Empirical Processes (MATH-522)

Lecture 2: Basic Concentration Inequalities

Myrto Limnios

MATH, Ecole Polytéchnique Fédérale de Lausanne

February 25, 2025

What we saw last week

1. We recalled fundamental definitions and properties for metric spaces, to be able to understand modes of convergence of random sequences of random maps
2. We recalled the continuous mapping Theorem, LLN and CLT
3. We defined empirical process indexed by arbitrary sets
4. We stated the first important extensions of uniform limit theorems
5. We presented important learning examples in statistics

Empirical measure

Suppose the r.v. X to be valued in a multidimensional space, e.g., a generic Euclidean space $\mathcal{X} \subseteq \mathbb{R}^d$, with $d \geq 2$.

Based on an independent random sample, $X_1, \ldots, X_n$, we defined the *empirical measure* by

$$P_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i} \ , \tag{1}$$

where $\delta_X(A)$ is the Dirac measure of the event $\{X \in A\}$. In particular

$$\begin{aligned} P_n(A) \quad &:= \quad \frac{1}{n} \sum_{i=1}^{n} 1_A(X_i) \\ &= \quad \frac{1}{n} \{\text{number of observations } i \leq n\colon \quad X_i \in A\} \ , \end{aligned}$$

for any Borel subset $A \subset \mathcal{X}$. We refer to the empirical measure indexed by a collection of subsets $\mathcal{C}$ of $\mathcal{X}$ by

$$\boxed{\{P_n(C), \quad C \in \mathcal{C}\}} .$$

Empirical measure indexed by a function class

In some applications, it is more convenient to consider the index set to be a class of functions, e.g., for averages. For any measurable function $h : \mathcal{X} \to \mathbb{R}$, consider

$$P_n h := \boxed{\frac{1}{n}\sum_{i=1}^{n} h(X_i)}\,, \qquad (2)$$

$h = \mathrm{Id}$

then the empirical measure indexed by a class of real-valued functions $\mathcal{H}$ is

$$\boxed{\{P_n h, \quad h \in \mathcal{H}\}}$$

considered as the empirical estimator of $\boxed{Ph := \int_{\mathcal{X}} h(x) P(\mathrm{d}x)}$ for a given $h \in \mathcal{H}$.

Empirical measure indexed by a function class

In some applications, it is more convenient to consider the index set to be a class of functions, e.g., for averages. For any measurable function $h : \mathcal{X} \to \mathbb{R}$, consider

$$P_n h := \frac{1}{n} \sum_{i=1}^{n} h(X_i) \ , \tag{2}$$

then the empirical measure indexed by a class of real-valued functions $\mathcal{H}$ is

$$\{P_n h, \quad h \in \mathcal{H}\}$$

considered as the empirical estimator of $Ph := \int_{\mathcal{X}} h(x) P(\mathrm{d}x)$ for a given $h \in \mathcal{H}$.

Remark

Notice that taking $\mathcal{H}$ to be the collection $\{1_C, \ C \in \mathcal{C}\}$ recovers the first definition.
It also recovers the empirical c.d.f.s in the univariate case, i.e., when $\mathcal{X} = \mathbb{R}$, by taking $\mathcal{C} = \{1\{(-\infty, x]\}, \ x \in \mathbb{R}\}$.

$\{\{x \mapsto 1(-\infty, x]\}, x \in \mathbb{R}\}$

Classical empirical estimators

- Empirical mean: $(1/n)\,\underline{\sum_{i=1}^n 1\{X_i \leq \overset{\in \mathbb{R}.}{x}}\}$ $\Longrightarrow$ GC/Donsker first uniform theorems for the empirical measures

$\Longrightarrow$ **But also we would like to establish (non)asymptotic properties of optimal empirical estimators from the index set ...**

Classical empirical estimators

- Empirical mean: $(1/n)\sum_{i=1}^n 1\{X_i \le x\} \Longrightarrow$ GC/Donsker first uniform theorems for the empirical measures

$\Longrightarrow$ **But also we would like to establish (non)asymptotic properties of optimal empirical estimators from the index set ...**

- Empirical median: $(1/n)\sum_{i=1}^n |X_i - m|$, with m being the median
- Empirical binary risk: $(1/n)\sum_{i=1}^n 1\{h(X_i) \neq Y_i\}$, with $h : \mathcal{X} \to \{0,1\}$ being a classifier
- Least-squares estimator: $(1/n)\sum_{i=1}^n (Y_i - h(X_i))^2$, with $h : \mathcal{X} \to \mathbb{R}$ being a regression function
- Empirical likelihood estimator: $(1/n)\sum_{i=1}^n \log \underline{p_\theta}(X_i)$, with p_θ density function indexed by a parameter of interest θ

We will model those estimators by a generic r.v. Z_n in this lecture.

What we will see today

We will focus on a **fixed estimator** modeled by a generic real-valued r.v. $\boxed{Z_n}$ and suppose that its mean $\underline{\mathbb{E}[Z_n] = PZ_n}$ exists and is finite (i.e. suppose Z_n to be integrable).

We want to provide upper-bounds of the probability

$$\boxed{\mathbb{P}\{|Z_n - \mathbb{E}Z_n| \geq t\}}$$ is very Small

+ ↘0 for $n \to +\infty$

for all $t > 0$.

What we will see today

- We will study how to derive probabilistic bounds to quantify the speed of the deviation of averages of r.v.s w.r.t their mean, known as basic *concentration inequalities*.
- We will focus on bounds that have exponential decay, under various assumptions on the moments of the r.v.s
- Importantly, those results allow to assess how averages concentrate around their mean for **fixed** sample size.

Today's outline

1. Basic methods for bounding sums of independent variables

2. Bounding the variance general functionals of independent variables

From Markov's inequality...

Theorem (Markov's inequality)

For any nonnegative real-valued r.v. Z, for all $t > 0$, one has

$$\underline{\mathbb{P}\{Z \geq t\}} \leq \frac{\mathbb{E}Z}{t} \ . \tag{3}$$

Proof.

Let $t>0$. $\{Z \geq t\}$: $Z \mathbb{1}\{Z \geq t\} \geq t \mathbb{1}\{Z \geq t\}$

$\mathbb{P}(Z \mathbb{1}\{Z \geq t\}) \geq \mathbb{P}(t \mathbb{1}\{Z \geq t\})$

→ $\mathbb{E}[Z] = \mathbb{E}[Z \mathbb{1}\{Z \geq t\}] + \mathbb{E}[Z \mathbb{1}\{Z < t\}]$

From Markov's inequality...

Theorem (Markov's inequality)

For any nonnegative real-valued r.v. Z, for all $t > 0$, one has

$$\mathbb{P}\{Z \geq t\} \leq \frac{\mathbb{E}Z}{t} \ . \tag{3}$$

Proof.

Let $t > 0$. Notice that $Z1\{Z \geq t\} \geq t1\{Z \geq t\}$, thus

$$\mathbb{E}[Z1\{Z \geq t\}] \geq \mathbb{E}[t1\{Z \geq t\}] = t\mathbb{P}\{Z \geq t\}$$

We also can write $\mathbb{E}Z = \mathbb{E}[Z1\{Z \geq t\}] + \mathbb{E}[Z1\{Z < t\}]$, where both expectations are positive, the result is obtained. □

From Markov's inequality...

Theorem (Markov's inequality)

For any nonnegative real-valued r.v. Z, for all $t > 0$, one has

$$\boxed{\mathbb{P}\{Z \geq t\} \leq \frac{\mathbb{E}Z}{t}\ .} \tag{3}$$

Proof.

Let $t > 0$. Notice that $Z1\{Z \geq t\} \geq t1\{Z \geq t\}$, thus

$$\mathbb{E}[Z1\{Z \geq t\}] \geq \mathbb{E}[t1\{Z \geq t\}] = t\mathbb{P}\{Z \geq t\}$$

We also can write $\mathbb{E}Z = \mathbb{E}[Z1\{Z \geq t\}] + \mathbb{E}[Z1\{Z < t\}]$, where both expectations are positive, the result is obtained. □

Remark

- We can always apply Markov's inequality to $\underbrace{|Z - \mathbb{E}Z|}$ as it is nonnegative a.s.
- It is only interesting when $\boxed{\mathbb{E}Z < \infty.}$

… to Chebyschev's inequality

We can easily obtain **sharper results**, following the proof technique of Markov's inequality, that comes at a price of **higher finite moments**.

Let $h : I \subseteq \mathbb{R} \to (0, \infty)$ be a **nondecreasing function**. Then we have that

$$\mathbb{P}\{Z \geq t\} \leq \mathbb{P}\{h(Z) \geq h(t)\} \overset{MI}{\leq} \frac{\mathbb{E}h(Z)}{h(t)} \ . \tag{4}$$

Why is this interesting?

... to Chebyschev's inequality

We can easily obtain **sharper results**, following the proof technique of Markov's inequality, that comes at a price of **higher finite moments**.

Let $h : I \subseteq \mathbb{R} \to (0, \infty)$ be a **nondecreasing function**. Then we have that

$$\mathbb{P}\{Z \geq t\} \leq \mathbb{P}\{h(Z) \geq h(t)\} \leq \frac{\mathbb{E}h(Z)}{h(t)} \, . \qquad (4)$$

$h : x \mapsto x^2$

Why is this interesting?

Lemma (Chebyschev's inequality)

Let a sample of n independent real-valued square-integrable r.v. $X_1, \ldots, X_n$. Then, for all $t > 0$

$$\mathbb{P}\left\{ \frac{1}{n} \left| \sum_{i=1}^{n} (X_i - \mathbb{E}X_i) \right| \geq t \right\} \leq \frac{\sigma^2}{nt^2} \, , \qquad (5)$$

where $\sigma^2 = (1/n) \sum_{i=1}^{n} \mathrm{Var}(X_i)$.

Remark

- One can take $h(t) = t^q$, with $q \in \mathbb{N}^*$, as soon as the q-th moment is finite. More generally, a large family of nondecreasing maps h can be considered as soon as they are valued in $(0, \infty)$.

- We will focus on exponential transforms, yielding the basis of *Cramér-Chernoff's method* to obtain sharp exponential bound of such deviation probabilities.

Cramér-Chernoff method

$h: \lambda \mapsto e^{\lambda Z}$

Let a generic real-valued r.v. Z, and let a parameter $\lambda \geq 0$. Observe that by Markov's Inequality

$$\underline{\mathbb{P}\{Z \geq t\}} = \underline{\mathbb{P}\{e^{\lambda Z} \geq e^{\lambda t}\}} \underset{\uparrow \text{MI}}{\leq} e^{-\lambda t}\underbrace{\mathbb{E}[e^{\lambda Z}]} \ , = \underbrace{e^{-\lambda t + \log(\mathbb{E}e^{\lambda Z})}}$$

minize w.r.t. λ ⤳ sharp expo-type upperbound

Definition

The *Cramér transform* of Z is defined by:

$$\psi_Z^* : t \in \mathbb{R} \mapsto \sup_{\lambda \geq 0}(\lambda t - \psi_Z(\lambda)) \ ,$$

where $\lambda \mapsto \log \mathbb{E}[e^{\lambda Z}] =: \psi_Z(\lambda)$ is the log-moment generating function of Z.

The goal is to derive the sharpest upperbound

$$\boxed{\mathbb{P}\{Z \geq t\} \leq e^{-\psi_Z^*(t)} \ ,}$$

The function ψ_Z^* known as the *Cramér transform* of Z. In fact (prove as exercise), we can consider

$$\psi_Z^* : t \mapsto \sup_{\lambda \in \mathbb{R}}(\lambda t - \psi_Z(\lambda)) \ .$$

Short reminder about moment generating functions

The moment generating function of a r.v. Z is defined by

$$\boxed{\lambda \mapsto \mathbb{E}[e^{\lambda Z}] \quad (= \exp\{\psi_Z(\lambda)\})}$$

- It is defined in an open neighborhood of 0 and real-valued
- It does not exist for all r.v.s Z (e.g. Cauchy distribution)
- **BUT it fully characterizes the distribution of Z**

Short reminder about moment generating functions

The moment generating function of a r.v. Z is defined by

$$\lambda \mapsto \mathbb{E}[e^{\lambda Z}] \quad (= \exp\{\psi_Z(\lambda)\})$$

- It is defined in an open neighborhood of 0 and real-valued
- It does not exist for all r.v.s Z (e.g. Cauchy distribution)
- **BUT it fully characterizes the distribution of Z**

Remark (Key fact)

Recall that the series expansion of $\underline{e^{\lambda Z}}$ around 0 gives

$$e^{\lambda Z} = 1 + \lambda Z + \frac{(\lambda Z)^2}{2} + \frac{(\lambda Z)^3}{3!} + \dots$$

Hence, if the moments exist, by linearity of the expectation,

$$\boxed{\mathbb{E}e^{\lambda Z} = 1 + \lambda \mathbb{E}Z + \frac{\lambda^2 \mathbb{E}Z^2}{2} + \frac{\lambda^3 \mathbb{E}Z^3}{3!} + \dots}$$

where the moments are w.r.t. the distribution of Z (e.g. probability mass function, continuous probability distribution, Stieljes integrals).

Explicit bounds for parametric distributions

Example

Prove the following inequalities.

1. Suppose Z to be a centered Gaussian r.v. with finite variance σ^2, then $\underline{\psi_Z(\lambda) = \lambda^2\sigma^2/2}$ and

$$\boxed{\mathbb{P}\{Z \geq t\} \leq e^{-t^2/(2\sigma^2)}} .$$

2. Suppose $\boxed{Z = Y - b}$ where Y is a Poisson r.v. with parameter b, i.e., $\mathbb{P}\{Y = k\} = e^{-b}b^k/k!$. Show that

$$\boxed{\psi_Z^*(t) = bh(t/b)} ,$$

with $h(u) = (1 + u)\log(1 + u) - u$ for all $u \geq -1$, and for all $t \leq b$,

$$\psi_{-Z}^*(t) = bh(-t/b) \ .$$

Remark (Key fact)

We now derive Chernoff's inequality when applied to a sum of **i.i.d.** centered r.v. $X_1, \dots, X_n$, then

$Z = \sum_{i \le n} X_i$

$$\psi_Z(\lambda) = \log \mathbb{E}[e^{\lambda Z}] = \log \mathbb{E}[e^{\lambda \sum_{i \le n} X_i}] \overset{\perp\!\!\!\perp \text{ r.v.s}}{=} \log \prod_{i \le n} \mathbb{E}[e^{\lambda X_i}] \overset{\text{identically distr.}}{=} n\psi_X(\lambda)\,. \tag{6}$$

In particular

$$\boxed{\psi_Z^*(t) = n\psi_X^*\left(\frac{t}{n}\right)}$$

yielding the general form.

Example

Consider an i.i.d. sample drawn from a Poisson distribution of parameter b (cf. previous slide) then, for all $t > 0$,

$$\mathbb{P}\left\{\sum_{i=1}^{n}(X_i - b) \ge t\right\} \le e^{-nbh(-t/nb)}\,.$$

Chernoff's inequality

Lemma

Let a sample of $\underline{n \text{ i.i.d.}}$ real-valued centered r.v. $\underline{X_1, \dots, X_n}$. Then, for all $\underline{t > 0}$

$$\mathbb{P}\left\{ \underline{\frac{1}{n}\left|\sum_{i=1}^{n} X_i\right|} \geq t \right\} \leq \underline{2e^{-n\overbrace{\psi_X^*(t)}}}\,, \tag{7}$$

where ψ_X^* is the Cramér transform of the r.v. X.

We will use the log-moment generating function ψ_X to characterize the decrease of the tails of distributions for real-valued r.v. X. For instance, Gaussian r.v. are characterized by exact squared decrease.

sub-Gaussian and sub-Gamma r.v.s

Three important classes are defined below.

Definition

A real-valued centered r.v. X is said to be *sub-Gaussian* with variance parameter ν if

$$\boxed{\psi_X(\lambda) \le \frac{\lambda^2 \nu}{2}} . \qquad (8)$$

« variance »

$$\psi_{\mathcal{N}(0,\sigma^2)}(\lambda) = \frac{\lambda^2\sigma^2}{2}$$

$\lambda \in \mathbb{R}$

We denote this class by $\mathcal{G}(\nu)$.

Definition

A real-valued centered r.v. X is said to be *sub-Gamma* on the right tail with variance factor ν and scale parameter c if

$$\boxed{\psi_X(\lambda) \le \frac{\lambda^2 \nu \; 1}{2(1 - c\lambda)}}, \qquad (9)$$

ex $Y \sim$ Gamma r.v

$Y - EY \sim$ sub-Gamma

for all $\boxed{0 < \lambda < 1/c.}$

$\lambda c \in (0,1)$

And similarly, X is said to be *sub-Gamma* on the left tail with variance factor ν and scale parameter c, if $-X$ is *sub-Gamma* of right tail with same parameters.

X is said to be *sub-Gamma* with variance factor ν and scale parameter c, if it is both on left and right tails with equal parameters. We denote that class by $\Gamma(\nu, c)$.

Remark

Notice that the moment-generating function of a centered sub-Gaussian r.v. is dominated by that of a centered Gaussian r.v.

Example (Exercise)

- Prove that for any $X \in \mathcal{G}(\nu)$, for all $t > 0$,

$$\underbrace{\mathbb{P}\{X > t\}} \vee \underbrace{\mathbb{P}\{-X > t\}} \leq \underbrace{e^{-t^2/(2\boxed{\nu})}} \overset{\nu \; \sigma^2 \; \nu}{} . \tag{10}$$

- Prove that for any $X \in \Gamma(\nu, c)$, for all $t > 0$,

$$\mathbb{P}\{X > \underbrace{\sqrt{2\nu t} + ct}\} \vee \mathbb{P}\{-X > \sqrt{2\nu t} + ct\} \leq e^{-t} . \tag{11}$$

We will see now how Cramér-Chernoff's method is a key tool for proving fundamental concentration inequalities for general formulations of averages based on finite sample of independent r.v.s. with **exponential decay**.

Hoeffding's inequality

We start with a fundamental bound of the log-moment generating function for bounded r.v.s.

Lemma (Hoeffding inequality, 63)

Consider the r.v. X to be centered, and bounded a.s. by $a \leq X \leq b$, with $a < b$. Then, for all $\lambda > 0$

$$\boxed{\psi_X(\lambda) \leq \frac{\lambda^2 (b-a)^2}{8}}. \tag{12}$$

$\nu = \frac{(b-a)^2}{4}$ $X \in \mathcal{G}\left(\frac{(b-a)^2}{4}\right)$

Proof.

By convexity of the exponential function $x \mapsto e^x$, we have for all $t > 0$,

$$\boxed{e^{tx} \le \frac{b-x}{b-a} e^{ta} + \frac{x-a}{b-a} e^{tb}}$$

$$\log \underbrace{\mathbb{E}e^{\lambda X}} \le \frac{\lambda^2 (b-a)^2}{8}$$

Let $x \in [a,b]$, $a < b$

Notice $1 - \frac{b-x}{b-a} = \frac{x-a}{b-a}$

By convexity on the e^t

$$e^{\underbrace{\frac{b-x}{b-a} ta + \frac{x-a}{b-a} tb}_{= tx}} \le \frac{b-x}{b-a} e^{ta} + \frac{x-a}{b-a} e^{tb}$$

Because X is centered we can write

$$\mathbb{E}e^{t\cancel{X}} \le \frac{b \overset{=0}{-\mathbb{E}X}}{b-a} e^{ta} - \frac{\overset{=0}{\mathbb{E}X} - a}{b-a} e^{tb} =: \boxed{e^{g(t)}}$$

The first and second derivatives of $g(t)$ equal to

$$g'(t) = a - a\frac{b-a}{be^{-t(b-a)} - a}$$

and

$$g''(t) = -\frac{ab(b-a)^2 e^{-t(b-a)}}{(be^{-t(b-a)} - a)^2}$$

That we can rewrite (exercise)

$$g'' : u \in [0,1] \mapsto (b-a)^2 u(1-u) \le \frac{(b-a)^2}{4} .$$

Because $g(0) = g'(0) = 0$, by Taylor's Theorem applied to $g(u)$ of order 2 at 0, there exists $t \in [0, \lambda]$ such that

$$\boxed{\psi_X(\lambda) = \frac{\lambda^2}{2} g''(t) \le \frac{t^2(b-a)^2}{8}}.$$

Remark

Notice that in that case X is sub-Gaussian of parameter $\nu = (b-a)^2/4$, i.e., $X \in \mathcal{G}((b-a)^2/4)$.

Theorem (Hoeffding tail inequality, 1963)

Let $X_1, \dots, X_n$, a sequence of n independent r.v., s.t. $a_i \le X_i \le b_i$ a.s., with $(a_i, b_i) \in \mathbb{R}^2$, for all $i \le n$. Then, for all $t > 0$,

$$\mathbb{P}\left\{\frac{1}{n}\left|\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])\right| \ge t\right\} \le \underbrace{2e^{-2t^2 \overset{n^2}{\wedge} / c^2}}, \quad \sigma^2 \tag{13}$$

with $c^2 = \sum_{i=1}^{n}(b_i - a_i)^2$.

Theorem 11 is a simple consequence of the Hoeffding's inequality recalled in Lemma 10, combined with Cramér-Chernoff's bound illustrated in the subsequent proof.

Proof.

Let $t > 0$, $\lambda > 0$ and consider the centered r.v. $\boxed{Z = \sum_{i=1}^{n} (X_i - \mathbb{E}[X_i])}$, with $n \in \mathbb{N}^*$. Then, using sequentially Cramér-Chernoff's method, Hoeffding's Lemma and the independence of the X_is, one has

$\lambda \mapsto e^{\lambda Z}$

$$\begin{aligned}
\mathbb{P}\{Z \geq t\} &= \mathbb{P}\left\{e^{\lambda Z} \geq e^{\lambda t}\right\} \\
&\leq e^{-\lambda t}\mathbb{E}[e^{\lambda Z}] = \mathbb{E}\left[e^{\lambda \Sigma (X_i - EX_i)}\right] = \mathbb{E}\left[\prod e^{\lambda (X_i - EX_i)}\right] \\
&\leq \exp\left\{-\lambda t + \frac{\lambda^2 t^2 c^2}{8}\right\} \\
&\leq \inf_{\lambda > 0} \underbrace{\exp\left\{-\lambda t + \frac{\lambda^2 t^2 c^2}{8}\right\}} .
\end{aligned}$$

MI

$c = \Sigma (b_i - a_i)^2 = \prod_{i \leq n} \mathbb{E}\left[e^{\lambda (X_i - EX_i)}\right]$

The bound is obtained with the optimal parameter $\boxed{\lambda^* = 4t/c^2.}$ □

$$\mathbb{P}\left(\left|\sum X_i - \mathbb{E}X_i\right| \geq t\right) \leq \underbrace{2e^{-\frac{2t^2}{c^2}}}_{=\delta} \quad , \quad c^2 = \sum_{i\leq n}(b_i-a_i)^2$$

- To better understand what it encompasses, define $\delta = 2\exp\{-2t^2/c^2\}$.
- Then with probability at least $1-\delta$, it is possible to control almost surely the deviation of the sample mean w.r.t. its expectation by inverting Eq. (13) as follows

LLN, CLT

$$\boxed{\frac{1}{n}\sum X_i \underset{n\to\infty}{\sim} EX}$$

$$\boxed{\underbrace{\frac{1}{n}\left|\sum_{i=1}^{n}(h(X_i)-\mathbb{E}[h(X_i)])\right|}\leq c\underbrace{\sqrt{\frac{\log(2/\delta)}{2n}}}}. \qquad O_P(1/\sqrt{n}) \qquad (14)$$

$\Longrightarrow$ **Compare it to Chebychev's inequality**

- To better understand what it encompasses, define $\delta = 2\exp\{-2t^2/c^2\}$.
- Then with probability at least $1-\delta$, it is possible to control almost surely the deviation of the sample mean w.r.t. its expectation by inverting Eq. (13) as follows

$$\frac{1}{n}\left|\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])\right| \le c\sqrt{\frac{\log(2/\delta)}{2n}}\,. \qquad (14)$$

$= \sum_{1\le i\le n}(b_i - a_i)^2$

$\Longrightarrow$ **Compare it to Chebychev's inequality**

- This bound expresses the importance of the spread effect for obtaining a *good* estimation of the expectation.
- It also provides an explicit bound, independent on the distribution of the sample, for which it is possible to exactly determine the sample size n required for the probabilistic control of the empirical bias.

Before deriving two important inequalities that refine that of Hoeffding, when the variance is small, we expose two important applications for Hoeffding's inequality.

(X, Y) label

input r.v

$Y \in \{0,1\}$

find the best classifier

$h : \mathcal{X} \to \{0,1\}$

$x \mapsto h(x) \in \{\text{cat}, \text{dog}\}$

$h^* \in \operatorname{argmin}_{h \in \mathcal{H}} \mathbb{E}[\mathbb{1}\{h(X) \neq Y\}]$

$\hat{h} \in \operatorname{argmin}_{h \in \mathcal{H}} \frac{1}{n} \sum_{i \leq n} \mathbb{1}\{h(X_i) \neq Y_i\}$ ←

Example (Binary Classification)

Considering the binary loss, and choosing $\underbrace{Z_i = 1\{g(X_i) \neq Y_i\}}$, yields with probability at least $1 - \delta$

$$\sup_{g \in G} \left| \frac{1}{n} \sum_{i=1}^{n} (Z_i - \mathbb{E}[Z_i]) \right| \leq \boxed{\sqrt{\frac{\log(2/\delta)}{2n}}} . \quad \text{rate } \mathcal{O}_{\mathbb{P}}(1/\sqrt{n}) \tag{15}$$

The tail bound is of order $\mathcal{O}_{\mathbb{P}}(n^{-1/2})$ that is classic for empirical estimators and processes, as will be shown throughout the chapters.

Example (Simple Rademacher averages)

Let an i.i.d. sequence of Rademacher variables $\varepsilon_1, \ldots, \varepsilon_n$ (symmetric and $\mathbb{P}(\varepsilon = 1) = \mathbb{P}(\varepsilon = -1) = 1/2$), and a sequence of real constants $a_1, \ldots, a_n$. Then

$$\mathbb{P}\left\{\left|\sum_{i=1}^{n} \varepsilon_i a_i\right| \geq t\right\} \leq 2e^{\underbrace{-t^2/2(\sum_{i\leq n} a_i^2)^2}} , \tag{16}$$

notice that $\boxed{\mathrm{Var}(\sum_{i=1}^{n} \varepsilon_i a_i) = \sum_{i=1}^{n} a_i^2.}$

Example (Simple Rademacher averages)

Let an i.i.d. sequence of Rademacher variables $\varepsilon_1, \dots, \varepsilon_n$ (symmetric and $\mathbb{P}(\varepsilon = 1) = \mathbb{P}(\varepsilon = -1) = 1/2$), and a sequence of real constants $a_1, \dots, a_n$. Then

$$\mathbb{P}\left\{\left|\sum_{i=1}^{n} \varepsilon_i a_i\right| \geq t\right\} \leq 2e^{-t^2/2(\sum_{i\leq n} a_i^2)^2} \;, \quad \leftarrow \; \mathbb{P}(|R_n| \geq t \mid X_1, \dots, X_n) \qquad (16)$$

notice that $\mathrm{Var}(\sum_{i=1}^{n} \varepsilon_i a_i) = \sum_{i=1}^{n} a_i^2$.

Example (Rademacher averages)

Let an i.i.d. sequence of Rademacher variables $\varepsilon_1, \dots, \varepsilon_n$, independent of the X's. The Rademacher average is defined by

$$\boxed{R_n = \sum_{i=1}^{n} \varepsilon_i X_i \;,} \leftarrow \qquad R_n(\underset{\substack{\in \\ \mathcal{H}}}{h}) = \sum_{i\leq n} \overset{\swarrow}{\varepsilon_i} \underbrace{h(X_i)} \qquad (17)$$

Hoeffding's inequality yields

$$\boxed{\mathbb{P}\{|R_n| \geq t\} \leq 2e^{-t^2/2c^2}}. \qquad (18)$$

Notice that Hoeffding's inequality does not characterize sub-Gaussian class, as the variance can be smaller than $\sum_{i=1}^{n}(b_i - a_i)^2$. In the case of Rademacher averages however, it exactly equals to the sample variance. As we will see later, Rademacher averages are key quantities that can be used to measure the *richness/complexity/size* of classes of functions.

Bennett's and Bernstein's inequalities.

- When higher moments of the r.v.s are small, and especially the variance, then sharper bounds can be obtained.
- We start by deriving a sub-Gaussian inequality, namely Bennett's inequality, that will help us prove Bernstein's under weaker assumptions on the r.v.s.

Bennett's and Bernstein's inequalities.

- When higher moments of the r.v.s are small, and especially the variance, then sharper bounds can be obtained.
- We start by deriving a sub-Gaussian inequality, namely Bennett's inequality, that will help us prove Bernstein's under weaker assumptions on the r.v.s.
- Before stating the theorems, we highlight a simple inequality bounding the moment-generating function. Let the sequence of $\underline{\text{independent r.v.s be } X_1, \dots, X_n}$, define $\underline{Z = \sum_{i=1}^n (X_i - \mathbb{E}X_i)}$.
  Then for any $\underline{\lambda > 0}$, we have:

$$\underline{\psi_Z(\lambda)} = \boxed{\sum_{i=1}^n \log \mathbb{E}e^{\lambda X_i} - \lambda \mathbb{E}X_i} \leq \sum_{i=1}^n (\underbrace{\mathbb{E}e^{\lambda X_i}}_{\text{Series expansion of the } e^{\lambda t}} \underline{- \lambda \mathbb{E}X_i - 1}) \ , \tag{19}$$

using that $\boxed{\log u \leq u - 1}$.

We will define in the following the function

$$h : u > 0 \mapsto (1+u)\log(1+u) - u$$

Bennett's inequality

Theorem

Let $X_1, \dots, X_n$, a sequence of square-integrable and independent r.v.s. Suppose there exists a constant $b > 0$ that bounds $X_i \leq b$ a.s. Then, for all $t > 0$,

$$\mathbb{P}\left\{\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) \geq t\right\} \leq \exp\left\{-\frac{\nu}{b^2}h\left(\frac{bt}{\nu}\right)\right\} \ , \qquad \nu/b \tag{20}$$

where $\nu = \sum_{i \leq n} \mathbb{E}X_i^2$ and $h(u) = (1+u)\log(1+u) - u$. for $u > 0$.

Proof.

Exercise. □

Bernstein's inequality

The following version of Bernstein's inequality **only requires bounded moments for the r.v.s.**

Theorem

Let $X_1, \ldots, X_n$, a sequence of n independent r.r.v.. Suppose that there exist the nonnegative constants ν, c such that

$$\boxed{\sum_{i \le n} \mathbb{E} X_i^2 \le \nu}$$

and

$$\boxed{\sum_{i=1}^{n} \mathbb{E}[(X_i)_+^q] \le \frac{q!}{2} \nu c^{q-2} \ ,}$$

for all integers $q \ge 3$. Then, for all $t \ge 0$,

$$\boxed{\mathbb{P}\left\{ \sum_{i=1}^{n} (X_i - \mathbb{E}[X_i]) \ge \sqrt{2\nu t} + ct \right\} \le e^{-t} \ .} \qquad (21)$$

And,

$$\boxed{\mathbb{P}\left\{ \sum_{i=1}^{n} (X_i - \mathbb{E}[X_i]) \ge t \right\} \le \exp\left\{ -\frac{t^2}{2(\nu + ct)} \right\} \ .} \qquad (22)$$

Exercise.

Denote by $Z = \sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])$. Show that for all $\lambda \in (0, 1/c)$ and $t > 0$,

$$\psi_Z(\lambda) = \frac{\nu\lambda^2}{2(1 - c\lambda)}$$

and that

$$\psi_Z(t)^* \geq \frac{\nu}{c^2} h_1\left(\frac{ct}{\nu}\right)$$

with $h_1 : u > 0 \mapsto 1 + u - \sqrt{1 + 2u}$. Use that $\log u \leq u - 1$, $u > 0$ to conclude.

□

Remark (Connexion between Bennett's and Bernstein's inequalities)

How to interpret this inequality? Show that

$$\boxed{h(u) \geq \frac{u^2}{2(1+u/3)}} \tag{23}$$

and conclude

$$\mathbb{P}\left\{\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) \geq t\right\} \leq \underbrace{\boxed{\exp\left\{-\frac{t^2}{2\nu + bt/3}\right\}}}_{\nu \gg b} . \quad \sim e^{-\frac{t^2}{2\nu}} \tag{24}$$

- Notice that both inequalities provide sub-Gaussian type of inequalities as soon as the variance ν is the dominant term in the denominator.
- Bennett's inequality can be seen as a version of Bernstein's inequality with strong assumptions on the r.v.s. as we shall see in the following version of Bernstein's inequality.

Bernstein's inequality for bounded variables

Lemma

Assume the sequence of independent r.v.s $X_1, \ldots, X_n$, $\mathbb{E}X_i = 0$, *such that there exists a constant* $c > 0$ *bounding the observations* $|X_i| \leq c$ *a.s., for all* $i \leq n$, *and of finite variance* $\nu^2 = \sum \mathrm{Var}(X_i)$. *Then, for any* $t > 0$,

$$\boxed{\mathbb{P}\left\{\left|\sum_{i=1}^{n} X_i\right| \geq t\right\} \leq 2\exp\left\{-\frac{t^2}{2\nu^2 + 2ct/3}\right\}}.$$

Proof.

Exercise, hint: use moment series decomposition of the moment-generating function. □

Today's outline

1. Basic methods for bounding sums of independent variables

2. Bounding the variance general functionals of independent variables

- We proved a series of fundamental tail bounds applied to sums of independent real-valued r.v. in particular.
- A necessary condition was to be able to bound a specific order of their moment by a constant, and the variance in particular.
- Of course, one could use Chebychev's inequality under the later assumption. We will see now, how to obtain exponential bounds, by using a very simple trick based on functions with *bounded differences* (it is an application of the *Efron-Stein inequality*).
- We define first functions with bounded differences, that we will apply to functions of the independent data of the very general form

$$\boxed{Z_n = h(X_1, \ldots, X_n)}. \qquad X_i\text{'s are independent}$$

Functions with Bounded differences

Definition

Let $\mathcal{X}$ a measurable set and $h : \mathcal{X}^n \to \mathbb{R}$ be a measurable function of n variables.
The function h satisfies the *bounded differences inequality*, if for the real constants $c_1, \dots, c_n$ and for all $i \leq n$, we have

$$\sup_{x_1,\dots,x_n,x_i' \in \mathcal{X}} |h(\cancel{x}_1, \dots, \cancel{x}_n) - h(\cancel{x}_1, \dots, \cancel{x}_i', \dots, \cancel{x}_n)| \leq c_i \ . \tag{25}$$

$$h(x_1,\dots,x_{i-1},\boxed{x_i},x_{i+1},\dots,x_n)$$

$$\updownarrow$$

$$h(\quad '' \quad , x_i', \quad '' \quad)$$

McDiarmid's inequality

Lemma (McDiarmid (89,98))

Let $X_1, \ldots, X_n$ a sequence of independent r.v.s valued in $\mathcal{X}$. Consider a function h satisfying the bounded difference inequality, with constants $c_1, \ldots, c_n \in \mathbb{R}$. Define $Z = h(X_1, \ldots, X_n)$, then for all $t > 0$:

$$\mathbb{P}\underbrace{\{|Z - \mathbb{E}[Z]| \geq t\}}_{} \leq 2e^{-2t^2/c^2}, \qquad (26)$$

$\uparrow$ “ σ^2 ” “ $\sum_{i \leq n}(b_i - a_i)^2$ ”

where $\boxed{c^2 = \sum_{i=1}^n c_i^2}$.

Proof.

We define a sequence $Y_1, \dots, Y_n$ of r.v.s by

$$\boxed{Y_i = \mathbb{E}[Z|X_1, \dots, X_i] =: \mathbb{E}_i[Z],} \quad i \leq n$$

such that the Y_i is a martingales w.r.t. the filtration induced by $\sigma(X_1, \dots, X_i)$.

Notice that by writing $\boxed{\Delta_i = \mathbb{E}_i[Z] - \mathbb{E}_{i-1}[Z],}$ the r.v. is centered. $\boxed{Z - \mathbb{E}Z = \sum_{i \leq n} \Delta_i.}$

Proof.

We define a sequence $Y_1, \dots, Y_n$ of r.v.s by

$$Y_i = \mathbb{E}[Z | X_1, \dots, X_i] =: \mathbb{E}_i[Z], \quad i \leq n$$

such that the Y_i is a martingales w.r.t. the filtration induced by $\sigma(X_1, \dots, X_i)$.

Notice that by writing $\Delta_i = \mathbb{E}_i[Z] - \mathbb{E}_{i-1}[Z]$, the r.v. is centered, $Z - \mathbb{E}Z = \sum_{i \leq n} \Delta_i$.

Let us check that the Δ's are bounded using the bounded difference assumption as follows. Fix the index i and write conditionally on the set $X_1 = x_1, \dots, X_{i-1} = x_{i-1}$, the Δ_i is a function of the ith r.v. X_i. Thus, for $x \in \mathcal{X}$

$$\begin{aligned}
|\Delta_i| &= |\overbrace{\mathbb{E}[Z|X_1 = x_1, \dots, X_i = x]} - \overbrace{\mathbb{E}[Z|X_1 = x_1, \dots, X_{i-1} = x_{i-1}]}| \\
&= \big|\mathbb{E}[h(\underbrace{x_1, \dots, x}_{i\text{'s}}, X_{i+1}, \dots, X_n) - h(\underbrace{x_1, \dots,}_{(i-1)} X_i, X_{i+1}, \dots, X_n)]\big| \\
&\leq \mathbb{E}[|h(x_1, \dots, x, X_{i+1}, \dots, X_n) - h(x_1, \dots, X_i, X_{i+1}, \dots, X_n)|] \\
&\leq \mathbb{E}\left[\sup_{x_1, \dots, x_n, x \in \mathcal{X}} |h(x_1, \dots, x, X_{i+1}, \dots, X_n) - h(x_1, \dots, X_i, X_{i+1}, \dots, X_n)|\right] \\
&\leq c_i. \qquad (27)
\end{aligned}$$

It remains to apply Hoeffding's inequality in Theorem 13.

□

What we saw today

- We studied how to derive probabilistic bounds to quantify the speed of the deviation of averages of r.v.s w.r.t their mean, known as basic *concentration inequalities*.
- We focused on bounds that have exponential decay, under various assumptions on the moments of the r.v.s
- We saz how those results allow to assess how averages concentrate around their mean for **fixed** sample size.

What we will see next week

$$\underbrace{\sup_{h \in \mathcal{H}} | P_n h - Ph |}_{\text{measurable}}$$

- We have studied so far how averages deviate from their mean when we fix the functional. The next chapter will provide us with sufficient theoretical tools to extend those results **uniformly** over a class of functionals/sets.
- We will focus on extending the notion of weak convergence for stochastic processes, that will result in a new set of characterizations requiring the definition of *outer-measure*.
- Next week, we will go from separable finite-dimensional metric spaces to non-separable metric spaces.

Main reference

- Stéphane Boucheron, Gábor Lugosi, and Pascal Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence.* Oxford University Press, Oxford Academic, 2013.