Learning at the Turing

The Alan Turing Institute

Mathematics of Machine Learning - Summer School

Lecture 2
Concentration Inequalities. Bounds in Probability

June 28, 2021

Patrick Rebeschini
Department of Statistics, University of Oxford

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

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

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

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 $a \le X \le b$ then (by Hoeffding's Lemma 2.1)

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

(Proposition 6.6)

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

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

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

Hoeffding's Lemma (Lemma 2.1)

Let X be a bounded random variable $a \leq X - \mathbf{E}X \leq b$. Then, for any $\lambda \in \mathbb{R}$,

$$\boxed{\mathbf{E}\, e^{\lambda(X-\mathbf{E}X)} \leq e^{\lambda^2(b-a)^2/8}}$$

Proof

- W.l.o.g., take $\mathbf{E}X = 0$. Let $\psi(\lambda) = \log \mathbf{E}\, e^{\lambda X}$

$$\psi'(\lambda) = \frac{\mathbf{E}[Xe^{\lambda X}]}{\mathbf{E}\, e^{\lambda X}} \qquad \psi''(\lambda) = \frac{\mathbf{E}[X^2 e^{\lambda X}]}{\mathbf{E}\, e^{\lambda X}} - \left(\frac{\mathbf{E}[Xe^{\lambda X}]}{\mathbf{E}\, e^{\lambda X}}\right)^2$$

- $\psi''(\lambda)$ is the variance of X under the distribution $\mathbf{Q}(\mathrm{d}x) = \frac{e^{\lambda x}}{\mathbf{E}\, e^{\lambda X}}\mathbf{P}(\mathrm{d}x)$

- $\psi''(\lambda) = \mathbf{Var}_{\mathbf{Q}}\left(X - \frac{a+b}{2}\right) \leq \mathbf{E}_{\mathbf{Q}}\left[\left(X - \frac{a+b}{2}\right)^2\right] \leq \frac{(b-a)^2}{4}$

- Fundamental Thm of Calculus: $\psi(\lambda) = \int_0^\lambda \int_0^\mu \psi''(\rho)\mathrm{d}\rho \mathrm{d}\mu \leq \frac{\lambda^2(b-a)^2}{8}$

□

Hoeffding's Inequality: Application to Learning Part I

Hoeffding's Inequality (Corollary 6.8)

Let $X_1, \dots, X_n \sim X$ 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:

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

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

Azuma (Lemma 6.10)

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

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

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

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

Proof: 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 $\mathbf{E}\Delta_i = 0$)

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

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,

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

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