

Handout 17: Asymptotic behavior of the posterior distribution ^a

Lecturer: Georgios P. Karagiannis

georgios.karagiannis@durham.ac.uk

Aim: We examine properties of the posterior distribution $\pi(d\theta|x_{1:n})$ as the number of observations increases $n \rightarrow \infty$. We show that, as $n \rightarrow \infty$,

- posterior beliefs about $\theta \in \Theta$ would become more and more concentrated around a “true” value θ^* (if exists)
- the unknown parameter $\theta \in \Theta$ has Normal distribution as asymptotic distribution (if $\Theta \subseteq \mathbb{R}^k$, $k \geq 1$).

References:

- Van der Vaart, A. W. (2000, Chapter 10). Asymptotic statistics. Cambridge series in statistical and probabilistic mathematics.
- Ferguson, T. S. (1996, Section 21). A course in large sample theory. Chapman and Hall/CRC.
- DeGroot, M. H. (1970, Chapter 10). Optimal statistical decisions (Vol. 82). John Wiley & Sons.

Web-applets

- https://georgios-stats-1.shinyapps.io/demo_conjugatepriors/
- https://georgios-stats-1.shinyapps.io/demo_conjugatejeffreyslplacepriors/
- https://georgios-stats-1.shinyapps.io/demo_mixturepriors/

^aAuthor: Georgios P. Karagiannis.

What is about?

We consider the Bayesian model $(f(x_{1:n}|\theta), \pi(d\theta))$ as

$$\begin{cases} x_i|\theta & \stackrel{\text{iid}}{\sim} f(d \cdot |\theta), & i = 1, \dots, n \\ \theta & \sim \pi(d \cdot) \end{cases} \quad (1)$$

where a sequence of observables $x_{1:n} = (x_1, \dots, x_n)$ are drawn from the parametric model $f(dx|\theta)$ with unknown parameter $\theta \in \Theta$. We study the behavior of the posterior distribution $\pi(d\theta|x_{1:n})$ with respect to the number of observables n , first when θ is a discrete parameter, and then when it is a continuous one.

All the theorems in this chapter are frequentist in character, namely we study the posterior laws under the assumption that the observables $x_{1:n}$ are a random IID sample from the sampling distribution $f(dx|\theta^*)$ for some fixed, non-random $\theta^* \in \Theta$.

1 Asymptotic consistency: when θ is discrete

In this section we focus in the case that $\theta \in \Theta$ is a discrete parameter, and Θ is a countable space, given the Bayesian model (1). In particular, the theorem below states that, for countable Θ , the posterior distribution function for $\theta \in \Theta$ ultimately degenerates to a step function with a single (unit) step at $\theta = \theta^*$, where θ^* is the real value of the unknown discrete parameter θ .

Theorem 1. (Discrete case) Assume the Bayesian model (1), let $x_{1:n} = (x_1, \dots, x_n)$ be a sequence of observables, $\theta \in \Theta$ be the unknown parameter with prior distribution $\pi(\theta)$, and posterior distribution $\pi(\theta|x_{1:n})$, where Θ is a countable parametric space. Suppose $\theta^* \in \Theta$ is the (only) true value of θ such that $\pi(\theta^*) > 0$, and $-KL(f(\cdot|\theta^*), f(\cdot|\theta)) := \int \log \frac{f(x|\theta)}{f(x|\theta^*)} f(dx|\theta^*) < 0$ for all $\theta \neq \theta^*$. Then

$$\lim_{n \rightarrow \infty} \pi(\theta|x_{1:n}) = \begin{cases} 1 & , \theta = \theta^* \\ 0 & , \theta \neq \theta^* \end{cases}.$$

Proof. It is

$$\pi(\theta|x_{1:n}) = \frac{f(x_{1:n}|\theta)\pi(\theta)}{\sum_{\theta \in \Theta} f(x_{1:n}|\theta)\pi(\theta)} = \frac{\frac{f(x_{1:n}|\theta)}{f(x_{1:n}|\theta^*)}\pi(\theta)}{\sum_{\theta \in \Theta} \frac{f(x_{1:n}|\theta)}{f(x_{1:n}|\theta^*)}\pi(\theta)}.$$

Due to exchangeability of $x_{1:n}$, it is

$$\pi(\theta|x_{1:n}) = \frac{[\prod_{i=1}^n \frac{f(x_i|\theta)}{f(x_i|\theta^*)}]\pi(\theta)}{\sum_{\theta \in \Theta} [\prod_{i=1}^n \frac{f(x_i|\theta)}{f(x_i|\theta^*)}]\pi(\theta)} = \frac{\exp(\overbrace{\sum_{i=1}^n \log \frac{f(x_i|\theta)}{f(x_i|\theta^*)}}^{=S_n(\theta)})\pi(\theta)}{\sum_{\theta \in \Theta} \exp(\sum_{i=1}^n \log \frac{f(x_i|\theta)}{f(x_i|\theta^*)})\pi(\theta)} = \frac{\exp(S_n(\theta))\pi(\theta)}{\sum_{\theta \in \Theta} \exp(S_n(\theta))\pi(\theta)}$$

Now about $S_n(\theta) = \sum_{i=1}^n \log \frac{f(x_i|\theta)}{f(x_i|\theta^*)}$. From the SLLN, as $n \rightarrow \infty$, it is

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \log \frac{f(x_i|\theta)}{f(x_i|\theta^*)} = E^{f(dx|\theta^*)}(\log \frac{f(x|\theta)}{f(x|\theta^*)}), \quad \text{a.s.} \quad (2)$$

By using Jensen's inequality and the fact that \log is concave, it is

$$\begin{aligned} E^{f(dx|\theta^*)}(\log \frac{f(x|\theta)}{f(x|\theta^*)}) &\leq \log E^{f(dx|\theta^*)}(\frac{f(x|\theta)}{f(x|\theta^*)}) = \log(1) = 0 \\ \implies E^{f(dx|\theta^*)}(\log \frac{f(x|\theta)}{f(x|\theta^*)}) &\leq 0 \end{aligned} \quad (3)$$

In the (3), the equality holds for $\theta = \theta^*$ a.s., and the inequality holds for $\theta \neq \theta^*$ a.s., since Θ is a countable space and $\theta^* \in \Theta$, θ^* is "distinguishable" from the others, according to Theorem 15. Notice that, for any $\theta \neq \theta^*$, (2) and (3) imply that

$$\lim_{n \rightarrow \infty} \frac{1}{n} S_n(\theta) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \log \frac{f(x_i|\theta)}{f(x_i|\theta^*)} < 0, \quad \text{a.s.}$$

which implies that

$$\lim_{n \rightarrow \infty} S_n(\theta) = -\infty, \quad \text{as}$$

Therefore, □

• for any $\theta \neq \theta^*$, it is

$$\lim_{n \rightarrow \infty} \pi(\theta|x_{1:n}) = \lim_{n \rightarrow \infty} \frac{\exp(S_n(\theta))\pi(\theta)}{\sum_{\theta \in \Theta} \exp(S_n(\theta))\pi(\theta)} = 0, \quad \text{a.s.}$$

– for $\theta = \theta^*$, it is

$$\lim_{n \rightarrow \infty} \pi(\theta^*|x_{1:n}) = 1 - \underbrace{\sum_{\theta \neq \theta^*} \lim_{n \rightarrow \infty} \pi(\theta|x_{1:n})}_{=0, \text{ for } \theta \neq \theta^*} = 1, \quad \text{a.s.}$$

Remark 2. It can be shown that if $\theta^* \notin \Theta$, the posterior degenerates onto the value in Θ which gives the parametric model closest θ^* .

Remark 3. One important condition we considered was that the real parameter value θ^* is unique. If there was another θ^{**} such that $f(x|\theta^{**}) = f(x|\theta^*)$, we would observe IID data when θ equaled θ^* or θ^{**} , and hence the data could not discriminate between the two values.

Remark 4. If θ is a continuous parameter, then $\pi(\theta|x_{1:n})$ is always zero for any finite sample, and so the above theorem 1 apply.

2 Asymptotic consistency and normality: when θ is continuous

In this section, we focus in the case that $\theta \in \Theta$ may be a continuous parameter, $\Theta \subset \mathbb{R}^k$ is compact with $k \geq 1$, given the Bayesian model (1). In particular, under specific conditions, we prove that as $n \rightarrow \infty$:

1. the posterior PDF of θ becomes more and more concentrated above an area around the true value θ^*
2. the limiting posterior distribution of θ is a Normal distribution with specific parameters, and remarkably
3. these conclusions do not depend on the choice of the prior distribution provided that $\pi(\theta^*) > 0$.

We recall asymptotic results of MLE in Appendix B.

Consider the Bayesian model $(f(x_{1:n}|\theta), \pi(d\theta))$ in (1), with $\theta \in \Theta$, where Θ is an open subset of \mathbb{R}^k . Assume a prior distribution $\pi(d\theta)$ with a PDF $\pi(\theta)$ where $\pi(\theta)$ continuous in Θ . To easy notation, we denote the likelihood as $L_n(\theta) := f(x_{1:n}|\theta) = \prod_{i=1}^n f(x_i|\theta)$. So the posterior PDF of θ is

$$\pi(\theta|x_{1:n}) = \frac{L_n(\theta)\pi(\theta)}{\int L_n(\theta)\pi(\theta)d\theta} = \frac{\prod_{i=1}^n f(x_i|\theta)\pi(\theta)}{\int \prod_{i=1}^n f(x_i|\theta)\pi(\theta)d\theta}$$

We present the Bernstein-von Mises theorem which (more-or-less) states that the posterior PDF $\pi(\theta|x_{1:n})$ is close to a normal density $N(\theta|\hat{\theta}_n, \frac{1}{n}\mathcal{J}(\theta^*)^{-1})$ centered at $\hat{\theta}_n$ (the MLE of (16)), with variance $\frac{1}{n}\mathcal{J}(\theta^*)^{-1}$ when θ^* is the true value, when the size of the number of the observables n is really big. Here, $\mathcal{J}(\theta)$ is the Fisher information, where

$$\mathcal{J}(\theta) = E^{f(\cdot|\theta)}((\nabla_{\theta} \log f(x|\theta))^T (\nabla_{\theta} \log f(x|\theta))) = -E^{f(\cdot|\theta)}(\nabla_{\theta}^2 \log f(x|\theta))$$

The following version of the theorem is due to Le Cam (1953). It equivalently states that the posterior PDF of (the linear transformation) $\vartheta = \sqrt{n}(\theta - \hat{\theta}_n)$ given the data

$$\pi(\vartheta|x_{1:n}) = \frac{L_n(\vartheta \frac{1}{\sqrt{n}} + \hat{\theta}_n)\pi(\vartheta \frac{1}{\sqrt{n}} + \hat{\theta}_n)}{\int L_n(\vartheta \frac{1}{\sqrt{n}} + \hat{\theta}_n)\pi(\vartheta \frac{1}{\sqrt{n}} + \hat{\theta}_n)d\vartheta}$$

approaches the PDF of $N(0, \mathcal{J}(\theta^*)^{-1})$ and $n \rightarrow \infty$. Remarkably (!!!), this limiting posterior distribution is independent of the prior distribution $\pi(\theta)$.

Theorem 5. (Bernstein-von Mises) Let x_1, x_2, \dots be IID random variables drawn from a sample distribution with density $f(x|\theta)$, $\theta \in \Theta$, and let $\theta^* \in \Theta$ denote the true value of θ . Let $L_n(\theta) = f(x_{1:n}|\theta)$ denote the likelihood. Assume that the prior density $\pi(\theta)$ is continuous and $\pi(\theta) > 0$ for all $\theta \in \Theta$. Under the following assumptions:

d1 Θ is an open subset of \mathbb{R}^k

d2 second partial derivatives of $f(x|\theta)$ with respect to θ exist and are continuous for all x , and may be passed under the integral operator in $\int f(x|\theta)d\theta$

d3 there is a function $K(x)$ such that $E^{f(x|\theta_0)}(K(x)) < \infty$ and each component of $\nabla_{\theta}^2 \log(f(x|\theta))$ is bounded in absolute value by $K(x)$ uniformly in some neighborhood of θ^*

76 **d4** $\mathcal{J}(\theta_0) = -\mathbb{E}^{f(\text{d}x|\theta_0)}(\nabla_{\theta}^2 \log(f(x|\theta)))$ is positive definite

77 **d5** (identifiability) $f(x|\theta) = f(x|\theta^*)$ a.s. then $\theta = \theta^*$

78 It is

$$79 \frac{L_n(\hat{\theta}_n + \vartheta \frac{1}{\sqrt{n}})}{L_n(\hat{\theta}_n)} \pi(\hat{\theta}_n + \vartheta \frac{1}{\sqrt{n}}) \xrightarrow{\text{a.s.}} \exp(-\frac{1}{2} \vartheta^{\top} \mathcal{J}(\theta_0) \vartheta) \pi(\theta^*), \quad (4)$$

80 where $\hat{\theta}_n$ is the strongly consistent sequence of roots of the likelihood equation (16) of Theorem 17. If, additionally,

$$81 \int_{\Theta} \frac{L_n(\hat{\theta}_n + \vartheta \frac{1}{\sqrt{n}})}{L_n(\hat{\theta}_n)} \pi(\hat{\theta}_n + \vartheta \frac{1}{\sqrt{n}}) d\vartheta \xrightarrow{\text{a.s.}} \int_{\Theta} \exp(-\frac{1}{2} \vartheta^{\top} \mathcal{J}(\theta^*) \vartheta) \pi(\theta^*) d\vartheta \quad (5)$$

82 then

$$83 \int_{\Theta} |\pi(\vartheta|x_{1:n}) - N(\vartheta|0, \mathcal{J}(\theta^*)^{-1})| d\theta \xrightarrow{\text{a.s.}} 0. \quad (6)$$

84 *Proof.* We prove: fist the existence of the the MLE, then (4), and finally (6).

85 • Existence of consistent roots

86 I gonna use Theorem 17 to prove that there exists a consistent sequence $\hat{\theta}_n$ of roots of (16), and hence I need
87 to show that its conditions are satisfied. Let $S_{\rho} = \{\theta : |\theta - \theta^*| \leq \rho\}$, with $\rho > 0$, be a neighborhood of θ^*
88 on which (d3) is satisfied. So for $\Theta = S_{\rho}$ (in Theorem 17) I have:

- 89 – Conditions (c1), (c2), (c5), of that theorem are automatic!
- 90 – Condition (c4) follows from continuity of $f(x|\theta)$ at θ . !
- 91 – Condition (c3), ok.... By Taylor's theorem, I expand $U(x, \theta) = \log(f(x|\theta)) - \log(f(x|\theta^*))$ around θ^*
92 as

$$93 U(x, \theta) = U(x, \theta^*) + \nabla_{\theta} \log(f(x|\theta^*))(\theta - \theta^*) \\ 94 + (\theta - \theta^*) \int_0^1 \int_0^1 v \nabla_{\theta}^2 \log(f(x|\theta_0 + uv(\theta - \theta^*))) du dv (\theta - \theta^*)$$

95 So because $U(x, \theta^*) = 0$, $\nabla_{\theta} \log(f(x|\theta^*))$ is integrable, and the components of $\nabla_{\theta}^2 \log(f(x|\theta))$ are
96 bounded by $K(x)$ uniformly on S_{ρ} , we get that $U(x, \theta)$ is bounbed on S_{ρ} . So (c3) holds. So

97 • Asymptotic Normality

98 Let

$$99 \ell_n(\theta) = \log(L_n(\theta)); \quad \dot{\ell}_n(\theta) = \nabla_{\theta} \log(L_n(\theta)); \quad \ddot{\ell}_n(\theta) = \nabla_{\theta}^2 \log(L_n(\theta))$$

100 By Taylor's Theorem 14, we expand $\ell_n(\theta)$ around $\hat{\theta}_n$ as

$$101 \ell_n(\theta) = \ell_n(\hat{\theta}_n) + \dot{\ell}_n(\hat{\theta}_n)(\theta - \hat{\theta}_n) + (\theta - \hat{\theta}_n)^{\top} I_n(\theta)(\theta - \hat{\theta}_n)$$

102 where

$$103 I_n(\theta) = -\frac{1}{n} \int_0^1 \int_0^1 v \ddot{\ell}_n(\hat{\theta}_n + uv(\theta - \hat{\theta}_n)) du dv \quad (7)$$

Because, it is $\dot{\ell}_n(\hat{\theta}_n) = 0$ a.s., we get:

$$\begin{aligned}\ell_n(\theta) &= \ell_n(\hat{\theta}_n) + (\theta - \hat{\theta}_n)^\top I_n(\theta)(\theta - \hat{\theta}_n) \iff \\ \frac{L_n(\theta)}{L_n(\hat{\theta}_n)} &= \exp(-(\theta - \hat{\theta}_n)^\top I_n(\theta)(\theta - \hat{\theta}_n)), \quad \text{a.s.}\end{aligned}$$

Let's work on the asymptotics of (7); it is:

$$\begin{aligned}\frac{1}{n} \ddot{\ell}_n(\theta) &= \frac{1}{n} \nabla_\theta^2 \log(L_n(\theta)) = \frac{1}{n} \nabla_\theta^2 \log\left(\prod_{i=1}^n f(x_i|\theta)\right) = \frac{1}{n} \sum_{i=1}^n \nabla_\theta^2 \log(f(x_i|\theta)) \\ &\xrightarrow{\text{a.s.}} \mathbb{E}^{f(x|\theta_0)}(\nabla_\theta^2 \log(f(x|\theta)))\end{aligned}\tag{8}$$

as $n \rightarrow \infty$ by SLLN. Also, it is

$$\mathbb{E}^{f(x|\theta_0)}(\nabla_\theta^2 \log(f(x|\theta))) = -\mathcal{J}(\theta_0)\tag{9}$$

by Lemma ?? . Hence, from (8) and (9), I get

$$\frac{1}{n} \ddot{\ell}_n(\theta) \xrightarrow{\text{a.s.}} -\mathcal{J}(\theta_0)\tag{10}$$

Therefore,

$$I_n(\hat{\theta}_n + \vartheta \frac{1}{\sqrt{n}}) = -\frac{1}{n} \int_0^1 \int_0^1 v \ddot{\ell}_n(\hat{\theta}_n + uv(\theta - \hat{\theta}_n)) du dv \xrightarrow{\text{a.s.}} \frac{1}{2} \mathcal{J}(\theta^*)\tag{11}$$

because of (10) and because of $\hat{\theta}_n \xrightarrow{\text{a.s.}} \theta^*$ from Theorem 17.

So back to what we wish to prove, and putting all these together, it is

$$\begin{aligned}\frac{L_n(\hat{\theta}_n + \vartheta \frac{1}{\sqrt{n}})}{L_n(\hat{\theta}_n)} \pi(\hat{\theta}_n + \vartheta \frac{1}{\sqrt{n}}) &= \exp(-\vartheta^\top I_n(\hat{\theta}_n + \vartheta \frac{1}{\sqrt{n}}) \vartheta) \pi(\hat{\theta}_n + \vartheta \frac{1}{\sqrt{n}}) \\ &\xrightarrow{\text{a.s.}} \exp(-\frac{1}{2} \vartheta^\top \mathcal{J}(\theta^*) \vartheta) \pi(\theta^*)\end{aligned}$$

because of (11) and $\hat{\theta}_n \xrightarrow{\text{a.s.}} \theta^*$.

Now, about the second part of the proof. If (5) then by dividing (4) and (5), I get

$$\pi(\vartheta|x_{1:n}) \xrightarrow{\text{a.s.}} \mathcal{N}(\vartheta|0, \mathcal{J}(\theta^*)^{-1})$$

for all $\theta \in \Theta$. Hence By Scheffe's Theorem 12 we get (6).

□

Remark 6. Note that Bernstein-von Mises Theorem 5 and Theorem A imply that the posterior distribution of $\vartheta = \sqrt{n}(\theta - \hat{\theta})$ given the data converges to the Normal distribution $\mathcal{N}(0, \mathcal{J}(\theta_0)^{-1})$ in Total Variation Norm, namely

$$\sup_{A \subset \Theta} |\pi(\vartheta \in A|x_{1:n}) - \mathcal{N}(\vartheta \in A|0, \mathcal{J}(\theta^*)^{-1})| dx \rightarrow 0, \quad \text{as } n \rightarrow \infty$$

Remark 7. Bernstein-von Mises Theorem 5 says that the posterior PDFs of $\sqrt{n} \mathcal{J}(\theta^*)^{1/2}(\theta - \hat{\theta}_n)$ converges to the PDF of Normal $\mathcal{N}(0, I_k)$. Hence, this imply the CDF of $\sqrt{n} \mathcal{J}(\theta^*)^{1/2}(\theta - \hat{\theta}_n)$ converges to the PDF of Normal $\mathcal{N}(0, I_k)$ as well. Namely

$$\sqrt{n} \mathcal{J}(\theta^*)^{1/2}(\theta - \hat{\theta}_n) \xrightarrow{D} z \quad \text{where } z \sim \mathcal{N}(0, I_k) \quad (\text{convergence in distribution}),$$

and $\mathcal{J}(\cdot) = [\mathcal{J}(\cdot)^{1/2}] [\mathcal{J}(\cdot)^{1/2}]^\top$.

Corollary. *If the conditions of Bernstein-von Mises Theorem 5 hold, and if $\mathcal{J}(\theta)$ is continuous at Θ , then*

$$\sqrt{n}\mathcal{J}(\hat{\theta}_n)^{-1/2}(\theta - \hat{\theta}_n) \xrightarrow{D} z, \quad \text{where } z \sim N(0, I_k) \quad (12)$$

this is the result was stated in Stat Concepts II notes (Term 2 2017).

Proof. From Bernstein-von Mises Theorem implies $\sqrt{n}(\theta - \hat{\theta}_n) \xrightarrow{D} N(0, \mathcal{J}(\theta^*)^{-1})$ or equiv.

$$Y_n = \sqrt{n}\mathcal{J}(\theta^*)^{1/2}(\theta - \hat{\theta}_n) \xrightarrow{D} Z, \quad (13)$$

with $Z \sim N(0, I_k)$. From Theorem 17 I get $\hat{\theta}_n \rightarrow \theta^*$ a.s.. Due to continuity of $\mathcal{J}(\theta)$, it is

$$X_n = \mathcal{J}(\hat{\theta}_n)^{1/2}\mathcal{J}(\theta^*)^{-1/2} \xrightarrow{\text{a.s.}} I_k \quad (14)$$

According to Slutsky's theorem¹ by multiplying (13),(14), I get $X_n Y_n \xrightarrow{D} Z$, i.e., $\sqrt{n}\mathcal{J}(\hat{\theta}_n)^{1/2}(\theta - \hat{\theta}_n) \xrightarrow{D} N(0, I_k)$. □

Asymptotic efficiency of Bayes Estimates

Consider the squared error loss $\ell(\theta, \delta) = (\theta - \delta)^\top (\theta - \delta)$ which implies the posterior expectation $\delta^\pi = E^{\pi(d|x_{1:n})}(\theta)$ as Bayes point estimator. Given that, we can interchange the limit and the expectation operator of $\vartheta = \sqrt{n}(\theta - \hat{\theta}_n)$, we get $\sqrt{n}(\delta^\pi - \hat{\theta}_n) \rightarrow 0$ meaning that δ^π and $\hat{\theta}_n$ are asymptotically equivalent; i.e. $\delta^\pi = \hat{\theta}_n$ a.s.. Hence

$$\sqrt{n}(\delta^\pi - \theta^*) = \sqrt{n}(\delta^\pi - \hat{\theta}_n) + \sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{D} N(0, \mathcal{J}(\theta^*)^{-1})$$

which means that the Bayes estimator is asymptotically efficient.

¹Sluky's theorem: If $Y_n \xrightarrow{D} Z$ and $X_n \xrightarrow{\text{a.s.}} c$, where $c \in \mathbb{R}^k$ is a constant, then $X_n Y_n \xrightarrow{D} cZ$

3 Implementation

Nowadays, Bayesian asymptotics can be mainly be used for theoretical purposes such as investigating the behavior of Bayesian procedures-methods-algorithms mainly addressing big dataset problems.

Below, we demonstrate a toy example for pedagogical purposes.

Example 8. Consider a Bayesian model

$$\begin{cases} x_i | \theta & \stackrel{\text{iid}}{\sim} \text{Bn}(\theta), & i = 1, \dots, n \\ \theta & \sim \text{Be}(a, b) \end{cases}$$

where $a > 0$, $b > 0$, and $n > 2$.

1. Find the asymptotic posterior distribution of θ as $n \rightarrow \infty$, by using Bernstein-von Mises Theorem.
2. What is the Bayes estimators under the square loss and under the 0-1 loss. How do they behave as $n \rightarrow \infty$.

Solution.

1. I will find the MLE by computing the roots of the equation (16). The likelihood is $f(x_{1:n}|\theta) = \prod_{i=1}^n \text{Bn}(x_i|\theta)$. Then

$$\frac{d}{d\theta} \log f(x_{1:n}|\theta) = \frac{d}{d\theta} \sum_{i=1}^n \log(\text{Bn}(x_i|\theta)) = \frac{n\theta - \sum_{i=1}^n x_i}{\theta(1-\theta)} \Rightarrow$$

$$\frac{d}{d\theta} \log f(x_{1:n}|\theta)|_{\theta=\hat{\theta}_n} = 0 \Rightarrow \hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n x_i$$

Now, I will find the Fisher Information; it is

$$\frac{d^2}{d\theta^2} \log(f(x|\theta)) = \frac{d^2}{d\theta^2} \log(\text{Bn}(x|\theta)) = -\frac{x}{\theta^2} - \frac{1-x}{(1-\theta)^2} \Rightarrow$$

$$\mathcal{J}(\theta) = -\mathbb{E}^{\text{Bn}(\theta)} \left(-\frac{x}{\theta^2} - \frac{1-x}{(1-\theta)^2} \right) = \frac{1}{\theta(1-\theta)} \quad (15)$$

So according to Bernstein-von Mises Theorem 5, it is

$$\pi(\theta|x_{1:n}) \xrightarrow{\text{a.s.}} \text{N}(\theta | \underbrace{\frac{1}{n} \sum_{i=1}^n x_i}_{=\hat{\theta}_n}, \underbrace{\frac{1}{n} \theta^* (1-\theta^*)}_{=\frac{1}{n} \mathcal{J}(\theta^*)^{-1}})$$

2. The exact posterior distr. is $\text{Be}(a + n\bar{x}, b + n - n\bar{x})$. The squared loss and the 0 – 1 loss imply the posterior mean $\delta_1(x_{1:n}) = \frac{n\bar{x}+a}{n+a+b}$ and posterior mode $\delta_2(x_{1:n}) = \frac{n\bar{x}+a-1}{n+a+b-2}$ as Bayes estimators correspondingly. Both converge to the MLE $\hat{\theta}_n = \bar{x}$ since $\lim_{n \rightarrow \infty} \delta_1(x_{1:n}) = \bar{x}$ and $\lim_{n \rightarrow \infty} \delta_2(x_{1:n}) = \bar{x}$.

Appendix

A An inventory of definitions

The following are definitions and statements used in proofs in Sections 1 and 2.

Definition 9. (Types of converge) Assume a probability triplet $\{\Omega, \mathcal{F}, P\}$, and a sequence of random quantities $\{x_n; n = 1, 2, \dots\}$, such that $x_n : \Omega \rightarrow \mathbb{R}^d$, $d > 0$. Then

- $\{x_n\}$ converges almost surely to a random quantify x if and only if

$$P(\lim_{n \rightarrow \infty} x_n = x) = 1.$$

It is demoted as $x_n \xrightarrow{\text{a.s.}} x$.

- $\{x_n\}$ converges in distribution to a random quantify x if and only if for all $\epsilon > 0$,

$$\lim_{n \rightarrow \infty} F_{x_n}(t) = F_x(t)$$

for all continuity points t of F in \mathbb{R} , where $F_{x_n}(t) = P(x_n \leq t)$ and $F_x(t) = P(x \leq t)$ are CDFs of x_n and x . It is demoted as $x_n \xrightarrow{D} x$.

- $\{x_n\}$ converges in total variation a random quantity x if and only if

$$\lim_{n \rightarrow \infty} \sup_{\forall B \subset \Theta} |P(x_n \in B) - P(x \in B)| = 0,$$

It is demoted as $x_n \xrightarrow{\text{T.V.}} x$.

Definition 10. (Upper semicontinuous) A real-valued function, $f(\theta)$, defined on Θ is said to be upper semicontinuous (u.s.c.) on Θ , if for all $\theta \in \Theta$ and for any sequence θ_n in Θ such that $\theta_n \rightarrow \theta$, we have

$$\lim_{n \rightarrow \infty} \sup f(\theta_n) \leq f(\theta)$$

Proposition 11. If $\pi_n(\cdot)$ and $\pi(\cdot)$ are the PDFs of x_n and x correspondingly, then

$$\sup_{\forall B \subset \Theta} |P(x_n \in B) - P(x \in B)| = \int \frac{1}{2} |\pi_n(t) - \pi(t)| dt$$

Theorem 12. (Scheffe convergence theorem²) If $f_n(\cdot)$ and $g(\cdot)$ are density functions such that for all $x \in \mathcal{X}$ $\lim_{n \rightarrow \infty} f_n(x) = g(x)$, then

$$\lim_{n \rightarrow \infty} \int_{\mathcal{X}} |f_n(x) - g(x)| dx = 0$$

(that is a point-wise convergence of densities)

Theorem If random variables x_n has density $f_n(x)$ and random variable x has density $g(x)$, and if $\lim_{n \rightarrow \infty} \int_{\mathcal{X}} |f_n(x) - g(x)| dx = 0$ then

$$\sup_{\forall A \subset \mathcal{X}} |P(x_n \in A) - P(x \in A)| \rightarrow 0, \quad \text{as } n \rightarrow \infty$$

that is called convergence in Total Variation.

Theorem 13. (A Strong law of large numbers (SLLN)) Let $\{x_i\}_{i=1}^n$ be a sequence of IID random quantities, with $E(x_i) = \mu < \infty$, and $\bar{x}_n = \frac{1}{n} \sum_{i=1}^n x_i$ then $\bar{x}_n \xrightarrow{\text{a.s.}} \mu$.

²This is not the original version, but it is what we need

Theorem 14. (Taylor's theorem) If $f : \mathbb{R}^d \rightarrow \mathbb{R}$, and if $\nabla^2 f(x) = \nabla(\nabla f(x))^\top$ is continuous in the ball $\{x \in \mathcal{X} : |x - x_0| < r\}$, then for $|t| < r$, it is

$$f(x_0 + t) = f(x_0) + \nabla f(x_0)t + t^\top \cdot \int_0^1 \int_0^1 u \nabla^2 f(x_0 + uvt) du dv \cdot t$$

Lemma 15. (Shannon-Kolmogorov Information Inequality) Let $f_0(x)$ and $f_1(x)$ be densities with respect to Lebesgue measure dx . Then

$$KL(f_0, f_1) = E^{f_0(dx)}(\log \frac{f_0(x)}{f_1(x)}) = \int_{\mathcal{X}} \log \frac{f_0(x)}{f_1(x)} f_0(x) dx \geq 0,$$

with equality if and only if $f_1(x) = f_0(x)$ a.s.

Lemma 16. (Passing the derivative under the integral operator) If $(\partial/\partial\theta)g(x, \theta)$ exists and is continuous in θ for all x and all θ in an open interval \mathcal{S} and if $|(\partial/\partial\theta)g(x, \theta)| \leq K(x)$ on \mathcal{S} where $\int K(x)dx < \infty$ and if $\int g(x, \theta)dx$ exists on \mathcal{S} , then

$$\frac{d}{d\theta} \int g(x, \theta) dx = \int \frac{d}{d\theta} g(x, \theta) dx$$

B Strong consistency of Maximum Likelihood Estimates

In frequentist statistics, given that $\nabla_\theta f(x|\theta)$ exists, one may seek to find the MLE $\hat{\theta}_n$ as the solution of the likelihood equation:

$$\hat{\theta}_n : \nabla_\theta \log f(x_{1:n}|\theta)|_{\theta=\hat{\theta}_n} = \sum_{i=1}^n \nabla_\theta \log f(x_i|\theta)|_{\theta=\hat{\theta}_n} = 0 \quad (16)$$

The following theorem states (more or less) that the MLE $\hat{\theta}_n$ in (16) is consistent.

Theorem 17. (Strong consistency of MLE) Let x_1, x_2, \dots be IID random variables with density $f(x|\theta)$ (with respect to measure dx), $\theta \in \Theta$, and let θ^* denote the true value of θ . If the following conditions are satisfied:

c1 Θ is a closed and bounded set in \mathbb{R}^k

c2 $f(x|\theta)$ is u.s.c. in θ for all $x \in \mathcal{X}$

c3 there is a function $K(x)$ such that $E^{f(x|\theta^*)}(|K(x)|) < \infty$ and

$$\log(f(x|\theta)) - \log(f(x|\theta^*)) \leq K(x), \quad \forall x, \forall \theta$$

c4 for all $\theta \in \Theta$ and sufficiency small $\rho > 0$, $\sup_{|\theta' - \theta| < \rho} f(x|\theta')$ is measurable in x

c5 (identifiability) $f(x|\theta) = f(x|\theta^*)$ a.s. then $\theta = \theta^*$

then, for any sequence of maximum-likelihood estimates $\hat{\theta}_n$ of θ , it is

$$\hat{\theta}_n \xrightarrow{a.s.} \theta^* \quad (17)$$

The following theorem states (more or less) that the MLE is asymptotically normal.

Theorem 18. (Cramer) Let x_1, x_2, \dots be IID random variables density $f(x|\theta)$ (with respect to some distribution $F(x|\theta)$), $\theta \in \Theta$, and let θ^* denote the true value of θ . If the conditions³ (d1)-(d5) stated in Theorem 5 (check in the next Theorem) are satisfied, then there exists a strongly consistent sequence $\hat{\theta}_n$ of roots of the likelihood equation (16) such that

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{D} N(0, \mathcal{I}(\theta^*)^{-1})$$

³The conditions of the theorem are the (d1)-(d5) in Theorem 5.