CPSC 532D — 3. CONCENTRATION INEQUALITIES

Danica J. Sutherland
University of British Columbia, Vancouver
Fall 2025

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* $\Pr(X \geq t) \leq \frac{1}{t}\mathbb{E}\,X$ *for all* $t > 0$.

Proof. We know $X \geq 0$. We also know, if $X \geq t$, then $X \geq t$. Combining those two statements, we can write $X \geq t\,\mathbb{1}(X \geq t)$. Now take the expectation of both sides of that inequality, giving $\mathbb{E}\,X \geq t\,\mathbb{E}\,\mathbb{1}(X \geq t) = t\Pr(X \geq t)$. 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, $\Pr(|X - \mathbb{E}\,X| \geq \varepsilon) \leq \frac{1}{\varepsilon^2}\operatorname{Var} X$.

Proof. $(X - \mathbb{E}\,X)^2$ is a nonnegative random variable; applying Markov gives $\Pr((X - \mathbb{E}\,X)^2 \geq t) \leq \frac{1}{t}\mathbb{E}(X - \mathbb{E}\,X)^2$. 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 $|\overline{X} - \mu| \leq \sigma/\sqrt{m\delta}$. 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. For instance, if the X_i are normal, $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/m)$ and so the $1 - \delta$ probability range on the deviation is actually $|\bar{X} - \mu| \leq \frac{\sigma}{\sqrt{m}} \cdot \Phi^{-1}(1 - \frac{\delta}{2})$, where Φ is the standard normal CDF. To compare the true value to the bound for some choices of δ:

δ	0.1	0.01	0.001	0.0001	0.00001
$\Phi^{-1}(1 - \frac{\delta}{2})$	1.6	2.6	3.3	3.9	4.4
$1/\sqrt{\delta}$	3.2	10.0	31.6	100.0	316.2
$\sqrt{2\log\frac{2}{\delta}}$	2.4	3.3	3.9	4.5	4.9

The last line, $\sqrt{2\log\frac{2}{\delta}}$, is the (*much* tighter) bound we'll obtain from (3.2) below.

Chebyshev's inequality is sharp, meaning that it can be an equality in certain cases; this happens for random variables of the form $\Pr(X = 0) = 1 - \delta$, $\Pr(X = 1/\sqrt{\delta}) =$

For more, visit https://cs.ubc.ca/~dsuth/532D/25w1/.

$\Pr(X = -1/\sqrt{\delta}) = \frac{1}{2}\delta$. 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 will make stronger assumptions than just the mean and variance to take this "typicality" into account.

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 $Y \sim \mathcal{N}(\mu, \sigma^2)$, 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 $1 - \delta$.

In fact, the only place we used the Gaussian assumption in this argument was in that $\mathbb{E}\, e^{\lambda(X-\mathbb{E}\,X)} \le e^{\frac{1}{2}\lambda^2\sigma^2}$. 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* $\sigma \ge 0$, abbreviated $\mathcal{SG}(\sigma)$, if its centred moment-generating function $\mathbb{E}[e^{\lambda(X-\mu)}]$ exists and satisfies that for all $\lambda \in \mathbb{R}$, $\mathbb{E}[e^{\lambda(X-\mu)}] \le e^{\frac{1}{2}\lambda^2\sigma^2}$.

Watch out with other sources; notation for subgaussians is not very standardized, in particular whether the parameter is σ or σ^2. (It's not even standardized whether it's written "subgaussian," "sub-Gaussian," or many other variants.)

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 $\mathcal{SG}(\sigma_2)$.

The final key "base" result we'll need is that any bounded variable is subgaussian:

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. 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 *is* $\mathcal{SG}(\sigma_1)$ *and* X_2 *is* $\mathcal{SG}(\sigma_2)$*, and the two are independent, then* $X_1 + X_2$ *is* $\mathcal{SG}\left(\sqrt{\sigma_1^2 + \sigma_2^2}\right)$.

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

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

Proposition 3.8 (Chernoff bound for subgaussians). *If* X *is* $\mathcal{SG}(\sigma)$*, then for any* $\varepsilon \ge 0$, $\Pr(X \ge \mathbb{E}\,X + \varepsilon) \le \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right)$*, or equivalently* $\Pr\left(X < \mathbb{E}\,X + \sigma\sqrt{2\log\frac{1}{\delta}}\right) > 1 - \delta$.

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 $\Pr(X \le \mathbb{E}\,X - \varepsilon)$. A union bound then immediately gives $\Pr(|X-\mu| \ge \varepsilon) \le 2\exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right)$.

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

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

Proof. By Propositions 3.6 and 3.7, $\frac{1}{m}\sum_{i=1}^{m} X_i$ is $\mathcal{SG}\left(\frac{1}{\sqrt{m}}\sqrt{\frac{1}{m}\sum_{i=1}^{m}\sigma_i^2}\right)$. The result then follows by applying Proposition 3.8. □

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

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

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

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

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*

This proof roughly follows Zhang [Zhang23, Lemma 2.15], but rearranged to be (I think) clearer. We start with a useful special case.

Wikipedia's proof is similar, but I think a little less clean. Other proofs are based more explicitly on convexity, but either don't get the tight constant [Har23, Section 21.3.1], use extremely 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.

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

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

This has derivatives

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

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

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

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

Thus the centred mgf satisfies

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

□

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. 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 $\mu = \mathbb{E}\, X$. First notice that

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

You can interchange this derivative and expectation just fine, since $e^{\lambda X}$ is continuously differentiable.

For $\lambda \geq 0$, recalling $0 \leq X \leq 1$ gives us that $f'(\lambda) \leq \mathbb{E}[Xe^\lambda] = \mu e^\lambda = g'(\lambda)$. Similarly, for $\lambda \leq 0$, we have $f'(\lambda) \geq \mathbb{E}[Xe^\lambda] = g'(\lambda)$. As $f(0) = 1 = g(0)$, it follows that $f(\lambda) \leq g(\lambda)$ 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.

[Har23] Nick Harvey. *A second course in randomized algorithms*. March 12, 2023.

[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.