Contents

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

$X = \sum_{i=1}^{n}$, $X_i \sim \mathrm{Bern}(1/2)$. $\mathbb{E}[X] = 5000$. 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 $X_1, ..., X_n$ be IID RVs with mean $\mathbb{E}[X_1] = \mu$. Let $\mathrm{Var}(X_1) = \sigma^2 < \infty$. Then $\frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n}(X_i - \mu) \underset{D}{\to} N(0,1)$, i.e.

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

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 $Q(\delta) = \int_\delta^\infty \frac{1}{\sqrt{2n}} e^{-x^2/2d}\,\mathrm{d}x$. 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) $\mathbb{P}(|X - \mu| \geq \varepsilon) \leq \frac{\mathrm{Var}(X)}{\varepsilon^2}$ for all $\varepsilon > 0$.

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

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

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

Question 3: Let $(X_1, ..., X_n), (Y_1, ..., Y_n)$ be IID. What is the longest common subsequence, i.e. $f(X_1, ..., X_n, Y_1, ..., Y_n) = \max\left\{k : \exists i_1, ..., i_k, j_1, ..., j_k \text{ s.t. } X_{i_j} = Y_{i_j}\ \forall j \in [k]\right\}$. 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 $\mathbb{E}_Y[\mathbb{E}_X[X \mid Y]] = \mathbb{E}_X[X]$.

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

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

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

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

Definition 0.8 The **conditional variance** of Y given X is the random variable

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

1. The Chernoff-Cramer method

1.1. The Chernoff bound and Cramer transform

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

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

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 $\varphi : \mathbb{R} \to \mathbb{R}_+$ 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 $\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i)$.

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

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

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 $\mathrm{Var}(X_1) < \infty$, 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 $Z_1, ..., Z_n$ independent, then

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

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. 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 $\mathbb{P}(Z \geq t) \leq \inf\big\{e^{-(\lambda t - \psi_Z(\lambda))} : \lambda > 0\big\}$, 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 $b = \sup_{\lambda > 0}\{\mathbb{E}[e^{\lambda Z}] < \infty\}$.
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. $\psi_Z(\alpha\lambda_1 + (1-\alpha)\lambda_2) = \log \mathbb{E}\big[e^{\alpha\lambda_1 Z} \cdot e^{(1-\alpha)\lambda_2 Z}\big] \leq \alpha \log \mathbb{E}\big[e^{\lambda_1 Z}\big] + (1 - \alpha) \log \mathbb{E}\big[e^{\lambda_2 Z}\big]$ 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

$$\mathbb{E}\big[e^{\lambda Z}\big] = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-x^2/2\sigma^2} e^{\lambda x} \,\mathrm{d}x$$

$$
\begin{aligned}
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x^2-2\lambda\sigma^2 x+\lambda^2\sigma^4)/2\sigma^2} e^{\lambda^2\frac{\sigma^2}{2}}\,\mathrm{d}x \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\lambda\sigma^2)^2/2\sigma^2} e^{\lambda^2\frac{\sigma^2}{2}}\,\mathrm{d}x \\
&= e^{\lambda^2\sigma^2/2}.
\end{aligned}
$$

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 $X_1, ..., X_n$ are independent with each $X_i \in \mathcal{G}(\nu_i)$ then $\sum_{i=1}^n X_i \in \mathcal{G}\left(\sum_{i=1}^n \nu_i\right)$.
3. If $X \in \mathcal{G}(\nu)$, then $\mathrm{Var}(X) \leq \nu$.

Proof. Exercise. □

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

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

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. $\mathbb{E}\left[e^{dX^2}\right] \leq 2$.

Proof (Hints).
- $(1 \Rightarrow 2)$: straightforward.
- $(2 \Rightarrow 3)$: Explain why we can assume $b = 1$. Use that $\mathbb{E}[Y] = \int_0^\infty \mathbb{P}(Y > t)\,\mathrm{d}t$ for $Y \geq 0$, and the Γ function.
- $(3 \Rightarrow 1)$: show that $\mathbb{E}[e^{\lambda X}] \leq \mathbb{E}\left[e^{\lambda(X-X')}\right]$ where X' is an IID copy of X. Show that $\mathbb{E}[(X-X')^{2q}] \leq \mathbb{E}[X^{2q}]$. Expand $\mathbb{E}\left[e^{\lambda(X-X')}\right]$ as a series. Conclude that $X \in \mathcal{G}(4c)$.
- $(3 \Leftrightarrow 4)$: exercise.

$\square$

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

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

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

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

$(3 \Rightarrow 1)$: $\mathbb{E}[e^{-\lambda X}] \cdot \mathbb{E}[e^{\lambda X}] = \mathbb{E}\left[e^{\lambda(X - X')}\right]$, 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

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

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

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

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. $\square$

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 $\psi_Y''(\lambda)\le(b-a)^2/4$ and $Y\in\mathcal{G}((b-a)^2/4)$.

Proof (Hints).

- Define a new distribution based on λ, which should be obvious after expanding $\psi_Y'(\lambda)$.
- To conclude the result, use a Taylor expansion at 0 of $\psi_Y(\lambda)$.

□

Proof. Let Y have distribution P. We have

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

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

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

Now

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

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

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

Finally, using a Taylor expansion at 0, we obtain

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

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

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

Theorem 1.20 (Hoeffding's Inequality) Let $X_1, ..., X_n$ 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. □

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. □

Remark 1.21 A drawback of Hoeffding's Inequality is that the bound does not involve $\mathrm{Var}(X_i)$ the variance 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 $X_1, ..., X_n$ be independent RVs with $\mathbb{E}[X_i] = 0$ and $|X_i| \leq c$ for all i. Let $\nu = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)$. 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 $\mathbb{E}[e^{\lambda X_i}] = 1 + \frac{\mathrm{Var}(X_i)}{c^2}(e^{\lambda c} - \lambda c - 1)$.
- Deduce that $\psi_{\sum_i X_i} \leq \nu_c^2(e^{\lambda c} - \lambda c - 1)$.
- Find an upper lower for $\psi^*_{\sum_i X_i}(t)$.

□

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

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

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 $t \geq 0 = \mathbb{E}\left[\sum_i X_i\right]$, 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 $\lambda = \frac{1}{c} \log\left(1 + t\frac{c}{v}\right) =: \lambda^*$. 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. $\square$

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 $Z \sim \mathrm{Pois}(\lambda)$, then $\psi_{Z-\nu}(\lambda) = \nu\left(e^\lambda - \lambda - 1\right)$.

2. The variance method

2.1. The Efron-Stein inequality

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

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

We want to study the concentration of $Z = f(X_1, ..., X_n)$ for independent X_i. If $Z = \sum_i X_i$, then $\mathrm{Var}\left(\sum_i X_i\right) = \sum_i \mathrm{Var}(X_i)$ if $\mathbb{E}\left[X_i X_j\right] = 0$ for all $i \neq j$, which holds if the X_i are independent.

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

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

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

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

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

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

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

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

Now

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

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

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

Hence, by the Law of Total Expectation,

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

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

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

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

Moreover, if $X_1', ..., X_n'$ are IID copies of $X_1, ..., X_n$, and $Z_i' = f\big(X_{1:(i-1)}, X_i', X_{(i+1):n}\big)$, 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,

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

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 $\operatorname{Var}^{(i)}(Z) = \frac{1}{2}\operatorname{Var}^{(i)}(Z - Z_i')$.
- For last part, use that $\operatorname{Var}(X) = \inf_a \mathbb{E}[(X - a)^2]$.

□

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

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

Thus we have

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

The equality with $\cdot_+$ and $\cdot_-$ follows since $Z - Z_i'$ is a symmetric RV. Finally, recall that $\operatorname{Var}(X) = \inf_a \mathbb{E}[(X - a)^2]$, with equality if $a = \mathbb{E}[X]$. So $\operatorname{Var}^{(i)}(Z) =$

$\inf_{Z_i} E^{(i)}\left((Z - Z_i)^2\right)$, with equality if $Z_i = E^{(i)} Z$. Taking expectations and summing completes the proof. □

2.2. Functions with bounded differences

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

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

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

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

Proof (Hints). Consider the random variable

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

□

Proof. Define

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

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 $\operatorname{Var}(Z) \leq \sum_{i=1}^n \mathbb{E}\left[(Z - Z_i)^2\right] \leq \frac{1}{4} \sum_{i=1}^n c_i^2$. □

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

Suppose $X_1, ..., X_n$ 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, $\operatorname{Var}(Z) \leq \frac{1}{4} n$.

Note that this bound is tight, e.g. when $X_i \sim \operatorname{Bern}(1/2)$, $Z \sim B(n, 1/2)$, which has variance $1/4$.

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

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

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, $\operatorname{Var}(Z) \leq n/2 = \Theta(n)$. 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

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

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, $\mathrm{Var}(Z) \le \binom{n}{2}/4 = \Theta(n^2)$. 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 $1 \le i \le n-1$, 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 $1 \le j \le i$. The Y_i are independent. Also, note that $\{Y_i\}_{i=1}^{n-1}$ determines the graph. Hence, $\chi(G) = g\left(Y_{1:(n-1)}\right)$ 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, $\mathrm{Var}(\chi(G)) \le (n-1)/4$, 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 $X_1, ..., X_n$ be real-valued random variables, and let $Z = f(X_1, ..., X_n)$. A Poincaré inequality is of the form $\mathrm{Var}(Z) \lesssim \mathbb{E}[\|\nabla f(X)\|^2]$. So we have a local property (smoothness) which gives a global property (bound on the variance).

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

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

$$\mathrm{Var}(Z) \le \mathbb{E}\left[\|\nabla f(X_{1:n})\|^2\right],$$

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

Proof (Hints).

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

□

Proof. Let $Z_i = \inf_{x_i'} f\left(X_{1:(i-1)}, x_i', X_{(i+1):n}\right)$. 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 $\overline{X}^{(i)} = \left(X_{1:(i-1)}, X_i', X_{(i+1):n}\right)$. Note that since f is separately convex and X_i' is a minimiser (so $f\left(X'_{(i)}\right) \le f(X)$),

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

By the Efron-Stein Inequality,

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

□

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:

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

σ_1 is convex (and so separately convex) since it is a supremum of linear functions. Since it is a norm, we have $\sigma_1(A+B) \le \sigma_1(A) + \sigma_1(B)$ and $\sigma_1(A-B) \ge |\sigma_1(A) - \sigma_1(B)|$. Fix A. Since f is convex, the supremum is achieved: let u, v achieve the supremum. Then

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

Now if X, X' are independent, $d(X, X') = \|X - X'\|_F \ge \sigma_1(X - X') \ge |\sigma_1(X) - \sigma_1(X')|$ 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-lipchitz function satisfies $\|\nabla f\| \le L$. So by the Convex Poincare Inequality, $\mathrm{Var}(\sigma_1(X)) \le 4$ (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 Poincare Inequality) Let $X_{1:n}$ be IID and standard Gaussian (i.e. each $X_i \sim N(0,1)$). Then for any continuously differentiable $f \in C^1(\mathbb{R}^n)$,

$$\mathrm{Var}(f(X_{1:n})) \le \mathbb{E}\left[\|\nabla f(X_{1:n})\|^2\right].$$

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 $\mathrm{Var}^{(i)}(Z) = \frac{1}{4}\left(f\left(S_n - \frac{\varepsilon_i}{\sqrt{n}} + \frac{1}{\sqrt{n}}\right) - f\left(S_n - \frac{\varepsilon_i}{\sqrt{n}} - \frac{1}{\sqrt{n}}\right)\right)^2$.
- 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 the central limit theorem to conclude the result.

$\square$

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

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

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,

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

Note $\mathrm{Var}^{(i)}$ 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

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

and so

$$\begin{aligned}
\mathrm{Var}^{(i)}(f(S_n)) &= \frac{1}{2}\Big(f\big(S_n - \varepsilon_i/\sqrt{n} + 1/\sqrt{n}\big) - \Big(\frac{1}{2} f\big(S_n - \varepsilon_i/\sqrt{n} + 1/\sqrt{n}\big) + \frac{1}{2} f\big(S_n - \varepsilon_i/\sqrt{n} - 1/\sqrt{n}\big)\Big)\Big)^2 \\
&+ \frac{1}{2}\Big(f\big(S_n - \varepsilon_i/\sqrt{n} - 1/\sqrt{n}\big) - \Big(\frac{1}{2} f\big(S_n - \varepsilon_i/\sqrt{n} + 1/\sqrt{n}\big) + \frac{1}{2} f\big(S_n - \varepsilon_i/\sqrt{n} - 1/\sqrt{n}\big)\Big)\Big)^2 \\
&= \frac{1}{4}\big(f\big(S_n - \varepsilon_i/\sqrt{n} + 1/\sqrt{n}\big) - f\big(S_n - \varepsilon_i/\sqrt{n} - 1/\sqrt{n}\big)\big)^2
\end{aligned}$$

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

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

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

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

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

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 $\mathrm{Var}\left(\sum_i X_i\right) = \sum_i \mathrm{Var}(X_i)$ uses the tensorisation property of variance. Also, the Efron-Stein Inequality

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

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 Poincare 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 Poincare Inequality, $\mathrm{Var}(\sigma_1(X)) \leq 1$, which is a good bound, given that it is known that $\mathbb{E}[\sigma_1(X)] = O\left(\sqrt{n} + \sqrt{d}\right)$.

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

3.1. Poincare constant

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

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

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

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

Proposition 3.10 The Poincare 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 $\theta \in \mathbb{R}^d$, $\mathrm{Var}(\langle X, \theta \rangle) \leq C_P(X)$. In particular, $\mathrm{Var}(X_i) \leq C_P(X)$ for all i.
3. If $X_1, ..., X_n$ 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 $i_1, ..., i_n, j \in S$, 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 A transition matrix $P \in \mathbb{R}^{d \times d}$ is said to be **reversible** if $P_{ij} = P_{ji}$ for all $1 \leq i, j \leq d$.

Definition 3.15 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 3.16 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 3.17 Let $P \in \mathbb{R}^{d \times d}$ be a reversible transition matrix with stationary distribution $\pi \in \mathbb{R}^d$. Let $f \in \mathbb{R}^d$. Then

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

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

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

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

□

Remark 3.18 If the transition matrix P is reversible, then $\Lambda = P - I$ is self-adjoint (with respect to $\langle\cdot,\cdot\rangle_\pi$), so has real eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$. By Proposition 3.17, 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 $f_1 = (1, ..., 1)$.

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

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

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 $\operatorname{Var}_\pi(f) \leq c \cdot \mathcal{E}(f,f)$ 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 $X \sim \mathrm{Unif}(A)$.

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

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

Theorem 4.9 (Chain Rule for Entropy) We have

$$H(X_{1:n}) = \mathbb{E}[-\log P(X_{1:n})] = \sum_{i=1}^n H\left(X_1 \mid X_{1:(i-1)}\right).$$

Proof (Hints). Straightforward. ☐

Proof. Since

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

we have

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

$$
\begin{aligned}
&= \sum_{i=1}^{n} \mathbb{E}\Big[-\log P\Big(X_i \mid X_{1:(i-1)}\Big)\Big] \\
&= \sum_{i=1}^{n} H\Big(X_1 \mid X_{1:(i-1)}\Big).
\end{aligned}
$$

□

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

Proof (Hints). Straightforward. □

Proof. We have

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

□

Proposition 4.11 (Chain Rule for Relative Entropy) Let P, Q be PMFs on A^n. Let $X_{1:n} \sim P$. Then

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

Proof (Hints). Straightforward. □

Proof. We have

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

□

Remark 4.12 The Chain Rule for Relative Entropy is similar to the chain rule for variance:

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

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

Lemma 4.13 (Conditioning Reduces Conditional Entropy) $H(X \mid Y, Z) \leq H(Y)$.

Proof (Hints). Straightforward. $\square$

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. $\square$

Theorem 4.14 (Han's Inequality) Let $X_{1:n}$ be discrete RVs. Then

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

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

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

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

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

Corollary 4.15 (Loomis-Whitney Inequality) The Loomis-Whitney inequality states that for finite $A \subseteq \mathbb{Z}^n$,

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

Proof (Hints). Straightforward. $\square$

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

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

$\square$

Lemma 4.16 Let Q, P be PMFs on a discrete set $A \times B \times C$. Then

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

Proof (Hints). Use convexity of relative entropy. $\square$

Proof. By convexity of relative entropy,

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

$$
\begin{aligned}
&= \sum_x Q(x) D\big(Q_{Y \mid X=x} \parallel P_Y\big) \\
&= D\big(Q_{Y \mid X} \parallel P_Y \mid Q_X\big).
\end{aligned}
$$

□

Theorem 4.17 (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

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

Equivalently,

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

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

Remark 4.18 Taking P to be uniform in Han's Inequality for Relative Entropy gives Han's Inequality for Shannon entropy.

Proof (Hints). Explain why $D(Q \parallel P) = D(Q_{X^{(i)}} \parallel P_{X^{(i)}}) + D\big(Q_{X_i \mid X^{(i)}} \parallel P_{X_i} \mid Q_{X^{(i)}}\big)$, then use Lemma 4.16. □

Proof. By the Chain Rule for Relative Entropy and Lemma 4.16,

$$
\begin{aligned}
D(Q \parallel P) &= D(Q_{X^{(i)}} \parallel P_{X^{(i)}}) + D\big(Q_{X_i \mid X^{(i)}} \parallel P_{X_i \mid X^{(i)}} \mid Q_{X^{(i)}}\big) \\
&= D(Q_{X^{(i)}} \parallel P_{X^{(i)}}) + D\big(Q_{X_i \mid X^{(i)}} \parallel P_{X_i} \mid Q_{X^{(i)}}\big) \\
&\geq D(Q_{X^{(i)}} \parallel P_{X^{(i)}}) + D\big(Q_{X_i \mid X_{1:(i-1)}} \parallel P_{X_i} \mid Q_{X_{1:(i-1)}}\big)
\end{aligned}
$$

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

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

□

Definition 4.19 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

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

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

Proposition 4.20 Let $X \sim P$, where $Q \ll P$ are PMFs on a countable alphabet A. Let $Z = \frac{Q(X)}{P(X)}$. Then

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

Proof (Hints). Straightforward. □

Proof. We have

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

□

Remark 4.21 In general, when Z is the Radon-Nikodym derivative $\frac{\mathrm{d}Q}{\mathrm{d}P}(X)$ and $X \sim P$, then $\mathrm{Ent}(Z) = D(Q \parallel P)$.

Theorem 4.22 (Tensorisation of Entropy) Let $X_1, ..., X_n$ be independent RVs taking values in a countable set A, and let $f : A^n \to \mathbb{R}_{\geq 0}$. Let $Z = f(X_{1:n}) = f(X)$. Then

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

where

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

Remark 4.23 Tensorisation of Entropy is analogous to the Efron-Stein Inequality.

Proof (Hints).

- Show that $\mathrm{Ent}(aZ) = a \,\mathrm{Ent}(Z)$, and so can assume WLOG that $\mathbb{E}[Z] = 1$, so Q is PMF.
- Show that

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

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

□

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

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

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

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

Now

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

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

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

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

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

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

4.2. Herbst's argument

Theorem 4.24 (Herbst's Argument) Suppose Z is a real-valued RV and $\mathbb{E}[e^{\lambda Z}] < \infty$ for all $\lambda > 0$. If there exists $\nu > 0$ such that for all $\lambda > 0$, $\mathrm{Ent}\left(e^{\lambda Z}\right) \leq \lambda^2 \frac{\nu}{2} \mathbb{E}[e^{\lambda Z}]$, then

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

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

□

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

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

But

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

So by the above expression for Ent,

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

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

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

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

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

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

Theorem 4.25 (McDiarmid's Inequality) Let $X = (X_1, ..., X_n)$, where the X_i are independent. Let f have bounded differences with constants c_i. Let $Z = f(X)$. Then for all $t > 0$,

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

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 $\frac{\operatorname{Ent}^{(i)}(e^{\lambda Z})}{\mathbb{E}[e^{\lambda Z} \mid X^{(i)}]} \le \frac{1}{8}\lambda^2 c_i^2$ (you should use an integral somewhere).
- Use Tensorisation of Entropy and Herbst's Argument to show that $Z - \mathbb{E}[Z]$ is sub-Gaussian 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

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

Write $f_{X^{(i)}}(x_i) = f\left(X_{1:(i-1)}, x_i, X_{(i+1):n}\right)$. 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 $\psi_i(\lambda) = \log \mathbb{E}\left[e^{\lambda Z} \mid X^{(i)}\right] - \lambda \mathbb{E}\left[Z \mid X^{(i)}\right]$. As in the proof of Herbst's Argument, we have

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

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

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

The third step is using Herbst's Argument: we have

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

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. □

4.3. Log-Sobolev inequalities on the hypercube

Notation 4.26 Let $X_1, ..., X_n$ 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, $\mathrm{Var}(Z) \leq \frac{1}{2} \mathbb{E}\left[\sum_{i=1}^{n} (Z - Z'_i)^2\right] =: \nu$, where $Z'_i = f\left(X_{1:(i-1)}, X'_i, X_{(i+1):n}\right)$ and X'_i is an independent copy of X_i. Define $\mathcal{E}(f)$ as

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

$$= \frac{1}{2}\mathbb{E}\left[\sum_{i=1}^{n}\left(f(X) - f\left(\overline{X}^{(i)}\right)\right)_+^2\right] =: \mathcal{E}(f),$$

where $\overline{X}^{(i)} = \left(X_{1:(i-1)}, -X_i, X_{(i+1):n}\right)$. $\frac{1}{2}\left(f(X) - f\left(\overline{X}^{(i)}\right)\right)$ 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 4.27 (Log-Sobolev Inequality for Bernoullis) Let X be uniformly distributed on $\{-1,1\}^n$ and $f : \{-1,1\}^n \to \mathbb{R}$. Then

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

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,

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

If the result was true for $n = 1$, then we would have $\mathrm{Ent}^{(i)}(Z^2) \leq \frac{1}{2}\left(f(X) - f\left(\overline{X}^{(i)}\right)\right)^2$ (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 $f\left(\overline{X}^{(i)}\right)^2$ with equal probability) and so $\mathrm{Ent}(Z^2) \leq 2\mathcal{E}(f)$. 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 convex). 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 4.28 Log-Sobolev Inequality for Bernoullis is stronger than Efron-Stein Inequality. Also, the constant 2 on the RHS is tight.

Theorem 4.29 (Gaussian Log-Sobolev Inequality) Let $X = (X_1, ..., X_n)$ be a vector of n independent RVs with each $X_i \sim N(0, 1)$, let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable. Then

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

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

Definition 4.30 $f : \mathbb{R}^n \to \mathbb{R}$ is ***L*-Lipschitz** if

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

Theorem 4.31 (Gaussian Concentration Inequality) Let $X = (X_1, ..., X_n)$ be a vector of n independent RVs with each $X_i \sim N(0, 1)$. Let $f : \mathbb{R}^n \to \mathbb{R}$ 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 $t > 0$,

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

Note that these bounds are independent of the dimension n.

Proof (Hints).

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

□

Proof. WLOG, we can assume f is continuously differentiable (otherwise, we can approximate f with a sequence of contiuously 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

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

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. □

Theorem 4.32 (Concentration on the Hypercube) Let $f : \{-1,1\}^n \to \mathbb{R}$ and let $X = (X_1, ..., X_n)$ be uniform on $\{-1,1\}^n$. Let $Z = f(X)$ and assume

$$\max_{x\in\{-1,1\}^n} \sum_{i=1}^{n} \left(f(x) - f\left(\overline{x}^{(i)}\right)\right)_+^2 > 0 \leq \nu$$

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

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

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

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

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 4.33

- If the same condition for the negative part $(\cdot)_-$ holds, then we get the analogous left tail bound.
- If $\max_{x\in\{-1,1\}^n} \sum_{i=1}^{n} \left(f(x) - f\left(\overline{x}^{(i)}\right)\right)^2 \leq \nu$, 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

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

so $Z-\mathbb{E}[Z]\in\mathcal{G}(\nu/4)$. McDiarmid's 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 McDiarmid's Inequality but under a weaker assumption.

- The Efron-Stein Inequality gives

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

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

4.4. The modified log-Sobolev inequality (MLSI)

Lemma 4.34 (Variational Principle for Entropy) For any non-negative random variable Y,

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

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

Proof (Hints). Use the inequality $\log x\leq x-1$. □

Proof. We have

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

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

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

□

Remark 4.35 This is an entropy analogue of $\operatorname{Var}(Y)=\inf_{a\in\mathbb{R}}\mathbb{E}\left[(Y-a)^2\right]$. 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 4.36 (Modified Log-Sobolev Inequality) Let $X_1,\ldots,X_n$ be independent RVs taking values on A. Let $f:A^n\to\mathbb{R}$ and $Z=f(X)$. Let $f_i:A^{n-1}\to\mathbb{R}$ be an arbitrary function and $Z_i=f_i\left(X^{(i)}\right)$ for each $i\in[n]$. Then

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

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:

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

Proof (Hints). Use Tensorisation of Entropy and the Variational Principle for Entropy, with $u = Y_i$ (conditional on $X^{(i)}$). □

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

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

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,

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

The result follows by summing and taking expectations. □

Theorem 4.37 (Relaxed Bounded Differences) Let $Z = f(X_1, ..., X_n)$ for independent RVs $X_1, ..., X_n$ which take values on A and $f : A^n \to \mathbb{R}$. Let

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

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). Straightforward. □

Proof. By the Modified Log-Sobolev Inequality,

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

The result follows by Herbst's Argument and the Chernoff Bound. □

Remark 4.38 If $Z_i = \sup_{x_i'} f\left(X_{1:(i-1)}, x_i', X_{(i+1):n}\right)$ and $\sum_{i=1}^n \left(Z - Z_i\right)^2 \leq \nu$, then we also obtain a left tail bound. If this condition holds for the supremum and the infimum, then $Z \in \mathcal{G}(\nu)$.

4.5. Concentration of convex Lipschitz functions

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

Theorem 4.39 Let $f : [0,1]^n \to \mathbb{R}$ be separately convex and 1-Lipschitz. Let $Z = f(X_1, ..., X_n)$ where $X_1, ..., X_n$ 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 $\sum_{i=1}^n \left(f(X) - f\left(X_{(i)}'\right)\right)^2$, where $X_{(i)}' = \left(X_{1:(i-1)}, X_i', X_{(i+1):n}\right)$ 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 $\sum_{i=1}^n \left(Z - Z_i\right)^2 \leq 1$, where $Z_i = \inf_{x_i'} f\left(X_{1:(i-1)}, x_i', X_{(i+1):n}\right)$. We have

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

where $X_{(i)}' = \left(X_{1:(i-1)}, X_i', X_{(i+1):n}\right)$ 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 $f\left(X_{(i)}'\right) \leq f(X)$),

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

since f is 1-Lipschitz. $\square$

Remark 4.40 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 4.41 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 $(\Omega, \mathcal{A}, P)$, where P is a probability measure.

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

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

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, A, P) be a countable probability space and Z be a random variable with $\mathbb{E}[|Z|] < \infty$. Then

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

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

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

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

Proof (Hints). Define

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

and show that $0 \leq D(Q \parallel P) + \log \mathbb{E}_P[e^Z] - \mathbb{E}_Q[Z]$. When is equality achieved? $\square$

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

$$0 \leq D(Q \parallel Q^*)$$

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

Hence $\log \mathbb{E}\big[e^Z\big] \geq \mathbb{E}_Q Z - D(Q \parallel P)$, with equality iff $Q = Q^*$. The result follows since $Q^* \ll P$. For the second statement, note that $D(Q \parallel P) \geq \mathbb{E}_Q[Z] - \log \mathbb{E}\big[e^Z\big]$, for any $Q \ll P$ and Z. There is equality if $Z(\omega) = \log \frac{Q(\omega)}{P(\omega)}$. (Note that $\mathbb{E}_Q[|Z|] = \mathbb{E}_Q\left[\left|\log \frac{Q}{P}\right|\right] < \infty$ 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 $\mathbb{E}_P\big[e^Z\big] = \infty$. □

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

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

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

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

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

□

Proof. By the Variational Representation for log-MGF and Relative Entropy,

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

$$\leq \sup_{t\geq 0}\left(\lambda\sqrt{2\nu t}-t\right).$$

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 $\lambda>0$,

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

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

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

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

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

□

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 $f: A^n\to\mathbb{R}$ 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 $\nu=C\sum_{i=1}^n c_i$.

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

$$
\begin{aligned}
\mathbb{E}_Q[Z] - \mathbb{E}_P[Z] &= \mathbb{E}[f(Y)] - \mathbb{E}[f(X)] \\
&= \mathbb{E}_\pi[f(Y) - f(X)] \quad \text{for all } \pi \in \Pi(P, Q) \\
&\leq \mathbb{E}_\pi\left[\sum_{i=1}^n c_i d(X_i, Y_i)\right] \\
&= \sum_{i=1}^n c_i \mathbb{E}_\pi[d(X_i, Y_i)] \\
&\leq \left(\sum_{i=1}^n c_i^2\right)^{1/2} \left(\sum_{i=1}^n \mathbb{E}_\pi[d(X_i, Y_i)]^2\right)^{1/2} \quad \text{by Cauchy-Schwarz}
\end{aligned}
$$

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

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