Algorithms for Big Data (Fall 25)

Lecture 1

Logistic, Course Overview and Backgrounds

Ali Vakilian (vakilian@vt.edu)

Virginia Tech

Logistics

- **Schedule:** Tue/Thu, 12:30–1:45 pm (MCB 240)
- **Instructor:** ALI VAKILIAN (vakilian@vt.edu)
- **Office Hours:** **Wed, 11am–noon (?)**
- **Website & Canvas:** All announcements, slides, and assignments will be posted online on the course website and Canvas.

Evaluations

- **Homework (45%):** Three problem sets released, each 15%, roughly every five weeks.
- **Final Project (45%):**
  - Proposal (5%),
  - Checkpoint meeting (5%),
  - Presentation (10%), and
  - Final report (25%).
- **Class Participation (10%):** Active engagement is expected during lectures, and each student must contribute by scribing **at least one lecture** of their choice.

Any volunteer for scribing next lecture?

Evaluations (contd.)

Assignments and Deadlines

- Release and due dates appear on the course calendar. **No late submissions**. All work is due at the posted time; no slip days or late penalties apply.

Final-Project & Options

An opportunity to explore an area of modern big data algorithms.

- **Project Style:**
  - **Survey:** Read 3--5 recent research papers and write a mini-survey that highlights common themes, contrasting approaches, and open questions.
  - **Implementation:** Build and benchmark two (or more) competing algorithms on realistic data sets; evaluate trade-offs in accuracy, speed, and memory.
  - **Research:** Propose and develop a new theoretical or empirical result under close mentorship from the instructor.

Projects can be done alone or in pairs.
Surveys must be completed individually.

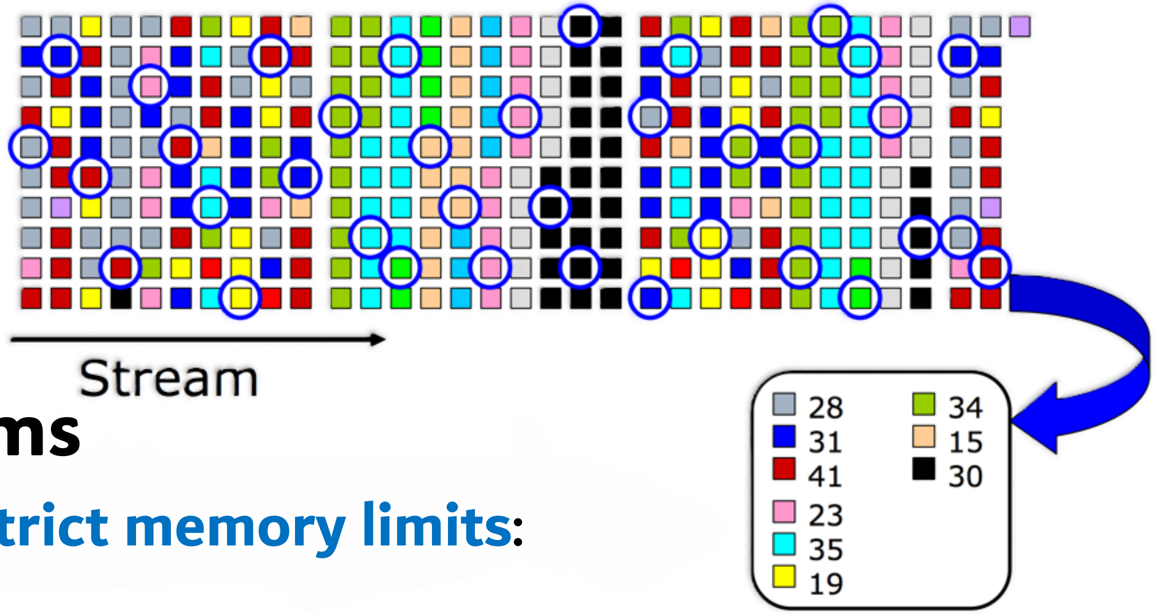
Evaluations (contd.)

Final-Project & Options

An opportunity to explore an area of modern big data algorithms.

- **Project Style**
- **Deliverables & Timeline**
  - **Proposal (5%)** – due **Week 5**. A 1-page PDF describing your topic, motivation, and an initial plan of study/research. Schedule a quick chat with the instructor to refine scope.
  - **Checkpoint Meeting (5%)** – due **Week 12**. A 15-minute meeting to review progress and adjust goals. Please bring preliminary results or a working demo.
  - **Presentations (10%)** – last three lecture slots. Each team/individual will give a **15-minute talk (+2 min Q&A)** to the class. These final sessions replace lectures and are a chance for peer learning.
  - **Final Report (25%)** – due reading day **(12/11/2025)**. A 6–8 page write-up summarizing motivation, methods, results, and future work. Submit both PDF and any code/data via Canvas.

What to Expect



- **Streaming & Sketching Algorithms**

A framework for handling **massive data** with **strict memory limits**:

- items arrive in a sequence,
- can be computed once (or in some cases a few times), and
- use low space, and output approximate answers

Examples (basic routines in data analytics):

- approximate counting with only $O(\log \log n)$ bits (Morris),
- distinct elements/cardinality (for 10^9 items, ~2% accuracy using 1.5kB) (HyperLogLog),
- heavy hitters with guarantees (Misra–Gries/CountSketch).

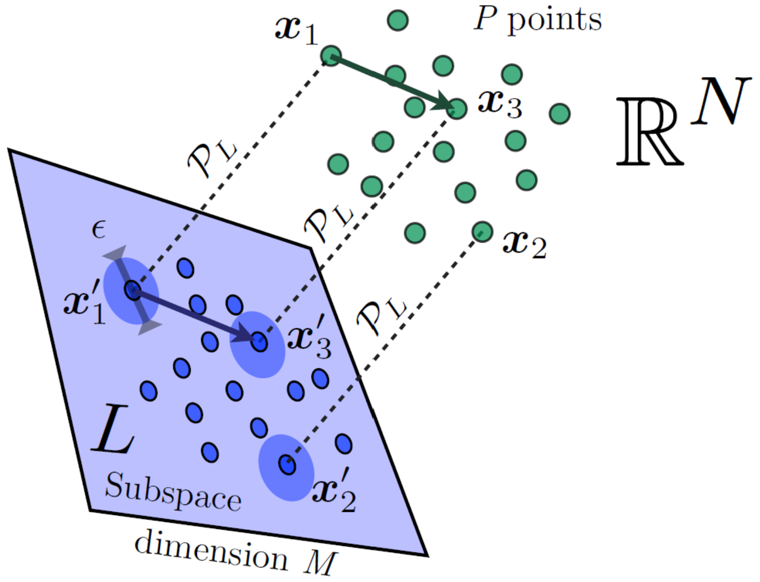
What to Expect

- **Streaming & Sketching Algorithms**
  - Vector statistics, such as frequency estimation and moment
  - Graph problems
  - Geometric problems
  - Linear sketching (e.g., CountMin and CountSketch)
  - Lowerbounds

What to Expect

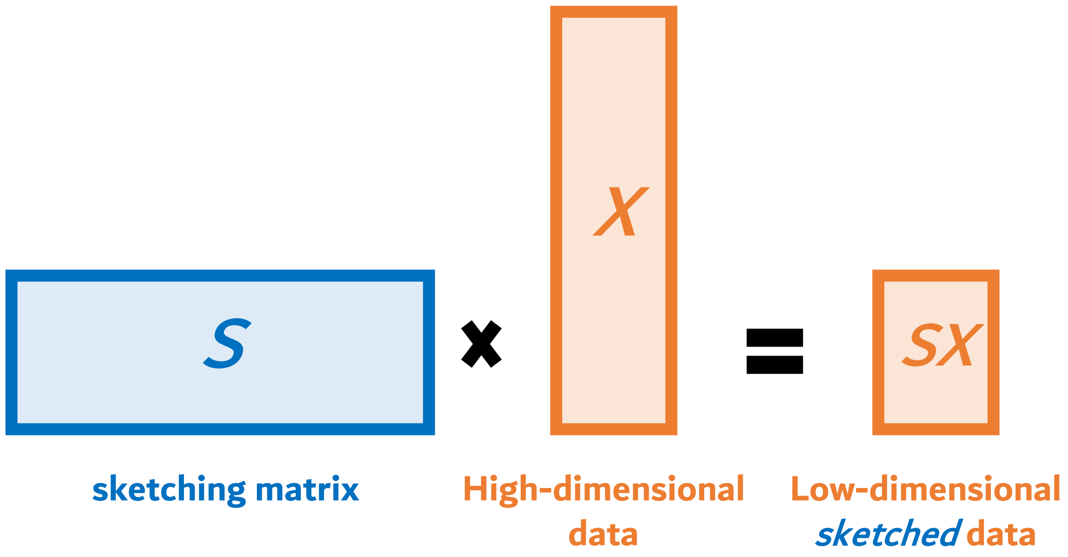
- **Dimensionality Reduction**

A technique to **compress high-dimensional** data into far fewer features while approximately preserving structure (e.g., distances or variance). It speeds up learning and visualization, cuts noise, and helps algorithms scale

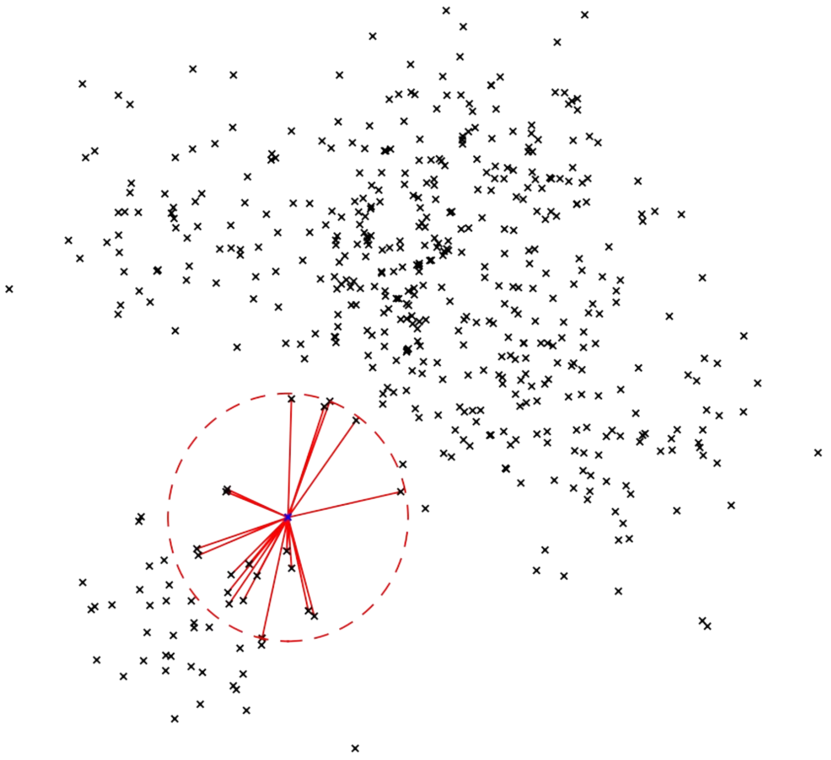


What to Expect

- **Numerical Linear Algebra at Scale**
  - Subspace embedding
  - Approximate matrix multiplication
  - Low-rank approximation and PCA



What to Expect



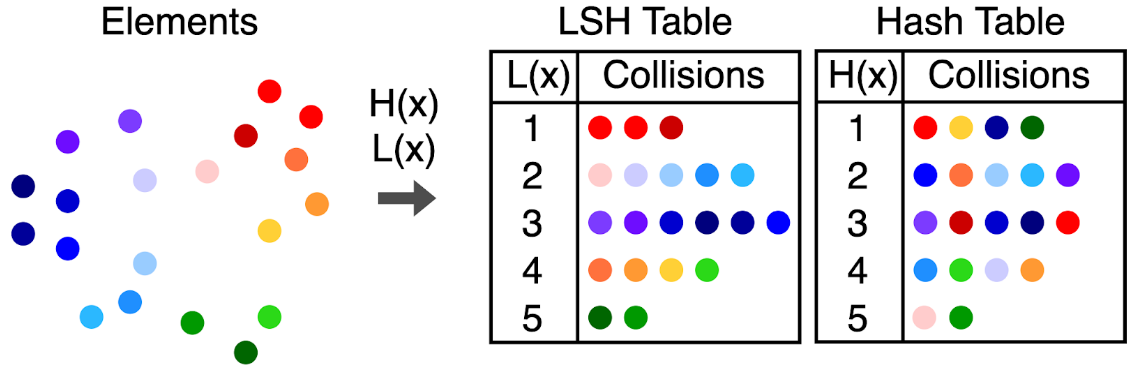
- **Nearest Neighbor Search**

Basic problem of finding the most similar vector to a query. In LLMs

- underpins RAG (fetching relevant chunks from vector stores),
- memory/caching and kNN-LM style prompts,
- and it also approximates/sparsifies attention (e.g., LSH/sparse routing)

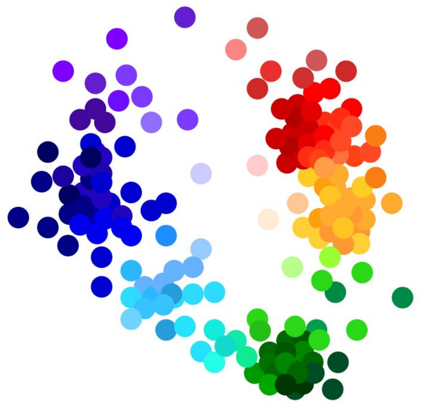
to cut quadratic costs, driving lower latency, larger contexts, and more efficient inference.

Locality Sensitive Hashing (LSH) [Indyk, Motwani'98]

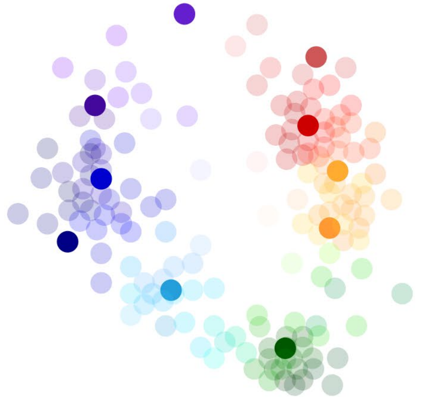


What to Expect

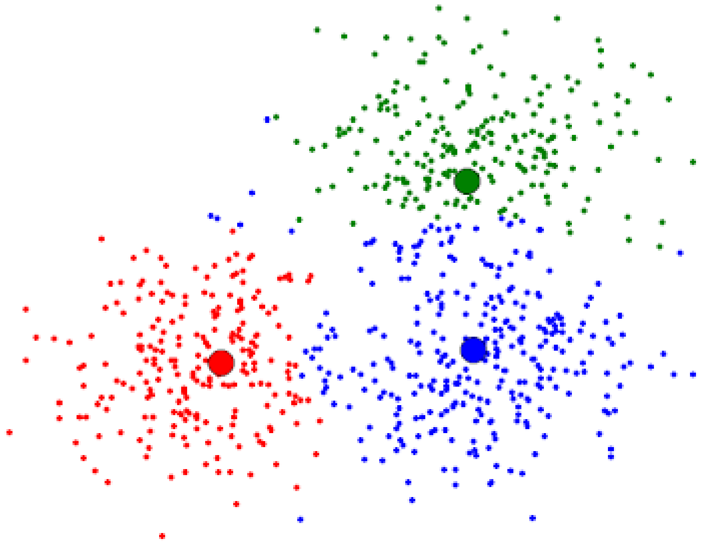
- **Clustering and its Coreset Constructions**
  - A technique of grouping unlabeled data points into coherent clusters based on similarity, revealing latent structure for summarization, anomaly detection, and retrieval.
  - Our focus is on clustering techniques such as **coreset constructions** for large scale data.



Massive Dataset



Coreset



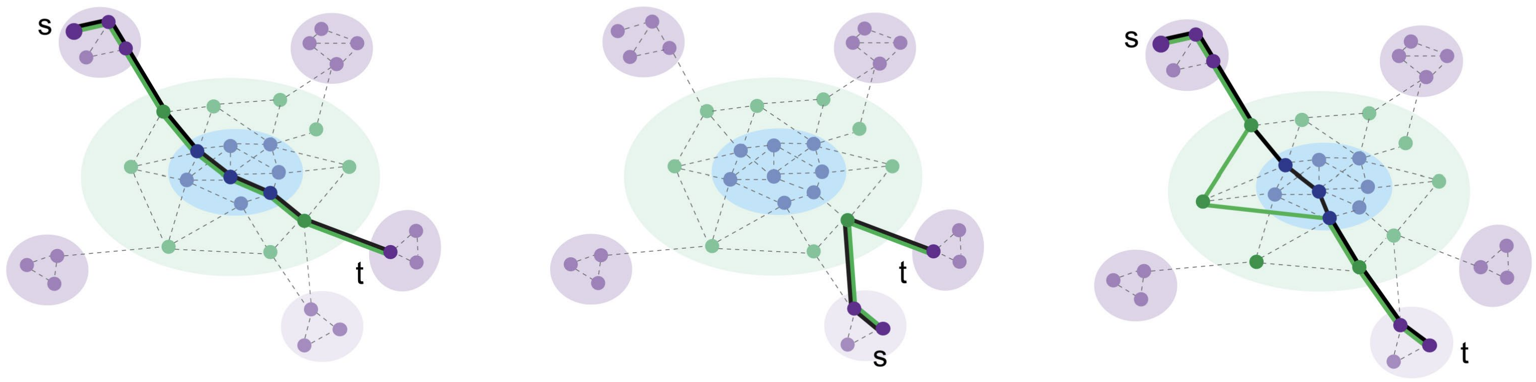
What to Expect

- **Sublinear Algorithms for Graph Problems**

They use sampling and sketches to estimate global properties (triangle counts, connectivity, PageRank) in sublinear time.

In ML pipelines, they enable web-scale candidate generation and monitoring, e.g.,

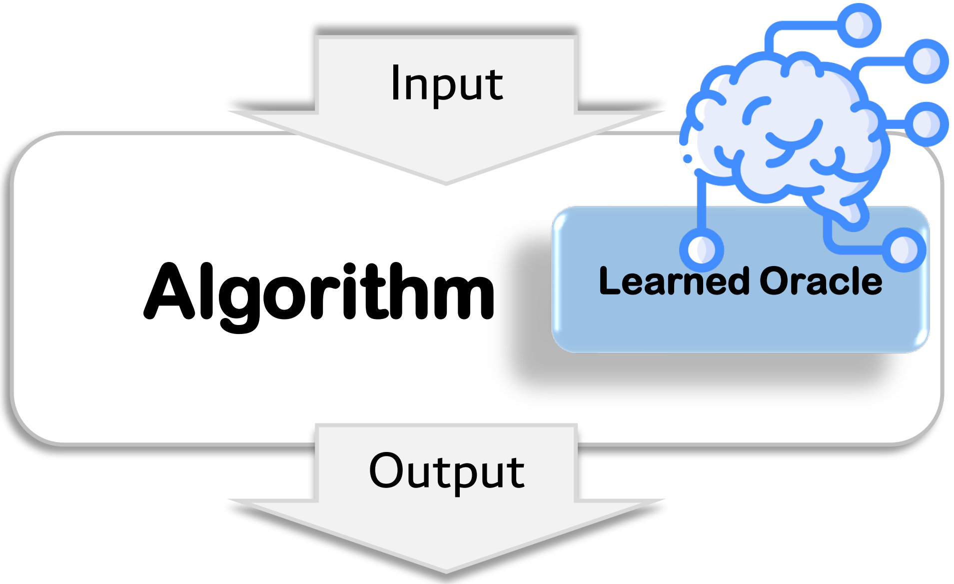
- approximate personalized PageRank for retrieval,
- graph sparsification and partitioning for faster GNNs,
- and fast statistics over evolving knowledge graphs for RAG.

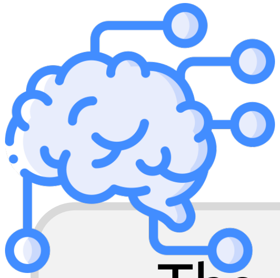


What to Expect

- **Learning-Augmented Algorithms/Data-Driven Algorithms**

I) **<u>Better performance</u>** when input has “learnable” pattern



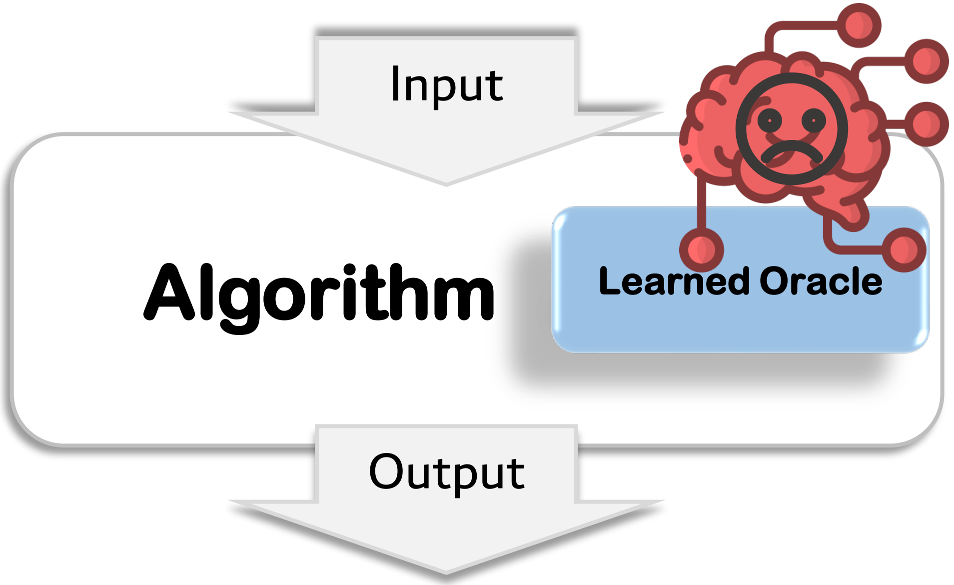


The algorithm has access to a **learned ORACLE** providing a certain type of **PREDICTIONS** about the instance in hand

What to Expect

- **Learning-Augmented Algorithms/Data-Driven Algorithms**

I) **<u>Better performance</u>** when input has “learnable” pattern



II) **<u>Similar worst-case guarantee</u>** as the best-known classical algorithms

What to Expect

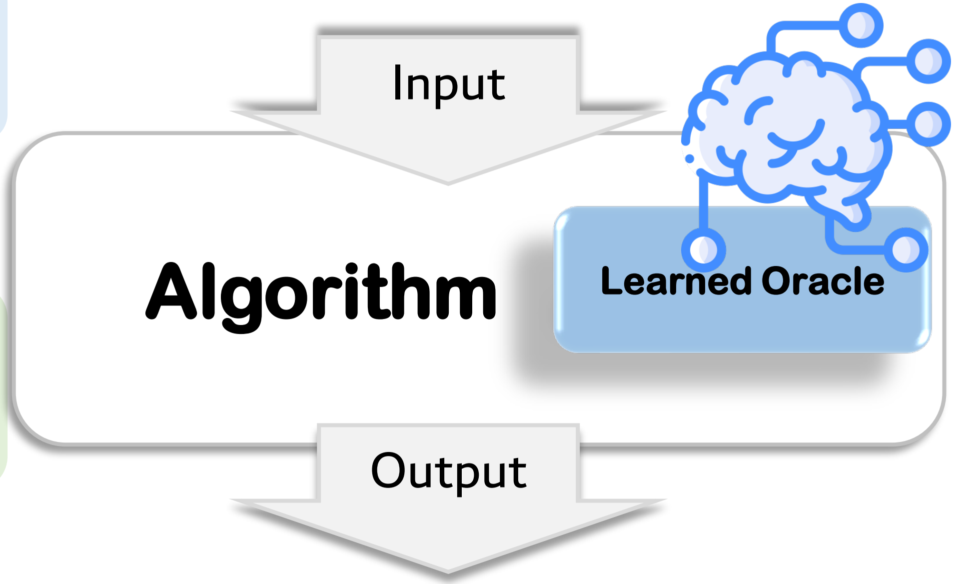
- **Learning-Augmented Algorithms/Data-Driven Algorithms**

ML

I) **Better performance** when input has "learnable" pattern

Algorithms

II) **Similar worst-case guarantee** as the best-known classical algorithms



Probabilities Refresher

Independence

- Two events A and B are *independent* if

$$\Pr[A \cap B] = \Pr[A] \times \Pr[B]$$

I. A collection of events are *independent*, if every subset obeys this equality.

II. A collection of events are *k-wise independent*, if every subset of size at most c obeys this equality. A commonly-used scenario is $k = 2$ which is referred to as *pairwise-independence* too.

Conditional Probability

- For $\Pr[B] > 0$, the probability of A given B is

$$\Pr[A|B] = \frac{\Pr[A \cap B]}{\Pr[B]}$$

Bayes' rule

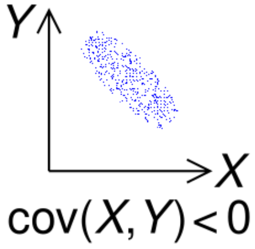
$$\Pr[B|A] = \Pr[A|B] \times \Pr[B]/\Pr[A]$$

More generally, for a partition of space, $\{B_i\}$

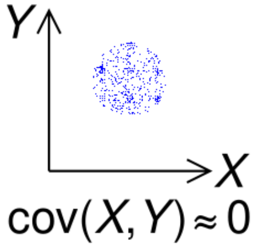
$$\Pr[B_j|A] = \frac{\Pr\left[A|B_j\right] \times \Pr[B_j]}{\sum_i \Pr[A|B_i] \times \Pr[B_i]}$$

Random Variables

- A random process that assign a number to each outcome of $\omega \in \Omega$.
  - Roll a dice, record the number of pips (e.g., $X = 6$)
  - Run Quicksort, record its runtime (e.g. $T = 100$)



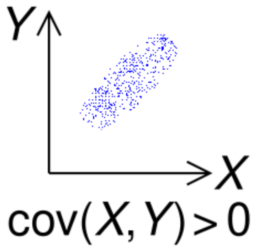
- Expectation: $\mathbb{E}[X] = \sum_{x \in \mathrm{range}(X)} x \cdot \Pr[X = x]$



- Variance: $\mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}X)^2]$

$\mathrm{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)]$

- Indicator Variable:



$\mathbf{I} = \mathbf{1}_E$ that is equal to 1 when event E occurs and 0 otherwise: $\mathbb{E}[\mathbf{1}_E] = \Pr[E]$

Random Variables

- **Independence of R.V. :** $X_1, \cdots, X_k$ are independent, if for every choice of real numbers $a_1, \cdots, a_k$,

$$\Pr[X_1 = a_1, \cdots, X_k = a_k] = \Pr[X_1 = a_1] \times \cdots \times \Pr[X_k = a_k]$$

- Expectation:

$$\mathbb{E}[X] = \sum_x x \cdot \Pr[X = x] \text{ (discrete)}, \ \mathbb{E}[X] = \int_{-\infty}^{+\infty} x \cdot f_X(x)\, dx \text{ (continuous)}$$

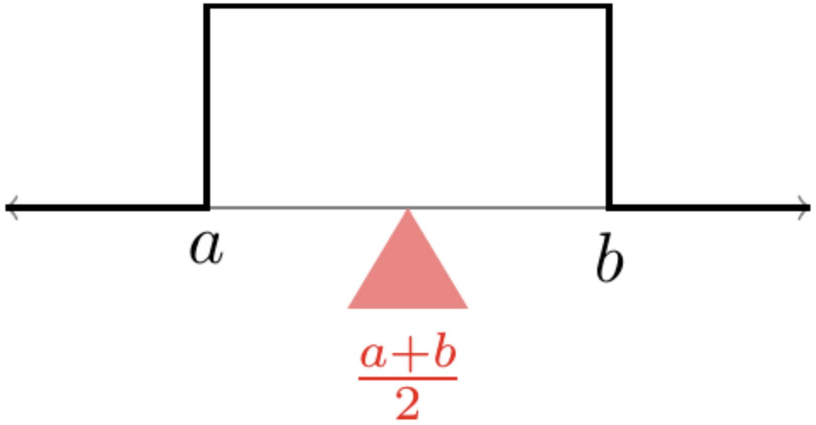
Example. (Expected Value of the Uniform Distribution) Let X be a Uniform(a, b) random variable. What is $\mathbb{E}[X]$?

Random Variables

Example. (Expected Value of the Uniform Distribution) Let X be a Uniform(a, b) random variable. What is $\mathbb{E}[X]$?

$$f(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases},$$

$$\begin{aligned} E[X] &= \int_a^b x \cdot \frac{1}{b-a}\, dx \\ &= \frac{1}{b-a} \left.\frac{x^2}{2}\right]_a^b \\ &= \frac{1}{b-a}(b^2 - a^2)\frac{1}{2} \\ &= \frac{a+b}{2}. \end{aligned}$$



Random Variables

- Two key facts:

1. **Linearity of expectation.** For any r.v. X and Y, and constants a, b:

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$$

In particular, **no independence** is required.

2. **Expectation of a function.** If g is any real function, then

$$\mathbb{E}[g(X)] = \sum_x g(x) \cdot \Pr[X = x]$$ (or the analogous integral in the continuous case)

Union Bound

- For events $E_1, \cdots, E_k$, $\Pr[E_1 \cup \cdots \cup E_k] \le \sum_{i \in [k]} \Pr[E_i]$

Example. (Erdös-Rényi random graph) Let B_n be the event that a graph randomly generated according to $G(n,p)$ model has at least one isolated node. Show that

$$\Pr[B_n] \le n(1-p)^{n-1},$$

And conclude that for any $\epsilon > 0$, if $p = p_n = (1+\epsilon)\frac{\ln n}{n}$, then

$$\lim_{n \to \infty} \Pr[B_n] = 0.$$

Union Bound Example

Markov and Chebyshev Inequalities

- **(Markov's inequality)** For a random variable $X > 0$, and value $a > 0$,
$$\Pr[X \geq a] \leq \frac{\mathbb{E}[X]}{a}$$

Only required non-negativity of X.

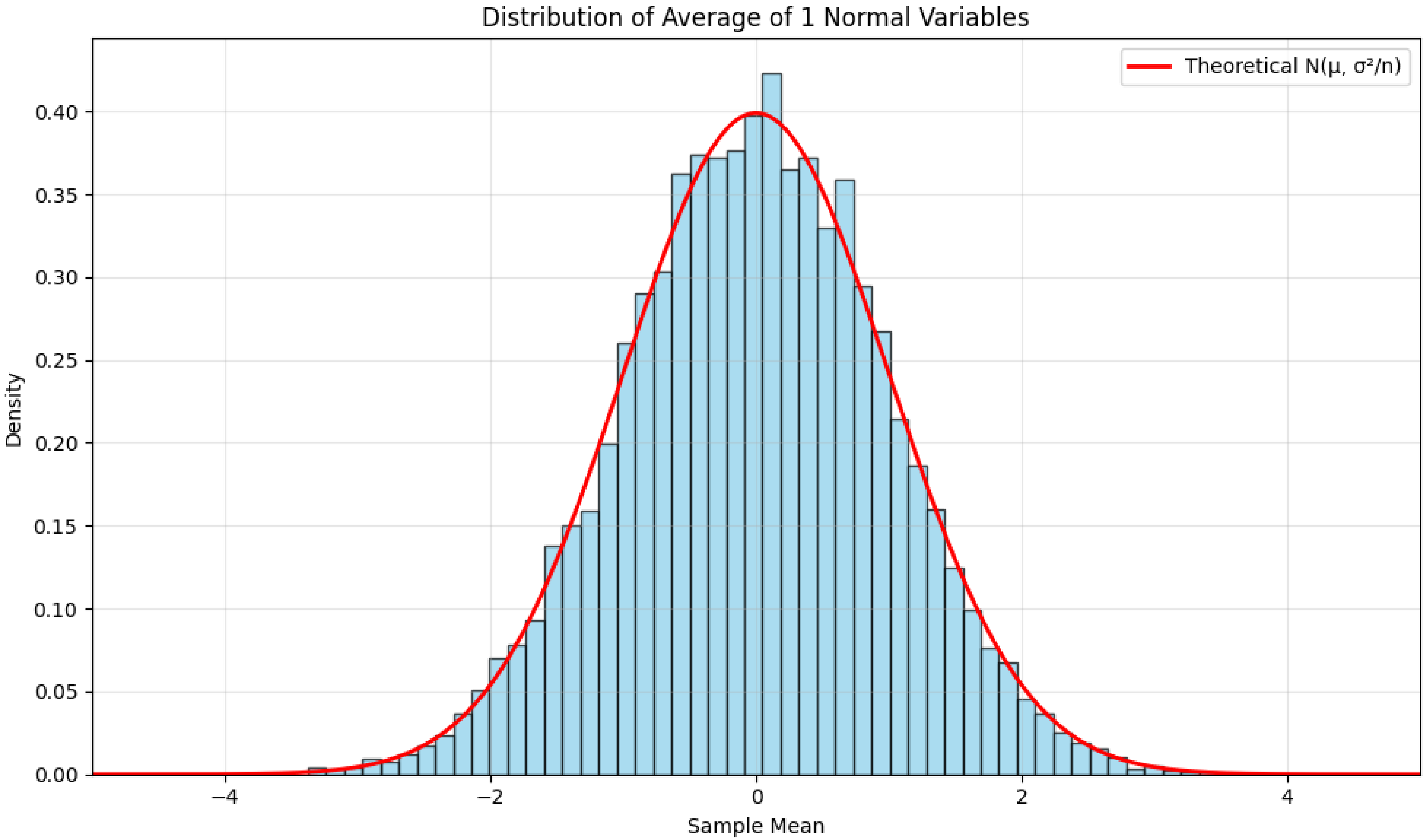
- **(Chebyshev's inequality)** For a random variable X, and a value $t > 0$,
$$\Pr[|X - \mathbb{E}[X]| \geq t] \leq \frac{\mathrm{Var}[X]}{t^2}$$

Proof. Apply Markov's on $Y = (X - \mathbb{E}[X])^2$.

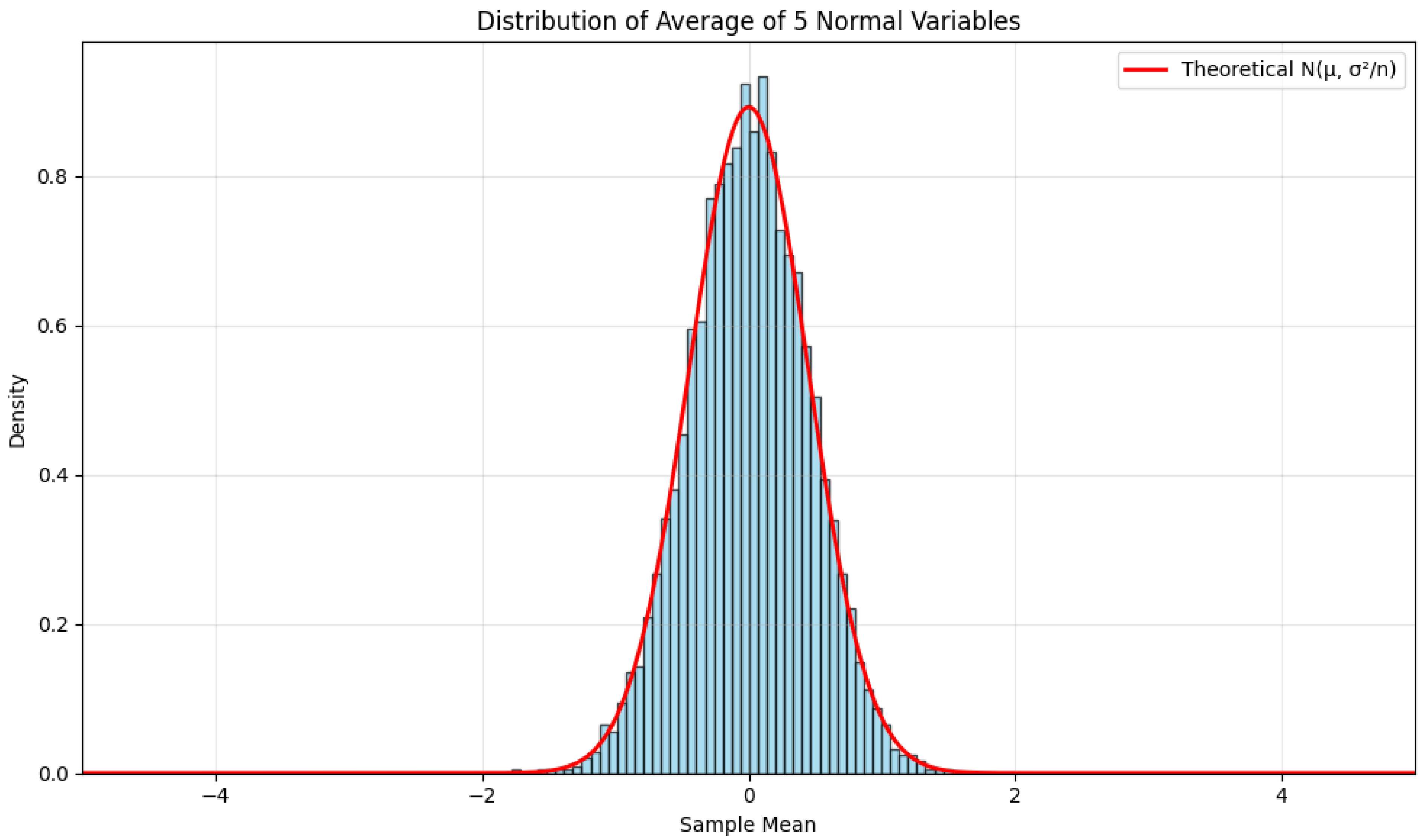
Chernoff and Hoeffding Bounds

- Concentration: As you add up many independent random variables, their average concentrated around the expected value more and more; typical deviations shrink like $1/\sqrt{n}$, so the distribution piles up tightly around the mean.

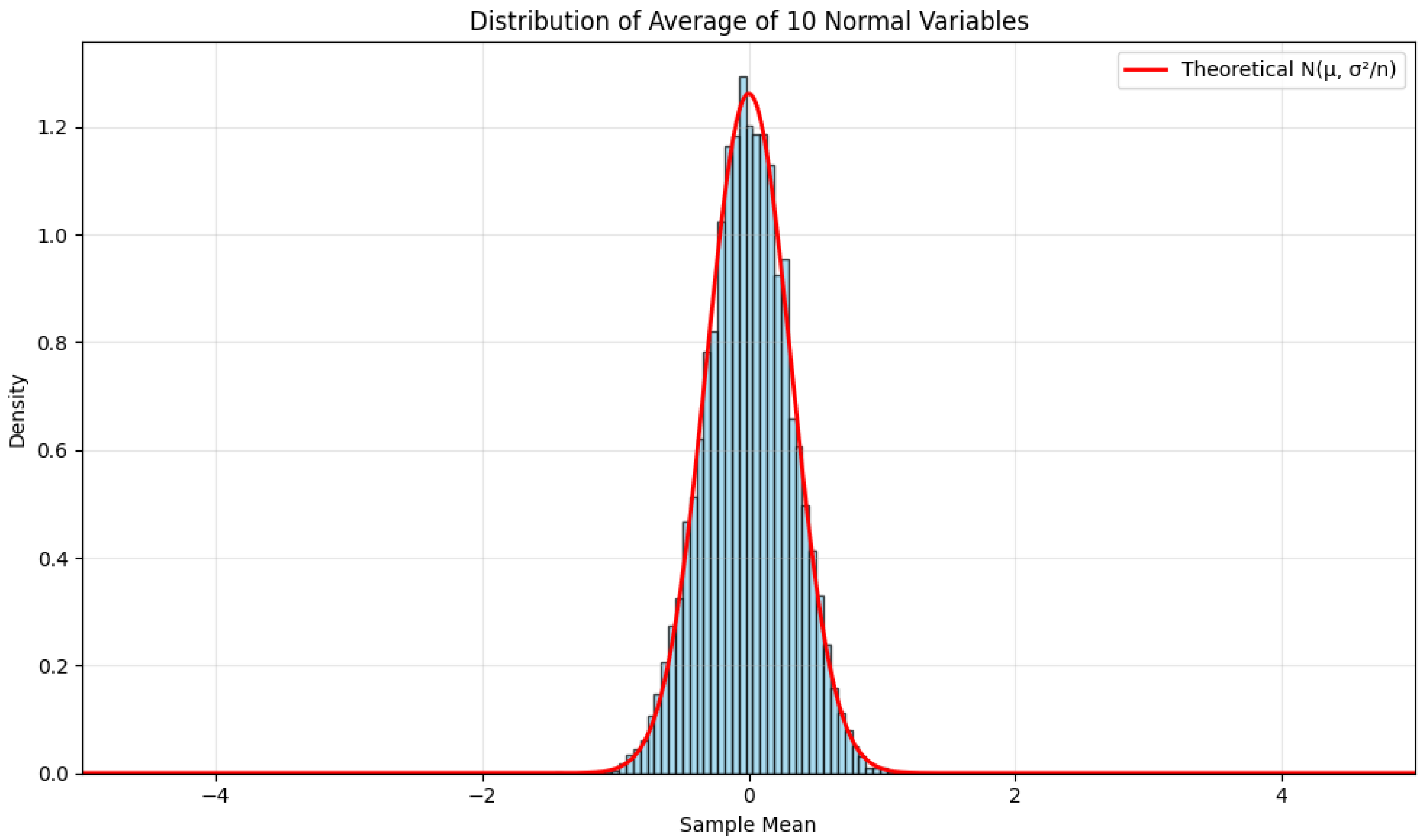
Sample mean of 1 normal distribution



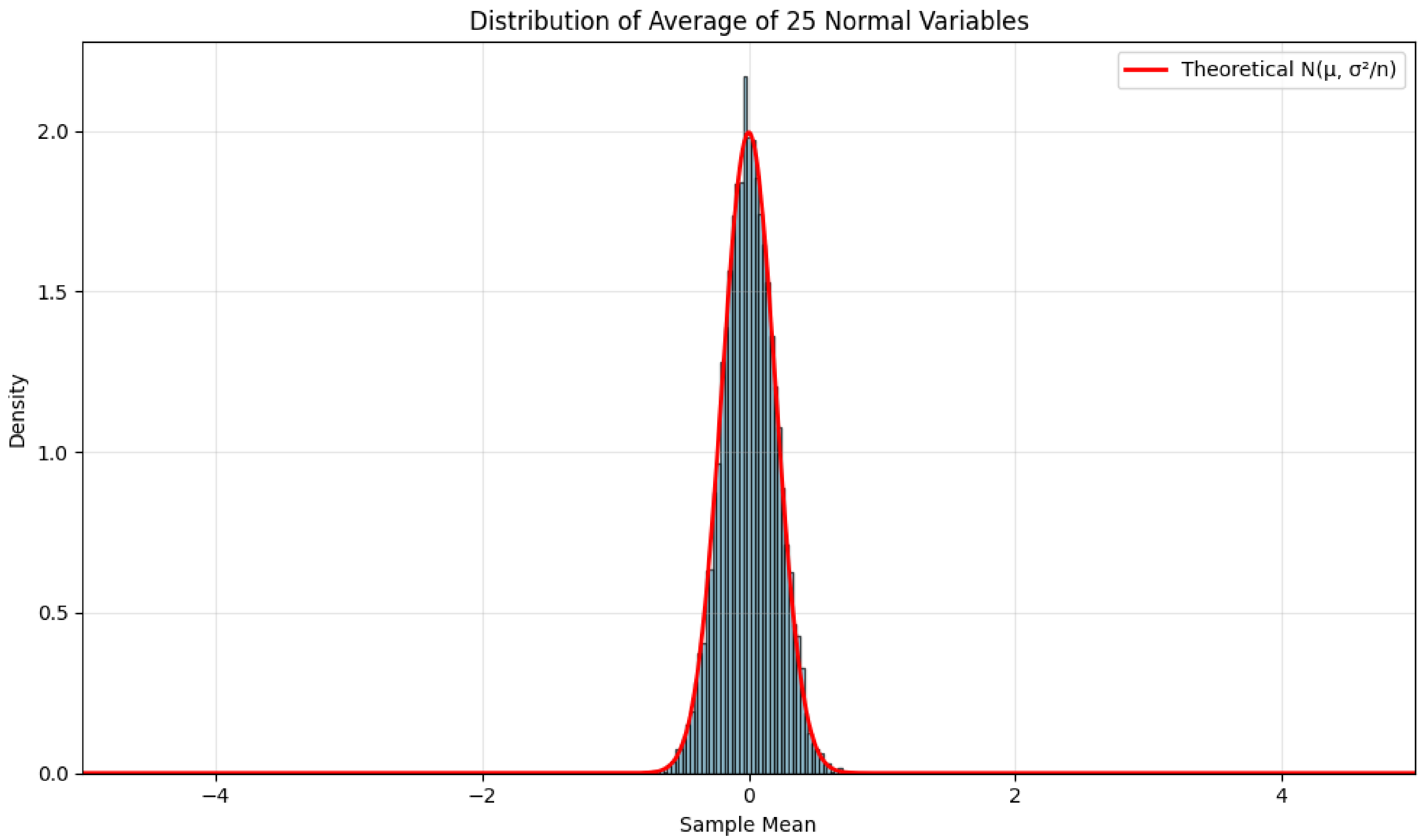
Sample mean of 5 normal distribution



Sample mean of 10 normal distribution



Sample mean of 25 normal distribution



Chernoff and Hoeffding Bounds

- **(Chernoff Bound)** Let $X = \sum_i^n X_i$ where the $X_i \in [0,1]$ are **independent**, and set $\mu = \mathbb{E}[X]$. Then for $0 < \varepsilon \leq 1$,
  - $\Pr[X \geq (1+\varepsilon)\mu] \leq \exp(-\frac{\varepsilon^2 \mu}{3})$, and
  - $\Pr[X \leq (1-\varepsilon)\mu] \leq \exp(-\frac{\varepsilon^2 \mu}{2})$.
- **(Hoeffding's Inequality)** Let $X = \sum_i^n X_i$ where the $X_i \in [a_i, b_i]$ are **independent**, and set $\mu = \mathbb{E}[X]$. Then for $t > 0$,
  - $\Pr[|X - \mu| \geq t] \leq 2\exp(-\frac{2t^2}{\sum_i^n (b_i - a_i)^2})$.

Linear Algebra Refresher

Vector Norms

- $x = (x_1, \cdots, x_d) \in \mathbb{R}^d$

$$\|x\|_0, \|x\|_1, \|x\|_2, \|x\|_\infty$$

In general, $\|x\|_p = \left(|x_1^p| + \cdots + |x_d^p|\right)^{1/p}$

Exercise (Norm Inequalities). For every $x \in \mathbb{R}^d$,

$$\|x\|_\infty \le \|x\|_2 \le \|x\|_1 \le \sqrt{d} \cdot \|x\|_2$$

Exercise. For any $x \in \mathbb{R}^d$,

$$\|x\|_{\log_2 d} \le 2\|x\|_\infty.$$

Proof. By Cauchy–Schwarz inequality.

Dot Product and Angles. For $x, y \in \mathbb{R}^d$, $x \cdot y = \langle x, y \rangle = x^\top y$

$$|x^\top y| \le \|x\|_2 \cdot \|y\|_2$$

Eigenvalue, Eigenvector, and PSD

- **Eigenvector.** A non-zero vector $v \in \mathbb{R}^d$ is an **eigenvector** of matrix $A \in \mathbb{R}^{d \times d}$ with **eigenvalue** $\lambda \in \mathbb{R}$, if $Av = \lambda v$. (directions that are invariant under A)
  - can be negative or complex
- **Positive Semidefinite (PSD) Matrices.** A symmetric matrix A is positive semidefinite if

$$x^\top A x \geq 0 \qquad \text{for all } x \in \mathbb{R}^d$$

Exercise. Let $A \in \mathbb{R}^{d \times d}$ be symmetric with eigenvalues $\lambda_1, \cdots, \lambda_d$. Then,

$$A \text{ is } \textbf{PSD} \Leftrightarrow \lambda_i \geq 0 \text{ for every } i.$$

Singular Values, SVD

- **Singular value.** σ_i is a singular value of $A \in \mathbb{R}^{m\times d}$, if λ_i is an eigenvalue of $A^\top A$, and $\sigma_i = \sqrt{\lambda_i}$.
  - always real and non-negative

Exercise. If λ is an eigenvalue of $A^\top A$, then $\lambda \geq 0$.

Proof. Consider the corresponding eigenvector x to λ, and argue with $\|A^\top x\|_2$

- **Singular value decomposition (SVD).** For every $A \in \mathbb{R}^{m\times d}$, there exist orthogonal matrices $U \in \mathbb{R}^{m\times m}, V \in \mathbb{R}^{d\times d}$ such that

$$A = U\,\Sigma\,V^\top,$$

where $\Sigma = diag(\sigma_1, \cdots, \sigma_r, 0, \cdots, 0)$, $\sigma_1 \geq \cdots, \geq \sigma_r > 0$. Recall r is the $rank(A)$.

Matrix Norm

- **Frobenius norm.** For a matrix $A \in \mathbb{R}^{m \times d}$,

$$\|A\|_F = \sqrt{\sum_{i \in [m]} \sum_{j \in [d]} \left|A_{i,j}\right|^2}$$

Alternatively, $\|A\|_F = \|vec(A)\|_2 = \sqrt{trace(A^* A)} = \sqrt{\sum_i \sigma_i^2}$

- **Spectral nom.** $\|A\|_2 = \sigma_1$

Exercise. For any matrix A, $\|A\|_2 \leq \|A\|_F \leq \sqrt{\mathrm{rank}(A)} \cdot \|A\|_2$