

# Lecture 12: Random design and minimax lower bounds

1 / 20

## Announcements

- Please fill out your oral exam availabilities (slots will be updated and you should sign up when), taking place 17.11./18.11. 9 am - 5 pm
  - mark *all slots* where you do not have a strict conflict, it's hard enough to figure out a
  - exams are 25 minutes long, each slot is 35 minutes
- Regarding material in lecture 10: you are expected to understand and explain everything except for what's marked as "bonus" (see updated slides)

### Plan for today

- Random design proofs
- Minimax lower bounds

2 / 20

## Recap: Random design

We considered the case where  $X_1, \dots, X_n$  are sampled i.i.d. from a distribution  $\mathbb{P}$  and where we care about the squared  $\mathcal{L}^2(\mathbb{P})$  norm of the prediction error (for regression) denoted by

$$\|\hat{f} - f^*\|_2^2 = \mathbb{E}(\hat{f}(X) - f^*(X))^2.$$

- Also note that  $\|\hat{f} - f^*\|_2^2 = \mathbb{E}_{x_1, \dots, x_n} \|f - f^*\|_n^2$ .
- We had two results to get a bound on  $\|\hat{f} - f^*\|_2^2$ , depending on two different assumptions, both using the population localized Rademacher complexity

### Definition (localized *population* Rademacher complexity)

We define  $\mathcal{R}_n(\mathcal{F}^*; \delta) = \frac{1}{n} \mathbb{E}_{X, \epsilon} \sup_{g \in \mathcal{F}^*, \|g\|_2 \leq \delta} \sum_{i=1}^n \epsilon_i g(x_i)$

- Last time we spent most of the time instantiating them on the example of linear regression
- Today, we'll go through the proof idea behind both

3 / 20

## Recap: For uniformly bounded functions

### Definition (pop. critical radius for unif.-bounded functions)

For  $b$ -uniformly bounded  $\mathcal{F}^*$ , we can then define the corresponding critical radius  $\bar{\delta}_n$  as the smallest  $\delta > 0$  satisfying

$$\mathcal{R}_n(\mathcal{F}^*; \delta) \leq \frac{\delta^2}{16b}$$

### Theorem 1 (Localized uniform law, MW Thm 14.1)

For star-shaped and  $b$ -uniformly bounded  $\mathcal{F}^*$ , let  $\bar{\delta}_n$  as defined above.

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

$$\sup_{\Delta \in \mathcal{F}^*} \|\Delta\|_2 - \|\Delta\|_n \leq c \bar{\delta}_n$$

4 / 20

## Recap: For functions satisfying moment condition

### Definition (pop. critical radius for moment-bounded functions)

Assume that for all  $f \in \mathcal{F}^*$  with  $\|f\|_2 \leq 1$  we have  $\mathbb{E}f^4(X) \leq C^2 \mathbb{E}f^2(X)$ . Then we can define  $\bar{\delta}_n$  to be the smallest  $\delta > 0$  satisfying the population critical inequality

$$128C \mathcal{R}_n(\mathcal{F}^*; \delta) \leq \delta^2$$

### Theorem 2 (modified MW Cor. 14.15)

Assume that  $\mathcal{F}^*$  star-shaped, and moment condition holds. Suppose  $n$  large enough such that  $\delta_n \leq 1$ , then w/ probability greater than

$1 - c_1 e^{-\frac{c_2 n \bar{\delta}_n^2}{\sigma^2 + C^2}}$  over the training set  $\mathcal{D}$ , the least-squares estimate  $\hat{f} \in \mathcal{F}$  satisfies

$$\|\hat{f} - f^*\|_2^2 \leq c(\bar{\delta}_n^2 + \delta_n^2)$$

5 / 20

## Proof of Theorem: Intermediate Lemma

We'll provide a proof sketch of an intermediate result which yields the same conclusion:

### Lemma (MW Lemma 14.9. + peeling)

Under the assumptions of Theorem 1, with probability at least  $1 - e^{-\frac{cn\bar{\delta}_n^2}{2}}$ , we have

$$\sup_{\substack{\|\Delta\|_2 \geq \bar{\delta}_n \\ \Delta \in \mathcal{F}^*}} \|\Delta\|_2^2 - \|\Delta\|_n^2 \leq \|\Delta\|_2 \bar{\delta}_n$$

- Note that again, we only need to bound  $\Delta$  with  $\|\Delta\|_2 \leq \bar{\delta}_n$  since else the statement trivially holds
- From the Lemma (and the Theorem), we obtain that  $\|\hat{f} - f^*\|_2 \leq O(\bar{\delta}_n)$  if  $\|\hat{f} - f^*\|_n \leq O(\bar{\delta}_n)$ .

6 / 20

## Proof ansatz: Self-bounding

- Previously for empirical norm error bound we used

$$\|\Delta\|_n^2 \leq \sup_{\substack{\|\Delta\|_2 \geq \delta_n \\ \Delta \in \mathcal{F}^*}} \frac{1}{n} \sum_{i=1}^n w_i \Delta(x_i) \leq c_1 \|\Delta\|_n \delta_n$$

- Recall that the first inequality is a result of the basic inequality, and the second follow from concentration of Lipschitz functions of (sub)Gaussians. Notably, we are “self-bounding” the middle term, that is, the upper bound depends on the norm  $\|\Delta\|_n$  (that then cancels with square on the LHS)
- Now to bound  $\|\Delta\|_2$ , we “piggy-back” on the optimality and basic inequality and bound the difference between  $\|\Delta\|_2^2$  and  $\|\Delta\|_n^2$  again with “self-bounding” but with the norm  $\|\Delta\|_2$

$$\sup_{\substack{\|\Delta\|_2 \geq \delta_n \\ \Delta \in \mathcal{F}^*}} \|\Delta\|_2^2 - \|\Delta\|_n^2 \leq \|\Delta\|_2 \bar{\delta}_n$$

To prove this we use 1. Symmetrization, Rademacher contraction and concentration and 2. A peeling argument akin to SRM

7 / 20

## Proof sketch of Lemma 1: Part I

For simplicity of the proof, assume  $b = 1$  and hence  $\|\Delta\|_2 \leq 1$  (general case follows from scaling arguments as last time)

1. Step: For fixed  $r = t\bar{\delta}_n$  with  $t \geq 1$ , bounding  $\sup_{\Delta \in \mathcal{F}^*, \|\Delta\|_2 \leq r} \|\Delta\|_2^2 - \|\Delta\|_n^2$  (MW Lemma 14.9.)

- for  $r = t\bar{\delta}_n$ , by symmetrization and Rademacher contraction:

$$\begin{aligned} \mathbb{E} \sup_{\Delta \in \mathcal{F}^*, \|\Delta\|_2 \leq r} \|\Delta\|_2^2 - \|\Delta\|_n^2 &\leq 2\mathbb{E} \sup_{\Delta \in \mathcal{F}^*, \|\Delta\|_2 \leq r} \frac{1}{n} \sum_{i=1}^n \epsilon_i \Delta^2(x_i) \\ &\leq 4\mathbb{E} \sup_{\Delta \in \mathcal{F}^*, \|\Delta\|_2 \leq r} \frac{1}{n} \sum_{i=1}^n \epsilon_i \Delta(x_i) \leq r\bar{\delta}_n \end{aligned}$$

where the last inequality follows from definition of  $\bar{\delta}_n$

- we then use Talagrand concentration/functional Bernstein (MW Thm 3.27) to derive that w.p  $\geq 1 - e^{-cn\bar{\delta}_n^2}$  we have  $\sup_{\Delta \in \mathcal{F}^*, \|\Delta\|_2 \leq r} \|\Delta\|_2^2 - \|\Delta\|_n^2 \leq \frac{r\bar{\delta}_n}{2}$  (see slide after next)

8 / 20

## Proof sketch of Lemma 1: Part II

2. Step: If we could plug in  $r = \|\Delta\|_2$  we'd be done, but above h.p. bound only holds for fixed  $r$ !

- Use peeling argument like before and split

$S := \{\sup_{\Delta \in \mathcal{F}^*, \|\Delta\|_2 \geq \bar{\delta}_n} \|\Delta\|_2^2 - \|\Delta\|_n^2 \geq \|\Delta\|_2 \bar{\delta}_n\}$  into sub-events:

$S_m = \{\|\Delta\|_2 \in [r_{m-1}, r_m]\}$  where  $r_m = t_m \bar{\delta}_n := 2^m \bar{\delta}_n$ . In particular, by uniform boundedness  $\|\Delta\|_2 \leq 1$ , we have that

$S \subset \bigcup_{m=1}^M \{S \cap S_m\}$  with  $M = 4 \log(1/\bar{\delta}_n)$

- using  $\sup_{\Delta \in \mathcal{F}^*, \|\Delta\|_2 \leq t \bar{\delta}_n} \|\Delta\|_2^2 - \|\Delta\|_n^2 \leq \frac{t \bar{\delta}_n^2}{2}$  w.p. at least  $1 - e^{-cn \bar{\delta}_n^2}$  with  $t = t_m$  and using union bound gives

$$\begin{aligned} \mathbb{P}(S) &\leq \sum_{m=1}^M \mathbb{P}(S \cap S_m) \leq \sum_{m=1}^M \mathbb{P}\left(\sup_{\Delta \in \mathcal{F}^*, \|\Delta\|_2 \leq t_m \bar{\delta}_n} \|\Delta\|_2^2 - \|\Delta\|_n^2 \geq \frac{t_m \bar{\delta}_n^2}{2}\right) \\ &\leq \sum_{m=1}^M e^{-cn \bar{\delta}_n^2} \leq e^{-cn \bar{\delta}_n^2 + \log M} \leq e^{-cn \bar{\delta}_n^2} \end{aligned}$$

9 / 20

## Some intuition on functional Bernstein Part I

Recall that we needed w.p.  $\geq 1 - e^{-cn \bar{\delta}_n^2}$  we have

$$\sup_{\Delta \in \mathcal{F}^*, \|\Delta\|_2 \leq r} \|\Delta\|_2^2 - \|\Delta\|_n^2 \leq \frac{r \bar{\delta}_n}{2} \quad (1)$$

for which we required following Theorem 3 instead of Theorem 4

### Theorem 3 (Functional Bernstein, MW Thm 3.27)

For  $\mathcal{F}$   $b$ -uniformly bounded, define random quantity

$\gamma = \sup_{f \in \mathcal{F}} \mathbb{E} f^2(X) + 2b \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n f(X_i)$ . Then

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n f(X_i) \geq \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n f(X_i) + \delta\right) \leq e^{-\frac{cn \delta^2}{\gamma + 4b\delta}}$$

### Theorem 4 (Functional Hoeffding, modified MW Thm 3.26)

Given deterministic quantity  $L^2 := \sup_{f \in \mathcal{F}} (\sup_x f(x) - \inf_x f(x))$

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n f(X_i) \geq \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n f(X_i) + \delta\right) \leq e^{-\frac{n \delta^2}{4L^2}}$$

10 / 20

## Some intuition on functional Bernstein Part II

We can use the theorems to get eq. 1 by plugging in

$$f(X_i) = \mathbb{E}\Delta^2(X) - \Delta^2(X_i)$$

Using Theorem 3 instead of Theorem 4 is a key technique to prove tight rates with localization:

- With Theorem 4, we'd only be able to achieve an upper bound of  $e^{-cnt^2\bar{\delta}_n^4} = e^{-cnr^2\bar{\delta}_n^2}$  when  $r = t\bar{\delta}_n$ . Caveats:
  - forces slow deviation terms of  $n^{-1/4}$  dominating possibly faster rate of  $\bar{\delta}_n$
  - dependence on  $t$ , would not allow peeling argument
- Instead, Theorem 3 gives us a probability of  $e^{-cn\bar{\delta}_n^2}$
- The key here is the denominator in the exponent of the probability. We typically have a *denominator*  $L^2$  in the exponent in  $\frac{n\bar{\delta}_n^4 t^2}{L^2}$ , where  $L$  is e.g. the Lipschitz constant, the length of the boundedness interval or, similarly, the subgaussian parameter.

11 / 20

## Some intuition on functional Bernstein Part III

- For bounding the empirical norm specifically (MW Lemma 13.23), using basic inequality the term we want to bound is a Lipschitz function of Gaussians, so that  $L = t\bar{\delta}_n$  was the (scaled) Lipschitz constant cancelling out  $t^2\bar{\delta}_n^2$ .
- Here, we are not dealing with Lipschitz functions of (sub)Gaussians, and using functional Hoeffding MW Thm 3.27 would lead to a constant denominator  $L$  which we can't afford. Instead, we can use a "Bernstein" refinement: if the variance of the summand (appearing as  $\gamma$ ) is bounded by a term comparable with  $r^2 = t^2\bar{\delta}_n^2$ , then we're also good

You can/should follow the exact mathematical steps how this happens for functional Bernstein in the proof of MW Lemma 14.9. that we skip in class.

12 / 20

## Proof sketch of Theorem 2

As always, if  $\|\hat{\Delta}\|_2^2 \leq \bar{\delta}_n^2$  we're done, hence focus on the case when  $\|\hat{\Delta}\|_2^2 \geq \bar{\delta}_n^2$ . For that case we use the following lemma

### Lemma (MW Thm 14.12.)

Assume  $\mathcal{F}^*$  star-shaped and moment conditions in Theorem hold. For any  $\hat{\Delta} \in \mathcal{F}^*$  with  $\hat{\Delta} \geq \bar{\delta}_n^2$  we have

$$\|\hat{\Delta}\|_2^2 \leq 2\|\hat{\Delta}\|_n^2 \quad w/ \text{prob} \geq 1 - e^{-c \frac{n\bar{\delta}_n^2}{c^2}}$$

We can now use this lemma in two events separately:

- $\bar{\delta}_n \geq \delta_n$ : Use fixed design bound on  $\|\hat{\Delta}\|_n^2$  with  $\bar{\delta}_n$  (follow directly using  $t = \frac{\bar{\delta}_n^2}{\delta_n^2} \geq 1$ ) and then Lemma to get desired bound on  $\|\hat{\Delta}\|_2^2$

13 / 20

## Proof sketch of Theorem (ctd)

- $\bar{\delta}_n < \delta_n$ : We want to bound  $\mathcal{E} = \{\|\hat{\Delta}\|_2^2 \geq 16\delta_n^2 + 2\bar{\delta}_n^2\}$  and further split this event into two cases and use for conditional probability  $\mathbb{P}(\mathcal{E}) = \mathbb{P}(\mathcal{E} \cap \mathcal{B}) + \mathbb{P}(\mathcal{E} \cap \mathcal{B}^c) \leq \mathbb{P}(\mathcal{E} \cap \mathcal{B}) + \mathbb{P}(\mathcal{B}^c)$ 
  - Event  $\mathcal{B} = \{\|\hat{\Delta}\|_n^2 \leq 8\delta_n^2\} \subset \{\|\hat{\Delta}\|_2^2 \geq 2\|\hat{\Delta}\|_n^2 + 2\bar{\delta}_n^2\}$  which is directly bounded using Lemma
  - Event  $\mathcal{B}^c = \{\|\hat{\Delta}\|_n^2 > 8\delta_n^2\}$  has probability  $\leq e^{-\frac{n\delta_n^2}{2\sigma^2}} \leq e^{-\frac{n\bar{\delta}_n^2}{2\sigma^2}}$  using fixed design bound where last inequality holds by the case assumption  $\bar{\delta}_n < \delta_n$ . □

The main idea behind the proof of Lemma is: use concentration on truncated (uniformly bounded) functions  $\hat{\Delta}_\tau$  in  $\mathcal{F}^*$  and **moment condition** allows us to bound  $\|\hat{\Delta}\|_2$  via  $c\|\hat{\Delta}_\tau\|_2$  with small constant  $c > 1$

14 / 20

## Proof of Lemma 2

Truncate  $\hat{\Delta}$  at  $\tau$  to  $\hat{\Delta}_\tau$  (i.e. all values beyond  $\pm\tau$  are set at  $\pm\tau$ ) then

1) show for truncated part  $\sup_{\|\hat{\Delta}_\tau\|_2 \leq \delta} \|\hat{\Delta}_\tau\|_n^2 - \|\hat{\Delta}_\tau\|_2^2 \leq \bar{\delta}_n^2$

with probability at least  $1 - c_1 e^{-c_2 \frac{n\bar{\delta}_n^2}{C\delta^2 + C^2\bar{\delta}_n}}$

2) choose  $\tau$  such that  $\hat{\Delta}_\tau$  and  $\hat{\Delta}$  are “close” (i.e. via moment condition, we get something like what we need for boundedness of  $\|\hat{\Delta}\|_2$ )

For 1) concentration of  $\|\hat{\Delta}_\tau\|_n^2$ :

- use functional Bernstein (McDiarmid-like) (for now bounded function!) as tail bound
- for expectation use symmetrization & contraction for Rademacher  $\rightarrow \leq \bar{\delta}_n$

For 2) approximation error of the truncation:

- Choice of  $\tau^2 = 4C^2$  yields  $\|\hat{\Delta}\|_2^2 - \|\hat{\Delta}_\tau\|_2^2 \leq \frac{1}{4}\|\hat{\Delta}\|_2^2$
- via Cauchy Schwartz 4–th order moment condition □

15 / 20

## Notion of optimality

- 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  $y = f^*(x) + w$  (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  $F(\mathbb{P})(x) = \mathbb{E}[Y|x] = f^*(x)$ .
- 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  $\hat{f}_{\mathcal{D}} = \mathcal{A}(\mathcal{D})$
- So far we've seen: Error bounds of the type  $\|\hat{f}_{\mathcal{D}} - f^*\|_2^2 \leq O(n^{-\alpha})$

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?

16 / 20

## Discussion on reasonable notion of optimality

- First we write for  $\mathcal{D} = \{z_1, \dots, z_n\}$  with  $Z_i \stackrel{i.i.d.}{\sim} \mathbb{P}$  that  $\mathcal{D} \sim \mathbb{P}^n$
- What the best algorithm could achieve on this distribution is  $\inf_{\mathcal{A}} \mathbb{E}_{\mathcal{D} \sim \mathbb{P}^n} \|\mathcal{A}(\mathcal{D}) - F(\mathbb{P})\|^2$  where the expectation is over the training data.
- However, this does not make sense because for any fixed  $\mathbb{P}$ , we could just output a constant *independent from the data*, that is  $\mathcal{A}(\mathcal{D}) \equiv F(\mathbb{P})$ .
- But of course this is a “stupid” algorithm, since you have to fix  $\mathbb{P}$  before seeing data, and you’ll do very poorly on  $\mathcal{D}$  that is drawn from a different distribution!
- Hence what we need is an algorithm that does well on all possible  $\mathbb{P}$  where  $\mathcal{D}$  could be sampled from. This is captured in the definition of the minimax risk

17 / 20

## Minimax risk

### Definition (Minimax risk)

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

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

- $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$  has i.i.d. samples from  $\mathbb{P}^n \rightarrow \mathcal{A}(\mathcal{D})$  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)

18 / 20

## 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

Next week: 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

19 / 20

## References

Random design bound proofs

- MW Chapter 3, 14

Minimax risk

- 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.*

20 / 20