Lecture 2: Uniform tail bound and McDiarmid

Plan today

1. Recap excess risk decomposition and Hoeffding's inequality
2. Concentration of functions of n dependent r.v. via bounded differences
3. McDiarmid inequality and uniform tail bound
4. Proof of McDiarmid via Doob martingales, Azuma-Hoeffding inequality

Recap last lecture: excess risk decomposition

- Recall we assume that $Z_i := (X_i, Y_i) \overset{\text{i.i.d.}}{\sim} \mathbb{P}$ with $Z_i \in \mathcal{Z}$ and evaluate a function f by the expected loss (population risk) $R(f) = \mathbb{E}\ell(Z; f)$
- The empirical risk is defined by $R_n(f) = \frac{1}{n}\sum_{i=1}^n \ell(Z_i; f)$ and for fixed f, we have $\mathbb{E}R_n(f) = R(f)$.
- We want to bound the excess risk

$$R(\widehat{f}_n) - R(f^\star) = R(\widehat{f}_n) - R_n(\widehat{f}_n) + \overbrace{R_n(\widehat{f}_n) - R_n(f^\star)}^{\leq 0 \text{ by optimality}} + R_n(f^\star) - R(f^\star)$$
$$\leq \underbrace{R(\widehat{f}_n) - R_n(\widehat{f}_n)}_{T_1} + \underbrace{R_n(f^\star) - R(f^\star)}_{T_2}$$

- Then we can use Hoeffding's inequality to bound T_2 if $X = \ell(Z, f)$ for fixed f is subgaussian (e.g. bounded loss)

$$\mathbb{P}(\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}X \geq t) \leq e^{-\frac{nt^2}{2\sigma^2}}$$

Proof of Hoeffding's inequality

Lemma (Hoeffding's inequality)

For i.i.d sub-Gaussian R.V. X_i, it holds that

$$\mathbb{P}(\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}X \geq t) \leq e^{-\frac{nt^2}{2\sigma^2}}$$

Proof:

1. We can apply Chernoff on the mean of n independent random variables with moment generating function

$$\mathbb{E}e^{\lambda(\frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}X_i))} = \prod_{i=1}^n \mathbb{E}e^{\frac{\lambda}{n}(X_i - \mu)} = [\mathbb{E}e^{\frac{\lambda}{n}(X_i - \mu)}]^n$$

2. Hence, the mean of n i.i.d. sub-Gaussian variables is sub-Gaussian with parameter $\frac{\sigma}{\sqrt{n}}$ since $\mathbb{E}e^{\lambda(\frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}X_i))} \leq e^{\frac{\lambda^2\sigma^2}{2n^2}n}$
3. yielding Hoeffding's inequality for the mean of iid sub-Gaussians

$$\mathbb{P}(\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}X \geq t) \leq e^{-\frac{nt^2}{2\sigma^2}}$$

Back to term T_1

- Problem: $R_n(\widehat{f}_n) = \frac{1}{n}\sum_{i=1}^n \ell(Z_i;\widehat{f}_n)$ *not* an emp. mean of i.i.d. R.V.! Can we still show some sort of concentration for $R_n(\widehat{f}_n)$?
- Crude bound: since by assumption algorithm searches in a model/function class $\mathcal{F}$, i.e. $\widehat{f}_n \in \mathcal{F}$, we can upper bound T_1 by

$$T_1 = R(\widehat{f}_n) - R_n(\widehat{f}_n) \leq \sup_{f\in\mathcal{F}} R(f) - R_n(f) =: g_n(Z_1,\ldots,Z_n)$$

- Recall for fixed f, $R_n(f)$ is an empirical mean of n i.i.d. random variables $\ell(Z_i;f)$ if Z_i are i.i.d. - this appears in T_2.
- As opposed to the *average of n* i.i.d. random variables, for T_1 we now bound another *function of n* i.i.d. random variables, that is $g_n : \mathcal{Z}^n \to \mathbb{R}$ - the supremum of an *empirical process* $R(f) - R_n(f)$
- While for T_2 we can show that empirical mean $R_n(f)$ concentrates around its expectation $\mathbb{E}R_n(f) = R(f)$, we now show:
if g_n satisfies some properties, g_n concentrates around $\mathbb{E}g_n(z)$!

Plan:

- Tail bound for bounded losses
  - what we can do with the tail bound
  - McDiarmid for g_n with bounded differences
  - Proof of tailbound with McDiarmid
- Proof of McDiarmid
  - Proof intuition: High-level steps
  - Proof of McDiarmid using Azuma-Hoeffding
    - Azuma-Hoeffding inequality tailored for this proof
    - Proof why assumptions for Azuma-Hoeffding hold

Tail bound for supremum of (bounded) empirical process

- Remember for $f \in \mathcal{F}$: $R_n(f) = \frac{1}{n}\sum_{i=1}^n \ell(Z_i, f)$ and $R(f) = \mathbb{E}\ell(Z, f)$
- We can now use McDiarmid's inequality on the sup. of empirical process $g_n(Z_1, \ldots, Z_n) = \sup_{f\in\mathcal{F}} R(f) - R_n(f)$ for bounded losses to prove the following tail bound

Theorem (Uniform tail bound)

For b-unif. bounded $\ell(\cdot, f)$, that is $\|\ell(\cdot; f)\|_\infty \leq b$ for all $f \in \mathcal{F}$, 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.

- Note that there are other results beyond boundedness (Lipschitz functions etc.), that are tighter particularly in the context of bounding suprema of empirical process - MW Chapter 3

"Recap": Bounding excess risk using the tail bound

Discuss with your neighbor: Assuming $\text{Res}(n, \mathcal{F}) := \mathbb{E}[\sup_{f\in\mathcal{F}} R(f) - R_n(f)]$ is bounded, how do we now bound T_1 with high probability?

We immediately obtain

$$\mathbb{P}(\sup_{f\in\mathcal{F}} R(f) - R_n(f) \leq \text{Res}(n, \mathcal{F}) + t) \geq 1 - e^{-\frac{nt^2}{2b^2}}$$

This is a "high probability" bound in the sense that with probability at least $1 - \delta$ we have

$$\sup_{f\in\mathcal{F}} R(f) - R_n(f) \leq b\sqrt{\frac{2\log(\frac{1}{\delta})}{n}} + \text{Res}(n, \mathcal{F})$$

McDiarmid inequality for g_n with bounded differences

Definition (bounded difference property)

Define for given $z, z' \in \mathcal{Z}^n$ a new vector $z^{\setminus k}$ with the k-th element from z' and all other from z: $z_j^{\setminus k} = \begin{cases} z_j & \text{if } j \neq k \\ z'_k & \text{if } j = k \end{cases}$. We say that $g_n : \mathcal{Z}^n \to \mathbb{R}$ satisfies the bounded difference inequality if for each $k = 1, \ldots, n$ it holds that

$$|g_n(z) - g_n(z^{\setminus k})| \leq \sigma_k \quad \text{for all } z, z' \in \mathcal{Z}^n$$

Theorem (McDiarmid, MW Cor. 2.21)

If $g_n : \mathcal{Z}^n \to \mathbb{R}$ satisfies the bounded difference condition and $Z \in \mathcal{Z}^n$ is a random vector with n independent entries, then

$$\mathbb{P}(g_n(Z) - \mathbb{E}g_n(Z) \geq t) \leq e^{-\frac{2t^2}{\sum_{k=1}^n \sigma_k^2}}$$

- Concentration with n is usually obtained via $t \sim n$ or via $\sigma_k \sim \frac{1}{n}$

Proof of tail bound using McDiarmid

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

Use McDiarmid by checking bounded differences assumption with $g_n(z) := \sup_{f \in \mathcal{F}} R_n(f) - R(f) = \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n h(z_i) - \mathbb{E}h$

- For b-uniformly bounded $\mathcal{H}$, we have for all $k = 1, \ldots, n$ and any $z, z' \in \mathcal{Z}^n$ that for any $h \in \mathcal{H}$

$$\frac{1}{n}\sum_i [h(z_i) - \mathbb{E}h] - \sup_{\tilde{h} \in \mathcal{H}} \frac{1}{n} \sum_i \left[\tilde{h}(z_i^{\setminus k}) - \mathbb{E}\tilde{h}\right]$$
$$\leq \frac{\sum_i h(z_i) - h(z_i^{\setminus k})}{n} = \frac{h(z_k) - h(z'_k)}{n} \leq \frac{2b}{n}$$

- Since it holds for all $h \in \mathcal{H}$, taking the sup on both sides yields

$$g_n(z) - g_n(z^{\setminus k}) = \sup_{h \in \mathcal{H}} \frac{1}{n}\sum_i [h(z_i) - \mathbb{E}h] - \sup_{\tilde{h} \in \mathcal{H}} \frac{1}{n} \sum_i \left[\tilde{h}(z_i^{\setminus k}) - \mathbb{E}\tilde{h}\right] \leq \frac{2b}{n}$$

- By symmetry it holds for $g_n(z^{\setminus k}) - g_n(z) \to |g_n(z) - g_n(z^{\setminus k})| \leq \frac{2b}{n}$
- Plugging in $\sigma_k = \frac{2b}{n}$ into McDiarmid then yields the result.

Intuition for proving McDiarmid

Theorem (McDiarmid, MW Cor. 2.21)

If $g_n : \mathcal{Z}^n \to \mathbb{R}$ satisfies the bounded difference condition with $\{\sigma_k\}_{k=1}^n$ and Z is a random vector with n independent entries, then

$$\mathbb{P}(g_n(Z) - \mathbb{E}g_n(Z) \geq t) \leq e^{-\frac{2t^2}{\sum_{k=1}^n \sigma_k^2}}$$

Proof intuition:

Re-writing g_n as a sum

- For any function $g_n : \mathcal{Z}^n \to \mathbb{R}$, even though we don't have an average / sum per se, we can write the difference as a sum (check for yourself)

$$g_n(Z) - \mathbb{E}g_n(Z) =: \sum_{j=1}^n D_j$$

where $D_j := \mathbb{E}[g_n(Z)|Z_1, \ldots, Z_j] - \mathbb{E}[g_n(Z)|Z_1, \ldots Z_{j-1}]$ for $j \geq 2$ and $D_1 = \mathbb{E}[g_n(Z)|Z_1] - \mathbb{E}[g_n(Z)]$

Proof intuition Part I

- For the special case of empirical mean $g_n(Z) = \frac{1}{n}\sum_{i=1}^n Z_i$ with Z_i independent and bounded, we get for all $j = 1, \ldots, n$

$$D_j = \frac{1}{n}\sum_{i=j}^n \mathbb{E}[Z_i|Z_1, \ldots, Z_j] - \frac{1}{n}\sum_{i=j-1}^n \mathbb{E}[Z_i|Z_1, \ldots, Z_{j-1}] = \frac{Z_j}{n} - \frac{\mathbb{E}Z}{n}$$

with all D_j independent and bounded (hence sub-Gaussian) and hence one can use Hoeffding's bound

- Can we use this for general g_n? For general $g_n(Z)$ independence of D_j does not hold!

Proof intuition Part II

- However, we can still show that instead of independent subgaussians D_j, we have
  - a martingale difference sequence D_j where
  - D_j is bounded "conditionally" on the past (and is still conditionally subgaussian)
- (informal) Then instead of *Hoeffding* that can be used on independent **bounded** R.V., we can use *Azuma-Hoeffding*, that shows

$$\mathbb{P}(\sum_{i=1}^{n} D_i \geq t) \leq \mathrm{e}^{-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}}$$

  for **bounded** martingale difference sequences where $D_i \in [a_i, b_i]$ a.s.

We now formalize the proof of part II.

Formal proof of McDiarmid: Part I

- Define shorthand $Z_1^i = (Z_1, \ldots, Z_i) \in \mathcal{Z}^i$ for random/real vectors
- We require the following assumptions on $\{D_j\}_{j=1}^{\infty}$ corresponding to is a martingale difference sequence

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 $\sigma(Z_1, \ldots, Z_j)$*
- *$\mathbb{E}[D_j | Z_1, \ldots, Z_{j-1}] = 0$ and $\mathbb{E}|D_j| < \infty$*
- *$D_j | Z_1, \ldots, Z_{j-1}$ 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

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

Formal proof of McDiarmid: Part II

Note: This version of Azuma-Hoeffding is tailored to our proof, Azuma-Hoeffding generally holds for any bounded martingale difference sequence wrt filtrations generally, the way we write it is to connect more directly with what's needed to prove McDiarmid.

The proof of McDiarmid now follows from Azuma-Hoeffding immediately if we can show that

Fact (Doob martingale differences)

For any g_n satisfying the bounded difference property with $\{\sigma_j\}_{j=1}^n$, the Doob martingale difference sequence $D_j := \mathbb{E}[g_n(Z)|Z_1, \ldots, Z_j] - \mathbb{E}[g_n(Z)|Z_1, \ldots Z_{j-1}]$ satisfies the bounded martingale difference assumptions.

First, measurability and conditional zero mean follows from the definition and the tower property. In the next slide we show boundedness.

Proof of Fact

- We can now prove that if g_n satisfies the bounded difference condition with $\{\sigma_j\}_{j=1}^n$, then for all $z_1^{j-1} \in \mathcal{Z}^{j-1}$ there exists a_j, b_j s.t. $D_j|Z_1^{j-1} = z_1^{j-1} \in [a_j, b_j]$ almost surely with $b_j - a_j \leq \sigma_j$
- We define shorthand (last equality follows by independence of Z_j): $\mathbb{E}[g_n(Z)|z_1^{j-1}] := \mathbb{E}[g_n(Z)|Z_1^{j-1} = z_1^{j-1}] = \mathbb{E}g_n(z_1^{j-1}, Z_j^n)$
- Further, by definition, for all $z_1^{j-1} \in \mathcal{Z}^{j-1}$ we have

$$D_j|Z_1^{j-1} = z_1^{j-1} \geq \inf_{z\in\mathcal{Z}} \mathbb{E}[g_n(Z)|z_1^{j-1}, Z_j = z] - \mathbb{E}[g_n(Z)|z_1^{j-1}] =: a_j$$

$$D_j|Z_1^{j-1} = z_1^{j-1} \leq \sup_{z\in\mathcal{Z}} \mathbb{E}[g_n(Z)|z_1^{j-1}, Z_j = z] - \mathbb{E}[g_n(Z)|z_1^{j-1}] =: b_j$$

- Hence $D_j|Z_1^{j-1} = z_1^{j-1} \in [a_j, b_j]$ and, by bounded diff. ass. on g_n:

$$\begin{aligned} b_j - a_j &= \sup_{z\in\mathcal{Z}} \mathbb{E}g_n(z_1^{j-1}, z, Z_{j+1}^n) - \inf_{z\in\mathcal{Z}} \mathbb{E}g_n(z_1^{j-1}, z, Z_{j+1}^n) \\ &\leq \sup_{z,z'\in\mathcal{Z}} \mathbb{E}|g_n(z_1^{j-1}, z, Z_{j+1}^n) - g_n(z_1^{j-1}, z', Z_{j+1}^n)| \leq \sigma_j \end{aligned}$$

Summary

- McDiarmid inequality for bounded difference
- uniform tail bound for T_1
- Proof McDiarmid: Hoeffding bound for sums of independent R.V. $\rightarrow$ martingale (difference) sequences and Azuma-Hoeffding inequality

Next up: Uniform law with symmetization and Rademacher complexity

References

Concentration bounds including Azuma-Hoeffding, McDiarmid

- MW Chapter 2
- *Boucheron, Lugosi, Massart*: Chapter 2

Martingales - any probability theory book, e.g.:

- *P. Billingsley. Probability and Measure*
- *R. Durrett. Probability: Theory and Examples (4th edition)*

(Bonus) More concentration bounds on suprema of empirical processes:

- MW Chapter 3
- *Ledoux, Talagrand: Probability for Banach spaces* for functional Bernstein