Lecture 4: Uniform law and Rademacher complexity

FAQ for muddiest point

Online learning

- added more motivation and explanation, also lecture note from Tewari, Kakade.

Azuma-Hoeffding

- How martingale properties allow the AH bound (will discuss now)

Questions on the uniform law - will discuss today

Plans for today

- Recap Azuma-Hoeffding proof
- Intuition for Rademacher complexity
- Proof of uniform law with symmetrization
- Application of Rademacher complexity: Finite function classes
    - Massart's lemma and its proof

Recap: From Hoeffding to Azuma-Hoeffding

Why use $D_i = \mathbb{E}[g_n(Z)|Z_1, \dots, Z_i] - \mathbb{E}[g_n(Z)|Z_1, \dots, Z_{i-1}]$ to decompose g_n? Azuma-Hoeffding is a *generalization* of Hoeffding (i.e. Azuma-Hoeffding implies Hoeffding), for functions of n independent R.V. instead of sum of n independent RV.

Decomposition of g_n

- For Hoeffding, in $\frac{1}{n}\sum_{i=1}^n X_i$ for X_i independent, each R.V. adds fresh randomness $\rightarrow$
- For AH, in decomposition $\sum_{i=1}^n D_i$, each D_i has the additional randomness that is due to addition of Z_i only. This is why we chose the particular D_i (property 1 next slide)

In addition the D_i are in some sense **bounded** (for McDiarmid, generally subgaussian is fine), so sth "like Hoeffding" should work:

- For Hoeffding, each summand is subgaussian $\rightarrow$
- For AH (for proving McDiarmid), each summand is conditionally a.s. bounded and hence also conditionally subgaussian (property 2)

Recap: Martingale properties to prove Azuma-Hoeffding

The following properties of this choice are what we need in the proof (these are the properties of martingale differences)

1. D_i is $\mathcal{F}_i$ measurable, i.e. D_i is a deterministic function given specific values for $Z_1, \ldots, Z_i$
2. For any values $z_1, \ldots, z_{i-1}$, for some a_i, b_i
   - the random variable $D_i | Z_1^{i-1} = z_1^{i-1}$ is bounded in an interval $[a_i, b_i]$ of length L_i and
   - $\mathbb{E}[D_i | Z_1^{i-1} = z_1^{i-1}] = 0$ and hence together we use the fact that r.v. bounded a.s. in $[a_i, b_i]$ are $\frac{b_i - a_i}{2}$ subgaussian to get

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

Further we use the tower property (TP): $\mathbb{E}[\mathbb{E}[X|Y, Z]|Y] = \mathbb{E}[X|Y]$

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

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

Recap: Uniform tail bound via Rademacher complexity

- Define $\mathcal{H} = \{h : h(\cdot) = \ell(\cdot; f) \quad \forall f \in \mathcal{F}\}$
- ϵ_i are i.i.d. Rademacher R.V.
- $Z = \{Z_i\}_{i=1}^n$ are training points $\overset{iid}{\sim} \mathbb{P}$

Definition (Rademacher complexity)

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

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

Intuition for Rademacher complexity

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

1. How does the empirical Rademacher complexity

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

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

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

2. What is the connection between R.C. and VC dimension?

3. (Why) is it easier to reason about than the original $\text{Res}(n, \mathcal{H}) = \mathbb{E} g_n(Z)$

Some answers

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

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 $\epsilon_1, \ldots, \epsilon_n$

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

- 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. $|\{(h(z_1), \ldots, h(z_n)) : h \in \mathcal{H}\}| = 2^n$, 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" $\{\epsilon_i\}_{i=1}^n \to \mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \sum_{i=1}^n \epsilon_i h(z_i)$

Caveats of the uniform law

- Requires boundedness of ℓ (for bounded differences)
  - for regression you also bound suprema of empirical processes, can use Gaussian complexity and Lipschitz-of-Gaussians rule (see MW 3)
  - or argue that ℓ bounded with high probability, cause X and hence $f(X)$ bounded for continuous f
- Super loose bound $\rightarrow \mathcal{F}$ needs to be algorithm / data dependent
  - we will see for regularized optimizers
  - structural risk minimization
- in second half of lectures we'll discuss a different way to bound the excess risk for regression $\rightarrow$ however even there, we will control suprema of empirical processes will be needed

Proof of uniform law - Step I: Tail bound

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.

We recapped the proof last lecture, using McDiarmid.

In particular, by the uniform tail bound, if we can prove that $\mathbb{E}[\sup_{f\in\mathcal{F}} R(f) - R_n(f)] \leq 2\mathcal{R}_n(\mathcal{H})$ then it immediately follows that

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

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

This proof step is called symmetrization

Proof of uniform law - Step II: Symmetrization

(i) For any H, $\sup_H \mathbb{E} H(Z) \le \mathbb{E}\sup_H H(Z)$ (Exercise)
(ii) $h(Z_i) - h(\tilde{Z}_i)$ is symmetric $\to$ multiplying by ϵ_i preserves distr.

$$
\begin{aligned}
\mathbb{E}_Z g_n(Z) &= \mathbb{E}_Z \sup_{h\in\mathcal{H}} \mathbb{E}h - \frac{1}{n}\sum_i h(Z_i) \\
&= \mathbb{E}_Z \sup_{h\in\mathcal{H}} \mathbb{E}_{\tilde{Z}} \frac{1}{n}\sum_{i=1}^n h(\tilde{Z}_i) - \frac{1}{n}\sum_{i=1}^n h(Z_i) \\
&\overset{(i)}{\le} \mathbb{E}_{Z,\tilde{Z}} \sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n [h(Z_i) - h(\tilde{Z}_i)] \\
&\overset{(ii)}{=} \mathbb{E}_{Z,\tilde{Z},\epsilon} \sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \epsilon_i [h(Z_i) - h(\tilde{Z}_i)] \\
&\le 2\mathbb{E}_{Z,\epsilon} \sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \epsilon_i h(Z_i) =: 2\mathcal{R}_n(\mathcal{H}) \square
\end{aligned}
$$

- Tight: $\frac{\mathcal{R}_n(\mathcal{H})}{2} \le \mathbb{E}\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_i h - \mathbb{E}h \le 2\mathcal{R}_n(\mathcal{H})$ (MW Prop 4.11.)

Classification setup

- Labels are now in discrete domain $y \in \{-1, +1\}$
- Given f, we predict the label of some x using $\hat{y} = sign(f(x)$
- Evaluation metric: $\ell((x,y); f) = \mathbb{1}_{\{yf(x)<0\}}$ and hence population risk: $R(f) = \mathbb{E}\ell((x,y);f) = \mathbb{P}(y \ne sign(f(x))$
- A fixed $f \in \mathcal{F}$ defines a labeling from domain $\mathcal{X} \to \{-1,+1\}$. For a *given set* $Z^n = \{Z_i = (x_i, y_i)\}_{i=1}^n$, the function space $\mathcal{F}$ induces a set in $\{-1,1\}^n$ that reads $\mathcal{F}(Z^n) = \{(f(Z_1), \ldots, f(Z_n)) : f \in \mathcal{F}\}$
- We again use notation $h(z) = \ell(z, f)$ and define

$$\mathcal{H}(Z^n) = \{(\ell(Z_1; f), \ldots, \ell(Z_n; f)) : f \in \mathcal{F}\}$$

Notice that $|\mathcal{F}(Z^n)| = |\mathcal{H}(Z^n)|$

Massart's lemma

Lemma (Massart)

For n points $Z^n := \{Z_1, \ldots, Z_n\}$, let all $h : \mathcal{Z} \to \{0,1\}$ and $\mathcal{H}(Z^n) := \{(h(Z_1), \ldots, h(Z_n)) : h \in \mathcal{H}\}$ with cardinality $|\mathcal{H}(Z^n)|$.

$$\tilde{\mathcal{R}}_n(\mathcal{H}(Z^n)) := \mathbb{E}_\epsilon \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \epsilon_i h(Z_i) \leq \sqrt{\frac{2 \log |\mathcal{H}(Z^n)|}{n}}$$

- Step 1: For Rademacher ϵ_i and any Z_1^n we have that $\theta_i := h(Z_i) \in \{0,1\}$, show $\frac{1}{n}\epsilon^\top \theta$ is zero-mean and $\frac{1}{\sqrt{n}}$ sub-gaussian (similar to Hoeffding proof). This follows from the fact that $[a_i, b_i]$ bounded r.v. are $[b_i - a_i]/2$ subgaussian
- Step 2: Use the fact from HW 1 that, for N zero-mean subgaussians $X_1, \ldots, X_N$ with sub-gaussian parameter σ

  $$\mathbb{E} \max_{i=1..N} X_i \leq \sqrt{2\sigma^2 \log N}$$

  Here, $N = \mathcal{H}(Z^n)$ the number of different vectors $(h(Z_1), \ldots, h(Z_n))$

References

Uniform law

- MW Chapter 4
- "Understanding machine learning" by Shalev-Shwartz, Ben-David, Chapter 26