Group Study and Seminar Series (Summer 20)

Minimax Lower Bounds

Presenter: Hao Liang

The Chinese University of Hong Kong, Shenzhen, China

July 2, 2020

Mainly based on:

Wainwright, M. J. (2019). High-dimensional statistics: A non-asymptotic viewpoint (Vol. 48). Cambridge University Press.

John, Duchi (2019). Lecture Notes for Statistics 311/Electrical Engineering 377.

Outline

Outline

Example 0: Gaussian location family

- $\left\{\mathcal{N}\left(\theta, \sigma^2\right), \theta \in \mathbb{R}\right\}$: a normal distribution family with fixed variance σ^2
- Data: a collection $Z = (Y_1, \ldots, Y_n)$ with Y_i i.i.d. $\sim \mathcal{N}\left(\theta, \sigma^2\right)$
- Method: estimate unknown θ^* via an estimator $\widehat{\theta}(Z)$
- Performance measure: risk $R(\widehat{\theta}, \theta^*)$
- How does $\widetilde{\theta}_n := \frac{1}{n}\sum_{i=1}^n Y_i$ perform?
- Upper bound provides worst-case performance guarantee

$$\sup_{\theta \in \mathbb{R}} R(\widetilde{\theta}_n, \theta) \leq \frac{\sigma^2}{n}$$

Example 0: Gaussian location family

- But how to answer the following questions?
  - Can this analysis be improved? Or does $\widetilde{\theta}_n$ actually satisfy better bounds?
  - Can any estimator improve upon the bound?
- Both questions ask about some form of optimality(switch orders?)
  - Optimality of an estimator
  - Optimality of a bound
- A positive answer consists in
  - Finding a better proof for $\widetilde{\theta}_n$
  - Finding a better estimator, together with a proof that it performs better

Example 0: Gaussian location family

- Lower bound may provide negative answer to both questions

$$\inf_{\widehat{\theta}} \sup_{\theta \in \mathbb{R}} \mathbb{E}_\theta \left[(\widehat{\theta} - \theta)^2 \right] \geq \Theta(\frac{\sigma^2}{n})$$

- Any estimator suffers risk at least $\Theta(\frac{\sigma^2}{n})$ in the worst case
- Recall that $\widetilde{\theta}_n$ suffers risk at most $\Theta(\frac{\sigma^2}{n})$ in the worst case
- Both the upper bound $\Theta(\frac{\sigma^2}{n})$ and the estimator $\widetilde{\theta}_n$ are not improvable!

Outline

Statistical decision theory

- $\mathcal{P} = \{\mathbb{P}_\theta | \theta \in \Omega\}$: A parametric family with parameter θ
- Data/samples: Y_i i.i.d. $\sim \mathbb{P}_\theta$ or $Y^n = (Y_1, \ldots, Y_n) \sim \mathbb{P}_\theta^n$
- Decision rule
  - (Point) Estimation: estimate θ^* via an estimator $\widehat{\theta}(Y^n)$, $\widehat{\theta} : \mathcal{X}^n \to \Omega$
  - (Hypothesis) Test: nature randomly choose index $J = j$, decide $j \in \{1, 2, ..., M\}$ via an test function $\psi(Y^n)$, where $Y^n \sim \mathbb{P}_{\theta^j}^n$
- Loss function $\rho\left(\widehat{\theta}, \theta^*\right)$
  - Absolute loss $\rho\left(\widehat{\theta}, \theta^*\right) = |\widehat{\theta} - \theta^*|$
  - Squared loss $\rho\left(\widehat{\theta}, \theta^*\right) = (\widehat{\theta} - \theta^*)^2$
- Risk $R(\widehat{\theta}, \theta^*) = \mathbb{E}_{\mathbb{P}}\left[\rho\left(\widehat{\theta}(Y^n), \theta^*\right)\right]$

Information theory

- Entropy $H(X) := \int_{\mathcal{X}} p_X(u) \log \frac{1}{p_X(u)} du$
- Relative entropy/KL divergence $D(\mathbb{P}_X \| \mathbb{P}_Y) := \int_{\mathcal{X}} p_X(u) \log \frac{p_X(u)}{p_Y(u)} du$
  - $D(\mathcal{N}(\theta_1, \sigma_1^2), \mathcal{N}(\theta_2, \sigma_2^2)) = \log \frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}$
- Mutual information $I(X;Y) := D(\mathbb{P}_{X,Y} \| \mathbb{P}_X \mathbb{P}_Y)$
  - KL divergence form: $I(X;Y) = \mathbb{E}_X D(\mathbb{P}_{Y|X} \| \mathbb{P}_Y) = \sum_x \mathbb{P}(x) D(\mathbb{P}_{Y|X=x} \| \mathbb{P}_Y)$
- Fano's inequality provides a lower bound on the error in a M-ary testing problem

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

Outline

Basic framework

- Given a class of distributions $\mathcal{P}$ and $\theta : \mathcal{P} \to \Omega$ is a functional mapping distributions to a parameter $\theta(\mathbb{P})$
- For parametric classes, $\theta(\mathbb{P})$ uniquely determines the underlying distribution $\mathbb{P}$, write $\mathcal{P} = \{\mathbb{P}_\theta | \theta \in \Omega\}$ (e.g. Gaussian location family)
- The viewpoint of estimating functionals here is more general than a parametric family (e.g. estimating the mode of the density $\theta(\mathbb{P}) = \arg\max_{x\in[0,1]} f(x)$)

Minimax risk

- Given a sample $X \sim \mathbb{P}_{\theta^*}$, θ^* fixed but unknown
- The goal of an estimator $\widehat{\theta}$ is to estimate θ^* based on X, write also $\widehat{\theta} \equiv \widehat{\theta}(X)$
- Let $\rho : \Omega \times \Omega \to [0, \infty)$ be a semi-metric, consider r.v. $\rho\left(\widehat{\theta}, \theta^*\right)$
- Taking expectations over X yields the deterministic quantity $R(\widehat{\theta}, \theta^*) := \mathbb{E}_{\mathbb{P}}\left[\rho\left(\widehat{\theta}, \theta^*\right)\right]$
- Typically referred to as the risk function associated with $\widehat{\theta}$

Minimax risk

- Goal: $\min_{\widehat{\theta}} R(\widehat{\theta}, \theta^*), \forall\theta^*$?
- Multi-objective optimization problem
- Two ways to deal with this issue: Bayesian approach and minimax approach
  - Bayesian approach: taking average over parameters

$$\inf_{\widehat{\theta}} \mathbb{E}_{\theta^* \sim \pi}[R(\widehat{\theta}, \theta^*)]$$

  - Minimax approach: adversarial perspective

$$\inf_{\widehat{\theta}} \sup_{\theta^*} R(\widehat{\theta}, \theta^*)$$

Minimax risk

- More generally

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

- The ρ-minimax risk

$$\mathfrak{M}(\theta(\mathcal{P}); \rho) := \inf_{\widehat{\theta}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\rho(\widehat{\theta}, \theta(\mathbb{P}))] \tag{2}$$

- Introduce a non-decreasing function $\Phi : [0, \infty) \to [0, \infty)$,

$$\mathfrak{M}(\theta(\mathcal{P}); \Phi \circ \rho) := \inf_{\widehat{\theta}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\Phi(\rho(\widehat{\theta}, \theta(\mathbb{P})))] \tag{3}$$

From estimation to testing

- Developing methods for lower bounding the minimax risk
- Reduction to the problem of obtaining lower bounds for certain testing problems
- Start with constructing such testing problems as follows:
- Suppose that $\left\{\theta^1, \ldots, \theta^M\right\} \subseteq \theta(\mathcal{P})$ is a 2δ-separated set, i.e., $\rho\left(\theta^j, \theta^k\right) \geq 2\delta$ for all $j \neq k$
- For each θ^j, choose some representative distribution $\mathbb{P}_{\theta^j}$ for which $\theta\left(\mathbb{P}_{\theta^j}\right) = \theta^j$

From estimation to testing

- Generate a random variable Z by the following procedure:
  - Sample a random integer J from the uniform distribution over the index set $[M] := \{1, \ldots, M\}$
  - Given $J = j$, sample $Z \sim \mathbb{P}_{\theta^j}$
- let $\mathbb{Q}$ denote the joint distribution of the pair (Z, J), then the marginal distribution over Z is $\overline{\mathbb{Q}} := \frac{1}{M} \sum_{j=1}^{M} \mathbb{P}_{\theta^j}$
- Consider the M-ary hypothesis testing problem of determining J based on a sample Z
- A testing function for this problem is a mapping $\psi : \mathcal{Z} \to [M]$

From estimation to testing

- The probability of error of ψ is $\mathbb{Q}[\psi(Z) \neq J]$, can be used to obtain lower bound

Proposition 1.

(From estimation to testing) For any non-decreasing function Φ and choice of 2δ-separated set, the minimax risk is lower bounded as

$$\mathfrak{M}(\theta(\mathcal{P}), \Phi \circ \rho) \geq \Phi(\delta) \inf_{\psi} \mathbb{Q}[\psi(Z) \neq J], \tag{4}$$

where the infimum ranges over all test functions.

Remarks

- The r.h.s. of the bound involves two terms, and both of them depends δ
  - The function Φ is decreasing in δ
  - As δ increases, M decreases
  - The underlying testing problem becomes easier, $\mathbb{Q}[\psi(Z) \neq J]$ decreases
- Choose a sufficiently small δ^* to ensure that this testing error is at least 0.5,

$$\mathfrak{M}(\theta(\mathcal{P}), \Phi \circ \rho) \geq \frac{1}{2}\Phi\left(\delta^*\right) \tag{5}$$

Proof

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

$$\mathbb{E}_{\mathbb{P}}[\Phi(\rho(\widehat{\theta}, \theta))] \overset{\text{(i)}}{\geq} \Phi(\delta)\mathbb{P}[\Phi(\rho(\widehat{\theta}, \theta)) \geq \Phi(\delta)] \overset{\text{(ii)}}{\geq} \Phi(\delta)\mathbb{P}[\rho(\widehat{\theta}, \theta) \geq \delta]. \tag{6}$$

- It suffices to lower bound $\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{P}[\rho(\widehat{\theta}, \theta(\mathbb{P})) \geq \delta]$
- Recall that $\mathbb{Q}$ denotes the joint distribution over the pair (Z, J),

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{P}[\rho(\widehat{\theta}, \theta(\mathbb{P})) \geq \delta] \geq \frac{1}{M} \sum_{j=1}^{M} \mathbb{P}_{\theta^j} \left[\rho\left(\widehat{\theta}, \theta^j\right) \geq \delta\right] = \mathbb{Q}[\rho\left(\widehat{\theta}, \theta^J\right) \geq \delta]. \tag{7}$$

- Reduced to lower bounding $\mathbb{Q}\left[\rho\left(\widehat{\theta}, \theta^J\right) \geq \delta\right]$

Proof

- Construct a test based on the estimator $\widehat{\theta}$ via

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

- Suppose that the true parameter is θ^j, then the event $\left\{\rho\left(\theta^j, \widehat{\theta}\right) < \delta\right\}$ ensures that the test is correct

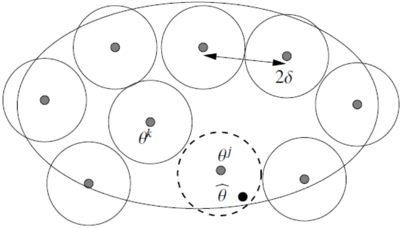

Figure: geometry of this argument

Proof

- Conditioned on $J = j$, $\left\{\rho\left(\theta^j, \widehat{\theta}\right) < \delta\right\} \subseteq \{\psi(Z) = j\}$, implying

$$\mathbb{P}_{\theta j}\left[\rho\left(\widehat{\theta}, \theta^j\right) \geq \delta\right] \geq \mathbb{P}_{\theta j}[\psi(Z) \neq j] \tag{8}$$

- Taking averages over index j,

$$\mathbb{Q}\left[\rho\left(\widehat{\theta}, \theta^J\right) \geq \delta\right] = \frac{1}{M}\sum_{j=1}^{M} \mathbb{P}_{\theta^j}\left[\rho\left(\widehat{\theta}, \theta^j\right) \geq \delta\right] \geq \mathbb{Q}[\psi(Z) \neq J] \tag{9}$$

- Combined with the earlier argument, $\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}}[\Phi(\rho(\widehat{\theta}, \theta))] \geq \Phi(\delta)\mathbb{Q}[\psi(Z) \neq J]$
- Take the infimum over all estimators $\widehat{\theta}$ on the l.h.s., and the the infimum over the induced set of tests on the r.h.s.
- Finally notice that the full infimum over all tests can only be smaller, from which the claim follows

Some divergence measures

- Three important measures
  - Total variation (TV) distance $\|\mathbb{P}-\mathbb{Q}\|_{\mathrm{TV}} := \sup_{A\subseteq X} |\mathbb{P}(A)-\mathbb{Q}(A)| = \frac{1}{2}\int_{\mathcal{X}} |p(x)-q(x)|dx$
  - KL divergence $D(\mathbb{P}\|\mathbb{Q}) = \int_{\mathcal{X}} p(x)\log\frac{p(x)}{q(x)}dx$
  - Squared Hellinger distance $H^2(\mathbb{P}\|\mathbb{Q}) := \int_{\mathcal{X}} (\sqrt{p(x)}-\sqrt{q(x)})^2 dx = 2-2\int_{\mathcal{X}}\sqrt{p(x)q(x)}dx$
- The second and third distance can be used to upper bound TV distance
  - Pinsker's inequality $\|\mathbb{P}-\mathbb{Q}\|_{\mathrm{TV}} \le \sqrt{\frac{1}{2}D(\mathbb{Q}\|\mathbb{P})}$
  - Le Cam's inequality $\|\mathbb{P}-\mathbb{Q}\|_{\mathrm{TV}} \le H(\mathbb{P}\|\mathbb{Q})\sqrt{1-\frac{H^2(\mathbb{P}\|\mathbb{Q})}{4}}$
- These inequalities are useful when dealing with product distributions

Some divergence measures

- Let $\mathbb{P}^{1:n} = \bigotimes_{i=1}^{n} \mathbb{P}_i$ be the product distribution of $(\mathbb{P}_1, \ldots, \mathbb{P}_n)$ defined on $\mathcal{X}^n$
- What's the expression of $\mathrm{Div}(\mathbb{Q}^{1:n} \| \mathbb{P}^{1:n})$ in terms of $\mathrm{Div}(\mathbb{Q}_i \| \mathbb{P}_i)$?
- The TV distance behaves badly: difficult to decouple
- The KL divergence exhibits a very attractive decoupling property,

$$D\left(\mathbb{P}^{1:n} \| \mathbb{Q}^{1:n}\right) = \sum_{i=1}^{n} D\left(\mathbb{P}_i \| \mathbb{Q}_i\right) \tag{10}$$

- The squared Hellinger distance does not decouple in a simple way, but

$$\frac{1}{2} H^2\left(\mathbb{P}^{1:n} \| \mathbb{Q}^{1:n}\right) = 1 - \prod_{i=1}^{n}\left(1 - \frac{1}{2} H^2\left(\mathbb{P}_i \| \mathbb{Q}_i\right)\right) \tag{11}$$

Some divergence measures

- In the i.i.d. case where $\mathbb{P}_i = \mathbb{P}_1$ and $\mathbb{Q}_i = \mathbb{Q}_1$ for all i

$$D\left(\mathbb{P}^{1:n} \| \mathbb{Q}^{1:n}\right) = nD\left(\mathbb{P}_1 \| \mathbb{Q}_1\right) \tag{12}$$

$$\frac{1}{2}H^2\left(\mathbb{P}^{1:n} \| \mathbb{Q}^{1:n}\right) = 1 - \left(1 - \frac{1}{2}H^2\left(\mathbb{P}_1 \| \mathbb{Q}_1\right)\right)^n \leq \frac{1}{2}nH^2\left(\mathbb{P}_1 \| \mathbb{Q}_1\right) \tag{13}$$

- Combined with the previous inequalities,
  - Pinsker's inequality in the i.i.d. case

$$\|\mathbb{P}^{1:n} - \mathbb{Q}^{1:n}\|_{\mathrm{TV}} \leq \sqrt{\frac{n}{2}D(\mathbb{P}_1 \| \mathbb{Q}_1)} \tag{14}$$

  - Le Cam's inequality in the i.i.d. case

$$\|\mathbb{P}^{1:n} - \mathbb{Q}^{1:n}\|_{\mathrm{TV}} \leq H(\mathbb{P}^{1:n} \| \mathbb{Q}^{1:n})\sqrt{1 - \frac{H^2(\mathbb{P}^{1:n} \| \mathbb{Q}^{1:n})}{4}} \leq H(\mathbb{P}^{1:n} \| \mathbb{Q}^{1:n}) \leq \sqrt{n}H(\mathbb{P}_1 \| \mathbb{Q}_1) \tag{15}$$

Outline

Binary test and Le Cam's method

- Recall the reduction from estimation to testing

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

- Le Cam's method is to consider the simplest type of testing problem - binary hypothesis test, which involves only two distributions
- In a binary testing problem with equally weighted hypotheses, $Z \sim \bar{Q} := \frac{1}{2}\mathbb{P}_0 + \frac{1}{2}\mathbb{P}_1$
- For a given decision rule $\psi : \mathcal{Z} \to \{0, 1\}$, the associated probability of error is

$$\mathbb{Q}[\psi(Z) \neq J] = \frac{1}{2}\mathbb{P}_0[\psi(Z) \neq 0] + \frac{1}{2}\mathbb{P}_1[\psi(Z) \neq 1] \tag{16}$$

Bayes error and TV distance

- Take the infimum over all decision rules yields Bayes error
- Recall the definition of TV distance $\|\mathbb{P} - \mathbb{Q}\|_{\mathrm{TV}} := \sup_{A \subseteq X} |\mathbb{P}(A) - \mathbb{Q}(A)|$
- There is a one-to-one correspondence between ψ and partitions (A, A^c) of the space $\mathcal{X}$
  $A = \{x \in X \mid \psi(x) = 1\}$
- The Bayes risk can be expressed in terms of $\|\mathbb{P}_1 - \mathbb{P}_0\|_{\mathrm{TV}}$

$$
\begin{aligned}
\inf_{\psi} \mathbb{Q}[\psi(Z) \neq J] &= \frac{1}{2} \inf_{\psi} (\mathbb{P}_0[\psi(Z) \neq 0] + \mathbb{P}_1[\psi(Z) \neq 1]) \\
&= \frac{1}{2} \inf_{A \subseteq \mathcal{X}} (\mathbb{P}_0[A] + \mathbb{P}_1[A^c]) \\
&= \frac{1}{2} \{1 - \|\mathbb{P}_1 - \mathbb{P}_0\|_{\mathrm{TV}}\}
\end{aligned}
$$

Le Cam's method

Proposition 2.

(Le Cam's bound) For any pair of distributions $\mathbb{P}_0, \mathbb{P}_1 \in \mathcal{P}$ *s.t.* $\rho\left(\theta\left(\mathbb{P}_0\right), \theta\left(\mathbb{P}_1\right)\right) \geq 2\delta$*,*

$$\mathfrak{M}(\theta(\mathcal{P}), \Phi \circ \rho) \geq \frac{\Phi(\delta)}{2} \left\{1 - \|\mathbb{P}_1 - \mathbb{P}_0\|_{\mathrm{TV}}\right\} \tag{17}$$

- Two extremes
  - Worst case: $\mathbb{P}_1 = \mathbb{P}_0$, hypotheses are completely indistinguishable
  - Best case: $\|\mathbb{P}_1 - \mathbb{P}_0\|_{\mathrm{TV}} = 1$, $\mathbb{P}_0$ and $\mathbb{P}_1$ have disjoint supports

Example 1: Gaussian location family

- $\{\mathcal{N}(\theta, \sigma^2), \theta \in \mathbb{R}\}$: a normal distribution family with fixed variance σ^2
- Goal: Estimate θ
  - Metric: either $|\widehat{\theta} - \theta|$ or $(\widehat{\theta} - \theta)^2$
  - Data: a collection $Z = (Y_1, \ldots, Y_n) \sim \mathcal{N}(\theta, \sigma^2)^{1:n} = P_\theta^n$
- Apply the two-point Le Cam bound with the distributions $\mathbb{P}_0^n$ and $\mathbb{P}_\theta^n$
  - Set $\theta = 2\delta$ s.t. the two means are 2δ-separated
  - Bound $\|\mathbb{P}_\theta^n - \mathbb{P}_0^n\|_{\mathrm{TV}}$

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

Example 1: Gaussian location family

- Apply the two-point Le Cam bound with the distributions $\mathbb{P}_0^n$ and $\mathbb{P}_\theta^n$
  - Setting $\delta = \frac{1}{2}\frac{\sigma}{\sqrt{n}}$ thus yields

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

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

- Although the pre-factors $1/12$ and $1/24$ are not optimal, the scalings $\sigma/\sqrt{n}$ and σ^2/n are sharp/order optimal
- Matching upper bound: the sample mean $\widetilde{\theta}_n := \frac{1}{n}\sum_{i=1}^n Y_i$ satisfies the bounds

$$\sup_{\theta \in \mathbb{R}} \mathbb{E}_\theta\left[|\widetilde{\theta}_n - \theta|\right] = \sqrt{\frac{2}{\pi}}\frac{\sigma}{\sqrt{n}} \quad \text{and} \quad \sup_{\theta \in \mathbb{R}} \mathbb{E}_\theta\left[\left(\widetilde{\theta}_n - \theta\right)^2\right] = \frac{\sigma^2}{n}$$

Example 2: Uniform location family

- Mean-squared error decaying as n^{-1} is typical for parametric problems, but faster rates is possible for some other problems
- $\{\mathbb{U}_\theta, \theta \in \mathbb{R}\}$: $\mathbb{U}_\theta$ is uniform over the interval $[\theta, \theta + 1]$
- Impossible to use Pinsker's inequality to control the TV norm!
- Consider $H^2\left(\mathbb{U}_\theta \| \mathbb{U}_{\theta'}\right)$, recall that

$$H^2(\mathbb{P}\|\mathbb{Q}) = 2 - 2\int_{\mathcal{X}} \sqrt{p(x)q(x)}dx$$

  - It suffices to consider the case $\theta' > \theta$
  - If $\theta' > \theta + 1$, then $H^2\left(\mathbb{U}_\theta \| \mathbb{U}_{\theta'}\right) = 2$
  - Otherwise, $H^2\left(\mathbb{U}_\theta \| \mathbb{U}_{\theta'}\right) = 2 - 2\int_{\theta'}^{\theta+1} dt = 2\left|\theta' - \theta\right|$

Example 2: Uniform location family

- Apply the Le Cam bound with the distributions $\mathbb{U}_\theta^n$ and $\mathbb{U}_{\theta'}^n$
  - Take a pair θ, θ' s.t. $|\theta' - \theta| = 2\delta := \frac{1}{4n}$
  - $\|\mathbb{U}_\theta^n - \mathbb{U}_{\theta'}^n\|_{\mathrm{TV}}^2 \leq H^2\left(\mathbb{U}_\theta^n \| \mathbb{U}_{\theta'}^n\right) \leq nH^2\left(\mathbb{U}_\theta \| \mathbb{U}_{\theta'}\right) = n2\left|\theta' - \theta\right| = \frac{1}{2}$
  - $\inf_{\widehat{\theta}} \sup_{\theta \in \mathbb{R}} \mathbb{E}_\theta\left[(\widehat{\theta} - \theta)^2\right] \geq \frac{\delta^2}{2}\left\{1 - \sqrt{\frac{1}{2}}\right\} = \frac{\left(1 - \frac{1}{\sqrt{2}}\right)}{128} \frac{1}{n^2}$
- Contrasted with the n^{-1} rate, this lower bound has faster n^{-2} rate!
- Matching upper bound: the estimator $\widetilde{\theta} = \min\{Y_1, \ldots, Y_n\}$ satisfies the bound $\sup_{\theta \in \mathbb{R}} \mathbb{E}\left[(\tilde{\theta} - \theta)^2\right] \leq \frac{2}{n^2}$
- Estimating the location parameter of uniform location family is easier.

Outline

Fano's method

- Le Cam's method reduces the estimation problem to binary test, how about M-ary hypothesis testing problem?
- In information theory, Fano's inequality lower bounds the error probability in such problems

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

- Combined with the reduction in Proposition 1

Proposition 3.

(Fano's bound) Let $\{\theta^1, \ldots, \theta^M\}$ *be a* 2δ*-separated set in the* ρ *semi-metric on* $\Theta(\mathcal{P})$*,*

$$\mathfrak{M}(\theta(\mathcal{P}); \Phi \circ \rho) \geq \Phi(\delta) \left\{ 1 - \frac{I(Z;J) + \log 2}{\log M} \right\} \tag{18}$$

Remarks on Fano's method

- Consider the behavior of the different terms of r.h.s. as $\delta \to 0^+$
  - The 2δ-separation criterion becomes milder, $M \equiv M(2\delta)$ increases
  - $J \in [M(2\delta)]$ can take on a larger number of potential values, $I(Z; J)$ decreases
  - Decreasing δ sufficiently may ensure that

$$\frac{I(Z; J) + \log 2}{\log M} \leq \frac{1}{2} \tag{19}$$

  - $\mathfrak{M}(\theta(\mathcal{P}); \Phi \circ \rho) \geq \frac{1}{2}\Phi(\delta)$
- Two technical and possibly challenging steps
  - Specify 2δ-separated sets with large cardinality $M(2\delta)$, metric entropy theory
  - Compute or upper bound $I(Z; J)$, non-trivial
- Using convexity of KL divergence and the mixture representation

$$I(Z; J) = \frac{1}{M} \sum_{j=1}^{M} D\left(\mathbb{P}_{\theta^j} \| \overline{\mathbb{Q}}\right) \leq \frac{1}{M^2} \sum_{j,k=1}^{M} D\left(\mathbb{P}_{\theta^j} \| \mathbb{P}_{\theta^k}\right) \tag{20}$$

Example 3: Gaussian location model via Fano's method

- Consider the 2δ-separated set $\left\{\theta^1, \theta^2, \theta^3\right\} = \{0, 2\delta, -2\delta\}$
- $D\left(\mathbb{P}_{\theta^j}^{1:n} \| \mathbb{P}_{\theta^k}^{1:n}\right) = \frac{n}{2\sigma^2}\left(\theta^j - \theta^k\right)^2 \le \frac{8n\delta^2}{\sigma^2}$ for all $j, k = 1, 2, 3$
- $I\left(Z; J\right) \le \frac{8n\delta^2}{\sigma^2}$
- Choosing $\delta^2 = \frac{\sigma^2}{80n}$ ensures that $\frac{8n\delta^2/\sigma^2 + \log 2}{\log 3} = \frac{0.1 + \log 2}{\log 3} < 0.75$
- The Fano's bound with $\Phi(t) = t^2$ implies

$$\sup_{\theta \in \mathbb{R}} \mathbb{E}_\theta\left[(\widehat{\theta} - \theta)^2\right] \ge \frac{\delta^2}{4} = \frac{1}{320}\frac{\sigma^2}{n}$$

Bounds based on local packings

- Construct a 2δ-separated set contained within Ω s.t. for some $c > 0$

$$\sqrt{D\left(\mathbb{P}_{\theta^j} \| \mathbb{P}_{\theta^k}\right)} \leq c\sqrt{n}\delta \quad \text{for all } j \neq k \tag{21}$$

- The bound (20) then implies that $I(Z; J) \leq c^2 n\delta^2$, and hence the bound (19) will hold if

$$\log M(2\delta) \geq 2\left\{c^2 n\delta^2 + \log 2\right\} \tag{22}$$

- The minimax risk is lower bounded as $\mathfrak{M}(\theta(\mathcal{P}), \Phi \circ \rho) \geq \frac{1}{2}\Phi(\delta)$

Example 4: Minimax risks for linear regression

- Standard linear regression model: $y = \mathbf{X}\theta^* + w$
- $\mathbf{X} \in \mathbb{R}^{n \times d}$: fixed design matrix, $w \sim \mathcal{N}\left(0, \sigma^2 \mathbf{I}_n\right)$: observation noise
- Metric: prediction norm $\rho_{\mathbf{X}}\left(\widehat{\theta}, \theta^*\right) := \frac{\left\|\mathbf{X}(\widehat{\theta} - \theta^*)\right\|_2}{\sqrt{n}}$, θ^*: vary over $\mathbb{R}^d$
- Consider the set $\{\gamma \in \operatorname{range}(\mathbf{X}) \mid \|\gamma\|_2 \leq 4\sqrt{n}\delta\}$
- let $\left\{\gamma^1, \ldots, \gamma^M\right\}$ be a $2\sqrt{n}\delta$-packing in the ℓ_2-norm, $r = \operatorname{rank}(\mathbf{X})$
- Lemma 5.7 in HDS book implies such a packing with $\log M \geq r \log 2$ elements

Example 4: Minimax risks for linear regression

- A collection of vectors of the form $\gamma^j = \mathbf{X}\theta^j$ for some $\theta^j \in \mathbb{R}^d$ s.t.

$$\frac{\left\|\mathbf{X}\theta^j\right\|_2}{\sqrt{n}} \leq 4\delta, \quad \text{for each } j \in [M] \tag{23}$$

$$2\delta \leq \frac{\left\|\mathbf{X}\left(\theta^j - \theta^k\right)\right\|_2}{\sqrt{n}} \leq 8\delta, \text{ for each } j \neq k \in [M] \times [M] \tag{24}$$

- Let $\mathbb{P}_{\theta^j}$ denote the distribution of y when $\theta^* = \theta^j$, then $\mathbb{P}_{\theta^j} = \mathcal{N}\left(\mathbf{X}\theta^j, \sigma^2\mathbf{I}_n\right)$
- $D\left(\mathbb{P}_{\theta^j} \| \mathbb{P}_{\theta^k}\right) = \frac{1}{2\sigma^2}\left\|\mathbf{X}\left(\theta^j - \theta^k\right)\right\|_2^2 \leq \frac{32n\delta^2}{\sigma^2}$
- Condition (21) holds with $c = \frac{\sqrt{32}}{\sigma}$

Example 4: Minimax risks for linear regression

- Need to lower bound $\log M \geq r \log 2 \geq 2(c^2 n\delta^2 + \log 2)$
- Choose $\delta^2 = \frac{\sigma^2}{64}\frac{r}{n}$, then condition (22) holds for sufficiently large r since
  - $D\left(\mathbb{P}_{\theta^j} \| \mathbb{P}_{\theta^k}\right) \leq \frac{r}{2}$
  - $2(c^2 n\delta^2 + \log 2) = 2(\frac{r}{2} + \log 2) = r + \log 2$
- Set $\Phi(t) = t^2$

$$\inf_{\widehat{\theta}} \sup_{\theta \in \mathbb{R}^d} \mathbb{E}\left[\frac{1}{n}\|\mathbf{X}(\widehat{\theta} - \theta)\|_2^2\right] \geq \frac{1}{128}\frac{\mathrm{rank}(\mathbf{X})\sigma^2}{n}$$

- This lower bound is sharp up to constant pre-factors