Notes on Bayesian Asymptotics

Adam N. Smith

July 31, 2021

This note reviews some of the key results in Bayesian asymptotics. We consider the following questions: Where do posteriors concentrate mass as the sample size gets large? Are posteriors consistent in the frequentist sense? What shape does the limiting posterior have? We start with a general result on the consistency of posterior distributions (Doob's theorem), and then present results on the asymptotic normality for parametric models (Bernstein-von Mises theorem).

1 Posterior Consistency

Frequentist consistency provides guarantees about the limiting location of an estimator. Specifically, an estimator is said to be consistent if it converges to the true parameter as the sample size grows. We start by examining the consistency of posterior distributions – i.e., the concentration of posterior mass in neighborhoods of the true parameter. The first result is referred to as Doob's theorem, which shows that under very mild assumptions, the posterior will concentrate its mass around the truth.

First, some preliminary notation. Let P_θ denote a distribution indexed by parameters $\theta \in \Theta$. The idea of consistency requires being precise about parameters being "close", so let $d(\theta, a)$ be the metric of Θ which induces ϵ-neighborhoods of the form $\mathcal{N}_\epsilon(a) = \{\theta : d(\theta, a) < \epsilon\}$. We can now state Doob's theorem.

Theorem 1 (Doob 1949)**.** *Assume the sampling model P_θ is identifiable in the sense that $\theta \neq \theta'$ implies $P_\theta \neq P_{\theta'}$. Then there exists $\Theta_\star \subseteq \Theta$ with $\Pi(\Theta_\star) = 1$ such that for each $\theta_\star \in \Theta_\star$, if $\mathbf{X}_n = (X_1, \ldots, X_n)$ are iid $P_{\theta_\star}$, then for all $\epsilon > 0$, we have*

$$\lim_{n\to\infty} \mathbb{P}(\theta \in \mathcal{N}_\epsilon(\theta_\star)|\mathbf{X}_n) = 1.$$

Proof. See chapter 7.4.1 of Schervish (1995), chapter 10.4 of van der Vaart (1998), chapter 1.3 of Ghosh and Ramamoorthi (2003), or Miller (2018). □

In words, Doob's theorem says that the *for any prior distribution*, the posterior is guaranteed to concentrate in a neighborhood of the true parameter $\theta_\star \in \Theta_\star$ as long as $\Theta_\star$ has strictly positive measure under the prior. Or, in other words, the posterior is consistent everywhere except for a set of values having measure zero under the prior. Doob's theorem thus provides a form of "internal consistency". What is remarkable is its generality: Bayes estimators are guaranteed to be consistent only under an assumption of identifiability.

A major criticism of Doob's theorem is that it allows for consistency to fail on a null set, and the size of this null set actually depends on the prior. To see why this is a concern, consider the following example: let $X_1, \ldots, X_n \sim N(\theta, 1)$ where $\theta \in \Theta = \mathbb{R}$ and define the prior $\pi(\theta) = \delta_0$ which is a point mass at 0. Since the posterior is also a point mass at zero, it is inconsistent on $\Theta_0 = \mathbb{R} \backslash \{0\}$ (i.e., everywhere except zero!) but because Θ_0 has measure zero under the prior, the posterior is still consistent by Doob's theorem. In this sense, Doob's theorem can be perceived as an unsatisfactory result because the posterior is technically consistent but practically useless. What is also concerning is that the size of the null sets can be quite large in nonparametric settings, even for "reasonably" chosen priors. See examples provided by Freedman (1963) and Diaconis and Freedman (1986).

General posterior consistency results with stronger prior conditions are given by Schwartz (1965), Barron et al. (1999), and Ghosal et al. (1999b). Specifically, Schwartz (1965) replaces the identifiability condition with a testing condition, and also requires the prior to place enough positive mass on each Kullback-Leibler (KL) neighborhood of the true density. Barron et al. (1999) and Ghosal et al. (1999b) build on the KL prior property and provide consistency results in the Hellinger (i.e., total variation) distance. The KL property of priors is now known to be a key property in establishing posterior consistency (although on its own, it is still not a sufficient condition for posterior consistency). For example, Ghosal et al. (1999a) show that the type of inconsistency outlined in Diaconis and Freedman (1986) for estimating location parameters can be avoided if the prior satisfies the KL property. Ghosal and van der Vaart (2017) provide examples of nonparametric priors satisfying the KL property, including Pólya trees, Dirichlet process mixtures, Bernstein polynomial priors, and Gaussian process priors.

2 Asymptotic Normality for Parametric Models

We now turn to the question of the shape of the limiting posterior distribution. We focus on the simplest case of smooth parametric models with iid data. Note that regularity conditions ensuring well-behaved likelihood and prior will be much stronger than the assumptions underling posterior consistency stated above. For notation, let $\mathbf{X}_n = (X_1, \ldots, X_n)$ denote the data, $p(\mathbf{X}_n|\theta)$ the likelihood, and

$$\hat{\theta}_n = \arg\max_{\theta} p(\mathbf{X}_n|\theta)$$

the MLE. Also define the Fisher information matrix as

$$\mathcal{I}_n(s) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log p(\mathbf{X}_n|\theta)\right)^2\Bigg|_{\theta=s}\right] = -\mathbb{E}\left[\left(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p(\mathbf{X}_n|\theta)\right)\Bigg|_{\theta=s}\right]$$

where the expectation is taken with respect to the data.

Our benchmark for establishing asymptotic results will be the asymptotic results for the MLE. Under some regularity conditions, the MLE will have the following Gaussian limiting distribution:

$$\sqrt{n}(\hat{\theta}_n - \theta)|\theta = \theta_\star \overset{d}{\to} N(0, \mathcal{I}_n(\theta_\star)^{-1}). \tag{1}$$

Note that here we are explicitly including the conditioning argument $\theta = \theta_\star$ to remind ourselves that in classical large-sample theory, the parameter is fixed and so the randomness of $\sqrt{n}(\hat{\theta}_n - \theta)$ is inherited only through the randomness in the estimator $\hat{\theta}_n$, which in turn inherits its randomness through the data $\mathbf{X}_n$.

The objective of this section is to outline the Bayesian analogue of (1). That is, we would like to show that the posterior distribution of $\sqrt{n}(\theta - \hat{\theta}_n)$ is equal to a Gaussian distribution in the limit. In the Bayesian approach, however, we will be conditioning on the data $\mathbf{X}_n$ so the MLE is fixed. Therefore, $\sqrt{n}(\theta - \hat{\theta}_n)$ is random only because of the random parameter θ. The desired result is that:

$$\sqrt{n}(\theta - \hat{\theta}_n)|\mathbf{X}_n \overset{d}{\to} N(0, \mathcal{I}_n(\theta_\star)^{-1}) \tag{2}$$

which implies an asymptotic equivalence between the limiting posterior and the limiting distribution of the MLE. This result dates back to Laplace (1809) but is now known as the Bernstein-von Mises theorem.

Heuristic Argument We follow Bernardo and Smith (1994) and first outline a heuristic argument for why the posterior should look like a normal density in the limit. Take a second-order Taylor series expansion of the log likelihood and the log prior around their respective maxima, $\hat{\theta}_n$ (the MLE) and $\hat{\theta}_0$ (the prior mode):

$$\begin{aligned}
\log p(\mathbf{X}_n|\theta) &= \log p(\mathbf{X}_n|\hat{\theta}_n) + \frac{1}{2}(\theta - \hat{\theta}_n)'\Big(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p(X_i|\theta)\Big)\Big|_{\theta=\hat{\theta}_n}(\theta - \hat{\theta}_n) + R_n \\
&= \log p(\mathbf{X}_n|\hat{\theta}_n) - \frac{1}{2}(\theta - \hat{\theta}_n)'\Lambda_n(\hat{\theta}_n)(\theta - \hat{\theta}_n) + R_n \\
\log \pi(\theta) &= \log \pi(\hat{\theta}_0) + \frac{1}{2}(\theta - \hat{\theta}_0)'\Big(\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log \pi(\theta)\Big)\Big|_{\theta=\hat{\theta}_0}(\theta - \hat{\theta}_0) + R_0 \\
&= \log \pi(\hat{\theta}_0) - \frac{1}{2}(\theta - \hat{\theta}_0)'\Lambda_0(\theta - \hat{\theta}_0) + R_0
\end{aligned}$$

where

$$\begin{aligned}
\Lambda_n(\hat{\theta}_n) &= \left(-\frac{\partial^2}{\partial\theta^2}\log p(x|\theta)\right)\Bigg|_{\theta=\hat{\theta}_n} \\
\Lambda_0 &= \left(-\frac{\partial^2}{\partial\theta^2}\log \pi(\theta)\right)\Bigg|_{\theta=\hat{\theta}_0}
\end{aligned}$$

are the observed and prior information matrices, respectively. Note that the first-order terms are absent in the Taylor series expansions above because the gradients evaluated at the maxima are equal to zero by definition. Under some regularity conditions that ensure the likelihood and prior are sufficiently smooth, then R_n and R_0 will both be small so we can write the posterior as:

$$\begin{aligned}
\pi(\theta|\mathbf{X}_n) &\propto \exp\Bigg\{\log p(\mathbf{X}_n|\theta) + \log\pi(\theta)\Bigg\} \\
&\approx \exp\Bigg\{\log p(\mathbf{X}_n|\hat{\theta}_n) - \frac{1}{2}(\theta - \hat{\theta}_n)'\Lambda_n(\hat{\theta}_n)(\theta - \hat{\theta}_n) + \log\pi(\hat{\theta}_0) - \frac{1}{2}(\theta - \hat{\theta}_0)\Lambda_0(\theta - \hat{\theta}_0)\Bigg\} \\
&\propto \exp\Bigg\{-\frac{1}{2}(\theta - \hat{\theta}_n)'\Lambda_n(\hat{\theta}_n)(\theta - \hat{\theta}_n) - \frac{1}{2}(\theta - \hat{\theta}_0)'\Lambda_0(\theta - \hat{\theta}_0)\Bigg\} \\
&\propto \exp\Bigg\{-\frac{1}{2}(\theta - \tilde{\theta}_n)'\tilde{\Lambda}_n(\theta - \tilde{\theta}_n)\Bigg\}
\end{aligned}$$

where $\tilde{\Lambda}_n = \Lambda_n(\hat{\theta}_n) + \Lambda_0$ is the posterior precision and $\tilde{\theta}_n = \tilde{\Lambda}_n^{-1}(\Lambda_n(\hat{\theta}_n)\hat{\theta}_n + \Lambda_0\hat{\theta}_0)$ is the posterior mean. This shows that the posterior has the kernel of a $N(\tilde{\theta}_n, \tilde{\Lambda}_n^{-1})$ distribution with $\tilde{\theta}_n \to \hat{\theta}_n$ and $\tilde{\Lambda}_n \to \Lambda_n(\hat{\theta}_n)$ as n gets large.

Formal Statement We now provide a more formal statement of the Bernstein-von Mises theorem. The proof is quite involved and so we only outline a sketch. Details can be found in chapter 7.4.2 of Schervish (1995), chapter 10.2 of van der Vaart (1998), or chapter 1.4 of Ghosh and Ramamoorthi (2003).

Theorem 2 (Bernstein-von Mises)**.** *Let $\mathbf{X}_n = (X_1, \ldots, X_n)$ be iid $p(x|\theta_\star)$ for some $\theta_\star \in \Theta$ where $\Theta \subset \mathbb{R}^k$. Let $\ell_n = \log p(\mathbf{X}_n|\theta)$ and define $\psi = \Sigma_n^{-1/2}(\theta - \hat{\theta}_n)$ where $\Sigma_n = [-\ell_n''(\hat{\theta}_n)]^{-1}$ is the inverse of the observed information matrix. Then under some regularity conditions, the posterior density of ψ converges to the density of a $N(0, I_k)$ distribution. That is, for each subset $B \subseteq \Theta$ and each $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}\left(\sup_{\psi\in B}\left|\pi_{\psi|\mathbf{X}_n}(\psi|\mathbf{X}_n) - \phi(\psi)\right| > \epsilon\right) = 0 \tag{3}$$

where $\phi(\psi) = (2\pi)^{-k/2}\exp\left(-\frac{1}{2}\|\psi\|^2\right)$.

Proof. A very rough sketch of the proof is as follows. First, define the posterior of the model parameters θ:

$$\pi_{\theta|\mathbf{X}_n}(\theta|\mathbf{X}_n) = \frac{p(\mathbf{X}_n|\theta)\pi(\theta)}{\int_\Theta p(\mathbf{X}_n|\theta)\pi(\theta)d\theta} = \frac{p(\mathbf{X}_n|\theta)\pi(\theta)}{m(\mathbf{X}_n)}.$$

Then by the change of variables theorem we can then write the posterior of the transformed variables ψ as:

$$\pi_{\psi|\mathbf{X}_n}(\psi|\mathbf{X}_n) = \pi_{\theta|\mathbf{X}_n}(\psi|\mathbf{X}_n)|\Sigma_n|^{1/2}.$$

The proof amounts to showing that this object converges to the standard normal density function. This is facilitated by multiplying and dividing by the likelihood evaluated at the MLE and then splitting this posterior into a product of two pieces.

$$\begin{aligned}\pi_{\psi|\mathbf{X}_n}(\psi|\mathbf{X}_n) &= \frac{p(\mathbf{X}_n|\hat{\theta}_n)}{p(\mathbf{X}_n|\hat{\theta}_n)}\pi_{\psi|\mathbf{X}_n}(\psi|\mathbf{X}_n) \\ &= \left(\frac{p(\mathbf{X}_n|\hat{\theta}_n)\pi(\hat{\theta} + \Sigma_n^{1/2}\psi)|\Sigma_n|^{1/2}}{m(\mathbf{X}_n)}\right)\left(\frac{p(\mathbf{X}_n|\hat{\theta}_n + \Sigma_n^{1/2}\psi)}{p(\mathbf{X}_n|\hat{\theta}_n)}\right).\end{aligned}$$

Then some heavy math follows (see details in Schervish, 1995, pg. 437-441) to show that the first term converges to the appropriate normalizing constant and the second term converges to the kernel of a normal density.

$$\frac{p(\mathbf{X}_n|\hat{\theta}_n)\pi(\hat{\theta}_n + \Sigma_n^{1/2}\psi)|\Sigma_n|^{1/2}}{m(\mathbf{X}_n)} \to (2\pi)^{-k/2}$$

$$\frac{p(\mathbf{X}_n|\hat{\theta}_n + \Sigma_n^{1/2}\psi)}{p(\mathbf{X}_n|\hat{\theta}_n)} \to \exp\left(-\frac{1}{2}\|\psi\|^2\right)$$

So together, the product converges to a standard normal density. □

The Berstein-von Mises theorem guarantees that inferences based on the posterior distribution are "asymptotically correct" in the frequentist sense. That is, confidence intervals based on the sampling distribution of an efficient estimator and credible sets coincide asymptotically. This result also provides formal assurance that the influence of the prior will vanish in large samples. However, it's also worth remembering that the Bernstein-von Mises result above applies to a very simple empirical setting: iid data, correctly specified (finite-dimensional) parametric models, and likelihoods and priors that are sufficiently well-behaved. Proving similar results for more complicated models and data settings can be challenging. A few extensions to other parametric settings exist, including misspecified models (Kleijn and Vaart, 2012), non-iid data, parameters on the boundary of the parameter space (Bochkina and Green, 2014), and models based on pseudo-likelihoods and generalized posteriors (Miller, 2019).

Proving Bernstein-von Mises theorems for infinite-dimensional models is challenging (Freedman, 1999). To see why, consider our heuristic argument above based on a Taylor series expansion of the log-likelihood. In a loose sense, we need the prior has to be "locally constant" in order for its influence to wash away with large n. We therefore need to look for regions of the parameter space that are large enough so that posterior concentrates and also small enough so that prior is approximately constant. As parameter space gets larger it becomes harder to satisfy both of these criteria. That said, this continues to be an active area of research with many recent papers providing Bernstein-von Mises results in semiparametric and nonparametric settings (Castillo, 2012; Bickel and Kleijn, 2012; Rousseau, 2016; Ročková, 2020; Ray and Vaart, 2021). A summary of recent developments can also be found in chapter 12 of Ghosal and van der Vaart (2017).

References

Barron, A., Schervish, M. J., and Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *The Annals of Statistics*, 27(2):536–561.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory.* John Wiley & Sons.

Bickel, P. J. and Kleijn, B. J. K. (2012). The semiparametric Bernstein–von Mises theorem. *The Annals of Statistics*, 40(1):206–237.

Bochkina, N. A. and Green, P. J. (2014). The Bernstein–von Mises theorem and nonregular models. *The Annals of Statistics*, 42(5):1850–1878.

Castillo, I. (2012). A semiparametric Bernstein–von Mises theorem for Gaussian process priors. *Probability Theory and Related Fields*, 152(1-2):53–99.

Diaconis, P. and Freedman, D. (1986). On the consistency of bayes estimates. *The Annals of Statistics*, 14(1):1–26.

Doob, J. L. (1949). Application of the theory of martingales. In *Actes du Colloque International Le Calcul des Probabilités et ses applications (Lyon, 28 Juin – 3 Juillet, 1948)*, pages 23–27. Paris CNRS.

Freedman, D. (1999). Wald Lecture: On the Bernstein-von Mises theorem with infinite-dimensional parameters. *The Annals of Statistics*, 27(4):1119–1141.

Freedman, D. A. (1963). On the asymptotic behavior of bayes' estimates in the discrete case. *The Annals of Mathematical Statistics*, 34(4):1386–1403.

Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. (1999a). Consistent semiparametric Bayesian inference about a location parameter. *Journal of Statistical Planning and Inference*, 77(2):181–193.

Ghosal, S., Ghosh, J. K., and Ramamoorthi, R. V. (1999b). Posterior consistency of Dirichlet mixtures in density estimation. *The Annals of Statistics*, 27(1):143–158.

Ghosal, S. and van der Vaart, A. (2017). *Fundamentals of Bayesian Nonparametric Inference.* Cambridge University Press.

Ghosh, J. K. and Ramamoorthi, R. V. (2003). *Bayesian Nonparametrics.* Springer Series in Statistics.

Kleijn, B. and Vaart, A. v. d. (2012). The Bernstein-Von-Mises theorem under misspecification. *Electronic Journal of Statistics*, 6(none):354–381.

Laplace, P.-S. (1809). Mémoire sur les intégrales définies et leur application aux probabilités, et spécialement à la recherche du milieu qu'il faut choisir entre les résultats des observations. *Mémoires presentés à L'Académie des Sciences, Paris.*

Miller, J. W. (2018). A detailed treatment of Doob's theorem. *Working Paper.*

Miller, J. W. (2019). Asymptotic normality, concentration, and coverage of generalized posteriors. *Working Paper.*

Ray, K. and Vaart, A. v. d. (2021). On the Bernstein-von Mises theorem for the Dirichlet process. *Electronic Journal of Statistics*, 15(1).

Ročková, V. (2020). On semi-parametric inference for BART. In *Proceedings of the 37th International Conference on Machine Learning*. Vienna, Austria.

Rousseau, J. (2016). On the Frequentist Properties of Bayesian Nonparametric Methods. *Annual Review of Statistics and Its Application*, 3(1):211–231.

Schervish, M. J. (1995). *Theory of Statistics.* Springer.

Schwartz, L. (1965). On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 4(1):10–26.

van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge University Press.