Concentration of Measure with Applications in Information Theory, Communications, and Coding

Maxim Raginsky (UIUC) and **Igal Sason** (Technion)

A Tutorial, Presented at 2015 IEEE International Symposium on Information Theory (ISIT)
Hong Kong, June 2015

Part 2 of 2

The Plan

1. Prelude: The Chernoff bound
2. The entropy method
3. Logarithmic Sobolev inequalities
4. Transportation cost inequalities
5. Some applications

Given:

- $X_1, X_2, \ldots, X_n$: independent random variables
- $Z = f(X^n)$, for some real-valued f

Problem: derive sharp bounds on the deviation probabilities

$$\mathbb{P}[Z - \mathbb{E}Z \geq t], \qquad \text{for } t \geq 0$$

Benchmark:

- $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$
- $Z = f(X^n) = \frac{1}{n}(X_1 + \ldots + X_n)$ — sample mean

$$\mathbb{P}[Z \geq t] \leq \exp\left(-\frac{nt^2}{2\sigma^2}\right)$$

Goal:

- extend to other distributions besides Gaussians
- extend to nonlinear f

Prelude: The Chernoff Bound

Review: The Chernoff Bound

Define:

- logarithmic moment-generating function $\psi(\lambda) \triangleq \log \mathbb{E}[e^{\lambda(Z-\mathbb{E}Z)}]$
- its Legendre dual $\psi^*(t) \triangleq \sup_{\lambda \geq 0} \{\lambda t - \psi(\lambda)\}$

$$\begin{aligned} \mathbb{P}[Z - \mathbb{E}Z \geq t] &= \mathbb{P}[e^{\lambda(Z-\mathbb{E}Z)} \geq e^{\lambda t}] \\ &\leq e^{-\lambda t}\mathbb{E}[e^{\lambda(Z-\mathbb{E}Z)}] \qquad \text{Markov's inequality} \\ &= e^{-\{\lambda t - \psi(\lambda)\}} \end{aligned}$$

Optimize over λ: $\quad \mathbb{P}[Z - \mathbb{E}Z \geq t] \leq e^{-\psi^*(t)}$

Sanity check:

$$Z \sim N(0, \sigma^2) \qquad \psi(\lambda) = \frac{\lambda^2\sigma^2}{2} \qquad \psi^*(t) = \frac{t^2}{2\sigma^2}$$

$$\mathbb{P}[Z \geq t] \leq e^{-t^2/2\sigma^2}$$

Chernoff Bound: Subgaussian Random Variables

Definition. A real-valued r.v. Z is σ^2-*subgaussian* if

$$\psi(\lambda) \leq \frac{\lambda^2 \sigma^2}{2}$$

Immediate: if Z is σ^2-subgaussian, then

$$\begin{aligned} \psi^*(t) &= \sup_{\lambda \geq 0} \{\lambda t - \psi(\lambda)\} \\ &\geq \sup_{\lambda \geq 0} \left\{\lambda t - \lambda^2 \sigma^2 / 2\right\} \\ &= t^2 / 2\sigma^2 \end{aligned}$$

giving the Gaussian tail bound

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

How do we establish subgaussianity?

Review: Hoeffding's Lemma

Any almost surely bounded r.v. is subgaussian:

If there exist $-\infty < a \leq b < \infty$ such that $Z \in [a, b]$ a.s., then

$$\psi(\lambda) \leq \frac{\lambda^2 (b-a)^2}{8}.$$

Wassily Hoeffding

Corollary. If $Z \in [a, b]$ a.s., then

$$\mathbb{P}[Z - \mathbb{E}Z \geq t] \leq \exp\left(-\frac{2t^2}{(b-a)^2}\right)$$

Proof (of Corollary). By Hoeffding's lemma, Z is subgaussian with $\sigma^2 = (b-a)^2/4$.

Hoeffding's Lemma: An Alternative Proof

1. Assume without loss of generality that $\mathbb{E}Z = 0$.
2. Compute the first two derivatives of ψ:

$$\psi'(\lambda) = \frac{\mathbb{E}[Ze^{\lambda Z}]}{\mathbb{E}[e^{\lambda Z}]} \qquad \psi''(\lambda) = \frac{\mathbb{E}[Z^2 e^{\lambda Z}]}{\mathbb{E}[e^{\lambda Z}]} - \left(\frac{\mathbb{E}[Ze^{\lambda Z}]}{\mathbb{E}[e^{\lambda Z}]}\right)^2$$

3. Tilted distribution:

$$P = \mathcal{L}(Z) \longmapsto Q \qquad \frac{\mathrm{d}Q}{\mathrm{d}P}(Z) = \frac{e^{\lambda Z}}{\mathbb{E}_P[e^{\lambda Z}]}.$$

   Then $\psi'(\lambda) = \mathbb{E}_Q[Z]$, $\psi''(\lambda) = \mathrm{Var}_Q[Z]$.

4. $Z \in [a,b]$ P-a.s. $\Longrightarrow Z \in [a,b]$ Q-a.s. $\Longrightarrow \mathrm{Var}_Q[Z] \leq \frac{(b-a)^2}{4}$
5. Calculus:

$$\psi(\lambda) = \int_0^\lambda \int_0^\tau \psi''(\rho)\,\mathrm{d}\rho\,\mathrm{d}\tau \leq \frac{\lambda^2(b-a)^2}{8}. \qquad \blacksquare$$

The Entropy Method

Exponential Tilting and (Relative) Entropy

Back to our setting:

- $Z = f(X)$, X an arbitrary r.v.
- Want to prove subgaussianity of Z, so need to analyze

$$\psi(\lambda) = \log \mathbb{E}[e^{\lambda(Z-\mathbb{E}Z)}] = \log \mathbb{E}[e^{\lambda(f(X)-\mathbb{E}f(X))}].$$

- Let $P = \mathcal{L}(X)$, introduce tilted distribution $P^{\lambda f}$:

$$\frac{\mathrm{d}P^{\lambda f}}{\mathrm{d}P}(X) = \frac{e^{\lambda f(X)}}{\mathbb{E}[e^{\lambda f(X)}]}.$$

- We will relate $\psi(\lambda)$ to the relative entropy $D(P^{\lambda f} \| P)$.

Exponential Tilting and (Relative) Entropy

- Tilting of $P = \mathcal{L}(X)$:

$$\frac{\mathrm{d}P^{\lambda f}}{\mathrm{d}P}(X) = \frac{e^{\lambda f(X)}}{\mathbb{E}[e^{\lambda f(X)}]} \equiv \frac{e^{\lambda(f(X) - \mathbb{E} f(X))}}{e^{\psi(\lambda)}}.$$

- Relative entropy:

$$\begin{aligned} D(P^{\lambda f} \| P) &= \int \mathrm{d}P^{\lambda f} \log \frac{\mathrm{d}P^{\lambda f}}{\mathrm{d}P} \\ &= \int \mathrm{d}P^{\lambda f} \left(\lambda(f - \mathbb{E}_P f) - \psi(\lambda)\right) \\ &= \frac{\lambda \mathbb{E}_P[(f - \mathbb{E}_P f) e^{\lambda(f - \mathbb{E}_P f)}]}{e^{\psi(\lambda)}} - \psi(\lambda) \\ &= \lambda \psi'(\lambda) - \psi(\lambda) \end{aligned}$$

- With a bit of foresight,

$$\lambda \psi'(\lambda) - \psi(\lambda) = \lambda^2 \left(\frac{\psi'(\lambda)}{\lambda} - \frac{\psi(\lambda)}{\lambda^2} \right) = \lambda^2 \frac{\mathrm{d}}{\mathrm{d}\lambda} \left(\frac{\psi(\lambda)}{\lambda} \right)$$

The Herbst Argument

- Tilting of $P = \mathcal{L}(X)$: $\dfrac{\mathrm{d}P^{\lambda f}}{\mathrm{d}P}(X) = \dfrac{e^{\lambda f(X)}}{\mathbb{E}[e^{\lambda f(X)}]}$
- Relative entropy:

$$D(P^{\lambda f} \| P) = \lambda^2 \frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\frac{\psi(\lambda)}{\lambda}\right)$$

- Since $\lim_{\lambda \to 0} \psi(\lambda)/\lambda = 0$ (by l'Hôpital), we have

$$\psi(\lambda) = \lambda \int_0^\lambda \frac{D(P^{\rho f} \| P)}{\rho^2}\, \mathrm{d}\rho.$$

- Suppose now that $P = \mathcal{L}(X)$ and f are such that

$$D(P^{\rho f} \| P) \le \frac{\rho^2 \sigma^2}{2}, \qquad \forall \rho \ge 0$$

for some σ^2. Then

$$\psi(\lambda) \le \lambda \int_0^\lambda \frac{\rho^2 \sigma^2}{2\rho^2}\, \mathrm{d}\rho = \frac{\lambda^2 \sigma^2}{2}.$$

The Herbst Argument

Lemma (Herbst, 1975). Suppose that $Z = f(X)$ is such that

$$D(P^{\lambda f} \| P) \leq \frac{\lambda^2 \sigma^2}{2}, \qquad \forall \lambda \geq 0.$$

Then Z is σ^2-subgaussian, and so

$$\mathbb{P}[f(X) - \mathbb{E}f(X) \geq t] \leq e^{-t^2/2\sigma^2}, \qquad \forall t \geq 0.$$

Ira Herbst

For the sake of completeness, we … prove a theorem which states roughly that the potential of an intrinsically hypercontractive Schrödinger operator must increase at least quadratically at infinity. … [I]ts complete proof requires information in an unpublished letter of 1975 from I. Herbst to L. Gross. [C]ertain steps in the argument … can be written down abstractly.

— from a 1984 paper by B. Simon and E.B. Davies

The Herbst Converse

Lemma (R. van Handel, 2014). Suppose $Z = f(X)$ is $\sigma^2/4$-subgaussian. Then

$$D(P^{\lambda f} \| P) \le \frac{\lambda^2 \sigma^2}{2}, \qquad \forall \lambda \ge 0.$$

Proof. Let $\widetilde{\mathbb{E}}[\cdot] \triangleq \mathbb{E}_{P^{\lambda f}}[\cdot]$.

$$\begin{aligned}
D(P^{\lambda f} \| P) &= \widetilde{\mathbb{E}}\left[\log \frac{\mathrm{d}P^{\lambda f}}{\mathrm{d}P}\right] \overset{\text{(Jensen)}}{\le} \log \widetilde{\mathbb{E}}\left[\frac{\mathrm{d}P^{\lambda f}}{\mathrm{d}P}\right] \\
&= \log \widetilde{\mathbb{E}}\left[\frac{e^{\lambda(f(X) - \mathbb{E}f(X))}}{\mathbb{E}[e^{\lambda(f(X) - \mathbb{E}f(X))}]}\right] \\
&= \log \mathbb{E}\left[e^{2\lambda(f - \mathbb{E}f)}\right] - \log\Big\{\underbrace{\mathbb{E}[e^{\lambda(f - \mathbb{E}f)}]}_{\ge 1}\Big\}^2 \\
&\le \frac{(2\lambda)^2 \sigma^2/4}{2} = \frac{\lambda^2\sigma^2}{2} \qquad \blacksquare
\end{aligned}$$

The Herbst Argument: What Is It Good For?

- Subgaussianity of $Z = f(X)$ is equivalent to $D(P^{\lambda f} \| P) = O(\lambda^2)$, but what does that give us?
- Recall: we are interested in *high-dimensional* settings

$$Z = f(X^n) = f(X_1, \ldots, X_n)$$
$$X_1, \ldots, X_n \text{ — independent r.v.'s}$$

  Thus, P is a product measure:

$$P = P_1 \otimes P_2 \otimes \ldots \otimes P_n, \qquad P_i \triangleq \mathcal{L}(X_i)$$

- The relative entropy *tensorizes*!! — we can break the (hard) n-dimensional problem into n (hopefully) easier 1-dimensional problems.

Tensorization

$$X_1, \ldots, X_n \text{ — independent r.v.'s}$$
$$P = \mathcal{L}(X^n) = P_1 \otimes \ldots \otimes P_n, \qquad P_i = \mathcal{L}(X_i)$$

- Recall the Efron–Stein–Steele inequality:

$$\mathrm{Var}_P[f(X^n)] \le \sum_{i=1}^n \mathbb{E}\left[\mathrm{Var}_P[f(X^n)|\bar{X}^i]\right]$$

— tensorization of variance

- Tensorization of relative entropy: for an *arbitrary* probability measure Q on X^n,

$$D(Q\|P) \le \sum_{i=1}^n \underbrace{D(Q_{X_i|\bar{X}^i} \| P_{X_i|\bar{X}^i} | Q_{\bar{X}^i})}_{\text{conditional divergence}}$$

— independence of the X_i's is key

Tensorization: A Quick Proof

$$
\begin{aligned}
&D(Q\|P) \\
&= \sum_{i=1}^n D(Q_{X_i|X^{i-1}} \| P_{X_i|X^{i-1}} | Q_{X^{i-1}}) && \text{(chain rule)} \\
&= \sum_{i=1}^n \mathbb{E}_Q \left[\log \frac{\mathrm{d}Q_{X_i|X^{i-1}}}{\mathrm{d}P_{X_i|X^{i-1}}} \right] \\
&= \sum_{i=1}^n \mathbb{E}_Q \left[\log \frac{\mathrm{d}Q_{X_i|X^{i-1}}}{\mathrm{d}Q_{X_i|\bar{X}^i}} \right] + \sum_{i=1}^n \mathbb{E}_Q \left[\log \frac{\mathrm{d}Q_{X_i|\bar{X}^i}}{\mathrm{d}P_{X_i|X^{i-1}}} \right] \\
&= -\sum_{i=1}^n \mathbb{E}_Q \left[\log \frac{\mathrm{d}Q_{X_i|\bar{X}^i}}{\mathrm{d}Q_{X_i|X^{i-1}}} \right] + \sum_{i=1}^n \mathbb{E}_Q \left[\log \frac{\mathrm{d}Q_{X_i|\bar{X}^i}}{\mathrm{d}P_{X_i|\bar{X}^i}} \right] && \text{(independence)} \\
&= -\sum_{i=1}^n D(Q_{X_i|\bar{X}^i} \| Q_{X_i|X^{i-1}} | Q_{\bar{X}^i}) + \sum_{i=1}^n D(Q_{X_i|\bar{X}^i} \| P_{X_i} | Q_{\bar{X}^i}) \\
&\le \sum_{i=1}^n D(Q_{X_i|\bar{X}^i} \| P_{X_i} | Q_{\bar{X}^i}) \\
&\equiv D^-(Q\|P) && \text{— erasure divergence (Verdú–Weissman, 2008)}
\end{aligned}
$$

Tensorization and Tilting

- Recall: we are interested in $D(Q\|P)$, where

$$P = P_1 \otimes \ldots \otimes P_n, \qquad \mathrm{d}Q = \frac{e^{\lambda f}}{\mathbb{E}_P[e^{\lambda f}]}\,\mathrm{d}P$$

- Then

$$\begin{aligned}\frac{\mathrm{d}Q_{\bar{X}^i}}{\mathrm{d}P_{\bar{X}^i}}(\bar{x}^i) &= \int_{\mathsf{X}} P_i(\mathrm{d}x_i)\frac{e^{\lambda f(x_1,\ldots,x_{i-1},x_i,x_{i+1},\ldots,x_n)}}{\mathbb{E}_P[e^{\lambda f(X^n)}]} \\ &= \frac{\mathbb{E}_P[e^{\lambda f(X^n)}|\bar{X}^i = \bar{x}^i]}{\mathbb{E}_P[e^{\lambda f(X^n)}]}\end{aligned}$$

therefore

$$\frac{\mathrm{d}Q_{X_i|\bar{X}^i=\bar{x}^i}}{\mathrm{d}P_{X_i}}(x_i) = \frac{e^{\lambda f(x_1,\ldots,x_{i-1},x_i,x_{i+1},\ldots,x_n)}}{\mathbb{E}_P[e^{\lambda f(x_1,\ldots,x_{i-1},X_i,x_{i+1},\ldots,x_n)}]}$$

— remember, $\bar{x}^i$ is fixed.

Tensorization and Tilting

$$\frac{\mathrm{d}P^{\lambda f}_{X_i|\bar{X}^i=\bar{x}^i}}{\mathrm{d}P_{X_i}}(x_i) = \frac{e^{\lambda f(x_1,\ldots,x_{i-1},x_i,x_{i+1},\ldots,x_n)}}{\mathbb{E}_{P_i}\big[e^{\lambda f(x_1,\ldots,x_{i-1},X_i,x_{i+1},\ldots,x_n)}\big]}$$

- For a fixed $\bar{x}^i$, define the function

$$f_i(\cdot|\bar{x}^i) : \mathsf{X} \to \mathbb{R}$$

$$\begin{aligned} f_i(x_i|\bar{x}^i) &= f(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) \\ &= f_i(x_i) \qquad\qquad \text{(just shorthand, always remember } \bar{x}^i\text{)} \end{aligned}$$

- Now observe that $Q_{X_i|\bar{X}^i=\bar{x}^i}$ is the tilting of $P_i \equiv P_{X_i}$:

$$\mathrm{d}Q_{X_i|\bar{X}^i=\bar{x}^i} = \frac{e^{\lambda f_i}}{\mathbb{E}_{P_i}\big[e^{\lambda f_i}\big]}\,\mathrm{d}P_i$$

Tensorization and Tilting

Lemma. If $X_1, \ldots, X_n$ are independent r.v.'s with joint law $P = P_1 \otimes \ldots \otimes P_n$, where $P_i = \mathcal{L}(X_i)$, then for any $f : \mathsf{X}^n \to \mathbb{R}$

$$D(P^{\lambda f} \| P) \le \sum_{i=1}^{n} \widetilde{\mathbb{E}}\left[D(P_i^{\lambda f_i} \| P_i) \right],$$

where $\widetilde{\mathbb{E}}[\cdot] \triangleq \mathbb{E}_{P^{\lambda f}}[\cdot]$ and $f_i(\cdot) = f_i(\cdot|\bar{X}^i)$ for each i.

Proof.

$$\begin{aligned}
D(P^{\lambda f} \| P) &\le \sum_{i=1}^{n} D(P^{\lambda f}_{X_i|\bar{X}^i} \| P_{X_i} | P^{\lambda f}_{\bar{X}^i}) \\
&= \sum_{i=1}^{n} \widetilde{\mathbb{E}} \left[\log \frac{\mathrm{d}P^{\lambda f}_{X_i|\bar{X}^i}}{\mathrm{d}P_{X_i}} \right] \\
&= \sum_{i=1}^{n} \widetilde{\mathbb{E}} \left[\log \frac{\mathrm{d}P_i^{\lambda f_i}}{\mathrm{d}P_i} \right] = \sum_{i=1}^{n} \widetilde{\mathbb{E}} \left[D(P_i^{\lambda f_i} \| P_i) \right] \quad \blacksquare
\end{aligned}$$

The Entropy Method: Divide and Conquer

- Want to derive a subgaussian tail bound

$$\mathbb{P}[f(X^n) - \mathbb{E}f(X^n) \geq t] \leq e^{-t^2/2\sigma^2}, \qquad \forall t \geq 0$$

  where $X_1, \ldots, X_n$ are independent r.v.'s.

- Suppose we can prove there exist constants $c_1, \ldots, c_n$, such that

$$(\star) \qquad D(P_i^{\lambda f_i} \| P_i) \leq \frac{\lambda^2 c_i^2}{2}, \qquad i \in \{1, \ldots, n\}.$$

  Then

$$D(P^{\lambda f} \| P) \overset{\text{(tensor.)}}{\leq} \frac{\lambda^2 \sum_{i=1}^n c_i^2}{2} \overset{\text{(Herbst)}}{\Longrightarrow} \sigma^2 = \sum_{i=1}^n c_i^2$$

Now we "just" need to prove $(\star)$!!

Logarithmic Sobolev Inequalities

Log-Sobolev in a Nutshell

- Goal: control the relative entropy $D(P^{\lambda f} \| P)$.
- A *log-Sobolev inequality* ties together:
  - (i) the underlying probability measure P
  - (ii) a function class $\mathcal{A}$ (containing f of interest)
  - (iii) an "energy" functional $E : \mathcal{A} \to \mathbb{R}$ such that

$$E(\alpha f) = \alpha E(f), \qquad \forall f \in \mathcal{A},\ \alpha \geq 0$$

  and looks like this:

$$D(P^{f} \| P) \leq \frac{c}{2} E^2(f), \qquad \forall f \in \mathcal{A}.$$

- In that case, if $E(f) \leq L$, then

$$D(P^{\lambda f} \| P) \leq \frac{c}{2} E^2(\lambda f) = \frac{c}{2} \lambda^2 E^2(f) \leq \frac{\lambda^2 c L^2}{2}$$

- The name comes from an analogy with *Sobolev inequalities* in functional analysis.

The Bernoulli Log-Sobolev Inequality

Theorem (Gross, 1975). Let $X_1, \dots, X_n$ be i.i.d. Bern(1/2) random variables. Then, for any function $f : \{0,1\}^n \to \mathbb{R}$,

$$D(P^f \| P) \le \frac{1}{8} \frac{\mathbb{E}[|Df(X^n)|^2 e^{f(X^n)}]}{\mathbb{E}[e^{f(X^n)}]},$$

where

$$Df(x^n) \triangleq \sqrt{\sum_{i=1}^{n} |f(x^n) - f(\underbrace{x^n \oplus e_i}_{\text{flip } i\text{th bit}})|^2},$$

and $\mathbb{E}[\cdot]$ is w.r.t. $P = (\text{Bern}(1/2))^{\otimes n}$.

Leonard Gross

Remarks:

- This is not the original form of the inequality from Gross' 1975 paper, but they are equivalent.
- Note that $D(\lambda f) = \lambda D(f)$ for all $\lambda \ge 0$.

Bernoulli LSI: Proof Sketch

- Consider first $n = 1$ and $f : \{0,1\} \to \mathbb{R}$ with $a = f(0)$ and $b = f(1)$.
- In that case, the log-Sobolev inequality reads

$$\frac{e^a}{e^a + e^b} \log \frac{2e^a}{e^a + e^b} + \frac{e^b}{e^a + e^b} \log \frac{2e^b}{e^a + e^b} \le \frac{1}{8}(b-a)^2.$$

  Proof: Elementary (but tedious) exercise in calculus.
- Tensorization:

$$\begin{aligned} D(P^f \| P) &\le \sum_{i=1}^n \widetilde{\mathbb{E}}[D(P_i^{f_i} \| P_i)] \qquad \text{where } \tilde{\mathbb{E}}[h(X^n)] = \frac{\mathbb{E}[h(X^n)e^{f(X^n)}]}{\mathbb{E}[e^{f(X^n)}]} \\ &\le \frac{1}{8} \sum_{i=1}^n \widetilde{\mathbb{E}}\left[\left| f(X^{i-1}, 0, X_{i+1}^n) - f(X^{i-1}, 1, X_{i+1}^n) \right|^2 \right] \\ &= \frac{1}{8} \widetilde{\mathbb{E}}\left[|Df(X^n)|^2 \right] = \frac{1}{8} \frac{\mathbb{E}[|Df(X^n)|^2 e^{f(X^n)}]}{\mathbb{E}[e^{f(X^n)}]} \qquad \blacksquare \end{aligned}$$

The Gaussian Log-Sobolev Inequality

Theorem (Gross, 1975). Let $X_1, \ldots, X_n$ be i.i.d. $N(0,1)$ random variables. Then, for any smooth function $f : \mathbb{R}^n \to \mathbb{R}$,

$$D(P^f \| P) \le \frac{1}{2} \frac{\mathbb{E}\left[\|\nabla f(X^n)\|_2^2 \, e^{f(X^n)}\right]}{\mathbb{E}[e^{f(X^n)}]},$$

where all expectations are w.r.t. $P = \mathcal{L}(X^n) = N(0, I_n)$.

Leonard Gross

Remarks:

- This is not the original form of the inequality from Gross' 1975 paper, but they are equivalent.
- Equivalent forms of Gaussian LSI have been obtained independently by A. Stam (1959) and by P. Federbush (1969).
- The contribution of Stam (via the entropy power inequality) was first pointed out by E.A. Carlen in 1991.

Proof(s) of the Gaussian LSI

- There are *many* ways of proving the Gaussian log-Sobolev inequality.
- Original proof by Gross: apply the Bernoulli LSI to

$$f\left(\frac{X_1+\ldots+X_n-n/2}{\sqrt{n/4}}\right), \qquad X_i \stackrel{\text{i.i.d.}}{\sim} \mathrm{Bern}(1/2)$$

  then use the Central Limit Theorem:

$$\frac{X_1+\ldots+X_n-n/2}{\sqrt{n/4}} \rightsquigarrow N(0,1) \qquad \text{as } n \to \infty$$

  - via Markov semigroups
  - via hypercontractivity (E. Nelson)
  - via Stam's inequality for entropy power and Fisher info.
  - via I-MMSE relation (cf. Raginsky and Sason)
  - ...

Application of Gaussian LSI

Theorem (Tsirelson–Ibragimov–Sudakov, 1976). Let $X_1, \ldots, X_n$ be independent $N(0,1)$ random variables. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a function which is L-Lipschitz:

$$|f(x^n) - f(y^n)| \le L\|x^n - y^n\|_2, \qquad \forall x^n, y^n \in \mathbb{R}^n.$$

Then $Z = f(X^n)$ is L^2-subgaussian:

$$\log \mathbb{E}[e^{\lambda(f(X^n) - \mathbb{E}f(X^n))}] \le \frac{\lambda^2 L^2}{2}, \qquad \forall \lambda \ge 0.$$

Remarks:

- The original proof did not rely on the Gaussian LSI.
- It is a striking result: f can be an arbitrary nonlinear function, and the subgaussian constant is *independent* of the dimension n.

Tsirelson–Ibragimov–Sudakov: Proof via LSI

- By an approximation argument, can assume that f is differentiable. Since it is L-Lipschitz, $\|\nabla f\|_2 \le L$.
- By the Gaussian LSI, for any $\lambda \ge 0$,

$$\begin{aligned} D(P^{\lambda f} \| P) &\le \frac{1}{2} \frac{\mathbb{E}\left[\|\lambda \nabla f(X^n)\|_2^2 \, e^{\lambda f(X^n)}\right]}{\mathbb{E}\left[e^{\lambda f(X^n)}\right]} \\ &= \frac{\lambda^2}{2} \frac{\mathbb{E}\left[\|\nabla f(X^n)\|_2^2 \, e^{\lambda f(X^n)}\right]}{\mathbb{E}\left[e^{\lambda f(X^n)}\right]} \\ &\le \frac{\lambda^2 L^2}{2}. \end{aligned}$$

- By the Herbst argument,

$$\log \mathbb{E}\left[e^{\lambda(f(X^n) - \mathbb{E} f(X^n))}\right] \le \frac{\lambda^2 L^2}{2}. \qquad \blacksquare$$

A Gaussian Concentration Bound

The Tsirelson–Ibragimov–Sudakov inequality gives us

Corollary. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(0,1)$, and let $f : \mathbb{R}^n \to \mathbb{R}$ be an L-Lipschitz function. Then

$$\mathbb{P}\left[f(X^n) - \mathbb{E}f(X^n) \geq t\right] \leq e^{-t^2/2L^2}.$$

Proof. Use the Chernoff bound.

Remarks:

- This is an example of *dimension-free concentration*: the tail bound does not depend on n.
- Applying the same result to $-f$ and using the union bound, we get

$$\mathbb{P}\left[|f(X^n) - \mathbb{E}f(X^n)| \geq t\right] \leq 2e^{-t^2/2L^2}.$$

Deriving Log-Sobolev (1)

- Are there *systematic ways* to derive log-Sobolev?
- The usual (probabilistic) approach (a subtle art):
  - Construct a continuous-time Markov process $\{X_t\}_{t\in\geq 0}$ with stationary distribution P and Markov generator

$$\mathbb{L}f(x) \triangleq \lim_{t\downarrow 0} \frac{\mathbb{E}[f(X_t)|X_0 = x] - f(x)}{t}$$

  - Use the structure of $\mathbb{L}$ to obtain an inequality of the form

$$D(P^f \| P) \leq \frac{c}{2} \frac{\mathcal{E}(e^f, f)}{\mathbb{E}_P[e^f]},$$

    where $\mathcal{E}(g,h) \triangleq -\mathbb{E}_P[f(X)\mathbb{L}g(X)]$ is the *Dirichlet form.*
  - Extract Γ by looking for a bound of the form

$$\mathcal{E}(e^f, f) \leq \mathbb{E}_P[|\Gamma f(X)|^2 e^{f(X)}]$$

  - Different choices of $\mathbb{L}$ (for the same P) will yield different Γ's, and hence different log-Sobolev inequalities.

Deriving Log-Sobolev (2)

- An alternative (information-theoretic) approach: based on a recent paper of A. Maurer (2012).
- Exploits a representation of $D(P^{\lambda f} \| P)$ in terms of the variance of $f(X)$ under the tilted distributions

$$\mathrm{d}P^{sf} = \frac{e^{sf}}{\mathbb{E}_P[e^{sf}]}\,\mathrm{d}P.$$

- An interpretation in terms of statistical physics: think of $-f$ as energy and of $s \geq 0$ as inverse temperature. Then

$$\mathrm{Var}^{sf}[f(X)] \triangleq \frac{\mathbb{E}_P[f^2(X)e^{sf(X)}]}{\mathbb{E}_P[e^{sf(X)}]} - \left(\frac{\mathbb{E}_P[f(X)e^{sfX()}]}{\mathbb{E}_P[e^{sf(X)}]}\right)^2$$

  gives the "thermal fluctuations" of $-f$ at temp. $T = 1/s$.
- Infinite-temperature limit ($T \to \infty$): recover $\mathrm{Var}_P[f(X)]$.

Entropy via Thermal Fluctuations

Theorem (A. Maurer, 2012). Let X be a random variable with law P. Then for any real-valued function f and any $\lambda \geq 0$

$$D(P^{\lambda f} \| P) = \int_0^\lambda \int_t^\lambda \mathrm{Var}^{sf}[f(X)] \, \mathrm{d}s \, \mathrm{d}t,$$

where $\mathrm{Var}^{sf}[f(X)]$ is the variance of $f(X)$ under the tilted distribution P^{sf}.

Andreas Maurer

Recall:

$$\mathrm{Var}^{sf}[f(X)] = \frac{\mathbb{E}[f^2(X) e^{sf(X)}]}{\mathbb{E}[e^{sf(X)}]} - \left(\frac{\mathbb{E}[f(X) e^{sf(X)}]}{\mathbb{E}[e^{sf(X)}]} \right)^2 \equiv \psi''(s)$$

where $\psi(s) = \log \mathbb{E}[e^{s(f(X) - \mathbb{E}f(X))}]$

Proof

- Recall
$$D(P^{\lambda f}\|P) = \lambda\psi'(\lambda) - \psi(\lambda), \qquad \text{where } \psi(\lambda) = \log \mathbb{E}[e^{\lambda(f-\mathbb{E}f)}].$$

- Since $\psi(0) = \psi'(0) = 0$, we have
$$\begin{aligned} \lambda\psi'(\lambda) &= \int_0^\lambda \psi'(\lambda)\,\mathrm{d}t \\ \psi(\lambda) &= \int_0^\lambda \psi'(t)\,\mathrm{d}t. \end{aligned}$$

- Substitute:
$$\begin{aligned} D(P^{\lambda f}\|P) &= \int_0^\lambda (\psi'(\lambda) - \psi'(t))\,\mathrm{d}t \\ &= \int_0^\lambda \int_t^\lambda \psi''(s)\,\mathrm{d}s\,\mathrm{d}t \\ &= \int_0^\lambda \int_t^\lambda \mathrm{Var}^{sf}[f(X)]\,\mathrm{d}s\,\mathrm{d}t. \qquad \blacksquare \end{aligned}$$

From Thermal Fluctuations to Log-Sobolev

Theorem. Let $\mathcal{A}$ be a class of functions of X, and suppose that there is a mapping $\Gamma : \mathcal{A} \to \mathbb{R}$, such that:

1. For all $f \in \mathcal{A}$ and $\alpha \geq 0$, $\Gamma(\alpha f) = \alpha \Gamma(f)$.
2. There exists a constant $c > 0$, such that

$$\mathrm{Var}^{\lambda f}[f(X)] \leq c|\Gamma(f)|^2, \qquad \forall f \in \mathcal{A},\ \lambda \geq 0.$$

Then

$$D(P^{\lambda f} \| P) \leq \frac{\lambda^2 c |\Gamma(f)|^2}{2}, \qquad \forall f \in \mathcal{A}, \lambda \geq 0.$$

Proof.

$$D(P^{\lambda f} \| P) \leq c|\Gamma(f)|^2 \int_0^\lambda \int_t^\lambda \mathrm{d}s \, \mathrm{d}t = \frac{c|\Gamma(f)|^2 \lambda^2}{2} \qquad \blacksquare$$

From Thermal Fluctuations to Log-Sobolev: Example 1

Let's use Maurer's method to derive the Bernoulli LSI.

- For any $f : \{0,1\} \to \mathbb{R}$, define

$$\Gamma(f) \triangleq |f(0) - f(1)|.$$

- Since f is obviously bounded, for every $s \geq 0$ we have

$$\mathrm{Var}^{sf}[f(X)] \leq \frac{(f(0) - f(1))^2}{4} \equiv \frac{|\Gamma f|^2}{4}.$$

- Finally,

$$\begin{aligned} D(P^f \| P) &= \int_0^1 \int_t^1 \mathrm{Var}^{sf}[f(X)] \, \mathrm{d}s \, \mathrm{d}t \\ &\leq \frac{|\Gamma(f)^2}{4} \int_0^1 \int_t^1 \mathrm{d}s \, \mathrm{d}t \\ &= \frac{1}{8} |\Gamma f|^2. \end{aligned}$$

- For $n > 1$, use tensorization.

From Thermal Fluctuations to Log-Sobolev: Example 2

Let's use Maurer's method to derive McDiarmid's inequality.

- We will use tensorization, so let's first consider $n = 1$.
- We are interested in all functions $f : \mathsf{X} \to \mathbb{R}$, such that

$$\sup_{x \in \mathsf{X}} f(x) - \inf_{x \in \mathsf{X}} f(x) \le c$$

  for some $c < \infty$.
- Define $\Gamma(f) \triangleq \sup_{x \in \mathsf{X}} f(x) - \inf_{x \in \mathsf{X}} f(x)$.
- Any f satisfies $f(X) \in [\inf f, \sup f]$. If $\Gamma(f) < \infty$, then $[\inf f, \sup f]$ is a bounded interval.
- In that case, for any P,

$$\mathrm{Var}^{sf}[f(X)] \le \frac{(\sup f - \inf f)^2}{4} = \frac{|\Gamma(f)|^2}{4}.$$

  Using the integral representation of the divergence, we get

$$D(P^{\lambda f} \| P) \le \frac{\lambda^2 c^2}{8}, \qquad \text{if } \sup f - \inf f \le c.$$

Proof of McDiarmid (cont.)

- So far, we have obtained

$$(\star) \qquad D(P^{\lambda f} \| P) \le \frac{\lambda^2 c^2}{8}, \qquad \text{if } \sup f - \inf f \le c.$$

- Let $X_i \sim P_i, 1 \le i \le n$, be independent r.v.'s.
- Consider $f : \mathsf{X}^n \to \mathbb{R}$ that has bounded differences:

$$\sup_{\bar{x}^i} \left(\sup_{x_i} f(x^{i-1}, x_i, x_{i+1}^n) - \inf_{x_i} f(x^{i-1}, x_i, x_{i+1}^n) \right) \le c_i$$

for all i, for some constants $0 \le c_1, \ldots, c_n < \infty$.

- For each i, apply $(\star)$ to $f_i(\cdot) \equiv f(x^{i-1}, \cdot, x_{i+1}^n)$:

$$D(P_i^{\lambda f_i} \| P_i) \le \frac{1}{8} \left(\sup_{x_i} f(x^{i-1}, x_i, x_{i+1}^n) - \inf_{x_i} f(x^{i-1}, x_i, x_{i+1}^n) \right)^2$$

[recall, $f_i(\cdot)$ depends on $\bar{x}^i$].

Proof of McDiarmid (cont.)

- Now we tensorize: for $P = P_1 \otimes \ldots \otimes P_n$,

$$
\begin{aligned}
&D(P^{\lambda f} \| P) \\
&\le \sum_{i=1}^{n} \widetilde{\mathbb{E}}\left[D(P_i^{\lambda f_i} \| P_i) \right] \\
&\le \frac{\lambda^2}{8} \widetilde{\mathbb{E}}\Bigg[\underbrace{\sum_{i=1}^{n} \Big(\sup_{x_i} f(X^{i-1}, x_i, X_{i+1}^n) - \inf_{x_i} f(X^{i-1}, x_i, X_{i+1}^n) \Big)^2}_{=|\Gamma(f)(X^n)|^2} \Bigg]
\end{aligned}
$$

Theorem. Let $X_1, \ldots, X_n \in \mathsf{X}$ be independent r.v.'s with joint law $P = P_1 \otimes \ldots \otimes P_n$. Then, for any function $f : \mathsf{X}^n \to \mathbb{R}$,

$$D(P^f \| P) \le \frac{1}{8} \left\| |\Gamma f|^2 \right\|_\infty,$$

where

$$\Gamma f(x^n) = \left\{ \sum_{i=1}^n \Big(\underbrace{\sup_{x_i} f(x^{i-1}, x_i, x_{i+1}^n) - \inf_{x_i} f(x^{i-1}, x_i, x_{i+1}^n)}_{=\Gamma_i f(\bar{x}^i)} \Big)^2 \right\}^{1/2}$$

Remarks:

- McDiarmid: if f has bounded differences with $c_1, \ldots, c_n$, then

$$f(X^n) \text{ is } \frac{\sum_{i=1}^n c_i^2}{4}\text{-subgaussian}$$

- Since $\left\| |\Gamma f|^2 \right\|_\infty \le \sum_{i=1}^n \left\| |\Gamma_i f|^2 \right\|_\infty$, the above theorem is stronger than McDiarmid.

Transportation-Cost Inequalities

Concentration and the Lipschitz Property

A common theme:

- Let $f : \mathbb{R}^n \to \mathbb{R}$ be 1-Lipschitz w.r.t. the Euclidean norm:

$$|f(x^n) - f(y^n)| \le \|x^n - y^n\|_2.$$

  Let $X_1, \dots, X_n$ be i.i.d. $N(0,1)$ r.v.'s. Then

$$\mathbb{P}\left[|f(X^n) - \mathbb{E}f(X^n)| \ge t\right] \le 2e^{-t^2/2}.$$

- Let $f : \mathsf{X}^n \to \mathbb{R}$ be 1-Lipschitz w.r.t. a weighted Hamming metric:

$$|f(x^n) - f(y^n)| \le d_{\mathbf{c}}(x^n, y^n), \qquad d_{\mathbf{c}}(x^n, y^n) \triangleq \sum_{i=1}^n c_i \mathbf{1}\{x_i \ne y_i\}$$

  (this is equivalent to bounded differences). Let $X_1, \dots, X_n$ be independent r.v.'s. Then

$$\mathbb{P}\left[|f(X^n) - \mathbb{E}f(X^n)| \ge t\right] \le 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right).$$

The Setting: Probability in Metric Spaces

- Let (X, d) be a metric space.
- A function $f : \mathsf{X} \to \mathbb{R}$ is *L-Lipschitz* (w.r.t. d) if

$$|f(x) - f(y)| \le L d(x, y), \qquad \forall x, y \in \mathsf{X}.$$

- Notation: $\mathrm{Lip}_L(\mathsf{X}, d)$ – the class of all L-Lipschitz functions.

Question: What conditions does a probability measure P on X have to satisfy, so that $f(X)$, $X \sim P$, is σ^2-subgaussian for every $f \in \mathrm{Lip}_1(\mathsf{X}, d)$?

Some Definitions

- A *coupling* of two probability measures P and Q on X is any probability measure π on $\mathsf{X} \times \mathsf{X}$, such that

$$(X, Y) \sim \pi \qquad \Longrightarrow \qquad X \sim P, Y \sim Q.$$

- $\Pi(P, Q)$: family of all couplings of P and Q.

Definition. For $p \geq 1$, the L^p *Wasserstein distance* between P and Q is given by

$$W_p(P, Q) \triangleq \inf_{\pi \in \Pi(P,Q)} \left(\mathbb{E}_\pi[d^p(X, Y)]\right)^{1/p}.$$

Kantorovich–Rubinstein formula. For any two P, Q,

$$W_1(P, Q) = \sup_{f \in \mathrm{Lip}_1(\mathsf{X}, d)} \left|\mathbb{E}_P[f] - \mathbb{E}_Q[f]\right|.$$

Optimal Transportation

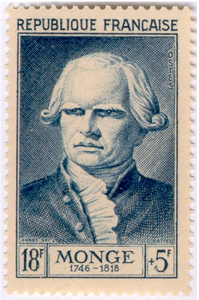
Monge-Kantorovich Optimal Transportation Problem. Given two probability measures P, Q on a common space X and a cost function $c : \mathsf{X} \times \mathsf{X} \to \mathbb{R}_+$,

$$\text{minimize} \qquad \mathbb{E}_\pi[c(X,Y)]$$

over all couplings $\pi \in \Pi(P,Q)$.

Interpretation:

- P and Q: initial and final distributions of some material (say, sand) in space
- $c(x,y)$: cost of transporting a grain of sand from location x to location y
- $\pi(\mathrm{d}y|x)$: randomized strategy for transporting from location x
- Wasserstein distances: transportation cost is some power of a metric



Gaspard Monge

Leonid Kantorovich

Transportation Cost Inequalities

Definition. A probability measure P on a metric space (X, d) satisfies an L^p *transportation cost inequality* with constant c, or $T_p(c)$ for short, if

$$W_p(P, Q) \leq \sqrt{2cD(Q\|P)}, \qquad \forall Q.$$

Preview:

- We will be primarily concerned with $T_1(c)$ and $T_2(c)$ inequalities.
- $T_1(c)$ is easier to work with, while $T_2(c)$ has strong properties (dimension-free concentration).

Examples of TC Inequalities: 1

- X: arbitrary space with trivial metric $d(x, y) = \mathbf{1}\{x \neq y\}$
- Then

$$\begin{aligned} W_1(P, Q) &= \inf_{\pi \in \Pi(P,Q)} \mathbb{E}_\pi[\mathbf{1}\{X \neq Y\}] \\ &\equiv \inf_{\pi \in \Pi(P,Q)} \mathbb{P}_\pi[X \neq Y] \\ &= \|P - Q\|_{\mathrm{TV}} \qquad \text{– total variation distance} \end{aligned}$$

  (proof by explicit construction of optimal coupling)
- Any P satisfies $T_1(1/4)$:

$$\|P - Q\|_{\mathrm{TV}} \le \sqrt{\frac{1}{2} D(Q\|P)}$$

  (Csiszár–Kemperman–Kullback–Pinsker)

Examples of TC Inequalities: 2

- $\mathsf{X} = \{0, 1\}$ with trivial metric $d(x, y) = \mathbf{1}\{x \neq y\}$
- $P = \mathrm{Bern}(p)$
- Distribution-dependent refinement of Pinsker:

$$\|P - Q\|_{\mathrm{TV}} \leq \sqrt{2c(p) D(Q\|P)},$$

  where

$$c(p) = \frac{p - \bar{p}}{2(\log p - \log \bar{p})}, \qquad \bar{p} \triangleq 1 - p$$

  (Ordentlich–Weinberger, 2005)
- Thus, $P = \mathrm{Bern}(p)$ satisfies $T_1(c(p))$, and the constant is optimal:

$$\inf_Q \frac{D(Q\|P)}{\|Q - P\|_{\mathrm{TV}}^2} = \frac{1}{2c(p)}.$$

Examples of TC Inequalities: 3

- $\mathsf{X} = \mathbb{R}^n$ with $d(x, y) = \|x - y\|_2$
- The Gaussian measure $P = N(0, I_n)$ satisfies $T_2(1)$:

  $$W_2(P, Q) \le \sqrt{2D(Q\|P)}$$

  (Talagrand, 1996)
- Note: the constant is *independent* of n!!

Michel Talagrand

Transportation and Concentration

Theorem (Bobkov–Götze, 1999). Let P be a probability measure on a metric space (X, d). Then the following two statements are equivalent:

1. $f(X)$, $X \sim P$, is σ^2-subgaussian for every $f \in \mathrm{Lip}_1(\mathsf{X}, d)$.
2. P satisfies $T_1(c)$ on (X, d):

$$W_1(P, Q) \le \sqrt{2\sigma^2 D(Q\|P)}, \qquad \forall Q \ll P.$$

Sergey Bobkov

Friedrich Götze

Remarks:

- Connection between concentration and transportation inequalities was first pointed out by Marton (1986).
- Remarkable result: concentration phenomenon for Lipschitz functions can be expressed purely in terms of probabilistic and metric structures.

Proof of Bobkov–Götze

(Ultra-light version, due to R. van Handel)

- Statement 1 of the theorem is equivalent to

$$(\star) \qquad \sup_{\lambda \geq 0} \sup_{f \in \mathrm{Lip}_1(\mathsf{X},d)} \left\{ \log \mathbb{E}_P \left[e^{\lambda (f - \mathbb{E}_P f)} \right] - \frac{\lambda^2 \sigma^2}{2} \right\} \leq 0.$$

- Use Gibbs variational principle:

$$\log \mathbb{E}_P[e^h] = \sup_Q \left\{ \mathbb{E}_Q[h] - D(Q \| P) \right\}, \qquad \forall h \text{ s.t. } e^h \in L_1(P)$$

(supremum achieved by the tilted distribution $Q = P^h$).

- Then $(\star)$ is equivalent to

$$(\star\star) \quad \sup_{\lambda \geq 0} \sup_{f \in \mathrm{Lip}_1(\mathsf{X},d)} \sup_Q \left\{ \lambda \left(\mathbb{E}_Q[f] - \mathbb{E}_P[f] \right) - D(Q \| P) - \frac{\lambda^2 \sigma^2}{2} \right\} \leq 0$$

Proof of Bobkov–Götze (cont.)

$$(\star\star)\quad \sup_{\lambda\ge 0}\sup_{f\in \mathrm{Lip}_1(\mathsf{X},d)}\sup_{Q}\left\{\lambda\left(\mathbb{E}_Q[f]-\mathbb{E}_P[f]\right)-D(Q\|P)-\frac{\lambda^2\sigma^2}{2}\right\}\le 0$$

- Interchange the order of suprema:

$$\sup_{\lambda\ge 0}\sup_{f\in \mathrm{Lip}_1(\mathsf{X},d)}\sup_{Q}[\ldots]=\sup_{Q}\sup_{\lambda\ge 0}\sup_{f\in \mathrm{Lip}_1(\mathsf{X},d)}[...]$$

- Then

$$(\star\star)\Longleftrightarrow \sup_{Q}\sup_{\lambda\ge 0}\left\{\lambda W_1(P,Q)-D(Q\|P)-\frac{\lambda^2\sigma^2}{2}\right\}\le 0$$

(by Kantorovich–Rubinstein)

$$\Longleftrightarrow \sup_{Q}\left\{\frac{W_1(P,Q)}{2\sigma^2}-D(Q\|P)\right\}\le 0$$

(optimize over λ) $\blacksquare$

Tensorization of Transportation Inequalities

- At first sight, all we have is another equivalent characterization of concentration of Lipschitz functions.
- However, transportation inequalities tensorize!!
- Proof of tensorization is through a beautiful result on couplings by Katalin Marton.

The Marton Coupling

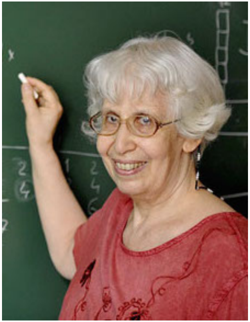
Theorem (Marton, 1986). Let (X_i, P_i), $1 \le i \le n$, be probability spaces. Let $w_i : \mathsf{X}_i \times \mathsf{X}_i \to \mathbb{R}_+$, $1 \le i \le n$, be positive weight functions, and let $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ be a convex function. Suppose that, for each i,

$$\inf_{\pi \in \Pi(P_i, Q)} \varphi\left(\mathbb{E}_\pi[w_i(X,Y)]\right) \le 2\sigma^2 D(Q \| P_i), \ \forall Q.$$

Then the following holds for the product measure $P = P_1 \otimes \ldots \otimes P_n$ on the product space $\mathsf{X} = \mathsf{X}_1 \otimes \ldots \otimes \mathsf{X}_n$:

$$\inf_{\pi \in \Pi(P, Q)} \sum_{i=1}^n \varphi\left(\mathbb{E}_\pi[w_i(X_i, Y_i)]\right) \le 2\sigma^2 D(Q \| P)$$

for every Q on X.



Katalin Marton

Proof idea: Chain rule for relative entropy + conditional coupling + induction.

Tensorization of Transportation Cost Inequalities

Theorem. Let (X_i, P_i, d_i), $1 \le i \le n$, be probability metric spaces. If for some $1 \le p \le 2$ each P_i satisfies $T_p(c)$ on (X_i, d_i), then the product measure $P = P_1 \otimes \ldots \otimes P_n$ on $\mathsf{X} = \mathsf{X}_1 \times \ldots \times \mathsf{X}_n$ satisfies $T_p(cn^{2/p-1})$ w.r.t. the metric

$$d_p(x^n, y^n) \triangleq \left(\sum_{i=1}^{n} d_i^p(x_i, y_i) \right)^{1/p}.$$

Remarks:

- If each P_i satisfies $T_1(c)$, then $P = P_1 \otimes \ldots \otimes P_n$ satisfies $T_1(cn)$ with respect to the metric $\sum_i d_i$. Note: constant deteriorates with n.
- If each P_i satisfies $T_2(c)$, then P satisfies $T_2(c)$ with respect to $\sqrt{\sum_i d_i^2}$. Note: constant is independent of n.

Proof of Tensorization

- By hypothesis, for each i,

$$\underbrace{\inf_{\pi\in\Pi(P_i,Q)} \left(\mathbb{E}_\pi[d_i^p(X,Y)]\right)^{2/p}}_{W_{p,d_i}^2(P_i,Q)} \leq 2cD(Q\|P_i), \qquad \forall Q.$$

- $1 \leq p \leq 2 \implies \varphi(u) = u^{2/p}$ is convex. Take $w_i = d_i^p$. Then

$$\inf_{\pi\in\Pi(P_i,Q)} \varphi\left(\mathbb{E}_\pi[w_i(X,Y)]\right) \leq 2cD(Q\|P_i), \qquad \forall Q.$$

- By Marton's coupling,

$$(\star) \qquad \inf_{\pi\in\Pi(P,Q)} \sum_{i=1}^{n} \left(\mathbb{E}_\pi[d_i^p(X_i,Y_i)]\right)^{2/p} \leq 2cD(Q\|P), \qquad \forall Q.$$

Proof of Tensorization (cont.)

- We have shown that if P_i satisfies $T_p(c)$ w.r.t. d_i, for each i, then

$$(\star) \qquad \inf_{\pi \in \Pi(P,Q)} \sum_{i=1}^{n} \left(\mathbb{E}_\pi[d_i^p(X_i, Y_i)]\right)^{2/p} \le 2cD(Q\|P), \qquad \forall Q.$$

- We will now prove LHS$(\star) \ge n^{1-2/p} W_{p,d_p}^2(P,Q)$.
- For any $\pi \in \Pi(P,Q)$,

$$\begin{aligned}
&\left(\mathbb{E}_\pi\left[\sum_{i=1}^{n} d_i^p(X_i, Y_i)\right]\right)^{2/p} \\
&\qquad \le \left(\sum_{i=1}^{n} \mathbb{E}_\pi[d_i^p(X_i, Y_i)]\right)^{2/p} \qquad \text{(concavity of } t \mapsto t^{1/p}\text{)} \\
&\qquad \le n^{2/p-1} \sum_{i=1}^{n} \left(\mathbb{E}_\pi[d_i^p(X_i, Y_i)]\right)^{2/p} \qquad \text{(convexity of } t \mapsto t^{2/p}\text{)}
\end{aligned}$$

Take infimum over all $\pi \in \Pi(P,Q)$, and we are done. ■

(Yet Another) Proof of McDiarmid's Inequality

- Product probability space: $(\mathsf{X}_1 \times \ldots \times \mathsf{X}_n, P_1 \otimes \ldots \otimes P_n)$
- Given $c_1, \ldots, c_n \geq 0$, equip X_i with $d_{c_i}(x_i, y_i) \triangleq c_i \mathbf{1}\{x_i \neq y_i\}$.
- Then any P_i on X_i satisfies $T_1(c_i^2/4)$:

$$W_{1,d_{c_i}}(P_i, Q) \equiv c_i \|P_i - Q\|_{\mathrm{TV}} \leq \sqrt{\frac{c_i^2}{2} D(Q\|P_i)}$$

  (by rescaling Pinsker).
- By Marton coupling, $P = P_1 \otimes \ldots \otimes P_n$ satisfies T_1 with constant $(1/4)\sum_{i=1}^n c_i^2$ with respect to the metric

$$d_{\mathbf{c}}(x^n, y^n) \triangleq \sum_{i=1}^n d_{c_i}(x_i, y_i) = \sum_{i=1}^n c_i \mathbf{1}\{x_i \neq y_i\}.$$

- By Bobkov–Götze, this is equivalent to subgaussian property of all $f \in \mathrm{Lip}_1(\mathsf{X}_1 \times \ldots \times \mathsf{X}_n, d_{\mathbf{c}})$:

$$\underbrace{\mathbb{P}[|f(X^n) - \mathbb{E}f(X^n)| \geq t] \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right)}_{\text{McDiarmid}}, \quad \forall\, \underbrace{f \in \mathrm{Lip}_1(d_{\mathbf{c}})}_{\text{bdd. diff.}}$$

Some Applications

The Blowing-Up Lemma

- Consider a product space Y^n equipped with Hamming metric $d(y^n, z^n) = \sum_{i=1}^n \mathbf{1}\{y_i \neq z_i\}$
- For a set $A \subseteq \mathsf{Y}^n$ and for $r \in \{0, 1, \ldots, n\}$, define its r-blowup

$$[A]_r \triangleq \left\{ z^n \in \mathsf{Y}^n : \min_{y^n \in A} d(z^n, y^n) \leq r \right\}$$

The following result, in a different (asymptotic) form was first proved by Ahlswede–Gács–Körner (1976); a simple proof was given by Marton (1986):

Let $Y_1, \ldots, Y_n$ be independent r.v.'s taking values in Y. Then for every set $A \subseteq \mathsf{Y}^n$ with $P_{Y^n}(A) > 0$

$$P_{Y^n}\left\{[A]_r\right\} \geq 1 - \exp\left[-\frac{2}{n} \left(r - \sqrt{\frac{n}{2} \log \frac{1}{P_{Y^n}(A)}} \right)_+^2 \right]$$

Informally, any set in a product space can be "blown up" to engulf most of the probability mass.

Marton's Proof

- Let $P_i = \mathcal{L}(Y_i)$, $1 \le i \le n$; $P = P_1 \otimes \ldots \otimes P_n \equiv \mathcal{L}(Y^n)$.
- By tensorization, P satisfies the TC inequality

$$(\star) \quad W_1(P, Q) \le \sqrt{\frac{n}{2} D(Q \| P)}, \qquad \forall Q \in \mathcal{P}(\mathsf{Y}^n),$$

where

$$W_1(P, Q) = \inf_{\pi \in \Pi(P,Q)} \mathbb{E}_\pi \left[\sum_{i=1}^n \mathbf{1}\{Y_i \neq Z_i\} \right], \qquad Y^n \sim P,\ Z^n \sim Q$$

- For any $B \subseteq \mathsf{Y}^n$ with $P(B) > 0$, consider conditional distribution

$$P_B(\cdot) \triangleq \frac{P(\cdot \cap B)}{P(B)}.$$

Then $D(P_B \| P) = \log \frac{1}{P(B)}$, and therefore

$$(\star\star) \quad W_1(P, P_B) \le \sqrt{\frac{n}{2} \log \frac{1}{P(B)}}.$$

Marton's Proof

- $(\star\star)$ $W_1(P, P_B) \le \sqrt{\frac{n}{2}\log\frac{1}{P(B)}}$
- Apply $(\star\star)$ to $B = A$ and $B = [A]_r^c$:

$$W_1(P, P_A) \le \sqrt{\frac{n}{2}\log\frac{1}{P(A)}}, \quad W_1(P, P_{[A]_r^c}) \le \sqrt{\frac{n}{2}\log\frac{1}{1 - P([A]_r)}}$$

- Then

$$\begin{aligned}
\sqrt{\frac{n}{2}\log\frac{1}{P(A)}} + \sqrt{\frac{n}{2}\log\frac{1}{1 - P([A]_r)}} &\ge W_1(P_A, P) + W_1(P, P_{[A]_r^c}) \\
&\ge W_1(P_A, P_{[A]_r^c}) \quad \text{(triangle ineq.)} \\
&\ge \min_{y^n \in A, z^n \in [A]_r^c} d(y^n, z^n) \\
&\ge r. \quad \text{(def. of } [\cdot]_r\text{)}
\end{aligned}$$

- Rearrange to finish the proof. $\blacksquare$

The Blowing-Up Lemma: Consequences

- Consider a DMC $(\mathsf{X}, \mathsf{Y}, T)$ with input alphabet X, output alphabet Y, transition probabilities $T(y|x), (x, y) \in \mathsf{X} \times \mathsf{Y}$
- (n, M, ε)-code $\mathcal{C}$: encoder $f : \{1, \ldots, M\} \to \mathsf{X}^n$, decoder $g : \mathsf{Y}^n \to \{1, \ldots, M\}$ with

$$\max_{1 \le j \le M} \mathbb{P}[g(Y^n) \neq j | X^n = f(j)] \le \varepsilon.$$

  Equivalently, $\mathcal{C} = \{(u_j, D_j)\}_{j=1}^M$, where
  - $u_j = f(j) \in \mathsf{X}^n$ — codewords
  - $D_j = g^{-1}(j) = \{y^n \in \mathsf{Y}^n : g(y^n) = j\}$ — decoding sets

$$T^n(D_j|u_j) \ge 1 - \varepsilon, \qquad j = 1, \ldots, M.$$

Lemma. There exists $\delta_n > 0$, such that

$$T^n\Big([D_j]_{n\delta_n} \Big| X^n = u_j \Big) \ge 1 - \frac{1}{n}, \qquad j = 1, \ldots, M.$$

Proof

- Choose

$$\delta_n = \frac{1}{n}\left\lceil n\left(\sqrt{\frac{\log n}{2n}} + \sqrt{\frac{1}{2n}\log\frac{1}{1-\varepsilon}}\right)\right\rceil$$

- For each j, apply Blowing-Up Lemma to the product measure

$$P_j(y^n) = \prod_{i=1}^{n} T(y_i|u_j(i)), \quad \text{where } \underbrace{u_j = (u_j(1), \ldots, u_j(n)) = f(j)}_{j\text{th codeword}}.$$

With $r = n\delta_n$, this gives

$$T^n\Big([D_j]_{n\delta_n} \Big| X^n = u_j \Big) \geq 1 - \exp\left[-\frac{2}{n}\left(n\delta_n - \sqrt{\frac{n}{2}\log\frac{1}{1-\varepsilon}}\right)_+^2\right]$$

$$\geq 1 - \frac{1}{n}. \qquad \blacksquare$$

From Blowing-Up Lemma to Strong Converses

- Informally, the Blowing-Up Lemma shows that “any bad code contains a good subcode” (Ahlswede and Dueck, 1976).
- Consider an (n, M, ε)-code $\mathcal{C} = \{(u_j, D_j)\}_{j=1}^{M}$.
- Each decoding set D_j can be “blown up” to a set $\tilde{D}_j \subseteq \mathsf{Y}^n$ with

$$T^n(\tilde{D}_j | u_j) \geq 1 - \frac{1}{n}.$$

- The object $\tilde{\mathcal{C}} = \{(u_j, \tilde{D}_j)\}_{j=1}^{M}$ is not a code (since the sets $\tilde{D}_j$ are no longer disjoint), but a random coding argument can be used to extract an (n, M', ε') “subcode” with M' slightly smaller than M and $\varepsilon' < \varepsilon$. Then one can apply the usual (weak) converse to the subcode.
- Similar ideas can be used in multiterminal settings (starting with Ahlswede–Gács–Körner).

Example: Capacity-Achieving Channel Codes

The set-up

- DMC $(\mathsf{X}, \mathsf{Y}, T)$ with capacity

$$C = C(T) = \max_{P_X} I(X;Y)$$

- (n, M)-code: $\mathcal{C} = (f, g)$ with encoder $f : \{1, \dots, M\} \to \mathsf{X}^n$ and decoder $g : \mathsf{Y}^n \to \{1, \dots, M\}$

Capacity-achieving codes:

A sequence $\{\mathcal{C}_n\}_{n=1}^{\infty}$, where each $\mathcal{C}_n$ is an (n, M_n)-code, is capacity-achieving if

$$\lim_{n\to\infty} \frac{1}{n} \log M_n = C.$$

Capacity-Achieving Channel Codes

Capacity-achieving input and output distributions:

$$P_X^* \in \arg\max_{P_X} I(X;Y) \qquad \text{(may not be unique)}$$

$$P_X^* \xrightarrow{\quad T \quad} P_Y^* \qquad \text{(always unique)}$$

Theorem (Shamai–Verdú, 1997). Let $\{\mathcal{C}_n\}$ be any capacity-achieving code sequence with vanishing error probability. Then

$$\lim_{n\to\infty} \frac{1}{n} D\left(P_{Y^n}^{(\mathcal{C}_n)} \middle\| P_{Y^n}^*\right) = 0,$$

where $P_{Y^n}^{(\mathcal{C}_n)}$ is the output distribution induced by the code $\mathcal{C}_n$ when the messages in $\{1, \ldots, M_n\}$ are equiprobable.

Capacity-Achieving Channel Codes

$$\lim_{n\to\infty} \frac{1}{n} D(P_{Y^n} \| P^*_{Y^n}) = 0$$

Main message: channel output sequences induced by good code "resemble" i.i.d. sequences drawn from the CAOD P^*_Y

Useful implications: estimate performance characteristics of good channel codes by their expectations w.r.t. $P^*_{Y^n} = (P^*_Y)^n$

- often much easier to compute explicitly
- bound estimation accuracy using large-deviation theory (e.g., Sanov's theorem)

Question: what about good codes with nonvanishing error probability?

Codes with Nonvanishing Error Probability

Y. Polyanskiy and S. Verdú, "Empirical distribution of good channel codes with non-vanishing error probability" (2012)

1. Let $\mathcal{C} = (f, g)$ be any (n, M, ε)-code for T:

$$\max_{1 \le j \le M} \mathbb{P}\left[g(Y^n) \neq j \middle| X^n = f(j) \right] \le \varepsilon.$$

   Then $D\big(P_{Y^n}^{(\mathcal{C})} \big\| P_{Y^n}^{*}\big) \le nC - \log M + o(n)$.*

2. If $\{\mathcal{C}_n\}_{n=1}^{\infty}$ is a capacity-achieving sequence, where each $\mathcal{C}_n$ is an (n, M_n, ε)-code for some fixed $\varepsilon > 0$, then

$$\lim_{n \to \infty} \frac{1}{n} D\left(P_{Y^n}^{(\mathcal{C}_n)} \middle\| P_{Y^n}^{*} \right) = 0.$$

* In some cases, the $o(n)$ term can be improved to $O(\sqrt{n})$.

Relative Entropy at the Output of a Code

Consider a DMC $T : \mathsf{X} \to \mathsf{Y}$ with $T(\cdot|\cdot) > 0$, and let

$$c(T) = 2 \max_{x \in \mathsf{X}} \max_{y, y' \in \mathsf{Y}} \left| \log \frac{T(y|x)}{T(y'|x)} \right|$$

Theorem (Raginsky–Sason, 2013). Any (n, M, ε)-code $\mathcal{C}$ for T, where $\varepsilon \in (0, 1/2)$, satisfies

$$D\Big(P^{(\mathcal{C})}_{Y^n} \Big\| P^*_{Y^n}\Big) \le nC - \log M + \log \frac{1}{\varepsilon} + c(T) \sqrt{\frac{n}{2} \log \frac{1}{1 - 2\varepsilon}}$$

Remark:

- Polyanskiy and Verdú show that

$$D\Big(P^{(\mathcal{C})}_{Y^n} \Big\| P^*_{Y^n}\Big) \le nC - \log M + a\sqrt{n}$$

  for some constant $a = a(\varepsilon)$.

Proof Sketch (1)

Fix $x^n \in \mathsf{X}^n$ and study concentration of the function

$$h_{x^n}(y^n) = \log \frac{\mathrm{d}P_{Y^n|X^n=x^n}}{\mathrm{d}P^{(\mathcal{C})}_{Y^n}}(y^n)$$

around its expectation w.r.t. $P_{Y^n|X^n=x^n}$:

$$\mathbb{E}[h_{x^n}(Y^n)|X^n = x^n] = D\Big(P_{Y^n|X^n=x^n}\Big\|P^{(\mathcal{C})}_{Y^n}\Big)$$

Step 1: Because $T(\cdot|\cdot) > 0$, the function $h_{x^n}(y^n)$ is 1-Lipschitz w.r.t. scaled Hamming metric

$$d(y^n, \bar{y}^n) = c(T) \sum_{i=1}^{n} \mathbf{1}\{y_i \neq \bar{y}_i\}$$

Proof Sketch (2)

Step 1: Because $T(\cdot|\cdot) > 0$, the function $h_{x^n}(y^n)$ is 1-Lipschitz w.r.t. scaled Hamming metric

$$d(y^n, \bar{y}^n) = c(T) \sum_{i=1}^{n} \mathbf{1}\{y_i \neq \bar{y}_i\}$$

Step 2: Any product probability measure μ on (Y^n, d) satisfies

$$\log \mathbb{E}_\mu \left[e^{tf(Y^n)} \right] \leq \frac{nc(T)^2 t^2}{8}$$

for any f with $\mathbb{E}_\mu f = 0$ and $\|F\|_{\mathrm{Lip}} \leq 1$.

Proof: Tensorization of T_1 (Pinsker), followed by appeal to Bobkov–Götze.

Proof Sketch (3)

$$h_{x^n}(y^n) = \log \frac{\mathrm{d}P_{Y^n|X^n=x^n}}{\mathrm{d}P_{Y^n}^{(\mathcal{C})}}(y^n)$$

$$\mathbb{E}[h_{x^n}(Y^n)|X^n = x^n] = D\left(P_{Y^n|X^n=x^n} \middle\| P_{Y^n}^{(\mathcal{C})}\right)$$

Step 3: For any x^n, $\mu = P_{Y^n|X^n=x^n}$ is a product measure, so

$$\mathbb{P}\left(h_{x^n}(Y^n) \geq D\left(P_{Y^n|X^n=x^n} \middle\| P_{Y^n}^{(\mathcal{C})}\right) + r\right) \leq \exp\left(-\frac{2r^2}{nc(T)^2}\right)$$

Use this with $r = c(T)\sqrt{\frac{n}{2}\log\frac{1}{1-2\varepsilon}}$:

$$\mathbb{P}\left(h_{x^n}(Y^n) \geq D\left(P_{Y^n|X^n=x^n} \middle\| P_{Y^n}^{(\mathcal{C})}\right) + c(T)\sqrt{\frac{n}{2}\log\frac{1}{1-2\varepsilon}}\right) \leq 1 - 2\varepsilon$$

Remark: Polyanskiy–Verdú show $\mathrm{Var}[h_{x^n}(Y^n)|X^n = x^n] = O(n)$.

Proof Sketch (4)

Recall:

$$\mathbb{P}\left(h_{x^n}(Y^n) \geq D\left(P_{Y^n|X^n=x^n} \middle\| P_{Y^n}^{(\mathcal{C})}\right) + c(T)\sqrt{\frac{n}{2}\log\frac{1}{1-2\varepsilon}} \right) \leq 1-2\varepsilon$$

Step 4: Same as Polyanskiy–Verdú, appeal to Augustin's strong converse (1966) to get

$$\log M \leq \log\frac{1}{\varepsilon} + D\left(P_{Y^n|X^n} \middle\| P_{Y^n}^{(\mathcal{C})} \middle| P_{X^n}^{(\mathcal{C})}\right) + c(T)\sqrt{\frac{n}{2}\log\frac{1}{1-2\varepsilon}}$$

$$\begin{aligned} &D\left(P_{Y^n}^{(\mathcal{C})} \middle\| P_{Y^n}^{*}\right) \\ &\quad = D\left(P_{Y^n|X^n} \middle\| P_{Y^n}^{*} \middle| P_{X^n}^{(\mathcal{C})}\right) - D\left(P_{Y^n|X^n} \middle\| P_{Y^n}^{(\mathcal{C})} \middle| P_{X^n}^{(\mathcal{C})}\right) \\ &\quad \leq nC - \log M + \log\frac{1}{\varepsilon} + c(T)\sqrt{\frac{n}{2}\log\frac{1}{1-2\varepsilon}} \qquad \blacksquare \end{aligned}$$

Relative Entropy at the Output of a Code

Theorem (Raginsky–Sason, 2013). Let $(\mathsf{X}, \mathsf{Y}, T)$ be a DMC with $C > 0$. Then, for any $0 < \varepsilon < 1$, any (n, M, ε)-code $\mathcal{C}$ for T satisfies

$$
\begin{aligned}
& D\big(P^{(\mathcal{C})}_{Y^n} \big\| P^*_{Y^n}\big) \le nC - \log M \\
& \quad + \sqrt{2n}\,(\log n)^{3/2} \left(1 + \sqrt{\frac{1}{\log n}\,\log\left(\frac{1}{1-\varepsilon}\right)}\right) \left(1 + \frac{\log|\mathsf{Y}|}{\log n}\right) \\
& \quad + 3\log n + \log\big(2|\mathsf{X}||\mathsf{Y}|^2\big).
\end{aligned}
$$

Proof idea:

- Apply Blowing-Up Lemma to the code, then extract a good subcode.

Remark:

- Polyanskiy and Verdú show that

$$
D\big(P^{(\mathcal{C})}_{Y^n} \big\| P^*_{Y^n}\big) \le nC - \log M + b\sqrt{n}\log^{3/2} n
$$

for some constant $b > 0$.

Concentration of Lipschitz Functions

Theorem (Raginsky–Sason, 2013). Let $(\mathsf{X}, \mathsf{Y}, T)$ be a DMC with $c(T) < \infty$. Let $d : \mathsf{Y}^n \times \mathsf{Y}^n \to \mathbb{R}_+$ be a metric, and suppose that $P_{Y^n|X^n=x^n}$, $x^n \in \mathsf{X}^n$, as well as $P^*_{Y^n}$, satisfy $\mathrm{T}_1(c)$ for some $c > 0$.

Then, for any $\varepsilon \in (0, 1/2)$, any (n, M, ε)-code $\mathcal{C}$ for T, and any function $f : \mathsf{Y}^n \to \mathbb{R}$ we have

$$
\begin{aligned}
& P^{(\mathcal{C})}_{Y^n}\Big(|f(Y^n) - \mathbb{E}[f(Y^{*n})]| \ge r \Big) \\
& \le \frac{4}{\varepsilon} \, \exp\left(nC - \ln M + a\sqrt{n} - \frac{r^2}{8c\|f\|^2_{\mathrm{Lip}}} \right), \; \forall \, r \ge 0
\end{aligned}
$$

where $Y^{*n} \sim P^*_{Y^n}$, and $a \triangleq c(T)\sqrt{\frac{1}{2}\ln\frac{1}{1-2\varepsilon}}$.

Proof Sketch

Step 1: For each $x^n \in \mathsf{X}^n$, let $\phi(x^n) \triangleq \mathbb{E}[f(Y^n)|X^n = x^n]$. Then, by Bobkov–Götze,

$$\mathbb{P}\Big(|f(Y^n) - \phi(x^n)| \geq r \Big| X^n = x^n \Big) \leq 2 \exp\left(- \frac{r^2}{2c\|f\|^2_{\mathrm{Lip}}} \right)$$

Step 2: By restricting to a subcode $\mathcal{C}'$ with codewords $x^n \in \mathsf{X}^n$ satisfying $\phi(x^n) \geq \mathbb{E}[f(Y^{*n})] + r$, we can show that

$$r \leq \|f\|_{\mathrm{Lip}} \sqrt{2c\left(nC - \log M' + a\sqrt{n} + \log \frac{1}{\varepsilon}\right)},$$

with $M' = M P^{(\mathcal{C})}_{X^n}\Big(\phi(X^n) \geq \mathbb{E}[f(Y^{*n})] + r\Big)$. Solve to get

$$P^{(\mathcal{C})}_{X^n}\Big(|\phi(X^n) - \mathbb{E}[f(Y^{*n})]| \geq r \Big) \leq 2e^{nC - \log M + a\sqrt{n} + \log\frac{1}{\varepsilon} - \frac{r^2}{2c\|f\|^2_{\mathrm{Lip}}}}$$

Step 3: Apply union bound. ■

Empirical Averages at the Code Output

- Equip Y^n with the Hamming metric

$$d(y^n, \bar{y}^n) = \sum_{i=1}^{n} \mathbf{1}\{y_i \neq \bar{y}_i\}$$

- Consider functions of the form

$$f(y^n) = \frac{1}{n} \sum_{i=1}^{n} f_i(y_i),$$

  where $|f_i(y_i) - f_i(\bar{y}_i)| \le L\mathbf{1}\{y_i \neq \bar{y}_i\}$ for all $i, y_i, \bar{y}_i$. Then $\|f\|_{\mathrm{Lip}} \le L/n$.

- Since $P_{Y^n|X^n=x^n}$ for all x^n and $P^*_{Y^n}$ are product measures on Y^n, they all satisfy $\mathrm{T}_1(n/4)$ (by tensorization)

- Therefore, for any (n, M, ε)-code and any such f we have

$$P^{(\mathcal{C})}_{Y^n}\big(|f(Y^n) - \mathbb{E}[f(Y^{*n})]| \ge r\big) \le \frac{4}{\varepsilon} \exp\left(nC - \log M + a\sqrt{n} - \frac{nr^2}{2L^2}\right)$$

Concentration of Measure

Information-Theoretic Converse

- Concentration phenomenon in a nutshell: if a subset of a metric probability space does not have too small of a probability mass, then its blowups will eventually take up most of the probability mass.
- Question: given a set whose blowups eventually take up most of the probability mass, how small can this set be?

This question was answered by Kontoyiannis (1999) as a consequence of a general information-theoretic converse.

Converse Concentration of Measure: The Set-Up

- Let X be a finite set, together with a distortion function $d : \mathsf{X} \times \mathsf{X} \to \mathbb{R}_+$ and a mass function $M : \mathsf{X} \to (0, \infty)$.
- Extend to product space X^n:

$$d_n(x^n, y^n) \triangleq \sum_{i=1}^{n} d(x_i, y_i)$$

$$M_n(x^n) \triangleq \prod_{i=1}^{n} M(x_i)$$

$$M_n(A) \triangleq \sum_{x^n \in A} M_n(x^n), \qquad \forall A \subseteq \mathsf{X}^n$$

- Blowups:

$$A \subseteq \mathsf{X}^n \qquad \longrightarrow \qquad [A]_r \triangleq \left\{ x^n \in \mathsf{X}^n : \min_{y^n \in A} d_n(x^n, y^n) \le r \right\}$$

Converse Concentration of Measure

- Let P be a probability measure on X. Define

$$R_n(\delta) \triangleq \min_{P_{X^n Y^n}} \Big\{ I(X^n; Y^n) + \mathbb{E} \log M_n(Y^n) : \\ P_{X^n} = P^{\otimes n},\ \mathbb{E}[d_n(X^n, Y^n)] \le n\delta \Big\}$$

Theorem (Kontoyiannis). Let $A_n \subseteq \mathsf{X}^n$ be an arbitrary set. Then

$$\frac{1}{n} \log M_n(A_n) \ge R(\delta),$$

where

$$\delta \triangleq \frac{1}{n} \mathbb{E} \left[\min_{y^n \in A_n} d_n(X^n, y^n) \right] \quad \text{and} \quad R(\delta) \triangleq \lim_{n \to \infty} \frac{R_n(\delta)}{n} \equiv R_1(\delta).$$

Remark:

- It can be shown that $R_1(\delta) = \inf_{n \ge 1} \frac{R_n(\delta)}{n}$.

Proof

- Define the mapping $\varphi_n : \mathsf{X}^n \to \mathsf{X}^n$ via

$$\varphi_n(x^n) \triangleq \arg\min_{y^n \in A_n} d_n(x^n, y^n)$$

  and let $Y^n = \varphi_n(X^n)$, $Q_n = \mathcal{L}(Y^n)$.

- Then

$$\begin{aligned}
\log M_n(A_n) &= \log \sum_{y^n \in A_n} M_n(y^n) \\
&\geq \log \sum_{y^n \in A_n :\, Q_n > 0} Q_n(y^n) \frac{M_n(y^n)}{Q_n(y^n)} \\
&\geq \sum_{y^n \in A_n} Q_n(y^n) \log \frac{M_n(y^n)}{Q_n(y^n)} \\
&= -\sum_{y^n \in A_n} Q_n(y^n) \log Q_n(y^n) + \sum_{y^n \in A_n} Q(y^n) \log M_n(y^n) \\
&= H(Y^n) + \mathbb{E} \log M(Y^n) \\
&= I(X^n; Y^n) + \mathbb{E} \log M(Y^n) \\
&\geq R_n(\delta).
\end{aligned}$$

■

Converse Concentration of Measure

- Consider a sequence of sets $\{A_n\}_{n=1}^{\infty}$ with

$$(\star) \qquad P^{\otimes n}\left([A_n]_{n\delta}\right) \xrightarrow{n\to\infty} 1.$$

- Apply Kontoyiannis' converse to the mass function $M = P$, to get the following:

Corollary. If the sequence $\{A_n\}$ satisfies $(\star)$, then

$$\liminf_{n\to\infty} \frac{1}{n} \log P^{\otimes n}(A_n) \geq R(\delta),$$

where the "concentration exponent" is

$$\begin{aligned} R(\delta) &= \min_{P_{XY}} \Big\{ I(X;Y) + \mathbb{E} \log P(Y) :\ P_X = P,\ \mathbb{E}[d(X,Y)] \leq \delta \Big\} \\ &\equiv -\max_{P_{XY}} \Big\{ H(Y|X) + D(P_Y \| P) :\ P_X = P,\ \mathbb{E}[d(X,Y)] \leq \delta \Big\}. \end{aligned}$$

Example of The Concentration Exponent

Theorem (Raginsky–Sason, 2013). Let $P = \mathrm{Bern}(p)$. Then

$$R(\delta) \begin{cases} \le -\varphi(p)\delta^2 - (1-p)h\left(\frac{\delta}{1-p}\right), & \text{if } \delta \in [0, 1-p] \\ = \log p, & \text{if } \delta \in [1-p, 1] \end{cases}$$

where

$$\varphi(p) = \frac{1}{1-2p} \log \frac{1-p}{p}$$

and $h(\cdot)$ is the binary entropy function.

Remarks:

- The upper bound is not tight, but in this case $R(\delta)$ can be evaluated numerically (cf. Kontoyiannis, 2001).
- The proof is by a coupling argument.

Summary

- Three related methods for obtaining sharp concentration inequalities in high dimension:
  1. The entropy method
  2. Log-Sobolev inequalities
  3. Transportation-cost inequalities
- All three methods crucially rely on *tensorization*:
  - Breaking the original high-dimensional problem into low-dimensional pieces, exploiting low-dimensional structure to control entropy locally, assembling local information into a global bound.
- Tensorization is a consequence of *independence*.
- Applications to information theory:
  - Exploit the problem structure to isolate independence (e.g., output distribution of a DMC for any fixed input block).

What We Had to Skip

- Log-Sobolev inequalities and hypercontractivity
- Log-Sobolev inequalities when Herbst fails (e.g., Poisson measures)
- Connections to isoperimetric inequalities
- HWI inequalities: tying together relative entropy, Wasserstein distance, Fisher information
- Concentration inequalities for functions of dependent random variables

For this and more, consult our monograph: M. Raginsky and I. Sason, *Concentration of Measure Inequalities in Info. Theory, Comm. and Coding*, FnT, 2nd edition, 2014.

Recent Books and Surveys - Concentration Inequalities

1. N. Alon and J. H. Spencer, *The Probabilistic Method*, Wiley, 3rd edition, 2008.
2. S. Boucheron, G. Lugosi, and P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*, Oxford Press, 2013.
3. F. Chung and L. Lu, *Complex Graphs and Networks*, vol. 106, Regional Conference Series in Mathematics, Wiley, 2006.
4. D. P. Dubhashi and A. Panconesi, *Concentration of Measure for the Analysis of Randomized Algorithms*, Cambridge Press, 2009.
5. M. Ledoux, *The Concentration of Measure Phenomenon*, Mathematical Surveys and Monographs, vol. 89, AMS, 2001.
6. C. McDiarmid, "Concentration," *Probabilistic Methods for Algorithmic Discrete Mathematics*, pp. 195–248, Springer, 1998.
7. M. Raginsky and I. Sason, *Concentration of Measure Inequalities in Info. Theory, Comm. and Coding*, FnT, 2nd edition, 2014.
8. R. van Handel, *Probability in High Dimension*, lecture notes, 2014.