Algorithmic Foundations of Learning

Lecture 6
Sub-Gaussian Concentration Inequalities
Bounds in Probability

Patrick Rebeschini

Department of Statistics
University of Oxford

From Bounds in Expectations to Bounds in Probability

Recall: $\mathcal{L} \circ \{Z_1, \ldots, Z_n\} = \{(\ell(a, Z_1), \ldots, \ell(a, Z_n)) : a \in \mathcal{A}\}$

- **Bounds in expectation:**

$$\mathbf{E}\, r(A^\star) - r(a^\star) \leq 4\, \mathbf{E}\, \texttt{Rad}(\mathcal{L} \circ \{Z_1, \ldots, Z_n\}) \leq \begin{cases} \bullet \text{ regression} \\ \quad \text{(lecture 3)} \\ \bullet \text{ classification (VC dim.)} \\ \quad \text{(lecture 4)} \\ \bullet \text{ covering num., chaining} \\ \quad \text{(lecture 5)} \end{cases}$$

(lecture 2)

- **Bounds in probability:** (lecture 6!)

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

Can use bounds for $\mathbf{E}\, r(A^\star) - r(a^\star)$ and still get probability $\geq 1 - \delta$

Concentration inequalities

Concentration phenomenon

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

- Already seen manifestation **(Problem 1.1)**: if $X_1, \ldots, X_n$ are i.i.d. mean μ:

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

  E.g., **variance** ($p = 2$) captures how close random variable is to its mean

  These notions of "closeness" capture **size** of fluctuations

- We need notion of "closeness" that captures **distribution** of fluctuations:

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

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

Markov's Inequality and Chernoff's bounds

Markov's inequality is the main result to prove tail inequalities

Markov's Inequality (Proposition 6.1)

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: $X = X1_{X \geq \varepsilon} + X1_{X < \varepsilon} \geq \varepsilon 1_{X \geq \varepsilon}$, where we used that $X \geq 0$

Chernoff's Bound (Proposition 6.2)

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: Exponentiate and apply Markov's inequality: $\mathbf{P}(X \geq \varepsilon) = \mathbf{P}(e^{\lambda X} \geq e^{\lambda\varepsilon}) \leq \frac{\mathbf{E}\,e^{\lambda X}}{e^{\lambda\varepsilon}}$

Concentration Inequality for Sums of i.i.d. Variables

Let $\psi^\star(\varepsilon) := \sup_{\lambda \geq 0}(\lambda\varepsilon - \psi(\lambda))$ be the **convex conjugate** of $\psi : \mathbb{R}_+ \to \mathbb{R}$.

Optimal Chernoff's Bound: Convex Conjugate (Proposition 6.3)

Let $\mathbf{E}\, e^{\lambda(X - \mathbf{E}X)} \leq e^{\psi(\lambda)}$ for any $\lambda \geq 0$. Then,

$$\mathbf{P}(X - \mathbf{E}X \geq \varepsilon) \leq e^{-\psi^\star(\varepsilon)}$$

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

Concentration Inequality for Sums of i.i.d. Variables (Lemma 6.4)

Let $X_1, \ldots, X_n \sim X$ be i.i.d. with $\mathbf{E}\, e^{\lambda(X - \mathbf{E}X)} \leq e^{\psi(\lambda)}$ for any $\lambda \geq 0$. Then,

$$\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbf{E}X \geq \varepsilon\right) \leq e^{-n\psi^\star(\varepsilon)}$$

$$\mathbf{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbf{E}X < (\psi^\star)^{-1}\left(\frac{\log(1/\delta)}{n}\right)\right) \geq 1 - \delta$$

Sub-Gaussian Random Variables

Sub-Guassian (Definition 6.5)

A random variable X is *sub-Gaussian* if for every $\lambda \in \mathbb{R}$ we have

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

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

- **Gaussian**: if $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\mathbf{E}\, e^{\lambda(X - \mathbf{E}X)} = e^{\sigma^2 \lambda^2 / 2}$
- **Bounded r.v.'s**: if $a \leq X \leq b$ then (by Hoeffding's Lemma 2.1)

$$\mathbf{E}\, e^{\lambda(X - \mathbf{E}X)} \leq e^{\lambda^2 (b-a)^2 / 8} \quad \Longrightarrow \quad \sigma^2 = \frac{(b-a)^2}{4}$$

(Proposition 6.6)

Let X be sub-Gaussian with variance proxy σ^2. Then,

$$\boxed{\mathbf{P}(X - \mathbf{E}X > \varepsilon) \leq e^{-\varepsilon^2 / (2\sigma^2)}}$$

Tail bound equivalent to bound on moment generating function **(Problem 2.9)**

Hoeffding's Inequality: Application to Learning Part I

Hoeffding's Inequality (Corollary 6.8)

Let $X_1, \ldots, X_n \sim X$ be i.i.d. sub-Gaussian random variables with variance proxy σ^2. Then, for any $n \in \mathbb{N}_+$ and any $\varepsilon \geq 0$ we have

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

Proof: $\frac{1}{n}\sum_{i=1}^{n} X_i$ is sub-Gaussian with variance proxy σ^2/n

Application to Learning (Proposition 6.9)

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

Proof: Union bound $\mathbf{P}(\sup_{a\in\mathcal{A}}\{R(a) - r(a)\} \geq \varepsilon) \leq \sum_{a\in\mathcal{A}} \mathbf{P}(R(a) - r(a) \geq \varepsilon) \leq |\mathcal{A}| e^{-2n\varepsilon^2/c^2}$

Bound is trivial for $|\mathcal{A}| = \infty$. We need to develop more sophisticated tools...

Azuma's Lemma

Martingale method:

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

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

Azuma (Lemma 6.10)

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: For every $k \in [n]$, by the tower property and the "take out what is known" property:

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

The proof follows by induction

McDiarmid's Inequality

Notion of "sensitivity" to changes in the coordinates: **discrete derivatives**

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

McDiarmid (Theorem 6.11)

Let $X_1, \ldots, X_n$ be independent. Then, $f(X_1, \ldots, X_n)$ is sub-Gaussian with variance proxy $\frac{1}{4}\sum_{i=1}^n \|\delta_i f\|_\infty^2$ and

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

Proof: We have $A_i \leq \Delta_i \leq B_i$, with

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

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

Apply Hoeffding's Lemma conditionally on $X_1, \ldots, X_{i-1}$ (note that $\mathbf{E}\Delta_i = 0$)

$$\mathbf{E}[e^{\lambda \Delta_i} | X_1, \ldots, X_{i-1}] \leq e^{\lambda^2 \sigma_i^2 / 2} \qquad \text{with } \sigma_i^2 = \frac{(B_i - A_i)^2}{4}$$

Proof follow by Azuma's Lemma

McDiarmid's Inequality: Application to Learning Part II

(Theorem 6.13)

Assume that the loss function ℓ is bounded in the interval $[0, c]$. Then,

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

Proof: Define

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

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,

$$\begin{aligned}\delta_k f(z) = \sup_u \left\{ \sup_{a \in \mathcal{A}} \left[g_k(a, z) - \frac{\ell(a, u)}{n}\right] + \sup_{a \in \mathcal{A}} \left[-g_k(a, z) + \frac{\ell(a, u)}{n}\right]\right\} \\ - \inf_u \left\{ \sup_{a \in \mathcal{A}} \left[g_k(a, z) - \frac{\ell(a, u)}{n}\right] + \sup_{a \in \mathcal{A}} \left[-g_k(a, z) + \frac{\ell(a, u)}{n}\right]\right\}.\end{aligned}$$

Using $0 \leq \ell(a, u) \leq c$, the above yields $\delta_k f(z) \leq \frac{2c}{n}$. Proof follows by McDiarmid's Theorem