Exponential Concentration II: Part B

SDS 391P.6, Spring 2026

Pratik Patil

1 Motivation

In the last class, we introduced the entropy method as an exponential analog of variance tensorization. The punchline was that if we can control the entropy of an exponential weight $e^{\lambda Z}$, then Herbst's argument turns that control into a quadratic bound on the centered log-MGF, which then yields Gaussian tails by Chernoff. This gave a clean route to McDiarmid-type concentration via a *resampling energy*

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

When V admits a uniform bound, we immediately get sub-Gaussian tails for $Z - \mathbb{E}Z$.

This approach is quite general (it requires only independence), but it has two practical limitations:

- The variance proxy V is expressed in terms of resampling differences $Z - Z^{(i)}$. In many smooth problems, we would prefer to work with calculus derivatives (gradients) rather than discrete perturbations.
- The cleanest conclusion used a uniform bound $V \leqslant v$ almost surely. This excludes important examples where the natural proxy fluctuates (e.g. quadratic forms, or functions with unbounded gradients).

The goal of this lecture is to explain a second, complementary viewpoint: for certain special distributions (Gaussian, Rademacher, bounded variables under convexity), one can replace resampling differences by a *Dirichlet energy* (e.g. $\|\nabla f\|_2^2$). The key input is a functional inequality called a (modified) log-Sobolev inequality (MLSI). Conceptually, MLSIs are to entropy what Poincaré inequalities are to variance: they control a convexity gap by an energy.

At a high level, the last class gave the recipe:

$$\text{entropy tensorization} \;+\; \text{(discrete) one-dimensional entropy bound} \;\Longrightarrow\; \text{sub-Gaussian tails.}$$

In this lecture, we keep the same template, but we change the one-dimensional entropy bound: instead of bounding entropy by a resampling square $(Z - Z')^2$, we bound it by a *squared derivative* $|f'|^2$ (or $\|\nabla f\|_2^2$). Recall we also did a similar discrete-to-continuous upgrade for variance tensorization.

2 Recap: entropy, Herbst, and bounded differences

The recap here is intentionally brief; see the last lecture note for more details.

2.1 Entropy as a convexity gap

For $a, t > 0$, define the scalar divergence

$$D_{\mathrm{KL}}(a\|t) \;:=\; a(\log a - \log t) - (a - t) \;\geqslant\; 0.$$

This is the Bregman divergence for $\phi(u) = u \log u$.

For a nonnegative random variable Y with $\mathbb{E}[Y \log Y] < \infty$, the concentration entropy is

$$\mathrm{Ent}(Y) \;:=\; \mathbb{E}[Y \log Y] - (\mathbb{E}Y)\log(\mathbb{E}Y) \;=\; \mathbb{E}\big[D_{\mathrm{KL}}(Y\|\mathbb{E}Y)\big] \;\geqslant\; 0.$$

2.2 Why entropy controls the log-MGF

If Z is real and $Y = e^{\lambda Z}$, then (Lecture 07) one computes

$$\frac{\mathrm{Ent}(e^{\lambda Z})}{\mathbb{E}e^{\lambda Z}} \;=\; \lambda\, \psi_Z'(\lambda) - \psi_Z(\lambda), \qquad \psi_Z(\lambda) = \log \mathbb{E}e^{\lambda Z}.$$

Herbst's argument is the calculus step that integrates this identity: if for all λ,

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

then $\psi_Z(\lambda) \leqslant \lambda^2 v/2$, hence Z is v-sub-Gaussian.

2.3 Tensorization of entropy

If $X = (X_1, \dots, X_n)$ has independent coordinates and $Y = Y(X) \geqslant 0$, define $\mathrm{Ent}_i(Y)$ by taking entropy only in coordinate i (conditioning on the others). Then entropy tensorizes:

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

This is the entropy analogue of variance tensorization (Efron–Stein–Steele).

2.4 Discrete MLSI and bounded differences

Combining tensorization with a univariate discrete inequality yields the multivariate discrete MLSI: for $Z = f(X)$ and $Z^{(i)} = f(X_1, \dots, X_i', \dots, X_n)$,

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

Define the resampling energy

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

If $V \leqslant v$ a.s., then $\log \mathbb{E} e^{\lambda(Z-\mathbb{E}Z)} \leqslant \lambda^2 v/2$, hence

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

A key special case is bounded differences: if changing only coordinate i changes f by at most c_i, then $|Z - Z^{(i)}| \leqslant c_i$, hence $V \leqslant \sum_i c_i^2$ a.s. and

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

Example: bin packing revisited. Let $X_1, \dots, X_n \in [0,1]$ be i.i.d. item sizes. Let Z be the minimum number of bins needed to pack them so that each bin has total load $\leqslant 1$. Changing one item size can change the optimal number of bins by at most 1: in the worst case we can always place the new item into a fresh bin. Thus $c_i = 1$ for all i, and bounded differences yields

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

This is a useful meta-message: for many combinatorial optimization problems, the optimum may not change much with one coordinate change and thus leads to exponential concentration.

3 Modified log-Sobolev inequalities (MLSI)

The discrete MLSI above is completely distribution-free, but it measures sensitivity by resampling. For smooth distributions (and smooth functions), it is natural to ask: can we replace resampling differences by derivatives?

3.1 MLSI in one dimension

Let X be a real random variable, and let $f : \mathbb{R} \to \mathbb{R}$ be a (nice) function. A typical modified log-Sobolev inequality has the form

$$\mathrm{Ent}\big(e^{f(X)}\big) \;\leqslant\; \frac{\rho}{2}\,\mathbb{E}\big[(f'(X))^2\, e^{f(X)}\big], \tag{1}$$

where $\rho > 0$ depends only on the law of X (not on f).

Just as the Poincaré inequality bounds a variance by an average squared derivative, MLSI bounds entropy by a squared-derivative energy, but with an exponential weight $e^{f(X)}$. Equivalently, writing $h = e^{f/2}$, inequality (1) becomes

$$\mathrm{Ent}(h(X)^2) \;\leqslant\; 2\rho\,\mathbb{E}\big[(h'(X))^2\big],$$

which resembles the more classical log-Sobolev inequality.

Once we have (1), we can plug in $f = \lambda Z$ and combine it with Herbst's argument. In effect, MLSI is a one-dimensional machine that produces quadratic log-MGF control.

3.2 Tensorization and Lipschitz concentration

Suppose $X_1, \dots, X_n$ are independent copies of X, and $F : \mathbb{R}^n \to \mathbb{R}$ is smooth. Apply entropy tensorization and then apply the one-dimensional MLSI (1) conditionally in each coordinate. One obtains the multivariate MLSI:

$$\operatorname{Ent}\big(e^{F(X)}\big) \;\leqslant\; \frac{\rho}{2}\,\mathbb{E}\big[\|\nabla F(X)\|_2^2\, e^{F(X)}\big]. \tag{2}$$

Now assume F is L-Lipschitz in Euclidean norm, i.e. $\|\nabla F(x)\|_2 \leqslant L$ almost everywhere. Then (2) yields

$$\operatorname{Ent}\big(e^{\lambda F(X)}\big) \;\leqslant\; \frac{\rho\lambda^2 L^2}{2}\,\mathbb{E}e^{\lambda F(X)}.$$

By homogeneity of entropy, the same holds for $F(X) - \mathbb{E}F(X)$, so Herbst gives the quadratic centered log-MGF bound

$$\log \mathbb{E}e^{\lambda(F(X)-\mathbb{E}F(X))} \;\leqslant\; \frac{\rho\lambda^2 L^2}{2}.$$

Therefore $F(X) - \mathbb{E}F(X)$ is (ρL^2)-sub-Gaussian, and

$$\mathbb{P}\{|F(X) - \mathbb{E}F(X)| \geqslant t\} \;\leqslant\; 2\exp\Big(-\frac{t^2}{2\rho L^2}\Big). \tag{3}$$

The pipeline is the same as last lecture: tensorization → entropy bound → Herbst → tails, but the variance proxy is now a bound on $\|\nabla F\|_2^2$ rather than a bound on the resampling energy V. Which proxy is more convenient depends on the problem: for discrete or combinatorial problems V is often easier; for smooth geometric problems $\|\nabla F\|_2$ is.

4 Examples of MLSI

The abstract recipe above becomes meaningful once we know which distributions satisfy which MLSIs. Here are three canonical examples:

4.1 Rademacher: an MLSI on the hypercube

Let $\varepsilon \in \{-1, +1\}$ be Rademacher. Applying the univariate discrete MLSI from Lecture 07 to $U = f(\varepsilon)$ yields

$$\operatorname{Ent}\big(e^{f(\varepsilon)}\big) \;\leqslant\; \frac{1}{4}\,(f(+1) - f(-1))^2\,\mathbb{E}e^{f(\varepsilon)}.$$

This is a discrete "derivative" inequality: the role of f' is played by the jump $f(+1) - f(-1)$. Tensorizing over $\varepsilon_1, \dots, \varepsilon_n$ gives an MLSI on the Boolean hypercube in terms of the discrete gradient.

Even in a purely discrete setting, the MLSI viewpoint is already present: the energy is the sum of squared discrete derivatives, and this is exactly what drives concentration.

4.2 Gaussian: Gross MLSI and Gaussian concentration

Let $g \sim \mathcal{N}(0,1)$. The Gaussian log-Sobolev inequality implies: for sufficiently nice $f : \mathbb{R} \to \mathbb{R}$,

$$\mathrm{Ent}\big(e^{f(g)}\big) \;\leqslant\; \frac{1}{2}\,\mathbb{E}\big[(f'(g))^2\, e^{f(g)}\big]. \tag{4}$$

Tensorizing yields the multivariate form: if $g \sim \mathcal{N}(0, I_n)$, then

$$\mathrm{Ent}\big(e^{F(g)}\big) \;\leqslant\; \frac{1}{2}\,\mathbb{E}\big[\|\nabla F(g)\|_2^2\, e^{F(g)}\big]. \tag{5}$$

Tsirelson–Ibragimov–Sudakov inequality. If F is L-Lipschitz, then (3) with $\rho = 1$ gives

$$\log \mathbb{E} e^{\lambda(F(g)-\mathbb{E}F(g))} \leqslant \frac{\lambda^2 L^2}{2}, \qquad \mathbb{P}\{|F(g) - \mathbb{E}F(g)| \geqslant t\} \leqslant 2e^{-t^2/(2L^2)}.$$

This is one of the cleanest examples of a dimension-free concentration phenomenon: the ambient dimension n never enters, only the Lipschitz constant does.

4.3 Convex MLSI for bounded random variables

Now let X be any random variable supported on $[a, b]$. There is a one-dimensional MLSI that holds for convex functions: if $f : [a, b] \to \mathbb{R}$ is convex and differentiable, then

$$\mathrm{Ent}\big(e^{f(X)}\big) \;\leqslant\; \frac{(b-a)^2}{2}\,\mathbb{E}\big[(f'(X))^2\, e^{f(X)}\big]. \tag{6}$$

Tensorizing yields the multivariate extension: if $X_1, \ldots, X_n$ are independent in $[a, b]$ and $F : \mathbb{R}^n \to \mathbb{R}$ is separately convex (convex in each coordinate), then

$$\mathrm{Ent}\big(e^{F(X)}\big) \;\leqslant\; \frac{(b-a)^2}{2}\,\mathbb{E}\big[\|\nabla F(X)\|_2^2\, e^{F(X)}\big].$$

It is important to note that the convexity restriction leads to an *upper tail* inequality for $F(X) - \mathbb{E}F(X)$, but not necessarily a matching lower tail bound. Indeed, $-F$ is concave, so the same argument does not apply to the lower tail.

Example: spectral norm of a random matrix revisited. The largest singular value is convex and 1-Lipschitz (in Frobenius norm) as a function of the entries. Thus (6) (tensorized) yields a sub-Gaussian *upper tail* for the spectral norm of a random matrix with independent bounded entries. This is a powerful "black-box" result: the distribution can be arbitrary inside $[a, b]$, and the geometry (convex + Lipschitz) does most of the work.

5 Summary and look ahead

It is useful to organize what we have proved so far into three parallel "tensorization stories":

- Variance tensorizes $\Rightarrow$ Efron–Stein-type variance bounds. These are second-moment statements: they tell us the scale but not the tail shape.
- Entropy tensorizes $\Rightarrow$ via discrete MLSI + Herbst we get distribution-free sub-Gaussian concentration under bounded differences (and related resampling-energy conditions).

- MLSI/log-Sobolev inequalities $\Rightarrow$ for special measures, entropy can be controlled by a derivative energy, leading to sharp, dimension-free concentration for Lipschitz functions (Gaussian case), and one-sided concentration for convex Lipschitz functions under boundedness (convex MLSI).

In the next unit, we will emphasize applications of these concentration tools in high-dimensional probability and statistics (random matrices, empirical averages, etc.). Later (time permitting), we may return to deeper geometric perspectives on concentration, including isoperimetric and transportation methods (e.g. Talagrand-type inequalities).

Source material

Parts of this lecture are based on references: Boucheron et al. (2013); van Handel (2016); Tropp (2023), in addition to the author's accumulated experience working on related topics.

References

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities - A Nonasymptotic Theory of Independence*. Oxford University Press.

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.