Book review:
High-Dimensional Statistics (Martin J. Wainwright)

Chapter 15: Minimax Lower Bounds

Suehyun Kim
29 April 2026

Causal Inference Lab.
Seoul National University

Outline

Chapter 15: Minimax lower bounds

- Basic framework
- Divergence measures
- From estimation to testing
- Le Cam's two-point method
- Fano's method

Outline

Basic framework

Notation

- $\mathcal{P}$: Class of distributions
- $\theta(\mathbb{P}) \in \Omega$: Functional on the space $\mathcal{P}$, i.e. $\theta : \mathcal{P} \to \Omega$

Goal Estimate $\theta(\mathbb{P})$ based on samples drawn from the unknown distribution $\mathbb{P}$.

- In certain cases, $\theta(\mathbb{P})$ uniquely determines $\mathbb{P}$, so that $\theta(\mathbb{P}_0) = \theta(\mathbb{P}_1) \iff \mathbb{P}_0 = \mathbb{P}_1$.
  - e.g. finite-dimensional parametric classes
- However, we may also be interested in functionals that do *not* specify the distribution.
  - e.g. smoothness of density $\mathbb{P} \mapsto \theta(\mathbb{P}) = \int_0^1 (f'(t))^2 dt$, mode of density $\theta(\mathbb{P}) = \arg\max_x f(x)$

Question Is there a *fundamental limit* to how well this goal can be achieved, *regardless of the estimation procedure*?

Basic framework

Problem setup

- Suppose we observe a random variable $X \sim \mathbb{P}$ such that $\theta(\mathbb{P}) = \theta^*$. Our goal is to estimate θ^* using X.
- To do so, we construct an estimator $\hat{\theta}$, which is a measurable function from $\mathcal{X} \to \Omega$.
- In order to assess the quality of estimators, we consider the quantity $\rho(\hat{\theta}, \theta^*)$, where $\rho : \Omega \times \Omega \to [0, \infty)$ is a semi-metric.
  - Note that $\rho(\hat{\theta}, \theta^*)$ is a random variable, since $\hat{\theta}$ is random (as a function of X).
- By taking expectations, we obtain the **risk function** $\mathbb{E}_{\mathbb{P}}[\rho(\hat{\theta}, \theta^*)]$.
  - The risk function is a function of the true parameter θ^*, and hence it measures the quality of an estimator at this point as a deterministic quantity.

Minimax risks

Comparison of estimators

Note that comparing estimators in a pointwise sense is pointless, since for any fixed θ^*, the optimal estimator would simply return θ^*, regardless of the observed data X. Therefore, it is necessary to consider the entire parameter space Ω.

1. **Minimax approach**: For a given estimator $\hat{\theta}$, compute the worst-case risk

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}}[\rho(\hat{\theta}, \theta(\mathbb{P}))].$$

2. **Bayesian approach**: View the unknown parameter θ^* as a random variable, and take the expectation of the risk with respect to a prior distribution.

Minimax risks

The estimator that is optimal in the minimax sense defines the **minimax risk**, namely

$$\mathfrak{M}(\theta(\mathcal{P}); \rho) := \inf_{\hat{\theta}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}}[\rho(\hat{\theta}, \theta(\mathbb{P}))].$$

- The infimum ranges over all possible estimators, i.e. measurable functions of the data.
- When the estimator is based on n iid samples from $\mathbb{P}$, we use the notation $\mathfrak{M}_n$.

We are often interested in evaluating minimax risks with respect to a *squared norm*. Hence, we generalize the notion by accommodating an increasing function $\Phi : [0, \infty) \to [0, \infty)$ and defining

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

- To obtain minimax risks wrt the mean-squared error, we set $\Phi(t) = t^2$.

Lower bounds

Lower bounds capture *intrinsic limits of a problem*, which helps in understanding the optimality of an estimator.

- Upper bounds (e.g., concentration inequalities): How well can we perform a task?
- Lower bounds: What is the best performance achievable by any method?

Question In what context do these lower bounds arise?

Example Cramer-Rao bound (information inequality)

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

Although this is not a minimax bound, it is still worthwhile to consider what the RHS quantifies.

Outline

Chapter 15: Minimax lower bounds

- Basic framework
- Divergence measures
- From estimation to testing
- Le Cam's two-point method
- Fano's method

Divergence measures

Before proceeding, we review different types of divergence measures between probability measures and their relationships.

1. **Total variation (TV) distance**
2. **Kullback-Leibler (KL) divergence**
3. **Hellinger distance**

cf. These three divergence measures are all instances of f-divergences.

Divergence measures

Total variation (TV) distance

$$\|\mathbb{P} - \mathbb{Q}\|_{TV} \coloneqq \sup_{A \subset \mathcal{X}} |\mathbb{P}(A) - \mathbb{Q}(A)|$$

- TV distance is the maximum difference in probabilities over all events. Hence, it frequently appears in probabilistic statements.
- Equivalent formalization:

$$\|\mathbb{P} - \mathbb{Q}\|_{TV} = \frac{1}{2} \int_x |p(x) - q(x)| \nu(dx)$$

- However, the TV distance tensorizes poorly, as we will see later.

Divergence measures

Kullback-Leibler (KL) divergence

$$D(\mathbb{Q} \parallel \mathbb{P}) \coloneqq \int_x q(x) \log \frac{q(x)}{p(x)} \nu(dx)$$

- Unlike the TV distance, the KL divergence *is not a metric*. Importantly, note that $D(\mathbb{Q} \parallel \mathbb{P}) \neq D(\mathbb{P} \parallel \mathbb{Q})$.
- KL divergence is closely related to notions of information and entropy:

$$D(\mathbb{Q} \parallel \mathbb{P}) = \underbrace{\mathbb{E}_{\mathbb{Q}}[-\log p(x)]}_{\text{cross entropy}} - \underbrace{H(\mathbb{Q})}_{\text{entropy}}.$$

- However, it diverges when one distribution is not dominated by the other.
  - i.e. it can be inappropriate when we want to measure distribution with different supports, such as $\mathcal{U}[\theta, \theta + 1]$.

Divergence measures

Squared Hellinger distance

$$H^2(\mathbb{P} \parallel \mathbb{Q}) := \int \left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2 \nu(dx)$$

- The squared Hellinger distance is the $L^2(\nu)$-norm between the square-root density functions.
- It is a valid metric, and it takes values in the interval $[0, 2]$.
- Geomteric interpretation: Euclidean distance between two points (distributions) on an infinite-dimensional sphere.

Relationships between divergence measures

When deriving minimax lower bounds, we will work with propositions stated in terms of the TV distance. However, the TV distance behaves badly, in the sense that it is difficult to express $\|\mathbb{P}^{1:n} - \mathbb{Q}^{1:n}\|_{TV}$ in terms of the individual distances $\|\mathbb{P}_i - \mathbb{Q}_i\|_{TV}$.

- Decoupling property of the KL divergence:

$$D(\mathbb{P}^{1:n} \parallel \mathbb{Q}^{1:n}) = \sum_{i=1}^{n} D(\mathbb{P}_i \parallel \mathbb{Q}_i).$$

- Squared Hellinger distance:

$$\frac{1}{2}H^2(\mathbb{P}^{1:n} \parallel \mathbb{Q}^{1:n}) = 1 - \prod_{i=1}^{n}\left(1 - \frac{1}{2}H^2(\mathbb{P}_i \parallel \mathbb{Q}_i)\right).$$

  - In the iid case, $\frac{1}{2}H^2(\mathbb{P}^{1:n} \parallel \mathbb{Q}^{1:n}) = 1 - \left(1 - \frac{1}{2}H^2(\mathbb{P}_1 \parallel \mathbb{Q}_1)\right)^n \leq \frac{1}{2}n\,H^2(\mathbb{P}_1 \parallel \mathbb{Q}_1)$.

Relationships between divergence measures

The following lemmas connect the TV distance with the KL divergence and the Hellinger distance:

Lemma 15.2 (Pinsker-Csiszár-Kullback inequality)

For all distributions $\mathbb{P}$ and $\mathbb{Q}$,

$$\|\mathbb{P} - \mathbb{Q}\|_{TV} \leq \sqrt{\frac{1}{2} D(\mathbb{Q} \parallel \mathbb{P})}.$$

Lemma 15.3 (Le Cam's inequality)

For all distributions $\mathbb{P}$ and $\mathbb{Q}$,

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

Outline

Chapter 15: Minimax lower bounds

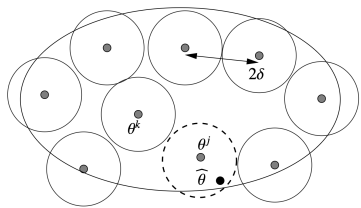
- Basic framework
- Divergence measures
- From estimation to testing
- Le Cam's two-point method
- Fano's method

From estimation to testing

Our aim is to develop methods for lower bounding the minimax risk. One way to do so is to reduce the problem to obtaining lower bounds in an M-ary hypothesis testing problem.

Setup

- Suppose that we have a 2δ-separated set $\{\theta^1, \dots, \theta^M\} \subset \theta(\mathcal{P})$, so that $\rho(\theta^j, \theta^k) \geq 2\delta$ for all $j \neq k$.
- For each θ^j, choose a representative distribution $\mathbb{P}_{\theta_j}$.
- Consequently, consider the M-ary hypothesis testing problem defined by the family of distributions $\{\mathbb{P}_{\theta^j}, j = 1, \dots, M\}$.



From estimation to testing

In particular, we generate a random variable Z by the following procedure:

(1) Sample a random integer J from the uniform distribution over $[M] := \{1, \dots, M\}$.

(2) Given $J = j$, sample $Z \sim \mathbb{P}_{\theta^j}$.

Given a sample Z, we consider the M-ary hypothesis testing problem of determining the randomly chosen index J.

- Denote by $\mathbb{Q}$ the joint distribution of the pair (Z, J).
- Then, the marginal distribution over Z is given by $\bar{\mathbb{Q}} := \frac{1}{M} \sum_{j=1}^{M} \mathbb{P}_{\theta^j}$.
- A **testing function** is a mapping $\psi : \mathcal{Z} \to [M]$, and the associated probability of error is given by $\mathbb{Q}[\psi(Z) \neq J]$.

From estimation to testing

The error probability of the hypothesis testing problem can be used to obtain a lower bound on the minimax risk as follows:

Proposition 15.1 (From estimation to testing)

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

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

where the infimum ranges over test functions.

Remarks.

- The term $\Phi(\delta)$ is maximized by choosing δ as large as possible.
- The testing error $\inf_{\psi} \mathbb{Q}[\psi(Z) \neq J]$ would increase as $\delta \to 0+$, since the underlying testing problem becomes more difficult.

From estimation to testing

Proof of Prop 15.1.

- For any $\mathbb{P} \in \mathcal{P}$ with parameter $\theta = \theta(\mathbb{P})$, we have

$$\mathbb{E}_{\mathbb{P}}\big[\Phi(\rho(\hat{\theta}, \theta))\big] \overset{\text{(i)}}{\geq} \Phi(\delta)\, \mathbb{P}\big[\rho(\hat{\theta}, \theta) \geq \delta\big] \overset{\text{(ii)}}{\geq} \Phi(\delta)\, \mathbb{P}\big[\rho(\hat{\theta}, \theta(\mathbb{P})) \geq \delta\big],$$

  where (i) follows from Markov's inequality, and (ii) follows from the increasing nature of Φ.

- Thus, it suffices to lower bound the quantity $\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{P}\big[\rho(\hat{\theta}, \theta(\mathbb{P})) \geq \delta\big]$.

- Note that

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

  so we have reduced the problem to lower bounding the quantity $\mathbb{Q}\big[\rho(\hat{\theta}, \theta^J) \geq \delta\big]$.

From estimation to testing

Proof of Prop 15.1. (continued)

- Observe that any estimator $\hat{\theta}$ can be used to define a test:

$$\psi(Z) := \operatorname*{arg\,min}_{\ell \in [M]} \rho(\theta^\ell, \hat{\theta}).$$

- Suppose the true parameter is θ^j. Then the event $\{\rho(\theta^j, \hat{\theta}) < \delta\}$ ensures that the test ψ is correct.
- This implies that

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

- Combined with our earlier argument, we have shown that

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathbb{P}}\big[\Phi(\rho(\hat{\theta}, \theta))\big] \geq \Phi(\delta)\, \mathbb{Q}\big[\psi(Z) \neq J\big].$$

- Finally, we take the infimum on both sides; the full infimum on the RHS can only be smaller. □

Outline

Chapter 15: Minimax lower bounds

- Basic framework
- Divergence measures
- From estimation to testing
- Le Cam's two-point method
- Fano's method

Binary testing and TV distance

The first approach for computing $\inf_\psi \mathbb{Q}[\psi(Z) \neq J]$ is **Le Cam's two-point method**, which reduces the problem to a binary testing problem of the simplest form.

More specifically, we establish the connection between binary testing and the TV distance:

- Consider a binary testing problem with equally weighted hypotheses, so that we observe Z drawn according to $\bar{\mathbb{Q}} := \frac{1}{2}\mathbb{P}_0 + \frac{1}{2}\mathbb{P}_1$.
- For a given decision rule $\psi : \mathcal{Z} \to \{0, 1\}$, the associated probability of error is given by

$$\mathbb{Q}[\psi(Z) \neq J] = \frac{1}{2}\mathbb{P}_0[\psi(Z) \neq 0] + \frac{1}{2}\mathbb{P}_1[\psi(Z) \neq 1].$$

- In the binary case, the following relationship holds:

$$\inf_\psi \mathbb{Q}[\psi(Z) \neq J] = \frac{1}{2}\left\{1 - \|\mathbb{P}_1 - \mathbb{P}_0\|_{TV}\right\}.$$

Binary testing and TV distance

Proof of $\inf_\psi \mathbb{Q}[\psi(Z) \neq J] = \frac{1}{2}\{1 - \|\mathbb{P}_1 - \mathbb{P}_0\|_{TV}\}$.

- Note that any decision rule ψ is uniquely determined by the set $A = \{x \in \mathcal{X} \mid \psi(x) = 1\}$.
- Thus, we have

$$\sup_\psi \mathbb{Q}[\psi(Z) = J] = \sup_{A \subseteq \mathcal{X}} \left\{ \frac{1}{2}\mathbb{P}_1(A) + \frac{1}{2}\mathbb{P}_0(A^c) \right\} = \frac{1}{2} \sup_{A \subseteq \mathcal{X}} \{\mathbb{P}_1(A) - \mathbb{P}_0(A)\} + \frac{1}{2}.$$

- The claim follows directly from the definition of the TV norm, since we have $\sup_\psi \mathbb{Q}[\psi(Z) = J] = 1 - \inf_\psi \mathbb{Q}[\psi(Z) \neq J]$.

Remark. We have $0 \leq \inf_\psi \mathbb{Q}[\psi(Z) \neq J] \leq \frac{1}{2}$, where the upper bound $\frac{1}{2}$ is achieved when $\mathbb{P}_1 = \mathbb{P}_0$ so that the hypotheses are completely indistinguishable.

Le Cam's two-point method

Plugging this result into **Proposition 15.1** provides one avenue for deriving minimax lower bounds:

Proposition (Le Cam's two-point method)

For any pair of distributions $\mathbb{P}_0, \mathbb{P}_1 \in \mathcal{P}$ such that $\rho(\theta(\mathbb{P}_0), \theta(\mathbb{P}_1)) \geq 2\delta$, we have

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

Le Cam's two-point method: Examples

Example 15.4 (Gaussian location family)

- For a fixed variance σ^2, let $\mathbb{P}_\theta$ be $\mathcal{N}(\theta, \sigma^2)$-distributed.
- We consider the problem of estimating θ under the absolute error $|\hat{\theta} - \theta|$ or the squared error $(\hat{\theta} - \theta)^2$ using iid samples drawn from $\mathcal{N}(\theta, \sigma^2)$.
- Consider $\mathbb{P}_0^n$ and $\mathbb{P}_\theta^n$, where $\theta = 2\delta$ for some $\delta > 0$ to be chosen later appropriately.
- In order to apply the two-point Le Cam bound, we need to bound the TV distance $\|\mathbb{P}_\theta^n - \mathbb{P}_0^n\|_{TV}$.
- One can show that

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

Le Cam's two-point method: Examples

Example 15.4 (Gaussian location family; continued)

- Setting $\delta = \frac{\sigma}{2\sqrt{n}}$ yields

$$\inf_{\hat{\theta}} \sup_{\theta \in \mathbb{R}} \mathbb{E}_\theta[|\hat{\theta} - \theta|] \geq \frac{\delta}{2}\left\{1 - \frac{1}{2}\sqrt{e-1}\right\} \geq \frac{\delta}{6} = \frac{\sigma}{12\sqrt{n}},$$

and

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

Remark. Although the pre-factors $1/12$ and $1/24$ are not optimal, the scalings $\sigma/\sqrt{n}$ and σ^2/n are sharp. For instance, the sample mean $\bar{\theta}_n := \frac{1}{n}\sum_{i=1}^n Y_i$ satisfies the bounds

$$\sup_{\theta \in \mathbb{R}} \mathbb{E}_\theta[|\bar{\theta}_n - \theta|] = \sqrt{\frac{2}{\pi}}\frac{\sigma}{\sqrt{n}}, \quad \sup_{\theta \in \mathbb{R}} \mathbb{E}_\theta[(\bar{\theta}_n - \theta)^2] = \frac{\sigma^2}{n}.$$

Le Cam's two-point method: Examples

Example 15.5 (Uniform location family)

- We consider the problem of estimating θ in the uniform location family $\{\mathbb{U}_\theta : \theta \in \mathbb{R}\}$, where $\mathbb{U}_\theta$ is uniform over $[\theta, \theta + 1]$.
- Here, we *cannot* use the KL divergence to bound the TV distance, since $D(\mathbb{U}_\theta \,\|\, \mathbb{U}_{\theta'}) = \infty$ whenever $\theta \neq \theta'$; we use the Hellinger distance instead.
- One can easily compute that

$$H^2(\mathbb{U}_\theta \,\|\, \mathbb{U}_{\theta'}) = \begin{cases} 2 & |\theta' - \theta| > 1 \\ 2\,|\theta' - \theta| & \text{otherwise.} \end{cases}$$

- By the property of Hellinger distance, we have $\frac{1}{2} H^2(\mathbb{U}_\theta^n \,\|\, \mathbb{U}_{\theta'}^n) \leq \frac{n}{2} \cdot 2|\theta' - \theta| = \frac{1}{4}$.

Le Cam's two-point method: Examples

Example 15.5 (Uniform location family; continued)

- In conjunction with Lemma 15.3, we obtain

$$\|\mathbb{U}_\theta^n - \mathbb{U}_{\theta'}^n\|_{TV}^2 \le H^2(\mathbb{U}_\theta^n \,\|\, \mathbb{U}_{\theta'}^n) \le \frac{1}{2}.$$

- Therefore, the minimax risk is lower bounded as

$$\inf_{\hat{\theta}} \sup_{\theta \in \mathbb{R}} \mathbb{E}_\theta[(\hat{\theta} - \theta)^2] \ge \frac{1}{128}\left(1 - \frac{1}{\sqrt{2}}\right)\frac{1}{n^2}.$$

Remarks.

- Whereas n^{-1}-order decay of the mean-squared error is typical for parametric problems with certain regularity conditions, here we have established a faster n^{-2} rate.
- In fact, this n^{-2} rate is optimal, achieved for instance by the estimator $\hat{\theta} = \min\{Y_1, \dots, Y_n\}$.

Le Cam's convex hull method

Thus far, we have compared two distributions in order to apply the two-point method.

We can generalize this approach by taking the convex hulls of two classes of distributions, which leads to a smaller separation between the classes and thus better lower bounds.

Lemma 15.9 (Le Cam)

Consider two subsets $\mathcal{P}_0$ and $\mathcal{P}_1$ that are 2δ-separated, so that

$$\rho(\theta(\mathbb{P}_0), \theta(\mathbb{P}_1)) \geq 2\delta \quad \text{for all } \mathbb{P}_0 \in \mathcal{P}_0 \text{ and } \mathbb{P}_1 \in \mathcal{P}_1.$$

Then, any estimator $\hat{\theta}$ has worst-case risk at least

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

Remark. Note that this lemma does *not* involve the function Φ.

Le Cam's convex hull method: Examples

In order to see how taking the convex hulls can decrease the TV distance, we revisit the Gaussian location model example.

Example 15.10 (Sharpened bounds for Gaussian location family)

- Set $\theta = 2\delta$ as before, and consider the two families $\mathcal{P}_0 = \{\mathbb{P}_0^n\}$ and $\mathcal{P}_1 = \{\mathbb{P}_\theta^n, \mathbb{P}_{-\theta}^n\}$.
- Note that the mixture distribution $\bar{\mathbb{P}} := \frac{1}{2}\mathbb{P}_\theta^n + \frac{1}{2}\mathbb{P}_{-\theta}^n$ belongs to $\mathrm{conv}(\mathcal{P}_1)$.
- Using similar techniques, one can derive

$$\|\bar{\mathbb{P}} - \mathbb{P}_0^n\|_{TV}^2 \leq \frac{1}{4}\left\{e^{\frac{1}{2}\left(\frac{\sqrt{n}\theta}{\sigma}\right)^4} - 1\right\} = \frac{1}{4}\left\{e^{\frac{1}{2}\left(\frac{2\sqrt{n}\delta}{\sigma}\right)^4} - 1\right\}.$$

Le Cam's convex hull method: Examples

Example 15.10 (Sharpened bounds for Gaussian location family; continued)

- Setting $\delta = \frac{\sigma t}{2\sqrt{n}}$, the convex hull Le Cam bound yields

$$\inf_{\hat{\theta}} \sup_{\theta \in \mathbb{R}} \mathbb{E}_\theta[|\hat{\theta} - \theta|] \geq \frac{\sigma}{4\sqrt{n}} \sup_{t>0} \left\{ t \left(1 - \frac{1}{2} \sqrt{e^{t^4/2} - 1} \right) \right\} \geq \frac{3}{20} \frac{\sigma}{\sqrt{n}}.$$

This bound is an improvement over our original bound from Example 15.4, which has the pre-factor of $\frac{1}{12} \approx 0.08$, as opposed to $\frac{3}{20} = 0.15$ obtained from this analysis.

Outline

Chapter 15: Minimax lower bounds

- Basic framework
- Divergence measures
- From estimation to testing
- Le Cam's two-point method
- Fano's method

KL divergence and mutual information

Now we introduce **Fano's method**, which exploits results from information theory to provide a lower bound for the testing error $\inf_\psi \mathbb{Q}[\psi(Z) \neq J]$.

- Recall that we seek to lower bound the error probability in a hypothesis testing problem, where we draw Z from $\{\mathbb{P}_{\theta^1}, \dots, \mathbb{P}_{\theta^M}\}$ by choosing an index $J \in [M]$ uniformly at random.
- The observation then follows a mixture distribution $\bar{\mathbb{Q}} := \frac{1}{M}\sum_{j=1}^{M} \mathbb{P}_{\theta^j}$.
- The difficulty of this problem would depend on the amount of dependece between the observation Z and the unknown random index J; the problem would be pointless if, in the extreme case, Z were independent of J.
- This leads to the concept of **mutual information**, which quantifies the dependence between distributions.

KL divergence and mutual information

Mutual information (between two r.v.s)

$$I(Z;J) \coloneqq D(\mathbb{Q}_{Z,J} \parallel \mathbb{Q}_Z \mathbb{Q}_J)$$

- In words, mutual information is the *KL divergence of the joint distribution and the product of marginals*.
- By standard properties of the KL divergence, we have $I(Z,J) \geq 0$, where the equality holds iff Z and J are independent.
- Under our setup, we have $I(Z;J) = \frac{1}{M} \sum_{j=1}^{M} D(\mathbb{P}_{\theta_j} \parallel \bar{\mathbb{Q}})$. By the convexity of KL divergence, we obtain

$$I(Z;J) \leq \frac{1}{M^2} \sum_{j,k=1}^{M} D(\mathbb{P}_{\theta_j} \parallel \mathbb{P}_{\theta_k}). \qquad (\bigstar)$$

Fano lower bound on minimax risk

The **Fano lower bound** controls the error probability in an M-ary testing problem, applicable when J is *uniformly distributed over the index set*:

$$\mathbb{P}[\psi(Z) \neq J] \geq 1 - \frac{I(Z;J) + \log 2}{\log M}.$$

Combining with our earlier results, we obtain a minimax lower bound.

Fano lower bound on minimax risk

Proposition 15.12

Let $\{\theta^1, \dots, \theta^M\}$ be a 2δ-separated set in the ρ semi-metric on $\Theta(\mathcal{P})$, and suppose that J is uniformly distributed over the index set $\{1, \dots, M\}$, and $(Z \mid J = j) \sim \mathbb{P}_{\theta_j}$.

Then for any increasing function $\Phi : [0, \infty) \to [0, \infty)$, the minimax risk is lower bounded as

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

where $I(Z;J)$ is the mutual information between Z and J.

Remarks. As $\delta \to 0+$,

- The 2δ-separation criterion becomes milder, so that $M \equiv M(2\delta)$ increases.
- The mutual information $I(Z;J)$ will decrease, since $J \in [M(2\delta)]$ can take on a larger number of potential values.

Bounds based on local packings

One way to lower bound the minimax risk as $\mathfrak{M}(\theta(\mathcal{P}); \Phi \circ \rho) \geq \frac{1}{2}\Phi(\delta)$ is to construct a 2δ-separated local packing in Ω satisfying the following:

(i) For some quantity c, the KL divergences satisfy the *uniform* upper bound

$$\sqrt{D(\mathbb{P}_{\theta_j} \,\|\, \mathbb{P}_{\theta_k})} \leq c\sqrt{n}\,\delta \quad \text{for all } j \neq k.$$

(ii) The size of the packing satisfies

$$\log M(2\delta) \geq 2\{c^2 n\delta^2 + \log 2\}.$$

This approach is referred to as the *Generalized Fano* method (although it is rather a substantial weakening of the Fano bound).

Fano's method: Examples

Example 15.13 (Gaussian location model via Fano method)

- Consider the 2δ-separated set of parameters $\{\theta^1, \theta^2, \theta^3\} = \{0, 2\delta, -2\delta\}$.
- Since $\mathbb{P}_{\theta_j} = \mathcal{N}(\theta^j, \sigma^2)$, we have

$$D(\mathbb{P}_{\theta_j}^{1:n} \,\|\, \mathbb{P}_{\theta_k}^{1:n}) = \frac{n}{2\sigma^2}(\theta^j - \theta^k)^2 \leq \frac{8n\delta^2}{\sigma^2} \quad \text{for all } j, k = 1, 2, 3.$$

- The KL divergence bound ($\bigstar$) ensures that $I(Z; J_\delta) \leq \frac{8n\delta^2}{\sigma^2}$, and choosing $\delta^2 = \frac{\sigma^2}{80n}$ ensures that $\frac{2n\delta^2/\sigma^2 + \log 2}{\log 3} < 0.75$.
- Consequently, the Fano bound with $\Phi(t) = t^2$ implies that

$$\sup_{\theta \in \mathbb{R}} \mathbb{E}_\theta[(\hat{\theta} - \theta)^2] \geq \frac{\delta^2}{4} = \frac{1}{320}\frac{\sigma^2}{n}.$$

Other approaches to bound $I(Z;J)$

So far, we have used the convexity-based upper bound ($\bigstar$) on $I(Z;J)$. If the conditional distribution of Z given J is Gaussian, we can use the following alternative bound:

Lemma 15.17

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

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

Other approaches to bound $I(Z;J)$

Moreover, the *Yang–Barron version* of Fano's method relies on the covering number of the model class, rather than constructing a local packing:

Lemma 15.18 (Yang–Barron method)

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

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

This approach is particularly useful for nonparametric problems, where the covering number can be controlled via metric entropy bounds.