Concentration Inequalities - Based off Varun Jog's Part III class

Anqi Li

May 2024

1 Chernoff-Crámer Method

(💡) Write the *log-MGF* of RV Z as $\psi_Z(\lambda) = \log \mathbb{E}e^{\lambda Z}$. Then we have

$$\mathbb{P}[Z \geq t] \leq \frac{\mathbb{E}e^{\lambda Z}}{e^{\lambda t}} = \exp(-(\lambda t - \psi_Z(\lambda))) \leq \exp(-\psi_Z^*(\lambda))$$

where $\psi_Z^*(t) = \sup_{\lambda \geq 0}(\lambda t - \psi_Z(\lambda))$.

Example 1.1. When $Z \sim N(0, \sigma^2)$ then $\psi_Z^*(t) = \frac{t^2}{2\sigma^2}$ which recovers Chernoff's:

$$\mathbb{P}[|Z - \mathbb{E}[Z]| \geq t] \leq e^{-\frac{t^2}{2\sigma^2}}.$$

This same bound holds for $Z \sim \mathcal{D}$ where $\mathcal{D}$ is *sub-Gaussian with parameter* σ^2 i.e. $\psi_Z(\lambda) \leq \frac{\lambda^2\sigma^2}{2}$. It is not too difficult to check that $\mathrm{Var}[Z] \leq \sigma^2$ in this case. (💡: Taylor expand $e^{\lambda Z}$)

Claim 1.2 (Hoeffding's)**.** *Let X be such that $\mathbb{E}X = 0$ and $X \in [a, b]$ almost surely. Then $\psi_X''(\lambda) \leq \frac{(b-a)^2}{4}$. In particular X is sub-Gaussian with parameter $\frac{(b-a)^2}{4}$.*

That is, bounded RV's are sub-Gaussian. ("Obviously", since the tails become 0)

Proof. Just compute the double derivative: (first derivative is $\psi_X'(\lambda) = \frac{\mathbb{E}Xe^{\lambda X}}{\mathbb{E}e^{\lambda X}}$)

$$\begin{aligned}\psi_X''(\lambda) &= \frac{\mathbb{E}X^2e^{\lambda X}\mathbb{E}e^{\lambda X} - (\mathbb{E}Xe^{\lambda X})^2}{(\mathbb{E}e^{\lambda X})^2} \\ &= \int y^2(e^{-\psi_Y(\lambda)}d^{\lambda y})dP - \left(\int y(e^{-\psi_Y(\lambda)}d^{\lambda y})dP\right)^2\end{aligned}$$

and the desired conclusion follows (easy to upper bound variance of RV supported on $[a, b]$ by $\frac{(b-a)^2}{4}$) by considering the *tilited measure* $dP_\lambda = (e^{-\psi_Y(\lambda)}d^{\lambda y})dP$. The second part follows from Taylor expansion of $\psi_X(\lambda)$. □

Corollary 1.3 (Chernoff)**.** *Let X_i be independent random variables supported on $[a_i, b_i]$ for $1 \leq i \leq n$. Then for any $t > 0$,*

$$\mathbb{P}\left[\sum_{i=1}^n (X_i - \mathbb{E}X_i) \geq t\right] \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Theorem 1.4 (Bennett)**.** *Let X_i be independent random variables supported on $[-c_i, c_i]$ for $1 \leq i \leq n$. Suppose $\mathbb{E}[X_i] = 0$, $\mathrm{Var}[X_i] = \sigma_i^2$ and $\nu = \sum_o \sigma_i^2$. Then for $S = \sum_i X_i$, we have*

$$\mathbb{P}[S \geq t] \geq \exp\left(-\frac{t^2}{2(\nu + \max_i c_i t/3)}\right).$$

We incorporate variance information here, so when $\nu \ll \max_i c_i t$ we get better bounds than Chernoff.

(💡: Poisson tails. The whole point is to show that the tails of $\frac{X_i}{c}$ behave like $\mathrm{Poi}(\sigma_i^2/C) - \sigma_i^2$. The key fact is if $Y = Z - \xi$ for $Z \sim \mathrm{Poi}(\xi)$ then $\psi_Y^*(t) = \xi h_1(t/\xi)$ where $h_1(\cdot)$ is the binary entropy function.)

Proof. Taylor expand

$$\begin{aligned}
\psi_{X_i}(\lambda) = \mathbb{E}[e^{\lambda X_i}] &= \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \mathbb{E}[X_i^k] \\
&\le 1 + \sum_{k=2}^{\infty} \frac{\lambda^k}{k!} \mathbb{E}[c^{k-2} X_i^2] \\
&= 1 + \sum_{k=2}^{\infty} \frac{\lambda^k c^{k-2} \sigma_i^2}{k!} \\
&= 1 + \frac{\sigma_i^2}{c^2}(e^{\lambda c} - 1 - \lambda c) \\
&\le \exp\left(\frac{\sigma_i^2}{c^2} \left(e^{\lambda c} - 1 - \lambda c \right) \right).
\end{aligned}$$

In particular, let $Y = Z - \xi$ for $Z \sim \mathrm{Poi}(\frac{\nu^2}{c})$ so that $\psi_S(\lambda) \le \psi_Y(\lambda c)$. This in turn implies $\psi_S^*(t) \ge \frac{1}{c}\psi_Y^*(t) = \frac{\nu}{c^2} h_1\left(\frac{ct}{\nu^2}\right)$ and the final step is to use the numerical inequality $h_1(x) \ge \frac{x^2}{1+\frac{x}{3}}$. □

2 Tensorization of variance – Efron Stein

(💡) Let $Z = f(X_1, \dots, X_n)$ for i.i.d. X_i. We endeavor to write

$$Z - \mathbb{E}Z = \sum_i \Delta_i$$

for uncorrelated Δ_i so that $\mathrm{Var}[Z] \le \sum_i \mathbb{E}\Delta_i^2$ by "tensorization".

The most natural way to find Δ_i is to use a *Doob's martingale*: think of $\{X_i\}$ as a stream with X_i revealed at time i and we let Δ_i encode the additional information revealed at time i. Intuitively Δ_i and Δ_j are uncorrelated. Precisely, let $E_i[Z] := \mathbb{E}[Z \mid X_1, \dots, X_i]$ then for $\Delta_i = E_i[Z] - E_{i-1}[Z]$ we have the telescoping sum $Z - \mathbb{E}[Z] = \sum_i \Delta_i$ with $\mathbb{E}[\Delta_i \Delta_j] = \mathbb{E}[\Delta_i \mathbb{E}[\Delta_j \mid X_1, \dots X_i]] = 0$. This way, $\mathrm{Var}[Z] \le \sum_i \mathbb{E}\Delta_i^2 \le \sum_i \mathbb{E}[(Z - \mathbb{E}[Z \mid X_1, \dots, X_{i-1}, X_i, \dots, X_n])^2]$ because we can use Jensen's inequality and the identity $E_i[Z - \mathbb{E}[Z \mid X_1, \dots, X_{i-1}, X_i, \dots, X_n] = E_{i-1}[Z]$.

We simplify using $\mathrm{Var}[Y \mid X] = \mathbb{E}[(Y - \mathbb{E}[Y \mid X])^2 \mid X]$ so that we get

$$\mathrm{Var}[Z] \le \mathbb{E}\left[\sum_i \mathrm{Var}[Z \mid X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n] \right].$$

Finally, we also have

$$\begin{aligned}
\mathbb{E}\left[\sum_i \mathrm{Var}[Z \mid X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n] \right] &= \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}[Z - f(X_1, \dots, X_{i-1}, X_i', X_{i+1}, \dots, X_n))^2] \\
&= \mathbb{E} \inf_{Z_i : (X_1, \dots, X_{i-1}, X_{i+1}, \dots X_n)\text{-measurable}} \mathbb{E}[(Z - Z_i)^2]
\end{aligned}$$

where X_i' is an independent copy of X_i and the second equality is just the variational/least squares characterization of variance.

Theorem 2.1 (Bounded differences V1)**.** *Let X_i be independent random variables on Ω and suppose $f\colon \Omega^n \to \mathbb{R}$ satisfies*

$$\sup_{x_1,\dots,x_n,x_i'} |f(x_1,\dots,x_n) - f(x_1,\dots,x_i,x_{i+1},\dots,x_n)| \le c_i$$

for some $c_i > 0$ for each $i \in [n]$. Then

$$\mathrm{Var}[Z] \le \frac{1}{4}\sum_{i=1}^n c_i^2.$$

Remark 2.2. When applying bounded differences in graph settings, it is worthwhile to consider whether the edge exposure martingale or vertex exposure martingale gives better bounds.

Theorem 2.3 (Convex Poincaré)**.** *Let $X_1,\dots,X_n$ be independent random variables taking values in $[0,1]$ and let $f\colon [0,1]^n \to \mathbb{R}$ be a separately convex function whose partial derivatives exist. Then $Z = f(X_1,\dots,X_n)$ satisfies*

$$\mathrm{Var}[Z] \le \mathbb{E}\|\nabla f\|^2.$$

Proof. It suffices to note that

$$\sum_{i=1}^n \mathbb{E}[Z - \inf_{X_i} f(X_1,\dots,X_{i-1},X_i',X_{i+1},\dots,X_n))^2] \le \sum_{i=1}^n \mathbb{E}\left[\left(\frac{\partial f}{\partial x_i}(X_i)\right)^2 (X_i - X_i')^2\right] \le \mathbb{E}[\|\nabla f\|^2].$$

□

Example 2.4. Let $M \in \mathbb{R}^{n\times n}$ be a random matrix with entries M_{ij} independent and taking values on $[0,1]$. Then its largest singular value satisfies $\mathrm{Var}[\sigma_1(M)] \le 4$.

Theorem 2.5 (Gaussian Poincaré)**.** *Let $X = (X_1,\dots,X_n)$ be a vector of i.i.d. standard Gaussian random variables. Let $f\colon \mathbb{R}^n \to \mathbb{R}^n$ be any continuously differentiable function. Then $Z = f(X_1,\dots,X_n)$ satisfies*

$$\mathrm{Var}[Z] \le \mathbb{E}\|\nabla f\|^2.$$

We give two proofs of this, because making a mock qual for Liane made me actually read up about Markov semi-groups. The first proof illustrates *tensorization.* (💡: Efron-Stein is fundamentally about splitting an n-dimensional variance into a sum of 1-dimensional variances; so we can just "tensorize" the 1-dimensional bound)

Proof 1 of Theorem 2.5. As described above, it suffices to prove for $X \sim \mathcal{N}(0,1)$, we have

$$\mathrm{Var}[f(X)] \le \mathbb{E}[\|\nabla f(x)\|^2].$$

This 1-dimensional Gaussian Poincaré is really a central limit statement, in which let $\{\varepsilon_i\}_{i=1}^n$ be i.i.d. Rademacher random variables so that for $S_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i$ we have $f(S_n) \xrightarrow{n\to\infty} f(X)$. To finish up just Taylor expand

$$\left|f\left(S_n - \frac{\varepsilon_i}{\sqrt{n}} + \frac{1}{\sqrt{n}}\right) - f\left(S_n - \frac{\varepsilon_i}{\sqrt{n}} - \frac{1}{\sqrt{n}}\right)\right| \le \frac{2|f'(S_n)|}{\sqrt{n}} + \frac{2\sup_x |f''(x)|}{n}$$

and apply Efron-Stein. □

Before we give proof 2, we remind ourselves some notations regarding Markov semi-groups. Let (X_t) be a Markov process on some state space, let P_t be its semi-group and let L be the generator of this semi-group. Suppose X_t has an invariant measure μ.

Definition 2.6. The bilinear form $\mathcal{E}$, also known as the *Dirichlet form*, is defined to be

$$\mathcal{E}(f,g) := -(f, Lg) = -\int f L g d\mu.$$

We will always assume that (X_t) is reversible so that L is self-adjoint.

Definition 2.7. The covariance w.r.t. the invariant measure Cov_μ is defined to be

$$\mathrm{Cov}_\mu(f,g) := \int fg d\mu - \int f d\mu \int g d\mu.$$

Lemma 2.8 (Covariance Lemma)**.** $\mathrm{Cov}_\mu(f,g) = \int_0^\infty \mathcal{E}(f, P_t g) dt$.

We define the *Ornstein-Uhlenbeck process* given by the SDE:

$$dX_t = -X_t dt + \sqrt{2} dB_t.$$

Alternatively it is realized by the Brownian motion as

$$X_t = e^{-t} X_0 + e^{-t} B_{e^{2t}-1}.$$

It turns out that the generator and Dirichlet form for this process is given by $(Lf)(x) = f''(x) - xf'(x)$ and $\mathcal{E}(f,g) = \mathbb{E}_Z[f'(Z)g'(Z)]$. Furthermore the invariant distribution of the OU process is $\mathcal{N}(0,1)$.

Now generalize this to multiple dimensions the obvious way. Note that we also have the identity $\nabla(P_t g) = e^{-t}(P_t \nabla g)$ which is obvious from the BM representation of the OU process.

Proof 2 of Theorem 2.5. This literally follows from Lemma 2.8 and Cauchy-Schwarz, along with the fact that P_t is contracting in $L^2(\mu)$ which itself is a consequence of Jensen's. We just compute away:

$$\begin{aligned}
\mathrm{Cov}_\mu(f,g) &= \int_0^\infty \mathcal{E}(f, P_t g) dt \\
&= \int_0^\infty (\mathbb{E}_\mu[\nabla f \cdot \nabla(P_t g)] dt \\
&= \int_0^\infty (\mathbb{E}_\mu[\nabla f \cdot e^{-t}(P_t(\nabla g))] dt \\
&= \int_0^\infty e^{-t} \langle \nabla f, P_t(\nabla g) \rangle dt \\
&\le \int_0^\infty e^{-t} \|\nabla f\|_{L^2(\mu)} \|P_t(\nabla g)\|_{L^2(\mu)} dt \\
&\le \int_0^\infty e^{-t} \|\nabla f\|_{L^2(\mu)} \|\nabla g\|_{L^2(\mu)} dt \\
&= \|\nabla f\|_{L^2(\mu)} \|\nabla g\|_{L^2(\mu)}.
\end{aligned}$$

□

3 Tensorization of entropy – Entropy method

The aim of this section is to bootstrap what we saw in the previous section to get stronger tail bounds. To that end we will work with *entropy* instead of *variance*. In parallel with the Poincaré inequalities we will instead prove *log-Sobolev inequalities*.

3.1 Motivating examples of entropy tensorization: Han's inequality

Definition 3.1 (Shannon entropy)**.** Let X be a RV with distribution $P(X = x) = p(x)$. Then

$$H(X) := \mathbb{E}[-\log p(X)] = -\sum_x p(x) \log p(x).$$

We have the *chain rule*:

$$H(X_1, \dots, X_n) = \sum_{i=1}^n H(X_i \,|\, X_1, \dots, X_{i-1}).$$

Definition 3.2 (KL divergence)**.** Let P and Q be two probability measures on countable Ω. Then

$$D(Q \parallel P) = \sum_x q(x) \frac{q(x)}{p(x)}.$$

There is also a chain rule for KL-divergence; but to simplify notation we write $X_{1:i} = (X_1, \dots, X_i)$:

$$D(Q \parallel P) = \sum_{i=1}^n D(Q_{X_1 \,|\, X_{1:i-1}} \parallel P_{X_1 \,|\, X_{1:i-1}} \,|\, Q_{X_{1:i-1}}).$$

The chain rule illustrates tensorization. We have an "analogue" for Efron-Stein in these contexts as:

Theorem 3.3 (Han's inequality for Shannon entropy)**.** *For arbitrary discrete random variables $X_1, \dots, X_n$,*

$$H(X_{1:n}) \le \frac{1}{n-1} \sum_{i=1}^n H(X_1, \dots, X_{i-1}, X_{i+1}, X_n).$$

Theorem 3.4 (Han's inequality for KL-divergence)**.** *Let Ω be a countable set and let P and Q be* product *probability distributions on Ω^n. Then*

$$D(Q \parallel P) \ge \frac{1}{n-1} \sum_{i=1}^n D(Q_{X^{(i)}} \parallel P_{X^{(i)}}),$$

where $Q_{X^{(i)}} = Q_1 \otimes \cdots \otimes Q_{i-1} \otimes Q_{i+1} \otimes \cdots \otimes Q_n$.

The main idea in the proof of Theorem 3.3 is that conditioning only reduces entropy, which itself is basically Jensen's. And then we can deduce Theorem 3.4 from Theorem 3.3. Here's a more conceptual way to understand Theorem 3.3. Let's write Theorem 3.3 as

$$\sum_{i=1}^n H(X_{1:n}) - H(X^{(i)}) \le H(X_{1:n}).$$

Let us also define $f \colon \{0,1\}^n \to \mathbb{R}$ by $f(z_1, \dots, z_n) = H(X_i : z_i = 1)$ then we can artificially rewrite the above as

$$\sum_{i=1}^n (f(x_1, \dots, x_n) - \min_{x_i} f(x_1, \dots, x_n)) \le f(x_1, \dots, x_n) \tag{1}$$

which is reminiscent of the kind of self-bounding conditions arising in bounded differences. To recap, we can prove bounded differences for *self-bounding* functions i.e. $f \colon \Omega^n \to \mathbb{R}$ such that for each i there exists some $f_i \colon \Omega^{(i)} \to \mathbb{R}$ such that:

- $0 \le f(x) - f_i(x^{(i)}) \le 1$,
- $\sum_{i=1}^n (f(x_1, \dots, x_n) - f_i(x_1, \dots, x_n)) \le f(x_1, \dots, x_n)$.

It turns out that this kind of self-bounding condition is true for all *sub-modular functions* so we can in fact prove Han's for all sub-modular functions.

Lemma 3.5. *For a set $\mathcal{S}$ of random variables, let $H(\mathcal{S}) = H(X_i : i \in \mathcal{S})$ be the joint entropy. Then for any sets $\mathcal{A}, \mathcal{B}$ of random variables, the* sub-modularity *of H refers to the fact that*

$$H(A \cup B) + H(A \cap B) \leq H(A) + H(B).$$

Indeed to prove (1) just apply submodularity to $A = [i]$ and $B = [n] \setminus \{i\}$ and then we'll get a telescoping sum.

Example 3.6 (Isoperimetry on Boolean hypercube)**.** For any $A \subset \{0,1\}^n$ we have $|E(A, \overline{A})| \geq |A|(n - \log_2 |A|)$. (💡: Let $X \sim_{\mathsf{unif}} A$, then $H(X) - H(X^{(i)})$ tells us what happens along direction i; in particular we check $H(X) - H(X^{(i)}) = \frac{2}{|A|}\#\{\text{edges in direction } i\}$, then apply Han's.)

3.2 Bounded differences with exponential tails: Introduction to Herbst's argument

Definition 3.7. We define the *entropy* of a RV X as

$$\mathrm{Ent}[X] = \mathbb{E}[X \log X] - \mathbb{E}[X]\mathbb{E}[\log X].$$

Remark 3.8. This is related to KL divergence as follows: if Z is the *Radon-Nikodym derivative* of $Q \ll P$ and $X \sim P$, then $\mathrm{Ent}[Z] = D(Q \parallel P)$.

Lemma 3.9 (Entropy Tensorizes)**.** *Same set-up as before, so $X = f(X_{1:n})$. Then*

$$\mathrm{Ent}[Z] \leq \mathbb{E}\left[\sum_{i=1}^{n} \mathrm{Ent}^{(i)}[Z]\right].$$

Proof 1. We make use of Remark 3.8 so define a measure Q with pmf given by $q(x) = p(x)f(x)$ so that $D(Q \parallel P) = \mathrm{Ent}[Z]$. We can compute that Theorem 3.4 basically implies the tensorization we desire. □

Proof 2. Begin with the telescoping sum

$$\mathrm{Ent}[Z] = \sum_{i=1}^{n} \mathbb{E}[ZU_i]$$

where $U_i := \log \frac{\mathbb{E}[Z \mid X_1, \ldots, X_i]}{\mathbb{E}[Z \mid X_1, \ldots, X_{i-1}]}$ and then invoke the *dual characterization of entropy* which gives

$$\mathrm{Ent}[Z] = \sup_{X : \mathbb{E}[e^X] = 1} \mathbb{E}[ZX] \tag{2}$$

and this itself is a consequence of Young's inequality. (by homogeneity we may assume that $\mathbb{E}[Z] = 1$; now use $uv \leq u \log u - u + e^v$, so that

$$\mathbb{E}[ZX] \leq \mathbb{E}[Z \log Z] - 1 + \mathbb{E}[e^X] \leq \mathbb{E}[Z \log Z] = \mathrm{Ent}[Z]$$

and the converse follows by letting $Z_N = \min\{\max\{Z, N^{-1}\}, N\}$ and letting $X = \log \frac{Z_N}{\mathbb{E}Z_N}$.) □

Lemma 3.10 (Herbst's argument)**.** *Let Z be an integrable random variable such that for some $\nu > 0$, we have for every $\lambda > 0$,*

$$\mathrm{Ent}[e^{\lambda Z}] \leq \frac{\lambda^2 \nu}{2} \mathbb{E} e^{\lambda Z}.$$

Then, for every $\lambda > 0$,

$$\psi_{Z - \mathbb{E}Z}(\lambda) \leq \frac{\lambda^2 \nu}{2}.$$

Proof. We have that the log-Laplace transform expands $\psi_{Z-\mathbb{E}Z}(\lambda) = \log \mathbb{E}[e^{\lambda Z}] - \lambda \mathbb{E}[Z]$.

(💡) Check that

$$\frac{d}{d\lambda} \frac{\psi_{Z-\mathbb{E}[Z]}(\lambda)}{\lambda} = \frac{\mathrm{Ent}[e^{\lambda Z}]}{\lambda^2 \mathbb{E}[e^{\lambda Z}]}.$$

(Alternatively, $\lambda\psi_0'(\lambda) - \psi_0(\lambda) = \frac{\mathrm{Ent}[e^{\lambda Z}]}{\mathbb{E}[e^{\lambda Z}]}$ where $\psi_0(\lambda) = \psi_{Z-\mathbb{E}[Z]}(\lambda)$.)

L'Hopital's shows that $\frac{\psi_{Z-\mathbb{E}[Z]}(\lambda)}{\lambda} \to 0$ as $\lambda \to 0$. Now integrate: $\int_0^\lambda \frac{d}{d\lambda} \frac{\psi_{Z-\mathbb{E}[Z]}(\lambda)}{\lambda} d\lambda \le \int_0^\lambda \frac{\nu}{2}$. □

Theorem 3.11 (Bounded differences V2)**.** *Let X_i be independent random variables on Ω and suppose $f\colon \Omega^n \to \mathbb{R}$ satisfies*

$$\sup_{x_1,\dots,x_n,x_i'} |f(x_1,\dots,x_n) - f(x_1,\dots,x_i,x_{i+1},\dots,x_n)| \le c_i$$

for some $c_i > 0$ for each $i \in [n]$. Let $\nu = \frac{1}{4}\sum_{i=1}^n c_i^2$. Then for $t \ge 0$, we have

$$\mathbb{P}[|Z - \mathbb{E}Z| > t] \le e^{-\frac{t^2}{2\nu}}.$$

Proof.

- Lemma 3.10 implies we we need to bound $\frac{\mathrm{Ent}[e^{\lambda Z}]}{\mathbb{E}[e^{\lambda Z}]}$ which we can do so using Claim 1.2:

$$\frac{\mathrm{Ent}[e^{\lambda Z}]}{\mathbb{E}[e^{\lambda Z}]} = \lambda\psi_0'(\lambda) - \psi_0(\lambda) = \int_0^\lambda \theta\psi_0''(\theta)d\theta \le \frac{c_i^2\lambda^2}{8}.$$

- Tensorization of entropy $\implies$ tensorization of Laplace transforms. (💡) Plug in $e^{\lambda Z}$ for the random variable in Lemma 3.9.

$$\begin{aligned}
\mathbb{E}[e^{\lambda Z}] &\le \mathbb{E}\left[\sum_{i=1}^n \mathrm{Ent}^{(i)}[e^{\lambda Z}]\right] \\
&\le \mathbb{E}\left[\sum_{i=1}^n \frac{c_i^2\lambda^2}{4} E^{(i)}[e^{\lambda Z}]\right] \\
&\le \mathbb{E}[e^{\lambda Z}] \cdot \frac{\nu\lambda^2}{2}
\end{aligned}$$

- Apply Lemma 3.10 again and finish up with the Chernoff-Crámer argument.

□

3.3 log-Sobolev inequalities (LSI)

As in the case of the Poincaré inequality, we will give two proofs of the log-Sobolev inequality. After giving the semi-group proof we will also introduce a unified way to think about these inequalities.

Theorem 3.12 (LSI for Bernoulli RV)**.** *Let $X_1,\dots,X_n$ be i.i.d. Bernoulli RV. Then*

$$\mathrm{Ent}[f^2] \le 2\mathbb{E}[\|\nabla f\|^2].$$

Proof. Tensorize $\frac{a^2\log a^2}{2} + \frac{b^2\log b^2}{2} - \left(\frac{a^2+b^2}{2}\right)\log\left(\frac{a^2+b^2}{2}\right) \le \frac{(a-b)^2}{2}$. □

Theorem 3.13 (LSI for Gaussian RV)**.** *Let $X = (X_1,\dots,X_n)$ be a vector of n independent standard normal random variables. Then*

$$\mathrm{Ent}[f^2] \le 2\mathbb{E}[\|\nabla f\|^2].$$

Proof 1 of Theorem 3.13. Use CLT and combine with Theorem 3.12. □

First, recall that both variance and entropy are both special cases of *ϕ-entropy*.

Definition 3.14. For a given convex function ϕ and a random variable Z, we define the *ϕ-entropy* to be

$$\mathrm{Ent}^{\phi}[Z] = \mathbb{E}[\phi(Z)] - \phi(\mathbb{E}[Z]).$$

Lemma 3.15 (Semi-group expression for ϕ-entropies)**.** *Insert some technical conditions on ϕ and f to make everything well-defined:*

$$\mathrm{Ent}^{\phi}[f] = \int_0^{\infty} \mathbb{E}[\phi''(P_t f)|\nabla P_t f|^2]dt.$$

Proof. (💡) Define $\alpha(t) := \mathbb{E}[\phi(P_t f)]$. Note that $\alpha(0) = \mathbb{E}[\phi(f)]$ and

$$\alpha(\infty) = \mathbb{E}[\lim_{t\to\infty} \phi(P_t f)] = \phi(\mathbb{E}[f]).$$

This means $\mathrm{Ent}^{\phi}[f] = \alpha(0) - \alpha(\infty)$.

Finish up with chain rule and integration by parts:

$$\begin{aligned}
\alpha'(t) &= \mathbb{E}[\phi'(P_t f)L(P_t f)] \\
&= \mathbb{E}[\langle \nabla\phi'(P_t f), \nabla(P_t f)\rangle] \\
&= \mathbb{E}[\phi''(P_t f)|\nabla P_t f|^2].
\end{aligned}$$

□

Proof 2 of Theorem 3.13. When $\phi = x\log x$, note that Lemma 3.15 gives

$$\begin{aligned}
\mathrm{Ent}[f] &= \int_0^{\infty} \mathbb{E}\left[\frac{|\nabla P_t[f]|^2}{P_t[f]}\right] dt \\
&= \int_0^{\infty} e^{-2t}\mathbb{E}\left[\frac{|P_t[\nabla f]|^2}{P_t[f]}\right] dt \\
&\overset{\ddagger}{=} \int_0^{\infty} e^{-2t}\mathbb{E}\left[\frac{\left|P_t\left[\frac{\nabla f}{\sqrt{f}}\cdot\sqrt{f}\right]\right|^2}{P_t[f]}\right] dt \\
&\leq \int_0^{\infty} e^{-2t}\mathbb{E}\left[\frac{P_t\left[\frac{\|\nabla f\|^2}{f}\right]\cdot P_t[f]}{P_t[f]}\right] dt \\
&\leq \frac{1}{2}\mathbb{E}\left[\frac{\|\nabla f\|^2}{f}\right]
\end{aligned}$$

where the inequality follows from Cauchy-Schwarz. This implies that $\mathrm{Ent}[f^2] \leq 2\mathbb{E}[\|\nabla f\|^2]$. □

Remark 3.16. We can write both Poincaré's and LSI in a unified form as:

- $\mathrm{Var}[P_t f] \leq e^{-2t}\mathrm{Var}[f]$
- $\mathrm{Ent}[P_t f] \leq e^{-2t}\mathrm{Ent}[f]$.

Example 3.17. We can get the Gaussian concentration inequality for Lipschitz functions by combining Herbst's argument with Theorem 3.13 for $e^{\lambda Z/2}$. Precisely, if $f\colon \mathbb{R}^d \to \mathbb{R}$ is Lipschitz then for $Z = f(X_{1:n})$ we have

$$\mathbb{P}[|f(Z) - \mathbb{E}[Z]| \geq t] \leq 2\exp(-ct^2).$$

Here is another proof. First WLOG $\mathbb{E}[Z] = 0$. The goal is to show $\psi_Z(\lambda) \le C\lambda^2$. Let Y be an independent copy of X, then Jensen's gives $\mathbb{E}[e^{-\lambda f(y)}] \ge 1$ so that

$$\mathbb{E}[e^{\lambda f(X)}] \le \mathbb{E}[e^{\lambda(f(X)-f(Y))}].$$

Now, write $f(X) - f(Y) = \int_0^{\pi/2} \frac{d}{d\theta} f(Y\cos\theta + X\sin\theta)d\theta$. Now, write

$$\begin{aligned}\mathbb{E}[e^{\lambda(f(X)-f(Y))}] &\le \frac{2}{\pi}\int_0^{\pi/2} \mathbb{E}\left[\exp\left(\frac{\pi\lambda}{2}\frac{d}{d\theta}f(Y\cos\theta + X\sin\theta)\right)\right] \\ &\le \frac{2}{\pi}\int_0^{\pi/2} \mathbb{E}\left[\exp\left(\frac{\pi\lambda}{2}\nabla f(Y\cos\theta + X\sin\theta)\cdot(-Y\sin\theta + X\cos\theta)\right)\right]\end{aligned}$$

but $(-Y\sin\theta + X\cos\theta) \sim \mathcal{N}(0,1)^d$ for fixed $Y\cos\theta + X\sin\theta$ and furthermore $-Y\sin\theta + X\cos\theta$ is independent of $Y\cos\theta + X\sin\theta$. This gives the desired tails.

3.4 Modified log-Sobolev inequalities (MLSI)

The motivation for this is in Example 3.17 we actually only need LSI for $e^{\lambda Z/2}$. The MLSI shold be thought of as a tensorization specific to the Laplace transform.

Lemma 3.18 (Modified log-Sobolev inequality)**.** *Let $X_1, \dots, X_n$ be independent random variables, let $Z = f(X_1, \dots, X_n)$ and $Z_i = f_i(X^{(i)})$ where f_i is an arbitrary function on $X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n$, andlet $\phi(x) = e^x - x - 1$. Then for all $\lambda \in \mathbb{R}$, we have*

$$\mathrm{Ent}(e^{\lambda Z}) \le \sum_{i=1}^n \mathbb{E}[e^{\lambda Z}\phi(-\lambda(Z - Z_i))].$$

Remark 3.19. As states this MLSI will only be able to give us right tails and not left tails, because we will crucially always assume that $Z - Z_i > 0$ so that we can use the bound $\phi(-x) \le x^2/2$. That is, a working mathematicians' version of MLSI is

$$\mathrm{Ent}(e^{\lambda Z}) \le \lambda^2 \sum_{i=1}^n \mathbb{E}[e^{\lambda Z}(Z - Z_i)^2]. \tag{3}$$

We need to introduce another variational formulation of Ent: (i.e. expected value minimizes expected Bregman divergence which is $f(y) - f(x) - f'(x)(y-x)$)

$$\mathrm{Ent}[Z] = \inf_{u>0} \mathbb{E}[Z(\log Z - \log u) - (Z - u)]. \tag{4}$$

Proof of Lemma 3.18. It suffices to note that for $Y = e^{\lambda Z}$ and $Y_i = e^{\lambda Z_i}$, using (4) we have

$$\begin{aligned}\mathrm{Ent}^{(i)}[e^{\lambda Z}] &\le E^{(i)}[Y(\log Y - \log Y_i) - (Y - Y_i)] \\ &\le E^{(i)}[e^{\lambda Z}\lambda(Z - Z_i) - (e^{\lambda Z} - e^{\lambda Z_i})] \\ &= E^{(i)}[e^{\lambda Z}\phi(-(Z - Z_i))]\end{aligned}$$

and use tensorization of entropy. □

The first application is a stronger form of bounded difference.

Theorem 3.20 (Bounded Difference V3)**.** *Let $X_1, \dots, X_n$ be independent random variables and let $Z_i = \inf_{x_i} f(X_{1:i-1}, x_i, X_{i+1:n})$. Suppose there is $\nu > 0$ such that*

$$\sum_{i=1}^{n} (Z - Z_i)^2 \le \nu.$$

Then for all $t > 0$,

$$\mathbb{P}[Z - \mathbb{E}Z \ge t] \le e^{-t^2/2\nu}.$$

Example 3.21. Let $A = (a_{ij})$ be a symmetric $n \times n$ matrix with each a_{ij} being independent $\mathrm{Ber}(-1, +1)$. Let $\lambda_1(A)$ be the largest eigenvalue of A. Then

$$\mathbb{P}[\lambda_1(A) - \mathbb{E}\lambda_1(A) \ge t] \le e^{-\frac{t^2}{8}}.$$

Note that it is important here that we had an "average" type of bound for

$$v = \max_{x \in \{-1,+1\}^n} \sum_{i=1}^{n} (f(x) - f(\overline{x}^{(i)}))^2.$$

Perhaps a more explicit way to spell this out is to think of bounded difference V2 as giving concentration when the function f has bounded

$$\sum_{i=1}^{n} \max_{x \in \{\pm 1\}^n} (f(x_1, x_2, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n) - f(x_1, x_2, \dots, x_{i-1}, -x_i, x_{i+1}, \dots, x_n))^2$$

while this version gives concentration when f is bounded in terms of

$$\max_{x \in \{\pm 1\}^n} \sum_{i=1}^{n} (f(x_1, x_2, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n) - f(x_1, x_2, \dots, x_{i-1}, -x_i, x_{i+1}, \dots, x_n))^2$$

and the position of the max was critical so that we can use the same test vector (alá Courant-Fischer) for all the coordinates and the apply the Cauchy-Schwarz inequality.

Next, we bootstrap Theorem 2.3 to have sub-Gaussian tails.

Theorem 3.22 (Right tails of convex, 1-Lipschitz functions)**.** *Let $f \colon [0,1]^n \to \mathbb{R}$ be a separately convex, 1-Lipschitz function. Let $Z = f(X_1, \dots, X_n)$ where X_i are independent and supported on $[0,1]$. Then for all $t > 0$,*

$$\mathbb{P}[Z - \mathbb{E}Z \ge t] \le e^{-t^2/2}.$$

Proof. It suffices to note that by convexity of f we have

$$\sum_{i=1}^{n} (Z - Z_i)^2 \le \sum_{i=1}^{n} \left(\frac{d}{dx_i} f(X) \right)^2 = \|\nabla f\|^2 \le 1.$$

□

Remark 3.23. Naïvely using 1-Lipschitz alone would have given $\sum_{i=1}^{n} (Z - Z_i)^2 \le n$, and so convexity is crucial here. Furthermore, because we need $Z - Z_i \ge 0$ (why?), this method does not give left tails.

4 Transportation method

Recall Remark 3.8 and (2) which combined implies that we have the following variational characterization of log-Laplace transform/log MGF in terms of the KL-divergence:

$$\psi_{Z-\mathbb{E}[Z]}(\lambda) = \sup_{Q \ll P} [\lambda(\mathbb{E}_Q[Z] - \mathbb{E}_P[Z]) - D(Q \parallel P)].$$

Proof. In (2), let $Z = dQ/dP$ be the Radon-Nikodym derivative and let $U = \lambda(Z - \mathbb{E}[Z]) - \psi_{Z-\mathbb{E}[Z]}(\lambda)$. □

(💡) The key idea is that if for all $Q \ll P$ we have $\mathbb{E}_Q Z - \mathbb{E}_P Z \leq \sqrt{2\nu D(Q \parallel P)}$, then it follows that $\psi_{Z-\mathbb{E}_P[Z]}(\lambda) \leq \frac{\lambda^2 \nu}{2}$.

To prove $\mathbb{E}_Q Z - \mathbb{E}_P Z \leq \sqrt{2\nu D(Q \parallel P)}$ we need to find a good coupling between P and Q.

Example 4.1. Let P be a product measure on Ω^n, then for any $A, B \subset \Omega^n$ measurable we have that the Hamming distance between A and B satisfy

$$d_H(A, B) \leq \sqrt{\frac{n}{2} \cdot \frac{1}{P(A)}} + \sqrt{\frac{n}{2} \cdot \frac{1}{P(B)}}.$$