Lecture 1: Background on Probability and Linear Algebra

08-26-2025 Lecturer: Prof. Ali Vakilian | Scribe: Ali Vakilian

1 Motivation

The topics covered in this course, and more broadly, most developments in algorithms for massive data, rely on two key concepts:

1. **Probability.** Concentration inequalities tell us when a random variable or sample behaves like its expectation.
2. **Linear Algebra.** Matrix factorizations and spectral bounds sit at the heart of dimensionality reduction and numerical subroutines.

This lecture captures the core facts you will invoke throughout the course.

2 Probability Refresher

We assume familiarity with basic probability spaces and random variables (r.v.s). Unless stated otherwise, all r.v.s are defined on the same probability space Ω.

2.1 Independence, Conditional Probability, Random Variables

Independence. Two events A and B are *independent* if

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

A collection $(E_1, \ldots, E_k)$ is *mutually independent* if every sub-collection obeys this equality. *Pairwise independence*, which requires independence for every *pair* of events, is *strictly weaker* than full (mutual) independence. In upcoming lectures we will often work with pairwise independence or with the slightly broader notion of *c-wise independence* (the special case $c = 2$ coincides with pairwise independence).

Conditional Probability. For $\Pr[B] > 0$ the probability of A *given* B is

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

Key consequences include the *Law of Total Probability* $\Pr[A] = \sum_i \Pr[A \mid B_i] \Pr[B_i]$ for a partition $\{B_i\}$, and *Bayes' Rule*

$$\Pr[B_j \mid A] = \frac{\Pr[A \mid B_j] \Pr[B_j]}{\sum_i \Pr[A \mid B_i] \Pr[B_i]}.$$

Random Variables (an informal view). A random variable[1] is simply a rule that assigns a number to each outcome $\omega \in \Omega$. Think "roll a die, record the number of pips" or "run Quicksort algorithm, record the running time". We denote the set of possible values by $\mathrm{range}(X)$ and write

$$\Pr[X = x] \quad \text{for } x \in \mathrm{range}(X).$$

The most useful derived quantities are the *expectation* $\mathbb{E}[X] = \sum_x x \Pr[X = x]$ (or $\int x \, d\Pr$ in the continuous case) and the *variance* $\mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}X)^2]$.

A handy special case is an *indicator variable* $\mathbf{1}_E$ that equals 1 when event E occurs and 0 otherwise; then $\mathbb{E}[\mathbf{1}_E] = \Pr[E]$.

Independence of random variables. Random variables $X_1, \ldots, X_k$ are *independent* if, for every choice of real numbers $a_1, \ldots, a_k$,

$$\Pr\big[X_1 = a_1, X_2 = a_2, \ldots, X_k = a_k\big] = \prod_{i=1}^{k} \Pr\big[X_i = a_i\big].$$

Intuitively, knowing the outcome of any subset of the variables tells us nothing about the rest. (For two events this reduces to the familiar rule $\Pr[XY] = \Pr[X]\Pr[Y]$ when X and Y are indicator variables.)

Expectation. The *expectation* or *mean* of a random variable X is its long-run average value:

$$\mathbb{E}[X] \;=\; \sum_x x \Pr[X = x] \quad \text{(discrete)} \qquad \text{or} \qquad \mathbb{E}[X] \;=\; \int_{-\infty}^{\infty} x \, f_X(x) \, dx \quad \text{(continuous)}.$$

Two key facts we will use constantly:

1. **Linearity of expectation.** For any random variables X and Y and constants a, b,

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

   in particular, no independence required.

2. **Expectation of a function.** If g is any real function then $\mathbb{E}[g(X)] = \sum_x g(x) \Pr[X = x]$ (or the analogous integral in the continuous case).

These properties let us break complicated expressions into simple, computable pieces; we will see many examples of this in later lectures.

2.2 Union Bound

Theorem 2.1 (Union Bound). *For events $E_1, \ldots, E_k$,* $\Pr\left[\bigcup_{i=1}^{k} E_i\right] \le \sum_{i=1}^{k} \Pr[E_i]$.

Proof. By induction on k. For $k = 1$ equality holds trivially. For $k = 2$,

$$\Pr\left[E_1 \cup E_2\right] = \Pr\left[E_1\right] + \Pr\left[E_2\right] - \Pr\left[E_1 \cap E_2\right] \le \Pr\left[E_1\right] + \Pr\left[E_2\right],$$

[1] Formally, a random variable is a *measurable* function $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B})$, where $\mathcal{B}$ is the Borel σ-algebra on $\mathbb{R}$.

following exclusion-inclusion identity for two sets and the fact that probabilities are non-negative. Assume the claim for $k-1$ events. Write

$$\begin{aligned}\Pr\Big[\bigcup_{i=1}^{k} E_i\Big] &= \Pr\Big[\Big(\bigcup_{i=1}^{k-1} E_i\Big) \cup E_k\Big] \\ &\le \Pr\Big[\bigcup_{i=1}^{k-1} E_i\Big] + \Pr[E_k] && \triangleright \text{ set } E' = \bigcup_{i=1}^{k-1} E_i \text{ and apply union bound on } E' \text{ and } E_k \\ &\le \sum_{i=1}^{k} \Pr[E_i] && \triangleright \text{ by induction hypothesis on } E'\end{aligned}$$

□

2.3 Markov and Chebyshev

Theorem 2.2 (Markov's inequality)**.** *Let $X \ge 0$ be any random variable and $a > 0$. Then $\Pr[X \ge a] \le \frac{\mathbb{E}[X]}{a}$.*

Proof. Observe that $\mathbb{E}[X] \ge \mathbb{E}[\mathbf{1}_{\{X \ge a\}}\, a] = a \Pr[X \ge a]$. □

Theorem 2.3 (Chebyshev inequality)**.** *Let X be any random variable with finite variance. For $t > 0$, $\Pr\big[|X - \mathbb{E}X| \ge t\big] \le \frac{\mathrm{Var}[X]}{t^2}$.*

Proof. Apply Markov 2.2 to $Y = (X - \mathbb{E}X)^2$: $\Pr[Y \ge t^2] \le \frac{\mathbb{E}[Y]}{t^2}$ where $\mathbb{E}[Y] = \mathrm{Var}[X]$. □

2.4 Chernoff and Hoeffding Bounds

Before proving the main inequalities we recall a standard tool.

Lemma 2.4 (Moment Generating Function (MGF) Trick)**.** *For any r.v. X and $\lambda > 0$, $\Pr[X \ge a] = \Pr[e^{\lambda X} \ge e^{\lambda a}] \le e^{-\lambda a}\, \mathbb{E}[e^{\lambda X}]$ by Markov.*

Theorem 2.5 (Chernoff Bound, multiplicative)**.** *Let $X = \sum_{i=1}^{n} X_i$ where the $X_i \in [0,1]$ are independent and set $\mu = \mathbb{E}[X]$. Then for $0 < \varepsilon \le 1$,*

$$\Pr[X \ge (1+\varepsilon)\mu] \le \exp\big(-\tfrac{\varepsilon^2\mu}{3}\big), \qquad \Pr[X \le (1-\varepsilon)\mu] \le \exp\big(-\tfrac{\varepsilon^2\mu}{2}\big).$$

Proof. We prove the upper tail; the lower tail is analogous. Let $\lambda > 0$ (to be chosen) and apply Lemma 2.4:

$$\Pr[X \ge (1+\varepsilon)\mu] \le e^{-\lambda(1+\varepsilon)\mu}\, \mathbb{E}\big[e^{\lambda X}\big] = e^{-\lambda(1+\varepsilon)\mu} \prod_{i=1}^{n} \mathbb{E}\big[e^{\lambda X_i}\big] \quad \text{(independence)}.$$

Because $0 \le X_i \le 1$, $\mathbb{E}[e^{\lambda X_i}] \le 1 + (e^{\lambda} - 1)\mathbb{E}[X_i]$ (by convexity of e^x). Thus

$$\mathbb{E}[e^{\lambda X}] \le \exp\big((e^{\lambda} - 1)\mu\big).$$

Combine and set $\lambda = \ln(1+\varepsilon)$ to minimize the bound; algebra yields the stated exponent $-\varepsilon^2\mu/3$. □

Theorem 2.6 (Hoeffding's Inequality)**.** *Let $X_i \in [a_i, b_i]$ be independent with $S_n = \sum_{i=1}^{n} X_i$ and $\mathbb{E}[S_n] = \mu$. For any $t > 0$,*

$$\Pr\big[|S_n - \mu| \ge t\big] \le 2\exp\Big(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\Big).$$

Sketch. Apply Lemma 2.4 to $\pm(S_n - \mu)$ and use the fact that the MGF of each centered X_i is bounded by $\exp\big(\frac{\lambda^2(b_i - a_i)^2}{8}\big)$ (Hoeffding's lemma). Optimizing over λ produces the claimed bound. □

Takeaway 2.1

Markov & Chebyshev provide *polynomial* tails under minimal assumptions; Chernoff & Hoeffding give *exponential* tails when independence (and boundedness) hold.

3 Linear Algebra Refresher

3.1 Vector Norms

For a vector $x \in \mathbb{R}^d$ we use three standard norms

$$\|x\|_2 = \Big(\sum_{i=1}^{d} x_i^2\Big)^{1/2}, \qquad \|x\|_1 = \sum_{i=1}^{d} |x_i|, \qquad \|x\|_\infty = \max_{1\le i\le d} |x_i|.$$

Lemma 3.1 (Norm inequalities). *For every $x \in \mathbb{R}^d$,*

$$\|x\|_\infty \;\le\; \|x\|_2 \;\le\; \|x\|_1 \;\le\; \sqrt{d}\,\|x\|_2.$$

Proof. First note $|x_i| \le \sqrt{\sum_j x_j^2} = \|x\|_2$, which gives $\|x\|_\infty \le \|x\|_2$. For the last inequality apply Cauchy–Schwarz:

$$\|x\|_1 = \sum_i |x_i| \cdot 1 \;\le\; \|x\|_2\,\|\mathbf{1}\|_2 \;=\; \sqrt{d}\,\|x\|_2.$$

□

3.2 Dot Product and Angles

For $x, y \in \mathbb{R}^d$ the dot product $x^\top y$ measures the cosine of the angle between them. The Cauchy–Schwarz inequality states $|x^\top y| \le \|x\|_2\,\|y\|_2$, with equality if and only if x and y are colinear.

3.3 Singular Values and the SVD

Informal picture. Any linear map $A : \mathbb{R}^n \to \mathbb{R}^m$ can be viewed as *rotate* $\to$ *stretch* $\to$ *rotate*. The stretching factors are the *singular values* of A.

Singular Value Decomposition (formal). For every $A \in \mathbb{R}^{m\times n}$ there exist orthogonal matrices $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ such that

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

where $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_r, 0, \ldots, 0)$, $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_r > 0$. The non–zero σ_i are the *singular values* of A and r is called the *rank*.

Rank. Because each non–zero singular value contributes an independent column direction, we have

$$\mathrm{rank}(A) \;=\; r \;\le\; \min\{m, n\}.$$

3.4 Eigenvalues, Eigenvectors and Positive Semidefinite Matrices

Eigenvalues and eigenvectors (informal view). For most vectors a square matrix changes both length and direction. An *eigenvector* keeps its direction and is scaled by a factor called the *eigenvalue*.

Formally, a non-zero vector $v \in \mathbb{R}^n$ is an eigenvector of $A \in \mathbb{R}^{n\times n}$ with eigenvalue $\lambda \in \mathbb{R}$ if $Av = \lambda v$.

Positive semidefinite (PSD) matrices. A symmetric matrix A is *positive semidefinite* if

$$x^\top Ax \;\geq\; 0 \quad \text{for all } x \in \mathbb{R}^n.$$

Proposition 3.2 (PSD characterised by eigenvalues)**.** *Let $A \in \mathbb{R}^{n\times n}$ be symmetric with eigenvalues $\lambda_1, \ldots, \lambda_n$. Then*

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

Proof. *($\Rightarrow$)* If A is PSD and (λ, v) is any eigenpair with $\|v\|_2 = 1$ then $\lambda = v^\top Av \geq 0$.
($\Leftarrow$) If all eigenvalues are non–negative, write $A = \sum_{i=1}^n \lambda_i\, u_i u_i^\top$ with an orthonormal eigenbasis $\{u_i\}$. For $x = \sum_i \alpha_i u_i$ we have

$$x^\top Ax \;=\; \sum_{i=1}^n \lambda_i \alpha_i^2 \;\geq\; 0.$$

Hence A is PSD. □

3.5 Spectral and Frobenius Norms

Definitions. For $A \in \mathbb{R}^{m\times n}$ with singular values $\sigma_1 \geq \cdots \geq \sigma_r > 0$ set

$$\|A\|_2 \;=\; \sigma_1, \qquad \|A\|_F^2 \;=\; \sum_{i=1}^r \sigma_i^2.$$

Lemma 3.3. *For any matrix A,*

$$\|A\|_2 \;\leq\; \|A\|_F \;\leq\; \sqrt{\operatorname{rank}(A)}\, \|A\|_2.$$

Proof. Treat the vector of singular values $(\sigma_1, \ldots, \sigma_r)$ and apply Lemma 3.1 with $d = r$:

$$\|\sigma\|_\infty = \sigma_1 = \|A\|_2, \quad \|\sigma\|_2 = \|A\|_F, \quad \|\sigma\|_1 = \sum_{i=1}^r \sigma_i \leq \sqrt{r}\, \|\sigma\|_2.$$

The middle inequality yields $\|A\|_F \leq \sqrt{r}\, \|A\|_2$ and $r = \operatorname{rank}(A)$. □

Takeaway 3.1

Spectral norm controls worst–case distortion; Frobenius norm captures average distortion. Many sketching results bound both simultaneously.

4 Further Reading

For further readings on related topics, refer to

- Chandra Chekuri, *Background on Probability*, lecture notes.
- Krzysztof Onak, *Useful Probabilistic Inequalities*, lecture notes.
- Martin Wainwright, *High-Dimensional Statistics*, Chapter 2 (concentration).
- Joel Tropp, *Introduction to Matrix Concentration Inequalities*.