High-Dimensional Probability

Daniele Zago

November 20, 2021

Contents

Lecture 1: Concentration inequalities

2021-11-20

The object of the first lectures is trying to characterize deviations of sums of random variables X_i w.r. to their expected value $\mathbb{E}$. These *concentration inequalities* take for instance the form of

$$\mathbb{P}(|S-\mu|>t) \leq \text{Bound},$$

where the bound is tighter than what we usually obtain using the standard inequalities that are presented in a first course in probability. In particular, we are not looking for asymptotic results as in the central limit theorem, but rather for estimates which are valid for any sample size N.

1.1 Hoeffding's inequality

Let us begin by recalling two standard inequalities which are going to be especially useful in the following sections.

Thm. 1 (Markov's inequality)
Let $X \geq 0$ be a random variable with finite expected value, then

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t}, \quad \textit{for all } t>0.$$

A straightforward consequence of Markov's inequality can be obtained by replacing the random variable X with $|X-\mu|$ and squaring both sides inside the probability operator, which yields the following inequality.

Corollary 1 (Chebyshev's inequality)
If X is a random variable with finite variance, $\mathbb{V}[X]<\infty$, then

$$\mathbb{P}(|X-\mu| \geq t) \leq \frac{\mathbb{V}[X]}{t^2}.$$

Remark Many of the arguments that we make in this lecture will be based on the following trick: for any random variable X and for any $\lambda>0$,

$$\begin{aligned} \mathbb{P}(X-\mu \geq t) &= \mathbb{P}(e^{\lambda(X-\mu)} \leq e^{\lambda t}) && \text{(monotone)} \\ &\leq e^{-\lambda t}\mathbb{E}[e^{\lambda(X-\mu)}] && \text{(Markov)} \end{aligned}$$

Now, since it holds for any choice of $\lambda>0$ we can obtain the tightest bound by optimizing w.r. to λ,

$$\mathbb{P}(X-\mu \geq t) \leq \inf_{\lambda>0} e^{-\lambda t}\mathbb{E}[e^{\lambda(X-\mu)}],$$

and since X is usually a sum of random variables, its characteristic function can be decomposed into a product and evaluated quite easily.

Thm. 2 (Hoeffding's inequality)
Let $X_1, \dots, X_N$ be i.i.d Rademacher($\frac{1}{2}$) random variables and $a_1, \dots, a_N \in \mathbb{R}$, then for any $t > 0$ we have

$$\mathbb{P}\Big(\sum_{i=1}^N a_i X_i \geq t\Big) \leq \exp\left(-\frac{t^2}{2\|a\|_2^2}\right)$$

Sample size Unlike standard concentration inequalities based on the central limit theorem, this inequality gives an exact bound for any value of N.

Tightness Moreover, we can see that the tail behaviour, i.e. $\mathbb{P}(Y \geq t)$, is square-exponential in t, which means that this bound is extremely tight.

Proof.
Suppose that $\|a\|_2 = 1$, otherwise we can rescale t accordingly. For $\lambda > 0$, we have

$$\begin{aligned}
\mathbb{P}\Big(\sum_{i=1}^N a_i X_i \geq t\Big) &\overset{\text{Markov}}{\leq} e^{-\lambda t}\mathbb{E}[e^{\lambda \sum_{i=1}^N a_i X_i}] \\
&= e^{-\lambda t} \prod_{i=1}^N \underbrace{\mathbb{E}[e^{\lambda a_i X_i}]}_{\frac{1}{2}e^{\lambda a_i} + \frac{1}{2}e^{-\lambda a_i}} && \text{(Indep.)} \\
&= e^{-\lambda t} \prod_{i=1}^N \cosh(\lambda a_i) && (\tfrac{1}{2}e^x + \tfrac{1}{2}e^{-x} = \cosh(x)) \\
&\leq e^{-\lambda t} e^{\frac{\lambda^2}{2}\sum_{i=1}^N a_i^2} && (\cosh(x) \leq e^{\frac{x^2}{2}}, \text{ see here})
\end{aligned}$$

Now, if we want to find the optimal bound, $\lambda_{\text{opt}} = \inf_{\lambda > 0} e^{-\lambda t + \frac{\lambda^2}{2}\|a\|_2^2}$, we first notice that the function inside the exponent is parabolic in λ,

$$f(\lambda) = -\lambda t + \frac{\lambda^2}{2}\|a\|_2^2 \overset{\text{parabola}}{\Longrightarrow} \lambda_{\text{opt}} = \frac{t}{\|a\|_2^2} \implies f(\lambda_{\text{opt}}) = -\frac{t^2}{2\|a\|_2^2}.$$

Therefore, by substituting the optimal λ we obtain the proof of Hoeffding's inequality,

$$\mathbb{P}\Big(\sum_{i=1}^N a_i X_i \geq t\Big) \leq e^{-\frac{t^2}{2\|a\|_2^2}}.$$

$\square$

Exercise Restate Hoeffding's inequality for $X_1, \dots, X_N \overset{\text{iid}}{\sim} \text{Ber}(\frac{1}{2})$, using the fact that $Z_i = 2X_i - 1$ with $Z_i \sim \text{Rademacher}(\frac{1}{2})$.

Exercise Use Hoeffding's inequality for Bernoulli random variables to prove that by tossing a coin N times we have the exact bound

$$\mathbb{P}\Big(\text{at least } \frac{3}{4} \text{ heads}\Big) \leq e^{-N/8}.$$

Remark We can get a double bound from the above 2 by using $\mathbb{P}(|S| \geq t) \leq \mathbb{P}(S \geq t) + \mathbb{P}(-S \geq t)$, and observing that the Rademacher r.v. is symmetric $S = -S$. Therefore, both bounds are equal and the following two-sided inequality can be stated.

Thm. 3 (Two-sided Hoeffding's inequality)
Let $X_1, \ldots, X_N$ be i.i.d Rademacher r.v.'s, then for all $t \geq 0$ and for all $a \in \mathbb{R}^N$,

$$\mathbb{P}\Big(\Big|\sum_{i=1}^{N} a_i X_i\Big| \geq t\Big) \leq 2\exp\left(-\frac{t^2}{2\|a\|_2^2}\right).$$

We now turn to the more general problem of bounded random variables, which include as a special case the setting of Bernoulli r.v.'s with varying parameter p_i.

Thm. 4 (Hoeffding's inequality for bounded r.v.'s)
Let $X_1, X_2, \ldots, X_N$ be independent but not identically distributed r.v.'s, such that $X_i \in [m_i, M_i]$ and $\mathbb{E}[X_i] < \infty$. Then, for all $t \geq 0$ the following inequality holds,

$$\mathbb{P}\Big(\sum_{i=1}^{N}(X_i - \mathbb{E}[X_i]) \geq t\Big) \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{N}(M_i - m_i)^2}\right).$$

Proof.
(Exercise 2.2.7 in the book) The difficult part is achieving the constant 2 in the numerator, therefore we start with a different constant and then use a trick to get it. Let $\lambda > 0$, then by the same argument as before we can write

$$\begin{aligned}\mathbb{P}(\sum_{i=1}^{N}(X_i - \mathbb{E}[X_i]) \geq t) &\leq e^{-\lambda t}\mathbb{E}[e^{\lambda\sum_i X_i - \mathbb{E}[X_i]}] \\ &= e^{-\lambda t}\prod_i \mathbb{E}[e^{\lambda(X_i - \mathbb{E}[X_i])}] \\ &\leq e^{-\lambda t + \sum_i \lambda(M_i - m_i)}\end{aligned}$$

This is not as easy to optimize as before since we don't have a quadratic form, therefore we need a subtle trick to transform it into a more easily handled problem.

Trick In order to replace "$\cosh x \leq e^{x^2/2}$" we can use the following trick: Let Y be a r.v. with $\mathbb{E}[Y] = 0$ (our case of $X - \mathbb{E}[X]$) and $Y \in [a, b]$, then for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}[e^{\lambda Y}] \leq e^{\lambda^2\frac{(b-a)^2}{2}}.$$

This is based on a symmetrization of Y by introducing another independent random variable $Y' \overset{\mathrm{d}}{=} Y$ and $Z \sim \text{Rademacher}(\frac{1}{2})$ from which we have $\mathbb{E}[e^{-\lambda Y'}] \overset{\text{Jens.}}{\leq} e^{-\lambda\mathbb{E}[Y]} = 1$, therefore

$$\mathbb{E}[e^{\lambda Y}] \leq \mathbb{E}[e^{\lambda Y}]\cdot\mathbb{E}[^{-\lambda Y'}] = \mathbb{E}[e^{\lambda(Y-Y')}] = \mathbb{E}[e^{\lambda Z(Y-Y')}] = \mathbb{E}[\cosh(\lambda(Y-Y'))] \leq \mathbb{E}[e^{\lambda^2\frac{(Y-Y')^2}{2}}] = e^{\frac{\lambda^2(b-a)^2}{2}}.$$

Using this trick, we can optimize the equation using

$$\mathbb{P}\Big(\sum_{i=1}^{N}(X_i - \mathbb{E}[X_i]) \geq t\Big) \leq e^{-\lambda t}\prod_i e^{\lambda^2 \frac{(M_i-m_i)^2}{2}}$$

$$= \exp\Big(-\lambda t + \frac{\lambda^2}{2}\sum_i \frac{(M_i-m_i)^2}{2}\Big).$$

We can optimize with $\lambda > 0$ and get the minimum with a different constant than 2. Finding this other minimum requires more work.

□

Example (Book 2.2.9 – Boosting a randomized algorithm)
We have an algorithm that gives the right answer out of two classes with a probability $\frac{1}{2}+\delta$, with $\delta > 0$. We run this algorithm N (odd) times and take the majority vote to get the final classification.

Problem Find the minimal N such that $\mathbb{P}(\text{correct answer}) \geq 1-\varepsilon$ for $\varepsilon \in (0,1)$ fixed.

Solution Consider the following r.v. $X_1, \dots, X_N$ be the indicator of the wrong answer

$$X_i = \begin{cases} 1 & \text{if } i^{\text{th}} \text{ run is wrong} \\ 0 & \textit{otherwise} \end{cases}$$

then, using thm. 4 with $t = N\delta$, $M_i = 1$ and $m_i = 0$ we can bound the probability of wrong answer as

$$\mathbb{P}\Big(X_1 + \ldots + X_N \geq \frac{N}{2}\Big) = \mathbb{P}\Big(\sum_{i=1}^{N}(X_i - (\frac{1}{2}-\delta)) \geq N\delta\Big) \overset{4}{\leq} \exp\left(-\frac{2N^{\cancel{2}}\delta^2}{\cancel{N}}\right).$$

Therefore, in order to have the required bounded probability we need

$$-2N\delta^2 \leq \log\varepsilon \iff \boxed{N \geq \frac{1}{2\delta^2}\log\frac{1}{\varepsilon}}.$$

1.2 Chernoff's inequality

Consider the last Hoeffding's inequality (thm. 4), then for a sum of random variables we can write the Gaussian tail using the CLT as approximately

$$\mathbb{P}(|Z| \geq t) \leq 2e^{-\frac{t^2}{2}}.$$

Chernoff's inequality is useful in regimes of sums in order to prove a bound that is again independent from the central limit theorem. The following theorem is a merged result of Theorem 2.3.1, Exercise 2.3.2 and Exercise 2.3.5 in the book.

Thm. 5 (Chernoff's inequality)
Let $X_1, \ldots, X_N$ be such that $X_i \overset{iid}{\sim} Bern(p_i)$ and consider the cumulative sum $S_N = \sum_i X_i$ with expected value $\mu = \mathbb{E}[S_N] = \sum_i p_i$. Then, the following inequalities hold:

$$\mathbb{P}(S_N \geq t) \leq e^{-\mu} \cdot \left(\frac{e\mu}{t}\right)^t \qquad \text{for } t > \mu,$$

$$\mathbb{P}(S_N \leq t) \leq e^{-\mu} \left(\frac{e\mu}{t}\right)^t \qquad \text{for } t < \mu,$$

$$\mathbb{P}(|S_N - \mu| \geq \delta\mu) \leq 2e^{-C\mu\delta^2} \qquad \text{for } \delta \in (0, 1],$$

where C is a universal constant (i.e. does not depend on the other quantities).

Proof.

1. The first step is always the same, let $\lambda > 0$ then

$$\mathbb{P}(S_N \geq t) = \mathbb{P}(e^{\lambda S_N} \geq e^{\lambda t}) \leq e^{-\lambda t}\mathbb{E}[e^{\lambda S_N}] = e^{-\lambda t} \prod_i \mathbb{E}[e^{\lambda X_i}]. \tag{1}$$

   Now for a Bernoulli random variable, $\mathbb{E}[e^{\lambda X_i}] = (1 - p_i)e^0 + p_i e^\lambda = 1 + (e^\lambda - 1)p_i$, and we use the following identity:

$$1 + x \leq e^x \quad \text{for all } x > 0,$$

   to write

$$\mathbb{E}[e^{\lambda X_i}] = 1 + \overbrace{(e^\lambda - 1)p_i}^{x} \leq \exp\big((e^\lambda - 1)p_i\big).$$

   Going back to (1), we have the following bound for any $\lambda > 0$,

$$\mathbb{P}(S_N \geq t) \leq e^{-\lambda t} e^{(e^\lambda - 1)\sum_i p_i} = e^{-\lambda t + \mu(e^\lambda - 1)}.$$

   Again, by optimizing over λ we find that the tightest bound from (1) is given by

$$f(\lambda) = -\lambda t + \mu(e^\lambda - 1) \implies \lambda_{\text{opt}} = \operatorname*{argmin}_{\lambda > 0} f(\lambda) = \log \frac{t}{\mu},$$

   from which we obtain the first Chernoff bound,

$$\mathbb{P}(S_N \geq t) \leq e^{-\mu} \left(\frac{e\mu}{t}\right)^t.$$

2. For the second inequality, proceed as before using

$$\mathbb{P}(S_N \leq t) \overset{\lambda \geq 0}{=} \mathbb{P}(e^{-\lambda S_N} \geq e^{-\lambda t}).$$

3. We can obtain the bound on $\mathbb{P}(|S_N - \mu| \geq \delta\mu)$ by using the fact that

$$\mathbb{P}(|S_N - \mu| \geq \delta\mu) \leq \mathbb{P}(S_N - \mu \geq \delta\mu) + \mathbb{P}(S_N - \mu \leq -\delta\mu) \overset{(1),(2)}{\leq} \ldots$$

$\square$

Thm. 6 (Poisson tail)
Let $X \sim Pois(\gamma)$ with $\gamma > 0$, and

$$\mathbb{P}(X = k) = e^{-\gamma}\frac{\gamma^k}{k!}, \quad \text{for } k = 0, 1, \ldots$$

Let now $t > \gamma$, then

$$\mathbb{P}(X \geq t) \leq e^{-\gamma}\left(\frac{e\gamma}{t}\right)^t \tag{A}$$

Remark This bound is extremely useful and is similar to Chernoff's bound thm. 5, which works instead for a sum of random variables.

Proof.

Exercise Prove equation (A) using the basic trick $\mathbb{P}(X \geq t) \leq e^{-\lambda t}\mathbb{E}[e^{\lambda X}]$, which can be computed explicitly, and then optimize over $\lambda > 0$. Briefly comment why this bound is optimal.

$\square$