Randomized numerical linear algebra Bootcamp, Part I

Daniel Kressner

Institute of Mathematics

Jane Street

ziemer OPHTHALMOLOGY

MORAT

Mathematics of Randomized linear Algebra Techniques

EPFL ETH zürich

EPFL

go.epfl.ch/moratdaniel

Randomized numerical linear algebra...

In a nutshell:

Leveraging randomization
to solve (deterministic) linear algebra problems.

DECEMBER 26, 2024 | 10 MIN READ

Probability Probably Doesn't Exist

All of statistics and much of science depends on probability—an astonishing achievement, considering no one's really sure what it is

BY DAVID SPIEGELHALTER & NATURE MAGAZINE

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.
- ..and more to come!

Additional material for bootcamp lectures

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.

Numerical Linear Algebra

The usual suspects

Linear systems

$$Ax = b$$

Eigenvalue problems

$$Ax = \lambda x$$

Least-squares problems

$$\min_x \|Ax - b\|_2$$

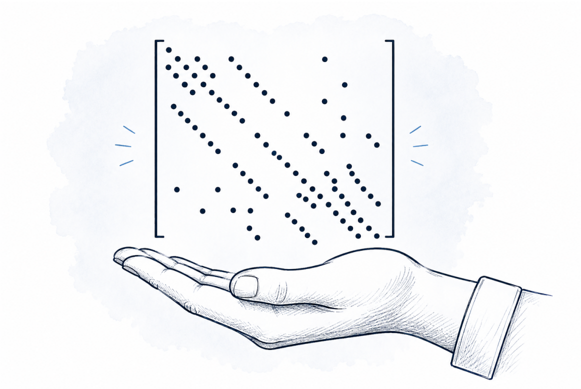
Low-rank approximation

$$A \approx UV^T$$

Matrix functions

$$f(A), f(A)b, \mathrm{trace}(f(A))$$

LARGE matrices

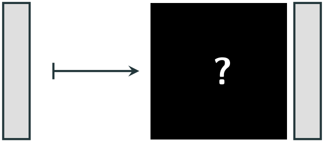


Finite element discretizations, large graphs, ... give rise to LARGE *sparse* matrices. If you are lucky, the entire matrix is handed to you, and you can conveniently store and access all its nonzero entries.

LARGE matrices: Matrix-vector access

$A: x \mapsto Ax$

$x \mapsto Ax$

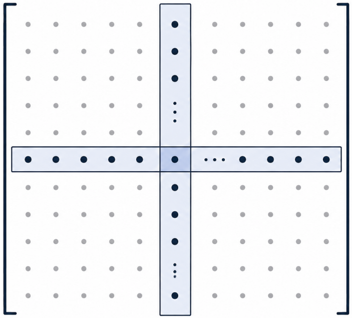


Matrix A can only be accessed through matrix-vector products Ax. Sometimes, A can also be accessed through $A^T x$.

Examples: Solution operators, $A = f(B)$, Jacobians/Hessians in opt.

Difficult/expensive to get access to more than a handful of individual entries (e.g., the diagonal of A).

LARGE matrices: Entrywise access



Individual entries/rows/columns of the matrix can be queried. Typical scenario: Matrix A is dense and each entry requires some computation.

Examples: Kernel matrices, distance/similarity matrices, nonlocal discretizations / operators.

Usually very expensive to even perform a single matvec with A)

Powering up

Given initial vector x_0, consider iteration

$$\tilde{x}_{k+1} = \begin{bmatrix} 1 & & & & & \\ & 0.9 & & & & \\ & & -0.9 & & & \\ & & & \ddots & & \\ & & & & 0.1 & \\ & & & & & -0.1 \end{bmatrix} x_k \qquad x_{k+1} = \frac{\tilde{x}_{k+1}}{\|\tilde{x}_{k+1}\|_2}$$

$$x_0 = \begin{bmatrix} * \\ \vdots \\ * \end{bmatrix} \qquad * \neq 0$$

Does this iteration converge?

$$\begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Given **symmetric** matrix A with initial vector x_0, consider iteration

$$\tilde{x}_{k+1} = Ax_k, \qquad x_{k+1} = \frac{1}{\|\tilde{x}_{k+1}\|_2}\tilde{x}_{k+1}$$

Does this iteration converge?

$A = U \Lambda U^T$ $\quad \langle u_1, x_0 \rangle \neq 0$

spectral decomp.

Randomization in the early days of Scientific Computing

Power method:

$$x_k \sim A^k x_0$$

Theory: Converges to dominant eigenvector u_1 only if init vector x_0 satisfies

$$\langle x_0, u_1 \rangle \neq 0.$$

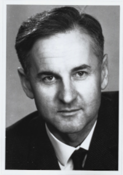
Practice: Preferably

$$|\langle x_0, u_1 \rangle| \not\ll 1.$$

u is not known a priori:

$\nexists$ reasonable deterministic choice for x_0!

```
procedure random(l); comment globals(n, z, x);
    value l;
    integer l;
    comment fills random elements into column l of matrix x;
    begin
        integer k, f, g, h, y, m;
        m := entier(z/4096 + 0.0001);
        y := z − 4096×m;
        for k := 1 step 1 until n do
        begin
            f := 1001×y + 3721;
            g := entier(f/4096 + 0.0001);
            f := f − 4096×g;
            h := 1001×(m + y) + g + 173;
            g := entier(h/4096 + 0.0001);
            h := h − 4096×g;
            m := h;
            y := f;
            x[k, l] := (m + y/4096)/2048 − 1;
        end for k;
        z := 4096×m + y;
end random;
```



Rutishauser'1970 (Handbook Series Linear Algebra): Algol 60 code (pre-pre-LAPACK) used for "randomly" initializing power method / subspace iteration

Generic 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 fixed $y \in \mathbb{R}^n$, 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.
- Given 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.

$\Sigma \in \mathbb{R}^{r \times r}$

What do these actually statements mean?

Why do they hold?

$$\begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{array}{cc} & r \\ n & \begin{bmatrix} \Sigma X_1 \\ 0 \end{bmatrix} \end{array}$$

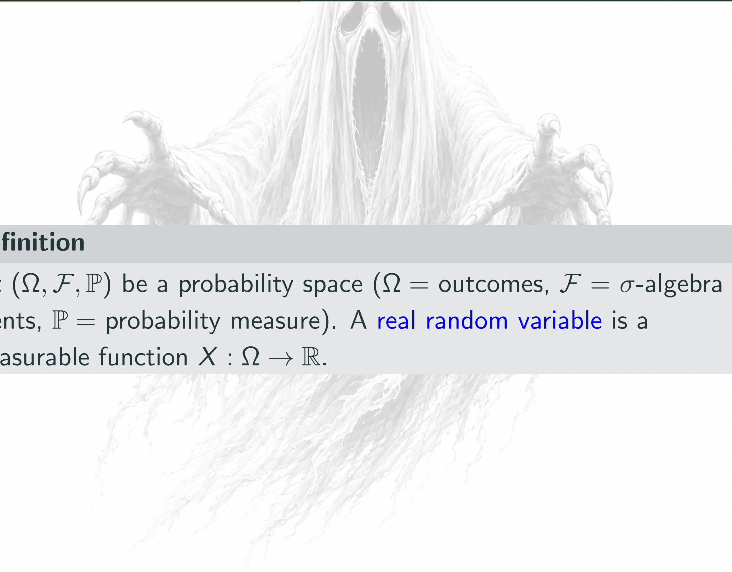
Robustification/quantification via randomness

- Given random $\boldsymbol{\alpha} \in \mathbb{R}$, what is the probability that $|\boldsymbol{\alpha}| < \epsilon$ for $\epsilon > 0$?
- Given random $\boldsymbol{\omega} \in \mathbb{R}^n$, what is the probability that $\|\boldsymbol{\omega}\|_2 < \epsilon$ for $\epsilon > 0$?
- Given fixed $y \in \mathbb{R}^n$ and random $\boldsymbol{\omega}$ what is the probability that $|\langle \boldsymbol{\omega}, y \rangle| < \epsilon \|y\|_2$ for $\epsilon > 0$?.
- Given random $\boldsymbol{\Omega} \in \mathbb{R}^{n \times n}$, what is the probability that $\|\boldsymbol{\Omega}^{-1}\|_2 \leq C$ for $C > 0$?
- Given fixed $m \times n$ matrix A of *approximate* rank r, how well does the span of $A\boldsymbol{\Omega}$ approximate the range of A for random (tall+skinny) $\boldsymbol{\Omega}$?

What does "random" actually mean?

Recap of random variables

Definition of real random variable



Definition

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space ($\Omega =$ outcomes, $\mathcal{F} = \sigma$-algebra of events, $\mathbb{P} =$ probability measure). A real random variable is a measurable function $X : \Omega \to \mathbb{R}$.

Андрей Николаевич Колмогоров

Discrete random variable

Distribution of discrete random variable X taking values in $\Omega = \{1, 2, \ldots\}$ fully specified by

$$\mathbb{P}(X = i) = p_i.$$

$p_1, p_2, \ldots$ discrete probabilities with $p_i \geq 0$ and $p_1 + p_2 + \cdots = 1$.

For arbitrary event $E \in 2^\Omega$:

$$\mathbb{P}(X \in E) = \sum_{i \in E} p_i.$$

Example: fair die

$$\Pr(\text{even}) = \Pr(2) + \Pr(4) + \Pr(6)$$
$$= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$$

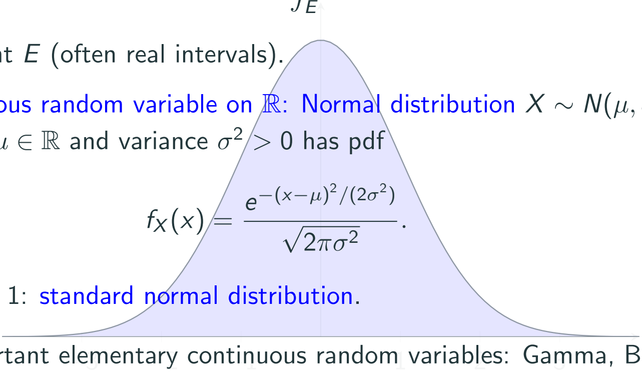
#1 discrete random variable: Rademacher random variable X

$$\Omega = \{+1, -1\}, \quad \Pr(X = -1) = \Pr(X = 1) = 1/2.$$

Distribution of any *continuous* random variable X is fully specified by its pdf (probability density function) f_X:

$$\Pr(X \in E) = \int_E f_X(x)\,\mathrm{d}x$$

for any event E (often real intervals).



#1 continuous random variable on $\mathbb{R}$: Normal distribution $X \sim N(\mu, \sigma^2)$ with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$ has pdf

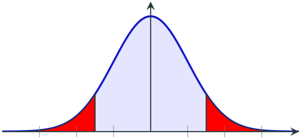
$$f_X(x) = \frac{e^{-(x-\mu)^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}}.$$

$\mu = 0, \sigma^2 = 1$: standard normal distribution.

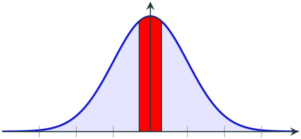
Other important elementary continuous random variables: Gamma, Beta, exponential, Cauchy, χ^2.

Two frequent questions asked on random variables

Tail bound



Small ball



$$\boxed{X \sim N(0,1)}$$

$$\Pr\{|X| \geq t\} \leq 2e^{-t^2/2}$$

$$\Pr\{|X| \leq \epsilon\} \leq \sqrt{\frac{2}{\pi}}\epsilon$$

The main trouble with random variables

Often, random variables arise from compositions of functions with elementary random variables. Only in *rare* cases is it possible to write down the pdf.

Still doable: Let $X \sim N(0,1)$. Pdf of X^2:

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

This is called χ_1^2 distribution (chi squared with one degree of freedom).

Techniques to deal with more difficult situations:

- Massage complicated random quantities one wants to understand until familiar random quantities pop up.
- Focus on moments rather than pdfs.
- Arithmetic of tail behavior (subexponential, sub-Gaussian, ... $\rightsquigarrow$ Alice's talk).

Recap of random vectors

X, Y independent ?

$$\begin{pmatrix} X \\ Y \end{pmatrix}$$

Two real random variables

Attention!

Beware: Two random quantities may be coupled!

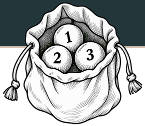
- Two continuous r.v. X, Y are specified via their *joint pdf* $f_{X,Y}(x, y)$ with $f_{X,Y} \geq 0$, $\int_{\mathbb{R}^2} f_{X,Y} = 1$. Example:

$$\mathbb{P}(X \leq a \text{ and } Y \leq b) = \int_{(-\infty,a]\times(-\infty,b]} f_{X,Y}(x, y)\,(\mathrm{d}x \times \mathrm{d}y),$$

- Two discrete r.v. X, Y are specified via their *joint distribution*

$$\mathbb{P}(X = i \text{ and } Y = j), \quad i, j = 1, 2, \ldots.$$

Marginal vs. conditional distribution



Experiment. A bag contains three balls labelled $1, 2, 3$. Draw two balls successively, *without replacement*.

$$X = \text{first draw}, \qquad Y = \text{second draw}.$$

Before seeing the first draw

By symmetry,

$$\boxed{\Pr(Y = j) = \frac{1}{3}, \qquad j = 1, 2, 3.}$$

This is the **marginal distribution** of Y.

Now suppose $X = 2$

The ball labelled 2 is gone, so

$$\boxed{\Pr(Y = j \mid X = 2) = \begin{cases} \frac{1}{2}, & j = 1, 3, \\ 0, & j = 2. \end{cases}}$$

This is the **conditional distribution** of Y.

Marginal vs. conditional distribution, independence

Consider two continuous r.v. X, Y with joint pdf $f_{X,Y}$.

- The pdfs of the *marginal* distributions of X and Y are
$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}y, \qquad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)\,\mathrm{d}x.$$
- For $f_Y(y) > 0$, the pdf of the distribution of X *conditioned* on $Y = y$ is
$$f_{X \mid Y=y}(x) = \frac{f_{X,Y}(x,y)}{f_Y(y)}.$$

Independence

X and Y are called independent if their joint pdf factorizes:

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

In this case, $f_{X \mid Y=y}(x) = f_X(x)$, so conditioning on Y does not change the distribution of X.

Real random vectors

Definition

A random vector $X = \left[X_1, \ldots, X_n\right]$ is just a tuple of random variables $X_1, \ldots, X_n$ specified through their joint distribution.

If $X_1, \ldots, X_n$ are independent it is sufficient to specify their marginal distributions.

If, additionally, $X_1, \ldots, X_n$ have the same distribution, we say that $X_1, \ldots, X_n$ are iid.

Three popular random vectors: 1. Rademacher

For a Rademacher random vector $X = [X_1, \dots, X_n]$, the entries X_i are iid Rademacher variables, that is,

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

Some trivial facts:

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

A nontrivial (and important) question:
Behavior of Rademacher sum $\langle X, v \rangle$ for fixed vector v.

Three popular random vectors: 2. Gaussian

If $X_1, \ldots, X_n$ are iid $N(0,1)$ then $X = [X_1, \ldots, X_n]$ is called a Gaussian random vector (also: standard normal random vector)

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

$X \sim N(0, \Sigma)$ $\quad f(x) \sim \exp\left(-\frac{x^T \Sigma x}{2}\right)$

$Q^T X \sim N(0, Q^T \Sigma Q)$

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

- $\langle X, a \rangle \sim N(0, \|a\|_2^2)$ for 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 $Y = \langle X, a \rangle / \|a\|_2$ to zero? Because $Y \sim N(0, 1)$, we know

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

Oblivious to a!

Three popular random vectors: 3. 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

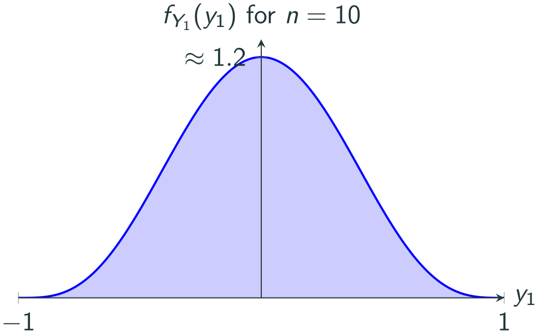
Y is uniformly distributed on the sphere S^{n-1}. We write $Y \sim U(S^{n-1})$.

The components of Y are *not* independent.

What is the distribution of Y_1 *conditioned* on $Y_2, \dots, Y_n$?

What is the marginal distribution of Y_1?

3. Uniform distribution on sphere

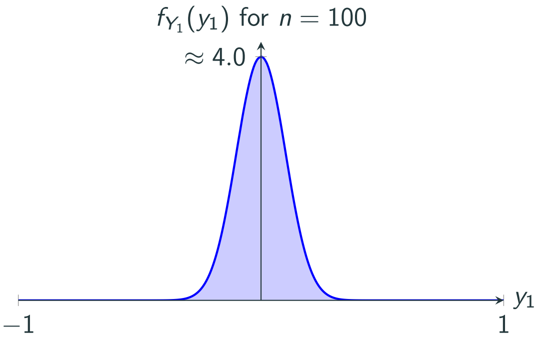


The marginal distribution of Y_1 has pdf

$$f_{Y_1}(y_1) = \begin{cases} 0 & \text{if } |y_1| > 1, \\ c_n(1 - y_1^2)^{(n-3)/2} & \text{otherwise.} \end{cases}$$

Constant c_n determined by density integrating to 1: $c_n \leq \sqrt{(n-1)/2\pi}$.

3. Uniform distribution on sphere



The marginal distribution of Y_1 has pdf

$$f_{Y_1}(y_1) = \begin{cases} 0 & \text{if } |y_1| > 1, \\ c_n(1 - y_1^2)^{(n-3)/2} & \text{otherwise.} \end{cases}$$

Constant c_n determined by density integrating to 1: $c_n \leq \sqrt{n/2\pi}$.

Back to the power method

Consider power method for SPD A with eigenvalues $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_n > 0$:

$$\tilde{x}_{k+1} = Ax_k, \qquad x_{k+1} = \frac{1}{\|\tilde{x}_{k+1}\|_2}\tilde{x}_{k+1}.$$

Converges to normalized eigenvector u_1 belonging to λ_1 for almost every starting vector x_0 .

Convergence speed:

$$\boxed{\tan\angle(u_1, x_k) \leq \left(\frac{\lambda_2}{\lambda_1}\right)^k \tan\angle(u_1, x_0).}$$

Note that $\tan\angle(u_1, x_0) \leq \|x_0\|/|\langle u_1, x_0\rangle|$

Back to the power method

Now suppose that $x_0 \sim N(0, I_n)$. Then

$$\frac{\|x_0\|_2}{|\langle u_1, x_0\rangle|} = \frac{1}{|\langle u_1, y\rangle|} \stackrel{d}{=} \frac{1}{|Y_1|},$$

where Y is uniformly distributed on $\mathbb{S}^{n-1}$. Because of $|f_{Y_1}(y_1)| \leq \sqrt{\frac{n-1}{2\pi}}$,

$$\Pr\{|Y_1| \leq \eta\} \leq 2\eta\sqrt{\frac{n-1}{2\pi}} = \eta\sqrt{\frac{2(n-1)}{\pi}}, \qquad \eta > 0.$$

We obtain, with probability at least $1-\delta$,

$$\boxed{\frac{\|x_0\|_2}{|\langle u_1, x_0\rangle|} \leq \frac{1}{\delta}\sqrt{\frac{2(n-1)}{\pi}}}.$$

Thus, to guarantee $\tan\angle(u_1, x_k) \leq \epsilon$ with probability at least $1-\delta$, it suffices to choose

$$k = O\big(\big(\log n + \log(\epsilon^{-1}) + \log(\delta^{-1})\big)/\log(\lambda_1/\lambda_2)\big).$$

No deterministic construction of x_0 is known that comes even close!

A short note on 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.

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.