Deck 5: Some uses of scalar concentration inequalities

Math 7870: Topics in Randomized Numerical Linear Algebra

Spring 2026

Akil Narayan

Example 1: Randomized trace estimators

Let $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ be symmetric. Our goal is to estimate the trace of $\boldsymbol{A}$,

$$\operatorname{tr}(\boldsymbol{A}) = \sum_{j \in [n]} A_{j,j} = \sum_{j \in [n]} \lambda_j(\boldsymbol{A}).$$

The idea for accomplishing this fairly transparent: the trace is a sum of a symmetric set of entries of a matrix.

So consider a centered random vector $\boldsymbol{x} \in \mathbb{R}^n$, i.e., $\mathbb{E}\boldsymbol{x} = \boldsymbol{0}$. Then,

$$\mathbb{E}\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} = \sum_{i,j \in [n]} A_{i,j} \mathbb{E}(x_i x_j) = \sum_{i,j \in [n]} \left(\operatorname{cov}(\boldsymbol{x}) \odot \boldsymbol{A}\right)_{i,j}.$$

Therefore, the mean of $\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x}$ is the sum of $\boldsymbol{A}$-elementwise-weighted covariance of $\boldsymbol{x}$.

To recover the trace, we need a special type of random vector.

A random vector $\boldsymbol{x}$ is *isotropic* if $\mathbb{E}\boldsymbol{x}\boldsymbol{x}^T = \boldsymbol{I}$.

The (Hutchinson) randomized trace estimator

Here's the ("Hutchinson") randomized trace estimation procedure:

Input: N, an isotropic distribution for $\boldsymbol{x}$, and $\boldsymbol{A}$

0. Initialize $Z_0 = 0$, $j = 0$.
1. Generate an isotropic random vector $\boldsymbol{x}$.
2. Compute $Z_0 \leftarrow Z_0 + \frac{1}{N}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}$.
3. Set $j \leftarrow j + 1$.
4. If $j < N$, return to step 1. Otherwise, return Z_0.

NB: We need not have $\boldsymbol{A}$, but instead just access to operator queries $\boldsymbol{x} \mapsto \boldsymbol{A}\boldsymbol{x}$.

Let Z be the random variable $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}$. If Z_j, $j \geqslant 1$ are iid copies of Z, the Hutchinson trace estimator produces,

$$Z_0 = \frac{1}{N}\sum_{j\in[N]} Z_j, \quad \mathbb{E}Z_0 = \mathbb{E}Z = \sum_{i,j}(\mathbb{E}\boldsymbol{x}\boldsymbol{x}^T \odot \boldsymbol{A})_{i,j} = \sum_{i,j}(\boldsymbol{I} \odot \boldsymbol{A})_{i,j} = \mathrm{tr}(\boldsymbol{A}), \quad \mathrm{Var} Z_0 = \frac{1}{N}\mathrm{Var} Z.$$

Isotropic distribution examples

There are two common choices for isotropic distributions:

- $\boldsymbol{x} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n)$. Then $\mathrm{Var} Z = 2\|\boldsymbol{A}\|_F^2$.
- $\boldsymbol{x}$ has iid Rademacher entries: $x_j \sim \mathrm{Unif}(\{-1, 1\})$. Then $\mathrm{Var} Z = 4\sum_{i<j} |A_{i,j}|^2$.

To see the Rademacher result: Suppose $\boldsymbol{x}$ has iid Rademacher entries. Then:

$$
\begin{aligned}
\mathrm{Var} Z &= \mathbb{E}\left[\left(\sum_{i,j}(x_i x_j) A_{i,j}\right)^2\right] - \mathbb{E} Z^2 \\
&= \mathbb{E}\left[\sum_{i,j,k,\ell} x_i x_j x_k x_\ell A_{i,j} A_{k,\ell}\right] - \left(\sum_j A_{j,j}\right)^2 \\
&= \mathbb{E}\left[\sum_i x_i^4 A_{i,i}^2 + \sum_{i \neq k} x_i^2 x_j^2 A_{i,i} A_{k,k} + \sum_{i \neq j} x_i^2 x_j^2 A_{i,j}^2 + \sum_{i \neq j} x_i^2 x_j^2 A_{i,j}^2\right] - \sum_i A_{i,i}^2 - \sum_{i \neq j} A_{i,i} A_{j,j} \\
&= 2\sum_{i \neq j} A_{i,j}^2.
\end{aligned}
$$

Using moments

These actually provide significant intuition about when trace estimators succeed. Consider the *intrinsic dimension* of an spd matrix $\boldsymbol{A}$:

$$\mathrm{idim}(\boldsymbol{A}) = \frac{\mathrm{tr}(\boldsymbol{A})}{\|\boldsymbol{A}\|_2} = \frac{\sum_{j \in [n]} \lambda_j(\boldsymbol{A})}{\max_{j \in [n]} \lambda_j(\boldsymbol{A})} \in [1, n]$$

The intrinsic dimension is one way to define a "continuous" version of rank/dimension of a matrix.

Chaining together the previous results and using $\|\boldsymbol{A}\|_F^2 \leqslant \|\boldsymbol{A}\|_2 \mathrm{tr}(\boldsymbol{A})$, we conclude for either distribution above and Chebyshev's inequality:

$$\Pr\left(|Z_0 - \mathrm{tr}(\boldsymbol{A})| > \epsilon \mathrm{tr}(\boldsymbol{A})\right) \leqslant \frac{2}{N\epsilon^2 \mathrm{idim}(\boldsymbol{A})}.$$

I.e., this method works well for matrices with large intinsic dimension.

Concentration

Of course, we know we can do better than Chebyshev's inequality!

Given $\boldsymbol{A}$:

$$|Z - \mathbb{E}Z| \leqslant \sum_{i \neq j} |A_{i,j}| \quad \text{wp1}.$$

So $Z - \mathbb{E}Z$ is a bounded random variable.

Applying Hoeffding's inequality for bounded random variables: Taking N samples, we have

$$\Pr\left(|Z_0 - \mathrm{tr}(\boldsymbol{A})| \gtrsim \epsilon \mathrm{tr}(\boldsymbol{A})\right) \leqslant 2 \exp\left(-\frac{N\epsilon^2}{\mathrm{Var} Z}\right),$$

I.e., if $N \gtrsim r \log N \frac{\mathrm{Var} Z}{\epsilon^2}$, then this succeeds with probability $\gtrsim 1 - N^{-r}$, which is *much* better than Chebyshev's inequality!

Why are trace estimators useful?

If I have easy access to entries of $\boldsymbol{A}$, it's easy to exactly compute $\operatorname{tr}(\boldsymbol{A})$.

These estimators are useful when the entries of $\boldsymbol{A}$ are not easily available, but instead when $\boldsymbol{x} \mapsto \boldsymbol{A}\boldsymbol{x}$ is available.

For example:

- When the entries of $\boldsymbol{A}$ are simply not available.
- When computing $\operatorname{tr}\left[(z\boldsymbol{I} - \boldsymbol{A})^{-1}\right]$
- When computing $\operatorname{tr} f(\boldsymbol{A})$, often using truncated Taylor approximations of f (i.e., polynomials in $\boldsymbol{A}$.

However, in any of these situations $\boldsymbol{x} \mapsto \boldsymbol{A}\boldsymbol{x}$ could be used to recover entries of $\boldsymbol{A}$ by choosing $\boldsymbol{x}$ appropriately.

Therefore, a randomized trace estimator would only be useful if we can choose N much smaller than the dimension of $\boldsymbol{A}$.

Gaussian sampling

We hope an exponential concentration argument, similar to the Rademacher case, works for Gaussian isotropic vectors.

However, note that if $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$, then

$$Z = \boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x} = \sum_i A_{i,i} x_i^2 + \sum_{i \neq j} A_{i,j} x_{i,j}^2,$$

The important observation is that, at least part of this quantity, x_i^2, is *not* a normal random variable. Indeed:

$$\Pr(x_i^2 \geqslant t) = \Pr(|x_i| \geqslant \sqrt{t}) \sim e^{-t},$$

i.e., this random variable is <u>not</u> sub-Gaussian.
(It's exactly a Gamma/exponential random variable.)

To handle these types of random variables, we require some extra study of concentration.

Back to MGF's

When studying concentration in the simplest setting, we have the setup:

$$Z_n = \frac{1}{n} S_n := \frac{1}{n} \sum_{i \in [n]} X_i \qquad X_i \stackrel{\text{iid}}{\sim} X, \qquad \mathbb{E} X = 0$$

And we want to know tail probabilities of Z_n.
Recall, for any $s > 0$:

$$\Pr(Z_n \geqslant t) = \Pr(S_n \geqslant nt) \leqslant e^{-nst} \left[M_X^n(s) \right]$$

While we've introduced this for $s > 0$, taking $s < 0$ bounds $\Pr(Z_n \leqslant t)$, so that considering all $s \in \mathbb{R}$ bounds both tails.

In particular, this implies:

$$\begin{aligned} \Pr(Z_n \geqslant t) &\leqslant \inf_{s \in \mathbb{R}} e^{-nst} M_X^n(s) \\ &= \exp\left(n \left(\inf_s -st + \log M_X(s) \right) \right) \\ &= \exp\left(-n \sup_s \left[st - \log M_X(s) \right] \right). \end{aligned}$$

CGF's and rate functions

With a little rearrangement, we've shown the following:

$$\frac{1}{n}\log \Pr(Z_n \geqslant t) \leqslant -I(t), \qquad\qquad I(t) = \sup_s (st - \log M_X(s)).$$

In the language of large deviations, $I(t)$ is the *rate function*, which characterizes tail probabilities.

Equivalently, define the function K_X,

$$K_X(s) := \log M_X(s) = \log \mathbb{E} e^{sX},$$

which is the *cumulant generating function* (CGF) of X.

Then $I(t)$ is related to $K_X(s)$ by:

$$I_n(t) = \sup_s \left(st - K_X(s)\right)$$

The Legendre-Fenchel transform

$$\log \Pr\left(Z_n \geqslant t\right) \leqslant -I(t), \qquad\qquad I(t) = \sup_s \left(st - K_X(s)\right)$$

The operation connecting I to K itself has a name:

Let $f : \mathbb{R} \to \mathbb{R}$. The function f^* defined as,

$$f^*(t) := \sup_s \left(st - f(s)\right),$$

is called the *Legendre-Fenchel transform/dual* or *convex conjugate* of f.

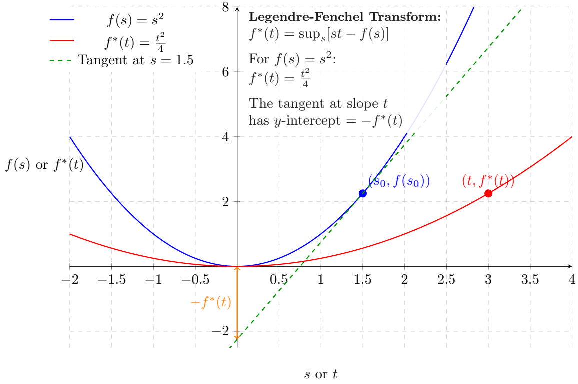
So with all this new language:

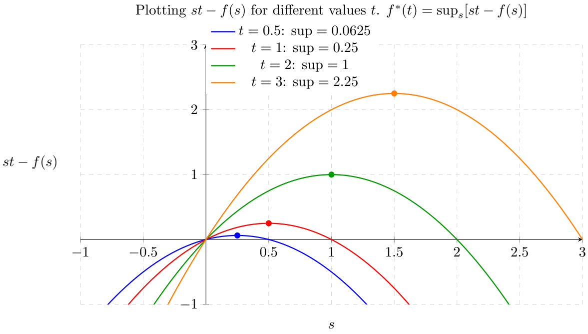
> The rate function (corresponding to the tail probability we seek to compute) is the convex conjugate of the CGF of X: $I(t) = K_X^*(t)$.

Or:

> The tail probability is the exponential of the negative convex conjugate of the CGF of X: $\Pr(Z_n \geqslant t) \leqslant \exp(-nI(t))$

Legendre-Fenchel visualization





Legendre-Fenchel transform

$$f^*(t) := \sup_s \left(st - f(s)\right), \qquad I(t) = K_X^*(t), \qquad \Pr(Z_n \geqslant t) \leqslant \exp(-nI(t))$$

The body of knowledge on Legendre-Fenchel transforms is large. Here are some punch lines that are somewhat relevant for us:

- Convex conjugates of convex functions are convex.
- CGF's are convex functions $\Longrightarrow$ Rate functions are convex.
- The Legendre-Fenchel transform is an involution on convex functions: $f^{**} = f$.
- We know the convex conjugate for the MGF of many named ("classical") distributions.
- $X \sim \mathcal{N}(0, \sigma^2) \Longrightarrow K_X(s) = \frac{1}{2}\sigma^2 s^2 \Longrightarrow K_X^*(t) = \frac{1}{2\sigma^2} t^2$.
- So: If X is $\mathcal{N}(0, \sigma^2)$, then:

$$I(t) = -\frac{1}{n} \log \Pr(Z_n \geqslant t) \geqslant nK_X^*(t) = \frac{nt^2}{2\sigma^2} \quad \Longrightarrow \quad \Pr(Z_n \geqslant t) \leqslant \exp\left(-\frac{nt^2}{2\sigma^2}\right).$$

Exponential distributions

We can apply this general technique to a new, relevant situation: If $X \sim \mathcal{N}(0,1)$, then X^2 has tail probabilities behaving like $f_{X^2}(x) \sim \exp(-|x|)$.

So, suppose Y has density $f_Y(y) = \frac{1}{2}\exp(-|y|)$. Let's compute the tail probability for an iid sum:

$$Z_n = \frac{1}{n}\sum_{i\in[n]} Y_i, \qquad\qquad Y_i \overset{\text{iid}}{\sim} Y.$$

As before, we have:

$$I(t) := K_Y^*(t), \qquad\qquad \Pr(Z_n \geqslant t) \leqslant \exp(-nI(t)).$$

We directly compute that $M_Y(s) = \frac{1}{1-s^2}$, $|s| < 1$.
Of particular note here: $M_Y(s)$ is defined only on a subset of $\mathbb{R}$. It does not exist for $|s| \geqslant 1$.

We therefore have $K_Y(s) = -\log(1-s^2)$, which is smooth (and convex), and using

$$K_Y^*(t) = ts^* - K_Y(s^*), \qquad\qquad s^*(t) = \left(K'\right)^{-1}(t),$$

we identify

$$K_Y^*(t) = \log\left(\frac{2}{e}\right) + \sqrt{1+t^2} + \log\left(\frac{1}{t^2}\left(\sqrt{1+t^2}-1\right)\right).$$

Exponential tail probabilities

We can now identify tail probability behavior:

$$Z_n = \frac{1}{n} \sum_{i \in [n]} Y_i, \qquad Y_i \overset{\text{iid}}{\sim} Y, \qquad f_Y(y) = \frac{1}{2} \exp(-|y|)$$

$$K_Y^*(t) = \log\left(\frac{2}{e}\right) + \sqrt{1+t^2} + \log\left(\frac{1}{t^2}\left(\sqrt{1+t^2} - 1\right)\right).$$

Through Taylor series ($\sqrt{1+x^2} \sim 1 + x^2/2$ for small x), one can show:

$$\begin{aligned} t \ll 1 \quad &\Longrightarrow \quad K_Y^*(t) \sim -1 + \frac{1}{2} t^2 \\ t \gg 1 \quad &\Longrightarrow \quad K_Y^*(t) \sim t. \end{aligned}$$

I.e.:

$$\begin{aligned} t \ll 1 \quad &\Longrightarrow \quad \Pr(Z_n \geqslant t) \sim \exp\left(-\frac{nt^2}{2}\right) \\ t \gg 1 \quad &\Longrightarrow \quad \Pr(Z_n \geqslant t) \sim \exp\left(-nt\right). \end{aligned}$$

And one can show that these bounds can be stitched together to cover all $t > 0$.

Sub-exponential distributions

In order to state the generalization of this result, we consider the appropriate class of exponential-like distributions.

Definition 1 (Sub-exponential distributions). *A random variable X has sub-exponential distribution if there is a constant $C > 0$ such that:*

$$\Pr(|X| \geqslant t) \leqslant 2 \exp(-Ct).$$

As expected, there is an equivalent definition that is more useful for tail probabilties:

Theorem (Equivalent definitions of sub-exponential distributions). *The definition above for a sub-exponential distribution X is equivalent to: there exists a $c > 0$ such that*

$$K_X(s) \leqslant c^2 s^2, \qquad |s| \leqslant \frac{1}{c}$$

(Recall for $f_X \sim \exp(-|x|)$, we had $K_X(s) \leqslant s^2$ for $|s| < 1$.)

Sub-Gaussian distributions are sub-exponential distributions.

In particular: Sub-Gaussian distributions have $K_X(s) \sim s^2$ for all $s \in \mathbb{R}$. Sub-exponential distributions require this only in a neighborhood of the origin.

Bernstein's inequality

Sub-exponential distributions allows us to codify our previous observation that such random variables have tail probabilities behaving like e^{-t^2} when t is small, but e^{-t} when t is large.

Theorem (Bernstein's inequality). *Let X_i, $i \in [n]$ be independent and centered sub-exponential random variables. Then there is a $c > 0$ such that:*

$$\Pr\left(\left|\frac{1}{n}\sum_{i\in[n]} X_i\right| \geqslant t\right) \leqslant 2\exp\left(-cnh_X(t)\right), \qquad h_X(t) = \min\left\{\frac{t^2}{K^2}, \frac{t}{K}\right\},$$

where K is the maximum sub-exponential norm of the X_i, defined as:

$$K = \max_{i\in[n]} \inf\left\{t > 0 \;\middle|\; \mathbb{E}\exp\left(|X|/t\right) \leqslant 2\right\}.$$

Bernstein's inequality says that for sub-exponential random variables:

- Deviations close to the mean ("small deviations") have a Gaussian, CLT-type probability
- Deviations far from the mean ("large deviations") have an exponential-type probability.

Another Bernstein inequality

There are a few versions of Bernstein's inequality. Here's one that applies to bounded random variables:

Theorem (Bernstein's inequality, II). *Let X_i, $i \in [n]$ be independent, centered, and bounded random variables, with $|X_i| \leqslant K$ for all $i \in [n]$. Then:*

$$\Pr\left(\left|\sum_{i\in[n]} X_i\right| \geqslant t\right) \leqslant 2\exp\left(-\frac{t^2/2}{\sum_{i\in[n]} \mathbb{E}X_i^2 + Kt/2}\right).$$

This form explicitly shows for bounded random variables the tail probabilities are $\sim \exp(-t)$ for large t and $\sim \exp(-t^2)$ for small t.

Without belaboring the point: now that we know that "Gaussian^2" random variables have exponential concentration, we could use this to get a exponential success probability on the isotropic Gaussian randomized trace estimator.

Example 2: The Johnson-Lindenstrauss Lemma

A very nice, "straightforward" corollary of all this is a celebrated result in dimension reduction:

Lemma (Johnson-Lindenstrauss, Gaussian version)**.** *Let* $\boldsymbol{X} \in \mathbb{R}^{n \times m}$ *be a deterministic matrix,*

$$\boldsymbol{X} = [\boldsymbol{x}_1, \, \boldsymbol{x}_2, \, \cdots \, \boldsymbol{x}_m] \, .$$

Let $\boldsymbol{A} \in \mathbb{R}^{k \times n}$ *have iid standard normal entries. Define:*

$$\boldsymbol{Y} = \frac{1}{\sqrt{k}} \boldsymbol{A} \boldsymbol{X} = [\boldsymbol{y}_1, \, \boldsymbol{y}_2, \, \cdots \, \boldsymbol{y}_m] \in \mathbb{R}^{k \times m}$$

Then:

$$k \gtrsim \frac{(r+2) \log m}{\epsilon^2} \quad \stackrel{\text{w/ probability} \gtrsim 1 - m^{-r}}{\Longrightarrow} \quad (1-\epsilon)\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 \leqslant \|\boldsymbol{y}_i - \boldsymbol{y}_j\|_2 \leqslant (1+\epsilon)\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2,$$

for all $i, j \in [m]$*.*

I.e., with dimension complexity *logarithmic* in the number of samples, arbitrary point clouds can be near-isometrically projected.

Johnson-Lindenstrauss proof sketch, I

We have all the necessary ingredients to prove this directly:

First, let $\boldsymbol{z} \in \mathbb{R}^n$, $\|\boldsymbol{z}\|_2 = 1$, be deterministic and arbitrary. We'll first show that $\boldsymbol{y} := \frac{1}{\sqrt{k}} \boldsymbol{A}\boldsymbol{z}$ has norm similar to $\boldsymbol{z}$.

We have that $\boldsymbol{y}$ is a centered, normally distributed random vector, with,

$$\begin{aligned}
\mathrm{Cov}\left(\boldsymbol{y}\right)_{i,j} &= \mathbb{E} y_i y_j \\
&= \frac{1}{k} \sum_{q,\ell \in [n]} \mathbb{E} A_{i,q} A_{j,\ell} z_q z_\ell \\
&= \frac{1}{k} \delta_{i,j} \sum_{q \in [n]} z_q^2 = \frac{1}{k} \delta_{i,j}.
\end{aligned}$$

Therefore, $\|\boldsymbol{y}\|_2^2 = \sum_{j \in [k]} y_j^2$ is a sum of k $\mathcal{N}(0, 1/k)$-squared iid random variables, each of which is sub-exponential.

Johnson-Lindenstrauss proof sketch, II

Let $W_i = y_i^2$, so that $\mathbb{E}\|\boldsymbol{y}_i^2\|_2 = \mathbb{E}\sum_{i\in[k]} W_i = 1$, with W_i iid subexponential, having subexponential norm $\sim 1/\sqrt{k}$.

We are interested in small deviations from the mean. For small $\epsilon > 0$, we have our concentration estimate:

$$\Pr\left(\left|\sum_{i\in[k]} W_i - 1\right| \geqslant \epsilon\right) \lesssim \exp(-k\epsilon^2)$$

I.e., the following occurs with probability at least $1 - ce^{-k\epsilon^2}$:

$$(1-\epsilon) \leqslant \|\boldsymbol{y}\|_2^2 \leqslant (1+\epsilon),$$

with $\|\boldsymbol{z}\|_2^2 = 1$.

Johnson-Lindenstrauss proof sketch, III

We've shown that with probability $1 - ce^{-k\epsilon^2}$:

$$(1-\epsilon) \leqslant \left\| \frac{\frac{1}{\sqrt{k}}\boldsymbol{A}\boldsymbol{z}}{\|\boldsymbol{z}\|_2^2} \right\|_2^2 \leqslant (1+\epsilon),$$

Now choose $\boldsymbol{z} = \boldsymbol{x}_i - \boldsymbol{x}_j$ for all $i, j \in [m]$ with $i \neq j$. Then:

$$(1-\epsilon) \leqslant \left\| \frac{\frac{1}{\sqrt{k}}\boldsymbol{A}(\boldsymbol{x}_i - \boldsymbol{x}_j)}{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2} \right\|_2^2 \leqslant (1+\epsilon),$$

holds, by the union bound, for all $i \neq j$ with probability at least $1 - cm^2 e^{-k\epsilon^2}$.

Pick $k \gtrsim (r+2)\frac{\log m}{\epsilon^2}$ to ensure this happens with probability at least,

$$1 - cm^2 \exp\left(-(r+2)\log m\right) = 1 - cm^{-r}.$$

NB: The whole thing rested on us being able to bound tail probabilities of sub-exponential random variables.