CPSC 532D — 3. CONCENTRATION INEQUALITIES

Danica J. Sutherland
University of British Columbia, Vancouver
Fall 2023

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

1 MARKOV

We'll start with the surprisingly simple basis for everything:

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

Proof. We know $X \geq 0$. We also know, if $X \geq t$, then $X \geq t$. So $X \geq t\,\mathbb{1}(X \geq t)$. Take the expectation of both sides, 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 2 (Chebyshev's inequality). *For any* X, $\Pr(|X - \mathbb{E}\,X| \geq \varepsilon) \leq \frac{\operatorname{Var} X}{\varepsilon^2}$.

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 $1 - \delta$, $|X - \mathbb{E}\,X| \leq \sqrt{\operatorname{Var}[X]/\delta}$.

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 $\left|\overline{X} - \mu\right| \leq \frac{\sigma}{\sqrt{m\delta}}$. 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 (2) below we'll obtain $\overline{X} - \mu \leq \frac{\sigma}{\sqrt{m}}\sqrt{2\log\frac{1}{\delta}}$. 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, but 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/23w1/.

2 Chernoff bounds

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

$$\Pr(X \geq \mathbb{E}\,X + \varepsilon) = \Pr\left(e^{\lambda(X-\mathbb{E}\,X)} \geq e^{\lambda\varepsilon}\right) \leq e^{-\lambda\varepsilon}\,\mathbb{E}\,e^{\lambda(X-\mathbb{E}\,X)}, \tag{1}$$

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 $e^t = 1 + t + \frac{t^2}{2!} + \frac{t^3}{3!} + \cdots$, 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. *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 $\mathcal{N}(0,1)$. 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 $Y = \mathcal{N}(0, \sigma^2)$, note that this is equivalent to σX, and

$$e^{\lambda Y} = e^{\lambda(\sigma X)} = e^{(\sigma\lambda)X} = e^{\frac{1}{2}\sigma^2\lambda^2}.$$

We just subtract off the mean anyway, so allowing $\mu \neq 0$ is immediate. □

Plugging Proposition 3 into (1) gives that for any $\lambda > 0$,

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

This holds for any λ, but we'd like the tightest bound, so 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 $\lambda = \varepsilon/\sigma^2 > 0$, giving the bound

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

or equivalently $X \leq \mu + \sigma\sqrt{2\log\frac{1}{\delta}}$ with probability at least $1 - \delta$.

3 Sub-Gaussian 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 *sub-Gaussian*:

DEFINITION 4. A random variable X with mean $\mu = \mathbb{E}[X]$ is called *sub-Gaussian with parameter* σ, written $X \in \mathcal{SG}(\sigma)$, if its centred moment-generating function $\mathbb{E}[e^{\lambda(X-\mu)}]$ exists and satisfies that for all $\lambda \in \mathbb{R}$, $\mathbb{E}[e^{\lambda(X-\mu)}] \le e^{\frac{1}{2}\lambda^2\sigma^2}$.

Watch out with other sources; notation for sub-Gaussians 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 5 (Hoeffding's lemma). *A real-valued random variable bounded in $[a, b]$ is $\mathcal{SG}\left(\frac{b-a}{2}\right)$.*

Proof. See Section 3.1; we'll probably skip this in class. □

Here are some useful properties about building sub-Gaussian variables:

PROPOSITION 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. $\mathbb{E}[e^{\lambda(X_1+X_2-\mathbb{E}[X_1+X_2])}] = \mathbb{E}[e^{\lambda(X_1-\mathbb{E}X_1)}]\,\mathbb{E}[e^{\lambda(X_2-\mathbb{E}X_2)}]$. Bounding each with the definition, 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 7. *If $X \in \mathcal{SG}(\sigma)$, then $aX \in \mathcal{SG}(|a|\,\sigma)$ for any $a \in \mathbb{R}$.*

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

PROPOSITION 8 (Chernoff bound for sub-Gaussians). *If $X \in \mathcal{SG}(\sigma)$, then $\Pr(X \ge \mathbb{E}X + \varepsilon) \le \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right)$ for $\varepsilon \ge 0$.*

Proof. Exactly as the argument leading from (1) to (2). □

Since $-X$ is also $\mathcal{SG}(\sigma)$ by Proposition 7, the same bound holds for a lower deviation $\Pr(X \le \mathbb{E}X - t)$. A union bound then immediately gives $\Pr(\left|X - \mu\right| \ge t) \le 2\exp\left(-\frac{t^2}{2\sigma^2}\right)$.

PROPOSITION 9 (Hoeffding). *If $X_1, \dots, X_m$ are independent and each $\mathcal{SG}(\sigma_i)$ with mean μ_i, for all $\varepsilon \ge 0$*

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

Proof. By Propositions 6 and 7, $\frac{1}{m}\sum_{i=1}^{m} X_i \in \mathcal{SG}\left(\frac{1}{m}\sqrt{\sum_{i=1}^{n}\sigma_i^2}\right)$. Then apply Proposition 8. □

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

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

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

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

The form of Hoeffding's inequality stated in the last lecture follows immediately from Proposition 5 and (Hoeffding'):

proposition 10. *If $X_1, \dots, X_m$ are independent with mean μ and each a.s. bounded in (a, b), then with probability at least $1 - \delta$,*

$$\frac{1}{m}\sum_{i=1}^{m} X_i \le \mu + (b-a)\sqrt{\frac{1}{2m}\log\frac{1}{\delta}}.$$

3.1 *Proof of Hoeffding's Lemma*

I don't know a fully satisfying proof for this lemma that I feel is totally worth teaching.

The convexity-based proof in Section 3.2 is straightforward to follow, but I don't find it very insightful.

The exponential tilting argument of Section 3.3 has a couple of steps that take some verification; it's *slightly* more insightful.

It's also possible to show $\mathcal{SG}(b-a)$ with a nice symmetrization-type argument, as in Examples 2.3 and 2.4 of Wainwright [Wai19], but this is still a little longer than I'd like and not tight. It also requires understanding symmetrization, which is important and we'll cover it in class soon, but not obvious the first time you see it.

Romaní [Rom21] gives an "in-between" argument showing $\mathcal{SG}\left(\frac{b-a}{\sqrt{2}}\right)$. This uses aspects of all three proofs, but quickly, and it might be the best single one to read, but I still don't really love it (plus it's not tight).

3.2 *Convexity*

This proof follows Lemma B.7 of Shalev-Shwartz and Ben-David [SSBD], which is essentially the same as Lemma D.1 of Mohri, Rostamizadeh, and Talkwalkar [MRT]. This proof is straightforward but relies on either a clever but totally opaque change of variables (as below) or computing some pretty-unpleasant derivatives (as in [MRT]). (Lemma 2.15 of [Zhang23] might be better?) It's also not clear to me that it provides any actual insight.

proposition 5 (Hoeffding's lemma). *A real-valued random variable bounded in $[a, b]$ is $\mathcal{SG}\left(\frac{b-a}{2}\right)$.*

Proof. Assume WLOG that $\mathbb{E}\, X = 0$.

First note that $f_\lambda(x) = e^{\lambda x}$ is a convex function for any λ, e.g. $f_\lambda''(x) = \lambda^2 e^{\lambda x} > 0$ for all x. This implies that it lies below its chords, i.e. for $\alpha \in [0, 1]$

$$f_\lambda(\alpha a + (1-\alpha)b) \le \alpha f_\lambda(a) + (1-\alpha) f_\lambda(b).$$

We can rewrite this with $x = \alpha a + (1-\alpha)b$, $\alpha = \frac{b-x}{b-a}$ so that

$$e^{\lambda x} \le \frac{b-x}{b-a} e^{\lambda a} + \frac{x-a}{b-a} e^{\lambda b}.$$

Taking an expectation over x, we get

$$\mathbb{E}\, e^{\lambda X} \le \frac{b - \mathbb{E}\, X}{b-a} e^{\lambda a} + \frac{\mathbb{E}\, X - a}{b-a} e^{\lambda b} = \frac{b}{b-a} e^{\lambda a} - \frac{a}{b-a} e^{\lambda b},$$

using that $\mathbb{E}\, X = 0$.

Now, do a very opaque change of variables to $h = \lambda(b-a)$, $p = -\frac{a}{b-a}$, $L(h) = -hp + \log(1 - p + pe^h)$. If we write

$$e^{L(h)} = e^{-hp}(1 - p + pe^h) = e^{\lambda a}\left(1 + \frac{a}{b-a} - \frac{a}{b-a} e^{\lambda(b-a)}\right) = \frac{b}{b-a} e^{\lambda a} - \frac{a}{b-a} e^{\lambda b},$$

we've recovered the right-hand side above. So to prove the lemma, it remains to show that $L(h) \le \frac{1}{8} h^2$.

Notice that $L(0) = 0 + \log(1 - p + p) = 0$.

Also, $L'(h) = -p + \frac{pe^h}{1-p+pe^h}$, so $L'(0) = -p + \frac{p}{1-p+p} = 0$.

Finally, $L''(h) = \frac{pe^h}{1-p+pe^h} - \left(\frac{pe^h}{1-p+pe^h}\right)^2 = t(1-t)$, letting $t = \frac{pe^h}{1-p+pe^h}$. Because $\mathbb{E}\, X = 0$, we must have $a \le 0 \le b$, and so $p \in [0,1]$. Thus $1 - p \in [0,1]$, $pe^h \ge 0$, and so $t = \frac{pe^h}{1-p+pe^h} \in [0,1]$. Thus $0 \le t(1-t) \le \frac{1}{4}$, so $0 \le L''(h) \le \frac{1}{4}$ for all h.

Using Taylor's theorem, this means at last that for any h, there is some h' such that $L(h) = L(0) + L'(0)h + \frac{1}{2}L''(h')h^2 = L''(h') \le \frac{1}{2}\frac{1}{4}h^2$. Thus $L(h) \le \frac{1}{8}h^2$, and $\mathbb{E}\, e^{\lambda X} \le e^{L(\lambda(b-a))} \le e^{\frac{1}{2}\lambda^2\left(\frac{b-a}{2}\right)^2}$. □

3.3 *Exponential tilting*

This proof uses an "exponential tilting" argument, as in Lemma 2.2 of Boucheron, Lugosi, and Massart [BLM13] or Lemma 1 of Raginsky [Rag14]. It's tight, but it requires a few details to be fully rigorous (which neither of these sources do out), some of which aren't totally obvious. It's also not super-intuitive; the variable transformation here is related to the mgf, but it seems to rely on basically a coincidence where I'm not sure if there's a deeper meaning.

We'll need the following lemma, which might be of independent interest.

LEMMA 11. *Suppose* X *is a.s. bounded in* $[a,b]$. *Then* $\operatorname{Var} X \le \frac{(b-a)^2}{4}$.

Proof. Note that

$$\begin{aligned}
\mathbb{E}(X-c)^2 &= \mathbb{E}\left((X-\mu) + (\mu - c)\right)^2 \\
&= \mathbb{E}(X-\mu)^2 + (\mu-c)^2 + 2(\mu-c)\underbrace{\mathbb{E}(X-\mu)}_{0} \\
&= \operatorname{Var} X + (\mu - c)^2.
\end{aligned}$$

Thus $\operatorname{Var} X = \mathbb{E}(X-c)^2 - (\mu-c)^2 \le \mathbb{E}(X-c)^2$ for any c. In particular, take $c = (a+b)/2$. Since $X \in [a,b]$, we know that $|X-c| \le (b-a)/2$, and so

$$\operatorname{Var} X \le \mathbb{E}(X-c)^2 \le \frac{(b-a)^2}{4}. \qquad \square$$

This bound is tight for a random variable which takes value a half the time and value b the other half; you can check this with direct evaluation, or note that for this random variable both inequalities in the proof are exactly equalities.

proposition 5 (Hoeffding's lemma). *A real-valued random variable bounded in $[a,b]$ is $\mathcal{SG}\left(\frac{b-a}{2}\right)$.*

Proof. Assume WLOG that $\mathbb{E}\, X = 0$.

Define an auxiliary random variable Y_λ with the same support as X such that for any A, $\Pr(Y_\lambda \in A) = \mathbb{E}[\mathbb{1}(x \in A)e^{\lambda X}]/\mathbb{E}[e^{\lambda X}]$. If X has a density, then Y_λ's density is just proportional to multiplying X's density by $e^{\lambda x}$. This is known as exponential tilting.

It follows that for any function f, $\mathbb{E}[f(Y_\lambda)] = \mathbb{E}[f(X)e^{\lambda X}]/\mathbb{E}[e^{\lambda X}]$. (Write f as a [limit of] linear combinations of indicators of events, known as *simple functions*.) We then have that

$$\mathbb{E}[Y_\lambda] = \frac{\mathbb{E}[Xe^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} \qquad \mathbb{E}[Y_\lambda^2] = \frac{\mathbb{E}[X^2e^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]}.$$

Here I'm skipping some details about proving that you can interchange the derivative and expectation; you can, but it's actually slightly trickier to prove than usual, requiring e.g. Theorem 3 here.

Define $\psi(\lambda) = \log \mathbb{E}[e^{\lambda(X-\mathbb{E}X)}]$ to be the log-mgf of X. Now take derivatives:

$$\psi'(\lambda) = \frac{\mathbb{E}[Xe^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} = \mathbb{E}[Y_\lambda]$$

$$\psi''(\lambda) = \frac{\mathbb{E}[X^2e^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} - \frac{\mathbb{E}[Xe^{\lambda X}]\mathbb{E}[Xe^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]^2} = \mathbb{E}[Y_\lambda^2] - \mathbb{E}[Y_\lambda]^2 = \operatorname{Var}[Y_\lambda].$$

Since $Y_\lambda \in [a,b]$ almost surely, this means that $\psi''(\lambda) \le \frac{(b-a)^2}{4}$ for any λ.

Notice that $\psi(0) = \log \mathbb{E}\, e^0 = 0$ and $\psi'(0) = \frac{\mathbb{E}[Xe^0]}{\mathbb{E}\, e^0} = \mathbb{E}\, X$, which we assumed was 0. Thus, we have that

$$\psi(\lambda) = \psi(0) + \int_0^\lambda \psi'(s)\,ds = \psi(0) + \int_0^\lambda \left[\psi''(0) + \int_0^s \psi''(t)dt\right] ds$$

$$= \int_0^\lambda \int_0^s \psi''(t)\,dt\,ds$$

$$\le \frac{(b-a)^2}{4} \int_0^\lambda \int_0^s 1\,dt\,ds = \frac{(b-a)^2}{4} \int_0^\lambda s\,ds = \frac{(b-a)^2}{4} \times \frac{1}{2}\lambda^2.$$

Taking the exponential of both sides, we've shown as desired that

$$\mathbb{E}[e^{\lambda(X-\mathbb{E}X)}] \le e^{\frac{1}{2}\lambda^2\left(\frac{b-a}{2}\right)^2}. \qquad \square$$

REFERENCES

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

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

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

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

[SSBD] 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*. 2023 pre-publication version.