10

Bayes Procedures

In this chapter Bayes estimators are studied from a frequentist perspective. Both posterior measures and Bayes point estimators in smooth parametric models are shown to be asymptotically normal.

10.1 Introduction

In Bayesian terminology the distribution $P_{n,\theta}$ of an observation $\vec{X}_n$ under a parameter θ is viewed as the conditional law of $\vec{X}_n$ given that a random variable $\overline{\Theta}_n$ is equal to θ. The distribution Π of the "random parameter" $\overline{\Theta}_n$ is called the *prior distribution*, and the conditional distribution of $\overline{\Theta}_n$ given $\vec{X}_n$ is the *posterior distribution*. If $\overline{\Theta}_n$ possesses a density π and $P_{n,\theta}$ admits a density $p_{n,\theta}$ (relative to given dominating measures), then the density of the posterior distribution is given by Bayes' formula

$$p_{\overline{\Theta}_n \mid \vec{X}_n = x}(\theta) = \frac{p_{n,\theta}(x)\,\pi(\theta)}{\int p_{n,\theta}(x)\,d\Pi(\theta)}.$$

This expression may define a probability density even if π is not a probability density itself. A prior distribution with infinite mass is called *improper*.

The calculation of the posterior measure can be considered the ultimate aim of a Bayesian analysis. Alternatively, one may wish to obtain a "point estimator" for the parameter θ, using the posterior distribution. The posterior mean $\mathrm{E}(\overline{\Theta}_n \mid \vec{X}_n) = \int \theta\, p_{\overline{\Theta}_n \mid \vec{X}_n}(\theta)\,d\theta$ is often used for this purpose, but other location estimators are also reasonable.

A choice of point estimator may be motivated by a loss function. The *Bayes risk* of an estimator T_n relative to the loss function ℓ and prior measure Π is defined as

$$\int \mathrm{E}_\theta \ell(T_n - \theta)\,d\Pi(\theta) = \mathrm{E}\ell(T_n - \overline{\Theta}_n).$$

Here the expectation $\mathrm{E}_\theta \ell(T_n - \theta)$ is the risk function of T_n in the usual set-up and is identical to the conditional risk $\mathrm{E}\big(\ell(T_n - \overline{\Theta}_n) \mid \overline{\Theta}_n = \theta\big)$ in the Bayesian notation. The corresponding *Bayes estimator* is the estimator T_n that minimizes the Bayes risk. Because the Bayes risk can be written in the form $\mathrm{EE}\big(\ell(T_n - \overline{\Theta}_n) \mid \vec{X}_n\big)$, the value $T_n = T_n(x)$ minimizes, for every fixed x, the "posterior risk"

$$\mathrm{E}\big(\ell(T_n - \overline{\Theta}_n) \mid \vec{X}_n = x\big) = \frac{\int \ell(T_n - \theta)\, p_{n,\theta}(x)\,d\Pi(\theta)}{\int p_{n,\theta}(x)\,d\Pi(\theta)}.$$

Minimizing this expression may again be a well-defined problem even for prior densities of infinite total mass. For the loss function $\ell(y) = \|y\|^2$, the solution T_n is the posterior mean $\mathrm{E}(\overline{\Theta}_n \mid \vec{X}_n)$, for absolute loss $\ell(y) = \|y\|$, the solution is the posterior median.

Other Bayesian point estimators are the posterior mode, which reduces to the maximum likelihood estimator in the case of a uniform prior density; or a maximum probability estimator, such as the center of the smallest ball that contains at least posterior mass $1/2$ (the "posterior shorth" in dimension one).

If the underlying experiments converge, in a suitable sense, to a Gaussian location experiment, then all these possibilities are typically asymptotically equivalent. Consider the case that the observation consists of a random sample of size n from a density p_θ that depends smoothly on a Euclidean parameter θ. Thus the density $p_{n,\theta}$ has a product form, and, for a given prior Lebesgue density π, the posterior density takes the form

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

Typically, the distribution corresponding to this measure converges to the measure that is degenerate at the true parameter value θ_0, as $n \to \infty$. In this sense Bayes estimators are usually consistent. A further discussion is given in sections 10.2 and 10.4. To obtain a more interesting limit, we rescale the parameter in the usual way and study the sequence of posterior distributions of $\sqrt{n}(\overline{\Theta}_n - \theta_0)$, whose densities are given by

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

If the prior density π is continuous, then $\pi(\theta_0+h/\sqrt{n})$, for large n, behaves like the constant $\pi(\theta_0)$, and π cancels from the expression for the posterior density. For densities p_θ that are sufficiently smooth in the parameter, the sequence of models $(P_{\theta_0+h/\sqrt{n}} : h \in \mathbb{R}^k)$ is locally asymptotically normal, as discussed in Chapter 7. This means that the likelihood ratio processes $h \mapsto \prod_{i=1}^n p_{\theta_0+h/\sqrt{n}}/p_{\theta_0}(X_i)$ behave asymptotically as the likelihood ratio process of the normal experiment $\big(N(h, I_{\theta_0}^{-1}) : h \in \mathbb{R}^k\big)$. Then we may expect the preceding display to be asymptotically equivalent in distribution to

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

where $dN(\mu, \Sigma)$ denotes the density of the normal distribution. The expression in the preceding display is exactly the posterior density for the experiment $\big(N(h, I_{\theta_0}^{-1}) : h \in \mathbb{R}^k\big)$, relative to the (improper) Lebesgue prior distribution. The expression on the right shows that this is a normal distribution with mean X and covariance matrix $I_{\theta_0}^{-1}$.

This heuristic argument leads us to expect that the posterior distribution of $\sqrt{n}(\overline{\Theta}_n - \theta_0)$ "converges" under the true parameter θ_0 to the posterior distribution of the Gaussian limit experiment relative to the Lebesgue prior. The latter is equal to the $N(X, I_{\theta_0}^{-1})$-distribution, for X possessing the $N(0, I_{\theta_0}^{-1})$-distribution. The notion of convergence in this statement is a complicated one, because a posterior distribution is a conditional, and hence stochastic, probability measure, but there is no need to make the heuristics precise at this point. On the other hand, the convergence should certainly include that "nice" Euclidean-valued functionals applied to the posterior laws converge in distribution in the usual sense.

Consequently, a sequence of Bayes point estimators, which can be viewed as location functionals applied to the posterior distributions, should converge to the corresponding Bayes point estimator in the limit experiment. Most location estimators (all reasonable ones) map symmetric distributions, such as the normal distribution, into their center of symmetry. Then, the Bayes point estimator in the limit experiment is X, and we should expect Bayes point estimators to converge in distribution to the random vector X, that is, to a $N(0, I_{\theta_0}^{-1})$-distribution under θ_0. In particular, they are asymptotically efficient and asymptotically equivalent to maximum likelihood estimators (under regularity conditions).

A remarkable fact about this conclusion is that the limit distribution of a sequence of Bayes estimators does not depend on the prior measure. Apparently, for an increasing number of observations one's prior beliefs are erased (or corrected) by the observations. To make this true an essential assumption is that the prior distribution possesses a density that is smooth and positive in a neighborhood of the true value of the parameter. Without this property the conclusion fails. For instance, in the case in which one rigorously sticks to a fixed discrete distribution that does not charge θ_0, the sequence of posterior distributions of $\overline{\Theta}_n$ cannot even be consistent.

In the next sections we make the preceding heuristic argument precise. For technical reasons we separately consider the distributional approximation of the posterior distributions by a Gaussian one and the weak convergence of Bayes point estimators.

Even though the heuristic extends to convergence to other than Gaussian location experiments, we limit ourselves in this chapter to the locally asymptotically normal case. More precisely, we even assume that the observations are a random sample $X_1, \ldots, X_n$ from a distribution P_θ that admits a density p_θ with respect to a measure μ on a measurable space $(\mathcal{X}, \mathcal{A})$. The parameter θ is assumed to belong to a measurable subset Θ of $\mathbb{R}^k$ that contains the true parameter θ_0 as an interior point, and we assume that the maps $(\theta, x) \mapsto p_\theta(x)$ are jointly measurable.

All theorems in this chapter are frequentist in character in that we study the posterior laws under the assumption that the observations are a random sample from P_{θ_0} for some fixed, nonrandom θ_0. The alternative, which we do not consider, would be to make probability statements relative to the joint distribution of $(X_1, \ldots, X_n, \overline{\Theta}_n)$, given a fixed prior marginal measure for $\overline{\Theta}_n$ and with P_θ^n being the conditional law of $(X_1, \ldots, X_n)$ given $\overline{\Theta}_n$.

10.2 Bernstein–von Mises Theorem

The heuristic argument in the preceding section indicates that posterior distributions in differentiable parametric models converge to the Gaussian posterior distribution $N(X, I_{\theta_0}^{-1})$. The Bernstein–von Mises theorem makes this approximation rigorous and actually yields the approximation in a stronger sense than discussed so far. In Chapter 7 it is seen that the observation X in the limit experiment is the asymptotic analogue of the "locally sufficient" statistics

$$\Delta_{n,\theta_0} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} I_{\theta_0}^{-1} \dot{\ell}_{\theta_0}(X_i),$$

where $\dot{\ell}_\theta$ is the score function of the model. The Bernstein–von Mises theorem asserts that the total variation distance between the posterior distribution of $\sqrt{n}(\overline{\Theta}_n - \theta_0)$ and the random distribution $N(\Delta_{n,\theta_0}, I_{\theta_0}^{-1})$ converges to zero. Because $\Delta_{n,\theta_0} \rightsquigarrow X$, this has as a

consequence that the posterior distribution of $\sqrt{n}(\overline{\Theta}_n - \theta_0)$ converges, in any reasonable sense, in distribution to $N(X, I_{\theta_0}^{-1})$.

The conditions of the following version of the Bernstein–von Mises theorem are remarkably weak. Besides differentiability in quadratic mean of the model, it is assumed that there exists a sequence of uniformly consistent tests for testing $H_0 : \theta = \theta_0$ against $H_1 : \|\theta - \theta_0\| \geq \varepsilon$, for every $\varepsilon > 0$. In other words, it must be possible to separate the true value θ_0 from the complements of balls centered at θ_0. Because the theorem implies that the posterior distributions eventually concentrate on balls of radii $M_n/\sqrt{n}$ around θ_0, for every $M_n \to \infty$, this separation hypothesis appears to be very reasonable. Even more so, since, as is noted in Lemmas 10.4 and 10.6, under continuity and identifiability of the model, separation by tests of $H_0 : \theta = \theta_0$ from $H_1 : \|\theta - \theta_0\| \geq \varepsilon$ for a single (large) $\varepsilon > 0$ already implies separation for every $\varepsilon > 0$. Furthermore, if Θ is compact and the model continuous and identifiable, then even the separation condition is superfluous (because it is automatically satisfied).†

10.1 ***Theorem (Bernstein-von Mises).*** *Let the experiment $(P_\theta : \theta \in \Theta)$ be differentiable in quadratic mean at θ_0 with nonsingular Fisher information matrix I_{θ_0}, and suppose that for every $\varepsilon > 0$ there exists a sequence of tests ϕ_n such that*

$$P_{\theta_0}^n \phi_n \to 0, \qquad \sup_{\|\theta - \theta_0\| \geq \varepsilon} P_\theta^n (1 - \phi_n) \to 0. \tag{10.2}$$

Furthermore, let the prior measure be absolutely continuous in a neighborhood of θ_0 with a continuous positive density at θ_0. Then the corresponding posterior distributions satisfy

$$\left\| P_{\sqrt{n}(\overline{\Theta}_n - \theta_0) \mid X_1, \ldots, X_n} - N\big(\Delta_{n,\theta_0}, I_{\theta_0}^{-1}\big) \right\| \overset{P_{\theta_0}^n}{\to} 0.$$

Proof. Throughout the proof we rescale the parameter θ to the local parameter $h = \sqrt{n}(\theta - \theta_0)$. Let Π_n be the corresponding prior distribution on h (hence $\Pi_n(B) = \Pi(\theta_0 + B/\sqrt{n})$), and for a given set C let Π_n^C be the probability measure obtained by restricting Π_n to C and next renormalizing. Write $P_{n,h}$ for the distribution of $\vec{X}_n = (X_1, \ldots, X_n)$ under the original parameter $\theta_0 + h/\sqrt{n}$, and let $P_{n,C} = \int P_{n,h}\, d\Pi_n^C(h)$. Finally, let $\overline{H}_n = \sqrt{n}(\overline{\Theta}_n - \theta_0)$, and denote the posterior distributions relative to Π_n and Π_n^C by $P_{\overline{H}_n \mid \vec{X}_n}$ and $P^C_{\overline{H}_n \mid \vec{X}_n}$, respectively.

The proof consists of two steps. First, it is shown that the difference between the posterior measures relative to the priors Π_n and $\Pi_n^{C_n}$, for C_n the ball with radius M_n, is asymptotically negligible, for any $M_n \to \infty$. Next it is shown that the difference between $N(\Delta_{n,\theta_0}, I_{\theta_0}^{-1})$ and the posterior measures relative to the priors $\Pi_n^{C_n}$ converges to zero in probability, for some $M_n \to \infty$.

For U, a ball of fixed radius around zero, we have $P_{n,U} \lhd\rhd P_{n,0}$, because $P_{n,h_n} \lhd\rhd P_{n,0}$ for every bounded sequence h_n, by Theorem 7.2. Thus, when showing convergence to zero in probability, we may always exchange $P_{n,0}$ and $P_{n,U}$.

† Recall that a test is a measurable function of the observations taking values in the interval [0, 1]; in the present context this means a measurable function $\phi_n : \mathcal{X}^n \mapsto [0, 1]$.

Let C_n be the ball of radius M_n. By writing out the conditional densities we see that, for any measurable set B,

$$P_{\bar{H}_n \mid \vec{X}_n}(B) - P^{C_n}_{\bar{H}_n \mid \vec{X}_n}(B) = P_{\bar{H}_n \mid \vec{X}_n}\big(C_n^c \cap B\big) - P_{\bar{H}_n \mid \vec{X}_n}\big(C_n^c\big)\, P^{C_n}_{\bar{H}_n \mid \vec{X}_n}(B).$$

Taking the supremum over B yields the bound

$$\left\| P_{\bar{H}_n \mid \vec{X}_n} - P^{C_n}_{\bar{H}_n \mid \vec{X}_n} \right\| \le 2 P_{\bar{H}_n \mid \vec{X}_n}\big(C_n^c\big).$$

The right side will be shown to converge to zero in mean under $P_{n,U}$ for U a ball of fixed radius around zero. First, by assumption and because $P_{n,U} \lhd P_{n,0}$,

$$P_{n,U} P_{\bar{H}_n \mid \vec{X}_n}\big(C_n^c\big) = P_{n,U} P_{\bar{H}_n \mid \vec{X}_n}\big(C_n^c\big)(1 - \phi_n) + o(1).$$

Manipulating again the expressions for the posterior densities, we can rewrite the first term on the right as

$$\frac{\Pi_n\big(C_n^c\big)}{\Pi_n(U)} P_{n,C_n^c} P_{\bar{H}_n \mid \vec{X}_n}(U)(1 - \phi_n) \le \frac{1}{\Pi_n(U)} \int_{C_n^c} P_{n,h}(1 - \phi_n)\, d\Pi_n(h).$$

For the tests given in the statement of the theorem, the integrand on the right converges to zero pointwise, but this is not enough. By Lemma 10.3, there automatically exist tests ϕ_n for which the convergence is exponentially fast. For the tests given by the lemma the preceding display is bounded above by

$$\frac{1}{\Pi_n(U)} \int_{\|h\| \ge M_n} e^{-c(\|h\|^2 \wedge n)}\, d\Pi_n(h).$$

Here $\Pi_n(U) = \Pi(\theta_0 + U/\sqrt{n})$ is bounded below by a term of the order $1/\sqrt{n^k}$, by the positivity and continuity of the density π at θ_0. Splitting the integral into the domains $M_n \le \|h\| \le D\sqrt{n}$ and $\|h\| \ge D\sqrt{n}$ for $D \le 1$ sufficiently small that $\pi(\theta)$ is uniformly bounded on $\|\theta - \theta_0\| \le D$, we see that the expression is bounded above by a multiple of

$$\int_{\|h\| \ge M_n} e^{-c\|h\|^2} dh + \sqrt{n}^k e^{-cD^2 n}.$$

This converges to zero as $n, M_n \to \infty$.

In the second part of the proof, let C be the ball of fixed radius M around zero, and let $N^C(\mu, \Sigma)$ be the normal distribution restricted and renormalized to C. The total variation distance between two arbitrary probability measures P and Q can be expressed in the form $\|P - Q\| = 2\int (1 - p/q)^+ dQ$. It follows that

$$\begin{aligned}
&\tfrac{1}{2}\left\| N^C\big(\Delta_{n,\theta_0}, I_{\theta_0}^{-1}\big) - P^C_{\bar{H}_n \mid \vec{X}_n} \right\| \\
&\quad = \int \left(1 - \frac{dN^C\big(\Delta_{n,\theta_0}, I_{\theta_0}^{-1}\big)(h)}{1_C(h)\, p_{n,h}(\vec{X}_n)\pi_n(h) / \int_C p_{n,g}(\vec{X}_n)\pi_n(g)\, dg}\right)^+ dP^C_{\bar{H}_n \mid \vec{X}_n}(h) \\
&\quad \le \iint \left(1 - \frac{p_{n,g}(\vec{X}_n)\pi_n(g) dN^C\big(\Delta_{n,\theta_0}, I_{\theta_0}^{-1}\big)(h)}{p_{n,h}(\vec{X}_n)\pi_n(h)\, dN^C\big(\Delta_{n,\theta_0}, I_{\theta_0}^{-1}\big)(g)}\right)^+ dN^C\big(\Delta_{n,\theta_0}, I_{\theta_0}^{-1}\big)(g)\, dP^C_{\bar{H}_n \mid \vec{X}_n}(h),
\end{aligned}$$

because $(1 - \mathrm{E}Y)^+ \le \mathrm{E}(1 - Y)^+$. This can be further bounded by replacing the third occurrence of $N^C(\Delta_{n,\theta_0}, I_{\theta_0}^{-1})$ by a multiple of the uniform measure λ_C on C. By the dominated-convergence theorem, the double integral on the right side converges to zero in mean under $P_{n,C}$ if the integrand converges to zero in probability under the measure

$$P_{n,C}(dx)\, P^C_{\bar{H}_n \mid \vec{X}_n = x}(dh)\, \lambda_C(dg) = \Pi_n^C(dh)\, P_{n,h}(dx)\, \lambda_C(dg).$$

(Note that $P_{n,C}$ is the marginal distribution of $\vec{X}_n$ under the Bayesian model with prior Π_n^C.) Here Π_n^C is bounded up to a constant by λ_C for every sufficiently large n. Because $P_{n,h} \lhd\rhd P_{n,0}$ for every h, the sequence of measures on the right is contiguous with respect to the measures $\lambda_C(dh)\, P_{n,0}(dx)\, \lambda_C(dg)$. The integrand converges to zero in probability under the latter measure by Theorem 7.2 and the continuity of π at θ_0.

This is true for every ball C of fixed radius M and hence also for some $M_n \to \infty$. ■

10.3 ***Lemma.*** *Under the conditions of Theorem 10.1, there exists for every $M_n \to \infty$ a sequence of tests ϕ_n and a constant $c > 0$ such that, for every sufficiently large n and every $\|\theta - \theta_0\| \ge M_n/\sqrt{n}$,*

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

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

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

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

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

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

Because, by the differentiability of the model, $P_\theta \dot{\ell}_{\theta_0}^L - P_{\theta_0} \dot{\ell}_{\theta_0}^L = \big(P_{\theta_0} \dot{\ell}_{\theta_0}^L \dot{\ell}_{\theta_0}^T + o(1) \big)(\theta - \theta_0)$, the first term on the right is bounded below by $c\|\theta - \theta_0\|$ for some $c > 0$, for every θ that is sufficiently close to θ_0, say for $\|\theta - \theta_0\| < \varepsilon$. If $\omega_n = 0$, then the second term (without the minus sign) is bounded above by $\sqrt{M_n/n}$. Consequently, for every $c\|\theta - \theta_0\| \ge 2\sqrt{M_n/n}$, and hence for every $\|\theta - \theta_0\| \ge M_n/\sqrt{n}$ and every sufficiently large n,

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

by Hoeffding's inequality (e.g., Appendix B in [117]), for a sufficiently small constant C.

Next, consider the range $\|\theta - \theta_0\| > \varepsilon$ for an arbitrary fixed $\varepsilon > 0$. By assumption there exist tests ϕ_n such that

$$P_{\theta_0}^n \phi_n \to 0, \qquad \sup_{\|\theta - \theta_0\| > \varepsilon} P_\theta^n(1 - \phi_n) \to 0.$$

It suffices to show that these tests can be replaced, if necessary, by tests for which the convergence to zero is exponentially fast. Fix k large enough such that $P_{\theta_0}^k \phi_k$ and $P_\theta^k(1-\phi_k)$ are smaller than $1/4$ for every $\|\theta - \theta_0\| > \varepsilon$. Let $n = mk + r$ for $0 \le r < k$, and define $Y_{n,1}, \ldots, Y_{n,m}$ as ϕ_k applied in turn to $X_1, \ldots, X_k$, to $X_{k+1}, \ldots, X_{2k}$, and so forth. Let $\overline{Y}_{n,m}$ be their average and then define $\omega_n = 1\{\overline{Y}_{n,m} \ge 1/2\}$. Because $\mathrm{E}_\theta Y_{n,j} \ge 3/4$ for every $\|\theta - \theta_0\| > \varepsilon$ and every j, Hoeffding's inequality implies that

$$P_\theta^n(1-\omega_n) = P_\theta(\overline{Y}_{n,m} < 1/2) \le e^{-2m(\frac{1}{2}-\frac{3}{4})^2} \le e^{-m/8}.$$

Because m is proportional to n, this gives the desired exponential decay. Because $\mathrm{E}_{\theta_0} Y_{n,j} \le 1/4$, the expectations $P_{\theta_0}^n \omega_n$ are similarly bounded. ■

The Bernstein–von Mises theorem is sometimes written with a different "centering sequence." By Theorem 8.14 any sequence of standardized asymptotically efficient estimators $\sqrt{n}(\hat{\theta}_n - \theta)$ is asymptotically equivalent in probability to $\Delta_{n,\theta}$. Because the total variation distance

$$\left\| N\left(\Delta_{n,\theta}, I_\theta^{-1}\right) - N\left(\sqrt{n}(\hat{\theta}_n - \theta), I_\theta^{-1}\right) \right\|$$

is bounded by a multiple of $\left\| \Delta_{n,\theta} - \sqrt{n}(\hat{\theta}_n - \theta) \right\|$, any such sequence $\sqrt{n}(\hat{\theta}_n - \theta)$ may replace $\Delta_{n,\theta}$ in the Bernstein–von Mises theorem. By the invariance of the total variation norm under location and scale changes, the resulting statement can be written

$$\left\| P_{\overline{\Theta}_n \mid X_1, \ldots, X_n} - N\left(\hat{\theta}_n, \frac{1}{n} I_\theta^{-1}\right) \right\| \xrightarrow{P_\theta^n} 0.$$

Under regularity conditions this is true for the maximum likelihood estimators $\hat{\theta}_n$. Combining this with Theorem 5.39 we then have, informally,

$$P_{\overline{\Theta}_n \mid \hat{\theta}_n} \approx N\left(\hat{\theta}_n, \frac{1}{n} I_{\hat{\theta}_n}^{-1}\right) \quad \text{and} \quad P_{\hat{\theta}_n \mid \overline{\Theta}_n} \approx N\left(\overline{\Theta}_n, \frac{1}{n} I_{\overline{\Theta}_n}^{-1}\right),$$

since conditioning $\hat{\theta}_n$ on $\overline{\Theta}_n = \theta$ gives the usual "frequentist" distribution of $\hat{\theta}_n$ under θ. This gives a remarkable symmetry.

Le Cam's version of the Bernstein–von Mises theorem requires the existence of tests that are uniformly consistent for testing $H_0 : \theta = \theta_0$ versus $H_1 : \|\theta - \theta_0\| \ge \varepsilon$, for every $\varepsilon > 0$. Such tests certainly exist if there exist estimators T_n that are uniformly consistent, in that, for every $\varepsilon > 0$,

$$\sup_\theta P_\theta\left(\|T_n - \theta\| \ge \varepsilon\right) \to 0.$$

In that case, we can define $\phi_n = 1\left\{\|T_n - \theta_0\| \ge \varepsilon/2\right\}$. Thus the condition of the Bernstein–von Mises theorem that certain tests exist can be replaced by the condition that uniformly consistent estimators exist. This is often the case. For instance, the next lemma shows that this is the case for a Euclidean sample space $\mathcal{X}$ provided, for F_θ the distribution functions corresponding to the P_θ,

$$\inf_{\|\theta - \theta'\| > \varepsilon} \|F_\theta - F_{\theta'}\|_\infty > 0.$$

For compact parameter sets, this is implied by identifiability and continuity of the maps $\theta \mapsto F_\theta$. We generalize and formalize this in a second lemma, which shows that uniformity on compact subsets is always achievable if the model $(P_\theta : \theta \in \Theta)$ is differentiable in quadratic mean at every θ and the parameter θ is identifiable.

A class of measurable functions $\mathcal{F}$ is a *uniform Glivenko-Cantelli* class (in probability) if, for every $\varepsilon > 0$,

$$\sup_P P_P\big(\|\mathbb{P}_n - P\|_{\mathcal{F}} > \varepsilon\big) \to 0.$$

Here the supremum is taken over all probability measures P on the sample space, and $\|Q\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} |Qf|$. An example is the collection of indicators of all cells $(-\infty, t]$ in a Euclidean sample space.

10.4 ***Lemma.*** *Suppose that there exists a uniform Glivenko–Cantelli class $\mathcal{F}$ such that, for every $\varepsilon > 0$,*

$$\inf_{d(\theta,\theta')>\varepsilon} \|P_\theta - P_{\theta'}\|_{\mathcal{F}} > 0. \tag{10.5}$$

Then there exists a sequence of estimators that is uniformly consistent on Θ for estimating θ.

10.6 ***Lemma.*** *Suppose that Θ is σ-compact, $P_\theta \neq P_{\theta'}$ for every pair $\theta \neq \theta'$, and the maps $\theta \mapsto P_\theta$ are continuous for the total variation norm. Then there exists a sequence of estimators that is uniformly consistent on every compact subset of Θ.*

Proof. For the proof of the first lemma, define $\hat\theta_n$ to be a point of (near) minimum of the map $\theta \mapsto \|\mathbb{P}_n - P_\theta\|_{\mathcal{F}}$. Then, by the triangle inequality and the definition of $\hat\theta_n$, $\|P_{\hat\theta_n} - P_\theta\|_{\mathcal{F}} \le 2\|\mathbb{P}_n - P_\theta\|_{\mathcal{F}} + 1/n$, if the near minimum is chosen within distance $1/n$ of the true infimum. Fix $\varepsilon > 0$, and let δ be the positive number given in condition (10.5). Then

$$P_\theta\big(d(\hat\theta_n, \theta) > \varepsilon\big) \le P_\theta\Big(\big\|P_{\hat\theta_n} - P_\theta\big\|_{\mathcal{F}} \ge \delta\Big) \le P_\theta\Big(2\|\mathbb{P}_n - P_\theta\|_{\mathcal{F}} \ge \delta - \frac{1}{n}\Big).$$

By assumption, the right side converges to zero uniformly in θ.

For the proof of the second lemma, first assume that Θ is compact. Then there exists a uniform Glivenko-Cantelli class that satisfies the condition of the first lemma. To see this, first find a sequence $A_1, A_2, \ldots$ of measurable sets that separates the points P_θ. Thus, for every pair $\theta, \theta' \in \Theta$, if $P_\theta(A_i) = P_{\theta'}(A_i)$ for every i, then $\theta = \theta'$. A separating collection exists by the identifiability of the parameter, and it can be taken to be countable by the continuity of the maps $\theta \mapsto P_\theta$. (For a Euclidean sample space, we can use the cells $(-\infty, t]$ for t ranging over the vectors with rational coordinates. More generally, see the lemma below.) Let $\mathcal{F}$ be the collection of functions $x \mapsto i^{-1}1_{A_i}(x)$. Then the map $h : \Theta \mapsto \ell^\infty(\mathcal{F})$ given by $\theta \mapsto (P_\theta f)_{f\in\mathcal{F}}$ is continuous and one-to-one. By the compactness of Θ, the inverse $h^{-1} : h(\Theta) \mapsto \Theta$ is automatically uniformly continuous. Thus, for every $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\big\|h(\theta) - h(\theta')\big\|_{\mathcal{F}} \le \delta \quad \text{implies} \quad d(\theta, \theta') \le \varepsilon.$$

This means that (10.5) is satisfied. The class $\mathcal{F}$ is also a uniform Glivenko-Cantelli class, because by Chebyshev's inequality,

$$P_P\big(\|\mathbb{P}_n - P\|_{\mathcal{F}} > \varepsilon\big) \le \sum_f P_P\big(|\mathbb{P}_n f - Pf| > \varepsilon\big) \le \sum_i \frac{1}{n\varepsilon^2 i^2}.$$

This concludes the proof of the second lemma for compact Θ.

To remove the compactness condition, write Θ as the union of an increasing sequence of compact sets $K_1 \subset K_2 \subset \cdots$. For every m there exists a sequence of estimators $T_{n,m}$ that is uniformly consistent on K_m, by the preceding argument. Thus, for every fixed m,

$$a_{n,m} := \sup_{\theta \in K_m} P_\theta\left(d(T_{n,m}, \theta) \ge \frac{1}{m}\right) \to 0, \qquad n \to \infty.$$

Then there exists a sequence $m_n \to \infty$ such that $a_{n,m_n} \to 0$ as $n \to \infty$. It is not hard to see that $\hat{\theta}_n = T_{n,m_n}$ satisfies the requirements. ■

As a consequence of the second lemma, if there exists a sequence of tests ϕ_n such that (10.2) holds for some $\varepsilon > 0$, then it holds for every $\varepsilon > 0$. In that case we can replace the given sequence ϕ_n by the minimum of ϕ_n and the tests $1\{\|T_n - \theta_0\| \ge \varepsilon/2\}$ for a sequence of estimators T_n that is uniformly consistent on a sufficiently large subset of Θ.

10.7 ***Lemma.*** *Let the set of probability measures $\mathcal{P}$ on a measurable space $(\mathcal{X}, \mathcal{A})$ be separable for the total variation norm. Then there exists a countable subset $\mathcal{A}_0 \subset \mathcal{A}$ such that $P_1 = P_2$ on $\mathcal{A}_0$ implies $P_1 = P_2$ for every $P_1, P_2 \in \mathcal{P}$.*

Proof. The set $\mathcal{P}$ can be identified with a subset of $L_1(\mu)$ for a suitable probability measure μ. For instance, μ can be taken a convex linear combination of a countable dense set. Let $\mathcal{P}_0$ be a countable dense subset, and let $\mathcal{A}_0$ be the set of all finite intersections of the sets $p^{-1}(B)$ for p ranging over a choice of densities of the set $\mathcal{P}_0 \subset L_1(\mu)$ and B ranging over a countable generator of the Borel sets in $\mathbb{R}$.

Then every density $p \in \mathcal{P}_0$ is $\sigma(\mathcal{A}_0)$-measurable by construction. A density of a measure $P \in P - P_0$ can be approximated in $L_1(\mu)$ by a sequence from P_0 and hence can be chosen $\sigma(\mathcal{A}_0)$-measurable, without loss of generality.

Because $\mathcal{A}_0$ is intersection-stable (a "π-system"), two probability measures that agree on $\mathcal{A}_0$ automatically agree on the σ-field $\sigma(\mathcal{A}_0)$ generated by $\mathcal{A}_0$. Then they also give the same expectation to every $\sigma(\mathcal{A}_0)$-measurable function $f : \mathcal{X} \mapsto [0, 1]$. If the measures have $\sigma(\mathcal{A}_0)$-measurable densities, then they must agree on $\mathcal{A}$, because $P(A) = \mathrm{E}_\mu 1_A p = \mathrm{E}_\mu \mathrm{E}_\mu\big(1_A \mid \sigma(\mathcal{A}_0)\big)p$ if p is $\sigma(\mathcal{A}_0)$-measurable. ■

10.3 Point Estimators

The Bernstein–von Mises theorem shows that the posterior laws converge in distribution to a Gaussian posterior law in total variation distance. As a consequence, any location functional that is suitably continuous relative to the total variation norm applied to the sequence of

posterior laws converges to the same location functional applied to the limiting Gaussian posterior distribution. For most choices this means to X, or a $N(0, I_{\theta_0}^{-1})$-distribution.

In this section we consider more general Bayes point estimators that are defined as the minimizers of the posterior risk functions relative to some loss function. For a given loss function $\ell : \mathbb{R}^k \mapsto [0, \infty)$, let T_n, for fixed $X_1, \ldots, X_n$, minimize the posterior risk

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

It is not immediately clear that the minimizing values T_n can be selected as a measurable function of the observations. This is an implicit assumption, or otherwise the statements are to be understood relative to outer probabilities. We also make it an implicit assumption that the integrals in the preceding display exist, for almost every sequence of observations.

To derive the limit distribution of $\sqrt{n}(T_n - \theta_0)$, we apply general results on M-estimators, in particular the argmax continuous-mapping theorem, Theorem 5.56.

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

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

with strict inequality for at least one M.[†] This is true, for instance, for loss functions of the form $\ell(h) = \ell_0\big(\|h\|\big)$ for a nondecreasing function $\ell_0 : [0, \infty) \mapsto [0, \infty)$ that is not constant on $(0, \infty)$. Furthermore, we suppose that ℓ grows at most polynomially: For some constant $p \ge 0$,

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

10.8 *Theorem.* *Let the conditions of Theorem 10.1 hold, and let ℓ satisfy the conditions as listed, for a p such that $\int \|\theta\|^p\, d\Pi(\theta) < \infty$. Then the sequence $\sqrt{n}(T_n - \theta_0)$ converges under θ_0 in distribution to the minimizer of $t \mapsto \int \ell(t-h)\, dN(X, I_{\theta_0}^{-1})(h)$, for X possessing the $N(0, I_{\theta_0}^{-1})$-distribution, provided that any two minimizers of this process coincide almost surely. In particular, for every nonzero, subconvex loss function it converges to X.*

Proof. We adopt the notation as listed in the first paragraph of the proof of Theorem 10.1. The last assertion of the theorem is a consequence of Anderson's lemma, Lemma 8.5.

The standardized estimator $\sqrt{n}(T_n - \theta_0)$ minimizes the function

$$t \mapsto Z_n(t) = \frac{\int \ell(t-h)\, p_{n,h}(\vec{X}_n)\, d\Pi_n(h)}{\int p_{n,h}(\vec{X}_n)\, d\Pi_n(h)} = P_{\overline{H}_n \mid \vec{X}_n} \ell_t,$$

where ℓ_t is the function $h \mapsto \ell(t-h)$. The proof consists of three parts. First it is shown that integrals over the sets $\|h\| \ge M_n$ can be neglected for every $M_n \to \infty$. Next, it is proved that the sequence $\sqrt{n}(T_n - \theta_0)$ is uniformly tight. Finally, it is shown that the stochastic processes $t \mapsto Z_n(t)$ converge in distribution in the space $\ell^\infty(K)$, for every compact K, to the process

$$t \mapsto Z(t) = \int \ell(t-h)\, dN\big(X, I_{\theta_0}^{-1}\big)(h).$$

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

The sample paths of this limit process are continuous in t, in view of the subexponential growth of ℓ and the smoothness of the normal density. Hence the theorem follows from the argmax theorem, Corollary 5.58.

Let C_n be the ball of radius M_n for a given, arbitrary sequence $M_n \to \infty$. We first show that, for every measurable function f that grows subpolynomially of order p,

$$P_{\overline{H}_n \mid \vec{X}_n}(f 1_{C_n^c}) \overset{P_{n,0}}{\to} 0. \tag{10.9}$$

To see this, we utilize the tests ϕ_n for testing $H_0 : \theta = \theta_0$ that exist by assumption. In view of Lemma 10.3, these may be assumed without loss of generality to satisfy the stronger property as given in the statement of this lemma. Furthermore, they can be constructed to be nonrandomized (i.e., to have range $\{0, 1\}$). Then it is immediate that $(P_{\overline{H}_n \mid \vec{X}_n} f)\phi_n$ converges to zero in $P_{n,0}$-probability for every measurable function f. Next, by writing out the posterior densities, we see that, for U a fixed ball around the origin,

$$\begin{aligned} P_{n,U} P_{\overline{H}_n \mid \vec{X}_n}\big(f 1_{C_n^c}\big)(1 - \phi_n) &= \frac{1}{\Pi_n(U)} \int_{C_n^c} f(h) P_{n,h}\big[P_{\overline{H}_n \mid \vec{X}_n}(U)(1 - \phi_n)\big] \, d\Pi_n(h) \\ &\le \frac{1}{\Pi_n(U)} \int_{C_n^c} \big(1 + \|h\|^p\big) e^{-c(\|h\|^2 \wedge n)} \, d\Pi_n(h). \end{aligned}$$

Here $\Pi_n(U)$ is bounded below by a term of the order $1/\sqrt{n}^k$, by the positivity and continuity at θ_0 of the prior density π. Split the integral over the domains $M_n \le \|h\| \le D\sqrt{n}$ and $\|h\| \ge D\sqrt{n}$, and use the fact that $\int \|\theta\|^p \, d\Pi(\theta) < \infty$ to bound the right side of the display by terms of the order $e^{-AM_n^2}$ and $\sqrt{n}^{k+p} e^{-Bn}$, for some $A, B > 0$. These converge to zero, whence (10.9) has been proved.

Define $\bar{\ell}(M)$ as the supremum of $\ell(h)$ over the ball of radius M, and $\underline{\ell}(M)$ as the infimum over the complement of this ball. By assumption, there exists $\delta > 0$ such that $\eta := \underline{\ell}(2\delta) - \bar{\ell}(\delta) > 0$. Let U be the ball of radius δ around 0. For every $\|t\| \ge 3M_n$ and sufficiently large M_n, we have $\ell(t - h) - \ell(-h) \ge \eta$ if $h \in U$, and $\ell(t - h) - \ell(-h) \ge \underline{\ell}(2M_n) - \bar{\ell}(M_n) \ge 0$ if $h \in U^c \cap C_n$, by assumption. Therefore,

$$\begin{aligned} Z_n(t) - Z_n(0) &= P_{\overline{H}_n \mid \vec{X}_n}\Big[\big(\ell(t - h) - \ell(-h)\big)\big(1_U + 1_{U^c \cap C_n} + 1_{C_n^c}\big)\Big] \\ &\ge \eta P_{\overline{H}_n \mid \vec{X}_n}(U) - P_{\overline{H}_n \mid \vec{X}_n}\big(\ell(-h) 1_{C_n^c}\big). \end{aligned}$$

Here the posterior probability $P_{\overline{H}_n \mid \vec{X}_n}(U)$ of U converges in distribution to $N(X, I_{\theta_0}^{-1})(U)$, by the Bernstein–von Mises theorem. This limit is positive almost surely. The second term in the preceding display converges to zero in probability by (10.9). Conclude that the infimum of $Z_n(t) - Z_n(0)$ over the set of t with $\|t\| \ge 3M_n$ is bounded below by variables that converge in distribution to a strictly positive variable. Thus this infimum is positive with probability tending to one. This implies that the probability that $t \mapsto Z_n(t)$ has a minimizer in the set $\|t\| \ge 3M_n$ converges to zero. Because this is true for any $M_n \to \infty$, it follows that the sequence $\sqrt{n}(T_n - \theta_0)$ is uniformly tight.

Let C be the ball of fixed radius M around 0, and fix some compact set $K \subset \mathbb{R}^k$. Define stochastic processes

$$\begin{aligned} Z_{n,M}(t) &= P_{\overline{H}_n \mid \vec{X}_n}(\ell_t 1_C), \qquad & W_{n,M} &= N\big(\Delta_{n,\theta_0}, I_{\theta_0}^{-1}\big)(\ell_t 1_C), \\ & & W_M &= N\big(X, I_{\theta_0}^{-1}\big)(\ell_t 1_C). \end{aligned}$$

The function $h \mapsto \ell(t-h)1_C(h)$ is bounded, uniformly if t ranges over the compact K. Hence, by the Bernstein–von Mises theorem, $Z_{n,M} - W_{n,M} \xrightarrow{P} 0$ in $\ell^\infty(K)$ as $n \to \infty$, for every fixed M. Second, by the continuous-mapping theorem, $W_{n,M} \rightsquigarrow W_M$ in $\ell^\infty(K)$, as $n \to \infty$, for fixed M. Next $W_M \xrightarrow{P} Z$ in $\ell^\infty(K)$ as $M \to \infty$, or equivalently $C \uparrow \mathbb{R}^k$. Conclude that there exists a sequence $M_n \to \infty$ such that the processes $Z_{n,M_n} \rightsquigarrow Z$ in $\ell^\infty(K)$. Because, by (10.9), $Z_n(t) - Z_{n,M_n}(t) \xrightarrow{P} 0$, we finally conclude that $Z_n \rightsquigarrow Z$ in $\ell^\infty(K)$. ■

*10.4 Consistency

A sequence of posterior measures $P_{\overline{\Theta}_n \mid X_1,\ldots,X_n}$ is called *consistent* under θ if under P_θ^∞-probability it converges in distribution to the measure δ_θ that is degenerate at θ, in probability; it is *strongly consistent* if this happens for almost every sequence $X_1, X_2, \ldots$.

Given that, usually, ordinarily consistent point estimators of θ exist, consistency of posterior measures is a modest requirement. If we could know θ with almost complete accuracy as $n \to \infty$, then we would use a Bayes estimator only if this would also yield the true value with similar accuracy. Fortunately, posterior measures are usually consistent. The following famous theorem by Doob shows that under hardly any conditions we already have consistency under almost every parameter.

Recall that Θ is assumed to be Euclidean and the maps $\theta \mapsto P_\theta(A)$ to be measurable for every measurable set A.

10.10 ***Theorem (Doob's consistency theorem).*** *Suppose that the sample space $(\mathcal{X}, \mathcal{A})$ is a subset of Euclidean space with its Borel σ-field. Suppose that $P_\theta \neq P_{\theta'}$ whenever $\theta \neq \theta'$. Then for every prior probability measure Π on Θ the sequence of posterior measures is consistent for Π-almost every θ.*

Proof. On an arbitrary probability space construct random vectors $\overline{\Theta}$ and $X_1, X_2, \ldots$ such that $\overline{\Theta}$ is marginally distributed according to Π and such that given $\overline{\Theta} = \theta$ the vectors $X_1, X_2, \ldots$ are i.i.d. according to P_θ. Then the posterior distribution based on the first n observations is $P_{\overline{\Theta} \mid X_1,\ldots,X_n}$. Let Q be the distribution of $(X_1, X_2, \ldots, \overline{\Theta})$ on $\mathcal{X}^\infty \times \Theta$.

The main part of the proof consists of showing that there exists a measurable function $h : \mathcal{X}^\infty \mapsto \Theta$ with

$$h(x_1, x_2, \ldots) = \theta, \qquad Q\text{-a.s..} \tag{10.11}$$

Suppose that this is true. Then, for any bounded, measurable function $f : \Theta \mapsto \mathbb{R}$, by Doob's martingale convergence theorem,

$$\begin{aligned} \mathrm{E}\big(f(\overline{\Theta}) \mid X_1, \ldots, X_n\big) &\to \mathrm{E}\big(f(\overline{\Theta}) \mid X_1, X_2, \ldots\big) \\ &= f\big(h(X_1, X_2, \ldots)\big), \qquad Q\text{-a.s..} \end{aligned}$$

By Lemma 2.25 there exists a countable collection $\mathcal{F}$ of bounded, continuous functions f that are determining for convergence in distribution. Because the countable union of the associated null sets on which the convergence of the preceding display fails is a null set, we have that

$$P_{\overline{\Theta} \mid X_1,\ldots,X_n} \rightsquigarrow \delta_{h(X_1,X_2,\ldots)}, \qquad Q\text{-a.s..}$$

This statement refers to the marginal distribution of $(X_1, X_2, \ldots)$ under Q. We wish to translate it into a statement concerning the P_θ^∞-measures. Let $C \subset \mathcal{X}^\infty \times \Theta$ be the intersection of the sets on which the weak convergence holds and on which (15.9) is valid. By Fubini's theorem

$$1 = Q(C) = \iint 1_C(x, \theta)\, dP_\theta^\infty(x)\, d\Pi(\theta) = \int P_\theta^\infty(C_\theta)\, d\Pi(\theta),$$

where $C_\theta = \{x : (x, \theta) \in C\}$ is the horizontal section of C at height θ. It follows that $P_\theta^\infty(C_\theta) = 1$ for Π-almost every θ. For every θ such that $P_\theta^\infty(C_\theta) = 1$, we have that $(x, \theta) \in C$ for P_θ^∞-almost every sequence $x_1, x_2, \ldots$ and hence

$$P_{\Theta \mid X_1 = x_1, \ldots, X_n = x_n} \rightsquigarrow \delta_{h(x_1, x_2, \ldots)} = \delta_\theta.$$

This is the assertion of the theorem.

In order to establish (15.9), call a measurable function $f : \Theta \mapsto \mathbb{R}$ *accessible* if there exists a sequence of measurable functions $h_n : \mathcal{X}^n \mapsto \mathbb{R}$ such that

$$\iint \bigl|h_n(x) - f(\theta)\bigr| \wedge 1\, dQ(x, \theta) \to 0.$$

(Here we abuse notation in viewing h_n also as a measurable function on $\mathcal{X}^\infty \times \Theta$.) Then there also exists a (sub)sequence with $h_n(x) \to f(\theta)$ almost surely under Q, whence every accessible function f is almost everywhere equal to an $\mathcal{A}^\infty \times \{\emptyset, \Theta\}$-measurable function. This is a measurable function of $x = (x_1, x_2, \ldots)$ alone. If we can show that the functions $f(\theta) = \theta_i$ are accessible, then (15.9) follows. We shall in fact show that every Borel measurable function is accessible.

By the strong law of large numbers, $h_n(x) = \sum_{i=1}^n 1_A(x_i) \to P_\theta(A)$ almost surely under P_θ^∞, for every θ and measurable set A. Consequently, by the dominated convergence theorem,

$$\iint \bigl|h_n(x) - P_\theta(A)\bigr|\, dQ(x, \theta) \to 0.$$

Thus each of the functions $\theta \mapsto P_\theta(A)$ is accessible.

Because $(\mathcal{X}, \mathcal{A})$ is Euclidean by assumption, there exists a countable measure-determining subcollection $\mathcal{A}_0 \subset \mathcal{A}$. The functions $\theta \mapsto P_\theta(A)$ are measurable by assumption and separate the points of Θ as A ranges over $\mathcal{A}_0$, in view of the choice of $\mathcal{A}_0$ and the identifiability of the parameter θ. This implies that these functions generate the Borel σ-field on Θ, in view of Lemma 10.12.

The proof is complete once it is shown that every function that is measurable in the σ-field generated by the accessible functions (which is the Borel σ-field) is accessible. From the definition it follows easily that the set of accessible functions is a vector space, contains the constant functions, is closed under monotone limits, and is a lattice. The desired result therefore follows by a monotone class argument, as in Lemma 10.13. ■

The merit of the preceding theorem is that it imposes hardly any conditions, but its drawback is that it gives the consistency only up to null sets of possible parameters (depending on the prior). In certain ways these null sets can be quite large, and examples have

been constructed where Bayes estimators behave badly. To guarantee consistency under every parameter it is necessary to impose some further conditions. Because in this chapter we are mainly concerned with asymptotic normality of Bayes estimators (which implies consistency with a rate), we omit a discussion.

10.12 ***Lemma.*** *Let $\mathcal{F}$ be a countable collection of measurable functions $f : \Theta \subset \mathbb{R}^k \mapsto \mathbb{R}$ that separates the points of Θ. Then the Borel σ-field and the σ-field generated by $\mathcal{F}$ on Θ coincide.*

Proof. By assumption, the map $h : \Theta \mapsto \mathbb{R}^{\mathcal{F}}$ defined by $h(\theta) f = f(\theta)$ is measurable and one-to-one. Because $\mathcal{F}$ is countable, the Borel σ-field on $\mathbb{R}^{\mathcal{F}}$ (for the product topology) is equal to the σ-field generated by the coordinate projections. Hence the σ-fields generated by h and $\mathcal{F}$ (viewed as Borel measurable maps in $\mathbb{R}^{\mathcal{F}}$ and $\mathbb{R}$, respectively) on Θ are identical. Now h^{-1}, defined on the range of h, is automatically Borel measurable, by Proposition 8.3.5 in [24], and hence Θ and $h(\Theta)$ are Borel isomorphic. ■

10.13 ***Lemma.*** *Let $\mathcal{F}$ be a linear subspace of $\mathcal{L}_1(\Pi)$ with the properties*

(i) if $f, g \in \mathcal{F}$, then $f \wedge g \in \mathcal{F}$;

(ii) if $0 \le f_1 \le f_2 \le \cdots \in \mathcal{F}$ and $f_n \uparrow f \in \mathcal{L}_1(\Pi)$, then $f \in \mathcal{F}$;

(iii) $1 \in \mathcal{F}$.

Then $\mathcal{F}$ contains every $\sigma(\mathcal{F})$-measurable function in $\mathcal{L}_1(\Pi)$.

Proof. Because any $\sigma(\mathcal{F})$-measurable nonnegative function is the monotone limit of a sequence of simple functions, it suffices to prove that $1_A \in \mathcal{F}$ for every $A \in \sigma(\mathcal{F})$. Define $\mathcal{A}_0 = \{A : 1_A \in \mathcal{F}\}$. Then $\mathcal{A}_0$ is an intersection-stable Dynkin system and hence a σ-field. Furthermore, for every $f \in \mathcal{F}$ and $\alpha \in \mathbb{R}$, the functions $n(f - \alpha)^+ \wedge 1$ are contained in $\mathcal{F}$ and increase pointwise to $1_{\{f > \alpha\}}$. It follows that $\{f > \alpha\} \in \mathcal{A}_0$. Hence $\sigma(\mathcal{F}) \subset \mathcal{A}_0$. ■

Notes

The Bernstein–von Mises theorem has that name, because, as Le Cam and Yang [97] write, it was first discovered by Laplace. The theorem that is presented in this chapter is considerably more elegant than the results by these early authors, and also much better than the result in Le Cam [91], who revived the theorem in order to prove results on superefficiency. We adapted it from Le Cam [96] and Le Cam and Yang [97].

Ibragimov and Hasminskii [80] discuss the convergence of Bayes point estimators in greater generality, and also cover non-Gaussian limit experiments, but their discussion of the i.i.d. case as discussed in the present chapter is limited to bounded parameter sets and requires stronger assumptions. Our treatment uses some elements of their proof, but is heavily based on Le Cam's Bernstein–von Mises theorem. Inspection of the proof shows that the conditions on the loss function can be relaxed significantly, for instance allowing exponential growth.

Doob's theorem originates in [39]. The potential null sets of inconsistency that it leaves open really exist in some situations particularly if the parameter set is infinite dimensional,

and have attracted much attention. See [34], which is accompanied by evaluations of the phenomenon by many authors, including Bayesians.

PROBLEMS

1. Verify the conditions of the Bernstein–von Mises theorem for the experiment where P_θ is the Poisson measure of mean θ.
2. Let P_θ be the k-dimensional normal distribution with mean θ and covariance matrix the identify. Find the a posteriori law for the prior $\Pi = N(\tau, \Lambda)$ and some nonsingular matrix Λ. Can you see directly that the Bernstein–von Mises theorem is true in this case?
3. Let P_θ be the Bernoulli distribution with mean θ. Find the posterior distribution relative to the beta-prior measure, which has density

$$\theta \mapsto \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}1_{(0,1)}(\theta).$$

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