47-837: Statistical Foundations of OR Lecture 1

Basic Concentration Inequalities

Lecturer: Karan Singh

This course is about both negotiating uncertainty and using randomness in decision making. Most modern-day decision making systems (think, for example, ride-share) operate while dealing with multiple sources of randomness: Can we describe them well using deterministic systems? Do the particulars of uncertainties they are subject to actually matter? What are the aggregate statistical properties of such algorithms? The goal here is to provide just enough foundational knowledge so that the students can dive deeply into these data-driven optimization and algorithm design.

In general, we will weave through a potpourri of topics. We will start with *basic concentration inequalities* in the first week and discuss *Bayesian statistics and causal inference* the week after. After switching to a couple of weeks on *statistical learning* and *sequential prediction*, we will survey *statistical fairness* in the decision-making context. The final topic will be the *information-theoretic limits* of performance for data-driven algorithms.

Contents

1. Convexity

Definition 1 A set $S \subseteq$ vector space V is **convex** if for all $x, y \in S$ and $\lambda \in [0,1]$, $\lambda x + (1-\lambda)y \in S$.

Note that $S_{x,y} = \{\lambda x + (1-\lambda)y : \lambda \in [0,1]\}$ is a line segment that connects x and y. In other words, a convex set contains all line segments whose endpoints lie inside the set.

Definition 2 A function $f : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X}$ is a convex set, is said to be **convex** if $\forall x, y \in \mathcal{X}$, $\forall \lambda \in [0,1]$, $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$.

In other words, for a convex function, the function value at any interpolation is (weakly) less than the interpolated function values.

Definition 3 The convex hull of a set $X \subseteq$ some vector space V is the smallest convex set that contains X.

Here, the notion of smallness is with respect to inclusion, that is, A is smaller than B if $A \subset B$. Therefore, incomparable sets exist; nevertheless, the smallest is well defined. If $X = \{x_1, x_2, ...x_m\}$ is finite, the convex hull becomes $\left\{\sum_{i=1}^m \alpha_i x_i : \sum_{i=1}^m \alpha_i = 1, \alpha_i \geq 0, \forall i \in [m]\right\}$.

2. Probability

A probability measure is described by $(\Omega, \mathcal{F}, \Pr)$, where Ω is the outcome space, $\mathcal{F} \subseteq 2^\Omega$ is the set of measurable subsets of Ω, and $P : \mathcal{F} \to [0,1]$ assigns probability to the sets in $\mathcal{F}$. The three rules for P are: it must be non-negative, it must assign measure one to Ω, and it must be additive on any collection of countably many disjoint sets in $\mathcal{F}$. In this course, it is okay to think of $\mathcal{F} = 2^\Omega$; this is true, for example, for discrete outcomes. We will avoid measurability issues by assuming that you know what expectation is.

Lemma 4 For any two real-valued random variables X, Y, $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.

Recall that a collection of random variables $\{X_i\}_{i \in \mathcal{J}}$ is independent if $\Pr(X_i \in S_i)_{i \in \mathcal{J}} = \Pi_{i \in \mathcal{J}} \Pr(X_i \in S_i)$. A weaker condition is pairwise independence, which only requires that for all pairs $X_1, X_2 \in \{X_i\}_{i \in \mathcal{J}}$, we have $\Pr(X_1 \in S_1, X_2 \in S_2) = \Pr(X_2 \in S_1) \Pr(X_2 \in S_2)$.

Recall that the variance of X is defined as $\mathbb{V}[X] = \mathbb{E}\big[(X - \mathbb{E}[X])^2\big]$. Using this, we observe that for independent random variables X, Y, we have

$$\mathbb{V}[X+Y] = \mathbb{E}\big[(X - \mathbb{E}X + Y - \mathbb{E}Y)^2\big] = \mathbb{V}[X] + \mathbb{V}[Y] + 2\underbrace{\mathbb{E}[X - \mathbb{E}X]}_{=0}\mathbb{E}[Y - \mathbb{E}Y] = \mathbb{V}[X] + \mathbb{V}[Y].$$

In fact, the additivity of variance, over any number of random variables, only requires pairwise independence.

Exercise. Prove that $\mathbb{V}[X] = \mathbb{E}\big[X^2\big] - (\mathbb{E}[X])^2 = \min_{a \in \mathbb{R}} \mathbb{E}\big[(X-a)^2\big] = \frac{1}{2}\mathbb{E}\Big[(X - X')^2\Big]$, where X' is an independent copy of X.

3. Approximate Caratheodory's Theorem

In convex geometry, Caratheodory's theorem states that, given a set of any size in $\mathbb{R}^d$, any point in its convex hull can be written as a convex combination of $d+1$ points. Notice that there is no dependence on the number of points in the original set. In two dimensions, you can convince

yourself that this is true by noting that any convex polygon can be perfectly decomposed into non-overlapping triangles by connecting the vertices in the correct way; thus, any point in a convex polygon lies in a triangle supported on the vertices of the polygon.

Such theorems are results about minimal representations: how many elements are needed to represent any point from a rich set? We will prove another such result that uses far fewer points, in fact, independent of the dimension, given a small slack.

Theorem 5 (Approximate Caratheodory's) Let X be a set in $\mathbb{R}^d$ contained in the unit ball $\mathbb{B}_2 = \{x : \|x\|_2 \leq 1\}$. For any $k \geq 0$ and y in the convex hull of X, there exist k points $z_1, ...z_k \in X$ such that

$$\left\| y - \frac{1}{k} \sum_{i=1}^{k} z_k \right\|_2 \leq \frac{1}{\sqrt{k}}.$$

Thus, if we are willing to tolerate ε slack in how well x is approximated, $\frac{1}{\varepsilon^2}$ points suffice. In fact, note that our convex combination is also *special*, being a simple average.

Proof. Since y is the convex hull, we know that $y = \sum_{i=1}^{m} \alpha_i x_i$ for some $x_1, x_2, ...x_m \in X$, where $\alpha_i \geq 0$ sum to one. Consider a random variable Z_1 that picks x_i with probability α_i for all $i \in [m]$. Notice that by definition $\mathbb{E}[Z_1] = y$. Consider $Z_2, ...Z_k$ additional independent copies of Z_1. Now

$$\mathbb{E}\left\| \frac{1}{k} \sum_{i=1}^{k} Z_i - y \right\|_2^2 = \frac{1}{k^2} \sum_{i=1}^{k} \mathbb{E}\|Z_i - y\|_2^2 = \frac{1}{k}\left(\mathbb{E}\|Z_1\|_2^2 - \|y\|_2^2\right) \leq \frac{1}{k}.$$

Thus, there must exist some realization of $Z_1, ...Z_k$ such that $\left\| \frac{1}{k} \sum_{i=1}^{k} Z_i - y \right\|_2^2 \leq \frac{1}{k}$. □

3.1. Application: Choice Models

Human behavior is tricky and often inconsistent. Choice models study statistical models of the same. For example, in the Plackett-Luce model, with a universe of n products, it is assumed that a tester given items i and j will prefer i over j with probability $w_i/(w_i + w_j)$. Given a lot of such empirical observations, the task is to recover the underlying quality w_i of the items.

Here, we consider a (nonparametric) model for ranking. Imagine a universe of n products, where it is assumed that the testers are identical and randomly pick a ranking of all products, i.e., $\Pr(\pi) = \sum_{i=1}^{n!} \alpha_i \mathbf{1}_{\pi=\pi_i}$, where we π_i's are all $n!$ permutations. But instead of receiving direct observations of many such stochastic rankings, we receive a summary statistic: a matrix $Z \in [0,1]^{n \times n}$ where Z_{ij} represents the fraction of times the product i earned the rank j. Farias, Jagathabula and Shah considered the following question: can we produce a choice model that explains the observed matrix? Their main result is that a linearly sparse model is sufficient for this purpose. We will recover this as an implication.

We begin with an application of Caratheodory's theorem. A nonnegative matrix is said to be *doubly stochastic* if row and column sums are one. Z is such a matrix. Each permutation π can

be written as a permutation matrix $\Pi \in \{0,1\}^{n\times n}$ where Π_{ij} if and only if the element i occurs in position j in π. For example $(2,4,3,1)$ can be encoded as

$$\begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

Notice that such matrices are also doubly stochastic, since there is exactly one 1 in each row and column. A very fundamental result is as follows:

Theorem 6 (Birkhoff-von-Neumann) The convex hull of all permutation matrices is precisely the set of doubly stochastic matrices.

By Caratheodory's, any such Z can be written as the convex combination of $(n-1)^2+1$ permutation matrices. Thus, any Z can be explained away with a distribution supported on just $(n-1)^2+1$ matrices, an exponentially small fraction of the original $n!$ parameters. Perhaps a bound of n^2+1 is easier. The precise number comes about by noting that the set of doubly stochastic matrices is $(n-1)^2$-dimensional, given the $(n-1)\times(n-1)$ principal sub-matrix, one can always reconstruct the last row and column due to the row and column sum constraints.

Using Approximate Caratheodory's, we can guarantee that there exists a convex combination Z' of n/ε^2 permutation matrices such that $\|Z-Z'\|_F = \sqrt{\sum_{i,j\in[n]^2}\left(Z_{ij}-Z'_{ij}\right)^2} \le \varepsilon$. The additional n factor arises because a permutation matrix has Frobenius norm $\sqrt{n}$, and both sides of the inequality in Approximate Caratheodory's grow linearly with the scale.

4. Moment Generation Function

Given a random variable X, $\Psi_{X(t)} = \mathbb{E}[e^{tX}]$ is defined to be its moment generating function. The MGFs, upon their existence, encode all the moments and vise versa.

Proposition 7 $\frac{d^k}{dt^k}\Psi_{X(t)}|_{t=0} = \mathbb{E}[X^k e^{tX}]|_{t=0} = \mathbb{E}[X^k]$.

Bernoulli Distribution. $X \sim \mathrm{Be}(p)$ is $\{0,1\}$-valued with $\Pr(X=1)=p$.

$$\Psi_{\mathrm{Be}(p)}(t) = (1-p) + pe^t.$$

Rademacher Distribution. $X \sim \{\pm 1\}$ is $\{\pm1\}$-valued with $\Pr(X=1)=\frac{1}{2}$.

$$\Psi_{\{\pm1\}}(t) = \frac{1}{2}(e^t+e^{-t}) = 1+\frac{t^2}{2!}+\frac{t^4}{4!}+\ldots \le 1+\sum_{k\ge1}\frac{t^{2k}}{2^k k!} = e^{t^2/2}.$$

Gaussian Distribution. $Z \,\mathrm{sim}\, \mathcal{N}(\mu,\sigma^2)$ if $Z=\sigma X+\mu$ and $X\sim\mathcal{N}(0,1)$ where $f_{\mathcal{N}(0,1)}(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$. Recall that $\Gamma(k)=\int_0^\infty t^{k-1}e^{-t}dt$ and $\Gamma(k+1)\le k^k$. Now, a few basic facts, where we assume $t,k\ge1$.

$$\Psi_X(t) = \frac{1}{\sqrt{2\pi}} \int e^{tx - x^2/2} dx = \frac{e^{\sigma^2 t^2/2}}{\sqrt{2\pi}} \int e^{-(x-t)^2/2} dx = e^{t^2/2}$$

$$\Psi_{\frac{3X^2}{8}}(1) = \frac{1}{\sqrt{2\pi}} \int e^{3x^2/8 - x^2/2} dx = \frac{1}{\sqrt{2\pi}} \int e^{-x^2/8} dx = 2$$

$$\Pr(X \geq t) = \frac{1}{\sqrt{2\pi}} \int_t^\infty e^{-x^2/2} dx \leq \frac{1}{\sqrt{2\pi}} \int_t^\infty \frac{x}{t} e^{-x^2/2} dx = \frac{1}{\sqrt{2\pi} t} e^{-t^2/2} \qquad \text{(Mill's inequality)}$$

$$\mathbb{E}[|X|^k] = \frac{1}{\sqrt{2\pi}} \int x^k e^{-x^2/2} dx \underbrace{=}_{z = x^2/2} \frac{2^{\frac{k}{2}}}{2\sqrt{k\pi}} \int z^{k/2} e^{-z} dz = \frac{2^{k/2}}{2\sqrt{k\pi}} \Gamma(k/2+1) \leq C p^{p/2}$$

5. Why high probability bounds? Why not CLT?

In the future, we will be interested in high-probability tail bounds, where the magnitude of the random deviation scales as $\log 1/\delta$, instead of inverse polynomially in $1/\delta$. This might be hard to appreciate now, but ultimately, only bounds of the former sort will be useful to control the maximum of a bunch of random variables, which is what we are ultimately building towards.

If $X_1, ...X_n \sim \mathcal{N}(0,1)$, then $\frac{1}{n}\sum_{i=1}^n x_i \sim \mathcal{N}\left(0, \frac{1}{n}\right)$, and hence $\Pr\left(\frac{1}{n}\sum_{i=1}^n x_i \geq t\right) \leq e^{-nt^2/2}$.

Theorem 8 (Central Limit Theorem) Consider independent random variables $X_1, X_2...$ with mean μ and variance σ^2, then

$$\frac{\sum_{i=1}^n (X_i - \mu)}{\sigma\sqrt{n}} \to \mathcal{N}(0,1)$$

that is, their density functions at any level t converge, as $n \to \infty$.

CLT says that, in principle, for sufficiently large n, appropriately normalized sums behave as if each component is a Gaussian. The tail bound for a Gaussian decay exponentially. What stops us from reducing everything to the Gaussian case? CLT is an asymptotic statement and does not immediately imply anything for a finite number of samples. There is a way to repair this.

Theorem 9 (Berry-Essen) Consider independent random variables $X_1, X_2...$ with mean μ and variance σ^2, then for any t,

$$\left| \Pr\left(\frac{\sum_{i=1}^n (X_i - \mu)}{\sigma\sqrt{n}} \geq t \right) - \Pr(\mathcal{N}(0,1) \geq t) \right| \leq \frac{C\mathbb{E}[|X - \mathbb{E}X|^3]}{\sigma^3\sqrt{n}}.$$

But now, we get $\Pr\left(\frac{1}{n}\sum_{i=1}^n x_i - \mu \geq t\right) \leq e^{-nt^2/2} + \frac{C\mathbb{E}[|X-\mathbb{E}X|^3]}{\sigma^3\sqrt{n}}$, and thus, for the RHS to be at most δ, $n \geq 1/\delta^2$, ruling out a dependence of solely $\log 1/\delta$.

6. Subgaussian Random Variables

Theorem 10 (Markov's) For a nonnegative RV X, for any $t > 0$,

$$\Pr(X \geq t) \geq \frac{\mathbb{E}[X]}{t}.$$

Proof. Observe $X \geq X\mathbf{1}_{X \geq t} \geq t\mathbf{1}_{X \geq t}$. Take expectations on both sides and note $\mathbb{E}[\mathbf{1}_A] = \Pr(A)$. □

Exercise. One advantage of using indicator variables and expectations is that your proofs hold as-is for both continuous and discrete random variables, in fact, for mixed ones too. Similarly use indicate variables to prove that $\Pr\left(\cup_{\{i \in \mathcal{J})} A_i\right) \leq \sum_{i \in \mathcal{J}} \Pr(A_i)$.

Exercise. Using indicator variables and expectations, prove, for any a nonnegative random variable X, that $\mathbb{E}[X] = \int^{\infty} \Pr(X \geq t) dt$.

Theorem 11 (Chebyshev's) For a RV X with mean mu and variance σ^2, for any $t > 0$,

$$\Pr(|X - \mu| \geq t) \leq \frac{\sigma^2}{t^2}.$$

Proof. Apply Markov's starting from $\Pr(|X - \mu| \geq t) = \Pr((X - \mu)^2 \geq t^2)$. □

This implies that $\Pr\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geq t\right) \leq \frac{\sigma^2}{nt^2}$, but once again the number of samples is inverse polynomial in $1/\delta$. To get exponential tail bounds, we define a class of Gaussian-like random variables that have a Gaussian-like tail. Notice that this similarity stops at the tail and does not extend to the body of the distribution unlike CLT. Then, we will prove that this class of RVs is suitably closed under interesting operations, such as averaging.

Definition 12 A random variable X is said to be σ^2-SubGaussian if $\mathbb{E}\left[e^{t(X - \mathbb{E}X)}\right] \leq e^{\sigma^2 t^2/2}$.

An important note of caution here is that σ^2, being defined in terms of the MGF, in general is not equal to the variance of the SubGaussian distribution. Although this equality does hold for the Gaussian and Rademacher families, in the broadest terms it is merely an upper bound on the variance.

Lemma 13 (Hoeffding's) Any random variable with support in $[a, b]$ is $\frac{(b-a)^2}{4}$-SubGaussian.

Proof. In fact, we will prove a weaker result, but using a technique that we will reuse in the future.

$$\mathbb{E}_{X[e^{t(X - \mathbb{E}X)}]} \leq \mathbb{E}_{X,X'}\left[e^{t(X - X')}\right] = \mathbb{E}_{X,X'}\mathbb{E}_{\eta \sim \{\pm 1\}}\left[e^{t\eta(X - X')}\right] \leq \mathbb{E}_{X,X'} e^{t^2 (X - X')^2/2} \leq e^{t^2 (b-a)^2/2}$$

Here, the first inequality introduces an independent copy of the random variable X' and follows due to Jensen's, namely that $f(\mathbb{E}X) \leq \mathbb{E}f(X)$ for any convex function f. The only equality follows by noting that $X - X'$ and $X - X'$ have the same distribution. The second inequality involves the MGF of a Rademacher random variable. Overall, we have only proved the result to be $(b - a)^2$

-SubGaussian, but we will use the idea of introducing Rademacher random variables later on, and this is a good first introduction. □

Using this lemma, any bounded distribution, e.g. Bernoulli, is SubGaussian. Before we move on to concentration inequalities, let us give alternative characterizations of SubGaussian random variables.

Theorem 14 The following four conditions are equivalent, in the sense that all four constants $\sigma_1, \sigma_2, \sigma_3, \sigma_4$ are within (universal) constant factors of one another.

1. $\mathbb{E}\left[e^{t(X-\mathbb{E}X)}\right] \le e^{\sigma_1^2 t^2/2}$
2. $\Pr(|X-\mathbb{E}X| \ge t) \le 2e^{-t^2/2\sigma_2^2}$
3. $\left(\mathbb{E}[|X-\mathbb{E}X|^p]\right)^{1/p} \le \sigma_3\sqrt{p}$ for all $p \ge 1$
4. $\mathbb{E}\left[e^{(X-\mathbb{E}X)^2/\sigma_4^2}\right] \le 2$

Proof. (1⇒2) We start with a union bound, and then apply Markov's on the MGF. Because this move is valid for any $t \ge 0$, we choose the optimal $t = \varepsilon/\sigma_1^2$ by minimizing the quadratic.

$$\begin{aligned}\Pr(|X-\mathbb{E}X| \ge \varepsilon) &\le 2\Pr(X-\mathbb{E}X \ge \varepsilon) = 2\min_{t\ge 0}\Pr\left(e^{t(X-\mathbb{E}X)} \ge e^{t\varepsilon}\right)\\ &\le 2e^{-\max_{t\ge 0}\{t\varepsilon - t^2\sigma_1^2/2\}} = 2e^{-\varepsilon^2/2\sigma_1^2}\end{aligned}$$

(2⇒3) Recall the Γ function and that $\Gamma(k+1) \le k^k$, and then substitute $t = \left(2\sigma_2^2 z\right)^{p/2}$.

$$\begin{aligned}\mathbb{E}|X-\mathbb{E}X|^p &= \int_0^\infty \Pr(|X-\mathbb{E}X|^p \ge t)dt = 2\int_0^\infty e^{-t^{2/p}/2\sigma_2^2}dt\\ &= 2^{p/2+1}\sigma_2^p\int_0^\infty e^{-z}z^{p/2-1}dz = 2^{p/2+1}\sigma_2^p\Gamma(p/2)\end{aligned}$$

(3⇒4) Consider the power series expansion, where we use $k! \ge (k/e)^k$ and substitute $\sigma_4 = 2\sqrt{e}\sigma_3$.

$$\mathbb{E}\left[e^{(X-\mathbb{E}X)^2/\sigma_4^2}\right] = 1 + \sum_{k\ge 1}\frac{\mathbb{E}[(X-\mathbb{E}X)^{2k}]}{\sigma_4^{2k}k!} \le 1 + \sum_{k\ge 1}\frac{\left(\sigma_3^2 2k\right)^k}{\left(\sigma_4^2\frac{k}{e}\right)^k} = \frac{1}{1-2(\sigma_3/\sigma_4)^2 e} = 2$$

(4⇒1) This is messy. First, by Taylor's theorem, $e^x \le 1 + x + e^{|x|}\frac{x^2}{2}$. Taking expectation on both sides, we get

$$\begin{aligned}\mathbb{E}\left[e^{t(X-\mathbb{E}X)}\right] &\le 1 + \frac{t^2}{2}\mathbb{E}\left[(X-\mathbb{E}X)^2 e^{t|X-\mathbb{E}X|}\right]\\ &\le 1 + \frac{t^2}{2}\mathbb{E}\left[(X-\mathbb{E}X)^2 e^{\sigma_4^2 t^2/2}e^{(X-\mathbb{E}X)^2/2\sigma_4^2}\right] && \because tx \le \frac{\sigma_4^2 t^2 + x^2/\sigma_4^2}{2}\\ &\le 1 + \frac{\sigma_4^2 t^2}{2}e^{\sigma_4^2 t^2/2}\mathbb{E}\left[e^{(X-\mathbb{E}X)^2/\sigma_4^2}\right] \le (1+\sigma_4^2 t^2)e^{\sigma_4^2 t^2/2} && \because x^2 \le e^{x^2/2}\\ &\le e^{3\sigma_4^2 t^2/2} && \because (1+x) \le e^x.\end{aligned}$$

□

Two useful corollaries follow.

Corollary 14.1 If $X_1 \sim \text{SubGaussian}(\sigma_1^2)$ and $X_2 \sim \text{SubGaussian}(\sigma_2^2)$ are independent random variables, then $X_1 + X_2 \sim \text{SubGaussian}(\sigma_1^2 + \sigma_2^2)$.

Proof. This follows from noting that for independent RVs X, Y, $\mathbb{E}\left[e^{t(X+Y)}\right] = \mathbb{E}[e^{tX}]\mathbb{E}[e^{tY}]$ and using the MGF characterization. □

Corollary 14.2 If $X_1 \sim \text{SubGaussian}(\sigma_1^2), X_2 \sim \text{SubGaussian}(\sigma_2^2)$, then $X_1 + X_2 \sim \text{SubGaussian}\left((\sigma_1 + \sigma_2)^2\right)$.

Proof. Using the moment characterization, for zero-mean random variables, we get

$$\left(\mathbb{E}[|X_1 + X_2|^p]\right)^{1/p} \leq \left(\mathbb{E}[|X_1|^p]\right)^{1/p} + \left(\mathbb{E}[|X_2|^p]\right)^{1/p} \leq (\sigma_1 + \sigma_2)\sqrt{p},$$

where the first (triangle) inequality follows from the fact that $\|X\|_{L^p} = \left(\mathbb{E}[|X|^p]\right)^{1/p}$ is a (semi) norm. □

7. Concentration Inequalities for Linear Forms

Proposition 15 Let $\left\{X_i \sim \text{SubGaussian}(\sigma_i^2)\right\}_{i=1}^n$ be a collection of n independent random variables. Then for any $a \in \mathbb{R}^n$, $t \geq 0$, we have

$$\Pr\left(\left|\sum_{i=1}^n a_i X_i - \mathbb{E}\left[\sum_{i=1}^n a_i X_i\right]\right| \geq t\right) \leq 2e^{-\frac{t^2}{2\sum_{i=1}^n a_i^2 \sigma_i^2}}.$$

Proof. This follows from $\sum_{i=1}^n a_i X_i \sim \text{SubGaussian}\left(\sum_{i=1}^n a_i^2 \sigma_i^2\right)$, using the penultimate corollary. □

A common use case is that of averages, that is, when $a_i = 1/n$.

Corollary 15.1 Let $\left\{X_i \sim \text{SubGaussian}(\sigma^2)\right\}_{i=1}^n$ be a collection of n identically distributed independent random variables. Then for any $a \in \mathbb{R}^n$, $t \geq 0$, we have

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^n X - \mathbb{E}[X]\right| \geq t\right) \leq 2e^{-\frac{nt^2}{2\sigma^2}}.$$

Thus, for any $\delta > 0$, with probability at least $1 - \delta$, we have

$$\left|\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}[X]\right| \leq \sqrt{\frac{2\sigma^2 \log \frac{2}{\delta}}{n}}.$$

The moment generating function was seemingly essential for this bound, but the specific choice $x \mapsto e^{tX}$ might feel like an accident. Who is to say that we cannot do better? The following result from the theory of large deviations provides some moral support.

Theorem 16 (Cramer) For any sequence of independent and identically sampled random variables $X_1, X_2, ...$ with $\Psi(t) = \mathbb{E}\left[e^{t(X-\mathbb{E}X)}\right]$, we have for any $t \geq 0$ that

$$\lim_{n\to\infty} \frac{1}{n} \log \Pr\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}X \geq \varepsilon\right) \geq -\max_{t\geq 0}\{t\varepsilon - \log \Psi(t)\}.$$

8. Maximal Inequalities

Lemma 17 Let $\left\{X_i \sim \text{SubGaussian}(\sigma^2)\right\}_{i=1}^{n}$ be a collection of n zero-mean random variables. Then, for any $t \geq 0$, we have that

$$\mathbb{E}\left[\max_{i\leq n}|X_i|\right] \leq C\sigma\sqrt{\log n}, \text{ and } \Pr\left(\max_{i\leq n}|X_i| \geq t\right) \leq 2ne^{-\frac{t^2}{2\sigma^2}}.$$

Proof. The high probability bound follows from a union bound. We will try to prove

$$\Pr\left(\max_{i\leq n}|X_i| \geq t\right) \leq 2e^{-\frac{t^2}{8\sigma^2\log n}}.$$

If so, then $\mathbb{E}[\max_{i\leq n}|X_i|] \leq \int_0^\infty 2e^{-\frac{t^2}{8\sigma^2\log n}}dt = C\sigma\sqrt{\log n}$ using the standard Gaussian integral. If $n \leq e^{t^2/4\sigma^2}$, substituting this into the union bound suffices. In the other case, the RHS of the needed inequality is at least one, which is a tautology. □

9. Looking Ahead: Nonlinear Functions

In a few weeks, we will revisit concentration inequalities and arrive at a meta-principle: as long as $X_1, ...X_n$ are independent and $f(X_1, ...X_n)$ is not too sensitive in any of its arguments, we have that

$$\Pr(|f(X_1, ...X_n) - \mathbb{E}f(X_1, ...X_n)| > t) \leq e^{-cnt^2}.$$

Formulating this precisely, for different notions of *sensitivities*, will be both easy and difficult. At least one rudimentary form will automatically follow from a Martingale version of the argument we just developed. More sophisticated versions will link to deep mathematical ideas, like isoperimetric and Poincare inequalities, which, for example, link variance to the expected size of the gradient norm.

10. References

1. Primary reference: Chapters 0, 1 and 2 in Vershynin.
2. Alternative: Up to Section 1.4 in Rigollet and Hutter.

3. To appreciate measurability-related issues: Chapter 1 in Kantor, Matousek, Samal.
4. Concentration of Measure from the perspective of convex geometry: Naor.