5 Chernoff bound

Chernoff bounds give us much better tail bounds than the second moment method when applied to sums of independent random variables. This is one of the most useful bounds in probabilistic combinatorics.

The proof technique of bounding the exponential moments is perhaps just as important as the resulting bounds themselves. We will see this proof method come up again later on when we prove martingale concentration inequalities. The method allows us to adapt the proof of the Chernoff bound to other distributions. Let us give the proof in the most basic case for simplicity and clarity.

Theorem 5.0.1. Let $S_n = X_1 + \cdots + X_n$ where $X_i \in \{-1, 1\}$ uniformly iid. Let $\lambda > 0$. Then

$$\mathbb{P}(S_n \geq \lambda\sqrt{n}) \leq e^{-\lambda^2/2}$$

Note that in contrast, $\operatorname{Var} S_n = n$, so Chebyshev's inequality would only give a tail bound $\leq 1/\lambda^2$

Proof. Let $t \geq 0$. Consider the **moment generating function**

$$\mathbb{E}\left[e^{tS_n}\right] = \mathbb{E}\left[e^{t\sum_i X_i}\right] = \mathbb{E}\left[\prod_i e^{tX_i}\right] = \prod_i \mathbb{E}\left[e^{tX_i}\right] = \left(\frac{e^{-t}+e^t}{2}\right)^n.$$

We have (by comparing Taylor series coefficients $\frac{1}{(2n)!} \leq \frac{1}{n!2^n}$), for all $t \geq 0$,

$$\frac{e^{-t}+e^t}{2} \leq e^{t^2/2}.$$

By Markov's inequality,

$$\mathbb{P}(S_n \geq \lambda\sqrt{n}) \leq \frac{\mathbb{E}\left[e^{tS}\right]}{e^{t\lambda\sqrt{n}}} \leq e^{-t\lambda\sqrt{n}+t^2n/2}$$

Set $t = \lambda/\sqrt{n}$ gives the bound. □

Remark 5.0.2. The technique of considering the moment generating function can be thought morally as taking an appropriately high moment. Indeed, $\mathbb{E}[e^{tS}] = \sum_{n\geq 0} \mathbb{E}[S^n]t^n/n!$ contains all the moments data of the random variable.

The second moment method (Chebyshev + Markov) can be thought of as the first iteration of this idea. By taking fourth moments (now requiring 4-wise independence of the summands), we can obtain tail bounds of the form $\lesssim \lambda^{-4}$. And similarly with higher moments.

In some applications, where one cannot assume independence, but can estimate high moments, the above philosophy can allow us to prove good tail bounds as well.

Also by symmetry, $\mathbb{P}(S_n \le -\lambda\sqrt{n}) \le e^{-\lambda^2/2}$. Thus we have the following two-sided tail bound.

Corollary 5.0.3. $\mathbb{P}(|S_n| \ge \lambda\sqrt{n}) \le 2e^{-\lambda^2/2}$

Remark 5.0.4. It is easy to adapt the above proof so that each X_i is a mean-zero random variable taking $[-1,1]$-values, and independent (but not necessarily identical) across all i. Indeed, by convexity, we have $e^{tx} \le \frac{1-x}{2}e^{-t} + \frac{1+x}{2}e^{t}$ for all $x \in [-1,1]$ by convexity, so that $\mathbb{E}[e^{tX}] \le \frac{e^t+e^{-t}}{2}$. In particular, we obtain the following tail bounds on the binomial distribution.

Theorem 5.0.5. Let each X_i be an independent random variable taking values in $[-1,1]$ and $\mathbb{E}X_i = 0$. Then $S_n = X_1 + \cdots + X_n$ satisfies

$$\mathbb{P}(S_n \ge \lambda\sqrt{n}) \le e^{-\lambda^2/2}.$$

Corollary 5.0.6. Let X be a sum of n independent Bernoulli's (not necessarily the same probability). Let $\mu = \mathbb{E}X$ and $\lambda > 0$. Then Then

$$\mathbb{P}(X \ge \mu + \lambda\sqrt{n}) \le e^{-\lambda^2/2} \quad \text{and} \quad \mathbb{P}(X \le \mu - \lambda\sqrt{n}) \le e^{-\lambda^2/2}$$

The quality the Chernoff compares well to that of the normal distribution. For the standard normal $Z \sim N(0,1)$, one has $\mathbb{E}[e^{tZ}] = e^{t^2/2}$ and so

$$\mathbb{P}(Z \ge \lambda) = \mathbb{P}(e^{tZ} \ge e^{t\lambda}) \le e^{-t\lambda}\mathbb{E}[e^{tX}] = e^{-t\lambda + t^2/2}$$

Set $t = \lambda$ and get

$$\mathbb{P}(Z \ge \lambda) \le e^{-\lambda^2/2}$$

And this is actually pretty tight, as, for $\lambda \to \infty$,

$$\mathbb{P}(Z \ge \lambda) = \frac{1}{\sqrt{2\pi}} \int_{\lambda}^{\infty} e^{-t^2/2}\, dt \sim \frac{e^{-\lambda^2/2}}{\sqrt{2\pi\lambda}}$$

The same proof method allows you to prove bounds for other sums of random variables, suitable for whatever application you have in mind. See Alon–Spencer Appendix A for some calculations.

For example, for a sum of independent Bernoulli's with small means, we can improve on the above estimates as follows

Theorem 5.0.7. Let X be the sum of independent Bernoulli random variables (not necessarily same probability). Let $\mu = \mathbb{E}X$. For all $\epsilon > 0$,

$$\mathbb{P}(X \geq (1+\epsilon)\mu) \leq e^{-((1+\epsilon)\log(1+\epsilon)-\epsilon)\mu} \leq e^{-\frac{\epsilon^2}{1+\epsilon}\mu}$$

and

$$\mathbb{P}(X \leq (1-\epsilon)\mu) \leq e^{-\epsilon^2\mu/2}.$$

Remark 5.0.8. The bounds for upper and lower tails are necessarily asymmetric, when the probabilities are small. Why? Think about what happens when $X \sim \text{Bin}(n, c/n)$, which, for a constant $c > 0$, converges as $n \to \infty$ to a Poisson distribution with mean c, whose value at k is $c^k e^k/k! = e^{-\Theta(k \log k)}$ and not $e^{-\Omega(k^2)}$ as one might naively predict by an incorrect application of the Chernoff bound formula.

Nonetheless, both formulas tell us that both tails exponentially decay like ϵ^2 for small values of ϵ, say, $\epsilon \in [0, 1]$.

5.1 Discrepancy

Theorem 5.1.1. Let $\mathcal{F}$ be a collection of m subsets of $[n]$. Then there exists some assignment $[n] \to \{-1, 1\}$ so that the sum on every set in $\mathcal{F}$ is at most $2\sqrt{n \log m}$ in absolute value.

Proof. Put ±1 iid uniformly at random on each vertex. On each edge, the probability that the sum exceeds $2\sqrt{n\log m}$ in absolute value is, by Chernoff bound, less than $2e^{-2\log m} = 2/m^2$. By union bound over all m edges, with probability greater than $1 - 2/m \geq 0$, no edge has sum exceeding $2\sqrt{n \log m}$. □

Remark 5.1.2. In a beautiful landmark paper titled *Six standard deviations suffice*, Spencer (1985) showed that one can remove the logarithmic term by a more sophisticated semi-random assignment algorithm.

Theorem 5.1.3 (Spencer (1985))**.** Let $\mathcal{F}$ be a collection of n subsets of $[n]$. Then there exists some assignment $[n] \to \{-1, 1\}$ so that the sum on every set in $\mathcal{F}$ is at most $6\sqrt{n}$ in absolute value.
More generally, if $\mathcal{F}$ be a collection of $m \geq n$ subsets of $[n]$, then we can replace $6\sqrt{n}$ by $11\sqrt{n \log(2m/n)}$.

Remark 5.1.4. More generally, Spencer proves that the same holds if vertices have $[0,1]$-valued weights.

The idea, very roughly speaking, is to first generalize from $\{-1,1\}$-valued assignments to $[-1,1]$-valued assignments. Then the all-zero vector is a trivially satisfying assignment. We then randomly, in iterations, alter the values from 0 to other values in $[-1,1]$, while avoiding potential violations (e.g., edges with sum close to $6\sqrt{n}$ in absolute value), and finalizing a color of a color when its value moves to either -1 and 1.

Spencer's original proof was not algorithmic, and he suspected that it could not be made efficiently algorithmic. In a breakthrough result, Bansal (2010) gave an efficient algorithm for producing a coloring with small discrepancy. Another very nice algorithm with another beautiful proof of the algorithmic result was given by Lovett and Meka (2015).

Here is a famous conjecture on discrepancy.

Conjecture 5.1.5 (Komlós)**.** There exists some absolute constant K so that for every set of vectors $v_1, \dots, v_m$ in the unit ball in R^n, there exists signs $\epsilon_1, \dots, \epsilon_m \in \{-1,1\}$ such that

$$\epsilon_1 v_1 + \cdots + \epsilon_m v_m \in [-K,K]^n.$$

Banaszczyk (1998) proved the bound $K = O(\sqrt{\log n})$ in a beautiful paper using deep ideas from convex geometry.

Spencer's theorem's implies the Komlós conjecture if all vectors v_i have the form $n^{-1/2}(\pm 1, \dots, \pm 1)$ (or more generally when all coordinates are $O(n^{-1/2})$). The deduction is easy when $m \le n$. When $m > n$, we use the following observation.

Lemma 5.1.6. Let $v_1, \dots, v_m \in \mathbb{R}^n$. Then there exists $a_1, \dots, a_m \in [-1,1]^m$ with $|\{i : a_i \notin \{-1,1\}\}| \le n$ such that

$$a_1 v_1 + \cdots + a_m v_m = 0$$

Proof. Find $(a_1, \dots, a_m) \in [-1,1]^m$ satisfying and as many $a_i \in \{-1,1\}$ as possible. Let $I = \{i : a_i \notin \{-1,1\}\}$. If $|I| > n$, then we can find some nontrivial linear combination of the vectors $v_i, i \in I$, allowing us to to move $(a_i)_{i \in I}$'s to new values, while preserving $a_1 v_1 + \cdots + a_m v_m = 0$, and end up with at one additional a_i taking $\{-1,1\}$-value. □

Letting $a_1, \dots, a_m$ and $I = \{i : a_i \notin \{-1,1\}\}$ as in the Lemma, we then take $\epsilon_i = a_i$ for all $i \notin I$, and apply a corollary of Spencer's theorem to find $\epsilon_i \in \{-1,1\}^n$, $i \in I$ with

$$\sum_{i \in I} (\epsilon_i - a_i) v_i \in [-K,K]^n,$$

which would yield the desired result. The above step can be deduced from Spencer's theorem by first assuming that each $a_i \in [-1,1]$ has finite binary length (a compactness argument), and then rounding it off one digit at a time during Spencer's theorem, starting from the least significant bit (see Corollary 8 in Spencer's paper for details).

5.2 Hajós conjecture counterexample

We begin by reviewing some classic result from graph theory. Recall some definitions:

- H is an **induced subgraph** of G if H can be obtained from G by removing vertices;
- H is a **subgraph** if G if H can be obtained from G by removing vertices and edges;
- H is a **subdivision** of G if H can be obtained from a subgraph of G by contracting induced paths to edges;
- H is a **minor** of G if H can be obtained from a subgraph of G by by contracting edges to vertices.

Kuratowski's theorem (1930). Every graph without $K_{3,3}$ and K_5 as subdivisions as subdivision is planar.

Wagner's theorem (1937). Every graph free of $K_{3,3}$ and K_5 as minors is planar.

(There is a short argument shows that Kuratowski and Wagner's theorems are equivalent.)

Four color theorem (Appel and Haken 1977) Every planar graph is 4-colorable.

Corollary: Every graph without $K_{3,3}$ and K_5 as minors is 4-colorable.

The condition on K_5 is clearly necessary, but what about $K_{3,3}$? What is the "real" reason for 4-colorability.

Hadwidger posed the following conjecture, which is one of the biggest open conjectures in graph theory.

Conjecture 5.2.1 (Hadwiger 1936)**.** For every $t \geq 1$, every graph without a K_{t+1} minor is t-colorable.

$t = 1$ trivial

$t = 2$ nearly trivial (if G is K_3-minor-free, then it's a tree)

$t = 3$ elementary graph theoretic arguments

$t = 4$ is equivalent to the 4-color theorem (Wagner 1937)

$t = 5$ is equivalent to the 4-color theorem (Robertson–Seymour–Thomas 1994; this work won a Fulkerson Prize)

$t \geq 6$ remains open

Let us explore a variation of Hadwiger's conjecture:

Hajós conjecture. (1961) Every graph without a K_{t+1}-subdivision is t-colorable.

Hajós conjecture is true for $t \leq 3$. However, it turns out to be found in general. Catlin (1979) constructed counterexamples for all $t \geq 6$ ($t = 4, 5$ are still open).

It turns out that Hajós conjecture is not just false, but very false.

Erdős–Fajtlowicz (1981) showed that almost every graph is a counterexample (it's a good idea to check for potential counterexamples among random graphs!)

To be continued

Theorem 5.2.2. With probability $1 - o(1)$, $G(n, 1/2)$ has no K_t-subdivision with $t = \lceil 10\sqrt{n} \rceil$.

From Theorem 4.3.3 we show that, with high probability, $G(n, 1/2)$ has independence number $\sim 2\log_2 n$ and hence chromatic number $\geq (1 + o(1))\frac{n}{2\log_2 n}$. Thus the above result shows that $G(n, 1/2)$ is whp a counterexample to Hajós conjecture.

Proof. If G had a K_t-subdivision, say with $S \subset V$, $|S| = t$, then at most $n - t \leq n$ of the edges in the subdivision can be paths with at least two edges (since they must use distinct vertices outside S). So S must induce at least $\binom{t}{2} - n \geq \frac{3}{4}\binom{t}{2}$ edges in G.

By Chernoff bound, for fixed t-vertex subset S

$$\mathbb{P}\left(e(S) \geq \frac{3}{4}\binom{t}{2}\right) \leq e^{-t^2/10}.$$

Taking a union bound over all t-vertex subsets S, and noting that

$$\binom{n}{t} e^{-t^2/10} < n^t e^{-t^2/10} \leq e^{-10n + O(\sqrt{n}\log n)} = o(1)$$

we see that whp no such S exists, so that this $G(n, 1/2)$ whp has no K_t-subdivision □

Remark 5.2.3. One can ask the following quantitative question regarding Hadwidger's conjecture:

Can we show that every graph without a K_{t+1}-minor can be properly colored with a small number of colors?

Wagner (1964) showed that every graph without K_{t+1}-minor is 2^{t-1} colorable.

Here is the proof: assume that the graph is connected. Take a vertex v and let L_i be the set of vertices with distance exactly i from v. The subgraph induced on L_i has no K_t-minor, since otherwise such a K_t-minor would extend to a K_{t+1}-minor with v. Then by induction L_i is 2^{t-2}-colorable (check base cases), and using alternating colors for even and odd layers L_i yields a proper coloring of G.

This bound has been improved over time. The best current bound was proved this past summer. Postle (2020+) showed that if every graph with no K_t-minor is $O(t(\log\log t)^6)$-colorable.

For more on Hadwiger's conjecture, see Seymour's survey (2016).