Randomized Matrix Computations Lecture 2

Daniel Kressner
Chair for Numerical Algorithms and HPC
Institute of Mathematics, EPFL
daniel.kressner@epfl.ch

Schedule for semester

- Oct 4–Oct 18: Homework 1
- Nov 8–Nov 22: Homework 2
- Around Nov 22: Project assignment and start of work on projects

This lecture

- Todo from last time: Uniform distribution on sphere and power method, random matrices.
- Expectation
- Moments and tail bounds

Expectation

- Definition and basic properties
- Expectation and convexity

Literature:

Tropp'2023 Joel A. Tropp. *Probability Theory & Computational Mathematics*, Lecture notes, Caltech, 2023.

pdf available on Moodle

Expectation: Definition

Theorem. Given probability space $(\Omega, \mathcal{F}, \mathbb{P})$, the expectation of a real *non-negative* random variable $X : \Omega \to \mathbb{R}$ is defined as

$$\mathbb{E}[X] := \int_0^\infty \mathbb{P}(X > t)\,\mathrm{d}t.$$

- For non-negative X, the case $\mathbb{E}[X] = \infty$ is usually admitted.
- If X is not non-negative, we will always assume that X is integrable, that is, $\mathbb{E}[|X|] < \infty$. Then we define

$$\mathbb{E}[X] = \mathbb{E}[X_+] - \mathbb{E}[X_-], \quad X_+ = \max\{0, X\}, \quad X_- = -\min\{0, X\}.$$

- Recall that $F_X(t) = \mathbb{P}(X \leq t)$. Thus,

$$\mathbb{E}[X] = \int_0^\infty (1 - F_X(t))\,\mathrm{d}t - \int_{-\infty}^0 F_X(t)\,\mathrm{d}t.$$

Expectation: Simpler formulas

- For a discrete random variable X with measure $\mu_X = \sum_{i=1}^n p_i \delta_{a_i}$, a rearrangement of summation gives

$$\mathbb{E}[X] = \sum_{i=1}^{n} a_i p_i.$$

- For a continuous random variable X with density f_X, a change of variable gives

$$\mathbb{E}[X] = \int_{\mathbb{R}} x \cdot f_X(x)\,\mathrm{d}x.$$

Both formulas can be unified by using the Lebesgue integral wrt probability measure of X:

$$\mathbb{E}[X] = \int_{\mathbb{R}} x\,\mu_X(\mathrm{d}x).$$

Defying expectations: Cauchy random variables

EFY: Consider independent $X, Y \sim N(0,1)$. Show that $Z = X/Y$ has the pdf

$$f_Z(Z) = \frac{1}{\pi(1+x^2)}$$

on $\mathbb{R}$. Good luck!

This function is not integrable and, hence, the expectation of Z is not defined.

Z is the canonical example of a Cauchy random variable.

Expectation: Law of the unconscious statistician

Let $h : \mathbb{R} \to \mathbb{R}$ be measurable such that $\mathbb{E}[h(X)]$ is well defined.

- For a discrete random variable X with measure $\mu_X = \sum_{i=1}^n p_i \delta_{a_i}$, a rearrangement of summation gives

$$\mathbb{E}[X] = \sum_{i=1}^{n} h(a_i) p_i .$$

- For a continuous random variable X with density f_X, a change of variable gives

$$\mathbb{E}[X] = \int_{\mathbb{R}} h(x) \cdot f_X(x) \, \mathrm{d}x .$$

Both formulas can be unified by using the Lebesgue integral wrt probability measure of X:

$$\mathbb{E}[X] = \int_{\mathbb{R}} h(x) \, \mu_X(\mathrm{d}x) .$$

$\mathbb{E}[h(X)]$ is also called a moment (often reserved for $h(x) = x^p$ for $p \in \mathbb{N}$).

Properties of expectation

For integrable real random variables X, Y (on the same probability space, but not necessarily independent), the following hold:

1. If $X \leq Y$ (almost surely) then $\mathbb{E}[X] \leq \mathbb{E}[Y]$.
2. If $X = Y$ (almost surely) then $\mathbb{E}[X] = \mathbb{E}[Y]$.
3. If X is non-negative and $\mathbb{E}[X] = 0$ then $X = 0$ (almost surely).
4. $\mathbb{E}[\alpha X + \beta Y] = \alpha\mathbb{E}[X] + \beta\mathbb{E}[X]$ for every $\alpha, \beta \in \mathbb{R}$,
5. $\mathbb{E}[X \cdot Y] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$ if X, Y *are* independent.

These properties follow from basic properties of Lebesgue integrals, except for the last one (which follows from Fubini).

EFY: If $a \leq X \leq b$ then $a \leq \mathbb{E}[X] \leq b$.

This is the basis of another flavor of the probabilistic method (see Exercises 1).

Simple examples

- For $X \sim N(0,1)$, $\mathbb{E}[X] = 0$ and $\mathbb{E}[X^2] = 1$.
  What is $\mathbb{E}[X^p]$ for general $p \in \mathbb{N}$?
- For Rademacher X, $\mathbb{E}[X] = 0$ and $\mathbb{E}[X^2] = 1$.
  What is $\mathbb{E}[X^p]$ for general $p \in \mathbb{N}$?
- For independent $X, Y \sim N(0,1)$, $\mathbb{E}[X \cdot Y] = 0$
- EFY: Let X be either a Gaussian or a Rademacher random vector. Show that

  $$\mathbb{E}[X^T A X] = \mathrm{trace}(A) = a_{11} + a_{22} + \cdots + a_{nn}$$

  for fixed $A \in \mathbb{R}^{n \times n}$. What are the decisive properties of X used in your arguments?

Expectation and convexity

Let $I \subseteq \mathbb{R}$ be an interval (finite or infinite). Then $\varphi : I \to \mathbb{R}$ is called convex if

$$\varphi((1-\tau)x + \tau y) \leq (1-\tau)\varphi(x) + \tau\varphi(y), \quad \forall \tau \in [0,1],\ x, y \in I.$$

φ is called concave if $-\varphi$ is convex.

EFY: Recap examples of convex and concave functions.

An important property of a convex function $\varphi : I \to \mathbb{R}$ on an open interval I is:

$$\varphi(y) \geq \varphi(a) + \varphi'(a) \cdot (y - a), \quad \forall a, y \in I,$$

provided that φ is differentiable at a.[1]

[1] If φ is not differentiable at a, the formula still holds with $\varphi'(a)$ replaced by a subgradient of φ at a.

Jensen’s inequality for random variable

Theorem. Let $\varphi : I \to \mathbb{R}$ be convex on an open interval $I \subseteq \mathbb{R}$ and bounded from below. Let X be an integrable, real random variable that takes values in I. Then

$$\mathbb{E}[\varphi(X)] \geq \varphi(\mathbb{E}[X]).$$

Proof. For simplicity, suppose that φ is differentiable. Setting $y = X(\omega)$ and $a = \mathbb{E}[X]$ in the “important property” gives

$$\varphi(X(\omega)) \geq \varphi(\mathbb{E}[X]) + \varphi'(\mathbb{E}[X]) \cdot (X(\omega) - \mathbb{E}[X]), \quad \forall \omega \in \Omega.$$

Taking expectations on both sides completes the proof. $\diamond$

Two important examples:

$$\mathbb{E}[X^2] \geq \big(\mathbb{E}[X]\big)^2, \quad \mathbb{E}[\exp(X)] \geq \exp\big(\mathbb{E}[X]\big).$$

Expectation of random vectors

For a random vector $X = [X_1, \dots, X_n]^\top \in \mathbb{R}^n$, expectation $\mathbb{E}[X]$ is simply defined entry-wise:

$$\mathbb{E}[X] = \int_{\mathbb{R}^n} x \mu_X(\mathrm{d}x) = \begin{bmatrix} \int_{\mathbb{R}^n} x_1 \mu_X(\mathrm{d}x) \\ \vdots \\ \int_{\mathbb{R}^n} x_n \mu_X(\mathrm{d}x), \end{bmatrix} = \begin{bmatrix} \mathbb{E}_{X_1}[X_1] \\ \vdots \\ \mathbb{E}_{X_n}[X_n] \end{bmatrix}.$$

where $\mathbb{E}_{X_j}$ denotes expectation wrt the marginal distribution of X_j. Properties like linearity are thus inherited directly from the scalar case.

Law of the unconscious statistician: For a multivariate measurable function $h : \mathbb{R}^n \to \mathbb{R}$, it holds that

$$\mathbb{E}[h(X)] = \int_{\mathbb{R}^n} h(x) \mu_X(\mathrm{d}x).$$

In particular for a continuous random vector with joint density f_X, we have

$$\mathbb{E}[h(X)] = \int_{\mathbb{R}^n} h(x) f_X(x) \mathrm{d}x.$$

Jensen's inequality for random vector

Recall that $\varphi : C \to \mathbb{R}$ on convex set C is called convex if φ is convex along any line in C.

Theorem. For a convex function $\varphi : C \to \mathbb{R}$ bounded from below on a convex open set C, we have

$$\mathbb{E}[\varphi(X)] \geq \varphi(\mathbb{E}[X]).$$

Example: $\mathbb{E}[\|X\|_2^2] \geq \|\mathbb{E}[X]\|_2^2$.

Proof. Assuming that φ is differentiable at $a \in C$, we have the "important" property

$$\varphi(y) \geq \varphi(a) + \nabla\varphi(a)^T(y - a), \quad \forall a, y \in C,$$

provided that φ is differentiable at a. Taking expectations on both sides for $a = \mathbb{E}[X]$ again completes the proof. ◇

Moments and Tails

- From moments to tails
- From tails to moments
- Subgaussian random variables
- Sub-exponential random variables

Literature:

Tropp'2023 Joel A. Tropp. *Probability Theory & Computational Mathematics*, Lecture notes, Caltech, 2023.

Vershynin'2018 Roman Vershynin. *High-Dimensional Probability*, CUP, 2018.

Wainwright'2019 Martin J. Wainwright. *High-Dimensional Statistics*, CUP, 2019.

pdf available on Moodle

Types of moments

- Polynomial moments:

$$\mathbb{E}[X^n] = \int_{\mathbb{R}} x^n \mu_X(\mathrm{d}x), \quad n = 0, 1, 2, \ldots$$

  assuming that the expectation is well-defined.

- Exponential moments:

$$\mathbb{E}[\exp(\theta X)] = \int_{\mathbb{R}} e^{\theta x} \mu_X(\mathrm{d}x), \quad \theta \in \mathbb{R}.$$

- If the polynomial moments do not grow too quickly, $\mathbb{E}[\exp(\theta X)]$ is finite and

$$\mathbb{E}[\exp(\theta X)] = \sum_{n=0}^{\infty} \frac{\theta^n}{n!} \mathbb{E}[X^n],$$

  $\theta \mapsto E[\exp(\theta X)]$ is called the moment generating function (mgf).

Types of tails

There are different types of tails: For a real random variable X:

- Right tail probability $\mathbb{P}\{X \geq t\}$
- Left tail probability $\mathbb{P}\{X \leq t\}$
- Two-sided tail probability $\mathbb{P}\{|X| \geq t\}$

Often, it is convenient to first center the random variable, that is, consider $X - \mathbb{E}X$ instead of X.

In the following we will see: There is a direct relation between the polynomial moments and tail bounds.

Markov's inequality

Theorem. For a real nonnegative random variable X, it holds that

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

Note: Only expected value needed, but bound usually quite poor.
Proof for continuous r.v. Using that $x/t \geq 1$ for $x \geq t$ we obtain

$$\begin{aligned}\mathbb{P}\{X \geq t\} &= \int_t^\infty f_X(x)\,\mathrm{d}x \leq \int_t^\infty \frac{x}{t} f_X(x)\,\mathrm{d}x \\ &\leq \int_0^\infty \frac{x}{t} f_X(x)\,\mathrm{d}x = \mathbb{E}[X/t] = \frac{\mathbb{E}[X]}{t}.\end{aligned}$$

Boosting Markov's inequality

Let $\varphi : \mathbb{R} \to \mathbb{R}$ be an *increasing, non-negative* function. Then $X \geq t$ implies $\varphi(X) \geq \varphi(t)$. This implies $\mathbb{P}\{X \geq t\} \leq \mathbb{P}\{\varphi(X) \geq \varphi(t)\}$. Applying Markov's inequality to the rhs (with X / t replaced by $\varphi(X)$ / $\varphi(t)$) gives

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

Three important cases:

- $\varphi(x) = (x - \mathbb{E}X)^2 \rightsquigarrow$ Chebyshev's inequality:

$$\mathbb{P}\{|X - \mathbb{E}X| \geq t\} \leq \frac{\mathrm{Var}[X]}{t^2}, \quad \mathrm{Var}[X] = \mathbb{E}\big[(X - \mathbb{E}X)^2\big]$$

- $\varphi(x) = |x|^p$ for $p \in \mathbb{R}, p \geq 1 \rightsquigarrow$

$$\mathbb{P}\{|X| \geq t\} \leq \frac{\mathbb{E}[|X|^p]}{t^p}, \quad \forall t > 0.$$

  $\exists$ polynomial moment $\Rightarrow$ polynomial decay

- mgf → Chernoff (later)

Polynomial decay ⇒ ∃ polynomial moment

Theorem. Let X be non-negative real r.v. and $\varphi : \mathbb{R}_+ \to \mathbb{R}$ increasing, cont. differentiable. Then

$$\mathbb{E}[\varphi(X)] = \varphi(0) + \int_0^\infty \mathbb{P}\{X \geq t\}\varphi'(t)\,\mathrm{d}t.$$

Proof for cont. r.v.

$$\begin{aligned}
\mathbb{E}[\varphi(X)] &= \int_0^\infty \varphi(x) f_X(x)\,\mathrm{d}x = \varphi(0) + \int_0^\infty (\varphi(x) - \varphi(0)) f_X(x)\,\mathrm{d}x \\
&= \varphi(0) + \int_0^\infty \int_0^x \varphi'(t) f_X(x)\,\mathrm{d}t\,\mathrm{d}x \\
&= \varphi(0) + \int_0^\infty \int_t^\infty \varphi'(t) f_X(x)\,\mathrm{d}x\,\mathrm{d}t \\
&= \varphi(0) + \int_0^\infty \mathbb{P}\{X \geq t\}\varphi'(t)\,\mathrm{d}t.
\end{aligned}$$

◇

Polynomial decay ⇒ ∃ polynomial moment

Suppose that the r.v. X has polynomial decay, that is, there is a constant C s.t.

$$\mathbb{P}\{|X| \geq t\} \leq C \cdot t^{-p}, \quad \forall t > 0.$$

By the theorem, this implies for every $q < p$:

$$\begin{aligned}
\mathbb{E}[|X|^q] &= \int_0^\infty \mathbb{P}\{X \geq t\} q t^{q-1} \, \mathrm{d}t \\
&\leq \int_0^1 1 \cdot q t^{q-1} \, \mathrm{d}t + \int_1^\infty \mathbb{P}\{X \geq t\} q t^{q-1} \, \mathrm{d}t \\
&\leq 1 + \int_1^\infty C q t^{q-p-1} \, \mathrm{d}t = 1 + \frac{Cq}{p-q}.
\end{aligned}$$

An excursion to L_p spaces

If $\mathbb{E}[|X|^p] < \infty$, we say that $X \in L_p$ and compute the corresponding semi-norm as

$$\|X\|_{L_p} := \left(\mathbb{E}[|X|^p]\right)^{1/p}.$$

We require $p \geq 1$ but not that p is an integer. Important properties for r.v. X, Y:

- $p \leq q$ implies $\|X\|_p \leq \|X\|_q$ (monotonicity, consequence of Jensen)
- $\mathbb{E}[|XY|] \leq \|X\|_p \|Y\|_q$ for $p^{-1} + q^{-1} = 1$ (Hölder)
- $\|X + Y\|_p \leq \|X\|_p + \|Y\|_p$ (Minkowski)
- If $\|X\|_p = 0$ then $X = 0$ almost surely.

For $X, Y \in L_2$ we can define an pseudo-inner product $\langle X, Y\rangle := \mathbb{E}[XY]$. The covariance is defined as

$$\mathrm{Cov}(X, Y) = \langle X - \mathbb{E}X, Y - \mathbb{E}Y\rangle = \mathbb{E}[XY] - (\mathbb{E}X)(\mathbb{E}Y).$$

If $\mathrm{Cov}(X, Y) = 0$ then X, Y are called uncorrelated. “independent” implies “uncorrelated” but not vice versa!

Chernoff's inequalities

- Setting $\varphi(x) = e^{\theta X}$, $\theta > 0$, in the boosted Markov inequality gives

$$\mathbb{P}\{X \geq t\} \leq \mathbb{E}[\exp(\theta X)]e^{-\theta t}, \quad t \in \mathbb{R}.$$

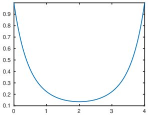
As this holds for any $\theta > 0$,

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

This is Chernoff's inequality. It requires access to the mgf (or good bounds for it) and an optimal/good choice of θ.

Example: For $X \sim N(0, \sigma^2)$, we have $\mathbb{E}[\exp(\theta X)] = e^{\sigma^2\theta^2/2}$. Chernoff gives:

$$\mathbb{P}\{X \geq t\} \leq \inf_{\theta > 0} \exp(\sigma^2\theta^2/2 - \theta t)$$



By differentiating, optimal $\theta_* = t/\sigma^2 \rightsquigarrow$

$$\mathbb{P}\{X \geq t\} \leq \exp(-t^2/(2\sigma^2)).$$

Quadratic-exponential decay! Nearly optimal bound (up to factor $1/(\sqrt{2\pi}t)$ – Mills inequality).

Sub-Gaussian random variables

Definition A real r.v. X is called sub-Gaussian with parameter $\sigma > 0$ if

$$\mathbb{E}\Big[\exp\big(\theta(X - \mathbb{E}X)\big)\Big] \le e^{\sigma^2\theta^2/2}, \quad \forall \theta \in \mathbb{R}.$$

We just saw that $X \sim N(0, \sigma^2)$ is sub-Gaussian with parameter σ and obtained a tail bound. By the same arguments, *any* sub-Gaussian r.v. X with parameter σ satisfies the same tail bound:

$$\mathbb{P}\{(X - \mathbb{E}X) \ge t\} \le \exp(-t^2/(2\sigma^2)). \tag{1}$$

If two-sided bound is needed: Union bound $\rightsquigarrow$

$$\mathbb{P}\{|X - \mathbb{E}X| \ge t\} \le 2\exp(-t^2/(2\sigma^2)).$$

The tail bound (1) is an equivalent characterization:
(1) implies sub-Gaussian(σ).

Properties of sub-Gaussians

- Additivity. EFY: Assume that X_1 is sub-Gaussian(σ_1) and X_2 is sub-Gaussian(σ_2). If X_1, X_2 are independent then

$$X_1 + X_2 \text{ is sub-Gaussian}(\sqrt{\sigma_1^2 + \sigma_2^2})$$

  If X_1, X_2 are not necessarily independent then

$$X_1 + X_2 \text{ is sub-Gaussian}(2\sqrt{\sigma_1^2 + \sigma_2^2})$$

- Moment characterization. If X is sub-Gaussian then there exists $\gamma \geq 0$ s.t.

$$\mathbb{E}[X^{2k}] \leq \frac{(2k)!}{2^k k!}\gamma \tag{2}$$

  Can be proven by majorization and using moments of Gaussian (see Exercises). The moment bound (2) is an equivalent characterization:
  (2) implies sub-Gaussian for some σ.

  EFY: Show that (2) is equivalent to

$$\|X\|_{L_p} \leq C\sqrt{p}, \quad p = 1, 2, \ldots,$$

  for some constant C.

Bounded random variables are sub-Gaussian

Let X be bounded, that is, X is supported on an interval $[a, b]$ with $-\infty < a < b < +\infty$. Then X is sub-Gaussian. To see this, assume w.l.o.g. that $\mathbb{E}X = 0$ and let Y be an independent copy of X. Then, using Jensen,

$$\begin{aligned} \mathbb{E}_X\big[\exp(\theta X)\big] &= \mathbb{E}_X\big[\exp(\theta(X - \mathbb{E}Y))\big] \leq \mathbb{E}_X\mathbb{E}_Y\big[\exp(\theta(X - Y))\big] \\ &= \mathbb{E}_{(X,Y)}\big[\exp(\theta(X - Y))\big] = \mathbb{E}_{(X,Y)}\mathbb{E}_\epsilon\big[\exp(\theta\epsilon(X - Y))\big], \end{aligned}$$

where ϵ is Rademacher and the symmetry of $X - Y$ implies that $X - Y$ and $\epsilon(X - Y)$ have the same distribution. Using the Taylor expansion of the exponential, it follows that

$$\mathbb{E}[e^{\alpha\epsilon}] = \frac{1}{2}(e^{-\alpha} + e^{\alpha}) \leq e^{\alpha^2/2}, \quad \forall \alpha \in \mathbb{R}.$$

Thus,

$$\mathbb{E}_X\big[\exp(\theta X)\big] \leq \mathbb{E}_{(X,Y)}\big[\exp(\theta^2(X - Y)^2/2)\big] \leq \exp(\theta^2(b - a)^2/2)$$

A more refined argument [Wainwright'2019] shows that X is, in fact, sub-Gaussian($(b - a)/2$).

Hoeffding's inequality

The power of sub-Gaussians shines when considering an independent sum

$$Y = X_1 + X_2 + \cdots + X_n, \text{where } X_j \text{ are independent sub-Gaussian}(\sigma_j)$$

By additivity, Y is sub-Gaussian$(\sqrt{\sigma_1^2 + \cdots + \sigma_n^2})$. The tail bound for sub-Gaussian implies:

Theorem (Hoeffding's inequality). For X_j defined above,

$$\mathbb{P}\Big\{ \sum_{j=1}^{n} (X_j - \mathbb{E}X_j) \geq t \Big\} \leq \exp\Big(- \frac{t^2}{2(\sigma_1^2 + \cdots + \sigma_n^2)} \Big).$$

For the special case when X_j are bounded in $[a, b]$, this implies

$$\mathbb{P}\Big\{ \sum_{j=1}^{n} (X_j - \mathbb{E}X_j) \geq t \Big\} \leq \exp\Big(- \frac{2t^2}{n(b-a)^2} \Big).$$

EFY: What is the implication of this result for a Rademacher sum $\sum_j \epsilon_j \alpha_j$?

Sub-exponential random variables

In contrast to the sum, the product of two sub-Gaussians is not sub-Gaussian but sub-exponential only.

Definition A real r.v. X is called sub-exponential with nonnegative parameters (ν, b) if

$$\mathbb{E}\Big[\exp\big(\theta(X - \mathbb{E}X)\big)\Big] \leq e^{\nu^2\theta^2/2}, \quad \forall|\theta| < 1/b.$$

- Clearly, sub-Gaussian(σ) is sub-exponential with parameters $(\sigma, 0)$ but the opposite is not true.
- Biggest difference: There is no need for the mgf to be defined for all θ. In fact, the existence of the mgf in a neighborhood of 0 is sufficient for sub-exponential.
- $X \sim \chi_1^2$ is sub-exponential with parameters $(2, 4)$ but not sub-Gaussian
- More on sub-exponential later...