

## Lecture 9: Localization and non-parametric regression

1 / 17

### Recap: Caveats of using uniform law to bound the generalization gap

- First of all, notice the “slow” uniform excess risk bound holds for any  $\mathcal{F}$ , including ones for which  $f^* \notin \mathcal{F}$ !
- Further, in our argument using uniform law, we used optimality of  $\hat{f}_n$  only once

$$R(\hat{f}_n) - R(f^*) = R(\hat{f}_n) - R_n(\hat{f}_n) + \overbrace{R_n(\hat{f}_n) - R_n(f^*)}^{\leq 0 \text{ by optimality}} + R_n(f^*) - R(f^*)$$

Today and next few classes: using *localized complexities* to prove tighter bounds for *particular estimator*:  
global minimizer of *square loss for regression!*

- Idea: By using **optimality of  $\hat{f}$**  instead of uniform bound
  1. circumvent uniform boundedness of losses
  2. can upper bound uniformly in a more restricted function space

2 / 17

## Recap: (Non-)parametric regression setting

We now set up constrained regression using the square loss that we use in the next few classes

- Fixed design, i.e. only care about prediction on training inputs  $x_1, \dots, x_n$
- Gaussian observation noise  $\sigma W = Y - f^*(X)$  with  $W \sim \mathcal{N}(0, 1)$
- Minimizer of the square loss:  
 $\hat{f} = \arg \min_{f \in \mathcal{F}} R_n(f) := \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2$   
(and later minimizer of penalized square loss  
 $\hat{f} = \arg \min_{f \in \mathcal{F}} R_n(f) := \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{F}}$ )
- Evaluation: Prediction error of some  $f$  on fixed design points

$$\|f - f^*\|_n^2 = \frac{1}{n} \sum_{i=1}^n (f(x_i) - f^*(x_i))^2 = \mathbb{E}_Y R_n(f) - \sigma^2 = R(f) - R(f^*)$$

3 / 17

## Basic inequality circumventing boundedness and more

Optimality of  $\hat{f}$  yields the *basic inequality*

$$R_n(\hat{f}) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2 \leq \frac{1}{n} \sum_{i=1}^n (y_i - f^*(x_i))^2 = R_n(f^*) \tag{1}$$
$$\|\hat{f} - f^*\|_n^2 \leq \frac{2\sigma}{n} \sum_{i=1}^n w_i (\hat{f}(x_i) - f^*(x_i))$$

where  $w_i$  are i.i.d. standard normals.

- Upper bounding by the sup on the RHS and taking expectations on both sides, and defining  $\mathcal{F}^* = \mathcal{F} - f^* = \{f - f^* : f \in \mathcal{F}\} \rightarrow$   
 $\mathbb{E} \|\hat{f} - f^*\|_n^2 \leq 2\sigma \tilde{\mathcal{G}}_n(\mathcal{F}^*(x_1^n)) :=$   
 $\mathbb{E}_{w_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)} \sup_{g \in \mathcal{F}^*} \frac{2\sigma}{n} \sum_{i=1}^n w_i g(x_i)$
- This expectation resembles the empirical R.C. of the set  $\mathcal{F}^*(x_1^n) \subset \mathbb{R}^n$  - just with Rademacher replaced by Gaussian R.V. Indeed this is the **empirical Gaussian complexity**  $\tilde{\mathcal{G}}_n(\mathcal{F}^*(x_1^n))$ , of the same "order" as R.C.: as the following sandwich relationship holds (proved in HW 2), for each  $\mathbb{T}$ :  $\frac{1}{2 \log n} \tilde{\mathcal{G}}_n(\mathbb{T}) \leq \tilde{\mathcal{R}}_n(\mathbb{T}) \leq \sqrt{\frac{\pi}{2}} \tilde{\mathcal{G}}_n(\mathbb{T})$

4 / 17

## Motivation for localized Gaussian complexity

- Let's assume that we could magically use some concentration argument to show that  $\|\hat{f} - f^*\|_n^2$  is close to its expectation. Since

$$\mathbb{E}\|\hat{f} - f^*\|_n^2 \leq 2\sigma \tilde{\mathcal{G}}_n(\mathcal{F}^*(x_1^n)), \quad (2)$$

all we'd need to bound is the empirical G.C. What happens when we try to do that?

Discuss with neighbor:

- Compute the bound for the example of finding the best sparse linear fit

$$\hat{f} = \arg \min_{f \in \mathcal{F}_{lin,s}} \|y - f(x_1^n)\|_n^2$$

where  $\mathcal{F}_{lin,s} = \{f(\cdot) = \langle \theta, x \rangle : \|\theta\|_0 \leq s\}$ , equivalent to finding

$$\hat{\theta} = \arg \min_{\{\theta : \|\theta\|_0 \leq s\}} \|y - X\theta\|_n^2$$

- What's an upper bound of

$$\tilde{\mathcal{G}}_n(\mathcal{F}^*(x_1^n)) := \mathbb{E}_{w_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)} \sup_{g \in \mathcal{F}_{lin,s}^*} \frac{1}{n} \sum_{i=1}^n w_i g(x_i)?$$

5 / 17

## Solution

- First note that generally  $\tilde{\mathcal{G}}_n(\mathcal{F}^*(x_1^n)) = \tilde{\mathcal{G}}_n(\mathcal{F}(x_1^n))$  since for fixed  $f^*$ , we have  $\mathbb{E}_{w_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)} \frac{1}{n} \sum_{i=1}^n w_i f^*(x_i) = 0$
- Now observe that 
$$\tilde{\mathcal{G}}_n(\mathcal{F}_{lin,s}(x_1^n)) = \mathbb{E}_{w_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)} \sup_{\|\theta\|_0 \leq s} \left\langle \frac{2\sigma}{n} \sum_{i=1}^n w_i x_i, \theta \right\rangle$$
- Since  $\theta$  is unbounded, if for any vector  $w$ ,  $\sum_{i=1}^n w_i x_i \neq 0$  (which is true unless  $X = 0$ ) this expectation is unbounded, because  $\theta$  is unbounded. Hence, this empirical Gaussian complexity is generally unbounded and we can't use it!

**Problem: We used optimality but still have sup over huge function space  $\mathcal{F}^*$ !**

**The trick is to notice eq. 1 restricts function space!**

## Idea behind Localization I

- Define  $\hat{\Delta} = \hat{f} - f^*$  and recall  $\mathcal{F}^* = \{f - f^* : f \in \mathcal{F}\}$
- Let's assume that  $\mathcal{F}^*$  is **star-shaped**, i.e. for any  $f \in \mathcal{F}^*$ , we have  $\alpha f \in \mathcal{F}^*$  for all  $\alpha \in [0, 1]$
- The basic inequality then reads  $\|\hat{\Delta}\|_n^2 \leq \frac{2\sigma}{n} \sum_{i=1}^n w_i \hat{\Delta}(x_i)$

Note that there is some “self-bounding” here and what we did before is to throw away the fact that we don't have to take the *sup* over all  $\hat{\Delta} \in \mathcal{F}^*$  for the RHS. Instead, localization takes the following route:

1. Space to control is smaller than all of  $\mathcal{F}^*$  since for any  $\delta_n$ , either
  - (i)  $\|\hat{\Delta}\|_n \leq \delta_n$  or
  - (ii) if  $\|\hat{\Delta}\|_n \geq \delta_n$  then still  $\|\hat{\Delta}\|_n^2 \leq \frac{2\sigma}{n} \sum_{i=1}^n w_i \hat{\Delta}(x_i)$  by basic inequality
2. Further for case (ii), if we can show that w.h.p.

$$\frac{2\sigma}{n} \sum_{i=1}^n w_i \hat{\Delta}(x_i) \leq 4\|\hat{\Delta}\|_n \delta_n \quad (3)$$

for all  $\|\hat{\Delta}\|_n \geq \delta_n$  then we can plug that into RHS of (ii) to obtain  $\|\hat{\Delta}\|_n \leq 4\delta_n$  w.h.p.

7 / 17

## Idea behind Localization II

For which  $\delta_n$  is (2) true?

- a. By star-shaped assumption on  $\mathcal{F}^*$  and  $\|\hat{\Delta}\|_n \geq \delta_n$  (used in step(a)):

$$\begin{aligned} \iff \sup_{\|\hat{\Delta}\|_n \geq \delta_n, \hat{\Delta} \in \mathcal{F}^*} \frac{\sigma}{n} \sum_{i=1}^n w_i \frac{\hat{\Delta}(x_i)}{\|\hat{\Delta}\|_n} &= \sup_{\|\hat{\Delta}\|_n \geq \delta_n, \hat{\Delta} \in \mathcal{F}^*} \frac{\sigma}{n} \sum_{i=1}^n w_i \underbrace{\frac{\hat{\Delta}(x_i) \delta_n}{\|\hat{\Delta}\|_n}}_{=:\tilde{\Delta}} \frac{1}{\delta_n} \\ \stackrel{(a)}{=} \sup_{\|\tilde{\Delta}\|_n = \delta_n, \tilde{\Delta} \in \mathcal{F}^*} \frac{\sigma}{n} \sum_{i=1}^n w_i \frac{\tilde{\Delta}(x_i)}{\delta_n} &\leq \sup_{\|\tilde{\Delta}\|_n \leq \delta_n, \tilde{\Delta} \in \mathcal{F}^*} \frac{\sigma}{n} \sum_{i=1}^n w_i \frac{\tilde{\Delta}(x_i)}{\delta_n} \end{aligned}$$

The key here is that only the normalized error shows up in expression

- b. Now assume that  $\delta_n$  satisfies the following “recursive” inequality:

$$\mathbb{E} \sup_{\substack{\|\tilde{\Delta}\|_n \leq \delta_n \\ \tilde{\Delta} \in \mathcal{F}^*}} \frac{\sigma}{n} \sum_{i=1}^n w_i \tilde{\Delta}(x_i) \leq \delta_n^2 \quad (4)$$

This is called the critical inequality.

8 / 17

## Idea behind Localization III

- c. Then, if we can show concentration of the above supremum to its expectation, we'd have w.h.p.

$$\sup_{\substack{\|\tilde{\Delta}\|_n \leq \delta_n \\ \tilde{\Delta} \in \mathcal{F}^*}} \frac{\sigma}{n} \sum_{i=1}^n w_i \hat{\Delta}(x_i) \leq \mathbb{E} \sup_{\substack{\|\hat{\Delta}\|_n \leq \delta_n \\ \hat{\Delta} \in \mathcal{F}^*}} \frac{\sigma}{n} \sum_{i=1}^n w_i \hat{\Delta}(x_i) + \delta_n^2 \leq 2\delta_n^2$$

eq. 3 then follows immediately.

- Together with the basic inequality, you now obtain:  $\|\hat{\Delta}\|_n \leq 4\sqrt{t}\delta_n$  holds for any  $t \geq 1$  w.h.p. if  $\delta_n$  is the **smallest**  $\delta > 0$  such that  $\sigma\tilde{\mathcal{G}}_n(\mathcal{F}^*; \delta) \leq \delta^2$
- All that's left to do: see that  $\delta_n$  exists for b. and show c.

9 / 17

## Localized Gaussian complexity & critical radius

In the above argument, a key object was the “localized” (extra bound on the  $\|\cdot\|_n$  norm) version of the empirical Gaussian complexity:

### Definition (Localized (empirical) Gaussian complexity)

The localized Gaussian complexity around  $f^*$  of scale  $\delta$  is

$$\sigma\tilde{\mathcal{G}}_n(\mathcal{F}^*; \delta) := \sigma\tilde{\mathcal{G}}_n(\mathcal{F}^*(x_1^n) \cap \mathbb{B}_n(\delta)) = \mathbb{E} \sup_{\|\hat{\Delta}\|_n \leq \delta_n, \hat{\Delta} \in \mathcal{F}^*} \frac{\sigma}{n} \sum_{i=1}^n w_i \hat{\Delta}(x_i)$$

And in step b we had an assumption on  $\delta_n$

### Lemma (Critical radius (MW 13.6.))

For any star-shaped  $\mathcal{F}$ , it holds that  $\frac{\tilde{\mathcal{G}}_n(\mathcal{F}; \delta)}{\delta}$  is non-increasing and the critical inequality

$$\frac{\tilde{\mathcal{G}}_n(\mathcal{F}; \delta)}{\delta} \leq \frac{\delta}{\sigma}$$

has a smallest solution  $\delta_n > 0$  that we call the critical quantity/radius.

10 / 17

# Illustration of localized Gaussian complexity

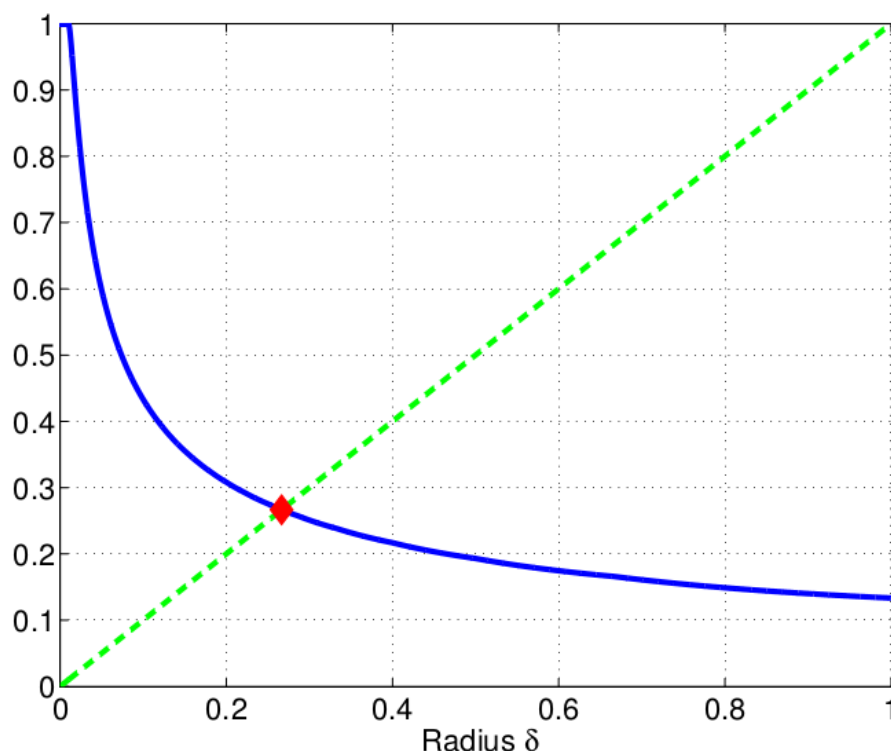


Figure 1: Blue solid:  $f(\delta) = \frac{\tilde{\mathcal{G}}_n(\mathcal{F}; \delta)}{\delta}$ , Green dashed:  $f(\delta) = \delta$

11 / 17

## Prediction error bound for constrained $^2$ -loss minimizer

With these concepts we can now formally establish our error bound

### Theorem (Prediction error bound, MW Thm 13.5.)

Assume  $\mathcal{F}^*$  is star-shaped and let  $\delta_n$  be any positive solution to the critical inequality eq. 4. Then, we have for the square loss minimizer  $\hat{f}$  for any  $t \geq 1$

$$\mathbb{P}(\|\hat{f} - f^*\|_n^2 \geq 16t\delta_n^2) \leq e^{-\frac{nt\delta_n^2}{2\sigma^2}}$$

- Plugging in  $t = O(\log \frac{1}{\delta})$  yields that probability at least  $1 - \delta$  we have  $\|\hat{f} - f^*\|_n^2 \leq O(\log(\frac{1}{\delta})\delta_n^2)$ . For small prob,  $\delta_n^2 \geq O(1/n)$ .
- As  $f^*$  is unknown, can replace  $\tilde{\mathcal{G}}_n(\mathcal{F}^*; \delta)$  by  $\tilde{\mathcal{G}}_n(\mathcal{F} - \mathcal{F}; \delta)$  (or its star hull MW Eq (13.21.)) to define critical radius  $\delta_n$
- Note: the notation for  $t$  is different from MW Thm 13.5.
- Proof follows by proof of (modified) c. and noting that

$$g_n(w) = \sup_{\|\hat{\Delta}\|_n \leq \sqrt{t}\delta_n, \hat{\Delta} \in \mathcal{F}^*} \frac{\sigma}{n} \sum_{i=1}^n w_i \hat{\Delta}(x_i)$$

Gaussians and using MW Thm 2.26 (next slide, skipped in class)

12 / 17

## Proof of error bound: tail bounding $g_n(w)$ (skipped)

We now establish the tail bound for  $g_n(w)$

1.  $g_n(w)$  as a function of  $w_i \sim \mathcal{N}(0, 1)$  is  $\frac{\sigma\sqrt{t\delta_n}}{\sqrt{n}}$ -Lipschitz (check yourself) so that  $\mathbb{P}(g_n(w) \geq \mathbb{E}g_n(w) + s) \leq e^{-\frac{ns^2}{2\sigma^2 t\delta_n^2}}$  (see MW Thm 2.26). Note this is only possible because we consider the localized (i.e. bounded domain) version of the supremum!
2. Furthermore  $\mathbb{E}g_n(w) = \sigma\tilde{\mathcal{G}}_n(\mathcal{F}; \sqrt{t\delta_n})$
3. The map  $\delta \rightarrow \frac{\tilde{\mathcal{G}}_n(\mathcal{F}; \delta)}{\delta}$  is non-increasing by MW Lemma 13.6.
4. By 2. and assumption on  $\delta_n$  we have  $\sigma\frac{\tilde{\mathcal{G}}_n(\mathcal{F}; \sqrt{t\delta_n})}{\sqrt{t\delta_n}} \leq \sigma\frac{\tilde{\mathcal{G}}_n(\mathcal{F}; \delta_n)}{\delta_n} \leq \delta_n$  and setting  $s = t\delta_n^2$ , we obtain

$$\begin{aligned} & \mathbb{P}\left(\sup_{\|\hat{\Delta}\|_n \leq \sqrt{t\delta_n}} \frac{\sigma}{n} \sum_{i=1}^n w_i \hat{\Delta}(x_i) \geq 2t\delta_n^2\right) \\ & \leq \mathbb{P}\left(\sup_{\|\hat{\Delta}\|_n \leq \sqrt{t\delta_n}} \frac{\sigma}{n} \sum_{i=1}^n w_i \hat{\Delta}(x_i) \geq \sigma\tilde{\mathcal{G}}_n(\mathcal{F}; \sqrt{t\delta_n}) + t\delta_n^2\right) \leq e^{-\frac{nt\delta_n^2}{2\sigma^2}} \square \end{aligned}$$

13 / 17

## Application 1: $\ell_0$ -constrained sparse linear regression

Going back to finding the best sparse linear fit

$$\hat{f} = \arg \min_{f \in \mathcal{F}_{lin,s}} \|y - X\theta\|_n^2$$

with  $\mathcal{F}_{lin,s} = \{f(\cdot) = \langle \theta, x \rangle : \|\theta\|_0 \leq s\}$  Even though  $\theta$  and the square loss are unbounded, we can *now* use get the prediction bound

- In HW 2 we prove  $\tilde{\mathcal{G}}_n(\mathcal{F}_{lin,s}; \delta) \leq O\left(\delta\sqrt{\frac{s \log(ed/s)}{n}}\right)$  when  $\lambda_{\max}\left(\frac{X_S^\top X_S}{n}\right)$  bounded for all subsets  $S$  of size  $s$
- Hence the critical radius has to satisfy  $\frac{\tilde{\mathcal{G}}_n(\mathcal{F}_{lin,s}; \delta)}{\delta} = \sqrt{\frac{s \log(ed/s)}{n}} \leq \frac{\delta}{\sigma}$
- Thus using the theorem, plugging in  $\delta_n^2$  that achieves equality, we can obtain with probability at least  $1 - \delta$

$$\|\hat{f} - f^*\|_n^2 \leq O\left(\frac{s \log(ed/s) \log 1/\delta}{n}\right)$$

Also see MW Example 13.16.

14 / 17

## General functions via Dudley's integral

Corollary (Dudley's integral & critical quantity - MW Cor. 13.7.)

If  $\mathcal{F}^*$  is star-shaped, any  $\delta \in [0, \sigma]$  such that

$$\frac{16}{\sqrt{n}} \int_{\frac{\delta^2}{4\sigma}}^{\delta} \sqrt{\log \mathcal{N}(t; \mathcal{F}^*(x_1^n) \cap \mathbb{B}_n(\delta), \|\cdot\|_n)} dt \leq \frac{\delta^2}{4\sigma}$$

satisfies the critical inequality.

Proof via chaining for localized Gaussian complexity for a  $\frac{\delta^2}{4\sigma}$  cover

$$\tilde{\mathcal{G}}_n(\mathcal{F}^*; \delta) \leq \frac{16}{\sqrt{n}} \int_{\frac{\delta^2}{4\sigma}}^{\delta} \sqrt{\log \mathcal{N}(t; \mathcal{F}^*(x_1^n) \cap \mathbb{B}_n(\delta), \|\cdot\|_n)} dt + \frac{\delta^2}{4\sigma}$$

(skipped in class). Note that any  $\delta$  satisfying

$\frac{16}{\sqrt{n}} \int_0^{\delta} \sqrt{\log \mathcal{N}(t; \mathcal{F}^*(x_1^n) \cap \mathbb{B}_n(\delta), \|\cdot\|_n)} dt \leq \frac{\delta^2}{4\sigma}$  also satisfies the inequality in the corollary (often use this easier integral)

The following examples correspond to MW Ex 13.10. and 13.11.

15 / 17

## Application 2: General functions via Dudley's integral

1.  $\mathcal{F}_L$ : Lipschitz functions on  $[0, 1]$  and  $f(0) = 0$  has  $\log \mathcal{N}(\epsilon) \leq O(\frac{L}{\epsilon})$

$$\frac{1}{\sqrt{n}} \int_0^{\delta} \sqrt{\log \mathcal{N}(t; \mathcal{F}_L(x_1^n), \|\cdot\|_n)} dt \leq \frac{1}{\sqrt{n}} \int_0^{\delta} \left(\frac{L}{t}\right)^{\frac{1}{2}} dt \leq \sqrt{\frac{L\delta}{n}} \stackrel{(!)}{\leq} \frac{\delta^2}{4\sigma^2}$$

→ Rearranging terms yields  $\|\hat{f} - f^*\|_n^2 \leq \delta_n(\mathcal{F}_L)^2 = O(\frac{L\sigma^2}{n})^{\frac{2}{3}}$

Recall how for Lipschitz functions, the “unlocalized” Dudley bound for the R.C. from last lecture plugged into eq. 2 would yield

$\mathbb{E}\|\hat{f} - f^*\|_n^2 \leq O(\frac{1}{n^{1/2}})$  → can't expect faster decay for  $\|\hat{f} - f^*\|_n^2$  this “unlocalized” route!

2.  $\mathcal{F}_{1,c}$ :  $f \in \mathcal{F}_1$  **and** convex, has  $\log \mathcal{N}(\epsilon) \leq O((\frac{1}{\epsilon})^{\frac{1}{2}})$  (no proof)

$$\frac{1}{\sqrt{n}} \int_0^{\delta} \sqrt{\log \mathcal{N}(t; \mathcal{F}_{1,c}(x_1^n), \|\cdot\|_n)} dt \leq \frac{1}{\sqrt{n}} \int_0^{\delta} \left(\frac{1}{t}\right)^{\frac{1}{4}} dt \leq \frac{\delta^{3/4}}{\sqrt{n}} \stackrel{(!)}{\leq} \frac{\delta^2}{4\sigma^2}$$

→ Rearranging terms yields  $\delta_n(\mathcal{F}_{1,c})^2 = O((\frac{\sigma^2}{n})^{\frac{4}{5}})$

16 / 17

# References

Non-parametric regression:

- MW Chapter 13