)$ upper brackets at discretization level ϵ_n, let Π_n be the discrete uniform measure on the set $\mathcal{P}_n = \{u_j / \int u_j\, d\nu\colon j = 1, \ldots, N_n\}$. Next, as overall prior form the mixture $\Pi = \sum_{n \in \mathbb{N}} \lambda_n \Pi_n$, for (λ_n) a strictly positive probability vector on $\mathbb{N}$.

Normalizing the upper brackets $u_{j,n}$ is necessary to ensure that the prior is supported on genuine probability densities, as upper brackets are functions with (typically) integrals larger than unity. The final prior Π is not supported on the target model $\mathcal{P}$, but this does not pose a challenge for applying the rate theorems.

Theorem 8.15 (Prior on nets) *If $\epsilon_n \downarrow 0$ is a sequence such that $\log N_{]}(\epsilon_n, \mathcal{P}, d_H) \leq n\epsilon_n^2$ for every n and $n\epsilon_n^2/\log n \to \infty$ and the weights λ_n are strictly positive with $\log \lambda_n^{-1} = O(\log n)$, then the posterior converges relative to d_H at the rate ϵ_n almost surely, at every $p_0 \in \mathcal{P}$.*

Proof Let $\mathcal{P}_n$ be the support of Π_n and let $\mathcal{Q} = \cup_{n=1}^\infty \mathcal{P}_n$, whence $\Pi(\mathcal{Q}) = 1$ by construction. It suffices to control the Hellinger entropy of $\mathcal{Q}$ and verify the prior mass condition.

If $p \le u$ and $\|\sqrt{u} - \sqrt{p}\|_2 = d_H(p, u) < \epsilon$, then by the triangle inequality $1 \le \|\sqrt{u}\|_2 \le \|\sqrt{u} - \sqrt{p}\|_2 + \|\sqrt{p}\|_2 \le \epsilon + 1$, and hence

$$d_H\Big(p, \frac{u}{\int u\, d\nu}\Big) \le d_H(p, u) + d_H\Big(u, \frac{u}{\int u\, d\nu}\Big) = \|\sqrt{u} - \sqrt{p}\|_2 + \|\sqrt{u}\|_2 - 1 \le 2\epsilon.$$

Therefore, by construction every point of $\mathcal{P}$ is within $2\epsilon_n$ distance of some point in $\mathcal{P}_n$. Since for $j > n$, every $p \in \mathcal{P}_j$ is within distance $2\epsilon_j \le 2\epsilon_n$ of $\mathcal{P}$, it follows that $\mathcal{P}_n$ is a $4\epsilon_n$-net over $\mathcal{P}_j$. Consequently, $\mathcal{P}_n$ is a $4\epsilon_n$-net over $\cup_{j\ge n}\mathcal{P}_j$ and hence $\cup_{j\le n}\mathcal{P}_j$ is a $4\epsilon_n$-net over $\mathcal{Q}$. The cardinality of $\cup_{j\le n}\mathcal{P}_j$ is bounded above by $nN_n \le \exp(\log n + n\epsilon_n^2) \le e^{2n\epsilon_n^2}$, for sufficiently large n. This verifies the entropy condition (ii) of Theorem 8.9 with the sieve taken equal to $\mathcal{Q}$ and ϵ_n taken equal to four times the present ϵ_n. For this sieve the remaining mass condition (iii) is trivially satisfied.

If u is the upper limit of the ϵ_n-bracket containing p_0, then

$$\frac{p_0}{u/\int u\, d\nu} \le \int u\, d\nu \le (1 + \epsilon_n)^2 \le 2.$$

It follows that for large n, the set of points p such that $d_H^2(p, p_0)\|p_0/p\|_\infty \le 8\epsilon_n^2$ contains at least the function $u/\int u\, d\nu$ and hence has prior mass at least

$$\lambda_n \frac{1}{N_n} \ge \exp[-n\epsilon_n^2 - O(\log n)] \ge e^{-2n\epsilon_n^2},$$

for large n. In virtue of (8.8), this verifies the prior mass condition (i) of Theorem 8.9 for ϵ_n, a multiple of the present ϵ_n and with the neighborhoods $B_k(p_0; \epsilon_n)$ instead of $B_2(p_0; \epsilon_n)$, for any $k \ge 2$. Thus Theorem 8.9 gives the contraction rate ϵ_n in the strong sense. □

Example 8.16 (Smooth densities) Suppose that $\mathcal{P}$ consists of all densities p such that $\sqrt{p}$ belongs to a fixed multiple of the unit ball of the Hölder class $\mathfrak{C}^\alpha[0, 1]$, for some fixed $\alpha > 0$ (see Definition C.4). By Proposition C.5, the entropy of this unit ball relative to the uniform norm, and hence also relative to the weaker $\mathbb{L}_2(\nu)$-norm, is bounded by a multiple of $\epsilon^{-1/\alpha}$. Thus the Hellinger bracketing entropy (see Appendix C) of $\mathcal{P}$ possesses the same upper bound, and $\epsilon_n \asymp n^{-\alpha/(2\alpha+1)}$ satisfies the relation $\log N_{[\,]}(\epsilon_n, \mathcal{P}, d_H) \le n\epsilon_n^2$ and so does the upper bracketing entropy. Thus the prior based on upper brackets achieves the posterior contraction rate $n^{-\alpha/(2\alpha+1)}$. This rate is known to be the frequentist optimal rate for point estimators of a density in a Hölder ball.

Example 8.17 (Monotone densities) Suppose that $\mathcal{P}$ consists of all monotone decreasing densities on a compact interval in $\mathbb{R}$, bounded above by a fixed constant. Since the square root of a monotone density is again monotone, the bracketing entropy of $\mathcal{P}$ for the Hellinger distance is bounded by the bracketing $\mathbb{L}_2$-entropy of the set of monotone functions, which grows like ϵ^{-1}, in view of Proposition C.8. This leads to an $n^{-1/3}$-rate of contraction of the posterior. Again this agrees with the optimal frequentist rate for the problem, and hence it cannot be improved.

Theorem 8.15 implicitly requires that the model $\mathcal{P}$ is totally bounded for d_H. When $\mathcal{P}$ is a countable union of totally bounded models, a simple modification works, provided that we use a sequence of priors. To this end, suppose that there exist subsets $\mathcal{P}_n \uparrow \mathcal{P}$ with finite upper bracketing numbers, and let ϵ_n be numbers such that $\log N_](\epsilon_n, \mathcal{P}_n, d_H) \le n\epsilon_n^2$, for every n. Now construct Π_n as before with $\mathcal{P}$ replaced by $\mathcal{P}_n$. Next, do not form a mixture prior, but use Π_n itself as the prior distribution (different for different sample sizes). Then, as before, the corresponding posteriors achieve the contraction rate ϵ_n (see Theorem 8.24).

The entropy condition (ii) of Theorem 8.9 is in terms of metric entropy only, whereas presently we use brackets. The brackets give control over likelihood ratios, and are used for verifying the prior mass condition only. If likelihood ratios can be controlled by other means, then ordinary entropy will do also in Theorem 8.15. For instance, it suffices that the densities are uniformly bounded away from zero and infinity; the quotients p_0/p are then uniformly bounded. In the absence of such uniform bounds, ordinary metric entropy can still be used if the maximum of the set of densities is integrable, but a small (logarithmic) price has to be paid in the contraction rate.

To show this, let m be a ν-integrable *envelope function* for the set of densities $\mathcal{P}$; thus $p(x) \le m(x)$ for every $p \in \mathcal{P}$ and every x. Let $\{p_{1,n}, \ldots, p_{N_n,n}\}$ be a minimal ϵ_n-net over $\mathcal{P}$, and put $g_{j,n} = (p_{j,n}^{1/2} + \epsilon_n m^{1/2})^2/c_{j,n}$, where $c_{j,n}$ is the constant that normalizes $g_{j,n}$ to a probability density. Let Π_n be the uniform discrete measure on $g_{1,n}, \ldots, g_{N_n,n}$, and let $\Pi = \sum_{n=1}^{\infty} \lambda_n \Pi_n$ be a convex combination of the Π_n.

Theorem 8.18 *Assume that $\mathcal{P}$ has a ν-integrable envelope m, and construct Π as indicated for $\epsilon_n \downarrow 0$ with $n\epsilon_n^2/\log n \to \infty$ satisfying $\log N(\epsilon_n, \mathcal{P}, d_H) \le n\epsilon_n^2$, and strictly positive weights λ_n with $\log \lambda_n^{-1} = O(\log n)$. Then the posterior contracts at the rate $\epsilon_n(\log n)^{1/2}$ in probability, relative to the Hellinger distance d_H.*

Proof It can be verified that for $p \in \mathcal{P}$ with $d_H(p, p_{j,n}) \le \epsilon_n$,

$$d_H(p, g_{j,n}) = O(\epsilon_n), \qquad \frac{p}{g_{j,n}} = O(\epsilon_n^{-2}). \tag{8.16}$$

Then Lemma B.2 implies that $K(p; g_{j,n}) = O(\epsilon_n^2 \log_- \epsilon_n)$. This verifies the prior mass condition (i) of Theorem 8.11 with ϵ_n replaced by a multiple of $\epsilon_n(\log n)^{1/2}$, since $\log_- \epsilon_n = O(\log n)$ by the condition $n\epsilon_n^2/\log n \to \infty$. The rest of the proof is the same as the proof of Theorem 8.15. □

8.3 General Observations

In this section we return to the general setup of the introduction of the present chapter, involving experiments $(\mathfrak{X}^{(n)}, \mathscr{X}^{(n)}, P_\theta^{(n)}: \theta \in \Theta_n)$, with parameter spaces Θ_n, observations $X^{(n)}$ and true parameters $\theta_{n,0} \in \Theta_n$. The following theorem applies whenever exponentially powerful tests exist, and it is not otherwise restricted to a particular structure of the experiments.

For each n, let d_n and e_n be two semimetrics on Θ_n with the property: there exist universal constants $\xi, K > 0$ such that for every $\epsilon > 0$ and every $\theta_{n,1} \in \Theta_n$ with $d_n(\theta_{n,1}, \theta_{n,0}) > \epsilon$, there exists a test ϕ_n such that

$$P^{(n)}_{\theta_{n,0}}\phi_n \le e^{-Kn\epsilon^2}, \qquad \sup_{\theta\in\Theta_n: e_n(\theta,\theta_{n,1})<\xi\epsilon} P^{(n)}_\theta(1-\phi_n) \le e^{-Kn\epsilon^2}. \tag{8.17}$$

Typically, we have $d_n \le e_n$, and in many cases we choose $d_n = e_n$, but using two semimetrics adds flexibility. Apart from this, the condition is a direct generalization of (8.2) to general experiments.

As in the i.i.d. case, the behavior of posterior distributions depends on the concentration rate of the prior at the true parameter and the *Le Cam dimension* of the parameter set. For K and $V_{k,0}$ the Kullback-Leibler divergence and variation, let

$$B_{n,0}(\theta_{n,0},\epsilon) = \Big\{\theta\in\Theta_n: K(p^{(n)}_{\theta_{n,0}}; p^{(n)}_\theta) \le n\epsilon^2\Big\}, \tag{8.18}$$

$$B_{n,k}(\theta_{n,0},\epsilon) = \Big\{\theta\in\Theta_n: K(p^{(n)}_{\theta_{n,0}}; p^{(n)}_\theta) \le n\epsilon^2,\ V_{k,0}(p^{(n)}_{\theta_{n,0}}; p^{(n)}_\theta) \le n^{k/2}\epsilon^k\Big\}, \quad (k>0). \tag{8.19}$$

Theorem 8.19 (Contraction rate) *Let d_n and e_n be semimetrics on Θ_n for which there exist tests satisfying* (8.17)*, and let $\Theta_{n,1}\subset\Theta_n$ be arbitrary. If for constants ϵ_n with $n\epsilon_n^2 \ge 1$, for every sufficiently large $j\in\mathbb{N}$,*

$$\text{(i)}\ \frac{\Pi_n(\theta\in\Theta_{n,1}: j\epsilon_n < d_n(\theta,\theta_{n,0}) \le 2j\epsilon_n)}{\Pi_n(B_{n,0}(\theta_{n,0},\epsilon_n))} \le e^{Kn\epsilon_n^2 j^2/2}, \tag{8.20}$$

$$\text{(ii)}\ \sup_{\epsilon>\epsilon_n} \log N\Big(\xi\epsilon, \{\theta\in\Theta_{n,1}: d_n(\theta,\theta_{n,0}) < 2\epsilon\}, e_n\Big) \le n\epsilon_n^2, \tag{8.21}$$

then $\Pi_n(\theta\in\Theta_{n,1}: d_n(\theta,\theta_{n,0}) \ge M_n\epsilon_n \mid X^{(n)}) \to 0$, in $P^{(n)}_{\theta_{n,0}}$-probability, for every $M_n\to\infty$. Furthermore, if (8.20) *holds with $B_{n,0}(\theta_{n,0},\epsilon_n)$ replaced by $B_{n,k}(\theta_{n,0},\epsilon_n)$ for some $k>1$, then*

(a) *If all $X^{(n)}$ are defined on the same probability space, $\epsilon_n \gtrsim n^{-\alpha}$ for some $\alpha\in(0,1/2)$ and $k(1-2\alpha)>2$, then the contraction holds also in the almost sure sense.*
(b) *If $n\epsilon_n^2\to\infty$, then $\Pi_n(\theta\in\Theta_{n,1}: d_n(\theta,\theta_{n,0}) \ge M\epsilon_n \mid X^{(n)}) = O_P(e^{-n\epsilon_n^2})$, for sufficiently large M.*

Theorem 8.19 employs subsets $\Theta_{n,1}\subset\Theta_n$ to alleviate the entropy condition (8.21), but returns an assertion about the posterior distribution on $\Theta_{n,1}$ only. A complementary assertion about the sets $\Theta_{n,2} = \Theta_n\setminus\Theta_{n,1}$ may be obtained either by a direct argument or by the following result.

Theorem 8.20 (Remaining mass) *For any sets $\Theta_{n,2}\subset\Theta_n$ we have $\Pi_n(\Theta_{n,2}\mid X^{(n)})\to 0$, in $P^{(n)}_{\theta_{n,0}}$-probability if, for arbitrary ϵ_n, either (iii) or (iii′) holds:*

$$\text{(iii)}\ \frac{\Pi_n(\Theta_{n,2})}{\Pi_n(B_{n,0}(\theta_{n,0},\epsilon_n))} = o(e^{-D_n n\epsilon_n^2}), \qquad \textit{for some } D_n\to\infty,$$

$$\text{(iii}'\text{)}\ \frac{\Pi_n(\Theta_{n,2})}{\Pi_n(B_{n,k}(\theta_{n,0},\epsilon_n))} = o(e^{-2n\epsilon_n^2}), \qquad \textit{for some } k>1.$$

Under (iii′) the convergence to zero can be strengthened to almost sure convergence and to be of order $O_P(e^{-n\epsilon_n^2})$ under conditions (a)–(b) of Theorem 8.19.

The proofs of Theorems 8.19 and 8.20 are very similar to the proof of Theorem 8.11, apart from changes in the notation and the fact that we cover d_n-shells by e_n-balls to exploit the two metrics. The lower bound on the norming constant of the posterior distribution is based on the following lemma, which is proved in the same way as Lemmas 6.26 (when $k = 0$) and 8.10 (when $k > 1$).

Lemma 8.21 (Evidence lower bound) *For any $D > 0$, and $\epsilon \geq n^{-1/2}$,*

$$P^{(n)}_{\theta_{n,0}}\left(\int \frac{p^{(n)}_\theta}{p^{(n)}_{\theta_{n,0}}}\,d\Pi_n(\theta) \leq \Pi_n(B_{n,k}(\theta_{n,0},\epsilon))e^{-(1+D)n\epsilon^2}\right) \leq \begin{cases} 2/(1+D), & \text{if } k = 0,\\ (D\sqrt{n}\epsilon)^{-k}, & \text{if } k > 1.\end{cases}$$

As in the i.i.d. case weaker, but simpler theorems are obtained by using an absolute lower bound on the prior mass and by replacing the local by the global entropy. In particular, conditions (8.20) and (8.21) are implied by, for some $C > 0$,

$$\Pi_n(B_{n,0}(\theta_{n,0},\epsilon_n)) \geq e^{-Cn\epsilon_n^2}, \tag{8.22}$$

$$\log N(\xi\epsilon_n, \Theta_{n,1}, e_n) \leq n\epsilon_n^2. \tag{8.23}$$

Condition (8.22) leads to (8.20) (for $Kj^2/2 \geq C$) by bounding the numerator in the latter condition by the trivial bound one. Simple sufficient conditions in terms of only $\Pi_n(\Theta_{n,2})$ for (iii) and (iii′) in Theorem 8.20 can be similarly derived.

Bounding entropy to construct tests is not always the most fruitful strategy. As sometimes ad hoc tests can be easily constructed, a theorem using testing conditions is useful.

Theorem 8.22 *Given a semimetric d_n on Θ_n and subsets $\Theta_{n,1} \subset \Theta_n$, suppose that for $k > 1$ and ϵ_n with $n\epsilon_n^2 \geq 1$ condition* (8.20) *holds and there exist tests ϕ_n such that, for some $K > 0$ and sufficiently large j,*

$$P^{(n)}_{\theta_{n,0}}\phi_n \to 0, \qquad \sup_{\theta\in\Theta_{n,1}: j\epsilon_n < d_n(\theta,\theta_{n,0}) \leq 2j\epsilon_n} P^{(n)}_\theta(1-\phi_n) \lesssim e^{-Kj^2n\epsilon_n^2}, \tag{8.24}$$

Then $\Pi_n(\theta \in \Theta_{n,1}: d_n(\theta,\theta_{n,0}) \geq M_n\epsilon_n \mid X^{(n)}) \to 0$, in $P^{(n)}_{\theta_{n,0}}$-probability, for every $M_n \to \infty$.

8.3.1 Independent Observations

For $X^{(n)}$ consisting of n independent observations, with laws $P_{n,\theta,1},\ldots,P_{n,\theta,n}$, tests satisfying (8.17) always exist relative to $d_n = e_n$ equal to the *root average square Hellinger metric*

$$d_{n,H}(\theta_1,\theta_2) = \sqrt{\frac{1}{n}\sum_{i=1}^n d_H^2(P_{n,\theta_1,i}, P_{n,\theta_2,i})}. \tag{8.25}$$

(See Proposition D.9.) Furthermore, by Lemma B.8, the Kullback-Leibler divergence is additive and the variation $V_{k,0}$ subadditive up to an $O(n^{k/2-1})$ factor (for $k \geq 2$). It follows that the neighborhoods $B_{n,0}(\theta_{n,0}, \epsilon)$ and $B_{n,k}(\theta_{n,0}, \epsilon)$ contain the sets

$$B^*_{n,0}(\theta_{n,0}, \epsilon) = \Big\{\theta \in \Theta_n : \frac{1}{n}\sum_{i=1}^{n} K(P_{i,\theta_{n,0},n}; P_{i,\theta,n}) \leq \epsilon^2\Big\},$$

$$B^*_{n,k}(\theta_{n,0}, \epsilon) = \Big\{\theta \in \Theta_n : \frac{1}{n}\sum_{i=1}^{n} K(P_{i,\theta_{n,0},n}; P_{i,\theta,n}) \leq \epsilon^2, \frac{1}{n}\sum_{i=1}^{n} V_{k,0}(P_{i,\theta_{n,0},n}; P_{i,\theta,n}) \leq d_k\epsilon^k\Big\}.$$

The choices $k = 0$ or $k = 2$ (with $d_2 = 1$) are convenient and good enough for many applications. This leads to the following special case of Theorem 8.19.

Theorem 8.23 *Let $P^{(n)}_\theta$ be the product of measures $P_{n,\theta,1}, \ldots, P_{n,\theta,n}$, and suppose that there exists a partition $\Theta_n = \Theta_{n,1} \cup \Theta_{n,2}$ such that for sequences $\bar\epsilon_n, \epsilon_n \geq n^{-1/2}$ and all sufficiently large j,*

$$\text{(i)}\ \frac{\Pi_n(\theta \in \Theta_{n,1} : j\epsilon_n < d_{n,H}(\theta, \theta_{n,0}) \leq 2j\epsilon_n)}{\Pi_n(B^*_{n,0}(\theta_{n,0}, \epsilon_n))} \leq e^{n\epsilon_n^2 j^2/4}, \tag{8.26}$$

$$\text{(ii)}\ \sup_{\epsilon > \epsilon_n} \log N\Big(\epsilon/36, \{\theta \in \Theta_{n,1} : d_{n,H}(\theta, \theta_{n,0}) < \epsilon\}, d_{n,H}\Big) \leq n\epsilon_n^2, \tag{8.27}$$

$$\text{(iii)}\ \frac{\Pi_n(\Theta_{n,2})}{\Pi_n(B^*_{n,k}(\theta_{n,0}, \bar\epsilon_n))} = o(e^{-D_{n,k} n\bar\epsilon_n^2}), \tag{8.28}$$

for $k = 0$ and some $D_{n,0} \to \infty$, or for some $k \geq 2$ and $D_{n,k} = 2$. Then the rate of posterior contraction at $\theta_{n,0}$ relative to $d_{n,H}$ is ϵ_n.

The root average square Hellinger distance $d_{n,H}$ always works with independent observations, but it is not always the best choice of metric to derive contraction rates. Rates relative to a stronger metric d_n may be obtainable by applying Theorem 8.19 directly. This is possible as soon as tests as in (8.17) exist for these metrics (and metrics e_n). An important example where this is relevant is nonparametric regression with normal errors, as considered in Section 8.3.2.

Discrete Uniform Priors

As in the i.i.d. case, priors attaining the minimax rate of contraction may be constructed using bracketing methods. Consider a sequence of models $\{P^{(n)}_\theta : \theta \in \Theta\}$ with fixed parameter set Θ, where $P^{(n)}_\theta$ has a density $(x_1, \ldots, x_n) \mapsto \prod_{i=1}^n p_{\theta,i}(x_i)$ with respect to a product measure $\otimes_{i=1}^n \nu_i$. For a given n and $\epsilon > 0$ define the *componentwise Hellinger upper bracketing number* $N_{]}^{n\otimes}(\epsilon, \Theta_n, d_{n,H})$ for a set $\Theta_n \subset \Theta$ as the smallest number N of vectors $(u_{j,1}, \ldots, u_{j,n})$ of ν_j-integrable nonnegative functions $u_{j,i}$ (for $j = 1, 2, \ldots, N$), with the property that for every $\theta \in \Theta_n$ there exists some j such that $p_{\theta,i} \leq u_{j,i}$ for all $i = 1, 2, \ldots, n$ and $\sum_{i=1}^n d_H^2(p_{\theta,i}, u_{j,i}) \leq n\epsilon^2$.

Given a sequence of sets $\Theta_n \uparrow \Theta$ and $\epsilon_n \to 0$ such that $\log N_{]}^{n\otimes}(\epsilon_n, \Theta_n, d_{n,H}) \leq n\epsilon_n^2$, let $\{(u_{j,1}, \ldots, u_{j,n}) \colon j = 1, 2, \ldots, N\}$ be a minimal componentwise Hellinger upper bracketing for Θ_n. Let Π_n be the uniform discrete measure on the joint densities

$$p_j^{(n)}(x_1, \ldots, x_n) = \prod_{i=1}^{n} \frac{u_{j,i}(x_i)}{\int u_{j,i}\, d\nu_i}.$$

Theorem 8.24 *If $\Theta_n \uparrow \Theta$ are sets satisfying $\log N_{]}^{n\otimes}(\epsilon_n, \Theta_n, d_{n,H}) \leq n\epsilon_n^2$ for some ϵ_n with $n\epsilon_n^2 \to \infty$ and the prior is constructed as indicated, then the posterior contraction rate relative to $d_{n,H}$ is ϵ_n at any $\theta_0 \in \Theta$.*

Proof The collection $\{p_j^{(n)} \colon j = 1, 2, \ldots, N\}$ forms a sieve for the models $\{P_\theta^{(n)} \colon \theta \in \Theta\}$. Formally, we can add the measures $p_j^{(n)}$ to the model, identified by new parameters $j = 1, \ldots, N_n$, and apply Theorem 8.23 with the partition of the enlarged parameter space in $\Theta_{n,1} = \{1, \ldots, N_n\}$ the set of new parameters and $\Theta_{n,2} = \Theta_n \setminus \Theta_{n,1}$.

The covering number of $\Theta_{n,1} = \{1, \ldots, N_n\}$ certainly does not exceed its cardinality, which is $N_{]}^{n\otimes}(\epsilon_n, \Theta_n, d_{n,H})$ by construction. Therefore (8.27), with ϵ_n a multiple of the present ϵ_n, holds by assumption. Condition (8.28) holds trivially as $\Pi_n(\Theta_{n,2}) = 0$.

Because $\Theta_n \uparrow \Theta$, any $\theta_0 \in \Theta$ is in Θ_n for all sufficiently large n. For such large n, let j_0 be the index for which $p_{\theta_0,i} \leq u_{j_0,i}$ for every i, and $\sum_{i=1}^n d_H^2(p_{\theta_0,i}, u_{j_0,i}) \leq n\epsilon_n^2$. If p is a probability density, u is an integrable function such that $u \geq p$, and $v = u/\int u\, d\nu$, then

$$d_H^2(p, v) = 2 - \frac{2\int \sqrt{pu}\, d\nu}{\sqrt{\int u\, d\nu}} \leq \frac{1 + \int u\, d\nu - 2\int \sqrt{pu}\, d\nu}{\sqrt{\int u\, d\nu}} \leq \frac{d_H^2(p, u)}{\sqrt{\int u\, d\nu}}.$$

Because the likelihood ratios satisfy $p_{\theta_0,i}/v_{j_0,i} \leq \int u_{j_0,i}\, d\nu \leq 2$, it follows with the help of Lemma B.2 that $n^{-1}\sum_{i=1}^n K(p_{\theta_0,i}; v_{j_0,i}) \lesssim \epsilon_n^2$ and $n^{-1}\sum_{i=1}^n V_2(p_{\theta_0,i}; v_{j_0,i}) \lesssim \epsilon_n^2$. As $\otimes_{i=1}^n v_{j_0,i}$ receives prior probability equal to $N^{-1} \geq e^{-n\epsilon_n^2}$, the prior mass condition (8.26) holds, for a multiple of the present ϵ_n. □

Finite-dimensional Models

Application of Theorem 8.23 to finite-dimensional models yields the following result.

Theorem 8.25 *Let $X^{(n)}$ consist of independent observations $X_1, \ldots, X_n$ following densities $p_{\theta,i}$, indexed by parameters in $\Theta \subset \mathbb{R}^k$ such that for constants $\alpha, c, C > 0$, and $c_i \leq C_i$ with $c \leq \bar{c}_n \leq \bar{C}_n \leq C$, for every $\theta, \theta_1, \theta_2 \in \Theta$,*

$$K(p_{\theta_0,i}; p_{\theta,i}) \leq C_i \|\theta - \theta_0\|^{2\alpha}, \qquad V_{2,0}(p_{\theta_0,i}; p_{\theta,i}) \leq C_i \|\theta - \theta_0\|^{2\alpha}, \tag{8.29}$$

$$c_i \|\theta_1 - \theta_2\|^{2\alpha} \leq d_H^2(p_{\theta_1,i}, p_{\theta_2,i}) \leq C_i \|\theta_1 - \theta_2\|^{2\alpha}. \tag{8.30}$$

If the prior Π possesses a bounded density that is bounded away from zero on a neighborhood of θ_0, then the posterior converges at the rate $n^{-1/(2\alpha)}$ with respect to the Euclidean metric.

Proof By assumption (8.30), it suffices to show that the posterior contraction rate with respect to $d_{n,H}$, as defined in (8.25) is $n^{-1/2}$. Now by Proposition C.2,

$$N(\epsilon/18, \{\theta \in \Theta: d_{n,H}(\theta,\theta_0) < \epsilon\}, d_{n,H})$$
$$\leq N\Big(\frac{\epsilon^{1/\alpha}}{(36C)^{1/(2\alpha)}}, \{\theta \in \Theta: \|\theta-\theta_0\| < \frac{(\sqrt{2}\epsilon)^{1/\alpha}}{c^{1/(2\alpha)}}\}, \|\cdot\|\Big) \leq 3^k\Big(\frac{72C}{c}\Big)^{k/(2\alpha)}.$$

This verifies (8.27). For (8.26), note that

$$\frac{\Pi(\theta: d_{n,H}(\theta,\theta_0) \leq j\epsilon)}{\Pi(B^*_{n,2}(\theta_0,\epsilon))} \leq \frac{\Pi(\theta: \|\theta-\theta_0\| \leq \big(2j^2\epsilon^2/c\big)^{1/(2\alpha)})}{\Pi(\theta: \|\theta-\theta_0\| \leq \big(\epsilon^2/(2C)\big)^{1/(2\alpha)})} \leq Aj^{k/\alpha},$$

for sufficiently small $\epsilon > 0$, where A is a constant depending on d, c, C and the upper and lower bound on the prior density. It follows that (8.26) is satisfied for ϵ_n equal to a large multiple of $n^{-1/2}$. □

For regular parametric families, the conditions of the theorem are satisfied for $\alpha = 1$, and the usual $n^{-1/2}$ rate is obtained. When the densities have discontinuities or other singularities depending on the parameter (for instance the uniform distribution on $(0,\theta)$, or the examples discussed in Chapters V and VI of Ibragimov and Has'minskiĭ, 1981), then the conditions of Theorem 8.25 hold for $\alpha < 1$, and contraction rates faster than $n^{-1/2}$ pertain.

In an unbounded parameter space the Hellinger distance cannot be bounded below by a power of the Euclidean distance and hence the lower inequality of assumption (8.30) fails. In this case Theorem 8.22 may be applied after first proving that the posterior distribution concentrates on bounded sets. By the consistency theorems of Chapter 6 the latter can be achieved by showing the existence of uniformly exponentially consistent tests for $\theta = \theta_0$ against the complement of a bounded set. Often such tests exist by bounds on the log affinity, in view of Lemma D.11.

8.3.2 Gaussian Regression with Fixed Design

The observation $X^{(n)}$ in the Gaussian, fixed-design regression model consists of independent random variables $X_1, \ldots, X_n$ distributed as $X_i = \theta(z_i) + \varepsilon_i$, for an unknown regression function θ, deterministic covariates $z_1, \ldots, z_n$, and $\varepsilon_i \overset{\text{iid}}{\sim} \text{Nor}(0,\sigma^2)$, for $i = 1, \ldots, n$. Write $\|\theta\|_{2,n}$ for the $\mathbb{L}_2$-norm of a function $\theta: \mathfrak{Z} \to \mathbb{R}$ relative to the empirical measure of the covariates:

$$\|\theta\|_{2,n}^2 = \frac{1}{n}\sum_{i=1}^n \theta^2(z_i).$$

By Lemma 2.7, the average square Hellinger distance and the Kullback-Leiber divergence and variation on the distributions of the observations are all bounded above by the square of the $\|\cdot\|_{2,n}$-norm. However, for the Hellinger distance this can be reversed (in general) only if the regression functions are bounded. This makes it better to deduce a rate of contraction directly from the general rate theorem, Theorem 8.19, with an ad hoc construction of tests, rather than to refer to Theorem 8.23.

Theorem 8.26 (Regression) *Let $P_\theta^{(n)}$ be the product of the measures* $\mathrm{Nor}(\theta(z_i), \sigma^2)$, *for $i = 1, \ldots, n$. If there exists a partition $\Theta_n = \Theta_{n,1} \cup \Theta_{n,2}$ such that for a sequence $\epsilon_n \to 0$ with $n\epsilon_n^2 \geq 1$ and all sufficiently large j,*

$$\text{(i)}\ \frac{\Pi_n(\theta \in \Theta_{n,1}\colon j\epsilon_n < \|\theta - \theta_{n,0}\|_{2,n} \leq 2j\epsilon_n)}{\Pi_n(\theta \in \Theta_n\colon \|\theta - \theta_{n,0}\|_{2,n} < \epsilon_n)} \leq e^{n\epsilon_n^2 j^2/16}, \tag{8.31}$$

$$\text{(ii)}\ \sup_{\epsilon > \epsilon_n} \log N\Big(\epsilon/2, \{\theta \in \Theta_{n,1}\colon \|\theta - \theta_{n,0}\|_{2,n} < 2\epsilon\}, \|\cdot\|_n\Big) \leq n\epsilon_n^2, \tag{8.32}$$

$$\text{(iii)}\ \frac{\Pi_n(\Theta_{n,2})}{\Pi_n(\theta \in \Theta_n\colon \|\theta - \theta_{n,0}\|_{2,n} < \epsilon_n)} = o(e^{-2n\epsilon_n^2}), \tag{8.33}$$

then the rate of posterior contraction at $\theta_{n,0}$ relative to $\|\cdot\|_{2,n}$ is ϵ_n.

Proof We combine Theorems 8.19 and 8.20, with $k = 2$. Because by Lemma 2.7 the neighborhoods $B_{n,2}(\theta_{n,0}, \epsilon)$ in these theorems contain balls $\{\theta\colon \|\theta - \theta_{n,0}\|_{2,n} < \epsilon\}$, it suffices to exhibit tests as in (8.17), with d_n and e_n taken equal to the metric induced by $\|\cdot\|_{2,n}$. A likelihood ratio test for testing $\theta_{n,0}$ versus $\theta_{n,1}$ with appropriate cutoff is shown to do the job in Lemma 8.27(i) below, with $\xi = 1/2$ and $K = 1/32$. □

The following lemma verifies the existence of tests as in (8.17). Part (i) is used in the proof of the preceding theorem. Part (ii) may be used to extend the theorem to the case of unknown error variance. If the prior of σ has bounded support $[0, \bar\sigma]$, for a constant $\bar\sigma$, then Theorem 8.19 applies with $d_n = e_n$ and $d_n((\theta_1, \sigma_1), (\theta_2, \sigma_2))$ taken equal to a multiple of $\|\theta_1 - \theta_2\|_{2,n} + |\sigma_1 - \sigma_2|$.

Lemma 8.27 *For $\theta \in \mathbb{R}^n$ and $\sigma > 0$, let $P_{\theta,\sigma}^{(n)} = \mathrm{Nor}_n(\theta, \sigma^2 I)$, and let $\|\theta\|$ be the Euclidean norm.*

(i) *For any $\theta_0, \theta_1 \in \mathbb{R}^n$ and $\sigma > 0$, there exists a test ϕ such that $P_{\theta_0,\sigma}^{(n)}\phi \vee P_{\theta,\sigma}^{(n)}(1 - \phi) \leq e^{-\|\theta_0 - \theta_1\|^2/(32\sigma^2)}$, for every $\theta \in \mathbb{R}^n$ with $\|\theta - \theta_1\| \leq \|\theta_0 - \theta_1\|/2$.*

(ii) *For any $\theta_0, \theta_1 \in \mathbb{R}^n$ and $\sigma_0, \sigma_1 > 0$, there exists a test ϕ such that $P_{\theta_0,\sigma_0}^{(n)}\phi \vee P_{\theta,\sigma}^{(n)}(1 - \phi) \leq e^{-K[\|\theta_0 - \theta_1\|^2 + n|\sigma_0 - \sigma_1|^2]/(\sigma_0^2 \vee \sigma_1^2)}$, for every $\theta \in \mathbb{R}^n$ and $\sigma > 0$ with $\|\theta - \theta_1\| \leq \|\theta_0 - \theta_1\|/2$ and $|\sigma - \sigma_1| \leq |\sigma_0 - \sigma_1|/2$, and a universal constant K.*

Proof (i). By shifting the observation $X \sim \mathrm{Nor}(\theta, \sigma^2 I)$ by $-\theta_0$ and dividing by σ, we can reduce to the case that $\theta_0 = 0$ and $\sigma = 1$. If $\|\theta - \theta_1\| \leq \|\theta_1\|/2$, then $\|\theta\| \geq \|\theta_1\|/2$ and hence $\theta_1^\top\theta = (\|\theta\|^2 + \|\theta_1\|^2 - \|\theta - \theta_1\|^2)/2 \geq \|\theta_1\|^2/2$. Therefore, the test $\phi = \mathbb{1}\{\theta_1^\top X > D\|\theta_1\|\}$ satisfies, with Φ the standard normal distribution function,

$$P_{\theta_0}^{(n)}\phi = 1 - \Phi(D), \qquad P_\theta^{(n)}(1 - \phi) = \Phi((D\|\theta_1\| - \theta_1^\top\theta)/\|\theta_1\|) \leq \Phi(D - \rho),$$

for $\rho = \|\theta_1\|/2$. The infimum over D of $1 - \Phi(D) + \Phi(D - \rho)$ is attained at $D = \rho/2$, for which $D - \rho = -\rho/2$. We substitute this in the preceding display and use the bound $1 - \Phi(x) \leq e^{-x^2/2}$, valid for $x \geq 0$ (see Lemma K.6(i)).

(ii). By shifting by $-\theta_0$ we again reduce to the case that $\theta_0 = 0$.

By the proof as under (i), the test $\phi = \mathbb{1}\{\theta_1^\top X > D\|\theta_1\|\}$ can be seen to have error probabilities satisfying $P_{\theta_0,\sigma_0}^{(n)}\phi \leq 1 - \Phi(D/\sigma_0)$ and $P_{\theta,\sigma}^{(n)}(1-\phi) \leq \Phi((D-\rho)/\sigma)$, for $\rho = \|\theta_1\|/2$. For $D = \rho/2$, both error probabilities are bounded by $e^{-\|\theta_0-\theta_1\|^2/(32\sigma_0^2\vee\sigma_1^2)}$.

Assume first that $\sigma_1 > \sigma_0$. The test $\psi = \mathbb{1}\{n^{-1}\|X\|_1 - c\sigma_0 > D\}$, where $c = \sqrt{2/\pi}$ is the absolute mean of a standard normal variable, has error probability of the first kind satisfying $P_{\theta_0,\sigma_0}^{(n)}\psi = \mathrm{P}(\|Z\|_1 - \mathrm{E}\|Z\|_1 > Dn/\sigma_0)$ for $Z \sim \mathrm{Nor}_n(0, I)$. Since $|\,\|X-\theta\|_1 - \|X\|_1| \leq \|\theta\|_1 \leq \sqrt{n}\|\theta\|$, by the Cauchy-Schwarz inequality, the error probability of the second kind satisfies

$$\begin{aligned}P_{\theta,\sigma}^{(n)}(1-\psi) &\leq P_{\theta,\sigma}^{(n)}(n^{-1}\|X-\theta\|_1 < D + c\sigma_0 + n^{-1/2}\|\theta\|)\\ &\leq \mathrm{P}(\|Z\|_1 - \mathrm{E}\|Z\|_1 \leq Dn/\sigma + nc(\sigma_0-\sigma)/\sigma + \sqrt{n}\|\theta\|/\sigma).\end{aligned}$$

For $D = c(\sigma_1-\sigma_0)/4$, the argument in the last probability is bounded above by $-cn(\sigma_1 - \sigma_0)/(4\sigma) + \sqrt{n}(3/2)\|\theta_1\|/\sigma$, whenever $|\sigma - \sigma_1| \leq |\sigma_0 - \sigma_1|/2$ and $\|\theta - \theta_1\| \leq \|\theta_1\|/2$, which is further bounded above by $-cn(\sigma_1 - \sigma_0)/(8\sigma) \leq -cn(\sigma_1 - \sigma_0)/(12\sigma_1)$, if also $\|\theta_1\| \leq \sqrt{n}(\sigma_1 - \sigma_0)/12$. Because $\|Z\|_1 - \mathrm{E}\|Z\|_1$ is the sum of independent, mean-zero sub-Gaussian variables, it satisfies a sub-Gaussian tail bound of the type $\mathrm{P}(|\,\|Z\|_1 - \mathrm{E}\|Z\|_1| > x) \leq e^{-K'x^2/n}$, for some constant $K' > 0$. Conclude that the error probabilities of ψ are bounded above by $e^{-Kn(\sigma_1-\sigma_0)^2/(\sigma_0\vee\sigma_1)^2}$ for some constant $K > 0$.

For $\sigma_1 < \sigma_0$, a test of the form $\psi = \mathbb{1}\{n^{-1}\|X\|_1 - c\sigma_0 < D\}$ can be seen to satisfy the same error bounds.

Finally, consider two cases. If $\|\theta_0 - \theta_1\| > \sqrt{n}|\sigma_0 - \sigma_1|/12$, then we use only the test ϕ, and it has the desired error bounds, for sufficiently small K. If $\|\theta_0 - \theta_1\| \leq \sqrt{n}|\sigma_0 - \sigma_1|/12$, then we use only the test ψ, and it has again the desired error bounds. □

8.3.3 Markov Chains

Let $P_\theta^{(n)}$ stand for the law of $(X_0, \ldots, X_n)$, for a discrete time Markov process $(X_n: n \in \mathbb{Z})$ with stationary transition probabilities, in a given state space $(\mathfrak{X}, \mathscr{X})$. Assume that the transition kernel is given by a transition density p_θ relative to a given dominating measure ν, and let $Q_{\theta,i}$ be the distribution of X_i. Thus $(x, y) \mapsto p_\theta(y|\,x)$ are measurable maps, and $p_\theta(\cdot|\,x)$ is a probability density relative to ν, for every x.

Assume that there exist finite measures $\underline{\mu}$ and $\overline{\mu}$ on $(\mathfrak{X}, \mathscr{X})$ such that, for some $k, l \in \mathbb{N}$, every $\theta \in \Theta$, and every $x \in \mathfrak{X}$ and $A \in \mathscr{X}$,

$$\underline{\mu}(A) \leq \frac{1}{k}\sum_{j=1}^{k} \mathrm{P}_\theta(X_j \in A|\,X_0 = x), \qquad \mathrm{P}_\theta(X_l \in A|\,X_0 = x) \leq \overline{\mu}(A). \tag{8.34}$$

Here P_θ is a generic notation for a probability law governed by θ. This holds in particular if there exists $r \in \mathbb{L}_1(\nu)$ such that $r \lesssim p_\theta(\cdot|\,x) \lesssim r$ for every x, with the measures $\underline{\mu}$ and $\overline{\mu}$ equal to multiples of $A \mapsto \int_A r\,d\nu$. For $\mu \in \{\underline{\mu}, \overline{\mu}\}$ define semimetrics $d_{H,\mu}$ by

$$d_{H,\mu}^2(\theta_1, \theta_2) = \int\!\!\int \Big[\sqrt{p_{\theta_1}(y|\,x)} - \sqrt{p_{\theta_2}(y|\,x)}\Big]^2\,d\nu(y)\,d\mu(x). \tag{8.35}$$

It is shown in Proposition D.14 that tests as in (8.17) exist for $d_n = d_{H,\underline{\mu}}$ and $e_n = d_{H,\overline{\mu}}$.

The following lemma relates the Kullback-Leibler divergence and variation on the laws $P_\theta^{(n)}$ to the corresponding characteristics of the transition densities. For the variation we assume that the Markov chain is α-mixing, for the divergence itself this is not necessary. The kth *α-mixing coefficient* α_k under θ_0 is defined as

$$\alpha_k = \sup_{m\in\mathbb{N}} \sup_{A,B\in\mathscr{X}} |\mathrm{P}_{\theta_0}(X_m \in A, X_{m+k} \in B) - \mathrm{P}_{\theta_0}(X_m \in A)\mathrm{P}_{\theta_0}(X_{m+k} \in B)|.$$

Lemma 8.28 *For every $s > 2$ and $D_s = 8s\sum_{h=0}^\infty \alpha_h^{1-2/s}$,*

(i) $K(p_{\theta_0}^{(n)}; p_\theta^{(n)}) \le n\sup_{i\ge 0}\int K(p_{\theta_0}(\cdot|\,x); p_\theta(\cdot|\,x))\,dQ_{\theta_0,i}(x) + K(q_{\theta_0,0}; q_{\theta,0})$.

(ii) $V_2(p_{\theta_0}^{(n)}; p_\theta^{(n)}) \le nD_s \sup_{i\ge 0}\big[\int V_s(p_{\theta_0}(\cdot|\,x); p_\theta(\cdot|\,x))\,dQ_{\theta_0,i}(x)\big]^{2/s} + 2V_2(q_{\theta_0,0}; q_{\theta,0})$.

Proof Assertion (i) is immediate from the expression for the likelihood:

$$\log\frac{p_{\theta_0}^{(n)}}{p_\theta^{(n)}}(X^{(n)}) = \sum_{i=1}^n \log\frac{p_{\theta_0}}{p_\theta}(X_i|\,X_{i-1}) + \log\frac{q_{\theta_0,0}}{q_{\theta,0}}(X_0) =: \sum_{i=1}^n Y_i + Z_0.$$

The time series $Y_i = \log(p_{\theta_0}/p_\theta)(X_i|\,X_{i-1})$ is α-mixing with mixing coefficients α_{k-1}. Consequently, the variance $\mathrm{var}(\sum_{i=1}^n Y_i)$ of their sum can be further bounded by the desired expression, in view of Lemma K.5. □

Let $\Theta_{0,1} \subset \Theta$ be the set of parameter values such that $K(q_{\theta_0,0}; q_{\theta,0})$ and $V_2(q_{\theta_0,0}; q_{\theta,0})$ are bounded by 1. Then by the preceding lemma, for large n and $n\epsilon^2 \ge 4$, the neighborhoods $B_{n,0}(\theta_0,\epsilon)$ and $B_{n,2}(\theta_0,\epsilon)$ contain the sets, respectively,

$$B_0^*(\theta_0,\epsilon) = \Big\{\theta\in\Theta_{0,1}: \sup_i \int K(p_{\theta_0}(\cdot|\,x); p_\theta(\cdot|\,x))\,dQ_{\theta_0,i}(x) \le \tfrac{1}{2}\epsilon^2\Big\},$$

$$B_s^*(\theta_0,\epsilon) = \Big\{\theta\in B_0^*(\theta_0,\epsilon): \sup_i \int V_s(p_{\theta_0}(\cdot|\,x); p_\theta(\cdot|\,x))\,dQ_{\theta_0,i}(x) \le (2D_s)^{s/2}\epsilon^s\Big\}.$$

Therefore, Theorem 8.19 implies the following result.

Theorem 8.29 *Let $P_\theta^{(n)}$ be the distribution of a Markov chain $(X_0, X_1, \ldots, X_n)$ with transition densities p_θ satisfying* (8.34) *and marginal densities $q_{\theta,i}$. If there exists a partition $\Theta = \Theta_{n,1}\cup\Theta_{n,2}$ such that for sequences $\epsilon_n, \bar\epsilon_n \ge 2n^{-1/2}$ and every sufficiently large j,*

$$\text{(i)}\quad \frac{\Pi_n(\theta\in\Theta_{n,1}: d_{H,\underline{\mu}}(\theta,\theta_0) \le j\epsilon_n)}{\Pi_n(B_0^*(\theta_0,\epsilon_n))} \le e^{Kn\epsilon_n^2 j^2/8}, \tag{8.36}$$

$$\text{(ii)}\quad \sup_{\epsilon>\epsilon_n} \log N\Big(\epsilon/16, \{\theta\in\Theta_{n,1}: d_{H,\underline{\mu}}(\theta,\theta_0) < \epsilon\}, d_{H,\overline{\mu}}\Big) \le n\epsilon_n^2, \tag{8.37}$$

$$\text{(iii)}\quad \frac{\Pi_n(\Theta_{n,2})}{\Pi_n(B_0^*(\theta_0,\bar\epsilon_n))} = o(e^{-M_n n\bar\epsilon_n^2}), \qquad \textit{for some } M_n\to\infty, \tag{8.38}$$

then the posterior rate of contraction at θ_0 relative to $d_{H,\underline{\mu}}$ is ϵ_n. If under θ_0 the chain is α-mixing with coefficients α_h satisfying $\sum_{k=0}^{\infty} \alpha_k^{1-2/s} < \infty$ for some $s > 2$, then this conclusion remains true if (iii) is replaced by

$$\text{(iii}'\text{)}\ \frac{\Pi_n(\Theta_{n,2})}{\Pi_n(B_s^*(\theta_0, \bar{\epsilon}_n))} = o(e^{-2n\bar{\epsilon}_n^2}). \tag{8.39}$$

Condition (iii′) may yield a slightly faster rate, but at the cost of introducing mixing numbers and higher moments of the log likelihood. Condition (iii) is easier to apply, but we finish the section with some remarks on the verification of (iii′).

As usual, there is a trade-off between mixing and moment conditions: for quickly mixing sequences, with convergence of the series $\sum_k \alpha_k^{1-2/s}$ for small s, existence of lower-order moments suffices.

From the Markov property it can be seen that, with $Q_{\theta_0,\infty}$ any measure,

$$\alpha_k \le 2 \sup_x d_{TV}(\mathrm{P}_{\theta_0}(X_k \in \cdot \,|\, X_0 = x), Q_{\theta_0,\infty}).$$

A Markov chain is said to be *uniformly ergodic* if the right side of this display, with $Q_{\theta_0,\infty}$ an invariant probability measure of the chain, tends to zero as $k \to \infty$. The convergence is then automatically exponentially fast: the right side is bounded by a multiple of c^k for some constant $c < 1$ (see Meyn and Tweedie 1993, Theorem 16.0.2). Then $\sum_{k=0}^{\infty} \alpha_k^{1-2/s} < \infty$ for every $s > 2$, and (iii′) of the preceding theorem is applicable with an arbitrary fixed $s > 2$.

A simple (but strong) condition for uniform ergodicitiy is:

$$\sup_{x_1,x_2 \in \mathfrak{X}} d_{TV}(p_{\theta_0}(\cdot \,|\, x_1), p_{\theta_0}(\cdot \,|\, x_2)) < 2.$$

This follows by integrating the display with respect to x_2 relative to the stationary measure Q_{θ_0}, and next applying condition (16.8) of Theorem 16.0.2 of Meyn and Tweedie (1993).

8.3.4 White Noise Model

For $\theta \in \Theta \subset \mathbb{L}_2[0,1]$, let $P_\theta^{(n)}$ be the distribution on $\mathfrak{C}[0,1]$ of the stochastic process $X^{(n)}$ defined structurally relative to a standard Brownian motion W as

$$X_t^{(n)} = \int_0^t \theta(s)\,ds + \frac{1}{\sqrt{n}} W_t, \qquad 0 \le t \le 1.$$

This model arises as an approximation of many different types of experiments, including density estimation and nonparametric regression (see Nussbaum 1996 and Brown and Low 1996).

An equivalent experiment is obtained by expanding $dX^{(n)}$ on an arbitrary orthonormal basis $e_1, e_2, \ldots$ of $\mathbb{L}_2[0,1]$, giving the variables $X_{n,i} = \int e_i(t)\, dX_t^{(n)}$. (The "differentials" $dX^{(n)}$ cannot be understood as elements of $\mathbb{L}_2[0,1]$ and hence "expanding" must be understood in the generalized sense of Itô integrals.) These variables are independent and normally distributed with means the coefficients $\theta_i := \int_0^1 e_i(t)\theta(t)\,dt$ in the expansion of $\theta = \sum_i \theta_i e_i$ relative to the basis (e_i), and variance $1/n$. Observing $X^{(n)}$ or the infinite vector $(X_{n,1}, X_{n,2}, \ldots)$ are equivalent by sufficiency of the latter vector in the experiment consisting of observing $X^{(n)}$. The parameter $(\theta_1, \theta_2, \ldots)$ belongs to ℓ_2 and the norms $\|\theta\|_2$ of the function $\theta \in \mathbb{L}_2[0,1]$ and the vector $(\theta_1, \theta_2, \ldots) \in \ell_2$ are equal.

By Lemma D.16, likelihood ratio tests satisfy the requirements of (8.17), for the metrics $d_n = e_n$ taken equal to the $\mathbb{L}_2$-norm. The Kullback-Leibler divergence and variation $V_{2,0}$ turn out to be multiples of the $\mathbb{L}_2$-norm as well.

Lemma 8.30 *For every $\theta, \theta_0 \in \Theta \subset \mathbb{L}_2[0,1]$,*

(i) $K(P_{\theta_0}^{(n)}; P_\theta^{(n)}) = \frac{1}{2} n\|\theta - \theta_0\|_2^2$.
(ii) $V_{2,0}(P_{\theta_0}^{(n)}; P_\theta^{(n)}) = n\|\theta - \theta_0\|_2^2$.

Proof The log likelihood ratio for the sequence formulation of the model takes the form

$$\log \frac{p_{\theta_0}^{(n)}}{p_\theta^{(n)})(X^{(n)}} = n \sum_{i=1}^{\infty} (\theta_0 - \theta)_i X_{n,i} - \frac{n}{2}\|\theta_0\|_2^2 + \frac{n}{2}\|\theta\|_2^2.$$

The mean and variance under θ_0 of this expression are the Kullback-Leibler divergence and variation. □

Lemma 8.30 implies that $B_{n,2}(\theta_0, \epsilon) = \{\theta \in \Theta \colon \|\theta - \theta_0\|_2 \leq \epsilon\}$, and we obtain the following corollary to Theorem 8.19.

Theorem 8.31 *Let $P_\theta^{(n)}$ be the distribution on $\mathfrak{C}[0,1]$ of the solution to the diffusion equation $dX_t = \theta(t)\,dt + n^{-1/2}\,dW_t$ with $X_0 = 0$. If there exists a partition $\Theta = \Theta_{n,1} \cup \Theta_{n,2}$ such that for ϵ_n with $n\epsilon_n^2 \geq 1$, and for every $j \in \mathbb{N}$,*

$$\text{(i)}\quad \frac{\Pi_n(\theta \in \Theta_{n,1} \colon \|\theta - \theta_0\|_2 \leq j\epsilon_n)}{\Pi_n(\theta \in \Theta \colon \|\theta - \theta_0\|_2 \leq \epsilon_n)} \leq e^{n\epsilon_n^2 j^2/64}, \tag{8.40}$$

$$\text{(ii)}\quad \sup_{\epsilon > \epsilon_n} \log N(\epsilon/8, \{\theta \in \Theta_{n,1} \colon \|\theta - \theta_0\|_2 < \epsilon\}, \|\cdot\|_2) \leq n\epsilon_n^2, \tag{8.41}$$

$$\text{(iii)}\quad \frac{\Pi_n(\Theta_{n,2})}{\Pi_n(\theta \in \Theta \colon \|\theta - \theta_0\|_2 \leq \epsilon_n)} = o(e^{-n\epsilon_n^2}), \tag{8.42}$$

then the posterior rate of contraction relative to $\|\cdot\|_2$ at θ_0 is given by ϵ_n.

The prior that models the coordinates θ_i as independent Gaussian variables is conjugate in this model, and allows explicit calculation of the posterior distribution. In Example 8.6 and Section 9.5.4 we calculate the rate of contraction by a direct method.

8.3.5 Gaussian Time Series

For $f \in \mathcal{F} \subset \mathbb{L}_2(-\pi, \pi)$, let $P_f^{(n)}$ be the distribution of $(X_1, \ldots, X_n)$ for a stationary, mean-zero Gaussian time series $(X_t \colon t \in \mathbb{Z})$ with spectral density f. Denote the corresponding autocovariance function by

$$\gamma_f(h) = \int_{-\pi}^{\pi} e^{ih\lambda} f(\lambda)\,d\lambda.$$

Assume that $\|\log f\|_\infty \le M$ and $\sum_{h=0}^\infty h\gamma_f^2(h) \le N$, for every $f \in \mathcal{F}$ and given constants M and N. The first condition is reasonable as it is known that the structure of the time series changes dramatically if the spectral density approaches zero or infinity. The second condition imposes smoothness on the spectral density.

Under these restrictions Proposition D.15 shows that condition (8.17) is satisfied, for d_n the metric of $\mathbb{L}_2(-\pi,\pi)$, and e_n the uniform metric on the spectral densities, and constants ξ and K depending on M and N only.

The following lemma relates the Kullback-Leibler divergence and variation to the $\mathbb{L}_2$-norm on the spectral densities, and implies that the neighborhoods $B_{n,2}(f_0,\epsilon)$ contain $\mathbb{L}_2$-neighborhoods.

Lemma 8.32 *There exists a constant C depending only on M, such that for every spectral densities f, g with $\|\log f\|_\infty \le M$ and $\|\log g\|_\infty \le M$,*

(i) $K(p_f^{(n)}; p_g^{(n)}) \le Cn\|f-g\|_2^2$.
(ii) $V_{2,0}(p_f^{(n)}; p_g^{(n)}) \le Cn\|f-g\|_2^2$.

Proof The covariance matrix $T_n(f)$ of $X^{(n)} = (X_1,\ldots,X_n)$ given the spectral density f is linear in f, and $T_n(1) = 2\pi I$. Using the matrix identities $\det(AB^{-1}) = \det(I + B^{-1/2}(A-B)B^{-1/2})$ and $A^{-1} - B^{-1} = A^{-1}(A-B)B^{-1}$, we can write the log likelihood ratio in the form

$$\log\frac{p_f^{(n)}}{p_g^{(n)}}(X^{(n)}) = -\tfrac{1}{2}\log\det\left(I + T_n(g)^{-1/2}T_n(f-g)T_n(g)^{-1/2}\right) \\ -\tfrac{1}{2}(X^{(n)})^\top T_n(f)^{-1}T_n(g-f)T_n(g)^{-1}X^{(n)}.$$

As the mean and variance of a quadratic form in a random vector X with mean zero and covariance matrix Σ take the forms $\mathrm{E}(X^\top AX) = \mathrm{tr}(\Sigma A)$ and $\mathrm{var}(X^\top AX) = \mathrm{tr}(\Sigma A\Sigma A) + \mathrm{tr}(\Sigma A\Sigma A^\top)$, it follows that

$$\begin{aligned} 2K(p_f^{(n)}; p_g^{(n)}) &= -\log\det\left(I + T_n(g)^{-1/2}T_n(f-g)T_n(g)^{-1/2}\right) \\ &\quad - \mathrm{tr}(T_n(g-f)T_n(g)^{-1}), \\ 4V_{2,0}(p_f^{(n)}; p_g^{(n)}) &= \mathrm{tr}(T_n(g-f)T_n(g)^{-1}T_n(g-f)T_n(g)^{-1}) \\ &\quad + \mathrm{tr}(T_n(g-f)T_n(g)^{-1}T_n(f)T_n(g)^{-1}T_n(g-f)T_n(f)^{-1}). \end{aligned}$$

For $\|A\|$ the *Frobenius norm* (given by $\|A\|^2 = \sum_k\sum_l a_{k,l}^2 = \mathrm{tr}(AA^\top)$) and $|A|$ the operator norm (given by $|A| = \sup\{\|Ax\|: \|x\| = 1\}$) of a matrix A, we have $\mathrm{tr}(A^2) \le \|A\|^2$ and $\|AB\| \le |A|\|B\|$. Furthermore $-\frac{1}{2}\mathrm{tr}(A^2) \le \log\det(I+A) - \mathrm{tr}(A) \le 0$ for any nonnegative-definite matrix A, as a result of the inequality $-\frac{1}{2}\mu^2 \le \log(1+\mu) - \mu \le 0$ for $\mu \ge 0$. Because $x^\top T_n(f)x = \int_{-\pi}^{\pi}|\sum_{k=1}^n x_k e^{ik\lambda}|^2 f(\lambda)\,d\lambda$, we have that the map $f \mapsto T_n(f)$ is monotone in the sense of positive-definiteness. In particular, $T_n(1)\inf_{-\pi<x<\pi} f(x) \le T_n(f) \le T_n(1)\|f\|_\infty$. Because $T_n(1)$ is equal to 2π times the identity matrix, we can derive that

$$|T_n(f)| \le 2\pi\|f\|_\infty \qquad |T_n(f)^{-1}| \le (2\pi)^{-1}\|1/f\|_\infty.$$

(For the second inequality we use that $|A^{-1}| \le c^{-1}$ if $\|Ax\| \ge c\|x\|$ for all x.) Furthermore,

$$\|T_n(f)\|^2 = \sum_{|h|<n} (n-|h|)\gamma_f^2(h) \le 2\pi n \int_{-\pi}^{\pi} f^2(\lambda)\,d\lambda. \tag{8.43}$$

Given the preceding bounds and the identity $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, derivation of the desired bounds on the mean and variance of $\log(p_f^{(n)}/p_g^{(n)})(X^{(n)})$ is now straightforward. □

The lemma allows us to reduce the prior mass condition to $\mathbb{L}_2$-neighborhoods. The following theorem gives the rate of contraction of the posterior distribution for the spectral density, relative to the $\mathbb{L}_2$-distance. The theorem is an immediate corollary of Theorem 8.19 and the preceding lemma.

Theorem 8.33 (Gaussian time series) *Let $P_f^{(n)}$ be the distribution of $(X_1,\ldots,X_n)$ for a stationary Gaussian time series $(X_t: t\in\mathbb{Z})$ with spectral density $f\in\mathcal{F}$. If there exist constants M and N such that $\|\log f\|_\infty \le M$ and $\sum_h |h|\gamma_f^2(h) \le N$ for every $f\in\mathcal{F}$, and there exist partitions $\mathcal{F} = \mathcal{F}_{n,1}\cup\mathcal{F}_{n,2}$ such that for numbers $\epsilon_n \ge n^{-1/2}$ and every $j\in\mathbb{N}$,*

(i) $\dfrac{\Pi(f\in\mathcal{F}_{n,1}: \|f-f_0\|_2 \le j\epsilon_n)}{\Pi(f\in\mathcal{F}: \|f-f_0\|_2\le\epsilon_n)} \lesssim e^{Kn\epsilon_n^2 j^2/8}$,

(ii) $\sup_{\epsilon\ge\epsilon_n} \log N(\xi\epsilon/2, \{f\in\mathcal{F}_{n,1}: \|f-f_0\|_2\le\epsilon\}, \|\cdot\|_\infty) \le n\epsilon_n^2$,

(iii) $\dfrac{\Pi(\mathcal{F}_{n,2})}{\Pi(f\in\mathcal{F}: \|f-f_0\|_2\le\epsilon_n)} = o(e^{-n\epsilon_n^2})$,

then the posterior rate of contraction relative to the $\mathbb{L}_2$-norm is ϵ_n at every $f_0\in\mathcal{F}$.

8.4 Lower Bounds

The theory so far gives upper bounds for posterior contraction rates, without a guarantee that the actual posterior contraction rate at a given true parameter is not faster. In applications, the theory is typically applied to obtain a rate of contraction for every parameter (uniformly) in a given model. If this compares to the minimax rate, then clearly the rate is sharp in the minimax sense. However, such a minimax rate refers to a worst parameter in a model, and the rate at a specific parameter may well be faster. To study this we define *lower bounds* on contraction rates as follows.

We adopt the general setting of a given sequence $(\mathfrak{X}^{(n)}, \mathscr{X}^{(n)}, P_\theta^{(n)}: \theta\in\Theta_n)$ of statistical experiments, with observations $X^{(n)}$, a prior distribution Π_n on the parameter set Θ_n, and a semimetric d on Θ_n.

Definition 8.34 (Contraction rate, lower bound) A sequence δ_n is a lower bound for the contraction rate at the parameter θ_0 with respect to the semimetric d if $\Pi_n(\theta: d(\theta,\theta_0) \le m_n\delta_n \,|\, X^{(n)}) \to 0$ in $P_{\theta_0}^{(n)}$-probability as $n\to\infty$, for any $m_n \to 0$.

If δ_n is a lower bound for the posterior contraction rate at θ_0, then a ball around θ_0 of radius slightly smaller than δ_n is too small to hold any appreciable posterior mass. Naturally, it is desirable to obtain a lower bound that is as large as possible, ideally one that matches an upper bound ϵ_n up to a constant, or a slowly decaying factor such a (negative) power of logarithm. A contraction rate ϵ_n and lower bound δ_n together imply that the posterior distribution puts asymptotically all its mass inside an annular region of the type $\{\theta \in \Theta: m_n\delta_n \leq d(\theta, \theta_0) \leq M_n\epsilon_n\}$, for any sequences $m_n \to 0$ and $M_n \to \infty$. (In nonparametric situations the two sequences m_n and M_n can often be chosen fixed, but the current definition including the two sequences yields the correct rate $n^{-1/2}$ in parametric situations, and also the correct rate in other situations where the rescaled posterior distribution concentrates asymptotically on an unbounded set.)

A simple, but potentially effective, tool to establish a lower convergence rate is given by Theorem 8.20. This shows that the posterior mass of sets of asymptotically vanishing prior probability tends to zero. Applied in the present context, this gives the following sufficient criterion.

Theorem 8.35 (Contraction rate, lower bound) *A sequence δ_n is (certainly) a lower bound for the rate of contraction if, for some ϵ_n and $k > 1$,*

$$\frac{\Pi_n(\theta: d(\theta, \theta_0) \leq \delta_n)}{\Pi_n(B_{n,k}(\theta_0, \epsilon_n))} = o(e^{-2n\epsilon_n^2}), \tag{8.44}$$

for $B_{n,k}(\theta_0, \epsilon)$ the neighborhoods of the true parameter defined in (8.19).

The sequence ϵ_n may be arbitrary, but will typically be (an upper bound on) the rate of contraction, for which the prior mass $\Pi_n(B_{n,k}(\theta_{n,0}, \epsilon_n))$ is at least $e^{-Cn\epsilon_n^2}$, under the prior mass condition (8.22). The condition on δ_n is then satisfied if $\Pi_n(\theta: d(\theta, \theta_0) \leq \delta_n) \ll e^{-(C+2)n\epsilon_n^2}$.

This criterion is particularly attractive if the neighborhoods $B_{n,k}(\theta_{n,0}, \epsilon_n)$ can be replaced by balls for the metric d. If these neighborhoods contain the balls $\{\theta: d(\theta, \theta_0) < c\epsilon_n\}$, then δ_n is a lower bound on the contraction rate if, for some ϵ_n and constants with $C_2 > C_1 + 2c^{-2}$,

$$\Pi_n(\theta: d(\theta, \theta_0) \leq \delta_n) \leq e^{-C_2 n\epsilon_n^2} \leq e^{-C_1 n\epsilon_n^2} \leq \Pi_n(\theta: d(\theta, \theta_0) \leq \epsilon_n). \tag{8.45}$$

This requires a lower and an upper bound on the prior mass near the true parameter and may be difficult to obtain. However, it is not unreasonable to expect these bounds to hold for δ_n of (nearly) the same order as ϵ_n, as infinite-dimensional priors tend to concentrate near boundaries of balls rather than spread continuously. Intuitively this can be understood from the fact that high-dimensional balls have a large surface.

8.5 Misspecification

Consider placing a prior on a (dominated) model $\mathcal{P}$ that does not contain or approximate the true density p_0 of the observations. As the posterior distributions also concentrate on $\mathcal{P}$, they cannot be consistent. From a frequentist point of view studying such *model misspecification* is relevant, because a model is typically only perceived as an approximation to reality. From an objective Bayesian point of view model misspecification is of interest, because in principle the Bayesian paradigm does not restrict the prior, and hence this may rule out a given density p_0. For both it is relevant, because a model may be adopted for convenience only, such as Gaussian errors in nonparametric regression.

The main result of this section shows that the posterior distributions, even though inconsistent, will still settle down eventually, near a Kullback-Leibler projection p^* of p_0 into the model. Furthermore, the rate at which the posterior concentrates near p^* is characterized by similar quantities as before.

We restrict to i.i.d. observations following the Bayesian model $X_1, \ldots, X_n \mid p \overset{\text{iid}}{\sim} p$ and $p \sim \Pi_n$, where Π_n is a prior distribution on a set $\mathcal{P}$ of probability densities relative to a dominating measure ν on a sample space $(\mathfrak{X}, \mathscr{X})$. We adapt the entropy and the neighborhoods for the prior mass condition to misspecified models, as follows.

First we define the *covering number for testing under misspecification* of $\mathcal{P}$ relative to a semimetric d, denoted by $N_{t,\text{mis}}(\epsilon, \mathcal{P}, d; p_0, p^*)$, as the minimal number N of sets in a partition $\mathcal{P}_1, \ldots, \mathcal{P}_N$ of $\{p \in \mathcal{P}: \epsilon < d(p, p^*) < 2\epsilon\}$ such that, for every partitioning set

$$\inf_{p \in \text{conv}(\mathcal{P}_l)} \sup_{0<\alpha<1} -\log P_0\Big(\frac{p}{p^*}\Big)^\alpha \geq \tfrac{1}{2}\epsilon^2. \tag{8.46}$$

The logarithm of this number is referred to as “entropy.”

Second, we consider modified Kullback-Leibler neighborhoods of p, given by

$$B(p^*, \epsilon; p_0) = \Big\{p \in \mathcal{P}: P_0 \log \frac{p^*}{p} \leq \epsilon^2,\ P_0(\log \frac{p^*}{p})^2 \leq \epsilon^2\Big\}. \tag{8.47}$$

Theorem 8.36 *If there exist partitions $\mathcal{P} = \mathcal{P}_{n,1} \cup \mathcal{P}_{n,2}$ and a constant $C > 0$ such that for some $p_n^* \in \mathcal{P}$ with $K(p_0; p_n^*) < \infty$ and $P_0(p/p_n^*) < \infty$ for all $p \in \mathcal{P}$ and constants $\bar\epsilon_n \leq \epsilon_n$ with $n\bar\epsilon_n^2 \geq 1$,*

$$\text{(i)}\ \frac{\Pi_n(p \in \mathcal{P}_{n,1}: j\epsilon_n < d(p, p_n^*) \leq 2j\epsilon_n)}{\Pi_n(B(p_n^*, \bar\epsilon_n; p_0))} \leq e^{n\bar\epsilon_n^2 j^2/8}, \tag{8.48}$$

$$\text{(ii)}\ \sup_{\epsilon \geq \epsilon_n} \log N_{t,\text{mis}}(\epsilon, \mathcal{P}_{n,1}, d; p_0, p_n^*) \leq n\epsilon_n^2, \tag{8.49}$$

$$\text{(iii)}\ \frac{\int_{\mathcal{P}_{n,2}} (P(p_0/p_n^*))^n \, d\Pi_n(p)}{\Pi_n(B(p_n^*, \bar\epsilon_n; p_0))} = o(e^{-2n\bar\epsilon_n^2}), \tag{8.50}$$

then $\Pi_n(p \in \mathcal{P}: d(p, p_n^*) \geq M_n\epsilon_n \mid X_1, \ldots, X_n) \to 0$ in P_0^n-probability, for every $M_n \to \infty$, and for every sufficiently large $M_n = M$ if $n\epsilon_n^2 \to \infty$.

Proof For $p \in \mathcal{P}$, the function $q(p) = (p_0/p_n^*)p$ is integrable by assumption, and hence defines a finite measure. Its Hellinger transform with p_0 satisfies $\rho_\alpha(p_0; q(p)) = P_0(p/p_n^*)^{1-\alpha}$. By Lemma D.6, for any set $\mathcal{P}_l$ of densities,

$$\rho_\alpha\Big(p_0^n; \text{conv}(q(p)^n: p \in \mathcal{P}_l)\Big) \leq \sup_{p \in \text{conv}(\mathcal{P}_l)} \rho_\alpha(p_0; q(p))^n = \sup_{p \in \text{conv}(\mathcal{P}_l)} \Big[P_0\Big(\frac{p}{p_n^*}\Big)^{1-\alpha}\Big]^n.$$

If we take the infimum over $\alpha \in (0, 1)$ across this inequality, then on the right side the order of the infimum can be exchanged with the supremum over p, by the minimax theorem Theorem L.5, applicable, because $\alpha \mapsto P_0(p/p_n^*)^{1-\alpha}$ is convex and can be continuously extended to the compact set [0, 1], and $p \mapsto P_0(p/p_n^*)^{1-\alpha}$ is concave. Thus when applied to a set $\mathcal{P}_l$ as in the definition of the covering number for testing under misspecification, the right and hence left side is bounded above by $e^{-n\epsilon^2/2}$. It follows that the sets

$\mathcal{Q}_l := \{q(p)^n : p \in \mathcal{P}_l\}$ in the induced partition of $\mathcal{Q} := \{q(p)^n : \epsilon < d(p, p_n^*) \leq 2\epsilon\}$ satisfy (D.5) with $K = n/2$. Therefore by Theorem D.4, applied with $\log N(\epsilon)$ equal to the right side of (8.49) and $\epsilon = M\epsilon_n$, there exist tests ϕ_n such that

$$P_0^n \phi_n \leq e^{n\epsilon_n^2} \frac{e^{-nM^2\epsilon_n^2/2}}{1 - e^{-nM^2\epsilon_n^2/2}}, \qquad Q(p)^n(1 - \phi_n) \leq e^{-nM^2\epsilon_n^2/2}.$$

Next we partition $\{p \in \mathcal{P}_{n,1} : d(p, p_n^*) > M\epsilon_n\}$ into the "shells" $\mathcal{S}_{n,j} = \{p \in \mathcal{P}_{n,1} : M\epsilon_n j < d(p, p_n^*) \leq M\epsilon_n(j+1)\}$ for $j \in \mathbb{N}$. For every j we obtain, by Fubini's theorem,

$$P_0^n\Big[\int_{\mathcal{S}_{n,j}} \prod_{i=1}^n \frac{p}{p_n^*}(X_i)\, d\Pi_n(p)(1 - \phi_n)\Big] \leq \int_{\mathcal{S}_{n,j}} Q(p)^n(1 - \phi_n)\, d\Pi_n(p)$$
$$\leq e^{-KnM^2\epsilon_n^2 j^2/2}\Pi_n(\mathcal{S}_{n,j}).$$

For A_n the event $\{\int \prod_{i=1}^n (p/p_n^*)(X_i)\, d\Pi_n(p) \geq e^{-2Cn\bar\epsilon_n^2}\Pi_n(B(p_n^*, \bar\epsilon_n; p_0))\}$ the posterior probability of the set $\{p : d(p, p_n^*) > M\epsilon_n\}$ is bounded above by

$$\phi_n + \mathbb{1}\{A_n^c\} + \frac{1}{e^{-Cn\bar\epsilon_n^2}\Pi_n(B(p_n^*, \bar\epsilon_n; p_0))} \int_{d(p,p_n^*) > M\epsilon_n} \prod_{i=1}^n \frac{p}{p_n^*}(X_i)\, d\Pi_n(p)(1 - \phi_n).$$

The expected value under P_0^n of the third term is bounded by

$$\sum_{j \geq M} \frac{e^{-KnM^2\epsilon_n^2 j^2/2}\Pi_n(\mathcal{S}_{n,j})}{e^{-2Cn\bar\epsilon_n^2}\Pi_n(B_2(p_n^*, \bar\epsilon_n; p_0))} + \frac{\int_{\mathcal{P}_{n,2}} (P(p_0/p_n^*))^n\, d\Pi_n(p)}{e^{-2Cn\bar\epsilon_n^2}\Pi_n(B_2(p_n^*, \bar\epsilon_n; p_0))}.$$

We finish as in the proof of Theorem 8.11, where we substitute Lemma 8.37 below for Lemma 8.10. □

Lemma 8.37 *For any $p^* \in \mathcal{P}$ with $P_0 \log(p_0/p^*) < \infty$, every probability measure Π on $\mathcal{P}$, and every constants $\epsilon, C > 0$, with P_0^n-probability at least $1 - (C\sqrt{n}\epsilon)^{-2}$,*

$$\int \prod_{i=1}^n \frac{p}{p^*}(X_i)\, d\Pi(p) \geq \Pi(B(p^*, \epsilon; p_0))e^{-(1+C)n\epsilon^2}.$$

Since $P_0 \log(p_n^*/p) = K(p_0; p) - K(p_0; p^*)$, the first inequality in the definition of the neighborhoods $B(p^*, \epsilon; p_0)$ can be written in the form

$$K(p_0; p) \leq K(p_0; p^*) + \epsilon^2.$$

Therefore the neighborhood contains only p that are within ϵ^2 of the Kullback-Leibler divergence $K(p_0; p^*)$, which may be regarded the minimal divergence of p_0 to the model. For $p^* = p_0$ the prior mass condition (8.48) reduces to the analogous condition employed previously for correctly specified models.

On the other hand, the entropy condition (8.49) is more abstract than the corresponding condition in the correctly specified case. The technical motivation is the same, existence of good tests, but presently the tests concern the null hypothesis p_0 versus the alternatives $q(p)$ defined in the proof of Theorem 8.36, which may have total mass bigger than one. The form of the condition is explained in Appendix D.1, and also that (8.49) can be valid only if p_n^* is at minimal Kullback-Leibler divergence of p_0 to $\mathcal{P}_{n,1}$, which is not otherwise included in the conditions of the theorem.

The testing entropy can be bounded in terms of ordinary entropy relative to appropriate semimetrics. For convex models $\mathcal{P}$, this is possible for convex semimetrics d satisfying

$$d^2(p, p^*) \leq \sup_{0<\alpha<1} -\log P_0\Big(\frac{p}{p^*}\Big)^\alpha. \tag{8.51}$$

More generally, this is possible for semimetrics that satisfy the following strengthening of (8.51): for some fixed constants $c, C > 0$ and for every $m \in \mathbb{N}$, $\lambda_1, \ldots, \lambda_m \geq 0$ with $\sum_i \lambda_i = 1$ and every $p, p_1, \ldots, p_m \in \mathcal{P}$ with $d(p, p_i) \leq c\, d(p, p^*)$ for all i,

$$\sum_i \lambda_i \big[d^2(p_i, p^*) - C d^2(p_i, p)\big] \leq \sup_{0<\alpha<1} -\log P_0\Big(\frac{\sum_i \lambda_i p_i}{p^*}\Big)^\alpha. \tag{8.52}$$

Evaluating this inequality with $m = 1$ and $p = p_1$ leads back to (8.51).

Lemma 8.38 *We have $N_{t,\mathrm{mis}}(\epsilon, \mathcal{P}, d; p_0, p^*) \leq N(A\epsilon, \{p \in \mathcal{P}: \epsilon < d(p, p^*) < 2\epsilon\}, d)$, for all $\epsilon > 0$, and some constant $A > 0$, if one of the following conditions holds:*

(i) (8.51) *is valid for every $p \in \mathrm{conv}(\mathcal{P})$ and the map $p \mapsto d^2(p, p')$ is defined and convex on* $\mathrm{conv}(\mathcal{P})$*, for every $p' \in \mathcal{P}$.*
(ii) (8.52) *holds.*

Proof (ii). For a given constant $A > 0$ we can cover the set $\mathcal{P}_\epsilon := \{p \in \mathcal{P}: \epsilon < d(p, p^*) < 2\epsilon\}$ with $N = N(A\epsilon, \mathcal{P}_\epsilon, d)$ balls of radius $A\epsilon$. If the centers of these balls are not contained in $\mathcal{P}_\epsilon$, then we can replace them by N balls of radius $2A\epsilon$ with centers in $\mathcal{P}_\epsilon$ whose union also covers $\mathcal{P}_\epsilon$. It suffices to show that (8.46) is valid for $\mathcal{P}_l$ equal to a typical ball B in this cover. Choose $2A < 1$. If $p \in \mathcal{P}_\epsilon$ is the center of B and $p_i \in B$ for every i, then $d(p_i, p^*) \geq d(p, p^*) - 2A\epsilon$ by the triangle inequality, and hence by assumption (8.52) the left side of (8.46) is bounded below by $\sum_i \lambda_i((\epsilon - 2A\epsilon)^2 - C(2A\epsilon)^2)$. This is bounded below by $\epsilon^2/2$ for sufficiently small A.

(i). It suffices to show that condition (i) implies condition (ii) for $d/2$ instead of d. By repeatedly applying the triangle inequality we find

$$\begin{aligned}\sum_i \lambda_i d^2(p_i, p^*) &\leq 3\sum_i \lambda_i\Big[d^2(p_i, p) + d^2\Big(p, \sum_j \lambda_j p_j\Big) + d^2\Big(\sum_j \lambda_j p_j, p^*\Big)\Big]\\ &\leq 6\sum \lambda_i d^2(p_i, p) + 3d^2\Big(\sum_j \lambda_j p_j, p^*\Big),\end{aligned}$$

by the convexity of d^2. It follows that

$$d^2\Big(\sum_i \lambda_i p_i, p^*\Big) \geq \frac{1}{3}\sum_i \lambda_i d^2(p_i, p^*) - 2\sum_i \lambda_i d^2(p_i, p).$$

If (8.51) holds for $p = \sum_i \lambda_i p_i$, then we obtain (8.52) with d^2 replaced by $d^2/4$ (or even $d^2/3$) and $C = 6$. □

8.5.1 Convex Models

The case of convex models $\mathcal{P}$ is of interest, in particular for non- or semiparametrics, and permits some simplification. For a convex model the point of minimal Kullback-Leibler divergence (if it exists) is automatically unique (up to a P_0-null set). Moreover, the expectations $P_0(p/p^*)$ are automatically finite, and condition (8.52) is automatically satisfied for the *weighted Hellinger distance*, with square,

$$d_{H,w}^2(p_1, p_2) = \int (\sqrt{p_1} - \sqrt{p_2})^2 \frac{p_0}{p^*}\, d\nu.$$

Theorem 8.39 (Convex model) *If $\mathcal{P}$ is convex and $K(p_0; p^*) \leq K(p_0, p)$, for $p^* \in \mathcal{P}$ and every $p \in \mathcal{P}$, then $P_0(p/p^*) \leq 1$ for every $p \in \mathcal{P}$, and there exists a constant $A > 0$ such that $N_{t,\mathrm{mis}}(\epsilon, \mathcal{P}, d_{H,w}; p_0, p^*) \leq N(A\epsilon, \{p \in \mathcal{P}: \epsilon < d_{H,w}(p, p^*) < 2\epsilon\}, d_{H,w})$, for all $\epsilon > 0$.*

Proof For $p \in \mathcal{P}$, define a family of convex combinations $\{p_\lambda: \lambda \in [0, 1]\} \subset \mathcal{P}$ by $p_\lambda = \lambda p + (1 - \lambda)p^*$. Since $p^* \in \mathcal{P}$ minimizes the Kullback-Leibler divergence with respect to p_0 over $\mathcal{P}$, for every $\lambda \in [0, 1]$,

$$0 \leq f(\lambda) := P_0 \log \frac{p^*}{p_\lambda} = -P_0 \log\Bigl(1 + \lambda\Bigl(\frac{p}{p^*} - 1\Bigr)\Bigr).$$

For every fixed $y \geq 0$ the function $\lambda \mapsto \log(1 + \lambda y)/\lambda$ is nonnegative and increases monotonically to y as $\lambda \downarrow 0$. The function is bounded in absolute value by 2 for $y \in [-1, 0]$ and $\lambda \leq \frac{1}{2}$. Therefore, by the monotone and dominated convergence theorems applied to the positive and negative parts of the integrand on the right side the right derivative of f at 0 is given by $f'(0+) = 1 - P_0(p/p^*)$. Now $f'(0+) \geq 0$, since $0 = f(0) \leq f(\lambda)$ for every λ, whence $P_0(p/p^*) \leq 1$.

Because $-\log x \geq 1 - x$, we have $-\log P_0(p/p^*)^{1/2} \geq 1 - P_0(p/p^*)^{1/2}$. Now

$$\int (\sqrt{p^*} - \sqrt{p})^2 \frac{p_0}{p^*}\, d\mu = 1 + P_0 \frac{p}{p^*} - 2P_0 \sqrt{\frac{p}{p^*}} \leq 2 - 2P_0 \sqrt{\frac{p}{p^*}},$$

by the first part of the proof. It follows that the semimetric $d_{H,w}/2$ satisfies (8.51), whence the theorem follows by Lemma 8.38(i). □

8.5.2 Nonparametric Regression

Let P_f be the distribution of a random vector (X, Y) following the regression model $Y = f(X) + e$, for independent random variables X and e taking values in a measurable space $(\mathfrak{X}, \mathscr{X})$ and in $\mathbb{R}$, respectively, and an unknown measurable function $f: \mathfrak{X} \to \mathbb{R}$ belonging to some model $\mathcal{F}$. We form the posterior distribution for f in the Bayesian model $(X_1, Y_1), \ldots, (X_n, Y_n) | f \overset{\mathrm{iid}}{\sim} P_f$ and $f \sim \Pi$, where the corresponding errors are assumed to have either a standard normal or Laplace distribution and the covariates a fixed distribution. The latter covariate distribution cancels from the posterior distribution, and hence can be assumed known. On the other hand, the normal or Laplace distribution for the errors is considered a working model only, but is essential to the definition of the posterior distribution. We investigate the posterior distribution under the assumption that the observations in reality follow the model $Y = f_0(X) + e_0$, for a given regression function f_0, and an error e_0 that has mean or median (in the normal and Laplace cases) zero, but may not be Gaussian or Laplace. Thus the error distribution is typically misspecified. We also allow that the true regression function f_0 is misspecified in that it does not belong to the model $\mathcal{F}$, but this is of secondary interest, as in our nonparametric situation $\mathcal{F}$ will typically be large.

The main finding is that misspecification of the error distribution does not have serious consequences for estimation of the regression function. Thus the nonparametric Bayesian approach possesses the same robustness to misspecification as minimum contrast estimation using least squares or minimum absolute deviation. The results below require that the error distribution be light-tailed in the normal case, but not so for the Laplace posterior. Thus the tail robustness of minimum absolute deviation versus the nonrobustness of least squares appears to extend also to Bayesian regression.

As in the general theory the true distribution of an observation is denoted P_0. This should not be confused with P_f for $f = 0$, which does not appear in the following.

Normal Errors

If the working (misspecified) error distribution is standard normal, and a typical observation follows the model $Y = f_0(X) + e_0$ for a a true error e_0 with mean zero, then

$$\log \frac{p_{f_0}}{p_f}(X, Y) = \tfrac{1}{2}(f - f_0)^2(X) - e_0(f - f_0)(X),$$
$$K(p_0; p_f) = \tfrac{1}{2}P_0(f - f_0)^2.$$

It follows that the Kullback-Leibler divergence of p_0 to the model is minimized at p_{f^*} for $f^* \in \mathcal{F}$ minimizing the map $f \mapsto P_0(f - f_0)^2$. In particular $f^* = f_0$, in the case that the model for the regression functions is correctly specified ($f_0 \in \mathcal{F}$). If $\mathcal{F}$ is convex and closed in $\mathbb{L}_2$, then a unique minimizer f^* exists by the Hilbert space projection theorem, and is characterized by the inequalities $P_0(f_0 - f^*)(f^* - f) \geq 0$, for every $f \in \mathcal{F}$. If the minimizer f^* is also the minimizer over the linear span of $\mathcal{F}$, then the latter inequalities are equalities.

Replace the "neighborhoods" $B(p_{f^*}, \epsilon; p_0)$ in the prior mass condition (8.48) by

$$\tilde{B}(p_{f^*}, \epsilon; f_0) = \Bigl\{ f \in \mathcal{F}: \|f - f_0\|_2^2 \leq \|f^* - f_0\|_2^2 + \epsilon^2, \|f - f^*\|_2^2 \leq \epsilon^2 \Bigr\}.$$

If $P_0(f - f^*)(f^* - f_0) = 0$ for every $f \in \mathcal{F}$, then this reduces to an $\mathbb{L}_2$-ball around f^*, since $\|f - f_0\|_2^2 = \|f - f^*\|_2^2 + \|f^* - f_0\|_2^2$, by Pythagoras's theorem.

Theorem 8.40 *Let $\mathcal{F}$ be a class of uniformly bounded functions $f: \mathfrak{X} \to \mathbb{R}$ such that either $f_0 \in \mathcal{F}$ or $\mathcal{F}$ is convex and closed in $\mathbb{L}_2$. If f_0 is uniformly bounded,* $\mathrm{E}_0 e_0 = 0$ *and* $\mathrm{E}_0 e^{M|e_0|} < \infty$ *for every $M > 0$, and ϵ_n are positive numbers with $n\epsilon_n^2 \to \infty$ such that for a constant $C > 0$,*

$$\Pi(\tilde{B}(f^*, \epsilon_n; f_0)) \geq e^{-Cn\epsilon_n^2},$$
$$\log N(\epsilon_n, \mathcal{F}, \|\cdot\|_{P_0,2}) \leq n\epsilon_n^2,$$

then $\Pi_n(f \in \mathcal{F}: P_0(f-f^*)^2 \ge M\epsilon_n^2 | X_1,\ldots,X_n) \to 0$ *in* P_0^n*-probability, for every sufficiently large constant* M.

The theorem is a corollary of Theorem 8.36 and the following lemma, which shows that (8.52) is satisfied for a multiple of the $\mathbb{L}_2$-distance on $\mathcal{F}$, and that the neighborhoods (8.52) contain the neighborhoods induced by the map $f \mapsto p_f$ and the present neighborhoods (with ϵ replaced by a multiple).

Lemma 8.41 *Let* $\mathcal{F}$ *be a class of uniformly bounded functions* $f\colon \mathfrak{X} \to \mathbb{R}$ *such that either* $f_0 \in \mathcal{F}$ *or* $\mathcal{F}$ *is convex and closed in* $\mathbb{L}_2(P_0)$. *If* f_0 *is uniformly bounded,* $\mathrm{E}_0 e_0 = 0$ *and* $\mathrm{E}_0 e^{M|e_0|} < \infty$ *for every* $M > 0$, *then there exist positive constants* C_1, C_2, C_3 *such that, for all* $m \in \mathbb{N}$, $f, f_1, \ldots, f_m \in \mathcal{F}$ *and* $\lambda_1,\ldots,\lambda_m \ge 0$ *with* $\sum_i \lambda_i = 1$,

$$P_0\Big(\log\frac{p_{f^*}}{p_f}\Big)^2 \le C_1 P_0(f-f^*)^2,$$
$$\sup_{0<\alpha<1} -\log P_0\Big(\frac{\sum_i \lambda_i p_{f_i}}{p_{f^*}}\Big)^{\alpha} \ge C_2 \sum_i \lambda_i \Big(P_0(f_i-f^*)^2 - C_3 P_0(f-f_i)^2\Big).$$

Proof The second term on the right in

$$\log\frac{p_{f^*}}{p_f}(X,Y) = \tfrac{1}{2}\big[(f_0-f)^2-(f_0-f^*)^2\big](X) + e_0(f^*-f)(X) \tag{8.53}$$

has mean zero by assumption. The first term has expectation $\frac{1}{2}P_0(f^*-f)^2$ if $f_0 = f^*$, as is the case if $f_0 \in \mathcal{F}$. Furthermore, if $\mathcal{F}$ is convex the minimizing property of f^* implies that $P_0(f_0-f^*)(f^*-f) \ge 0$ for every $f \in \mathcal{F}$ and then the expectation of the first term on the right is bounded above by $\frac{1}{2}P_0(f^*-f)^2$. Therefore, in both cases $P_0\log(p_{f^*}/p_f) \ge \frac{1}{2}P_0(f-f^*)^2$.

From (8.53) we also have, with M a uniform upper bound on $\mathcal{F}$ and f_0,

$$P_0(\log\frac{p_f}{p_{f^*}})^2 \le 2P_0\big[(f^*-f)^2(2f_0-f-f^*)^2\big] + 2P_0e_0^2 P_0(f^*-f)^2,$$
$$P_0(\log\frac{p_f}{p_{f^*}})^2(\frac{p_f}{p_{f^*}})^{\alpha} \le 2P_0\big[(f^*-f)^2(2f_0-f-f^*)^2 + 2e_0^2(f^*-f)^2\big]e^{2\alpha(M^2+M|e_0|)}.$$

Both right sides can be further bounded by a constant times $P_0(f-f^*)^2$, where the constant depends on M and the distribution of e_0 only.

In view of Lemma B.6 with $p = p_{f^*}$ and $q_i = p_{f_i}$, we see that there exists a constant $C > 0$ depending on M only such that for all $\lambda_i \ge 0$ with $\sum_i \lambda_i = 1$,

$$\Big|1 - P_0\Big(\frac{\sum_i \lambda_i p_{f_i}}{p_{f^*}}\Big)^{\alpha} - \alpha P_0 \log\frac{p_{f^*}}{\sum_i \lambda_i p_{f_i}}\Big| \le 2\alpha^2 C \sum_i \lambda_i P_0(f_i-f^*)^2. \tag{8.54}$$

By Lemma B.6 with $\alpha = 1$ and $p = p_f$ and similar arguments also, for any $f \in \mathcal{F}$,

$$\Big|1 - P_0\Big(\frac{\sum_i \lambda_i p_{f_i}}{p_f}\Big) - P_0\log\frac{p_f}{\sum_i \lambda_i p_{f_i}}\Big| \le 2C\sum_i \lambda_i P_0(f_i-f)^2.$$

For $\lambda_i = 1$ this becomes $|1 - P_0(p_{f_i}/p_f) - P_0 \log(p_f/p_{f_i})| \leq 2C P_0(f_i - f)^2$. Subtracting the convex combination of these inequalities from the preceding display gives

$$\Big| P_0 \log \frac{p_f}{\sum_i \lambda_i p_{f_i}} - \sum_i \lambda_i P_0 \log \frac{p_f}{p_{f_i}} \Big| \leq 4C \sum_i \lambda_i P_0(f_i - f)^2.$$

By the fact that $\log(ab) = \log a + \log b$ for every $a, b > 0$, this inequality remains true if f on the left is replaced by f^*. We combine the resulting inequality with (8.54) to find that

$$\begin{aligned} 1 - P_0\Big(\frac{\sum_i \lambda_i p_{f_i}}{p_{f^*}}\Big)^\alpha &\geq \alpha \sum_i \lambda_i P_0 \log \frac{p_{f^*}}{p_{f_i}} - 2\alpha^2 C \sum_i \lambda_i P_0(f^* - f_i)^2 \\ &\quad - 4C \sum_i \lambda_i P_0(f_i - f)^2 \\ &\geq \Big(\frac{\alpha}{2} - 2\alpha^2 C\Big) \sum_i \lambda_i P_0(f^* - f_i)^2 - 4C \sum_i \lambda_i P_0(f_i - f)^2. \end{aligned}$$

In the last step we used that $P_0 \log(p_{f^*}/p_f) \geq \frac{1}{2} P_0(f - f^*)^2$. For sufficiently small $\alpha > 0$ and suitable constants C_2, C_3 the right side is bounded below by the right side of the lemma. Finally the left side of the lemma can be bounded by the supremum over $\alpha \in (0, 1)$ of the left side of the last display, since $-\log x \geq 1 - x$ for every $x > 0$. □

Laplace Errors

If the working (misspecified) error distribution is Laplace, and a typical observation follows the model $Y = f_0(X) + e_0$ for a true error e_0 with median zero, then

$$\begin{aligned} &\log \frac{p_{f_0}}{p_f}(X, Y) = (|e_0 + f_0(X) - f(X)| - |e_0|), \\ &K(p_0; p_f) = P_0 \Phi(f - f_0), \qquad \Phi(v) := P_0(|e_0 - v| - |e_0|). \end{aligned}$$

The function Φ is minimized over $v \in \mathbb{R}$ at the median of e_0, which is zero by assumption. Thus the Kullback-Leibler divergence $f \mapsto K(p_0; p_f)$ is minimized over $f \in \mathcal{F}$ at f_0, if $f_0 \in \mathcal{F}$. If $\mathcal{F}$ is a compact, convex subset of $\mathbb{L}_1(P_0)$, then in any case there exists $f^* \in \mathcal{F}$ that minimizes the Kullback-Leibler divergence.

If the distribution of e_0 is smooth, then so is the function Φ. Because it is minimal at $v = 0$ it is reasonable to expect that, for v in a neighborhood of 0 and some positive constant C_0,

$$\Phi(v) = P_0(|e_0 - v| - |e_0|) \geq C_0 |v|^2. \tag{8.55}$$

Because Φ is convex, it is also reasonable to expect that its second derivative, if it exists, is strictly positive.

Define neighborhoods $\tilde{B}(f^*, \epsilon; f_0)$ as in Theorem 8.40. The following theorem is a corollary of Theorem 8.36, in the same way as Theorem 8.40. (See Kleijn and van der Vaart 2006 for a detailed proof.)

Theorem 8.42 *Let $\mathcal{F}$ be a class of uniformly bounded functions $f: \mathfrak{X} \to \mathbb{R}$ and suppose that either $f_0 \in \mathcal{F}$ and* (8.55) *hold, or $\mathcal{F}$ is convex and compact in $\mathbb{L}_1(P_0)$ and Φ is twice continuously differentiable with strictly positive second derivative. If f_0 is uniformly* *bounded, e_0 has median 0, and ϵ_n are positive numbers with $n\epsilon_n^2 \to \infty$ such that for a constant $C > 0$,*

$$\Pi(\tilde{B}(f^*, \epsilon_n; f_0)) \geq e^{-Cn\epsilon_n^2},$$
$$\log N(\epsilon_n, \mathcal{F}, \|\cdot\|_{P_0,2}) \leq n\epsilon_n^2,$$

then $\Pi_n(f \in \mathcal{F}: P_0(f - f^*)^2 \geq M\epsilon_n^2 | X_1, \ldots, X_n) \to 0$ in P_0^n-probability, for every sufficiently large constant M.

8.6 α-Posterior

The α-posterior distribution is defined for i.i.d. observations in (6.19). In the context of general statistical experiments $(\mathfrak{X}, \mathscr{X}, P_\theta: \theta \in \Theta)$ with observations X, the *α-posterior distribution* is defined as

$$\Pi^{(\alpha)}(\theta \in B | X) = \frac{\int_B p_\theta^\alpha(X)\, d\Pi_n(\theta)}{\int p_\theta^\alpha(X)\, d\Pi_n(\theta)}. \tag{8.56}$$

Let $R_\beta(p; q) = -\log \rho_\beta(p; q)$ denote the *Renyi divergence* between densities p and q, as defined in (B.5). By Lemma B.5 this is an upper bound on the square Hellinger distance. The following theorem gives a bound on the expected Renyi divergence from a true density under the posterior distribution.

Theorem 8.43 *For any numbers $\epsilon > 0$, $\beta \in (0, 1)$, $\gamma \geq 0$, for $\alpha = (\gamma + \beta)/(\gamma + 1)$,*

$$P_{\theta_0} \int R_\beta(p_{\theta_0}; p_\theta)\, d\Pi^{(\alpha)}(\theta | X) \leq (\gamma + 1)\epsilon^2\alpha - \log \Pi(\theta: K(p_{\theta_0}; p_\theta) < \epsilon^2).$$

Proof The nonnegativity of the Kullback-Leibler divergence implies that for any given nonnegative measurable function $v: \Theta \to \mathbb{R}$ and every probability density w relative to Π,

$$-\int (\log w) w\, d\Pi + \int (\log v) w\, d\Pi \leq \log \int v\, d\Pi. \tag{8.57}$$

Furthermore, equality is attained for any function w with $w \propto v$. Applying this twice, first for a fixed observation X with $v(\theta) \propto (p_\theta/p_{\theta_0})^\delta(X)$ and $\delta = 1$, and second with $v(\theta) \propto (p_\theta/p_{\theta_0})^\beta(X)/\rho_\beta(p_{\theta_0}; p_\theta)$, we find for any probability density w relative to Π,

$$-\int (\log w) w\, d\Pi + \delta \int \Bigl(\log \frac{p_\theta}{p_{\theta_0}}(X)\Bigr) w(\theta)\, d\Pi(\theta) \leq \log \int \Bigl(\frac{p_\theta}{p_{\theta_0}}\Bigr)^\delta (X)\, d\Pi(p_\theta),$$
$$-\int (\log w) w\, d\Pi + \beta \int \Bigl(\log \frac{p_\theta}{p_{\theta_0}}(X)\Bigr) w(\theta)\, d\Pi(\theta) - \int \log \rho_\beta(p_{\theta_0}; p_\theta)\, w(\theta)\, d\Pi(\theta)$$
$$\leq \log c_\beta(X),$$

where $c_\beta(X) = \int (p_\theta/p_{\theta_0})^\beta(X)/\rho_\beta(p_{\theta_0}; p_\theta)\, d\Pi(\theta)$ is the normalizing constant. By the convexity of the negative of the logarithm, Jensen's inequality and Fubini's theorem, the expectation of the right side of the first inequality is at most $\log \int \rho_\delta(p_{\theta_0}; p_\theta)\, d\Pi(\theta) \leq 0$. Similarly the expectation of the right side of the second inequality is nonnegative. We replace the right sides by 0 and add the second inequality to γ times the first inequality. The resulting inequality can be reorganized into

$$\int R_\beta(p_\theta, p_{\theta_0})\, w(\theta)\, d\Pi(\theta) \le (\gamma+1)\int (\log w) w\, d\Pi - (\gamma\delta+\beta)\int \Bigl(\log \frac{p_\theta}{p_{\theta_0}}(X)\Bigr) w(\theta)\, d\Pi(\theta).$$

By (8.57) the right side is minimized with respect to probability densities w, for fixed X, by $w(\theta) \propto p_\theta^\alpha(X)$. For this minimizing function $w(\theta)\, d\Pi(\theta)$ in the left side becomes $d\Pi_\alpha(\theta|\, X)$. It follows that

$$\begin{aligned}
&\frac{1}{\gamma+1} P_{\theta_0} \int R_\beta(p_{\theta_0}; p_\theta)\, d\Pi_\alpha(\theta|\, X)\\
&\quad\le P_{\theta_0} \inf_w \Bigl[\int (\log w) w\, d\Pi - \alpha \int \Bigl(\log \frac{p_\theta}{p_{\theta_0}}(X)\Bigr) w(\theta)\, d\Pi(\theta)\Bigr]\\
&\quad\le \inf_w \Bigl[\int (\log w) w\, d\Pi + \alpha \int P_{\theta_0}\Bigl(\log \frac{p_{\theta_0}}{p_\theta}(X)\Bigr) w(\theta)\, d\Pi(\theta)\Bigr]\\
&\quad= \inf_w \Bigl[\int (\log w) w\, d\Pi - \int \Bigl(\log e^{-\alpha K(p_{\theta_0}; p_\theta)}\Bigr) w(\theta)\, d\Pi(\theta)\Bigr]\\
&\quad= -\log \int e^{-\alpha K(p_{\theta_0}; p_\theta)}\, d\Pi(\theta).
\end{aligned}$$

Here the last step follows again by (8.57). Finally we use Markov's inequality to bound this by $(\gamma+1)^{-1}$ times the right side of the theorem. □

An attractive feature of the preceding theorem is that it makes no assumption on the sampling model: the family p_θ is completely general. The theorem will typically be applied with t such that the two terms on its right side have the same order of magnitude. If $\Pi(\theta\colon K(p_{\theta_0}; p_\theta) < t^2) \ge \exp(-t^2)$, then the order of magnitude is t^2. This is a version of the prior mass condition.

The parameter α is necessarily strictly smaller than 1, and hence the theorem applies to the pseudo-posterior, but not to the true posterior distribution. The theorem shows that for the pseudo-posterior distribution, the prior mass condition alone gives posterior concentration.

Example 8.44 (Pseudo-posterior, i.i.d. observations) The Rényi and Kullback-Leibler divergences are additive on product measures. Therefore, if p_θ in the theorem are replaced by product densities p_θ^n and ϵ by $\sqrt{n}\epsilon_n$, then we obtain

$$P_{\theta_0}^n \int R_\beta(p_{\theta_0}; p_\theta)\, d\Pi_n^{(\alpha)}(\theta|\, X_1,\ldots,X_n) \le (\gamma+1)\epsilon_n^2\alpha - \frac{1}{n}\log \Pi(\theta\colon K(p_{\theta_0}; p_\theta) < \epsilon_n^2).$$

Under the prior mass condition $\Pi(\theta\colon K(p_{\theta_0}; p_\theta) < \epsilon_n^2) \ge \exp(-n\epsilon_n^2)$, the right side is of the order ϵ_n^2. On the left side we can lower bound R_β by $\min(\beta, 1-\beta) d_H^2$, in view of Lemma B.5. Thus, by Markov's inequality the "usual" measure of posterior contraction $\Pi_n(d_H(p_{\theta_0}, p_\theta) > M_n\epsilon_n|\, X_1,\ldots,X_n)$ is bounded by $(M_n\epsilon_n)^2$ times the left side, and hence tends to zero for any $M_n \to \infty$.

It follows that the usual prior mass condition (8.4) alone guarantees a contraction rate of the α-posterior.

8.7 Historical Notes

Posterior contraction rates for i.i.d. observations were studied almost simultaneously by Ghosal et al. (2000) and Shen and Wasserman (2001). While the first used metric entropy to quantify the size of a parameter space and showed the fundamental role of tests, as in Schwartz's theory of consistency and inspired by Birgé (1983a), the second employed stronger conditions, involving bracketing numbers and entropy integrals to ensure that the likelihood ratio is uniformly exponentially small in view of results of Wong and Shen (1995); see Problem 8.5 for a formulation of posterior contraction rate using bracketing entropy integral. That the prior mass condition need not involve the Kullback-Leibler variation, but only the divergence, is a small innovation in the results presented here relative to the results in Ghosal et al. (2000). The idea of constructing priors based on finite approximation of a compact parameter space was first used by Ghosal et al. (1997) in the context of constructing default priors for nonparametric problems; see Theorem 6.48. Ghosal et al. (2000) refined their approach through the use of brackets and controlled ratios to obtain bounds for prior concentration. The resulting conditions for the posterior contraction rate are essentially the same as the conditions for convergence of maximum likelihood estimators or minimum contrast estimators, derived by Wong and Shen (1995), Theorem 1. Contraction rates in finite-dimensional models were studied by various authors, including Le Cam (1986) and Ibragimov and Has'minskiĭ (1981) and Ghosal et al. (1995), who also treated non-regular models. Posterior contraction rates in non-i.i.d. settings, including the special cases of independent, non-identically distributed observations, Markov chains, white noise model and Gaussian time series, were studied by Ghosal and van der Vaart (2007a). Zhao (2000) studied conjugate priors in the white noise model, of the type considered in Example 8.6, but also block-based priors, which have certain advantages. Posterior contraction rates under misspecification were studied by Kleijn and van der Vaart (2006). The information theoretic approach to posterior contraction rate was developed in Zhang (2006), Catoni (2004) and Kruijer and van der Vaart (2013). A martingale-based approach in the spirit of Theorem 6.24 was considered by Walker et al. (2007); see Problem 8.11. Although the approach does not require the explicit construction of a sieve, existence of a sieve is implied. Xing (2010) introduced an elegant approach, where the size of a model is measured by the so called "Hausdorff entropy" introduced by Xing and Ranneby (2009), which also takes into account the role of the prior in size calculations; see Problem 8.13.

Problems

8.1 When θ is multidimensional, show that Lemma 8.2 holds if $\|\mathrm{E}(\theta|\, X^{(n)}) - \theta_0\| = O_p(n^{-1/2})$ and for each component θ_j of θ, $\mathrm{var}(\theta_j|\, X^{(n)}) = O_p(n^{-1})$.

8.2 Verify the calculations in Example 8.4.

8.3 If $\Pi_n(\theta' : d(\theta_0, \theta') < \epsilon_n | X^{(n)}) \to 1$ in $P_{\theta_0}^{(n)}$-probability (respectively, a.s.) and an estimator $\hat{\theta}_n$ is defined as the near maximizer (up to a tolerance that tends to zero) of $\theta \mapsto \Pi_n(\theta' : d(\theta, \theta') < \epsilon_n | X^{(n)})$, show that $d(\hat{\theta}_n, \theta_0) \leq 2\epsilon_n$ in $P_{\theta_0}^{(n)}$-probability (respectively, a.s.).

8.4 Show the following analog of Theorem 8.8 for a general family indexed by an abstract parameter $\theta \in \Theta$: If Θ is a convex set with a semi-distance d_n on it (possibly depending on n), d_n^s is a convex function in one argument keeping the other fixed for some $s > 0$, d_n is bounded above by some universal constant K and $\Pi_n\left(d_n(\theta, \theta_0) \geq \epsilon_n | X^{(n)}\right) = O_p(\epsilon_n^2)$, then $d_n(\hat{\theta}_n, \theta_0) = O_p(\epsilon_n)$, where $\hat{\theta}_n = \int \theta \, d\Pi_n(\theta | X^{(n)})$.

8.5 (Shen and Wasserman 2001) Let $X_1, X_2, \ldots \overset{\text{iid}}{\sim} p$, $p \in \mathcal{P}$, and $p \sim \Pi$. Let d be a semi-distance on $\mathcal{P}$ and let p_0 stand for the true density. Let $t_n > 0$, $s_n \geq r_n > 0$ be sequences such that $\Pi(p : K(p_0, p) \leq t_n, V_{2,0}(p_0, p) \leq t_n) \gtrsim e^{-2nt_n}$ and $P_0^*(\sup\{\mathbb{P}_n \log(p/p_0) : d(p, p_0) \geq s_n\} > -cs_n^2) \to 0$ for some $c > 0$. (An empirical process inequality [cf. Wong and Shen 1995] implies that for the Hellinger distance, the condition holds if the bracketing entropy integral satisfies the inequality $\int_{s_n^2}^{s_n} \sqrt{\log N_{[\,]}(u, \mathcal{P}, d_H)}\, du \lesssim s_n^2$.) Show that the posterior contracts at the rate $\max(r_n, \sqrt{t_n})$ at p_0 in P_0^n-probability.

8.6 (Ghosal et al. 2000) Show that for $B = \{p : d_H^2(p, p_0) \|p_0/p\|_\infty \leq \epsilon^2\}$ there exists a universal constant $C > 0$, such that

$$P_0^n\Bigl(\int \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi(p) \leq \Pi(B) e^{-3n\epsilon^2}\Bigr) \leq e^{-Cn\epsilon^2}. \tag{8.58}$$

8.7 (Exponential rate) Verify that the proof of Theorem 8.9 yields the stronger assertions that $\Pi_n(p \in \mathcal{P}_{n,1} : d(p, p_0) > M\epsilon_n | X_1, \ldots, X_n) = O_P(e^{-n\epsilon_n^2})$, for every sufficiently large constant M, under its assumptions (i) and (ii), and $\Pi_n(\mathcal{P}_{n,2} | X_1, \ldots, X_n) = O_P(e^{-n\epsilon_n^2})$, under its assumptions (i) an (iii), both in P_0^n-probability. Combine this with the preceding problem to see that given uniformly bounded likelihood ratios these assertions are also true in mean, whence also $P_0^n \Pi_n(p : d(p, p_0) > M\epsilon_n | X_1, \ldots, X_n) = O(e^{-n\epsilon_n^2})$.

8.8 (Ghosal et al. 2000) For a given function m let $\Phi^{-1}(\epsilon) = \sup\{M : \Phi(M) \geq \epsilon\}$ be the inverse function of $\Phi(M) = P_0 m \mathbb{1}\{m \geq M\}/M$. Then for every $\epsilon \in (0, 0.44)$ and probability measure Π on the set

$$\Bigl\{P : p_0/p \leq m, 18h^2(P, P_0)\Bigl(1 + \log_+ \frac{\sqrt{P_0 m}}{h(P, P_0)} + \Phi^{-1}(h^2(P, P_0))\Bigr) \leq \epsilon^2\Bigr\}$$

show that for a universal constant $B > 0$,

$$P_0^n\Bigl(\int \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi(P) \leq e^{-2n\epsilon^2}\Bigr) \leq e^{-Bn\epsilon^2}. \tag{8.59}$$

8.9 (Ghosal et al. 2000) Suppose that for a given function m with $P_0 m < \infty$ and $\Phi^{-1}(\epsilon) = \sup\{M : \Phi(M) \geq \epsilon\}$ the inverse of the function $\Phi(M) = P_0 m \mathbb{1}\{m \geq M\}/M$,

$$\Pi_n\bigl(P : 18d_H^2(P, P_0)\bigl(1 + \log_+ \bigl(\sqrt{P_0 m}/d_H(P, P_0)\bigr) + \Phi^{-1}(d_H^2(P, P_0))\bigr) \leq \epsilon_n^2, p_0/p \leq m\bigr)$$

is at least $e^{-n\epsilon_n^2 C}$ for some $C > 0$. If conditions (i) and (ii) of Theorem 8.9 hold and in addition $\sum_n e^{-Bn\epsilon_n^2} < \infty$ for every $B > 0$, then $\Pi_n(P\colon d(P, P_0) \geq M\epsilon_n | X_1, \ldots, X_n) \to 0$ almost surely $[P_0^n]$, for sufficiently large M.

8.10 Prove (8.16).

8.11 (Walker et al. 2007) Let $X_1, X_2, \ldots \overset{\text{iid}}{\sim} p$, $p \in \mathcal{P}$, $p \sim \Pi$, and let p_0 stand for the true density. Let $\epsilon_n \to 0$ be a positive sequence such that $\Pi(p\colon K(p_0, p) \leq \epsilon_n^2, V_2(p_0, p) \leq \epsilon_n^2) \gtrsim e^{-cn\epsilon_n^2}$ for some $c > 0$. Assume that there exists a sequence $\delta_n \leq \rho\epsilon_n$, $\rho < 1$, such that $n\delta_n^2 \to \infty$ and that there exists a countable partition $\{\mathcal{P}_{j,n}\colon j \in \mathbb{N}\}$ of $\mathcal{P}$, each with d_H-diameter at most δ_n such that $e^{-n\delta_n^2/16}\sum_{j=1}^{\infty}\sqrt{\Pi(\mathcal{P}_{j,n})} \to 0$. Then the posterior contracts at the rate $\max(r_n, \sqrt{t_n})$ at p_0 in P_0^n-probability with respect to d_H.

8.12 If the conditions of Problem 8.11 hold, show that there exists a sequence $\mathcal{P}_n \subset \mathcal{P}$ such that the conditions of Theorem 8.9 also hold.

8.13 (Xing 2010) Let $\mathcal{P}$ be a set of densities, Π a prior on $\mathcal{P}$, $\alpha \geq 0$, $\epsilon > 0$ and $\mathcal{P}' \subset \mathcal{P}$. Define the *Hausdorff entropy* by

$$J(\epsilon, \mathcal{P}', \Pi, \alpha) = \log\inf\left\{\sum \Pi(\mathcal{P}_j)^\alpha\colon \cup_{j\in\mathbb{N}}\mathcal{P}_j \supset \mathcal{P}', \operatorname{diam}(\mathcal{P}_j) \leq \epsilon\right\}.$$

Note that for $\alpha = 0$, this exactly reduces to the ordinary metric entropy. Show that in condition (ii) of Theorem 8.9, metric entropy can be replaced by the Hausdorff entropy for any $0 < \alpha < 1$ to reach the same conclusion. (For $\alpha = 1/2$, the result implies the conclusion of Problem 8.11.)

8.14 (Kleijn and van der Vaart 2006) Show that for the well-specified case, the condition of prior concentration (8.48) in Theorem 8.36 reduces to the prior mass condition (i) in Theorem 8.11.

8.15 (Kleijn and van der Vaart 2006) If all points in a finite subset $\mathcal{P}^* \subset \mathcal{P}$ are at minimal Kullback-Leibler divergence, prove the following version of Theorem 8.36. Redefine covering numbers for testing under misspecification $N_{t,\text{mis}}(\epsilon, \mathcal{P}, d; P_0, \mathcal{P}^*)$ as the minimal number N of convex sets $B_1, \ldots, B_N$ of probability measures on $(\mathfrak{X}, \mathscr{X})$ needed to cover the set $\{P \in \mathcal{P}\colon \epsilon < d(P, \mathcal{P}^*) < 2\epsilon\}$ such that

$$\sup_{P^*\in\mathcal{P}^*}\inf_{P\in B_i}\sup_{0<\alpha<1} -\log P_0\left(\frac{p}{p^*}\right)^\alpha \geq \frac{\epsilon^2}{4}.$$

Suppose that there exists a sequence of strictly positive numbers ϵ_n with $\epsilon_n \to 0$ and $n\epsilon_n^2 \to \infty$ and a constant $L > 0$, such that for all n and all $\epsilon > \epsilon_n$:

$$\inf_{P^*\in\mathcal{P}^*}\Pi(B(\epsilon_n, P^*; P_0)) \geq e^{-Ln\epsilon_n^2}, \tag{8.60}$$

$$N_{t,\text{mis}}(\epsilon, \mathcal{P}, d; P_0, \mathcal{P}^*) \leq e^{n\epsilon_n^2}. \tag{8.61}$$

Then for every sufficiently large constant $M > 0$, $\Pi_n(P \in \mathcal{P}\colon d(P, \mathcal{P}^*) \geq M\epsilon_n | X_1, \ldots, X_n \to 0$ as $n \to \infty$ in P_0^n-probability.

8.16 (Kleijn and van der Vaart 2006) Consider a mixture model $p_F(x) = \int \psi(x; z)\,dF(z)$, $F \in \mathfrak{F}$, where ψ is continuous in z for all x and $\mathfrak{F}$ is a compact subset of $\mathfrak{M}$. Let p_0 be a density such that $K(p_0; p_F) < \infty$ for some $F \in \mathfrak{F}$. Then there exists $F^* \in \mathfrak{F}$ which minimizes $K(p_0; p_F)$ over $F \in \mathfrak{F}$, and F^* is unique if the family $\{p_F : F \in \mathfrak{F}\}$ is identifiable. (For example, the normal location model is identifiable [Teicher 1961].)

8.17 (Kleijn and van der Vaart 2006) Consider the setup of the last problem. Assume that p_0/p_{F^*} is bounded and $\Pi(B(\epsilon, P_{F^*}; P_0)) > 0$ for every $\epsilon > 0$. Then $P_0^n \Pi_n(F: d(P_F, P_{F^*}) \geq \epsilon | X_1, \ldots, X_n) \to 0$, where d is the weighted Hellinger distance defined by $d^2(P_{F_1}, P_{F_2}) = \frac{1}{2}\int(\sqrt{p_{F_1}} - \sqrt{p_{F_2}})^2(p_0/p_{F^*})\, d\mu$.

8.18 Suppose that for every n and $D > 0$ there exists a partition $\mathcal{P}_n = \mathcal{P}_{n,1} \cup \mathcal{P}_{n,2}$ such that (ii) of Theorem 8.11 holds and $\Pi_n(\mathcal{P}_{n,2}) = o(e^{-Dn\bar{\epsilon}_n^2})$. Then there also exist partitions such that (ii) and (iii) hold. [Hint: if $a_n(D) \to 0$ for every $D > 0$, then there exists $D_n \to \infty$ with $a_n(D_n) \to 0$.]