

Contents

1. The Chernoff-Cramer method	3
1.1. The Chernoff bound and Cramer transform	3
1.2. Hoeffding's and related inequalities	7
2. The variance method	9
2.1. The Efron-Stein inequality	9
2.2. Functions with bounded differences	12
3. Poincaré inequalities	13
3.1. Poincaré constant	17
4. The entropy method	19
4.1. Entropy, chain rules and Han's inequality	19
4.2. Herbst's argument	25
4.3. Log-Sobolev inequalities on the hypercube	27
4.4. The modified log-Sobolev inequality (MLSI)	31
4.5. Concentration of convex Lipschitz functions	33
5. The transport method	33
5.1. Concentration via Marton's argument	36
5.2. Talagrand's inequality	41
6. Log-concave random variables	45
6.1. One-dimensional log-concave random variables	47

Question: toss a fair coin $n = 10000$ times. How many heads?

$X = \sum_{i=1}^n X_i$, $X_i \sim \text{Bern}(1/2)$. $\mathbb{E}[X] = 5000$. But $\mathbb{P}(X = 5000) = \binom{10^4}{5000} \cdot 2^{-10^4} \approx 0.008$.

By WLLN, $\mathbb{P}(X \in [5000 - n\varepsilon, 5000 + n\varepsilon]) \approx 1$.

Theorem 0.1 (Central Limit Theorem) Let X_1, \dots, X_n be IID RVs with mean $\mathbb{E}[X_1] = \mu$. Let $\text{Var}(X_1) = \sigma^2 < \infty$. Then $\frac{1}{\sigma\sqrt{n}} \sum_{i=1}^n (X_i - \mu) \xrightarrow{D} N(0, 1)$, i.e.

$$\mathbb{P}\left(\frac{1}{\sigma\sqrt{n}} \sum_{i=1}^n (X_i - \mu) \in A\right) \rightarrow \int_A \frac{1}{\sqrt{2n}} e^{-x^2/2} dx$$

for all A .

So $\mathbb{P}\left(X \in \left[\frac{n}{2} - \frac{\sqrt{n}}{2} Q^{-1}(\delta), \frac{n}{2} + \frac{\sqrt{n}}{2} Q^{-1}(\delta)\right]\right) \geq 1 - \delta$, for n large enough, where $Q(\delta) = \int_{\delta}^{\infty} \frac{1}{\sqrt{2n}} e^{-x^2/2} dx$. We have $Q^{-1}(x) \propto \sqrt{\log \frac{1}{x}}$. So interval has length $\propto \sqrt{n} \sqrt{\log \frac{1}{\delta}}$.

Theorem 0.2 (Chebyshev's Inequality) $\mathbb{P}(|X - \mu| \geq \varepsilon) \leq \frac{\text{Var}(X)}{\varepsilon^2}$ for all $\varepsilon > 0$.

Corollary 0.3 $\mathbb{P}\left(\left|\sum_{i=1}^n (X_i) - \frac{n}{2}\right| \geq t\right) \leq \frac{\text{Var}(\sum_{i=1}^n X_i)}{t^2} = n \frac{\sigma^2}{t^2} \leq \delta$ where $t = \sqrt{n}\sigma/\sqrt{\delta}$.

So $\mathbb{P}(X \in [\frac{n}{2} - \frac{n}{2}\sqrt{\delta}, \frac{n}{2} + \frac{n}{2}\sqrt{\delta}]) \geq 1 - \delta$.

Question 2: we have N coupons. Each day receive one uniformly at random independent of the past. How many days until all coupons received?

We have $X = \sum_{i=1}^n X_i$, $X_i \sim \text{Geom}(\frac{1}{n})$. $\mathbb{E}[X] = \sum_i \mathbb{E}[X_i] \approx n \log n$ (verify this).

Question 3: Let $(X_1, \dots, X_n), (Y_1, \dots, Y_n)$ be IID. What is the longest common subsequence, i.e. $f(X_1, \dots, X_n, Y_1, \dots, Y_n) = \max\{k : \exists i_1, \dots, i_k, j_1, \dots, j_k \text{ s.t. } X_{i_j} = Y_{j_j} \forall j \in [k]\}$. Computing f is NP-hard. f is smooth.

Principle: a smooth function of many independent random variables concentrates around its mean.

Theorem 0.4 (Law of Total Expectation) We have $\mathbb{E}_Y[\mathbb{E}_X[X | Y]] = \mathbb{E}_X[X]$.

Theorem 0.5 (Tower Property of Conditional Expectation) We have $\mathbb{E}[\mathbb{E}[Z | X, Y] | Y] = \mathbb{E}[Z | Y]$.

Theorem 0.6 We have $\mathbb{E}[f(Y)X | Y] = f(Y)\mathbb{E}[X | Y]$.

Theorem 0.7 (Holder's Inequality) Let $p \geq 1$ and $1/p + 1/q = 1$. Then

$$\mathbb{E}[|XY|] \leq \mathbb{E}[|X|^p]^{1/p} \cdot \mathbb{E}[|Y|^q]^{1/q}.$$

Theorem 0.8 (Cauchy-Schwarz) A special case of Holder's inequality:

$$\mathbb{E}[|XY|] \leq \mathbb{E}[X^2]^{1/2} \cdot \mathbb{E}[Y^2]^{1/2}.$$

Definition 0.9 The **conditional variance** of Y given X is the random variable

$$\text{Var}(Y | X) := \mathbb{E}[(Y - \mathbb{E}[Y | X])^2 | X].$$

1. The Chernoff-Cramer method

1.1. The Chernoff bound and Cramer transform

Theorem 1.1 (Weak Law of Large Numbers) Let X_1, \dots, X_n be IID RVs with mean $\mathbb{E}[X_1] = \mu$. Then, for all $\varepsilon > 0$,

$$\mathbb{P}\left(\left|\frac{1}{n} \sum_{i=1}^n X_i - \mu\right| > \varepsilon\right) \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Theorem 1.2 (Markov's Inequality) Let Y be a non-negative RV. For any $t \geq 0$,

$$\mathbb{P}(Y \geq t) \leq \frac{\mathbb{E}[Y]}{t}.$$

Proof (Hints). Split Y using indicator variables. □

Proof. We have $Y = Y \cdot \mathbb{I}_{\{Y \geq t\}} + Y \cdot \mathbb{I}_{\{Y < t\}} \geq t \cdot \mathbb{I}_{\{Y \geq t\}}$. Taking expectations gives the result. □

Corollary 1.3 Let $\varphi : \mathbb{R} \rightarrow \mathbb{R}_+$ be non-decreasing, then

$$\mathbb{P}(Y \geq t) \leq \mathbb{P}(\varphi(Y) \geq \varphi(t)) \leq \frac{\mathbb{E}[\varphi(Y)]}{\varphi(t)}.$$

For $\varphi(t) = t^2$, we can use $\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \text{Var}(X_i)$.

Corollary 1.4 (Chebyshev's Inequality) For any RV Y and $t > 0$,

$$\mathbb{P}(|Y - \mathbb{E}[Y]| \geq t) \leq \frac{\text{Var}(Y)}{t^2}.$$

Proof (Hints). Straightforward. □

Proof. Take $Z = |Y - \mathbb{E}[Y]|$ and use Corollary 1.3 with $\varphi(t) = t^2$. □

Exercise 1.5 Prove WLLN, assuming that $\text{Var}(X_1) < \infty$, using Chebyshev's inequality.

Remark 1.6 If higher moments exist, we can use them in a similar way: let $\varphi(t) = t^q$ for $q > 0$, then for all $t > 0$,

$$\mathbb{P}(|Z - \mathbb{E}[Z]| \geq t) \leq \frac{\mathbb{E}[|Z - \mathbb{E}[Z]|^q]}{t^q}.$$

We can then optimise over q to pick the lowest bound on $\mathbb{P}(|Z - \mathbb{E}[Z]| \geq t)$. Note that **Chebyshev's Inequality** is the most popular form of this bound due to the additivity of variance.

Definition 1.7 The **moment generating function (MGF)** of F is

$$F(\lambda) := \mathbb{E}[e^{\lambda Z}] = \sum_{k=0}^{\infty} \frac{\lambda^k \mathbb{E}[Z^k]}{k!}.$$

Definition 1.8 The **log-MGF** of Z is $\psi_Z(\lambda) = \log F(\lambda)$.

Note that $\psi_Z(\lambda)$ is additive: if $Z = \sum_{i=1}^n Z_i$, with Z_1, \dots, Z_n independent, then

$$\psi_Z(\lambda) = \log(\mathbb{E}[e^{\lambda Z}]) = \sum_{i=1}^n \log \mathbb{E}[e^{\lambda Z_i}] = \sum_{i=1}^n \psi_{Z_i}(\lambda).$$

Definition 1.9 The **Cramer transform** of Z is

$$\psi_Z^*(t) = \sup\{\lambda t - \psi_Z(\lambda) : \lambda > 0\}.$$

Proposition 1.10 (Chernoff Bound) Let Z be an RV. For all $t > 0$,

$$\mathbb{P}(Z \geq t) \leq e^{-\psi_Z^*(t)}.$$

Proof (Hints). Use Corollary 1.3. □

Proof. By Corollary 1.3, we have

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

Taking the infimum over all $\lambda > 0$ gives $\mathbb{P}(Z \geq t) \leq \inf\{e^{-(\lambda t - \psi_Z(\lambda))} : \lambda > 0\}$, which gives the result. □

Remark 1.11 Our goal is to obtain an upper bound on $\psi_Z(\lambda)$, as this will give exponential concentration. The function $\psi_{Z - \mathbb{E}[Z]}(\lambda)$ gives upper bounds on $\mathbb{P}(Z - \mathbb{E}[Z] \geq t)$, the function $\psi_{-Z + \mathbb{E}[Z]}(\lambda)$ gives upper bounds on $\mathbb{P}(Z - \mathbb{E}[Z] \leq -t)$.

Proposition 1.12

1. $\psi_Z(\lambda)$ is convex and infinitely differentiable on $(0, b)$, where $b = \sup\{\lambda > 0 : \psi_Z(\lambda) < \infty\}$.
2. $\psi_Z^*(t)$ is non-negative and convex.
3. If $t > \mathbb{E}[Z]$, then $\psi_Z^*(t) = \sup_{\lambda \in \mathbb{R}}\{\lambda t - \psi_Z(\lambda)\}$, the **Fenchel-Legendre** dual.

Proof (Hints).

1. Differentiability proof omitted. For convexity, use **Holder's Inequality**.
2. Straightforward (note that each $t \mapsto \lambda t - \psi_Z(\lambda)$ is linear).
3. Straightforward. □

Proof.

1. $\psi_Z(\alpha\lambda_1 + (1 - \alpha)\lambda_2) = \log \mathbb{E}[e^{\alpha\lambda_1 Z} \cdot e^{(1 - \alpha)\lambda_2 Z}] \leq \alpha \log \mathbb{E}[e^{\lambda_1 Z}] + (1 - \alpha) \log \mathbb{E}[e^{\lambda_2 Z}]$ by Holder's inequality. The differentiability proof is omitted.
2. $\lambda t - \psi_Z(\lambda)|_{\lambda=0} = 0$, so $\psi_Z^*(t) \geq 0$ by definition. Convexity follows since it is a supremum of linear functions.
3. By convexity and Jensen's inequality, $\mathbb{E}[e^{\lambda Z}] \geq e^{\lambda \mathbb{E}[Z]}$. So for $\lambda < 0$, $\lambda t - \psi_Z(\lambda) \leq \lambda(t - \mathbb{E}[Z]) < 0 = \lambda t - \psi_Z(\lambda)|_{\lambda=0}$. □

Example 1.13 Let $Z \sim N(0, \sigma^2)$. Then the MGF of Z is

$$\begin{aligned}
\mathbb{E}[e^{\lambda Z}] &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/2\sigma^2} e^{\lambda x} dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x^2 - 2\lambda\sigma^2 x + \lambda^2\sigma^4)/2\sigma^2} e^{\lambda^2\sigma^2/2} dx \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x - \lambda\sigma^2)^2/2\sigma^2} e^{\lambda^2\sigma^2/2} dx \\
&= e^{\lambda^2\sigma^2/2}.
\end{aligned}$$

So $\psi_Z(\lambda) = \frac{\lambda^2\sigma^2}{2}$. By Proposition 1.12, for $t > 0 = \mathbb{E}[Z]$, the Cramer transform is

$$\psi_Z^*(t) = \sup_{\lambda \in \mathbb{R}} \{\lambda t - \lambda^2\sigma^2/2\} =: \sup_{\lambda \in \mathbb{R}} g(\lambda).$$

We have $g'(\lambda) = t - \lambda\sigma^2 = 0$ iff $\lambda = t/\sigma^2$. So $\psi_Z^*(t) = t^2/\sigma^2 - \sigma^2 t^2/2\sigma^4 = t^2/2\sigma^2$. So Chernoff Bound gives

$$\mathbb{P}(Z \geq t) \leq e^{-t^2/2\sigma^2}.$$

Definition 1.14 Let X be an RV with $\mathbb{E}[X] = 0$. X is **sub-Gaussian** with variance parameter ν if

$$\psi_X(\lambda) \leq \frac{\lambda^2\nu}{2} \quad \forall \lambda \in \mathbb{R},$$

i.e. if its log MGF is less than that of a normally distributed random variable with mean 0 and variance ν . The set of all such sub-Gaussian variables is denoted $\mathcal{G}(\nu)$.

Proposition 1.15 For any sub-Gaussian RV X ,

1. If $X \in \mathcal{G}(\nu)$, then $\mathbb{P}(X \geq t), \mathbb{P}(X \leq -t) \leq e^{-t^2/2\nu}$ for all $t > 0$.
2. If X_1, \dots, X_n are independent with each $X_i \in \mathcal{G}(\nu_i)$ then $\sum_{i=1}^n X_i \in \mathcal{G}(\sum_{i=1}^n \nu_i)$.
3. If $X \in \mathcal{G}(\nu)$, then $\text{Var}(X) \leq \nu$.

Proof. Exercise. □

Definition 1.16 The **Gamma function** is defined as

$$\Gamma(z) := \int_0^\infty t^{z-1} e^{-t} dt.$$

Theorem 1.17 Let $\mathbb{E}[X] = 0$. TFAE for suitable choices of ν, b, c, d :

1. $X \in \mathcal{G}(\nu)$.
2. $\mathbb{P}(X \geq t), \mathbb{P}(X \leq -t) \leq e^{-t^2/2b}$ for all $t > 0$.
3. $\mathbb{E}[X^{2q}] \leq q!c^q$ for all $q \geq \mathbb{N}$.
4. $\mathbb{E}[e^{dX^2}] \leq 2$.

Proof (Hints).

- (1 \Rightarrow 2): straightforward.
- (2 \Rightarrow 3): Explain why we can assume $b = 1$. Use that $\mathbb{E}[Y] = \int_0^\infty \mathbb{P}(Y > t) dt$ for $Y \geq 0$, and the Γ function.

- (3 \Rightarrow 1): show that $\mathbb{E}[e^{\lambda X}] \leq \mathbb{E}[e^{\lambda(X-X')}]$ where X' is an IID copy of X . Show that $\mathbb{E}[(X - X')^{2q}] \leq 2^{2q} \cdot \mathbb{E}[X^{2q}]$. Expand $\mathbb{E}[e^{\lambda(X-X')}]$ as a series. Conclude that $X \in \mathcal{G}(4c)$.
- (3 \Leftrightarrow 4): exercise.

□

Proof. (1 \Rightarrow 2) instantly follows (with $b = \nu$) by Proposition [1.15](#).

(2 \Rightarrow 3): WLOG, $b = 1$. Otherwise consider $\tilde{X} = X/\sqrt{b}$. Recall that for $Y \geq 0$, $\mathbb{E}[Y] = \int_0^\infty \mathbb{P}(Y > t) dt$. Now

$$\begin{aligned}
\mathbb{E}[X^{2q}] &= \int_0^\infty \mathbb{P}(X^{2q} > t) dt = \int_0^\infty \mathbb{P}(|X| > t^{1/2q}) dt \\
&\leq 2 \int_0^\infty e^{-t^{1/q}/2} dt \\
&= 2 \cdot 2^q \cdot q \int_0^\infty u^{q-1} e^{-u} du \\
&= 2 \cdot 2^q \cdot q \cdot \Gamma(q) \\
&= 2^{q+1} \cdot q! \leq c^q q!
\end{aligned}$$

for some constant c , where we use the substitution $t^{1/q}/2 = u$, so $t = (2u)^q$, so $dt = 2^q q u^{q-1} du$.

(3 \Rightarrow 1): $\mathbb{E}[e^{-\lambda X}] \cdot \mathbb{E}[e^{\lambda X}] = \mathbb{E}[e^{\lambda(X-X')}]$, where X' is an IID copy of X . By Jensen's inequality, $\mathbb{E}[e^{-\lambda X}] \geq e^{-\lambda \mathbb{E}[X]} = 1$. So

$$\mathbb{E}[e^{\lambda X}] \leq \mathbb{E}[e^{\lambda(X-X')}] = \sum_{q=0}^\infty \frac{\lambda^{2q} \mathbb{E}[(X - X')^{2q}]}{(2q)!}$$

(we can ignore odd powers since $X - X'$ is a symmetric RV: $X - X'$ has the same distribution as $X' - X$). Now

$$\mathbb{E}[(X - X')^{2q}] = \sum_{k=0}^{2q} \binom{2q}{k} \mathbb{E}[X^k] \mathbb{E}[(X')^{2q-k}] \leq \sum_{k=0}^{2q} \binom{2q}{k} \mathbb{E}[X^{2q}] = 2^{2q} \cdot \mathbb{E}[X^{2q}],$$

by Holder's inequality with $p = 2q/k$ and $q = 2q/(2q - k)$ for each k . Thus,

$$\begin{aligned}
\mathbb{E}[e^{\lambda X}] &\leq \sum_{q=0}^\infty \frac{\lambda^{2q} \mathbb{E}[X^{2q}] \cdot 2^{2q}}{(2q)!} \\
&\leq \sum_{q=0}^\infty \frac{\lambda^{2q} c^q q! 2^{2q}}{(2q)!} \\
&\leq \sum_{q=0}^\infty \frac{\lambda^{2q} \cdot c^q 2^q}{q!} = \sum_{q=0}^\infty \frac{(\lambda^2 \cdot 2c)^q}{q!} = e^{2\lambda^2 c},
\end{aligned}$$

where we used that $(2q)!/q! = \prod_{j=1}^q (q+1)! \geq 2^q \cdot q!$. Hence $\psi_X(\lambda) = 2\lambda^2 c = \frac{\lambda^2 \cdot 4c}{2}$, hence $X \in \mathcal{G}(4c)$.

(3 \Leftrightarrow 4): exercise. □

1.2. Hoeffding's and related inequalities

Lemma 1.18 (Hoeffding's Lemma) Let Y be a RV with $\mathbb{E}[Y] = 0$ and $Y \in [a, b]$ almost surely. Then $\psi_Y''(\lambda) \leq (b-a)^2/4$ and $Y \in \mathcal{G}((b-a)^2/4)$.

Proof (Hints).

- Define a new distribution based on λ , which should be obvious after expanding $\psi_Y'(\lambda)$.
 - Show that $\psi_Y''(\lambda)$ is equal to the variance of this distribution, and obtain the upper bound on $\psi_Y''(\lambda)$ by using the shift-invariance of the variance.
 - To conclude the result, use a Taylor expansion at 0 of $\psi_Y(\lambda)$.
-

Proof. Let Y have distribution P . We have

$$\psi_Y'(\lambda) = \frac{\mathbb{E}_{Y \sim P}[Y e^{\lambda Y}]}{\mathbb{E}_{Y \sim P}[e^{\lambda Y}]} = \mathbb{E}_{Y \sim P} \left[Y \cdot \frac{e^{\lambda Y}}{\mathbb{E}[e^{\lambda Y}]} \right] = \mathbb{E}_{Y \sim P_\lambda}[Y],$$

where if P is discrete, then P_λ is the discrete distribution with PMF

$$P_\lambda(y) = \frac{e^{\lambda y} P(y)}{\sum_z P(z) e^{\lambda z}} = \frac{e^{\lambda y} P(y)}{\mathbb{E}[e^{\lambda Y}]},$$

and if P is continuous with PDF f , then P_λ is the continuous distribution with PDF

$$f_\lambda(y) = \frac{e^{\lambda y} f(y)}{\int_{-\infty}^{\infty} f(z) e^{\lambda z} dz} = \frac{e^{\lambda y} f(y)}{\mathbb{E}[e^{\lambda Y}]},$$

Now

$$\begin{aligned} \psi_Y''(\lambda) &= \frac{\mathbb{E}_{Y \sim P}[Y^2 e^{\lambda Y}] \cdot \mathbb{E}_{Y \sim P}[e^{\lambda Y}] - \mathbb{E}_{Y \sim P}[Y e^{\lambda Y}]^2}{\mathbb{E}_{Y \sim P}[e^{\lambda Y}]^2} \\ &= \mathbb{E}_{Y \sim P} \left[Y^2 \frac{e^{\lambda Y}}{\mathbb{E}_{Y \sim P}[e^{\lambda Y}]} \right] - \mathbb{E} \left[Y \frac{e^{\lambda Y}}{\mathbb{E}_{Y \sim P}[e^{\lambda Y}]} \right]^2 \\ &= \mathbb{E}_{Y \sim P_\lambda}[Y^2] - \mathbb{E}_{Y \sim P_\lambda}[Y]^2 = \text{Var}_{Y \sim P_\lambda}(Y). \end{aligned}$$

Note that if $Y \in [a, b]$, then $|Y - \frac{b-a}{2}|^2 \leq (b-a)^2/4$. So we have

$$\text{Var}_{Y \sim P_\lambda}(Y) = \text{Var}_{Y \sim P_\lambda}(Y - (b-a)/2) \leq \mathbb{E}_{Y \sim P_\lambda} \left[\left(Y - \frac{b-a}{2} \right)^2 \right] \leq \frac{(b-a)^2}{4}.$$

Finally, using a Taylor expansion at 0, we obtain

$$\psi_Y(\lambda) = \psi_Y(0) + \lambda'_Y(0)\lambda + \psi''_Y(\xi)\frac{\lambda^2}{2} = \psi''_Y(\xi)\frac{\lambda^2}{2} \leq \lambda^2 \frac{(b-a)^2}{8},$$

for some $\xi \in [0, \lambda]$, since $\mathbb{E}_{Y \sim P}[Y] = \mathbb{E}_{Y \sim P_0}[Y] = 0$. \square

Remark 1.19 The distribution P_λ in the above proof is called the **exponentially tilted** distribution.

Theorem 1.20 (Hoeffding's Inequality) Let X_1, \dots, X_n be independent RVs where each X_i takes values in $[a_i, b_i]$. Then for all $t \geq 0$,

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

Proof (Hints). Straightforward. \square

Proof. By Hoeffding's Lemma, $X_i - \mathbb{E}[X_i] \in \mathcal{G}((b_i - a_i)^2/4)$ for all i . By Proposition 1.15 (part 2), we have

$$\sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \in \mathcal{G}\left(\frac{1}{4} \sum_{i=1}^n (b_i - a_i)^2\right).$$

Hence, by Proposition 1.15 (part 1), we are done. \square

Remark 1.21 A drawback of Hoeffding's Inequality is that the bound does not involve $\text{Var}(X_i)$, and the variances could be much smaller than the upper bound of $(b_i - a_i)^2/4$. This is addressed by Bennett's inequality:

Theorem 1.22 (Bennett's Inequality) Let X_1, \dots, X_n be independent RVs with $\mathbb{E}[X_i] = 0$ and $|X_i| \leq c$ for all i . Let $\nu = \text{Var}(X_1) + \dots + \text{Var}(X_n)$. Then for all $t \geq 0$,

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geq t\right) \leq \exp\left(-\frac{\nu}{c^2} \cdot h_1\left(\frac{ct}{\nu}\right)\right),$$

where $h_1(x) = (1+x)\log(1+x) - x$ for $x > 0$.

Proof (Hints).

- Show that $\mathbb{E}[e^{\lambda X_i}] \leq 1 + \frac{\text{Var}(X_i)}{c^2}(e^{\lambda c} - \lambda c - 1)$.
- Deduce that $\psi_{\sum_i X_i} \leq \frac{\nu}{c^2}(e^{\lambda c} - \lambda c - 1)$.
- Find a lower bound for $\psi_{\sum_i X_i}^*(t)$.

\square

Proof. Denote $\sigma_i^2 = \text{Var}(X_i) = \mathbb{E}[X_i^2] - \mathbb{E}[X_i]^2 = \mathbb{E}[X_i^2]$. The MGF of X_i is

$$\begin{aligned} \mathbb{E}[e^{\lambda X_i}] &= \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \mathbb{E}[X_i^k] = 1 + \sum_{k=2}^{\infty} \frac{\lambda^k}{k!} \mathbb{E}[X_i^{k-2} X_i^2] \\ &\leq 1 + c^{k-2} \sum_{k=2}^{\infty} \frac{\lambda^k}{k!} \mathbb{E}[X_i^2] = 1 + \frac{\sigma_i^2}{c^2} \sum_{k=2}^{\infty} \frac{\lambda^k c^k}{k!} \end{aligned}$$

$$\begin{aligned}
&= 1 + \frac{\sigma_i^2}{c^2} \left(\sum_{k=0}^{\infty} \frac{\lambda^k c^k}{k!} - \lambda c - 1 \right) \\
&= 1 + \frac{\sigma_i^2}{c^2} (e^{\lambda c} - \lambda c - 1).
\end{aligned}$$

(We can apply the inequality since $\mathbb{E}[X_i^k] \geq \mathbb{E}[X_i]^k = 0$ by Jensen's inequality.) So $\psi_{X_i}(\lambda) = \log\left(1 + \frac{\sigma_i^2}{c^2}(e^{\lambda c} - \lambda c - 1)\right) \leq \frac{\sigma_i^2}{c^2}(e^{\lambda c} - \lambda c - 1)$. So by additivity of ψ , we have

$$\psi_{\sum_{i=1}^n X_i}(\lambda) \leq \frac{\nu}{c^2} e^{\lambda c} - \frac{\nu}{c^2} \lambda c - \frac{\nu}{c^2}.$$

So for $t \geq 0 = \mathbb{E}[\sum_i X_i]$, by Proposition 1.12,

$$\psi_{\sum_i X_i}^*(t) \geq \sup_{\lambda \in \mathbb{R}} \left\{ \lambda t - \frac{\nu}{c^2} e^{\lambda c} + \frac{\nu}{c} \lambda + \frac{\nu}{c^2} \right\} =: \sup_{\lambda \in \mathbb{R}} \{g(\lambda)\}$$

We have $g'(\lambda) = t - \frac{\nu}{c} e^{\lambda c} + \frac{\nu}{c}$ which is 0 iff $t + \frac{\nu}{c} = \frac{\nu}{c} e^{\lambda c}$, i.e. iff $\lambda = \frac{1}{c} \log(1 + t \frac{c}{\nu}) =: \lambda^*$. So

$$\begin{aligned}
\psi_{\sum X_i}^*(t) &\geq \frac{1}{c} t \log\left(1 + \frac{tc}{\nu}\right) - \frac{\nu}{c^2} \left(1 + \frac{tc}{\nu}\right) + \frac{\nu}{c^2} \log\left(1 + \frac{tc}{\nu}\right) + \frac{\nu}{c^2} \\
&= \frac{\nu}{c^2} \left(\left(1 + \frac{tc}{\nu}\right) \log\left(1 + \frac{tc}{\nu}\right) - \frac{tc}{\nu} \right) \\
&= \frac{\nu}{c^2} h_1\left(\frac{tc}{\nu}\right).
\end{aligned}$$

So we are done by the Chernoff Bound. □

Remark 1.23 We can show that $h_1(x) \geq \frac{x^2}{2(x/3+1)}$ for $x \geq 0$. So by Bennett's Inequality, we obtain

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geq t\right) \leq \exp\left(-\frac{t^2}{2(ct/3 + \nu)}\right),$$

which is **Bernstein's inequality**. If $\nu \gg ct$, then this yields a sub-Gaussian tail bound, and if $\nu \ll ct$, then this yields an exponential bound. So Bernstein misses a log factor.

Remark 1.24 If $Z \sim \text{Pois}(\lambda)$, then $\psi_{Z-\nu}(\lambda) = \nu(e^\lambda - \lambda - 1)$.

2. The variance method

2.1. The Efron-Stein inequality

Notation 2.1 Denote $X^{(i)} = (X_{1:(i-1)}, X_{(i+1):n})$ and for $i < j$, denote $X_{i:j} = (X_i, \dots, X_j)$.

Notation 2.2 Denote $E_i Z = \mathbb{E}[Z \mid X_{1:i}]$, $E_0 Z = \mathbb{E}[Z]$, $E^{(i)} = \mathbb{E}[Z \mid X^{(i)}]$, and $\text{Var}^{(i)}(Z) = \text{Var}(Z \mid X^{(i)})$.

We want to study the concentration of $Z = f(X_1, \dots, X_n)$ for independent X_i . If $Z = \sum_i X_i$, then $\text{Var}(\sum_i X_i) = \sum_i \text{Var}(X_i)$ if $\mathbb{E}[X_i X_j] = \mathbb{E}[X_i] \mathbb{E}[X_j]$ for all $i \neq j$, which holds if the X_i are independent.

Theorem 2.3 (Efron-Stein Inequality) Let X_1, \dots, X_n be independent and let $Z = f(X_1, \dots, X_n)$. Then

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

Proof (Hints).

- The [Law of Total Expectation](#) and [Tower Property of Conditional Expectation](#) will come in handy a lot...
- Let $\Delta_i = E_i Z - E_{i-1} Z$. Show that $\mathbb{E}[\Delta_i] = 0$.
- Show that the Δ_i are uncorrelated, i.e. $\mathbb{E}[\Delta_i \Delta_j] = \mathbb{E}[\Delta_i] \mathbb{E}[\Delta_j]$.
- Show that $\Delta_i = E_i(Z - E^{(i)} Z)$.

□

Proof. Let $\Delta_i = E_i Z - E_{i-1} Z$. By the [Law of Total Expectation](#), we have

$$\mathbb{E}[\Delta_i] = \mathbb{E}[\mathbb{E}[Z \mid X_{1:i}]] - \mathbb{E}[\mathbb{E}[Z \mid X_{1:(i-1)}]] = \mathbb{E}[Z] - \mathbb{E}[Z] = 0.$$

Also, note that $Z - \mathbb{E}[Z] = \mathbb{E}[Z \mid X_{1:n}] - \mathbb{E}[Z] = \sum_{i=1}^n \Delta_i$. We claim that the Δ_i are uncorrelated, i.e. $\mathbb{E}[\Delta_i \Delta_j] = \mathbb{E}[\Delta_i] \mathbb{E}[\Delta_j] = 0$ for $i \neq j$. Indeed, for $i < j$, by the [Law of Total Expectation](#), we can write

$$\mathbb{E}[\Delta_i \Delta_j] = \mathbb{E}[\mathbb{E}[\Delta_i \Delta_j \mid X_{1:i}]] = \mathbb{E}[\Delta_i \mathbb{E}[\Delta_j \mid X_{1:i}]],$$

since Δ_i is a function of $X_{1:i}$. But

$$\begin{aligned} \mathbb{E}[\Delta_j \mid X_{1:i}] &= \mathbb{E}(E_j Z - E_{j-1} Z \mid X_{1:i}) \\ &= \mathbb{E}[\mathbb{E}[Z \mid X_{1:j}] \mid X_{1:i}] - \mathbb{E}[\mathbb{E}[Z \mid X_{1:(j-1)}] \mid X_{1:i}] \\ &= \mathbb{E}[Z \mid X_{1:i}] - \mathbb{E}[Z \mid X_{1:i}] = E_i Z - E_i Z = 0, \end{aligned}$$

where on the third line we used the [Tower Property of Conditional Expectation](#). Hence, the Δ_i are uncorrelated, which implies

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

Now

$$\begin{aligned} E_i(E^{(i)} Z) &= \mathbb{E}[E^{(i)} Z \mid X_{1:i}] \\ &= \mathbb{E}[E^{(i)} Z \mid X_{1:(i-1)}, X_i] \\ &= \mathbb{E}[\mathbb{E}[Z \mid X^{(i)}] \mid X_{1:(i-1)}] \end{aligned}$$

$$\begin{aligned}
&= \mathbb{E}[Z \mid X_{1:(i-1)}] \\
&= E_{i-1}Z,
\end{aligned}$$

where on the third line we used that X_i and $X^{(i)}$ are independent, and on the fourth line we used the [Tower Property of Conditional Expectation](#). So we can rewrite $\Delta_i = E_i Z - E_{i-1} Z = E_i(Z - E^{(i)} Z)$, and so by Jensen's inequality

$$\begin{aligned}
\Delta_i^2 &= (E_i(Z - E^{(i)} Z))^2 = \mathbb{E}[Z - E^{(i)} Z \mid X_{1:i}]^2 \\
&\leq \mathbb{E}[(Z - E^{(i)} Z)^2 \mid X_{1:i}] = E_i((Z - E^{(i)} Z)^2).
\end{aligned}$$

Hence, by the [Law of Total Expectation](#),

$$\begin{aligned}
\text{Var}(Z) &= \sum_{i=1}^n \mathbb{E}[\Delta_i^2] \leq \sum_{i=1}^n \mathbb{E}[E_i((Z - E^{(i)} Z)^2)] \\
&= \sum_{i=1}^n \mathbb{E}[\mathbb{E}[(Z - E^{(i)} Z)^2 \mid X_{1:i}]] = \sum_{i=1}^n \mathbb{E}[(Z - E^{(i)} Z)^2].
\end{aligned}$$

Finally, we have $\mathbb{E}[E^{(i)}(Z - E^{(i)} Z)^2] = \mathbb{E}[\text{Var}(Z \mid X^{(i)})] = \mathbb{E}[\text{Var}^{(i)}(Z)]$, which gives the equality in the theorem statement. \square

Theorem 2.4 (Efron-Stein Inequality) Let X_1, \dots, X_n be independent and f be square integrable. Let $Z = f(X_1, \dots, X_n)$. Then

$$\text{Var}(Z) \leq \mathbb{E}\left[\sum_{i=1}^n (Z - E^{(i)} Z)^2\right] =: \nu.$$

Moreover, if X'_1, \dots, X'_n are IID copies of X_1, \dots, X_n , and $Z'_i = f(X_{1:(i-1)}, X'_i, X_{(i+1):n})$, then

$$\nu = \frac{1}{2} \mathbb{E}\left[\sum_{i=1}^n (Z - Z'_i)^2\right] = \mathbb{E}\left[\sum_{i=1}^n (Z - Z'_i)_+^2\right] = \mathbb{E}\left[\sum_{i=1}^n (Z - Z'_i)_-^2\right],$$

where $X_+ = \max\{0, X\}$ and $X_- = \max\{-X, 0\}$. Moreover,

$$\nu = \sum_{i=1}^n \inf_{Z_i} \mathbb{E}[(Z - Z_i)^2],$$

where the infimum is over all $X^{(i)}$ -measurable and square-integrable RVs Z_i .

Proof (Hints).

- First part is straightforward.
- For second part, show that $\text{Var}^{(i)}(Z) = \frac{1}{2} \text{Var}^{(i)}(Z - Z'_i)$.
- For last part, use that $\text{Var}(X) = \inf_a \mathbb{E}[(X - a)^2]$.

\square

Proof. The first part follows instantly from the [Efron-Stein Inequality](#) by linearity of expectation. Now $\text{Var}(X) = \frac{1}{2} \text{Var}(X - Y)$, if X and Y are IID. Conditional on $X^{(i)}$, Z and Z'_i are independent. Hence, since $\mathbb{E}[Z] = \mathbb{E}[Z'_i]$,

$$\text{Var}^{(i)}(Z) = \frac{1}{2} \text{Var}^{(i)}(Z - Z'_i) = \frac{1}{2} \mathbb{E}^{(i)}[(Z - Z'_i)^2].$$

Thus we have

$$\nu = \frac{1}{2} \sum_{i=1}^n \mathbb{E}[(Z - Z'_i)^2].$$

The equality with \cdot_+ and \cdot_- follows since $Z - Z'_i$ is a symmetric RV. Finally, recall that $\text{Var}(X) = \inf_a \mathbb{E}[(X - a)^2]$, with equality if $a = \mathbb{E}[X]$. So $\text{Var}^{(i)}(Z) = \inf_{Z'_i} \mathbb{E}^{(i)}[(Z - Z'_i)^2]$, with equality if $Z'_i = \mathbb{E}^{(i)} Z$. Taking expectations and summing completes the proof. \square

2.2. Functions with bounded differences

Definition 2.5 $f : A^n \rightarrow \mathbb{R}$ has the **bounded differences (b.d.)** property if

$$\sup_{(x, x'_i) \in A^{n+1}} |f(x_{1:(i-1)}, x_i, x_{(i+1):n}) - f(x_{1:(i-1)}, x'_i, x_{(i+1):n})| \leq c_i \quad \forall i \in [n].$$

So changing one of the coordinates changes the value of the function at most by a constant.

Corollary 2.6 Let X_1, \dots, X_n be independent and $Z = f(X_{1:n})$ have bounded differences with constants c_i . Then $\text{Var}(Z) \leq \frac{1}{4} \sum_{i=1}^n c_i^2$.

Proof (Hints). Consider the random variable

$$Z_i = \frac{1}{2} \left(\sup_{x_i \in A} f(X_{1:(i-1)}, x_i, X_{(i+1):n}) + \inf_{x_i \in A} f(X_{1:(i-1)}, x_i, X_{(i+1):n}) \right).$$

\square

Proof. Define

$$Z_i = \frac{1}{2} \left(\sup_{x_i \in A} f(X_{1:(i-1)}, x_i, X_{(i+1):n}) + \inf_{x_i \in A} f(X_{1:(i-1)}, x_i, X_{(i+1):n}) \right)$$

Z_i is a function of $X^{(i)}$. We have $|Z - Z_i| \leq c_i/2$. By the final part of the [Efron-Stein Inequality](#), we have $\text{Var}(Z) \leq \sum_{i=1}^n \mathbb{E}[(Z - Z_i)^2] \leq \frac{1}{4} \sum_{i=1}^n c_i^2$. \square

Example 2.7 (Bin packing) Given $x_1, \dots, x_n \in [0, 1]$, what is the minimum number k of bins B_j into which $\sum_{x \in B_j} x \leq 1$ for each $j = 1, \dots, k$?

Suppose X_1, \dots, X_n be independent and let $Z = f(X_{1:n})$ be the minimum number of bins. Note that changing any one x_i changes f by at most 1, so f has bounded differences with constants $c_i = 1$. So by the [Efron-Stein Inequality](#), $\text{Var}(Z) \leq \frac{1}{4}n$.

Note that this bound is tight, e.g. when $X_i \sim \text{Bern}(1/2)$, $Z \sim B(n, 1/2)$, which has variance $n \cdot \frac{1}{2} \cdot \frac{1}{2}$.

Example 2.8 (Longest common sub-sequence) Let $X_{1:n}$ and $Y_{1:n}$ be independent sequences of coin flips. Let

$$Z = f(X_{1:n}, Y_{1:n}) = \max \left\{ k : \exists i_1 < \dots < i_k, j_1 < \dots < j_k \text{ s.t. } X_{i_\ell} = Y_{j_\ell} \forall \ell \in [k] \right\}$$

Note that changing any one coin flip changes Z by at most 1, so f has bounded differences with constants $c_i = 1$, so by the [Efron-Stein Inequality](#), $\text{Var}(Z) \leq n/2 = \Theta(n)$. Since it is known that $\mathbb{E}[Z] = \Theta(n)$, the deviations from the mean are small compared to the mean.

Example 2.9 (Chromatic numbers of graphs) Let G be an **Erdos-Renyi random graph** with n vertices, i.e. each $\{i, j\} \in E(G)$ with probability p (independently). The **chromatic number** $\chi(G)$ of G is the smallest number of colors on the vertices such that there are no two adjacent vertices with the same colour. For $i < j$, let $X_{ij} = \mathbb{1}_{\{\{i,j\} \in E\}}$. We have

$$\chi(G) = f\left(\{X_{ij}\}_{1 \leq i < j \leq n}\right),$$

for some (complicated) function f . Since adding or removing an edge changes $\chi(G)$ by at most 1, f has bounded differences with constants $c_{ij} = 1$. By [Efron-Stein Inequality](#), $\text{Var}(Z) \leq \binom{n}{2}/4 = \Theta(n^2)$. It is known that $\mathbb{E}[\chi(G)] \approx n/\log n$, so the bound on the variance is not useful when applying [Chebyshev's Inequality](#). However:

Now for each $1 \leq i \leq n-1$, let $Y^{(i)}$ be a random vector taking values in $\{0, 1\}^i$ where $Y_j^{(i)} = \mathbb{1}_{\{\{i+1, j\} \in E\}}$ for each $1 \leq j \leq i$. The Y_i are independent. Also, note that $\{Y^{(i)}\}_{i=1}^{n-1}$ determines the graph. Hence, $\chi(G) = g(Y^{(1)}, \dots, Y^{(n-1)})$ for some (complicated) function g . g has bounded differences with constants 1 (e.g. by considering giving vertex $i+1$ a new colour). Then by [Efron-Stein Inequality](#), $\text{Var}(\chi(G)) \leq (n-1)/4$, which is a tighter bound. This yields a useful application of [Chebyshev's Inequality](#), which shows that $\chi(G)$ is close to its mean value.

3. Poincaré inequalities

Let X_1, \dots, X_n be real-valued random variables, and let $Z = f(X_1, \dots, X_n)$. A Poincaré inequality is of the form $\text{Var}(Z) \lesssim \mathbb{E}[\|\nabla f(X)\|^2]$. So we have a local property (smoothness) which gives a global property (bound on the variance).

Definition 3.1 Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is **separately convex** if it is convex if all of its individual arguments.

Theorem 3.2 (Convex Poincaré Inequality) Let $X_{1:n}$ be independent RVs supported on $[0, 1]$ and $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be separately convex with partial derivatives that exist. Let $Z = f(X_{1:n})$. Then

$$\text{Var}(Z) \leq \mathbb{E}[\|\nabla f(X_{1:n})\|^2],$$

where $\|\cdot\| = \|\cdot\|_2$ is the Euclidean norm.

Proof (Hints).

- Let $Z_i = \inf_{x'_i} f(X_{1:(i-1)}, x'_i, X_{(i+1):n})$. Let X'_i be the value for which the infimum is achieved (why is it achieved?).
- Use that $|Z - Z_i|^2 \leq |X_i - X'_i|^2 \cdot \left(\frac{\partial f}{\partial x_i}(X)\right)^2$ (since X'_i is a minimiser).

□

Proof. Let $Z_i = \inf_{x'_i} f(X_{1:(i-1)}, x'_i, X_{(i+1):n})$. Let X'_i be the value for which the infimum is achieved (since f is continuous and the domain $[0, 1]^n$ we consider is compact). Denote $\bar{X}^{(i)} = (X_{1:(i-1)}, X'_i, X_{(i+1):n})$. Note that since f is separately convex and X'_i is a minimiser (so $f(X'_i) \leq f(X)$),

$$|Z - Z_i|^2 = |f(X_{1:n}) - f(\bar{X}^{(i)})|^2 \leq |X_i - X'_i|^2 \cdot \left(\frac{\partial f}{\partial x_i}(X_{1:n})\right)^2.$$

By the [Efron-Stein Inequality](#),

$$\begin{aligned} \text{Var}(Z) &\leq \sum_{i=1}^n \mathbb{E}[(Z - Z_i)^2] \\ &\leq \sum_{i=1}^n \mathbb{E}\left[(X_i - X'_i)^2 \left(\frac{\partial f}{\partial x_i}(X_{1:n})\right)^2\right] \\ &\leq \sum_{i=1}^n \mathbb{E}\left[\left(\frac{\partial f}{\partial x_i}(X_{1:n})\right)^2\right] = \mathbb{E}[\|\nabla f(X_{1:n})\|^2]. \end{aligned}$$

□

Example 3.3 Let $X \in \mathbb{R}^{n \times d}$ be a random matrix with $X_{i,j} \in [-1, 1]$ independent. The spectral norm (or ℓ_2 -operator norm) of X is its largest singular value:

$$\sigma_1(X) = \sup\{\|Xu\| : u \in \mathbb{R}^d, \|u\| = 1\} = \sup_{u \in \mathbb{R}^n, \|u\|=1} \sup_{v \in \mathbb{R}^d, \|v\|=1} \langle u, Xv \rangle.$$

σ_1 is convex (and so separately convex) since it is a supremum of linear functions. Since it is a norm, we have $\sigma_1(A + B) \leq \sigma_1(A) + \sigma_1(B)$ and $\sigma_1(A - B) \geq |\sigma_1(A) - \sigma_1(B)|$. Fix A . Since X ranges over a compact set, the supremum is achieved: let u, v achieve the supremum. Then

$$\begin{aligned} \sigma_1(A) = \langle v, Xu \rangle &\leq \|v\| \cdot \|Xu\| \quad \text{by Cauchy-Schwarz} \\ &\leq \|v\| \cdot \|u\| \left(\sum_{i,j} X_{i,j}^2\right)^{1/2} = \left(\sum_{i,j} X_{i,j}^2\right)^{1/2} = \|X\|_F. \end{aligned}$$

Now if X, X' are independent, $d(X, X') = \|X - X'\|_F \geq \sigma_1(X - X') \geq |\sigma_1(X) - \sigma_1(X')|$ where d is the Euclidean distance between vectorised X and X' (i.e. Frobenius norm). So σ_1 is a 1-Lipschitz function, and note that an L -Lipschitz function satisfies $\|\nabla f\| \leq L$. So by the [Convex Poincaré Inequality](#), $\text{Var}(\sigma_1(X)) \leq 4$ (the RHS is 4, not

1, since X_{ij} take values in $[-1, 1]$ instead of $[0, 1]$). Note that this is independent of the dimension of X !

Theorem 3.4 (Gaussian Poincaré Inequality) Let $X_{1:n}$ be IID and standard Gaussian (i.e. each $X_i \sim N(0, 1)$). Then for any continuously differentiable $f \in C^1(\mathbb{R}^n)$,

$$\text{Var}(f(X_{1:n})) \leq \mathbb{E}[\|\nabla f(X_{1:n})\|^2].$$

Proof (Hints).

- Show, using the [Efron-Stein Inequality](#), that it is sufficient to prove the result for $n = 1$.
- You may assume that $f \in C^2(\mathbb{R})$ (f is twice continuously differentiable) and has compact support.
- Using the definition of conditional variance, show that $\text{Var}^{(i)}(f(S_n)) = \frac{1}{4} \left(f\left(S_n - \frac{\varepsilon_i}{\sqrt{n}} + \frac{1}{\sqrt{n}}\right) - f\left(S_n - \frac{\varepsilon_i}{\sqrt{n}} - \frac{1}{\sqrt{n}}\right) \right)^2$, where $S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i$ and ε_i are IID Rademacher random variables (taking values in $\{-1, 1\}$ with equal probability).
- Use Taylor's theorem to find an upper bound for

$$\left| f\left(S_n - \frac{\varepsilon_i}{\sqrt{n}} + \frac{1}{\sqrt{n}}\right) - f\left(S_n - \frac{\varepsilon_i}{\sqrt{n}} - \frac{1}{\sqrt{n}}\right) \right|$$

- Use [Efron-Stein Inequality](#) for $f(S_n)$ and the central limit theorem to conclude the result.

□

Proof. Assume the result holds for the $n = 1$ case, i.e. $\text{Var}(f(X)) \leq \mathbb{E}[f'(X)^2]$ for $X \sim N(0, 1)$. Then by the [Efron-Stein Inequality](#) and [Law of Total Expectation](#),

$$\begin{aligned} \text{Var}(Z) &\leq \mathbb{E} \left[\sum_{i=1}^n \text{Var}^{(i)}(f(X_{1:n})) \right] \\ &\leq \mathbb{E} \left[\sum_{i=1}^n \mathbb{E} \left[\left(\frac{\partial f}{\partial x_i}(X_{1:n}) \right)^2 \mid X^{(i)} \right] \right] \\ &= \mathbb{E} \left[\sum_{i=1}^n \left(\frac{\partial f}{\partial x_i}(X_{1:n}) \right)^2 \right] = \mathbb{E}[\|\nabla f(X_{1:n})\|^2]. \end{aligned}$$

So it suffices to prove the result for $n = 1$: WLOG, assume $\mathbb{E}[\|\nabla f(X)\|^2] < \infty$. Let ε_i be IID Rademacher random variables (taking values in $\{-1, 1\}$ with equal probability). Consider $S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i$. It suffices to prove the case when $f \in C^2(\mathbb{R})$ (f is twice continuously differentiable) and has compact support. So f' and f'' are bounded. By the [Efron-Stein Inequality](#),

$$\text{Var}(f(S_n)) \leq \mathbb{E} \left[\sum_{i=1}^n \text{Var}^{(i)}(S_n) \right].$$

Note $\text{Var}^{(i)}$ here is conditional on $\varepsilon^{(i)}$. We have $S_n = S_n - \varepsilon_i/\sqrt{n} \pm 1/\sqrt{n}$ with equal probabilities. Note that $S_n - \varepsilon_i/\sqrt{n}$ is a function of $\varepsilon^{(i)}$. We have

$$\mathbb{E}^{(i)}[f(S_n)] = \frac{1}{2}f(S_n - \varepsilon_i/\sqrt{n} + 1/\sqrt{n}) + \frac{1}{2}f(S_n - \varepsilon_i/\sqrt{n} - 1/\sqrt{n})$$

and so

$$\begin{aligned} \text{Var}^{(i)}(f(S_n)) &= \frac{1}{2} \left(f(S_n - \varepsilon_i/\sqrt{n} + 1/\sqrt{n}) - \left(\frac{1}{2}f(S_n - \varepsilon_i/\sqrt{n} + 1/\sqrt{n}) + \frac{1}{2}f(S_n - \varepsilon_i/\sqrt{n} - 1/\sqrt{n}) \right) \right)^2 \\ &\quad + \frac{1}{2} \left(f(S_n - \varepsilon_i/\sqrt{n} - 1/\sqrt{n}) - \left(\frac{1}{2}f(S_n - \varepsilon_i/\sqrt{n} + 1/\sqrt{n}) + \frac{1}{2}f(S_n - \varepsilon_i/\sqrt{n} - 1/\sqrt{n}) \right) \right)^2 \\ &= \frac{1}{4} (f(S_n - \varepsilon_i/\sqrt{n} + 1/\sqrt{n}) - f(S_n - \varepsilon_i/\sqrt{n} - 1/\sqrt{n}))^2 \end{aligned}$$

Let K be an upper bound for $|f''|$. Then

$$\begin{aligned} &|f(S_n + (1 - \varepsilon_i)/\sqrt{n}) - f(S_n - (1 + \varepsilon_i)/\sqrt{n})| \\ &= \left| f(S_n) + \frac{1 - \varepsilon_i}{\sqrt{n}} f'(S_n - \varepsilon_i/\sqrt{n}) + \frac{(1 - \varepsilon_i)^2}{2n} f''(S_n - \varepsilon_i/\sqrt{n} + \xi_{i,m}) \right. \\ &\quad \left. - f(S_n) + \frac{1 + \varepsilon_i}{\sqrt{n}} f'(S_n - \varepsilon_i/\sqrt{n}) - \frac{(1 + \varepsilon_i)^2}{2n} f''(S_n - \varepsilon_i/\sqrt{n} + \xi_{i,m}^{(2)}) \right| \\ &\leq \left| \frac{2}{\sqrt{n}} f'(S_n) \right| + 2K/n. \end{aligned}$$

Thus, $\text{Var}^{(i)}(f(S_n)) \leq (|f'(S_n)/\sqrt{n}| + K/n)^2$. Hence,

$$\text{Var}(f(S_n)) \leq \mathbb{E} \left[\sum_{i=1}^n (|f'(S_n)/\sqrt{n}| + K/n)^2 \right] = \mathbb{E}[f'(S_n)^2] + 2 \frac{K}{\sqrt{n}} \mathbb{E}[|f'(S_n)|] + \frac{K^2}{n}$$

As $n \rightarrow \infty$, $\text{Var}(f(S_n)) \rightarrow \text{Var}(X)$, $X \sim N(0, 1)$ by the central limit theorem. Also, $\mathbb{E}[f'(S_n)^2] \rightarrow \mathbb{E}[f'(X)^2]$ by the central limit theorem. So in the limit, $\text{Var}(f(X)) \leq \mathbb{E}[f'(X)^2]$. \square

Remark 3.5 The above proof uses a **tensorisation** argument. Tensorisation roughly means decomposing a high-dimensional function into a sum of lower-dimensional functions. E.g. the formula $\text{Var}(\sum_i X_i) = \sum_i \text{Var}(X_i)$ uses the tensorisation property of variance. Also, the [Efron-Stein Inequality](#)

$$\text{Var}(Z) \leq \sum_{i=1}^n \mathbb{E}[\text{Var}^{(i)}(Z)].$$

can be thought of as an example of the tensorisation of variance.

Remark 3.6 If f is L -Lipschitz, i.e. $|f(x) - f(y)| \leq L \cdot \|x - y\|$, then $\|\nabla f\| \leq L$. The [Gaussian Poincaré Inequality](#) holds for L -Lipschitz functions (with L^2 on the RHS).

Example 3.7 Recall from earlier that the operator norm σ_1 is 1-Lipschitz. If $X \in \mathbb{R}^{n \times d}$ with each $X_{ij} \sim N(0, 1)$ IID, then by the Gaussian Poincaré Inequality, $\text{Var}(\sigma_1(X)) \leq 1$, which is a good bound, given that it is known that $\mathbb{E}[\sigma_1(X)] = O(\sqrt{n} + \sqrt{d})$.

Example 3.8 Let $X_1, \dots, X_n \sim N(0, 1)$ be independent. Let $Z = f(X) = \max_i X_i$. We have $\nabla f = (0, \dots, 1, \dots, 0)$ where 1 is at the index of the maximum. Hence, by the Gaussian Poincaré Inequality, $\text{Var}(Z) \leq 1$, which is a good bound, given it is known that $\mathbb{E}[Z_n] \approx \log n$.

3.1. Poincaré constant

Definition 3.9 Let X be an RV taking values in \mathbb{R}^d . We say X satisfies the Poincaré inequality with constant C if

$$\text{Var}(f(X)) \leq C \cdot \mathbb{E}[\|\nabla f(X)\|^2] \quad \forall f \in C^1(\mathbb{R}^d).$$

The smallest such constant $C_P(X)$ is the **Poincaré constant** of X :

$$C_P(X) = \sup_{f \in C^1(\mathbb{R}^d)} \frac{\text{Var}(f(X))}{\mathbb{E}[\|\nabla f(X)\|^2]}.$$

Proposition 3.10 The Poincaré constant satisfies the following properties:

1. $C_P(aX + b) = a^2 C_P(X)$ for constants $a \in \mathbb{R}, b \in \mathbb{R}^d$.
2. For any unit vector $\theta \in \mathbb{R}^d$, $\text{Var}(\langle X, \theta \rangle) \leq C_P(X)$. In particular, $\text{Var}(X_i) \leq C_P(X)$ for all i .
3. If X_1, \dots, X_n are independent, then

$$C_P(X_{1:n}) = \max_i C_P(X_i).$$

4. If $C_P(X) < \infty$, then X has connected support.

Proof. Exercise. □

Remark 3.11 The constant $1/C_P(X)$ is called the **spectral gap**.

Definition 3.12 We say $\{X_n\}_{n \in \mathbb{N}}$ is a **(time homogenous) Markov chain** on a finite state space S (which WLOG we can take to be $[d]$) if

$$\mathbb{P}(X_{n+1} = j \mid X_{1:n} = i_{1:n}) = \mathbb{P}(X_{n+1} = j \mid X_n = i_n)$$

for all n and $i_1, \dots, i_n, j \in S$, i.e. if X_{n+1} is conditionally independent of $X_{1:(n-1)}$ given X_n for all n .

Definition 3.13 The **transition matrix** $P \in \mathbb{R}^{d \times d}$ of the Markov chain is defined by

$$P_{ij} = \mathbb{P}(X_{n+1} = j \mid X_n = i),$$

and its **discrete generator** is $\Lambda := P - I$.

Definition 3.14 Let P be the transition matrix of a Markov chain. A row vector $\pi \in \mathbb{R}^d$ (which represents a distribution on $[d]$) on state space S is called **stationary** if $\pi_j = \sum_i \pi_i P_{ij}$ for all j (i.e. $\pi P = \pi$).

Definition 3.15 Given a Markov chain with stationary distribution $\pi \in \mathbb{R}^d$ and $f, g \in \mathbb{R}^d$, the **Dirichlet form** is defined as

$$\mathcal{E}(f, g) := -\langle f, \Lambda g \rangle_\pi,$$

where $\langle x, y \rangle_\pi = \sum_{i=1}^d x_i y_i \pi_i$.

Proposition 3.16 Let $P \in \mathbb{R}^{d \times d}$ be a reversible transition matrix with stationary distribution $\pi \in \mathbb{R}^d$. Assume the **reversibility** condition:

$$\pi_i P_{ij} = \pi_j P_{ji} \quad \forall i, j \in [d].$$

Let $f \in \mathbb{R}^d$. Then

$$\mathcal{E}(f, f) = \frac{1}{2} \mathbb{E}_{X_{n+1}, X_n \sim \pi} [(f(X_{n+1}) - f(X_n))^2],$$

which is the **discrete gradient** (we may view f as a function $i \mapsto f_i$).

Proof (Hints). Use that $\sum_j P_{ij} = 1$ for all i to split the sum $\sum_i f_i^2 \pi_i$ into two sums. \square

Proof. Since $\sum_j P_{ij} = 1$ for all i , we have

$$\begin{aligned} \mathcal{E}(f, f) &= \langle f, (I - P)f \rangle_\pi = \sum_i f_i^2 \pi_i - \sum_i f_i \pi_i \sum_j P_{ij} f_j \\ &= \frac{1}{2} \left(\sum_{i,j} f_i^2 \pi_i P_{ij} + \sum_{i,j} f_j^2 \pi_j P_{ji} - 2 \sum_{i,j} \pi_i P_{ij} f_i f_j \right) \\ &= \frac{1}{2} \sum_{i,j} \pi_i P_{ij} (f_i - f_j)^2 \\ &= \frac{1}{2} \sum_{i,j} \mathbb{P}(X_{n+1} = j \mid X_n = i) \mathbb{P}(X_n = i) (f_i - f_j)^2 \\ &= \frac{1}{2} \sum_{i,j} \mathbb{P}(X_{n+1} = j, X_n = i) (f(i) - f(j))^2 \\ &= \frac{1}{2} \mathbb{E} [(f(X_{n+1}) - f(X_n))^2]. \end{aligned}$$

\square

Remark 3.17 If the transition matrix P is reversible, then $\Lambda = P - I$ is self-adjoint with respect to $\langle \cdot, \cdot \rangle_\pi$ (i.e. $\langle \Lambda f, g \rangle_\pi = \langle f, \Lambda g \rangle_\pi$), so has real eigenvalues $\lambda_1 \geq \dots \geq \lambda_n$. By Proposition 3.16, we have $\langle f, -\Lambda f \rangle_\pi \geq 0$, so $-\Lambda$ is positive semi-definite, and so all $\lambda_i \leq 0$. Since $\sum_j \Lambda_{ij} = 0$ for all i , we have $\lambda_1 = 0$, corresponding to eigenvector $f_1 = (1, \dots, 1)$.

Now $\lambda_2 = \sup_{f: \langle f, f_1 \rangle_\pi = 0} \frac{\langle f, \Lambda f \rangle_\pi}{\langle f, f \rangle_\pi}$, so

$$\mathcal{E}(f, f) = -\langle f, \Lambda f \rangle_\pi \geq -\lambda_2 \langle f, f \rangle_\pi = -\lambda_2 \mathbb{E}_\pi [f(X_1)^2] = -\lambda_2 \text{Var}_\pi(f) = (\lambda_1 - \lambda_2) \text{Var}_\pi(f)$$

for all $f \in \mathbb{R}^d$ such that $\mathbb{E}_\pi[f(X_1)] = \langle f, f_1 \rangle_\pi = 0$. There is equality if $f = f_2$, the eigenvector corresponding to λ_2 .

The best constant, c , in the inequality $\text{Var}_\pi(f) \leq c \cdot \mathcal{E}(f, f)$ is $c = \frac{1}{\lambda_1 - \lambda_2}$, the spectral gap.

4. The entropy method

4.1. Entropy, chain rules and Han's inequality

In the following section, let A be a discrete (countable) alphabet and let X be an RV on A .

Definition 4.1 The **Shannon entropy** of X with PMF P is

$$H(X) = \mathbb{E}[-\log P(X)] = - \sum_{x \in A} \mathbb{P}(X = x) \log \mathbb{P}(X = x),$$

where we use the convention $0 \log 0 = 0$.

Example 4.2 The entropy of $X \sim \text{Bern}(p)$ is $H(X) = -p \log p - (1-p) \log(1-p)$.

Remark 4.3 Note that for $x_1^n \in A^n$, $P^n(x_1^n) = e^{-n \frac{1}{n} \sum_{i=1}^n -\log P(x_i)}$ (P^n is the product distribution). So $P^n(X_1^n) = e^{-n \frac{1}{n} \sum_{i=1}^n -\log P(X_i)} \approx e^{-nH(X_i)}$ for IID X_i , by the **Weak Law of Large Numbers**.

Proposition 4.4 Properties of Shannon entropy:

- H is non-negative.
- $H(\cdot)$ is concave as a functional of P .
- If $|A| < \infty$, then $H(X) \leq \log|A|$ with equality if $X \sim \text{Unif}(A)$.

Proof. Exercise. □

Definition 4.5 For PMFs Q, P on A , Q is **absolutely continuous** with respect to P , written $Q \ll P$, if $P(x) = 0 \Rightarrow Q(x) = 0$ for all $x \in A$.

Definition 4.6 Let Q, P be PMFs on A such that $Q \ll P$ (which means if $P(x) = 0$, then $Q(x) = 0$). The **relative entropy** between Q and P is

$$D(Q \parallel P) = \mathbb{E}_Q \left[\log \frac{Q(X)}{P(X)} \right] = \sum_{x \in A} Q(x) \log \frac{Q(x)}{P(x)}$$

if $Q \ll P$, and $D(Q \parallel P) = \infty$ otherwise. We use the convention that $0 \log \frac{0}{0} = 0$.

Proposition 4.7 Properties of relative entropy:

- $D(Q \parallel P) \geq 0$.
- $D(Q \parallel P)$ is convex in both arguments.
- If $X \sim P$ where P is the uniform distribution on A , and $Y \sim Q$, then $D(Q \parallel P) = H(X) - H(Y)$.

Proof. Exercise. □

Definition 4.8 The **conditional entropy** of X given Y is

$$\begin{aligned}
H(X | Y) &= \mathbb{E}[-\log P_{X|Y}(X | Y)] = - \sum_{x,y} P(x,y) \log P(x | y) \\
&= \sum_y \mathbb{P}(Y = y) H(X | Y = y)
\end{aligned}$$

Theorem 4.9 (Chain Rule for Entropy) We have

$$H(X_{1:n}) = \mathbb{E}[-\log P(X_{1:n})] = \sum_{i=1}^n H(X_i | X_{1:(i-1)}).$$

Proof (Hints). Straightforward. □

Proof. Since

$$\mathbb{P}(X_{1:n} = x_{1:n}) = \mathbb{P}(X_1 = x_1) \mathbb{P}(X_2 = x_2 | X_1 = x_1) \cdots \mathbb{P}(X_n = x_n | X_{1:(n-1)} = x_{1:(n-1)}),$$

we have

$$\begin{aligned}
H(X_{1:n}) &= \mathbb{E}[-\log P(X_{1:n})] = \mathbb{E}\left[\sum_{i=1}^n -\log P(X_i | X_{1:(i-1)})\right] \\
&= \sum_{i=1}^n \mathbb{E}[-\log P(X_i | X_{1:(i-1)})] \\
&= \sum_{i=1}^n H(X_i | X_{1:(i-1)}).
\end{aligned}$$

□

Proposition 4.10 (Conditioning Reduces Entropy) $H(X | Y) \leq H(X)$.

Proof (Hints). Straightforward. □

Proof. We have

$$\begin{aligned}
H(X) - H(X | Y) &= \mathbb{E}\left[\log \frac{1}{P(X)} + \log P(X | Y)\right] \\
&= \mathbb{E}\left[\log \frac{P(X | Y)P(Y)}{P(X)P(Y)}\right] = D(P_{X,Y} \| P_X P_Y) \geq 0.
\end{aligned}$$

□

Definition 4.11 Similarly to the definition of relative entropy, the **conditional relative entropy** of X and Y given Z is

$$D(X \| Y | Z) = \sum_z \mathbb{P}(Z = z) D(X | Z = z \| Y | Z = z).$$

We also write e.g.

$$D(Q_{Y|X} \| P_Y | Q_X) = \sum_x \mathbb{P}(X = x) D(Q_{Y|X=x} \| P_Y).$$

Proposition 4.12 (Chain Rule for Relative Entropy) Let P, Q be PMFs on A^n . Let $X_{1:n} \sim Q$. Then

$$\begin{aligned} D(Q_{X_{1:n}} \parallel P_{X_{1:n}}) &= \sum_{i=1}^n \mathbb{E}_{Q_{X_{1:(i-1)}}} \left[D(Q_{X_i \mid X_{1:(i-1)}} \parallel P_{X_i \mid X_{1:(i-1)}}) \right] \\ &=: \sum_{i=1}^n D(Q_{X_i \mid X_{1:(i-1)}} \parallel P_{X_i \mid X_{1:(i-1)}} \mid Q_{X_{1:(i-1)}}) \end{aligned}$$

Proof (Hints). Straightforward. □

Proof. We have

$$\begin{aligned} D(Q_{X_{1:n}} \parallel P_{X_{1:n}}) &= \mathbb{E}_Q \left[\log \frac{Q(X_{1:n})}{P(X_{1:n})} \right] \\ &= \mathbb{E}_Q \left[\sum_{i=1}^n \log \frac{Q_{X_i \mid X_{1:(i-1)}}(X_i \mid X_{1:(i-1)})}{P_{X_i \mid X_{1:(i-1)}}(X_i \mid X_{1:(i-1)})} \right] \\ &= \sum_{i=1}^n \mathbb{E}_{Q_{X_{1:(i-1)}}} \left[D(Q_{X_i \mid X_{1:(i-1)}} \parallel P_{X_i \mid X_{1:(i-1)}}) \right] \end{aligned}$$

□

Remark 4.13 The Chain Rule for Relative Entropy is similar to the chain rule for variance:

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

$\Delta_i = \mathbb{E}[Z \mid X_{1:i}] - \mathbb{E}[Z \mid X_{1:(i-1)}]$, which led to the Efron-Stein Inequality.

Lemma 4.14 (Conditioning Reduces Conditional Entropy) $H(X \mid Y, Z) \leq H(X \mid Y)$.

Proof (Hints). Straightforward. □

Proof. $H(X \mid Y, Z) = \sum_z \mathbb{P}(Z = z) H(X \mid Y, Z = z) \leq \sum_z \mathbb{P}(Z = z) H(X \mid Z = z) = H(X \mid Z)$ by Conditioning Reduces Entropy. □

Theorem 4.15 (Han's Inequality) Let $X_{1:n}$ be discrete RVs. Then

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

Proof (Hints). Show that $H(X_{1:n}) \leq H(X^{(i)}) + H(X_i \mid X_{1:(i-1)})$. □

Proof. By the Chain Rule for Entropy and Conditioning Reduces Entropy,

$$\begin{aligned} H(X_{1:n}) &= H(X^{(i)}) + H(X_i \mid X^{(i)}) \\ &\leq H(X^{(i)}) + H(X_i \mid X_{1:(i-1)}) \end{aligned}$$

Summing over i , we obtain $nH(X_{1:n}) \leq \sum_{i=1}^n H(X^{(i)}) + H(X_{1:n})$ by the chain rule. □

Corollary 4.16 (Loomis-Whitney Inequality) The Loomis-Whitney inequality states that for finite $A \subseteq \mathbb{Z}^n$,

$$|A| \leq \prod_{i=1}^n |A^{(i)}|^{1/(n-1)}$$

Proof (Hints). Straightforward. □

Proof. Let $X_{1:n}$ be uniform on A . Then $\log|A| = H(X_{1:n})$. By Han's Inequality,

$$H(X_{1:n}) \leq \frac{1}{n-1} \sum_{i=1}^n H(X^{(i)}) \leq \frac{1}{n-1} \sum_{i=1}^n \log|A^{(i)}|$$

□

Lemma 4.17 Let Q, P be PMFs on a discrete set $A \times B \times C$. Then

$$D(Q_{Y|X,Z} \parallel P_Y \mid Q_{X,Z}) \geq D(Q_{Y|X} \parallel P_Y \mid Q_X)$$

Proof (Hints). Use convexity of relative entropy. □

Proof. By convexity of relative entropy,

$$\begin{aligned} D(Q_{Y|X,Z} \parallel P_Y \mid Q_{X,Z}) &= \sum_{x,z} Q_{X,Z}(x,z) D(Q_{Y|X=x,Z=z} \parallel P_Y) \\ &= \sum_x Q(x) \sum_z Q(z|x) D(Q_{Y|X=x,Z=z} \parallel P_Y) \\ &\geq \sum_x Q(x) D\left(\sum_z Q(z|x) Q_{Y|X=x,Z=z} \parallel P_Y\right) \\ &= \sum_x Q(x) D(Q_{Y|X=x} \parallel P_Y) \\ &= D(Q_{Y|X} \parallel P_Y \mid Q_X). \end{aligned}$$

□

Theorem 4.18 (Han's Inequality for Relative Entropy) Suppose Q, P are PMFs on A^n , and assume that $P = P_1 \otimes \cdots \otimes P_n$. Then

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

Equivalently,

$$D(Q \parallel P) \leq \sum_{i=1}^n D(Q_{X_i|X^{(i)}} \parallel P_{X_i} \mid Q_{X^{(i)}})$$

(this is tensorisation of $D(\cdot \parallel \cdot)$).

Remark 4.19 Taking P to be uniform in Han's Inequality for Relative Entropy gives Han's Inequality for Shannon entropy.

Proof (Hints). Explain why $D(Q \parallel P) = D(Q_{X^{(i)}} \parallel P_{X^{(i)}}) + D(Q_{X_i | X^{(i)}} \parallel P_{X_i | Q_{X^{(i)}}})$, then use Lemma 4.17. \square

Proof. By the Chain Rule for Relative Entropy and Lemma 4.17,

$$\begin{aligned} D(Q \parallel P) &= D(Q_{X^{(i)}} \parallel P_{X^{(i)}}) + D(Q_{X_i | X^{(i)}} \parallel P_{X_i | Q_{X^{(i)}}}) \\ &= D(Q_{X^{(i)}} \parallel P_{X^{(i)}}) + D(Q_{X_i | X^{(i)}} \parallel P_{X_i} \mid Q_{X^{(i)}}) \quad \text{since } P \text{ is a product distribution} \\ &\geq D(Q_{X^{(i)}} \parallel P_{X^{(i)}}) + D(Q_{X_i | X_{1:(i-1)}} \parallel P_{X_i} \mid Q_{X_{1:(i-1)}}) \end{aligned}$$

Summing over i , we obtain $nD(Q \parallel P) \geq \sum_{i=1}^n D(Q_{X^{(i)}} \parallel P_{X^{(i)}}) + D(Q \parallel P)$ by the Chain Rule for Relative Entropy, hence

$$\begin{aligned} D(Q \parallel P) &\geq \frac{1}{n-1} \sum_{i=1}^n D(Q_{X^{(i)}} \parallel P_{X^{(i)}}) \\ &= \frac{1}{n-1} \sum_{i=1}^n (D(Q \parallel P) - D(Q_{X_i | X^{(i)}} \parallel P_{X_i} \mid Q_{X^{(i)}})) \\ \Leftrightarrow \frac{n}{n-1} D(Q \parallel P) - D(Q \parallel P) &\leq \frac{1}{n-1} \sum_{i=1}^n D(Q_{X_i | X^{(i)}} \parallel P_{X_i} \mid Q_{X^{(i)}}) \end{aligned}$$

\square

Definition 4.20 There is another notion of entropy. Let $Z \geq 0$ almost surely. Let $\varphi(x) = x \log x$ for $x > 0$ and $\varphi(0) = 0$. The **entropy** of Z is defined as

$$\text{Ent}(Z) = \mathbb{E}[\varphi(Z)] - \varphi(\mathbb{E}[Z]),$$

Note the similarity to the definition $\text{Var}(Z) = \mathbb{E}[Z^2] - \mathbb{E}[Z]^2$. Also, since φ is convex, $\text{Ent}(Z)$ is non-negative by Jensen's inequality.

Proposition 4.21 Let $X \sim P$, where $Q \ll P$ are PMFs on a countable alphabet A . Let $Z = \frac{Q(X)}{P(X)}$. Then

$$\text{Ent}(Z) = D(Q \parallel P).$$

Proof (Hints). Straightforward. \square

Proof. We have

$$\begin{aligned} \text{Ent}(Z) &= \mathbb{E}_P \left[\frac{Q(X)}{P(X)} \log \frac{Q(X)}{P(X)} \right] - \left(\mathbb{E}_P \frac{Q(X)}{P(X)} \right) \log \mathbb{E}_P \left[\frac{Q(X)}{P(X)} \right] \\ &= D(Q \parallel P) - 1 \log 1 = D(Q \parallel P). \end{aligned}$$

\square

Remark 4.22 In general, when Z is the Radon-Nikodym derivative $\frac{dQ}{dP}(X)$ and $X \sim P$, then $\text{Ent}(Z) = D(Q \parallel P)$.

Theorem 4.23 (Tensorisation of Entropy) Let X_1, \dots, X_n be independent RVs taking values in a countable set A , and let $f : A^n \rightarrow \mathbb{R}_{\geq 0}$. Let $Z = f(X_{1:n}) = f(X)$. Then

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

where

$$\begin{aligned} \text{Ent}^{(i)}(Z) &= E^{(i)}[Z \log Z] - E^{(i)}[Z] \log E^{(i)}[Z] \\ &= \mathbb{E}[Z \log Z \mid X^{(i)}] - \mathbb{E}[Z \mid X^{(i)}] \log \mathbb{E}[Z \mid X^{(i)}]. \end{aligned}$$

Remark 4.24 Tensorisation of Entropy is analogous to the Efron-Stein Inequality.

Proof (Hints).

- Let $X \sim P = P_1 \otimes \cdots \otimes P_n$. Let $Q(x) = f(x)P(x)$.
- Show that $\text{Ent}(aZ) = a \text{Ent}(Z)$, and so can assume WLOG that $\mathbb{E}[Z] = 1$, so Q is PMF.
- Use Han's Inequality for Relative Entropy on Q and P .
- Show that

$$Q_{X_i \mid X^{(i)}}(x_i \mid x^{(i)}) = \frac{P(x_i)f(x)}{\mathbb{E}[f(X) \mid X^{(i)} = x^{(i)}]}.$$

- Show that $Q^{(i)}(x^{(i)}) = P^{(i)}(x^{(i)})\mathbb{E}[f(X) \mid X^{(i)} = x^{(i)}]$, and so that $\mathbb{E}[D(Q_{X_i \mid X^{(i)}} \parallel P_{X_i} \mid Q_{X^{(i)}})] = \mathbb{E}_P[\text{Ent}^{(i)}(f(X))]$.

□

Proof. Let $X \sim P = P_1 \otimes \cdots \otimes P_n$. Let $Q(x) = f(x)P(x)$. Since

$$\text{Ent}(aZ) = a\mathbb{E}[Z \log Z] + a\mathbb{E}[Z \log a] - a\mathbb{E}[Z] \log \mathbb{E}[Z] - a\mathbb{E}[Z] \log a = a \text{Ent}(Z),$$

we may assume WLOG that $\mathbb{E}[Z] = 1$, and so Q is a valid PMF. By Han's Inequality for Relative Entropy,

$$D(Q \parallel P) \leq \sum_{i=1}^n \mathbb{E}[D(Q_{X_i \mid X^{(i)}} \parallel P_{X_i} \mid Q_{X^{(i)}})]$$

Now

$$\begin{aligned} Q_{X_i \mid X^{(i)}}(x_i \mid x^{(i)}) &= \frac{Q_X(x)}{Q_{X^{(i)}}(x^{(i)})} = \frac{P(x)f(x)}{\sum_{x'_i \in A} Q(x_{1:(i-1)}, x'_i, x_{(i+1):n})} \\ &= \frac{P_i(x_i)P^{(i)}(x^{(i)})f(x)}{\sum_{x'_i \in A} P_i(x'_i)P^{(i)}(x^{(i)})f(x^{(i)}, x'_i)} \\ &= \frac{P_i(x_i)f(x)}{\mathbb{E}[f(X) \mid X^{(i)} = x^{(i)}]} \end{aligned}$$

(write $f(x^{(i)}, x'_i) = f(x_{1:(i-1)}, x'_i, x_{(i+1):n})$). By definition,

$$\mathbb{E}[D(Q_{X_i \mid X^{(i)}} \parallel P_{X_i} \mid Q_{X^{(i)}})]$$

$$= \sum_{x^{(i)} \in A^{n-1}} Q^{(i)}(x^{(i)}) \sum_{x_i \in A} \frac{P_i(x_i) f(x)}{\mathbb{E}[f(X) \mid X^{(i)} = x^{(i)}]} \log \frac{f(x)}{\mathbb{E}[f(X) \mid X^{(i)} = x^{(i)}]}$$

But $Q^{(i)}(x^{(i)}) = P^{(i)}(x^{(i)}) \mathbb{E}[f(X) \mid X^{(i)} = x^{(i)}]$. So,

$$\begin{aligned} & \mathbb{E}[D(Q_{X_i \mid X^{(i)}} \parallel P_{X_i} \mid Q_{X^{(i)}})] \\ &= \sum_{x^{(i)} \in A^{n-1}} P^{(i)}(x^{(i)}) \left(\sum_{x_i \in A} P_i(x_i) f(x) \log f(x) - \mathbb{E}[f(X) \mid X^{(i)} = x^{(i)}] \log \mathbb{E}[f(X) \mid X^{(i)} = x^{(i)}] \right) \\ &= \sum_{x^{(i)} \in A^{n-1}} P^{(i)}(x^{(i)}) (\mathbb{E}[f(X) \log f(X) \mid X^{(i)} = x^{(i)}] - \mathbb{E}[f(X) \mid X^{(i)} = x^{(i)}] \log \mathbb{E}[f(X) \mid X^{(i)} = x^{(i)}]) \\ &= \mathbb{E}_P[\text{Ent}^{(i)}(f(X))] \end{aligned}$$

So $\text{Ent}(f(X)) = D(Q \parallel P) \leq \sum_{i=1}^n \mathbb{E}[\text{Ent}^{(i)}(f(X))]$. \square

4.2. Herbst's argument

Theorem 4.25 (Herbst's Argument) Let Z be a real-valued RV such that $\mathbb{E}[e^{\lambda Z}] < \infty$ for all $\lambda > 0$. Suppose there exists $\nu > 0$ such that for all $\lambda > 0$,

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

Then

$$\psi_{Z - \mathbb{E}[Z]}(\lambda) = \log \mathbb{E}[e^{\lambda(Z - \mathbb{E}[Z])}] \leq \lambda^2 \frac{\nu}{2} \quad \forall \lambda > 0.$$

Proof (Hints).

- Show that $\frac{\text{Ent}(e^{\lambda Z})}{\mathbb{E}[e^{\lambda Z}]} = \lambda^2 G'(\lambda)$, where $G(\lambda) = \frac{1}{\lambda} \psi_{Z - \mathbb{E}[Z]}(\lambda)$.
- Given an upper bound for $\int_0^\lambda G'(t) dt$ (explain using a Taylor expansion of $\psi_{Z - \mathbb{E}[Z]}$ why this integral is valid).

\square

Proof. Write $\psi = \psi_{Z - \mathbb{E}[Z]}$. We have

$$\begin{aligned} \text{Ent}(e^{\lambda Z}) &= \lambda \mathbb{E}[e^{\lambda Z} \cdot Z] - \mathbb{E}[e^{\lambda Z}] \log \mathbb{E}[e^{\lambda Z}] \\ &= \mathbb{E}[e^{\lambda Z}] \left(\lambda \mathbb{E} \left[\frac{Z e^{\lambda Z}}{\mathbb{E}[e^{\lambda Z}]} \right] - \log \mathbb{E}[e^{\lambda Z}] \right) \end{aligned}$$

But

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

So by the above expression for Ent ,

$$\begin{aligned}\frac{\text{Ent}(e^{\lambda Z})}{\mathbb{E}[e^{\lambda Z}]} &= [\lambda\psi'(\lambda) + \lambda\mathbb{E}[Z] - \lambda\mathbb{E}[Z] - \psi(\lambda)] \\ &= \lambda^2 \left(\frac{1}{\lambda}\psi'(\lambda) - \frac{1}{\lambda^2}\psi(\lambda) \right) = \lambda^2 G'(\lambda)\end{aligned}$$

where $G(\lambda) = \frac{1}{\lambda}\psi(\lambda)$. Also, by assumption,

$$\frac{\text{Ent}(e^{\lambda Z})}{\mathbb{E}[e^{\lambda Z}]} \leq \lambda^2 \frac{\nu}{2}$$

By Taylor's theorem, $G(\lambda) = \frac{1}{\lambda}(\psi(0) + \lambda\psi'(0) + O(\lambda^2)) = \frac{1}{\lambda}O(\lambda^2) = O(\lambda) \rightarrow 0$ as $\lambda \rightarrow 0$. Hence, we may integrate $G'(\theta)$ from 0 to λ :

$$\begin{aligned}G(\lambda) &= \int_0^\lambda G'(\theta) d\theta \leq \int_0^\lambda \frac{\nu}{2} d\theta \quad \text{since } \theta^2 G'(\theta) \leq \theta^2 \frac{\nu}{2} \\ &= \lambda \frac{\nu}{2}\end{aligned}$$

So $\psi(\lambda) \leq \lambda^2 \frac{\nu}{2}$. □

Theorem 4.26 (Bounded Differences Inequality) Let $X = (X_1, \dots, X_n)$, where the X_i are independent. Let f have bounded differences with constants c_i . Let $Z = f(X)$. Then for all $t > 0$,

$$\mathbb{P}(Z - \mathbb{E}[Z] \geq t), \mathbb{P}(Z - \mathbb{E}[Z] \leq -t) \leq e^{-2t^2 / \sum_{i=1}^n c_i^2} = e^{-t^2 / 2\nu},$$

where $\nu = \frac{1}{4} \sum_{i=1}^n c_i^2$.

Proof (Hints).

- Use [Hoeffding's Lemma](#) and an equality from the proof of [Herbst's Argument](#) to show that $\frac{\text{Ent}^{(i)}(e^{\lambda Z})}{\mathbb{E}[e^{\lambda Z} | X^{(i)}]} = \lambda\psi'_i(\lambda) - \psi_i(\lambda) \leq \frac{1}{8}\lambda^2 c_i^2$ (you should use an integral somewhere), where $\psi_i(\lambda) = \log \mathbb{E}[e^{\lambda(Z - \mathbb{E}[Z])} | X^{(i)}]$.
- Use [Tensorisation of Entropy](#) and [Herbst's Argument](#) to show that $Z - \mathbb{E}[Z]$ has sub-Gaussian right tail with parameter ν .
- Why does the result also hold for $-f$?

□

Proof. The first step is tensorisation of entropy: by [Tensorisation of Entropy](#), we have

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

Write $f_{X^{(i)}}(x_i) = f(X_{1:(i-1)}, x_i, X_{(i+1):n})$. Conditional on $X^{(i)}$, $f_{X^{(i)}}$ takes values on an interval of length $\leq c_i$ by the bounded differences property.

The second step is to apply [Hoeffding's Lemma](#). Let $\psi_i(\lambda) = \log \mathbb{E}[e^{\lambda(Z - \mathbb{E}[Z])} | X^{(i)}]$. As in the proof of [Herbst's Argument](#), we have

$$\frac{\text{Ent}(e^{\lambda Z})}{\mathbb{E}[e^{\lambda Z}]} = \lambda \psi'_{Z-\mathbb{E}[Z]}(\lambda) - \psi_{Z-\mathbb{E}[Z]}(\lambda).$$

Note that this holds for the random variable $Z \mid X^{(i)} = x^{(i)}$, for any value of $x^{(i)}$. By [Hoeffding's Lemma](#), we have $\psi''_i(\lambda) \leq c_i^2/4$, and so

$$\begin{aligned} \frac{\text{Ent}^{(i)}(e^{\lambda Z})}{\mathbb{E}[e^{\lambda Z} \mid X^{(i)}]} &= \lambda \psi'_i(\lambda) - \psi_i(\lambda) = \int_0^\lambda \theta \psi''_i(\theta) d\theta \\ &\leq \int_0^\lambda \theta \frac{c_i^2}{4} d\theta \\ &= \frac{1}{8} \lambda^2 c_i^2 \end{aligned}$$

The third step is using [Herbst's Argument](#): we have

$$\begin{aligned} \text{Ent}(e^{\lambda Z}) &\leq \mathbb{E} \left[\sum_{i=1}^n \text{Ent}^{(i)}(e^{\lambda Z}) \right] \leq \mathbb{E} \left[\sum_{i=1}^n \frac{1}{8} \lambda^2 c_i^2 \mathbb{E}[e^{\lambda Z} \mid X^{(i)}] \right] \\ &= \frac{1}{2} \lambda^2 \cdot \frac{1}{4} \sum_{i=1}^n c_i^2 \mathbb{E}[e^{\lambda Z}] \end{aligned}$$

by [Law of Total Expectation](#). By [Herbst's Argument](#), we have

$$\psi_{Z-\mathbb{E}[Z]}(\lambda) \leq \frac{\lambda^2 \nu}{2} \quad \forall \lambda > 0,$$

and so the [Chernoff Bound](#) gives $\mathbb{P}(Z - \mathbb{E}[Z]) \leq e^{-t^2/2\nu}$. Now noting that $-f$ also has bounded differences with the same constants, we obtain the left-tail bound. \square

4.3. Log-Sobolev inequalities on the hypercube

Notation 4.27 Let X_1, \dots, X_n be IID and uniform on $\{-1, 1\}$, so $X = X_{1:n}$ is uniform on the hypercube $\{-1, 1\}^n$. Let $Z = f(X)$. By [Efron-Stein Inequality](#), $\text{Var}(Z) \leq \frac{1}{2} \mathbb{E} \left[\sum_{i=1}^n (Z - Z'_i)^2 \right] =: \nu$, where $Z'_i = f(X_{1:(i-1)}, X'_i, X_{(i+1):n})$ and X'_i is an independent copy of X_i . Define $\mathcal{E}(f)$ as

$$\begin{aligned} \nu &= \frac{1}{4} \mathbb{E} \left[\sum_{i=1}^n (f(X) - f(\bar{X}^{(i)}))^2 \right] \\ &= \frac{1}{2} \mathbb{E} \left[\sum_{i=1}^n (f(X) - f(\bar{X}^{(i)}))_+^2 \right] =: \mathcal{E}(f), \end{aligned}$$

where $\bar{X}^{(i)} = (X_{1:(i-1)}, -X_i, X_{(i+1):n})$. $\frac{1}{2} (f(X) - f(\bar{X}^{(i)}))$ looks like a discrete partial derivative in the i -th direction. So $\mathcal{E}(f)$ is a discrete analogue of $\mathbb{E}[\|\nabla f(X)\|^2]$.

Theorem 4.28 (Log-Sobolev Inequality for Bernoullis) Let X be uniformly distributed on $\{-1, 1\}^n$ and $f : \{-1, 1\}^n \rightarrow \mathbb{R}$. Then

$$\text{Ent}(f^2(X)) \leq 2 \cdot \mathcal{E}(f).$$

Proof (Hints).

- Use [Tensorisation of Entropy](#) to show that it is enough to prove the result for $n = 1$.
- Based on the one-dimensional inequality that needs to be shown, construct a suitable function $h(a, b)$. Let $g(a) = h(a, b)$ for fixed b . Show that $g(b) = 0$, $g'(b) = 0$, and $g''(a) \leq 0$ for all $a \geq b$.

□

Proof. Let $Z = f(X)$. By [Tensorisation of Entropy](#),

$$\text{Ent}(Z^2) \leq \mathbb{E} \left[\sum_{i=1}^n \text{Ent}^{(i)}(Z^2) \right]$$

If the result was true for $n = 1$, then we would have $\text{Ent}^{(i)}(Z^2) \leq \frac{1}{2} \left(f(X) - f(\overline{X}^{(i)}) \right)^2$ (since when $X^{(i)}$ is fixed, we may think of Z^2 as being a function of X_i , and this function is $f(X)^2$ or $f(\overline{X}^{(i)})^2$ with equal probability) and so $\text{Ent}(Z^2) \leq 2\mathcal{E}(f)$. So it suffices to prove the $n = 1$ case. Let $f(1) = a$, $f(-1) = b$. In the $n = 1$ case, the inequality we want to show is

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

We may assume $a, b \geq 0$, since $\frac{(b-a)^2}{2} \geq \frac{(|b|-|a|)^2}{2}$. Also, by symmetry, WLOG we assume $a \geq b$. For fixed $b \geq 0$, define

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

Since $h(b) = 0$, it is enough to show that $h'(b) = 0$ and $h''(a) \leq 0$ (so h is concave). We have

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

Hence, $h'(b) = 0$. Also,

$$h''(a) = 1 + \log \frac{2a^2}{a^2 + b^2} - \frac{2a^2}{a^2 + b^2} \leq 0,$$

since $\log x \leq x - 1$. □

Remark 4.29 [Log-Sobolev Inequality for Bernoullis](#) is stronger than [Efron-Stein Inequality](#). Also, the constant 2 on the RHS is tight.

Theorem 4.30 (Gaussian Log-Sobolev Inequality) Let $X = (X_1, \dots, X_n)$ be a vector of n independent RVs with each $X_i \sim N(0, 1)$, let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be continuously differentiable. Then

$$\text{Ent}(f^2(X)) \leq 2 \cdot \mathbb{E}[\|\nabla f(X)\|^2].$$

Proof. Exercise (use tensorisation and the central limit theorem). \square

Definition 4.31 $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is **L -Lipschitz** if

$$|f(x) - f(y)| \leq L \cdot \|x - y\| \quad \forall x, y \in \mathbb{R}^n.$$

An L -Lipschitz function f satisfies $\|\nabla f(x)\| \leq L$ for all $x \in \mathbb{R}^n$.

Theorem 4.32 (Gaussian Concentration Inequality) Let $X = (X_1, \dots, X_n)$ be a vector of n independent RVs with each $X_i \sim N(0, 1)$. Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be L -Lipschitz and $Z = f(X)$. Then $Z - \mathbb{E}[Z] \in \mathcal{G}(L^2)$, i.e. for all $\lambda \in \mathbb{R}$,

$$\psi_{Z - \mathbb{E}[Z]}(\lambda) \leq \frac{\lambda^2 L^2}{2},$$

and so for all $t > 0$,

$$\mathbb{P}(Z - \mathbb{E}[Z] \geq t) \leq e^{-t^2/2L^2}, \quad \text{and} \quad \mathbb{P}(Z - \mathbb{E}[Z] \leq -t) \leq e^{-t^2/2L^2}.$$

Note that these bounds are independent of the dimension n .

Proof (Hints).

- Explain why we can assume f is continuously differentiable (think sequences).
- Use that $\|\nabla f(X)\| \leq L$ and the Gaussian Log-Sobolev Inequality on $e^{\lambda f/2}$ to obtain an upper bound that is a suitable assumption for Herbst's Argument.

\square

Proof. WLOG, we can assume f is continuously differentiable (otherwise, we can approximate f with a sequence of continuously differentiable functions which converge to f). Note that $\|\nabla f(X)\| \leq L$. By the Gaussian Log-Sobolev Inequality for $e^{\lambda f/2}$, we have

$$\begin{aligned} \text{Ent}(e^{\lambda f(X)}) &\leq 2 \cdot \mathbb{E} \left[\left\| \nabla e^{\lambda f(X)/2} \right\|^2 \right] \\ &= 2 \cdot \mathbb{E} \left[\left\| \frac{\lambda}{2} \nabla(f(X)) \cdot e^{\lambda f(X)/2} \right\|^2 \right] \\ &= \frac{\lambda^2}{2} \mathbb{E} [e^{\lambda f(X)} \|\nabla f(X)\|^2] \\ &\leq \frac{\lambda^2 L^2}{2} \mathbb{E} [e^{\lambda f(X)}] \end{aligned}$$

So by Herbst's Argument,

$$\psi_{Z - \mathbb{E}[Z]}(\lambda) \leq \frac{\lambda^2 L^2}{2},$$

and the Chernoff Bound gives the right tail bound. The left tail bound follows from the fact that $-f$ is also L -Lipschitz. \square

Theorem 4.33 (Concentration on the Hypercube) Let $f : \{-1, 1\}^n \rightarrow \mathbb{R}$ and let $X = (X_1, \dots, X_n)$ be uniform on $\{-1, 1\}^n$. Let $Z = f(X)$ and assume

$$\max_{x \in \{-1,1\}^n} \sum_{i=1}^n (f(x) - f(\bar{x}^{(i)}))_+^2 \leq \nu$$

for some $\nu > 0$. Then for all $t > 0$,

$$\mathbb{P}(Z - \mathbb{E}[Z] \geq t) \leq e^{-t^2/\nu},$$

i.e. Z has a sub-Gaussian right tail with variance parameter $\nu/2$.

Proof (Hints).

- Explain why $\frac{e^{z/2} - e^{y/2}}{(z-y)/2} \leq e^{z/2}$ for $z > y$.
- Use the [Log-Sobolev Inequality for Bernoullis](#) on an appropriate function to obtain an upper bound that is a suitable assumption for [Herbst's Argument](#).

□

Proof. We use the [Log-Sobolev Inequality for Bernoullis](#) for the function $e^{\lambda f/2}$: for $\lambda > 0$, we have

$$\begin{aligned} \text{Ent}(e^{\lambda f(X)}) &\leq \frac{1}{2} \mathbb{E} \left[\sum_{i=1}^n \left(e^{\lambda f(X)/2} - e^{\lambda f(\bar{X}^{(i)}/2} \right)^2 \right] \\ &= \mathbb{E} \left[\sum_{i=1}^n \left(e^{\lambda f(X)/2} - e^{\lambda f(\bar{X}^{(i)}/2} \right)_+^2 \right] \end{aligned}$$

Since for $z > y$, $\frac{e^{z/2} - e^{y/2}}{(z-y)/2} \leq e^{z/2}$ (by convexity of \exp),

$$\begin{aligned} \text{Ent}(e^{\lambda f(X)}) &\leq \mathbb{E} \left[\sum_{i=1}^n \frac{\lambda^2}{2^2} \left(f(X) - f(\bar{X}^{(i)}) \right)_+^2 \cdot e^{\lambda f(X)} \right] \\ &\leq \frac{\nu \lambda^2}{4} \mathbb{E}[e^{\lambda f(X)}]. \end{aligned}$$

By [Herbst's Argument](#), we thus have $\psi_{Z - \mathbb{E}[Z]}(\lambda) \leq \frac{\lambda^2 \nu/2}{2}$ for all $\lambda > 0$, and the [Chernoff Bound](#) gives $\mathbb{P}(Z - \mathbb{E}[Z] \geq t) \leq e^{-t^2/\nu}$. □

Remark 4.34

- If the same condition for the negative part $(\cdot)_-$ holds, then we get the analogous left tail bound.
- If $\max_{x \in \{-1,1\}^n} \sum_{i=1}^n (f(x) - f(\bar{x}^{(i)}))^2 \leq \nu$, then $Z - \mathbb{E}[Z] \in \mathcal{G}(\nu/2)$. In fact, more careful analysis shows that $Z - \mathbb{E}[Z] \in \mathcal{G}(\nu/4)$.
- If f has bounded differences with constants c_i where $\sum_{i=1}^n c_i^2 \leq \nu$, then f also satisfies

$$\max_{x \in \{-1,1\}^n} \sum_{i=1}^n (f(x) - f(\bar{x}^{(i)}))^2 \leq \nu$$

so $Z - \mathbb{E}[Z] \in \mathcal{G}(\nu/4)$. [Bounded Differences Inequality](#) also gives $Z - \mathbb{E}[Z] \in \mathcal{G}(\nu/4)$ under stronger assumptions. So we are able to prove a result that is as strong as [Bounded Differences Inequality](#) but under a weaker assumption.

- The [Efron-Stein Inequality](#) gives

$$\text{Var}(Z) \leq \mathbb{E} \left[\sum_{i=1}^n (Z - Z'_i)^2_+ \right] = \frac{1}{2} \mathbb{E} \left[\sum_{i=1}^n (Z - \bar{Z}^{(i)})^2 \right] \leq \nu/2$$

if $\mathbb{E} \left[\sum_{i=1}^n (Z - \bar{Z}^{(i)})^2 \right] \leq \nu$. Note that this is a weaker result, but makes a weaker assumption than [Concentration on the Hypercube](#).

4.4. The modified log-Sobolev inequality (MLSI)

Lemma 4.35 (Variational Principle for Entropy) For any non-negative random variable Y ,

$$\text{Ent}(Y) = \inf_{u>0} \mathbb{E}[Y(\log Y - \log u) - (Y - u)]$$

and the infimum is achieved at $u = \mathbb{E}[Y]$.

Proof (Hints). Use the inequality $\log x \leq x - 1$ and show that the difference is non-positive for all $u > 0$. \square

Proof. We have

$$\begin{aligned} \text{Ent}(Y) - \mathbb{E}[Y \log Y + Y \log u - (Y - u)] &= \mathbb{E} \left[Y \log \frac{u}{\mathbb{E}[Y]} + Y - u \right] \\ &\leq \frac{\mathbb{E}[Y]}{\mathbb{E}[Y]} u - \mathbb{E}[Y] + \mathbb{E}[Y] - u = 0 \end{aligned}$$

since $\log x \leq x - 1$. For $u = \mathbb{E}[Y]$,

$$\mathbb{E}[Y \log Y] - \mathbb{E}[Y \log u + Y - u] = \text{Ent}(Y).$$

\square

Remark 4.36 This is an entropy analogue of $\text{Var}(Y) = \inf_{a \in \mathbb{R}} \mathbb{E}[(Y - a)^2]$. In fact, for any convex function φ , we can prove that the infimum

$$\inf_{u>0} \mathbb{E}[\varphi(Y) - \varphi(u) - \varphi'(u)(Y - u)]$$

is achieved when $u = \mathbb{E}[Y]$. The [Variational Principle for Entropy](#) is a special case for $\varphi(x) = x \log x$.

Theorem 4.37 (Modified Log-Sobolev Inequality) Let X_1, \dots, X_n be independent RVs taking values on A . Let $f : A^n \rightarrow \mathbb{R}$ and $Z = f(X)$. Let $f_i : A^{n-1} \rightarrow \mathbb{R}$ be an arbitrary function and $Z_i = f_i(X^{(i)})$ for each $i \in [n]$. Then

$$\text{Ent}(e^{\lambda Z}) \leq \sum_{i=1}^n \mathbb{E}[e^{\lambda Z} \varphi(-\lambda(Z - Z_i))] \quad \forall \lambda > 0,$$

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

For $\lambda > 0$ and $Z \geq Z_i$, we may use the inequality $\varphi(-x) \leq x^2/2$ for $x \geq 0$ to give a simpler upper bound:

$$\text{Ent}(e^{\lambda Z}) \leq \frac{\lambda^2}{2} \sum_{i=1}^n \mathbb{E}[e^{\lambda Z} (Z - Z_i)^2].$$

Proof (Hints). Use [Tensorisation of Entropy](#) and the [Variational Principle for Entropy](#), with $u = Y_i = e^{\lambda Z_i}$ (conditional on $X^{(i)}$). \square

Proof. Let $Y = e^{\lambda Z}$ and $Y_i = e^{\lambda Z_i}$. By [Tensorisation of Entropy](#),

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

We will bound each of the n terms on the RHS. Conditional on $X^{(i)}$, take $u = Y_i$ (note that $u > 0$). By the [Variational Principle for Entropy](#),

$$\begin{aligned} \text{Ent}^{(i)}(Y) &\leq \mathbb{E} \left[Y \log \frac{Y}{Y_i} - (Y - Y_i) \mid X^{(i)} \right] \\ &= \mathbb{E} [e^{\lambda Z} \lambda (Z - Z_i) - (e^{\lambda Z} - e^{\lambda Z_i}) \mid X^{(i)}] \\ &= \mathbb{E} [e^{\lambda Z} (\lambda (Z - Z_i) + e^{-\lambda (Z - Z_i)} - 1) \mid X^{(i)}] \\ &= \mathbb{E} [e^{\lambda Z} \varphi(-\lambda (Z - Z_i)) \mid X^{(i)}]. \end{aligned}$$

The result follows by summing and taking expectations. \square

Theorem 4.38 (Relaxed Bounded Differences) Let $Z = f(X_1, \dots, X_n)$ for independent RVs X_1, \dots, X_n which take values on A and $f : A^n \rightarrow \mathbb{R}$. Let

$$Z_i = \inf_{x'_i} f(X_{1:(i-1)}, x'_i, X_{(i+1):n}).$$

Suppose that

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

almost surely for some $\nu > 0$. Then for all $t > 0$,

$$\mathbb{P}(Z - \mathbb{E}[Z] \geq t) \leq e^{-t^2/2\nu}.$$

Proof (Hints). By the [Modified Log-Sobolev Inequality](#). \square

Proof. By the [Modified Log-Sobolev Inequality](#),

$$\text{Ent}(e^{\lambda Z}) \leq \frac{\lambda^2}{2} \mathbb{E} \left[e^{\lambda Z} \sum_{i=1}^n (Z - Z_i)^2 \right] \leq \frac{\lambda^2 \nu}{2} \mathbb{E}[e^{\lambda Z}]$$

The result follows by [Herbst's Argument](#) and the [Chernoff Bound](#). \square

Remark 4.39 If $Z_i = \sup_{x'_i} f(X_{1:(i-1)}, x'_i, X_{(i+1):n})$ and $\sum_{i=1}^n (Z - Z_i)^2 \leq \nu$, then we also obtain a left tail bound. If this condition holds for the supremum and the infimum, then $Z - \mathbb{E}[Z] \in \mathcal{G}(\nu)$.

4.5. Concentration of convex Lipschitz functions

Let $f : [0, 1]^n \rightarrow \mathbb{R}$ be separately convex and 1-Lipschitz. The [Convex Poincaré Inequality](#) says that $\text{Var}(f(X)) \leq \mathbb{E}[\|\nabla f(X)\|^2] \leq 1$.

Theorem 4.40 Let $f : [0, 1]^n \rightarrow \mathbb{R}$ be separately convex and 1-Lipschitz. Let $Z = f(X_1, \dots, X_n)$ where X_1, \dots, X_n are independent and are supported on $[0, 1]$. Then for all $t > 0$,

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

so $Z - \mathbb{E}[Z]$ has a sub-Gaussian right tail.

Proof (Hints).

- You may assume the partial derivatives of f exist.
- Find an appropriate upper bound for $\sum_{i=1}^n \left(f(X) - f(X'_{(i)})\right)^2$, where $X'_{(i)} = (X_{1:(i-1)}, X'_i, X_{(i+1):n})$ and X'_i is the value for which the infimum is achieved (why is the infimum achieved?).
- Conclude using [Relaxed Bounded Differences](#).

□

Proof. We may assume the partial derivatives of f exist (by approximating f as a sequence of differentiable functions if necessary). By [Relaxed Bounded Differences](#), it is enough to show that $\sum_{i=1}^n (Z - Z_i)^2 \leq 1$, where $Z_i = \inf_{x'_i} f(X_{1:(i-1)}, x'_i, X_{(i+1):n})$. We have

$$\sum_{i=1}^n (Z - Z_i)^2 = \sum_{i=1}^n \left(f(X) - f(X'_{(i)})\right)^2,$$

where $X'_{(i)} = (X_{1:(i-1)}, X'_i, X_{(i+1):n})$ and X'_i is the value for which the infimum is achieved. (The infimum is achieved since f is continuous and $[0, 1]$ is compact) By convexity and the fact that X'_i is a minimiser (so $f(X'_{(i)}) \leq f(X)$),

$$\begin{aligned} \sum_{i=1}^n \left(f(X) - f(X'_{(i)})\right)^2 &\leq \sum_{i=1}^n (X_i - X'_i)^2 \left(\frac{\partial}{\partial x_i} f(X)\right)^2 \\ &\leq \sum_{i=1}^n \left(\frac{\partial}{\partial x_i} f(X)\right)^2 \\ &= \|\nabla f(X)\|^2 \leq 1 \end{aligned}$$

since f is 1-Lipschitz. □

Remark 4.41 The proof wouldn't work for a left-tail bound, since $-f$ is concave not convex. The entropy method does not seem to give a left tail.

Remark 4.42 The naive bound using just the Lipschitz-ness of f would give $\sum_{i=1}^n (Z - Z_i)^2 \leq n$, so convexity gives a big improvement.

5. The transport method

Definition 5.1 Let Ω be a countable set and \mathcal{A} be a collection of subsets of Ω which is a σ -algebra. A **probability space** is (Ω, \mathcal{A}, P) , where P is a probability measure.

Definition 5.2 A **real-valued RV** Z is a map $\Omega \rightarrow \mathbb{R}$. We define

$$\mathbb{P}(Z \in A) = \sum_{\omega \in \Omega: Z(\omega) \in A} P(\omega)$$

for $A \subseteq \mathbb{R}$. We define $\mathbb{E}[Z] = \sum_{\omega \in \Omega} P(\omega)Z(\omega)$. If $Q \ll P$, write $\mathbb{E}_Q[Z] = \sum_{\omega \in \Omega} Q(\omega)Z(\omega)$.

Theorem 5.3 (Variational Representation for log-MGF and Relative Entropy) Let (Ω, \mathcal{A}, P) be a countable probability space and Z be a random variable with $\mathbb{E}[|Z|] < \infty$. Then

$$\log \mathbb{E}[e^Z] = \log \mathbb{E}_P[e^Z] = \sup_{Q \ll P} (\mathbb{E}_Q[Z] - D(Q \parallel P))$$

where the supremum is taken over all probability measures Q that are absolutely continuous with respect to P such that $\mathbb{E}_Q[|Z|] < \infty$.

Conversely, fix $Q \ll P$. Then

$$D(Q \parallel P) = \sup_Z (\mathbb{E}_Q[Z] - \log \mathbb{E}_P[e^Z]),$$

where the supremum is over all RVs Z such that $\mathbb{E}_P[|Z|], \mathbb{E}_Q[|Z|] < \infty$.

Proof (Hints).

- For first statement, define

$$Q^*(\omega) = \frac{e^{Z(\omega)} P(\omega)}{\mathbb{E}_P[e^Z]}$$

and show that $D(Q \parallel P) + \log \mathbb{E}_P[e^Z] - \mathbb{E}_Q[Z] = D(Q \parallel Q^*)$.

- For second statement, show that $D(Q \parallel P) \geq \mathbb{E}_Q[Z] - \log \mathbb{E}[e^Z]$ for any $Q \ll P$ and Z , with equality if $Z(\omega) = \log \frac{Q(\omega)}{P(\omega)}$.

□

Proof. Define

$$Q^*(\omega) = \frac{e^{Z(\omega)} P(\omega)}{\mathbb{E}_P[e^Z]}.$$

Note that Q^* is a valid PMF. For any $Q \ll P$ such that $\mathbb{E}_Q[|Z|] < \infty$, we have

$$\begin{aligned} 0 &\leq D(Q \parallel Q^*) \\ &= \mathbb{E}_{Y \sim Q} \left[\log \frac{Q(Y)}{Q^*(Y)} \right] \\ &= \mathbb{E}_{Y \sim Q} \left[\log \left(\frac{Q(Y)}{P(Y)} \frac{P(Y)}{Q^*(Y)} \right) \right] \end{aligned}$$

$$\begin{aligned}
&= \mathbb{E}_{Y \sim Q} \left[\log \frac{Q(Y)}{P(Y)} \right] + \mathbb{E}_Q \left[\log \frac{P(Y) \mathbb{E}_{Z \sim P}[e^Z]}{P(Y) e^Z} \right] \\
&= D(Q \parallel P) + \log \mathbb{E}_P[e^Z] - \mathbb{E}_Q[Z]
\end{aligned}$$

Hence $\log \mathbb{E}[e^Z] \geq \mathbb{E}_Q Z - D(Q \parallel P)$, with equality iff $Q = Q^*$. The result follows since $Q^* \ll P$. For the second statement, note that $D(Q \parallel P) \geq \mathbb{E}_Q[Z] - \log \mathbb{E}[e^Z]$, for any $Q \ll P$ and Z . There is equality if $Z(\omega) = \log \frac{Q(\omega)}{P(\omega)}$. (Note that $\mathbb{E}_Q[|Z|] = \mathbb{E}_Q[|\log \frac{Q}{P}|] < \infty$ since $D(Q \parallel P) < \infty$ and the negative part of $x \log x$ is finitely bounded.) Note it can be shown that the result holds when $D(Q \parallel P) = \infty$ and when $\mathbb{E}_P[e^Z] = \infty$. \square

Corollary 5.4 For all $\lambda \in \mathbb{R}$, we have

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

Theorem 5.5 (Marton's Argument) Let P be a PMF and $Z \sim P$. If there exists $\nu > 0$ such that

$$\mathbb{E}_Q[Z] - \mathbb{E}_P[Z] \leq \sqrt{2\nu D(Q \parallel P)}$$

for all PMFs Q such that $Q \ll P$, then

$$\psi_{Z - \mathbb{E}[Z]}(\lambda) = \log \mathbb{E}_P[e^{\lambda(Z - \mathbb{E}_P[Z])}] \leq \frac{\lambda^2 \nu}{2} \quad \forall \lambda > 0,$$

(and so also $\mathbb{P}(Z - \mathbb{E}[Z] \geq t) \leq e^{-t^2/2\nu}$ by the [Chernoff Bound](#)). Conversely, if there exists $\nu > 0$ such that $\psi_{Z - \mathbb{E}[Z]}(\lambda) = \log \mathbb{E}_P[e^{\lambda(Z - \mathbb{E}_P[Z])}] \leq \frac{\lambda^2 \nu}{2}$ for all $\lambda > 0$, then

$$\mathbb{E}_Q[Z] - \mathbb{E}_P[Z] \leq \sqrt{2\nu D(Q \parallel P)}$$

for all $Q \ll P$.

Proof (Hints).

- Show that $\log \mathbb{E}_P[e^{\lambda(Z - \mathbb{E}[Z])}] \leq \sup_{t \geq 0} (\lambda \sqrt{2\nu t} - t)$.
- For converse, may assume that $\mathbb{E}_Q[Z] - \mathbb{E}_P[Z] \geq 0$ (why?). The proof is similar to the first proof.

\square

Proof. By the [Variational Representation for log-MGF and Relative Entropy](#),

$$\begin{aligned}
\log \mathbb{E}_P[e^{\lambda(Z - \mathbb{E}[Z])}] &= \sup_{Q \ll P} (\lambda(\mathbb{E}_Q[Z] - \mathbb{E}_P[Z]) - D(Q \parallel P)) \\
&\leq \sup_{Q \ll P} (\lambda \sqrt{2\nu D(Q \parallel P)} - D(Q \parallel P)) \\
&\leq \sup_{t \geq 0} (\lambda \sqrt{2\nu t} - t).
\end{aligned}$$

Let $f(t) = \lambda \sqrt{2\nu t} - t$. Then $f'(t) = 0$ iff $t = \frac{\lambda^2 \nu}{2}$, and so the $\sup_{t \geq 0} f(t) = \frac{\lambda^2 \nu}{2}$.

For the converse, we may assume that $\mathbb{E}_Q[Z] - \mathbb{E}_P[Z] \geq 0$, since otherwise we are trivially done. By [Variational Representation for log-MGF and Relative Entropy](#), for all $\lambda > 0$,

$$D(Q \parallel P) \geq \lambda(\mathbb{E}_Q[Z] - \mathbb{E}_P[Z]) - \log \mathbb{E}_P e^{\lambda(Z - \mathbb{E}_P[Z])} \geq \lambda(\mathbb{E}_Q[Z] - \mathbb{E}_P[Z]) - \frac{\lambda^2 \nu}{2}$$

Taking the supremum over $\lambda > 0$, we obtain

$$D(Q \parallel P) \geq \sup_{\lambda > 0} \left(\lambda(\mathbb{E}_Q[Z] - \mathbb{E}_P[Z]) - \frac{\lambda^2 \nu}{2} \right)$$

Differentiating the RHS, we see that it is maximised when $\lambda = \frac{1}{\nu}(\mathbb{E}_Q[Z] - \mathbb{E}_P[Z])$, and so

$$D(Q \parallel P) \geq \frac{(\mathbb{E}_Q[Z] - \mathbb{E}_P[Z])^2}{2\nu}.$$

□

5.1. Concentration via Marton's argument

Definition 5.6 Let P, Q be distributions on A . Let $X \sim P$ and $Y \sim Q$. A **coupling** π is a joint distribution on (X, Y) such that X has marginal P (w.r.t π) and Y has marginal Q (w.r.t. π). Write $\Pi(P, Q)$ for the set of all couplings.

Example 5.7 $P \otimes Q$ is the independent coupling.

Lemma 5.8 $f : A^n \rightarrow \mathbb{R}$ such that $f(y) - f(x) \leq \sum_{i=1}^n c_i d(x_i, y_i)$ for some constants c_i and distance $d(\cdot, \cdot)$. Let $X \sim P_1 \otimes \cdots \otimes P_n =: P$, $Z = f(X)$. Let $C > 0$ be such that

$$\inf_{\pi \in \Pi(P, Q)} \sum_{i=1}^n \mathbb{E}_\pi [d(X_i, Y_i)]^2 \leq 2CD(Q \parallel P).$$

for all $Q \ll P$. Then

$$\mathbb{P}(Z - \mathbb{E}[Z] \geq t) \leq e^{-t^2/2\nu},$$

where $\nu = C \sum_{i=1}^n c_i^2$.

Proof (Hints). Let $Q \ll P$ and $Y \sim Q$. Show that

$$\mathbb{E}_Q[Z] - \mathbb{E}_P[Z] \leq \left(\sum_{i=1}^n c_i^2 \right)^{1/2} \left(\sum_{i=1}^n \mathbb{E}_\pi [d(X_i, Y_i)]^2 \right)^{1/2},$$

and conclude the result using [Marton's Argument](#). □

Proof. Let $Q \ll P$ and $Y \sim Q$. Then

$$\begin{aligned} \mathbb{E}_Q[Z] - \mathbb{E}_P[Z] &= \mathbb{E}[f(Y)] - \mathbb{E}[f(X)] \\ &= \mathbb{E}_\pi[f(Y) - f(X)] \quad \text{for all } \pi \in \Pi(P, Q) \end{aligned}$$

$$\begin{aligned}
&\leq \mathbb{E}_\pi \left[\sum_{i=1}^n c_i d(X_i, Y_i) \right] \\
&= \sum_{i=1}^n c_i \mathbb{E}_\pi [d(X_i, Y_i)] \\
&\leq \left(\sum_{i=1}^n c_i^2 \right)^{1/2} \left(\sum_{i=1}^n \mathbb{E}_\pi [d(X_i, Y_i)]^2 \right)^{1/2} \quad \text{by Cauchy-Schwarz}
\end{aligned}$$

So

$$\mathbb{E}_Q[Z] - \mathbb{E}_P[Z] \leq \left(\sum_{i=1}^n c_i^2 \right)^{1/2} \left(\inf_{\pi \in \Pi(P, Q)} \sum_{i=1}^n \mathbb{E}_\pi [d(X_i, Y_i)]^2 \right)^{1/2}$$

Since

$$\inf_{\pi \in \Pi(P, Q)} \sum_{i=1}^n \mathbb{E}_\pi [d(X_i, Y_i)]^2 \leq 2CD(Q \parallel P)$$

we have $\mathbb{E}_Q[Z] - \mathbb{E}_P[Z] \leq \sqrt{2\nu D(Q \parallel P)}$, where $\nu = C \sum_{i=1}^n c_i^2$. The result follows by [Marton's Argument](#). \square

Definition 5.9 Let $X \sim P$ and $Y \sim Q$. The **transportation cost** from Q to P w.r.t a distance $d(\cdot, \cdot)$ is

$$\inf_{\pi \in \Pi(P, Q)} \mathbb{E}_\pi [d(X, Y)].$$

Definition 5.10 Let P and Q be distributions on the same space (Ω, \mathcal{A}) . The **total variation distance** between P and Q is

$$d_{\text{TV}}(P, Q) := \sup_{A \in \mathcal{A}} |P(A) - Q(A)|.$$

Proposition 5.11 Let $A^* = \{\omega \in \Omega : P(\omega) \geq Q(\omega)\}$. We have the alternative expressions

$$\begin{aligned}
d_{\text{TV}}(P, Q) &= \frac{1}{2} \sum_{\omega \in \Omega} |P(\omega) - Q(\omega)| = \sum_{\omega \in \Omega} (P(\omega) - Q(\omega))_+ \\
&= P(A^*) - Q(A^*) = 1 - \sum_{\omega \in \Omega} \min\{P(\omega), Q(\omega)\}.
\end{aligned}$$

Proof (Hints).

- For second equality, consider the $+$ and $-$ parts.
- For the first equality, show \leq by splitting sum over A and A^c for $A \in \mathcal{A}$, show \geq by considering $A^* = \{\omega : P(\omega) \geq Q(\omega)\}$.
- For the third equality, show the fourth expression is equal to the third.

\square

Proof. For the first inequality: for any $A \in \mathcal{A}$, by the triangle inequality,

$$\begin{aligned}
\sum_{\omega \in \Omega} |P(\omega) - Q(\omega)| &= \sum_{\omega \in A} |P(\omega) - Q(\omega)| + \sum_{\omega \in A^c} |P(\omega) - Q(\omega)| \\
&\geq P(A) - Q(A) + Q(A^c) - P(A^c) = 2(P(A) - Q(A))
\end{aligned}$$

and similarly $\sum_{\omega \in \Omega} |P(\omega) - Q(\omega)| \geq 2(Q(A) - P(A))$. Conversely,

$$\begin{aligned}
d_{\text{TV}}(P, Q) &\geq P(A^*) - Q(A^*) \\
&= \sum_{\omega \in \Omega} (P(\omega) - Q(\omega))_+ = \frac{1}{2} \sum_{\omega \in \Omega} |P(\omega) - Q(\omega)|,
\end{aligned}$$

since $\sum_{\omega \in \Omega} (P(\omega) - Q(\omega))_+ = \sum_{\omega \in \Omega} (P(\omega) - Q(\omega))_-$. For the third inequality,

$$\begin{aligned}
1 - \sum_{\omega \in \Omega} \min\{P(\omega), Q(\omega)\} &= \sum_{\omega \in \Omega} P(\omega) - \min\{P(\omega), Q(\omega)\} \\
&= \sum_{\omega \in \Omega} (P(\omega) - Q(\omega))_+
\end{aligned}$$

□

Lemma 5.12 Let P and Q be distributions on the same space. Then if $X \sim P$ and $Y \sim Q$,

$$\inf_{\pi \in \Pi(P, Q)} \mathbb{P}_\pi(X \neq Y) = d_{\text{TV}}(P, Q) \in [0, 1].$$

Proof (Hints). Show that LHS \geq RHS by taking a supremum and infimum and using that $|\mathbb{1}_{\{x \in A\}} - \mathbb{1}_{\{y \in A\}}| \leq \mathbb{1}_{\{x \neq y\}}$, then consider

$$\pi(\omega_1, \omega_2) = \begin{cases} \min\{P(\omega), Q(\omega)\} & \text{if } \omega_1 = \omega_2 = \omega \\ \frac{1}{d_{\text{TV}}(P, Q)} (P(\omega_1) - Q(\omega_1))(Q(\omega_2) - P(\omega_2)) & \text{if } (\omega_1, \omega_2) \in A^* \times (A^*)^c \\ 0 & \text{otherwise.} \end{cases}$$

□

Proof. Let $\pi \in \Pi(P, Q)$ and $A \in \mathcal{A}$. Since $|\mathbb{I}_{\{X \in A\}} - \mathbb{I}_{\{Y \in A\}}| \leq \mathbb{I}_{\{X \neq Y\}}$ We have

$$\begin{aligned}
|P(A) - Q(A)| &= |\mathbb{E}_\pi[\mathbb{I}_{\{X \in A\}} - \mathbb{I}_{\{Y \in A\}}]| \\
&\leq \mathbb{E}_\pi[|\mathbb{I}_{\{X \in A\}} - \mathbb{I}_{\{Y \in A\}}|] \\
&\leq \mathbb{E}[\mathbb{I}_{\{X \neq Y\}}] \quad \text{pointwise} \\
&= \mathbb{P}(X \neq Y).
\end{aligned}$$

Taking the supremum over all $A \in \mathcal{A}$ and the infimum over all couplings gives $d_{\text{TV}}(P, Q) \leq \inf_{\pi \in \Pi(P, Q)} \mathbb{P}(X \neq Y)$. We will construct π such that $\mathbb{P}(X \neq Y) = d_{\text{TV}}(P, Q)$. Intuitively, we want to place as much mass as possible on the “diagonal”, i.e. make $\pi(\omega, \omega)$ as large as possible.

For $(\omega_1, \omega_2) \in \Omega \times \Omega$, let

$$\pi(\omega_1, \omega_2) = \begin{cases} \min\{P(\omega), Q(\omega)\} & \text{if } \omega_1 = \omega_2 = \omega \\ \frac{1}{d_{\text{TV}}(P, Q)}(P(\omega_1) - Q(\omega_1))(Q(\omega_2) - P(\omega_2)) & \text{if } (\omega_1, \omega_2) \in A^* \times (A^*)^c \\ 0 & \text{otherwise.} \end{cases}$$

Clearly, $\mathbb{P}_\pi(X = Y) = \sum_{\omega \in \Omega} \pi(\omega, \omega) = \sum_{\omega \in \Omega} \min\{P(\omega), Q(\omega)\}$, and so by Proposition 5.11, $\mathbb{P}_\pi(X \neq Y) = 1 - \sum_{\omega \in \Omega} \min\{P(\omega), Q(\omega)\} = d_{\text{TV}}(P, Q)$. Also, π is indeed a valid coupling:

$$\begin{aligned} \sum_{\omega_1 \in \Omega} \pi(\omega_1, \omega_2) &= \sum_{\omega_1 \in A^*} (P(\omega_1) - Q(\omega_1)) \frac{Q(\omega_2) - P(\omega_2)}{d_{\text{TV}}(P, Q)} \mathbb{I}_{\{\omega_2 \in (A^*)^c\}} + \min\{P(\omega_2), Q(\omega_2)\} \\ &= Q(\omega_2), \end{aligned}$$

and similarly $\sum_{\omega_2 \in \Omega} \pi(\omega_1, \omega_2) = P(\omega_1)$. □

Definition 5.13 The minimising coupling

$$\pi(\omega_1, \omega_2) = \begin{cases} \min\{P(\omega), Q(\omega)\} & \text{if } \omega_1 = \omega_2 = \omega \\ \frac{1}{d_{\text{TV}}(P, Q)}(P(\omega_1) - Q(\omega_1))(Q(\omega_2) - P(\omega_2)) & \text{if } (\omega_1, \omega_2) \in A^* \times (A^*)^c \\ 0 & \text{otherwise.} \end{cases}$$

in the proof of Lemma 5.12 is called the **optimal total variation coupling**.

Lemma 5.14 (Pinsker's Inequality) Let P and Q be PMFs such that $Q \ll P$. Then

$$d_{\text{TV}}(P, Q)^2 \leq \frac{1}{2} D(Q \parallel P).$$

Proof (Hints). Let $Y(\omega) = \frac{Q(\omega)}{P(\omega)}$ and $Z = \mathbb{I}_{\{Y \geq 1\}}$. Use Hoeffding's Lemma and Marton's Argument. □

Proof. Let $Y(\omega) = \frac{Q(\omega)}{P(\omega)}$. Let $Z = \mathbb{I}_{\{Y \geq 1\}}$. By Hoeffding's Lemma,

$$\psi_{Z - \mathbb{E}[Z]}(\lambda) \leq \frac{\lambda^2}{8}.$$

But then by Marton's Argument,

$$\mathbb{E}_Q[Z] - \mathbb{E}_P[Z] \leq \sqrt{2 \cdot \frac{1}{4} \cdot D(Q \parallel P)},$$

i.e. $d_{\text{TV}}(P, Q) = Q(A) - P(A) \leq \sqrt{\frac{1}{2} \cdot D(Q \parallel P)}$, where $A = \{\omega \in \Omega : Q(\omega) \geq P(\omega)\}$, by Proposition 5.11. □

Theorem 5.15 (Marton's Transport Cost Inequality) Let $P = P_1 \otimes \cdots \otimes P_n$ and $Q \ll P$. Let $X \sim P$ and $Y \sim Q$. Then

$$\inf_{\pi \in \Pi(P, Q)} \sum_{i=1}^n \mathbb{E}_\pi [\mathbb{I}_{\{X_i \neq Y_i\}}]^2 = \inf_{\pi \in \Pi(P, Q)} \sum_{i=1}^n \mathbb{P}_\pi(X_i \neq Y_i)^2 \leq \frac{1}{2} D(Q \parallel P).$$

Proof. We use induction on n . The $n = 1$ case follows from Lemma 5.12 and Pinsker's Inequality. Assume that for every $n \leq k$, there exists a coupling π_n on $(X_{1:n}, Y_{1:n})$ such that $\sum_{i=1}^n \mathbb{P}(X_i \neq Y_i)^2 \leq \frac{1}{2}D(Q \parallel P)$. We will extend it to a coupling π_{k+1} on $(X_{1:(k+1)}, Y_{1:(k+1)})$. Write

$$\sum_{i=1}^{k+1} \mathbb{P}(X_i \neq Y_i)^2 = \sum_{i=1}^k \mathbb{P}(X_i \neq Y_i)^2 + \mathbb{P}(X_{k+1} \neq Y_{k+1})^2$$

For fixed $y_{1:k}$, let $\pi_{y_{1:k}} \in \Pi(P_{X_{k+1}}, Q_{Y_{k+1} | Y_{1:k}=y_{1:k}})$ be the optimal total variation coupling of X_{k+1} and $Y_{k+1} | Y_{1:k} = y_{1:k}$. Define

$$\begin{aligned} \pi_{k+1}(x_{1:(k+1)}, y_{1:(k+1)}) &:= \pi_k(x_{1:k}, y_{1:k}) \cdot \pi_{y_{1:k}}(x_{k+1}, y_{k+1}) \\ &= \mathbb{P}(X_{1:k} = x_{1:k}, Y_{1:k} = y_{1:k}) \mathbb{P}(X_{k+1} = x_{k+1}) \mathbb{P}(Y_{k+1} = y_{k+1} | X_{k+1} = x_{k+1}) \end{aligned}$$

This new coupling has two properties:

1. Given $(X_{1:k}, Y_{1:k})$, the distribution of (X_{k+1}, Y_{k+1}) depends only on $Y_{1:k}$, i.e. $X_{1:k} - Y_{1:k} - (X_{k+1}, Y_{k+1})$ form a Markov chain.
2. Also, X_{k+1} is independent of $(X_{1:k}, Y_{1:k})$.

These properties imply that $(X_{k+1}, Y_{k+1}) | X_{1:k} = x_{1:k}, Y_{1:k} = y_{1:k} \sim \pi_{y_{1:k}}$. Hence,

$$\begin{aligned} \mathbb{P}(X_{k+1} \neq Y_{k+1} | X_{1:k} = x_{1:k}, Y_{1:k} = y_{1:k}) &= d_{\text{TV}}(P_{X_{k+1}}, Q_{Y_{k+1} | Y_{1:k}=y_{1:k}}) \\ &\leq \sqrt{\frac{1}{2}D(Q_{Y_{k+1} | Y_{1:k}=y_{1:k}} \parallel P_{X_{k+1}})} \end{aligned}$$

by the $n = 1$ result. Taking expectation over π_k on the LHS gives

$$\begin{aligned} \mathbb{P}(X_{k+1} \neq Y_{k+1}) &= \mathbb{E}_{\pi_k}[\mathbb{P}(X_{k+1} \neq Y_{k+1} | X_{1:k}, Y_{1:k})] \\ &\leq \mathbb{E}_{Q_{Y_{1:k}}} \left[\sqrt{\frac{1}{2}D(Q_{Y_{k+1} | Y_{1:k}} \parallel P_{X_{k+1}})} \right] \end{aligned}$$

Squaring and using Jensen's inequality gives

$$\begin{aligned} \mathbb{P}(X_{k+1} \neq Y_{k+1})^2 &\leq \frac{1}{2} \mathbb{E}_{Q_{Y_{1:k}}} [D(Q_{Y_{k+1} | Y_{1:k}} \parallel P_{X_{k+1}})] \\ &= \frac{1}{2} D(Q_{Y_{k+1} | Y_{1:k}} \parallel P_{X_{k+1}} | Q_{Y_{1:k}}) \end{aligned}$$

By the induction hypothesis,

$$\begin{aligned} \sum_{i=1}^{k+1} \mathbb{P}(X_i \neq Y_i)^2 &\leq \frac{1}{2} (D(Q_{Y_{1:k}} \parallel P_{X_{1:k}}) + D(Q_{Y_{k+1} | Y_{1:k}} \parallel P_{X_{k+1}} | Q_{Y_{1:k}})) \\ &= \frac{1}{2} D(Q_{Y_{1:(k+1)}} \parallel P_{X_{1:(k+1)}}) \end{aligned}$$

by the Chain Rule for Relative Entropy. □

Remark 5.16 We can recover the [Bounded Differences Inequality](#) from [Marton's Transport Cost Inequality](#): the conditions of Lemma [5.8](#) are satisfied with $C = \frac{1}{4}$, since f having bounded differences with constant c_i implies

$$f(y) - f(x) \leq \sum_{i=1}^n c_i d(x_i, y_i),$$

where $d(x_i, y_i) = \mathbb{I}_{\{x_i \neq y_i\}}$. This gives the concentration bound.

5.2. Talagrand's inequality

Definition 5.17 Marton's divergence is

$$d_2^2(Q, P) = \mathbb{E}_{X \sim P} \left[\left(1 - \frac{Q(X)}{P(X)} \right)_+^2 \right] = \sum_{\omega: P(\omega) > 0} \frac{(P(\omega) - Q(\omega))_+^2}{P(\omega)}.$$

Lemma 5.18 Let P and Q be distributions on the same space (Ω, \mathcal{A}) . Then

$$\inf_{\pi \in \Pi(P, Q)} \mathbb{E}_{(X, Y) \sim \pi} [\mathbb{P}(X \neq Y \mid X)^2] = d_2^2(Q, P).$$

Proof (Hints).

- For \geq , explain why $\mathbb{P}(X = Y \mid X = x) \leq \min\{1, Q(x)/P(x)\}$, then take expectation.
- Showing equality, by showing that the optimal total variation coupling minimises the LHS, is left as an exercise.

□

Proof. We have

$$\mathbb{P}(X = Y \mid X = x) = \frac{\mathbb{P}(X = x, Y = x)}{\mathbb{P}(X = x)} \leq \min\left\{1, \frac{Q(x)}{P(x)}\right\}.$$

So for any coupling π ,

$$\mathbb{E}_\pi [\mathbb{P}(X \neq Y \mid X)^2] \geq \mathbb{E}_P \left[\left(1 - \min\left\{1, \frac{Q(X)}{P(X)}\right\} \right)^2 \right] = \mathbb{E}_P \left[\left(1 - \frac{Q(X)}{P(X)} \right)_+^2 \right] = d_2^2(Q, P).$$

Showing equality is left as an exercise.

□

Lemma 5.19 (Pinsker's Inequality for Marton Divergence) Let P, Q be distributions on the same space (Ω, \mathcal{A}) with $Q \ll P$. Then

$$d_2^2(Q, P) \leq 2D(Q \parallel P).$$

Proof (Hints).

- Let $h(t) = (1 - t) \log(1 - t) + t$ for $t \leq 1$, expression $D(Q \parallel P)$ using h (as an expectation w.r.t P).
- Show that $h(t) \geq 0$ and by considering derivatives, show that $h(t) \geq t^2/2$ for all $t \in [0, 1]$.

□

Proof. Let $h(t) = (1-t)\log(1-t) + t$ for $t \leq 1$ and $q(X) = \frac{Q(X)}{P(X)}$. Then

$$D(Q \parallel P) = \mathbb{E}_{X \sim P}[h(1 - q(X))].$$

We have $h(t) = -(1-t)\log(1 + \frac{t}{1-t}) + t \geq -t + t \geq 0$ since $\log x \leq x - 1$. Also, $h(t) \geq t^2/2$ for $t \in [0, 1]$: indeed, $h(0) = 0^2/2$, and $h'(t) = -1 - \log(1-t) + 1 = -\log(1-t)$, thus

$$\frac{d}{dt} \left(h(t) - \frac{t^2}{2} \right) = -\log(1-t) - t \geq (1-t) + 1 - t = 0.$$

So we have

$$\begin{aligned} D(Q \parallel P) &= \mathbb{E}[h(1 - q(X))] \geq \mathbb{E}[h((1 - q(X))_+)] \\ &\geq \mathbb{E} \left[\frac{(1 - q(X))_+^2}{2} \right] = \frac{1}{2} d_2^2(Q, P). \end{aligned}$$

where first inequality is since $h \geq 0$ and $h(0) = 0$. \square

Theorem 5.20 (Marton's Conditional Transport Cost Inequality) Let $X = (X_1, \dots, X_n)$, $X \sim P = P_1 \otimes \dots \otimes P_n$, and let $Q \ll P$. Then

$$\inf_{\pi \in \Pi(P, Q)} \sum_{i=1}^n \mathbb{E}_{\pi} [\mathbb{P}(X_i \neq Y_i \mid X)^2] \leq 2D(Q \parallel P).$$

Proof. We use induction on n . The $n = 1$ case follows by Lemma 5.18 and Pinsker's Inequality for Marton Divergence. Now assume that for every $n \leq k$, there exists a $\pi_n \in \Pi(P, Q)$ such that $\sum_{i=1}^n \mathbb{E}_{\pi_n} [\mathbb{P}(X_i \neq Y_i \mid X)^2] \leq 2D(Q_{X_{1:n}} \parallel P_{X_{1:n}})$. We will find a coupling π_{k+1} (extended from π_k) such that

$$\begin{aligned} \sum_{i=1}^k \mathbb{E}_{\pi_{k+1}} [\mathbb{P}(X_i \neq Y_i \mid X_{1:(k+1)})^2] + \mathbb{E}_{\pi_{k+1}} [\mathbb{P}(X_{k+1} \neq Y_{k+1} \mid X_{1:(k+1)})^2] &= \sum_{i=1}^{k+1} \mathbb{E}_{\pi_{k+1}} [\mathbb{P}(X_i \neq Y_i \mid X_{1:(k+1)})^2] \\ &\leq D(Q_{Y_{1:(k+1)}} \parallel P_{X_{1:(k+1)}}) \end{aligned}$$

For fixed $y_{1:k}$, let $\pi_{y_{1:k}}$ be the optimal total variation coupling of X_{k+1} and $Y_{k+1} \mid Y_{1:k} = y_{1:k}$. Let

$$\begin{aligned} \pi_{k+1}(x_{1:(k+1)}, y_{1:(k+1)}) &= \pi_k(x_{1:k}, y_{1:k}) \cdot \pi_{y_{1:k}}(x_{k+1}, y_{k+1}) \\ &= \mathbb{P}(X_{1:k} = x_{1:k}, Y_{1:k} = y_{1:k}) \cdot \mathbb{P}(X_{k+1} = x_{k+1}) \cdot \mathbb{P}(Y_{k+1} = y_{k+1} \mid X_{k+1} = x_{k+1}), \end{aligned}$$

where the probabilities in the last line are w.r.t. the new coupling π_{k+1} . This coupling has two properties:

- $X_{1:k} - Y_{1:k} - (X_{k+1}, Y_{k+1})$ form a Markov chain, i.e. given $(X_{1:k}, Y_{1:k})$, the distribution of (X_{k+1}, Y_{k+1}) only depends on $Y_{1:k}$.
- X_{k+1} is independent of $(X_{1:k}, Y_{1:k})$.

These properties imply that given $X_{1:k} = x_{1:k}, Y_{1:k} = y_{1:k}$, we have $(X_{k+1}, Y_{k+1}) \sim \pi_{y_{1:k}}$. By the induction hypothesis,

$$\begin{aligned} \sum_{i=1}^k \mathbb{E}_{\pi_{k+1}} \left[\mathbb{P}(X_i \neq Y_i \mid X_{1:(k+1)})^2 \right] &= \sum_{i=1}^k \mathbb{E}_{\pi_{k+1}} \left[\mathbb{P}(X_i \neq Y_i \mid X_{1:k})^2 \right] \text{ by second property} \\ &= \sum_{i=1}^k \mathbb{E}_{\pi_k} \left[\mathbb{P}(X_i \neq Y_i \mid X_{1:k})^2 \right] \\ &\leq 2D(Q_{Y_{1:k}} \parallel P_{X_{1:k}}). \end{aligned}$$

We want to show

$$\mathbb{E} \left[\mathbb{P}(X_{k+1} \neq Y_{k+1} \mid X_{1:(k+1)})^2 \right] \leq 2D(Q_{Y_{k+1} \mid Y_{1:k}} \parallel P_{X_{k+1} \mid Q_{Y_{1:k}}})$$

From the $n = 1$ case (and since the optimal total variation coupling $\pi_{y_{1:k}}$ is a minimiser in Lemma 5.18), we know that

$$\mathbb{E}_{\pi_{y_{1:k}}} \left[\mathbb{P}(X_{k+1} \neq Y_{k+1} \mid X_{k+1}, Y_{1:k} = y_{1:k})^2 \right] \leq 2D(Q_{Y_{k+1} \mid Y_{1:k} = y_{1:k}} \parallel P_{X_{k+1}}).$$

By the two properties of π_{k+1} ,

$$\mathbb{P}(X_{k+1} \neq Y_{k+1} \mid X_{k+1}, Y_{1:k} = y_{1:k}) = \mathbb{P}(X_{k+1} \neq Y_{k+1} \mid X_{1:(k+1)}, Y_{1:k} = y_{1:k})$$

Taking $\mathbb{E}_{Y_{1:k}}(\cdot)$ in the above, we obtain

$$\begin{aligned} \mathbb{E} \left[\mathbb{P}(X_{k+1} \neq Y_{k+1} \mid X_{1:(k+1)}, Y_{1:k})^2 \right] &= \mathbb{E} \left[\mathbb{P}(X_{k+1} \neq Y_{k+1} \mid X_{k+1}, Y_{k+1})^2 \right] \\ &\leq 2D(Q_{Y_{k+1} \mid Y_{1:k}} \parallel P_{X_{k+1} \mid Q_{Y_{1:k}}}) \end{aligned}$$

The LHS is equal to

$$\begin{aligned} &\mathbb{E} \left[\mathbb{E} \left[\mathbb{I}_{\{X_{k+1} \neq Y_{k+1}\}} \mid X_{1:(k+1)}, Y_{1:k} \right]^2 \mid X_{1:(k+1)} \right] \\ &\geq \mathbb{E} \left[\mathbb{E} \left[\mathbb{E} \left[\mathbb{I}_{\{X_{k+1} \neq Y_{k+1}\}} \mid X_{1:(k+1)}, Y_{1:k} \right] \mid X_{1:(k+1)} \right]^2 \right] \text{ by Jensen} \\ &= \mathbb{E} \left[\mathbb{E} \left[\mathbb{I}_{\{X_{k+1} \neq Y_{k+1}\}} \mid X_{1:(k+1)} \right]^2 \right] \text{ by tower property} \\ &= \mathbb{E} \mathbb{P}(X_{k+1} \neq Y_{k+1} \mid X_{1:(k+1)})^2 \end{aligned}$$

So

$$\begin{aligned} &\sum_{i=1}^k \mathbb{E} \mathbb{P}(X_i \neq Y_i \mid X_{1:(k+1)})^2 + \mathbb{E} \mathbb{P}(X_{k+1} \neq Y_{k+1} \mid X_{1:k})^2 \\ &\leq 2D(Q_{Y_{1:k}} \parallel P_{X_{1:k}}) + 2D(Q_{Y_{k+1} \mid Y_{1:k}} \parallel P_{X_{k+1} \mid Q_{Y_{1:k}}}) \\ &= 2D(Q \parallel P) \end{aligned}$$

by the Chain Rule for Relative Entropy. □

Definition 5.21 $f : A^n \rightarrow \mathbb{R}$ satisfies the **one-sided bounded differences** property if

$$f(y) - f(x) \leq \sum_{i=1}^n \mathbb{I}_{\{x_i \neq y_i\}} c_i(x) \quad \forall x, y \in A^n,$$

where $c_i : A^n \rightarrow \mathbb{R}_{\geq 0}$.

Remark 5.22 We can't apply results for bounded differences on functions with this property, since it is a weaker property.

Remark 5.23 By **Relaxed Bounded Differences**, if $\sum_{i=1}^n (Z_i - Z)^2 \leq \nu$, where $Z_i = \sup_{x_i} f(X_{1:(i-1)}, x_i, X_{(i+1):n})$, then $\mathbb{P}(Z - \mathbb{E}[Z] \leq -t) \leq e^{-t^2/2\nu}$. Under one-sided bounded differences,

$$0 \leq \sum_{i=1}^n (Z_i - Z)^2 \leq \sum_{i=1}^n c_i(X)^2 \leq \sup_{x \in A^n} \sum_{i=1}^n c_i(x)^2 =: \nu_\infty,$$

so we obtain the left-tail bound $\mathbb{P}(Z - \mathbb{E}[Z] \leq -t) \leq e^{-t^2/2\nu_\infty}$. But now if $Z_i = \inf_{x_i} f(X_{1:(i-1)}, x_i, X_{(i+1):n})$, with infimum achieved at $(X')^{(i)} = (X_{1:(i-1)}, x'_i, X_{(i+1):n})$, then

$$0 \leq \sum_{i=1}^n (Z - Z_i)^2 \leq \sum_{i=1}^n c_i((X')^{(i)})^2.$$

We generally can't say that this is $\leq \sup_{x \in A^n} \sum_{i=1}^n c_i(x)^2$, so can't immediately deduce a right tail bound.

However, the transport method gives us a right-tail bound with a better parameter $\nu = \mathbb{E}[\sum_{i=1}^n c_i(X)^2] \leq \nu_\infty$.

Theorem 5.24 (Talagrand's One-sided Bounded Differences Inequality) Let $X = (X_1, \dots, X_n) \sim P_1 \otimes \dots \otimes P_n$, X_i independent. Let $f : A^n \rightarrow \mathbb{R}$ be a function with one-sided bounded differences with associated functions c_i . Let $Z = f(X)$ and let $\nu = \mathbb{E}[\sum_{i=1}^n c_i(X)^2]$. Then

$$\psi_{Z - \mathbb{E}[Z]}(\lambda) \leq \frac{\lambda^2 \nu}{2} \quad \forall \lambda > 0$$

which implies that

$$\mathbb{P}(Z - \mathbb{E}[Z] \geq t) \leq e^{-t^2/2\nu} \quad \forall t > 0.$$

Proof (Hints).

- For $Q \ll P$ and $\pi \in \Pi(P, Q)$, show that, using **Law of Total Expectation**,

$$\mathbb{E}_Q[Z] - \mathbb{E}_P[Z] \leq \sum_{i=1}^n \mathbb{E}_\pi[c_i(X) \mathbb{P}(X_i \neq Y_i \mid X)],$$

where $\mathbb{P}(X_i \neq Y_i \mid X) = \mathbb{E}_\pi[\mathbb{I}_{\{X_i \neq Y_i\}} \mid X]$.

- Apply Cauchy-Schwarz twice.

- Conclude using [Marton's Conditional Transport Cost Inequality](#) and [Marton's Argument](#).

□

Proof. Let $Q \ll P$. Then for all $\pi \in \Pi(P, Q)$,

$$\begin{aligned}
\mathbb{E}_Q[Z] - \mathbb{E}_P[Z] &= \mathbb{E}_\pi[f(Y) - f(X)] \\
&\leq \mathbb{E}_\pi \left[\sum_{i=1}^n c_i(X) \mathbb{I}_{\{X_i \neq Y_i\}} \right] \quad \text{by assumption} \\
&= \sum_{i=1}^n \mathbb{E}_\pi \mathbb{E}_\pi [\mathbb{I}_{\{X_i \neq Y_i\}} c_i(X) \mid X] \quad \text{by [Law of Total Expectation](#)} \\
&= \sum_{i=1}^n \mathbb{E}_\pi [c_i(X) \mathbb{P}(X_i \neq Y_i \mid X)] \\
&\leq \sum_{i=1}^n (\mathbb{E}_\pi [c_i(X)^2])^{1/2} (\mathbb{E}_\pi [\mathbb{P}(X_i \neq Y_i \mid X)^2])^{1/2} \quad \text{by Cauchy-Schwarz} \\
&\leq \left(\sum_{i=1}^n \mathbb{E}_\pi [c_i(X)^2] \right)^{1/2} \left(\sum_{i=1}^n \mathbb{E} [\mathbb{P}(X_i \neq Y_i \mid X)^2] \right)^{1/2} \quad \text{by Cauchy-Schwarz}
\end{aligned}$$

where we write $\mathbb{P}(X_i \neq Y_i \mid X) = \mathbb{E}_\pi [\mathbb{I}_{\{X_i \neq Y_i\}} \mid X]$. By [Marton's Conditional Transport Cost Inequality](#),

$$\inf_{\pi \in \Pi(P, Q)} \sum_{i=1}^n \mathbb{E} [\mathbb{P}(X_i \neq Y_i \mid X)^2] \leq 2D(Q \parallel P).$$

which implies that

$$\mathbb{E}_Q[Z] - \mathbb{E}_P[Z] \leq \sqrt{\nu \cdot 2 \cdot D(Q \parallel P)}$$

and so by [Marton's Argument](#), $\psi_{Z - \mathbb{E}[Z]}(\lambda) \leq \frac{\lambda^2 \nu}{2}$ for all $\lambda > 0$, which gives the right tail bound by the [Chernoff Bound](#). □

6. Log-concave random variables

Definition 6.1 A continuous random variable $X \in \mathbb{R}^n$ with density function ρ is **log-concave** if $\log \rho$ is concave, i.e. if

$$\rho(\lambda x + (1 - \lambda)y) \geq \rho(x)^\lambda \rho(y)^{1-\lambda}$$

for all $x, y \in \mathbb{R}^n$ and $\lambda \in [0, 1]$.

Definition 6.2 A **convex body** is a non-empty, convex, compact set. The **diameter** of a convex body K is $\text{Diam}(K) = \sup_{x, y \in K} \|x - y\|_2$.

Example 6.3 The Gaussian

$$\frac{1}{(2\pi)^n \det(\Sigma)^{1/2}} e^{-(x \Sigma^{-1} x)/2},$$

the exponential $\alpha e^{-\|x\|}$ and the uniform distribution on convex bodies are log-concave distributions.

Theorem 6.4 (Log-concave Poincaré inequality) Let X be log-concave, supported on a convex body $K \subseteq \mathbb{R}^n$. Then X satisfies the Poincaré inequality with Poincaré constant

$$C_P(X) \leq \text{Diam}(K)^2 \cdot C_n,$$

for some absolute C_n depending only on n ; that is,

$$\text{Var}(f(X)) \leq \text{Diam}(K)^2 \cdot C_n \cdot \mathbb{E}[\|\nabla f(X)\|^2],$$

for all $f \in C^1(\mathbb{R}^n)$.

Proof. WLOG $\mathbb{E}[f(X)] = 0$. We have

$$\text{Var}(f(X)) = \frac{1}{2} \text{Var}(f(X) - f(Y)) = \frac{1}{2} \mathbb{E}[(f(X) - f(Y))^2],$$

where Y is an independent copy of X . Hence,

$$\begin{aligned} \text{Var}(f(X)) &= \frac{1}{2} \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} |f(y) - f(x)|^2 \rho(x) \rho(y) \, dx \, dy \\ &= \frac{1}{2} \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} \left| \int_{[0,1]} \nabla f(ty + (1-t)x) \cdot (y-x) \, dt \right|^2 \rho(x) \rho(y) \, dx \, dy \\ &\leq \frac{\text{Diam}(K)^2}{2} \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} \int_{[0,1]} \|\nabla f(ty + (1-t)x)\|^2 \, dt \rho(x) \rho(y) \, dx \, dy \quad \text{by Cauchy-Schwarz} \\ &= \frac{\text{Diam}(K)^2}{2} \int_{[0,1]} \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} \|\nabla f(ty + (1-t)x)\|^2 \rho(x) \rho(y) \, dx \, dy \, dt \end{aligned}$$

First consider the case when $t \approx \frac{1}{2}$. We use the bound $\min\{\rho(x), \rho(y)\} \leq \rho(ty + (1-t)x)$ (due to concavity), which implies

$$\begin{aligned} \rho(x) \rho(y) &\leq \rho(ty + (1-t)x) \max\{\rho(x), \rho(y)\} \\ &\leq \rho(ty + (1-t)x) (\rho(x) + \rho(y)). \end{aligned}$$

So

$$\begin{aligned} &\int_{\mathbb{R}^n} \int_{\mathbb{R}^n} \|\nabla f(ty + (1-t)x)\|^2 \rho(x) \rho(y) \, dx \, dy \\ &\leq \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} \|\nabla f(ty + (1-t)x)\|^2 \rho(ty + (1-t)x) (\rho(x) + \rho(y)) \, dx \, dy \\ &\leq \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} \|\nabla f(u)\|^2 \rho(u) \rho(x) \frac{du \, dx}{t^n} + \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} \|\nabla f(u)\|^2 \rho(u) \rho(y) \frac{du}{(1-t)^n} \, dy \\ &= \left(\frac{1}{t^n} + \frac{1}{(1-t)^n} \right) \mathbb{E}[\|\nabla f(X)\|^2]. \end{aligned}$$

using the substitutions $ty + (1-t)x = u$ (so $t^n dy = du$), $ty + (1-t)x = v$ (so $(1-t)^n dx = dv$).

In the case $t \gg 1/2$ or $t \ll 1/2$, then

$$\rho(x)\rho(y) \leq \rho(ty + (1-t)x) \cdot \rho((1-t)y + tx)$$

hence

$$\begin{aligned} & \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} \|\nabla f(ty + (1-t)x)\|^2 \rho(x)\rho(y) dx dy \\ & \leq \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} \|\nabla f(ty + (1-t)x)\|^2 \rho(ty + (1-t)x) \rho((1-t)y + tx) dy dx \\ & = \int_{\mathbb{R}^n} \int_{\mathbb{R}^n} \|\nabla f(u)\|^2 \rho(u)\rho(v) \frac{du dv}{|t^2 - (1-t)^2|^n} \\ & = \frac{1}{|t^2 - (1-t)^2|^n} \mathbb{E}[\|\nabla f(X)\|^2] \end{aligned}$$

since the map $(x, y) \mapsto (tx + (1-t)y, (1-t)x + ty)$ is represented by the matrix $\begin{bmatrix} t & 1-t \\ 1-t & t \end{bmatrix}$ which has determinant $|t^2 - (1-t)^2|$. So $dx dy = \frac{du dv}{|t^2 - (1-t)^2|^n}$.

Combining these, we obtain

$$\begin{aligned} \text{Var}(f(X)) & \leq \frac{\text{Diam}(K)^2}{2} \mathbb{E}[\|\nabla f(X)\|^2] \int_{[0,1]} \min\left\{\frac{1}{t^n} + \frac{1}{(1-t)^n}, \frac{1}{|t^2 - (1-t)^2|^n}\right\} dt \\ & \leq \frac{\text{Diam}(K)^2}{2} C_n \mathbb{E}[\|\nabla f(X)\|^2]. \end{aligned}$$

□

Remark 6.5 Let $X \sim \text{Unif}(A)$, $A \subseteq \mathbb{R}^n$. The Poincaré constant $C_p(X)$ measures the **conductance** of A , which is large if A has a bottleneck.

6.1. One-dimensional log-concave random variables

Definition 6.6 Let X be an RV on \mathbb{R} with density function f . The **differential entropy** of X is

$$h(X) = - \int_{-\infty}^{\infty} f(x) \log f(x) dx = \mathbb{E}[-\log f(X)].$$

Definition 6.7 Let X, Y be an RVs on \mathbb{R} with density functions f, g . The **differential relative entropy** of X and Y is

$$D(f \parallel g) = D(X \parallel Y) = \int_{-\infty}^{\infty} f(x) \log \frac{f(x)}{g(x)} dx = \mathbb{E}\left[\log \frac{f(X)}{g(X)}\right] \geq 0.$$

Lemma 6.8 Let Y be an RV with density f on \mathbb{R} with variance $\text{Var}(Y) = \sigma^2$. Let $Z \sim N(\mathbb{E}[Y], \sigma^2)$. Then

$$h(Y) \leq h(Z) = \frac{1}{2} \log(2\pi e \sigma^2).$$

In other words, normally distributed random variables maximise differential entropy.

Proof (Hints).

- Explain why we can assume $\mathbb{E}[Y] = 0$ WLOG.
- Use non-negativity of differential relative entropy.

□

Proof. WLOG, $\mathbb{E}[Y] = 0$ (since entropy is invariant under constant shifts). Let $\varphi_{\sigma^2}(x) := \frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/2\sigma^2}$. We have

$$\begin{aligned} 0 \leq D(f \parallel \varphi_{\sigma^2}) &= \int_{-\infty}^{\infty} f(x) \log f(x) dx + \frac{1}{2} \log(2\pi\sigma^2) + \int_{-\infty}^{\infty} \frac{x^2}{2\sigma^2} f(x) dx \\ &= -h(Y) + \frac{1}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \mathbb{E}[Y^2] \\ &= -h(Y) + \frac{1}{2} \log(2\pi\sigma^2) + \frac{1}{2} = \frac{1}{2} \log(2\pi e \sigma^2). \end{aligned}$$

It is straightforward to show that $h(Z) = \frac{1}{2} \log(2\pi e \sigma^2)$.

□

Definition 6.9 A random variable X is **isotropic** if $\mathbb{E}[X] = 0$ and $\text{Var}(X) = 1$.

Lemma 6.10 Let X be log-concave and isotropic, with density function ρ on \mathbb{R} . Then

$$\rho(0) \geq \frac{1}{\sqrt{2\pi e}}.$$

Proof (Hints). Write $\rho(0) = e^{(\log(\rho(\int_{-\infty}^{\infty} \rho(x)x dx)))}$ and use log-concavity.

□

Proof. We have

$$\begin{aligned} \rho(0) &= \rho\left(\int_{-\infty}^{\infty} \rho(x)x dx\right) = e^{\log \rho(\int_{-\infty}^{\infty} \rho(x)x dx)} \geq e^{\int_{-\infty}^{\infty} \rho(x) \log \rho(x) dx} \\ &= e^{-h(\rho)} \geq \frac{1}{\sqrt{2\pi e}}, \end{aligned}$$

where the first inequality is by log-concavity (we use that $\int_{-\infty}^{\infty} \rho(x) dx = 1$), and the second is by Lemma 6.8.

□

Remark 6.11 It can be shown that for log-concave ρ , $\max_x \rho(x) \leq c$ for some absolute constant c . So the above lemma says that $\rho(0)$ and $\max_x \rho(x)$ are comparable.

Proposition 6.12 Let X be log-concave, isotropic, with density function ρ on \mathbb{R} . Then for all $x \geq 3/\rho(0)$,

$$\rho(x) \leq \rho(0) e^{-\frac{\rho(0)}{3} \log(2)x} \leq e^{-x \log(2)/(3\sqrt{2\pi e})}$$

Proof (Hints).

- Let x_m denote the mode of X (why is this unique?). Can assume WLOG that $x_m > 0$. WLOG, $x_m > 0$. Let $x_0 = \frac{2}{\rho(0)} + x_m$. Why is $x_0 \geq x_m$?
- By writing 1 as an integral, show that $x_m \leq 1/\rho(0)$ (justify using log-concavity).
- Use the same idea to show that $\rho(x_0) \leq \rho(0)/2$.
- For $x \geq 3/\rho(0)$, write $x_0 = \frac{x_0}{x} \cdot x + (1 - \frac{x_0}{x}) \cdot 0$ (why is this a valid convex combination?). Use log-concavity and combine the above inequalities to obtain the result.

□

Proof. Write x_m for the mode of X (this is unique since X is log-concave). WLOG, $x_m > 0$ (the proof is similar if $x_m < 0$). Define $x_0 := \frac{2}{\rho(0)} + x_m$. We have $x_0 \geq x_m$ by Lemma 6.10. First note that

$$1 = \int_{-\infty}^{\infty} \rho(x) dx \geq \int_0^{x_m} \rho(x) dx \geq x_m \rho(0)$$

by log-concavity. Hence, $x_m \leq 1/\rho(0)$. Also,

$$1 = \int_{-\infty}^{\infty} \rho(x) dx \geq \int_{x_m}^{x_0} \rho(x) dx \geq \rho(x_0)(x_0 - x_m) = \rho(x_0) \frac{2}{\rho(0)}$$

where the last inequality is because ρ has one mode (unimodal). Hence, $\rho(x_0) \leq \rho(0)/2$. So we have $x \geq \frac{3}{\rho(0)} \geq \frac{2}{\rho(0)} + x_m = x_0$, so we write $x_0 = \frac{x_0}{x} \cdot x + (1 - \frac{x_0}{x}) \cdot 0$. By log-concavity,

$$\rho(x_0) \geq \rho(x)^{x_0/x} \cdot \rho(0)^{1-x_0/x}.$$

Exponentiating both sides by x/x_0 , we get

$$\begin{aligned} \rho(x) &\leq \frac{\rho(x_0)^{x/x_0}}{\rho(0)^{x/x_0-1}} = \rho(0) \left(\frac{\rho(x_0)}{\rho(0)} \right)^{x/x_0} \leq \rho(0) \left(\frac{1}{2} \right)^{x/x_0} \leq \rho(0) 2^{-\rho(0)x/3} \\ &= \rho(0) e^{-\rho(0) \log(2)x/3}. \end{aligned}$$

The final inequality is by Lemma 6.10.

□

Remark 6.13 If ρ is log-concave and isotropic then so is $x \mapsto \rho(-x)$, so we can obtain a left tail bound as well.