Algorithmic Foundations of Learning **Lecture 6**

Sub-Gaussian Concentration Inequalities. Bounds in Probability

Lecturer: Patrick Rebeschini *Version: December 8, 2021*

6.1 Introduction

In the previous lectures we developed tools to control the *expected value* of suprema of empirical processes, which we used, in particular, to bound quantities of the form $\mathbf{E}\sup_{a\in\mathcal{A}}\{r(a)-R(a)\}$.

Now we address the problem of deriving bounds *in probability*. Recall from Lecture 1 that we are interested in finding `UpperTailStats`, a strictly decreasing function of ε, such that for any $\varepsilon \geq 0$,

$$\mathbf{P}\Big(\sup_{a\in\mathcal{A}}\{r(a)-R(a)\} \geq \mathbf{E}\sup_{a\in\mathcal{A}}\{r(a)-R(a)\} + \varepsilon\Big) \leq \frac{1}{2}\boxed{\texttt{UpperTailStats}(\varepsilon)}$$

or, equivalently, for any $\delta \in [0,1]$,

$$\mathbf{P}\Big(\sup_{a\in\mathcal{A}}\{r(a)-R(a)\} < \mathbf{E}\sup_{a\in\mathcal{A}}\{r(a)-R(a)\} + \boxed{\texttt{UpperTailStats}^{-1}(2\delta)}\Big) \geq 1-\delta.$$

Recall that the supremum we are interested about is a function of the training data, namely,

$$\sup_{a\in\mathcal{A}}\{r(a)-R(a)\} = f(Z_1,\dots,Z_n).$$

Hence, establishing the bounds above falls within the scope of establishing concentration inequalities for a deterministic function f of random variables $Z_1,\dots,Z_n$, namely, find $\texttt{UpperTail}_f$, a strictly decreasing function of ε such that, for any $\varepsilon \geq 0$,

$$\mathbf{P}\Big(f(Z_1,\dots,Z_n) - \mathbf{E}\,f(Z_1,\dots,Z_n) \geq \varepsilon\Big) \leq \boxed{\texttt{UpperTail}_f(\varepsilon)}$$

or, equivalently, for any $\delta \in [0,1]$,

$$\mathbf{P}\Big(f(Z_1,\dots,Z_n) - \mathbf{E}\,f(Z_1,\dots,Z_n) < \boxed{\texttt{UpperTail}_f^{-1}(\delta)}\Big) \geq 1-\delta.$$

6.2 Concentration Phenomenon

The phenomenon of concentration of random variables can informally be stated as follows [1]:

> *If $X_1,\dots,X_n$ are independent (or weakly dependent) random variables, then the random variable $f(X_1,\dots,X_n)$ is "close" to its mean $\mathbf{E}[f(X_1,\dots,X_n)]$ provided that the function $x_1,\dots,x_n \to f(x_1,\dots,x_n)$ is not too "sensitive" to any of the coordinates x_i.*

This phenomenon takes various incarnations, depending on the chosen notions of "closeness" and "sensitivity". We have already seen this phenomenon in action. In **Problem 1.1** in the Problem Sheets we noted that if $X_1,\dots,X_n$ are i.i.d. random variables with mean μ and variance σ^2, then, for any $p \geq 2$ we have

$$\left\{\mathbf{E}\left[\left(\frac{1}{n}\sum_{i=1}^n X_i - \mu\right)^p\right]\right\}^{1/p} \leq \frac{c_p}{\sqrt{n}},$$

for a constant c_p that depends on p and on the distribution of X, but is independent of n. If we define the function $x_1, \ldots, x_n \to f(x_1, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^n x_i$, then the above can be rewritten as

$$\left\{ \mathbf{E}\left[\left(f(X_1, \ldots, X_n) - \mathbf{E}f(X_1, \ldots, X_n) \right)^p \right] \right\}^{1/p} \leq \frac{c_p}{\sqrt{n}}.$$

Here, as notion of "closeness" we choose the metric $Z, \widetilde{Z} \to (\mathbf{E}[(Z-\widetilde{Z})^p])^{1/p}$, which gives rise to the standard deviation $\sqrt{\mathbf{Var} f(X_1, \ldots, X_n)}$ for $p = 2$. As far as the "sensitivity" to changes in the coordinates, note that if we perturb one coordinate at a time we get

$$|f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, \tilde{x}_i, x_{i+1}, \ldots, x_n)| = \frac{|x_i - \tilde{x}_i|}{n},$$

which is small as a function of n (assuming that $|x_i - \tilde{x}_i|$ is controlled).

In the following, we will generalize these ideas and introduce a set of tools to systematically capture the concentration phenomenon. In particular, motivated by the discussion in Section 6.1, we will derive concentration inequalities where the notion of "closeness" is not just given by a quantity that captures the *size* of the fluctuations, as for $Z, \widetilde{Z} \to (\mathbf{E}[(Z-\widetilde{Z})^p])^{1/p}$, but by a quantity that gives a sharper control on the *distribution* of the fluctuations, as for $Z, \widetilde{Z} \to \mathbf{P}(Z - \widetilde{Z} \geq \varepsilon)$.

6.3 Markov's Inequality

The main tool that allows to derive concentration inequalities is given by Markov's inequality.

Proposition 6.1 (Markov's Inequality) *For any non-negative random variable X we have, for any $\varepsilon \geq 0$,*

$$\boxed{\mathbf{P}(X \geq \varepsilon) \leq \frac{\mathbf{E}X}{\varepsilon}}$$

Proof: As for any event E we have $1 = \mathbf{1}_E + \mathbf{1}_{E^C}$, it follows that $X = X\mathbf{1}_{X \geq \varepsilon} + X\mathbf{1}_{X < \varepsilon} \geq \varepsilon \mathbf{1}_{X \geq \varepsilon}$, where we used that $X \geq 0$. Taking the expectation we find $\mathbf{E}X \geq \varepsilon \mathbf{P}(X \geq \varepsilon)$. ■

Markov's inequality proves more useful when combined with a characterizations of the event $\{X \geq \varepsilon\}$ in terms of the exponential function, as we show next. Recall that taking exponentials is the same proof technique that we used to prove maximal inequalities in Lecture 2 (and, in particular, to prove Massart's Lemma).

6.4 Chernoff's Bound

The Chernoff's bound is a very useful technique that allows to translate a bound on the moment generating function into a bound on the tail probabilities. The Chernoff's bound for a random variable X is obtained by applying Markov's inequality to the random variable $e^{\lambda X}$.

Proposition 6.2 (Chernoff's Bound) *For any random variable X and any $\lambda \geq 0$ we have, for any $\varepsilon \in \mathbb{R}$,*

$$\boxed{\mathbf{P}(X \geq \varepsilon) \leq e^{-\lambda \varepsilon}\, \mathbf{E}\, e^{\lambda X}}$$

Proof: For $\lambda > 0$ the function $x \to e^{\lambda x}$ is invertible and increasing so the event $\{X \geq \varepsilon\}$ is equivalent to the event $\{e^{\lambda X} \geq e^{\lambda \varepsilon}\}$ and, by Markov's inequality,

$$\mathbf{P}(X \geq \varepsilon) = \mathbf{P}(e^{\lambda X} \geq e^{\lambda \varepsilon}) \leq \frac{\mathbf{E}\, e^{\lambda X}}{e^{\lambda \varepsilon}}.$$

The bound holds trivially for $\lambda = 0$, as $\mathbf{P}(X \geq \varepsilon) \leq 1$. ■

The Chernoff's bound applied to the centered random variable $X - \mathbf{E}X$ yields a bound on the *upper tail* of the random variable X, that is, a bound on the probability that a random variable X exceeds its mean $\mathbf{E}X$ by a fixed amount, namely,

$$\mathbf{P}(X - \mathbf{E}X \geq \varepsilon) \leq e^{-\lambda \varepsilon}\, \mathbf{E}\, e^{\lambda (X - \mathbf{E}X)}.$$

To get a bound on the *lower tail*, we can apply the Chernoff's bound to the random variable $-(X - \mathbf{E}X)$, as

$$\mathbf{P}(X - \mathbf{E}X \leq -\varepsilon) = \mathbf{P}(-(X - \mathbf{E}X) \geq \varepsilon) \leq e^{-\lambda \varepsilon}\, \mathbf{E}\, e^{-\lambda (X - \mathbf{E}X)}.$$

Given an upper and lower bound, we can obtain a bound on the magnitude of the fluctuation using the union bound:

$$\mathbf{P}(|X - \mathbf{E}X| \geq \varepsilon) = \mathbf{P}(\{X - \mathbf{E}X \leq -\varepsilon\} \cup \{X - \mathbf{E}X \geq \varepsilon\}) \leq \mathbf{P}(X - \mathbf{E}X \leq -\varepsilon) + \mathbf{P}(X - \mathbf{E}X \geq \varepsilon).$$

6.5 Optimal Chernoff's Bound: Convex Conjugate

Chernoff's bound shows that upper-tail bounds can be derived if one has access to bounds on the moment generating function $\lambda \to \mathbf{E}\, e^{\lambda (X - \mathbf{E}X)}$ holding for $\lambda \geq 0$ (analogously, lower-tail bounds can be derived if one has access to upper bounds on the moment generating function for $\lambda \leq 0$). In the following, given a function $\psi : \mathbb{R}_+ \to \mathbb{R}$ we let $\psi^\star : \mathbb{R}_+ \to \mathbb{R}$ be its convex conjugate (a.k.a. Legendre-Fenchel transform) defined as

$$\psi^\star(x) := \sup_{\lambda \geq 0} (\lambda x - \psi(\lambda)).$$

The convex conjugate function naturally appears when one optimizes Chernoff's bound with respect to the parameter λ.

Proposition 6.3 (Optimal Chernoff's Bound)
Let X be a random variable with $\mathbf{E}\, e^{\lambda (X - \mathbf{E}X)} \leq e^{\psi(\lambda)}$ for any $\lambda \geq 0$. Then, for any $\varepsilon \geq 0$ and $\delta \in [0, 1]$ we have

$$\boxed{\mathbf{P}(X - \mathbf{E}X \geq \varepsilon) \leq e^{-\psi^\star(\varepsilon)}} \qquad \boxed{\mathbf{P}(X - \mathbf{E}X < (\psi^\star)^{-1}(\log(1/\delta))) \geq 1 - \delta}$$

Let X be a random variable with $\mathbf{E}\, e^{-\lambda (X - \mathbf{E}X)} \leq e^{\psi(\lambda)}$ for any $\lambda \geq 0$. Then, for any $\varepsilon \geq 0$ and $\delta \in [0, 1]$ we have

$$\boxed{\mathbf{P}(X - \mathbf{E}X \leq -\varepsilon) \leq e^{-\psi^\star(\varepsilon)}} \qquad \boxed{\mathbf{P}(X - \mathbf{E}X > -(\psi^\star)^{-1}(\log(1/\delta))) \geq 1 - \delta}$$

Proof: By optimizing Chernoff's bound we have

$$\mathbf{P}(X - \mathbf{E}X \geq \varepsilon) \leq \inf_{\lambda \geq 0} e^{-\lambda \varepsilon}\, \mathbf{E}\, e^{\lambda (X - \mathbf{E}X)} \leq \inf_{\lambda \geq 0} e^{-\lambda \varepsilon + \psi(\lambda)} = e^{-\sup_{\lambda \geq 0} (\lambda \varepsilon - \psi(\lambda))} = e^{-\psi^\star(\varepsilon)}.$$

The lower tail follows analogously by noting that $\mathbf{P}(X - \mathbf{E}X \leq -\varepsilon) = \mathbf{P}(-(X - \mathbf{E}X) \geq \varepsilon)$. ■

6.6 Sums of i.i.d. Variables via Optimized Chernoff's Bound

Proposition 6.3 immediately yields a concentration inequality for the average of i.i.d. random variables. Henceforth, we only state results for upper tails, as we only need upper-tail bounds for our applications.

Lemma 6.4 *Let $X_1, \dots, X_n \sim X$ be i.i.d. random variables with $\mathbf{E}\, e^{\lambda(X-\mathbf{E}X)} \le e^{\psi(\lambda)}$ for any $\lambda \ge 0$. Then, for any $n \in \mathbb{N}_+$, any $\varepsilon \ge 0$ and $\delta \in [0,1]$ we have*

$$\boxed{\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mathbf{E}X \ge \varepsilon\right) \le e^{-n\psi^\star(\varepsilon)}} \qquad \boxed{\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mathbf{E}X < (\psi^\star)^{-1}\left(\frac{\log(1/\delta)}{n}\right)\right) \ge 1-\delta}$$

Proof: We have $\mathbf{E}\, e^{\lambda \frac{1}{n}\sum_{i=1}^n (X_i - \mathbf{E}X)} = \prod_{i=1}^n \mathbf{E}\, e^{\frac{\lambda}{n}(X_i - \mathbf{E}X)} \le e^{n\psi(\lambda/n)} \equiv e^{\varphi(\lambda)}$, where $\varphi(\lambda) := n\psi(\lambda/n)$. The proof follows by applying Proposition 6.3 noticing that

$$\varphi^\star(\varepsilon) = \sup_{\lambda \ge 0}(\lambda\varepsilon - \varphi(\lambda)) = n \sup_{\lambda \ge 0}\left(\frac{\lambda}{n}\varepsilon - \psi\left(\frac{\lambda}{n}\right)\right) = n \sup_{\lambda \ge 0}(\lambda\varepsilon - \psi(\lambda)) = n\psi^\star(\varepsilon).$$

■

Hence, an average of i.i.d. random variables concentrates exponentially fast around its mean with rate $\psi^\star(\varepsilon)$, where ε is the threshold on the upper tail, if and only if with probability at least $1-\delta$ the fluctuations are of order $(\psi^\star)^{-1}(\log(1/\delta)/n)$. We will see that different classes of random variables, as characterized by the function ψ giving upper bounds on the moment generating functions, give rise to different decay behaviors associated to the function $(\psi^\star)^{-1}$.

6.7 Sub-Gaussian Random Variables

An important class of random variables is characterized by the bound $\mathbf{E}\, e^{\lambda(X-\mathbf{E}X)} \le e^{a\lambda^2}$ for any $\lambda \in \mathbb{R}$ (hence, yielding both upper and lower tail bounds) for a given $\alpha > 0$. This class is called *sub-Gaussian*. Recall that for a Gaussian distribution $X \sim \mathcal{N}(\mu, \sigma^2)$ we have $\mathbf{E}\, e^{\lambda(X-\mathbf{E}X)} = e^{\sigma^2\lambda^2/2}$ for every $\lambda \in \mathbb{R}$.

Definition 6.5 (Sub-Gaussian random variable) *A random variable X is* sub-Gaussian *if*

$$\boxed{\mathbf{E}\, e^{\lambda(X-\mathbf{E}X)} \le e^{\sigma^2\lambda^2/2} \qquad \text{for any } \lambda \in \mathbb{R}}$$

for a given constant $\sigma^2 > 0$ called variance proxy.

Sub-Gaussian random variables have a moment generating function that is uniformly bounded above by the moment generating function of a Gaussian random variable. The class of sub-Gaussian random variables extends beyond Gaussian random variables. In fact, it is quite large. In particular, any bounded random variable is sub-Gaussian with variance proxy depending on the size of its support. We have already seen this. By Hoeffding's lemma, Lemma 2.1, if $a \le X \le b$ then $\mathbf{E}\, e^{\lambda X} \le e^{\lambda^2(b-a)^2/8}$ for any $\lambda \ge 0$. So a random variable bounded in $[a,b]$ is sub-Gaussian with variance proxy $\sigma^2 = \frac{(b-a)^2}{4}$.

In the case of sub-Gaussian random variables, Proposition 6.3 yields the following result.

Proposition 6.6 (Sub-Gaussian upper tail bound) *Let X be sub-Gaussian with variance proxy σ^2. Then, for any $\varepsilon \geq 0$ and $\delta \in [0,1]$ we have*

$$\boxed{\mathbf{P}(X - \mathbf{E}X \geq \varepsilon) \leq e^{-\varepsilon^2/(2\sigma^2)}} \qquad \boxed{\mathbf{P}\left(X - \mathbf{E}X < \sqrt{2\sigma^2 \log(1/\delta)}\right) \geq 1 - \delta}$$

Sub-Gaussian random variables can be equivalently characterized from their moment generating functions and from their tail bounds, up to constants.

Proposition 6.7 *Let X be a random variable such that for any $\varepsilon \geq 0$ we have*

$$\mathbf{P}(|X - \mathbf{E}X| \geq \varepsilon) \leq 2e^{-\varepsilon^2/(2\sigma^2)}.$$

Then, for any $\lambda \in \mathbb{R}$ we have

$$\mathbf{E}\, e^{\lambda(X - \mathbf{E}X)} \leq e^{4\sigma^2\lambda^2}.$$

Proof: See **Problem 2.9** in the Problem Sheets. ■

6.8 Hoeffding's Inequality

The property of sub-Gaussianity is preserved by linear operations with independent random variables: if $X_1, \ldots, X_n$ are independent sub-Gaussian random variables with variance proxy $\sigma_1^2, \ldots, \sigma_n^2$, then $\gamma_1 X_1 + \cdots + \gamma_n X_n$ is sub-Gaussian with variance proxy $\gamma_1^2\sigma_1^2 + \cdots + \gamma_n^2\sigma_n^2$. This fact immediately yields the following result, which is a corollary of Lemma 6.4 in the case of sub-Gaussian random variables.

Corollary 6.8 (Hoeffding's Inequality) *Let $X_1, \ldots, X_n \sim X$ be i.i.d. sub-Gaussian random variables with variance proxy σ^2. Then, for any $\varepsilon \geq 0$ and $\delta \in [0,1]$ we have*

$$\boxed{\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mathbf{E}X \geq \varepsilon\right) \leq e^{-n\varepsilon^2/(2\sigma^2)}} \qquad \boxed{\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mathbf{E}X < \sqrt{\frac{2\sigma^2 \log(1/\delta)}{n}}\right) \geq 1 - \delta}$$

Corollary 6.8 takes the name of *Hoeffding's inequality* in the case of bounded random variables, i.e., $a \leq X_i \leq b$, which leads to the bound in the corollary with $\sigma^2 = \frac{(b-a)^2}{4}$ by Hoeffding's Lemma, Lemma 2.1.

6.9 Hoeffding's Inequality: Back to Learning Part I

Hoeffding's Inequality can be used to derive a bound in probability for the excess risk $r(A^\star) - r(a^\star)$ obtained by the Empirical Risk Minimization rule $A^\star$ when the loss function ℓ is bounded and $|\mathcal{A}| < \infty$.

Proposition 6.9 *Assume that the loss function ℓ is bounded in the interval $[0, c]$. Then, for any $\delta \in [0,1]$,*

$$\boxed{\mathbf{P}\left(r(A^\star) - r(a^\star) < c\sqrt{\frac{2\log(2|\mathcal{A}|/\delta)}{n}}\right) \geq 1 - \delta}$$

Proof: Recall the following upper bound from Lecture 1:

$$r(A^\star) - r(a^\star) \leq \underbrace{\sup_{a\in\mathcal{A}}\{r(a) - R(a)\} + \sup_{a\in\mathcal{A}}\{R(a) - r(a)\}}_{\texttt{Statistics}}.$$

For any $a \in \mathcal{A}$, Hoeffding's inequality (Corollary 6.8) yields, for $\varepsilon \geq 0$,

$$\mathbf{P}(R(a) - r(a) \geq \varepsilon) \leq e^{-2n\varepsilon^2/c^2},$$

so that, by applying a union bound,

$$\begin{aligned}\mathbf{P}\left(\sup_{a\in\mathcal{A}}\{R(a) - r(a)\} \geq \varepsilon\right) &= \mathbf{P}\left(\left\{\text{There exists } a \in \mathcal{A} \text{ such that } R(a) - r(a) \geq \varepsilon\right\}\right) = \mathbf{P}\left(\bigcup_{a\in\mathcal{A}}\left\{R(a) - r(a) \geq \varepsilon\right\}\right)\\ &\leq \sum_{a\in\mathcal{A}} \mathbf{P}(R(a) - r(a) \geq \varepsilon) \leq |\mathcal{A}|e^{-2n\varepsilon^2/c^2}.\end{aligned}$$

Analogously, we find $\mathbf{P}(\sup_{a\in\mathcal{A}}\{r(a) - R(a)\} \geq \varepsilon) \leq |\mathcal{A}|e^{-2n\varepsilon^2/c^2}$. Using that if $\varepsilon \geq 0$, then

$$\{\alpha + \beta \geq \varepsilon\} \subseteq \Big(\{\alpha \geq \varepsilon/2\} \cup \{\beta \geq \varepsilon/2\}\Big),$$

we find, by the union bound,

$$\begin{aligned}\mathbf{P}(r(A^\star) - r(a^\star) \geq \varepsilon) &\leq \mathbf{P}\left(\sup_{a\in\mathcal{A}}\{r(a) - R(a)\} + \sup_{a\in\mathcal{A}}\{R(a) - r(a)\} \geq \varepsilon\right)\\ &\leq \mathbf{P}\left(\sup_{a\in\mathcal{A}}\{r(a) - R(a)\} \geq \frac{\varepsilon}{2}\right) + \mathbf{P}\left(\sup_{a\in\mathcal{A}}\{R(a) - r(a)\} \geq \frac{\varepsilon}{2}\right) \leq 2|\mathcal{A}|e^{-n\varepsilon^2/(2c^2)}.\end{aligned}$$

Equivalently, for any $\delta \in [0, 1]$,

$$\mathbf{P}\left(r(A^\star) - r(a^\star) < c\sqrt{\frac{2\log(2|\mathcal{A}|/\delta)}{n}}\right) \geq 1 - \delta.$$

■

The bound that we just obtained is only meaningful in the case $|\mathcal{A}| < \infty$. When $|\mathcal{A}| = \infty$, the bound yields $\mathbf{P}(r(A^\star) - r(a^\star) < \infty) \geq 1 - \delta$, which is always true as $\mathbf{P}(r(A^\star) - r(a^\star) \leq 2) = 1$ (it follows as a consequence of the fact that we assume the loss function ℓ to be bounded in the interval $[0, 1]$).

To address the case $|\mathcal{A}| = \infty$ we need to develop concentration inequalities that hold for more general functions of random variables, not just for their average. In particular, we are interested in concentration inequalities for the function $f : \mathcal{Z}^n \to \mathbb{R}$ defined as

$$f(Z_1, \ldots, Z_n) = \sup_{a\in\mathcal{A}}\{R(a) - r(a)\}.$$

6.10 Azuma's Lemma

Lemma 6.4 (and Hoeffding's Inequality in the particular case of bounded random variables) yields a concentration inequality for the average of random variables $\frac{1}{n}\sum_{i=1}^n X_i$. We now develop a tool that holds for a more general function of random variables $f(X_1, \ldots, X_n)$. The approach that we are going to describe is called the *martingale method*. As this course does not assume prior exposure to measure theory and

martingales, we present a simplified exposition of the main ideas. Full details of the general theory can be found in [1].

The main idea behind the martingale method is based on the following decomposition of the quantity of interest into a sum of so-called *martingale increments*:

$$f(X_1, \ldots, X_n) - \mathbf{E}f(X_1, \ldots, X_n) = \sum_{i=1}^{n} \Delta_i,$$

where

$$\Delta_i := \mathbf{E}[f(X_1, \ldots, X_n)|X_1, \ldots, X_i] - \mathbf{E}[f(X_1, \ldots, X_n)|X_1, \ldots, X_{i-1}],$$

with the convention $\Delta_1 := \mathbf{E}[f(X_1, \ldots, X_n)|X_1] - \mathbf{E}[f(X_1, \ldots, X_n)]$.

The following result shows that if the martingale increments are *conditionally* sub-Gaussian with certain variance proxies, then the sum is also sub-Gaussian with variance proxy given by the sum of the variance proxies.

Lemma 6.10 (Azuma) *Let $\mathbf{E}[e^{\lambda\Delta_i}|X_1, \ldots, X_{i-1}] \leq e^{\lambda^2\sigma_i^2/2}$ for each $i \in [n]$. Then, the sum $\sum_{i=1}^n \Delta_i$ is sub-Gaussian with variance proxy $\sum_{i=1}^n \sigma_i^2$.*

Proof: By the tower property and the "take out what is known" property of conditional expectations, we get, for every $k \in [n]$,

$$\mathbf{E}e^{\lambda \sum_{i=1}^k \Delta_i} = \mathbf{E}\mathbf{E}[e^{\lambda \sum_{i=1}^k \Delta_i}|X_1, \ldots, X_{k-1}] = \mathbf{E}e^{\lambda \sum_{i=1}^{k-1} \Delta_i}\mathbf{E}[e^{\lambda\Delta_k}|X_1, \ldots, X_{k-1}] \leq e^{\lambda^2\sigma_k^2/2}\mathbf{E}e^{\lambda \sum_{i=1}^{k-1} \Delta_i},$$

and the proof follows by induction. ■

6.11 McDiarmid's Inequality

Azuma's Lemma immediately yields the following result, which embodies the concentration phenomenon presented in Section 6.2 when the notion of "closeness" is captured by the upper tail of the distribution $Z, \widetilde{Z} \to \mathbf{P}(Z - \widetilde{Z} \geq \varepsilon)$ and the notion of "sensitivity" to changes in the coordinates is captured by the discrete derivatives defined as follows, for any $i \in [n]$ and any $x = (x_1, \ldots, x_n)$:

$$\delta_i f(x) := \sup_z f(x_1, \ldots, x_{i-1}, z, x_{i+1}, \ldots, x_n) - \inf_z f(x_1, \ldots, x_{i-1}, z, x_{i+1}, \ldots, x_n).$$

In the following, we use the notation $\|\delta_i f\|_\infty := \sup_x |\delta_i f(x)|$.

Theorem 6.11 (McDiarmid) *Let $X_1, \ldots, X_n$ be independent random variables. The random variable $f(X_1, \ldots, X_n)$ is sub-Gaussian with variance proxy $\frac{1}{4}\sum_{i=1}^n \|\delta_i f\|_\infty^2$. In particular,*

$$\boxed{\mathbf{P}(f(X_1, \ldots, X_n) - \mathbf{E}f(X_1, \ldots, X_n) \geq \varepsilon) \leq e^{-2\varepsilon^2/\sum_{i=1}^n \|\delta_i f\|_\infty^2}}$$

Proof: Note that we have $A_i \leq \Delta_i \leq B_i$, with

$$B_i := \mathbf{E}\left[\sup_z f(X_1, \ldots, X_{i-1}, z, X_{i+1}, \ldots, X_n) - f(X_1, \ldots, X_n)\middle| X_1, \ldots, X_{i-1}\right],$$

$$A_i := \mathbf{E}\left[\inf_z f(X_1, \ldots, X_{i-1}, z, X_{i+1}, \ldots, X_n) - f(X_1, \ldots, X_n)\middle| X_1, \ldots, X_{i-1}\right],$$

where we used the independence of X_i and $X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n$. Applying the Hoeffding's Lemma 2.1 conditionally on $X_1, \dots, X_{i-1}$, we get (note that $\mathbf{E}\Delta_i = 0$)

$$\mathbf{E}[e^{\lambda \Delta_i} | X_1, \dots, X_{i-1}] \leq e^{\lambda^2 \sigma_i^2 / 2},$$

with $\sigma_i^2 = \frac{(B_i - A_i)^2}{4}$. Hence, we can apply Azuma's Lemma 6.10 and get that the quantity $f(X_1, \dots, X_n) - \mathbf{E}f(X_1, \dots, X_n) = \sum_{i=1}^n \Delta_i$ is sub-Gaussian with variance proxy given by $\sum_{i=1}^n \sigma_i^2 = \frac{1}{4}\sum_{i=1}^n (B_i - A_i)^2 \leq \frac{1}{4}\sum_{i=1}^n \|\delta_i f\|_\infty^2$. The proof follows by Proposition 6.6. ■

Remark 6.12 (Improved concentration inequalities) *McDiarmid's inequality represents a first instance of concentration inequalities for generic functions f, making explicit the concentration phenomenon illustrated in Section 6.2. Using techniques more sophisticated than the martingale method, it is possible to establish much more refined results, for instance, improving the variance proxy from depending on $\sum_{i=1}^n \|\delta_i f\|_\infty^2$ to depending on $\| \sum_{i=1}^n |\delta_i f|^2 \|_\infty$ (this result is achieved by the so-called Bounded Difference Inequality). For the applications that we are going to consider in learning theory, these refinements are not needed, as already attested by the proof of the next theorem. For an in-depth exposition of concentration inequalities, along with the main ideas that lead to their development, we refer to [1].*

6.12 McDiarmid's Inequality: Back to Learning Part II

McDiarmid's inequality allows us to improve the shortcomings of Proposition 6.9 when $|\mathcal{A}| = \infty$. The following result can be combined with the bounds on the expected Rademacher complexity $\mathbf{E}\,\mathtt{Rad}(\mathcal{L} \circ \{Z_1, \dots, Z_n\})$ developed in the previous lectures. We recall the notation

$$\mathcal{L} \circ \{Z_1, \dots, Z_n\} := \{(\ell(a, Z_1), \dots, \ell(a, Z_n)) \in \mathbb{R}^n : a \in \mathcal{A}\}.$$

Theorem 6.13 *Assume that the loss function ℓ is bounded in the interval $[0, c]$. Then, for any $\delta \in [0, 1]$,*

$$\boxed{\mathbf{P}\left(r(A^\star) - r(a^\star) < \mathbf{E}\underbrace{\left[\sup_{a \in \mathcal{A}} \{r(a) - R(a)\} + \sup_{a \in \mathcal{A}} \{R(a) - r(a)\} \right]}_{\mathtt{Statistics}} + c\sqrt{2\frac{\log(1/\delta)}{n}} \right) \geq 1 - \delta}$$

In particular,

$$\boxed{\mathbf{P}\left(r(A^\star) - r(a^\star) < 4\,\mathbf{E}\,\mathtt{Rad}(\mathcal{L} \circ \{Z_1, \dots, Z_n\}) + c\sqrt{2\frac{\log(1/\delta)}{n}} \right) \geq 1 - \delta}$$

Proof: Recall the following upper bound from Lecture 1:

$$r(A^\star) - r(a^\star) \leq \underbrace{\sup_{a \in \mathcal{A}} \{r(a) - R(a)\} + \sup_{a \in \mathcal{A}} \{R(a) - r(a)\}}_{\mathtt{Statistics}}.$$

As $R(a) = \frac{1}{n}\sum_{i=1}^n \ell(a, Z_i)$, define the function

$$z = (z_1, \dots, z_n) \longrightarrow f(z) = \sup_{a \in \mathcal{A}} \left[r(a) - \frac{1}{n}\sum_{i=1}^n \ell(a, z_i) \right] + \sup_{a \in \mathcal{A}} \left[\frac{1}{n}\sum_{i=1}^n \ell(a, z_i) - r(a) \right].$$

For each $k \in [n]$ define $g_k(a,z) = r(a) - \frac{1}{n}\sum_{i\in[n]\setminus\{k\}} \ell(a,z_i)$. Then,

$$\begin{aligned}\delta_k f(z) &:= \sup_u f(z_1,\dots,z_{k-1},u,z_{k+1},\dots,z_n) - \inf_u f(z_1,\dots,z_{k-1},u,z_{k+1},\dots,z_n)\\ &= \sup_u \left\{ \sup_{a\in\mathcal{A}} \left[g_k(a,z) - \frac{\ell(a,u)}{n}\right] + \sup_{a\in\mathcal{A}} \left[-g_k(a,z) + \frac{\ell(a,u)}{n}\right]\right\}\\ &\quad - \inf_u \left\{ \sup_{a\in\mathcal{A}} \left[g_k(a,z) - \frac{\ell(a,u)}{n}\right] + \sup_{a\in\mathcal{A}} \left[-g_k(a,z) + \frac{\ell(a,u)}{n}\right]\right\}.\end{aligned}$$

Using that $0 \le \ell(a,u) \le c$, the above yields $\delta_k f(z) \le 2c/n$, which implies $\|\delta_k f\|_\infty = \sup_z |\delta_k f(z)| \le 2c/n$. By McDiarmid's Theorem 6.11, the random variable

$$f(Z_1,\dots,Z_n) = \sup_{a\in\mathcal{A}}\{r(a) - R(a)\} + \sup_{a\in\mathcal{A}}\{R(a) - r(a)\}$$

is sub-Gaussian with variance proxy $\frac{1}{4}\sum_{i=1}^n \|\delta_i f\|_\infty^2 = c^2/n$, and

$$\mathbf{P}\Big(f(Z_1,\dots,Z_n) - \mathbf{E} f(Z_1,\dots,Z_n) \ge \varepsilon \Big) \le e^{-n\varepsilon^2/(2c^2)}.$$

The results follow by setting $e^{-n\varepsilon^2/(2c^2)} = \delta$ and by using Proposition 2.11. ■

Remark 6.14 *Note that the proof of Theorem 6.13 deals with the* `Statistics` *term at once, without treating the two addends* $\sup_{a\in\mathcal{A}}\{r(a) - R(a)\}$ *and* $\sup_{a\in\mathcal{A}}\{R(a) - r(a)\}$ *separately, as done in the analysis provided in Section 1.2.1 in Lecture 1. Should we have followed the analysis in Lecture 1, we would have ended up with the choices*

$$\begin{aligned}\boxed{\texttt{UpperTailStats}(\varepsilon)} &= 2e^{-2n\varepsilon^2/c^2},\\ \boxed{\texttt{UpperTailStats}^{-1}(\delta)} &= c\sqrt{\frac{\log(2/\delta)}{2n}},\\ \boxed{\texttt{ExpectationStats}} &= 4\,\mathbf{E}\,\texttt{Rad}(\mathcal{L}\circ\{Z_1,\dots,Z_n\}),\end{aligned}$$

and we would have obtained the same bound as in the statement of Theorem 6.13 with the term $\log(1/\delta)$ *replaced by* $\log(2/\delta)$*, which is worse.*

A statement analogous to the one in Theorem 6.13 can be obtained by replacing the (deterministic) quantity $\mathbf{E}\,\texttt{Rad}(\mathcal{L}\circ\{Z_1,\dots,Z_n\})$ with the (random) quantity $\texttt{Rad}(\mathcal{L}\circ\{Z_1,\dots,Z_n\})$, a.k.a. the *empirical* Rademacher complexity. See also Remark 2.12. See **Problem 2.8** in the Problem Sheets.

References

[1] Ramon van Handel. Probability in high dimension. Technical report, PRINCETON UNIV NJ, 2014.