

Lecture 11: Prediction error bounds for random design

1 / 22

Announcements

- HW 2 released Tuesday, due 11.11. 23:59
- For oral exam: You should understand and be able to explain all proofs in the slides (even those we skipped in class), except the ones that come after the references

Plan for today

Prediction error bounds for random design (non)parametric regression

- for uniformly bounded errors
 - Naive approach using uniform law
 - using localization
- for certain function classes where moment conditions hold

2 / 22

Recap: Fixed design bound

Definition (Localized (empirical) Gaussian complexity)

The localized Gaussian complexity around f^* of scale δ is

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

Definition (Critical radius)

For \mathcal{F}^* star-shaped, we call $\delta_n > 0$ the *critical quantity/radius* which is the smallest solution $\delta > 0$ satisfying

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

This exists due to MW Lemma 13.6.

Theorem (Prediction error bound, MW Thm 13.5.)

If \mathcal{F}^* is star-shaped, we have for the square loss minimizer \hat{f} for any $t \geq 1$

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

3 / 22

Random design

- So far, we only controlled $\|\hat{f} - f^*\|_n^2$ w.h.p. over observation noise w

$$\begin{aligned} \|\hat{f} - f^*\|_n^2 &= R(\hat{f}) - R(f^*) = \mathbb{E}_w \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2 - \mathbb{E}_w \frac{1}{n} \sum_{i=1}^n w_i^2 \\ &= \frac{1}{n} \sum_{i=1}^n (\hat{f}(x_i) - f^*(x_i))^2 \end{aligned}$$

- can be bounded using empirical Gaussian complexities via basic inequality \rightarrow basic inequality

How does the error look like on the whole domain \mathcal{X} ?

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

$$\begin{aligned} \|f - f^*\|_2^2 &= R(f) - R(f^*) = \mathbb{E}_{X,W} (Y - f(X))^2 - \mathbb{E} W^2 \\ &= \mathbb{E}_X (f(X) - f^*(X))^2 = \mathbb{E}_{x_1, \dots, x_n} \|f - f^*\|_n^2 \end{aligned}$$

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

4 / 22

For uniformly bounded functions directly use uniform law?

Maybe use $\|\hat{f} - f^*\|_2^2 - \|\hat{f} - f^*\|_n^2 \leq \sup_{f \in \mathcal{F}} \|f - f^*\|_2^2 - \|f - f^*\|_n^2$
and then plug in previous bound on $\|\hat{f} - f^*\|_n^2$?

Definition (Rademacher complexity - recap)

Given a function class \mathcal{H} and distribution \mathbb{P} on its domain \mathcal{Z} , we define the Rademacher complexity as

$$\mathcal{R}_n(\mathcal{H}) = \mathbb{E}_{\epsilon, z} \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \epsilon_i h(z_i)$$

Theorem (Uniform law - recap)

For b -unif. bounded \mathcal{H} with $\mathcal{R}_n(\mathcal{H}) = \mathbb{E} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \epsilon_i h(z_i)$

$$\mathbb{P}\left(\sup_{h \in \mathcal{H}} \mathbb{E} h - \frac{1}{n} \sum_{i=1}^n h(z_i) \geq 2\mathcal{R}_n(\mathcal{H}) + t\right) \leq e^{-\frac{nt^2}{2b^2}}$$

w/ prob. over the training data.

Partner-Q: If you use the uniform law for e.g. linear functions $f(x) = \langle w, x \rangle$ with $\|x\|_2 \leq D$, $\|w\|_2 \leq B$ and bounded noise, what's the best-case generalization gap upper bound you could get for

5 / 22

Solution:

It suffices to bound $\mathbb{E}_X (Y - \hat{f}(X))^2 = \|\hat{f} - f^*\|_2^2 + \sigma^2$ using a uniform law on the generalization error with the square loss

$$R(f) - R_n(f) := \mathbb{E}_X (Y - \hat{f}(X))^2 - \|y - \hat{f}(x_1^n)\|_2^2$$

First of all, in this setting, by assumption, the loss is uniformly bounded since $|y_i - f(x_i)| \leq D'$ is bounded by some constant D' .

- Define $\tilde{\mathcal{F}}(z_1^n) = \{(y_1 - f(x_1), \dots, y_n - f(x_n)) : f \in \mathcal{F}\} \subset \mathbb{R}^n$
- Then for the square function $\ell_{sq}(u) = u^2$ for $|u| \leq D'$ we have $|\ell_{sq}(u) - \ell_{sq}(u')| \leq |u^2 - u'^2| \leq |u - u'| |u + u'| \leq 2D' |u - u'|$, i.e. ℓ_{sq} is $2D'$ -Lipschitz
- Then, analogous to the SVM example, we have $\mathcal{H}(z_1^n) = \ell_{sq} \circ \tilde{\mathcal{F}}(z_1^n)$ and $\tilde{\mathcal{R}}_n(\mathcal{H}(z_1^n)) \leq 2D' \tilde{\mathcal{R}}_n(\tilde{\mathcal{F}}(z_1^n)) = 2D' \tilde{\mathcal{R}}_n(\mathcal{F}(z_1^n))$ using Rademacher contraction, where \mathcal{F} is the space of bounded linear functions
- Analogously to the SVM excess risk bound, the uniform law yields a squared error bound of order $O(1/\sqrt{n}) \rightarrow$ highly suboptimal!

6 / 22

Motivating the localized uniform law

(Note: in the sequel, we sometimes write g for $f - f^*$ instead of $\hat{\Delta}$)

Can we do better? \rightarrow Indeed, can again use localization! We'll discuss two approaches: (i) for uniformly bounded \mathcal{F}^* (ii) when $f \in \mathcal{F}^*$ satisfy certain moment condition. For both, we need

Definition (localized *population* Rademacher complexity)

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

For case (ii) we further have

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

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

7 / 22

Population vs. empirical critical radius.

The following lemma connects the empirical and pop. critical radii

Lemma (critical radii, proposition MW 14.25)

Let \mathcal{F}^* be star-shaped and b -uniformly bounded. Then for some universal constants $c < 1 < C$, it holds with probability at least $1 - c_0 e^{-c_1 n \frac{\bar{\delta}_n^2}{b}}$ that $c\bar{\delta}_n \leq \delta_n \leq C\bar{\delta}_n$.

- A proof of this lemma for $b = 1$ can be found in MW Section 14.5.
- Given the lemma, we will use our earlier calculations for δ_n to bound $\bar{\delta}_n$ using the same order and vice versa.
- Now we're ready to state the localized uniform law

Theorem (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

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

8 / 22

Precise statement of localized uniform law

- Note that the condition is not too strong: if $\bar{\delta}_n \asymp 1/n$, i.e. we have the best possible achievable rate, then the inequality is still true for small enough c (only slightly depending on n), since $\log \log n$ is “almost constant”. For $\bar{\delta}_n \geq \omega(1/n)$, this condition always holds for large enough n .

Example I:

- Let's go back to the question in the beginning of today's lecture. The task was to bound $\|\hat{f} - f^*\|_2^2$ for linear functions $\mathcal{F}_{lin} = \{f : f(x) = \langle \theta, x \rangle \text{ for } \|\theta\|_2 \leq B\}$ in terms of $\|\hat{f} - f^*\|_n^2$
- We assume that the covariates/inputs are on a bounded domain $\|x\|_2 \leq D$ and the distribution is such that $\mathbb{E}x = 0$ and $\mathbb{E}xx^T = \Sigma$ with invertible Σ . We also define a whitened random vector $w = \Sigma^{-1/2}x$ for which $\langle x, \theta \rangle = \langle w, \sqrt{\Sigma}\theta \rangle$ for any θ .
- Further for all $\Delta \in \mathcal{F}_{lin}^*$ there exists $\|\theta\|_2 \leq B$ with $\Delta(x) = \langle \Delta_\theta, x \rangle$ where $\Delta_\theta = \theta - \theta^* \in \mathbb{R}^d$ and $\|\Delta_\theta\|_2 \leq 2B$.

9 / 22

Example I: Linear regression on bounded domain

- Note that on the bounded domain, functions in \mathcal{F}_{lin} are uniformly bounded by $b = BD$ and further for any $\Delta \in \mathcal{F}_{lin}^*$ we have

$$\|\Delta\|_2^2 = \Delta_\theta^T \mathbb{E}xx^T \Delta_\theta = \|\sqrt{\Sigma}\Delta_\theta\|_2^2$$

with $\mathcal{L}^2(\mathbb{P})$ norm on the LHS and Euclidean norm on the RHS

- Note that by the fact that $\bar{\delta}_n \approx \delta_n$ w.h.p., using the fixed design prediction error bound we get that $\|\hat{f} - f^*\|_n^2 \leq O(\bar{\delta}_n)$
- Now we compute the population critical radius. First note that we can rewrite the localized population Rademacher complexity

$$\begin{aligned} \mathcal{R}_n(\mathcal{F}_{lin}^*; \delta) &= \mathbb{E} \sup_{\substack{\Delta \in \mathcal{F}_{lin}^* \\ \|\Delta\|_2 \leq \delta}} \frac{1}{n} \sum_{i=1}^n \epsilon_i \Delta(x_i) = \mathbb{E} \sup_{\substack{\|\sqrt{\Sigma}\Delta_\theta\|_2 \leq \delta \\ \|\Delta_\theta\|_2 \leq 2B}} \left\langle \frac{1}{n} \sum_{i=1}^n \epsilon_i x_i, \Delta_\theta \right\rangle \\ &= \mathbb{E} \sup_{\substack{\|\sqrt{\Sigma}\Delta_\theta\|_2 \leq \delta \\ \|\Delta_\theta\|_2 \leq 2B}} \left\langle \frac{1}{n} \sum_{i=1}^n \epsilon_i w_i, \sqrt{\Sigma}\Delta_\theta \right\rangle \leq \delta \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^n \epsilon_i w_i \right\|_2 \end{aligned}$$

where in the last inequality by removing one of the constraints, we sup over a bigger set (crude but enough to make the point)

10 / 22

Example I: Linear regression on bounded domain (ctd)

- Then by Jensen's inequality and cyclicity of the trace, we obtain $\mathbb{E} \|\frac{1}{n} \sum_{i=1}^n \epsilon_i w_i\|_2 \leq \frac{1}{n} \sqrt{\mathbb{E} \sum_{i=1}^n \|w_i\|_2^2} = \frac{1}{n} \sqrt{n \text{trace}(\mathbb{E} w w^\top)} = \sqrt{\frac{d}{n}}$, where we use the identity covariance of w . Hence $\bar{\delta}_n = cBD\sqrt{\frac{d}{n}}$
- Using the Theorem, we then get $\|\hat{f} - f^*\|_2^2 \leq 2\|\hat{f} - f^*\|_n^2 + 2c^2\bar{\delta}_n^2 \leq \tilde{c}\bar{\delta}_n^2 \in O(\frac{d}{n})$ for some universal constant \tilde{c} (without trying to be tight).
- Compared to what we can get with the uniform law, in terms of n , this is much faster than $O(\frac{1}{\sqrt{n}})$

11 / 22

Example II: Sobolev spaces MW Ex. 14.6.

Sobolev space (1-smooth functions as appearing in HW 2)

$$\mathcal{F}_{sob} := \{f : [0, 1] \rightarrow \mathbb{R} : f(0) = 0, f' \in \mathcal{L}^2([0, 1])\}$$

- In MW Example 2.16 we saw how it corresponds to an RKHS with bounded kernel function $\mathcal{K}(x, x') = \min\{x, x'\}$. functions in \mathcal{F}_{sob} with RKHS norm at most 1 are uniformly bounded as well, since $\|f\|_\infty = \sup_{\|f\|_{\mathcal{F}} \leq 1} \langle f, \mathcal{K}(\cdot, x) \rangle = \|\mathcal{K}(\cdot, x)\|_{\mathcal{F}} = \sqrt{\mathcal{K}(x, x)}$.
- In last lecture, we noted how the polynomial eigenvalue decay of the kernel matrix for these spaces leads to a critical radius of $\delta_n = c \left(\frac{\sigma^2}{n}\right)^{1/3}$ for some constant c
- Hence, $\|\Delta\|_n \leq \delta_n$ and by Lemma we have $\bar{\delta}_n$ of the same order
- Therefore, using the Theorem, we then finally obtain that $\|\hat{f} - f^*\|_2^2 \leq O\left(\left(\frac{\sigma^2}{n}\right)^{2/3}\right)$

12 / 22

Proof idea for localized uniform law

Recall in the proof for empirical prediction error:

- For localization we used the basic inequality for the empirical error
- There we had LHS $\|g\|_n^2$ with $g \in \mathcal{F}^*$ which we self-bounded by $\delta_n \|g\|_n$ when $\|g\|_n > \delta_n$
- We can do something similar here: we choose $\|g\|_2^2 - \|g\|_n^2$ as our RHS and will also “self-upper-bound” it
- Observe that the binomial formula yields for any $g \in \mathcal{F}^*$

$$\|g\|_2 - \|g\|_n = \frac{\|g\|_2^2 - \|g\|_n^2}{\|g\|_2 + \|g\|_n}$$

- Hence the proof goes through by studying the two cases
 - a) if $\|g\|_2 \leq \bar{\delta}_n$, then $\frac{\|g\|_2^2 - \|g\|_n^2}{\|g\|_2 + \|g\|_n} \leq \bar{\delta}_n$
 - b) or if $\|g\|_2 \geq \bar{\delta}_n$ (uniformly for all $g \in \mathcal{F}^*$), then $\sup_{g \in \mathcal{F}^*, \|g\|_2 \geq \bar{\delta}_n} \|g\|_2^2 - \|g\|_n^2 \leq \|g\|_2 \bar{\delta}_n$ w.h.p.

We give intuition for the proof of b)

13 / 22

Proof of b): case $\|g\|_2 \geq \bar{\delta}_n$ (next class)

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

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

- symmetrization and Rademacher contraction for $r \geq \bar{\delta}_n$

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

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

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

14 / 22

Proof of b): case $\|g\|_2 \geq \bar{\delta}_n$ (next class)

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

- Use peeling argument like before and split

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

$S_m = \{\|g\|_2 \in [t_{m-1}, t_m]\}$ where $t_m = 2^m \bar{\delta}_n$. In particular, by uniform boundedness $\|g\|_2 \leq 1$, we have that $S \subset \bigcup_{m=1}^M \{S \cap S_m\}$ with $M = 4 \log(1/\bar{\delta}_n)$

- using $\sup_{g \in \mathcal{F}^*, \|g\|_2 \leq r} \|g\|_2^2 - \|g\|_n^2 \leq \frac{r \bar{\delta}_n}{2}$ with $r = t_m$ and using union bound gives

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

15 / 22

Going beyond uniform boundedness

So far, need boundedness assumptions \rightarrow too restrictive! In particular for linear regression, bounded noise and domain etc. seem excessive! For regression specifically, can we get rid of it?

Yes, using a specific moment condition instead

Assumption (Moment condition)

For all $f \in \mathcal{F}^*$ with $\|f\|_2 \leq 1$ we have $\mathbb{E}f^4(X) \leq C^2 \mathbb{E}f^2(X)$

We also need to define a population critical radius

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

Define $\bar{\delta}_n$ to be the smallest $\delta > 0$ satisfying the population critical inequality

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

16 / 22

Prediction error bound for random design

Further recall that δ_n is the critical radius defined via localized Gaussian complexity, i.e. smallest $\delta > 0$ satisfying $\frac{\tilde{\mathcal{G}}_n(\mathcal{F}; \delta)}{\delta} \leq \frac{\delta}{\sigma}$.

Theorem (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

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

Proof sketch of Theorem:

Write $\hat{\Delta} = \hat{f} - f^*$. 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 $\|\hat{\Delta}\|_2 \geq \bar{\delta}_n$ we have

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

17 / 22

Proof sketch of Theorem (ctd)

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 $\|\hat{\Delta}\|_2^2$
- $\bar{\delta}_n < \delta_n$: 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 $c > 1$

18 / 22

Proof of Lemma (skipped in class)

Truncate $\hat{\Delta}$ at τ 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 τ 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 □

19 / 22

Example 3: (Unbounded) Linear regression

We can now apply this bound to obtain error bounds for linear regression **without the bounded domain and parameter assumption.**

i.e. $\mathcal{F}_{lin} = \{f : f(x) = \langle \theta, x \rangle, \theta \in \mathbb{R}^d\}$

- The moment e.g. holds with Gaussian covariates (unbounded!)
 - Then, for $\Delta \in \mathcal{F}^*$, we can write $\Delta(x) = (f - f^*)(x) = \langle x, \theta - \theta^* \rangle$ is Gaussian for all $\theta \in \mathbb{R}^d$
 - Because $\Delta(x)$ is a Gaussian random variable for all $\Delta \in \mathcal{F}^*$, we have $\mathbb{E}(\Delta^4(x)) = 3[\mathbb{E}(\Delta^2(x))]^2$ by definition of the Gaussian distribution. Then we automatically have for all $\Delta \in \mathcal{F}^*$ for which $\|\Delta\|_2 \leq 1$, that moment condition is satisfied with $C^2 = 3$.
- The moment condition also holds when x has independent entries with bounded 2nd & 4th moment (see MW Exercise 14.6)
- Using the derivations in Example 1, we again have $\bar{\delta}_n \approx \delta_n = O(\frac{d}{n})$ so that using the fixed design and the last theorem we have $\|\hat{f} - f^*\|_2^2 \leq O(\frac{d}{n})$

Obviously, you could also use closed-form solutions and random matrix theory to analyze least-squares regression instead of empirical process theory. See more examples in the MW book

20 / 22

Caveats of random design bounds

Generally rely on additional assumptions

- above (MW Thm 14.12, Cor. 14.15): moment conditions
- uniformly bounded error (MW Thm 14.1.)

Beyond regression

- MW Theorem 14.20: One can also obtain excess risk bounds for random design for losses $\ell(z, y)$ with $z = f(x)$ that are
 - L -Lipschitz in the first argument
 - γ -strongly convex losses with respect to $\mathcal{L}^2(\mathbb{P})$ norm

again using functional Hoeffding + contraction

21 / 22

References

Localized uniform laws and Random design non-parametric regression

- MW Chapter 14

22 / 22