CONCENTRATION INEQUALITIES

JOHN THICKSTUN

Theorem. *(Markov) For any $h : \mathbb{R} \to [0, \infty)$ and $a > 0$,*

$$P(h(X) \geq a) \leq \frac{\mathbb{E}h(X)}{a}.$$

Proof. Because $h(X) \geq 0$, clearly $a\mathbb{1}_{h(X)\geq a} \leq h(X)$. By monotonicity of expectation,

$$aP(h(X) \geq a) = a\mathbb{E}(\mathbb{1}_{h(X)\geq a}) = \mathbb{E}(a\mathbb{1}_{h(X)\geq a}) \leq \mathbb{E}h(X).$$

□

If $h(x) = |x|$ then we have $P(|X| \geq a) \leq \frac{\mathbb{E}|X|}{a}$. This is the classic Markov inequality. Note that without further assumptions, the bound is sharp. In particular, consider the deterministic random variable $X = c$. In the case when $c \geq a$,

$$P(|X| \geq a) = \mathbb{1}_{c\geq a} = \frac{\mathbb{E}|X|}{a}.$$

We can get different bounds by investigating various choices of h.

Theorem. *(Chebyshev) If* $\operatorname{Var} X < \infty$ *then*

$$P(|X| \geq a) \leq \frac{\mathbb{E}X^2}{a^2}.$$

Proof. Take $h(x) = x^2$ and by Markov's inequality,

$$P(|X| \geq a) = P(X^2 \geq a^2) = P(h(X) \geq a^2) \leq \frac{\mathbb{E}h(X)}{a^2}.$$

□

If we consider $Y = X - \mathbb{E}X$ then we get a central inequality

$$P(|X - \mathbb{E}X| \geq a) = P(|Y| \geq a) \leq \frac{\mathbb{E}Y^2}{a^2} = \frac{\operatorname{Var} X}{a^2}.$$

Again this bound is tight. Suppose X is distributed on $\{-a, 0, a\}$ according to

$$p(x) = \begin{cases} p & \text{if } x = -a \\ p & \text{if } x = a \\ 1 - 2p & \text{if } x = 0 \end{cases}$$

Then

$$\frac{\operatorname{Var} X}{a^2} = \frac{2pa^2}{a^2} = 2p = P(|X| \geq a).$$

The same style of argument implies a bound for each moment $\mathbb{E}|X|^k$. Finding the k that gives us the best bound, gives us the following inequality.

Proposition. *If $\mathbb{E}|X|^k < \infty$ for all k, then for all $a > 0$*

$$P(|X - \mathbb{E}X| \geq a) \leq \min_{k>0} \frac{\mathbb{E}|X|^k}{a^k}.$$

We can also get bounds using an exponential, sometimes called the exponential Chebyshev inequalities

$$P(X \geq a) \leq e^{-ta}\mathbb{E}e^{tX}.$$

There is some intuition that, because the mgf $M_X(t) \equiv \mathbb{E}e^{tX}$ encodes all the moments of X, these bounds will be at least as good as (better than?) the moment bounds. But I don't have a good grasp of this. Minimizing these bounds over t gives us an abstract Chernoff bound.

Theorem. *(Chernoff) If M_X exists and $\epsilon > 0$ then*

$$P(X \geq \epsilon) \leq \min_t e^{-t\epsilon}\mathbb{E}e^{tX}.$$

Considering random variables with additional structure will let us get sharper results. First let's consider what happens when X is bounded; i.e. $X \in [a, b]$.

Lemma. *(Hoeffding) Let $\mathbb{E}X = 0$ with $X \in [a, b]$ almost surely. Then*

$$\mathbb{E}e^{tX} \leq \exp\left(\frac{1}{8}t^2(b-a)^2\right).$$

Proof. Because $X \in [a, b]$, we can write X as a convex combination $X = \alpha b + (1-\alpha)a$ and in particular

$$\alpha = \frac{X-a}{b-a}, \qquad (1-\alpha) = \frac{b-X}{b-a}.$$

By convexity of the exponential,

$$e^{tX} \leq \alpha e^{tb} + (1-\alpha)e^{ta} = \frac{X-a}{b-a}e^{tb} + \frac{b-X}{b-a}e^{ta}.$$

Define $L(h) = -hp + \log(1 - p + pe^h)$ where $h = t(b-a)$ and $p = -a/(b-a)$. Because $\mathbb{E}X = 0$,

$$\mathbb{E}e^{tX} \leq -\frac{a}{b-a}e^{tb} + \frac{b}{b-a}e^{ta} = e^{L(h)}.$$

By Taylor's theorem, there is some $z \in (0, h)$ with

$$L(h) = L(0) + hL'(0) + \frac{1}{2}h^2L''(z).$$

Note that $L(0) = L'(0)$, and $L''(z) \leq 1/4$ (see `https://en.wikipedia.org/wiki/Hoeffding%27s_inequality#Proof_of_Hoeffding.27s_Lemma`). We conclude that

$$L(h) = \frac{1}{2}h^2L''(z) \leq \frac{1}{8}h^2 = \frac{1}{8}t^2(b-a)^2. \qquad \square$$

Combining Hoeffding's Lemma with the abstract Chernoff bound described above gives us our first version of the Chernoff-Hoeffding bound.

Theorem. *(Chernoff-Hoeffding) If $\mathbb{E}X = 0$ with $X \in [a, b]$ almost surely then*

$$P(X \geq \epsilon) \leq \exp\left(-\frac{2\epsilon^2}{(b-a)^2}\right).$$

Proof. If X is restricted to $[a, b]$ then clearly M_X exists. By Chernoff's bound and Hoeffding's lemma respectively,

$$P(X \geq \epsilon) \leq \min_t e^{-t\epsilon}\mathbb{E}e^{tX} \leq \min_t e^{-t\epsilon}\exp\left(\frac{1}{8}t^2(b-a)^2\right)$$

$$= \min_t \exp\left(\frac{1}{8}t^2(b-a)^2 - t\epsilon\right) = \exp\left(\min_t \frac{1}{8}t^2(b-a)^2 - t\epsilon\right).$$

The quadratic is minimized when

$$\frac{1}{4}t(b-a)^2 - \epsilon = 0.$$

In particular we will get the sharpest bound when $t = 4\epsilon/(b-a)^2$. In this case

$$P(X \geq \epsilon) \leq \exp\left(\frac{2\epsilon^2}{(b-a)^2} - \frac{4\epsilon^2}{(b-a)^2}\right) = \exp\left(-\frac{2\epsilon^2}{(b-a)^2}\right).$$

□

Corollary. *Let $\mu = \mathbb{E}X$ with $X \in [a, b]$ almost surely. Then*

$$P(X \leq \mathbb{E}X + \epsilon) \leq \exp\left(-\frac{2\epsilon^2}{(b-a)^2}\right)$$

Proof. Let $Y = X - \mathbb{E}X$. Then $\mathbb{E}Y = 0$ and $Y \in [a - \mathbb{E}X, b - \mathbb{E}X]$. Then

$$P(X \geq \mathbb{E}X + \epsilon) = P(Y \geq \epsilon) \leq$$

□

We can also get a relative bound.

Corollary. *(Relative Chernoff-Hoeffding) Let $\mu = \mathbb{E}X$ with $X \in [a, b]$ almost surely. Then*

$$P(X \leq (1-\epsilon)\mu) \leq \exp\left(-\frac{\mu\epsilon^2}{2(b-a)^2}\right)$$

Proof. Let $Y = X - \mu$. Then $\mathbb{E}Y = 0$ and by the preceding theorem,

$$P(X \leq (1-\epsilon)\mu) = P(Y \leq -\epsilon\mu) \leq \exp\left(-\frac{\mu^2\epsilon^2}{(b-a)^2}\right)$$

□

Now let's turn our attention to a different kind of structure. Suppose our random variable is a sum S_n of random variables X_1^n. We can bound the expectation of this sum by the triangle inequality:

$$\mathbb{E}|S_n| \leq \sum_{i=1}^n \mathbb{E}|X_i|.$$

This and Markov's inequality give us a weak concentration statement:

$$P(|S_n| \geq a) \leq \frac{1}{a} \sum_{i=1}^n \mathbb{E}|X_i|.$$

This bound is sharp without further assumptions. For example, if X is uniform on $\{-1, 1\}$ and $X_i = X$ then

$$P(|S_n| \geq n) = 1 = \frac{1}{n} \sum_{i=1}^n 1.$$

The problem is that when X_i are correlated, they can combine produce large deviations in the sum. When they are perfectly correlated our sum yields no additional structure at all and our analysis reduces to the analysis of an arbitrary random variable. If we reduce the correlation of X_i we would expect their sum to produce a more even outcome, allowing a sharper analysis. This is the concentration of measure phenomenon.

If X_1^n are uncorrelated then $\operatorname{Var} S_n = \sum \operatorname{Var} X_i$ and by Chebyshev,

$$P(|S_n - \mathbb{E}S_n| \geq a) \leq \frac{1}{a^2} \operatorname{Var} S_n = \frac{1}{a^2} \sum_{i=1}^n \operatorname{Var} X_i.$$

We can proceed in this way, deriving bounds corresponding to higher order moments $\mathbb{E}|S_n|^k$, and get a sharp bound considering these bounds together. Terry Tao analyzes this approach in this blog post[1]. Or we use Chernoff's method and consider exponential bounds.

If X_1^n are independent we can proceed much more confidently.

Theorem. *(Chernoff-Hoeffding) Let X_1^n be independent, $\mathbb{E}X_i = 0$, and $a \leq X \leq b$. If $S_n = \sum_{i=1}^n X_i$ then*

$$P(S_n \geq \epsilon) \leq \exp\left(\frac{-2\epsilon^2}{n(b-a)^2}\right).$$

Proof. By Markov's inequality and independence,

$$P(S_n \geq \epsilon) = P(e^{tS_n} \geq e^{t\epsilon}) \leq e^{-t\epsilon}\mathbb{E}e^{tS_n} = e^{-t\epsilon} \prod_{i=1}^n \mathbb{E}e^{tX_i}.$$

By Hoeffding's lemma,

$$P(S_n \geq \epsilon) \leq e^{-t\epsilon} \prod_{i=1}^n \exp\left(\frac{n}{8}t^2(b-a)^2\right) = \exp\left(\frac{1}{8}t^2 n(b-a)^2 - t\epsilon\right).$$

[1] https://terrytao.wordpress.com/2010/01/03/254a-notes-1-concentration-of-measure/

Optimizing over t gives us a Chernoff bound when

$$\frac{n}{4}t(b-a)^2 - \epsilon = 0.$$

Solving for t, we find that $t = 4\epsilon/n(b-a)^2$ and

$$P(S_n \geq \epsilon) \leq \min_t \exp\left(\frac{1}{8}t^2 n(b-a)^2 - t\epsilon\right) = \exp\left(\frac{2\epsilon^2}{n(b-a)^2} - \frac{4\epsilon^2}{n(b-a)^2}\right).$$

□

What can we say if our variables are not independent? We can relax this assumption and replace it with a martingale condition.

Theorem. *(Azuma-Hoeffding) Suppose (X_i) is a bounded martingale difference sequence; i.e. $|X_i| \leq c$. If $S_n = \sum_{i=1}^n X_i$ then*

$$P(S_i \geq \epsilon) \leq \exp\left(\frac{-2\epsilon^2}{nc^2}\right).$$

Proof. By the same reasoning we used to prove Chernoff-Hoeffding,

$$P(S_n \geq \epsilon) \leq e^{-t\epsilon}\mathbb{E}e^{tS_n}$$

But we must take more care at this point in the argument. Consider $\mathbb{E}e^{tS_k}$ for $k \in \{1, \ldots, n\}$. If $k > 1$ then by the tower property of conditional expectation,

$$\mathbb{E}e^{tS_k} = \mathbb{E}\mathbb{E}\left(e^{tS_k}|S_{k-1}\right).$$

Pulling out the known quantity from the expectation given S_{k-1},

$$\mathbb{E}\left(e^{tS_k}|S_{k-1}\right) = \mathbb{E}\left(e^{tS_{k-1}+tX_k}|S_{k-1}\right) = e^{tS_{k-1}}\mathbb{E}\left(e^{tX_k}|S_{k-1}\right).$$

Note that $\mathbb{E}\left(X_k|S_{n-1}\right) = 0$ because (X_i) is a martingale difference sequence. Therefore Hoeffding's lemma applies and in particular

$$\mathbb{E}\left(e^{tX_k}|S_{k-1}\right) \leq e^{t^2c^2/2}.$$

This is not random, so we can pull it out of the expectation:

$$\mathbb{E}e^{tS_k} = e^{t^2c^2/2}\mathbb{E}\mathbb{E}\left(e^{tS_{k-1}}|S_{k-1}\right) = e^{t^2c^2/2}\mathbb{E}e^{tS_{k-1}}.$$

If $k = 1$ then again by Hoeffding,

$$\mathbb{E}e^{tS_1} = \mathbb{E}e^{tX_k} \leq e^{t^2c^2/2}.$$

Therefore by induction,

$$\mathbb{E}e^{tS_n} \leq \prod_{i=1}^n e^{t^2c^2/2} = e^{nt^2c^2/2}.$$

Substituting in this bound above, we find that

$$P(S_n \geq \epsilon) \leq \exp\left(nt^2c^2 - t\epsilon\right).$$

And the proof proceeds as in Chernoff-Hoeffding. □

So far we've limited our discussion to sums of random variables. We will now turn our attention to non-linear structures $f(X_1, \dots, X_n)$.

Theorem. *(McDiarmid) Suppose X_1^n are independent, $f : \mathbb{R}^n \to \mathbb{R}$ and for all $x_1^n \in \mathbb{R}^n$,*

$$\sup_{x_i^*} |f(x_1^n) - f(x_1, \dots, x_{i-1}, x_i^*, x_{i+1}, \dots, x_n)| \leq c.$$

Then $f(X_1^n)$ converges to its expectation and

$$P\left(f(X_1^n) - \mathbb{E}f(X_1^n) \geq \epsilon\right) \leq \exp\left(-\frac{2\epsilon^2}{nc^2}\right).$$

Proof. Consider the Doob martingale (B_i) defined by $B_0 = \mathbb{E}f(X_1^n)$ and

$$B_i = \mathbb{E}\left(f(X_1^n)|X_1^i\right).$$

Then $M_i = B_i - B_{i-1}$ is a martingale difference sequence and

$$\sum_{i=1}^n M_i = f(X_1^n) - \mathbb{E}f(X_1^n).$$

Let $X_i' \sim X_i$, X_i' independent of X_i, and define $X_1^{n'} \equiv (X_1, \dots, X_i', \dots, X_n)$. Because X_i', X_{i+1}^n are independent of X_1^i,

$$B_{i+1} - B_i = \mathbb{E}\left(f(X_1^n)|X_1^i\right) - \mathbb{E}\left(f(X_1^n)|X_1^{i-1}\right)$$
$$= \mathbb{E}\left(f(X_1^n)|X_1^i\right) - \mathbb{E}\left(f(X_1^{n'})|X_1^i\right) = \mathbb{E}\left(f(X_1^n) - f(X_1^{n'})|X_1^i\right).$$

It follows by our hypothesis on f that

$$|B_{i+1} - B_i| = \left|\mathbb{E}\left(f(X_1^n) - f(X_1^{n'})|X_1^i\right)\right| \leq \mathbb{E}\left(\left|f(X_1^n) - f(X_1^{n'})\right||X_1^i\right) \leq c.$$

□