6.7720/18.619/15.070 Lecture 6
Beyond the Second Moment: Chernoff Bounds

Kuikui Liu

February 19, 2025

Acknowledgements & Disclaimers *In the process of writing these notes, we consulted materials created by Guy Bresler and David Gamarnik, who taught previous iterations of this course. We are grateful for the discussions we had with them. Please be advised that these notes have not been subjected to the usual scrutiny reserved for formal publications. If you do spot an error, please contact the instructor.*

1 Weaknesses of Second Moment Tail Bounds

In the previous lectures, we've already seen how methods based purely on estimating expectations and variances can already yield extremely nontrivial results. In this lecture, we begin studying concentration inequalities which are stronger than what the second moment method and Chebyshev's Inequality can give. These stronger concentration inequalities typically go under the name of *Chernoff bounds*.

To get a sense of when Chebyshev will fail, consider the classic combinatorial optimization problem MAXCUT: Given a graph $G = (V, E)$, the goal is to find a partition $(S, \overline{S})$ of V into two nonempty pieces, i.e. a *cut*, maximizing the number of cut edges

$$\left|E\left(S, \overline{S}\right)\right| = \#\{uv \in E : u \in S, v \notin S\}.$$

We would like to get a handle on the random variable $\mathsf{MAXCUT}(G)$ for a randomly chosen graph $G \sim \boldsymbol{G}(n, 1/2)$. Certainly, for each fixed cut $(S, \overline{S})$, it is easy to compute the expectation and variance of the number of cut edges $\left|E\left(S, \overline{S}\right)\right|$, since we can use the independence afforded by the model $\boldsymbol{G}(n, 1/2)$. Hence, we can apply Chebyshev to deduce a 1% bound on the probability that $\left|E\left(S, \overline{S}\right)\right|$ deviates from its expectation by more than a constant factor.

Now, there are complex dependencies between the number of cut edges $\left|E\left(S, \overline{S}\right)\right|$ as we vary over the cuts $(S, \overline{S})$, and so the best tool at our disposal for controlling the *maximum* cut is the Union Bound. However, there are exponentially many possible cuts $(S, \overline{S})$, and so we would need an inverse exponential bound on the deviation probabilities for this strategy to have any chance of working. Unfortunately, Chebyshev simply isn't strong enough to furnish such bounds.

Note that Chebyshev's Inequality is tight in general. The reason it fails in the above application is primarily because it doesn't actually take full advantage of joint independence of $\boldsymbol{G}(n, 1/2)$. Indeed, the variance of $\left|E\left(S, \overline{S}\right)\right|$ is the same even if we allow the edges of G to be *pairwise independent*, which is significantly weaker. In general, the more independence you have among a collection of random variables, then the better concentrated their sum will be.

2 Concentration Inequalities Beyond Chebyshev

Throughout this section, we let $\{X_i\}_{i=1}^{\infty}$ be a sequence of independent and identically distributed (i.i.d.) random variables with mean μ and variance $\sigma^2 < \infty$. We also write $\overline{X}_n \stackrel{\mathsf{def}}{=} \frac{1}{n}\sum_{i=1}^{n} X_i$, which has mean μ and variance $\frac{\sigma^2}{n}$.

- **Law of Large Numbers (LLN):** This is the qualitative statement that averages of i.i.d. random variables concentrate around their mean. Formally, the *weak* version says that for every $\epsilon > 0$,

$$\lim_{n\to\infty} \Pr\left[\left|\overline{X}_n - \mu\right| < \epsilon\right] = 1.$$

We proved this in the previous lecture using Chebyshev's Inequality. Notably, it does not require the full power of joint independence; pairwise independence is sufficient. There is also a *strong* version of the law of large numbers which we won't discuss here.

- **Central Limit Theorem (CLT):** In many scenarios, we need more precise quantitative information on the deviation probability. This is furnished by the famous Central Limit Theorem, which stipulates pointwise convergence of the cumulative distribution function

$$\Pr\left[\frac{\sqrt{n}}{\sigma}\left(\overline{X}_n - \mu\right) \leq t\right] \to \Pr_{g\sim\mathcal{N}(0,1)}[g \leq t] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t} \exp\left(-\frac{g^2}{2}\right) dg, \qquad \forall t \in \mathbb{R},$$

where $\mathcal{N}(0,1)$ denotes the standard Gaussian distribution. Note that the rescaling by $\frac{\sqrt{n}}{\sigma}$ ensures that the resulting random variable has unit variance.

Standard bounds on the cumulative distribution function of $\mathcal{N}(0,1)$ say that

$$\frac{t}{t^2+1}\cdot\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{t^2}{2}\right) \leq \Pr_{g\sim\mathcal{N}(0,1)}[g > t] \leq \frac{1}{t}\cdot\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{t^2}{2}\right), \qquad \forall t > 0.$$

Hence, if we set t to be of order $\frac{\sqrt{n}}{\sigma}\cdot\mu$, then we expect to have the following *large deviation inequality*

$$\Pr\left[\left|\overline{X}_n - \mu\right| \geq \epsilon\mu\right] \text{ “}\lesssim\text{ ” } \exp\left(-\frac{n\mu^2}{\sigma^2}\right)$$

from the Central Limit Theorem. This exponential decay in n is extremely useful in practice since it is effective even when applying the Union Bound to a large number of random variables. However, at the moment, the above inequality is not legitimate due the error term in the convergence. In this lecture, we will establish such an inequality in the case of *bounded* random variables. These types of inequalities are often called *Chernoff bounds* in the literature.

Theorem 2.1 (Chernoff–Hoeffding Inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables, where for each $i = 1, \ldots, n$, X_i is bounded in the interval $[a_i, b_i]$ with probability 1. Then*

$$\Pr\left[\overline{X}_n - \mathbb{E}\left[\overline{X}_n\right] \geq t\right] \leq \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right), \qquad \forall t \geq 0.$$

Remark 1. Note that the quantity $\frac{1}{n^2}\sum_{i=1}^n \left(\frac{b_i - a_i}{2}\right)^2$ is only a *proxy* for the variance of $\overline{X}_n$. In general, it is an upper bound, with equality if and only if X_i takes the extremal values a_i, b_i each with probability 1/2. In an ideal world, we'd instead get a tail bound like

$$\Pr\left[\overline{X}_n - \mathbb{E}\left[\overline{X}_n\right] \geq t\right] \text{ “}\leq\text{ ” } \exp\left(-\frac{t^2}{2\operatorname{Var}\left(\overline{X}_n\right)}\right), \qquad \forall t \geq 0,$$

which would match the Gaussian case. We will see more refined inequalities of this flavor in the next few lectures, although many random variables do *not* satisfy such an inequality. We note that for $\{0,1\}$-valued random variables, there are user-friendly inequalities which are sharper than Theorem 2.1.

Note that we can apply the above theorem to $-\overline{X}_n$ as well to obtain the same upper bound on the lower tail $\Pr\left[\overline{X}_n - \mathbb{E}\left[\overline{X}_n\right] \leq -t\right]$. In particular, combining these two inequalities yields (and, for convenience, specializing to the i.i.d. case),

$$\Pr\left[\left|\overline{X}_n - \mathbb{E}\left[\overline{X}_n\right]\right| \geq t\right] \leq 2\cdot\exp\left(-\frac{2nt^2}{(b-a)^2}\right). \tag{1}$$

Before we prove Theorem 2.1, let us first apply it to the optimization problem from the beginning of the lecture.

Theorem 2.2. *We have*

$$\Pr_{G\sim\boldsymbol{G}(n,1/2)}\left[\mathsf{MAXCUT}(G) \geq \left(1 + \frac{1}{2\sqrt{n}}\right)\cdot\frac{n^2}{8}\right] \leq 2^{-n}.$$

Proof. Fix $\emptyset \subsetneq S \subsetneq V$, and for each pair $u \in S, v \in \overline{S}$, let X_{uv} denote the indicator random variable for whether or not $\{u, v\}$ is an edge in the random graph; note that these are all independent. Then expected size of the corresponding cut $\left(S, \overline{S}\right)$ is

$$\mathbb{E}_{G \sim \boldsymbol{G}(n,1/2)}\left[\left|E_G\left(S, \overline{S}\right)\right|\right] = \sum_{u \in S, v \in \overline{S}} \mathbb{E}[X_{uv}] = \frac{|S| \cdot |V \setminus S|}{2} \leq \max_{1 \leq k \leq n-1} \frac{k \cdot (n-k)}{2} \leq \frac{n^2}{8}.$$

Applying Theorem 2.1 and again using the fact that $|S| \cdot |V \setminus S| \leq n^2/4$

$$\Pr_{G \sim \boldsymbol{G}(n,1/2)}\left[\left|E_G\left(S, \overline{S}\right)\right| \geq \frac{n^2}{8} + t\right] \leq \exp\left(-\frac{2t^2}{|S| \cdot |V \setminus S|}\right) \leq \exp\left(-8(t/n)^2\right).$$

Setting $t = \frac{1}{2}n^{3/2}$ and then applying a Union Bound over all $\emptyset \subsetneq S \subsetneq V$, we obtain

$$\Pr_{G \sim \boldsymbol{G}(n,1/2)}\left[\mathsf{MAXCUT}(G) \geq \left(1 + \frac{1}{2\sqrt{n}}\right) \cdot \frac{n^2}{8}\right] \leq 2^n \exp\left(-2n\right) \leq 2^{-n}.$$

□

Remark 2. A matching lower bound for $\mathsf{MAXCUT}(G)$ is also known, namely

$$\Pr_{G \sim \boldsymbol{G}(n,1/2)}\left[\mathsf{MAXCUT}(G) \geq \left(1 + \Omega\left(\frac{1}{\sqrt{n}}\right)\right) \cdot \frac{n^2}{8}\right] \geq 1 - o(1).$$

2.1 The Moment Generating Function: Proof of Theorem 2.1

Observe that by generalizing the proof of Chebyshev's Inequality, we can bound the probability that a random variable X deviates from its expectation by leveraging *higher moments* of X: If X has $\mathbb{E}[X] = 0$, then by Markov's Inequality,

$$\Pr\left[X \geq t\right] \leq \Pr\left[|X|^k \geq t^k\right] \leq \frac{\mathbb{E}\left[|X|^k\right]}{t^k}, \qquad \forall t > 0.$$

Notice the $1/t^k$ decay rate in t, which is much better than Chebyshev when $k > 2$. Of course, one needs a good bound on the moment $\mathbb{E}\left[|X|^k\right]$, but this can often be obtained e.g. if X is a sum of sufficiently independent random variables.

If we have control over all of the moments of X, then we can encapsulate all of this information into a single *generating function*.

Definition 1 (Moment Generating Function)**.** *The* moment generating function *of a random variable X is defined as*

$$s \mapsto \mathbb{E}\left[\exp\left(s \cdot X\right)\right] = \sum_{k=0}^{\infty} \frac{s^k \cdot \mathbb{E}\left[X^k\right]}{k!}.$$

Remark 3. There are random variables for which its moment generating function can only be defined in a bounded interval of s. Of course, for bounded random variables, the moment generating function is well-defined everywhere. We actually already saw this generating function (slightly reparametrized) when we discussed the extinction probability for Galton–Watson branching processes.

To prove Theorem 2.1, we will need to bound the moment generating function of a bounded random variable. This is furnished by the following lemma.

Lemma 2.3 (Hoeffding's Lemma)**.** *Let X be a random variable taking values in the bounded interval $[a, b]$. Then we have the following upper bound on the moment generating function of X:*

$$\mathbb{E}\left[\exp\left(s \cdot (X - \mathbb{E}[X])\right)\right] \leq \exp\left(\frac{s^2}{2} \cdot \left(\frac{b-a}{2}\right)^2\right).$$

Proof. This lemma can be proved through direct calculations; we instead give a more conceptual proof. By shifting X and the interval $[a, b]$, we may assume that $\mathbb{E}[X] = 0$. Hence, our claim is equivalent to showing that

$$\psi_X(s) \stackrel{\text{def}}{=} \log \mathbb{E}[\exp(s \cdot X)] \leq \frac{s^2}{2} \cdot \left(\frac{b-a}{2}\right)^2.$$

The function $\psi_X(s)$ is known as the *cumulant generating function* of X. It is also fundamental, and we will see generalizations of it in a future lecture.

Claim 2.4. *For each $s \in \mathbb{R}$, define a new random variable Y_s whose law is given as follows:*

$$\Pr[Y_s = z] = \frac{\Pr[X = z] \cdot \exp(s \cdot z)}{\mathbb{E}\left[\exp(s \cdot X)\right]}, \qquad \forall z \in \mathbb{R}.$$

The law of Y_s is known as an exponential tilt *of the law of X. Then*

$$\psi_X'(s) = \mathbb{E}[Y_s]$$
$$\psi_X''(s) = \mathrm{Var}(Y_s).$$

In particular, ψ_X is convex on $\mathbb{R}$.

Proof. By the Chain Rule and linearity of expectation, we have

$$\psi_X'(s) = \frac{\frac{d}{ds}\mathbb{E}\left[\exp(s \cdot X)\right]}{\mathbb{E}\left[\exp(s \cdot X)\right]} = \frac{\mathbb{E}\left[\frac{d}{ds}\exp(s \cdot X)\right]}{\mathbb{E}\left[\exp(s \cdot X)\right]} = \frac{\mathbb{E}\left[X \cdot \exp(s \cdot X)\right]}{\mathbb{E}\left[\exp(s \cdot X)\right]} = \mathbb{E}[Y_s].$$

Differentiating again, we obtain

$$\psi_X''(s) = \frac{\mathbb{E}\left[X^2 \cdot \exp(s \cdot X)\right]}{\mathbb{E}\left[\exp(s \cdot X)\right]} - \frac{\mathbb{E}\left[X \cdot \exp(s \cdot X)\right]^2}{\mathbb{E}\left[\exp(s \cdot X)\right]^2} = \mathbb{E}[Y_s^2] - \mathbb{E}[Y_s]^2 = \mathrm{Var}(Y_s).$$

Since the variance of a random variable is always nonnegative, we deduce that ψ_X is convex. □

We will also need the following standard lemma from convex analysis, whose proof is given in Appendix A.

Claim 2.5. *Let $f, g : \mathbb{R} \to \mathbb{R}$ be smooth functions. If $f - g$ is convex, and there exists a point $x^* \in \mathbb{R}$ such that $f(x^*) \geq g(x^*)$ and $f'(x^*) = g'(x^*)$, then $f(x) \geq g(x)$ for all $x \in \mathbb{R}$.*

To conclude the proof, we verify the conditions of Claim 2.5 for ψ_X and $f(s) = \frac{s^2}{2} \cdot \left(\frac{b-a}{2}\right)^2$. Clearly, $f(0) = f'(0) = \psi_X(0) = 0$. Furthermore, $\psi_X'(0) = \mathbb{E}[Y_0] = \mathbb{E}[X] = 0$ by assumption. Finally, observe that their difference has second derivative

$$f''(s) - \psi_X''(s) = \left(\frac{b-a}{2}\right)^2 - \mathrm{Var}(Y_s),$$

where we again used Claim 2.4 for the last step. But Y_s has the same support as X, and in particular, is bounded in the interval $[a, b]$. Hence, $\mathrm{Var}(Y_s) \leq \left(\frac{b-a}{2}\right)^2$ for all $s \in \mathbb{R}$. This certifies the required convexity of $f - \psi_X$. Applying Claim 2.5 concludes the proof. □

Proof of Theorem 2.1. The high-level recipe for how nearly all Chernoff-type bounds are proved is as follows:

- Apply Markov's Inequality to the random variable $\exp(s \cdot X)$ for some parameter $s > 0$.
- Invoke a bound on the moment generating function of X.
- Optimize over the choice of s.

In the setting of Theorem 2.1, fix an arbitrary $s > 0$. We have

$$\begin{aligned}
\Pr\left[\overline{X}_n - \mathbb{E}\left[\overline{X}_n\right] \geq t\right] &= \Pr\left[\exp\left(s \cdot \left(\overline{X}_n - \mathbb{E}\left[\overline{X}_n\right]\right)\right) \geq \exp(s \cdot t)\right] \\
&\leq \frac{\mathbb{E}\left[\exp\left(s \cdot \left(\overline{X}_n - \mathbb{E}\left[\overline{X}_n\right]\right)\right)\right]}{\exp(s \cdot t)} && \text{(Markov's Inequality)} \\
&\leq \exp(-s \cdot t) \cdot \prod_{i=1}^{n} \exp\left(\frac{s}{n} \cdot (X_i - \mathbb{E}[X_i])\right) && \text{(Independence)} \\
&\leq \exp\left(\frac{s^2}{2n^2} \cdot \sum_{i=1}^{n} \left(\frac{b_i - a_i}{2}\right)^2 - s \cdot t\right). && \text{(Lemma 2.3)}
\end{aligned}$$

Since this holds for all s, to obtain the sharpest result, we choose s to minimize the right-hand side. This yields

$$\begin{aligned}
\Pr[X \geq t] &\leq \exp\left(\inf_{s>0}\left\{\frac{s^2}{2n^2} \cdot \sum_{i=1}^{n} \left(\frac{b_i - a_i}{2}\right)^2 - s \cdot t\right\}\right) \\
&= \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right),
\end{aligned}$$

where the infimum is attained at $s = \frac{4tn^2}{\sum_{i=1}^{n}(b_i - a_i)^2}$. □

3 Quantitative CLT and Anticoncentration

Theorem 3.1 (Berry–Esseen Theorem)**.** *Let $\{X_i\}_{i=1}^{\infty}$ be a sequence of independent and identically distributed (i.i.d.) random variables with mean μ and variance $\sigma^2 < \infty$. Suppose in addition our random variables also have finite skewness $\gamma \stackrel{\text{def}}{=} \mathbb{E}_X\left[\frac{|X-\mu|^3}{\sigma^3}\right] < \infty$. Then there exists a universal constant $C > 0$ such that we have the uniform convergence estimate*

$$\sup_{t \in \mathbb{R}} \left|\Pr\left[\frac{\sqrt{n}}{\sigma}\left(\overline{X}_n - \mu\right) \leq t\right] - \Pr_{g \sim \mathcal{N}(0,1)}[g \leq t]\right| \leq \frac{C\gamma}{\sqrt{n}}.$$

Proving this theorem is out of the scope of this lecture. However, let us use it to understand the typical order of fluctuations of the random variable $\overline{X}_n$. As we previously discussed, neither the Central Limit Theorem nor the Berry–Esseen Theorem can be used to directly obtain the exponential decay in n we saw in Eq. (1), since there is an additive error term decaying at the much slower rate of $O(1/\sqrt{n})$. However, the advantage of the precision of Theorem 3.1 is that it also gives *lower bounds* on the mass of the tails. To see this, observe that for any $t \geq 0$,

$$\Pr\left[\left|\overline{X}_n - \mu\right| \geq t \cdot \frac{\sigma}{\sqrt{n}}\right] \geq 2 \cdot \Pr_{g \sim \mathcal{N}(0,1)}[g \geq t] - O(1/\sqrt{n}).$$

In particular, setting t to be an absolute constant, we obtain

$$\Pr\left[\left|\overline{X}_n - \mu\right| \geq \Omega\left(\sqrt{\mathrm{Var}\left(\overline{X}_n\right)}\right)\right] \geq \Omega(1) > 0.$$

In other words, the deviation $\left|\overline{X}_n - \mu\right|$ really does fluctuate at the order of the standard deviation of $\overline{X}_n$. This is certainly not true for general random variables with finite variance. In the literature, this type of inequality is often referred to as an *anticoncentration inequality*.

A Unfinished Proofs

Proof of Claim 2.5. By replacing f with $f - g$ and replacing g with the function which is identically 0, it suffices to show that $f(x) \geq 0$ globally if f is convex and satisfies $f(x^*) \geq 0, f'(x^*) = 0$ for some point $x^* \in \mathbb{R}$. One way to see this is that global convexity of f combined with $f'(x^*) = 0$ imply that x^* is a minimizer of f. Hence, $f(x) \geq f(x^*) \geq 0$ for all $x \in \mathbb{R}$. Another way to see this

is to use the fact that f is lower bounded by its tangent line at any point. In particular, using the tangent at x^*, we obtain

$$f(x) \geq f(x^*) + f'(x^*) \cdot (x - x^*) = f(x^*) \geq 0$$

for all $x \in \mathbb{R}$. □