Lecture-20: Bayesian lower bounds

1 Bayesian HCR and CR lower bounds

The drawback of the HCR and CR lower bounds is that they are confined to unbiased estimators. For the minimax settings, there is no sound reason to restrict to unbiased estimators. In fact, it is often wise to trade bias with variance in order to achieve a smaller overall risk.

Next we discuss a lower bound, known as the Bayesian Cramér-Rao (BCR) lower bound or the van Trees inequality, for a Bayesian setting that applies to all estimators. To apply to the minimax setting, one just needs to choose an appropriate prior. Here we continue the previous line of thinking and derive it from the data processing argument.

Exercise 1.1 (Chain rule for χ^2-divergence). Show that for any pair of measures $P_{X,Y}$ and $Q_{X,Y}$ we have

$$\chi^2(P_{X,Y}\|Q_{X,Y}) = \chi^2(P_X\|Q_X) + \mathbb{E}_{X\sim Q_X}\left[\left(\frac{dP_X}{dQ_X}\right)^2 \chi^2(P_{Y|X}\|Q_{Y|X})\right], \tag{1}$$

regardless of the versions of conditional distributions $P_{Y|X}$ and $Q_{Y|X}$ one chooses.

Exercise 1.2 (Data processing inequality for f-divergence). For any Markov chain $X \to Y \to Z$, a pair of measures $P_{X,Y,Z}$ and $Q_{X,Y,Z}$ with common Markov kernel $P_{Z|Y} = Q_{Z|Y}$, a convex map $f : (0,\infty) \to \mathbb{R}_+$, and arbitrary function $g : \mathcal{X} \times \mathcal{Z} \to \mathbb{R}$, we have

$$D_f(P_{X,Y}\|Q_{X,Y}) \geqslant D_f(P_{X,Z}\|Q_{X,Z}) \geqslant D_f(P_{g(X,Z)}\|Q_{g(X,Z)}). \tag{2}$$

Definition 1.3 (Push forward operator). For any $\delta > 0$, we define a push forward operator $T_\delta : \mathcal{M}(\mathbb{R}) \to \mathcal{M}(\mathbb{R})$ that applies δ shift to measurable sets. Specifically, $T_\delta\mu \in \mathcal{M}(\mathcal{X})$ for any measure $\mu \in \mathcal{M}(\mathcal{X})$, and is defined as $(T_\delta\mu)(-\infty, x] \triangleq \mu(-\infty, x]$ for any $x \in \mathbb{R}$.

Theorem 1.4 (Bayesian HCR lower bound). *Consider statistical decision theory simple setting for $\Theta \triangleq \mathbb{R}$ with statistical model $\mathcal{P}(\Theta)$ such that for any $P_\theta \in \mathcal{P}(\Theta)$ there exists a relative density $p_\theta \in \mathcal{M}(\mathcal{X})$ with respect to a dominant measure $\mu \in \mathcal{M}(\mathcal{X})$. Further, we assume a prior $\pi \in \mathcal{M}(\Theta)$ that admits a relative density with respect to Lebesgue measure, and two distributions $P, Q \in \mathcal{M}(\Theta \times \mathcal{X})$ such that $dQ_{\theta,X} \triangleq d\pi(\theta)dP_\theta(X)$ and $dP_{\theta,X} \triangleq d(T_\delta\pi)(\theta)dP_{\theta-\delta}(X)$. Then, the Bayes risk satisfies the* Bayesian HCR lower bound

$$R^*_\pi \triangleq \inf_{\hat\theta} \mathbb{E}_\pi(\hat\theta - \theta)^2 \geqslant \sup_{\delta\neq 0} \frac{\delta^2}{\chi^2(P_{\theta,X}\|Q_{\theta,X})}.$$

Proof. We observe that for measures P, Q, their respective relative densities p, q exist with respect to product measure of Lebesgue measure on Θ and dominant measure $\mu \in \mathcal{M}(\mathcal{X})$ such that for all (θ, x),

$$q(\theta, x) \triangleq \pi'(\theta)p_\theta(x), \qquad\qquad p(\theta, x) \triangleq \pi'(\theta - \delta)p_\theta(x).$$

Consider an estimator $\hat\theta(X, U)$ for observation X and external randomness U, then we observe that $\theta \to X \to \hat\theta$ is a Markov chain, and for any joint distribution (θ, X), the Markov kernel $P_{\theta|X}$ is common. Applying data processing inequality from Exercise 1.2 and variational representation of χ^2-divergence from Exercise 1.1, we obtain

$$\chi^2(P_{\theta,X}\|Q_{\theta,X}) \geqslant \chi^2(P_{\theta,\hat\theta}\|Q_{\theta,\hat\theta}) \geqslant \chi^2(P_{\theta-\hat\theta}\|Q_{\theta-\hat\theta}) \geqslant \frac{(\mathbb{E}_P[\theta - \hat\theta] - \mathbb{E}_Q[\theta - \hat\theta])^2}{\mathrm{Var}_Q(\hat\theta - \theta)}.$$

We observe that $Q_X(x) = \int_\Theta d\pi(\theta) P_\theta(x)$ and $P_X(x) = \int_\Theta d\pi(\theta-\delta) P_{\theta-\delta}(x)$. By substitution of variables, we observe that $P_X = Q_X$ and thus $\mathbb{E}_P\hat{\theta} = \mathbb{E}_Q\hat{\theta}$. On the other hand, $\mathbb{E}_P\theta = \mathbb{E}_Q\theta + \delta$. Furthermore, $\mathbb{E}_\pi(\hat{\theta}-\theta)^2 \geqslant \mathrm{Var}_Q(\hat{\theta}-\theta)$ with equality iff $\mathbb{E}_\pi\hat{\theta} = \mathbb{E}_\pi\theta$. Since this applies to any estimator, the result follows. □

Definition 1.5 (Fisher information). For any measure $\pi \in \mathcal{M}(\mathbb{R})$ such that $\pi(x) \triangleq \pi(-\infty, x]$ for all $x \in \mathbb{R}$, and the relative density $\pi'(x) \triangleq \frac{d\pi(x)}{dx}$ with respect to Lebesgue measure exists, we define its *Fisher information* as

$$J(\pi') \triangleq \mathbb{E}_{X\sim\pi}\Big(\frac{d}{dx}\ln\pi'(X)\Big)^2 = \int_{\mathbb{R}} dx \frac{(\pi''(x))^2}{\pi'(x)}.$$

Corollary 1.6 (Bayesian CR lower bound). *Under the conditions of Theorem 1.4 and suitable regularity conditions for the local expansion of χ^2-divergence such that $\chi^2(T_\delta\pi\|\pi) = (J(\pi)+o(1))\delta^2$ and $\chi^2(P_{\theta-\delta}\|P_\theta) = (J_F(\theta)+o(1))\delta^2$, the Bayes risk satisfies the* Bayesian CR lower bound

$$R^*_\pi \geqslant \frac{1}{J(\pi) + \mathbb{E}_{\theta\sim\pi} J_F(\theta)}.$$

Proof. We can lower bound the supremum in Theorem 1.4 by evaluating the small-δ limit. Recognizing that $P_\theta = T_\delta\pi, Q_\theta = \pi$ and $P_{X|\theta} = P_{\theta-\delta}, Q_{X|\theta} = P_\theta$, applying the chain rule for the χ^2-divergence in Exercise 1.1, and applying the local expansion of χ^2-divergence we obtain the result. □

Example 1.7 (GLM). Consider an *i.i.d.* observation sample $X : \Omega \to \mathcal{X}^m$ under GLM with common Gaussian distribution $\mathcal{N}(\theta,1)$ and consider the prior $\theta \sim \pi \triangleq \mathcal{N}(0,s)$. To apply the Bayesian HCR bound, we note that $\bar{X} \triangleq \frac{1}{m}\sum_{i=1}^m X_i$ is a sufficient statistic for X, and apply the chain rule to obtain

$$\chi^2(P_{\theta,X}\|Q_{\theta,X}) = \chi^2(P_{\theta,\bar{X}}\|Q_{\theta,\bar{X}}) = \chi^2(P_\theta\|Q_\theta) + \mathbb{E}_Q\left[\chi^2(P_{\bar{X}|\theta}\|Q_{\bar{X}|\theta})\Big(\frac{dP_\theta}{dQ_\theta}\Big)^2\right].$$

From the definition of P and Q, we obtain that $Q_\theta = \mathcal{N}(0,s), Q_{\bar{X}|\theta} = \mathcal{N}(\theta, \frac{1}{m})$, and $P_\theta = \mathcal{N}(\delta,s), P_{\bar{X}|\theta} = \mathcal{N}(\theta-\delta, \frac{1}{m})$. Using the χ^2-divergence for Gaussians, we get

$$\chi^2(P_{\theta,X}\|Q_{\theta,X}) = e^{\frac{\delta^2}{s}} - 1 + e^{\frac{\delta^2}{s}}(e^{m\delta^2}-1) = e^{\delta^2(m+\frac{1}{s})} - 1.$$

We can write the Bayesian HCR lower bound as

$$R^*_\pi \geqslant \sup_{\delta\neq 0} \frac{\delta^2}{e^{\delta^2(m+\frac{1}{s})}-1} \geqslant \lim_{\delta\to 0} \frac{\delta^2}{e^{\delta^2(m+\frac{1}{s})}-1} = \frac{s}{sm+1}.$$

In view of the Bayes risk found, we see that in this case the Bayesian HCR and Bayesian Cramér-Rao lower bounds are exact.

2 Bayesian CR lower bounds and extensions

We give the rigorous statement of the Bayesian Cramér-Rao lower bound and discuss its extensions and consequences. For the proof, we take a more direct approach as opposed to the data-processing argument, based on asymptotic expansion of the χ^2-divergence.

Definition 2.1 (Bayesian score). Under the Bayesian setting, let π be a prior density on Θ and $P_\theta \in \mathcal{M}(\mathcal{X})$ be the observation distribution such that $P_\theta \ll \mu$ for some dominating measure $\mu \in \mathcal{M}(\mathcal{X})$ such that $p_\theta \triangleq \frac{dP_\theta}{d\mu}$. We define *Bayesian score* $V : \Theta\times\mathcal{X} \to \mathbb{R}$ defined as $V(\theta,x) \triangleq \nabla_\theta \ln(p_\theta(x)\pi(\theta))$ for (θ,x).

Theorem 2.2 (BCR lower bound). *Consider a statistical decision theory simple setting for $\Theta \triangleq \mathbb{R}$ under the following conditions.*

(a) Let π be a differentiable prior density with compact support $[\theta_0,\theta_1]$, vanishing on the boundary, and finite Fisher information $J(\pi)$.

(b) Let $P_\theta \ll \mu$ for some dominating measure $\mu \in \mathcal{M}(\mathcal{X})$ with relative density $p_\theta \triangleq \frac{dP_\theta}{d\mu}$ differentiable in θ for μ-almost every x.
(c) Let $\int_{\mathcal{X}} d\mu(x) \nabla_\theta p_\theta(x) = 0$ for π-almost every θ.
Then the Bayes quadratic risk $R^*_\pi \triangleq \inf_{\hat{\theta}} \mathbb{E}(\theta - \hat{\theta})^2$ satisfies Bayesian CR lower bound

$$R^*_\pi \geqslant \frac{1}{\mathbb{E}_{\theta\sim\pi} J_F(\theta) + J(\pi)}.$$

Proof. Since Bayes estimator is always deterministic, without loss of any generality, we assume that the estimator $\hat{\theta} \triangleq \hat{\theta}(X)$ is deterministic. For each x, integration by parts yields

$$\int_{\theta_0}^{\theta_1} d\theta(\hat{\theta}(x) - \theta)\nabla_\theta(p_\theta(x)\pi(\theta)) = \int_{\theta_0}^{\theta_1} p_\theta(x)\pi(\theta)d\theta.$$

From the definition of Bayesian score, recalling the fact that integrating both sides over $d\mu(x)$, we obtain the following expectation over the joint distribution of (θ, X),

$$\mathbb{E}[(\hat{\theta} - \theta)V(\theta, X)] = \int_{\Theta\times\mathcal{X}} d\mu(x)d\theta(\hat{\theta}(x) - \theta)p_\theta(x)\pi(\theta)\nabla_\theta \ln(p_\theta(x)\pi(\theta)) = \int_{\Theta\times\mathcal{X}} d\mu(x)d\theta p_\theta(x)\pi(\theta) = 1.$$

Applying Cauchy-Schwarz, we have $\mathbb{E}(\hat{\theta} - \theta)^2 \mathbb{E}V(\theta, X)^2 \geqslant 1$. We further observe that

$$\mathbb{E}V(\theta, X)^2 = \mathbb{E}(\nabla_\theta \ln p_\theta(X))^2 + \mathbb{E}(\nabla_\theta \ln \pi(\theta))^2 + 2\mathbb{E}[\nabla_\theta \ln p_\theta(X)\nabla_\theta \ln \pi(\theta)].$$

The proof is completed by noting that $\mathbb{E}[\nabla_\theta \ln p_\theta(X)\nabla_\theta \ln \pi(\theta)] = \int_{\Theta\times\mathcal{X}} d\mu(x)\nabla_\theta p_\theta(x)\nabla_\theta \pi(\theta) = 0$, since $\int_{\mathcal{X}} d\mu(x)\nabla_\theta p_\theta(x) = 0$ for π-almost every θ. □

Theorem 2.3 (Multivariate BCR). *Consider a statistical decision theory simple setting for $\Theta \triangleq \mathbb{R}^d$ under the following conditions.*
(a) Let π be a product prior density defined as $\pi(\theta) \triangleq \prod_{i=1}^d \pi_i(\theta_i)$ where for each $i]in[d]$, the density π_i is differential with compact support $[\theta_{0,i}, \theta_{1,i}]$, vanishing on the boundary, and has finite Fisher information $J(\pi_i)$.
(b) Let $P_\theta \ll \mu$ for some dominating measure $\mu \in \mathcal{M}(\mathcal{X})$ with relative density $p_\theta \triangleq \frac{dP_\theta}{d\mu}$ differentiable in θ for μ-almost every x.
(c) Let $\int_{\mathcal{X}} d\mu(x)\nabla_\theta p_\theta(x) = 0$ for π-almost every θ.
We define the Fisher information matrices by

$$J_F(\theta) \triangleq \mathbb{E}_\theta[\nabla_\theta \ln p_\theta(X)\nabla_\theta \ln p_\theta(X)^\top], \qquad J(\pi) \triangleq \operatorname{diag}(J(\pi_1), \ldots, J(\pi_d)).$$

Then, the quadratic Bayes risk is lower bounded in terms of two Fisher information matrices as

$$R^*_\pi \triangleq \inf_{\hat{\theta}} \mathbb{E}_\pi \left\|\hat{\theta} - \theta\right\|^2 \geqslant \operatorname{tr}(\mathbb{E}_{\theta\sim\pi} J_F(\theta) + J(\pi))^{-1}.$$

Proof. From conditions, we observe that π is defined over the box $\prod_{i=1}^d [\theta_{0,i}, \theta_{1,i}]$. We fix a deterministic estimator $\hat{\theta} \triangleq (\hat{\theta}_1(X), \ldots, \hat{\theta}_d(X))$ and a non-zero $u \in \mathbb{R}^d$. Let e_k be the unit vector in kth dimension. For each $i, k \in [d]$, integration by parts yields

$$\int_{\theta_{0,i}}^{\theta_{1,i}} (\hat{\theta}_k(x) - \theta_k)\nabla_{\theta_i}(p_\theta(x)\pi(\theta))d\theta_i = \langle e_i, e_k\rangle \int_{\theta_{0,i}}^{\theta_{1,i}} d\theta_i p_\theta(x)\pi(\theta).$$

Integrating both sides over $\prod_{j\neq i} d\theta_j$ and $d\mu(x)$, multiplying by u_i, and summing over i, we obtain

$$\mathbb{E}(\hat{\theta}_k(X) - \theta_k)\langle u, \nabla \ln p_\theta(X)\pi(\theta)\rangle = \langle u, e_k\rangle.$$

Defining $\Sigma \triangleq \mathbb{E}\nabla \ln(p_\theta(X)\pi(\theta))(\nabla \ln(p_\theta(X)\pi(\theta)))^\top = \mathbb{E}_{\theta\sim\pi} J_F(\theta) + J(\pi)$, and applying Cauchy-Schwarz and optimizing over u yields

$$\mathbb{E}(\hat{\theta}_k(X) - \theta_k)^2 \geqslant \sup_{u\neq 0} \frac{\langle u, e_k\rangle}{u^\top \Sigma u} = \Sigma^{-1}_{kk}.$$

Summing over k completes the proof. □

Remark 1. The multivariate Bayesian CR lower bound depends on the choice of prior density.

- The above versions of the BCR bound assume a prior density that vanishes at the boundary. If we choose a uniform prior, the same derivation leads to a similar lower bound known as the Chernoff-Rubin-Stein inequality, which also suffices for proving the optimal minimax lower bound.
- For the purpose of the lower bound, it is advantageous to choose a prior density with the minimum Fisher information. The optimal density with a compact support is known to be a squared cosine density. That is, the following minimum is attained by $g(u) = \cos^2 \frac{\pi u}{2}$,

$$\min_{g \text{ on } [-1,1]} J(g) = \pi^2.$$

- Suppose the goal is to estimate a smooth *functional* $T(\theta)$ of the unknown parameter θ, where $T : \mathbb{R}^d \to \mathbb{R}^s$ is differentiable with Jacobian matrix $\nabla T(\theta) \in \mathbb{R}^{s\times d}$ defined as $[\nabla T(\theta)]_{ij} \triangleq \frac{\partial T_i(\theta)}{\partial \theta_j}$ for all $i, j \in [s] \times [d]$. Then under the same condition of Theorem 2.3, we have the following Bayesian Cramér-Rao lower bound for functional estimation

$$\inf_{\hat{T}} \mathbb{E}_\pi \left\| \hat{T}(X) - T(\theta) \right\|_2^2 \geqslant \operatorname{tr} \mathbb{E}_{\theta\sim\pi} \nabla T(\theta) (\mathbb{E}_{\theta\sim\pi} J_F(\theta) + J(\pi))^{-1} \mathbb{E}_{\theta\sim\pi} \nabla T(\theta)^\top.$$

Theorem 2.4. *Assume that the map $\theta \mapsto J_F(\theta)$ is continuous. Denote the minimax squared error for* i.i.d. *sample $X : \Omega \to \mathcal{X}^m$ under common distribution $P_\theta \in \mathcal{M}(\mathcal{X})$ as $R_m^* \triangleq \inf_{\hat{\theta}} \sup_{\theta\in\Theta} \mathbb{E}_\theta \left\| \hat{\theta} - \theta \right\|^2$. Then, as $m \to \infty$, we have*

$$R_m^* \geqslant \frac{1 + o(1)}{m} \sup_{\theta\in\Theta} \operatorname{tr} J_F^{-1}(\theta).$$

Proof. Fix $\theta \in \Theta \subseteq \mathbb{R}^d$. Then for all sufficiently small δ, we have $B_\infty(\theta, \delta) \triangleq \theta + [-\delta, \delta]^d \subset \Theta$. Consider marginal prior densities $\pi_i(\theta_i) \triangleq \frac{1}{\delta} g\left(\frac{\theta - \theta_i}{\delta}\right)$, where g is the squared cosine density. Then the product distribution $\pi \triangleq \prod_{i=1}^d \pi_i$ satisfies the assumption of Theorem 2.3. By the scaling rule of Fisher information, we obtain $J(\pi_i) = \frac{1}{\delta^2} J(g) = \frac{\pi^2}{\delta^2}$. Thus $J(\pi) = \frac{\pi^2}{\delta^2} I_d$. It is known that the continuity of $\theta \mapsto J_F(\theta)$ implies $\int_{\mathcal{X}} d\mu(x) \nabla_\theta p_\theta(x) = 0$. Therefore, we can apply Theorem 2.3 to get the BCR lower bound. Further, lower bounding the minimax risk by the Bayes risk and applying the additivity property of Fisher information, we get

$$R_m^* \geqslant \frac{1}{m} \operatorname{tr} \left(\mathbb{E}_{\theta\sim\pi} J_F(\theta) + \frac{\pi^2}{m\delta^2} I_d \right)^{-1}.$$

Finally, choosing $\delta = m^{-1/4}$ and applying the continuity of $J_F(\theta)$ in θ, we get the result. □