Last lecture we discussed systematic methods to find the best inequalities between different f-divergence via their joint range. We showed that examining the binary cases is sufficient to derive optimal inequalities. In this lecture we will further discuss lower bounds for statistical estimation using f-divergences.

Outline:

- Variational representation of f-divergences.
  - Convexity.
  - Lower semi-continuity.
- (Specializing to χ^2) Lower bounds for statistical estimation.
  - Hammersley-Chapman-Robbins (HCR) lower bound.
  - Cramér-Rao (CR) lower bound.
  - Bayesian Hammersley-Chapman-Robbins (HCR) lower bound.
  - Bayesian Cramér-Rao (CR) lower bound.

29.1 Hammersley-Chapman-Robbins (HCR) lower bound

In this section, we derive a useful statistical lower bound by applying the variational representation of f-divergence in Section 7.5. Specifically, we will focus on the χ^2-divergence for probability distributions P and Q on $\mathbb{R}$.[1] By limiting the choice of function h to affine functions, the equality (7.27) becomes an inequality. In particular, let $h(x) = ax + b$ and optimize over $a, b \in \mathbb{R}$, we have

$$\chi^2(P\|Q) \geq \sup_{a,b\in\mathbb{R}} \left\{ 2(a\mathbb{E}_P(X) + b) - \mathbb{E}_Q[(aX+b)^2] - 1 \right\} = \frac{(\mathbb{E}_P[X] - \mathbb{E}_Q[X])^2}{\mathrm{Var}_Q(X)}. \tag{29.1}$$

Note: The inequality (29.1) can be interpreted as follows: On the left hand side of the inequality we have the χ^2-divergence, a measure of the dissimilarity between two distributions. Looking at the right hand side we see that if the two distributions are centered at very distant locations, then the right hand side will be large. Due to (29.1), this will lead to a bigger χ^2-divergence something that was in fact expected.

The reason that the variance with respect to the Q distribution appears in the denominator is to quantify how different the two means are *relatively*. Indeed, the standard deviation must appear as a normalizing factor because the LHS is a numerical number. Also, the bound only involves the variance under Q not P, which is consistent with the asymmetry of χ^2-divergence.

[1] This can always be assumed by allowing the likelihood ratio function $\frac{dP}{dQ}$ which is a sufficient statistic.

Using (7.27) we now derive the HCR lower bound on the variance of an estimator (possibly randomized). To this end, assume that data $X \sim P_\theta$, where $\theta \in \Theta \subset \mathbb{R}$. We use quadratic cost to quantify the difference between the real and the predicted parameter, i.e., $\ell(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$. Then the risk of estimator $\hat{\theta}$ when the real parameter is θ is given by $R_\theta(\hat{\theta}) = \mathbb{E}_\theta[(\theta - \hat{\theta})^2]$. Now, fix $\theta \in \Theta$. For any other $\theta' \in \Theta$ we will use (29.1) with $Q_X = P_\theta$ and $P_X = P_{\theta'}$. As a result we have that

$$\chi^2(P_{\theta'} \| P_\theta) = \chi^2(Q_X \| P_X) \geq \chi^2(P_{\hat{\theta}} \| Q_{\hat{\theta}}) \geq \frac{(\mathbb{E}_\theta[\hat{\theta}] - \mathbb{E}_{\theta'}[\hat{\theta}])^2}{\mathrm{Var}_\theta(\hat{\theta})} \tag{29.2}$$

Where the first inequality arises by using the data processing inequality and the second inequality by (29.1). Finally, by swapping the denominator with the left hand side and taking the supremum over all $\theta' \neq \theta$, and since $\mathrm{Var}_\theta(\hat{\theta})$ is not a function of θ', we derive the final result.

Theorem 29.1 (Hammersley-Chapman-Robbins (HCR) lower bound)**.** *For the quadratic loss, any estimator $\hat{\theta}$ satisfies*

$$R_\theta(\hat{\theta}) \geq \mathrm{Var}_\theta(\hat{\theta}) \geq \sup_{\theta' \neq \theta} \frac{(\mathbb{E}_\theta[\hat{\theta}] - \mathbb{E}_{\theta'}[\hat{\theta}])^2}{\chi^2(P_{\theta'} \| P_\theta)}, \quad \forall \theta \in \Theta. \tag{29.3}$$

When $\{P_\theta : \theta \in \Theta\}$ have different support, consider the following version: Fix $\epsilon \in (0,1)$. Similar to (29.2), let us apply χ^2-data processing to the pairs $Q_X = \bar{\epsilon} P_\theta + \epsilon P_{\theta'}$ and $P_X = P_{\theta'}$. By linearity of expectation, we get

$$\chi^2(P_{\theta'} \| \bar{\epsilon} P_\theta + \epsilon P_{\theta'}) \geq \bar{\epsilon}^2 \frac{(\mathbb{E}_\theta[\hat{\theta}] - \mathbb{E}_{\theta'}[\hat{\theta}])^2}{\mathrm{Var}_{\bar{\epsilon} P_\theta + \epsilon P_{\theta'}}(\hat{\theta})} \tag{29.4}$$

Note that the LHS is equal to $\epsilon \bar{\epsilon} D_{f_\epsilon}(P_{\theta'} \| P_\theta)$, which is a f-divergence defined by $f_\epsilon(x) = \frac{(x-1)^2}{\epsilon x + \bar{\epsilon}}$. Applying its local expansion from Theorem **TODO**, we get

$$D_{f_\epsilon}(P_{\theta'} \| P_\theta) = I(\theta)(\theta' - \theta)^2(1 + o(1)), \quad \theta' \to \theta.$$

where we used the fact that $f''_\epsilon(1) = 2$.

Using the fact that $\mathrm{Var}_{\bar{\epsilon} P_\theta + \epsilon P_{\theta'}}(\hat{\theta}) = \bar{\epsilon}\, \mathrm{Var}_\theta(\hat{\theta}) + \epsilon\, \mathrm{Var}_{\theta'}(\hat{\theta}) + 2\epsilon\bar{\epsilon}(\mathbb{E}_\theta[\hat{\theta}] - \mathbb{E}_{\theta'}[\hat{\theta}])^2$, by first sending $\theta' \to \theta$ followed by $\epsilon \to 0$, we conclude from (29.4) that, for unbiased $\hat{\theta}$,

$$\mathrm{Var}_\theta(\hat{\theta}) \geq \frac{1}{I(\theta)}.$$

29.2 Cramér-Rao (CR) lower bound

We now derive the Cramér-Rao lower bound as a consequence of the HCR lower bound. To this end, we restrict the problem to unbiased estimators, where an estimator $\hat{\theta}$ is said to be unbiased if $\mathbb{E}_\theta[\hat{\theta}] = \theta$ for all $\theta \in \Theta$. Then by applying the HCR lower bound we have that

$$R_\theta(\hat{\theta}) = \mathrm{Var}_\theta(\hat{\theta}) \geq \sup_{\theta' \neq \theta} \frac{(\theta - \theta')^2}{\chi^2(P_{\theta'} \| P_\theta)} \geq \lim_{\theta' \to \theta} \frac{(\theta - \theta')^2}{\chi^2(P_{\theta'} \| P_\theta)}. \tag{29.5}$$

As $\theta' \to \theta$, we expect the denominator will go to zero quadratically as the numerator does. Recall that

$$\chi^2(P_{\theta'} \| P_\theta) = \int \frac{(P_\theta - P_{\theta'})^2}{P_\theta}.$$

Then by using the Taylor expansion for P_θ around θ' we get that

$$P_\theta - P_{\theta'} = (\theta - \theta')\frac{dP_\theta}{d\theta} + o[(\theta - \theta')^2],$$

for θ near θ'. Combining the above while ignoring the little-o terms we get that

$$\chi^2(P_{\theta'} \| P_\theta) = (\theta - \theta')^2 \int \frac{(\frac{dP_\theta}{d\theta})^2}{P_\theta}.$$

Plugging back in (29.5) we get the well-known Cramér-Rao (CR) lower bound.

Theorem 29.2. *For any unbiased estimator $\hat{\theta}$ and any $\theta \in \Theta$*

$$\mathrm{Var}_\theta(\hat{\theta}) \geq \frac{1}{I(\theta)},$$

where $I(\theta)$ is the Fisher information given by

$$I(\theta) = \int \frac{(\frac{dP_\theta}{d\theta})^2}{P_\theta}.$$

An intuitive interpretation of $I(\theta)$ is that it is a measure of the information the data contains for the estimation of the parameter when its true value is θ.

Example 29.1 (GLM)**.** Let $\theta \in \mathbb{R}$ and $X \sim P_\theta = \mathcal{N}(\theta, 1)$. Define the standard normal density by $\varphi(x)$. Then the density of P_θ is $p_\theta(x) = \varphi(x - \theta)$. Next we calculate the Fisher information. By shifting x to θ, note that

$$I(\theta) = \int \frac{(\frac{\partial p_\theta(x)}{\partial\theta})^2}{p_\theta(x)} dx = \int (x - \theta)^2 \varphi(x - \theta) dx = 1.$$

Thus, $I(\theta) \equiv I(0) = 1$ for all $\theta \in \Theta$. In general, this is for any location model where $X = \theta + Z$, the Fisher information is the same everywhere.

Remark 29.1. Another useful way of seeing the Fisher information is the following:

$$I(\theta) = \int \frac{(\frac{\partial P_\theta(x)}{\partial\theta})^2}{P_\theta(x)} \partial x = \mathbb{E}_\theta\left[\left(\frac{\frac{\partial P_\theta(X)}{\partial\theta}}{P_\theta(X)}\right)^2\right] = \mathbb{E}_\theta\left[\left(\frac{\partial \log P_\theta(X)}{\partial\theta}\right)^2\right] = \mathrm{Var}_\theta\left[\frac{\partial \log P_\theta(X)}{\partial\theta}\right],$$

where the last equality holds after noticing that

$$\mathbb{E}_\theta\left[\frac{\partial \log P_\theta(X)}{\partial\theta}\right] = 0.$$

29.3 Fisher information

The Fisher information is a way of measuring the amount of information that an observable random variable X carries about an unknown, deterministic parameter θ upon which the distribution of the observation X depends. Assume the probability density function of random variable X conditional on the value of θ is p_θ. The Fisher information is defined as

Definition 29.1 (Fisher information)**.** The Fisher information of the parameteric family of densitities $\{p_\theta : \theta \in \Theta\}$ (with respect to μ) at θ is

$$I(\theta) = \mathbb{E}\left[\left(\frac{\partial \log p_\theta}{\partial \theta}\right)^2\right] = \int \left(\frac{\partial p_\theta}{\partial \theta}\right)^2 \frac{1}{p_\theta} d\mu. \tag{29.6}$$

Theorem 29.3 (Fisher information)**.** *Assume that p_θ is twice differentiable with respect to θ and satisfies the regularity condition:*

$$\int \frac{\partial^2 p_\theta}{\partial \theta^2} d\mu = \frac{\partial^2}{\partial \theta^2} \int p_\theta d\mu = 0.$$

The Fisher information can be written as

$$I(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2 \log p_\theta}{\partial \theta^2}\right]$$

Proof. Since

$$\frac{\partial^2 \log p_\theta}{\partial \theta^2} = \frac{\frac{\partial^2 p_\theta}{\partial \theta^2}}{p_\theta} - \left(\frac{\frac{\partial p_\theta}{\partial \theta}}{p_\theta}\right)^2 = \frac{\frac{\partial^2 p_\theta}{\partial \theta^2}}{p_\theta} - \left(\frac{\partial \log p_\theta}{\partial \theta}\right)^2$$

and

$$\mathbb{E}\left[\frac{\partial^2 p_\theta}{\partial \theta^2}\frac{1}{p_\theta}\right] = 0$$

by assumption, we have

$$I(\theta) = \mathbb{E}_\theta\left[\left(\frac{\partial}{\partial \theta}\log p_\theta\right)^2\right] = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial \theta^2}\log p_\theta\right].$$

□

Theorem 29.4 (Fisher information: mutiple sample)**.** *Suppose random sample $X_1, \ldots, X_n$ independently and identically drawn from a distribution p_θ. The Fisher information $I_n(\theta)$ provided by random samples $X_1, \ldots, X_n$ is*

$$I_n(\theta) = nI(\theta),$$

where $I(\theta)$ is Fisher information provided by a single sample X_1.

Proof. We first denote the joint pdf of $X_1, \ldots, X_n$ as

$$p_\theta(x_1, \ldots, x_n) = \prod_{i=1}^n p_\theta(x_i).$$

Then the Fisher information $I_n(\theta)$ provided by $X_1, \ldots, X_n$ is

$$I_n(\theta) = \mathbb{E}_\theta\left[\left(\frac{\partial p_\theta(X_1, \ldots, X_n)}{\partial \theta}\right)^2\right] = \int \cdots \int \left(\frac{\partial p_\theta(x_1, \ldots, x_n)}{\partial \theta}\right)^2 p_\theta(x_1, \ldots, x_n) dx_1 dx_2 \ldots dx_n,$$

which is an n-dimensional integral. Thus, by Theorem 29.3, the Fisher information provided by $X_1, \ldots, X_n$ can be calculated as

$$I_n(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2 \log p_\theta(X_1, \ldots, X_n)}{\partial \theta^2}\right] = -\mathbb{E}_\theta\left[\sum_{i=1}^n \frac{\partial^2 \log p_\theta(X_i)}{\partial \theta^2}\right] = -\sum_{i=1}^n \mathbb{E}_\theta\left[\frac{\partial^2 \log p_\theta(X_i)}{\partial \theta^2}\right] = nI(\theta).$$

□

29.4 Variations of HCR/CR lower bound

This section contains the following three versions of HCP/CR lower bound:

- Multiple Samples Version
- Multivariate Version
- Functional Version

29.4.1 Multiple-sample version

Suppose θ is some unknown, deterministic parameter and $X_1, \ldots, X_n$ are n random variables iid drawn from the distribution P_θ. The estimate $\hat{\theta}$ comes from $X_1, \ldots, X_n$. The relationships is shown as follows:

$$\theta \to X_1, \ldots, X_n \to \hat{\theta}.$$

Then the risk is lower bound by

$$R_\theta(\hat{\theta}) \geq \mathrm{Var}_\theta(\hat{\theta}) \geq \frac{(\mathbb{E}_\theta \hat{\theta} - \mathbb{E}_{\theta'} \hat{\theta})^2}{\chi^2(P_{\theta'}^{\otimes n} \| P_\theta^{\otimes n})}.$$

For the HCR lower bound,

$$R_\theta(\hat{\theta}) \geq \sup_{\theta \neq \theta'} \frac{(\theta - \theta')^2}{(1 + \chi^2(P_\theta \| P_{\theta'}))^n - 1} \overset{\theta' \to \theta}{\geq} \frac{1}{nI(\theta)}.$$

29.4.2 Multivariate Version

We next show the multi-dimensional version of

$$\chi^2(P \| Q) \geq \frac{(\mathbb{E}_P X - \mathbb{E}_Q X)^2}{\mathrm{Var}_Q X}.$$

Suppose P, Q are two distributions defined on $\mathbb{R}^p$, then

$$\chi^2(P \| Q) = \sup_{g: \mathbb{R}^p \to \mathbb{R}} [2\mathbb{E}_P g(X) - \mathbb{E}_Q g^2(X) - 1].$$

Furthter, if $g(X) = \langle a, X \rangle + 1$, then

$$\chi^2(P \| Q) \geq 2\mathbb{E}_P \langle a, X \rangle + 1 - \mathbb{E}_Q(\langle a, X \rangle + 1)^2.$$

If we further assume $\mathbb{E}_Q X = 0$, then we have

$$\chi^2(P \| Q) \geq 2 \langle a, \mathbb{E}_P X \rangle - a^T \mathbb{E}_Q[XX^T]a.$$

Therefore, we finally have

$$\chi^2(P \| Q) \geq (\mathbb{E}_P X - \mathbb{E}_Q X)^T \underset{Q}{\overset{-1}{\mathrm{cov}}}(X)(\mathbb{E}_P X - \mathbb{E}_Q X)$$

Let the loss function $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|_2^2$ and $\hat{\theta}$ be the unbiased estimate of θ, *i.e.*, $\mathbb{E}_\theta \hat{\theta} = \theta$. Then

$$(\theta' - \theta)^T \underset{\theta}{\overset{-1}{\mathrm{cov}}}(\hat{\theta})(\theta' - \theta) \leq \chi^2(P_{\theta'} \| P_\theta) \overset{\theta' \to \theta}{=} (\theta' - \theta)^T \mathbf{I}(\theta)(\theta' - \theta) + \|\theta' - \theta\|_2^2,$$

where the equality follows from the Taylor expansion and Fisher information matrix is given as

$$\mathbf{I}(\theta) = \int \frac{\nabla P_\theta (\nabla P_\theta)^T}{P_\theta}.$$

If we take $\theta' = \theta + \epsilon u$ for an arbitrary unit vector u and $\epsilon \to 0$, we have

$$u^T \underset{\theta}{\overset{-1}{\operatorname{cov}}}(\hat{\theta}) u \leq u^T \mathbf{I}(\theta) u,$$

which is equivalent to

$$\underset{\theta}{\operatorname{cov}}(\hat{\theta}) \succeq \mathbf{I}^{-1}(\theta),$$

and further indicates

$$R_\theta(\hat{\theta}) = \operatorname{tr}(\underset{\theta}{\operatorname{cov}}(\hat{\theta})) \geq \operatorname{tr}(\mathbf{I}^{-1}(\theta)). \tag{29.7}$$

Then we have

$$\mathbb{E}\|\theta - \hat{\theta}\|_2^2 = \sum_{i=1}^{p} \mathbb{E}(\hat{\theta}_i - \theta_i)^2 \geq \sum_{i=1}^{p} \frac{1}{I_i}, \tag{29.8}$$

where $I_i \triangleq \mathbf{I}_{ii}(\theta)$, since

$$\sum_{i=1}^{p} \frac{1}{I_i(\theta)} \leq \operatorname{tr}(I^{-1}(\theta)).$$

Note that if we apply the one-dimensional CRLB for each coordinate we would get (29.8) which is weaker than (29.7).

Finally, similar to Theorem 29.3, assuming the corresponding regularity of the Hessian, the Fisher information matrix can be written as

$$\mathbf{I}(\theta) = \mathbb{E}_\theta[(\nabla \log P_\theta)(\nabla \log P_\theta)^T] = \underset{\theta}{\operatorname{cov}}(\nabla \log P_\theta) = -\left(\mathbb{E}_\theta\left[\frac{\partial^2 \log P_\theta}{\partial\theta_i \partial\theta_j}\right]\right).$$

29.4.3 Functional Version

Assume that θ is an unknown parameter, that random variable X comes from the distribution P_θ and that $\hat{T}(X)$ is an estimation for $T(\theta)$, where $T : \Theta \to \mathbb{R}$. The relationship is shown as follows:

$$\theta \to X \to \hat{T}.$$

If we further assume $\hat{T}(\theta)$ is an unbiased estimation for $T(\theta)$, then

$$\operatorname{Var}_\theta(\hat{T}) \geq \frac{\|\nabla T\|_2^2}{I(\theta)}$$

29.5 Bayesian Cramér-Rao Lower Bound via data processing inequality

The class will introduce two methods of proving Bayesian Cramér-Rao lower bound.

- Method 1: $\chi^2 \to$ Bayesian HCR $\to$ Bayesian CR (next).
- Method 2: Classical Method.

The notation used in this section is shown as follows:

- $\Theta = \mathbb{R}$
- $\ell(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$.
- π is a "nice" prior on $\mathbb{R}$

The relationship can be described by the following Markov chain:

$$\pi \to \theta \to X \to \hat{\theta}.$$

Theorem 29.5 (Bayesian Cramér-Rao Lower Bound)**.** *Assuming suitable regularity conditions, then*

$$R^* \geq R_\pi^* = \inf_{\hat{\theta}} \mathbb{E}_\pi(\theta, \hat{\theta})^2 \geq \frac{1}{\mathbb{E}_{\theta\sim\pi} I(\theta) + I(\pi)},$$

where R_π^* is the Bayes risk and $I(\pi) = \int \frac{\pi'^2}{\pi}$ is the Fisher information of the prior.

Proof. Consider the following comparison of experiments:

$$Q : \pi \longrightarrow \theta \xrightarrow{P_\theta = Q_{X|\theta}} X \longrightarrow \hat{\theta},$$

$$P : \tilde{\pi} \longrightarrow \theta \xrightarrow{\tilde{P}_\theta = P_{X|\theta}} X \longrightarrow \hat{\theta}.$$

Then

$$\begin{aligned}
\chi^2(P_{\theta X} \| Q_{\theta X}) &\geq \chi^2(P_{\theta\hat{\theta}} \| Q_{\theta\hat{\theta}}) && \text{data processing inequality} \\
&\geq \chi^2(P_{\theta-\hat{\theta}} \| Q_{\theta-\hat{\theta}}) && \text{data processing inequality} \\
&\geq \frac{(\mathbb{E}_P(\theta - \hat{\theta}) - \mathbb{E}_Q(\theta - \hat{\theta}))^2}{\mathrm{Var}_\pi(\hat{\theta} - \theta)}. && \text{by (??)}
\end{aligned}$$

Let T_δ denote the pushforward of shifting by δ, that is, $T_\delta(P_A) = P_{A+\delta}$. Let us choose

$$Q_\theta = \pi, Q_{X|\theta} = P_\theta, P_\theta = T_\delta \pi, P_{X|\theta} = P_{\theta-\delta},$$

then $P_X = Q_X$ which further indicates $P_{\hat{\theta}} = Q_{\hat{\theta}}$ and the mean of $\hat{\theta}$ under distribution of P equals to the mean under the distribution under Q. Hence $\mathbb{E}_P(\theta - \hat{\theta}) - \mathbb{E}_Q(\theta - \hat{\theta}) = \delta$! For the Bayesian HCR lower bound,

$$R_\pi^* \geq \sup_{\delta \neq 0} \frac{\delta^2}{\chi^2(P_{X\theta} \| Q_{X\theta})} \geq \lim_{\delta\to 0} \frac{\delta^2}{\chi^2(P_{X\theta} \| Q_{X\theta})} = \frac{1}{I(\pi) + \mathbb{E}_{\theta\sim\pi}[I(\theta)]}. \tag{29.9}$$

The last step is justified as follows:

$$\begin{aligned}
\chi^2(P_{X\theta} \| Q_{X\theta}) &= \int \frac{(P_{X\theta} - Q_{X\theta})^2}{Q_{X\theta}} = \int \frac{[P_\theta(P_{X|\theta} - Q_{X|\theta}) + (P_\theta - Q_\theta)Q_{X|\theta}]^2}{Q_{X\theta}} \\
&= \int \frac{P_\theta^2}{Q_\theta} \int \frac{(P_{X|\theta} - Q_{X|\theta})^2}{Q_{X|\theta}} + \int \frac{(P_\theta - Q_\theta)^2}{Q_\theta^2} + 2 \int \frac{P_\theta(P_\theta - Q_\theta)}{Q_\theta} \int (P_{X|\theta} - Q_{X|\theta}) \\
&= \chi^2(P_\theta \| Q_\theta) + \mathbb{E}\left[\chi^2(P_{X|\theta} \| Q_{X|\theta}) \cdot \left(\frac{P_\theta}{Q_\theta}\right)^2\right]
\end{aligned}$$

Then applying

- $\chi^2(P_\theta \| Q_\theta) = \chi^2(T_{\delta\pi} \| \pi) = \delta^2[I(\pi) + o(1)]$ by Taylor expansion,
- $\chi^2(P_{X|\theta} \| Q_{X|\theta}) = [I(\theta) + o(1)]\delta^2$ by Taylor expansion,

we obtain (29.9). □

To end this part, we give a classical proof of the Bayesian Cramér-Rao Lower Bound (cf. [GL95a]):

Alternative Proof of Theorem 29.5. Note that

$$\int \hat{\theta}(x) \frac{\partial}{\partial \theta}(P_\theta(x)\pi(\theta)) \, \mathrm{d}\theta = 0, \tag{29.10}$$

$$\int \theta \frac{\partial}{\partial \theta}(P_\theta(x)\pi(\theta)) \, \mathrm{d}\theta = -\int P_\theta(x)\pi(\theta) \, \mathrm{d}\theta, \tag{29.11}$$

where the first equation follows from the regularity condition, and the second equation follows from integration by part.

Therefore,

$$\begin{aligned}
\mathbb{E}\left[(\hat{\theta}(X) - \theta)\frac{\partial \log(P_\theta(X)\pi(\theta))}{\partial \theta}\right] &= \int \mu(\mathrm{d}x) \int (\hat{\theta}(x) - \theta) \frac{\partial (P_\theta(x)\pi(\theta))}{\partial \theta} \frac{P_\theta(x)\pi(\theta)}{P_\theta(x)\pi(\theta)} \mathrm{d}\theta \\
&= \int \mu(\mathrm{d}x) \int P_\theta(x)\pi(\theta)\mathrm{d}\theta \\
&= 1,
\end{aligned}$$

where the second line follows from (29.10) and (29.11).

By Cauchy-Schwarz inequality,

$$1 = \mathbb{E}\left[(\hat{\theta}(X) - \theta)\frac{\partial \log(P_\theta(X)\pi(\theta))}{\partial \theta}\right] \le \mathbb{E}\left[(\hat{\theta}(X) - \theta)^2\right] \mathbb{E}\left[\left(\frac{\partial \log(P_\theta(X)\pi(\theta))}{\partial \theta}\right)^2\right].$$

Hence

$$\mathbb{E}\left[(\hat{\theta}(X) - \theta)^2\right] \ge \frac{1}{\mathbb{E}\left[\left(\frac{\partial \log P_\theta(X)}{\partial \theta} + \frac{\partial \log \pi(\theta)}{\partial \theta}\right)^2\right]} = \frac{1}{\mathbb{E}[I(\theta)] + I(\pi)}.$$

□

29.6 Information Bound

In this section, we introduce the local version of the minimax lower bound. The local minimax risks is defined in a quadratic form: $\inf_{\hat{\theta}} \sup_{|\theta - \theta_0| \le \epsilon} \mathbb{E}(\hat{\theta} - \theta)^2$. Further, we have

$$\begin{aligned}
\inf_{\hat{\theta}} \sup_{|\theta - \theta_0| \le \epsilon} \mathbb{E}(\hat{\theta} - \theta)^2 &\ge \frac{1}{I(\theta) + n\mathbb{E}_{\theta\sim\pi}[I(\theta)]} \\
&= \frac{1 + o(1)}{n\mathbb{E}_{\theta\sim\pi}[I(\theta)]}
\end{aligned}$$

If $\theta \mapsto I(\theta)$ is continuous, then

$$\mathbb{E}_{\theta\sim\pi}[I(\theta)] = I(\theta_0) + o(1) = \frac{1 + o(1)}{nI(\theta)}.$$

Assume the random variable Z coming from the distribution π, $Z \sim \pi$. Let $I(Z) \triangleq I(\pi)$. For constant $\alpha, \beta \neq 0$, then $I(Z+\alpha) = I(Z)$ and $I(\beta Z) = \frac{I(Z)}{\beta^2}$. If the π has the distribution of form $\cos^2 \frac{\pi x}{2}$, then $\min_{\pi:[-1,1]} I(\pi) = \pi^2$. If the distribution π has the form of $\cos^2 \frac{\pi(x-\theta_0)}{2\epsilon}$, then $I(\theta) = \frac{\pi^2}{\epsilon}$. Then we have

$$\inf_{\hat{\theta}} \sup_{|\theta_0 - \theta| \leq \epsilon} \mathbb{E}(\hat{\theta} - \theta)^2 \geq R_\pi^* \geq \frac{1}{n\mathbb{E}_{\theta\sim\pi}[I(\theta)] + I(\pi)}.$$

Now if we pick $\epsilon = n^{-1/4}$, we have

$$R^* \geq \inf_{\hat{\theta}} \sup_{|\theta - \theta_0| \leq n^{-1/4}} \mathbb{E}_\theta(\theta - \hat{\theta})^2 \geq \frac{1}{nI(\theta) + o(\sqrt{n})} \overset{Optimize}{\Longrightarrow} R^* \geq \frac{1 + o(1)}{n \inf_{\theta_0 \in \Theta} I(\theta_0)}.$$

29.7 Example: Gaussian Location Model (GLM)

Let $X_i = \theta + Z_i$, where $Z_i \sim \mathcal{N}(0,1)$, and $\theta \sim \pi = \mathcal{N}(0,s)$. Given i.i.d. observations $X = (X_1, X_2, \cdots, X_n)$, we have

$$\begin{aligned}
\chi^2(P_{\theta X}||Q_{\theta X}) &= \chi^2(P_{\theta \bar{X}}||Q_{\theta \bar{X}}) \\
&= \chi^2(P_\theta||Q_\theta) + \mathbb{E}_Q\left[\left(\frac{P_\theta}{Q_\theta}\right)^2 \chi^2(P_{\bar{X}|\theta}||Q_{\bar{X}|\theta})\right] \\
&= (e^{\delta^2/s} - 1) + e^{\delta^2/s}(e^{n\delta^2} - 1) \\
&= e^{\delta^2(n+\frac{1}{s})} - 1.
\end{aligned}$$

The first line follows from the fact that $\bar{X}$ is a sufficient statistic ($\theta \to \bar{X} \to X$), and the information processing inequality. The second line follows from Lecture 7 (last equation, Page 5). The third line follows from

$$\chi^2\left(\mathcal{N}(\theta, \sigma^2)||\mathcal{N}(\theta + \delta, \sigma^2)\right) = e^{\delta^2/\sigma^2} - 1.$$

Therefore, by Bayesian HCR and Bayesian Cramér-Rao Lower Bound:

$$R_\pi^* \geq \sup_{\delta \neq 0} \frac{\delta^2}{e^{\delta^2(n+\frac{1}{s})} - 1} = \lim_{\delta \to 0} \frac{\delta^2}{e^{\delta^2(n+\frac{1}{s})} - 1} = \frac{1}{n + \frac{1}{s}} = \frac{s}{sn+1}.$$

In this case, the lower bound is exact! (It has been verified that $R_\pi^* = \frac{s}{sn+1}$.) The minimax lower bound is $R^* \geq \sup_s R_\pi^* = \frac{1}{n}$.

29.8 An Alternative Information Inequality

If we choose a uniform prior in Theorem **??**, the resulting lower bound is zero since the Fisher information of uniform distribution is infinity. Nevertheless, it is possible to obtain an alternative information inequality involving $\mathbb{E}_{\theta\sim\text{uniform}}[I(\theta)]$; however, it should be pointed out that the lower bound applies to the minimax risk (not Bayes risk with respect to uniform prior) since the proof in act involves two prior: uniform on the interval and uniform over the two endpoints.

Theorem 29.6. *Assume the usual regularity condition:*

$$\int \frac{\partial p_\theta}{\partial x} dx = 0.$$

Then

$$R^* = \inf_{\hat{\theta}} \sup_{\theta \in [\theta_0-\epsilon, \theta_0+\epsilon]} \mathbb{E}_\theta[(\theta - \hat{\theta})^2] \geq \frac{1}{(\epsilon^{-1} + \sqrt{n\bar{I}})^2}$$

where $\overline{I}$ *denotes the average Fisher information:*

$$\overline{I} = \frac{1}{2\epsilon} \int_{\theta_0-\epsilon}^{\theta_0+\epsilon} I(\theta) \, \mathrm{d}\theta.$$

Proof. See Problem 2 in Homework 1. □

Remark 29.2. Theorem 29.6 is a strict improvement of the inequality of Chernoff-Rubin-Stein:[2]

$$\inf_{\hat{\theta}} \sup_{\theta \in [\theta_0-\epsilon, \theta_0+\epsilon]} \mathbb{E}_\theta[(\theta - \hat{\theta})^2] \geq \max_{0<\delta<1} \min \left\{ \frac{\delta^2}{4}, \frac{1-\epsilon}{n\bar{I}} \right\} = \frac{1}{(\epsilon^{-1} + \sqrt{n\bar{I}+1})^2}.$$

Both this and Theorem 29.6 suffice to prove the optimal minimax lower bound.

29.9 Maximum Likelihood Estimator (MLE) and asymptotic efficiency

We *sketch* the analysis of MLE in the classical large-sample asymptotics. Let $X = (X_1, X_2, \cdots, X_n) \stackrel{\text{i.i.d.}}{\sim} P_{\theta_0}$, define maximum likelihood estimator:

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta \in \Theta} L_\theta(X),$$

where

$$L_\theta(X) = \log P_\theta^{\otimes n}(X) = \sum_{i=1}^n \log P_\theta(X_i).$$

Intuition:

$$\mathbb{E}_{\theta_0}\left[L_\theta(X) - L_{\theta_0}(X)\right] = \mathbb{E}_{\theta_0}\left[\sum_{i=1}^n \log \frac{P_\theta(X_i)}{P_{\theta_0}(X_i)}\right] = -nD(P_{\theta_0}||P_\theta) \leq 0.$$

So as long as $\theta_0 \neq \theta$, $L_\theta(X) - L_{\theta_0}(X)$ is a random walk with negative drift. From here the consistency of MLE follows upon assuming appropriate regularity conditions.

Assuming more conditions one can obtain asymptotic normality and $\sqrt{n}$-consistency of MLE. Next, we derive a local quadratic approximation of the log-likelihood function. By Taylor expansion,

$$L_\theta(X) = L_{\theta_0}(X) + \sum_{i=1}^n \left.\frac{\partial \log P_\theta(X_i)}{\partial \theta}\right|_{\theta=\theta_0} (\theta - \theta_0) + \frac{1}{2} \sum_{i=1}^n \left.\frac{\partial^2 \log P_\theta(X_i)}{\partial \theta^2}\right|_{\theta=\theta_0} (\theta - \theta_0)^2 + o((\theta - \theta_0)^2). \tag{29.12}$$

Recall that

$$\mathbb{E}\left[\frac{\partial \log P_\theta(X_i)}{\partial \theta}\right] = 0, \quad \mathbb{E}\left[\left(\frac{\partial \log P_\theta(X_i)}{\partial \theta}\right)^2\right] = -\mathbb{E}\left[\frac{\partial^2 \log P_\theta(X_i)}{\partial \theta^2}\right] = I(\theta).$$

[2] This is given in [**?**, Lemma 1] without proof, which Chernoff credited to Rubin and Stein.

By the Central Limit Theorem,

$$\frac{1}{\sqrt{nI(\theta_0)}} \sum_{i=1}^{n} \frac{\partial \log P_\theta(X_i)}{\partial \theta} \xrightarrow{\text{d.}} \mathcal{N}(0,1).$$

By the Weak Law of Large Numbers,

$$\sum_{i=1}^{n} \frac{\partial^2 \log P_\theta(X_i)}{\partial \theta^2} = -nI(\theta_0) + o_P(n).$$

Substituting these quantities into (29.12), we obtain a local quadratic approximation of the log-likelihood function:

$$L_\theta(X) \approx L_{\theta_0}(X) + \sqrt{nI(\theta_0)} \cdot Z \cdot (\theta - \theta_0) - \frac{1}{2} nI(\theta_0)(\theta - \theta_0)^2,$$

where $Z \sim \mathcal{N}(0,1)$. Maximizing the right-hand side, we obtain:

$$\hat{\theta}_{\mathrm{MLE}} \approx \theta_0 + \frac{Z}{\sqrt{nI(\theta_0)}}.$$

Therefore, MLE achieves the locally minimax lower bound $R^* \geq \frac{1+o(1)}{nI(\theta_0)}$ (see Section 7.5 in Lecture 7).

Remark 29.3. The general asymptotic theory of MLE and achieving information bound is due to Hájek and LeCam.

29.10 Bayesian Lower Bounds for Functional Estimation

Next, we derive the Bayesian Cramér-Rao lower bound for functional estimation $\widehat{T}(X)$.

Theorem 29.7. *Let $T : \mathbb{R}^p \to \mathbb{R}$, and*

$$\begin{array}{ccc} \theta & \rightarrow & X \\ \downarrow & & \downarrow \\ T(\theta) & & \widehat{T}(X) \end{array}$$

Then we have

$$R_\pi^* \geq (\nabla T)' I^{-1} \nabla T.$$

Proof. By similar arguments in previous lectures,

$$\chi^2(P_{\theta X} || Q_{\theta X}) \geq \chi^2(P_{T-\widehat{T}} || Q_{T-\widehat{T}}) \geq \frac{\left(\mathbb{E}_P[T - \widehat{T}] - \mathbb{E}_Q[T - \widehat{T}]\right)^2}{\mathrm{Var}_Q[T - \widehat{T}]}. \tag{29.13}$$

Let $Q(\theta) = \pi(\theta)$, and $P(\theta) = \pi(\theta - \epsilon u)$, where $u \in \mathbb{R}^p$. In order to make the marginal distribution of $P_X = Q_X$, let $P_\theta(x) = Q_{\theta - \epsilon u}(x)$. Hence the numerator and the denominator in (29.13) satisfy:

$$\begin{aligned} \left(\mathbb{E}_P[T - \widehat{T}] - \mathbb{E}_Q[T - \widehat{T}]\right)^2 &= (\mathbb{E}_P[T] - \mathbb{E}_Q[T])^2 \\ &= \left(\int \pi(\theta) T(\theta + \epsilon u)\, \mathrm{d}\theta - \int \pi(\theta) T(\theta)\, \mathrm{d}\theta\right)^2 \\ &= \left(\int \pi(\theta) \langle \nabla T, \epsilon u \rangle + o(\epsilon)\right)^2 \\ &= \epsilon^2 \langle \mathbb{E}_\pi \nabla T, u \rangle^2 + o(\epsilon^2), \end{aligned} \tag{29.14}$$

$$\mathrm{Var}_Q[T-\widehat{T}] \le \mathbb{E}_Q[(T-\widehat{T})^2] = R_\pi. \tag{29.15}$$

The left-hand side of (29.13) satisfies

$$\begin{aligned}
\chi^2(P_{\theta X}||Q_{\theta X}) &= \chi^2(P_\theta||Q_\theta) + \mathbb{E}_Q\left[\chi^2(P_{X|\theta}||Q_{X|\theta})\left(\frac{P_\theta}{Q_\theta}\right)^2\right] \\
&= \int \frac{(\pi(\theta-\epsilon u)-\pi(\theta))^2}{\pi(\theta)}\mathrm{d}\theta + \mathbb{E}_\pi\left[\int \frac{(Q_{\theta-\epsilon u}(x)-Q_\theta(x))^2}{Q_\theta(x)}\mathrm{d}x\left(\frac{\pi(\theta-\epsilon u)}{\pi(\theta)}\right)^2\right] \\
&= \int \frac{\epsilon^2 u'(\nabla\pi)(\nabla\pi)'u}{\pi(\theta)}\mathrm{d}\theta + \mathbb{E}_\pi\left[\int \frac{\epsilon^2 u'(\nabla_\theta Q)(\nabla_\theta Q)'u}{Q_\theta(x)}\mathrm{d}x\right] + o(\epsilon^2) \\
&= \epsilon^2 u'\left(I(\pi)+\mathbb{E}_\pi[I(\theta)]\right)u + o(\epsilon^2).
\end{aligned} \tag{29.16}$$

Substituting (29.14), (29.15), and (29.16) into (29.13), we have

$$R_\pi^* \ge \frac{\langle \mathbb{E}_\pi \nabla T, u\rangle^2}{u'\left(I(\pi)+\mathbb{E}_\pi[I(\theta)]\right)u}$$

Locally, $\mathbb{E}_\pi \nabla T(\theta) \approx \nabla T(\theta_0)$, and $I(\pi)+\mathbb{E}_\pi[I(\theta)] \approx I(\theta_0)$. Hence

$$\boxed{R_\pi^* \ge \sup_u \frac{\langle \nabla T(\theta_0), u\rangle^2}{u' I(\theta_0) u} = (\nabla T(\theta_0))' I^{-1}(\theta_0)\nabla T(\theta_0).}$$

The maximum is attained when $u = I^{-1}(\theta_0)\nabla T(\theta_0)$.[3] □

Remark 29.4. The maximum likelihood estimator satisfies $T(\hat\theta_{\mathrm{MLE}}) = T(\theta_0 + \frac{1}{\sqrt{n}}Z)$, where $Z \sim \mathcal{N}(0, I^{-1}(\theta_0))$. Hence

$$T(\hat\theta_{\mathrm{MLE}}) \sim N\left(T(\theta_0), \frac{1}{n}(\nabla T(\theta_0))' I^{-1}(\theta_0)(\nabla T(\theta_0))\right).$$

The maximum likelihood estimator again asymptotically achieves the locally minimax lower bound.

29.11 Example: Classical asymptotics of entropy estimation

Corollary 29.1. *Let $X_1, \cdots, X_n \overset{i.i.d.}{\sim} p \in \mathcal{M}_k$, where $\mathcal{M}_k$ denotes the set of probability distributions over $[k] = \{1, \dots, k\}$. Then the minimax quadratic risk of entropy estimation satisfies*

$$R^* = \inf_{\widehat{H}} \sup_{P\in\mathcal{M}_k} \mathbb{E}[(\widehat{H}-H)^2] = \frac{1}{n}\left(\max_{p\in\mathcal{M}_k} V(p) + o(1)\right), \quad n\to\infty$$

where

$$\begin{aligned}
H(p) &= \sum_{i=1}^k p_i \log\frac{1}{p_i} = \mathbb{E}\left[\log\frac{1}{p(X)}\right], \\
V(p) &= \mathrm{Var}\left(\log\frac{1}{p(X)}\right)
\end{aligned}$$

[3]This can be shown, for example, by letting $\tilde u = I^{-\frac{1}{2}}(\theta_0)u$.

Note: $\max_{p\in\mathcal{M}_k} V(p) \le \log^2 k$ for all $k \ge 3$ (see [PPV10a, Eq. (464)]).

Proof. We have $H : \Theta \to \mathbb{R}^+$, where $\theta = (p_1, p_2, \cdots, p_{k-1})$ (recall that $p_k = 1 - p_1 - \cdots - p_{k-1}$.) Therefore,

$$\frac{\partial H}{\partial p_i} = \log \frac{p_k}{p_i}, \quad i = 1, 2, \cdots, k-1.$$

Next, we compute the Fisher Information matrix:

$$I(\theta)_{ij} = -\mathbb{E}\left[\frac{\partial^2 \log p(X)}{\partial p_i \partial p_j}\right] = \begin{cases} \frac{1}{p_i} + \frac{1}{p_k} & \text{if } i = j \\ \frac{1}{p_k} & \text{if } i \neq j \end{cases}.$$

Therefore,

$$I(\theta) = \begin{bmatrix} \frac{1}{p_1} & & \\ & \ddots & \\ & & \frac{1}{p_{k-1}} \end{bmatrix} + \frac{1}{p_k}\mathbf{1}\mathbf{1}'.$$

By the Matrix Inversion Lemma,[4] we have

$$I^{-1}(\theta) = \begin{bmatrix} p_1 & & \\ & \ddots & \\ & & p_{k-1} \end{bmatrix} + \begin{bmatrix} p_1 \\ \vdots \\ p_{k-1} \end{bmatrix} \begin{bmatrix} p_1 & \cdots & p_{k-1} \end{bmatrix}.$$

Therefore,

$$\begin{aligned}
\nabla H' I^{-1}(\theta) \nabla H &= \sum_{i=1}^{k-1} p_i \log^2 \frac{p_k}{p_i} - \left(\sum_{i=1}^{k-1} p_i \log \frac{p_k}{p_i}\right)^2 \\
&= \sum_{i=1}^{k} p_i \log^2 \frac{1}{p_i} + \log^2 \frac{1}{p_k} - 2\sum_{i=1}^{k} p_i \log \frac{1}{p_i} \log \frac{1}{p_k} - \left(\left(\sum_{i=1}^{k} p_i \log \frac{1}{p_i}\right) - \log \frac{1}{p_k}\right)^2 \\
&= \sum_{i=1}^{k} p_i \log^2 \frac{1}{p_i} - \left(\sum_{i=1}^{k} p_i \log \frac{1}{p_i}\right)^2 \\
&= \mathbb{E}\left[\log^2 \frac{1}{p(X)}\right] - \left(\mathbb{E}\left[\log \frac{1}{p(X)}\right]\right)^2 = \mathrm{Var}\left[\log \frac{1}{p(X)}\right] = V(p).
\end{aligned}$$

Given n samples, the Fisher Information matrix is $nI(\theta)$. By Theorem 29.7,

$$R^* \ge \frac{1 + o(1)}{n} \nabla H' I^{-1}(\theta) \nabla H = \frac{1 + o(1)}{n} V(p). \qquad \square$$

[4] $(A + UCV)^{-1} = A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}$.

§ 30. Applications to statistical decision theory

In this lecture we discuss applications of information theory to statistical decision theory. Although this lecture only focuses on statistical lower bound (converse result), let us remark in passing that the impact of information theory on statistics is far from being only on proving impossibility results. Many procedures are based on or inspired by information-theoretic ideas, e.g., those based on metric entropy, pairwise comparison, maximum likelihood estimator and analysis, minimum distance estimator (Wolfowitz), maximum entropy estimators, EM algorithm, minimum description length (MDL) principle, etc.

We discuss two methods: LeCam-Fano (hypothesis testing) method and the rate-distortion (*mutual information*) method.

We begin with the decision-theoretic setup of statistical estimation. The general paradigm is the following:

$$\underbrace{\theta}_{\text{parameter}} \to \underbrace{X}_{\text{data}} \to \underbrace{\hat{\theta}}_{\text{estimator}}$$

The main ingredients are

- Parameter space: $\Theta \ni \theta$
- Statistical model: $\{P_{X|\theta} : \theta \in \Theta\}$, which is a collection of distributions indexed by the parameter
- Estimator: $\hat{\theta} = \hat{\theta}(X)$
- Loss function: $\ell(\theta, \hat{\theta})$ measures the inaccuracy.

The goal is make random variable $\ell(\theta, \hat{\theta})$ small either in probability or in expectation, uniformly over the unknown parameter θ. To this end, we define the **minimax risk**

$$R^* = \inf_{\hat{\theta}} \sup_{\theta \in \Theta} \mathbb{E}_\theta[\ell(\theta, \hat{\theta})].$$

Here $\mathbb{E}_\theta$ denotes averaging with respect to the randomness of $X \sim P_\theta$.

Ideally we want to compute R^* and find the minimax optimal estimator that achieves the minimax risk. This tasks can be very difficult especially in high dimensions, in which case we will be content with characterizing the minimax rate, which approximates R^* within multiplicative universal constant factors, and the estimator that achieves a constant factor of R^* will be called rate-optimal.

As opposed to the worst-case analysis of the minimax risk, the Bayes approach is an average-case analysis by considering the average risk of an estimator over all $\theta \in \Theta$. Let the prior π be a probability distribution on Θ, from which the parameter θ is drawn. Then, the **average risk** (w.r.t π) is defined as

$$R_\pi(\hat{\theta}) = \mathbb{E}_{\theta \sim \pi} R_\theta(\hat{\theta}) = \mathbb{E}_{\theta, X} \ell(\theta, \hat{\theta}).$$

The **Bayes risk** for a prior π is the minimum that the average risk can achieve, i.e.

$$R_\pi^* = \inf_{\hat\theta} R_\pi(\hat\theta).$$

By the simple logic of "maximum $\geq$ average", we have

$$R^* \geq R_\pi^* \tag{30.1}$$

and in fact $R^* = \sup_{\pi \in \mathcal{M}(\Theta)} R_\pi^*$ whenever the minimax theorem holds, where $\mathcal{M}(\Theta)$ denotes the collection of all probability distributions on Θ. In other words, solving the minimax problem can be done by finding the least-favorable (Bayesian) prior. Almost all of the minimax lower bounds boil down to bounding from below the Bayes risk for some prior. When this prior is uniform on just two points, the method is known under a special name of (two-point) LeCam or LeCam-Fano method.

Note also that when $\ell(\theta, \hat\theta) = \|\theta - \hat\theta\|_2^2$ is the quadratic ℓ_2 risk, the optimal estimator achieving R_π^* is easy to describe: $\hat\theta^* = \mathbb{E}[\theta|X]$. This fact, however, is of limited value, since typically conditional expectation is very hard to analyze.

30.1 Fano, LeCam and minimax risks

We demonstrate the LeCam-Fano method on the following example:

- Parameter space $\theta \in [0, 1]$
- Observation model X_i – i.i.d. Bern(θ)
- Quadratic loss function:
$$\ell(\hat\theta, \theta) = (\hat\theta - \theta)^2$$
- Fundamental limit:
$$R^*(n) \triangleq \sup_{\theta_0 \in [0,1]} \inf_{\hat\theta} \mathbb{E}[(\hat\theta(X^n) - \theta)^2 | \theta = \theta_0]$$

A natural estimator to consider is the empirical mean:

$$\hat\theta_{emp}(X^n) = \frac{1}{n} \sum_i X_i$$

It achieves the loss

$$\sup_{\theta_0} \mathbb{E}[(\hat\theta_{emp} - \theta)^2 | \theta = \theta_0] = \sup_{\theta_0} \frac{\theta_0(1-\theta_0)}{n} = \frac{1}{4n}\,. \tag{30.2}$$

The question is how close this is to the optimal.

First, recall the *Cramer-Rao lower bound*: Consider an arbitrary statistical estimation problem $\theta \to X \to \hat\theta$ with $\theta \in \mathbb{R}$ and $P_{X|\theta}(dx|\theta_0) = f(x|\theta)\mu(dx)$ with $f(x|\theta)$ is differentiable in θ. Then for any $\hat\theta(x)$ with $\mathbb{E}[\hat\theta(X)|\theta] = \theta + b(\theta)$ and smooth $b(\theta)$ we have

$$\mathbb{E}[(\hat\theta - \theta)^2 | \theta = \theta_0] \geq b(\theta_0)^2 + \frac{(1 + b'(\theta_0))^2}{J_F(\theta_0)}\,, \tag{30.3}$$

where $J_F(\theta_0) = \mathrm{Var}[\frac{\partial \ln f(X|\theta)}{\partial \theta} | \theta = \theta_0]$ is the Fisher information (5.4). In our case, for any *unbiased* estimator (i.e. $b(\theta) = 0$) we have

$$\mathbb{E}[(\hat\theta - \theta)^2 | \theta = \theta_0] \geq \frac{\theta_0(1-\theta_0)}{n}\,,$$

and we can see from (30.2) that $\hat{\theta}_{emp}$ is optimal in the class of unbiased estimators.

Can biased estimators do better? The answer is yes. Consider

$$\hat{\theta}_{bias} = \frac{1-\epsilon_n}{n}\sum_i (X_i - \frac{1}{2}) + \frac{1}{2},$$

where choice of $\epsilon_n > 0$ "shrinks" the estimator towards $\frac{1}{2}$ and regulates the *bias-variance* tradeoff. In particular, setting $\epsilon_n = \frac{1}{\sqrt{n}+1}$ achieves the minimax risk

$$\sup_{\theta_0} \mathbb{E}[(\hat{\theta}_{bias} - \theta)^2 | \theta = \theta_0] = \frac{1}{4(\sqrt{n}+1)^2}, \tag{30.4}$$

which is better than the empirical mean (30.2), but only slightly.

How do we show that arbitrary biased estimators can not do significantly better? This is where LeCam-Fano method comes handy. Suppose some estimator $\hat{\theta}$ achieves

$$\mathbb{E}[(\hat{\theta} - \theta)^2 | \theta = \theta_0] \le \Delta_n^2 \tag{30.5}$$

for all θ_0. Then, setup the following probability space:

$$W \to \theta \to X^n \to \hat{\theta} \to \hat{W}$$

- $W \sim \text{Bern}(1/2)$
- $\theta = 1/2 + \kappa(-1)^W \Delta_n$ where $\kappa > 0$ is to be specified later
- X^n is i.i.d. $\text{Bern}(\theta)$
- $\hat{\theta}$ is the given estimator
- $\hat{W} = 0$ if $\hat{\theta} > 1/2$ and $\hat{W} = 1$ otherwise

The idea here is that we use our high-quality estimator to distinguish between two hypotheses $\theta = 1/2 \pm \kappa\Delta_n$. Notice that for probability of error we have:

$$\mathbb{P}[W \ne \hat{W}] = \mathbb{P}[\hat{\theta} > 1/2 | \theta = 1/2 - \kappa\Delta_n] \le \frac{\mathbb{E}[(\hat{\theta}-\theta)^2]}{\kappa^2\Delta_n^2} \le \frac{1}{\kappa^2}$$

where the last steps are by Chebyshev and (30.5), respectively. Thus, from Fano's inequality Theorem 6.3 we have

$$I(W;\hat{W}) \ge \left(1 - \frac{1}{\kappa^2}\right)\log 2 - h(\kappa^{-2}).$$

On the other hand, from data-processing and golden formula we have

$$I(W;\hat{W}) \le I(\theta; X^n) \le D(P_{X^n|\theta} \| \text{Bern}(1/2)^n | P_\theta)$$

Computing the last divergence we get

$$D(P_{X^n|\theta} \| \text{Bern}(1/2)^n | P_\theta) = nd(1/2 - \kappa\Delta_n \| 1/2) = n(\log 2 - h(1/2 - \kappa\Delta_n))$$

As $\Delta_n \to 0$ we have

$$h(1/2 - \kappa\Delta_n) = \log 2 - 2\log e \cdot (\kappa\Delta_n)^2 + o(\Delta_n^2).$$

So altogether, we get that for every fixed κ we have

$$\left(1-\frac{1}{\kappa^2}\right)\log 2 - h(\kappa^{-2}) \le 2n\log e\cdot(\kappa\Delta_n)^2 + o(n\Delta_n^2)\,.$$

In particular, by optimizing over κ we get that for some constant $c \approx 0.015 > 0$ we have

$$\Delta_n^2 \ge \frac{c}{n} + o(1/n)\,.$$

Together with (30.2), we have

$$\frac{0.015}{n} + o(1/n) \le R^*(n) \le \frac{1}{4n}\,,$$

and thus the empirical-mean estimator is *rate-optimal*.

We mention that for this particular problem (estimating mean of Bernoulli samples) the minimax risk is known exactly:

$$R^*(n) = \frac{1}{4(1+\sqrt{n})^2} \tag{30.6}$$

but obtaining this requires different methods.[1] In fact, even showing $R^*(n) = \frac{1}{4n} + o(1/n)$ requires careful priors on θ (unlike the simple two-point prior we used above).[2]

We demonstrated here the essense of the *Fano method* of proving lower (impossibility) bounds in statistical decision theory. Namely, given an estimation task we select a prior, uniform on finitely many θ's, which on one hand yields a rather small information $I(\theta; X)$ and on the other hand has sufficiently separated points which thus should be distinguishable by a good estimator. For more see [Yu97].

A natural (and very useful) generalization is to consider non-discrete prior P_θ, and use the following natural chain of inequalities

$$f(P_\theta, R) \le I(\theta;\hat\theta) \le I(\theta; X^n) \le \sup_{P_\theta} I(\theta; X^n)\,,$$

where

$$f(P_\theta, R) \triangleq \inf\{I(\theta;\hat\theta) : P_{\hat\theta|\theta} \text{ s.t. } \mathbb{E}[\ell(\theta,\hat\theta)] \le R\}$$

is the rate-distortion function. This method we discuss next.

[1] The easiest way to get this is to apply (30.1). . Fortunately, in this case if π is the β-distribution, computation of conditional expectation can be performed in closed form, and optimizing parameters of the β-distribution one recovers a lower bound that together with (30.4) establishes (30.6). Note that the resulting worst-case π is not uniform, and in fact $\beta \to \infty$ (i.e. π concentrates in a small region around $\theta = 1/2$).

[2] It follows from the following *Bayesian Cramer-Rao lower bound [GL95b]*: For any estimator $\hat\theta$ and for any prior $\pi(\theta)d\theta$ with smooth density π we have

$$\mathbb{E}_{\theta\sim\pi}[(\hat\theta(X)-\theta)^2] \ge \frac{(\log e)^2}{\mathbb{E}[J_F(\theta)] + J_F(\pi)}\,,$$

where $J_F(\theta)$ is as in (30.3), $J_F(\pi) \triangleq (\log e)^2 \int \frac{(\pi'(\theta))^2}{\pi(\theta)} d\theta$. Then taking π supported on a $n^{-1/4}$-neighborhood surrounding a given point θ_0 we get that $\mathbb{E}[J_F(\theta)] = \frac{n}{\theta_0(1-\theta_0)} + o(n)$ and $J_F(\pi) = o(n)$, yielding

$$R^*(n) \ge \frac{\theta_0(1-\theta_0)}{n} + o(1/n)\,.$$

This is a rather general phenomenon: Under regularity assumptions in any iid estimation problem $\theta \to X^n \to \hat\theta$ *with quadratic loss* we have

$$R^*(n) = \frac{1}{\inf_\theta J_F(\theta)} + o(1/n)\,.$$

30.2 Mutual information method

The main workhorse will be

1. Data processing inequality
2. Rate-distortion theory
3. Capacity and mutual information bound

To illustrate the mutual information method and its execution in various problems, we will discuss three vignettes:

- Denoise a vector;
- Denoise a sparse vector;
- Community detection.

Here's the main idea of the mutual information method. Fix some prior π and we turn to lower bound R^*_π. The unknown θ is distributed according to π. Let $\hat{\theta}$ be a Bayes optimal estimator that achieves the Bayes risk R^*_π.

The mutual information method consists of applying the data processing inequality to the Markov chain $\theta \to X \to \hat{\theta}$:

$$\inf_{P_{\tilde{\theta}|\theta}:\mathbb{E}\ell(\theta,\tilde{\theta})\le R^*_\pi} I(\theta;\hat{\theta}) \le I(\theta,\hat{\theta}) \overset{\text{dpi}}{\le} I(\theta;X). \tag{30.7}$$

Note that

- The leftmost quantity can be interpreted as the minimum amount of information required for an estimation task, which is reminiscent of rate-distortion function.
- The rightmost quantity can be interpreted as the amount of information provided by the data about the parameter. Sometimes it suffices to further upper-bound it by capacity of the channel $\theta \mapsto X$:
$$I(\theta;X) \le \sup_{\pi\in\mathcal{M}(\Theta)} I(\theta;X). \tag{30.8}$$
- This chain of inequalities is reminiscent of how we prove the converse in joint-source channel coding (Section 27.3), with the capacity-like upper bound and rate-distortion-like lower bound.
- Only the lower bound is related to the loss function.
- Sometimes we need a smart choice of the prior.

30.2.1 Denoising (Gaussian location model)

The setting is the following: given n noisy observations of a high-dimensional vector $\theta \in \mathbb{R}^p$,

$$X_i \overset{\text{i.i.d.}}{\sim} N(\theta, I_p), \quad i = 1, \ldots, n \tag{30.9}$$

The loss is simply the quadratic error: $\ell(\theta, \hat{\theta}) = \|\theta - \hat{\theta}\|_2^2$. Next we show that

$$R^* = \frac{p}{n}, \quad \forall p, n. \tag{30.10}$$

Upper bound. Consider the estimator $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$. Then $\bar{X} \sim N(\theta, \frac{1}{n}I_p)$ and clearly $\mathbb{E}\|\bar{X} - \theta\|_2^2 = p/n$.

Lower bound. Consider a Gaussian prior $\theta \sim \mathcal{N}(0, \sigma^2 I_p)$. Instead of evaluating the exact Bayes risk (MMSE) which is a simple exercise, let's implement the mutual information method (30.7). Given any estimator $\hat{\theta}$. Let $D = \mathbb{E}\|\hat{\theta} - \theta\|_2^2$. Then

$$\frac{p}{2}\log\frac{\sigma^2}{D/p} = \inf_{P_{\tilde{\theta}|\theta}:\mathbb{E}\|\theta-\tilde{\theta}\|_2^2 \le D} I(\theta; \hat{\theta}) \le I(\theta, \hat{\theta}) \le I(\theta; X) \overset{\text{suff stat}}{=} I(\theta; \bar{X}) = \frac{p}{2}\log\left(1 + \frac{\sigma^2}{1/n}\right).$$

where the left inequality follows from the Gaussian rate-distortion function (27.3) and the single-letterization result (Theorem 26.1) that reduces p dimensions to one dimension. Putting everything together we have

$$R^* \ge R_\pi^* \ge \frac{p\sigma^2}{1 + n\sigma^2}.$$

Optimizing over σ^2 (by sending it to ∞), we have $R^* \ge p/n$.

30.2.2 Denoising sparse vectors

Here the setting is identical to (30.9), expect that we have the prior knowledge that θ is *sparse*, i.e.,

$$\theta \in \Theta \triangleq \{\text{all } p\text{-dim } k\text{-sparse vectors}\} = \{\theta \in \mathbb{R}^p : \|\theta\|_0 \le k\}$$

where $\|\theta\|_0 = \sum_{i\in[p]} \mathbf{1}_{\{\theta_i \ne 0\}}$ is the sparisty (number of nonzeros) of θ.

The minimax rate of denoising k-sparse vectors is given by the following

$$R^* \asymp \frac{k}{n}\log\frac{ep}{k}, \quad \forall k, p, n. \tag{30.11}$$

Before proceeding to the proof, a quick observation is that we have the oracle lower bound $R^* \ge \frac{k}{n}$ follows from (30.10), since if the support of the θ is known which reduces the problem to k dimensions. Thus, the meaning of statement (30.11) is that the lack of knowledge of the support contributes (merely) a log factor.

To show this, again, by passing to sufficient statistics, it suffices to consider the observation $X \sim N(\theta, \frac{1}{n}I_p)$. For simplicity we only consider $n = 1$ below.

Upper bound. (Sketch) The rate is achieved by thresholding the observation X that only keep the large entries. The intuition is that since the ground truth θ has many zeros, we should kill the small entries in X. Since $\|Z\|_\infty \le (2+\epsilon)\sqrt{\log p}$ with high probability, hard thresholding estimator that sets all entries of X with magnitude $\le (2+\epsilon)\sqrt{\log p}$ achieves a mean-square error of $O(k\log p)$, which is rate optimal unless $k = \Omega(p)$, in which case we can simply apply the original X as the estimator.

Lower bound. In view of the oracle lower bound, it suffices to consider $k = O(p)$. Next we assume $k \le p/16$. Consider a p-dimensional Hamming sphere of radius k, i.e.

$$B = \{b \in \{0,1\}^p : w_H(b) = k\},$$

where $w_H(b)$ is the Hamming weights of b. Let b be drawn uniformly from the set B and $\theta = \tau b$, where $\tau = \sqrt{\frac{k}{100}\log\frac{p}{k}}$. Thus, we have the following Markov chain which represents our problem model,

$$b \to \theta \to X \to \hat{\theta} \to \hat{b}.$$

Note that the channel $\theta \to X$ is just p uses of the AWGN channel, with power $\frac{\tau^2 k}{p}$, and thus by Theorem 5.6 and single-letterization (Theorem 6.1) we have

$$I(\theta;\hat{\theta}) \le I(\theta;X) \le \frac{p}{2}\log\left(1+\frac{\tau^2 k}{p}\right) \le \sup_{\theta\in G}\frac{\log e}{2}\|\theta\|_2^2 = ck\tau^2\,,$$

for some $c > 0$. We note that related techniques have been used in proving lower bound for stable recovery in noiseless compressed sensing [PW12].

To give a lower bound for $I(\theta;\hat{\theta})$, consider

$$\hat{b} = \operatorname*{argmin}_{b\in B}\|\hat{\theta}-\tau b\|_2^2.$$

Since $\hat{b}$ is the minimizer of $\|\hat{\theta}-\tau b\|_2^2$, we have,

$$\|\tau\hat{b}-\theta\|_2 \le \|\tau\hat{b}-\hat{\theta}\|_2 + \|\theta-\hat{\theta}\|_2 \le 2\|\theta-\hat{\theta}\|_2.$$

Thus,

$$\tau^2 d_H(b,\hat{b}) = \|\tau\hat{b}-\theta\|_2^2 \le 4\|\theta-\hat{\theta}\|_2^2,$$

where d_H denotes the Hamming distance between b and $\hat{b}$. Suppose that $\mathbb{E}\|\hat{\theta}-\theta\|_2^2 = \epsilon\tau^2 k$. Then we have $\mathbb{E}d_H(b,\hat{b}) \le 4\epsilon k$. Our goal is to show that ϵ is at least a small constant by the mutual information method. First,

$$I(\hat{b};b) \ge \min_{\mathbb{E}d_H(b,\hat{b})\le 4\epsilon k} I(\hat{b};b).$$

Before we bound the RHS, let's first guess its behavior. Note that it is the rate-distortion function of the random vector b, which is uniform over B, the Hamming sphere of radius k, and each entry is Bern(k/p). Had the entries been iid, then rate-distortion theory ((27.1) and Theorem 26.1) would yield that the RHS is simply $p(h(k/p) - h(4\epsilon k/p))$. Next, following the proof of (27.1), we show that this behavior is indeed correct:

$$\begin{aligned}\min_{\mathbb{E}d_H(b,\hat{b})\le 4\epsilon k} I(\hat{b};b) &= H(b) - \max_{\mathbb{E}d_H(b,\hat{b})\le 4\epsilon k} H(b|\hat{b})\\ &= \log\binom{p}{k} - \max_{\mathbb{E}d_H(b,\hat{b})\le 4\epsilon k} H(b\oplus\hat{b}|\hat{b})\\ &\ge \log\binom{p}{k} - \max_{\mathbb{E}w_H(W)\le 4\epsilon k} H(W).\end{aligned}$$

The maximum-entropy problem is easy to solve:

$$\max_{\mathbb{E}w_H(W)=m, W\in\{0,1\}^p} H(W) = ph\left(\frac{m}{p}\right). \tag{30.12}$$

The solution is $W = \text{Bern}(m/p)^{\otimes p}$. One way to get this is to write $H(W) = p\log 2 - D(P_W \| \text{Bern}(1/2)^{\otimes p})$ and apply Theorem 14.3 with $X = w_H(W)$, to get that optimal $P_W(w) \sim \exp\{cw_H(w)\}$. In the end we get Combine this with the previous bound, we get

$$I(\hat{b}; b) \geq \log \binom{p}{k} - ph(\frac{4\epsilon k}{p}).$$

On the other hand, we have

$$I(\hat{b}; b) \leq I(\theta; Y) \leq c\tau^2 = c' k \log \frac{p}{k}.$$

Note that $h(\alpha) \asymp -\alpha \log \alpha$ for $\alpha < \frac{1}{4}$. WLOG, since $k \leq \frac{p}{16}$, we have $\epsilon \geq c_0$ for some universal constant c_0. Therefore

$$R^* \geq \epsilon\tau^2 k \gtrsim k \log \frac{p}{k}.$$

Combining with the result in the oracle lower bound, we have the desired.

$$R^* \gtrsim k + k \log \frac{p}{k}$$

or for general $n \geq 1$

$$R^* \gtrsim \frac{k}{n} \log \frac{ep}{k}.$$

Remark 30.1. Let $R^*_{k,p} = R^*$. For the case $k = o(p)$, the sharp asymptotics is

$$R^*_{k,p} \geq (2 + o_p(1)) k \log \frac{p}{k}.$$

To prove this result, we need to first show that for the case $k = 1$,

$$R^*_{1,p} \geq (2 + o_p(1)) \log p.$$

Next, show that for any k, the minimax risk is lower bounded by the Bayesian risk with the block prior. The block prior is that we divide the p-coordinate into k blocks, and pick one coordinate from each p/k-coordinate uniformly. With this prior, one can show

$$R^*_{k,p} \geq k R^*_{1,p/k} = (2 + o_p(1)) k \log \frac{p}{k}.$$

30.2.3 Community detection

We only consider the problem of a single hidden community. Given a graph of n vertices, a community is a subset of vertices where the edges tend to be denser than everywhere else. Specifically, we consider the *planted dense subgraph model* (i.e., the stochastic block model with a single community). Let the community C be uniformly drawn from all subsets of $[n]$ of cardinality k. The graph is generated by independently connecting each pair of vertices, with probability p if both belong to the community C^*, and with probability q otherwise. Equivalently, in terms of the adjacency matrix A, $A_{ij} \sim \text{Bern}(p)$ if $i, j \in C$ and $\text{Bern}(q)$ otherwise. Assume $p > q$. Thus the subgraph induced by C^* is likely to be denser than the rest of the graph. We are interested in the large-graph asymptotics, where both the network size n and the community size k grow to infinity.

Given the adjacency matrix A, the goal is to recover the hidden community C almost perfectly, i.e., achieving

$$\mathbb{E}[|\hat{C} \triangle C|] = o(k) \tag{30.13}$$

Given the network size n and the community size k, there exists a sharp condition on the edge density (p, q) that says the community needs to be sufficient denser than the outside. It turns out this is precisely described by the binary divergence $d(p\|q)$. Under the assumption that p/q is bounded, e.g., $p = 2q$, the information-theoretic necessary condition is

$$k \cdot d(p\|q) \to \infty \quad \text{and} \quad \liminf_{n\to\infty} \frac{kd(p\|q)}{\log \frac{n}{k}} \geq 2. \tag{30.14}$$

This condition is tight in the sense that if in the above "$\geq$" is replaced by "$>$", then there exists an estimator (e.g., the maximal likelihood estimator) that achieves (30.13).

Next we only prove the necessity of the second condition in (30.14), again using the mutual information method. Let ξ and $\hat{\xi}$ be the indicator vector of the community C and the estimator $\hat{C}$, respectively. Thus $\xi = (\xi_1, \dots, \xi_n)$ is uniformly drawn from the set $\{x \in \{0,1\}^n : w_H(x) = k\}$. Therefore ξ_i's are individually Bern(k/n). Let $\mathbb{E}[d_H(\xi, \hat{\xi})] = \epsilon_n k$, where $\epsilon_n \to 0$ by assumption. Consider the following chain of inequalities, which lower bounds the amount of information required for a distortion level ϵ_n:

$$\begin{aligned} I(A;\xi) &\overset{\text{dpi}}{\geq} I(\hat{\xi};\xi) \geq \min_{\mathbb{E}[d(\tilde{\xi},\xi)]\leq\epsilon_n k} I(\tilde{\xi};\xi) \geq H(\xi) - \max_{\mathbb{E}[d(\tilde{\xi},\xi)]\leq\epsilon_n k} H(\tilde{\xi}\oplus\xi) \\ &\overset{(30.12)}{=} \log\binom{n}{k} - nh\left(\frac{\epsilon_n k}{n}\right) \geq k\log\frac{n}{k}(1+o(1)), \end{aligned}$$

where the last step follows from the bound $\binom{n}{k} \geq \left(\frac{n}{k}\right)^k$, the assumption k/n is bounded away from one, and the bound $h(p) \leq -p\log p + p$ for $p \in [0,1]$.

On the other hand, to bound the mutual information, we use the golden formula Corollary 4.1 and choose a simple reference Q:

$$\begin{aligned} I(A;\xi) &= \min_Q D(P_{A|\xi}\|Q|P_\xi) \\ &\leq D(P_{A|\xi}\|\text{Bern}(q)^{\otimes\binom{n}{2}}|P_\xi) \\ &= \binom{k}{2} d(p\|q). \end{aligned}$$

Combining the last two displays yields $\liminf_{n\to\infty} \frac{(k-1)D(P\|Q)}{\log(n/k)} \geq 2$.

30.3 Assouad's method

Theorem 16.2 (Assouad's lemma) provides another method for lower bounding the minimax risk (especially popular for the high-dimensional questions, like density estimation). A high-level idea is that in the (two-point) LeCam method we attempt to find two values which have small $\text{TV}(P_{\theta_0}, P_{\theta_1})$ implying that the minimax risk is bounded by the distance between θ_0 and θ_1. Assouad's improvement is that if we manage to find 2^k pairs of such θ's and arrange then adjacent on the vertices of the hypercube, then the minimax risk is now bounded by k times the distance between adjacent θ's.

Here is a formal description:

- Step 0. Consider a statistical problem $\theta \in \Theta, X \sim P_\theta$ with a loss function $\ell(\theta, \hat{\theta})$ (note that this also models questions like estimating $f(\theta)$).

- Step 1. "Embedding the ϵ-hypercube in Θ". Suppose 2^k values $\theta_{b^k} \in \Theta, b^k \in \{0,1\}^k$ were selected so that one can convert any estimator $\hat{\theta}(X)$ into an estimator $\hat{B}^k$ so that for some $\epsilon > 0$:
$$\epsilon\, \mathbb{E}[d_H(\hat{B}^k, B^k)] \le \mathbb{E}[\ell(\theta, \hat{\theta}(X))]\,, \tag{30.15}$$
where we have the space
$$B^k \to \theta \to X \to \hat{\theta} \to \hat{B}^k\,, \qquad B_i \overset{\text{i.i.d.}}{\sim} \mathrm{Bern}(1/2), \theta = \theta_{B^k}, X \sim P_\theta \tag{30.16}$$
As an example, if Θ is a subset of a Hilbert space and $\ell(\hat{\theta}, \theta) = \|\hat{\theta} - \theta\|^2$, one chooses $\theta_{b^k} = \epsilon \sum_{i=1}^k b_i u_i$ where $u_1, \ldots, u_k$ are orthonormal in Θ.
- Step 2. "Bounding adjacent TV." Suppose furthermore that for any $b^k, \tilde{b}^k$ differing in one coordinate we have
$$\mathrm{TV}(P_{\theta_{b^k}}, P_{\theta_{\tilde{b}^k}}) \le c < 1\,. \tag{30.17}$$
- Step 3. Then we obtain a lower bound on the minimax risk:
$$\inf_{\hat{\theta}} \sup_{\theta\in\Theta} \mathbb{E}_{X\sim P_\theta}[\ell(\hat{\theta}(X), \theta)] \ge k\epsilon \frac{1-c}{2}\,. \tag{30.18}$$
Indeed, from Step 1 it is sufficient to lower-bound $\mathbb{E}[d_H(\hat{B}^k(X), B^k)]$ which is a sum of
$$\mathbb{P}[\hat{B}_i(X) \ne B_i] \ge \inf_f \mathbb{P}[f(X, B_{\sim i}) \ne B_i] \ge \frac{1-c}{2}$$
where in the first step we allowed "decoder of B_i" to depend on side-information $B_{\sim i} = (B_j, j \ne i)$, and in the second step we used (30.17). The proof of (30.18) is then completed by invoking (30.15).

As we described above, the key advantage here is the extra-factor k in (30.18) compared to the LeCam method.

Example 30.1. Say the data X is distributed according to P_θ parameterized by $\theta \in \mathbb{R}^k$ and let $\hat{\theta} = \hat{\theta}(X)$ be an estimator for θ. The goal is to minimize the maximal risk $\sup_{\theta\in\Theta} \mathbb{E}_\theta[\|\theta - \hat{\theta}\|_1]$. A lower bound (Bayesian) to this worst-case risk is the average risk $\mathbb{E}[\|\theta - \hat{\theta}\|_1]$, where θ is distributed to any prior. Consider θ uniformly distributed on the hypercube $\{0, \epsilon\}^k$ with side length ϵ embedded in the space of parameters. Then
$$\inf_{\hat{\theta}} \sup_{\theta\in\{0,\epsilon\}^k} \mathbb{E}[\|\theta - \hat{\theta}\|_1] \ge \frac{k\epsilon}{4} \min_{d_H(\theta,\theta')=1} (1 - \mathrm{TV}(P_\theta, P_{\theta'}))\,. \tag{30.19}$$
Explicitly, we have (WLOG assume $\epsilon = 1$).
$$\begin{aligned}
\mathbb{E}[\|\theta - \hat{\theta}\|_1] &\overset{(a)}{\ge} \frac{1}{2}\mathbb{E}[\|\theta - \hat{\theta}_{dis}\|_1] = \frac{1}{2}\mathbb{E}[d_H(\theta, \hat{\theta}_{dis})] \\
&\ge \frac{1}{2}\sum_{i=1}^k \min_{\hat{\theta}_i = \hat{\theta}_i(X)} \mathbb{P}[\theta_i \ne \hat{\theta}_i] \overset{(b)}{=} \frac{1}{4}\sum_{i=1}^k (1 - \mathrm{TV}(P_{X|\theta_i=0}, P_{X|\theta_i=1})) \\
&\overset{(c)}{\ge} \frac{k}{4} \min_{d_H(\theta,\theta')=1} (1 - \mathrm{TV}(P_\theta, P_{\theta'}))\,.
\end{aligned}$$

Here $\hat{\theta}_{dis}$ is the discretized version of $\hat{\theta}$, i.e. the closest point on the hypercube to $\hat{\theta}$ and so (a) follows from $|\theta_i - \hat{\theta}_i| \geq \frac{1}{2}\mathbf{1}_{\{|\theta_i - \hat{\theta}_i| > 1/2\}} = \frac{1}{2}\mathbf{1}_{\{\theta_i \neq \hat{\theta}_{dis,i}\}}$, (b) follows from the optimal binary hypothesis testing for θ_i given X, (c) follows from the convexity of TV: $\mathrm{TV}(P_{X|\theta_i=0}, P_{X|\theta_i=1}) = \mathrm{TV}(\frac{1}{2^{k-1}}\sum_{\theta:\theta_i=0} P_{X|\theta}, \frac{1}{2^{k-1}}\sum_{\theta:\theta_i=1} P_{X|\theta}) \leq \frac{1}{2^{k-1}}\sum_{\theta:\theta_i=0} \mathrm{TV}(P_{X|\theta}, P_{X|\theta\oplus e_i}) \leq \max_{d_H(\theta,\theta')=1} \mathrm{TV}(P_\theta, P_{\theta'})$. Alternatively, (c) also follows from by providing the extra information $\theta^{\backslash i}$ and allowing $\hat{\theta}_i = \hat{\theta}_i(X, \theta^{\backslash i})$ in the second line.

30.3.1 Assouad's lemma from the Mutual information method

One can integrate the Assouad's idea into the mutual information method. Consider, the probabilistic setting of (30.16). From the rate-distortion function of Bernoulli source (Section 27.1.1), we know that for any $\hat{B}^k$ and $\tau > 0$ there is some $\tau' > 0$ such that

$$I(B^k; X) \leq k(1-\tau)\log 2 \quad \Longrightarrow \quad \mathbb{E}[d_H(\hat{B}^k, B^k)] \geq k\tau' \,. \tag{30.20}$$

Here τ' is related to τ by $\tau \log 2 = h(\tau')$. Thus, if the "ϵ-hypercube embedding" has already been done, the bound similar to (30.18) will follow once we can bound $I(B^k; X)$ away from $k\log 2$.

Can we use the pairwise assumption (30.17) to do that? Yes! In fact we can recover exactly (30.18). Notice that thanks to the independence of B_i's we have[3]

$$I(B_i; X|B^{i-1}) = I(B_i; X, B^{i-1}) \leq I(B_i; X, B_{\backslash i}) = I(B_i; X|B_{\backslash i}) \,.$$

Applying the chain rule leads to the upper bound

$$I(B^k; X) = \sum_{i=1}^k I(B_i; X|B^{i-1}) \leq \sum_{i=1}^k I(B_i; X|B_{\backslash i}) \leq k\left(\log 2 - h\left(\frac{1-c}{2}\right)\right) \,,$$

where in the last step we also used a fact that for any $B \sim \mathrm{Bern}(1/2)$ we have

$$I(B; X) \leq \log 2 - h\left(\frac{1 - \mathrm{TV}(P_{X|B=0}, P_{X|B=1})}{2}\right) \,. \tag{30.21}$$

This implies the desired (30.18) by the mutual information method. To see (30.21), simply note that $I(B; X) = \mathbb{E}[\log 2 - h(\min_b P[B=b|X])] \leq \log 2 - h(\mathbb{E}[\min_b P[B=b|X]])$ by concavity and observe that $\mathbb{E}[\min_b P[B=b|X])] = \frac{1}{2}\int(P_{X|B=0} \wedge P_{X|B=1}) = \frac{1-\mathrm{TV}}{2}$.

In all, we may summarize Assouad's method as a convenient method for bounding $I(B^k; X)$ away from the full entropy ($k\log 2$) on the basis of distances between $P_{X|B^k}$ corresponding to adjacent B^k's.

[3] Equivalently, this also follows from the convexity of the mutual information in the channel (cf. Theorem 5.3).

Bibliography

[AFTS01] I.C. Abou-Faycal, M.D. Trott, and S. Shamai. The capacity of discrete-time memoryless rayleigh-fading channels. *IEEE Transaction Information Theory*, 47(4):1290 – 1301, 2001.

[Ahl82] Rudolf Ahlswede. An elementary proof of the strong converse theorem for the multiple-access channel. *J. Combinatorics, Information and System Sciences*, 7(3), 1982.

[AK01] András Antos and Ioannis Kontoyiannis. Convergence properties of functional estimates for discrete distributions. *Random Structures & Algorithms*, 19(3-4):163–193, 2001.

[Alo81] Noga Alon. On the number of subgraphs of prescribed type of graphs with a given number of edges. *Israel J. Math.*, 38(1-2):116–130, 1981.

[AN07] Shun-ichi Amari and Hiroshi Nagaoka. *Methods of information geometry*, volume 191. American Mathematical Soc., 2007.

[AS08] Noga Alon and Joel H. Spencer. *The Probabilistic Method.* John Wiley & Sons, 3rd edition, 2008.

[Ash65] Robert B. Ash. *Information Theory.* Dover Publications Inc., New York, NY, 1965.

[Ban92] Abhijit V Banerjee. A simple model of herd behavior. *The quarterly journal of economics*, 107(3):797–817, 1992.

[BC15] Sergey Bobkov and Gennadiy P Chistyakov. Entropy power inequality for the Rényi entropy. *IEEE Transactions on Information Theory*, 61(2):708–714, 2015.

[BF14] Ahmad Beirami and Faramarz Fekri. Fundamental limits of universal lossless one-to-one compression of parametric sources. In *Information Theory Workshop (ITW), 2014 IEEE*, pages 212–216. IEEE, 2014.

[Bir83] L. Birgé. Approximation dans les espaces métriques et théorie de l'estimation. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 65(2):181–237, 1983.

[Bla74] R. E. Blahut. Hypothesis testing and information theory. *IEEE Trans. Inf. Theory*, 20(4):405–417, 1974.

[BNO03] Dimitri P Bertsekas, Angelia Nedić, and Asuman E. Ozdaglar. *Convex analysis and optimization.* Athena Scientific, Belmont, MA, USA, 2003.

[Boh38] H. F. Bohnenblust. Convex regions and projections in Minkowski spaces. *Ann. Math.*, 39(2):301–308, 1938.

[Bre73] Lev M Bregman. Some properties of nonnegative matrices and their permanents. *Soviet Math. Dokl.*, 14(4):945–949, 1973.

[Bro86] L. D. Brown. Fundamentals of statistical exponential families with applications in statistical decision theory. In S. S. Gupta, editor, *Lecture Notes-Monograph Series*, volume 9. Institute of Mathematical Statistics, Hayward, CA, 1986.

[CB90] B. S. Clarke and A. R. Barron. Information-theoretic asymptotics of Bayes methods. *IEEE Trans. Inf. Theory*, 36(3):453–471, 1990.

[CB94] Bertrand S Clarke and Andrew R Barron. Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical planning and Inference*, 41(1):37–60, 1994.

[Ç11] Erhan Çinlar. *Probability and Stochastics.* Springer, New York, 2011.

[Cho56] Noam Chomsky. Three models for the description of language. *IRE Trans. Inform. Th.*, 2(3):113–124, 1956.

[CK81a] I. Csiszár and J. Körner. Graph decomposition: a new key to coding theorems. *IEEE Trans. Inf. Theory*, 27(1):5–12, 1981.

[CK81b] I. Csiszár and J. Körner. *Information Theory: Coding Theorems for Discrete Memoryless Systems.* Academic, New York, 1981.

[CS83] J. Conway and N. Sloane. A fast encoding method for lattice codes and quantizers. *IEEE Transactions on Information Theory*, 29(6):820–824, Nov 1983.

[Csi67] I. Csiszár. Information-type measures of difference of probability distributions and indirect observation. *Studia Sci. Math. Hungar.*, 2:229–318, 1967.

[CT06] Thomas M. Cover and Joy A. Thomas. *Elements of information theory, 2nd Ed.* Wiley-Interscience, New York, NY, USA, 2006.

[Doo53] Joseph L. Doob. *Stochastic Processes.* New York Wiley, 1953.

[Eli55] Peter Elias. Coding for noisy channels. *IRE Convention Record*, 3:37–46, 1955.

[Eli72] P. Elias. The efficient construction of an unbiased random sequence. *Annals of Mathematical Statistics*, 43(3):865–870, 1972.

[ELZ05] Uri Erez, Simon Litsyn, and Ram Zamir. Lattices which are good for (almost) everything. *IEEE Transactions on Information Theory*, 51(10):3401–3416, Oct. 2005.

[ES03] Dominik Maria Endres and Johannes E Schindelin. A new metric for probability distributions. *IEEE Transactions on Information theory*, 49(7):1858–1860, 2003.

[EZ04] U. Erez and R. Zamir. Achieving $\frac{1}{2}\log(1+\mathrm{SNR})$ on the AWGN channel with lattice encoding and decoding. *IEEE Trans. Inf. Theory*, IT-50:2293–2314, Oct. 2004.

[FHT03] A. A. Fedotov, P. Harremoës, and F. Topsøe. Refinements of Pinsker's inequality. *Information Theory, IEEE Transactions on*, 49(6):1491–1498, Jun. 2003.

[FJ89] G.D. Forney Jr. Multidimensional constellations. II. Voronoi constellations. *IEEE Journal on Selected Areas in Communications*, 7(6):941–958, Aug 1989.

[FK98] Ehud Friedgut and Jeff Kahn. On the number of copies of one hypergraph in another. *Israel J. Math.*, 105:251–256, 1998.

[FMG92] Meir Feder, Neri Merhav, and Michael Gutman. Universal prediction of individual sequences. *IEEE Trans. Inf. Theory*, 38(4):1258–1270, 1992.

[Gil10] Gustavo L Gilardoni. On pinsker's and vajda's type inequalities for csiszár's-divergences. *Information Theory, IEEE Transactions on*, 56(11):5377–5386, 2010.

[GKY56] I. M. Gel'fand, A. N. Kolmogorov, and A. M. Yaglom. On the general definition of the amount of information. *Dokl. Akad. Nauk. SSSR*, 11:745–748, 1956.

[GL95a] R. D. Gill and B. Y. Levit. Applications of the van Trees inequality: a Bayesian Cramér-Rao bound. *Bernoulli*, 1(1–2):59–79, 1995.

[GL95b] Richard D Gill and Boris Y Levit. Applications of the van Trees inequality: a Bayesian Cramér-Rao bound. *Bernoulli*, pages 59–79, 1995.

[Har] Sergiu Hart. Overweight puzzle. http://www.ma.huji.ac.il/˜hart/puzzle/overweight.html.

[Hoe65] Wassily Hoeffding. Asymptotically optimal tests for multinomial distributions. *The Annals of Mathematical Statistics*, pages 369–401, 1965.

[HV11] P. Harremoës and I. Vajda. On pairs of f-divergences and their joint range. *IEEE Trans. Inf. Theory*, 57(6):3230–3235, Jun. 2011.

[IS03] Y. I. Ingster and I. A. Suslina. *Nonparametric goodness-of-fit testing under Gaussian models*. Springer, New York, NY, 2003.

[KF+09] Christof Külske, Marco Formentin, et al. A symmetric entropy bound on the non-reconstruction regime of markov chains on galton-watson trees. *Electronic Communications in Probability*, 14:587–596, 2009.

[KO94] M.S. Keane and G.L. O'Brien. A Bernoulli factory. *ACM Transactions on Modeling and Computer Simulation*, 4(2):213–219, 1994.

[Kos63] VN Koshelev. Quantization with minimal entropy. *Probl. Pered. Inform*, 14:151–156, 1963.

[KS14] Oliver Kosut and Lalitha Sankar. Asymptotics and non-asymptotics for universal fixed-to-variable source coding. *arXiv preprint arXiv:1412.4444*, 2014.

[KV14] Ioannis Kontoyiannis and Sergio Verdú. Optimal lossless data compression: Non-asymptotics and asymptotics. *IEEE Trans. Inf. Theory*, 60(2):777–795, 2014.

[LC86] Lucien Le Cam. *Asymptotic methods in statistical decision theory*. Springer-Verlag, New York, NY, 1986.

[LM03] Amos Lapidoth and Stefan M Moser. Capacity bounds via duality with applications to multiple-antenna systems on flat-fading channels. *IEEE Transactions on Information Theory*, 49(10):2426–2467, 2003.

[Loe97] Hans-Andrea Loeliger. Averaging bounds for lattices and linear codes. *IEEE Transactions on Information Theory*, 43(6):1767–1773, Nov. 1997.

[Mas74] James Massey. On the fractional weight of distinct binary n-tuples (corresp.). *IEEE Transactions on Information Theory*, 20(1):131–131, 1974.

[MF98] Neri Merhav and Meir Feder. Universal prediction. *IEEE Trans. Inf. Theory*, 44(6):2124–2147, 1998.

[MP05] Elchanan Mossel and Yuval Peres. New coins from old: computing with unknown bias. *Combinatorica*, 25(6):707–724, 2005.

[MT10] Mokshay Madiman and Prasad Tetali. Information inequalities for joint distributions, with interpretations and applications. *IEEE Trans. Inf. Theory*, 56(6):2699–2713, 2010.

[OE15] O. Ordentlich and U. Erez. A simple proof for the existence of "good" pairs of nested lattices. *IEEE Transactions on Information Theory*, Submitted Aug. 2015.

[OPS48] BM Oliver, JR Pierce, and CE Shannon. The philosophy of pcm. *Proceedings of the IRE*, 36(11):1324–1331, 1948.

[Per92] Yuval Peres. Iterating von Neumann's procedure for extracting random bits. *Annals of Statistics*, 20(1):590–597, 1992.

[PPV10a] Y. Polyanskiy, H. V. Poor, and S. Verdú. Channel coding rate in the finite blocklength regime. *IEEE Trans. Inf. Theory*, 56(5):2307–2359, May 2010.

[PPV10b] Y. Polyanskiy, H. V. Poor, and S. Verdú. Feedback in the non-asymptotic regime. *IEEE Trans. Inf. Theory*, April 2010. submitted for publication.

[PPV11] Y. Polyanskiy, H. V. Poor, and S. Verdú. Minimum energy to send k bits with and without feedback. *IEEE Trans. Inf. Theory*, 57(8):4880–4902, August 2011.

[PV10] Y. Polyanskiy and S. Verdú. Arimoto channel coding converse and Rényi divergence. In *Proc. 2010 48th Allerton Conference*. Allerton Retreat Center, Monticello, IL, USA, September 2010.

[PW12] E. Price and D. P. Woodruff. Applications of the shannon-hartley theorem to data streams and sparse recovery. In *Proceedings of the 2012 IEEE International Symposium on Information Theory*, pages 1821–1825, Boston, MA, Jul. 2012.

[PW14] Y. Polyanskiy and Y. Wu. Peak-to-average power ratio of good codes for Gaussian channel. *IEEE Trans. Inf. Theory*, 60(12):7655–7660, December 2014.

[PW17] Yury Polyanskiy and Yihong Wu. Strong data-processing inequalities for channels and Bayesian networks. In Eric Carlen, Mokshay Madiman, and Elisabeth M. Werner, editors, *Convexity and Concentration. The IMA Volumes in Mathematics and its Applications, vol 161*, pages 211–249. Springer, New York, NY, 2017.

[Rad97] Jaikumar Radhakrishnan. An entropy proof of Bregman's theorem. *J. Combin. Theory Ser. A*, 77(1):161–164, 1997.

[Ree65] Alec H Reeves. The past present and future of PCM. *IEEE Spectrum*, 2(5):58–62, 1965.

[RSU01] Thomas J. Richardson, Mohammad Amin Shokrollahi, and Rüdiger L. Urbanke. Design of capacity-approaching irregular low-density parity-check codes. *IEEE Transactions on Information Theory*, 47(2):619–637, 2001.

[RU96] Bixio Rimoldi and Rüdiger Urbanke. A rate-splitting approach to the gaussian multiple-access channel. *Information Theory, IEEE Transactions on*, 42(2):364–375, 1996.

[RZ86] Ryabko B. Reznikova Zh. Analysis of the language of ants by information-theoretical methods. *Problemi Peredachi Informatsii*, 22(3):103–108, 1986. English translation: http://reznikova.net/R-R-entropy-09.pdf.

[SF11] Ofer Shayevitz and Meir Feder. Optimal feedback communication via posterior matching. *IEEE Trans. Inf. Theory*, 57(3):1186–1222, 2011.

[Sha48] C. E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27:379–423 and 623–656, July/October 1948.

[Sio58] Maurice Sion. On general minimax theorems. *Pacific J. Math*, 8(1):171–176, 1958.

[Smi71] J. G. Smith. The information capacity of amplitude and variance-constrained scalar Gaussian channels. *Information and Control*, 18:203 – 219, 1971.

[Spe15] Spectre. SPECTRE: Short packet communication toolbox. https://github.com/yp-mit/spectre, 2015. GitHub repository.

[Spi96] Daniel A. Spielman. Linear-time encodable and decodable error-correcting codes. *IEEE Transactions on Information Theory*, 42(6):1723–1731, 1996.

[Spi97] Daniel A. Spielman. The complexity of error-correcting codes. In *Fundamentals of Computation Theory*, pages 67–84. Springer, 1997.

[SV11] Wojciech Szpankowski and Sergio Verdú. Minimum expected length of fixed-to-variable lossless compression without prefix constraints. *IEEE Trans. Inf. Theory*, 57(7):4017–4025, 2011.

[SV16] Igal Sason and Sergio Verdu. f-divergence inequalities. *IEEE Transactions on Information Theory*, 62(11):5973–6006, 2016.

[TE97] Giorgio Taricco and Michele Elia. Capacity of fading channel with no side information. *Electronics Letters*, 33(16):1368–1370, 1997.

[Top00] F. Topsøe. Some inequalities for information divergence and related measures of discrimination. *IEEE Transactions on Information Theory*, 46(4):1602–1609, 2000.

[Tsy09] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Verlag, New York, NY, 2009.

[TV05] David Tse and Pramod Viswanath. *Fundamentals of wireless communication*. Cambridge University Press, 2005.

[UR98] R. Urbanke and B. Rimoldi. Lattice codes can achieve capacity on the AWGN channel. *IEEE Transactions on Information Theory*, 44(1):273–278, 1998.

[Vaj70] Igor Vajda. Note on discrimination information and variation (corresp.). *IEEE Transactions on Information Theory*, 16(6):771–773, 1970.

[Ver07] S. Verdú. *EE528–Information Theory, Lecture Notes*. Princeton Univ., Princeton, NJ, 2007.

[vN51] J. von Neumann. Various techniques used in connection with random digits. *Monte Carlo Method, National Bureau of Standards, Applied Math Series*, (12):36–38, 1951.

[Yek04] Sergey Yekhanin. Improved upper bound for the redundancy of fix-free codes. *IEEE Trans. Inf. Theory*, 50(11):2815–2818, 2004.

[Yos03] Nobuyuki Yoshigahara. *Puzzles 101: A Puzzlemaster's Challenge.* A K Peters, Natick, MA, USA, 2003.

[Yu97] Bin Yu. Assouad, Fano, and Le Cam. *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, pages 423–435, 1997.

[Zam14] Ram Zamir. *Lattice Coding for Signals and Networks.* Cambridge University Press, Cambridge, 2014.

[ZY97] Zhen Zhang and Raymond W Yeung. A non-Shannon-type conditional inequality of information quantities. *IEEE Trans. Inf. Theory*, 43(6):1982–1986, 1997.

[ZY98] Zhen Zhang and Raymond W Yeung. On characterization of entropy function via information inequalities. *IEEE Trans. Inf. Theory*, 44(4):1440–1452, 1998.