CPSC 532D — 3. CONCENTRATION INEQUALITIES

Danica J. Sutherland

University of British Columbia, Vancouver

Fall 2024

We'll now prove Hoeffding's inequality (Proposition 2.1), and learn a bunch of useful stuff along the way.

3.1 MARKOV

We'll start with the following surprisingly simple bound, which turns out to be the basis for just about everything:

Proposition 3.1 (Markov's inequality). *If* X *is a nonnegative-valued random variable, then* $\Pr(X \ge t) \le \frac{1}{t} \mathbb{E} X$ *for all* $t > 0$.

Proof. We know $X \ge 0$. We also know, if $X \ge t$, then $X \ge t$. Combining those two statements, we can write $X \ge t\, \mathbb{1}(X \ge t)$. Now take the expectation of both sides of that inequality, giving $\mathbb{E} X \ge t\, \mathbb{E}\, \mathbb{1}(X \ge t) = t \Pr(X \ge t)$. Rearrange. □

This was actually proved by Markov's PhD advisor Chebyshev. Luckily, though, Chebyshev has another inequality named after him:

Proposition 3.2 (Chebyshev's inequality). *For any* X, $\Pr(|X - \mathbb{E} X| \ge \varepsilon) \le \frac{1}{\varepsilon^2} \operatorname{Var} X$.

Proof. $(X - \mathbb{E} X)^2$ is a nonnegative random variable; applying Markov gives $\Pr((X - \mathbb{E} X)^2 \ge t) \le \frac{1}{t} \mathbb{E}(X - \mathbb{E} X)^2$. Change variables to $t = \varepsilon^2$. □

Equivalently, with probability at least $1 - \delta$, $|X - \mathbb{E} X| < \sqrt{\operatorname{Var}[X] / \delta}$.

Let's consider iid $X_1, \dots, X_m$, each with mean μ and variance σ^2. Then the random variable $\overline{X} = \frac{1}{m} \sum_{i=1}^{m} X_i$ has mean μ and variance σ^2/m, so Chebyshev gives that $\left|\overline{X} - \mu\right| \le \sigma/\sqrt{m\delta}$. This is $\mathcal{O}_p(1/\sqrt{m})$, as expected, so sometimes this is good enough.

But the dependence on δ is really quite bad compared to what we'd like. For instance, if the X_i are normal so that $\bar{X}$ is too, then in (3.2) below we'll obtain $\overline{X} - \mu \le \frac{\sigma}{\sqrt{m}} \sqrt{2 \log \frac{1}{\delta}}$. To emphasize the difference:

δ	0.1	0.01	0.001	0.0001	0.00001
$1/\sqrt{\delta}$	3.2	10.0	31.6	100.0	316.2
$\sqrt{2 \log \frac{1}{\delta}}$	2.2	3.0	3.7	4.3	4.8

Chebyshev's inequality is sharp, meaning that it can be an equality in certain cases; this happens for random variables of the form $\Pr(X = 0) = 1 - \delta$, $\Pr(X = 1/\sqrt{\delta}) = \Pr(X = -1/\sqrt{\delta}) = \frac{1}{2}\delta$. This X has mean 0 and variance 1, but it still has a big probability of being really far from zero. "Typical" random variables, like Gaussians, don't look like this. So here's another analysis that takes this into account.

For more, visit https://cs.ubc.ca/~dsuth/532D/24w1/.

3.2 CHERNOFF BOUNDS

Perhaps the most useful category of results are called Chernoff bounds; they're based on

$$\Pr(X \geq \mathbb{E}\,X + \varepsilon) = \Pr\left(e^{\lambda(X - \mathbb{E}\,X)} \geq e^{\lambda\varepsilon}\right) \leq e^{-\lambda\varepsilon}\,\mathbb{E}\,e^{\lambda(X-\mathbb{E}\,X)}, \tag{3.1}$$

where we applied Markov to the nonnegative random variable $\exp(\lambda(X - \mathbb{E}\,X))$ for any $\lambda > 0$.

The quantity $M_X(\lambda) = \mathbb{E}\,e^{\lambda(X-\mathbb{E}\,X)}$ is known as the centred *moment-generating function*; recalling that $e^t = 1 + t + \frac{t^2}{2!} + \frac{t^3}{3!} + \cdots$ and writing $\mu = \mathbb{E}\,X$, we have

$$M_X(\lambda) = \mathbb{E}\,e^{\lambda(X-\mu)} = 1 + \lambda\,\mathbb{E}[X-\mu] + \frac{\lambda^2}{2!}\,\mathbb{E}[(X-\mu)^2] + \frac{\lambda^3}{3!}\,\mathbb{E}[(X-\mu)^3] + \cdots.$$

So, taking the kth derivative of the centred mgf and then evaluating at $\lambda = 0$ gives $M_X^{(k)}(0) = \mathbb{E}[(X-\mu)^k]$.

Proposition 3.3. *If* $X \sim \mathcal{N}(\mu, \sigma^2)$, *then* $\mathbb{E}\,e^{\lambda(X-\mu)} = e^{\frac{1}{2}\lambda^2\sigma^2}$.

Proof. Let's start with $X \sim \mathcal{N}(0,1)$. We can write

$$\begin{aligned}
\mathop{\mathbb{E}}_{X\sim\mathcal{N}(0,1)} e^{\lambda X} &= \int \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} e^{\lambda x}\,\mathrm{d}x \\
&= \int \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2 + \lambda x - \frac{1}{2}\lambda^2 + \frac{1}{2}\lambda^2}\,\mathrm{d}x \\
&= e^{\frac{1}{2}\lambda^2} \int \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\lambda)^2}\,\mathrm{d}x \\
&= e^{\frac{1}{2}\lambda^2},
\end{aligned}$$

since the last integral is just the total probability density of an $\mathcal{N}(\lambda, 1)$ random variable. To handle $Y = \mathcal{N}(\mu, \sigma^2)$, note that this is equivalent to $\sigma X + \mu$, so

$$e^{\lambda(Y-\mathbb{E}\,Y)} = e^{\lambda(\sigma X + \mu - \mathbb{E}(\sigma X + \mu))} = e^{\lambda(\sigma X)} = e^{(\lambda\sigma)X} = e^{\frac{1}{2}\sigma^2\lambda^2}. \qquad \square$$

Plugging Proposition 3.3 into (3.1), for $X \sim \mathcal{N}(\mu, \sigma^2)$, it holds for any $\lambda > 0$ that

$$\Pr(X \geq \mu + \varepsilon) \leq e^{-\lambda\varepsilon} e^{\frac{1}{2}\sigma^2\lambda^2}.$$

The value of λ only appears on the right-hand side, not the left. So we might as well find the best value of λ to use: the one that gives the tightest bound. Let's optimize this in λ: noting that exp is monotonic, we can just check that $\frac{1}{2}\sigma^2\lambda^2 - \lambda\varepsilon$ has derivative $\sigma^2\lambda - \varepsilon$, which is zero when $\lambda = \varepsilon/\sigma^2 > 0$. (And this is indeed a max, since the second derivative is $\sigma^2 > 0$.) Plugging in that value of λ, we get the bound

$$\Pr(X \geq \mu + \varepsilon) \leq \exp\left(-\tfrac{\varepsilon^2}{2\sigma^2}\right). \tag{3.2}$$

Equivalently, $X < \mu + \sigma\sqrt{2\log\frac{1}{\delta}}$ with probability at least $1 - \delta$.

3.3 SUBGAUSSIAN VARIABLES

In fact, the only place we used the Gaussian assumption in this argument was in that $\mathbb{E}\,e^{\lambda(X-\mathbb{E}\,X)} \leq e^{\frac{1}{2}\lambda^2\sigma^2}$. So we can generalize the result to anything satisfying that condition, which we call *subgaussian*:

Definition 3.4. A random variable X with mean $\mu = \mathbb{E}[X]$ is called *subgaussian with parameter* $\sigma \geq 0$, written $X \in \mathcal{SG}(\sigma)$, if its centred moment-generating function $\mathbb{E}[e^{\lambda(X-\mu)}]$ exists and satisfies that for all $\lambda \in \mathbb{R}$, $\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\frac{1}{2}\lambda^2\sigma^2}$.

Watch out with other sources; notation for subgaussians is not very standardized, in particular whether the parameter is σ or σ^2. Also "$X \in \mathcal{SG}(\sigma)$" is kind of weird; probably "$\mathrm{Law}(X) \in \mathcal{SG}(\sigma)$" would be better, but oh well.

As we just saw, normal variables with variance σ^2 are $\mathcal{SG}(\sigma)$. Notice also that if $\sigma_1 < \sigma_2$, then anything that's $\mathcal{SG}(\sigma_1)$ is also $\mathcal{SG}(\sigma_2)$.

Proposition 3.5 (Hoeffding's lemma). *If* $\Pr(a \leq X \leq b) = 1$, X *is* $\mathcal{SG}\left(\frac{b-a}{2}\right)$.

Proof. See Section 3.3.1; we'll probably skip this in class. □

Here are some useful properties about building subgaussian variables:

Proposition 3.6. *If* $X_1 \in \mathcal{SG}(\sigma_1)$ *and* $X_2 \in \mathcal{SG}(\sigma_2)$ *are independent random variables, then* $X_1 + X_2 \in \mathcal{SG}(\sqrt{\sigma_1^2 + \sigma_2^2})$.

Proof. $\mathbb{E}[e^{\lambda(X_1+X_2-\mathbb{E}[X_1+X_2])}] = \mathbb{E}[e^{\lambda(X_1-\mathbb{E} X_1)}]\,\mathbb{E}[e^{\lambda(X_2-\mathbb{E} X_2)}]$ by independence. Bounding each expectation, this is at most $e^{\frac{1}{2}\lambda^2\sigma_1^2}\, e^{\frac{1}{2}\lambda^2\sigma_2^2} = e^{\frac{1}{2}\lambda^2\left(\sqrt{\sigma_1^2+\sigma_2^2}\right)^2}$. □

Proposition 3.7. *If* $X \in \mathcal{SG}(\sigma)$, *then* $aX + b \in \mathcal{SG}(|a|\,\sigma)$ *for any* $a, b \in \mathbb{R}$.

Proof.
$$\mathbb{E}[e^{\lambda(aX+b-\mathbb{E}[aX+b])}] = \mathbb{E}[e^{(a\lambda)(X-\mathbb{E} X)}] \leq e^{\frac{1}{2}(a\lambda)^2\sigma^2} = e^{\frac{1}{2}\lambda^2(|a|\sigma)^2}.$$
□

Proposition 3.8 (Chernoff bound for subgaussians). *If* $X \in \mathcal{SG}(\sigma)$, *then* $\Pr(X \geq \mathbb{E} X + \varepsilon) \leq \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right)$ *for* $\varepsilon \geq 0$.

Proof. Exactly as the argument leading from (3.1) to (3.2). □

Since $-X$ is also $\mathcal{SG}(\sigma)$ by Proposition 3.7, the same bound holds for a lower deviation $\Pr(X \leq \mathbb{E} X - t)$. A union bound then immediately gives $\Pr(\left|X - \mu\right| \geq t) \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right)$.

Proposition 3.9 (Hoeffding). *If* $X_1, \dots, X_m$ *are independent and each* $\mathcal{SG}(\sigma_i)$ *with mean* μ_i, *for all* $\varepsilon \geq 0$

$$\Pr\left(\frac{1}{m}\sum_{i=1}^m X_i \geq \frac{1}{m}\sum_{i=1}^m \mu_i + \varepsilon\right) \leq \exp\left(-\frac{m^2\varepsilon^2}{2\sum_{i=1}^m \sigma_i^2}\right).$$

Proof. By Propositions 3.6 and 3.7, $\frac{1}{m}\sum_{i=1}^m X_i \in \mathcal{SG}\left(\frac{1}{m}\sqrt{\sum_{i=1}^m \sigma_i^2}\right)$. Then apply Proposition 3.8. □

If the X_i have the same mean $\mu_i = \mu$ and parameter $\sigma_i = \sigma$, this becomes

$$\Pr\left(\frac{1}{m}\sum_{i=1}^m X_i \geq \mu + \varepsilon\right) \leq \exp\left(-\frac{m\varepsilon^2}{2\sigma^2}\right), \qquad \text{(Hoeffding)}$$

which can also be stated as that, with probability at least $1 - \delta$,

$$\frac{1}{m}\sum_{i=1}^{m} X_i < \mu + \sigma\sqrt{\frac{2}{m}\log\frac{1}{\delta}}. \qquad \text{(Hoeffding')}$$

The form of Hoeffding we saw before, Proposition 2.1, follows immediately from Proposition 3.5 and (Hoeffding').

3.3.1 *Proof of Hoeffding's lemma*

Wikipedia's proof is similar, but I think a little less clean. Other proofs are based more explicitly on convexity, but use either opaque changes of variable [SSBD14, Lemma B.7] or compute some pretty nasty derivatives [MRT18, Lemma D.1]. There's also a proof strategy based on "exponential tilting" (see [BLM13, Lemma 2.2], [Rag14, Lemma 1], or [Wai19, Exercise 2.4]) which is quite related but just overall a little more annoying. There are also proofs based on symmetrization (see [Wai19, Examples 2.3-2.4] or [Rom21]), which are nice but (a) have a worse constant and (b) require symmetrization, which is an important idea we'll cover soon but kind of hard to understand.

This proof roughly follows Zhang [Zhang23, Lemma 2.15].

Lemma 3.10. *Let* $X \sim \text{Bernoulli}(p)$. *Then* X *is* $\mathcal{SG}(1/2)$.

Proof. The logarithm of the (uncentred) moment-generating function of X is

$$\psi(\lambda) = \log \mathbb{E}\, e^{\lambda X} = \log\big((1-p)e^0 + pe^\lambda\big).$$

This has derivatives

$$\psi'(\lambda) = \frac{pe^\lambda}{(1-p)e^0 + pe^\lambda}$$

$$\psi''(\lambda) = \frac{pe^\lambda}{(1-p)e^0 + pe^\lambda} - \frac{(pe^\lambda)^2}{\big((1-p)e^0 + pe^\lambda\big)^2} = \psi'(\lambda)(1-\psi'(\lambda)).$$

Since the function $x(1-x)$ has maximum $1/4$, $\psi''(\lambda) \le 1/4$. By Taylor's theorem (in the Lagrange form), for any λ there exists some ξ_λ such that

$$\psi(\lambda) = \underbrace{\psi(0)}_{0} + \lambda\underbrace{\psi'(0)}_{p} + \frac{1}{2}\lambda^2\underbrace{\psi''(\xi_\lambda)}_{\le 1/4} \le \lambda p + \frac{1}{8}\lambda^2.$$

Thus the centred mgf satisfies

$$\mathbb{E}\, e^{\lambda(X - \mathbb{E}X)} = e^{-\lambda p}\,\mathbb{E}\, e^{\lambda X} \le e^{-\lambda p}\left(e^{\lambda p + \frac{1}{8}\lambda^2}\right) = e^{\frac{1}{8}\lambda^2}.$$ □

Proposition 3.5 (Hoeffding's lemma)**.** *If* $\Pr(a \le X \le b) = 1$, X *is* $\mathcal{SG}\left(\frac{b-a}{2}\right)$.

Proof. Using $(X-a)/(b-a)$ and Proposition 3.7, we need only consider $a = 0$, $b = 1$.

Let $f(\lambda) = \mathbb{E}\, e^{\lambda X}$ be the (uncentred) mgf of X, and $g(\lambda) = (1-\mu)e^0 + \mu e^\lambda$ that of a Bernoulli(μ) variable, where $\mu = \mathbb{E}\, X$. For $\lambda \ge 0$,

You can interchange this derivative and expectation, but it's trickier to prove than usual, requiring e.g. Theorem 3 here.

$$f'(\lambda) = \frac{d}{d\lambda}\mathbb{E}[e^{\lambda X}] = \mathbb{E}\left[\frac{d}{d\lambda}e^{\lambda X}\right] = \mathbb{E}[Xe^{\lambda X}] \le \mathbb{E}[Xe^\lambda] = \mu e^\lambda = g'(\lambda),$$

using in the inequality that $\lambda \ge 0$ and $0 \le X \le 1$. and that $0 \le X \le 1$. The same steps give $f'(\lambda) \ge g'(\lambda)$ for $\lambda \le 0$. As $f(0) = 1 = g(0)$, it follows that $f(\lambda) \le g(\lambda)$ everywhere. The conclusion follows by Lemma 3.10. □

References

[BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[MRT18] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talkwalkar. *Foundations of Machine Learning*. 2nd edition. MIT Press, 2018.

[Rag14] Maxim Raginsky. *Concentration inequalities*. September 2014.

[Rom21] Marc Romaní. *A short proof of Hoeffding's lemma*. May 1, 2021.

[SSBD14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[Wai19] Martin Wainwright. *High-dimensional statistics: a non-asymptotic viewpoint*. Cambridge University Press, 2019.

[Zhang23] Tong Zhang. *Mathematical Analysis of Machine Learning Algorithms*. Pre-publication version. 2023.