1 October 1, 2018

1.1 Martingales

We will mostly use martingales to get concentration results for Lipschitz functions $f(X_1, \ldots, X_n)$, where the X_i are iid random variables. We define

$$Z_k = \mathop{\mathbb{E}}_{X_{k+1},\ldots,X_n}[f(X_1, \ldots, X_n)] =: \Phi_k(X_1, \ldots, X_k)$$

Then the claim is that Z_k is a *martingale*, meaning that $\mathbb{E}[Z_{k+1} \mid Z_1, \ldots, Z_k] = Z_k$.

Proof. Recall that Z_k is some function, Φ_k, of the first k X variables. Then we have

$$\begin{aligned}
\mathbb{E}[Z_{k+1} \mid X_1, \ldots, X_k] &= \mathop{\mathbb{E}}_{X_{k+1}}[\Phi_{k+1}(X_1, \ldots, X_{k+1}) \mid X_1, \ldots, X_k] \\
&= \mathop{\mathbb{E}}_{X_{k+1}}\left[\mathop{\mathbb{E}}_{X_{k+2},\ldots,X_n}[f(X_1, \ldots, X_n) \mid X_1, \ldots, X_k]\right] \\
&= \mathop{\mathbb{E}}_{X_{k+1},\ldots,X_n}[f(X_1, \ldots, X_n) \mid X_1, \ldots, X_k] \\
&= \Phi_k(X_1, \ldots, X_k) = Z_k
\end{aligned}$$

□

Now, we assume that f is c_i-Lipschitz, meaning that for any $x_1, \ldots, x_n$ and any x_i', we have that

$$|f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_n)| \leq c_i$$

Proposition. *Under this assumption, $|Z_k - Z_{k-1}| \leq c_i$.*

Proof.

$$\begin{aligned}
Z_k - Z_{k-1} &= \mathop{\mathbb{E}}_{X_{k+1},\ldots,X_n}[f(X_1, \ldots, X_n)] - \mathop{\mathbb{E}}_{X_k',\ldots,X_n}[f(X_1, \ldots, X_k', \ldots, X_n)] \\
&= \mathop{\mathbb{E}}_{X_k',X_{k+1},\ldots,X_n}[f(X_1, \ldots, X_n) - f(X_1, \ldots, X_k', \ldots, X_n)]
\end{aligned}$$

Therefore,

$$|Z_k - Z_{k-1}| \leq \mathop{\mathbb{E}}_{X_k',X_{k+1},\ldots,X_n}[|f(X_1, \ldots, X_n) - f(X_1, \ldots, X_k', \ldots, X_n)|] \leq c_k$$

□

Theorem (Azuma's Inequality)**.** *For any martingale $0 = Z_0, Z_1, \ldots, Z_n$ with differences bounded as $|Z_k - Z_{k-1}| \leq c_k$ for all k, and for any $t > 0$, we have*

$$\Pr[Z_n \geq t] \leq e^{-\frac{t^2}{2\sum c_i^2}} \qquad \Pr[Z_n \leq -t] \leq e^{-\frac{t^2}{2\sum c_i^2}}$$

Proof. The proof is more or less the same as that of the Chernoff bound. By induction, we wish to prove that $\mathbb{E}[e^{\lambda Z_k}] \leq \exp(\frac{1}{2}\lambda^2 \sum_{i=1}^k c_i^2)$. Define the difference sequence $Y_k = Z_k - Z_{k-1}$. Then we have that

$$\begin{aligned}\mathop{\mathbb{E}}_{Y_1,\ldots,Y_{k+1}}\left[e^{\lambda Z_{k+1}}\right] &= \mathop{\mathbb{E}}_{Y_1,\ldots,Y_{k+1}}\left[e^{\lambda Z_k} e^{\lambda Y_{k+1}}\right] \\ &= \mathop{\mathbb{E}}_{Y_1,\ldots,Y_k}\left[\mathop{\mathbb{E}}_{Y_{k+1}}\left[e^{\lambda Z_k} e^{\lambda Y_{k+1}} \middle| Y_1,\ldots,Y_k\right]\right] \\ &= \mathop{\mathbb{E}}_{Y_1,\ldots,Y_k}\left[e^{\lambda Z_k} \mathop{\mathbb{E}}_{Y_{k+1}}\left[e^{\lambda Y_{k+1}} \middle| Y_1,\ldots,Y_k\right]\right]\end{aligned}$$

Now, we have that $|Y_{k+1}| \leq c_{k+1}$, and that $\mathbb{E}[Y_{k+1} \mid Y_1,\ldots,Y_k] = 0$. Therefore, as in the Chernoff bound,

$$\mathbb{E}\left[e^{\lambda Y_{k+1}} \middle| Y_1,\ldots,Y_k\right] \leq e^{\frac{1}{2}\lambda^2 c_{k+1}^2}$$

Therefore, continuing the above,

$$\mathop{\mathbb{E}}_{Y_1,\ldots,Y_{k+1}}\left[e^{\lambda Z_{k+1}}\right] \leq \mathop{\mathbb{E}}_{Y_1,\ldots,Y_k}\left[e^{\lambda Z_k} e^{\frac{1}{2}\lambda^2 c_{k+1}^2}\right] \leq e^{\frac{1}{2}\lambda^2 \sum_{i=1}^k c_i^2}$$

by induction. To prove Azuma's inequality, we now apply Markov's inequality to the exponential moment and optimize λ. □

An application of Azuma's inequality is the longest increasing subsequence problem. Let $X_i \in [0,1]$ be uniform and independent. The function f is the longest increasing subsequence of the X_i, namely

$$Z = f(X_1,\ldots,X_n) = \max\{k : \exists i_1 < i_2 < \cdots < i_k, X_{i_1} < \cdots < X_{i_k}\}$$

Then f is 1-Lipschitz. Therefore,

$$\Pr[|Z - \mathbb{E}[Z]| \geq t] \leq 2e^{-t^2/2n}$$

Recall the Erdős-Rényi model $G_{n,p}$, where each edge between n vertices appears independently with probability p.

Theorem (Shamir–Spencer)**.** *Let $Z = \chi(G_{n,p})$. Then*

$$\Pr[|Z - \mathbb{E}[Z]| \geq t] \leq e^{-t^2/2n}$$

Proof. The first attempt is to think of Z as a function of the $\binom{n}{2}$ random variables X_{ij}, where X_{ij} is the indicator random variable for the appearance of an edge between $i,j \in [n]$. Then f is 1-Lipschitz. But the number of variables is $\Theta(n^2)$, so the concentration we would get from this method would be of order $O(n)$, which is very poor for a function that takes values between 0 and n.

In order to do better, we will use the *vertex exposure martingale*. The key observation is that the modification of which edges are incident to a single vertex v can also not change the chromatic number by more than 1, since v can always

be assigned a new color. So for this martingale, at step i, we will reveal all the random variables X_{ji} for $j < i$. Then if we define

$$Z_k = \mathbb{E}[f(\vec{X})) \mid \{X_{ij} : 1 \leq i < j \leq k\}]$$

then we will get a martingale with 1-bounded differences. □

2 October 3, 2018

Whenever we define a sequence

$$Z_k = \mathop{\mathbb{E}}_{X_{k+1},\ldots,X_n}[f(X_1,\ldots,X_n) \mid X_1,\ldots,X_k]$$

it will always be a martingale (called the Doob martingale), regardless of how correlated the X_i are. However, independence is needed for the applications from last time, because if the X_i are not independent, then the property of bounded differences may fail. For instance, if all X_i are equal (i.e. perfectly correlated), and $\mathrm{Ber}(\frac{1}{2})$, and if we let f be the summation function, then $Z = \sum X_i$ will be either 0 or n with probability 1/2. We also have that $Z_0 = \mathbb{E}[Z] = n/2$, whereas $Z_1 = \mathbb{E}[Z \mid X_1] = Z$, so $|Z_1 - Z_0| = n/2$. Thus, even though f is 1-Lipschitz, we will have an arbitrarily large difference.

Generally speaking, concentration bounds stop holding when our variables are positively correlated. However, for negatively correlated variables, there are certain notions that allow one to prove concentration bounds. Here are some of the notions:

1. Pairwise negative correlation: for each i, j,

   $$\mathrm{Cov}(X_i, X_j) = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\,\mathbb{E}[X_j] \leq 0$$

   In this case, one can recover Chebyshev-type bounds for summation, since this property proves that $\mathrm{Var}(\sum X_i) \leq \sum \mathrm{Var}(X_i)$.

2. More generally, we can assume so-called cylinder negative dependence. For $\{0,1\}$-valued random variables, this means that for every set $S \subseteq [n]$,

   $$\mathbb{E}\left[\prod_{i\in S} X_i\right] \leq \prod_{i\in S} \mathbb{E}[X_i] \qquad \mathbb{E}\left[\prod_{i\in S}(1-X_i)\right] \leq \prod_{i\in S} \mathbb{E}[1-X_i]$$

   This suffices to get a Chernoff bound for $Z = \sum \alpha_i X_i$.

3. Negative association: we assume that for all disjoint $I, J \subseteq [n]$, and for any non-decreasing functions

   $$f : \{0,1\}^I \to \mathbb{R} \qquad g : \{0,1\}^J \to \mathbb{R}$$

   then

   $$\mathbb{E}[f(X_I)g(X_J)] \leq \mathbb{E}[f(X_I)]\,\mathbb{E}[g(X_J)]$$

It is conjectured that for any 1-Lipschitz of random variables satisfying negative association, we get Azuma-type bounds.

Such random variables appear naturally in matroids. For instance, if we pick a random spanning tree in a graph and define, for each edge, an indicator random variable for its appearance in the tree, then these variables satisfy negative association. In fact, they satisfy a stronger property, that they are *strongly Rayleigh* (which we won't define).

4. Pemantle–Peres '14 proved that 1-Lipschitz functions of strongly Rayleigh satisfy concentration with exponential tails.

5. Negative regression: for any disjoint $I, J \subseteq [n]$, and for any non-decreasing function $f : \{0,1\}^I \to \mathbb{R}$ and any $a \leq b \in \{0,1\}^J$,

$$\mathbb{E}[f(X_I) \mid X_J = a] \geq \mathbb{E}[f(X_I) \mid X_J = b]$$

Though this looks similar to negative association, no implications in either direction are known. Dubashi–Ranjan '96 claimed to prove exponential concentration for random variables satisfying negative regression, but there was a bug that was fixed by Garbe–Vondrák.

We now return to the coloring applications from last time.

Theorem (Bollobás)**.** *For $\alpha > 5/6$ and $p = n^{-\alpha}$, then there exists some u such that with high probability, $u \leq \chi(G_{n,p}) \leq u + 3$. In fact, it was later proved that one can get concentration on just two values.*

Proof. Set

$$u = \min\left\{a : \Pr(\chi(G_{n,p}) \leq a) > \frac{1}{n}\right\}$$

Define X to be the minimum number of vertices that need to be removed from G to make it u-colorable. We claim that X is 1-vertex-Lipschitz. Indeed, changing the edges incident to a single vertex v can only change X by 1, since we can always add v to the set of deleted vertices if necessary. Therefore,

$$\Pr[X \geq \mathbb{E}[X] + t] \leq e^{-t^2/2n}$$

We choose $t = \sqrt{2n \log n}$ so that $e^{-t^2/2n} = 1/n$. Therefore, we get that $\mathbb{E}[X] \leq t$, for otherwise the lower tail bound would contradict the definition of u:

$$\Pr(\chi(G) \leq u) = \Pr(X = 0) \leq \Pr(X \leq \mathbb{E}[X] - t) \leq 1/n$$

By the upper tail, we now get

$$\Pr(X \geq 2t) \leq \Pr(X \geq \mathbb{E}[X] + t) \leq \frac{1}{n}$$

Thus, with high probability, we can remove $O(\sqrt{n \log n})$ vertices to make G u-colorable. The remaining step is a simple combinatorial argument to show that these remaining vertices can be colored with only three colors.

Proposition. *For $G_{n,p}$ where $p = n^{-\alpha}$ with $\alpha > 5/6$, with high probability every subgraph on $\leq \sqrt{8n \log n}$ vertices can be 3-colored.*

Proof. We prove that with high probability, every such subgraph has $< 3t/2$ edges, which implies that the average degree in this subgraph is < 3, so a greedy 3-coloring algorithm will work. Counting the number of edges is a simple binomial approximation computation. □

□

3 October 5, 2018

3.1 The Efron–Stein Inequality

As before, let $Z = f(X_1, \ldots, X_n)$ where $X_1, \ldots, X_n$ are independent. Let

$$Z_k = \mathop{\mathbb{E}}_{X_{k+1},\ldots,X_n}[f(X_1, \ldots, X_n) \mid X_1, \ldots, X_k]$$

and

$$\Delta_k = Z_k - Z_{k-1}$$

Then as we saw,

$$\mathbb{E}[\Delta_k \mid X_1, \ldots, X_{k-1}] = 0$$

We can also compute

$$\begin{aligned}
\mathrm{Var}(Z) &= \mathbb{E}[(Z - \mathbb{E}[Z])^2] \\
&= \mathbb{E}[(Z_n - Z_0)^2] \\
&= \mathbb{E}\left[\left(\sum_{k=1}^{n} \Delta_k\right)^2\right] \\
&= \sum_{i=1}^{n} \mathbb{E}[\Delta_i^2] + 2\sum_{i<j} \mathbb{E}[\Delta_i \Delta_j]
\end{aligned}$$

Observe that for $i < j$,

$$\begin{aligned}
\mathbb{E}[\Delta_i \Delta_j] &= \mathop{\mathbb{E}}_{X_1,\ldots,X_i}\left[\mathop{\mathbb{E}}_{X_{i+1},\ldots,X_n}[\Delta_i \Delta_j | X_1, \ldots, X_i]\right] \\
&= \mathop{\mathbb{E}}_{X_1,\ldots,X_i}\left[\Delta_i \mathop{\mathbb{E}}_{X_{i+1},\ldots,X_n}[\Delta_j | X_1, \ldots, X_i]\right] \\
&= 0
\end{aligned}$$

Therefore, we can conclude the following result:

Lemma.

$$\mathrm{Var}(Z) = \sum_{i=1}^{n} \mathbb{E}[\Delta_i^2]$$

However, this is frequently not so helpful, because understanding Δ_i might be rather difficult; as an example, think back to the longest increasing subsequence example from before. The Efron–Stein inequality is often more helpful:

Theorem (Efron–Stein)**.** *Let $Z = f(X_1, \dots, X_n)$, where $X_1, \dots, X_n$ are independent. Then*

$$\mathrm{Var}(Z) \leq \sum_{i=1}^{n} \mathbb{E}\left[\left(Z - \mathop{\mathbb{E}}_{X_i}[Z \mid X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n]\right)^2\right]$$

Example. Suppose $X_1, \dots, X_n$ are iid $\mathrm{Ber}(\frac{1}{2})$ variables, and let Z be their sum mod 2. Then Z is also $\mathrm{Ber}(\frac{1}{2})$, and thus $\mathrm{Var}(Z) = \frac{1}{4}$. For any i,

$$\mathop{\mathbb{E}}_{X_i}[Z] = \frac{1}{2}$$

Therefore,

$$\mathbb{E}\left[\left(Z - \mathop{\mathbb{E}}_{X_i}[Z]\right)^2\right] = \frac{1}{4}$$

and thus the bound Efron–Stein gives in this case is $n/4$, which is pretty bad. We will soon see some examples where it's very good

Proof of Efron–Stein. We know that $\mathrm{Var}(Z) = \sum \mathbb{E}[\Delta_i^2]$. We have that

$$\Delta_i = Z_i - Z_{i-1} = \mathop{\mathbb{E}}_{X_{i+1}, \dots, X_n}\left[Z - \mathop{\mathbb{E}}_{X_i}[Z]\right]$$

where we've stopped explicitly writing the conditioning; note that this step fails if the X_i are dependent. Given $X_1, \dots, X_i$, we apply Jensen's inequality (or Cauchy–Schwarz) to get that

$$\Delta_i^2 \leq \mathop{\mathbb{E}}_{X_{i+1}, \dots, X_n}\left[\left(Z - \mathop{\mathbb{E}}_{X_i}[Z]\right)^2\right]$$

Therefore,

$$\mathop{\mathbb{E}}_{X_1, \dots, X_n}[\Delta_i^2] \leq \mathop{\mathbb{E}}_{X_1, \dots, X_n}\left[\left(Z - \mathop{\mathbb{E}}_{X_i}[Z]\right)^2\right]$$

□

There are several equivalent forms of the Efron–Stein inequality. For instance, if we let $Z_i' = f(X_1, \dots, X_{i-1}, X_i', X_{i+1}, \dots, X_n)$, where X_i' is an independent copy of X_i. Then the Efron–Stein inequality is equivalent to

$$\mathrm{Var}(Z) \leq \sum_{i=1}^{n} \mathbb{E}[(Z - Z_i)^2]$$

Equivalently,

$$\mathrm{Var}(Z) \leq \sum_{i=1}^{n} \mathbb{E}[(Z - Z_i')_+^2] = \sum_{i=1}^{n} \mathbb{E}[(Z - Z_i')_-^2]$$

where $(x)_+ = \max\{x, 0\}$ and $(x)_- = -\min\{x, 0\}$. Now, let Z_i be any random variable which is a function of $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$ only. Then

$$\mathrm{Var}(Z) \leq \sum_{i=1}^{n} \mathbb{E}[(Z - Z_i)^2]$$

This is because, by the variational property of the conditional expectation (namely that it is a projection operator), we see that the variable Z_i that would minimize the right-hand side is precisely the conditional expectation.

Let's return to the longest increasing subsequence problem, and use this last version. We will choose Z_i to be

$$Z_i = \min_{X_i} f(X_1, \ldots, X_n)$$

where f is the LIS function. If $X_1, \ldots, X_n$ are fixed, and $X_{i_1} < X_{i_2} < \cdots < X_{i_k}$ be an LIS. Then we observe first that $Z_i = Z$ for any $i \notin \{i_1, \ldots, i_k\}$, since changing the value of X_i can only increase the length of the LIS. Moreover, changing X_i for any i can only change Z by at most 1. So

$$\sum_{i=1}^{n} (Z - Z_i)^2 = \sum_{j=1}^{k} (Z - Z_{i_j})^2 \leq k$$

Note that the bound we get depends on the actual value of Z, since k is the length of the LIS for this choice of variables. Now, Efron–Stein tells us that

$$\mathrm{Var}(Z) \leq \mathbb{E}\left[\sum_{i=1}^{n} (Z - Z_i)^2\right] \leq \mathbb{E}[Z]$$

Thus, Chebyshev already tells us that the LIS is concentrated within $\pm O(n^{1/4})$, since we know that $\mathbb{E}[Z] = \Theta(\sqrt{n})$. This example leads us to the notion of self-bounding functions.

Definition. A function $f : \mathcal{X}^n \to \mathbb{R}_+$ is called self-bounding if there exist functions $f_i : \mathcal{X}^{n-1} \to \mathbb{R}_+$ such that

1. $\forall x \in \mathcal{X}^n$,

$$0 \leq f(x_1, \ldots, x_n) - f_i(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) \leq 1$$

2. $\forall x \in \mathcal{X}^n$,

$$\sum_{i=1}^{n} (f(x_1, \ldots, x_n) - f_i(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)) \leq f(x_1, \ldots, x_n)$$

Corollary (of Efron–Stein)**.** *If $Z = f(X_1, \ldots, X_n)$ where X_i are independent and f is self-bounding, then*

$$\mathrm{Var}(Z) \leq \mathbb{E}[Z]$$

The proof is the same as the one we did for the LIS, since the only properties we used there were those of self-bounding functions.

Many examples of self-bounding functions, including the LIS function, are so-called "configuration functions." These are functions that measure the largest size of some substructure among your variables. More formally a configuration function is a function of the form

$$f(x_1, \ldots, x_n) = \max\{k : \exists i_1 < i_2 < \cdots < i_k, (X_{i_1}, \ldots, X_{i_k}) \in P\}$$

where P is some set of patterns.

4 October 8, 2018

Recall the Efron–Stein inequality says

$$\mathrm{Var}(Z) \leq \sum_{i=1}^{n} \mathbb{E}\left[\left(Z - \mathop{\mathbb{E}}_{X_i}[Z]\right)^2\right]$$

where $Z = f(X_1, \ldots, X_n)$, and the $X_i$are independent.

We also defined self-bounding functions. $f : \mathcal{X}^n \to \mathbb{R}$ is self-bounding if there exist $f_i : \mathcal{X}^{n-1} \to \mathbb{R}$ with

$$0 \leq f(x) - f_i(x_{-i}) \leq 1$$

and

$$\sum_{i=1}^{n} (f(x) - f_i(x_{-i})) \leq f(x)$$

where $x_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$. We showed as a consequence of Efron–Stein that $\mathrm{Var}(Z) \leq \mathbb{E}[Z]$ if $Z = f(X_1, \ldots, X_n)$ where f is self-bounding.

Example. The VC (Vapnik–Chervonenkis) dimension is defined as follows. Given a space $\mathcal{X}$ and a family of subsets $\mathcal{A} \subseteq 2^{\mathcal{X}}$, we define

$$\mathrm{VC}_{\mathcal{A}}(x_1, \ldots, x_n) = \max\{|S| : \exists S \subseteq \{x_1, \ldots, x_n\} : S \text{ is shattered by } \mathcal{A}\}$$

where we say that S is shattered by $\mathcal{A}$ if for all $T \subseteq S$, there is some $A \in \mathcal{A}$ such that $A \cap S = T$. The VC dimension is very important in machine learning, where it captures a notion of complexity of the family $\mathcal{A}$.

Now assume that $x_1, \ldots, x_n$ are independently random in $\mathcal{X}$, and define $Z = \mathrm{VC}_{\mathcal{A}}(x_1, \ldots, x_n)$. Then we claim that Z is tightly concentrated. To see this, we claim that $\mathrm{VC}_{\mathcal{A}}$ is a self-bounding function (in fact, it's a configuration function, as defined last time). Indeed, if the maximal shattered set is $S \subseteq \{x_1, \ldots, x_n\}$, then we can only decrease Z by modifying of the $x_i \in S$, which suffices to get the self-bounding property, as we saw last time. Thus, we get that $\mathrm{Var}(Z) \leq \mathbb{E}[Z]$.

Example. We will now see an example of a self-bounding function that is not a configuration function, called Rademacher averages. Suppose we are given independently random vectors $\overrightarrow{X_1}, \ldots, \overrightarrow{X_n} \in [-1,1]^d$ (with any distributions). We are also given independently uniformly random $\varepsilon_1, \ldots, \varepsilon_n \in \{-1, +1\}$. We are interested in

$$Z = f\left(\overrightarrow{X_1}, \ldots, \overrightarrow{X_n}\right) = \mathop{\mathbb{E}}_{\varepsilon_1,\ldots,\varepsilon_n}\left[\max_{1\le j\le d} \sum_{i=1}^{n} \varepsilon_i X_{i,j}\right]$$

If $d = 1$ and the distribution of the X_i is uniformly ± 1, then we are just doing a random walk, so we get $\mathbb{E}[Z] = \Theta(\sqrt{n})$. For $d > 1$, if the vectors behave independently in each dimension, then we get $\mathbb{E}[Z] = \Theta(\sqrt{n}\log d)$, by a union bound.

We claim that f is self-bounding, so that $\mathrm{Var}(Z) \le \mathbb{E}[Z]$. Rewriting, we have that

$$f\left(\overrightarrow{X_1}, \ldots, \overrightarrow{X_n}\right) = \mathop{\mathbb{E}}_{\varepsilon_1,\ldots,\varepsilon_n}\left[\max_{1\le j\le d} \sum_{i=1}^{n} \varepsilon_i X_{i,j}\right]$$

Therefore, it makes sense to define

$$Z_i = f_i(\overrightarrow{X_1}, \ldots \overrightarrow{X_{i-1}}, \overrightarrow{X_{i+1}}, \ldots, X_n) = \mathop{\mathbb{E}}_{\varepsilon_1,\ldots,\varepsilon_n}\left[\max_{1\le j\le d} \sum_{\ell\neq i} \varepsilon_\ell X_{\ell,j}\right]$$

With this definition,

$$Z - Z_i = \mathop{\mathbb{E}}_{\varepsilon_1,\ldots,\varepsilon_n}\left[\max_j \sum_i \varepsilon_i X_{i,j} - \max_{j'} \sum_{\ell\neq i} \varepsilon_\ell X_{\ell,j}\right]$$

Denote by $j^*(\vec{\varepsilon})$ the maximizing j in the first expression, namely

$$j^*(\vec{\varepsilon}) = \arg\max_j \sum_i \varepsilon_i X_{i,j}$$

Then we have

$$Z - Z_i \le \mathop{\mathbb{E}}_{\varepsilon_1,\ldots,\varepsilon_n}\left[\sum_i \varepsilon_i X_{i,j^*(\vec{\varepsilon})} - \sum_{\ell\neq i} \varepsilon_\ell X_{\ell,j^*(\vec{\varepsilon})}\right] = \mathbb{E}[\varepsilon_i X_{i,j^*(\vec{\varepsilon})}] \le 1$$

Moreover,

$$\sum_i (Z - Z_i) \le \sum_i \mathop{\mathbb{E}}_{\varepsilon}[\varepsilon_i X_{i,j^*(\vec{\varepsilon})}] = Z$$

So the final thing to check is that $Z - Z_i \ge 0$. For this, we can pick the maximizing j' for the second term in $Z - Z_i$.

Example. Sometimes, the function you're interested in isn't self-bounding on the nose, but its variance can still be well bounded by applying the Efron–Stein inequality directly. One example is the maximum eigenvalue of a random symmetric matrix; for $i \leq j$, let X_{ij} be independently random in $[-1, 1]$ with variance σ^2, and set $X_{ji} = X_{ij}$ for $i > j$. We have Wigner's semicircle law, which says that as $n \to \infty$, the eigenvalues will have a density given by a semicircle, with the maximum eigenvalue around $2\sigma\sqrt{n}$. We want to understand how tightly concentrated it is.

Let Z be the maximum eigenvalue, which we can equivalently write as

$$Z = \sup_{\|u\|=1} u^T X u$$

Then Z will not be self-bounding. However, Efron–Stein tells us that

$$\mathrm{Var}(Z) \leq \sum_{i \leq j} \mathbb{E}[(Z - Z'_{ij})_+^2]$$

where Z'_{ij} is Z, but where X_{ij} has been replaced by an independent sample X'_{ij}. Let v be the top eigenvector of X, so that $Z = v^T X v$. Thus, we have that

$$(Z - Z'_{ij})_+ \leq \left(v^T X v - v^T X'_{(i,j)} v\right)_+ = 2|v_i v_j||X_{ij} - X'_{ij}|$$

Therefore, by Efron–Stein,

$$\mathrm{Var}(Z) \leq \sum_{i \leq j} \mathbb{E}[(Z - Z'_{ij})_+^2] \leq \sum_{i \leq j} (4|v_i v_j|)^2 \leq 16 \left(\sum_i v_i^2\right)^2 = 16$$

Thus, the maximal eigenvalue is concentrated in a constant-sized window (at least with polynomial tails, since we only get a variance bound).

Example (First-passage percolation)**.** Fix a graph G (usually just a grid) and two vertices u, v. We pick iid random edge weights $X_e \geq 0$ with $\mathbb{E}[X_e^2] = \sigma^2$. Our function Z is the weight of the lightest path from u to v, namely

$$Z = \inf_{P: u \to v \text{ path}} \sum_{e \in P} X_e$$

There are a lot of open questions, and even understanding $\mathbb{E}[Z]$ is hard. But for now, we will focus on $\mathrm{Var}(Z)$. From now on assume that G is finite, so that the infimum is in fact a minimum. Note that Z can increase only by modifying X_e for some $e \in P^*$, where P^* is the minimizing path for the assignment X_e. So define Z'_e like Z, except that X_e has been replaced by an independent copy X'_e. Since we can bound the way Z can increase, it makes sense to look at $(Z - Z'_e)_-$. We have that

$$|(Z - Z'_e)_-| \leq |(X_e - X'_e)_-| = (X'_e - X_e)_+ \leq X'_e$$

By Efron–Stein,

$$\begin{aligned}\mathrm{Var}(Z) &\le \sum_{e\in E} \mathbb{E}[(Z - Z'_e)^2_-] \\ &\le \mathbb{E}\left[\sum_{e\in P^*} (X'_e)^2\right] \\ &= \sigma^2\, \mathbb{E}[|P^*|]\end{aligned}$$

where we used the fact that P^* is independent of X'_e.

5 October 10, 2018

5.1 Information Theory

For now, we will deal only with discrete random variables, namely ones that take values in a countable set $\mathcal{X}$.

Definition. If X is a random variable with $\Pr(X = x) = p(x)$, then its *Shannon entropy* is defined as

$$H(X) = \sum_{x\in\mathcal{X}} p(x) \log \frac{1}{p(x)} = \mathop{\mathbb{E}}_{X}\left[\log \frac{1}{p(x)}\right]$$

with the convention $0 \log 0 = 0$.

The entropy measures, in some sense, the complexity of the distribution of X, or the amount of space one needs in an optimal encoding of X. One important property is that H is a concave functional; specifically, if X, Y are random variables and Z takes the value of X with probability α and the value of Y with probability $1 - \alpha$, then

$$H(Z) \ge \alpha H(X) + (1 - \alpha)H(Y)$$

This follows from the convexity of the function $x \mapsto x \log x$.

Definition. The *relative entropy*, also called the *Kullback–Leibler divergence*, is defined by

$$D(P\|Q) = \sum_{\substack{x\in\mathcal{X} \\ p(x)>0}} p(x) \log \frac{p(x)}{q(x)}$$

One interpretation is that if we use the optimal coding scheme for Q and use it to encode P, then $D(P\|Q)$ measures the extra wasted number of bits per symbol.

Proposition. *$D(P\|Q) \ge 0$ for all P, Q.*

Proof. Using the fact that $\log t \leq t - 1$, we get that

$$\begin{aligned} D(P\|Q) &= -\sum_{x:p(x)>0} p(x) \log \frac{q(x)}{p(x)} \\ &\geq -\sum_{x:p(x)>0} p(x) \left(\frac{q(x)}{p(x)} - 1 \right) \\ &= \sum_{x:p(x)>0} p(x) - \sum_{x:p(x)>0} q(x) \\ &\geq 0 \end{aligned}$$

Note that this proof also gives us that $D(P\|Q) = 0$ iff $P = Q$. □

Consider $\mathcal{X}$ finite and Q uniform, so that $q(x) = 1/|\mathcal{X}|$. Then in this setting,

$$D(P\|Q) = \sum p(x) \log(|\mathcal{X}|p(x)) = \log|\mathcal{X}| - H(X)$$

By non-negativity, we get that for any X, $H(X) \leq \log|\mathcal{X}|$. Thus, the uniform distribution maximizes entropy on a finite domain.

Given two random variables X, Y (that might be dependent), we define $H(X, Y)$ in the natural way, wherein we think of (X, Y) as a random variable on the domain $\mathcal{X} \times \mathcal{Y}$. The *mutual information* between X and Y is then defined as

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

It measures how much information one learns about Y from learning X (or vice versa). In particular, if $X = Y$, then $H(X, Y) = H(X)$, so $I(X;Y) = H(X)$. On the other hand, if X and Y are independent, then $H(X, Y) = H(X)+H(Y)$, so $I(X;Y) = 0$.

We can get another formula for $I(X;Y)$ as

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p_X(x)p_Y(y)} = D(P_{X,Y}\|P_X \otimes P_Y)$$

where $P_X \otimes P_Y$ is the distribution given by the denominator, which is just the distribution on the product set if we sample independent copies of X and Y. From this expression, we find that $I(X;Y) \geq 0$, since we expressed it as a KL divergence. Thus, entropy is subadditive:

$$H(X,Y) \leq H(X) + H(Y)$$

We can also define the *conditional entropy* by

$$H(X \mid Y) = H(X,Y) - H(Y)$$

Note that by the above, $H(X \mid Y) \leq H(X)$, so conditioning can only decrease entropy. Additionally, $H(X \mid Y) \geq 0$, since

$$H(X,Y) = \sum p(x,y) \log \frac{1}{p(x,y)} \geq \sum p(x,y) \log \frac{1}{\sum_x p(x,y)} = H(Y)$$

Consider

$$\begin{aligned} H(X,Y,Z)+H(Z) - H(X,Z) - H(Y,Z) = \\ &= \sum_{x,y,z} p(x,y,z) \log \frac{p_{X,Z}(x,z)p_{Y,Z}(y,z)}{p_{X,Y,Z}(x,y,z)p_Z(z)} \\ &= \sum_z p_Z(z) \sum_{x,y} p(x,y \mid z) \frac{\Pr(X=x \mid Z=z)\Pr(Y=y \mid Z=z)}{\Pr(X=x, Y=y \mid Z=z)} \\ &= -\sum_z I(X;Y \mid Z=z) \\ &\leq 0 \end{aligned}$$

where to go from the second to third lines we divide by $p_Z(z)^2$ in both numerator and denominator. For a set S of random variables, let $\widetilde{H}(S) = H(X_i : i \in S)$ be the joint entropy for the variables in S. Then if we are given variables $X_1, \ldots, X_k, Y_1, \ldots, Y_m, Z_1, \ldots, Z_\ell$, and letting A be the set of the X and Z variables, and B the set of Y and Z variables, we get from the above

$$\widetilde{H}(A \cup B) + \widetilde{H}(A \cap B) \leq \widetilde{H}(A) + \widetilde{H}(B)$$

This is called *submodularity* of entropy.

Lemma (Han's inequality)**.** *For random variables $X_1, \ldots, X_n$,*

$$\sum_{i=1}^n H(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) \geq (n-1)H(X_1, \ldots, X_n)$$

Equivalently,

$$\sum_{i=1}^n (H(X_1, \ldots, X_n) - H(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)) \leq H(X_1, \ldots, X_n)$$

Note that the latter expression is essentially the same as the defining property for self-bounding functions. Specifically, suppose we define a function $\widetilde{f}(z_1, \ldots, z_n)$ for $z_i \in \{0,1\}$ by

$$\widetilde{f}(z_1, \ldots, z_n) = H(X_i : z_i = 1)$$

Then we precisely get from the above that

$$\sum_{i=1}^n \left(\widetilde{f}(z_1, \ldots, z_n) - \min_{z_i} \widetilde{f}(z_1, \ldots, z_n) \right) \leq \widetilde{f}(z_1, \ldots, z_n)$$

Han's inequality is a consequence of the submodularity of the entropy. In fact, more generally, any non-negative non-decreasing submodular function has this

self-bounding property. Indeed, by restricting to the non-zero variables, we want to prove that

$$\sum_{i=1}^{n}(f([n]) - f([n] \setminus \{i\})) \leq f([n])$$

By submodularity applied to the sets $A = [i], B = [n] \setminus \{i\}$, we have that

$$\begin{aligned}\sum_{i=1}^{n}(f([n]) - f([n] \setminus \{i\})) &\leq \sum_{i=1}^{n}(f([i]) - f([i-1])) \\ &= f([n]) - f(\emptyset) \\ &\leq f([n])\end{aligned}$$

6 October 12, 2018

Remember that for a collection fo random variables $X_1, \ldots, X_n$ and for $S \subseteq [n]$, we defined $\widetilde{H}(S) = H(X_i : i \in S)$. We showed that $\widetilde{H}$ is submodular, where $f : 2^{[n]} \to \mathbb{R}$ is called submodular if for all S, T,

$$f(S \cup T) + f(S \cap T) \leq f(S) + f(T)$$

For the entropy, submodularity is equivalent to the fact that for variables X, Y, Z,

$$H(X, Y, Z) + H(Y) \leq H(X, Y) + H(Y, Z)$$

or equivalently

$$H(X, Z \mid Y) \leq H(X \mid Y) + H(Z \mid Y)$$

We also saw that submodularity implies the self-bounding property. Precisely, for any $f : 2^{[n]} \to \mathbb{R}_+$ that is submodular and non-decreasing, and for any $S \subseteq [n]$,

$$\sum_{i \in S}(f(S) - f(S \setminus \{i\})) \leq f(S)$$

We proved this by applying submodularity to $S = [i], T = [n] \setminus \{i\}$.

For the entropy, this property is called Han's inequality, and is usually written in the form

$$\sum_{i=1}^{n} H(X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) \geq (n-1)H(X_1, \ldots, X_n)$$

Definition (VC entropy)**.** Suppose $\mathcal{A} \subseteq 2^{\mathcal{X}}$ is a set system. Define the trace by

$$\mathrm{tr}_{\mathcal{A}}(x_1, \ldots, x_n) = \{A \cap \{x_1, \ldots, x_n\} : A \in \mathcal{A}\}$$

and then define the VC entropy by

$$h_{\mathcal{A}}(x_1, \ldots, x_n) = \log_2 |\mathrm{tr}_{\mathcal{A}}(x_1, \ldots, x_n)|$$

Note that if $x_1, \ldots, x_n$ are shattered, then $|\mathrm{tr}_{\mathcal{A}}(x_1, \ldots, x_n)| = n$, and thus $h_{\mathcal{A}}$ is always an upper bound for $\mathrm{VC}_{\mathcal{A}}$.

Lemma. *$h_{\mathcal{A}}$ is self-bounding.*

Proof. Consider the space $\{0,1\}^n$, and view $\mathrm{tr}_{\mathcal{A}}(x)$ as a subset of $\{0,1\}^n$. Let $Y = (Y_1, \ldots, Y_n) \in \{0,1\}^n$ be uniformly random on $\mathrm{tr}_{\mathcal{A}}(x)$. Since Y is uniform, we have that $H(Y)$ is log of the size of its support, namely $\log |\mathrm{tr}_{\mathcal{A}}(x)|$. So $H(Y) = (\log 2) h_{\mathcal{A}}(x)$.

Define $Y^{(i)} = (Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_n)$. Observe that $Y^{(i)}$ is supported on $\mathrm{tr}_{\mathcal{A}}(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$, though it may no longer be uniform on this support. However, we still get

$$H(Y^{(i)}) \leq \log |\mathrm{tr}_{\mathcal{A}}(x^{(i)})|$$

where $x^{(i)} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$. By Han's inequality,

$$\log |\mathrm{tr}_{\mathcal{A}}(x)| = H(Y) \geq \sum_{i=1}^{n} (H(Y) - H(Y^{(i)})) \geq \sum_{i=1}^{n} (\log |\mathrm{tr}_{\mathcal{A}}(x)| - \log |\mathrm{tr}_{\mathcal{A}}(x^{(i)})|)$$

Thus, $h_{\mathcal{A}}$ is self-bounding. □

In fact, this is even more general. Recall that the VC dimension is a configuration function, meaning that it measures the largest occuring pattern. Given any configuration function (with the property that the allowed patterns are downwards closed), we can turn it into a combinatorial entropy function, which measures the log of the number of occuring patterns. When we do this, we always get a self-bounding function.

6.1 Isoperimetry

Using Han's inequality, we can prove an isoperimetric inequality on the hypercube. We view the hypercube $\{0,1\}^n$ as a graph in the natural way, where two vertices are adjacent if their Hamming distance is 1. For a subset $A \subseteq \{0,1\}^n$, let $E(A)$ be the set of edges inside A, and $E(A, \overline{A})$ the number of edges between A and its complement. The isoperimetric question asks for the minimum value of $|E(A, \overline{A})|$ given $|A|$.

Theorem. *For any $A \subseteq \{0,1\}^n$,*

$$|E(A)| \leq \frac{1}{2} |A| \log_2 |A|$$

Therefore, by adding up all edges incident to A,

$$|E(A, \overline{A})| \geq |A|(n - \log_2 |A|)$$

This bound is tight, since if A is a subcube of dimension d, then $|A| = 2^d$ and $|E(A)| = \frac{1}{2} d \cdot 2^d$.

Proof. Let $X = (X_1, \ldots, X_n)$ be uniformly random on A, so that $H(X) = \log |A|$. As usual, let $X^{(i)}$ be X with the ith coordinate removed. The idea is that $H(X) - H(X^{(i)})$ precisely tells us what's going on in A with respect to the ith direction. More formally, we have that the probability density of X is $p(x) = 1/|A|$, whereas

$$p(x^{(i)}) = \sum_{x_i \in \{0,1\}} p(x^{(i)}, x_i)$$

Therefore,

$$\begin{aligned} H(X) - H(X^{(i)}) &= -\sum_x p(x) \log p(x) - \sum_{x^{(i)}} p(x^{(i)}) \log p(x^{(i)}) \\ &= -\sum_{x \in A} p(x) \left(\log p(x) - \log p(x^{(i)}) \right) \end{aligned}$$

We will see next time that in this sum, we get a contribution of $1/|A|$ for each edge in A in direction i

If $p(x^{(i)}) = 0$, then neither along the edge defined by $x^{(i)}$ is in A, so we may ignore it. If $p(x) = 1/|A|$ for $x_i = 0$ but $p(x') = 0$ for $x_i = 1$, meaning that only one point along this edge is in A, then $p(x^{(i)}) = 1/|A|$. □

7 October 15, 2018

Recall that we were trying to prove the edge-isoperimetric inequality, which asks us to bound the number of edges $|E(A)|$ induced by a given set $A \subseteq \{0,1\}^n$. Let X be a uniformly random variable on A, and as usual let $X^{(i)} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$. Let

$$p^{(i)}(x^{(i)}) = \sum_{x_i \in \{0,1\}} p(x^{(i)}, x)$$

be the induced probability distribution of $X^{(i)}$. We compute

$$\begin{aligned} H(X) - H(X^{(i)}) &= -\sum_{x \in A} p(x) \log p(x) + \sum_{x^{(i)}} p^{(i)}(x^{(i)}) \log p^{(i)}(x^{(i)}) \\ &= -\sum_x p(x) \left(\log p(x) - \log \sum_{x'_i \in \{0,1\}} p(x^{(i)}, x') \right) \\ &= -\sum_x p(x) \log \left(\frac{p(x)}{\sum_{x'_i \in \{0,1\}} p(x^{(i)}, x'_i)} \right) \end{aligned}$$

For a fixed point x, the quantity in parentheses is the conditional probability of inclusion in A, conditioned on the value of $x^{(i)}$. Thus,

$$H(X) - H(X^{(i)}) = -\sum_x p(x) \log p(x \mid x^{(i)})$$

The value of this conditional probability depends on how many of the two points agreeing on $x^{(i)}$ are contained in A. If both points $(x^{(i)}, 0)$ and $(x^{(i)}, 1)$ are in A, then this conditional distribution is uniform on the two points, so $p(x \mid x^{(i)}) = \frac{1}{2}$. If only one of the points, say x, is in A, then $p(x \mid x^{(i)}) = 1$. The final scenario doesn't actually matter, since in that case there is no contribution to the sum, since the $p(x)$ term is zero. Combining these two cases, we see that $\log p(x \mid x^{(i)})$ is minus the number of edges in A in the ith direction at location $x^{(i)}$. Note that each edge is counted twice, so

$$\sum_x p(x) \log p(x \mid x^{(i)}) = \frac{2}{|A|} \#(\text{edges in } A \text{ in direction } i)$$

Therefore,

$$\sum_{i=1}^{n} (H(X) - H(X^{(i)})) = \frac{2}{|A|} |E(A)|$$

By Han's inequality, we thus get that

$$\frac{2}{|A|} |E(A)| \leq H(X) = \log |A|$$

which is what we wanted to prove.

7.1 Other inequalities

There is also a version of Han's inequality for KL divergence.

Theorem. *Let P, Q be probability measures on $\mathcal{X}^n$, where P is a product measure $P_1 \otimes P_2 \otimes \cdots \otimes P_n$, and Q is arbitrary. Then*

$$D(Q\|P) \geq \frac{1}{n-1} \sum_{i=1}^{n} D(Q^{(i)}\|P^{(i)})$$

where $P^{(i)}, Q^{(i)}$ are the marginal distributions on $\mathcal{X}^{n-1}$ gotten by projecting along the ith coordinate.

Proof. We can write

$$D(Q\|P) = \sum_x q(x) \log \frac{q(x)}{p(x)} = \sum_x q(x) \log q(x) - \sum_x q(x) \log p(x)$$

We can also write $p(x) = \prod_{j=1}^{n} p_j(x_j)$, and thus

$$p^{(i)}(x^{(i)}) = \prod_{j \neq i} p_j(x_j)$$

As above, we can also write

$$q^{(i)}(x^{(i)}) = \sum_{x_i \in \mathcal{X}} q(x^{(i)}, x_i)$$

Therefore,

$$D(Q^{(i)}\|P^{(i)}) = \sum_{x^{(i)}} q^{(i)}(x^{(i)}) \log q^{(i)}(x^{(i)}) - \sum_{x^{(i)}} q^{(i)}(x^{(i)}) \log p^{(i)}(x^{(i)})$$

Han's inequality tells us that $H(Q) \leq \frac{1}{n-1} \sum_i H(Q^{(i)})$, and thus

$$\sum_x q(x) \log q(x) \geq \frac{1}{n-1} \sum_{i=1}^n \sum_{x^{(i)}} q^{(i)}(x^{(i)}) \log q^{(i)}(x^{(i)})$$

For the other term, we can write

$$\sum_x q(x) \log p(x) = \sum_x q(x) \sum_{i=1}^n \log p_i(x_i) = \sum_x q(x) \log(p^{(i)}(x^{(i)}) + \log p_i(x_i))$$

Thus,

$$\begin{aligned} n \sum_x q(x) \log p(x) &= \sum_{i=1}^n \sum_x q(x) \log(p^{(i)}(x^{(i)}) + \log p_i(x_i)) \\ &= \sum_{i=1}^n \sum_{x^{(i)}} q^{(i)}(x^{(i)}) \log p^{(i)}(x^{(i)}) + \sum_x q(x) \log p(x) \end{aligned}$$

and therefore

$$(n-1) \sum_x q(x) \log p(x) = \sum_{i=1}^n \sum_{x^{(i)}} q^{(i)}(x^{(i)}) \log p^{(i)}(x^{(i)})$$

Combining everything,

$$\begin{aligned} \sum_{i=1}^n D(Q^{(i)}\|P^{(i)}) &\leq (n-1) \sum_x q(x) \log q(x) - (n-1) \sum_x q(x) \log p(x) \\ &= (n-1) D(Q\|P) \end{aligned}$$

□

7.2 Another entropy function

The Shannon entropy is a function only of the distribution, and doesn't depend on the actual values of the random variable; in other words, the state space itself is meaningless. For some contexts, we actually want to change this.

Definition. For a non-negative random variable Z with $\mathbb{E}[Z] < \infty$, we define

$$\mathrm{Ent}(Z) = \mathbb{E}[Z \log Z] - \mathbb{E}[Z] \log \mathbb{E}[Z]$$

More generally, for a convex function Φ, we can consider $\mathbb{E}[\Phi(Z)] - \Phi(\mathbb{E}[Z])$, which is non-negative by Jensen's inequality. For $\Phi(x) = x \log x$, this gives the entropy above, whereas for $\Phi(x) = x^2$ we get the variance.

We can equivalently write

$$\mathrm{Ent}(Z) = \mathbb{E}\left[Z \log \frac{Z}{\mathbb{E}[Z]}\right]$$

Thus, if we scale Z by $\alpha > 0$, then we have

$$\mathrm{Ent}(\alpha Z) = \alpha \,\mathrm{Ent}(Z)$$

since the quotient in the log won't change. We can also relate the entropy to the KL divergence, as follows. Assume that $Z = f(X) \geq 0$ with $\mathbb{E}[Z] = 1$. Then we have that

$$\mathrm{Ent}(Z) = D(P_f \| P)$$

where P is the distribution of X, and

$$P_f(x) = f(x)P(X = x)$$

Note that P_f is a distribution, since

$$\sum_x P_f(x) = \sum_x f(x)P(x) = \mathbb{E}[Z] = 1$$

To prove this, we calculate

$$\begin{aligned} D(P_f \| P) &= \sum_x P_f(x) \log \frac{P_f(x)}{P(x)} \\ &= \sum_x f(x)P(x) \log f(x) \\ &= \mathop{\mathbb{E}}_{X \sim P}[f(X) \log f(X)] \\ &= \mathbb{E}[Z \log Z] \\ &= \mathrm{Ent}(Z) \end{aligned}$$

where $\mathrm{Ent}(Z) = \mathbb{E}[Z \log Z]$ since we normalized to $\mathbb{E}[Z] = 1$.

The most important inequality that we will use will be an analogue of Efron–Stein. Recall that Efron–Stein says that

$$\mathrm{Var}(Z) \leq \sum_{i=1}^n \mathbb{E}\left[\left(Z - \mathop{\mathbb{E}}_{X_i}[Z]\right)^2\right] =: \sum_{i=1}^n \mathbb{E}[\mathrm{Var}^{(i)}(Z)]$$

When viewed in this form, it makes sense to replace the variance by $\mathbb{E}[\Phi(Z)] - \Phi(\mathbb{E}[Z])$ for some other convex function Φ. We will only prove and use the inequality for $\Phi(x) = x \log x$. We define

$$\mathrm{Ent}^{(i)}(Z) = \mathop{\mathbb{E}}_{X_i}[Z \log Z] - \mathop{\mathbb{E}}_{X_i}[Z] \log \mathop{\mathbb{E}}_{X_i}[Z]$$

Next time, we will prove the subadditivity of entropy:

Theorem.

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

8 October 17, 2018

Recall that we were trying to prove an Efron–Stein inequality for entropy.

Theorem. *For $Z = f(X_1, \ldots, X_n)$ with X_i independent and $f \geq 0$,*

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

Proof. We can assume without loss of generality that $\mathbb{E}[Z] = 1$, since both sides scale linearly. Also, even though the statement is true for continuous random variables, we will only prove it for X_i discrete. As we saw last time, we can write

$$\mathrm{Ent}(Z) = D(P_f \| P)$$

where P is the law of $(X_1, \ldots, X_n)$, i.e.

$$P(x_1, \ldots, x_n) = \Pr(X_1 = x_1, \ldots, X_n = x_n)$$

and

$$P_f(x_1, \ldots, x_n) = f(x_1, \ldots, x_n) P(x_1, \ldots, x_n)$$

Since the X_i are independent, P is a product measure, so we can apply Han's inequality, which tells us that

$$D(P_f \| P) \leq \sum_{i=1}^{n} \Big(D(P_f \| P) - D(P_f^{(i)} \| P^{(i)}) \Big)$$

By definition,

$$P_f^{(i)}(x^{(i)}) = \sum_{y \in \mathcal{X}} f(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_n) P(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_n)$$

and therefore

$$\begin{aligned}
D(P_f^{(i)} \| P^{(i)}) &= \sum_{x^{(i)}} P_f^{(i)}(x^{(i)}) \log \frac{P_f^{(i)}(x^{(i)})}{P^{(i)}(x^{(i)})} \\
&= \sum_{x} f(x) P(x) \log \frac{\sum_y f(x_1, \ldots, y, \ldots, x_n) P(x_1, \ldots, y, \ldots, x_n)}{\sum_y P(x_1, \ldots, y, \ldots, x_n)} \\
&= \sum_{x} f(x) P(x) \log \left(\mathop{\mathbb{E}}_{X_i} [f(X) \mid X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n] \right) \\
&= \mathop{\mathbb{E}}_{X} \left[f(X) \log \mathop{\mathbb{E}}_{X_i} [f(X) \mid X^{(i)}] \right] \\
&= \mathop{\mathbb{E}}_{X} \left[Z \log \mathop{\mathbb{E}}_{X_i} [Z \mid X^{(i)}] \right]
\end{aligned}$$

So returning to Han's inequality, and using the fact that by our normalization $\mathrm{Ent}(Z) = \mathbb{E}[Z \log Z]$, we have that

$$\begin{aligned}\mathrm{Ent}(Z) = D(P_f \| P) &\leq \sum_{i=1}^{n} \Big(D(P_f \| P) - D(P_f^{(i)} \| P^{(i)}) \Big) \\ &= \sum_{i=1}^{n} \mathop{\mathbb{E}}_{X^{(i)}} \left[\mathop{\mathbb{E}}_{X_i}[Z \log Z] - \mathop{\mathbb{E}}_{X_i} \left[Z \log \mathop{\mathbb{E}}_{X_i}[Z] \right] \right] \\ &= \sum_{i=1}^{n} \mathop{\mathbb{E}}_{X^{(i)}} \left[\mathrm{Ent}^{(i)}(Z) \right]\end{aligned}$$

□

8.1 Continuous distributions

Many, but not quite all, of the things we are doing can be generalized for continuous distributions. Most importantly, there is no good generalization of the Shannon entropy; however, we can get by because the KL divergence can be defined for continuous distributions.

Definition. If P, Q are two measures on a probability space Ω, then we say that Q is absolutely continuous with respect to P (and write $Q \ll P$) if for each A such that $P(A) = 0$, then $Q(A) = 0$ too.

Fact. If Q is absolutely continuous with respect to P, then there is a random variable $Y = \mathrm{d}Q/\mathrm{d}P$, called the Radon–Nikodym derivative, so that for every function f,

$$\int_\Omega f(\omega)\, \mathrm{d}Q = \int_\Omega f(\omega) Y(\omega)\, \mathrm{d}P$$

Then if Q is absolutely continuous with respect to P, then we can define $D(Q\|P) = \mathrm{Ent}(Y)$, where $Y = \mathrm{d}Q/\mathrm{d}P$. This matches the discrete definition, since in that case $Y(x) = Q(x)/P(x)$, and $\mathbb{E}_P[Y] = 1$.

There are many duality relations one can prove for the entropy.

1. If $Y \geq 0$ with $\mathrm{Ent}(Y) < \infty$, then

$$\mathrm{Ent}(Y) = \sup\{\mathbb{E}[UY] : \mathbb{E}[e^U] = 1\}$$

Moreover, the supremum is achieved by $U = \log \frac{Y}{\mathbb{E}[Y]}$.

Proof. Assume $\mathbb{E}[e^U] = 1$, and consider $Q = e^U P$, where P is the default measure on Ω (the one according to which all the above expectations are

taken). Then consider

$$\begin{aligned}\mathrm{Ent}_Q(Ye^{-U}) &= \mathop{\mathbb{E}}_Q[Ye^{-U}\log(Ye^{-U})] - \mathop{\mathbb{E}}_Q[Ye^{-U}]\log\mathop{\mathbb{E}}_Q[Ye^{-U}] \\ &= \mathop{\mathbb{E}}_P[Y\log(Ye^{-U})] - \mathop{\mathbb{E}}_P[Y]\log\mathop{\mathbb{E}}_P[Y] \\ &= \mathbb{E}[Y\log Y] - \mathbb{E}[YU] - \mathbb{E}[Y]\log\mathbb{E}[Y] \\ &= \mathrm{Ent}(Y) - \mathbb{E}[YU]\end{aligned}$$

Since the left hand side is the entropy of some non-negative random variable with respect to some measure, it must be non-negative. Thus, we get that $\mathbb{E}[YU] \leq \mathrm{Ent}(Y)$, which is what we wanted to prove. □

2. Conversely, if U is a random variable such that for every $Y \geq 0$ with $\mathbb{E}[Y] = 1$ and $\mathrm{Ent}(Y) < \infty$,

$$\mathbb{E}[UY] \leq \mathrm{Ent}(Y)$$

then $\mathbb{E}[e^U] \leq 1$.

Proof. Given such a U, define

$$x_n = \mathbb{E}[e^{\min\{U,n\}}]$$

and

$$Y_n = \frac{1}{x_n}e^{\min\{U,n\}}$$

Then $\mathbb{E}[Y_n] = 1$, and by this truncation, we also have that $\mathrm{Ent}(Y_n) < \infty$. Then by our assumption,

$$\begin{aligned}\mathbb{E}\left[U\frac{1}{x_n}e^{\min\{U,n\}}\right] = \mathbb{E}[UY_n] &\leq \mathrm{Ent}(Y_n) \\ &= \mathbb{E}[Y_n\log Y_n] \\ &= \mathbb{E}\left[\frac{1}{x_n}e^{\min\{U,n\}}(\min\{U,n\} - \log x_n)\right] \\ &\leq \mathbb{E}\left[\frac{1}{x_n}e^{\min\{U,n\}}(U - \log x_n)\right]\end{aligned}$$

Comparing both sides, we get that $\log x_n \leq 0$, so that $x_n \leq 1$. Finally, by the monotone convergence theorem,

$$\mathbb{E}[e^U] = \lim_{n\to\infty}\mathbb{E}[e^{\min\{U,n\}}] = \lim_{n\to\infty} x_n \leq 1$$

□

3. Let $Z \geq 0$ with $\mathbb{E}[Z] < \infty$. Then for every $\lambda \in \mathbb{R}$,

$$\psi_{Z-\mathbb{E}Z}(\lambda) = \log\mathbb{E}[e^{\lambda(Z-\mathbb{E}Z)}] = \sup_{Q\ll P}\left[\lambda\left(\mathop{\mathbb{E}}_Q[Z] - \mathop{\mathbb{E}}_P[Z]\right) - D(Q\|P)\right]$$

where P is the default measure, i.e. the measure according to which the first expectation is taken. We will prove this next time.

9 October 19, 2018

We were discussing duality (or minimax) relations for various information-theoretic quantities. Recall that we stated the following three properties.

1. For all $Y \geq 0$,
$$\mathrm{Ent}(Y) = \sup\{\mathbb{E}[UY] : \mathbb{E}[e^U] = 1\}$$
2. If U is such that $\mathbb{E}[UY] \leq \mathrm{Ent}(Y)$ for all $Y \geq 0$ with $\mathrm{Ent}(Y) < \infty$, then $\mathbb{E}[e^U] \leq 1$.
3. If $Z \geq 0$ with $\mathbb{E}[Z] < \infty$, then for all $\lambda \in \mathbb{R}$,
$$\log \mathbb{E}[e^{\lambda(Z - \mathbb{E}\, Z)}] = \sup_{Q \ll P} \left[\lambda \left(\mathop{\mathbb{E}}_{Q} Z - \mathop{\mathbb{E}}_{P} Z\right) - D(Q\|P)\right]$$

We proved the first two last time.

Proof of (3). We will first prove that the left-hand side is at least the right-hand side. Consider any $Q \ll P$; we want to prove that

$$\log \mathbb{E}[e^{\lambda(Z - \mathbb{E}\, Z)}] \geq \lambda \left(\mathop{\mathbb{E}}_{Q} Z - \mathop{\mathbb{E}}_{P} Z\right) - D(Q\|P)$$

Let $Y = \mathrm{d}Q/\mathrm{d}P$, so that $\mathbb{E}_Q[\,\cdot\,] = \mathbb{E}_P[Y \cdot\,]$. Define

$$U = \lambda \left(Z - \mathop{\mathbb{E}}_{P} Z\right) - \log \mathop{\mathbb{E}}_{P}[e^{\lambda(Z - \mathbb{E}_P\, Z)}]$$

Then

$$\mathop{\mathbb{E}}_{P}[UY] = \mathop{\mathbb{E}}_{Q}[U] = \lambda \left(\mathop{\mathbb{E}}_{Q} Z - \mathop{\mathbb{E}}_{P} Z\right) - \log \mathop{\mathbb{E}}_{P}[e^{\lambda(Z - \mathbb{E}_P\, Z)}]$$

Moreover, we also know that $D(Q\|P) = \mathrm{Ent}(Y)$. By relation (1) above, we know that

$$\mathrm{Ent}(Y) \geq \mathop{\mathbb{E}}_{P}[UY]$$

since

$$\mathbb{E}[e^U] = \mathbb{E}[e^{\lambda(Z - \mathbb{E}\, Z) - \lambda(Z - \mathbb{E}\, Z)}] = 1$$

Putting this together, we find that

$$D(Q\|P) \geq \lambda \left(\mathop{\mathbb{E}}_{Q} Z - \mathop{\mathbb{E}}_{P} Z\right) - \log \mathop{\mathbb{E}}_{P}[e^{\lambda(Z - \mathbb{E}_P\, Z)}]$$

which rearranges to the inequality we wanted.

For the reverse inequality, define

$$U = \lambda \left(Z - \mathop{\mathbb{E}}_{P} Z\right) - \sup_{Q \ll P} \left[\lambda \left(\mathop{\mathbb{E}}_{Q} Z - \mathop{\mathbb{E}}_{P} Z\right) - D(Q\|P)\right]$$

Then we wish to prove that $\log \mathbb{E}[e^U] \leq 0$, or equivalently $\mathbb{E}[e^U] \leq 1$. By relation (2), it suffices to show that $\mathbb{E}[UY] \leq \mathrm{Ent}(Y)$ for all $Y \geq 0$ with $\mathbb{E}[Y] = 1, \mathrm{Ent}(Y) < \infty$. Fix such a Y. Define a measure $Q' = Y \cdot P$, which is a probability measure since $\mathbb{E}\, Y = 1$. Then

$$\begin{aligned}
\mathbb{E}[UY] &= \mathop{\mathbb{E}}_{Q'}[U] \\
&= \lambda\left(\mathop{\mathbb{E}}_{Q'} Z - \mathop{\mathbb{E}}_{P} Z\right) - \sup_{Q \ll P}\left[\lambda\left(\mathop{\mathbb{E}}_{Q} Z - \mathop{\mathbb{E}}_{P} Z\right) - D(Q\|P)\right] \\
&\leq \lambda\left(\mathop{\mathbb{E}}_{Q'} Z - \mathop{\mathbb{E}}_{P} Z\right) - \left[\lambda\left(\mathop{\mathbb{E}}_{Q'} Z - \mathop{\mathbb{E}}_{P} Z\right) - D(Q'\|P)\right] \\
&= D(Q'\|P) \\
&= \mathrm{Ent}(Y)
\end{aligned}$$

This is the result we need to apply relation (2), so we get the desired inequality. □

We will now see one further relation.

4. $$D(Q\|P) = \sup_{Z:\mathbb{E}[e^Z]<\infty}\left[\mathop{\mathbb{E}}_{Q} Z - \log \mathop{\mathbb{E}}_{P}[e^Z]\right]$$

   Proof. If $Q \not\ll P$, then $D(Q\|P)$ is defined to be ∞. To see that the right-hand side is also ∞, fix some event $A \subseteq \Omega$ so that $P(A) = 0$ but $Q(A) > 0$, and define $Z_n = n \cdot 1_A$. Then

   $$\mathop{\mathbb{E}}_{Q} Z_n - \log \mathop{\mathbb{E}}_{P}[e^{Z_n}] = n \cdot Q(A) \underset{n\to\infty}{\longrightarrow} \infty$$

   So we may assume $Q \ll P$, and set $Y = \mathrm{d}Q/\mathrm{d}P$. Then $\mathbb{E}_P\, Y = 1$, and $D(Q\|P) = \mathrm{Ent}(Y)$. By relation (1),

   $$\mathrm{Ent}(Y) = \sup_{U:\mathbb{E}_P[e^U]=1} \mathop{\mathbb{E}}_{P}[UY] = \sup_{Z:\mathbb{E}_P[e^Z]<\infty}\left(\mathop{\mathbb{E}}_{P}\left[ZY - Y \log \mathop{\mathbb{E}}_{P}[e^Z]\right]\right)$$

   where we can get from the third to second expression by rescaling as $U = Z - \log \mathbb{E}_P[e^Z]$. But the final expression is just $\mathbb{E}_Q\, z - \log \mathbb{E}_P[e^Z]$, as desired. □

Using these relations, we can also give a new, clean proof for the Efron–Stein-type inequality for entropy, which works for arbitrary random variables (not just discrete ones). Recall that the inequality said that

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

Proof. By definition,

$$\mathrm{Ent}(Z) = \mathbb{E}[Z \log Z] - \mathbb{E}[Z] \log \mathbb{E}[Z]$$

We will consider the expression $Z \log Z - Z \log \mathbb{E}[Z]$; eventually we will take an expectation of this to get $\mathrm{Ent}(Z)$. We can write this quantity as a telescoping sum:

$$Z \log Z - Z \log \mathbb{E}[Z] = \sum_{i=1}^{n} \left(Z \log \mathop{\mathbb{E}}_{X_{i+1},\dots,X_n}[Z] - Z \log \mathop{\mathbb{E}}_{X_i,\dots,X_n}[Z] \right)$$

For each term in this sum, set $T = \mathbb{E}_{X_{i+1},\dots,X_n}[Z]$, and $U = \log T - \mathbb{E}_{X_i}[T]$. Then we can write

$$Z \left(\log \mathop{\mathbb{E}}_{X_{i+1},\dots,X_n}[Z] - \log \mathop{\mathbb{E}}_{X_i,\dots,X_n}[Z] \right) = Z \left(\log T - \log \mathop{\mathbb{E}}_{X_i} T \right) = ZU$$

and when we condition on $X^{(i)}$, we have that $\mathbb{E}_{X_i}[e^U] = 1$. Applying relation (1) to $Z(X_i \mid X^{(i)})$, we get that

$$\mathrm{Ent}^{(i)}(Z) \geq \mathop{\mathbb{E}}_{X_i}[UZ] = Z \left(\log \mathop{\mathbb{E}}_{X_{i+1},\dots,X_n}[Z] - \log \mathop{\mathbb{E}}_{X_i,\dots,X_n}[Z] \right)$$

which is precisely the summand we're looking for. Taking the expectation over the remaining variables and summing over all i gives the desired result. □

9.1 Bregman Divergence

For f convex and differentiable, we define

$$D_f(x, y) = f(y) - f(x) - f'(x)(y - x)$$

It measures how big an error we see at y when we replace f by its linear approximation. It is always non-negative, by convexity.

Theorem. *If X is a random variable with values in an open interval $I \subseteq \mathbb{R}$, then*

$$\mathbb{E}[f(X)] - f(\mathbb{E}[X]) = \inf_{a \in I} \mathbb{E}[D_f(a, X)]$$

In particular, plugging in $f(x) = x \log x$,

$$\mathrm{Ent}(Y) = \inf_{a > 0} \mathbb{E}[Y \log Y - Y \log a + a - Y]$$

Proof. We can write

$$\begin{aligned} \mathbb{E}[D_f(\mathbb{E}\,X, X)] &= \mathbb{E}[f(X) - f(\mathbb{E}\,X) - f'(\mathbb{E}\,X)(X - \mathbb{E}\,X)] \\ &= \mathbb{E}[f(X)] - f(\mathbb{E}\,X) \end{aligned}$$

For any $a \in I$, we have

$$\begin{aligned}\mathbb{E}[D_f(a, X)] - \mathbb{E}[D_f(\mathbb{E}\,X, X)] &= \mathbb{E}[-f(a) - f'(a)(X - a) + f(\mathbb{E}\,X)] \\ &= f(\mathbb{E}\,X) - f(a) - f'(a)(\mathbb{E}\,X - a) \\ &= D_f(a, \mathbb{E}\,X) \geq 0\end{aligned}$$

This proves one direction of the inequality. However, plugging in $a = \mathbb{E}\,X$ gives equality, which proves that the infimum is attained, and thus proves the result. □

10 October 22, 2018

10.1 Logarithmic Sobolev Inequalities

Let $X_1, \ldots, X_n \in \{-1, 1\}$ be uniformly random, and set $Z = f(X_1, \ldots, X_n)$. Efron–Stein tells us that

$$\mathrm{Var}(Z) \leq \sum_{i=1}^{n} \mathbb{E}[\mathrm{Var}^{(i)}(Z) \mid X^{(i)}]$$

Equivalently, we can write

$$\mathrm{Var}(Z) \leq \sum_{i=1}^{n} \mathbb{E}\left[\frac{1}{2}\left(f(X) - f(X_i')\right)^2\right]$$

where X_i' is X with X_i replaced by an independent sample. Since X_i' equals X with probability $1/2$ and has its ith coordinate flipped with probability $1/2$, we can also write

$$\mathrm{Var}(Z) \leq \sum_{i=1}^{n} \mathbb{E}\left[\frac{1}{4}\left(f(X) - f(\overline{X}^{(i)})\right)^2\right]$$

where $\overline{X}^{(i)}$ is X with its ith coordinate flipped. If we define the discrete derivative of f by

$$\partial_i f(x) = \frac{1}{2}\left(f(x) - f(\overline{x}^{(i)})\right)$$

then this just be comes

$$\mathrm{Var}(Z) \leq \sum_{i=1}^{n} \mathbb{E}[\partial_i f(x)^2] = \mathbb{E}\left[\|\nabla f(X)\|^2\right]$$

where $\nabla f = (\partial_1 f, \ldots, \partial_n f)$. This is known as a Poincaré-type inequality. There is an analogous inequality for Gaussian random variables (which can in fact be proven from the above):

Theorem (Gaussian Poincaré inequality)**.** *If $X_1, \ldots, X_n$ are independent standard Gaussians, and $Z = f(X_1, \ldots, X_n)$ is a differentiable function of them, then*

$$\mathrm{Var}(Z) \leq \mathbb{E}\left[\|\nabla f(X)\|^2\right]$$

A similar inequality for the entropy is called a Log-Sobolev inequality.

Theorem (Log-Sobolev Gaussian Inequality)**.** *If $X_1, \ldots, X_n$ are independent standard Gaussians, and $Z = f(X_1, \ldots, X_n)$ is a differentiable function of them, then*

$$\frac{1}{2}\operatorname{Ent}(f^2(X)) \leq \mathbb{E}\left[\|\nabla f\|^2\right]$$

We will first prove the discrete version:

Theorem. *If $X_1, \ldots, X_n \in \{-1, 1\}$ are uniformly random, then*

$$\frac{1}{2}\operatorname{Ent}(f^2(X)) \leq \mathbb{E}\left[\|\nabla f(X)\|^2\right]$$

Suppose we set $g(x) = 1 + \varepsilon f(x)$. Then one can check that

$$\operatorname{Ent}(g^2) = 2\varepsilon^2 \mathbb{E}\left[f(X)^2\right] - 2\varepsilon^2 \left(\mathbb{E}[f(X)]\right)^2 - O(\varepsilon^3)$$

and therefore

$$\frac{1}{2\varepsilon^2}\operatorname{Ent}(g^2) = \operatorname{Var}(Z) - O(\varepsilon)$$

Thus, by sending $\varepsilon \to 0$, we can recover Efron–Stein.

Proof. We already know that

$$\begin{aligned}\operatorname{Ent}(f^2) &\leq \sum_{i=1}^n \mathbb{E}\left[\operatorname{Ent}^{(i)}\left(f^2\right)\right] \\ &= \sum_{i=1}^n \mathop{\mathbb{E}}_{X^{(i)}}\left[\mathop{\mathbb{E}}_{X_i}\left[f^2(X)\log f^2(X)\right] - \mathbb{E}\left[f^2(X)\right]\log \mathbb{E}\left[f^2(X)\right]\right]\end{aligned}$$

Let

$$\mathcal{E}(f) = \mathbb{E}\left[\|\nabla f(X)\|^2\right] = \frac{1}{4}\sum_{i=1}^n \mathbb{E}\left[\left(f(X) - f(\overline{X}^{(i)})\right)^2\right] = \sum_{i=1}^n \mathcal{E}_i(f)$$

So it suffices to prove that $\operatorname{Ent}^{(i)}(f^2) \leq 2\mathcal{E}_i(f)$.

Given $X^{(i)}$, the variable $Z = f(X)$ has only two values, say a, b, each with probability $1/2$. We have that

$$\operatorname{Ent}^{(i)}\left(f^2\right) = \frac{1}{2}a^2\log a^2 + \frac{1}{2}b^2\log b^2 - \frac{1}{2}\left(a^2+b^2\right)\log\frac{a^2+b^2}{2}$$

and

$$2\mathcal{E}_i(f) = \frac{1}{2}(a-b)^2$$

So we just need to prove the inequality between these two expressions for all $a, b \in \mathbb{R}$. We may assume without loss of generality that $a \geq b \geq 0$. So for fixed b, we may define a function

$$h(a) = \frac{1}{2}a^2\log a^2 + \frac{1}{2}b^2\log b^2 - \frac{1}{2}\left(a^2+b^2\right)\log\frac{a^2+b^2}{2} - \frac{1}{2}(a-b)^2$$

Observe that $h(b) = 0$. To prove that $h(a) \leq 0$ everywhere, we can compute its first and second derivatives. We have that

$$h'(a) = a \log \frac{2a^2}{a^2 + b^2} - (a - b)$$

Therefore, $h'(b) = 0$. Additionally,

$$h''(a) = 1 + \log \frac{2a^2}{a^2 + b^2} - \frac{2a^2}{a^2 + b^2}$$

Thus, since $\log x \leq x - 1$, we have that $h''(a) \leq 0$ everywhere. Thus, h is a concave function that starts with $h(b) = h'(b) = 0$, so it must always be non-positive for all $a \geq b$. □

Using this, we can prove the Gaussian Poincaré inequality.

Proof. We have independent Gaussians $X_1, \ldots, X_n$, and want to prove $\mathrm{Var}(Z) \leq \mathcal{E}(f)$. By Efron–Stein, it suffices to prove that

$$\mathbb{E}\left[\mathrm{Var}^{(i)}(Z)\right] \leq \mathbb{E}\left[\left|\frac{\partial f}{\partial x_i}(X)\right|^2\right]$$

Thus, it suffices to prove that for every fixed choice of $X^{(i)}$,

$$\mathrm{Var}^{(i)}(Z) \leq \mathop{\mathbb{E}}_{X_i}\left[\left|\frac{\partial f}{\partial x_i}(X)\right|^2\right]$$

In other words, it suffices to prove the Gaussian Poincaré inequality in one dimension. To do this, we will apply Efron–Stein again. We first approximate the Gaussian X_i by

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathcal{E}_i$$

where $\mathcal{E}_i \in \{-1, 1\}$ are uniformly random. By the central limit theorem, as $n \to \infty$, $S_n \xrightarrow{\mathcal{D}} N(0, 1)$. Let $f^{(i)}(x) = f(x^{(i)}, x)$. Then $\mathrm{Var}(f^{(i)}(S_n)) \to \mathrm{Var}^{(i)}(Z)$. By Efron–Stein,

$$\begin{aligned}\mathrm{Var}\left(f^{(i)}(S_n)\right) &\leq \sum_{j=1}^{n} \mathbb{E}\left[\mathrm{Var}_{\mathcal{E}_j}\left(f^{(i)}(S_n)\right)\right] \\ &= \sum_{j=1}^{n} \mathbb{E}\left[\left(f^{(i)}\left(S_n - \frac{\mathcal{E}_j}{\sqrt{n}} + \frac{1}{\sqrt{n}}\right) - f^{(i)}\left(S_n - \frac{\mathcal{E}_j}{\sqrt{n}} - \frac{1}{\sqrt{n}}\right)\right)^2\right]\end{aligned}$$

Assume f is twice differentiable and compactly supported (these assumptions can later be removed by a limiting argument). By Taylor's theorem, we can bound the quantity in brackets by

$$\left|f^{(i)}\left(S_n + \frac{1 - \mathcal{E}_j}{\sqrt{n}}\right) - f^{(i)}\left(S_n - \frac{1 + \mathcal{E}_j}{\sqrt{n}}\right)\right| \leq \frac{2}{\sqrt{n}}|f'(S_n)| + \frac{2K}{n}$$

where $K = \sup_x |f''(x)|$. Taking the limit as $n \to \infty$ gives the desired result. □

11 October 24, 2018

Recall that we were discussing the Boolean Log-Sobolev inequality.

Theorem. *If $Z = f(X)$, where $X \in \{-1,1\}^n$ is uniform, then*

$$\mathrm{Ent}(Z^2) \leq 2\mathcal{E}(f) = 2\,\mathbb{E}\left[\|\nabla f(X)\|^2\right]$$

The Gaussian analogue was the following:

Theorem. *If $X_1, \ldots, X_n$ are independent Gaussians and $Z = f(X)$ with f differentiable,*

$$\mathrm{Ent}(Z^2) \leq 2\,\mathbb{E}\left[\|\nabla f(X)\|^2\right]$$

Similarly, we also had a Gaussian version of the Poincaré inequality.

Theorem. *Under the same assumptions as the previous theorem,*

$$\mathrm{Var}(Z) \leq \mathbb{E}\left[\|\nabla f(X)\|^2\right]$$

Last time, we saw how to prove this by applying Efron–Stein twice, and approximating the Gaussian distribution by a binomial distribution. In much the same way, one can prove the Gaussian Log-Sobolev inequality by this approximation technique: we first use the subadditivity of entropy (which is the entropic analogue of Efron–Stein) to reduce to one dimension, and then use Boolean Log-Sobolev on the variables $S_n = n^{-1/2}\sum_j \varepsilon_j$.

Today, we will see how to use Log-Sobolev to obtain exponential tail bounds. The key idea is something called Herbst's Argument, which goes as follows. Suppose $Z = f(X)$, where $X \in \{-1,1\}^n$ is uniform. Instead of considering the exponential moment $F(\lambda) = \mathbb{E}[e^{\lambda Z}]$, we will consider the entropy of this random variable:

$$\begin{aligned}\mathrm{Ent}(e^{\lambda Z}) &= \mathbb{E}[e^{\lambda Z}\lambda Z] - \mathbb{E}[e^{\lambda Z}]\log\mathbb{E}[e^{\lambda Z}] \\ &= \lambda F'(\lambda) - F(\lambda)\log F(\lambda)\end{aligned}$$

where we use dominated convergence to switch the order of differentiation and expectation, to write $F'(\lambda) = \mathbb{E}[Ze^{\lambda Z}]$. Define $g(X) = e^{\frac{1}{2}\lambda f(X)} = e^{\frac{1}{2}\lambda Z}$. Then

by Log-Sobolev,

$$\begin{aligned}
\mathrm{Ent}(e^{\lambda Z}) = \mathrm{Ent}(g^2) &\le 2\mathcal{E}(g) \\
&= \frac{1}{2}\sum_{i=1}^{n} \mathbb{E}\left[\left(g(X) - g(\overline{X}^{(i)})\right)^2\right] \\
&= \sum_{i=1}^{n} \mathbb{E}\left[\left(e^{\frac{\lambda}{2}f(X)} - e^{\frac{\lambda}{2}f(\overline{X}^{(i)})}\right)_+^2\right] \\
&\le \sum_{i=1}^{n} \mathbb{E}\left[\left(\frac{\lambda}{2}\left(f(X) - f(\overline{X}^{(i)})\right)e^{\frac{\lambda}{2}f(X)}\right)_+^2\right] \\
&= \frac{\lambda^2}{4}\sum_{i=1}^{n} \mathbb{E}\left[\left(f(X) - f(\overline{X}^{(i)})\right)_+^2 e^{\lambda f(X)}\right]
\end{aligned}$$

where we use the convexity of the exponential function to bound the difference of exponentials by the first-order Taylor approximation at the larger point. From this point, our goal is to bound expressions of this form. As a simple example, suppose we assume that for every $x \in \{-1,1\}^n$,

$$\sum_{i=1}^{n}\left(f(x) - f(\overline{x}^{(i)})\right)_+^2 \le v$$

Note that this is analogous to the self-bounding condition, but is stronger, since we assume a uniform upper bound on the gradient, rather than an upper bound defined by the function itself. Under this assumption, we find that

$$\mathrm{Ent}(e^{\lambda Z}) \le \frac{\lambda^2}{4} v\,\mathbb{E}[e^{\lambda f(X)}] = \frac{1}{4}\lambda^2 v F(\lambda)$$

Thus, we derive a differential inequality for F, namely

$$\lambda F'(\lambda) - F(\lambda)\log F(\lambda) \le \frac{1}{4}\lambda^2 v F(\lambda)$$

This inequality can actually be solved. By rearranging terms, it becomes

$$\frac{v}{4} \ge \frac{F'(\lambda)}{\lambda F(\lambda)} - \frac{1}{\lambda^2}\log F(\lambda) = \frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\frac{\log F(\lambda)}{\lambda}\right)$$

In other words, $\log F(\lambda)/\lambda$ must grow sublinearly from zero. By l'Hôpital's rule,

$$\lim_{\lambda\to 0}\frac{\log F(\lambda)}{\lambda} = \lim_{\lambda\to 0}\frac{F'(\lambda)}{F(\lambda)} = \lim_{\lambda\to 0}\frac{\mathbb{E}[Ze^{\lambda Z}]}{\mathbb{E}[e^{\lambda Z}]} = \mathbb{E}[Z]$$

Therefore, we find that

$$\frac{\log F(\lambda)}{\lambda} \le \mathbb{E}[Z] + \frac{1}{4}v\lambda$$

Rearranging, we get what we called *sub-Gaussian* behavior, namely that the log of the exponential tail is bounded by a quadratic:

$$\log F(\lambda) \leq \lambda\,\mathbb{E}[Z] + \frac{1}{4}v\lambda^2$$

Applying Markov's inequality to the exponential moment, we derive the following exponential tail bounds:

$$\Pr(Z \geq \mathbb{E}[Z] + t) \leq e^{-t^2/v} \qquad \Pr(Z \leq \mathbb{E}[Z] - t) \leq e^{-t^2/v}$$

In exactly the same way, one can get a Gaussian concentration result:

Theorem (Tsirelson–Ibragimov–Sudakov)**.** *Suppose $X_1, \ldots, X_n$ are independent standard Gaussians, and $f : \mathbb{R}^n \to \mathbb{R}$ is L-Lipschitz (with respect to the Euclidean metric). Then*

$$\Pr(f(X) \geq \mathbb{E}[f(X)] + t) \leq e^{-t^2/2L^2} \qquad \Pr(f(X) \leq \mathbb{E}[f(X)] - t) \leq e^{-t^2/2L^2}$$

Proof. By a smoothing argument, we can assume that f is differentiable, and then the Lipschitz condition implies that $\|\nabla f(x)\| \leq L$ for all x. We also translate f so that $\mathbb{E}[f(X)] = 0$. Define, as before $g(x) = e^{\frac{\lambda}{2}f(x)}$, so that by Gaussian Log-Sobolev,

$$\mathrm{Ent}(g^2) \leq 2\,\mathbb{E}\left[\|\nabla g(X)\|^2\right]$$

We can compute that

$$\frac{\partial g}{\partial x_i} = e^{\frac{\lambda}{2}f(x)} \cdot \frac{\lambda}{2}\frac{\partial f}{\partial x_i}$$

and therefore

$$\|\nabla g(x)\|^2 = \sum_{i=1}^{n}\left(\frac{\partial g}{\partial x_i}\right)^2 = \frac{\lambda^2}{4}e^{\lambda f(x)}\|\nabla f(x)\|^2 \leq \frac{\lambda^2}{4}L^2 e^{\lambda F(x)}$$

Thus, we get the differential inequality

$$\lambda F'(\lambda) - F(\lambda)\log F(\lambda) \leq \frac{1}{2}\lambda^2 L^2 F(\lambda)$$

which is the same as before, up to the constant in front. Thus, the same calculation gives

$$\log F(\lambda) \leq \frac{1}{2}\lambda^2 L^2$$

which gives the desired tail bounds. □

As an application, we will consider the maximum of Gaussian variables. Define

$$Z = \max_{1 \leq i \leq n} X_i$$

where $(X_i)_i$ is a Gaussian vector, with covariance matrix Γ. Assume that for all i, $\mathbb{E}[X_i^2] \leq \sigma^2$. Then we claim that

$$\Pr(Z \geq \mathbb{E}[Z] + t) \leq e^{-t^2/2\sigma^2} \qquad \Pr(Z \leq \mathbb{E}[Z] - t) \leq e^{-t^2/2\sigma^2}$$

By a limiting argument, one can get a similar result for infinitely many Gaussian variables. To prove this, Gram-decompose Γ as $A^T A$, so that $X = AY$, where $Y_1, \ldots, Y_n$ are independent Gaussians. So

$$Z = \max_{1 \leq i \leq n} = \max_{1 \leq i \leq n} (AY)_i =: f(Y)$$

So we just need to check that f is Lipschitz. Indeed, we claim that f is σ-Lipschitz, since for every $u, v \in \mathbb{R}^n$, we have

$$\begin{aligned}
|f(u) - f(v)| &= \left| \max_i (Au)_i - \max_i (Av)_i \right| \\
&\leq \max_i |(Au)_i - (Av)_i| \\
&= \max_i \left| \sum_j A_{ij}(u-v)_j \right| \\
&\leq \max_i \left(\sum_j A_{ij}^2 \right)^{1/2} \|u - v\| \\
&\leq \sigma \|u - v\|
\end{aligned}$$

which gives the desired result.

12 October 26, 2018

12.1 The Entropy Method

Suppose, as usual, that X_i are independent variables valued in a domain $\mathcal{X}$, and $Z = f(X_1, \ldots, X_n)$. Define $Y = e^{\lambda Z}$, and we want to bound $\mathrm{Ent}(Y)$. As we saw, our main tool will be the subadditivity of entropy:

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

We saw that the Log-Sobolev inequality was useful to upper-bound these summands, in some specific settings (namely Boolean and Gaussian variables). Our ideal goal is to get an inequality of the form

$$\mathrm{Ent}(e^{\lambda Z}) \leq \frac{1}{2} \lambda^2 v \, \mathbb{E}[e^{\lambda Z}]$$

which we saw implies the sub-Gaussian bound

$$\log \mathbb{E}[e^{\lambda Z}] \leq \frac{1}{2} \lambda^2 v$$

Another way to see this implication is to define $\psi(\lambda) = \log \mathbb{E}[e^{\lambda Z}]$, at which point we can write

$$\mathrm{Ent}(e^{\lambda Z}) = e^{\psi(\lambda)} \cdot \lambda\psi'(\lambda) - e^{\psi(\lambda)} \cdot \psi(\lambda)$$

and we can thus derive the differential inequality

$$\lambda\psi'(\lambda) - \psi(\lambda) \leq \frac{1}{2}\lambda^2 v$$

and solve to get $\psi(\lambda) \leq \frac{1}{2}\lambda^2 v$. One way to guarantee this sort of inequality is the bounded differences assumption: suppose we assume that for all $x \in \mathcal{X}^n$, we have that

$$\sup_{x_i', x_i'' \in \mathcal{X}} |f(x_1, \ldots, x_i', \ldots, x_n) - f(x_1, \ldots, x_i'', \ldots, x_n)| \leq c_i$$

We will now apply the entropy method. As usual, we first use subadditivity to reduce to the one-dimensional case. We have

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

and we can write

$$\begin{aligned}\mathrm{Ent}^{(i)}(e^{\lambda Z}) &= \mathop{\mathbb{E}}_{X_i}[e^{\lambda Z}\lambda Z] - \mathop{\mathbb{E}}_{X_i}[e^{\lambda Z}] \log \mathop{\mathbb{E}}_{X_i}[e^{\lambda Z}] \\ &= e^{\psi_i(\lambda)} \cdot \lambda\psi_i'(\lambda) - e^{\psi_i(\lambda)} \cdot \psi_i(\lambda)\end{aligned}$$

where $\psi_i(\lambda) = \log \mathbb{E}_{X_i}[e^{\lambda Z}]$. By our assumption, $Z \mid X^{(i)}$ is in some interval $[a, b]$ with $b - a = c_i$.

Fact. For a random variable Y in an interval $[a, b]$ with $\mathbb{E}[Y] = 0$, and for $\psi_Y(\lambda) = \log \mathbb{E}[e^{\lambda Y}]$, we have

$$\psi_Y''(\lambda) \leq \frac{1}{4}(b-a)^2$$

Note that since $\psi_Y(0) = \psi_Y'(0) = 0$, this implies that $\psi_Y(\lambda) \leq \frac{1}{8}(b-a)^2\lambda^2$, which implies the Chernoff bound.

Proof. We can compute that

$$\begin{aligned}\psi_Y''(\lambda) &= \frac{\mathrm{d}^2}{\mathrm{d}\lambda^2} \log \mathbb{E}[e^{\lambda Y}] \\ &= \frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\frac{1}{\mathbb{E}[e^{\lambda Y}]}\,\mathbb{E}[Ye^{\lambda Y}]\right) \\ &= -\frac{1}{(\mathbb{E}[e^{\lambda Y}])^2}(\mathbb{E}[Ye^{\lambda Y}])^2 + \frac{1}{\mathbb{E}[e^{\lambda Y}]}\,\mathbb{E}[Y^2 e^{\lambda Y}]\end{aligned}$$

If we define a new measure Q by $\mathrm{d}Q = \frac{e^{\lambda Y}}{\mathbb{E}[e^{\lambda Y}]}\mathrm{d}P$, then we get

$$\psi_Y''(\lambda) = -\left(\mathop{\mathbb{E}}_{Q}[Y]\right)^2 + \mathop{\mathbb{E}}_{Q}[Y^2] = \mathrm{Var}_Q(Y)$$

Since Y is valued in $[a,b]$, $\mathrm{Var}_Q\, Y \le \frac{1}{4}(b-a)^2$. □

Using this fact and returning to the previous computation, we get that

$$\begin{aligned}\frac{\mathrm{Ent}^{(i)}(e^{\lambda Z})}{\mathbb{E}_{X_i}[e^{\lambda Z}]} &= \lambda\psi_i'(\lambda) - \psi_i(\lambda)\\ &= \int_0^\lambda t\psi_i''(t)\,\mathrm{d}t\\ &\le \int_0^\lambda t\cdot\frac{1}{4}(b-a)^2\mathrm{d}t\\ &= \frac{1}{8}\lambda^2(b-a)^2\end{aligned}$$

This implies that

$$\Pr[Z \ge \mathbb{E}[Z]+t] \le e^{-t^2/2\sum c_i^2} \qquad \Pr[Z \le \mathbb{E}[Z]-t] \le e^{-t^2/2\sum c_i^2}$$

To go beyond the bounded differences regime, we will need a Log-Sobolev inequality for more general random variables.

Theorem (Modified Log-Sobolev)**.** *Let $Z = f(X_1,\dots,X_n)$, where X_i are independently valued in $\mathcal{X}$, and let $Z_i = f_i(X^{(i)})$. Then for every $\lambda \in \mathbb{R}$,*

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

where $\Phi(x) = e^x - x - 1$.

Proof. From the connection to Bregman divergence, we know that

$$\mathrm{Ent}(Y) \le \mathbb{E}[Y\log Y - \mathbb{E}[Y]\log u - (Y-u)]$$

for any $u > 0$ and any random variable $Y \ge 0$. We apply subadditivity of entropy and this fact to write

$$\begin{aligned}\mathrm{Ent}(e^{\lambda Z}) &\le \sum_{i=1}^n \mathop{\mathbb{E}}_{X_i}\left[\mathrm{Ent}^{(i)}(e^{\lambda Z})\right]\\ &\le \sum_{i=1}^n \mathop{\mathbb{E}}_{X_i}\left[e^{\lambda Z}\lambda Z - e^{\lambda Z}\lambda Z_i - (e^{\lambda Z} - e^{\lambda Z_i})\right]\\ &= \sum_{i=1}^n \mathop{\mathbb{E}}_{X_i}\left[e^{\lambda Z}(\lambda Z - \lambda Z_i - 1 + e^{\lambda(Z_i - Z)})\right]\\ &= \sum_{i=1}^n \mathop{\mathbb{E}}_{X_i}\left[e^{\lambda Z}\Phi(-\lambda(Z-Z_i))\right]\end{aligned}$$

where we apply the fact above with $u = e^{\lambda Z_i}$, which is a constant once $X^{(i)}$ is fixed. □

Theorem. *Suppose X_i and Z are as above, and we define*

$$Z_i = \inf_{X_i} f(X_1, \ldots, X_n)$$

Suppose that for every $x \in \mathcal{X}^n$,

$$\sum_{i=1}^{n} (Z - Z_i)^2 \leq v$$

Therefore,

$$\Pr(Z \geq \mathbb{E}[Z] + t) \leq e^{-t^2/2v}$$

though we get no lower tail bounds.

Proof. By the modified Log-Sobolev inequality, we know that

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

We have the simple Taylor series bound

$$e^{-x+x-1} \leq \frac{1}{2} x^2$$

for $x \geq 0$. Therefore,

$$\begin{aligned} \mathrm{Ent}(e^{\lambda Z}) &\leq \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}[e^{\lambda Z} \lambda^2 (Z - Z_i)^2] \\ &\leq \frac{1}{2} \lambda^2 v \, \mathbb{E}[e^{\lambda Z}] \end{aligned}$$

and then the rest of the method proceeds as before. Note that since our approximation only holds for $x \geq 0$, we cannot apply this to get a lower tail bound. □

As an application, let's return to the problem of the maximum eigenvalue of a random symmetric matrix, where the entries are $X_{ij} = X_{ji}$, which are independent and valued in $[-1, 1]$. Let Z be the maximum eigenvalue of the matrix $(X_{ij})_{i,j} = 1^n$, and

$$Z_{ij} = \inf_{X_{ij}} Z$$

We saw previously that

$$\sum_{i,j=1}^{n} (Z - Z_{ij})^2 \leq 16$$

Therefore, we get the upper tail bound

$$\Pr(Z \geq \mathbb{E}[Z] + t) \leq e^{-t^2/32}$$

The lower tail is also true, but we don't get it yet.

13 October 29, 2018

13.1 The Entropy Method for Self-Bounding Functions

Recall that $f : \mathcal{X}^n \to \mathbb{R}_+$ is self-bounding if there exist functions $f_i : \mathcal{X}^{n-1} \to \mathbb{R}_+$ such that

$$0 \leq f(x) \leq f_i(x^{(i)}) \leq 1$$

and

$$\sum_{i=1}^{n} (f(x) - f_i(x^{(i)})) \leq f(x)$$

for all $x \in \mathcal{X}^n$.

Theorem. *Suppose $Z = f(X_1, \ldots, X_n)$, where the X_i are independent and f is self-bounding. Then*

$$\log \mathbb{E}[e^{\lambda(Z - \mathbb{E}\, Z)}] \leq \Phi(\lambda)\, \mathbb{E}[Z]$$

for all $\lambda \in \mathbb{R}$, where $\Phi(\lambda) = e^\lambda - \lambda - 1$.

Proof. We will apply the modified Log-Sobolev inequality. It says that

$$\mathrm{Ent}(e^{\lambda Z}) \leq \sum_{i=1}^{n} \mathbb{E}\left[e^{\lambda Z} \Phi(-\lambda(Z - Z_i))\right]$$

where $Z_i = f_i(X^{(i)})$. Observe that Φ is convex, and that $\Phi(0) = \Phi'(0) = 0$; this implies that for $0 \leq u \leq 1$,

$$\Phi(-\lambda u) \leq u\Phi(-\lambda)$$

Therefore,

$$\begin{aligned} \mathrm{Ent}(e^{\lambda Z}) &\leq \mathbb{E}\left[e^{\lambda Z} \Phi(-\lambda) \sum_{i=1}^{n} (Z - Z_i)\right] \\ &\leq \mathbb{E}\left[Z e^{\lambda Z}\right] \Phi(-\lambda) \end{aligned}$$

Writing $F(\lambda) = \mathbb{E}[e^{\lambda Z}]$ and $\mathrm{Ent}(e^{\lambda Z}) = \lambda F'(\lambda) - F(\lambda) \log F(\lambda)$, this becomes

$$\lambda F'(\lambda) - F(\lambda) \log F(\lambda) \leq \Phi(-\lambda) F'(\lambda)$$

or equivalently

$$(\lambda - \Phi(-\lambda)) \frac{F'(\lambda)}{F(\lambda)} \leq \log F(\lambda)$$

Define $G(\lambda) = \log F(\lambda) - \lambda\, \mathbb{E}[Z]$. Then

$$G'(\lambda) = \frac{F'(\lambda)}{F(\lambda)} - \mathbb{E}[Z]$$

and thus our inequality becomes

$$(\lambda - \Phi(-\lambda))(G'(\lambda) + \mathbb{E}[Z]) \leq G(\lambda) + \lambda\,\mathbb{E}[Z]$$

By the definition of Φ, this is just

$$(-e^{-\lambda} + 1)(G'(\lambda) + \mathbb{E}[Z]) \leq G(\lambda) + \lambda\,\mathbb{E}[Z]$$

or equivalently

$$(1 - e^{-\lambda})G'(\lambda) \leq G(\lambda) + (e^{-\lambda} + \lambda - 1)\,\mathbb{E}[Z]$$

Multiplying by e^{λ}, we get

$$(e^{\lambda} - 1)G'(\lambda) \leq e^{\lambda}G(\lambda) + (1 + \lambda e^{\lambda} - e^{\lambda})\,\mathbb{E}[Z]$$

We now divide by $(e^{\lambda} - 1)^2$ to get

$$\frac{G'(\lambda)}{e^{\lambda} - 1} - \frac{e^{\lambda}}{(e^{\lambda} - 1)^2}G(\lambda) \leq \frac{1 + \lambda e^{\lambda} - e^{\lambda}}{(e^{\lambda} - 1)^2}\,\mathbb{E}[Z]$$

Observe that the left-hand side is just the derivative of $G(\lambda)/(e^{\lambda} - 1)$, by the quotient rule. Moreover, the right hand side is the derivative of $\mathbb{E}[Z](e^{\lambda} - \lambda - 1)/(e^{\lambda} - 1)$. So we have

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\frac{G(\lambda)}{e^{\lambda} - 1}\right) \leq \frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\frac{e^{\lambda} - \lambda - 1}{e^{\lambda} - 1}\right)\mathbb{E}[Z]$$

We also have, by l'Hôpital's rule,

$$\lim_{\lambda \to 0} \frac{G(\lambda)}{e^{\lambda} - 1} = \lim_{\lambda \to 0} \frac{G'(\lambda)}{e^{\lambda}} = \lim_{\lambda \to 0} e^{-\lambda}\left(\frac{\mathbb{E}[Ze^{\lambda Z}]}{\mathbb{E}[e^{\lambda Z}]} - \mathbb{E}[Z]\right) = 0$$

and also

$$\lim_{\lambda \to 0} \frac{e^{\lambda} - \lambda - 1}{e^{\lambda} - 1} = \lim_{\lambda \to 0} \frac{e^{\lambda} - 1}{e^{\lambda}} = 0$$

So both sides of our inequality start out with value 0 at $\lambda = 0$. So we can integrate from 0 to λ to eliminate the derivatives and get

$$\frac{G(\lambda)}{e^{\lambda} - 1} \leq \frac{e^{\lambda} - \lambda - 1}{e^{\lambda} - 1}\,\mathbb{E}[Z]$$

for all $\lambda \geq 0$, which is exactly what we wanted to prove (by plugging in the definition of G).

To get the same result for $\lambda < 0$, we run the same argument, and everything works the same. □

Corollary. *For Z as above,*

$$\Pr(Z \geq \mathbb{E}[Z] + t) \leq e^{-\mathbb{E}[Z]h(\frac{t}{\mathbb{E}[Z]})} \qquad \Pr(Z \leq \mathbb{E}[Z] - t) \leq e^{-\mathbb{E}[Z]h(-\frac{t}{\mathbb{E}[Z]})}$$

where $h(u) = (1 + u)\log(1 + u) - u$.

Note that for u small, $h(u)$ behaves roughly quadratically, whereas for u large, it behaves like $u \log u$. Thus, for medium-range deviations, we get Gaussian tail bounds, namely bounds of the form $e^{-ct^2/\mathbb{E}[Z]}$, but for large deviations, we get bounds of the form $e^{-t \log \frac{t}{\mathbb{E}[Z]}}$. Indeed, such results are best possible: suppose $X_1, \ldots, X_n$ are Bernoulli variables with $\Pr(X_i = 1) = 1/n$, and f is the summation function. So $\mathbb{E}[Z] = 1$, and thus

$$\Pr(Z \geq n\,\mathbb{E}[Z]) = \frac{1}{n^n} = e^{-n \log n}$$

Thus, for such an example, the Gaussian bound is false for very large deviations, and an exponent of the form $t \log t$ is the best we can hope for.

Proof of Corollary. We know that

$$\mathbb{E}[e^{\lambda(Z - \mathbb{E}[Z])}] \leq e^{\Phi(\lambda)\,\mathbb{E}[Z]}$$

Therefore, by Markov's inequality,

$$\Pr(Z \geq \mathbb{E}[Z] + t) \leq \frac{\mathbb{E}[e^{\lambda(Z - \mathbb{E}[Z])}]}{e^{\lambda t}} \leq e^{\Phi(\lambda)\,\mathbb{E}[Z] - \lambda t}$$

To choose λ, we differentiate the exponent and set it equal to zero, getting $t = \Phi'(\lambda)\,\mathbb{E}[Z]$. Thus, we pick $\lambda = \log(1 + t/\,\mathbb{E}[Z])$. Plugging it in, we get

$$\Pr(Z \geq \mathbb{E}[Z] + t) \leq e^{(e^\lambda - \lambda - 1)\,\mathbb{E}[Z] - \lambda t} = e^{-\,\mathbb{E}[Z] h(\frac{t}{\mathbb{E}[Z]})}$$

and similarly for the lower tail. □

Slightly simpler, though weaker, tail bounds are as follows:

$$\Pr(Z \geq \mathbb{E}[Z] + t) \leq e^{-\frac{t^2}{2\,\mathbb{E}[Z] + 2t/3}} \qquad \Pr(Z \leq \mathbb{E}[Z] - t) \leq e^{-\frac{t^2}{2\,\mathbb{E}[Z]}}$$

13.2 Applications

Recall that we were considering combinatorial entropy functions, which were defined by

$$h(x_1, \ldots, x_n) = \log_2 |\mathrm{tr}_{\mathcal{A}}(x_1, \ldots, x_n)|$$

We saw that this function was self-bounding, so we get exponential tail bounds for it. However, it's also interesting to look directly at the bound we get for the exponential moment:

$$\log \mathbb{E}[e^{\lambda(h(X) - \mathbb{E}[h(X)])}] \leq \Phi(\lambda)\,\mathbb{E}[h(X)]$$

Setting $\lambda = \log 2$, we get

$$\log \mathbb{E}[2^{\log_2 |\mathrm{tr}(X)| - \mathbb{E}[h(X)]}] \leq (2 - \log 2 - 1)\,\mathbb{E}[h(X)]$$

or equivalently

$$\log \mathbb{E}[|\mathrm{tr}(X)|] - \log 2 \cdot \mathbb{E}[h(X)] \leq (1 - \log 2)\, \mathbb{E}[h(X)]$$

which simplifies to

$$\log \mathbb{E}[|\mathrm{tr}(X)|] \leq \mathbb{E}[\log_2 |\mathrm{tr}(X)|] = \frac{1}{\log 2}\, \mathbb{E}[\log |\mathrm{tr}(X)|]$$

Combining this with Jensen's inequality, we have

$$\mathbb{E}[\log |\mathrm{tr}(X)|] \leq \log \mathbb{E}[|\mathrm{tr}(X)|] \leq \frac{1}{\log 2}\, \mathbb{E}[\log |\mathrm{tr}(X)|]$$

In other words, for this variable, switching the expectation and the log costs a constant factor at worst. In general, such switches can be arbitrarily expensive.

14 October 31, 2018

14.1 Weakly self-bounding functions

Definition. We say that $f : \mathcal{X}^n \to \mathbb{R}_+$ is *strongly* (a, b)*-self-bounding* if there exist $f_i : \mathcal{X}^{n-1} \to \mathbb{R}_+$ such that

$$0 \leq f(x) - f_i(x^{(i)}) \leq 1$$

and

$$\sum_{i=1}^{n} (f(x) - f_i(x^{(i)})) \leq a f(x) + b$$

Definition. We say that $f : \mathcal{X}^n \to \mathbb{R}_+$ is *weakly* (a, b)*-self-bounding* if there exist $f_i : \mathcal{X}^{n-1} \to \mathbb{R}_+$ such that

$$\sum_{i=1}^{n} (f(x) - f_i(x^{(i)}))^2 \leq a f(x) + b$$

Recall that for Efron–Stein, the weakly (a, b)-self-bounding property suffices to imply that

$$\mathrm{Var}(f) \leq a\, \mathbb{E}[f] + b$$

Thus, this weakening is natural from the perspective of Efron–Stein.

Theorem. *If $Z = f(X_1, \ldots, X_n)$, where X_i are independent, f is weakly (a, b)-self-bounding, and $f_i(x^{(i)}) \leq f(x)$ for all x, then for any $\lambda \in [0, 2/a]$,*

$$\log \mathbb{E}[e^{\lambda(Z - \mathbb{E}[Z])}] \leq \frac{(a\, \mathbb{E}[Z] + b)\lambda^2}{2 - a\lambda}$$

which is what we call "sub-Gamma behavior". This implies the tail bound

$$\Pr(Z \geq \mathbb{E}[Z] + t) \leq e^{-\frac{t^2}{2(a\, \mathbb{E}[Z] + b) + at}}$$

for any $t > 0$.

Proof. Define $Z_i = f_i(X^{(i)})$, so that $Z - Z_i \geq 0$. Recall that we have the bound

$$\Phi(x) = e^x - x - 1 \leq \frac{1}{2}x^2$$

for all $x \leq 0$. The modified Log-Sobolev inequality says that

$$\begin{aligned}\mathrm{Ent}(e^{\lambda Z}) &\leq \sum_{i=1}^{n} \mathbb{E}[e^{\lambda Z}\Phi(-\lambda(Z - Z_i))] \\ &\leq \sum_{i=1}^{n} \mathbb{E}\left[e^{\lambda Z} \cdot \frac{1}{2}\lambda^2 (Z - Z_i)^2\right] \\ &\leq \frac{1}{2}\lambda^2 \,\mathbb{E}[e^{\lambda Z}(aZ + b)]\end{aligned}$$

Let $v = a\,\mathbb{E}[Z] + b$. Then this computation implies the differential inequality

$$\lambda F'(\lambda) - F(\lambda)\log F(\lambda) \leq \frac{1}{2}\lambda^2(aF'(\lambda) + bF(\lambda))$$

for $F(\lambda) = \mathbb{E}[e^{\lambda Z}]$. Letting $G(\lambda) = \log F(\lambda) - \lambda\,\mathbb{E}[Z]$, this becomes

$$\lambda G'(\lambda) - G(\lambda) \leq \frac{1}{2}a\lambda^2(G'(\lambda) + \mathbb{E}[Z]) + \frac{1}{2}b\lambda^2$$

Dividing by λ^2, we get

$$\frac{G'(\lambda)}{\lambda} - \frac{G(\lambda)}{\lambda^2} \leq \frac{1}{2}aG'(\lambda) + \frac{1}{2}v$$

or equivalently

$$\left(\frac{1}{\lambda} - \frac{1}{2}a\right)G'(\lambda) - \frac{1}{\lambda^2}G(\lambda) \leq \frac{1}{2}v$$

This is just

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\left(\frac{1}{\lambda} - \frac{1}{2}a\right)G(\lambda)\right) \leq \frac{1}{2}v$$

To integrate this, we need to know the value at zero. We already saw, by l'Hôpital's rule, that $\lim_{\lambda\to 0} G(\lambda) = \lim_{\lambda\to 0} G'(\lambda) = 0$, so the left-hand side is zero at $\lambda = 0$. Thus, we can integrate to get

$$\left(\frac{1}{\lambda} - \frac{1}{2}a\right)G(\lambda) \leq \frac{1}{2}v\lambda$$

or equivalently

$$G(\lambda) \leq \frac{v\lambda^2}{2 - a\lambda}$$

This is the desired bound on the exponential moment. The tail bound follows from general computations about sub-Gamma variables, which we didn't do in detail before, and are spelled out in the next theorem. □

Theorem. *If Z is sub-Gamma with parameters v, c, namely if we have a bound*

$$\psi(\lambda) = \mathbb{E}[e^{\lambda(Z-\mathbb{E}[Z])}] \leq \frac{v\lambda^2}{2(1-c\lambda)}$$

then

$$\psi^*(t) = \sup_{\lambda \in [0, \frac{1}{c})} \left(\lambda t - \frac{\lambda^2 v}{2(1-c\lambda)} \right) = \frac{v}{c^2} \left(1 + \frac{ct}{v} - \sqrt{1 + \frac{2ct}{v}} \right)$$

with the supremum attained at

$$\lambda^* = \frac{1}{c} \left(1 - \frac{1}{\sqrt{1 + \frac{2ct}{v}}} \right)$$

This implies the tail bound of

$$\Pr(Z \geq \mathbb{E}[Z] + t) \leq e^{-\psi^*(t)}$$

When t is small,

$$\psi^*(t) \approx \frac{t^2}{2v}$$

so the sub-Gamma bound matches the sub-Gaussian bound for small t, whereas for large t,

$$\psi^*(t) \approx \frac{t}{c}$$

A general bound that holds for all $t > 0$ is

$$\Pr(Z \geq \mathbb{E}[Z] + t) \leq e^{-\frac{t^2}{2(v+ct)}}$$

We can also obtain a lower tail bound for weakly (a, b)-self-bounding functions, though it is significantly more difficult and technical to prove.

Theorem. *Let*

$$c_- = \max \left\{ \frac{1-3a}{6}, 0 \right\}$$

and let $Z = f(X_1, \ldots, X_n)$, where X_i are independent, f is weakly (a, b)-self-bounding, and $0 \leq f(x) - f_i(x^{(i)}) \leq 1$ for all x. Then for any $t \in [0, \mathbb{E}[Z]]$,

$$\Pr(Z \leq \mathbb{E}[Z] - t) \leq e^{-\frac{t^2}{2(a\,\mathbb{E}[Z] + b + c_- t)}}$$

Proof. We have that

$$\frac{1}{x^2}\Phi(x) = \frac{e^x - x - 1}{x^2}$$

is increasing for $x \geq 0$. Therefore, for $\lambda < 0$,

$$\frac{\Phi(-\lambda(Z - Z_i))}{\lambda^2 (Z - Z_i)^2} \leq \frac{\Phi(-\lambda)}{\lambda^2}$$

which implies that

$$\Phi(-\lambda(Z - Z_i)) \leq \Phi(-\lambda)(Z - Z_i)^2$$

As usual, we now apply the modified Log-Sobolev inequality for $\lambda < 0$. We get that

$$\begin{aligned}\mathrm{Ent}(e^{\lambda Z}) &\leq \sum_{i=1}^{n} \mathbb{E}[e^{\lambda Z}\Phi(-\lambda(Z - Z_i))] \\ &\leq \mathbb{E}[e^{\lambda Z}\Phi(-\lambda)(Z - Z_i)^2] \\ &\leq \Phi(-\lambda)\,\mathbb{E}[e^{\lambda Z}(aZ + b)]\end{aligned}$$

Using the same notation as before, this implies the differential inequality

$$(\lambda - a\Phi(-\lambda))G'(\lambda) - G(\lambda) \leq \Phi(-\lambda)v$$

This differential inequality is harder to solve than the previous one, but it can be solved, and one can conclude

$$G(\lambda) \leq \frac{v\lambda^2}{2(1 + c\lambda)}$$

for $\lambda \in (-\theta, 0)$, where

$$\theta = \begin{cases} \frac{1}{a} & \text{if } c_- = 0 \\ \frac{1}{c_-}(1 - \sqrt{1 - 6c_-}) & \text{otherwise} \end{cases}$$

□

For an application of these results, we will consider concentration of the norm. Let

$$Z = g(X_1, \ldots, X_n) = \|\vec{X}\| = \left(\sum_{i=1}^{n} X_i^2\right)^{1/2}$$

Suppose that X_i are independent and valued in $[0, 1]$, but we don't know anything else about their distribution. We want to understand the concentration of Z. If we apply a general approach like Azuma (using the fact that g is 1-Lipschitz), then we will get concentration of order $\sqrt{n}$, which is not very good since $0 \leq Z \leq \sqrt{n}$. We can do better by applying the previous results.

Proposition. *Z is concentrated in a constant-sized window. Specifically,*

$$\Pr(Z \geq \mathbb{E}[Z] + t) \leq e^{-t^2/2} \qquad \Pr(Z \leq \mathbb{E}[Z] - t) \leq e^{-t^2/8}$$

Proof. The upper tail follows from the fact that g is convex, along with the following theorem.

Theorem. *If $Z = g(X_1, \ldots, X_n)$, where g is differentiable and convex with respect to each variable, and if $\|\nabla g\| \leq 1$, and if X_i are independent in $[0,1]$, then*

$$\Pr(Z \geq \mathbb{E}[Z] + t) \leq e^{-t^2/2}$$

Proof. Let

$$Z_i = \inf_{X_i} g(X_1, \ldots, X_n)$$

Then by convexity,

$$\sum_{i=1}^{n} (Z - Z_i)^2 \leq \sum_{i=1}^{n} \left(\left. \frac{\partial g}{\partial x_i} \right|_X |X - X_i| \right)^2 \leq \|\nabla g\|^2 \leq 1$$

We saw as an early application of the entropy method that if we have pointwise bounds on the sum of the squared differences, we automatically get sub-Gaussian upper tail bounds. □

For the lower tail, we need to use a different technique, which we'll see next time. □

15 November 2, 2018

Let's return to (a slight generalization of) what we were doing last time. Let $X_1, \ldots, X_n$ be independent and valued in $[0,1]$, and let

$$Z = g(X) = \|X\|_p$$

for $p \geq 2$. Then g is convex and 1-Lipschitz. This implies that

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

where $Z_i = \inf_{X_i} g(X)$. This, in turn, implies an exponential upper tail bound, as we saw earlier. Alternatively, this says that g is weakly $(0,1)$-self-bounding, which also gives the upper tail bound

$$\Pr(Z \geq \mathbb{E}[Z] + t) \leq e^{-t^2/2}$$

For the lower tail, we have the constant $c_- = 1/6$, and thus the results for weakly self-bounding functions give us

$$\Pr(Z \leq \mathbb{E}[Z] - t) \leq e^{-\frac{t^2}{2+t/3}}$$

which is not as good as we want; the standard deviation of Z is $O(1)$, so the additional t in the denominator becomes relevant for even small values of t, and this would only give us simple exponential lower tail bounds. We can do better, as follows.

We first claim that

$$0 \leq Z^2 - Z_i^2 \leq 1$$

Indeed,

$$Z^2 - Z_i^2 = \left(\sum_j X_j^p\right)^{2/p} - \left(\sum_{j \neq i} X_j^p\right)^{2/p}$$

Since $2/p \leq 1$, the function $y \mapsto y^{2/p}$ is concave. Thus, since $X_i \in [0,1]$, we get that $Z^2 - Z_i^2 \leq 1$. We also get that it's at least zero since the function $y \mapsto y^{2/p}$ is increasing.

Theorem. *Suppose $Z = g(X_1, \ldots, X_n)$, where X_i are independent and $g \geq 0$. Suppose also that*

$$\sum_{i=1}^n (Z - Z_i)^2 \leq v$$

for $v \geq 1/12$, and that

$$0 \leq Z^2 - Z_i^2 \leq 1$$

Then

$$\Pr(Z \leq \mathbb{E}[Z] - t) \leq e^{-t^2/8v}$$

for $t \in [0, \mathbb{E}[Z]]$.

For the norm application, $v = 1$, so we get a lower tail bound of $e^{-t^2/8}$.

Proof. Set $f(x) = g^2(x)$ and

$$f_i(x) = g_i^2(x) = \inf_{x_i} g^2(x)$$

We know by assumption that $0 \leq f(x) - f_i(x) \leq 1$. We also have

$$\begin{aligned}
\sum_{i=1}^n (f(x) - f_i(x))^2 &= \sum_{i=1}^n (g^2(x) - g_i^2(x))^2 \\
&= \sum_{i=1}^n (g(x) + g_i(x))^2 (g(x) - g_i(x))^2 \\
&\leq 4g^2(x) \sum_{i=1}^n (g(x) - g_i(x))^2 \\
&\leq 4vg^2(x) \\
&= 4vf(x)
\end{aligned}$$

Thus, f is weakly $(4v, 0)$-self-bounding. Plugging this fact into our general theorem, and using the fact that $v \geq 1/12$ so that $c_- = 0$, we get that

$$\Pr(f(X) \leq \mathbb{E}[f(X)] - t) \leq e^{-t^2/8v\,\mathbb{E}[f(X)]}$$

for $t \in [0, \mathbb{E}[g]]$. Therefore,

$$\begin{aligned}
\Pr(g(X) \leq \mathbb{E}[g(X)] - t) &\leq \Pr\left(g(X) \leq \sqrt{\mathbb{E}[g^2(X)]} - t\right) \\
&= \Pr\left(g^2(X) \leq \mathbb{E}[g^2(X)] - 2t\sqrt{\mathbb{E}[g^2(X)]} + t^2\right) \\
&\leq \Pr\left(g^2(X) \leq \mathbb{E}[g^2(X)] - t\sqrt{\mathbb{E}[g^2(X)]}\right) \\
&= \Pr(f(X) \leq \mathbb{E}[f(X)] - t\sqrt{\mathbb{E}[f(X)]}) \\
&\leq e^{-t^2/8v}
\end{aligned}$$

□

15.1 Janson's Inequality

Suppose we have a ground set N with $|N| = n$, and a family of subsets $\mathcal{I} \subseteq 2^N$. For each $i \in N$, we have a Bernoulli random variable X_i with parameter $\mathbb{E}[X_i] = p_i$. For each $\alpha \in \mathcal{I}$, we define

$$Y_\alpha = \prod_{i \in \alpha} X_i$$

Finally, we define

$$Z = \sum_{\alpha \in \mathcal{I}} Y_\alpha$$

As an application, N could be the edges of a random graph, $\mathcal{I}$ could consist of all potential copies of a subgraph, and Z counts the number of times this subgraph appears. We are intersted in bounding $\Pr(Z = 0)$.

Observe that Y_α and Y_β are independent as long as $\alpha \cap \beta = \emptyset$. Therefore,

$$\begin{aligned}
\mathrm{Var}(Z) &= \sum_{\alpha,\beta \in \mathcal{I}} \mathrm{Cov}(Y_\alpha, Y_\beta) \\
&= \sum_{\substack{\alpha,\beta \in \mathcal{I} \\ \alpha \cap \beta \neq \emptyset}} (\mathbb{E}[Y_\alpha Y_\beta] - \mathbb{E}[Y_\alpha]\,\mathbb{E}[Y_\beta]) \\
&\leq \sum_{\substack{\alpha,\beta \in \mathcal{I} \\ \alpha \cap \beta \neq \emptyset}} \mathbb{E}[Y_\alpha Y_\beta] =: \Delta
\end{aligned}$$

In most applications, we don't lose much by dropping the term $\mathbb{E}[Y_\alpha]\,\mathbb{E}[Y_\beta]$ since it'll generally be lower-order, so Δ will be a good approximation to $\mathrm{Var}(Z)$.

Theorem (Janson's Inequality)**.**

$$\Pr(Z = 0) \leq e^{-\frac{(\mathbb{E}[Z])^2}{2\Delta}}$$

This is "Poisson-type behavior", and in general, if Δ is on the same order as $\mathbb{E}[Z]$ (which is often what happens), it gives an exponential bound on $\Pr(Z = 0)$. Though there is a different proof, we will show how to derive it from the entropy method. Instead of the subadditivity of entropy, we will need some association inequalities for monotone functions (along the lines of the FKG inequality).

Theorem (Chebyshev's association inequality)**.** *Let $f, g : \mathbb{R} \to \mathbb{R}$ be non-decreasing, and X a real-valued random variable. Then*

$$\mathbb{E}[f(X)g(X)] \geq \mathbb{E}[f(X)]\,\mathbb{E}[g(X)]$$

Proof. Let X' be an independent copy of X. Then first observe that

$$\mathbb{E}[(f(X) - f(X'))(g(X) - g(X'))] \geq 0$$

This is because either $X \leq X'$ or $X' \leq X$, and in either case, the fact that f and g are both monotone increasing gives that the product is non-negative. On the other hand, we can expand

$$\begin{aligned} 0 &\leq \mathbb{E}[(f(X) - f(X'))(g(X) - g(X'))] \\ &= \mathbb{E}[f(X)g(X)] - \mathbb{E}[f(X)]\,\mathbb{E}[g(X')] - \mathbb{E}[f(X')g(X)] + \mathbb{E}[f(X')g(X')] \\ &= 2\,\mathbb{E}[f(X)g(X)] - 2\,\mathbb{E}[f(X)]\,\mathbb{E}[g(X)] \end{aligned}$$

which gives the desired bound. □

Theorem (Harris's Inequality)**.** *Let $f, g : \mathbb{R}^n \to \mathbb{R}$ be two functions that are both coordinate-wise non-decreasing. Let $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$, where X_i are independent $\mathbb{R}$-valued variables. Then*

$$\mathbb{E}[f(X)g(X)] \geq \mathbb{E}[f(X)]\,\mathbb{E}[g(X)]$$

Proof. The one-dimensional case is Chebyshev's inequality, above. For higher dimensions, we induct on n. We write

$$\begin{aligned} \mathbb{E}[f(X)g(X)] &= \mathop{\mathbb{E}}_{X_1,\ldots,X_{n-1}}\left[\mathop{\mathbb{E}}_{X_n}[f(X)g(X) \mid X_1, \ldots, X_{n-1}]\right] \\ &\geq \mathop{\mathbb{E}}_{X_1,\ldots,X_{n-1}}\left[\mathop{\mathbb{E}}_{X_n}[f(X) \mid X_1, \ldots, X_{n-1}]\mathop{\mathbb{E}}_{X_n}[g(X) \mid X_1, \ldots, X_{n-1}]\right] \\ &\geq \mathop{\mathbb{E}}_{X_1,\ldots,X_{n-1}}\left[\widetilde{f}(X_1, \ldots, X_{n-1})\widetilde{g}(X_1, \ldots, X_{n-1})\right] \\ &\geq \mathop{\mathbb{E}}_{X_1,\ldots,X_{n-1}}[\widetilde{f}(X_1, \ldots, X_{n-1})]\mathop{\mathbb{E}}_{X_1,\ldots,X_{n-1}}[\widetilde{g}(X_1, \ldots, X_{n-1})] \\ &= \mathbb{E}[f(X)]\,\mathbb{E}[g(X)] \end{aligned}$$

where we apply the inductive assumption to $\widetilde{f}, \widetilde{g}$; it is a simple matter to check that they too are coordinate-wise non-decreasing. □

Proof of Janson's inequality. Recall that $Z = \sum_{\alpha \in \mathcal{I}} Y_\alpha$. Let, as usual

$$G(\lambda) = \log \mathbb{E}[e^{\lambda(Z - \mathbb{E}[Z])}] = \log \mathbb{E}[e^{\lambda Z}] - \lambda \mathbb{E}[Z]$$

so that

$$\begin{aligned} G'(\lambda) &= \frac{\mathbb{E}[Ze^{\lambda Z}]}{\mathbb{E}[e^{\lambda Z}]} - \mathbb{E}[Z] \\ &= \sum_{\alpha \in \mathcal{I}} \frac{\mathbb{E}[Y_\alpha e^{\lambda Z}]}{\mathbb{E}[e^{\lambda Z}]} - \mathbb{E}[Z] \end{aligned}$$

Since Y_α is an indicator variable,

$$\mathbb{E}[Y_\alpha e^{\lambda Z}] = \mathbb{E}[e^{\lambda Z} \mid Y_\alpha = 1]\, \mathbb{E}[Y_\alpha]$$

Define

$$U_\alpha = \sum_{\beta \in \mathcal{I} : \beta \cap \alpha \neq \emptyset} Y_\alpha \qquad V_\alpha = \sum_{\beta \in \mathcal{I} : \beta \cap \alpha = \emptyset} Y_\beta$$

so that $Z = U_\alpha + V_\alpha$, and

$$\Delta = \mathbb{E}\left[\sum_\alpha Y_\alpha U_\alpha\right]$$

We can also write, by Harris's inequality,

$$\begin{aligned} \mathbb{E}[e^{\lambda Z} \mid Y_\alpha = 1] &= \mathbb{E}[e^{\lambda U_\alpha} e^{\lambda V_\alpha} \mid Y_\alpha = 1] \\ &\geq \mathbb{E}[e^{\lambda U_\alpha} \mid Y_\alpha = 1]\, \mathbb{E}[e^{\lambda V_\alpha} \mid Y_\alpha = 1] \\ &\geq \mathbb{E}[e^{\lambda U_\alpha} \mid Y_\alpha = 1]\, \mathbb{E}[e^{\lambda Z}] \\ &\geq e^{\mathbb{E}[\lambda U_\alpha \mid Y_\alpha = 1]}\, \mathbb{E}[e^{\lambda Z}] \end{aligned}$$

for all $\lambda < 0$. □

16 November 5, 2018

Recall that we were proving Janson's inequality. The setup was that we had a set system $\mathcal{I}$, and defined

$$Z = \sum_{\alpha \in \mathcal{I}} Y_\alpha$$

where $Y_\alpha = \prod_{i \in \alpha} X_i$, and X_i are independent Bernoulli variables. We also defined

$$\Delta = \mathbb{E}\left[\sum_{\alpha \cap \beta \neq \emptyset} Y_\alpha Y_\beta\right]$$

and then Janson's inequality asserted that

$$\Pr(Z \leq \mathbb{E}[Z] - t) \leq e^{-t^2/2\Delta}$$

Continuation of Proof. We defined

$$G(\lambda) = \log \mathbb{E}[e^{\lambda(Z-\mathbb{E}[Z])}]$$

so that

$$G'(\lambda) = \frac{\mathbb{E}[Ze^{\lambda Z}]}{\mathbb{E}[e^{\lambda Z}]} - \mathbb{E}[Z] = \sum_{\alpha\in\mathcal{I}} \frac{\mathbb{E}[Y_\alpha e^{\lambda Z}]}{\mathbb{E}[e^{\lambda Z}]} - \mathbb{E}[Z]$$

and we showed that

$$\mathbb{E}[e^{\lambda Z} \mid Y_\alpha = 1] \geq \mathbb{E}[e^{\lambda U_\alpha} \mid Y_\alpha = 1] \cdot \mathbb{E}[e^{\lambda Z}]$$

for $\lambda \leq 0$. By Jensen's inequality, this is at least

$$e^{\mathbb{E}[\lambda U_\alpha | Y_\alpha = 1]} \cdot \mathbb{E}[e^{\lambda Z}]$$

This implies that

$$\frac{\mathbb{E}[e^{\lambda Z} \mid Y_\alpha = 1]}{\mathbb{E}[e^{\lambda Z}]} \geq e^{\mathbb{E}[\lambda U_\alpha | Y_\alpha = 1]}$$

and therefore

$$\begin{aligned}
\frac{\mathbb{E}[Ze^{\lambda Z}]}{\mathbb{E}[e^{\lambda Z}]} &= \sum_{\alpha\in\mathcal{I}} \frac{\mathbb{E}[e^{\lambda Z} \mid Y_\alpha = 1]\,\mathbb{E}[Y_\alpha]}{\mathbb{E}[e^{\lambda Z}]} \\
&\geq \sum_{\alpha\in\mathcal{I}} e^{\mathbb{E}[\lambda U_\alpha | Y_\alpha = 1]}\,\mathbb{E}[Y_\alpha] \\
&= \mathbb{E}[Z] \sum_{\alpha\in\mathcal{I}} \frac{\mathbb{E}[Y_\alpha]}{\mathbb{E}[Z]} e^{\mathbb{E}[\lambda U_\alpha | Y_\alpha = 1]} \\
&\geq \mathbb{E}[Z] e^{\sum_{\alpha\in\mathcal{I}} \frac{\mathbb{E}[Y_\alpha]}{\mathbb{E}[Z]}\,\mathbb{E}[\lambda U_\alpha | Y_\alpha = 1]}
\end{aligned}$$

where in the last step we use Jensen again. Also observe that

$$\sum_{\alpha\in\mathcal{I}} \mathbb{E}[Y_\alpha]\,\mathbb{E}[\lambda U_\alpha \mid Y_\alpha = 1] = \sum_{\alpha\in\mathcal{I}} \mathbb{E}[\lambda U_\alpha Y_\alpha] = \sum_{\substack{\alpha\in\mathcal{I} \\ \beta:\beta\cap\alpha\neq\emptyset}} \mathbb{E}[\lambda Y_\beta Y_\alpha] = \lambda\Delta$$

Therefore, continuing the above computation, we find that

$$\frac{\mathbb{E}[Ze^{\lambda Z}]}{\mathbb{E}[e^{\lambda Z}]} \geq \mathbb{E}[Z] e^{\lambda\Delta/\mathbb{E}[Z]}$$

This implies that

$$G'(\lambda) = \frac{\mathbb{E}[Ze^{\lambda Z}]}{\mathbb{E}[e^{\lambda Z}]} - \mathbb{E}[Z] \geq \left(e^{\lambda\Delta/\,\mathbb{E}[Z]} - 1\right) \mathbb{E}[Z]$$

for $\lambda \leq 0$. Integrating from 0 to $\lambda < 0$, this implies that

$$\begin{aligned} G(\lambda) &\leq -\int_{\lambda}^{0} \left(e^{\lambda' \Delta / \mathbb{E}[Z]} - 1\right) \mathbb{E}[Z] \, \mathrm{d}\lambda' \\ &= \left(\frac{\mathbb{E}[Z]}{\Delta} e^{\lambda \Delta / \mathbb{E}[Z]} - \lambda - \frac{\mathbb{E}[Z]}{\Delta}\right) \mathbb{E}[Z] \\ &= \frac{\mathbb{E}[Z]^2}{\Delta} \left(e^{\lambda \Delta / \mathbb{E}[Z]} - \frac{\lambda \Delta}{\mathbb{E}[Z]} - 1\right) \\ &= \frac{\mathbb{E}[Z]^2}{\Delta} \Phi\left(\frac{\lambda \Delta}{\mathbb{E}[Z]}\right) \\ &\leq \frac{\mathbb{E}[Z]^2}{\Delta} \cdot \frac{1}{2} \left(\frac{\lambda \Delta}{\mathbb{E}[Z]}\right)^2 \\ &= \frac{1}{2} \lambda^2 \Delta \end{aligned}$$

This is a sub-Gaussian bound, so by the standard argument, it implies that

$$\Pr(Z \leq \mathbb{E}[Z] - t) \leq e^{-t^2/2\Delta}$$

as desired. □

As an application of Janson's inequality, we will consider the number of triangles in $G_{n,p}$. We have random variables X_e for each edge, which are independent $\mathrm{Ber}(p)$ variables. For a triple of vertices α, we have an indicator variable for the triangle appearing, $Y_\alpha = X_{e_1} X_{e_2} X_{e_3}$. Then the number of triangles in the random graph is

$$Z = \sum_{\alpha \in \mathcal{I}} Y_\alpha$$

Then

$$\mathbb{E}[Z] = \binom{n}{3} p^3$$

and

$$\mathrm{Var}(Z) = \sum_{\alpha, \beta} \mathrm{Cov}(Y_\alpha, Y_\beta) = \binom{n}{3} p^3 (1 - p^3) + 2 \binom{n}{4} \binom{4}{2} p^5 (1 - p)$$

while

$$\Delta = \binom{n}{3} p^3 + 2 \binom{n}{4} \binom{4}{2} p^5$$

Note that if $p \gg 1/n$, then $\mathbb{E}[Z] \to \infty$, so we expect many triangles. Janson's inequality implies that with high probability, we do indeed see triangles; indeed,

$$\Pr(Z = 0) \leq e^{-\mathbb{E}[Z]^2/2\Delta} \approx e^{-\frac{(n^3 p^3/6)^2}{n^3 p^3/3 + n^4 p^5}} = e^{-\frac{n^3 p^3}{12 + 36 n p^2}}$$

which is exponentially small.

16.1 Concentration vs. Isoperimetry

There is a close connection between concentration of measure and isoperimetric inequalities. The basic connection comes from Lévy's inequalities, as follows. Suppose $\mathcal{X}$ is a metric space with a probability measure P, and $f : \mathcal{X} \to \mathbb{R}$ is 1-Lipschitz. Let m be the median of f, and let $A = \{x \in \mathcal{X} : f(x) \leq m\}$, so that $\Pr(A) \geq \frac{1}{2}$. Since f is 1-Lipschitz, if $f(x) > m + t$, then $x \notin A_t$, where

$$A_t = \{y : d(y, A) \leq t\}$$

Lévy's inequalities, which follow immediately from this observation, say that

$$\Pr(f(X) \geq m + t) \leq \alpha(t) \qquad \Pr(f(X) < m - t) \leq \alpha(t)$$

where

$$\alpha(t) = \sup_{A:P(A)\geq\frac{1}{2}} \Pr(d(X, A) \geq t) = \sup_{A:P(A)\geq\frac{1}{2}} (1 - P(A_t))$$

Thus, the concentration of X about its median is connected to isoperimetry of large sets in $\mathcal{X}$. The connection goes both ways; we will first show how to derive an isoperimetric inequality from a concentration bound we already know. Recall the following theorem.

Theorem. *Suppose $Z = f(X_1, \ldots, X_n)$, where X_i are independent and f is 1-Lipschitz. Then*

$$\Pr(Z \geq \mathbb{E}[Z] + t) \leq e^{-2t^2/n} \qquad \Pr(Z \leq \mathbb{E}[Z] - t) \leq e^{-2t^2/n}$$

Note that this exponent is stronger by a factor of 4 from the naïve bound one gets from Azuma's inequality. The reason is that in the martingale for this process, not only do we have that the martingale differences Z_k satisfy $|Z_k| \leq 1$, but in fact, conditional on $X_1, \ldots, X_k$, we have that Z_{k+1} is in some interval of length 1, which saves us a factor of 4 in the variance. From this concentration bound, we can derive the following isoperimetric inequality.

Theorem. *If $\mathcal{X}^n$ is a product probability space, $A \subseteq \mathcal{X}^n$, and $t > 0$, then*

$$\Pr_X\left(d_H(X, A) \geq t + \sqrt{\frac{n}{2}\log\frac{1}{P(A)}}\right) \leq e^{-2t^2/n}$$

where d_H denotes the Hamming distance on $\mathcal{X}^n$ (i.e. the L^1 metric induced from the discrete metric on $\mathcal{X}$).

Proof. Let $f(x) = d_H(x, A)$, so that f is 1-Lipschitz with respect to d_H. Then by the concentration bound,

$$\Pr(d_H(X, A) \leq \mathbb{E}[d_H(X, A)] - t) \leq e^{-2t^2/n}$$

Therefore,

$$P(A) = \Pr_X(d_H(X, A) = 0) \leq e^{-2\,\mathbb{E}[d_H(X,A)]^2/n}$$

Thus,

$$\mathbb{E}[d_H(X,A)] \leq \sqrt{\frac{n}{2}\log\frac{1}{P(A)}}$$

Therefore, the upper tail bound immediately implies the desired bound:

$$\begin{aligned}\Pr_X\left(d_H(X,A) \geq t + \sqrt{\frac{n}{2}\log\frac{1}{P(A)}}\right) &\leq \Pr(d_H(X,A) \geq \mathbb{E}[d_H(X,A)] + t) \\ &\leq e^{-2t^2/n}\end{aligned}$$

□

17 November 7, 2018

17.1 Euclidean isoperimetric inequality

The following is a variant of the classical isoperimetric inequality.

Theorem. *Let A be a compact set in $\mathbb{R}^n$, with $\mathrm{vol}(A) = \mathrm{vol}(B)$, where B is the open Euclidean unit ball in $\mathbb{R}^n$. Then $\mathrm{vol}(A_t) \geq \mathrm{vol}(B_t)$, where $A_t = \{x : d(x,A) < t\}$.*

Note that we can derive the more standard version of the isoperimetric inequality, which speaks about the $(n-1)$-dimensional volume of ∂A, by looking at

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{vol}(A_t)\bigg|_{t=0} = \lim_{t\to 0}\frac{\mathrm{vol}(A_t) - \mathrm{vol}(A)}{t}$$

To prove this, we will need the Brunn–Minkowski inequality. Given $A, B \subset \mathbb{R}^n$, define

$$A + B = \{x + y : x \in A, y \in B\}$$

Then the Brunn–Minkowski inequality asserts that if A, B are compact, then

$$(\mathrm{vol}(A+B))^{1/n} \geq \mathrm{vol}(A)^{1/n} + \mathrm{vol}(B)^{1/n}$$

Proof of isoperimetry. Write $A_t = A + tB$. By Brunn–Minkowski,

$$\begin{aligned}\mathrm{vol}(A_t)^{1/n} &\geq \mathrm{vol}(A)^{1/n} + t\,\mathrm{vol}(B)^{1/n} \\ &= (1+t)\,\mathrm{vol}(B)^{1/n} \\ &= \mathrm{vol}((1+t)B)^{1/n} \\ &= \mathrm{vol}(B_t)^{1/n}\end{aligned}$$

□

So it only remains to prove the Brunn–Minkowski inequality, which we will first do in dimension 1. Observe that the statement of Brunn–Minkowski is invariant under translations of A and B, so we may assume that $\sup A = 0 = \inf B$. Therefore, $A \cup B \subseteq A + B$, and A and B are disjoint except at 0. Therefore,

$$\mathrm{vol}(A + B) \geq \mathrm{vol}(A \cup B) = \mathrm{vol}(A) + \mathrm{vol}(B)$$

To prove Brunn–Minkowski in higher dimensions, we will need the following inequality.

Theorem (Prékopa–Leindler)**.** *Suppose that $f, g, h : \mathbb{R}^n \to \mathbb{R}_+$ are measurable, and suppose there exists $\lambda \in (0, 1)$ such that for all $x, y \in \mathbb{R}^n$,*

$$h((1-\lambda)x + \lambda y) \geq f(x)^{1-\lambda} g(y)^{\lambda}$$

Then

$$\int h(x)\,\mathrm{d}x \geq \left(\int f(x)\,\mathrm{d}x\right)^{1-\lambda} \left(\int g(x)\,\mathrm{d}x\right)^{\lambda}$$

Proof. We first suppose $n = 1$. Assume for now that f, g are bounded, say $\sup f, \sup g \leq 1$. If we prove the inequality in this setting, then by rescaling it will be true for any pair of bounded functions, and then by taking monotone limits over truncations, it will be true for arbitrary non-negative functions. We now have

$$\int f(x)\,\mathrm{d}x = \int_0^1 \mathrm{vol}(\{x : f(x) \geq t\})\,\mathrm{d}t$$

and similarly for g. Note that by the assumption, if $f(x) \geq t$ and $g(y) \geq t$, then also $h((1-\lambda)x + \lambda y) \geq t$. In other words,

$$\{x : h(x) \geq t\} \supseteq (1-\lambda)\{x : f(x) \geq t\} + \lambda\{x : g(x) \geq t\}$$

Therefore,

$$\begin{aligned}
\int h(x)\,\mathrm{d}x &= \int_0^\infty \mathrm{vol}(\{x : h(x) \geq t\})\,\mathrm{d}t \\
&\geq \int_0^1 \mathrm{vol}(\{x : h(x) \geq t\})\,\mathrm{d}t \\
&\geq \int_0^1 \mathrm{vol}((1-\lambda)\{x : f(x) \geq t\} + \lambda\{x : g(x) \geq t\})\,\mathrm{d}t \\
&\geq \int_0^1 (1-\lambda)\,\mathrm{vol}(\{x : f(x) \geq t\})\,\mathrm{d}t + \int_0^1 \lambda\,\mathrm{vol}(\{x : g(x) \geq t\})\,\mathrm{d}t \\
&\geq (1-\lambda)\int f(x)\,\mathrm{d}x + \lambda \int g(x)\,\mathrm{d}x \\
&\geq \left(\int f(x)\,\mathrm{d}x\right)^{1-\lambda} \left(\int g(x)\,\mathrm{d}x\right)^{\lambda}
\end{aligned}$$

where we use the one-dimensional Brunn–Minkowski inequality, proved above, and in the final step use the AM–GM inequality.

For $n > 1$, we use induction. Let $x, y \in \mathbb{R}^{n-1}$ and $a, b \in \mathbb{R}$. Our assumption is that

$$h((1-\lambda)x + \lambda y, (1-\lambda)a + \lambda b) \geq f(x,a)^{1-\lambda} g(y,b)^{\lambda}$$

If we fix a, b, this becomes an inequality about functions $\mathbb{R}^{n-1} \to \mathbb{R}$. By the indutive hypothesis, this implies that

$$\int_{\mathbb{R}^{n-1}} h(z, (1-\lambda)a + \lambda b)\,\mathrm{d}z \geq \left(\int_{\mathbb{R}^{n-1}} f(x,a)\,\mathrm{d}x\right)^{1-\lambda} \left(\int_{\mathbb{R}^{n-1}} g(y,b)\,\mathrm{d}y\right)^{\lambda}$$

So if we define

$$F(a) = \int_{\mathbb{R}^{n-1}} f(x,a)\,\mathrm{d}x \qquad G(b) = \int_{\mathbb{R}^{n-1}} g(y,b)\,\mathrm{d}y \qquad H(c) = \int_{\mathbb{R}^{n-1}} h(z,c)\,\mathrm{d}z$$

then this just says that

$$H((1-\lambda)a + \lambda b) \geq F(a)^{1-\lambda} G(b)^{\lambda}$$

So the one-dimensional case implies that

$$\int H(c)\,\mathrm{d}c \geq \left(\int F(a)\,\mathrm{d}a\right)^{1-\lambda} \left(\int G(b)\,\mathrm{d}b\right)^{\lambda}$$

By Fubini's theorem, this is exactly the conclusion we wanted. □

Corollary. *For compact $A, B \subseteq \mathbb{R}^n$,*

$$\mathrm{vol}((1-\lambda)A + \lambda B) \geq \mathrm{vol}(A)^{1-\lambda}\,\mathrm{vol}(B)^{\lambda}$$

Proof. Take f, g, h to be the indicator functions of A, B, and $(1-\lambda)A + \lambda B$, respectively. Then this follows immediately from Prékopa–Leindler. □

Corollary. *The Brunn–Minkowski inequality is true.*

Proof. Take compact $A, B \subseteq \mathbb{R}^n$, and define

$$A' = \frac{1}{\mathrm{vol}(A)^{1/n}} A \qquad\qquad B' = \frac{1}{\mathrm{vol}(B)^{1/n}} B$$

so that $\mathrm{vol}(A') = \mathrm{vol}(B') = 1$. By the previous corollary, for any $\lambda \in (0,1)$,

$$\mathrm{vol}((1-\lambda)A' + \lambda B') \geq \mathrm{vol}(A')^{1-\lambda}\,\mathrm{vol}(B')^{\lambda} = 1$$

We now choose

$$\lambda = \frac{\mathrm{vol}(B)^{1/n}}{\mathrm{vol}(A)^{1/n} + \mathrm{vol}(B)^{1/n}}$$

so that the above implies

$$\begin{aligned} 1 &\leq \frac{1}{(\mathrm{vol}(A)^{1/n} + \mathrm{vol}(B)^{1/n})^n}\,\mathrm{vol}\left(\mathrm{vol}(A)^{1/n} A' + \mathrm{vol}(B)^{1/n} B'\right) \\ &= \frac{\mathrm{vol}(A+B)}{(\mathrm{vol}(A)^{1/n} + \mathrm{vol}(B)^{1/n})^n} \end{aligned}$$

which is precisely Brunn–Minkowski. □

17.2 Vertex isoperimetric inequality on the hypercube

Theorem (Harper)**.** *Let $A \subseteq \{0,1\}^n$, and define*

$$\partial_V(A) = \{v \in \{0,1\}^n \setminus A : \exists u \in A, u \sim v\}$$

Then

$$|\partial_V(A)| \geq \left|\partial_V(S_{|A|})\right|$$

where S_k consists of the first k vertices in the "simplicial ordering" of $\{0,1\}^n$, defined by $x \preceq y$ if either $\sum x_i < \sum y_i$ or if $\sum x_i = \sum y_i$ and x follows y lexicographically (i.e. for the smallest i with $x_i \neq y_i$, we have $x_i = 1$ and $y_i = 0$). In other words, S_k is a k-element approximation of the Hamming ball.

18 November 9, 2018

18.1 Talagrand's inequality

Let $\mathcal{X}^n$ be a product space, and let d_H be the Hamming metric on $\mathcal{X}^n$. In the case of $\mathcal{X} = \{0,1\}$, we saw that the vertex isoperimetric inequality corresponds to concentration of Lipschitz functions. Indeed, given $f : \{0,1\}^n \to \mathbb{R}$ that is 1-Lipschitz, let m be its median value and

$$A = \{x \in \{0,1\}^n : f(x) \leq m\}$$

so that

$$|A| \geq \frac{1}{2} \cdot 2^n = \sum_{k=0}^{\lfloor n/2 \rfloor} \binom{n}{k}$$

Then the vertex isoperimetric inequality implies that

$$|A_t| \geq \sum_{k=0}^{\lfloor n/2 \rfloor + t - 1} \binom{n}{k} \geq 1 - e^{-2t^2/n}$$

which implies the concentration bound

$$\Pr(f(X) \geq m + t) \leq e^{-2t^2/n}$$

One way to generalize this is to consider weighted Hamming distance. We fix some $\alpha \in \mathbb{R}^n_+$, and define, for $x, y \in \mathcal{X}^n$,

$$d_\alpha(x,y) = \sum_{i : x_i \neq y_i} \alpha_i$$

Then for any set A, we get that the function $f(x) = d_\alpha(x, A)$ will be α_i-Lipschitz in the ith coordinate. Therefore, by our concentration bounds for Lipschitz functions,

$$\Pr(F(X) \geq \mathbb{E}[f(X)] + t) \leq e^{-2t^2/\|\alpha\|^2} \qquad \Pr(F(X) \leq \mathbb{E}[f(X)] - t) \leq e^{-2t^2/\|\alpha\|^2}$$

and we can derive these bounds either from Azuma's inequality or from the entropy method. As we saw before, this can be used to get isoperimetric bounds for d_α. First, from the lower tail bound, we see that

$$\mathbb{E}[d_\alpha(X,A)] \leq \sqrt{\frac{1}{2}\|\alpha\|^2 \log \frac{1}{P(A)}}$$

We renormalize so that $\|\alpha\| = 1$. We set $u = \sqrt{\frac{1}{2}\log\frac{1}{P(A)}}$, and the upper tail bound implies

$$\Pr(d_\alpha(X,A) \geq t) \leq \Pr(d_\alpha(X,A) \geq \mathbb{E}[d_\alpha(X,A) + (t-u)]) \leq e^{-2(t-u)^2}$$

If $t \geq 2u$, then this bound is at most $e^{-t^2/2}$. On the other hand, if

$$t \leq 2u = \sqrt{2\log\frac{1}{P(A)}}$$

then we find that $P(A) \leq e^{-t^2/2}$. Combining the two cases, we find that

$$P(A) \cdot \Pr(d_\alpha(X,A) \geq t) \leq e^{-t^2/2}$$

Since this holds for any α, we also find that

$$\sup_{\substack{\alpha \geq 0 \\ \|\alpha\|=1}} [P(A) \cdot \Pr(d_\alpha(X,A) \geq t)] \leq e^{-t^2/2}$$

Talagrand's inequality is the following strengthening of this result.

Theorem (Talagrand)**.**

$$P(A) \cdot \Pr\left(\sup_{\substack{\alpha \geq 0 \\ \|\alpha\|=1}} d_\alpha(X,A) \geq t \right) \leq e^{-t^2/4}$$

Definition. The quantity above is called *Talagrand's convex distance*. Formally, for $x \in \mathcal{X}^n$ and $A \subseteq \mathcal{X}^n$,

$$d_T(x,A) = \sup_{\substack{\alpha \geq 0 \\ \|\alpha\|=1}} d_\alpha(x,A) = \sup_{\substack{\alpha \geq 0 \\ \|\alpha\|=1}} \inf_{y \in A} \sum_{i : x_i \neq y_i} \alpha_i$$

To gain intuition for this distance, consider the hypercube. In this case, we can equivalently write d_T as

$$d_T(x,A) = \sup_{\substack{\alpha \geq 0 \\ \|\alpha\|=1}} \inf_{y \in A} \sum_{i=1}^n \alpha_i |y_i - x_i| = \sup_{\|\alpha\|=1} \inf_{y \in A} \vec{\alpha} \cdot (\vec{y} - \vec{x})$$

where we can drop the absolute value by allowing α to take signs that cancel the signs of $y_i - x_i$. By taking H to be the halfspace defined by α, this is the same as

$$\begin{aligned} d_T(x, A) &= \sup\{d(x, H) : H \text{ is a halfspace containing } A\} \\ &= \sup\{d(x, H) : H \text{ is a halfspace containing } \operatorname{conv}(A)\} \\ &= d(x, \operatorname{conv}(A)) \end{aligned}$$

where we use the convex set separation theorem to obtain the final equality.

Before we prove Talagrand's inequality, let's look at some applications. The classical application is to configuration functions, where we have that $f(X_1, \ldots, X_n)$ is the maximal size of some allowed configuration appearing in $(X_1, \ldots, X_n)$. Let M be the median of f and pick $t > 0$. We define

$$A = \{x : f(x) \leq M - t\}$$

The for any x with $f(x) \geq M$, we claim that $d_T(x, A)$ is large. The key property that implies this is that if $f(x)$ is large, then there is some large set of coordinates certifying this fact. Specifically, (assuming that M is an integer), if $f(x) \geq M$, then there exist M coordinates so that unless we modify one of them, $f(\widetilde{x}) \geq M$ as well, where $\widetilde{x}$ is gotten from x by changing some coordinates. Moreover, assuming that the set of allowed patterns is downards closed (i.e. a subpattern is still a pattern), then changing each of the witness coordinates can decrease f by at most 1. So we choose $\alpha_i = 1/\sqrt{M}$ for these witness coordinates, and $\alpha_i = 0$ otherwise. Then

$$d_T(x, A) \geq \frac{t}{\sqrt{M}}$$

because we need to change at least t of the witness coordinates to bring f from $\geq M$ to $\leq M - t$. Talagrand's inequality says that

$$P(A) \Pr_X \left(d_T(X, A) \geq \frac{t}{\sqrt{M}} \right) \leq e^{-t^2/4M}$$

However,

$$\frac{1}{2} = \Pr(f(X) \geq M) \leq \Pr\left(d_T(X, A) \geq \frac{t}{\sqrt{M}} \right)$$

which implies an exponential bound on $P(A)$, which is a lower tail bound on f.

For the upper tail, we set $A = \{x : f(x) \leq M\}$. For any x with $f(x) \geq M + t$, we pick $\alpha_i = 1/\sqrt{M + t}$ for the witness coordinates, and 0 for the non-witness coordinates. Then we get that

$$d_T(x, A) \geq \frac{t}{\sqrt{M + t}}$$

and we get that

$$P(A) \cdot \Pr\left(d_T(x, A) \geq \frac{t}{\sqrt{M + t}} \right) \leq e^{-t^2/4(M+t)}$$

In this case, $P(A) = \frac{1}{2}$, while

$$\Pr(f(X) \geq M + t) \leq \Pr\left(d_T(x, A) \geq \frac{t}{\sqrt{M+t}}\right)$$

so we again obtain an exponential tail bound, though the exponent is slightly worse. In fact, this is the correct behavior; the upper tail is a bit weaker.

19 November 12, 2018

19.1 Talagrand's inequality

Recall the statement of Talagrand's inequality:

Theorem. *For any $A \subseteq \mathcal{X}^n$,*

$$P(A)\Pr(d_T(X, A) \geq t) \leq e^{-t^2/4}$$

where

$$d_T(x, A) = \sup_{\substack{\alpha \geq 0 \\ \|\alpha\|=1}} d_\alpha(x, A)$$

Today, we will prove this by using the results we've already developed for tail bounds of self-bounding functions. We can equivalently write the Talagrand convex distance as

$$d_T(x, A) = \sup_{\substack{\alpha \geq 0 \\ \|\alpha\|=1}} \inf_{y \in A} \sum_{i : x_i \neq y_i} \alpha_i$$

We would like to switch the supremum and infimum, but in order to do this, we need to replace the final term as some convex object, so that we can apply a minimax result. Let $\mathcal{M}(A)$ denote the set of probability measures on A; then this can equivalently be written as

$$d_T(x, A) = \sup_{\substack{\alpha \geq 0 \\ \|\alpha\|=1}} \inf_{\nu \in \mathcal{M}(A)} \sum_i \alpha_i \Pr_{Y \sim \nu}(x_i \neq Y_i)$$

We will now use Sion's Minimax Theorem. It says that if X, Y are compact and convex sets, and $f : X \times Y \to \mathbb{R}$ satisfies that $f(x, y)$ is convex in x and concave in y, then

$$\sup_{y \in Y} \inf_{x \in X} f(x, y) = \inf_{x \in X} \sup_{y \in Y} f(x, y)$$

In our case, we apply this with $X = \mathcal{M}(A)$ and Y the set of $\alpha \geq 0$ with $\|\alpha\| \leq 1$; our function f is actually linear in both α and ν, so it is certainly convex and concave as necessary. Note that extending to $\|\alpha\| \leq 1$ from $\|\alpha\| = 1$ doesn't change anything, since the expression monotone in α, so the supremum will still

be attained on the boundary $\|\alpha\| = 1$. Therefore, the minimax theorem applies, and we can write

$$d_T(x, A) = \inf_{\nu \in \mathcal{M}(A)} \sup_{\substack{\alpha \geq 0 \\ \|\alpha\| \leq 1}} \sum_i \alpha_i \Pr_{Y \sim \nu}(x_i \neq Y_i)$$

Note that now the inner supremum is very simple, since it's just the dot product of α with some fied vector $(\Pr(x_i \neq Y_i))_i$. By Cauchy–Schwarz, this quantity is maximized when α points in the same direction as this vector, i.e. when α_i is proportional to $\Pr_{Y \sim \nu}(x_i \neq Y_i)$. Thus,

$$d_T(x, A) = \inf_{\nu \in \mathcal{M}(A)} \left(\sum_i \Pr_{Y \sim \nu}(x_i \neq Y_i)^2 \right)^{1/2}$$

Using this formulation, we will prove the following lemma.

Lemma. *Let*

$$f(x) = (d_T(x, A))^2$$

Then f is weakly $(4, 0)$-self-bounding. Moreover, if

$$f_i(x) = \inf_{X_i} f(x)$$

then

$$0 \leq f(x) - f_i(x) \leq 1$$

Proof. Assume that $\inf_{X_i} f(x)$ is attained at $x_i^{(i)} \in \mathcal{X}$ (if the infimum is not attained anywhere, let this point be arbitrarily close to achieving the infimum, and take limits at the end). Then

$$f_i(x) = f(x^{(i)}, x_i^{(i)})$$

Suppose too that the infimum in

$$f_i(x) = f(x^{(i)}, x_i^{(i)}) = \inf_{\nu \in \mathcal{M}(A)} \left(\sum_j \Pr_{Y \sim \nu}(x_j^{(i)} \neq Y_j)^2 \right)$$

is attained at $\widetilde{\nu_i} \in \mathcal{M}(A)$, so that

$$f_i(x) = \sum_j \Pr_{Y \sim \widetilde{\nu_i}}(x_j^{(i)} \neq Y_j)^2 = \sum_{j \neq i} \Pr_{Y \sim \widetilde{\nu_i}}(x_j \neq Y_j)^2 + \Pr_{Y \sim \widetilde{\nu_i}}(x_i^{(i)} \neq Y_i)^2$$

Note that

$$\begin{aligned} f(x) &= \inf_{\nu \in \mathcal{M}(A)} \sum_j \Pr_{Y \sim \nu}(x_j \neq Y_j)^2 \\ &\leq \sum_j \Pr_{Y \sim \widetilde{\nu_i}}(x_j \neq Y_j)^2 \\ &= \sum_{j \neq i} \Pr_{Y \sim \widetilde{\nu_i}}(x_j^{(i)} \neq Y_j)^2 + \Pr_{Y \sim \widetilde{\nu_i}}(x_i \neq Y_i)^2 \end{aligned}$$

Therefore,

$$f(x) - f_i(x) \leq \Pr_{Y\sim\widetilde{\nu_i}}(x_i \neq Y_i)^2 - \Pr_{Y\sim\widetilde{\nu_i}}(x_i^{(i)} \neq Y_i)^2 \leq 1$$

since the difference of two numbers in $[0,1]$ is at most 1.

Next, write

$$\sqrt{f(x)} = d_T(x, A) = \inf_{\nu} \sup_{\alpha} \sum_i \alpha_i \Pr_{Y\sim\nu}(x_i \neq Y_i)$$

and suppose that this optimum is attained at $(\widehat{\nu}, \widehat{\alpha})$. Similarly, we can write

$$\begin{aligned}\sqrt{f_i(x)} &= \inf_{\nu} \sup_{\alpha} \left(\sum_{j\neq i} \alpha_j \Pr_{Y\sim\nu}(x_j \neq Y_j) + \alpha_i \Pr(Y \sim \nu)(x_i^{(i)} \neq Y_i) \right) \\ &\geq \inf_{\nu} \left(\sum_{j\neq i} \widehat{\alpha}_j \Pr_{Y\sim\nu}(x_j \neq Y_j) + \widehat{\alpha}_i \Pr_{Y\sim\nu}(x_i^{(i)} \neq Y_i) \right)\end{aligned}$$

Now, suppose this infimum is attained at $\widetilde{\nu_i}$. Then continuing, we have

$$\sqrt{f_i(x)} \geq \sum_{j\neq i} \widehat{\alpha}_j \Pr_{Y\sim\widehat{\nu_i}}(x_j \neq Y_j) + \widehat{\alpha}_i \Pr_{Y\sim\widehat{\nu_i}}(x_i^{(i)} \neq Y_i)$$

On the other hand,

$$\sqrt{f(x)} \leq \sum_j \widehat{\alpha}_j \Pr_{Y\sim\widehat{\nu_i}}(x_j \neq Y_j)$$

and thus

$$\sqrt{f(x)} - \sqrt{f_i(x)} \leq \widehat{\alpha}_i \left(\Pr_{Y\sim\widehat{\nu_i}}(x_i \neq Y_i) - \Pr_{Y\sim\widehat{\nu_i}}(x_i^{(i)} \neq Y_i) \right) \leq \widehat{\alpha}_i$$

Therefore,

$$\begin{aligned}\sum_{i=1}^n (f(x) - f_i(x))^2 &= \sum_{i=1}^n \left(\sqrt{f(x)} - \sqrt{f_i(x)} \right)^2 \left(\sqrt{f(x)} + \sqrt{f_i(x)} \right)^2 \\ &\leq 4f(x) \sum_{i=1}^n \widehat{\alpha}_i^2 \\ &= 4f(x)\end{aligned}$$

which is precisely the condition for being weakly $(4,0)$-self-bounding. □

Proof of Talagrand's inequality. Suppose $A \subseteq \mathcal{X}^n$, so we can write

$$A = \{x : d_T(x, A) = 0\}$$

Define $f(x) = d_T(x, A)^2$. By the lower tail bound for weakly self-bounding functions,

$$P(A) = \Pr_X(f(X) = 0) \leq e^{-\mathbb{E}[f(X)]/8}$$

Recall that in the proof of the upper tail bound for weakly self-bounding functions, we actually had a bound for the exponential moment, which implied

$$\log \mathbb{E}[e^{\lambda(f(X)-\mathbb{E}[f(X)])}] \leq \frac{2\lambda^2 \mathbb{E}[f(X)]}{1-2\lambda}$$

We plug in $\lambda = 1/10$, which implies

$$\mathbb{E}[e^{f(X)/10}] \leq e^{\mathbb{E}[f(X)]/8}$$

Therefore,

$$P(A)\,\mathbb{E}[e^{f(X)/10}] \leq 1$$

By Markov's inequality,

$$\Pr(d_T(X, A) \geq t) = \Pr(f(X) \geq t^2) \leq \frac{\mathbb{E}[e^{f(X)/10}]}{e^{t^2/10}}$$

and thus

$$p(A)\Pr(d_T(X, A) \geq t) \leq e^{-t^2/10}$$

which is Talagrand's inequality, except for the constant in the exponent. □

20 November 14, 2018

Recall that Talagrand's inequality says that

$$P(A)\Pr(d_T(X, A) \geq t) \leq e^{-t^2/4}$$

though we only proved an upper bound of $e^{-t^2/10}$. Today we will see some applications.

We discussed previously the concentration of convex Lipschitz functions. We have $X_1, \ldots, X_n \in [0, 1]$, and $Z = f(X)$ with f convex and 1-Lipschitz (with respect to the Euclidean metric); the key example was $f(x) = \|x\|_p$ for $p \geq 2$. We derived asymmetric upper and lower tail bounds; the upper tail was simple

$$\Pr(Z \geq \mathbb{E}[Z] + t) \leq e^{-t^2/2}$$

For the lower bound, we had to also assume that $0 \leq f^2(x) - f_i^2(x) \leq 1$, and then got a lower tail bound

$$\Pr(Z \leq \mathbb{E}[Z] - t) \leq e^{-t^2/8}$$

For these bounds, it actually sufficed to assume just that f was convex in each coordinate, which is weaker.

To apply Talagrand's inequality, we will need to assume the stronger condition that f is *quasi-convex*, which means that for every s, the set $\{x : f(x) \leq s\}$ is convex. Every convex function is quasi-convex, though the converse is not true.

Theorem. *If $X_1, \ldots, X_n$ are independent in $[0,1]$, and if $f : [0,1]^n \to \mathbb{R}$ is quasi-convex and 1-Lipschitz, then*

$$\Pr(Z \geq \mathrm{med}(Z) + t) \leq 2e^{-t^2/4} \qquad \Pr(Z \leq \mathrm{med}(Z) - t) \leq 2e^{-t^2/4}$$

where $Z = f(X)$ and $\mathrm{med}(Z)$ is the median of Z.

Remark. In all applications of Talagrand's inequality, we get concentration near the median, rather than near the mean. However, by Chebyshev's inequality, we can see that

$$|\mathrm{med}(Z) - \mathbb{E}[Z]| \leq \sqrt{2\,\mathrm{Var}(Z)}$$

for any random variable Z. Therefore, concentration around the mean and around the median are essentially equivalent.

Lemma. *Let $A \subseteq [0,1]^n$ be convex, and $x \in [0,1]^n$. Then*

$$D(x, A) = \inf_{y \in A} \|x - y\| \leq d_T(x, A)$$

Proof. Since A is convex, we have that

$$D(x, A) = \inf_{\nu \in \mathcal{M}(A)} \left\| x - \mathop{\mathbb{E}}_{Y \sim \nu} Y \right\|$$

and therefore

$$\begin{aligned} D(x,A)^2 &= \inf_{\nu \in \mathcal{M}(A)} \left\| x - \mathop{\mathbb{E}}_{Y \sim \nu} Y \right\|^2 \\ &= \inf_{\nu \in \mathcal{M}(A)} \left\| \mathop{\mathbb{E}}_{Y \sim \nu} [x - Y] \right\|^2 \\ &= \inf_{\nu \in \mathcal{M}(A)} \sum_{i=1}^n \mathop{\mathbb{E}}_{Y \sim \nu} [x_i - Y_i]^2 \\ &\leq \inf_{\nu \in \mathcal{M}(A)} \sum_{i=1}^n \Pr(x_i \neq Y_i)^2 \\ &= d_T(x,A)^2 \end{aligned}$$

□

Note that in the case where $x \in \{0,1\}^n$ and A is the convex hull of some points in $\{0,1\}^n$, then we actually get equality, which is the computation we did previously to understand Talagrand's convex distance in the case of the hypercube.

Proof of tail bounds. Let

$$A_s = \{x \in [0,1]^n : f(x) \leq s\}$$

Then A_s is convex, since f is quasi-convex. If $f(y) \geq s + t$, then since f is 1-Lipschitz, we know that $D(y, A_s) \geq t$, so by the lemma, $d_T(y, A_s) \geq t$. So picking $s = \text{med}(Z)$, Talagrand's inequality asserts that

$$P(A_s) \Pr(Z \geq s + t) \leq P(A_s) \Pr(d_T(X, A_s) \geq t) \leq e^{-t^2/4}$$

which implies the desired tail bound since $P(A_s) \geq \frac{1}{2}$. For the lower tail, the exact same argument applies, except that we pick $s = \text{med}(Z) - t$. □

20.1 Bin packing

Let $x_1, \ldots, x_n \in [0,1]$; we think of these as object of sizes in $[0,1]$. We wish to pack these into as few bins as possible, where each bin has size 1. More concretely, we define $f(x_1, \ldots, x_n)$ to be the minimal k so that $x_1, \ldots, x_n$ can be partitioned into k parts, where each part has sum ≤ 1. Observe that if we change x_i into x_i', then f can change by at most 1, since in the worst case we can put x_i' into a new bin. So f is 1-Lipschitz, which means that if $X_1, \ldots, X_n$ are chosen independently in $[0,1]$ and $Z = f(X)$, we get concentration of the form

$$\Pr(|Z - \mathbb{E}[Z]| \geq t) \leq 2e^{-2t^2/n}$$

This is concentration of order $\sqrt{n}$, which might be the truth; if each X_i is actually valued in $\{0,1\}$, then the number of bins is just the number of 1s, which does truly have variation of order $\sqrt{n}$.

However, if we assume that many of the X_i are small, then we expect better concentration. So we define

$$\Sigma = \sqrt{\sum_{i=1}^n \mathbb{E}[X_i^2]}$$

and want to prove concentration of order $O(\Sigma)$. Note that $Z \geq \sum_{i=1}^n X_i$; for an upper bound, note that we can always pack items so that all bins except one are at least half full (for otherwise, we could combine any two bins that are less than half full). Thus, we get an upper bound $Z \leq 2\sum_{i=1}^n X_i + 1$. So Z is of order $\sum X_i$, so for many distributions, $O(\Sigma)$ is small compared to $\mathbb{E}[Z]$.

Theorem.

$$\Pr(|Z - \text{med}(Z)| \geq t + 1) \leq 8e^{-\frac{t^2}{32\Sigma^2 + 16t}}$$

In particular, for $t = \Omega(\Sigma)$, this is an exponentially decaying bound.

Proof. Note that for any $x, y \in [0,1]^n$,

$$f(x) \leq f(y) + 2 \sum_{i : x_i \neq y_i} x_i + 1$$

Indeed, we first pack all the y_i that are equal to x_i as before, and for all the new x_i, we can pack them in at most $2\sum_{i:x_i \neq y_i} x_i + 1$ bins, using the same argument as above, by combining bins that are less than half full. Note that this bound is reminiscent of Talagrand's distance, since we have a sum over i with $x_i \neq y_i$; in order to actually apply Talagrand's inequality, we need to pick a vector α that captures this.

So we define

$$\alpha_i = \frac{x_i}{\sqrt{\sum_{i:x_i \neq y_i} x_i^2}}$$

for i where $x_i \neq y_i$, and 0 otherwise. Then for any s, if we define $A_s = \{y : f(y) \leq s\}$, we find that

$$f(x) \leq s + 2\|x\| d_T(x, A_s) + 1$$

Now, by splitting into cases depending on $\|X\|$, we see that

$$\begin{aligned}\Pr(f(X) \geq s+t+1) &\leq \Pr\left(f(X) \geq s+1+\frac{\|X\|}{\sqrt{2\Sigma^2+t}} \cdot t\right) + \\ &\quad + \Pr\left(\|X\| \geq \sqrt{2\Sigma^2+t}\right) \\ &\leq \Pr\left(d_T(x, A_s) \geq \frac{t}{\sqrt{2\Sigma^2+t}}\right) + \Pr\left(\sum_{i=1}^n X_i^2 \geq 2\Sigma^2 + t\right)\end{aligned}$$

Fix $s = \text{med}(Z)$. For the first quantity, Talagrand's inequality tells us that it is upper bounded by $2e^{-t^2/16(2\Sigma^2+t)}$. For the second quantity, we can simply apply a Chernoff bound to the variables X_i^2, since Σ^2 is the mean of $\sum X_i^2$. This gives an upper bound of $e^{-\frac{3}{8}(\Sigma^2+t)}$. One can check that the first term is always the dominant contribution, so one derives the desired bound. □

21 November 28, 2018

21.1 Transportation Method

Using the transportation method, we will again prove a concentration result for Lipschitz functions (which we've already proved in several other ways). The main tool in the transportation method is the following lemma.

Lemma (Transportation Lemma)**.** *Let Z be a random variable, P a probability distribution, and $I \subseteq \mathbb{R}$ an interval. Then the following are equivalent:*

- *For every $\lambda \in I$,*

$$\log \mathop{\mathbb{E}}_P \left[e^{\lambda(Z - \mathbb{E}[Z])}\right] \leq \Phi(\lambda)$$

  for some function Φ

- *For all $Q \ll P$,*
$$\mathop{\mathbb{E}}_{Q}[Z] - \mathop{\mathbb{E}}_{P}[Z] \leq \Phi^{*-1}(D(Q\|P))$$
where Φ^* is the Cramer dual of Φ.

The most important example is when $\Phi(\lambda) = \frac{1}{2}\lambda^2 v$, in which case $\Phi^{*-1}(x) = \sqrt{2vx}$.

Assume that $f : \mathcal{X}^n \to \mathbb{R}$ is c_i-Lipschitz in the ith coordinate, namely

$$f(y) - f(x) \leq \sum_i c_i \mathbf{1}_{x_i \neq y_i}$$

In order to prove concentration, we will prove the second equivalent condition, where $Z = f(X_1, \ldots, X_n)$ is a function of independent variables. We write

$$\begin{aligned}
\mathop{\mathbb{E}}_{Q}[Z] - \mathop{\mathbb{E}}_{P}[Z] &= \mathop{\mathbb{E}}_{Y\sim Q}[f(Y)] - \mathop{\mathbb{E}}_{X\sim P}[f(X)] \\
&\leq \sum_{i=1}^{n} c_i \Pr_{X\sim P, Y\sim Q}[X_i \neq Y_i] \\
&\leq \left(\sum_i c_i^2\right)^{1/2} \left(\sum_i \Pr(X_i \neq Y_i)^2\right)^{1/2}
\end{aligned}$$

To bound the last term, we will use the following tool.

Lemma (Marton's Transportation Inequality)**.** *For $P = \bigotimes_{i=1}^n P_i$ a product measure on $\mathcal{X}^n$, and $Q \ll P$, there exists a correlated pair of random variables (X, Y) so that $X \sim P$, $Y \sim Q$, and*

$$\sum_{i=1}^{n} \Pr(X_i \neq Y_i)^2 \leq \frac{1}{2} D(Q\|P)$$

Note that plugging this in to the above computation, we get that

$$\mathop{\mathbb{E}}_{Q}[Z] - \mathop{\mathbb{E}}_{P}[Z] \leq \sqrt{\frac{1}{2}\sum c_i^2 D(Q\|P)}$$

which gives the sub-Gaussian bound with $v = \frac{1}{4}\sum c_i^2$, and thus our standard concentration result.

To prove Marton's inequality, we will need the following standard definition.

Definition. The *total variation distance* between measures P, Q is defined by

$$V(P, Q) = \sup_A |P(A) - Q(A)|$$

Equivalently,

$$V(P, Q) = \inf_{X\sim P, Y\sim Q} \Pr(X \neq Y)$$

The one-dimensional case of Marton's inequality says that for some pair (X_1, Y_1), we have that

$$\Pr(X_1 \neq Y_1)^2 \leq \frac{1}{2} D(Q\|P)$$

which is equivalent to saying that

$$V(P,Q) \leq \sqrt{\frac{1}{2} D(Q\|P)}$$

This is a result known as Pinsker's inequality.

Proof of Pinsker's inequality. Recall that $V(P,Q) = \sup_A |P(A) - Q(A)|$. For any event $A \subseteq \Omega$, we can write

$$Q(A) - P(A) = \mathop{\mathbb{E}}_{Q}[\mathbf{1}_A] - \mathop{\mathbb{E}}_{P}[\mathbf{1}_A]$$

We will now apply the transportation lemma. Note that $Z = \mathbf{1}_A$ is a 0/1 random variable, which we already showed implies that

$$\log \mathbb{E}[e^{\lambda(Z - \mathbb{E}[Z])}] \leq \frac{1}{8}\lambda^2$$

by the Hoeffding bound. Therefore, setting $v = 1/4$ in the transportation lemma tells us that

$$\mathop{\mathbb{E}}_{Q}[Z] - \mathop{\mathbb{E}}_{P}[Z] \leq \sqrt{\frac{1}{2} D(Q\|P)}$$

Taking the supremum over all A gives the desired bound. □

We will now apply induction to obtain the full Marton bound from this base case.

Lemma (Induction lemma for transportation)**.** *Let $P = \bigotimes_{i=1}^n P_i, Q \ll P, d : \mathcal{X} \times X \to \mathbb{R}_+$, and $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ convex. Assume that for each $i \in [n]$, and for each $\nu \ll P_i$, there is a correlated pair (X_i, Y_i) with $X_i \sim P_i$ and $Y_i \sim \nu$ so that*

$$\Phi\left(\mathbb{E}[d(X_i, Y_i)]\right) \leq D(\nu\|P_i)$$

Then there exists a pair (X, Y) with $X \sim P, Y \sim Q$, and

$$\sum_{i=1}^n \Phi(\mathbb{E}[d(X_i, Y_i)]) \leq D(Q\|P)$$

Proof. We induct on n. The base case is trivial, since the conclusion is just the assumption for $n = 1$. Let P^{k-1} denote the restriction of P onto the first $k-1$ coordinates. Assume inductively that for any $Q' \ll P^{k-1}$, there exists a pair (X^{k-1}, Y^{k-1}) such that $X^{k-1} \sim P^{k-1}$ and $Y^{k-1} \sim Q'$ with

$$\sum_{i=1}^{k-1} \Phi(\mathbb{E}[d(X_i, Y_i)]) \leq D(Q'\|P)$$

Now let $Q \ll P^k$. So we can write $\mathrm{d}Q = g(x,t)\mathrm{d}P^k$ for some density function g, and $x \in \mathcal{X}^{k-1}, t \in \mathcal{X}$. Then we can write

$$\begin{aligned}
D(Q\|P^k) &= \int_{\mathcal{X}^k} \log g(x,t)\mathrm{d}Q(x,t) \\
&= \int_{\mathcal{X}} \int_{\mathcal{X}^{k-1}} g(x,t)\log g(x,t)\mathrm{d}P^k(x,t) \\
&= \int_{\mathcal{X}} \int_{\mathcal{X}^{k-1}} g(x \mid t)g_k(t)(\log g(x \mid t) + \log g_k(t))\mathrm{d}P^{k-1}(x)\mathrm{d}P(t) \\
&= \int_{\mathcal{X}} g_k(t)\log g_k(t) \int_{\mathcal{X}^{k-1}} g(x \mid t)\mathrm{d}P^{k-1}(x)\mathrm{d}P(t) + \\
&\qquad + \int_{\mathcal{X}} \int_{\mathcal{X}^{k-1}} g(x \mid t)g_k(t)\log g(x \mid t)\mathrm{d}P^{k-1}(x)\mathrm{d}P(t) \\
&= D(q_k\|P) + \int_{\mathcal{X}} \int_{\mathcal{X}^{k-1}} g(x \mid t)g_k(t)\log g(x \mid t)\mathrm{d}P^{k-1}(x)\mathrm{d}P(t)
\end{aligned}$$

where we write

$$g(x,t) = g(x \mid t)g_k(t) \qquad\qquad g_k(t) = \int_{\mathcal{X}^{k-1}} g(x,t)\mathrm{d}P^{k-1}(x)$$

and

$$\mathrm{d}q_k(t) = g_k(t)\mathrm{d}P(t)$$

To deal with the last term above, define, for every t, a new measure by

$$\mathrm{d}Q(x \mid t) = g(x \mid t)\mathrm{d}P^{k-1}(x)$$

This is the conditional measure obtained from Q by fixing the last coordinate to be equal to t. Then the last term above becomes

$$\int_{\mathcal{X}} g_k(t) \int_{\mathcal{X}^{k-1}} g(x \mid t)\log g(x \mid t)\mathrm{d}P^{k-1}(x)\mathrm{d}P(t) = \int_{\mathcal{X}} g_k(t)D(Q(\cdot|t)\|P^{k-1})\mathrm{d}P(t)$$

What we've obtained is the "chain rule" for the KL divergence, which says that

$$D(Q\|P^k) = D(q_k\|P) + \int_{\mathcal{X}} D(Q(\cdot \mid t)\|P^{k-1})\mathrm{d}q_k(t)$$

We now apply the inductive hypothesis to $Q' = Q(\cdot \mid t)$ for all $t \in \mathcal{X}$. Then we obtain, for every t, a pair $(X^{(t)}, Y^{(t)})$, with $X^{(t)} \sim P^{k-1}$ and $Y^{(t)} \sim Q(\cdot \mid t)$, satisfying

$$\sum_{i=1}^{k-1} \Phi(\mathbb{E}[d(X_i^{(t)}, Y_i^{(t)})]) \leq D(Q(\cdot \mid t)\|P^{k-1})$$

Moreover, the base case tells us that there is a pair (X_k, Y_k) such that

$$\Phi(\mathbb{E}[d(X_k, Y_k)]) \leq D(q_k\|P)$$

Then by the chain rule and Jensen's inequality, we can write

$$D(Q\|P) \geq \Phi(\mathbb{E}[d(X_k, Y_k)]) + \int_{\mathcal{X}} \sum_{i=1}^{k-1} \Phi(\mathbb{E}[d(X_i^{(t)}, Y_i^{(t)})]) \mathrm{d}q_k(t)$$
$$\geq \Phi(\mathbb{E}[d(X_k, Y_k)]) + \sum_{i=1}^{k-1} \Phi\left(\int_{\mathcal{X}} \mathbb{E}[d(X_i^{(t)}, Y_i^{(t)})] \mathrm{d}q_k(t) \right)$$

Thus, to get the desired pair (X, Y), we simply make them equal to (X_k, Y_k) in the final coordinate, and in the other coordinates, we simply condition on the outcome of Y_k; specifically, the distribution of $(X_i, Y_i) \mid Y_k = t$ is the pair $(X_i^{(t)}, Y_i^{(t)})$. Then we get the desired inequality, since the above right-hand side is exactly $\sum_i \Phi(\mathbb{E}[d(X_i, Y_i)])$. □

22 November 30, 2018

22.1 Conditional Transportation Inequalities

Let $Z = f(X_1, \ldots, X_n)$, and suppose that f satisfies the following Lipschitz-type condition:

$$f(y) - f(x) \leq \sum_{i=1}^{n} c_i(x) \mathbf{1}_{x_i \neq y_i}$$

Note that the new ingredient is that the Lipschitz constants c_i can depend on the point x. We would like to get a concentration bound that depends on some average of the c_i. Specifically, let's assume that

$$\sum_{i=1}^{n} \mathop{\mathbb{E}}_{X} \left[c_i^2(x) \right] \leq v$$

We would like to prove a tail bound of the form

$$\Pr(Z \geq \mathbb{E}[Z] + t) \leq e^{-t^2/2v}$$

To do this, we will again use the transportation method. Let $Q \ll P = \bigotimes_{i=1}^n P_i$, and consider

$$\begin{aligned}
\mathop{\mathbb{E}}_{Q}[Z] - \mathop{\mathbb{E}}_{P}[Z] &= \mathop{\mathbb{E}}_{X\sim P, Y\sim Q}[f(Y) - f(X)] \\
&\le \sum_{i=1}^n \mathbb{E}\left[c_i(X)\mathbf{1}_{X_i \neq Y_i}\right] \\
&= \sum_{i=1}^n \mathop{\mathbb{E}}_{X\sim P}\left[\mathop{\mathbb{E}}_{Y\sim Q}\left[c_i(X)\mathbf{1}_{X_i\neq Y_i} \mid X\right]\right] \\
&= \sum_{i=1}^n \mathop{\mathbb{E}}_{X\sim P}\left[c_i(X)\Pr_{Y\sim Q}(X_i \neq Y_i \mid X)\right] \\
&\le \left(\sum_{i=1}^n \mathop{\mathbb{E}}_{X\sim P}[c_i^2(X)]\right)^{1/2}\left(\sum_{i=1}^n \mathop{\mathbb{E}}_{X\sim P}\Pr_{Y\sim Q}(X_i \neq Y_i \mid X)^2\right)^{1/2}
\end{aligned}$$

by Cauchy–Schwarz. We can bound the first term by assumption, and for the second, we want to argue as last time. Namely, given $Q \ll P$, we want to prove the existence of a correlated pair (X, Y) with $X \sim P$ and $Y \sim Q$ so that

$$\mathop{\mathbb{E}}_{X\sim P}\left[\Pr_{Y\sim Q}(X_i \neq Y_i \mid X)^2\right] \le 2D(Q\|P)$$

Last time, we could do this by exploiting the connection between total variation distance and KL divergence, but this time we need to come up with an analogue of total variation distance that can deal with the conditioning.

Definition. For distributions P, Q on $\mathcal{X}$ that have densities p, q with respect to μ, define

$$d_2^2(Q, P) = \int_{\mathcal{X}} \frac{1}{p(x)}(p(x) - q(x))_+^2 \, \mathrm{d}\mu(x)$$

To see that this is well-defined (i.e. doesn't depend on μ), let ν be a different measure with $\mathrm{d}\mu = g\mathrm{d}\nu$. Then P, Q have densities gp and gq with respect to ν, and thus

$$\begin{aligned}
d_2^2(Q, P) &= \int_{\mathcal{X}} \frac{1}{g(x)p(x)}(g(x)p(x) - g(x)q(x))_+^2 \, \mathrm{d}\nu(x) \\
&= \int_{\mathcal{X}} \frac{1}{p(x)g(x)^2}(g(x)^2(p(x) - q(x))_+^2) \, \mathrm{d}\mu(x) \\
&= \int_{\mathcal{X}} \frac{1}{p(x)}(p(x) - q(x))_+^2 \, \mathrm{d}\mu(x)
\end{aligned}$$

Lemma.

$$d_2^2(Q, P) + d_2^2(P, Q) = \min_{\substack{(X,Y) \\ X\sim P, Y\sim Q}} \left\{\mathop{\mathbb{E}}_{X}\left[\Pr_Y(Y \neq X \mid X)\right] + \mathop{\mathbb{E}}_{Y}\left[\Pr_X(X \neq Y \mid Y)\right]\right\}$$

Proof. Let P, Q have densities p, q with respect to μ. We first claim that if $p(x) > 0$, then

$$\Pr(X = Y \mid X = x) \leq \frac{q(x)}{p(x)}$$

To see this, fix some measurable function $h(x) \geq 0$. Then

$$\begin{aligned}
\mathop{\mathbb{E}}_{X\sim P}[h(x)\Pr(X = Y \mid X)] &= \mathop{\mathbb{E}}_{X\sim P, Y\sim Q}[h(X)\mathbf{1}_{X=Y}] \\
&= \mathop{\mathbb{E}}_{X\sim P, Y\sim Q}[h(Y)\mathbf{1}_{X=Y}] \\
&\leq \mathop{\mathbb{E}}_{X\sim P, Y\sim Q}[h(Y)\mathbf{1}_{p(Y)>0}] \\
&= \mathop{\mathbb{E}}_{X\sim P}\left[h(X)\frac{q(X)}{p(X)}\right]
\end{aligned}$$

Since this holds for every measurable $h \geq 0$, we must have that

$$\frac{q(X)}{p(X)} - \Pr(X = Y \mid X) \geq 0$$

with probability 1 under X. Thus, it must hold for all x with $p(x) > 0$.

We therefore have that

$$\begin{aligned}
\mathop{\mathbb{E}}_{X\sim P}\left[\Pr_{Y\sim Q}(X \neq Y \mid X)^2\right] &\geq \mathop{\mathbb{E}}_{X\sim P}\left[\left(1 - \frac{q(X)}{p(X)}\right)_+^2\right] \\
&= \int_{\mathcal{X}} \left(1 - \frac{q(x)}{p(x)}\right)_+^2 p(x)\,\mathrm{d}\mu(x) \\
&= d_2^2(Q, P)
\end{aligned}$$

and the analogous inequality where we reverse the roles of P and Q. Therefore, we get that

$$\begin{aligned}
\inf_{(X,Y)\sim(P,Q)} \left(\mathop{\mathbb{E}}_{X\sim P}\left[\Pr_{Y\sim Q}(X \neq Y \mid X)^2\right] + \mathop{\mathbb{E}}_{Y\sim Q}\left[\Pr_{X\sim P}(X \neq Y \mid Y)^2\right]\right) \\
\geq d_2^2(Q, P) + d_2^2(P, Q)
\end{aligned}$$

To prove that the minimum is attained at this bound, we use the same coupling as before, where X and Y are correlated to be equal as often as possible; specifically, we first set

$$\Pr(X = Y = x) = \min\{p(x), q(x)\}$$

and have $X \neq Y$ on the remaining density to make the marginals equal P and Q. Under this coupling, observe that

$$\Pr(X \neq Y \mid X = x) = \begin{cases} 0 & q(x) > p(x) \\ \frac{p(x)-q(x)}{p(x)} & q(x) \leq p(x) \end{cases} = \frac{(p(x) - q(x))_+}{p(x)}$$

Therefore, this coupling gives an equality in the above inequality, which proves the lemma. □

23 December 3, 2018

23.1 Conditional Transportation

Recall that we proved a bound of the form

$$\mathop{\mathbb{E}}_{Y\sim Q}[f(Y)] - \mathop{\mathbb{E}}_{X\sim P}[f(X)] \leq \sqrt{\sum_i \mathbb{E}[c_i^2(X)]} \left(\sum_i \mathop{\mathbb{E}}_{X} \left[\Pr_Y(X_i \neq Y_i \mid X)^2 \right] \right)^{1/2}$$

We called the first term $\sqrt{v}$, and we need to bound the second term. For this, we need the following result.

Theorem (Marton's Conditional Transportation Inequality)**.** *For any $Q \ll P = \bigotimes_{i=1}^n P_i$ on $\mathcal{X}^n$, there exists a pair (X, Y) with $X \sim P, Y \sim Q$, and*

$$\sum_{i=1}^n \mathbb{E}\left[\Pr(X_i \neq Y_i \mid X)^2 + \Pr(X_i \neq Y_i \mid Y)^2 \right] \leq 2D(Q\|P)$$

Remark. The book includes a slightly different inequality (namely conditioning on X_i and Y_i rather than on X and Y), but that statement appears to not be the correct one.

Lemma (Conditional Induction Lemma)**.** *Let $\Phi : \mathbb{R}_+ \to \mathbb{R}_+$ be convex, $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ be measurable, and $Q \ll P = \bigotimes_{i=1}^n P_i$ be measures on $\mathcal{X}^n$. Assume that for every $i \in [n]$ and for every $\nu \ll P_i$, there exists a pair (X_i, Y_i) with $X_i \sim P_i, Y_i \sim \nu$ such that*

$$\mathop{\mathbb{E}}_{X_i}\left[\Phi\left(\mathop{\mathbb{E}}_{Y_i}[d(X_i, Y_i) \mid X_i] \right) \right] + \mathop{\mathbb{E}}_{Y_i}\left[\Phi\left(\mathop{\mathbb{E}}_{X_i}[d(X_i, Y_i) \mid Y_i] \right) \right] \leq D(\nu\|P_i)$$

Then there exists a pair (X, Y) with $X \sim P, Y \sim Q$ such that

$$\sum_{i=1}^n \mathop{\mathbb{E}}_{X_i}\left[\Phi(\mathbb{E}[d(X_i, Y_i) \mid X_i]) \right] + \mathop{\mathbb{E}}_{Y_i}\left[\Phi(\mathbb{E}[d(X_i, Y_i) \mid Y_i]) \right] \leq D(Q\|P)$$

and

$$\sum_{i=1}^n \mathop{\mathbb{E}}_{X}\left[\Phi\left(\mathop{\mathbb{E}}_{Y}[d(X_i, Y_i) \mid X] \right) \right] + \mathop{\mathbb{E}}_{Y}\left[\Phi\left(\mathop{\mathbb{E}}_{X}[d(X_i, Y_i) \mid Y] \right) \right] \leq D(Q\|P)$$

Remark. This statement is not proved in the book. It asserts that the proof is similar to that of the previous inductive lemma, but the conditioning adds a fair amount of technical difficulty.

Proof. Recall the chain rule for the KL divergence, which says that

$$D(Q\|P) = \int_{\mathcal{X}} D(Q(\,\cdot\mid t)\|P^{k-1})\,\mathrm{d}q_k(t) + D(q_k\|P_k)$$

where $Q \ll P = \bigotimes_{i=1}^{k} P_i$ and q_k is the marginal of Q on the kth coordinate. By the inductive hypothesis, applied to $Q(\cdot \mid t)$, there exists a pair $(X^{(t)}, Y^{(t)})$ with $X^{(t)} \sim P^{k-1}$ and $Y^{(t)} \sim Q(\cdot \mid t)$, such that

$$\sum_{i=1}^{k-1} \mathop{\mathbb{E}}_{X^{(t)}} \left[\Phi \left(\mathop{\mathbb{E}}_{Y^{(t)}} \left[d\left(X_i^{(t)}, Y_i^{(t)} \right) \middle| X^{(t)} \right] \right) \right] + \\ + \mathop{\mathbb{E}}_{Y^{(t)}} \left[\Phi \left(\mathop{\mathbb{E}}_{X^{(t)}} \left[d\left(X_i^{(t)}, Y_i^{(t)} \right) \middle| Y^{(t)} \right] \right) \right] \leq D(Q(\cdot \mid t) \| P^{k-1})$$

Also, from the base case of the induction (which is just the assumption in the theorem), applied to $q_k \ll P$, there exists a pair $(X_k, Y_k) \sim (P, q_k)$ so that

$$\mathop{\mathbb{E}}_{X_k} \left[\Phi \left(\mathop{\mathbb{E}}_{Y_k} [d(X_k, Y_k) | X_k] \right) \right] + \mathop{\mathbb{E}}_{Y_k} \left[\Phi \left(\mathop{\mathbb{E}}_{X_k} [d(X_k, Y_k) | Y_k] \right) \right] \leq D(q_k \| P)$$

We plug this into the chain rule to obtain the following.

$$\begin{aligned} D(Q\|P) &= \int_{\mathcal{X}} D(Q(\cdot \mid t)\|P^{k-1}) \, \mathrm{d}q_k(t) + D(q_k \| P_k) \\ &\geq \sum_{i=1}^{k-1} \int_{\mathcal{X}} \left(\mathop{\mathbb{E}}_{X^{(t)}} \left[\Phi \left(\mathop{\mathbb{E}}_{Y^{(t)}} \left[d\left(X_i^{(t)}, Y_i^{(t)} \right) \middle| X^{(t)} \right] \right) \right] + \right. \\ &\qquad \left. + \mathop{\mathbb{E}}_{Y^{(t)}} \left[\Phi \left(\mathop{\mathbb{E}}_{X^{(t)}} \left[d\left(X_i^{(t)}, Y_i^{(t)} \right) \middle| Y^{(t)} \right] \right) \right] \right) \mathrm{d}q_k(t) + \\ &\qquad + \mathop{\mathbb{E}}_{X_k} \left[\Phi \left(\mathop{\mathbb{E}}_{Y_k} [d(X_k, Y_k) | X_k] \right) \right] + \mathop{\mathbb{E}}_{Y_k} \left[\Phi \left(\mathop{\mathbb{E}}_{X_k} [d(X_k, Y_k) | Y_k] \right) \right] \end{aligned}$$

We are almost done, except that we need to condition on the entire vectors X, Y, rather than on a single coordinate, as in the final two terms above. Note that this is non-trivial, since Jensen's inequality implies an inequality between the two terms, but in the reverse direction.

In order to make this work, we need to inductively build the distribution (X, Y) on $\mathcal{X}^k \times \mathcal{X}^k$ in a careful way. The key property is that, as we did last time, we define this distribution based on (X_k, Y_k) and $(X^{(t)}, Y^{(t)})$ for each t, where we do it based on t being the value of Y_k. Therefore, given Y_k, we know the conditional distribution on all of the first $k-1$ coordinates of X and Y. Crucially, this means that, conditional on Y_k, X_k is independent of the first $k-1$ coordinates of both vectors. We continue in this way to Y_{k-1} and recursively backwards down the indices, to obtain a chain of conditional independences between variables. So consider first the term

$$\int_{\mathcal{X}} \mathop{\mathbb{E}}_{X^{(t)}} \left[\Phi \left(\mathop{\mathbb{E}}_{Y^{(t)}} \left[d(X_i^{(t)}, Y_i^{(t)}) \middle| X^{(t)} \right] \right) \right] \mathrm{d}q_k(t)$$

Note that integrating over t is the same as taking a conditional expectation over Y_k, by the definition of the coupling above. But conditional on Y_k, the inner term is independent of X_k, so we may freely include a conditioning on X_k on

the inner term. Thus, we are able to get the full vector in the conditioning. To do this we need to add X_k to the outer expectation, but by Jensen's inequality, we can then bring it inside the evaluation of Φ to the inner expectation. We apply the same argument to every term, and obtain the desired result. $\square$

24 December 5, 2018

Recall that we were proving the conditional inductive lemma. We were trying to construct a correlated pair (X, Y) on $\mathcal{X}^k \times \mathcal{X}^k$ so that $X \sim P, Y \sim Q$. To do so, we had inductively constructed pairs so that $X_k \sim P_k, Y_k \sim q_k$, and for each $t \in \mathcal{X}$, we had pairs $(X_i^{(t)}, Y_i^{(t)})_{i=1}^{k-1}$ so that $X^{(t)} \sim \bigotimes_{i=1}^{k-1} P_i$ and $Y^{(t)} \sim Q(\cdot \mid t)$. We then defined (X, Y) by first sampling (X_k, Y_k) as above, and then conditioning on the value $Y_k = t$, sampling $(X_i, Y_i) = (X_i^{(t)}, Y_i^{(t)})$. Crucially, conditioned on Y_k, we have that X_k is independent of $(X_i, Y_i)_{i=1}^{k-1}$. The inductive hypothesis was that

$$\begin{aligned} D(Q\|P) \geq \sum_{i=1}^{k-1} \mathbb{E}\left[\Phi\left(\mathbb{E}\left[d(X_i, Y_i) \mid X_1, \ldots, X_{k-1}, Y_k\right]\right)\right] + \\ + \sum_{i=1}^{k-1} \mathbb{E}\left[\Phi\left(\mathbb{E}\left[d(X_i, Y_i) \mid Y_1, \ldots, Y_{k-1}, Y_k\right]\right)\right] + \\ + \mathbb{E}\left[\Phi\left(\mathbb{E}\left[d(X_k, Y_k) \mid X_k\right]\right)\right] + \mathbb{E}\left[\Phi\left(\mathbb{E}\left[d(X_k, Y_k) \mid Y_k\right]\right)\right] \end{aligned}$$

The goal is to have every term above be conditioned on the whole X or Y vector. At the end of last lecture, we saw how the first term can be lower-bounded by a similar term where we condition on X_k instead of Y_k, by first introducing X_k into the conditioning by the conditional independence above, and then shifting the expectation over Y_k inside Φ by Jensen's inequality. The second term is already fine, since it's conditional on the entire vector Y. For the third term, observe that conditional on (X_k, Y_k), we have that $(X_1, \ldots, X_{k-1}) \sim \bigotimes_{i=1}^{k-1} P_i$, and in particular the vector $(X_1, \ldots, X_{k-1})$ is independent of (X_k, Y_k). So we may simply introduce these additional variables into the conditioning, and write

$$\mathbb{E}\left[\Phi\left(\mathbb{E}\left[d(X_k, Y_k) \mid X_k\right]\right)\right] = \mathbb{E}\left[\Phi\left(\mathbb{E}\left[d(X_k, Y_k) \mid X_1, \ldots, X_{k-1}, X_k\right]\right)\right]$$

Finally, for the final term, we again use the conditional independence, since we are already conditioning on Y_k. So all in all, we get that

$$\begin{aligned} D(Q\|P) \geq \sum_{i=1}^{k} \mathbb{E}\left[\Phi\left(\mathbb{E}\left[d(X_i, Y_i) \mid X_1, \ldots, X_{k-1}, X_k\right]\right)\right] + \\ + \sum_{i=1}^{k} \mathbb{E}\left[\Phi\left(\mathbb{E}\left[d(X_i, Y_i) \mid Y_1, \ldots, Y_{k-1}, Y_k\right]\right)\right] \end{aligned}$$

This finishes proving the induction. Recall that the goal was to prove the following theorem.

Theorem (Marton's conditional transportation inequality)**.** *For qny $Q \ll P = \bigotimes_{i=1}^n P_i$ on $\mathcal{X}^n$, there exists a pair (X,Y) with $X \sim P, Y \sim Q$, and*

$$\sum_{i=1}^n \left(\mathbb{E}\left[\Pr(X_i \neq Y_i \mid X)^2\right] + \mathbb{E}\left[\Pr(X_i \neq Y_i \mid Y)^2\right]\right) \leq 2D(Q\|P)$$

Proof. We prove this by induction on n. For the one-dimensional case, recall that we proved earlier that

$$d_2^2(P,Q)+d_2^2(Q,P) = \min_{\substack{(X,Y) \\ X\sim P, Y\sim Q}} \left\{ \mathop{\mathbb{E}}_X \left[\Pr_Y(X \neq Y \mid X)^2\right] + \mathop{\mathbb{E}}_Y \left[\Pr_X(X \neq Y \mid Y)^2\right] \right\}$$

To apply this, we will need a lemma analogous to Pinsker's inequality.

Lemma. *For any $Q \ll P$,*

$$d_2^2(P,Q) + d_2^2(Q,P) \leq 2D(Q\|P)$$

Proof. Let $\mathrm{d}Q = q\mathrm{d}P$. Then by the definition of d_2, with $\mu = P$, we get that

$$d_2^2(Q,P) = \int_{\mathcal{X}} (1-q(x))_+^2 \,\mathrm{d}P(x) \qquad d_2^2(P,Q) = \int_{\mathcal{X}} \frac{1}{q(x)}(q(x)-1)_+^2 \,\mathrm{d}P(x)$$

On the other hand,

$$D(Q\|P) = \int_{\mathcal{X}} q(x)\log q(x)\,\mathrm{d}P(x) = \int_{\mathcal{X}} h(q(x)-1)\mathrm{d}P(x)$$

where $h(t) = (1+t)\log(1+t) - t$. Note that $h(t) = \frac{1}{2}t^2 + O(t^3)$. Now we split the integration above into two cases, depending on the sign of $q(x)-1$; this gives

$$D(Q\|P) = \int_{\mathcal{X}} h\left((q(x)-1)_+\right)\mathrm{d}P(x) + \int_{\mathcal{X}} h(-(1-q(x))_+)\mathrm{d}P(x)$$

There are analytic bounds on h that one can get by the Taylor expansion. Specifically we have that for $t \in [-1,0]$, $h(t) \geq \frac{1}{2}t^2$, and for $t \geq 0$, $h(t) \geq \frac{1}{2}\frac{t^2}{t+1}$. Note that these bounds are very similar (and closely related) to the bounds in the exponent in the Chernoff bound. These bounds imply that

$$\begin{aligned} D(Q\|P) &\geq \frac{1}{2}\int_{\mathcal{X}} \frac{(q(x)-1)_+^2}{q(x)}\mathrm{d}P(x) + \frac{1}{2}\int_{\mathcal{X}} (1-q(x))_+^2 \mathrm{d}P(x) \\ &= \frac{1}{2}d_2^2(P,Q) + \frac{1}{2}d_2^2(Q,P) \end{aligned}$$

□

This proves the base case. For the induction, we apply the inductive lemma with $\Phi(t) = \frac{1}{2}t^2$ and $d(x,y) = \mathbf{1}_{x\neq y}$. □

Using this, we are able to prove the following concentration bound, which we stated in a previous lecture.

Theorem. *Let $X_1, \ldots, X_n$ be independent, and $Z = f(X_1, \ldots, X_n)$. Suppose that for any $x, y \in \mathcal{X}^n$,*

$$f(y) - f(x) \leq \sum_{i=1}^{n} c_i(x) \mathbf{1}_{x_i \neq y_i}$$

Then

$$\Pr(Z \geq \mathbb{E}[Z] + t) \leq e^{-t^2/2v} \qquad \Pr(Z \leq \mathbb{E}[Z] - t) \leq e^{-t^2/2v_\infty}$$

where

$$v = \sum_{i=1}^{n} \mathop{\mathbb{E}}_{X}[c_i^2(X)] \qquad v_\infty = \sup_x \sum_{i=1}^{n} c_i^2(x)$$

Proof. We saw, using Cauchy–Schwarz, that

$$\begin{aligned} \mathop{\mathbb{E}}_{Y \sim Q}[f(Y)] - \mathop{\mathbb{E}}_{X \sim P}[f(X)] &\leq \sqrt{v} \left(\sum_{i=1}^{n} \mathbb{E}[\Pr(X_i \neq Y_i \mid X)^2] \right)^{1/2} \\ &\leq \sqrt{v} \cdot \sqrt{2D(Q\|P)} \end{aligned}$$

where the final bound follows from Marton's inequality. By the transportation lemma, this implies that

$$\Pr(Z \geq \mathbb{E}[Z] + t) \leq e^{-t^2/2v}$$

for any $t > 0$.

For the lower tail, we need to consider $g(x) = -f(x)$. Then the condition we get on g is that

$$g(y) - g(x) \leq \sum_{i=1}^{n} c_i(y) \mathbf{1}_{x_i \neq y_i}$$

which implies, by the same argument, that

$$\mathop{\mathbb{E}}_{Q}[g(Y)] - \mathop{\mathbb{E}}_{P}[g(X)] \leq \sqrt{\mathop{\mathbb{E}}_{Q}\left[\sum_{i=1}^{n} c_i^2(Y)\right]} \left(\sum_{i=1}^{n} \mathbb{E}[\Pr(X_i \neq Y_i \mid Y)^2] \right)^{1/2}$$

The second term is again bounded by $\sqrt{2D(Q\|P)}$, but the first term is difficult to bound because Q is an arbitrary distribution. So the best we can do is bound it by $\sqrt{v_\infty}$, which gives the weaker lower tail. □