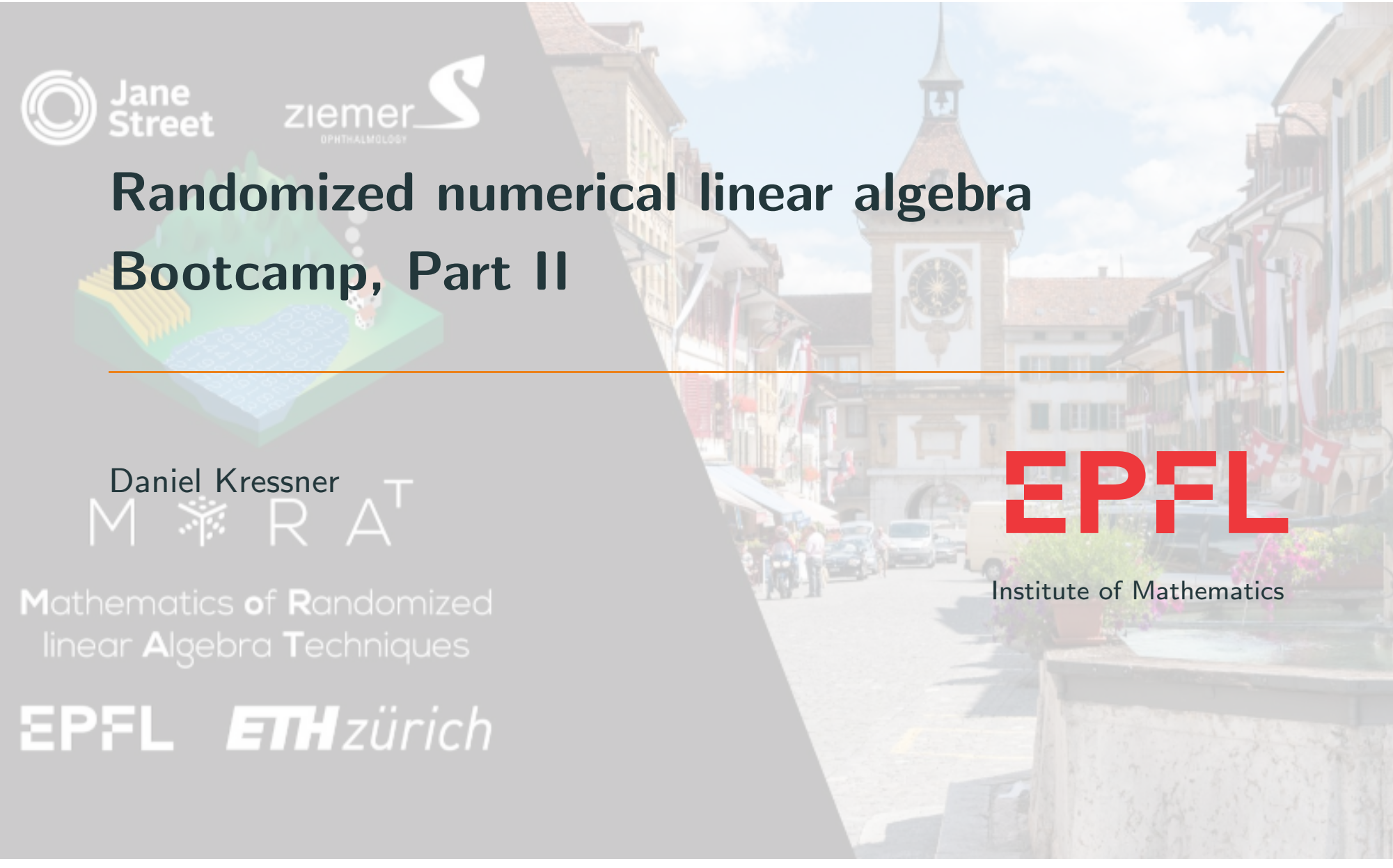


Randomized numerical linear algebra
Bootcamp, Part II

Daniel Kressner

Mathematics of Randomized linear Algebra Techniques

Institute of Mathematics

Expectation and moments

Expectation

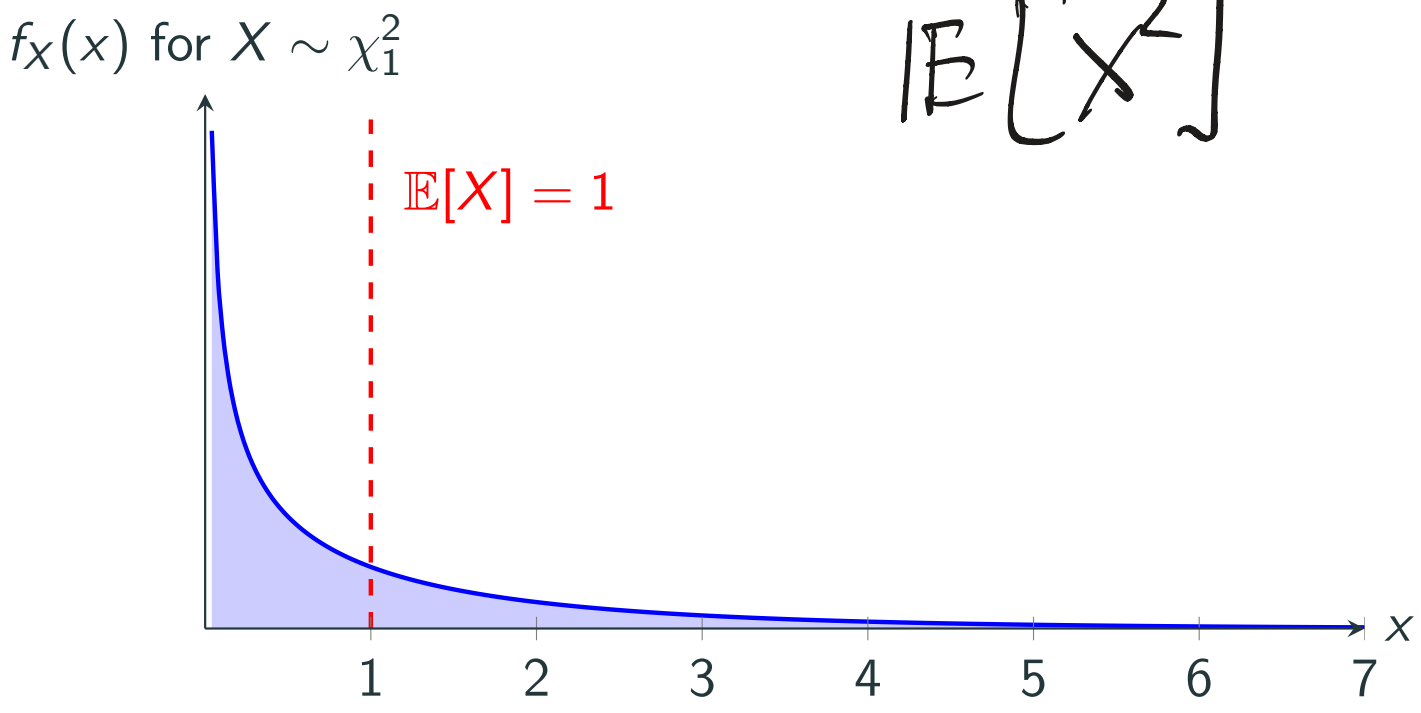
- For discrete random variable X with $\Pr[X = a_i] = p_i$:

$$\mathbb{E}[X] = \sum_i a_i p_i.$$

- For continuous random variable X with density f_X:

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

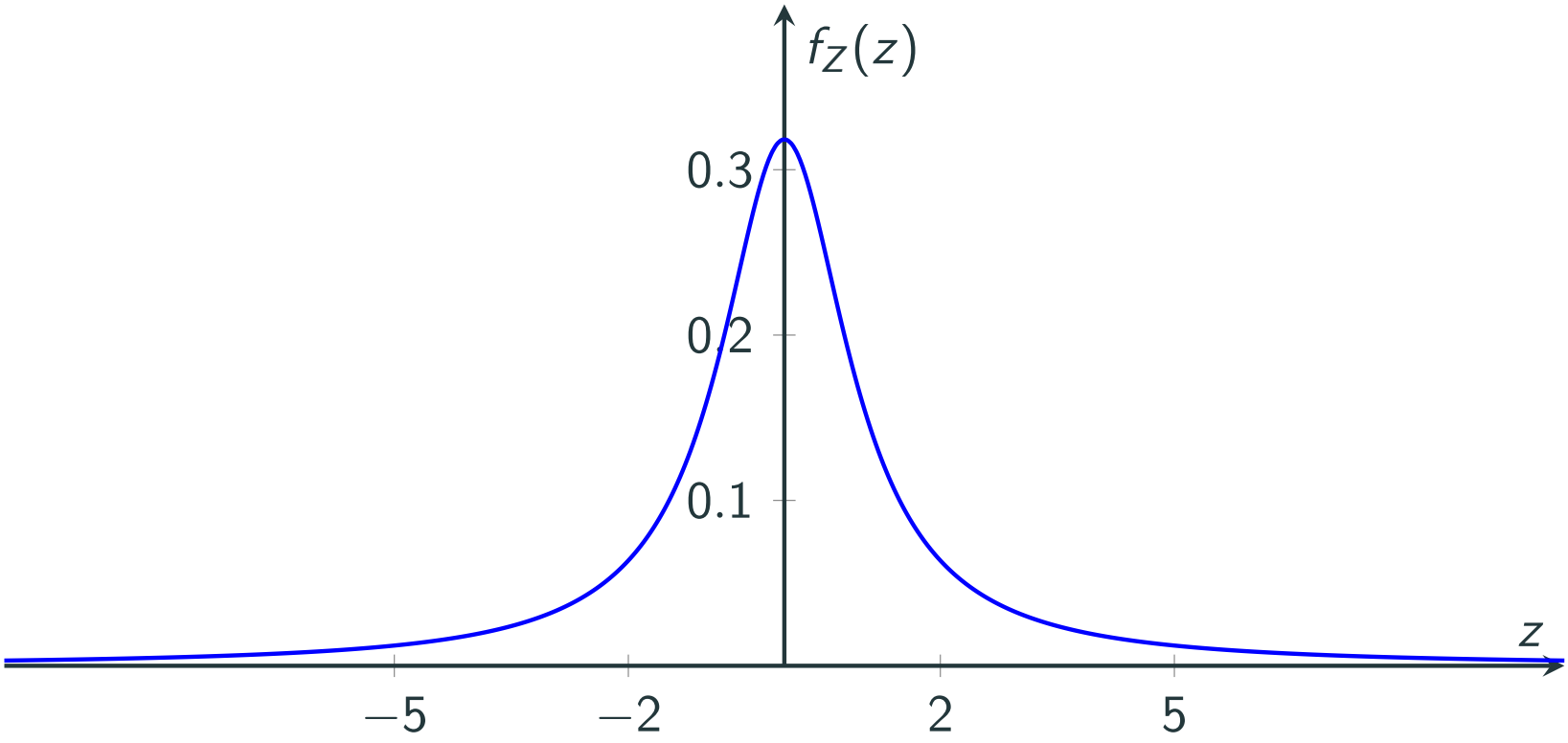
$\mathbb{E}[X^2]$



Defying expectations: Cauchy random variables

Consider independent $X, Y \sim N(0,1)$. Then $Z = X/Y$ has pdf

$$f_Z(z) = \frac{1}{\pi(1+z^2)}.$$



$z \cdot f_Z(z)$ is not integrable and, hence, the expectation of Z is not defined.

Z is canonical example of a Cauchy random variable.

Expectation: Law of the unconscious statistician

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

- For discrete random variable X with $\Pr[X = a_i] = p_i$,

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

- For continuous random variable X with pdf f_X,

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

$\mathbb{E}[X^p]$ is called the pth moment for $p = 1, 2, \ldots$.

$\mathbb{E}[(X - \mathbb{E}X)^p]$ is called the pth centered moment.

In particular, $\mathbb{E}[(X - \mathbb{E}X)^2]$ is the variance.

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. $\mathbb{E}[\alpha X + \beta Y] = \alpha\mathbb{E}[X] + \beta\mathbb{E}[Y]$ 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 integrals.

Jensen's inequality for random variables

For convex function $\varphi : I \to \mathbb{R}$ on an open interval I, one has

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

provided that φ is differentiable at a.

Theorem: Jensen's inequality

Let X be an integrable, real random variable that takes values in I. Then

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

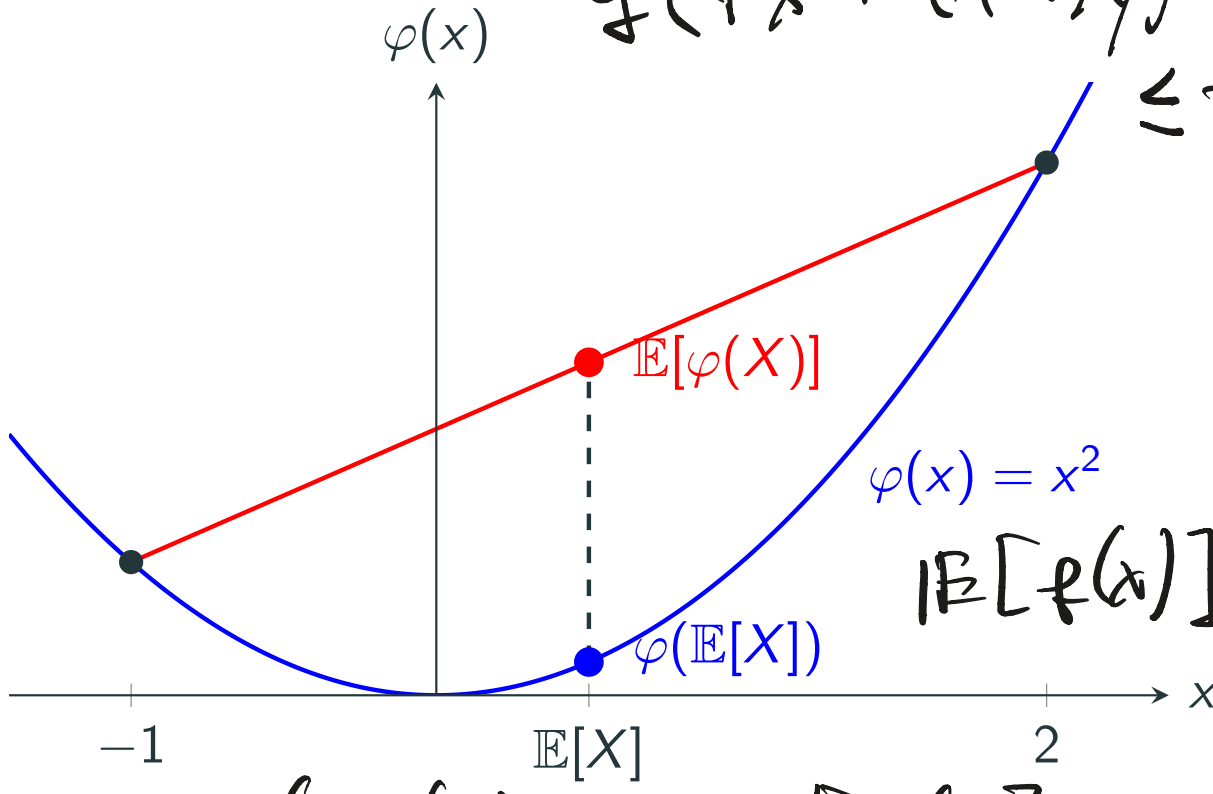
Proof. For simplicity, suppose that φ is differentiable. Setting $y = X(\omega)$ and $a = \mathbb{E}[X]$ in the property above gives

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

Taking expectations on both sides completes the proof. ◇

Two important examples:

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



$$X = \begin{cases} -1, & \text{with probability } 1/2, \\ 2, & \text{with probability } 1/2, \end{cases}$$

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

Expectation of random vectors

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

$$\mathbb{E}[X] = \begin{bmatrix} \mathbb{E}_{X_1}[X_1] \\ \vdots \\ \mathbb{E}_{X_n}[X_n] \end{bmatrix}.$$

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 continuous random vector with joint density f_X:

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

Jensen's inequality for random vectors

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

Theorem: Jensen's inequality

Let X be an integrable random vector taking values in a convex open set $C \subset \mathbb{R}^n$, and let $\varphi : C \to \mathbb{R}$ be convex, with $\mathbb{E}[\varphi(X)]$ well defined. Then

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

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

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

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

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

Moments and tails

Types of moments

- Polynomial moments:

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

  assuming that the expectation is well-defined.

- Exponential moments:

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

- If the polynomial moments do not grow too quickly, $\mathbb{E}[\exp(\theta X)]$ is finite for (sufficiently small) $\theta > 0$ 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 moment generating function (mgf).

Types of tails

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

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

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

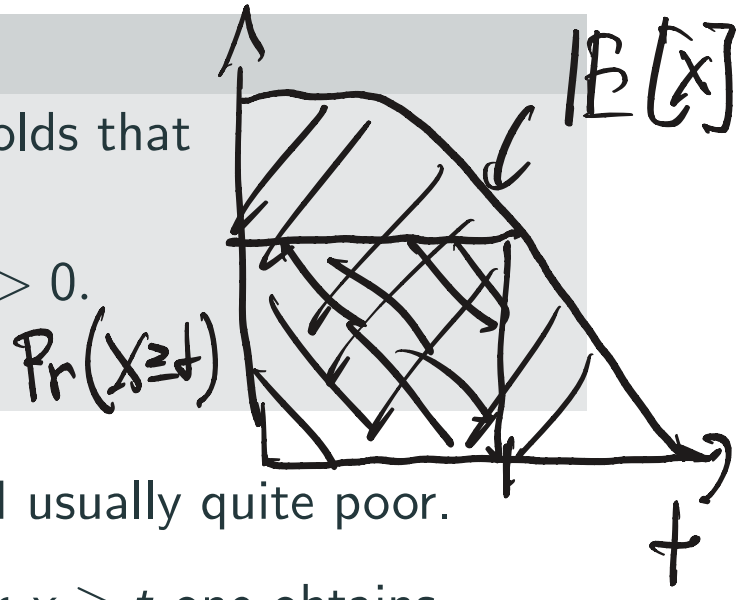
The better the polynomial moments behave, the better the tail behavior.

Markov's inequality

Theorem

For a real nonnegative random variable X, it holds that

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



Note: Little knowledge on X needed, but bound usually quite poor.

Proof for continuous r.v. Using that $x/t \geq 1$ for $x \geq t$ one obtains

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

◇

Equivalent formulation: $\Pr\{X \geq t \cdot \mathbb{E}[X]\} \leq t^{-1}$. A non-negative r.v. exceeds its expected value by a factor 10 with probability at most 10%.

Boosting Markov's inequality

Let Z be non-negative random variable and $\varphi : [0, \infty) \to [0, \infty)$ be increasing. Then $Z \geq t$ implies $\varphi(Z) \geq \varphi(t)$, and hence

$$\Pr\{Z \geq t\} \leq \Pr\{\varphi(Z) \geq \varphi(t)\} \leq \frac{\mathbb{E}[\varphi(Z)]}{\varphi(t)}, \qquad \varphi(t) > 0,$$

where used Markov's inequality (with X / t replaced by $\varphi(Z)$ / $\varphi(t)$).

Three important cases:

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

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

- $\varphi(z) = z^p$ for $p \geq 1$ and $Z = |X| \rightsquigarrow$

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

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

Averaging random variables

Let $X_1, \ldots, X_m$ be iid random variables with

$$\mathbb{E}[X_i] = \mu, \qquad \mathrm{Var}(X_i) = \sigma^2,$$

and consider sample average

$$\overline{X}_m := \frac{1}{m}(X_1 + X_2 + \cdots + X_m).$$

Then

$$\boxed{\mathbb{E}[\overline{X}_m] = \mu, \qquad \mathrm{Var}(\overline{X}_m) = \frac{\sigma^2}{m}.}$$

By Chebyshev's inequality,

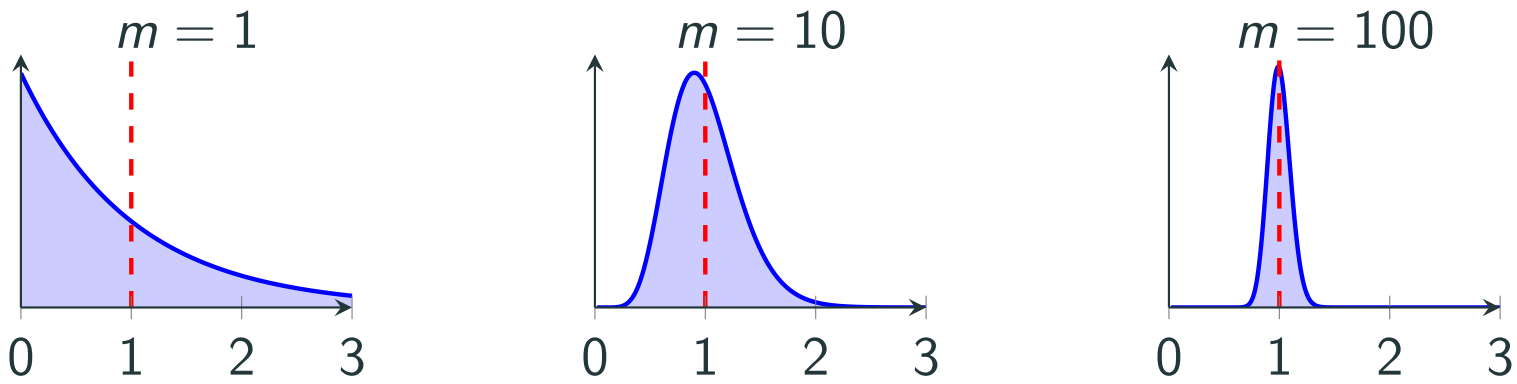
$$\boxed{\Pr\left\{|\overline{X}_m - \mu| \geq \epsilon\right\} \leq \frac{\sigma^2}{m\epsilon^2}.}$$

To guarantee $|\overline{X}_m - \mu| < \epsilon$ with probability at least $1 - \delta$, it suffices to take $m \geq \sigma^2 \delta^{-1} \epsilon^{-2}$.

Averaging random variables

Take

$$X_1, X_2, \ldots \overset{\text{iid}}{\sim} \text{Exp}(1), \qquad \mathbb{E}[X_i] = 1, \qquad \text{Var}(X_i) = 1.$$



The curse of Monte Carlo

$m = \mathcal{O}(\delta^{-1}\epsilon^{-2})$:

- δ^{-1}: Often simply an artifact of using Chebyshev.
- ϵ^{-2}: Unavoidable curse of Monte Carlo.

  CLT: For *any* iid X_i with finite expectation μ and variance σ^2,

$$\frac{\sqrt{m}(\overline{X}_m - \mu)}{\sigma} \xrightarrow{d} N(0,1).$$

  This implies that $|\overline{X}_m - \mu| \leq \epsilon$ can only hold with fixed success probability if $m \sim \epsilon^{-2}$.

Hence:

halve the error $\implies$ multiply number of samples by 4,

gain one decimal digit $\implies$ multiply number of samples by 100.

The curse of Monte Carlo

You will see ϵ^{-2} many times this week!

In randNLA:

- In many situations, $\epsilon = 1/2$ suffices: randomized range finding / low-rank approximation, sketched orthogonalization, ...
- When it hurts (e.g., in trace estimation), variance reduction techniques might help.

Chernoff's inequality

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

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

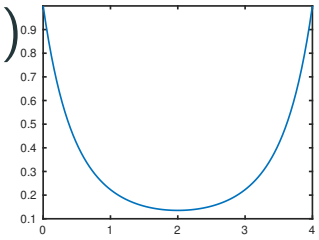
  As this holds for any $\theta > 0$,

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

  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:

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

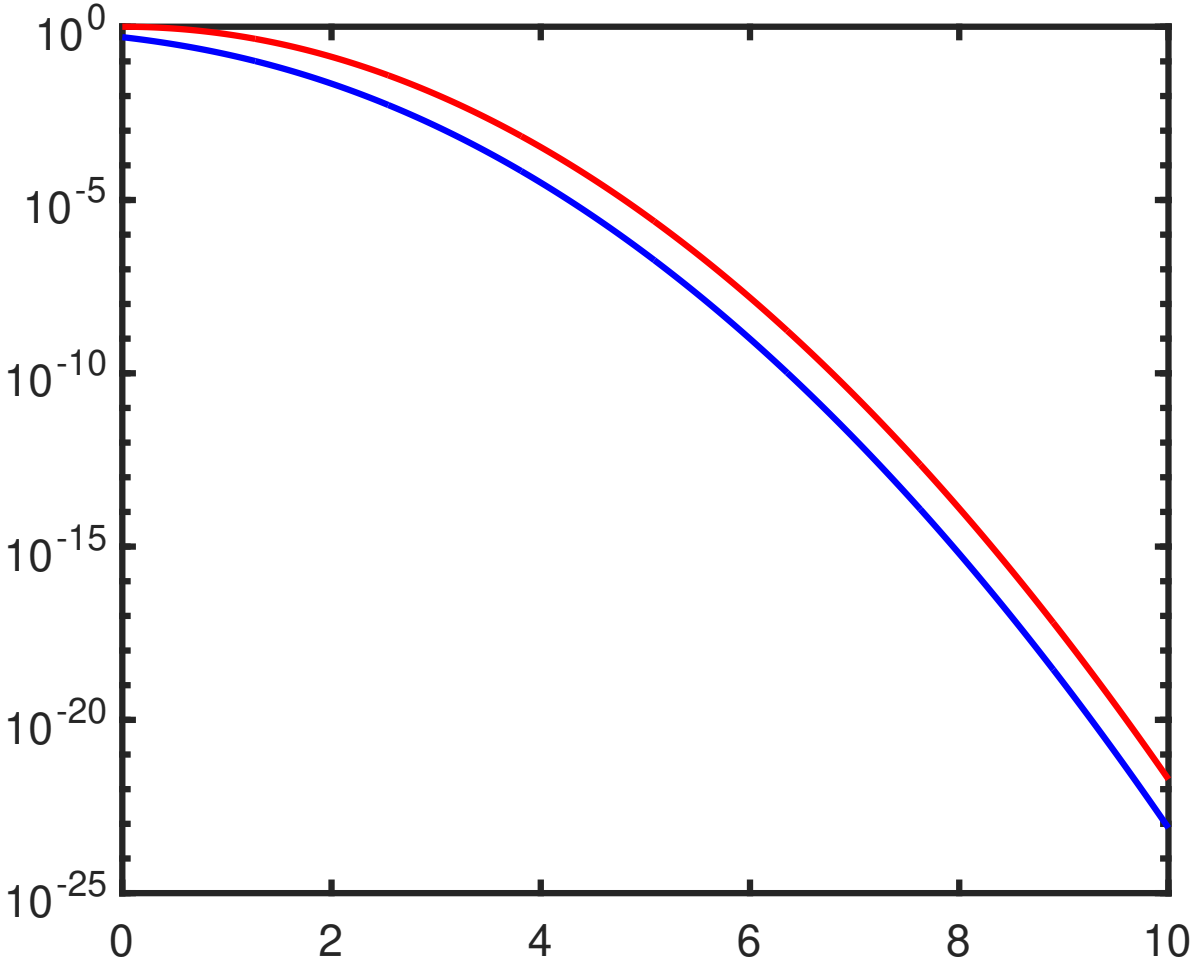


By differentiating, optimal $\theta_* = t/\sigma^2 \rightsquigarrow$

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

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

Chernoff's inequality



$\Pr\{X \geq t\}$ for $X \sim N(0,1)$ vs. Chernoff bound $\exp(-t^2/2)$

A first encounter with Johnson–Lindenstrauss

Chernoff for χ^2 r.v.

A chi-squared random variable with m degrees of freedom

$$Y \sim \chi^2_m, \qquad Y = X_1^2 + \cdots + X_m^2, \qquad X_1, \ldots, X_m \sim N(0,1) \text{ i.i.d.},$$

satisfies $\mathbb{E}[Y] = m$.

For $\lambda < 1/2$,

$$\mathbb{E}[e^{\lambda X_i^2}] = \frac{1}{\sqrt{1-2\lambda}}, \qquad \Longrightarrow \qquad \mathbb{E}[e^{\lambda Y}] = (1-2\lambda)^{-m/2}.$$

For $s > 0$, Chernoff's method gives

$$\begin{aligned}\Pr\{Y \geq m+s\} &\leq \inf_{0<\lambda<1/2} e^{-\lambda(m+s)}(1-2\lambda)^{-m/2} \\ &= \exp\left[-\frac{m}{2}\left(\frac{s}{m} - \log\left(1+\frac{s}{m}\right)\right)\right].\end{aligned}$$

For $0 \leq s < m$, using $u - \log(1+u) \geq \frac{u^2}{4}$ for $0 \leq u \leq 1$, we obtain

$$\boxed{\Pr\{Y \geq m+s\} \leq \exp\left(-\frac{s^2}{8m}\right), \qquad 0 \leq s < m.}$$

χ^2 concentration

The same Chernoff argument for the lower tail gives

$$\Pr\{Y \leq m - s\} \leq \exp\left(-\frac{s^2}{4m}\right) \leq \exp\left(-\frac{s^2}{8m}\right), \qquad 0 \leq s < m.$$

Therefore,

$$\boxed{\Pr\{|Y - m| \geq s\} \leq 2\exp\left(-\frac{s^2}{8m}\right), \qquad 0 \leq s < m.}$$

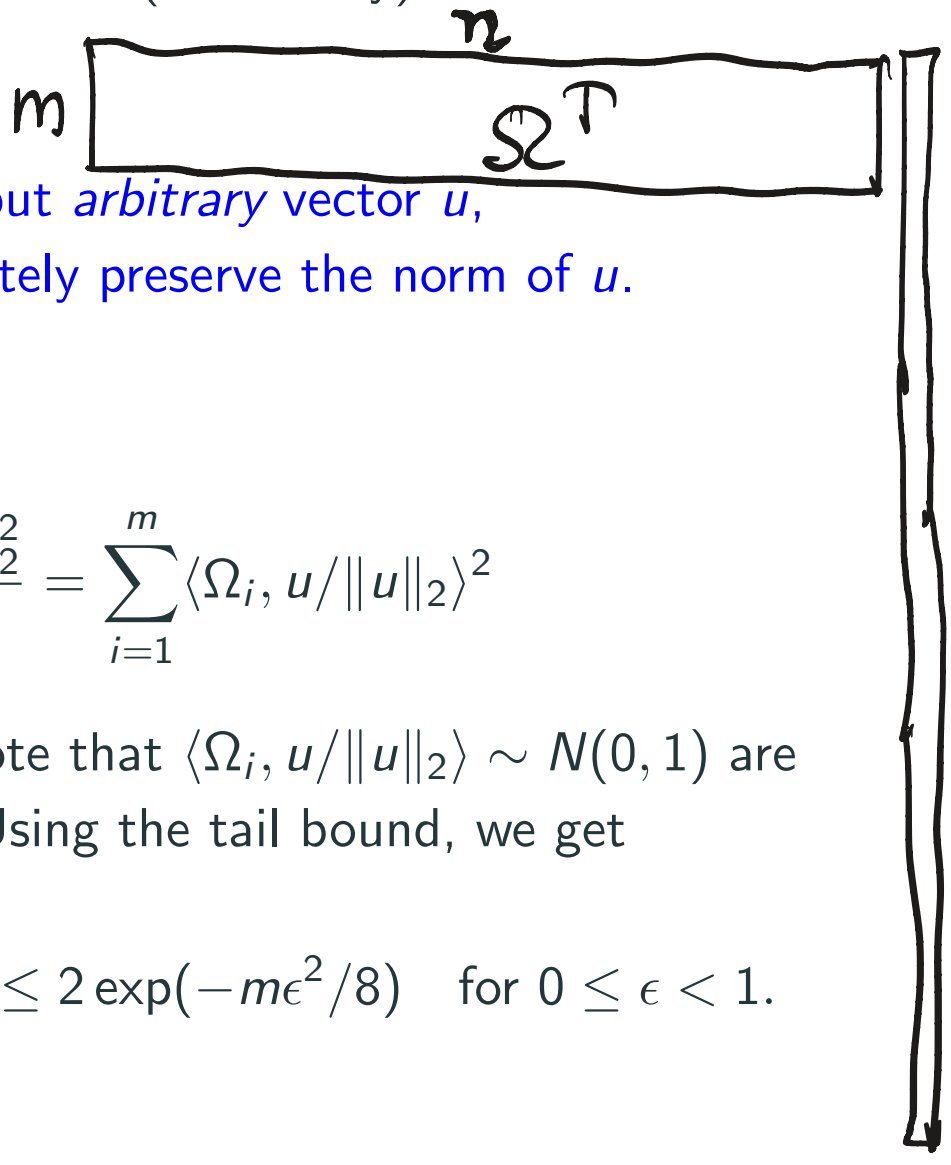
Setting $s = m\epsilon$ yields*:

$$\boxed{\Pr\left\{\left|\frac{X_1^2 + \cdots + X_m^2}{m} - 1\right| \geq \epsilon\right\} \leq 2\exp\left(-\frac{m\epsilon^2}{8}\right), \qquad 0 \leq \epsilon < 1.}$$

*For SOTA exponential tail bounds see [Laurent/Massart, Annals of Statistics'2000]

Johnson-Lindenstrauss embedding with Gaussians

Consider $\mathbb{R}^n$ with LARGE n. Consider (tall skinny) Gaussian random matrix $\Omega \in \mathbb{R}^{n \times m}$ with $m \ll n$.



Given fixed but *arbitrary* vector u,
$\frac{1}{\sqrt{m}}\Omega^\top u$ approximately preserve the norm of u.

To see this, consider ratio

$$Y := \frac{\|\Omega^\top u\|_2^2}{\|u\|_2^2} = \sum_{i=1}^{m} \langle \Omega_i, u/\|u\|_2 \rangle^2$$

where Ω_i is ith column of Ω. Note that $\langle \Omega_i, u/\|u\|_2 \rangle \sim N(0,1)$ are independent. Hence, $Y \sim \chi^2_m$. Using the tail bound, we get

$$\Pr\left\{ \left| \frac{\|\Omega^\top u\|_2^2}{m\|u\|_2^2} - 1 \right| \geq \epsilon \right\} \leq 2\exp(-m\epsilon^2/8) \quad \text{for } 0 \leq \epsilon < 1.$$

Johnson-Lindenstrauss embedding: A first encounter

We can rearrange this in a more common form: Rescale $\widetilde{\Omega} = \Omega/\sqrt{m}$. For any $0 \leq \epsilon < 1$ we have that

$$(1-\epsilon)\|u\|_2^2 \leq \|\widetilde{\Omega}^\top u\|_2^2 \leq (1+\epsilon)\|u\|_2^2$$

holds with probability at least $1 - 2\exp(-m\epsilon^2/8)$.

Now, consider (many) M fixed vectors $u_1, \ldots, u_M \in \mathbb{R}^n$. By the union bound,

$$(1-\epsilon)\|u_i\|_2^2 \leq \|\widetilde{\Omega}^\top u_i\|_2^2 \leq (1+\epsilon)\|u_i\|_2^2$$

holds for every i with probability at least $1 - 2M\exp(-m\epsilon^2/8)$.

Fix $\epsilon = 1/\sqrt{2}$. To attain failure probability δ, need to choose embedding dimension

$$m \geq 16(\log \delta^{-1} + \log 2M).$$

Importantly, m depends logarithmically on M (and does not depend on n at all)!

What is finite-set JL good for?

Given fixed data points $x_1, \ldots, x_M \in \mathbb{R}^n$ choose JL embedding that ϵ-preserves all $\binom{M}{2}$ pairwise distances

$$\|x_i - x_j\|_2, \quad i \neq j.$$

Many computations can be carried out in $\mathbb{R}^m$ instead of $\mathbb{R}^n$ without significant loss.

- Similarity search / nearest neighbors. Search for nearby points after projecting a high-dimensional database.

  $$n \gg m = O(\epsilon^{-2} \log M).$$

  [Indyk–Motwani'1998; Ailon–Chazelle'2009]
- Clustering. Distances relevant to k-means, k-medians, hierarchical clustering, etc. survive the projection.

  [Boutsidis–Zouzias–Mahoney–Drineas'2015;

  Makarychev–Makarychev–Razenshteyn'2023]
- Large image/text databases, geometric/combinatorial optimization, learning.