# Lecture-24: Large sample asymptotic

## 1 Statistical lower bound from data processing

We give an overview of the classical large-sample theory in the setting of *i.i.d.* observations focusing again on the minimax risk. We focus primarily on the quadratic risk and assume that $\Theta \subseteq \mathbb{R}^d$ is an open set. These results pertain to smooth parametric models in fixed dimensions, with the sole asymptotics being the sample size going to infinity. The main result is that, under suitable conditions, the minimax squared error of estimating $\theta$ based on *i.i.d.* sample $X : \Omega \to \mathcal{X}^m$ with common distribution $P_\theta \in \mathcal{P}(\Theta)$ and Fisher information matrix $J_F(\theta)$ satisfies

$$R_m^* \triangleq \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}[\|\hat{\theta} - \theta\|^2 \mid \theta] = \frac{1 + o(1)}{m} \sup_{\theta \in \Theta} \operatorname{tr} J_F^{-1}(\theta). \tag{1}$$

*Remark* 1. This is asymptotic characterization of the minimax risk with sharp constant. In high dimensions, such precise results are difficult and rare.
(a) We derive several statistical lower bounds from data processing argument.
(b) Specifically, we will take a comparison-of-experiment approach by comparing the actual model with a perturbed model.
(c) The performance of a given estimator can be then related to the $f$-divergence via the data processing inequality and the variational representation.

We start by discussing the Hammersley-Chapman-Robbins lower bound which implies the well-known Cramér-Rao lower bound. Because these results are restricted to unbiased estimators, we will also discuss their Bayesian version.

### 1.1 Hammersley-Chapman-Robbins (HCR) lower bound

**Theorem 1.1 (HCR lower bound).** *Consider the statistical decision theory simple setting with $\mathcal{Y} = \Theta = \Theta' \triangleq \mathbb{R}$, and quadratic loss function $\ell : (\theta, \hat{\theta}) \mapsto (\theta - \hat{\theta})^2$. The quadratic risk at any parameter $\theta \in \Theta$ satisfies*

$$R_\theta(\hat{\theta}) = \mathbb{E}_\theta(\theta - \hat{\theta})^2 \geqslant \operatorname{Var}_\theta(\hat{\theta}) \geqslant \sup_{\theta \neq \theta'} \frac{(\mathbb{E}_\theta[\hat{\theta}] - \mathbb{E}_{\theta'}[\hat{\theta}])^2}{\chi^2(P_{\theta'} \| P_\theta)}.$$

*Proof.* Since the conditional expectation minimizes quadratic risk, we have $\mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta} \mid \theta])g(\theta) \mid \theta] = 0$ for any function $g : \Theta \to \Theta'$. Further, from the property of conditional mean, the difference $\theta - \mathbb{E}[\hat{\theta} \mid \theta]$ is a function of $\theta$. Hence, it follows that

$$R_\theta(\hat{\theta}) = \mathbb{E}[(\theta - \hat{\theta})^2 \mid \theta] = \mathbb{E}[(\theta - \mathbb{E}[\hat{\theta} \mid \theta])^2 \mid \theta] + \mathbb{E}[(\hat{\theta} - \mathbb{E}[\hat{\theta} \mid \theta])^2 \mid \theta] \geqslant \operatorname{Var}_\theta(\hat{\theta}). \tag{2}$$

Conside a two subset $\{\theta, \theta'\} \subseteq \Theta$, and a random estimator $\hat{\theta}(X)$ defined by the Markov kernel $P_{\hat{\theta}(X)|X} : \mathcal{X} \to \mathcal{M}(\Theta')$. Consider the Markov chains $\theta \to X \to \hat{\theta}$ and $\theta' \to X \to \hat{\theta}$, where input is random observation $X$ and the output is random estimate $\hat{\theta}(X)$ with the common channel $P_{\hat{\theta}(X)|X}$. Corresponding to two different input distributions $Q_X \triangleq P_\theta$ and $P_X \triangleq P_{\theta'}$, we denote the marginal distribution of estimators as $Q_{\hat{\theta}} \triangleq \mathbb{E}_{X \sim Q_X} P_{\hat{\theta}(X)|X}$ and $P_{\hat{\theta}} \triangleq \mathbb{E}_{X \sim P_X} P_{\hat{\theta}(X)|X}$, respectively. From the data processing inequality for $f$-divergence and the variational representation of $\chi^2$-divergence, we obtain

$$\chi^2(P_X \| Q_X) \geqslant \chi^2(P_{\hat{\theta}} \| Q_{\hat{\theta}}) \geqslant \frac{(\mathbb{E}_\theta[\hat{\theta}] - \mathbb{E}_{\theta'}[\hat{\theta}])^2}{\operatorname{Var}_\theta(\hat{\theta})}.$$

□

**Corollary 1.2 (Cramér-Rao (CR) lower bound).** *Under the regularity conditions for parametric family, on (a) the existence of relative density, (b) the existence of continuous derivative of relative density with respect to parameter $\theta$, and (c) the uniform integrability of the ratio of square of derivative of the density and density, we have for any unbiased estimator $\hat{\theta}$ that satisfies $\mathbb{E}_\theta[\hat{\theta}] = \theta$ for all $\theta \in \Theta \subset \mathbb{R}$,*

$$\mathrm{Var}_\theta(\hat{\theta}) \geqslant \frac{1}{J_F(\theta)}. \tag{3}$$

*Proof.* From HCR lower bound in Theorem 1.1, we get $R_\theta(\hat{\theta}) = \mathrm{Var}_\theta(\hat{\theta}) \geqslant \sup_{\theta' \neq \theta} \frac{(\theta-\theta')^2}{\chi^2(P_{\theta'} \| P_\theta)}$. The result follows by lower bounding the supremum by the limit of $\theta' \to \theta$, and recalling the asymptotic quadratic expansion of $\chi^2$-divergence in the local neighborhood in terms of the Fisher information. □

**Exercise 1.3.** Show that for vector $y \in \mathbb{R}^d$ and a positive definite matrix $\Sigma \in \mathbb{R}^{d\times d}$, we have $\sup_{x\in\mathbb{R}^d : x\neq 0} \frac{\langle x,y\rangle^2}{x^\top \Sigma x} = y^\top \Sigma^{-1} y$, where the maxima is achieved at $x^* = \Sigma^{-1} y$.

*Remark* 2. We note the following for HCR lower bound and CR lower bound.

(a) The HCR lower bound is based on the $\chi^2$-divergence. We can write a lower bound version based on Hellinger distance which also implies the CR lower bound.

(b) Both the HCR and the CR lower bounds extend to the multivariate case as follows. Let $\hat{\theta}$ be an unbiased estimator of $\theta \in \Theta \subseteq \mathbb{R}^d$. Assume that its covariance matrix $\mathrm{Cov}_\theta(\hat{\theta}) \triangleq \mathbb{E}_\theta(\hat{\theta}-\theta)(\hat{\theta}-\theta)^\top$ is positive definite. Fix $a \in \mathbb{R}^d$. Applying HCR lower bound to estimand $T(\theta) \triangleq \langle a,\theta\rangle$ and estimator $\hat{T}(X) \triangleq \langle a, \hat{\theta}(X)\rangle$, we get

$$\chi^2(P_{\theta'} \| P_\theta) \geqslant \frac{(\mathbb{E}_\theta \langle a,\hat{\theta}\rangle - \mathbb{E}_{\theta'} \langle a,\hat{\theta}\rangle)^2}{\mathrm{Var}_\theta \langle a,\hat{\theta}\rangle} = \frac{\langle a, \theta - \theta'\rangle^2}{a^\top \mathrm{Cov}_\theta(\hat{\theta}) a}.$$

Since the choice of $a \in \mathbb{R}^d$ was arbitrary, the right hand side of the equation holds for all $a$. Taking supremum over $a$, it follows from Exercise 1.3 that

$$\chi^2(P_{\theta'} \| P_\theta) \geqslant (\theta - \theta')^\top \mathrm{Cov}_\theta(\hat{\theta})^{-1} (\theta - \theta').$$

(c) From the additivity property of the Fisher information, the Fisher information matrix for a sample of $m$ *i.i.d.* observations is equal to $mJ_F(\theta)$. Writing the Taylor series expansion of $\chi^2$-divergence in the neighborhood of $\theta \in \Theta \subseteq \mathbb{R}^d$, we get

$$(\theta' - \theta)^\top \left( mJ_F(\theta) - (\mathrm{Cov}_\theta(\hat{\theta}))^{-1} \right) (\theta' - \theta) + o(\|\theta' - \theta\|^2) \geqslant 0.$$

Taking the limit $\theta' \to \theta$, we obtain $mJ_F(\theta) - (\mathrm{Cov}_\theta(\hat{\theta}))^{-1} \succcurlyeq 0$, and taking trace we conclude that the squared error of any unbiased estimators satisfies

$$\mathbb{E}_\theta \|\hat{\theta} - \theta\|^2 = \mathrm{tr}\,\mathrm{Cov}_\theta(\hat{\theta}) \geqslant \frac{1}{m} \mathrm{tr}\, J_F^{-1}(\theta).$$

This is already very close to (1), except for the fundamental restriction of unbiased estimators.

## 1.2 Bayesian HCR and CR lower bounds

The drawback of the HCR and CR lower bounds is that they are confined 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.

**Exercise 1.4 (Chain rule for $\chi^2$-divergence).** Show that for any pair of measures $P_{X,Y}$ and $Q_{X,Y}$ we have

$$\chi^2(P_{X,Y}\|Q_{X,Y}) = \chi^2(P_X\|Q_X) + \mathbb{E}_{X\sim Q_X}\left[\Big(\frac{dP_X}{dQ_X}\Big)^2\chi^2(P_{Y|X}\|Q_{Y|X})\right], \tag{4}$$

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

**Exercise 1.5 (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

$$D_f(P_{X,Y}\|Q_{X,Y}) \geqslant D_f(P_{X,Z}\|Q_{X,Z}) \geqslant D_f(P_{g(X,Z)}\|Q_{g(X,Z)}). \tag{5}$$

**Definition 1.6 (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 $\delta$ 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 $(T_\delta\mu)(-\infty, x] \triangleq \mu(-\infty, x-\delta]$ for any $x \in \mathbb{R}$.

**Theorem 1.7 (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 $\pi'$ 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

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

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

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

Consider a random estimator $\hat\theta(X)$ for observation $X$ and external randomness with Markov kernel $P_{\hat\theta|X}$ such that $\theta \to X \to \hat\theta$ is a Markov chain. Consider joint distributions $P_{\theta,X}, Q_{\theta,X}$, apply data processing inequality from Exercise 1.5, and variational representation of $\chi^2$-divergence from Exercise 1.4, to obtain

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

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 $\mathbb{E}_{(\theta,X)\sim P}\hat\theta = \mathbb{E}_{(\theta,X)\sim Q}\hat\theta$. On the other hand, $\mathbb{E}_{(\theta,X)\sim P}\theta = \mathbb{E}_{(\theta,X)\sim Q}\theta + \delta$. Next, we focus on the denominator

$$\mathrm{Var}_{(\theta,X)\sim Q}(\hat\theta-\theta) = \mathbb{E}_{(\theta,X)\sim Q}(\theta-\hat\theta)^2 - (\mathbb{E}_{(\theta,X)\sim Q}(\theta-\hat\theta))^2 \leqslant \mathbb{E}_{(\theta,X)\sim Q}(\theta-\hat\theta)^2$$

with equality iff $\mathbb{E}_{(\theta,X)\sim Q}\theta = \mathbb{E}_{(\theta,X)\sim Q}\hat\theta$. Since this applies to any estimator, the result follows. □

**Definition 1.8 (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 1.9 (Bayesian CR lower bound).** *Under the conditions of Theorem 1.7 and suitable regularity conditions for the local expansion of $\chi^2$-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 1.7 by evaluating the small-$\delta$ 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 $\chi^2$-divergence in Exercise 1.4, and applying the local expansion of $\chi^2$-divergence we obtain the result. □

**Example 1.10 (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

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

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 $\chi^2$-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.