

Lecture 3: Azuma-Hoeffding and uniform law

1 / 18

Announcements:

- HW released tonight, due in two weeks on Thursday **8.10.22 23:59** on gradescope.
- Warning: HW is *long*, start early!
- Can discuss together, but write up your *own* solution and indicate who you've worked together with
- no late HW except in medical cases (with attest from doctor)
- Post questions on HW on moodle
- Please de-register once you know you are not going to continue the course!

2 / 18

Plan today

- Review of proof of uniform tail bound
- Proof of Azuma-Hoeffding (tailored for McDiarmid)
- Warm-up exercise: using Azuma-Hoeffding for online learning “excess risk”
- Uniform law with Rademacher complexity
- Intuition of Rademacher complexity

3 / 18

Recap: Main tail bound

- $\{Z_i\}_{i=1}^n$ are training points $\stackrel{iid}{\sim} \mathbb{P}$, estimator $\hat{f}_n \in \mathcal{F}$ trained on them
- We use Z both for the collection $Z = \{Z_i\}_{i=1}^n$ and a single random vector $Z \sim \mathbb{P}$ which should be clear from the context
- Goal: want to prove that $R(\hat{f}_n) - R_n(\hat{f}_n) \leq \sup_{f \in \mathcal{F}} \mathbb{E} \ell(Z; f) - \frac{1}{n} \sum_{i=1}^n \ell(Z_i; f) =: g_n(Z)$ small with probability at least $1 - \delta$

Theorem (Uniform tail bound)

For b -unif. bounded ℓ , it holds that

$$\mathbb{P}(\sup_{f \in \mathcal{F}} R(f) - R_n(f) \geq \mathbb{E}[\sup_{f \in \mathcal{F}} R(f) - R_n(f)] + t) \leq e^{-\frac{nt^2}{2b^2}}$$

where the probability is over the training data.

4 / 18

Recap: Proof of tail bound

Approach: Upper bound $\mathbb{P}(g_n(Z) - \mathbb{E}g_n(Z) \geq t)$ by following

1. If loss ℓ b -uniformly bounded, then $g_n = \sup_{f \in \mathcal{F}} \mathbb{E}\ell(Z, f) - \frac{1}{n} \sum_{i=1}^n \ell(Z_i, f)$ satisfies bounded difference property with $\sigma_i = \frac{2b}{n}$ for all i
2. For any g_n , we can decompose $g_n(Z) - \mathbb{E}g_n(Z) = \sum_{i=1}^n D_i$
 $D_i = \mathbb{E}[g_n(Z) | Z_1, \dots, Z_i] - \mathbb{E}[g_n(Z) | Z_1, \dots, Z_{i-1}]$
3. Then, D_i satisfies that for any z_1^{i-1} there are some a_i, b_i with $b_i - a_i \leq \sigma_i$ such that $D_i | Z_1^{i-1} = z_1^{i-1} \in [a_i, b_i]$.
4. show how for such D_i (bounded martingale diff sequence) we have $\sum_{i=1}^n D_i$ concentrates around its expectation $\mathbb{E}D_i = 0$, i.e.
$$\mathbb{P}(\sum_{i=1}^n D_i > t) \leq e^{-\frac{2t^2}{\sum_{i=1}^n \sigma_i^2}} \leq e^{-\frac{nt^2}{2b^2}} \text{ [Azuma Hoeffding]}$$

Note: 2-4 proves McDiarmid using Azuma-Hoeffding, 2-3 prove that assumptions for Azuma-Hoeffding hold.

Not shown, will show today: Azuma-Hoeffding

5 / 18

Recap: Azuma-Hoeffding vs. Hoeffding

- Hoeffding: Simple concentration for average of n independent sub-Gaussian (e.g bounded) Z_i

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n Z_i - \mathbb{E}Z > t\right) \leq e^{-\frac{nt^2}{2\sigma^2}}$$

- Azuma-Hoeffding: "Advanced" concentration for average of a martingale difference sequence $\{D_i\}_{i=1}^n$ bounded in intervals of length $\sigma = \frac{c}{n}$

$$\mathbb{P}\left(\sum_{i=1}^n D_i > t\right) \leq e^{-\frac{2t^2}{n\sigma^2}} = e^{-\frac{2nt^2}{c^2}}$$

6 / 18

Recap: Formal Azuma-Hoeffding statement

Assumption (Bounded martingale difference)

Let $\{D_j\}_{j=1}^{\infty}$ and $\{Z_j\}_{j=1}^{\infty}$ be two sequences of R.V. where for all j :

- D_j is measurable wrt the induced sigma algebra $\sigma(Z_1, \dots, Z_j)$
- $\mathbb{E}[D_j | Z_1, \dots, Z_{j-1}] = 0$ and $\mathbb{E}|D_j| < \infty$
- $D_j | Z_1, \dots, Z_{j-1}$ almost surely lies within an interval of length L_j

Theorem (Azuma-Hoeffding inequality, MW Cor 2.20)

If the sequences $\{D_j\}_{j=1}^{\infty}$ and $\{Z_j\}_{j=1}^{\infty}$ satisfy above assumptions, then

$$\mathbb{P}\left(\sum_{i=1}^n D_i \geq t\right) \leq e^{-\frac{2t^2}{\sum_{i=1}^n L_i^2}}$$

In the proof of Hoeffding, the D_i were independent and we used $\mathbb{E}e^{\lambda \sum_{i=1}^n D_i} = \prod_{i=1}^n \mathbb{E}e^{\lambda D_i}$. Here they're not, but we can similarly get a product - of the upper bounds using conditioning!

7 / 18

Proof of Azuma-Hoeffding

1. First of all, we have for all sequences z_1^{i-1} that for some $b_i - a_i \leq L_i$

$$\mathbb{E}[e^{\lambda D_i} | Z_1^{i-1} = z_1^{i-1}] \leq e^{\lambda^2 (b_i - a_i)^2 / 8} \leq e^{\lambda^2 L_i^2 / 8}$$

by the fact that R.V. bounded in an interval of length L_i are $L_i/2$ subgaussian (for the right constant check MW Exercise 2.4., for an easier proof for the wrong constant check MW Example 2.4.)

2. However, since D_i are measurable wrt $\sigma(Z_1, \dots, Z_i)$, we have by the "pull-out property" of conditional expectations, that for any

$$\begin{aligned} \mathbb{E}[v(D_i) | Z_1^i] &= v(D_i) \text{ and} \\ \mathbb{E}[v(D_i)u(Z_1^{i+1}) | Z_1^i] &= v(D_i)\mathbb{E}[u(Z_1^{i+1}) | Z_1^i] \end{aligned}$$

3. Now using the tower property (TP) of conditional expectations iteratively, we see that $\sum_{i=1}^n D_i$ is $\sqrt{\sum_{i=1}^n \frac{L_i^2}{4}}$ -subgaussian:

$$\mathbb{E}e^{\lambda \sum_{i=1}^n D_i} \stackrel{(TP)}{=} \mathbb{E}[\mathbb{E}[e^{\lambda \sum_{i=1}^{n-1} D_i} e^{\lambda D_n} | Z_1, \dots, Z_{n-1}]]$$

$$\stackrel{(3.)}{=} \mathbb{E}[e^{\lambda \sum_{i=1}^{n-1} D_i} \mathbb{E}[e^{\lambda D_n} | Z_1, \dots, Z_{n-1}]] \leq e^{\lambda^2 L_n^2 / 8} \mathbb{E}[e^{\lambda \sum_{i=1}^{n-1} D_i}] = e^{\lambda^2 \sum_{i=1}^n L_i^2 / 8}$$

8 / 18

Exercise Context I: Online learning setting

We now gain some more intuition for Azuma-Hoeffding by applying it to a different problem related to online learning

- Z_1, \dots, Z_n come in one at a time.
- At each point in time i you would like to output an estimator \hat{f}_{i-1} to predict on the next sample Z_i with small loss
- As a data scientist, we naturally consider functions that are trained using the previous examples Z_1, \dots, Z_{i-1} . More formally, we assume \hat{f}_{i-1} is a *deterministic function* of the previous samples Z_1, \dots, Z_{i-1} (e.g. ERM but *does not have to be!*), i.e. measurable with respect to sigma algebra $\sigma(Z_1, \dots, Z_{i-1})$.
- \hat{f}_0 can be any data-independent arbitrary estimator, e.g. a randomly initialized model.
- Assume the minimizer $\hat{f}_n := \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \ell(Z_i; f)$ exists

9 / 18

Exercise Context II: Online to batch conversion

- A standard quantity people want to keep small in online learning is the regret Reg_n , the average incurred loss of the sequence $\{\hat{f}_i\}_{i=1}^n$ with the loss of \hat{f}_n

$$\text{Reg}_n = \sum_{i=1}^n \ell(Z_i; \hat{f}_{i-1}) - \sum_{i=1}^n \ell(Z_i; \hat{f}_n)$$

- Note: Bounding the actual Reg_n is a whole area of research and in many cases, good online learning algorithms exist
- Online-to-batch conversion exploits online learning algorithms with small regret to get estimator based on batch Z_1, \dots, Z_n with good generalization. For example, one can consider a random estimator that samples from the sequence of online estimators $\{\hat{f}_i\}_{i=0}^{n-1}$ which
 - conditioned on the data are deterministic
 - has an average (over the sampling) risk of $\frac{1}{n} \sum_{i=1}^n R(\hat{f}_{i-1})$
- We will now prove a high probability bound on the “average” excess risk $\frac{1}{n} \sum_{i=1}^n R(\hat{f}_{i-1}) - R(f^*)$

10 / 18

Exercise: Bound on the average excess risk

With your neighbor, you'll prove that if Z_1, \dots, Z_n are i.i.d., with probability at least $1 - \delta$,

$$\frac{1}{n} \sum_{i=1}^n [R(\hat{f}_{i-1}) - R(f^*)] \leq \frac{1}{n} \text{Reg}_n + \sqrt{\frac{8 \log(1/\delta)}{n}} \quad (1)$$

with $R(f) = \mathbb{E}_Z \ell(Z; f)$ for $\ell \in [0, 1]$.

First interpret it: If the regret is of order \sqrt{n} (e.g. convex losses) or $\log n$ (e.g. strongly convex losses), we can get a $O(\frac{1}{\sqrt{n}})$ bound

Define $D_i = [\mathbb{E}_Z \ell(Z; \hat{f}_{i-1}) - \ell(Z_i; \hat{f}_{i-1})] + [\ell(Z_i; f^*) - \mathbb{E}_Z \ell(Z; f^*)]$

1. Step: Prove that $\{D_i\}_{i=1}^n$ satisfies the properties of a bounded martingale difference sequence and that by using optimality of \hat{f} ,

$$\frac{1}{n} \sum_{i=1}^n [R(\hat{f}_{i-1}) - R(f^*)] \leq \frac{1}{n} \text{Reg}_n + \frac{1}{n} \sum_{i=1}^n D_i$$

2. Step: Use Step 1 and Azuma-Hoeffding to prove the bound eq. 1

11 / 18

Solution: Proof of average excess risk bound

We use the following shorthands for simplicity:

- $R_n(\{\hat{f}_i\}_{i=0}^{n-1}) := \frac{1}{n} \sum_{i=1}^n \ell(Z_i; \hat{f}_{i-1})$
- $R(\{\hat{f}_i\}_{i=0}^{n-1}) := \frac{1}{n} \sum_{i=1}^n R(\hat{f}_{i-1}) = \frac{1}{n} \sum_{i=1}^n \mathbb{E} \ell(Z; \hat{f}_{i-1})$

1. Risk decomposition:

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n [R(\hat{f}_{i-1}) - R(f^*)] &\leq R(\{\hat{f}_i\}_{i=0}^{n-1}) - R_n(\{\hat{f}_i\}_{i=0}^{n-1}) + \underbrace{R_n(\{\hat{f}_i\}_{i=0}^{n-1}) - R_n(\hat{f}_n)}_{=\text{Reg}_n} \\ &\quad + \underbrace{R_n(\hat{f}_n) - R_n(f^*)}_{\leq 0 \text{ by optimality of } \hat{f}} + R_n(f^*) - R(f^*) \end{aligned}$$

2. D_i is a martingale difference sequence because

$$\mathbb{E}[D_i | Z_1^{i-1}] = 0$$

almost surely as Z_i is independent of \hat{f}_{i-1} and bounded a.s. by 4.

12 / 18

Using the tail bound requires bounding expectation

Define $\text{Res}(n, \mathcal{F}) := \mathbb{E} \sup_{f \in \mathcal{F}} \mathbb{E} \ell(Z, f) - \frac{1}{n} \sum_{i=1}^n \ell(Z_i, f)$.

The tail bound then says

$$\mathbb{P}(\sup_{f \in \mathcal{F}} \mathbb{E} \ell(Z, f) - \frac{1}{n} \sum_{i=1}^n \ell(Z_i, f) \geq \text{Res}(n, \mathcal{F}) + t) \leq e^{-\frac{nt^2}{2b^2}} \quad (2)$$

The next four sessions will be about how to bound $\text{Res}(n, \mathcal{F})$!

- Step I (today and next time): we obtain the *uniform law* by showing that $\text{Res}(n, \mathcal{F})$ is bounded by the Rademacher complexity
- Step II (next 2 weeks): We'll discuss how to bound the Rademacher complexity as a function of n, \mathcal{F}

13 / 18

Rademacher complexity and the uniform law

- Let ϵ_i be i.i.d. Rademacher R.V. (i.e. $+1, -1$ w.p. $\frac{1}{2}$)
- $Z = \{Z_i\}_{i=1}^n$ are training points $\stackrel{iid}{\sim} \mathbb{P}$

Definition (Rademacher complexity)

Given a function class \mathcal{H} and distribution \mathbb{P} on its domain \mathcal{Z} , we define the Rademacher complexity of \mathcal{H} wrt \mathbb{P} for sample size n 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 for the risk, MW Thm 4.10.)

For b -unif. bounded \mathcal{H} , with prob. over training data,

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

14 / 18

Rademacher complexity

- By using $\mathcal{H} = \{h : h(\cdot) = \ell(\cdot; f) \quad \forall f \in \mathcal{F}\}$ we get

$$\mathcal{R}_n(\mathcal{H}) = \mathbb{E}_{\epsilon, Z} \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \epsilon_i h(Z_i) = \mathbb{E}_{\epsilon, Z} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \epsilon_i \ell(Z_i, f)$$

and after showing $\text{Res}(n, \mathcal{F}) \leq 2\mathcal{R}_n(\mathcal{H})$, directly obtain our desired bound on $\sup_{f \in \mathcal{F}} R(f) - R_n(f)$

- Note if $\mathcal{R}_n(\mathcal{H}) = o(1)$, then $\sup_{f \in \mathcal{F}} R(f) - R_n(f) \xrightarrow{\text{a.s.}} 0$.
- Before the proof, we aim to gain some intuition for the quantity $\mathcal{R}_n(\mathcal{H})$ and how it may behave with different n and \mathcal{H}

15 / 18

Intuition for Rademacher complexity

Consider binary classification setting $\ell(z_i; f) = \mathbb{1}(f(x_i)y_i < 0)$.

With you neighbor discuss: How does the empirical Rademacher complexity

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

with $\mathcal{H} = \{h : h(\cdot) = \ell(\cdot; f) \quad \forall f \in \mathcal{F}\}$

depend on the factors \mathcal{F}, ℓ, n to control excess risk?

Some answers

- If \mathcal{F} larger $\rightarrow \mathcal{H}$ larger $\rightarrow \tilde{\mathcal{R}}_n(\mathcal{H})$ larger
- Similarly if ℓ has small variance $\rightarrow \tilde{\mathcal{R}}_n(\mathcal{H})$ is smaller (Lipschitz)
- As n grows, harder to fit $\rightarrow \tilde{\mathcal{R}}_n(\mathcal{H})$ smaller

16 / 18

Intuition (see figures in handwritten notes)

- Let's look $\tilde{\mathcal{R}}_n(\mathcal{H}) = \mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \sum_{i=1}^n \epsilon_i h(z_i)$ for fixed z_i and $h(z_i) = \ell(z_i; f)$ and see how it might decrease with n
- For simplicity, let $\mathcal{Z} = \mathbb{R}$, use e.g. $h(z) = \text{sgn} f(z)$ (you can do it more generally for ℓ)

- Let \mathcal{F} be “smooth” functions, given a draw/sample $\epsilon_1, \dots, \epsilon_n$

Which $f \in \mathcal{F}$ can achieve large $\sum_{i=1}^n \epsilon_i \ell(z_i, f)$?

- Maximizing $\tilde{\mathcal{R}}_n(\mathcal{H})$ requires for each $\{\epsilon_i\}_{i=1}^n$ matching “induced labeling” of f ($\{f(z_i)\}_{i=1}^n$)
- For small n , you can find a f for each sample of $\{\epsilon_i\}_{i=1}^n$ that matches in sign, i.e. $|\{(h(z_1), \dots, h(z_n)) : h \in \mathcal{H}\}| = 2^n$, then $\mathbb{E} \sup_{f \in \mathcal{F}} \sum_{i=1}^n \epsilon_i h(z_i) = 1$
- For large n , points are too dense, if \mathcal{F} need to be smooth, not that possible for some very “wiggly” $\{\epsilon_i\}_{i=1}^n \rightarrow \mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \sum_{i=1}^n \epsilon_i h(z_i)$

17 / 18

References

Azuma-Hoeffding

- MW Chapter 2

Online to batch conversion with Azuma-Hoeffding

- <https://home.ttic.edu/~tewari/lectures/lecture13.pdf>

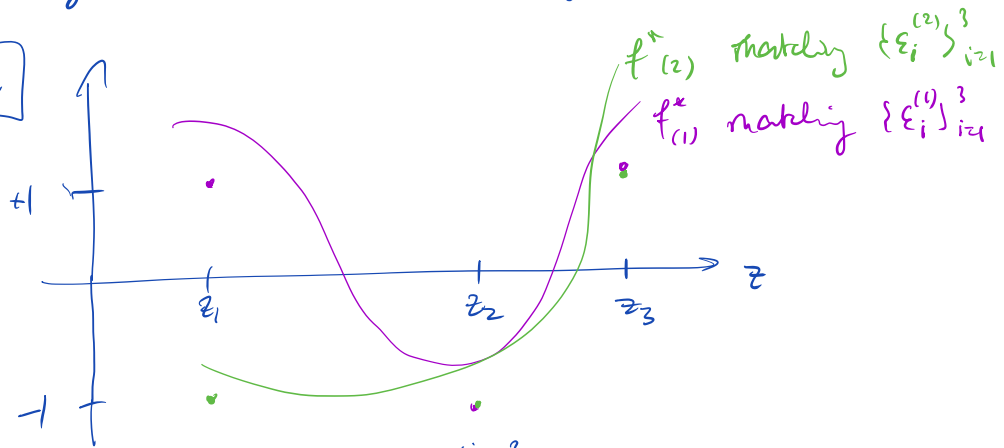
Uniform law and Rademacher complexity

- MW Chapter 4

18 / 18

Assuming "smooth" (not too wiggly) functions $f: \mathbb{R} \rightarrow \mathbb{R}$

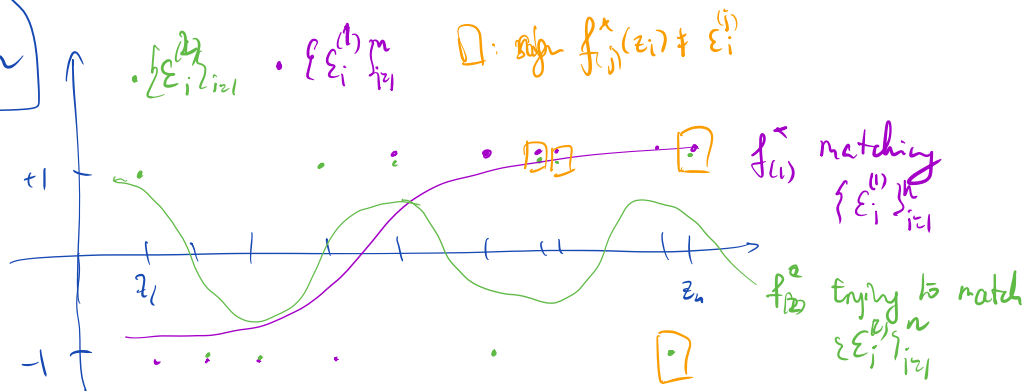
Small n



For all sequences $\{\epsilon_i^{(j)}\}_{i=1}^3$ you can find $f_{(j)}^*$ s.t. $\text{sign } f_{(j)}^*(z_i) = \epsilon_i^{(j)} \quad \forall i, j$

$$\Rightarrow \exists_{\epsilon} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \epsilon_i \text{sign } f(z_i) = 1 \quad (j \text{ indexes the sequence})$$

Larger n



For some sequences $\{\epsilon_i^{(j)}\}_{i=1}^n$ there's no function $f_{(j)}^*$ s.t. $\text{sign } f_{(j)}^*(z_i) = \epsilon_i^{(j)} \quad \forall i, j, n$

$$\Rightarrow \exists_{\epsilon} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum \epsilon_i \ll 1$$