Machine Learning for EDS

TUTORIAL WEEK 1

2023/2024

Problem 1

1. **(Markov inequality)** Let X be a real-valued non-negative random variable. Show that:

$$\forall a > 0, \quad \mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

2. **(Corollary for sub-Gaussian random variables)** A real-valued random variable X is said to be b-sub-Gaussian, $b > 0$, if for any $s \in \mathbb{R}$ it holds that: $\mathbb{E}[\exp(sX)] \leq \exp\left(\frac{s^2 b^2}{2}\right)$. It can be shown that any b-sub-Gaussian random variable X has $\mathbb{E}[X] = 0$ and $\mathbb{V}\mathrm{ar}(X) \leq b^2$. Assume that X is b-sub-Gaussian and let $a > 0$.

   (a) First show that

$$\forall s > 0, \qquad \mathbb{P}(X \geq a) \leq e^{\frac{s^2 b^2}{2} - sa}.$$

   (b) Deduce that

$$\mathbb{P}(X \geq a) \leq \exp\left(-\frac{a^2}{2b^2}\right).$$

   Hint: Notice that the left-hand side in (a) does not depend on s.

Problem 2

1. **(Sum of independent random variables)** Let $X_1, \ldots, X_n$ be n real-valued random variables. Assume that for any $i \leq n$, X_i is b_i-sub-Gaussian for some positive constant b_i.
   Let $S_n = X_1 + \ldots + X_n$.

   (a) Show that S_n is b-sub-Gaussian for some explicit positive constant b expressed in terms of $b_1, \ldots, b_n$.

   (b) Deduce the concentration inequalities

$$\forall a > 0, \quad \mathbb{P}(S_n \geq a) \leq \exp\left(-\frac{a^2}{2\sum_{i=1}^n b_i^2}\right).$$

$$\forall a > 0, \quad \mathbb{P}(S_n \leq -a) \leq \exp\left(-\frac{a^2}{2\sum_{i=1}^n b_i^2}\right).$$

   Remark: A concentration inequality is simply a bound on the probability that a random variable is below or above a certain value. For example, the Markov inequality proved in problem 1.1 is another example of a concentration inequality.

(c) Deduce a concentration inequality for $|S_n|$.

Hint: Notice that $\{|S_n| \geq a\}$ *and* $\{S_n \geq a\} \cup \{S_n \leq -a\}$ *are the same event.*

2. **(Gaussian case)**

(a) Let X be a Gaussian random variable $X \sim \mathcal{N}(\mu, \sigma^2)$. Show that $X - \mathbb{E}[X]$ is b-sub-Gaussian for some explicit positive constant b expressed in terms of σ.

(b) Deduce a concentration inequality for a sum of n independent Gaussian random variables $S_n = \sum_{i=1}^n \Big(X_i - \mathbb{E}[X_i]\Big)$, where $X_i \overset{i.i.d.}{\sim} \mathcal{N}(\mu_i, \sigma_i^2)$.

Problem 3 (Hoeffding's inequality)

Let X be a bounded random variable such that $X \in [c, d]$, for some constants $c < d$, and $\mathbb{E}[X] = 0$. The goal is to show that X is b-sub-Gaussian for some explicit parameter b, and deduce concentration inequalities for sum of independent bounded random variables.

1. First show that for any $x \in [c, d]$ and $t \in \mathbb{R}$,

$$e^{tx} \leq \frac{d-x}{d-c} e^{tc} + \frac{x-c}{d-c} e^{td}.$$

Hint: Use that $x = \frac{d}{d-c}x - \frac{c}{d-c}x = \frac{d-x}{d-c}c + \frac{x-c}{d-c}d$ and use the convexity of the function $f(x) = e^x$.

2. Deduce that

$$\mathbb{E}[e^{tX}] \leq \frac{d}{d-c} e^{tc} + \frac{-c}{d-c} e^{td}.$$

3. Letting $h := t(d-c)$, $p := \dfrac{-c}{d-c}$ and $L : h \mapsto L(h) := -hp + \ln(1 - p + pe^h)$, verify that

$$e^{L(h)} = \frac{d}{d-c} e^{tc} + \frac{-c}{d-c} e^{td}.$$

which implies that $\mathbb{E}[e^{tX}] \leq e^{L(h)}$.

4. Considering the function $L : h \mapsto L(h)$, show that $L(0) = L'(0) = 0$, and that for any h, $L''(h) \leq 1/4$.

5. Using a second order Taylor approximation, deduce that for any h:

$$L(h) \leq \frac{h^2}{8}.$$

Hint: in particular, use that for any twice differentiable function f, and any values x and a on the domain of f, we know there exists a value a^* between x and a, such that $f(x) = f(a) + f'(a)(x - a) + \frac{f''(a^*)}{2!}(x-a)^2$.

6. Conclude that X is b-sub-Gaussian for some explicit positive constant b expressed in terms of c and d.

7. How is the previous result modified if $\mathbb{E}[X] \neq 0$?

8. Letting $X_1, \ldots, X_n$ be n independent bounded random variables such that for any i, $X_i \in [c_i, d_i]$, deduce concentration inequalities for $S_n = \sum_{i=1}^n (X_i - \mathbb{E}[X_i])$ and $|S_n|$.

Problem 4 (Expectation of maximum of sub-Gaussian random variables)

The goal of this problem is to establish the following lemma.

Lemma: Let $Z_1, \ldots, Z_K$ be v-sub-Gaussian random variables for some parameter $v > 0$. Then,

$$\mathbb{E}\Big[\max_{1 \leq k \leq K} Z_k\Big] \leq v\sqrt{2\ln(K)}.$$

1. Define $M := \max_{1 \leq k \leq K} Z_k$. Using Jensen's inequality prove that for any $s > 0$

$$e^{s\mathbb{E}[M]} \leq \sum_{k=1}^{K} \mathbb{E}\Big[e^{sZ_k}\Big].$$

2. Deduce that for any $s > 0$

$$e^{s\mathbb{E}[M]} \leq K e^{s^2 v^2/2}.$$

3. Show that

$$\mathbb{E}[M] \leq \inf_{s>0} \Big\{\frac{\ln(K)}{s} + \frac{sv^2}{2}\Big\}.$$

4. Finally obtain that

$$\mathbb{E}\Big[\max_{1 \leq k \leq K} Z_k\Big] \leq v\sqrt{2\ln(K)}.$$