Deck 2: A Review of Probability

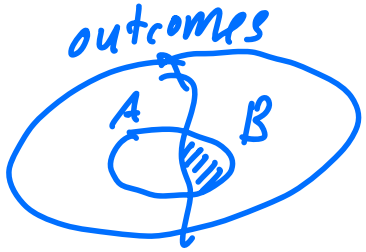
Math 7870: Topics in Randomized Numerical Linear Algebra

Spring 2026

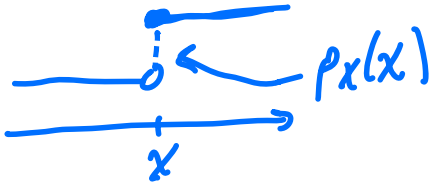
Akil Narayan

Notation

Some consistent notation we'll use:



- Probability of an event ω: $\Pr(\omega)$
- Probability events: A, B, etc.
- Conditional probability: $\Pr(A|B) = \Pr(A \cap B)/\Pr(B)$
- Scalar random variables (RV): uppercase Roman characters X, Y, etc.
- The (cumulative) distribution function for X: $F_X(x) = \Pr(X \leqslant x)$
- The mass function for discrete X: $p_X(x) = \Pr(X = x) = F_X(x) - \lim_{y \to x^-} F_X(y)$
- The density function for continuous X: $f_X(x) = F_X'(x)$
- Joint and conditional distributions: $F_{X,Y}(x, y)$, $F_{X|Y}(x|y)$.
- The expectation operator:



$$\mathbb{E}h(X) = \begin{cases} \int h(x) f_X(x) \mathrm{d}x, & X \text{ is continuous} \\ \sum_j h(x)_j p_X(x_j), & X \text{ is discrete} \end{cases}$$

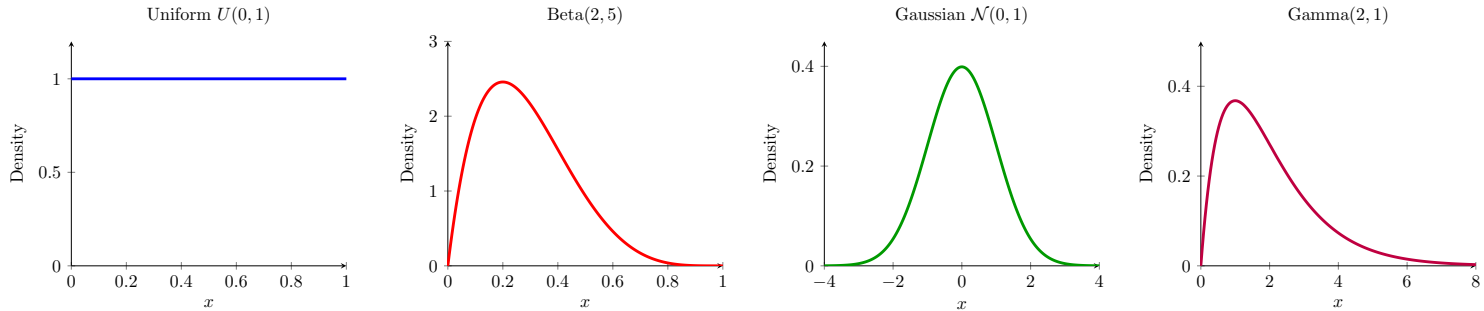
- Moments: $\mathbb{E}X$ is the mean, $\mathbb{E}(X - \mathbb{E}X)^2$ is the variance, etc.

$\mathrm{Var}(X)$

Standard probability distributions

Continuous:

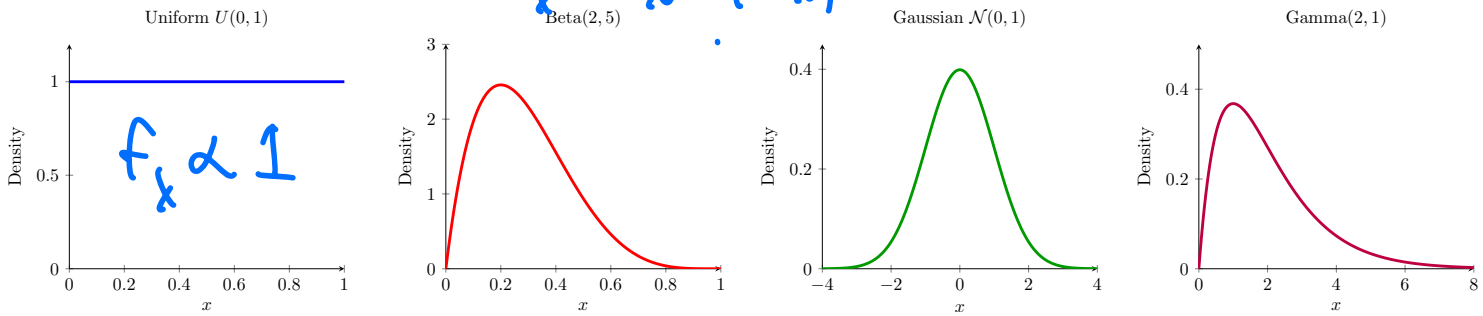
- Uniform:
- Beta
- Gaussian
- Gamma



Standard probability distributions

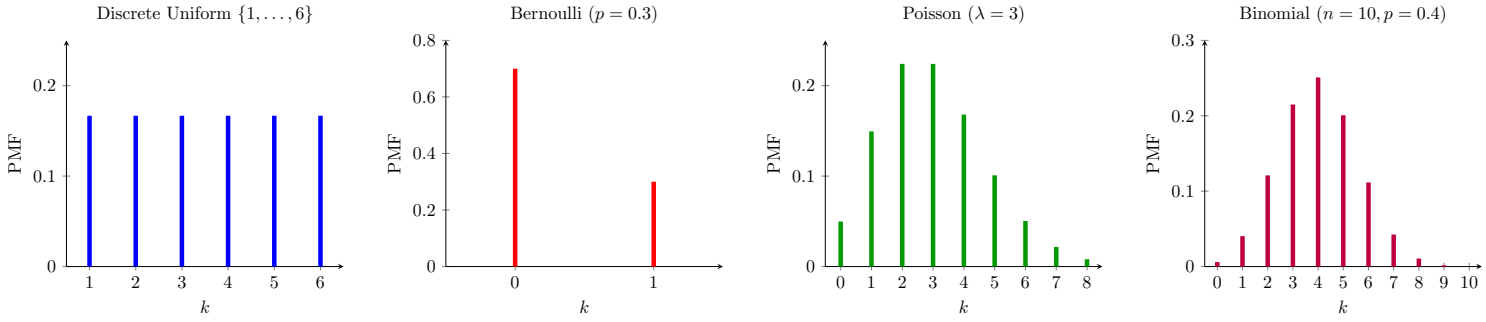
Continuous:

- Uniform:
- Beta
- Gaussian
- Gamma



Discrete:

- Uniform
- Poisson
- Bernoulli
- Binomial



Sampling from non-standard distributions is generally tricky, but is simple for discrete distributions.

Sampling from non-standard discrete distributions

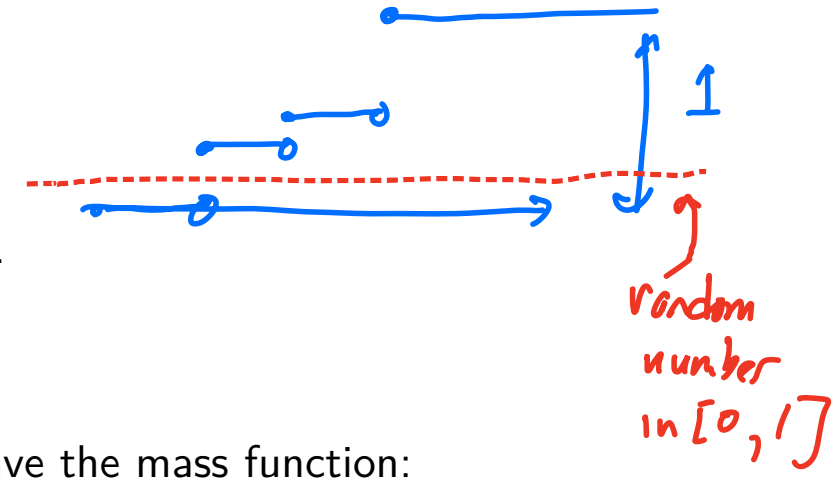
Let X be a finitely-supported (discrete) random variable, with ordered outcomes $x_1, \ldots, x_n$.

Given its distribution F_X, we can generate samples as follows:

Let Y be a (continuous) uniform RV on $[0, 1]$.

- Generate a realization Y_0 of Y.
- Find the smallest $j \in [n]$ such that $F_X(x_j) \geqslant Y_0$.
- Set $X_0 = x_j$.

Then X_0 is a random variable distributed according to X.



(This is called inverse transform sampling.)

Note that $F_X(x_k)$ is rather easy to compute if we only have the mass function:

$$F_X(x_k) = \sum_{\ell=1}^{k} p_X(x_\ell).$$

Asymptotics, I

Some crown jewels of probability theory are asymptotic results on the concentration of sums:

Let X_j be independent and identically distributed (iid) copies of X. The core question is

$$\text{What kind of quantity is } \sum_{j=1}^{n} X_j?$$

As written, this sum is unbounded in n. Normalizing by n gives a sample average.

Asymptotics, I

Some crown jewels of probability theory are asymptotic results on the concentration of sums:

Let X_j be independent and identically distributed (iid) copies of X. The core question is

$$\text{What kind of quantity is } \sum_{j=1}^{n} X_j?$$

As written, this sum is unbounded in n. Normalizing by n gives a sample average.

The Law of Large Numbers asserts that, if X has finite mean, then the sample average converges to the mean:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j\in[n]} X_j = \mathbb{E}X$$

This limit holds "strongly", meaning that the probability that the equality holds is 1.
This implies a "weak" statement, that for large n the probability of a small deviation of the sample average from $\mathbb{E}X$ converges to 1.

Asymptotics, II

The LLN encapsulates essentially the entire message of this course: If we don't know $\mathbb{E}X$, we can approximate it by taking an empirical average.

The main deficiency of the LLN: it provides no preasymptotic information. If $n = 1000$, we have no idea how good of an estimator the sample mean is.

To characterize discrepancy of the estimator for fixed n, we want to know the "size" of the following random variable:

$$\mathbb{E}X - \frac{1}{n} \sum_{j \in [n]} X_j$$

We expect that the random variable gets closer to 0 for larger n.

Asymptotics, II

The LLN encapsulates essentially the entire message of this course: If we don't know $\mathbb{E}X$, we can approximate it by taking an empirical average.

The main deficiency of the LLN: it provides no preasymptotic information. If $n = 1000$, we have no idea how good of an estimator the sample mean is.

To characterize discrepancy of the estimator for fixed n, we want to know the "size" of the following random variable:

$$\mathbb{E}X - \frac{1}{n}\sum_{j\in[n]} X_j$$

We expect that the random variable gets closer to 0 for larger n.

The first moment of this random variable is not informative: its mean is 0. What about its second moment?
Because this is a sum of iid random variables, we have,

$$\mathrm{Var}\left(\mathbb{E}X - \frac{1}{n}\sum_{j\in[n]} X_j\right) = \frac{1}{n}\mathrm{Var}(X).$$

Asymptotics, III

$$\text{StDev}\left(\mathbb{E}X - \frac{1}{n}\sum_{j\in[n]} X_j\right) = \frac{\text{const}}{\sqrt{n}} = \frac{\text{StDev}(X)}{\sqrt{n}}$$

There is a strong quantitative statement of this fact, the Central Limit Theorem (CLT).

Assume that X has finite first and second moments. Then for large n,

$$\sqrt{n}\left(\mathbb{E}X - \frac{1}{n}\sum_{j\in[n]} X_j\right) \xrightarrow{d} \mathcal{N}(0, \text{Var}(X)) \text{ (Normal random variable of mean 0 and variance Var}(X))$$

Above, $\xrightarrow{d}$ means converges "in distribution", i.e., its distribution function converges to the distribution function of a normal random variable.

Asymptotics, III

$$\text{StDev}\left(\mathbb{E}X - \frac{1}{n}\sum_{j\in[n]} X_j\right) = \frac{\text{const}}{\sqrt{n}} = \frac{\text{StDev}(X)}{\sqrt{n}}$$

There is a strong quantitative statement of this fact, the Central Limit Theorem (CLT).

Assume that X has finite first and second moments. Then for large n,

$$\sqrt{n}\left(\mathbb{E}X - \frac{1}{n}\sum_{j\in[n]} X_j\right) \xrightarrow{d} \mathcal{N}(0, \text{Var}(X)) \text{ (Normal random variable of mean 0 and variance Var}(X))$$

Above, $\xrightarrow{d}$ means converges "in distribution", i.e., its distribution function converges to the distribution function of a normal random variable.

In summary: the LLN tells us the limit of an empirical sum. The CLT tell us the (asymptotic) discrepancy between the empirical sum and its limit.

These are the simplest examples of (asymptotic) *concentration estimates*: they tell us about the limit of an empirical sum.

Random sums in NLA: a simple example, I

Here's a somewhat transparent example of why concentration estimates might be useful for NLA:

Consider $\boldsymbol{A} \in \mathbb{C}^{m\times k}$ and $\boldsymbol{B} \in \mathbb{C}^{k\times n}$.

Our scenario: suppose $k \gg m, n \gg 1$. Computing $\boldsymbol{AB}$ directly is (super) expensive.

Random sums in NLA: a simple example, I

Here's a somewhat transparent example of why concentration estimates might be useful for NLA:

Consider $\boldsymbol{A} \in \mathbb{C}^{m \times k}$ and $\boldsymbol{B} \in \mathbb{C}^{k \times n}$.

Our scenario: suppose $k \gg m, n \gg 1$. Computing $\boldsymbol{AB}$ directly is (super) expensive.

The matrix-matrix product $\boldsymbol{AB}$ is both

- an array of inner products (between columns of $\boldsymbol{A}^*$ and columns of $\boldsymbol{B}$)
- a sum of outer products (between columns of $\boldsymbol{A}$ and columns of $\boldsymbol{B}^*$)

This latter form is the more useful interpretation for us.

$$A = \begin{pmatrix} | & & | \\ a_1 & \cdots & a_k \\ | & & | \end{pmatrix} \qquad B = \begin{pmatrix} - & b_1^* & - \\ & \vdots & \\ - & b_k^* & - \end{pmatrix} \qquad AB = \sum_j a_j b_j^*$$

Random sums in NLA: a simple example, II

$\boldsymbol{A} \in \mathbb{C}^{m\times k}$ and $\boldsymbol{B} \in \mathbb{C}^{k\times n}$. Compute $\boldsymbol{AB}$.

Here's how we can mitigate the cost of summing over k in this example:

Let $(\boldsymbol{a}_j)_{j\in[k]}$ be the columns of $\boldsymbol{A}$, and let $(\boldsymbol{b}_j^*)_{j\in[k]}$ be the rows of $\boldsymbol{B}^*$:

$$\boldsymbol{A} = \begin{pmatrix} | & | & & | \\ \boldsymbol{a}_1 & \boldsymbol{a}_2 & \cdots & \boldsymbol{a}_k \\ | & | & & | \end{pmatrix}, \qquad \boldsymbol{B} = \begin{pmatrix} — & \boldsymbol{b}_1^* & — \\ — & \boldsymbol{b}_2^* & — \\ & \vdots & \\ — & \boldsymbol{b}_k^* & — \end{pmatrix}.$$

Random sums in NLA: a simple example, II

$\boldsymbol{A} \in \mathbb{C}^{m\times k}$ and $\boldsymbol{B} \in \mathbb{C}^{k\times n}$. Compute $\boldsymbol{AB}$.

$$AB = \sum_{j\in[k]} a_j b_j^*$$

Here's how we can mitigate the cost of summing over k in this example:

Let $(\boldsymbol{a}_j)_{j\in[k]}$ be the columns of $\boldsymbol{A}$, and let $(\boldsymbol{b}_j^*)_{j\in[k]}$ be the rows of $\boldsymbol{B}^*$:

$$\boldsymbol{A} = \begin{pmatrix} | & | & & | \\ \boldsymbol{a}_1 & \boldsymbol{a}_2 & \cdots & \boldsymbol{a}_k \\ | & | & & | \end{pmatrix}, \qquad \boldsymbol{B} = \begin{pmatrix} — & \boldsymbol{b}_1^* & — \\ — & \boldsymbol{b}_2^* & — \\ & \vdots & \\ — & \boldsymbol{b}_k^* & — \end{pmatrix}.$$

Define the $\mathbb{C}^{m\times n}$-valued random matrix $\boldsymbol{X}$ that has uniform mass on its support:

$$p_{\boldsymbol{X}}(k\boldsymbol{a}_j\boldsymbol{b}_j^*) = \frac{1}{k} \quad (\boldsymbol{X} \text{ uniformly at random selects a } k\text{-scaled outer product})$$

$$\mathbb{E}\,\underline{X} = \sum_j p_{\underline{X}}(k\underline{a}_j\underline{b}_j^*)\cdot k\underline{a}_j\underline{b}_j^* = \sum_j \frac{1}{k}\cdot k\,\underline{a}_j\underline{b}_j^* = AB$$

Random sums in NLA: a simple example, III

A direct computation reveals:

$$\mathbb{E}\boldsymbol{X} = \sum_{j\in[k]} p_{\boldsymbol{X}}(k\boldsymbol{a}_j\boldsymbol{b}_j^*)k\boldsymbol{a}_j\boldsymbol{b}_j^* = \frac{1}{k}\sum_{j\in[k]} k\boldsymbol{a}_j\boldsymbol{b}_j^* = \boldsymbol{AB}.$$

Random sums in NLA: a simple example, III

A direct computation reveals:

$$\mathbb{E}\boldsymbol{X} = \sum_{j\in[k]} p_{\boldsymbol{X}}(k\boldsymbol{a}_j\boldsymbol{b}_j^*)k\boldsymbol{a}_j\boldsymbol{b}_j^* = \frac{1}{k}\sum_{j\in[k]} k\boldsymbol{a}_j\boldsymbol{b}_j^* = \boldsymbol{AB}.$$

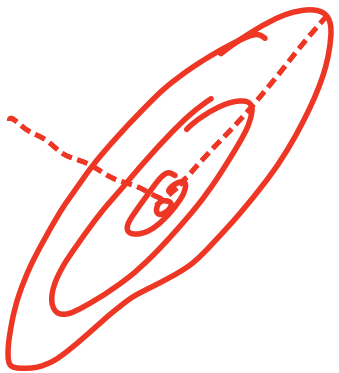
The above realization is the linchpin for a randomized method: Now let $\boldsymbol{X}_j$, $j \geqslant 1$ be iid copies of $\boldsymbol{X}$. Then for some $N \geqslant 1$, the (multivariate) LLN and CLT tell us:

$$\frac{1}{N}\sum_{q\in[N]} \boldsymbol{X}_q \xrightarrow{N\uparrow\infty} \boldsymbol{AB}, \qquad \sqrt{N}\,\mathrm{vec}\left(\boldsymbol{AB} - \frac{1}{N}\sum_{q\in[N]} \boldsymbol{X}_q\right) \overset{N=\infty}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma}$ satisfies,

$\Sigma = \mathrm{Cov}(\mathrm{vec}(X))$

$$\mathrm{tr}(\boldsymbol{\Sigma}) = V := k\sum_{i,j,q} |A_{i,q}B_{q,j}|^2 - \|\boldsymbol{AB}\|_F^2 \geqslant 0$$



Random sums in NLA: a simple example, III

A direct computation reveals:

$$\mathbb{E}\boldsymbol{X} = \sum_{j\in[k]} p_{\boldsymbol{X}}(k\boldsymbol{a}_j\boldsymbol{b}_j^*)k\boldsymbol{a}_j\boldsymbol{b}_j^* = \frac{1}{k}\sum_{j\in[k]} k\boldsymbol{a}_j\boldsymbol{b}_j^* = \boldsymbol{AB}.$$

The above realization is the linchpin for a randomized method: Now let $\boldsymbol{X}_j$, $j \geqslant 1$ be iid copies of $\boldsymbol{X}$. Then for some $N \geqslant 1$, the (multivariate) LLN and CLT tell us:

$$\frac{1}{N}\sum_{q\in[N]} \boldsymbol{X}_q \xrightarrow{N\uparrow\infty} \boldsymbol{AB}, \qquad \sqrt{N}\,\mathrm{vec}\left(\boldsymbol{AB} - \frac{1}{N}\sum_{q\in[N]} \boldsymbol{X}_q\right) \overset{N=\infty}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma}$ satisfies,

$$\mathrm{tr}(\boldsymbol{\Sigma}) = V := k\sum_{i,j,q} |A_{i,q}B_{q,j}|^2 - \|\boldsymbol{AB}\|_F^2 \geqslant 0$$

This is a useful result: if V is not too big, then for enormous k, we can take $N \ll k$ and obtain a reasonable estimate for $\boldsymbol{AB}$ with some understanding of the error.
(NB: $V = 0$ if $\boldsymbol{a}_j = \boldsymbol{a}$ and $\boldsymbol{b}_j = \boldsymbol{b}$ are j-independent vectors.)

A prototypical randomized matmat

Here is a prototype of a randomized algorithm for the matrix-matrix multiplication $\boldsymbol{AB}$, given $\boldsymbol{A}$ and $\boldsymbol{B}$:

0. Prescribe N. Set $i = 0$, $\boldsymbol{C} = \boldsymbol{0} \in \mathbb{C}^{m \times n}$.
1. If $i = N$, go to step 5.
2. Choose $j \in [k]$ uniformly at random.
3. Set $\boldsymbol{C} \leftarrow \frac{k}{N} \boldsymbol{a}_j \boldsymbol{b}_j^*$. $C \leftarrow C + \frac{k}{N} a_j b_j^*$
4. Set $i \leftarrow i + 1$. Go to step 1.
5. Return $\boldsymbol{C}$.

This algorithm comes with some understanding of error: $\boldsymbol{C} - \boldsymbol{AB}$ is centered, with total variance scaling like V/N.

Are we done?

Some outcomes of this problem:

- We crafted a randomized algorithm for approximating $\boldsymbol{AB}$.
- Randomness is an *algorithmic convenience*. It is not a model of noise of stochasticity.
- The algorithm can fail badly with nonzero probability: E.g., there is nonzero probability that we always choose index $j = 1$, resulting in $\boldsymbol{C} = k\boldsymbol{a}_1\boldsymbol{b}_1^* \neq \boldsymbol{AB}$.
- With high probability the algorithm is not exact: $\boldsymbol{C} \neq \boldsymbol{AB}$ with high probability.

$\underline{\underline{A}} = e_1^* = (1\ 0\ \cdots\ 0)$

$\underline{\underline{B}} = e_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$

Are we done?

Some outcomes of this problem:

- We crafted a randomized algorithm for approximating $\boldsymbol{AB}$.
- Randomness is an *algorithmic convenience*. It is not a model of noise of stochasticity.
- The algorithm can fail badly with nonzero probability: E.g., there is nonzero probability that we always choose index $j = 1$, resulting in $\boldsymbol{C} = k\boldsymbol{a}_1\boldsymbol{b}_1^* \neq \boldsymbol{AB}$.
- With high probability the algorithm is not exact: $\boldsymbol{C} \neq \boldsymbol{AB}$ with high probability.

We haven't really solved this problem:

- The LLN and CLT are not quantitative enough: We don't have a precise bound on how $\boldsymbol{C}$ deviates from $\boldsymbol{AB}$, we only have N-asymptotic results.
- Without this quantitative bound, we cannot construct reliable algorithms/software.