Lecture 11: Minimax lower bounds

Announcements

- Homework 2 was due last night, solutions out today
- Please fill out your oral exam availabilities sent out in email, taking place 20.11./21.11. 9 am - 5 pm
  - mark *all slots* where you do not have a strict conflict
  - exams are 20 minutes long

Recap: Upper bound for random design

We considered the non-parametric regression setting $Y = f^\star(X) + w$

We view X as random and take expectation also over X, i.e. for any $f \in \mathcal{L}^2(\mathbb{P})$, we have

$$\|f - f^\star\|_2^2 = R(f) - R(f^\star) = \mathbb{E}_{X,W}(Y - f(X))^2 - \mathbb{E}W^2$$
$$= \mathbb{E}_X(f(X) - f^\star(X))^2 = \mathbb{E}_{x_1,\ldots,x_n}\|f - f^\star\|_n^2$$

and want to bound $\|\widehat{f} - f^\star\|_2^2$ for an estimator $\widehat{f}$

Theorem (Localized uniform law, MW Thm 14.1)

For star-shaped and b−uniformly bounded $\mathcal{F}^\star$, let $\bar{\delta}_n$ be population critical radius. Then if $\bar{\delta}_n^2 > c\frac{\log[4\log(1/\bar{\delta}_n)]}{n}$ then w.p. at least $1 - c_1 e^{-c_2\frac{n\bar{\delta}_n^2}{b^2}}$ we have $\sup_{g\in\mathcal{F}^\star}\|g\|_2 - \|g\|_n \leq c\bar{\delta}_n$

For bounded domains, we can then plug in $g = \widehat{f} - f^\star$, use the h.p. upper bound for the empirical error $\|\widehat{f} - f^\star\|_n^2 \leq U(n)$ and obtain w.h.p

$$\|\widehat{f} - f^\star\|_2^2 \leq U(n) + c\bar{\delta}_n$$

Estimation task

- Let $\mathcal{P}$ be a set of probability distributions on $(\mathcal{X}, \mathcal{Y})$, can then view a quantity of interest to be a mapping F acting on a probability distribution (outputting a function or parameter)
- For today, we consider each $\mathbb{P}_{\mathcal{F}} \in \mathcal{P}$ defined via $y = f^\star(x) + w$ (either y or both x, y random), for different $f^\star \in \mathcal{F}$ but fixed distributions over x and noise w and the object of interest could be $F(\mathbb{P})(x) = \mathbb{E}[Y|x] = f^\star(x)$.
- View estimating procedure/algorithm for $F(\mathbb{P})$ as a mapping $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{F}$ from dataset to space of functions, where $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ with $(x_i, y_i) \sim \mathbb{P}$, outputting $\widehat{f}_{\mathcal{D}} = \mathcal{A}(\mathcal{D})$
- So far we've seen: Error bounds of the type $\|\widehat{f}_{\mathcal{D}} - f^\star\|_2^2 \leq O(n^{-\alpha})$

Pair-Q: Discuss with your neighbor: What is a reasonable notion of optimality of an algorithm that a practitioner might care about?
Today: Compare to what's the best possible (*optimal*) given the data?

Minimax risk

Definition (Minimax risk)

The minimax risk or error of estimating the mapping $F : \mathcal{P}_\mathcal{F} \to \mathcal{F}$ in some squared metric $\|\cdot\|^2$ is defined as

$$\mathfrak{M}(F(\mathcal{P}), \|\cdot\|^2) = \inf_{\mathcal{A}} \sup_{\mathbb{P}\in\mathcal{P}_\mathcal{F}} \mathbb{E}_{\mathcal{D}\sim\mathbb{P}^n} \|\mathcal{A}(\mathcal{D}) - F(\mathbb{P})\|^2$$

- $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ has i.i.d. samples from $\mathbb{P}^n \to \mathcal{A}(\mathcal{D})$ is random
- Note that more generally $\mathcal{F}$ can also be a parameter space for parameterized function classes (as we will see next lecture)
- Here $\mathcal{A}$ is not constrained to any particular procedure (could be minimization of risk but also something else) but "knows" to search in set $\mathcal{F}$ that induces $\mathcal{P}_\mathcal{F}$
- Here we consider deterministic (i.e. not random) algorithms $\mathcal{A}$
- could use as $\|\cdot\|$ standard metric of $\mathcal{F}$ (see MW Chapter 15)

Minimax lower bounds

What do we learn if we could obtain $\mathfrak{M}(F(\mathcal{P}), \|\cdot\|^2) \geq O(n^{-\alpha})$?

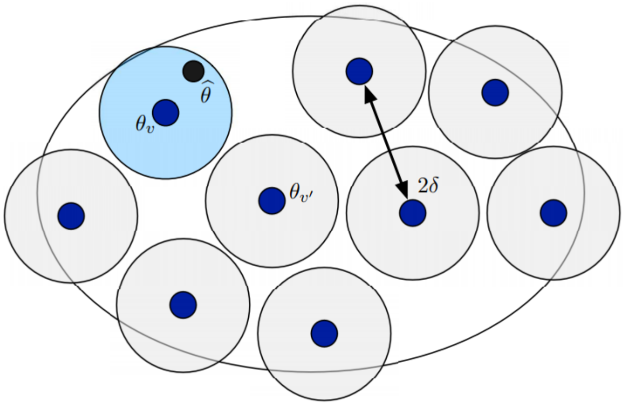
- no estimator (knowing $\mathcal{P}_\mathcal{F}$ or, equivalently, $\mathcal{F}$ and) can achieve smaller risk (for their resp. hardest case)
- if upper bound of an estimation procedure matches lower bound:
  - practically we don't need to waste time looking for "better"
  - if we want to do better in the worst case

This class: Find **lower bounds** for the minimax risk as large as possible for **given** $\mathcal{P}$, F

- From estimation to "testing" / classification
- Fano's method: bounding the probability of testing error via mutual information (MI)
- Upper bounding MI using Yang-Barron
- Examples: non-parametric regression on Sobolev functions

Main idea: From estimation to testing (intuition)

- Consider M finite functions f^i spread across $\mathcal{F}$ s.t. pairwise distances $> 2\delta$ (e.g. in a packing set of $\mathcal{F}$)
- If $\mathcal{A}$ can find $\widehat{f}$ (black dot) that is δ close to any true $f^\star \in \mathcal{F}$
  $\rightarrow$ if data is drawn from f^j, $\mathcal{A}$ induces a test that correctly identifies f^j by choosing the closest f^i (blue dot) to the estimated $\widehat{f}$
  $\rightarrow$ no "testing" error



- As we want a lower bound on estimation, can reverse the argument

$\rightarrow$ Problem reduces to: given n points, what's the smallest possible δ so that we can distinguish from which f^i the data was drawn?

Main idea: from estimation to testing

We sometimes write $\widehat{f}_{\mathcal{D}} = \mathcal{A}(\mathcal{D})$, omitting $\mathcal{A}$ subscript. Define

- For any M let $\{f^i\}_{i=1}^M$ be a set of functions in $\mathcal{F}$
- For each $\tilde{f} \in \mathcal{F}$, define $\mathbb{P}_{\tilde{f}}$ as a unique distribution with $F(\mathbb{P}_{\tilde{f}}) = \tilde{f}$
- Define the mixture distribution $\mathbb{Q}_M$ for $\mathcal{D}, J$ by defining
  1. J a uniform R.V. (flat "prior") with values in $[M] = \{1, \ldots, M\}$, i.e. $\mathbb{Q}_M(J = j) = \frac{1}{M}$ for all j
  2. and drawing random i.i.d. datapoints $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$ from $\mathbb{P}_{f^j}^n$, i.e. $\mathbb{Q}_M(\mathcal{D} | J = j) = \mathbb{P}_{f^j}^n$
- Decision / Testing functions of form $\psi : (\mathcal{X} \times \mathcal{Y})^n \rightarrow [M]$

Lemma (Estimation vs. testing, MW Prop 15.1)

Choose $\{f^i\}_{i=1}^{M(2\delta)}$ to be a 2δ-packing of $\mathcal{F}$ in the $\|\cdot\|$ metric so that $M(2\delta) \leq \mathcal{M}(2\delta; \mathcal{F}, \|\cdot\|)$, then

$$\inf_{\mathcal{A}} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathcal{D} \sim \mathbb{P}} \|\mathcal{A}(\mathcal{D}) - F(\mathbb{P})\|^2 \geq \delta^2 \inf_{\psi} \mathbb{Q}_{M(2\delta)}(\psi(\mathcal{D}) \neq J)$$

Proof of Lemma

Omitting $\mathbb{Q}_M$ subscript, define $\psi_{\mathcal{A}}(\mathcal{D}) := \arg\min_{i\in[M]} \|\mathcal{A}(\mathcal{D}) - f^i\|$

1. Markov's inequality yields
$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\|\mathcal{A}(\mathcal{D}) - F(\mathbb{P})\|^2 &\geq \delta^2 \mathbb{P}(\|\mathcal{A}(\mathcal{D}) - F(\mathbb{P})\|^2 \geq \delta^2) \\ &= \delta^2 \mathbb{P}(\|\mathcal{A}(\mathcal{D}) - F(\mathbb{P})\| > \delta)\end{aligned}$$
2. Key link between estimation and "testing" (via intuition sl. 8):
$$\mathbb{Q}(\{\|\mathcal{A}(\mathcal{D}) - f^i)\| \leq \delta\} | J = i) \leq \mathbb{Q}(\{\psi_{\mathcal{A}}(\mathcal{D}) = i\} | J = i)$$
because for any $f \in \mathcal{F}$ such that $\|f - f^i\| < \delta$, for any $j \neq i$ we have $\|f - f^j\| > \|f^j - f^i\| - \|f - f^i\| > \delta \to \psi_{\mathcal{A}}(\mathcal{D}) = i$
3. Then the Lemma follows by the distribution of J
$$\delta^{-2} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathcal{D}\sim\mathbb{P}}\|\mathcal{A}(\mathcal{D}) - F(\mathbb{P})\|^2 \overset{1.}{\geq} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}^n(\|\mathcal{A}(\mathcal{D}) - F(\mathbb{P})\| > \delta)$$
$$\geq \frac{1}{M} \sum_{i\in[M]} \mathbb{P}^n_{f^i}(\|\mathcal{A}(\mathcal{D}) - f^i\| > \delta) = \sum_{i\in[M]} \mathbb{Q}(J = i)\mathbb{Q}(\|\mathcal{A}(\mathcal{D}) - f^i\| > \delta | J =$$
$$\overset{2.}{\geq} \sum_{i\in[M]} \mathbb{Q}(J = i)\mathbb{Q}(\{\psi_{\mathcal{A}}(\mathcal{D}) \neq i\} | J = i) = \mathbb{Q}(\{\psi_{\mathcal{A}}(\mathcal{D}) \neq J\})$$

Lower bounding $\mathbb{Q}(\psi(\mathcal{D}) \neq J)$ with Fano's method

For simplicity assuming densities of joint and conditional distributions:

Definitions (Entropy and mutual information)

For any two R.V. X, Y with joint probability distribution $\mathbb{P}$ define

- the *entropy* $H(X, Y) = -\mathbb{E}_{\mathbb{P}} \log p(X, Y)$
- the *conditional entropy* $H(X|Y) = -\mathbb{E}_{\mathbb{P}} \log p(X|Y)$
- the *mutual information* $I(X, Y) = H(X) - H(X|Y)$

Intuitively (imprecise):

- $H(X|Y)$: uncertainty "left" about X if value of Y were known
- $I(X, Y)$: information of X in Y and vice versa

Theorem (Fano's method, MW Sec 15.4.)

For some $M \in \mathbb{N}$ and $\{f^i\}_{i=1}^M$, let $\mathbb{Q}_M$ be a mixture distribution as in slide 9. Then for any decision/testing function ψ, it holds that
$$\mathbb{Q}_M(\psi(\mathcal{D}) \neq J) \geq 1 - \frac{I(\mathcal{D}, J) + \log 2}{\log M}$$

Proof of Theorem (Fano's method)

Define Bernoulli $E_\psi = \mathbb{1}_{\psi(\mathcal{D})\neq J}$ with $\mathbb{Q}_M(E_\psi = 1) = \mathbb{Q}_M(\psi(\mathcal{D}) \neq J)$

1. We first establish *Fano's inequality* after which the proof is trivial

$$H(J|\mathcal{D}) \leq H(E_\psi) + \mathbb{Q}_M(\psi(\mathcal{D}) \neq J) \log(M-1)$$

- Proof: First, by Bayes' theorem and def. of conditional expectations

$$\underbrace{H(E_\psi|J,\mathcal{D})}_{=0} + H(J|\mathcal{D}) = H(J, E_\psi|\mathcal{D}) = H(J|E_\psi,\mathcal{D}) + \underbrace{H(E_\psi|\mathcal{D})}_{\leq H(E_\psi)}$$

- Proof then follows from
$H(J|E_\psi,\mathcal{D}) = \underbrace{H(J|E_\psi = 0,\mathcal{D})}_{=0}\mathbb{Q}(E_\psi = 0) + \underbrace{H(J|E_\psi = 1,\mathcal{D})}_{\leq \log(M-1)}\mathbb{Q}(E_\psi = 1)$

2. Since E_ψ Bernoulli $H(E_\psi) \leq \log 2$ for all ψ
and since J uniform $H(J) = \log M$
3. Using Fano's inequality and $H(J|\mathcal{D}) = H(J) - I(\mathcal{D}, J)$ yields Thm.

Fano's method to lower bound minimax risk

- We would like to ultimately plug in Fano's lower bound into the lemma.
- If we choose $\{f^i\}_{i=1}^{M(2\delta)}$ to be a 2δ-packing as in Lemma we can plug in $M = M(2\delta) \leq \mathcal{M}(2\delta; \mathcal{F}, \|\cdot\|)$ to get

$$\mathbb{Q}_{M(2\delta)}(\psi(\mathcal{D}) \neq J) \geq 1 - \frac{I(\mathcal{D}, J) + \log 2}{\log M(2\delta)}$$

- If δ is chosen such that $I(\mathcal{D}, J) \sim \log M(2\delta)$ then the Lemma implies a lower bound of order δ^2
- This might or might not be a tight lower bound (if it matches some algorithm dependent upper bound, you're in luck)

Upper bounding the mutual information

- To bound the mutual information we recall the

Definition (Kullback-Leibler divergence)

The KL divergence between any two probability distributions $\mathbb{P}, \mathbb{Q}$

$$KL(\mathbb{P} \parallel \mathbb{Q}) = \mathbb{E}_{\mathbb{P}} \log \frac{d\mathbb{P}}{d\mathbb{Q}}$$

- We can write $I(\mathcal{D}, J) = KL(\mathbb{Q} \parallel \mathbb{Q}_{\mathcal{D}}\mathbb{Q}_J)$ and then for q densities of $\mathbb{Q}$, we have

$$\begin{aligned}\mathbb{E}_J \mathbb{E}_{\mathcal{D}} \log \frac{q_{\mathcal{D}|J}}{q_{\mathcal{D}}} &= \mathbb{E}_J KL(\mathbb{Q}_{\mathcal{D}|J} \parallel \mathbb{Q}_{\mathcal{D}}) \\ &= \frac{1}{M}\sum_{i=1}^{M} KL(\mathbb{P}^n_{f^i} \parallel \frac{1}{M}\sum_{j=1}^{M} \mathbb{P}^n_{f^j})\end{aligned}$$

- The next theorem bounds the mutual information in Fano's method.

Theorem (Yang-Barron, MW Lemma 15.21)

$$I(\mathcal{D}, J) \leq \inf_{\epsilon > 0} \epsilon^2 + \log \mathcal{N}(\epsilon^2; \mathcal{P}^n, KL)$$

Summary: One recipe for minimax lower bounds

Recipe for using Yang-Barron + Fano to get lower bounds:

1. Choose ϵ such that $\epsilon^2 \geq \log \mathcal{N}(\epsilon^2; \mathcal{P}^n, KL)$
2. Choose δ such that $\log \mathcal{M}(2\delta; \mathcal{F}, \|\cdot\|) \geq 4\epsilon^2 + 2\log 2$
3. Hence $1 - \frac{I(\mathcal{D},J)+\log 2}{\log M(2\delta)} \geq \frac{1}{2}$ and via Fano's method

$$\inf_{\mathcal{A}} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}_{\mathbb{P}} \|\mathcal{A}(\mathcal{D}) - F(\mathbb{P})\|^2 \geq \frac{1}{2}\delta^2$$

Minimax prediction error for estimating Sobolev functions

Example: Sobolev functions $\mathcal{F} = \mathcal{W}_2^\alpha([0,1])$ with

- Consider the family of distributions $\mathcal{P}_\mathcal{F}$ generated via: $X \sim U([0,1])$ and $y = f^\star(x) + w$ with standard normal w and $f^\star \in \mathcal{W}_2^\alpha([0,1])$ so that conditional distribution $Y|x \sim \mathcal{N}(f(x), \sigma^2)$ (our non-parametric regression setting)
- We're interested in estimating $f^\star = \mathbb{E}_\mathbb{P}[Y|x]$ and evaluate it via the $\mathcal{L}^2([0,1])$ norm
- Recall *upper bounds* for constrained kernel regression
  - w.h.p. $\|\widehat{f} - f^\star\|_n^2 \leq O\left(\frac{\sigma^2}{n}\right)^{\frac{2\alpha}{2\alpha+1}}$ (HW 2)
  - $\widehat{f} - f^\star$ is uniformly bounded by reproducing property and Hilbert norm constraint → MW Thm 14.1. and MW Prop 14.25 yields $\|\widehat{f} - f^\star\|_{\mathcal{L}^2([0,1])}^2 \leq O\left(\frac{\sigma^2}{n}\right)^{\frac{2\alpha}{2\alpha+1}}$

Minimax prediction error for estimating Sobolev functions

Corollary (Minimax error for Sobolev function estimation)

Writing $\|\cdot\|_2 := \|\cdot\|_{\mathcal{L}^2([0,1])}^2$, we have for $\frac{n}{\sigma^2}$ larger than a constant

$$\mathfrak{M}(F(\mathcal{P}), \|\cdot\|_2^2) \geq O\left(\frac{\sigma^2}{n}\right)^{\frac{2\alpha}{2\alpha+1}}$$

Proof of Corollary

a) Writing out the conditional distribution we have for $n = 1$

$$KL(\mathbb{P}_f \,\|\, \mathbb{P}_g) = \frac{1}{2\sigma^2}\mathbb{E}_{\mathbb{P}_f} g^2(X) - f^2(X) + 2(f(X) - g(X))Y$$

$$= \frac{1}{2\sigma^2}\mathbb{E}_{\mathbb{P}_f} g^2(X) - f^2(X) + 2(f(X) - g(X))f(X) = \frac{\|f-g\|_2^2}{2\sigma^2}$$

b) For n samples we have an extra factor of n, since for $z_i = (x_i, y_i)$

$$KL(\mathbb{P}_f^n \,\|\, \mathbb{P}_g^n) = \int \prod_{i=1}^n p_f(z_i) \log \prod_{i=1}^n \frac{p_f(z_i)}{p_g(z_i)} \mu(dz^n)$$

$$= \sum_{i=1}^n \int p_f(z_i) \log \frac{p_f(z_i)}{p_g(z_i)} \mu(\mathrm{d}z_i) = n\frac{\|f-g\|_2^2}{2\sigma^2}$$

Proof ctd'

c) Hence $\mathcal{N}(\epsilon^2; \mathcal{P}^n, KL) = \mathcal{N}(\frac{\epsilon\sqrt{2\sigma^2}}{\sqrt{n}}; \mathcal{W}_2^\alpha([0,1]), \|\cdot\|_2)$

d) Using the result in next slide about covering number of Sobolev spaces

- Using $\log \mathcal{N}(\delta; \mathcal{W}_2^\alpha([0,1]), \|\cdot\|_2^2) = O(\frac{1}{\delta})^{1/\alpha}$ and 1. in slide 15 we require
$$\epsilon^2 \geq \Big(\frac{n}{2\sigma^2}\Big)^{\frac{1}{2\alpha}} \epsilon^{-1/\alpha} \quad \rightarrow \quad \epsilon^2 = O\Big(\frac{n}{\sigma^2}\Big)^{\frac{1}{2\alpha+1}}$$
- Recalling that $\mathcal{M}(2\delta) \geq \mathcal{N}(2\delta)$ and using 2. in slide 15, it suffices to require
$$\Big(\frac{1}{\delta}\Big)^{\frac{1}{\alpha}} \geq c\Big[\Big(\frac{n}{\sigma^2}\Big)^{\frac{1}{2\alpha+1}} + 2\log 2\Big] \quad \rightarrow \quad \delta^2 = O\Big(\frac{\sigma^2}{n}\Big)^{\frac{2\alpha}{2\alpha+1}}$$
for $\frac{\sigma^2}{n}$ smaller than a universal constant.

e) Hence by 3. (Fano's method) $\|\widehat{f} - f^\star\|^2_{\mathcal{L}^2([0,1])} \geq O\Big(\frac{\sigma^2}{n}\Big)^{\frac{2\alpha}{2\alpha+1}}$ □

Metric entropy for higher order Sobolev spaces (bonus)

Lemma (Metric entropy for α-order compact Sobolev spaces)

It holds that $\log \mathcal{N}(\delta; \mathcal{W}_2^\alpha([0,1]), \|\cdot\|_2^2) = O(\frac{1}{\delta})^{\frac{1}{\alpha}}$.

Proof steps

Define $\mathcal{E}_\alpha = \{\theta \in \ell_2(\mathbb{N}) : \sum_{j=1}^\infty j^{2\alpha}\theta_j^2 \leq 1\}$

1. First observation: $\mathcal{N}(\delta; \mathcal{W}_2^\alpha([0,1]), \|\cdot\|_2^2) = \mathcal{N}(\delta; \mathcal{E}_\alpha, \|\cdot\|_{\ell^2(\mathbb{N})})$
   - Note that by Mercer's Theorem, we can write for some orthonormal basis in $\|\cdot\|_2$ $\mathcal{W}_2^\alpha([0,1]) = \{f : f = \sum_{j=1}^\infty \theta_j\phi_j \text{ for } \theta \in \mathcal{E}_\alpha\}$
   - Kernel operator eigenvalues decay as $j^{2\alpha}$ (hinges on spectra of differential operators that we won't prove)
   - Because ϕ_j are orthonormal in $\|\cdot\|_2$ norm we have $\|f\|_2^2 = \|\theta_f\|^2_{\ell^2(\mathbb{N})}$
2. MW Example 5.12. proves $\log \mathcal{N}(\delta; \mathcal{E}_\alpha, \|\cdot\|_{\ell^2(\mathbb{N})}) \leq O\Big(\frac{1}{\delta}\Big)^{\frac{1}{\alpha}}$ □

References

Main source

- MW Chapter 15

Additional reading

- *John Duchi Information Theory (Stats 311) Lecture Notes*: Lectures 3, 5, 6
- *Bin Yu '97*: Assouad, Fano and LeCam, "Festschrift for Lucien LeCam" - overview of different minimax methods (including two we did not talk about)
- *Yang, Barron '99*: Information theoretic determination of minimax rates of convergence.