Mathematics of Machine Learning
Homework 8 - Solutions

Instructors: Afonso S. Bandeira & Nikita Zhivotovskiy
Course Coordinator: Pedro Abdalla Teixeira

April 30, 2021

Try to solve the questions before looking to the answers. Every item must be proved rigorously. Starred problems are harder.

Problem 1

Let $g_1, \ldots, g_n$ denote a collection of standard Gaussian random variables, not necessarily independent. The goal of this exercise is to provide a simple bound for the expectation of the maxima of Gaussian random variables

(a) Prove that, for every $\lambda > 0$, we have

$$\mathbb{E} e^{\lambda \max_{i=1,\ldots,n} g_i} \leq n e^{\lambda^2/2}$$

(b) Prove that, for every $\lambda > 0$,

$$e^{\lambda \mathbb{E} \max_{i=1,\ldots,n} g_i} \leq \mathbb{E} e^{\lambda \max_{i=1,\ldots,n} g_i}$$

(Hint: Apply Jensen's inequality)

(c) Prove that

$$\mathbb{E} \max_{i=1,\dots,n} g_i \leq C\sqrt{\log n},$$

where $C > 0$ is an absolute constant. (Hint: Follows the steps above and optimize in λ)

Solution 1

(a) Notice that $\mathbb{E}e^{\lambda \max_{i=1,\dots,n} g_i} = \mathbb{E} \max_{i=1,\dots,n} e^{\lambda g_i} \leq \mathbb{E} \sum_{i=1}^{n} e^{g_i}$. The result now follows from linearity of expectation and the fact that $\mathbb{E}e^{\lambda g_i} = e^{\lambda^2/2}$ for any standard Gaussian random variable g_i.

(b) We apply Jensen's inequality to function $f(x) := e^{\lambda x}$ (clearly convex), so we get

$$e^{\lambda \mathbb{E} \max_{i=1,\dots,n} g_i} = f(\mathbb{E} \max_{i=1,\dots,n} g_i) \leq \mathbb{E} f(\max_{i=1,\dots,n} g_i) = \mathbb{E}e^{\lambda \max_{i=1,\dots,n} g_i}.$$

(c) By letters "a" and "b", we get $e^{\lambda \mathbb{E} \max_{i=1,\dots,n} g_i} \leq ne^{\lambda^2/2}$. Now we take the logarithm on both sides,

$$\mathbb{E} \max_{i=1,\dots,n} g_i \leq \frac{\log n}{\lambda} + \frac{\lambda}{2}.$$

By arithmetic-geometric mean inequality, the right hand side is minimized when $\frac{\log n}{\lambda} = \frac{\lambda}{2}$, i.e, $\lambda = \sqrt{2 \log n}$. We have just proved the result with absolute constant $C = \sqrt{2}$.

This bound is remarkable, it does not require independent and it is asymptotically sharp for i.i.d Gaussian random variables.

Problem 2

Let X be a bounded random variable taking values in the interval $[a.b]$. The goal of this exercise is to prove a sharper version of Hoeffiding's lemma.

(a) Prove that

$$\mathbb{E}e^{\lambda X} \leq \frac{b - \mathbb{E}X}{b - a} e^{\lambda a} + \frac{\mathbb{E}X - a}{b - a} e^{\lambda b}.$$

(Hint: Use the convexity of the function $e^{\lambda x}$ and take expectation of both sides.)

(b) Let $\theta := \frac{-a}{b-a}$, $u := \lambda(b-a)$ and $f(u) := -\theta u + \log(1 - \theta + \theta e^u)$. Check that

$$\mathbb{E}e^{\lambda(X - \mathbb{E}X)} \leq e^{f(u)}.$$

(c) Use Taylor expansion to bound $e^{f(u)}$ by $e^{\frac{\lambda^2(b-a)^2}{8}}$.

(d) Derive a sharper Hoeffiding lemma.

Solution 2

(a) By convexity of the function $e^{\lambda x}$ we have

$$e^{\lambda x} \leq \frac{b - x}{b - a} e^{\lambda a} + \frac{x - a}{b - a} e^{\lambda b}.$$

Taking expectation of both sides we get the result.

(b) Without loss of generality assume that X has zero mean. Then letter "a" implies that $\mathbb{E}e^{\lambda X} \leq (1-\theta)e^{\lambda a} + \theta e^{\lambda b}$, the latter can be written as $(1 - \theta + \theta e^{\lambda(b-a)})e^{-\lambda\theta(b-a)} = e^{f(u)}$.

(c) We can easily check that $f(0) = 0$, $f'(0) = 0$. For the second derivative, $f''(u) = \frac{\theta e^u}{1-\theta+\theta e^u}(1 - \frac{\theta e^u}{1-\theta+\theta e^u}) \leq \frac{1}{4}$. By one dimensional Taylor theorem, for every real u there exists x such

that

$$f(u) = f(0) + uf'(0) + \frac{1}{2}u^2 f''(x) \leq \frac{u^2}{8}.$$

Recall that $u = \lambda(b-a)$ and plug it into $e^{\frac{u^2}{8}}$.

(d) We can conclude that, for every bounded random variable X taking values in the interval $[a, b]$, $\mathbb{E}e^{\lambda(X-\mathbb{E}X)} \leq \mathbb{E}e^{\frac{\lambda^2(b-a)^2}{8}}$.

Problem 3

A random variable X is said to be symmetric if X and $-X$ have the same distribution.

(a) Check that the standard Gaussian random variable is symmetric. Give an example of a random variable that is not symmetric.

(b) Let ε denote a Rademacher random variable, i.e, takes values ± 1 with probability $\frac{1}{2}$ each. Prove that if X is a symmetric random variable independent on ε, then εX and X have the same distribution.

(c) Prove via conditional expectation that if X is a symmetric random variable then all odd moments vanish, i.e, $\mathbb{E}X^k = 0$ for all odd k.

Solution 3

(a) Let g be a standard Gaussian random variable. It means that the random variable has density $\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ that is symmetric around 0. It follows that

$$\mathbb{P}(-g \geq t) = \mathbb{P}(g \leq -t) = \int_{-\infty}^{-t} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx = \int_{t}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx.$$

The latter term is exactly equal to $\mathbb{P}(g \geq t)$. Since it is valid for all $t \in \mathbb{R}$, we can conclude that g and $-g$ have the same distribution. There are many distributions that are not symmetric, for example the Bernoulli distribution. Assume that X takes values 0 and 1 with probability $\frac{1}{2}$ each, then $\mathbb{P}(X > 0) = \frac{1}{2}$ but $\mathbb{P}(-X > 0) = 0$.

(b) We use conditional probability to write,

$$\mathbb{P}(\varepsilon X \geq t) = \mathbb{P}(X \geq t|\varepsilon = 1)\mathbb{P}(\varepsilon = 1)+\mathbb{P}(-X \geq t|\varepsilon = -1)\mathbb{P}(\varepsilon = -1).$$

By independence, $\mathbb{P}(X \geq t|\varepsilon = 1) = \mathbb{P}(X \geq t)$. Similarly for $\mathbb{P}(-X \geq t|\varepsilon = -1) = \mathbb{P}(-X \geq t)$. By definition of Rademacher random variable $\mathbb{P}(\varepsilon = 1) = \mathbb{P}(\varepsilon = -1) = \frac{1}{2}$. Finally, because X is a symmetric random variable,

$$\mathbb{P}(\varepsilon X \geq t) = \frac{1}{2}[\mathbb{P}(X \geq t) + \mathbb{P}(-X \geq t)] = \mathbb{P}(X \geq t).$$

We conclude that X and εX have the same distribution.

(c) By letter "b", we know that $\mathbb{E}X^k = \mathbb{E}(\varepsilon X)^k = \mathbb{E}_X X^k \mathbb{E}_\varepsilon \varepsilon^k$. The latter equality is due to independence between X and ε. Now, it is easy to check that $\mathbb{E}\varepsilon^k = 0$ for odd k. Indeed, for odd k,

$$\mathbb{E}\varepsilon^k = 1^k\frac{1}{2} + (-1)^k\frac{1}{2} = 0.$$

In case it is not clear, $\mathbb{E}_X$ means expectation with respect to the random variable X.