Randomized Matrix Computations Lecture 1

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

Plan

- Organizational aspects!
- What is randomized NLA?
- Some fun questions
- Probability spaces, probabilistic method part 1.
- Random variables, random vectors, random matrices

Organizational aspects

- Lectures: Thursday 10-12, GC B330. First: September 12.
- Exercises: Friday 10-12, MA A331. First: September 13.
- Assessment of course: 2 graded homeworks (40%) and 1 project (60%). The project will be assessed in a short oral exam.
- Material: Slides, supplementary material, exercises will be posted on moodle. Password for self enrollment on moodle: rmc2024.
- `daniel.kressner@epfl.ch`, `hysan.lam@epfl.ch`
- Please feel encouraged to use the Ed Discussion board (link via moodle).

Randomization in numerical linear algebra...

... leads to new and cheap algorithms

... turns “statements that hold generically” into quantifiable results, guiding the analysis and improvement of algorithms

... replaces expensive components in classical algorithms by cheaper alternatives[1]

... offers increased flexibility to exploit structure

... regularizes ill-conditioned problems

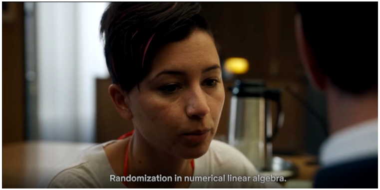
...

[1] hopefully, without spoiling reliability

Randomization in numerical linear algebra...

- ... leads to new and cheap algorithms
- ... turns “statements that hold generically” into quantifiable results, guiding the analysis and improvement of algorithms
- ... replaces expensive components in classical algorithms by cheaper alternatives
- ... offers increased flexibility to exploit structure
- ... regularizes ill-conditioned problems
- ... features prominently on Netflix (The Lincoln Lawyer S1E3, spotted by Petros Drineas)

Thesis? What is it about?



Randomization in NLA

Randomized numerical linear algebra: Surveys

- Murray et al.'2023. Randomized numerical linear algebra. A perspective on the field with an eye to software. `https://arxiv.org/abs/2302.11474v1`
- Martinsson/Tropp'2020. Randomized numerical linear algebra: Foundations and algorithms. Acta Numerica.
- Drineas/Mahoney'2018. Lectures on randomized numerical linear algebra. AMS.
- Kannan/Vempala'2017. Randomized algorithms in numerical linear algebra. Acta Numerica.
- Woodruff'2014. Sketching as a tool for numerical linear algebra, Foundations and Trends in Computer Science.
- Halko/Martinsson/Tropp'2011. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. SIAM Review.

Example for generically true statements

- A real number $\alpha \in \mathbb{R}$ is generically nonzero.
- The norm of a vector $x \in \mathbb{R}^n$ is generically nonzero.
- Given a fixed vector $y \in \mathbb{R}^n$, a vector $x \in \mathbb{R}^n$ generically satisfies $\langle x, y \rangle \neq 0$.
- An $n \times n$ matrix A is generically invertible.
- An $n \times n$ matrix A is generically diagonalizable.
- An $m \times n$ matrix A with $m \geq n$ is generically of rank n.
- Given a fixed $m \times n$ matrix A of rank r, the columns of AX span, for generic choices of $X \in \mathbb{R}^{n \times r}$, the range of A.

What do these actually statements mean?
Why do they hold?

Quantification of generic statements

- Given a random number $\alpha \in \mathbb{R}$, what is the probability that $|\alpha| > \epsilon$ for $\epsilon > 0$?
- Given a random vector $x \in \mathbb{R}^n$, what is the probability that $\|x\|_2 > \epsilon$ for $\epsilon > 0$?
- Given a random matrix $A \in \mathbb{R}^{n \times n}$, what is the probability that $\|A^{-1}\|_2 \leq C$ for $C > 0$?
- ...

What does “random” actually mean?

Basic Prob Foundations

- Probability spaces
- Real random variables
- Real random vectors

Literature:

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

pdf available on Moodle

Probability space

Definition. A probability space is a measure space $(\Omega, \mathcal{F}, \mathbb{P})$, where:

- The sample space Ω is an abstract set of points, called *sample points* or *outcomes*.
- The master σ-algebra $\mathcal{F}$ contains some subsets of Ω, called *events*.
- The probability measure $\mathbb{P} : \mathcal{F} \to [0, 1]$ is a finite measure that satisfies $\mathbb{P}(\Omega) = 1$. It assigns a probability to each event.

We will try to work with the bare minimum of measure theory needed for the purpose of these lectures.

Probability space: σ-algebra

A family of subsets $\mathcal{F} \subseteq \mathcal{P}(\Omega)$ is a σ-algebra on Ω if:

- $\varnothing \in \mathcal{F}$, $\Omega \in \mathcal{F}$
- $E \in \mathcal{F}$ implies $E^c := \Omega \setminus E$
- $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$ and $\bigcap_{i=1}^{\infty} A_i \in \mathcal{F}$ for $A_i \in \mathcal{F}$

Examples:

- $\{\varnothing, \Omega\}$ is a σ-algebra.
- $\mathcal{P}(\Omega)$ is a σ-algebra (called complete σ-algebra).
- Coin flips: $\Omega = \{\mathrm{H}, \mathrm{T}\}$ (head, tail).

$$\mathcal{P}(\Omega) = \{\varnothing, \{\mathrm{H}\}, \{\mathrm{T}\}, \{\mathrm{H}, \mathrm{T}\}\}$$

- Dice: $\Omega = \{1, \ldots, 6\}$. $\mathcal{P}(\Omega)$ has 2^6 elements.
- Major problem: For interval $\Omega = [0, 1]$, the complete σ-algebra $\mathcal{P}(\Omega)$ is not very useful because it contains subsets that are not measurable (assuming axiom of choice).

Probability space: σ-algebra

Given $\mathcal{S} \subseteq \mathcal{P}(X)$, the minimal σ-algebra $\sigma(\mathcal{S})$ is (loosely speaking) the smallest σ-algebra that contains $\mathcal{S}$.

Examples:

- For finite $\Omega = \{1, 2, \ldots, n\}$ and the set of singletons $\mathcal{S} = \{\{1\}, \{2\}, \ldots, \{n\}\}$, the smallest σ-algebra coincides with the complete σ-algebra: $\sigma(\mathcal{S}) = \mathcal{P}(\Omega)$.
  (This also holds for countable Ω.)
- For $\Omega = \mathbb{R}$, the Borel σ-algebra is generated by open intervals:

$$\mathcal{B}(\mathbb{R}) = \sigma\Big(\big\{(a, b) \colon a < b, a, b \in \mathbb{R}\big\}\Big)$$

Probability space: Measures

Let $\mathcal{F}$ be a σ-algebra on Ω. A function $\mu : \mathcal{F} \to [0, +\infty]$ is called a measure if:

1. $\mu(\varnothing) = 0$.
2. For mutually disjoint subsets $(A_i \in \mathcal{F} : i \in \mathbb{N})$,

$$\mu\Big(\dot{\bigcup_{i=1}^{\infty}} A_i\Big) = \sum_{i=1}^{\infty} \mu(A_i).$$

Basic properties:

- $A \subseteq B$ implies $\mu(A) \leq \mu(B)$
- $\mu(A \cup B) + \mu(A \cap B) = \mu(A) + \mu(B)$
- $\mu\big(\bigcup_{i=1}^{\infty} A_i\big) \leq \sum_{i=1}^{\infty} \mu(A_i)$

μ is called a finite measure if $\mu(\Omega) < \infty$.
μ is called a σ-finite measure if Ω can be covered by countably many $A_i \in \mathcal{F}$ with $\mu(A_i) < \infty$.

Recall that μ is a probability measure if $\mu(\Omega) = 1$.

Probability space: Discrete case

Typical measures for finitely many sample Ω points:

- Counting measure: $\mu(A) = \#A$ (cardinality of A)
- Uniform measure: $\mathbb{P}(A) = \#A/\#\Omega$ is a probability measure
- Weighted measure: For $\Omega = \{1, \dots, n\}$, define weights $w_i \geq 0$ and $\mu(A) = \sum_{i \in A} w_i$. This is a probability measure if $\sum_{i=1}^{n} w_i = 1$.

Example for a countable probability space: Flip a fair coin until it turns up heads. Outcome = number of flips until head appears.

$$\Omega = \mathbb{N}, \quad \mathcal{F} = \mathcal{P}(\mathbb{N}).$$

Probability measure defined from singleton outcome $\mathbb{P}(\{n\}) = 2^{-n}$ using additivity:

$$\mathbb{P}(E) = \sum_{n \in E} \mathbb{P}(\{n\}) = \sum_{n \in E} 2^{-n}.$$

EFY: What is the probability that head appears first after an even number of coin flips?

Probability space: Measures on the real line

Recall that $\mathcal{B}(\mathbb{R})$ denotes the Borel algebra.
A measure $\mu : \mathcal{B}(\mathbb{R}) \to [0, +\infty]$ is called a Borel measure on $\mathbb{R}$.

Dirac measure: For fixed $t \in \mathbb{R}$, define for $E \in \mathcal{B}(\mathbb{R})$,

$$\delta_t(E) := 1_E(t) := \begin{cases} 1, & t \in E, \\ 0, & t \notin E. \end{cases}$$

Then δ_t is a Borel (probability) measure on $\mathbb{R}$.

Lebesgue measure: The function $\lambda : \mathcal{B}(\mathbb{R}) \to [0, \infty]$ defined by

$$\lambda(E) := \inf \left\{ \sum_{i=1}^{\infty} |b_i - a_i| : E \subseteq \dot{\bigcup_{i=1}^{\infty}} (a_i, b_i] \right\}$$

is a Borel measure.

Clearly, $\lambda(\{a\}) = 0$, $\lambda([a, b]) = |b - a|$.

EFY: What is $\lambda(\mathbb{R})$? What is $\lambda(\mathbb{Q})$?

Uniform measure on $\Omega = [0, 1]$: $\mathbb{P}(E) = \lambda(E)/\lambda([0, 1]) = \lambda(E)$.

Probability space: Basic properties

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$:

- Recall that $\mathbb{P}(E \cup F) + \mathbb{P}(E \cap F) = \mathbb{P}(E) + \mathbb{P}(F)$.
- $\mathbb{P}(E_1 \cup \cdots \cup E_n) \leq \mathbb{P}(E_1) + \cdots + \mathbb{P}(E_n)$ (The union bound!).
- $\mathbb{P}(E^c) = 1 - P(E)$ for every $E \in \Omega$
- If $\mathbb{P}(E) = 1$ one says that the event E occurs almost surely.
- If $\mathbb{P}(E) = 0$ one says that the event E occurs almost never.

Probabilistic method (flavor 1): For an event $E \in \mathcal{F}$, the condition $\mathbb{P}(E) > 0$ implies $E \neq \varnothing$.

EFY: Given a unit circle in the plane so that a (measurable) subset of 23% of the circle is red and the rest is blue. Show that we can always inscribe a square in the circle so that all four vertices are blue. Hint: Choose the square at random, and show that there is a positive probability that its vertices are all blue.

Real random variable

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A real random variable is a *measurable* function $X : \Omega \to \mathbb{R}$.

Remark: A function X is called measurable if the pre-image of every Borel set is in $\mathcal{F}$. *For the purpose of this lecture, all functions are measurable.*

This allows us to define the law or distribution of the random variable X as the Borel measure

$$\mu_X(B) := \mathbb{P}(X^{-1}(B)) = \mathbb{P}(X \in B) \quad \forall B \in \mathcal{B}(\mathbb{R}).$$

This is a probability measure, that is, the random variable X pushes the distribution $\mathbb{P}$ of probability on the sample space Ω forward to a distribution μ_X of probability on the real line $\mathbb{R}$.

Examples:

- When flipping a coin, let $X = 1$ for head and $X = 0$ otherwise. Then $\mu_X = \delta_0/2 + \delta_1/2$ (Bernoulli 1/2 distribution).
- Assuming the uniform measure on $[0, 1]$, let X denote the position within the interval $[0, 1]$. Then $\mu_X = \lambda(\cdot \cap [0, 1])$. (uniform distribution)

EFY: Reflect on the quote “A random variable is neither random nor variable.” by Gian-Carlo Rota.

Real random variable: Distribution functions

Let X be a real random variable. Then

$$F_X(a) := \mathbb{P}(X \leq a) = \mu_X(-\infty, a], \quad a \in \mathbb{R}$$

is called the cumulative distribution function (cdf) of X.

Properties:

- Monotonicity: If $a \leq b$ then $F_X(a) \leq F_X(b)$.
- Right continuity: We have $\lim_{x \to a+} F_X(x) = F_X(a)$.
- $\mu_X(a, b] = F_X(b) - F_X(a)$.

Two flavors relevant in this course:

- Discrete random variables: $\mu_X = \sum_{i=1}^{\infty} p_i \delta_{a_i}$ for $a_i \in \mathbb{R}$ and $p_i \geq 0$ s.t. $p_1 + p_2 + \cdots = 1$.
- Continuous random variables: Law μ_X has a *density* (pdf) $f_X : \mathbb{R} \to \mathbb{R}$ with respect to the Lebesgue measure:

$$\mu_X(B) = \int_B f_X(x) \lambda(\mathrm{d}x) = \int_B f_X(x) \mathrm{d}x.$$

(See Chapter 4 of [Tropp'23] for Lebesgue integrals.)
Note that f_X is nonnegative and $\int_\Omega f_X(x) \mathrm{d}x = 1$.
Also, $F_X(a) = \int_{-\infty}^{a} f_X(x) \mathrm{d}x$ and, hence, $f_X = F_X'$.

Real random variable: Important examples

The two most important examples in this course:

- Rademacher distribution: $\mathbb{P}(X = 1) = 1/2$ and $\mathbb{P}(X = -1) = 1/2$.
- Normal distribution $X \sim N(m, \sigma^2)$ with mean $m \in \mathbb{R}$ and variance $\sigma^2 > 0$ has pdf

  $$f_X(x) = \frac{e^{-(x-m)^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}}.$$

  It is best to think of X as the identity function on the probability space $(\mathbb{R}, \sigma(\mathbb{R}), \mu_X)$.
  $m = 0, \sigma^2 = 1$: standard normal distribution.

Other important elementary continuous random variables include Gamma, Beta, exponential, Cauchy, χ^2. pdfs and many other properties for these elementary variables are explicitly known (Wikipedia, MathOverflow, ...).

Often, random variables arise from composition of functions with elementary random variables. Only in *rare* cases, pdfs are simple.

EFY: Let $X \sim N(0, 1)$. Prove that the pdf of X^2 is given by

$$f_{X^2}(x) = 0 \text{ for } x < 0, \quad f_{X^2}(x) = \frac{1}{\sqrt{2\pi}} x^{-1/2} e^{-x/2} \text{ for } x \geq 0.$$

This is called χ^2 distribution (with one degree of freedom).

Real random vectors

Let $X_1, \ldots, X_n$ be random variables defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Then $X = \left[X_1, \ldots, X_n\right]^\top$ is called a random vector.

- In the case of real random variables $X_1, \ldots, X_n : \Omega \to \mathbb{R}$, we will call X a real random vector and write $X \in \mathbb{R}^n$.
- There are direct extensions of the notion of Borel σ-algebra $\mathcal{B}(\mathbb{R}^n)$ and Borel measures $\mu : \mathcal{B}(\mathbb{R}^n) \to [0, +\infty]$. The Lebesgue measure $\lambda : \mathcal{B}(\mathbb{R}^n) \to [0, \infty]$ is the product Lebesgue measure.
- The distributions μ_{X_i} of the individual random variables are called *marginal distributions*. **IMPORTANT:** Generally, the marginal distributions are *not* sufficient to describe X. We need to prescribe a joint distribution

$$\mu_{X_1,\ldots,X_n}(B) = \mathbb{P}(X \in B) \quad \forall B \in \mathcal{B}(\mathbb{R}^n),$$

which is a Borel probability measure on $\mathcal{B}(\mathbb{R}^n) \to [0, 1]$.

Real random vectors: Specifying joint distributions

For simplicity, consider $n = 2$ and a real random vector $V = \left[X, Y\right]$.

To specify a *discrete joint distribution* ($\Omega = \{1, \ldots, k\}$), it suffices to prescribe the probabilities of the k^2 different singleton events:

$$\mathbb{P}(X = i \text{ and } Y = j), \quad i, j = 1, \ldots, k.$$

The marginal distributions are recovered by summing up, e.g., $\mathbb{P}(X = i) = \sum_{j=1}^{k} \mathbb{P}(X = i \text{ and } Y = j)$.

Real random vectors: Specifying joint distributions

To specify a *continuous joint distribution*, we can prescribe a joint cdf

$$F_{XY}(a,b) = \mathbb{P}\{X \le a \text{ and } Y \le b\} = \mu_{XY}\big((-\infty,a] \times (-\infty,b]\big)$$

or, more commonly, a joint pdf $f_{X,Y}(x,y) = \frac{\partial^2 F_{XY}(x,y)}{\partial x \partial y}$. We have $f_{X,Y} \ge 0$, $\int_{\mathbb{R}^2} f_{X,Y} = 1$, and

$$F_{XY}(a,b) = \int_{(-\infty,a]\times(-\infty,b]} f_{X,Y}(x,y)\,(\mathrm{d}x \times \mathrm{d}y),$$

The pdfs of the marginal distributions are recovered by integrating the other variable, e.g.,

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y.$$

Fubini-Tonelli theorem

Theorem. For σ-finite measure spaces $(\Omega_i, \mathcal{F}_i, \mu_i)$, $i = 1, 2$, consider a measurable function $f : \Omega_1 \times \Omega_2 \to \mathbb{R}$.

1. If $f \geq 0$,

$$\begin{aligned}
& \int_{\Omega_1 \times \Omega_2} f(x,y)(\mu_1 \times \mu_2)(\mathrm{d}x \times \mathrm{d}y) \\
= & \int_{\Omega_1} \Big(\int_{\Omega_2} f(x,y) \mu_2(\mathrm{d}y) \Big) \mu_1(\mathrm{d}x) \\
= & \int_{\Omega_2} \Big(\int_{\Omega_1} f(x,y) \mu_1(\mathrm{d}x) \Big) \mu_2(\mathrm{d}y)
\end{aligned}$$

2. Point 1 also holds when $\int_{\Omega_1 \times \Omega_2} |f(x,y)|(\mu_1 \times \mu_2)(\mathrm{d}x \times \mathrm{d}y) < \infty$.

Real random vectors: Independence

Two random variables are independent if

$$\mu_{XY} = \mu_X \times \mu_Y$$

or, equivalently,

$$F_{XY}(a, b) = F_X(a) F_Y(b).$$

In particular, the joint distribution is completely described by the marginal distributions.

- *Discrete* $\Omega = \{1, \ldots, k\}$: Independence equivalent to

$$\mathbb{P}(X = i \text{ and } Y = j) = \mathbb{P}(X = i) \cdot \mathbb{P}(Y = j), \quad i, j = 1, \ldots, k.$$

- *Continuous:* Independence equivalent to product pdf:

$$f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y).$$

Real random vectors: Rademacher

For a Rademacher random vector $X = [X_1, \ldots, X_n]$, the components are independent Rademacher variables X_i, that is,

$$\mathbb{P}(X_i = 1) = 1/2, \quad \mathbb{P}(X_i = -1) = 1/2, \quad i = 1, \ldots, n.$$

Some trivial facts:

- The probability that X is a vector of all ones is 2^{-n}.
- $\|X\|_2 = \sqrt{n}$.

EFY: Let $X \sim N(0, 1)$. Prove that $\operatorname{sign}(X)$ is Rademacher.

Example of nontrivial question: Behavior of Rademacher sum $\langle X, v \rangle$ for fixed vector v.

Real random vectors: Gaussian random

For a Gaussian random vector (also: standard normal random vector) $X = [X_1, \ldots, X_n]$, the components are independent $X_i \sim N(0, 1)$. Its density is given by

$$f_X(x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-x_i^2/2} = \frac{1}{(2\pi)^{n/2}} e^{-\|x\|_2^2/2}$$

We write $X \sim N(0, I_n)$.

Let $A \in \mathbb{R}^{n \times m}$ with $m \leq n$ s.t. $\mathrm{rank}(A) = m$. What is the distribution of random vector $Y = AX$ for $X \sim N(0, I_n)$?

Real random vectors: Normal random vectors

Change of variables in Lebesgue integrals yields the following result:

Let $X \in \mathbb{R}^n$ be a continuous random vector with pdf f_X. Let $g : \mathbb{R}^n \to \mathbb{R}^n$ be bijective and continuously differentiable. Then $Y = g(X)$ is a continuous random vector with pdf

$$f_Y(y) = f_X(g^{-1}(y))/|\det(J_g(g^{-1}(y)))|,$$

where J_g denotes the Jacobian of g.

Applied to Gaussian random vector $X \in \mathbb{R}^n$, this implies that $Y = AX$ has the distribution

$$\begin{aligned} f_Y(y) &= \frac{1}{(2\pi)^{n/2}|\det(A)|} \exp\left(-\frac{1}{2}\|A^{-1}y\|_2^2\right) \\ &= \frac{1}{(2\pi)^{n/2}\sqrt{\det C}} \exp\left(-\frac{1}{2}y^\top C^{-1} y\right) \end{aligned}$$

We write $Y \sim N(0, C)$ with the so called covariance matrix $C = AA^\top$.

EFY: This result continues to hold when A is an $m \times n$ matrix of rank m. Why?

EFY: Given $Y \sim N(0, C)$, what is the marginal distribution of Y_1?

Real random vectors: Properties of Gaussian random

Corollaries. Let $X \sim N(0, I_n)$. Then:

- $\langle X, a\rangle \sim N(0, 1/\|a\|_2^2)$ for a fixed vector $a \in \mathbb{R}^n$.
- $QX \sim N(0, I_n)$ for any fixed orthogonal matrix $Q \in \mathbb{R}^{n\times n}$.
- $U^T X \sim N(0, I_p)$ for any matrix $U \in \mathbb{R}^{n\times p}$ with orthonormal columns.

How close is $Y = \langle X, a\rangle$ to zero? Because $|f_Y(y)| \leq \frac{1}{\sqrt{2\pi}\|a\|_2}$, it follows that

$$\mathbb{P}\big(-\epsilon\|a\|_2 < \langle X, a\rangle < \epsilon\|a\|_2\big) \leq \frac{\sqrt{2}}{\sqrt{\pi}}\epsilon.$$

Important: Oblivious to a! This is the simplest example of a small-ball probability.

Such bounds on small-ball probabilities are only that simple to obtain if the pdf is explicitly available.

Real random vectors: Uniform distribution on sphere

Let $Y = X/\|X\|_2$ for $X \sim N(0, I_n)$. Then:

- $\|Y\|_2 = 1$ (almost surely)
- Y and QY have the same distribution for any orthogonal matrix Q

Hence, Y is uniformly distributed in the sphere S^{n-1}. We write $Y \sim U(S^{n-1})$.

- The components of Y are *not* independent.
- The marginal distribution of the first $k < n$ components is

$$f_{Y_1,\ldots,Y_k}(y) = \begin{cases} 0 & \text{if } \|y\|_2 > 1, \\ c_{k,n}(1 - \|y\|_2^2)^{(n-k)/2-1} & \text{otherwise.} \end{cases}$$

  The constant $c_{k,n}$ can be determined by the fact that the density integrates to 1:

$$c_{k,n} = \frac{\Gamma(n/2)}{\pi^{k/2}\Gamma((n-k)/2)}.$$

See [Muirhead'1982: Aspects of multivariate statistical theory].

Real random vectors: Uniform distribution on sphere

In particular,

$$f_{Y_1}(y) = c_{1,n}(1 - y)^{(n-3)/2}, \quad y < 1.$$

with $c_{1,n} \leq \sqrt{n/2\pi}$.

How close is Y_1 to zero?

$$\mathbb{P}(-\epsilon < Y_1 < \epsilon) \leq \frac{\sqrt{2n}}{\sqrt{\pi}}\epsilon.$$

Another small-ball probability!

A first analysis of the power method

Given a symmetric positive definite matrix A with eigenvalues $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_n > 0$, the power method

$$\tilde{v}_{k+1} = Av_k, \quad v_{k+1} = \tilde{v}_{k+1} / \| v_{k+1} \|_2,$$

converges for almost every starting vector v_0 to an eigenvector u_1 belonging to λ_1.

More specifically, we have (see course on Computational Linear Algebra) that

$$\tan \angle(u_1, v_k) \leq \left(\frac{\lambda_2}{\lambda_1}\right)^k \tan \angle(u_1, v_0).$$

Note that $\tan \angle(u_1, v_0) \leq 1/|u_1^T v_0|$ and $u_1^T v_0 \sim Y_1$ when choosing a random starting vector v_0 uniformly distributed on the sphere. Thus, the small-ball probability bound shows that $\tan \angle(u_1, v_0)$ is not much larger than $\sqrt{n}$ with high probability.

There is no known, reasonably cheap deterministic construction for v_0 that is guaranteed to have an equally favorable property.

Real random matrices

There is no conceptual difference between defining real random vectors and real random matrices. An $m \times n$ real random matrix is an (mn)-tuple of real random variables.

Most important examples:

- Gaussian random matrix: $a_{ij} \sim N(0, 1)$ iid
- Rademacher matrix: $a_{ij} \sim$ Rademacher iid
- Uniform on Stiefel: $A \sim U(\mathrm{St}(m, n))$ for $m \geq n$ = uniformly distributed on the Stiefel manifold $\mathrm{St}(m, n)$ of $m \times n$ matrices with orthonormal columns.

Random matrix theory is primarily concerned with studying the distribution of eigenvalues of random matrix models.

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

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

- 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 $F_X(t) = \mathbb{P}(X \le t)$. 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

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

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

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

Missing 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.

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 $a \leq \mathbb{E}[X] \leq b$.

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:

$$\varphi(y) \geq \varphi(a) + \varphi'(a) \cdot (y - a),$$

provided that φ is differentiable. If φ is not differentiable at a, the fomula still holds with $\varphi'(a)$ replaced by a subgradient of φ at a.

Jensen's inequality

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:

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

Expectation of random vectors

For a random vector $X \in \mathbb{R}^n$, the expectation $\mathbb{E}[X]$ is simply defined entry-wise.

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