Topics in learning theory*

Lecture 2: Tools from probability theory

Contents

1 On the Gaussian distribution

The goal of this section will be to have a first look at the concentration properties of random variables, i.e., to quantify their fluctuations around their expectation. Due to its universality, illustrated by the central limit theorem, the gaussian distribution is a natural first example that we'll shortly focus on. The results provided in this section will serve as a benchmark for further discussions. Let

$$\phi(t) := \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right),$$

be the density function of the gaussian $\mathcal{N}(0,1)$ distribution.

Theorem 1.1. *Suppose X follows the gaussian distribution $\mathcal{N}(0,1)$. Then, for all $t > 0$,*

$$\frac{t\phi(t)}{1+t^2} \leq \mathbb{P}\{X > t\} \leq \frac{\phi(t)}{t}. \tag{1.1}$$

Proof. Observe first that for any integrable function $f : \mathbb{R} \to (0, +\infty)$, and any $t > 0$,

$$\int_t^{+\infty} f(x)\,\mathrm{d}x \leq \frac{1}{t}\int_t^{+\infty} xf(x)\,\mathrm{d}x.$$

Applying this inequality to ϕ yields, for all $t > 0$,

$$\begin{aligned}\mathbb{P}\{X > t\} &= \int_t^{+\infty} \phi(x)\,\mathrm{d}x \\ &\leq \frac{1}{t}\int_t^{+\infty} x\phi(x)\,\mathrm{d}x \\ &= \frac{\phi(t)}{t}.\end{aligned}$$

For the second part, we integrate by parts the inequality that we have just obtained,

$$0 \leq -x\mathbb{P}\{X > x\} + \phi(x), \quad x > 0,$$

to get

$$\begin{aligned}0 &\leq \int_t^{+\infty} -x\mathbb{P}\{X > x\} + \phi(x)\,\mathrm{d}x \\ &= \frac{t^2}{2}\mathbb{P}\{X > t\} - \int_t^{+\infty} \frac{x^2}{2}\phi(x)\,\mathrm{d}x + \int_t^{+\infty} \phi(x)\,\mathrm{d}x \\ &= \frac{t^2}{2}\mathbb{P}\{X > t\} - \frac{t}{2}\phi(t) + \frac{1}{2}\int_t^{+\infty} \phi(x)\,\mathrm{d}x \\ &= \frac{(1+t^2)}{2}\mathbb{P}\{X > t\} - \frac{t}{2}\phi(t),\end{aligned}$$

where we have used that $\phi'(t) = -t\phi(t)$. □

The next inequality improves the previous upper bound for small values of $t > 0$.

Theorem 1.2. *Suppose X follows the gaussian distribution $\mathcal{N}(0,1)$. Then, for all $t > 0$,*

$$\mathbb{P}\{X > t\} \leq \frac{1}{2}\exp\left(-\frac{t^2}{2}\right). \tag{1.2}$$

Proof. Consider the function $F : \mathbb{R}_+ \to \mathbb{R}$, defined by

$$F(t) = \frac{1}{2}\exp\left(-\frac{t^2}{2}\right) - \mathbb{P}\{X > t\}.$$

We need to show that F takes only positive values. To that aim, observe that F is continuously differentiable on $\mathbb{R}_+$, that $F(0) = 0$, and that

$$\begin{aligned}F'(t) &= -\frac{t}{2}\exp\left(-\frac{t^2}{2}\right) - \frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{\sqrt{2\pi}}\int_t^{+\infty} \exp\left(-\frac{x^2}{2}\right)\mathrm{d}x \\ &= \left\{\frac{1}{\sqrt{2\pi}} - \frac{t}{2}\right\}\exp\left(-\frac{t^2}{2}\right).\end{aligned}$$

This already proves that F is positive on $[0, \sqrt{2/\pi}]$. Finally, for $t > \sqrt{2/\pi}$, using the same trick as for the upper bound in Theorem 1.1, we obtain

$$\begin{aligned}\mathbb{P}\{X > t\} &= \frac{1}{\sqrt{2\pi}}\int_t^{+\infty} \exp\left(-\frac{x^2}{2}\right)\mathrm{d}x \\ &\leq \frac{1}{\sqrt{2\pi}}\sqrt{\frac{\pi}{2}}\int_t^{+\infty} x\exp\left(-\frac{x^2}{2}\right)\mathrm{d}x \\ &= \frac{1}{2}\exp\left(-\frac{t^2}{2}\right),\end{aligned}$$

which implies that F is positive for $t > \sqrt{2/\pi}$ and completes the proof. □

Remark 1.3. *One can show that the constant $1/2$ in Theorem 1.2 is optimal in the sense that*

$$\sup_{t>0}\left\{\exp\left(\frac{t^2}{2\sigma^2}\right)\mathbb{P}\{X > t\}\right\} = \frac{1}{2}.$$

*Teaching material can be found at https://www.qparis-math.com/teaching.

Remark 1.4. *Theorems 1.1 and 1.2 apply to the $\mathcal{N}(m,\sigma^2)$ distribution for any $m\in\mathbb{R}$ and $\sigma^2>0$ by a simple scaling argument: If $X\sim\mathcal{N}(m,\sigma^2)$ then*

$$\frac{X-m}{\sigma}\sim\mathcal{N}(0,1).$$

2 The Cramér-Chernoff method

The results derived in the previous section rely heavily on the specific form of the gaussian density. In particular, one cannot reproduce the previous arguments in situations where the information on the distribution of X is of more general nature. The present section develops a tool to deal with more complex scenarios. We start with a very basic result.

Theorem 2.1 (Markov's inequality)**.** *Let X be a nonnegative random variable such that $\mathbb{E}X<+\infty$. Then, for all $t\geq 0$,*

$$t\mathbb{P}\{X\geq t\}\leq\mathbb{E}X.$$

Proof. We simply notice that,

$$t\mathbb{P}\{X\geq t\}=t\mathbb{E}\mathbf{1}\{X\geq t\}\leq\mathbb{E}[X\mathbf{1}\{X\geq t\}]\leq\mathbb{E}X,$$

for any $t\geq 0$. □

Exercise 2.2 (Chebychev's inequality)**.** *Using Markov's inequality, show that for any $\mathbb{R}$-valued random variable X such that $\mathbb{E}X^2<+\infty$ and any $t\geq 0$,*

$$t^2\mathbb{P}\{|X-\mathbb{E}X|\geq t\}\leq\mathbb{V}(X).$$

The simple idea used in the proof of Markov's inequality can be generalised and turned into a powerful method known as the Cramér-Chernoff method. To describe this method, consider a random variable X and any nonnegative and strictly increasing $\varphi:\mathbb{R}\to\mathbb{R}_+$. Then, for all $t\geq 0$,

$$\varphi(t)\mathbb{P}\{X\geq t\}=\varphi(t)\mathbb{P}\{\varphi(X)\geq\varphi(t)\}\leq\mathbb{E}\varphi(X),$$

by Markov's inequality. In particular, since $\mathbb{P}\{X\geq t\}$ does not depend on the choice of φ, it follows that

$$\mathbb{P}\{X\geq t\}\leq\inf_{\varphi\in\Phi}\varphi(t)^{-1}\mathbb{E}\varphi(X),\tag{2.1}$$

for any collection Φ of nonnegative and increasing functions $\varphi:\mathbb{R}\to\mathbb{R}_+$. Due to the specific algebraic properties of the exponential function, a very convenient choice for the class Φ is provided by the collection of all functions $x\mapsto e^{\theta x},\theta>0$.

Definition 2.3 (Moment generating function)**.** *The moment generating function*[1] (MGF) *of X, is defined, for all $\theta\in\mathbb{R}$, by*

$$M_X(\theta):=\mathbb{E}\exp(\theta X).$$

We'll denote, for $\theta\in\mathbb{R}$,

$$\Lambda_X(\theta)=\log M_X(\theta).$$

The most important insight, relative to the MGF, is summarised in the following result.

[1] Also referred to as the Laplace transform of the distribution of X

Theorem 2.4 (Cramér-Chernoff)**.** *For any real-valued random variable X and any $t\in\mathbb{R}$, defining*

$$\Lambda_X^*(t):=\sup_{\theta>0}\{\theta t-\Lambda_X(\theta)\},$$

we have

$$\mathbb{P}\{X>t\}\leq e^{-\Lambda_X^*(t)}.$$

Proof. For all $\theta>0$, using that thet function $x\mapsto e^{\theta x}$ is increasing, we deduce that for all $t\in\mathbb{R}$,

$$\begin{aligned}\mathbb{P}\{X>t\}&=\mathbb{P}\{\exp(\theta X)>\exp(\theta t)\}\\&\leq\exp(-\theta t)\,\mathbb{E}\exp(\theta X)\\&=\exp(-\theta t+\Lambda_X(\theta)),\end{aligned}$$

where the second line follows from Markov's inequality. The result follows by optimising the bound in $\theta>0$. □

Exercise 2.5 (Bernoulli distribution)**.** *For $p\in(0,1)$, consider a random variable ξ such that*

$$\mathbb{P}\{\xi=0\}=1-p\quad\text{and}\quad\mathbb{P}\{\xi=1\}=p,$$

and let $X=\xi-p$. For all $t\in(0,1-p)$, show that

$$\mathbb{P}\{X>t\}\leq e^{-h_p(t+p)},$$

where

$$h_p(u):=u\log\frac{u}{p}+(1-u)\log\frac{1-u}{1-p}.$$

Note that, for $t\geq 1-p$, we obviously have $\mathbb{P}\{X>t\}=0$.

The algebraic property $e^{x+y}=e^xe^y$ of the exponential function implies that the MGF behaves favorably in the context of independent random variables as described in the next exercise.

Exercise 2.6. *Let $X_1,\dots,X_n$ be independent random variables and set $X=X_1+\cdots+X_n$. Then, for all $\theta\in\mathbb{R}$,*

$$\Lambda_X(\theta)=\sum_{i=1}^n\Lambda_{X_i}(\theta).$$

In particular, if the variables $X_1,\dots,X_n$ are i.i.d., then for all $t\in\mathbb{R}$,

$$\Lambda_X^*(t)=n\Lambda_{X_1}^*\left(\frac{t}{n}\right).$$

Exercise 2.7 (Binomial distribution)**.** *Consider a variable ξ distributed according to the Binomial distribution with parameters $n\geq 1$ and $p\in(0,1)$, i.e.*

$$\mathbb{P}\{\xi=k\}=\binom{n}{k}p^k(1-p)^{n-k},$$

and let X be the centered random variable $X=\xi-np$. Using Exercises 2.5 and 2.6, show that for all $t\in(0,n(1-p))$,

$$\mathbb{P}\{X>t\}\leq e^{-nh_p\left(\frac{t}{n}+p\right)},$$

where h_p is as in Exercise 2.5. Note that for $t\geq n(1-p)$, we have $\mathbb{P}\{X>t\}=0$.

We end the section by studying the case of the gaussian distribution.

Lemma 2.8. *Let $m \in \mathbb{R}$ and $\sigma^2 > 0$. Suppose that X follows the gaussian distribution $\mathcal{N}(m, \sigma^2)$. Then, for all $\theta \in \mathbb{R}$,*

$$\log \mathbb{E} \exp(\theta\{X - m\}) = \frac{\theta^2\sigma^2}{2}.$$

Proof. Without loss of generality, suppose $m = 0$. Then, for $\theta \in \mathbb{R}$, we obtain

$$\begin{aligned}
&\mathbb{E} \exp(\theta X) \\
&= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left(\theta x - \frac{x^2}{2\sigma^2}\right) \mathrm{d}x \\
&= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left(\frac{\theta^2\sigma^2}{2} - \frac{(x - \sigma^2\theta)^2}{2\sigma^2}\right) \mathrm{d}x \\
&= \exp\left(\frac{\theta^2\sigma^2}{2}\right) \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp\left(-\frac{(x - \sigma^2\theta)^2}{2\sigma^2}\right) \mathrm{d}x \\
&= \exp\left(\frac{\theta^2\sigma^2}{2}\right),
\end{aligned}$$

which concludes the proof. □

Using the computation of the previous Lemma, applying Theorem 2.4 to the gaussian distribution yields that, for all $t > 0$,

$$\mathbb{P}\{X - m > t\} = \mathbb{P}\{X - m < -t\} \leq \exp\left(-\frac{t^2}{2\sigma^2}\right),$$

whenever $X \sim \mathcal{N}(m, \sigma^2)$. Comparing this bound with Theorems 1.1 and 1.2, we observe that the method presented in Theorem 2.4 gives a result only slightly weaker than if we had used the specific form of the gaussian density as in paragraph **??**. This advocates for a use of Theorem 2.4 in more general situations. A case of particular interest is the case of subgaussian random variables.

3 Subgaussian random variables

In the previous section we have shown how the deviations of a random variable X can be investigated by studying the behavior of its MGF. Motivated by the result of Lemma 2.8, we introduce a specific class of distributions that have lighter tails than gaussian distributions.

Definition 3.1. *A real-valued random variable X (or its distribution) is said to be subgaussian if there exists $s^2 > 0$ such that,*

$$\forall \theta \in \mathbb{R} : \quad \log \mathbb{E} \exp(\theta\{X - \mathbb{E}X\}) \leq \frac{\theta^2 s^2}{2}.$$

Whenever this holds, we'll denote $X \in \mathrm{SG}(s^2)$. The smallest $s^2 > 0$ for which $X \in \mathrm{SG}(s^2)$ is called the variance proxy of X, sometimes denoted $\|X\|_{\mathrm{vp}}^2$, and given by

$$\|X\|_{\mathrm{vp}}^2 = \sup_{\theta \neq 0} \frac{2}{\theta^2} \log \mathbb{E} \exp(\theta\{X - \mathbb{E}X\}).$$

According to Lemma 2.8, a gaussian variable is clearly subgaussian. Other examples of subgaussian variables are discussed below.

Remark 3.2. *The notation $\|X\|_{\mathrm{vp}}^2$ relates to the fact that the variance proxy is a squared semi-norm on the set of subgaussian random variables. More precisely, the reader may show as an exercise that the set of subgaussian random variables is indeed an $\mathbb{R}$-vector space and that the following properties hold.*
(1) *For any subgaussian variable X and any $\alpha \in \mathbb{R}$,*

$$\|\alpha X\|_{\mathrm{vp}} = |\alpha| \, \|X\|_{\mathrm{vp}}.$$

(2) *For any subgaussian variables X, Y (non necessarily independent)*

$$\|X + Y\|_{\mathrm{vp}} \leq \|X\|_{\mathrm{vp}} + \|Y\|_{\mathrm{vp}}.$$

(3) *For any subgaussian variable X,*

$$\|X\|_{\mathrm{vp}} = 0 \quad \Leftrightarrow \quad X = \mathbb{E}X \ \ a.s..$$

In particular, $\|.\|_{\mathrm{vp}}$ defines a norm on the space of centered subgaussian random variables.

Theorem 3.3. *Suppose that X is subgaussian. Then, for all $t \in \mathbb{R}$,*

$$\mathbb{P}\{X - \mathbb{E}X > t\} \vee \mathbb{P}\{X - \mathbb{E}X < -t\} \leq \exp\left(-\frac{t^2}{2\|X\|_{\mathrm{vp}}^2}\right),$$

where $a \vee b := \max\{a, b\}$.

Proof. By definition, X is subgaussian if and only if $-X$ is subgaussian and $\|X\|_{\mathrm{vp}}^2 = \|-X\|_{\mathrm{vp}}^2$. As a result, it is enough to prove the first inequality. Now, combining Theorem 2.4 and Definition 3.1, we obtain

$$\begin{aligned}
\mathbb{P}\{X - \mathbb{E}X > t\} &\leq \exp\left(-\sup_{\theta \geq 0}\left\{\theta t - \frac{\theta^2\|X\|_{\mathrm{vp}}^2}{2}\right\}\right) \\
&= \exp\left(-\frac{t^2}{2\|X\|_{\mathrm{vp}}^2}\right),
\end{aligned}$$

which completes the proof. □

The class of subgaussian random variables is much wider than the class of gaussian variables. The next result shows, for instance, that any bounded random variable is subgaussian.

Lemma 3.4 (Hoeffding's lemma)**.** *Let X be an $[a, b]$-valued random variable for $-\infty < a < b < +\infty$. Then, for all $\theta \in \mathbb{R}$,*

$$\log \mathbb{E} \exp(\theta\{X - \mathbb{E}X\}) \leq \frac{\theta^2(b - a)^2}{8}.$$

In other words $\|X\|_{\mathrm{vp}}^2 \leq (b - a)^2/4$.

Proof. Note that, by the convexity of the exponential function,

$$e^{\theta x} \leq \frac{x - a}{b - a} e^{\theta b} + \frac{b - x}{b - a} e^{\theta a},$$

for all $a \leq x \leq b$. Exploiting the fact that $\mathbb{E}[X - \mathbb{E}X] = 0$, and introducing the notation $p = -a/(b - a)$, we deduce that

$$\begin{aligned}
\mathbb{E} \exp(\theta\{X - \mathbb{E}X\}) &\leq \frac{b}{b - a} e^{\theta a} - \frac{a}{b - a} e^{\theta b} \\
&= (1 - p + pe^{\theta(b-a)}) e^{-\theta p(b-a)} \\
&= e^{f(u)},
\end{aligned}$$

where $u = \theta(b-a)$ and $f(u) = -pu + \log(1-p+pe^u)$. By straightforward computations, we get

$$f'(u) = -p + \frac{p}{p+(1-p)e^{-u}},$$

so that $f(0) = f'(0) = 0$. Moreover, for all $c \geq 0$,

$$f''(c) = \frac{p(1-p)e^{-c}}{(p+(1-p)e^{-c})^2} \leq \frac{1}{4}.$$

Thus, by the Taylor-Lagrange theorem, there exists $c \in [0,u]$ such that,

$$f(u) = f(0) + uf'(0) + \frac{u^2}{2}f''(c) \leq \frac{u^2}{8} = \frac{\theta^2(b-a)^2}{8},$$

which concludes the proof. □

Remark 3.5 (Variance vs variance proxy)**.** *As proven in Lemma 2.8, any gaussian random variable is subgaussian with variance proxy equal to its variance. However a random variable may be subgaussian with variance proxy strictly larger than its variance. For example, if X follows the Bernoulli distribution with parameter $p \in (0,1)$, $p \neq 1/2$, the variance of X is $p(1-p)$ while its variance proxy is*

$$\sup_{\theta \neq 0} \frac{2}{\theta^2}\log(pe^{\theta(1-p)} + (1-p)e^{-\theta p}) = \frac{1-2p}{2(\log(1-p) - \log p)},$$

which is strictly larger than $p(1-p)$. More generally, it may be proven as an exercise that inequality

$$\mathrm{var}(X) \leq \|X\|_{\mathrm{vp}}^2,$$

always holds.

4 Sums of independent subgaussians

We now investigate the concentration properties of sums of independent sub-gaussian variables.

Theorem 4.1 (Generalized Hoeffding inequality)**.** *Let $X_1, \dots, X_n$ be independent sub-gaussian variables. Then,*

$$\|\sum_{i=1}^n X_i\|_{\mathrm{vp}}^2 \leq \sum_{i=1}^n \|X_i\|_{\mathrm{vp}}^2. \tag{4.1}$$

In particular, for all $t > 0$,

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}X_i) > t\right\} \leq \exp\left(-\frac{n^2t^2}{2\sum_{i=1}^n \|X_i\|_{\mathrm{vp}}^2}\right).$$

Remark 4.2. *As usual, we deduce by symmetry that, under the same assumptions,*

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}X_i) < -t\right\} \leq \exp\left(-\frac{n^2t^2}{2\sum_{i=1}^n \|X_i\|_{\mathrm{vp}}^2}\right).$$

Remark 4.3. *Before the proof, note that (even without independence of the X_i's) we already know from the first lecture that $\sum_{i=1}^n X_i$ is sub-gaussian and that*

$$\|\sum_{i=1}^n X_i\|_{\mathrm{vp}} \leq \sum_{i=1}^n \|X_i\|_{\mathrm{vp}}.$$

Here, the statement shows that this inequality can be improved (add $\square^2$) if we assume the variables are independent.

Proof of Theorem 4.1. By independence, and for all $\theta \in \mathbb{R}$,

$$\begin{aligned}\log \mathbb{E}\exp\left(\theta\sum_{i=1}^n (X_i - \mathbb{E}X_i)\right) &= \sum_{i=1}^n \log \mathbb{E}\exp\left(\theta\{X_i - \mathbb{E}X_i\}\right) \\ &\leq \frac{\theta^2}{2}\sum_{i=1}^n \|X_i\|_{\mathrm{vp}}^2,\end{aligned}$$

where the inequality holds by definition of sub-gaussianity. Using the definition of sugaussianity again, we deduce (4.1). The last statement follows from the concentration property of sub-gaussian variables. □

Applying Hoeffding's lemma, we deduce the following result.

Corollary 4.4 (Classical Hoeffding inequality)**.** *If each X_i is $[a_i, b_i]$-valued for some $-\infty < a_i < b_i < +\infty$, then provided the X_i's are independent, a direct combination of Hoeffding's Lemma and Theorem 4.1 yields the inequality*

$$\begin{aligned}&\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}X_i) > t\right\} \vee \mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}X_i) < -t\right\} \\ &\leq \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right),\end{aligned} \tag{4.2}$$

known as Hoeffding's inequality.