

## Lecture 13: Minimax lower bounds

1 / 22

### Announcements

- Reminder to indicate all your availabilities for the oral exam slot if you haven't yet.
- Oral exam: You're expected to be able to explain all results and definitions/concepts that were covered in the slides and homework (minus the "bonus" slides, including the proofs "skipped in class")
- Next Tuesday: Last class before oral exam, presentation of multi-objective learning setting and interactive session on minimax lower bounds for that setting. Good practice for oral exam to explain to each other (basically what I'll ask you to do in oral exam)

2 / 22

## Recap: Optimality for (non-parametric) regression

- Consider the concrete setting where  $\mathbb{P}_{f^*} \in \mathcal{P}$  defined via  $y = f^*(x) + w$  (either  $y$  or both  $x, y$  random), for different  $f^* \in \mathcal{F}$  but fixed distributions over  $x$  and noise  $w$  and the object of interest could be the functional  $F : \mathcal{P} \rightarrow \mathcal{F}$  (mapping from distributions to space of functions), defined by

$$F(\mathbb{P})(x) = \mathbb{E}[Y|x] = f^*(x)$$

- The set of all such functions is denoted as  $\mathcal{P}_{\mathcal{F}} = \{\mathbb{P}_{f^*} : f^* \in \mathcal{F}\}$
- For parametric function spaces, the algorithms/procedures map to the parameter space  $\Theta$  instead of  $\mathcal{F}$  and that's how we'll use the theorems today
- View estimating procedure/algorithm for  $F(\mathbb{P})$  as a mapping  $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^n \rightarrow \mathcal{F}$  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  $\hat{f}_{\mathcal{D}} = \mathcal{A}(\mathcal{D})$
- the (worst-case) optimality of an algorithm can be captured by  $\inf_{\mathcal{A}} \sup_{f^* \in \mathcal{F}} \mathbb{E}_{\mathcal{D} \sim \mathbb{P}^n} \|\hat{f}_{\mathcal{D}} - f^*\|_2^2$  (the  $\mathcal{L}^2(\mathbb{P})$  loss)

3 / 22

## Recap: Minimax risk

- **In general**,  $F(\mathbb{P})$  can be arbitrary functional that takes any  $\mathbb{P}$  and maps to  $\mathcal{F}$ , and  $\mathcal{P}$  would be the family of all distributions  $\mathbb{P}$  that you can “imagine that your data is from” (encoding prior knowledge)

### Definition (Minimax risk)

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

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

- $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$  has i.i.d. samples from  $\mathbb{P}^n \rightarrow \mathcal{A}(\mathcal{D})$  is random
- Here  $\mathcal{A}$  is **not constrained to any particular procedure** but “knows”  $\mathcal{P}$  and naturally its range  $\mathcal{F}$
- In what follows you can generally think of our example setting above,  $\|\cdot\|$  as the  $\mathcal{L}^2(\mathbb{P})$  norm on functions in  $\mathcal{F} \subset \mathcal{L}^2(\mathbb{P})$ .

4 / 22

## Minimax lower bounds

- If we could obtain  $\mathfrak{M}(F(\mathcal{P}), \|\cdot\|^2) \geq O(n^{-\alpha})$ , this means no estimator (knowing  $\mathcal{P}_{\mathcal{F}}$  or, equivalently,  $\mathcal{F}$  and  $\mathbb{P}$ ) 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

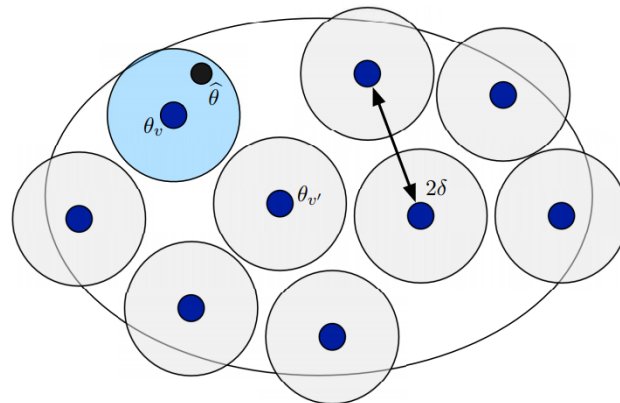
Today: Find **lower bounds** for the minimax risk as large as possible for **given**  $\mathcal{P}, F$

- From estimation to M-ary testing
- One method to obtain a lower bound (one more in the homework)
  - 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

5 / 22

## Main idea: From estimation to M-ary testing (intuition)

- Consider  $M$  finite functions  $f^i, i = 1, \dots, M$  spread across  $\mathcal{F}$  s.t. pairwise distances  $> 2\delta$  (i.e. is a packing of  $\mathcal{F}$ )
- If  $\mathcal{A}$  can use data  $\mathcal{D}$  to find  $\hat{f} = \mathcal{A}(\mathcal{D})$  (black dot) that is  $\delta$  close to  $F(\mathbb{P}) \rightarrow$  for data drawn from  $\mathbb{P}_{f^j}$ , we can use  $\mathcal{A}$  to correctly identify  $f^j$  by choosing the closest  $f^i$  (blue dot) to the estimated  $\hat{f}$   
 $\rightarrow$  this is a procedure that successfully tests the hypotheses that  $H_j : F(\mathbb{P}) = f^j$ .



- As we want a lower bound on estimation, can reverse the argument  
 $\rightarrow$  Problem reduces to: **given  $n$  points drawn from  $\mathbb{P}_{f^i}$  for some  $i$ , how far do they have to be apart (i.e.  $\delta$  lower bound) so that we can accept the right  $H_j$ ?**

6 / 22

## Main idea: from estimation to M-ary testing

- 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 joint distribution  $\mathbb{Q}_M$  over  $\mathcal{D}, J$  with  $J$  being a mixing variables:
  1.  $J$  a uniform R.V. (flat “prior”) with values in  $[M] = \{1, \dots, M\}$ , 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  $\{f^i\}_{i=1}^{M(2\delta)} \subset \mathcal{F}$  such that  $\min_{i \neq j} \|f_i - f_j\| \geq 2\delta$  so that  $M(2\delta) \leq \mathcal{M}(2\delta; \mathcal{F}, \|\cdot\|)$ , then

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

7 / 22

## 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. Denoting  $\mathbb{P}^n = \mathbb{P}^{\otimes n}$ , 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. 6):

$$\mathbb{Q}(\{\|\mathcal{A}(\mathcal{D}) - f^i\| \leq \delta\} | J = i) \leq \mathbb{Q}(\{\psi_{\mathcal{A}}(\mathcal{D}) = i\} | J = i)$$

because for any  $f \in \mathcal{F}$  such that  $\|f - f^i\| < \delta$ , for any  $j \neq i$  we have  $\|f - f^j\| > \|f^j - f^i\| - \|f - f^i\| > \delta \rightarrow \psi_{\mathcal{A}}(\mathcal{D}) = i$

3. Then the Lemma follows by the distribution of  $J$

$$\begin{aligned} \delta^{-2} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\mathcal{D} \sim \mathbb{P}} \|\mathcal{A}(\mathcal{D}) - F(\mathbb{P})\|^2 &\stackrel{1.}{\geq} \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{P}^n(\|\mathcal{A}(\mathcal{D}) - F(\mathbb{P})\| > \delta) \\ &\geq \frac{1}{M} \sum_{i \in [M]} \mathbb{P}_{f^i}^n(\|\mathcal{A}(\mathcal{D}) - f^i\| > \delta) = \sum_{i \in [M]} \mathbb{Q}(J = i) \mathbb{Q}(\|\mathcal{A}(\mathcal{D}) - f^i\| > \delta | J = i) \\ &\stackrel{2.}{\geq} \sum_{i \in [M]} \mathbb{Q}(J = i) \mathbb{Q}(\{\psi_{\mathcal{A}}(\mathcal{D}) \neq i\} | J = i) = \mathbb{Q}(\{\psi_{\mathcal{A}}(\mathcal{D}) \neq J\}) \quad \square \end{aligned}$$

8 / 22

## Lower bound methods

*Note:* In what follows we sometimes sloppily refer to such  $2\delta$ -separated sets as “packings” even though for  $2\delta$ -packings we require  $\|f_i - f_j\| > 2\delta$  with strict inequality.

There are now three different types of methods

- LeCam’s method: binary testing (a “packing” of two distributions) using the TV distance between them (see MW Chapter 15)
- Fano’s method: using  $M$ -ary testing with  $M > 2$  and a packing and mutual information/KL divergence (covered today)
- Assuoad’s method (doesn’t use the Lemma above, but the reduction via discretization is “built-in”): constructs a family of functions indexed by points on the hypercube, where far points on the hypercube corresponds to “far” functions. The problem then reduces to finding a point on the hypercube with small Hamming distance to the true one - or for a lower bound, prove that the minimax Hamming distance is large. You then translate that to a minimax distance over functions (covered next time)

9 / 22

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

For simplicity assuming densities of joint and conditional distributions:

### Definition (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 1 (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 7. Then for any decision/testing function  $\psi$ , it holds that

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

10 / 22

## Fano's method to lower bound minimax risk

- We can now get a lower bound on estimation by plugging in Fano's lower bound into the lemma and using a  $2\delta$ -packing of size  $M(2\delta)$
- If we choose  $\{f^i\}_{i=1}^{M(2\delta)}$  to be a  $2\delta$ -packing of  $\mathcal{P}$  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  $\delta$  is chosen such that  $I(\mathcal{D}, J) \sim \log M(2\delta)$  then the Lemma implies a lower bound of order  $\delta^2$
- This might or might not be a tight lower bound (if it matches some algorithm dependent upper bound, you're in luck)
- Next, we explore a coarse and more refined way to upper bound the mutual information

11 / 22

## Upper bounding the mutual information

- Denote by  $\mathbb{Q}_{\mathcal{D}}$  and  $\mathbb{Q}_J$  the marginal distribution over  $\mathcal{D}$  and  $J$  derived from the joint  $\mathbb{Q}_M$ .

### Definition (Kullback-Leibler divergence)

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

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

Further, by definition it follows that  $I(\mathcal{D}, J) = KL(\mathbb{Q} \parallel \mathbb{Q}_{\mathcal{D}}\mathbb{Q}_J)$  (check yourself). We can then establish

### Lemma (Intermediate bound)

For any family  $\{\mathbb{P}_i\}_{i=1}^M$  and  $\mathcal{D}, J$  following the distribution  $\mathbb{Q}_M$  defined above, we have

$$I(\mathcal{D}, J) \leq \min_{\mathbb{Q}} \max_{i=1, \dots, M} KL(\mathbb{P}_i^n \parallel \mathbb{Q})$$

In our case note that  $\mathbb{P}_i = \mathbb{P}_{f^i}$  and  $M = M(2\delta)$ .

12 / 22

## Exercise

To prove the intermediate bound we will need the following fact: For any family of distributions  $\{\mathbb{P}_i\}_{i=1}^M$  and a distribution  $\mathbb{Q}$  we have

$$\frac{1}{M} \sum_{i=1}^M KL(\mathbb{P}_i \parallel \frac{1}{M} \sum_{i=1}^M \mathbb{P}_i) \leq \frac{1}{M} \sum_{i=1}^M KL(\mathbb{P}_i \parallel \mathbb{Q}) \quad (1)$$

To familiarize yourself with the KL divergences, take 5 minutes to prove this fact with your neighbor!

Proof: For any  $\mathbb{Q}$  we can show (using the integral notation and densities is easiest) that

$$\begin{aligned} \frac{1}{M} \sum_{i=1}^M KL(\mathbb{P}_i \parallel \mathbb{Q}) &= \frac{1}{M} \sum_{i=1}^M KL(\mathbb{P}_i \parallel \frac{1}{M} \sum_{i=1}^M \mathbb{P}_i) + KL(\frac{1}{M} \sum_{i=1}^M \mathbb{P}_i \parallel \mathbb{Q}) \\ &\geq \frac{1}{M} \sum_{i=1}^M KL(\mathbb{P}_i \parallel \frac{1}{M} \sum_{i=1}^M \mathbb{P}_i) \end{aligned}$$

where the inequality follows from the fact that  $KL \geq 0$  by Jensen's inequality.

13 / 22

## Proof of intermediate bound

Denoting by  $q$  the (conditional and marginal) densities of the corresponding  $\mathbb{Q}$ , we have

$$\begin{aligned} KL(\mathbb{Q}_M \parallel \mathbb{Q}_D \mathbb{Q}_J) &= \mathbb{E}_J \mathbb{E}_{\mathcal{D}|J} \log \frac{q_{\mathcal{D}|J}}{q_D} = \mathbb{E}_J KL(\mathbb{Q}_{\mathcal{D}|J} \parallel \mathbb{Q}_D) \\ &= \frac{1}{M} \sum_{i=1}^M KL(\mathbb{P}_{f_i}^n \parallel \frac{1}{M} \sum_{j=1}^M \mathbb{P}_{f_j}^n) \\ &\leq \frac{1}{M} \sum_{i=1}^M KL(\mathbb{P}_{f_i}^n \parallel \mathbb{Q}) \text{ for any } \mathbb{Q} \\ &\leq \max_{i=1, \dots, M} KL(\mathbb{P}_{f_i}^n \parallel \mathbb{Q}) \text{ for any } \mathbb{Q} \end{aligned}$$

where the third line holds because of Inequality eq. 1 □

Note that this can work if for any two distributions in the space  $\mathcal{P}$ , we have bounded max KL divergence.

14 / 22

## Instantiating intermediate bound

Sometimes we can find a minimum  $\mathbb{Q}$  directly, that depends on whether you can control the minimax KL distance of any distribution to an element in your packing. A robust way to pick  $\mathbb{Q}$  yields the Yang-Barron inequality below.

- Define  $\mathcal{P}^n = \{\mathbb{P}^{\otimes n} : \mathbb{P} \in \mathcal{P}\}$
- The next theorem bounds the mutual information in Fano's method using the KL divergence

### Theorem 2 (Yang-Barron, MW Lemma 15.21)

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

For the proof we basically need to pick  $\mathbb{Q}$  as a mixture distribution of an *new*  $\epsilon$ -cover of  $\mathcal{P}$  in KL distance, and then picking the minimizing  $\epsilon$  variable of the RHS (see MW proof).

15 / 22

## Overall recipe to get minimax lower bounds ...

... using Yang-Barron + Fano to get lower bounds:

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

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

- We first use the pipeline in action for Sobolev functions
- Then we show the proof of Fano's method
- The example for sparse linear regression can be found in MW 15.16 (and you're expected to be able to explain it in the exam)

16 / 22

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

17 / 22

# 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

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

### Proof of Corollary

- a) First note that in our  $\mathcal{P}$ , the marginal distribution over  $X$  is the same for all  $f \in \mathcal{F}$ , so that we have for  $n = 1$

$$\begin{aligned} KL(\mathbb{P}_f \parallel \mathbb{P}_g) &= \frac{1}{2\sigma^2} \mathbb{E}_{\mathbb{P}_f} g^2(X) - f^2(X) + 2(f(X) - g(X))Y \\ &= \frac{1}{2\sigma^2} \mathbb{E}_{\mathbb{P}_f} g^2(X) - f^2(X) + 2(f(X) - g(X))f(X) = \frac{\|f - g\|_2^2}{2\sigma^2} \end{aligned}$$

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

$$\begin{aligned} KL(\mathbb{P}_f^n \parallel \mathbb{P}_g^n) &= \int \prod_{i=1}^n p_f(z_i) \log \prod_{i=1}^n \frac{p_f(z_i)}{p_g(z_i)} \mu(dz^n) \\ &= \sum_{i=1}^n \int p_f(z_i) \log \frac{p_f(z_i)}{p_g(z_i)} \mu(dz_i) = n \frac{\|f - g\|_2^2}{2\sigma^2} \end{aligned}$$

18 / 22

## 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) Now we can do step 1. in the lower bound pipeline: Using  $\log \mathcal{N}(\epsilon; \mathcal{W}_2^\alpha([0, 1]), \|\cdot\|_2) = O(\frac{1}{\epsilon})^{1/\alpha}$  we require

$$\epsilon^2 \geq \left(\frac{n}{2\sigma^2}\right)^{\frac{1}{2\alpha}} \epsilon^{-1/\alpha} \rightarrow \epsilon^2 = O\left(\frac{n}{\sigma^2}\right)^{\frac{1}{2\alpha+1}}$$

e) Following step 2. in the pipeline: Recalling that  $\mathcal{M}(2\delta) \geq \mathcal{N}(2\delta)$  it suffices to require

$$\left(\frac{1}{\delta}\right)^{\frac{1}{\alpha}} \geq c \left[ \left(\frac{n}{\sigma^2}\right)^{\frac{1}{2\alpha+1}} + 2 \log 2 \right] \rightarrow \delta^2 = O\left(\frac{\sigma^2}{n}\right)^{\frac{2\alpha}{2\alpha+1}}$$

satisfies the inequality in step 2. for  $\frac{n}{\sigma^2}$  larger than a universal constant.

f) Hence by 3. (Fano's method)  $\|\hat{f} - f^*\|_{\mathcal{L}^2([0,1])}^2 \geq O\left(\frac{\sigma^2}{n}\right)^{\frac{2\alpha}{2\alpha+1}} \quad \square$

19 / 22

## 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

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

• Proof then follows from noting that generally

$H(X|Y) = \mathbb{E}_Y H(X|Y=y)$ , we have

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

2. Since  $E_\psi$  Bernoulli  $H(E_\psi) \leq \log 2$  for all  $\psi$   
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.

20 / 22

## References

Minimax lower bounds

- MW Chapter 15

Additional reading

- *John Duchi “Information Theory” (Stats 311) Lecture Notes: Lectures 3, 5, 6*
- *Ryan Tibshirani “Statistical Learning” Lecture Notes: Minimax Theory for Nonparametric Regression*

Classical texts on minimax bounds:

- *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.*

21 / 22

## Metric entropy for higher order Sobolev spaces (bonus)

**Lemma (Metric entropy for  $\alpha$ -order compact Sobolev spaces)**

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

### 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  $\phi_j$  are orthonormal in  $\|\cdot\|_2$  norm we have  $\|f\|_2^2 = \|\theta_f\|_{\ell^2(\mathbb{N})}^2$

2. MW Example 5.12. proves  $\log \mathcal{N}(\delta; \mathcal{E}_\alpha, \|\cdot\|_{\ell^2(\mathbb{N})}) \leq O\left(\frac{1}{\delta}\right)^\frac{1}{\alpha}$   $\square$

22 / 22