# Lecture 2: Uniform tail bound and McDiarmid



## Announcements and lecture outline

Announcements:

- HW released tonight, due in two weeks on Thursday **12.10.22 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!

## 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

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

- Then via Chernoff, we proved Hoeffding's inequality that holds for the mean of i.i.d. sub-Gaussians

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

*Wakeup-Q*: How can we use Hoeffdings inequality to bound $T_2$?

## 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

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

- Instead of averages of $n$ i.i.d. random variables, the supremum of an *empirical process* $R(f) - R_n(f)$ is a general function $g_n : \mathcal{Z}^n \to \mathbb{R}$
- Instead of $R_n(f) \approx \mathbb{E}R_n(f) = R(f)$ for empirical means, if $g_n$ satisfies some properties, $g_n$ concentrates around $\mathbb{E}g_n(z)$!



## Specific case: $g_n$ satisfies bounded difference property

**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 $k = 1, \dots, n$ 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 $\sigma_k \sim \frac{1}{n}$

## Tail bound for supremum of (bounded) empirical process

- Remember for $f \in \mathcal{F}$: $R_n(f) = \frac{1}{n}\sum_{i=1}^n \ell(Z_i, f)$
- We can now use McDiarmid on the sup. of empirical process $g_n(z_1, \dots, z_n) = \sup_{f\in\mathcal{F}} R(f) - R_n(f)$ for bounded losses!

**Theorem (Uniform tail bound)**

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

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

*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
- This uniform tail bound can give us a (crude) high-probability bound and rate, if we can bound the expectation ($\rightarrow$ next class!)



## 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 $k = 1, \dots, n$ and any $z, z' \in \mathcal{Z}^n$ that for any $h \in \mathcal{H}$

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

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

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

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

# Proof sketch of 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 a 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

*Discuss with your neighbor*: 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

$\to$ $D_j$ are independent and sub-Gaussian so that one can use Hoeffding's bound on $D_j$. Can we use this for general $g_n$?

- Indeed, 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)

- For general $g_n(Z)$ independence of $D_j$ does not hold!

## Proof intuition Part II

- However, we can still show that
  - $D_j$ indepedendent $\rightarrow$ $D_j$ martingale difference, and hope that $D_j$ s.t.
  - $D_j$ "conditionally" bounded (and hence still in some way subgaussian)
- (informal) Then instead of *Hoeffding* that can be used on independent **bounded** R.V., we can use *Azuma-Hoeffding*, that shows

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

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

We now formalize the proof.



## "Recap": Martingale difference sequences

Let $\{Z_j\}_{j=1}^{\infty}$ be a sequence of R.V. and $\mathcal{F}_j := \sigma(Z_1, \ldots, Z_j)$,

Further, let $\{S_j\}_{j=1}^{\infty}$ be such that $S_j$ is measurable with respect to $\mathcal{F}_j$ (i.e. we say $\{S_j\}_{j=1}^{\infty}$ is *adapted to the filtration* $\{\mathcal{F}_j\}_{j=1}^{\infty}$)

**Definition (Martingale (difference))**

- $\{S_j, \mathcal{F}_j\}_{j=1}^{\infty}$ is a *martingale* if for all $j$, $\mathbb{E}|S_j| < \infty$ and $\mathbb{E}[S_{j+1}|\mathcal{F}_j] = S_j$
- Similarly, $\{D_j, \mathcal{F}_j\}_{j=1}^{\infty}$ is a *martingale difference sequence* if for all $j$, $\mathbb{E}|D_j| < \infty$ and $\mathbb{E}[D_{j+1}|\mathcal{F}_j] = 0$

- For any martingale $\{S_j, \mathcal{F}_j\}_{j=0}^{\infty}$, $D_j = S_j - S_{j-1}$ for $j \geq 1$ is a martingale difference sequence.
- Doob construction: given some function $g_n : \mathcal{Z}^n \rightarrow \mathbb{R}$, for a sequence of random variables $Z_1, \ldots, Z_n$, note that $S_j = \mathbb{E}[g_n(Z)|Z_1, \ldots, Z_j]$ fulfills exactly the above conditions if $\mathbb{E}|g_n(Z)| < \infty$. Then also $\mathbb{E}[D_{j+1}|\mathcal{F}_j] = 0$ for $D_j = S_j - S_{j-1}$
  *Check with your neighbor*

## Formal proof of McDiarmid

**Theorem (Azuma-Hoeffding inequality, MW Cor 2.20)**

*If for martingale difference sequence $\{(D_i, \mathcal{F}_i)\}_{i=1}^n$ it holds that $D_i|\mathcal{F}_{i-1}$ almost surely lies in an interval of length $L_i$ for all $i$, then*

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

*Note*: This version is slightly different than MW Cor 2.20 - this version does not require all $D_i|z_1^{i-1}$ to be in the same range $[a_i, b_i]$ for all $z_1^{i-1}$ - only the length matters

The proof of McDiarmid follows immediately if we can show that

- for any $g_n$ satisfying the bounded difference property with $\{\sigma_j\}_{j=1}^n$
- we have that $g_n(Z) - \mathbb{E}g_n(Z) = \sum_{j=1}^n D_j$ with $\{D_j, \mathcal{F}_j\}_{j=1}^n$ a martingale difference sequence and $D_i|\mathcal{F}_{i-1}$ almost surely lies in an interval of length $L_i$

We now show that this fact is true.



## Proof: Assumptions of Azuma-Hoeffding hold

- Define shorthand $Z_1^i = (Z_1, \ldots, Z_i) \in \mathcal{Z}^i$ for random/real vectors
- 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 $z_1^{j-1} \in \mathcal{Z}^{j-1}$ almost surely

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

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

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

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

## 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