Minimax Lower Bounds

Presented by Wanteng Ma

Department of Mathematics

Hong Kong University of Science and Technology

April 14, 2022

Motivation

Minimax lower bounds describe how fast we can learn a problem from data in the worst case scenario.

- Can we obtain matching lower bounds on estimation rates for
  - a specific procedure or algorithm?
  - **any possible algorithm**?
- Lower bounds of this minimax type can yield two different but complementary types of insight
  - If some computationally efficient estimators are statistically "optimal", then lower statistical errors are of little interest.
  - If lower bounds do not match the best known upper bounds. In this case, one has a strong motivation to study alternative estimators.

Definition

Given a class of distributions $\mathcal{P}$, with a functional $\theta(\mathbb{P})$ for $\mathbb{P} \in \mathcal{P}$. Our goal is to estimate $\theta(\mathbb{P})$ based on samples drawn from the unknown distribution $\mathbb{P}$.

1. The quantity $\theta(\mathbb{P})$ may uniquely determines the underlying distribution $\mathbb{P}$.
2. In other settings, it does not uniquely specify the distribution. For instance, consider density function f,
   - Estimating the quadratic functional

$$\mathbb{P} \mapsto \theta(\mathbb{P}) = \int_0^1 \left(f'(t)\right)^2 dt \in \mathbb{R}$$

   - For a class of unimodal density functions, consider estimating the mode of the density

$$\theta(\mathbb{P}) = \arg\max_{x \in [0,1]} f(x)$$

Definition

For any fixed θ^*, there is always a very good way to estimate it: ignore the data, and return θ^*. The resulting deterministic estimator has zero risk when evaluated at the fixed θ^*, but is likely to behave very poorly for other choices.

Definition (Minimax risks)

If any distribution $\mathbb{P}$ is with quantity $\theta = \theta(\mathbb{P})$, and $\rho(\theta, \theta')$ is a semi-metric, letting Φ be an increasing function on the non-negative real line, then the minimax risk is defined as

$$\mathcal{M}(\theta(\mathcal{P}); \Phi \circ \rho) := \inf_{\widehat{\theta}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}}[\Phi(\rho(\widehat{\theta}, \theta(\mathbb{P})))] = \inf_{\widehat{\theta}} \sup_{\mathbb{P} \in \mathcal{P}} R(\widehat{\theta}, \mathbb{P})$$

Definition

Definition (Bayes risks)

For a risk function and a given prior $\Lambda(\theta)$, Bayes risk is the lowest possible risk when the parameter is sampled from the prior[a]

$$R_{Bayes} = \inf_{\widehat{\theta}} r(\Lambda, \widehat{\theta}) = \inf_{\widehat{\theta}} \int R(\widehat{\theta}, \theta) d\Lambda(\theta)$$

[a] Erich L Lehmann and George Casella. *Theory of point estimation.* Springer Science & Business Media, 2006.

Bayes risk is **a nature lower bound** of minimax lower bound. One classic approach to get sharp Bayes risk lower bound is the van Trees inequality,

$$\int \mathrm{E}_\theta \left[(\widehat{\theta}(\mathbf{X}) - \theta)^2 \right] d\Lambda(\theta) \geq \frac{1}{\int \mathcal{I}(\theta) d\Lambda(\theta) + \mathcal{J}(\theta)},$$

where $\mathcal{I}(\theta)$ and $\mathcal{J}(\theta)$ are Fisher information of X and θ respectively.

From estimation to testing

Minimax lower bounds can be obtained via "reduction" to the problem of obtaining lower bounds for the probability of error in a certain testing problem.

Suppose that $\left\{\theta^1, \ldots, \theta^M\right\}$ is a 2δ-separated set. For each θ^j, there is a distribution $\mathbb{P}_{\theta^j}$ that can represent θ^j. Consider the M-ary hypothesis testing problem defined by the family of distributions $\{\mathbb{P}_{\theta^j}, j = 1, \ldots, M\}$.

1. Sample a random integer J from the uniform distribution over the index set $[M]$
2. Given $J = j$, sample $Z \sim \mathbb{P}_{\theta^j}$. Then we have $Z \sim \bar{\mathbb{Q}} = \frac{1}{M} \sum_{j=1}^{M} \mathbb{P}_{\theta^j}$
3. For a testing function $\psi : \mathcal{Z} \to [M]$, associated probability of error is given by $\mathbb{Q}[\psi(Z) \neq J]$

From estimation to testing

Proposition (From estimation to testing)

For any increasing function Φ and choice of 2δ-separated set, the minimax risk is lower bounded as

$$\mathcal{M}(\theta(\mathcal{P}), \Phi \circ \rho) \geq \Phi(\delta) \inf_{\psi} \mathbb{Q}[\psi(Z) \neq J]$$

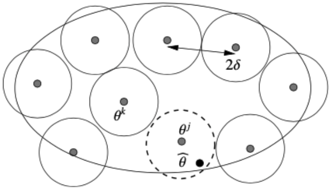


Figure 15.1 Reduction from estimation to testing using a 2δ-separated set in the space Ω in the semi-metric ρ. If an estimator $\widehat{\theta}$ satisfies the bound $\rho(\widehat{\theta}, \theta^j) < \delta$ whenever the true parameter is θ^j, then it can be used to determine the correct index j in the associated testing problem.

- $\mathbb{E}_{\mathbb{P}}[\Phi(\rho(\widehat{\theta}, \theta))] \geq \Phi(\delta)\mathbb{P}[\Phi(\rho(\widehat{\theta}, \theta)) \geq \Phi(\delta)] \geq \Phi(\delta)\mathbb{P}[\rho(\widehat{\theta}, \theta) \geq \delta]$

From estimation to testing

Thus, it suffices to lower bound the quantity $\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}[\rho(\widehat{\theta},\theta(\mathbb{P}))\geq\delta]$. From the definition of the mixture $\bar{\mathbb{Q}}$, by construction we have

- $$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}\left[\rho(\widehat{\theta},\theta(\mathbb{P}))\geq\delta\right] \geq \frac{1}{M}\sum_{j=1}^{M}\mathbb{P}_{\theta^j}\left[\rho\left(\widehat{\theta},\theta^j\right)\geq\delta\right] = \mathbb{Q}\left[\rho\left(\widehat{\theta},\theta^J\right)\geq\delta\right]$$

- For any estimation $\widehat{\theta}$, we have a corresponding test

  $$\psi(Z) := \arg\min_{\ell\in[M]} \rho\left(\theta^\ell,\widehat{\theta}\right)$$

  which subjects to $\left\{\rho\left(\widehat{\theta},\theta^J\right)<\delta\right\} \Rightarrow \{\psi(Z)=J\}$,

  i.e.,

  $$\mathbb{Q}\left[\rho\left(\widehat{\theta},\theta^J\right)\geq\delta\right] = \frac{1}{M}\sum_{j=1}^{M}\mathbb{P}_{\theta^j}\left[\rho\left(\widehat{\theta},\theta^j\right)\geq\delta\right] \geq \mathbb{Q}[\psi(Z)\neq J]$$

Some divergence measures

Our next step is to develop techniques for lower bounding the error probability, for which we require some background on divergence measures between probability distributions.

1. total variation (TV) distance: $\|\mathbb{P}-\mathbb{Q}\|_{\mathrm{TV}} := \sup_{A\subseteq\mathcal{X}} |\mathbb{P}(A)-\mathbb{Q}(A)|$
$$\|\mathbb{P}-\mathbb{Q}\|_{\mathrm{TV}} = \frac{1}{2}\int_{\mathcal{X}} |p(x)-q(x)|dv(x)$$
2. Kullback–Leibler (KL) divergence: $D(\mathbb{Q}\|\mathbb{P}) = \int_{\mathcal{X}} q(x)\log\frac{q(x)}{p(x)}dv(x)$
3. Hellinger distance: $H^2(\mathbb{P}\|\mathbb{Q}) := \int(\sqrt{p(x)}-\sqrt{q(x)})^2 dv(x)$

Lemma (Pinsker–Csisar–Kullback inequality)

$$\|\mathbb{P}-\mathbb{Q}\|_{\mathrm{TV}} \le \sqrt{\frac{1}{2}D(\mathbb{Q}\|\mathbb{P})}$$

Lemma (Le Cam's inequality)

$$\|\mathbb{P}-\mathbb{Q}\|_{\mathrm{TV}} \le H(\mathbb{P}\|\mathbb{Q})\sqrt{1-\frac{H^2(\mathbb{P}\|\mathbb{Q})}{4}}$$

Binary testing and Le Cam's method

For the divergence measures, we have the following properties for i.i.d cases

- $D\left(\mathbb{P}^{1:n} \| \mathbb{Q}^{1:n}\right)=\sum_{i=1}^{n} D\left(\mathbb{P}_{i} \| \mathbb{Q}_{i}\right)=n D\left(\mathbb{P}_{1} \| \mathbb{Q}_{1}\right)$
- $\frac{1}{2} H^{2}\left(\mathbb{P}^{1:n} \| \mathbb{Q}^{1:n}\right)=1-\left(1-\frac{1}{2} H^{2}\left(\mathbb{P}_{1} \| \mathbb{Q}_{1}\right)\right)^{n} \leq \frac{1}{2} n H^{2}\left(\mathbb{P}_{1} \| \mathbb{Q}_{1}\right)$

In a binary testing problem with equally weighted hypotheses, suppose Z is drawn from $\bar{\mathbb{Q}}:=\frac{1}{2} \mathbb{P}_{0}+\frac{1}{2} P_{1}$. We have:

$$\inf_{\psi} \mathbb{Q}[\psi(Z) \neq J]=\frac{1}{2}\left\{1-\left\|\mathbb{P}_{1}-\mathbb{P}_{0}\right\|_{\mathrm{TV}}\right\}$$

- Each decision rule ψ is equivalent to a decision region $\left(A, A^{c}\right)$. Thus

$$\begin{aligned}
\sup_{\psi} \mathbb{Q}[\psi(Z)=J] &=\sup_{A \subseteq \mathcal{X}}\left\{\frac{1}{2} \mathbb{P}_{1}(A)+\frac{1}{2} \mathbb{P}_{0}\left(A^{c}\right)\right\} \\
&=\frac{1}{2} \sup_{A \subseteq \mathcal{X}}\left\{\mathbb{P}_{1}(A)-\mathbb{P}_{0}(A)\right\}+\frac{1}{2}
\end{aligned}$$

Binary testing and Le Cam's method

- Since $\sup_\psi \mathbb{Q}[\psi(Z) = J] = 1 - \inf_\psi \mathbb{Q}[\psi(Z) \neq J]$, combined with the estimation-testing transform we have

Lemma (Le Cam's method for binary testing)

Consider a pair of 2δ-separated distributions $\mathbb{P}_0, \mathbb{P}_1 \in \mathcal{P}$. We have

$$\mathcal{M}(\theta(\mathcal{P}), \Phi \circ \rho) \geq \frac{\Phi(\delta)}{2} \{1 - \|\mathbb{P}_1 - \mathbb{P}_0\|_{\mathrm{TV}}\}$$

Example: (Gaussian location family) For the Gaussian location family $\mathbb{P}_\theta = N(\theta, \sigma^2)$ with fixed σ^2, use squared error to measure the risk.

- Set $\theta = 2\delta$ we have

$$\|\mathbb{P}_\theta^n - \mathbb{P}_0^n\|_{\mathrm{TV}}^2 \leq \frac{1}{4}\left\{e^{n\theta^2/\sigma^2} - 1\right\} = \frac{1}{4}\left\{e^{4n\delta^2/\sigma^2} - 1\right\}$$

- Choose $\delta = \frac{\sigma}{2\sqrt{n}}$ and use Le Cam's method:

$$\inf_{\widehat{\theta}} \sup_{\theta \in \mathbb{R}} \mathbb{E}_\theta\left[(\widehat{\theta} - \theta)^2\right] \geq \frac{\delta^2}{2}\left\{1 - \frac{1}{2}\sqrt{e-1}\right\} \geq \frac{\delta^2}{6} = \frac{1}{24}\frac{\sigma^2}{n}$$

Le Cam for functionals

Le Cam's method is also useful for nonparametric problems especially in density estimation. An important quantity in the Le Cam approach to such problems is the **Lipschitz constant** of θ w.r.t the Hellinger norm, given by

$$\omega(\epsilon;\theta,\mathcal{F}) := \sup_{f,g\in\mathcal{F}} \left\{ |\theta(f)-\theta(g)| \,\middle|\, H^2(f\|g) \le \epsilon^2 \right\}$$

Corollary (Le Cam for functionals)

For any increasing function Φ on the non-negative real line and any functional $\theta : \mathcal{F} \to \mathbb{R}$, we have

$$\inf_{\widehat{\theta}} \sup_{f\in\mathcal{F}} \mathbb{E}[\Phi(\widehat{\theta}-\theta(f))] \ge \frac{1}{4}\Phi\left(\frac{1}{2}\omega\left(\frac{1}{2\sqrt{n}};\theta,\mathcal{F}\right)\right)$$

- Setting $\epsilon^2 = \frac{1}{4n}$, we can find such a pair f, g that achieve $\omega(1/(2\sqrt{n}))$, thus

$$\left\|\mathbb{P}_f^n - \mathbb{P}_g^n\right\|_{\mathrm{TV}}^2 \le H^2\left(\mathbb{P}_f^n \| \mathbb{P}_g^n\right) \le nH^2\left(\mathbb{P}_f \| \mathbb{P}_g\right) \le \frac{1}{4}$$

Lower bounds for quadratic functionals

Now consider the class of twice-differentiable density functions: $\mathcal{F}_2([0,1]) := \left\{ f : [0,1] \to [c_0, c_1] \mid \|f''\|_\infty \le c_2 \text{ and } \int_0^1 f(x)dx = 1 \right\}$. The quadratic functional $f \mapsto \theta(f) := \int_0^1 \left(f'(x)\right)^2 dx$ measure the "smoothness" of the density.

- **Key idea**: constructing perturbation based on the uniform distribution f_0 to control the Lipschitz constant.
- Find a certain basis function $\phi(x)$ and define the shifted rescaled $\phi_j(x) = \frac{C}{m^2}\phi\left(m\left(x - x_j\right)\right)$ in an sub-interval $[x_j, x_{j+1}]$ with $x_j = \frac{j}{m}$
- We construct a new density $g(x) := 1 + \sum_{j=1}^m \phi_j(x)$ and control the Hellinger distance

$$\frac{1}{2}H^2\left(f_0 \| g\right) = 1 - \int_0^1 \sqrt{1 + \sum_{j=1}^m \phi_j(x)dx} \le cb_0 \frac{1}{m^4}$$

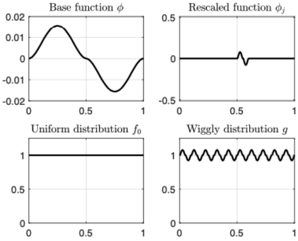
Lower bounds for quadratic functionals

- Choose $m^4 := 2cb_0 n$, and then we have $H^2\left(f_0 \| g\right) \leq \frac{1}{n}$ which satisfies the Corollary.
- For the functionals we have

$$\theta(g)=\int_0^1\left(\sum_{j=1}^m \phi_j'(x)\right)^2 dx = m\int_0^1\left(\phi_j'(x)\right)^2 dx=\frac{C^2 b_1}{m^2}.$$

- $|\theta(g)-\theta\left(f_0\right)| \geq \frac{K}{\sqrt{n}}$, which indicates that

$$\sup_{f\in\mathcal{F}_2}\mathbb{E}[|\widehat{\theta}(f)-\theta(f)|]\gtrsim n^{-1/2}$$



Le Cam's convex hull method

Taking the convex hulls of **two classes** of distribution can make the lower bound tighter. Consider two subsets that are 2δ-separated, in the sense that $\rho\left(\theta\left(\mathbb{P}_0\right), \theta\left(\mathbb{P}_1\right)\right) \geq 2\delta \quad$ for all $\mathbb{P}_0 \in \mathcal{P}_0$ and $\mathbb{P}_1 \in \mathcal{P}_1$.

Lemma (Le Cam)

For any 2δ-separated classes of distributions $\mathcal{P}_0$ and $\mathcal{P}_1$, any estimator $\widehat{\theta}$ has worst-case risk at least

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\rho(\widehat{\theta}, \theta(\mathbb{P}))] \geq \frac{\delta}{2} \sup_{\substack{\mathbb{P}_0\in\mathrm{conv}(\mathcal{P}_0)\\ \mathbb{P}_1\in\mathrm{conv}(\mathcal{P}_1)}} \left\{1 - \|\mathbb{P}_0 - \mathbb{P}_1\|_{\mathrm{TV}}\right\}$$

- Define the random variables $V_j(\widehat{\theta}) = \frac{1}{2\delta} \inf_{\mathbb{P}_j\in\mathcal{P}_j} \rho\left(\widehat{\theta}, \theta\left(\mathbb{P}_j\right)\right)$, for $j = 0, 1$. Then we have

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\rho(\widehat{\theta}, \theta(\mathbb{P}))] \geq \delta \left\{\mathbb{E}_{\mathbb{P}_0}\left[V_0(\widehat{\theta})\right] + \mathbb{E}_{\mathbb{P}_1}\left[V_1(\widehat{\theta})\right]\right\}$$

Le Cam's convex hull method

- Since the RHS is linear, we can take suprema over the convex hulls,

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\rho(\widehat{\theta},\theta(\mathbb{P}))] \geq \delta \sup_{\substack{\mathbb{P}_0\in\mathrm{conv}(\mathcal{P}_0)\\ \mathbb{P}_1\in\mathrm{conv}(\mathcal{P}_1)}} \left\{\mathbb{E}_{\mathbb{P}_0}\left[V_0(\widehat{\theta})\right] + \mathbb{E}_{\mathbb{P}_1}\left[V_1(\widehat{\theta})\right]\right\}$$

- By the triangle inequality, we have

$$\rho\left(\widehat{\theta},\theta\left(\mathbb{P}_0\right)\right) + \rho\left(\widehat{\theta},\theta\left(\mathbb{P}_1\right)\right) \geq \rho\left(\theta\left(\mathbb{P}_0\right),\theta\left(\mathbb{P}_1\right)\right) \geq 2\delta,$$

  i.e., $V_0(\widehat{\theta}) + V_1(\widehat{\theta}) \geq 1$.

- Studying the integral of density we have

$$\mathbb{E}_{\mathbb{P}_0}\left[V_0(\widehat{\theta})\right] + \mathbb{E}_{\mathbb{P}_1}\left[V_1(\widehat{\theta})\right] \geq 1 - \|\mathbb{P}_1 - \mathbb{P}_0\|_{\mathrm{TV}}$$

Example: (Gaussian location family) consider the two families $\mathcal{P}_0 = \{\mathbb{P}_0^n\}$ and $\mathcal{P}_1 = \left\{\mathbb{P}_\theta^n, \mathbb{P}_{-\theta}^n\right\}$. Note that the mixture distribution $\overline{\mathbb{P}} := \frac{1}{2}\mathbb{P}_\theta^n + \frac{1}{2}\mathbb{P}_{-\theta}^n$ belongs to $\mathrm{conv}\left(\mathcal{P}_1\right)$. Comparing $\overline{\mathbb{P}}$ and $\mathbb{P}_0$ we can get a sharper lower bound.

Optimal bounds for quadratic functionals

We resume the previous example for estimating quadratic functionals.

- For each binary vector $\alpha \in \{-1,1\}^m$, define $\mathbb{P}_\alpha$ with density $f_\alpha = 1 + \sum_{j=1}^m \alpha_j \phi_j(x)$.
- Define $\mathcal{P}_0 := \{\mathbb{U}^n\}$ and $\mathcal{P}_1 := \{\mathbb{P}_\alpha^n, \alpha \in \{-1,+1\}^m\}$.
- Let $\mathbb{Q} := 2^{-m} \sum_{\alpha \in \{-1,+1\}^m} \mathbb{P}_\alpha^n$ be the uniformly weighted mixture over all 2^m choices of $\mathbb{P}_\alpha^n$. Then,

$$\inf_{\substack{\mathbb{P}_j \in \mathrm{conv}(\mathcal{P}_j) \\ j=0,1}} \|\mathbb{P}_0 - \mathbb{P}_1\|_{\mathrm{TV}} \le \|\mathbb{U}^n - \mathbb{Q}\|_{\mathrm{TV}} \le H\left(\mathbb{U}^n \| \mathbb{Q}\right)$$

- One possible upper bound is given by

$$H^2\left(\mathbb{U}^n \| \mathbb{Q}\right) \le n^2 \sum_{j=1}^m \left(\int_0^1 \phi_j^2(x) dx\right)^2 \le b_0^2 \frac{n^2}{m^9}.$$

- Setting $m^9 = 4b_0^2 n^2$ yields that $\left\|\mathbb{U}^{1:n} - \mathbb{Q}\right\|_{\mathrm{TV}} \le H\left(\mathbb{U}^{1:n} \| \mathbb{P}^{1:n}\right) \le 1/2$

Optimal bounds for quadratic functionals

- Apply Le Cam's convex hull method:

$$\sup_{f \in \mathcal{F}_2} \mathbb{E}|\widehat{\theta}(f) - \theta(f)| \geq \delta/4 = \frac{C^2 b_1}{8m^2} \gtrsim n^{-4/9} \gg n^{-1/2}$$

- This lower bound turns out to be unimprovable.

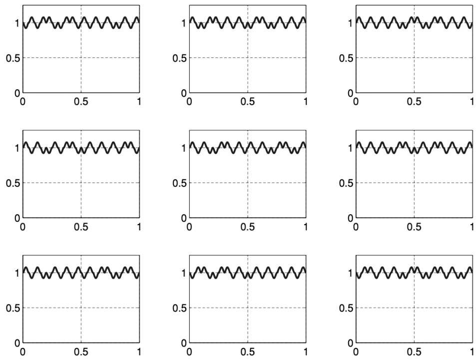


Figure 15.4 Illustration of some densities of the form $f_\alpha(x) = 1 + \sum_{j=1}^{m} \alpha_j \phi_j(x)$ for different choices of sign vectors $\alpha \in \{-1, 1\}^m$. Note that there are 2^m such densities in total.

Fano's method

For metrics like $\rho(f,g) = \int (f-g)^2$, Le Cam's method will usually not give a tight bound. Instead, we use Fano's method.

Lemma (Fano's inequality)

For any r.v J, Z, with $J \in \mathcal{X}$, $|\mathcal{X}| = M$, consider the error $e = \mathrm{I}\{\psi(Z) \neq J\}$, and $q(e) = \mathbb{P}[\psi(Z) \neq J]$. We let $H(e)$ denote the binary entropy. Then

$$H(J|Z) \leq H(e) + p(e)\log(M-1)$$

- If further J is uniformly distributed, we have $H(J) = \log M$. Thus we have
$$\mathbb{P}[\psi(Z) \neq J] \geq 1 - \frac{I(Z;J) + \log 2}{\log M}$$
which is critical to our analysis.
- The mutual information $I(Z;J) := D\left(Q_{Z,J} \| Q_Z Q_J\right)$
- In this case, $I(Z;J) = \frac{1}{M}\sum_{i=1}^{M} D\left(\mathbb{P}_{\theta^j} \| \overline{\mathbb{Q}}\right)$

Fano's method

Proposition (Fano's method)

Suppose that $\{\theta^1, \ldots, \theta^M\}$ is a 2δ-separated set in ρ on $\theta(\mathcal{P})$, and suppose that J is uniformly distributed over the index set $[M]$, and $(Z|J=j) \sim \mathbb{P}_{\theta^j}$. Then the minimax risk is lower bounded as

$$\mathcal{M}(\theta(\mathcal{P}); \Phi \circ \rho) \geq \Phi(\delta) \left\{ 1 - \frac{I(Z;J) + \log 2}{\log M} \right\}$$

From the convexity of the Kullback–Leibler divergence we can yield $I(Z;J) \leq \frac{1}{M^2} \sum_{j,k=1}^{M} D\left(\mathbb{P}_{\theta^j} \| \mathbb{P}_{\theta^k}\right)$. To use this proposition, we need to

- construct a 2δ-separated set such that all pairs of distributions $\mathbb{P}_{\theta^i}$ and $\mathbb{P}_{\theta^j}$ are close on average
- decrease δ sufficiently to ensure that $\frac{I(Z;J)+\log 2}{\log M} \leq \frac{1}{2}$

Bounds based on local packings

Example: (Gaussian location family)

- Consider a 2δ-separated set of real-valued parameters, e.g., $\left\{\theta^1, \theta^2, \theta^3\right\} = \{0, 2\delta, -2\delta\}$.
- $D\left(\mathbb{P}^{1:n}_{\theta^j} \| \mathbb{P}^{1:n}_{\theta^k}\right) = \frac{n}{2\sigma^2}\left(\theta^j - \theta^k\right)^2 \leq \frac{2n\delta^2}{\sigma^2}$. Thus the mutual information is bounded.
- Choosing $\delta^2 = \frac{\sigma^2}{20n}$ ensures that $\frac{2n\delta^2/\sigma^2 + \log 2}{\log 3} < 0.75$. We have

$$\sup_{\theta \in \mathbb{R}} \mathbb{E}_\theta \left[(\widehat{\theta} - \theta)^2\right] \geq \frac{\delta^2}{4} = \frac{1}{80}\frac{\sigma^2}{n}$$

In more general cases, we need to construct a local 2δ-packing in parameter space such that

- the KL divergences are bounded: $\sqrt{D\left(\mathbb{P}_{\theta^j} \| \mathbb{P}_{\theta^k}\right)} \leq c\sqrt{n}\delta$.
- The cardinality of the packing is large enough: $\log M \geq 2\left\{c^2 n\delta^2 + \log 2\right\}$

Minimax risks for linear regression

Consider the standard linear regression model $\mathbf{y} = \mathbf{X}\theta^* + \boldsymbol{w}$ with fixed design and Gaussian noise. Find the minimax risk in the prediction (semi-)norm $\rho_{\mathbf{X}}\left(\widehat{\theta}, \theta^*\right) := \frac{\left\|\mathbf{X}\left(\widehat{\theta}-\theta^*\right)\right\|_2}{\sqrt{n}}$

- Consider the set $\{\gamma \in \operatorname{range}(\mathbf{X}) \mid \|\gamma\|_2 \leq 4\delta\sqrt{n}\}$. Let $\left\{\gamma^1, \ldots, \gamma^M\right\}$ be a $2\delta\sqrt{n}$-packing in the ℓ_2-norm. We have $\log M > r \log 2$.
- We thus have a collection of vectors of the form $\gamma^j = \mathbf{X}\theta^j$ satisfying $2\delta \leq \frac{\left\|\mathbf{X}\left(\theta^j-\theta^k\right)\right\|_2}{\sqrt{n}} \leq 8\delta$.
- The construction forms a local 2δ-packing with

$$D\left(\mathbb{P}_{\theta^j} \| \mathbb{P}_{\theta^k}\right) = \frac{1}{2\sigma^2}\left\|\mathbf{X}\left(\theta^j - \theta^k\right)\right\|_2^2 \leq \frac{32n\delta^2}{\sigma^2}$$

- Thus, $\inf_{\widehat{\theta}} \sup_{\theta \in \mathbb{R}^d} \mathbb{E}\left[\frac{1}{n}\|\mathbf{X}(\widehat{\theta} - \theta)\|_2^2\right] \geq \frac{\sigma^2}{128}\frac{\operatorname{rank}(\mathbf{X})}{n}$
- This lower bound is sharp, and can be achieved by the usual linear least-squares estimate.

Minimax risk for sparse linear regression

Consider the high-dimensional linear regression model $\mathbf{y} = \mathbf{X}\theta^* + \boldsymbol{w}$ with a prior that θ^* is sparse and $s \ll d$. Find the minimax risk in the prediction (semi-)norm $\rho_{\mathbf{X}}\left(\widehat{\theta}, \theta^*\right) := \frac{\left\|\mathbf{X}\left(\widehat{\theta}-\theta^*\right)\right\|_2}{\sqrt{n}}$

- Consider the "ℓ_0" ball $\mathbb{T}^d(s) := \left\{\theta \in \mathbb{R}^d \mid \|\theta\|_0 \leq s, \|\theta\|_2 \leq 1\right\}$. We can construct a 1/2-packing of this set with:
$$\log M \geq c_1 s \log\left(\frac{ed}{s}\right) \text{ (use Gilbert–Varshamov lemma).}$$
- Use the same rescaling procedure to form a 2δ-packing such that $\left\|\theta^j - \theta^k\right\|_2 \leq 4\delta$. By sparsity we have
$$\sqrt{D\left(\mathbb{P}_{\theta^j} \| \mathbb{P}_{\theta^k}\right)} = \frac{1}{\sqrt{2}\sigma}\left\|\mathbf{X}\left(\theta^j - \theta^k\right)\right\|_2 \leq \frac{\gamma_{2s}}{\sqrt{2}\sigma} 4\delta$$
- The choice $\delta^2 = c_1 \frac{\sigma^2}{400\gamma_{2s}^2 \cdot n} s \log \frac{ed}{s}$ guarantees that $c_1 s \log \frac{ed}{s} \geq 128 \frac{\gamma_{2s}^2}{\sigma^2} n\delta^2 + 2\log 2$. Therefore, we have an unimprovable lower bound:

$$\mathcal{M}\left(\mathbb{S}^d(s); \|\cdot\|_2\right) \gtrsim \frac{\sigma^2}{\gamma_{2s}^2} \frac{s \log \frac{ed}{s}}{n}$$

Local packings with Gaussian entropy bounds

This is a more delicate upper bound for Gaussian mixture, and it is a consequence of the maximum entropy property of the multivariate Gaussian distribution.

Lemma (Gaussian entropy bounds)

Suppose J is uniformly distributed over $[M] = \{1, \ldots, M\}$ and that Z conditioned on $J = j$ has a Gaussian distribution with covariance $\mathbf{\Sigma}^j$. Then the mutual information is upper bounded as

$$I(Z; J) \leq \frac{1}{2}\left\{\log\det\operatorname{cov}(Z) - \frac{1}{M}\sum_{j=1}^{M}\log\det\left(\Sigma^j\right)\right\}$$

Example (Lower bounds for PCA) We study the lower bounds for principal component analysis in the spiked covariance ensemble case.

- Suppose a random vector $x \in \mathbb{R}^d$ is is generated via $x \stackrel{\mathrm{d}}{=} \sqrt{v}\xi\theta^* + w$, where v is the SNR, $\|\theta^*\|_2 = 1$, $\xi \sim \mathcal{N}(0, 1)$ and $w \sim \mathcal{N}(0, \mathbf{I}_d)$ are independent.

Lower bounds for PCA

- The covariance: $\Sigma := \mathbf{I}_d + v\left(\theta^* \otimes \theta^*\right)$, and the vector θ^* is the unique maximal eigenvector of the covariance matrix.
- We construct the local packing by first finding a $1/2$-packing of unit sphere $\left\{\Delta^1, \ldots, \Delta^M\right\}$ with $\log M \geq (d-1)\log 2 \geq d/2$. Then let

$$\theta^j(\mathbf{U}) = \sqrt{1-\delta^2}\begin{bmatrix} 1 \\ 0_{d-1} \end{bmatrix} + \delta \begin{bmatrix} 0 \\ \mathbf{U}\Delta^j \end{bmatrix}$$

  for any given orthonormal matrix $\mathbf{U}$. The corresponding covariance matrix is $\mathbf{\Sigma}^j(\mathbf{U}) := \mathbf{I}_d + v\left(\theta^j(\mathbf{U}) \otimes \theta^j(\mathbf{U})\right)$.
- For any fixed $\mathbf{U}$ we have $\mathbb{P}\left[\psi\left(Z_1^n(\mathbf{U})\right) \neq J \mid \mathbf{U}\right] \geq 1 - \frac{nI(Z(\mathbf{U});J)+\log 2}{d/2}$. Suppose $\mathbf{U}$ is chosen uniformly at random. Use Gaussian entropy bounds we have

$$\mathbb{E}_{\mathbf{U}}[I(Z(\mathbf{U});J] \leq \frac{1}{2}\{\log\det \underbrace{\mathbb{E}_{\mathbf{U}}(\mathrm{cov}(Z(\mathbf{U})))}_{:=\Gamma} - \log(1+v)\}$$

Lower bounds for PCA

- Computing the entries of the expected covariance matrix Γ we find that Γ is diagonal with $\Gamma_{11} = 1 + v - v\delta^2$, $\Gamma_{2:d,2:d} = \left(1 + \frac{\delta^2 v}{d-1}\right) \mathbf{I}_{d-1}$.
- Putting together the pieces, we have

$$\log\det\Gamma = (d-1)\log\left(1+\frac{v\delta^2}{d-1}\right) + \log\left(1+v-v\delta^2\right)$$

- Thus we have

$$\begin{aligned} 2\mathbb{E}_{\mathbf{U}}[I(Z(\mathrm{U});J)] &\le (d-1)\log\left(1+\frac{v\delta^2}{d-1}\right) + \log\left(1-\frac{v}{1+v}\delta^2\right) \\ &\le \left(v - \frac{v}{1+v}\right)\delta^2 \\ &= \frac{v^2}{1+v}\delta^2 \end{aligned}$$

- Then we yield $\mathcal{M}\left(\mathrm{PCA}; \mathbb{S}^{d-1}, \|\cdot\|_2^2\right) \gtrsim \min\left\{\frac{1+v}{v^2}\frac{d}{n}, 1\right\}$

Yang–Barron version of Fano's method

Our previous analysis is largely based on the construction of local packing. But what if the local packing is hard to find?

Lemma (Yang–Barronmethod)

Let $N_{\mathrm{KL}}(\epsilon;\mathcal{P})$ denote the ϵ-covering number of $\mathcal{P}$ in the square-root KL divergence. Then the mutual information is upper bounded as

$$I(Z;J) \leq \inf_{\epsilon>0} \left\{\epsilon^2 + \log N_{\mathrm{KL}}(\epsilon;\mathcal{P})\right\}$$

- Notice the fact that:

$$I(Z;J) = \frac{1}{M}\sum_{j=1}^{M} D\left(\mathbb{P}_{\theta^j} \| \overline{\mathbb{Q}}\right) \overset{\text{(i)}}{\leq} \frac{1}{M}\sum_{j=1}^{M} D\left(\mathbb{P}_{\theta^j} \| \mathbb{Q}\right) \leq \max_{j=1,\ldots,M} D\left(\mathbb{P}_{\theta^j} \| \mathbb{Q}\right)$$

- Since the bound holds for any distribution $\mathbb{Q}$, we let $\left\{\gamma^1,\ldots,\gamma^N\right\}$ be an ϵ-covering of Ω in the square-root KL pseudo-distance, and then set $\mathbb{Q} = \frac{1}{N}\sum_{k=1}^{N}\mathbb{P}_{\gamma^k}$.

Yang–Barron version of Fano's method

- By construction, for each j we can find γ^k such that $D\left(\mathbb{P}_{\theta^j} \| \mathbb{P}_{\gamma^k}\right) \leq \epsilon^2$, thus

$$
\begin{aligned}
D\left(\mathbb{P}_{\theta^j} \| \mathbb{Q}\right) &= \mathbb{E}_{\theta^j}\left[\log \frac{d\mathbb{P}_{\theta_j}}{\frac{1}{N}\sum_{\ell=1}^{N} d\mathbb{P}_{\gamma^\ell}}\right] \\
&\leq \mathbb{E}_{\theta^j}\left[\log \frac{d\mathbb{P}_{\theta_j}}{\frac{1}{N} d\mathbb{P}_{\gamma^k}}\right] \\
&= D\left(\mathbb{P}_{\theta^j} \| \mathbb{P}_{\gamma^k}\right) + \log N \\
&\leq \epsilon^2 + \log N.
\end{aligned}
$$

- To use this lemma, we need to find a pair $(\delta, \epsilon) \in \mathbb{R}_+^2$ such that

$$
\log M(\delta; \rho, \Omega) \geq 2\left\{\epsilon^2 + \log N_{\mathrm{KL}}(\epsilon; \mathcal{P}) + \log 2\right\}
$$

. Finding such a pair can be accomplished via a two-step procedure
    - [A] First, choose $\epsilon_n > 0$ such that $\epsilon_n^2 \geq \log N_{\mathrm{KL}}\left(\epsilon_n; \mathcal{P}\right)$
    - [B] Then choose the largest δ_n that satisfies the lower bound $\log M\left(\delta_n; \rho, \Omega\right) \geq 4\epsilon_n^2 + 2\log 2$.

Density estimation revisited

Example (Density estimation revisited) Let us return to the problem of density estimation in the Hellinger metric in $\mathcal{F}_2$ space.

- From classical theory, it is known that the metric entropy of the class $\mathcal{F}_2$ in L^2-norm scales as $\log N(\delta; \mathcal{F}_2, \|\cdot\|_2) \asymp (1/\delta)^{1/2}$ for $\delta > 0$ sufficiently small.
- Step A. Given n i.i.d. samples, the square-root Kullback–Leibler divergence is multiplied by a factor of $\sqrt{n}$. We can set $\epsilon_n^2 \gtrsim \left(\frac{\sqrt{n}}{\epsilon_n}\right)^{1/2}$, e.g., $\epsilon_n^2 \asymp n^{1/5}$ to satisfy the first inequality.
- Step B. Then the second condition can be satisfied by choosing δ_n that $\left(\frac{1}{\delta_n}\right)^{1/2} \gtrsim n^{2/5}$, i.e., $\delta_n^2 \asymp n^{-4/5}$.
- The minimax risk thus is lower bounded by the order $n^{-4/5}$.