Exponential Concentration II: Part A

SDS 391P.6, Spring 2026

Pratik Patil

1 Motivation

Last time, we developed exponential concentration for *sums* of independent random variables using the Laplace transform method. The key object was the centered log-moment generating function (centered CGF)

$$\psi_{S_n}(\lambda) = \log \mathbb{E} e^{\lambda(S_n - \mathbb{E}S_n)}.$$

For sums this approach is especially clean because independence gives additivity: if $S_n = \sum_{i=1}^n X_i$ and the X_i are independent, then

$$\psi_{S_n}(\lambda) = \sum_{i=1}^n \psi_{X_i}(\lambda).$$

So a high-dimensional question (a log-MGF for a sum) reduces to n one-dimensional bounds, followed by a Chernoff optimization in λ.

In many problems, however, the random quantity of interest is a *nonlinear* functional

$$Z = f(X_1, \ldots, X_n), \qquad \text{where } X_1, \ldots, X_n \text{ are independent.}$$

Examples that we have seen before (while bounding variances) include: norms, eigenvalues, maxima, and combinatorial statistics. For such Z, we have no easy factorization: the exponential moment $\mathbb{E}e^{\lambda f(X_1,\ldots,X_n)}$ does not decompose in any useful way. So we need a new structural idea.

A helpful analogy comes from the variance bounds we developed earlier. Variance is not additive for nonlinear functions either, but it *tensorizes*: one can control $\mathrm{Var}(Z)$ by summing one-coordinate conditional variances. This reduces an n-dimensional problem to a sequence of one-coordinate perturbations: we roughly only need to understand what happens to f when only one input coordinate is resampled.

Our goal in this lecture is to develop an *exponential* analogue of this idea. Instead of controlling a single number like $\mathrm{Var}(Z)$, we want to control the whole function

$$\lambda \mapsto \psi_Z(\lambda) = \log \mathbb{E} e^{\lambda(Z - \mathbb{E}Z)}.$$

It is tempting to hope that "sub-Gaussianity tensorizes" the way variance does. Unfortunately, the property $\psi_Z(\lambda) \leqslant \frac{v\lambda^2}{2}$ does not directly decompose cleanly coordinate-by-coordinate.

The key idea in this lecture is that the right tensorizing object is *entropy*. Entropy is designed to interact with exponentials, and it *does* tensorize under independence. A classic calculus step that goes by the *Herbst's argument* then converts an entropy bound into a quadratic CGF bound, which immediately yields Gaussian concentration. Calculus for the win again!

2 Entropy

Throughout, let $X = (X_1, \dots, X_n)$ have independent coordinates, and let

$$Z = f(X) \in \mathbb{R}, \qquad \widetilde{Z} := Z - \mathbb{E}Z.$$

Our ultimate target is the centered log-MGF (centered CGF)

$$\psi_Z(\lambda) := \log \mathbb{E}e^{\lambda \widetilde{Z}}.$$

2.1 Statistical divergences and convexity gaps

Let P and Q be probability measures on the same measurable space with $P \ll Q$. The *Kullback–Leibler divergence* (relative entropy) is

$$D_{\mathrm{KL}}(P\|Q) := \int \log\Big(\frac{dP}{dQ}\Big)\, dP = \mathbb{E}_P\Big[\log\Big(\frac{dP}{dQ}\Big)\Big] \in [0, \infty].$$

It is always nonnegative, and $D_{\mathrm{KL}}(P\|Q) = 0$ iff $P = Q$.

A useful convex-analytic notion is the *Bregman divergence*. Let $\phi : (0, \infty) \to \mathbb{R}$ be convex and differentiable. Define

$$D_\phi(a\|t) := \phi(a) - \phi(t) - \phi'(t)(a - t), \qquad a, t > 0.$$

Convexity implies $D_\phi(a\|t) \geqslant 0$ with equality iff $a = t$. Moreover, Jensen's inequality can be written as an average Bregman divergence:

$$\mathbb{E}[\phi(Y)] - \phi(\mathbb{E}Y) = \mathbb{E}\big[D_\phi(Y\|\mathbb{E}Y)\big].$$

Two special cases are worth keeping in mind.

- If $\phi(u) = u^2$, then $D_\phi(a\|t) = (a - t)^2$, and

  $$\mathbb{E}[Y^2] - (\mathbb{E}Y)^2 = \mathrm{Var}(Y).$$

  So variance is a Jensen/Bregman gap for u^2.
- If $\phi(u) = u \log u$, the same notion produces the entropy functional used in concentration.

2.2 The concentration entropy functional

Let $Y \geqslant 0$ with $\mathbb{E}Y < \infty$ and $\mathbb{E}[Y \log Y] < \infty$. Define the *concentration entropy* (entropy functional)

$$\mathrm{Ent}(Y) := \mathbb{E}[Y \log Y] - (\mathbb{E}Y) \log(\mathbb{E}Y). \tag{1}$$

This is *not* Shannon entropy of a distribution; it is a functional of a nonnegative random variable. It measures how far Y is from being constant: $\mathrm{Ent}(Y) \geqslant 0$ by Jensen, and $\mathrm{Ent}(Y) = 0$ iff Y is a.s. constant.

A basic property is homogeneity: for any $c > 0$,

$$\mathrm{Ent}(cY) = c\, \mathrm{Ent}(Y). \tag{2}$$

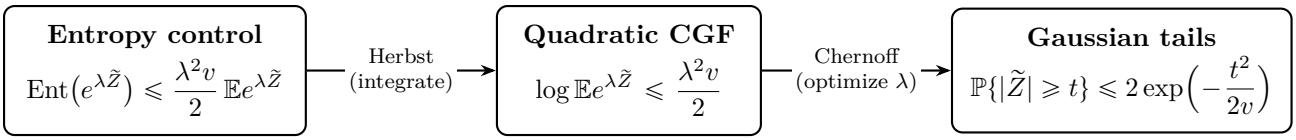

Figure 1: The main idea of the entropy-method: an entropy bound yields a quadratic bound on the centered log-MGF, which yields Gaussian concentration via Chernoff.

Entropy can be seen as KL via tilting. Given $Y \geqslant 0$ with $\mathbb{E}Y > 0$, define a tilted measure P_Y by

$$\frac{dP_Y}{dP} = \frac{Y}{\mathbb{E}Y}.$$

Then

$$D_{\mathrm{KL}}(P_Y \| P) = \frac{\mathrm{Ent}(Y)}{\mathbb{E}Y}, \qquad \text{equivalently} \qquad \mathrm{Ent}(Y) = (\mathbb{E}Y)\, D_{\mathrm{KL}}(P_Y \| P). \tag{3}$$

So $\mathrm{Ent}(Y)/\mathbb{E}Y$ measures how much the reweighting induced by Y changes the underlying law.

2.3 Why entropy appears in concentration

Take a real random variable Z and consider $Y = e^{\lambda Z}$. A direct computation gives

$$\mathrm{Ent}(e^{\lambda Z}) = \lambda\, \mathbb{E}[Z e^{\lambda Z}] - (\mathbb{E}e^{\lambda Z}) \log(\mathbb{E}e^{\lambda Z}). \tag{4}$$

Introduce the MGF and CGF

$$m_Z(\lambda) := \mathbb{E}e^{\lambda Z}, \qquad \kappa_Z(\lambda) := \log m_Z(\lambda).$$

Then (4) becomes

$$\frac{\mathrm{Ent}(e^{\lambda Z})}{m_Z(\lambda)} = \lambda\, \kappa_Z'(\lambda) - \kappa_Z(\lambda) = \lambda^2\, \frac{d}{d\lambda}\left(\frac{\kappa_Z(\lambda)}{\lambda}\right). \tag{5}$$

This identity is the bridge between entropy bounds and log-MGF bounds. Herbst's argument is essentially the act of integrating (5).

3 Herbst's argument

Herbst's argument converts an *entropy bound* for $e^{\lambda\widetilde{Z}}$ into a *quadratic centered CGF bound* for $\widetilde{Z}$. Once we have a quadratic bound on $\psi_Z(\lambda) = \log \mathbb{E}e^{\lambda\widetilde{Z}}$, Chernoff yields Gaussian tails. See Figure 1 for a visual summary.

3.1 The key differential identity

For the calculus step, it is cleanest to assume $\mathbb{E}Z = 0$ and work with $\psi_Z(\lambda) = \log \mathbb{E}e^{\lambda Z}$. (If $\mathbb{E}Z \neq 0$, replace Z by $\widetilde{Z} = Z - \mathbb{E}Z$; this does not change the entropy ratio $\mathrm{Ent}(e^{\lambda Z})/\mathbb{E}e^{\lambda Z}$ because of (2).)

Combining (5) with $\psi_Z = \kappa_Z$ (when $\mathbb{E}Z = 0$) gives: for $\lambda \neq 0$,

$$\frac{d}{d\lambda}\left(\frac{\psi_Z(\lambda)}{\lambda}\right) = \frac{1}{\lambda^2}\,\frac{\mathrm{Ent}(e^{\lambda Z})}{\mathbb{E}e^{\lambda Z}}. \tag{6}$$

This is the core of Herbst's argument: an upper bound on $\mathrm{Ent}(e^{\lambda Z})/\mathbb{E}e^{\lambda Z}$ becomes a differential inequality for $\psi_Z(\lambda)/\lambda$, which we can integrate.

3.2 Herbst's lemma

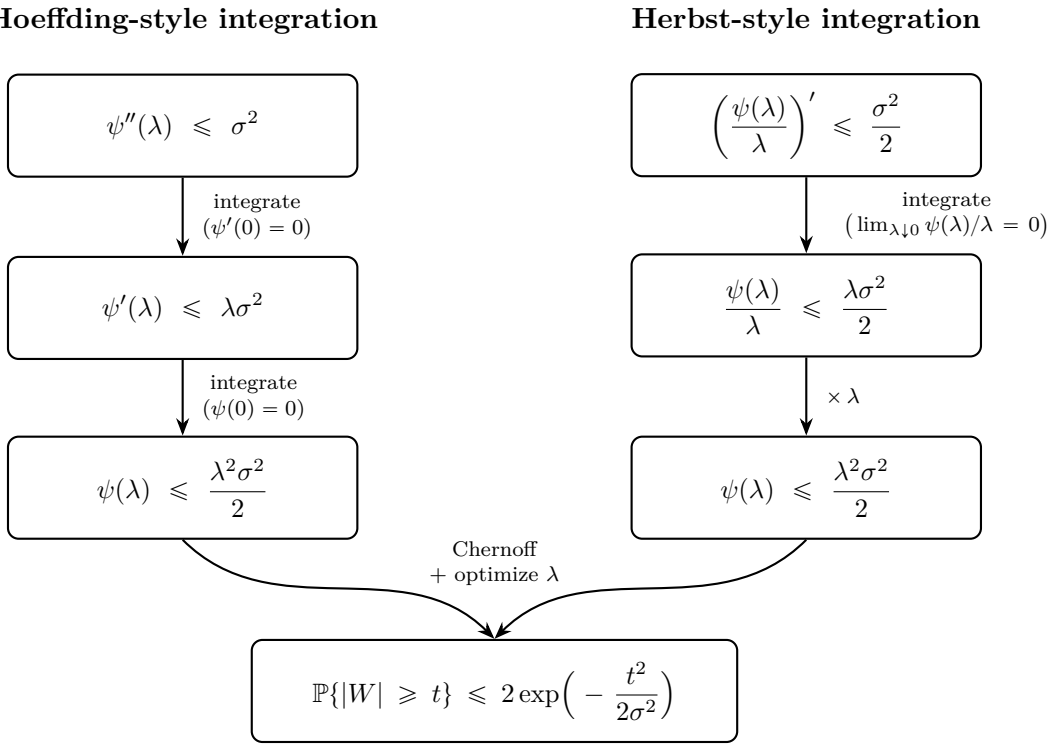

Figure 2: Two routes to the quadratic centered log-MGF bound $\psi(\lambda) \leqslant \lambda^2\sigma^2/2$ (for $\lambda > 0$), and the common Gaussian tail bound that follows.

Proposition 3.1 (Herbst)**.** *Let Z be a real random variable with $\mathbb{E}Z = 0$, and assume $\mathbb{E}e^{\lambda Z} < \infty$ for λ in a neighborhood of 0. Suppose there exists $v \geqslant 0$ such that for all $\lambda \in \mathbb{R}$,*

$$\mathrm{Ent}(e^{\lambda Z}) \;\leqslant\; \frac{\lambda^2 v}{2}\,\mathbb{E}e^{\lambda Z}.$$

Then for all $\lambda \in \mathbb{R}$,

$$\psi_Z(\lambda) = \log \mathbb{E}e^{\lambda Z} \;\leqslant\; \frac{\lambda^2 v}{2}.$$

Equivalently, Z is v-sub-Gaussian, and therefore

$$\mathbb{P}\{|Z| \geqslant t\} \leqslant 2\exp\left(-\frac{t^2}{2v}\right), \qquad t \geqslant 0.$$

Proof. By (6) and the assumed entropy bound, for $\lambda \neq 0$,

$$\frac{d}{d\lambda}\left(\frac{\psi_Z(\lambda)}{\lambda}\right) = \frac{1}{\lambda^2}\,\frac{\mathrm{Ent}(e^{\lambda Z})}{\mathbb{E}e^{\lambda Z}} \leqslant \frac{v}{2}.$$

Integrating from 0 to λ gives

$$\frac{\psi_Z(\lambda)}{\lambda} - \lim_{u\to 0}\frac{\psi_Z(u)}{u} \leqslant \int_0^\lambda \frac{v}{2}\,du = \frac{v\lambda}{2}.$$

The limit is 0 because $\psi_Z(0) = 0$ and $\psi_Z'(0) = \mathbb{E}Z = 0$. Therefore $\psi_Z(\lambda) \leqslant v\lambda^2/2$. The tail bound follows by applying Chernoff to Z and $-Z$. □

In the proof of Hoeffding-type bounds, we often bound ψ'' and integrate twice. Herbst's argument bounds $(\psi/\lambda)'$ and integrates once. Figure 2 summarizes these two routes to a quadratic CGF bound.

4 Tensorization of entropy

The main advantage of Herbst's lemma is that it reduces sub-Gaussian concentration to an *entropy bound*. This is only useful because entropy tensorizes under independence, while sub-Gaussianity itself does not.

Let $X = (X_1, \dots, X_n)$ have independent coordinates, and let $Y = Y(X) \geqslant 0$. For each coordinate i, define the operator that averages over X_i only:

$$\mathbb{E}_i[\cdot] := \mathbb{E}[\cdot \mid X_{-i}], \qquad X_{-i} := (X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n).$$

Independence implies that the $\mathbb{E}_i$ commute and that $\mathbb{E}_1 \cdots \mathbb{E}_n = \mathbb{E}$.

Define the *coordinatewise entropy* by

$$\mathrm{Ent}_i(Y) := \mathbb{E}_i\Big[Y\big(\log Y - \log \mathbb{E}_i Y\big)\Big] = \mathbb{E}_i\Big[Y \log\Big(\frac{Y}{\mathbb{E}_i Y}\Big)\Big]. \tag{7}$$

This is itself a random variable (it depends on X_{-i}). It is the entropy analogue of the conditional variance functional $\mathrm{Var}_i(Z) = \mathbb{E}_i[(Z - \mathbb{E}_i Z)^2]$ from Efron–Stein–Steele.

Theorem 4.1 (Tensorization of entropy)**.** *Let $X_1, \dots, X_n$ be independent and let $Y = Y(X_1, \dots, X_n) \geqslant 0$. Then*

$$\mathrm{Ent}(Y) \;\leqslant\; \mathbb{E}\left[\sum_{i=1}^{n} \mathrm{Ent}_i(Y)\right]. \tag{8}$$

Proof sketch. This mirrors the martingale proof of variance tensorization. Average coordinates one by one and telescope $\log Y$. Then use conditional Jensen with the joint convexity of the divergence $a \log(a/t) - a + t$ (the Bregman divergence for $u \log u$). □

Variance tensorizes as $\mathrm{Var}(Z) \leqslant \sum_i \mathbb{E}\mathrm{Var}_i(Z)$. Entropy tensorizes as $\mathrm{Ent}(Y) \leqslant \sum_i \mathbb{E}\,\mathrm{Ent}_i(Y)$. This parallel is the structural reason the entropy method works for nonlinear $f(X)$.

5 First application: discrete MLS and bounded differences

We now combine: (i) tensorization of entropy, (ii) a one-dimensional entropy inequality, and (iii) Herbst's argument.

Let $Z = f(X)$ and let $X' = (X'_1, \dots, X'_n)$ be an independent copy of X. For each i, define the resampled input and resampled output

$$X^{(i)} := (X_1, \dots, X_{i-1}, X'_i, X_{i+1}, \dots, X_n), \qquad Z^{(i)} := f(X^{(i)}).$$

Let $\mathbb{E}'[\cdot] := \mathbb{E}[\cdot \mid X]$ denote expectation over X' conditional on X.

5.1 Univariate discrete MLS

A symmetrization argument plus a numerical inequality yields a univariate entropy bound.

Lemma 5.1 (Discrete MLS in one dimension)**.** *Let U, U' be i.i.d. real random variables, and let $\lambda \in \mathbb{R}$. Then*

$$\mathrm{Ent}(e^{\lambda U}) \;\leqslant\; \frac{\lambda^2}{2}\,\mathbb{E}\Big[\mathbb{E}'\big[(U-U')^2\big]\, e^{\lambda U}\Big],$$

where $\mathbb{E}'$ averages over U' only.

Proof sketch. A symmetrization gives

$$\mathrm{Ent}(e^{\lambda U}) \leqslant \frac{\lambda}{2}\mathbb{E}\big[(e^{\lambda U} - e^{\lambda U'})(U-U')\big].$$

Then apply the numerical inequality

$$(a-b)(e^a - e^b) \leqslant \frac{1}{2}(a-b)^2(e^a + e^b),$$

with $a = \lambda U$ and $b = \lambda U'$, and use symmetry of (U, U') to simplify. □

5.2 Multivariate discrete MLS

Apply Theorem 4.1 to $Y = e^{\lambda Z}$, and then apply Lemma 5.1 conditionally in each coordinate. This yields:

Corollary 5.2 (Multivariate discrete MLS)**.** *For any $\lambda \in \mathbb{R}$,*

$$\mathrm{Ent}(e^{\lambda Z}) \;\leqslant\; \frac{\lambda^2}{2}\,\mathbb{E}\left[\Big(\sum_{i=1}^{n} \mathbb{E}'\big[(Z - Z^{(i)})^2\big]\Big)\, e^{\lambda Z}\right].$$

It is convenient to name the (random) energy proxy

$$V := \sum_{i=1}^{n} \mathbb{E}'\big[(Z - Z^{(i)})^2\big]. \tag{9}$$

This is the same resampling energy that appears in Efron–Stein-type variance bounds, but now it is paired with the exponential weight $e^{\lambda Z}$.

5.3 Exponential concentration under uniform energy bound

Theorem 5.3 (Uniform energy bound ⇒ sub-Gaussian tails)**.** *Let $Z = f(X_1, \ldots, X_n)$ and define V as in (9). If $V \leqslant v$ almost surely for some constant $v < \infty$, then*

$$\log \mathbb{E}e^{\lambda(Z - \mathbb{E}Z)} \leqslant \frac{\lambda^2 v}{2} \qquad (\lambda \in \mathbb{R}),$$

and consequently,

$$\mathbb{P}\{|Z - \mathbb{E}Z| \geqslant t\} \leqslant 2\exp\Big(-\frac{t^2}{2v}\Big), \qquad t \geqslant 0.$$

Proof sketch. Corollary 5.2 gives

$$\mathrm{Ent}(e^{\lambda Z}) \leqslant \frac{\lambda^2}{2}\mathbb{E}[Ve^{\lambda Z}] \leqslant \frac{\lambda^2 v}{2}\mathbb{E}e^{\lambda Z}.$$

By homogeneity of entropy, the same inequality holds with Z replaced by $\widetilde{Z} = Z - \mathbb{E}Z$ (and note $Z - \mathbb{E}Z$ has the same resampling differences as Z). Apply Herbst (Proposition 3.1) to $\widetilde{Z}$, then apply Chernoff for tails. □

5.4 Bounded differences

Assume there exist constants $c_1, \ldots, c_n$ such that for any $x, x' \in \mathbb{R}^n$ differing only in coordinate i,

$$|f(x) - f(x')| \leqslant c_i.$$

Then $|Z - Z^{(i)}| \leqslant c_i$, so $\mathbb{E}'[(Z - Z^{(i)})^2] \leqslant c_i^2$ and therefore $V \leqslant \sum_{i=1}^n c_i^2$ almost surely. Plugging into Theorem 5.3 yields:

Theorem 5.4 (Bounded differences via entropy)**.** *Let $X_1, \ldots, X_n$ be independent and let $Z = f(X_1, \ldots, X_n)$. Assume $|f(x) - f(x')| \leqslant c_i$ whenever x and x' differ only in coordinate i. Then for all $t \geqslant 0$,*

$$\mathbb{P}\{|Z - \mathbb{E}Z| \geqslant t\} \;\leqslant\; 2\exp\left(-\frac{t^2}{2\sum_{i=1}^n c_i^2}\right).$$

The classical McDiarmid inequality is often stated with a sharper numerical constant in the exponent. Obtaining that constant requires a refined one-dimensional entropy "range bound" (parallel to Hoeffding's lemma). For us, the main message so far is:

bounded coordinate sensitivity ⇒ entropy control ⇒ quadratic log-MGF bound ⇒ Gaussian concentration.

6 Look ahead

This lecture introduced the entropy method as a way to obtain sub-Gaussian concentration for nonlinear functions of independent inputs. The overall chain of argument is: tensorization of entropy → an entropy bound for $e^{\lambda \widetilde{Z}}$ → Herbst's argument → Gaussian tails.

Next time, we will develop sharper and more flexible one-dimensional entropy bounds through log-Sobolev and modified log-Sobolev inequalities for specific distributions. These inequalities tensorize cleanly and yield powerful concentration results for Lipschitz functions.

Source material

Parts of this lecture are based on references: Tropp (2023); van Handel (2016), in addition to the author's accumulated experience working on related topics.

References

Tropp, J. A. (2023). Probability in high dimensions. Caltech CMS Lecture Notes 2021-01.

van Handel, R. (2016). Probability in high dimension. Lecture Notes (Princeton University). APC 550.