Question 1: You toss a coin $10,000$ times. How many heads do you see?

Question 2: Coupon collector problem. Have N coupons and we need to collect them all. How many coupons do we need to sample to get all N?

Question 3: Largest common subsequence problem: have sequences $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ of iid Bern(1/2) random variables. What is the largest k such that there exist $i_1 < i_2 < \ldots < i_k$ and $j_1 < j_2 < \ldots, < j_k$ such that $X_{i_1} = Y_{j_1}, \ldots, X_{i_k} = Y_{j_k}$?

Question 1: we have various possible answers:

- 5,000. Indeed if we let X_i be the indicator of the event that we see heads on the ith toss, the number of heads is $S = \sum_{i=1}^{10000} X_i$ and $\mathbb{E}S = 5000$. But $\mathbb{P}(S = 5000) = \binom{10000}{5000} 2^{-1000} \approx 0.008$.
- Weak Law of Large Numbers: let $(X_i)_{i \geq 1}$ be iid with finite expectation μ and finite second moments. Then for every $\varepsilon > 0$,
$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right| > \varepsilon \right) \xrightarrow{n \to \infty} 0.$$
Therefore for large enough n, the number of heads lies in $[n(1/2 - \varepsilon), n(1/2 + \varepsilon)]$ with high probability. The main problem is that this is an asymptotic result - we don't know how large n should be.
- Central Limit Theorem: let $(X_i)_{i \geq 1}$ be iid with finite mean μ and finite second moment $\sigma^2 + \mu^2$. Then
$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu) \xrightarrow{d} \mathcal{N}(0, 1).$$
Therefore $\sum_{i=1}^{n} (X_i - \mu)$ has deviations of the order $\sqrt{n}\sigma$. Suppose we pretend 10000 is big: then
$$\begin{aligned} S = \sum_{i=1}^{10000} X_i &\in [5000 - Q^{-1}(0.005)\sqrt{100}/2, 5000 + Q^{-1}(0.005)\sqrt{100}/2] \\ &\approx [5000 \pm 128] \end{aligned}$$
with probability 0.99, where $Q(x) = \mathbb{P}(Z \geq x)$ for $Z \sim \mathcal{N}(0, 1)$. However we have the same issue again - is $n = 10000$ large enough?

We can however give some non-asymptotic answers to Question 1:

Proposition (Chebyshev's inequality)**.** Let X be any random variable with mean μ and variance σ^2. Then
$$\mathbb{P}(|X - \mu| > t) \leq \frac{\sigma^2}{t^2}.$$

With this, we have

$$\mathbb{P}\left(\left|\sum_{i=1}^{10000} X_i - 5000\right| > t\right) \leq \frac{10000 \times \frac{1}{4}}{t^2} = \frac{2500}{t^2}.$$

So in particular, if $t = 500$ the RHS is 0.01. So we have $S \in [4500, 5500]$ with probability 0.99. However note that this is a weaker result than what the Central Limit Theorem gives.

Question 2: the number of samples S is equal to $\sum_{i=1}^{N} X_i$ where $X_i \sim \text{Geo}(i/N)$. Thus $\mathbb{E}S = \sum_{i=1}^{N} \frac{N}{i} = N \sum_{i=1}^{N} \frac{1}{i} \approx N \log N$.

Question 3: we have a function $f(X_1, \ldots, X_n, Y_1, \ldots, Y_n)$ which gives the longest common subsequence. It turns out this function is “smooth” in a certain sense, for which we can use “Talagrand’s Principle”.

Chernoff-Crámer method

Theorem (Markov's inequality)**.** *Let Y be a non-negative random variable with finite expectation. Then for any $t > 0$ we have*

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

Proof. Note $t\mathbb{1}(Y \geq t) \leq Y$ and integrate. □

Corollary. *Let Y be a random variable. Suppose $\phi : \mathbb{R} \to \mathbb{R}_+$ is increasing and such that $\mathbb{E}|\phi(Y)| < \infty$. Then*

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

Note that for a random variable Z, letting $Y = |Z - \mathbb{E}Z|$ and $\phi : t \mapsto t^2$ gives Chebyshev's inequality $\mathbb{P}(|Z - \mathbb{E}Z| \geq t) \leq \frac{\mathrm{Var}(Z)}{t^2}$.

Could also take $\phi : t \mapsto t^q$ for any $q > 0$ to conclude $\mathbb{P}(|Z - \mathbb{E}Z| \geq t) \leq \frac{\mathbb{E}|Z-\mathbb{E}|^q}{t^q}$.

Consider instead $\phi : t \mapsto e^{\lambda t}$ for $\lambda > 0$. Then we get

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

Define $F(\lambda) = \mathbb{E}e^{\lambda Z}$, the *moment generating function* of Z. Define $\psi_Z(\lambda) = \log \mathbb{E}[e^{\lambda Z}]$. If $X_1, \ldots, X_n$ are independent and $Z = \sum_{i=1}^n X_i$ then it is clear that $\psi_Z(\lambda) = \sum_{i=1}^n \psi_{X_i}(\lambda)$. So we have

$$\mathbb{P}(Z \geq t) \leq \inf_{\lambda \geq 0} e^{\psi_Z(\lambda) - \lambda t}.$$

Now define $\psi_Z^*(t) = \sup_{\lambda \geq 0}(\lambda t - \psi_Z(\lambda))$ and write $\mathbb{P}(\lambda \geq t) \leq e^{-\psi_Z^*(t)}$. This is known as the *Chernoff bound*, and ψ_Z^* is known as the *Chernoff-Crámer transform*.

Properties of ψ_Z and ψ_Z^*

1. ψ_Z is convex and infinitely differentiable on $(0, b)$ where $b = \sup\{\lambda : \psi_Z(\lambda) < \infty\}$. Indeed

$$\begin{aligned} &F(\theta x + (1-\theta)y) \\ &= \mathbb{E}[e^{\theta x Z} e^{(1-\theta)yZ}] \\ &\leq \mathbb{E}[e^{xZ}]^\theta \mathbb{E}[e^{yZ}]^{1-\theta}. && \text{(Hölder with } 1/p = \theta, 1/q = 1-\theta) \end{aligned}$$

2. $\psi_Z^* \geq 0$ and it is convex (follows from the definition).

3. Suppose $t \geq \mathbb{Z}$. Then $\psi_Z^*(t) = \sup_\lambda(\lambda t - \psi_Z(\lambda))$. Indeed we'll show $\lambda t - \psi_Z(\lambda) \leq 0$ whenever $\lambda < 0$. We have

$$\begin{aligned}
&\mathbb{E}[e^{\lambda Z}] \geq e^{\lambda \mathbb{E} Z} && \text{(Jensen)} \\
&\implies \psi_Z(\lambda) \geq \lambda \mathbb{E} Z \\
&\implies \lambda t - \psi_Z(\lambda) \leq \lambda t - \lambda \mathbb{E} Z = \lambda(t - \mathbb{E} Z) \leq 0.
\end{aligned}$$

Example. Let $Z \sim \mathcal{N}(0, v)$. We want to upper bound $\mathbb{P}(Z \geq t)$ fir $t > 0$. We have

$$\begin{aligned}\mathbb{E}[e^{\lambda Z}] &= \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi v}} e^{-\frac{t^2}{2v}} e^{\lambda t} \mathrm{d}t \\ &= \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi v}} e^{-\frac{(t-\lambda v)^2}{2v}} e^{\frac{v^2\lambda^2}{2v}} \mathrm{d}t \\ &= e^{\frac{v\lambda^2}{2}}.\end{aligned}$$

Hence $\psi_Z^*(t) = \sup_\lambda \left(\lambda t - \frac{\lambda^2 v}{2}\right)$ (for $t > 0 = \mathbb{E}Z$). Differentiating we see the optimal value is $\lambda = t/v$. Plugging this in gives $\psi_Z^*(t) = \frac{t^2}{2v}$. Thus

$$\mathbb{P}(Z \leq t) \leq e^{-\frac{t^2}{2v}}.$$

Sub-Gaussian random variables

Definition. A random variable Y with $\mathbb{E}Y = 0$ is *sub-Gaussian* with variance parameter v if

$$\psi_Y(\lambda) < \frac{\lambda^2 v}{2} \ \forall \lambda \in \mathbb{R}.$$

The set of sub-Gaussian random variables with variance parameter v is denoted $\mathcal{G}(v)$.

1. It is clear from the above that if $Y \in \mathcal{G}(v)$ then $\mathbb{P}(Y \geq t) \leq e^{-t^2/2v}$ and $\mathbb{P}(Y \leq -t) \leq e^{-t^2/2v}$.
2. If $Y_i \in \mathcal{G}(v_i)$ for $i = 1, \ldots, n$ are independent then $\sum_{i=1}^n Y_i \in \mathcal{G}\left(\sum_{i=1}^n \nu_i\right)$ (immediate by additivity of $\psi_\cdot(t)$).
3. If $Y \in \mathcal{G}(v)$ then $\mathrm{Var}(Y) \leq v$ (see Example Sheet).

Theorem. *The following are equivalent for suitable v, b, c, d*

1. $Y \in \mathcal{G}(v)$;
2. $\max\{\mathbb{P}(Y \geq t), \mathbb{P}(Y \leq -t)\} \leq e^{-\frac{t^2}{2b}}$ *for all* $t > 0$;
3. $\mathbb{E}Y^{2q} \leq q! c^q$ *for all* $q \geq 1$;
4. $\mathbb{E}[e^{dY^2}] \leq 2$.

Proof. Not given. □

Lemma (Hoeffding's lemma)**.** *Let Y be supported on $[a, b]$ and suppose $\mathbb{E}Y = 0$. Then $\psi_Y''(\lambda) \leq \frac{(b-a)^2}{4}$, and so $Y \in \mathcal{G}\left(\frac{(b-a)^2}{4}\right)$.*

Proof. We have

$$\psi_Y'(\lambda) = \frac{\mathbb{E}[Ye^{\lambda Y}]}{\mathbb{E}[e^{\lambda Y}]} \implies \psi_Y''(\lambda) = \frac{\mathbb{E}[e^{\lambda Y}]\mathbb{E}[Y^2 e^{\lambda Y}] - (\mathbb{E}[Ye^{\lambda Y}])^2}{\mathbb{E}[e^{\lambda Y}]^2}.$$

So

$$\begin{aligned}\psi_Y''(\lambda) &= \int_{\mathbb{R}} y^2 \underbrace{\frac{e^{\lambda y}}{\mathbb{E}[e^{\lambda Y}]}\mathrm{d}\mu_Y(y)}_{:=\mathrm{d}Q(y)} - \left(\int_{\mathbb{R}} y \frac{e^{\lambda y}}{\mathbb{E}[e^{\lambda Y}]}\mathrm{d}\mu_Y(y)\right) \\ &= \mathrm{Var}_{Y\sim Q}(Y) \geq 0\end{aligned}$$

noting that Q is supported on $[a,b]$. If $Y \in [a,b]$ almost-surely then note

$$\mathrm{Var}(Y) = \mathrm{Var}\left(Y - \frac{a+b}{2}\right) \leq \mathbb{E}\left[\left(Y - \frac{a+b}{2}\right)^2\right] \leq \frac{(b-a)^2}{4}.$$

To finish, observe that $\psi_Y(\lambda) = \psi_Y(0) + \lambda\psi_Y'(0) + \frac{\lambda^2}{2}\psi_Y''(\theta)$ for some $\theta \in [0,\lambda]$. Thus $\psi_Y(\lambda) \leq \frac{\lambda^2(b-a)^2}{8}$. □

Theorem (Hoeffding's inequality)**.** *Let $Y_1, \ldots, Y_n$ be independent random variables with Y_i having support on $[a_i, b_i]$. Then*

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

Proof. Trivial by Hoeffding's lemma and additivity of the variance parameters. □

Theorem (Bennett's inequality)**.** *For $1 \leq i \leq n$, let X_i be independent random variables satisfying $\mathbb{E}X_i = 0$, $Var(X_i) = \sigma_i^2$ and let $v = \sum_{i=1}^n \sigma_i^2$. Further assume the X_i are bounded by some $C > 0$ almost-surely. Then*

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geq t\right) \leq \exp\left(-\frac{v}{C^2} h_1\left(\frac{Ct}{v}\right)\right)$$

where $h_1(x) = (1+x)\log(1+x) - x$ for $x > 0$. Furthermore, using the inequality $h_1(x) \geq \frac{x^2}{2(1+x/3)}$ we obtain

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

Example. Suppose $X_i \sim \mathrm{Bern}(p_n)$ are independent for $1 \leq i \leq n$. Then

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geq t\right) \leq \exp\left(-\frac{2t^2}{n}\right) \qquad \text{(Hoeffding)}$$

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geq t\right) \leq \exp\left(-\frac{t^2}{np_n(1-p_n) + t/3}\right). \qquad \text{(Bennett)}$$

Note that if $p_n \ll q$, e.g $p_n = 1/\sqrt{n}$, Hoeffding will stay the same, i.e of order $e^{-\frac{2t^2}{n}}$ (only depends on support, not variance). However, Bennet will be of the order $e^{-\frac{t^2}{\sqrt{n}+t/3}}$.

Proof. We have

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

This implies

$$\mathbb{E}^{\lambda S} \leq \exp\left(\frac{v}{C^2}(e^{\lambda C} - \lambda C - 1)\right)$$

and so

$$\psi_S(\lambda) \leq \underbrace{\frac{v}{C^2}(e^{\lambda C} - \lambda C - 1)}_{:=\tilde{\psi}(\lambda)}.$$

This means that

$$\psi_S^*(t) \geq \tilde{\psi}^*(t)$$

and

$$\mathbb{P}(S \geq t) \leq \exp(-\psi_S^*(t)) \leq \exp(-\tilde{\psi}^*(t)) = \exp\left(-\frac{v}{C^2}h_1\left(\frac{Ct}{v}\right)\right)$$

where the last equality is by a result from Example Sheet 1. ☐

Efron-Stein Inequality

We want to bound $\mathrm{Var}(Z)$ where $Z = f(X_1, \ldots, X_n)$ for independent X_i's (or even just uncorrelated). If $Z - \mathbb{E}Z = \sum_{i=1}^n \Delta_i$ for $\Delta_1, \ldots, \Delta_n$ uncorrelated and with 0 mean we have $\mathrm{Var}(Z) = \sum_{i=1}^n \mathbb{E}[\Delta_i^2]$. Define $\mathbb{E}_i Z = \mathbb{E}[Z|X_{1:i}]$[1] where $X_{1:i} = (X_1, \ldots, X_i)$.

Set $\Delta_i = \mathbb{E}_i Z - \mathbb{E}_{i-1} Z$. Then $Z - \mathbb{E}Z = \sum_{i=1}^n \Delta_i$. Also $\mathbb{E}\Delta_i = 0$ by the tower property of conditional expectation. Suppose $i < j$ so

$$\begin{aligned}
\mathbb{E}[\Delta_i \Delta_j] &= \mathbb{E}\left[\mathbb{E}[\Delta_i \Delta_j | X_{1:i}]\right] \\
&= \mathbb{E}[\Delta_i \mathbb{E}[\Delta_j | X_{1:i}]].
\end{aligned}$$

Note that $\mathbb{E}[\Delta_j|X_{1:i}] = \mathbb{E}[\mathbb{E}_j Z|X_{1:i}] - \mathbb{E}[\mathbb{E}_{j-1}Z|X_{1:i}] = \mathbb{E}_i Z - \mathbb{E}_i Z = 0$. Thus $\mathbb{E}[\Delta_i \Delta_j] = 0$ and so the Δ_i's are uncorrelated.

Thus $\mathrm{Var}(Z) = \sum_{i=1}^n \mathbb{E}[\Delta_i^2]$ regardless of the correlation between the X_i (though we still assume independence of the X_i going forward).

Define $\mathbb{E}^{(i)} Z = \mathbb{E}[Z|X_{1:i-1}, X_{i+1:n}]$. Then $\Delta_i = \mathbb{E}_i Z - \mathbb{E}_{i-1} Z = \mathbb{E}_i(Z - \mathbb{E}^{(i)} Z)$. Indeed we have $\mathbb{E}_i[\mathbb{E}^{(i)} Z] = \mathbb{E}[\mathbb{E}[Z|X^{(i)}]|X_{1:i}] = \mathbb{E}[\mathbb{E}[Z|X^{(i)}]|X_{1:i-1}]$ by independence and $\mathbb{E}[\mathbb{E}[Z|X^{(i)}]|X_{1:i-1}] = \mathbb{E}[Z|X_{1:i-1}]$ since $\sigma(X_{1:i-1}) \subseteq \sigma(X^{(i)})$.

Therefore
$$\Delta_i^2 = (\mathbb{E}_i(Z - \mathbb{E}^{(i)} Z))^2 \leq \mathbb{E}_i[(Z - \mathbb{E}^{(i)} Z)^2]$$
almost-surely by conditional Jensen.

Hence we have

$$\begin{aligned}
\mathrm{Var}(Z) &= \sum_{i=1}^n \mathbb{E}[\Delta_i^2] \\
&\leq \sum_{i=1}^n \mathbb{E}[(Z - \mathbb{E}^{(i)} Z)^2] \\
&= \sum_{i=1}^n \mathbb{E}[\mathbb{E}[(Z - \mathbb{E}^{(i)} Z)^2]|X^{(i)}] \\
&= \mathbb{E}\left[\sum_{i=1}^n \mathrm{Var}^{(i)}(Z)\right].
\end{aligned}$$

This is called the *Efron-Stein inequality.*

[1] For a rigorous definition of this conditional expectation see Part III Advanced Probability

To summarise:

Theorem (Efron-Stein Inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables and let $Z = f(X_1, \ldots, X_n)$ be a square integrable function of $X = X_{1:n}$. Then*

$$Var(Z) \leq \sum_{i=1}^{n} \mathbb{E}[(Z - \mathbb{E}^{(i)}Z)^2] = \underbrace{\sum_{i=1}^{n} Var^{(i)}(Z)}_{:=v}.$$

Proposition. Define $X_1', \ldots, X_n'$ to be independent copies of $X_1, \ldots, X_n$ respectively. Set $Z_i' = f(X^{(i)}, X_i')$. Then

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

Also

$$v = \inf_{Z_1, \ldots, Z_n} \sum_{i=1}^{n} \mathbb{E}[(Z - Z_i)^2]$$

where Z_i is some function of $X^{(i)}$.

Proof. Note that if X, Y are iid then

$$\mathrm{Var}(X) = \frac{1}{2}\mathbb{E}[(X - Y)^2] = \mathbb{E}[(X - Y)_+^2] = \mathbb{E}[(X - Y)_-^2]$$

since $(X - Y)_+, (X - Y)_-$ have the same distribution. For the final expression, note $\mathrm{Var}(X) = \inf_a \mathbb{E}[(X - a)^2]$. Then $\mathrm{Var}^{(i)}(Z) = \inf_{Z_i} \mathbb{E}[(Z - Z_i)^2 | X^{(i)}]$ where Z_i is $X^{(i)}$-measurable. $\square$

Functions with bounded-differences property

We say f satisfies the *bounded differences property* with constants $c_1, \ldots, c_n$ if

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

If $Z = f(X_1, \ldots, X_n)$ where the X_i are independent and f satisfying bounded differences, we'll show that $\mathrm{Var}(Z) \leq \sum_{i=1}^{n} \frac{c_i^2}{4}$. To see this, set

$$Z_i = \frac{1}{2}\left(\inf_{x_i} f(X^{(i)}, x_i) + \sup_{x_i} f(X^{(i)}, x_i)\right).$$

Then

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

Example. Let $X_1, \ldots, X_n$ be independent and supported on $[0,1]$. Define $f(X_{1:n})$ to be the smallest number of size 1 bins needed to "pack" $X_1, \ldots, X_n$. Note f satisfies the bounded differences property with $c_i = 1$ for all i. Therefore $\mathrm{Var}(Z) \leq \frac{n}{4}$. Suppose now the X_i are iid uniform on $[0,1]$. Then $\mathbb{E}f(X_1, \ldots, X_n) \approx Cn$ while the standard deviation is of order at most $\sqrt{n}$, giving tight confidence intervals for large n.

Example. Let $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ be iid Bernouilli with parameter $1/2$. Let $f(X_{1:n}, Y_{1:n})$ be the longest common subsequence between $X_{1:n}$ and $Y_{1:n}$. Then f satisfies boudned differences with $c_i = 1$ for all i. Thus $\mathrm{Var}(Z) \leq n/2$. It is known that $\mathbb{E}[Z] \sim [0.75n, 0.837n]$. So again Z is very concentrated about its mean for large n.

Example. The chromatic number $\chi(G)$ of a graph G is the smallest number of colours needed to colour vertices of G such that no two neighbouring vertices have the same colour. Let X_{ij} be iid Bernouilli of parameter p for $1 \leq i < j \leq n$. We construct a random graph G on vertex set $\{1, \ldots, n\}$ by saying $\{i, j\} \in E$ iff $X_{ij} = 1$. Take f such that $f(\{X_{ij}\}_{1 \leq i < j \leq n}) = \chi(G)$. Then f again satisfies bounded differences with $c_{ij} = 1$ for all $1 \leq i < j \leq n$. Hence $\mathrm{Var}(\chi(G)) \leq \frac{1}{4}\binom{n}{2}$. It is known that $\mathbb{E}[\chi(G)] \approx n/\log n$. This gives a poor confidence interval.

However, we can fix this bound by considering $Y_i = (X_{1,i+1}, \ldots, X_{i,i+1})$. Observe that $Y_1, \ldots, Y_{n-1}$ are independent and $\chi(G)$ is some function $\hat{f}$ of $Y_1, \ldots, Y_{n-1}$. It can be shown that we still have bounded differences with $c_1 = \ldots = c_{n-1} = 1$. This gives $\mathrm{Var}(\chi(G)) \leq \frac{n-1}{4}$ and thus we have a good confidence interval now.

Theorem (Convex Poincaré Inequality)**.** *Let $X_1, \ldots, X_n$ be independent and supported on $[0,1]$. Let f be a separately convex function (i.e convex in each variable) over $[0,1]^n$ which has partial derivatives. Then*

$$\mathit{Var}(f(X)) \leq \mathbb{E}[\|\nabla f(X)\|^2].$$

Remark. Jointly convex functions are separately convex so this inequality holds for such functions too.

Proof. We have

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

where Z_i is $X^{(i)}$-measurable. Let $Z_i = \inf_x f(X^{(i)}, x)$. Then

$$Z - Z_i = f(X_1, \ldots, X_n) - f(X_1, \ldots, X_{i-1}, x^*, x_{i+1}, \ldots, X_n) = f(X^{(i)}, X_i) - f(X^{(i)}, x^*) \geq 0$$

where x^* achieves the infimum of $f(X^{(i)}, x)$ over x. If g is convex then $g(y) \geq g(x) + g'(x)(y - x)$. Hence

$$f(X^{(i)}, X_i) - f(X^{(i)}, x^*) \leq \frac{\partial f}{\partial x_i}(X) \cdot (x^* - X_i).$$

Squaring gives

$$(Z - Z_i)^2 \leq \left[\frac{\partial f}{\partial x_i}(X)(x^* - X_i)\right]^2 \leq \left[\frac{\partial f}{\partial x_i}(X)\right]^2.$$

□

Example. Let $X \in \mathbb{R}^{n \times d}$ with $\mathbb{E}X_{ij} = 0$ for all i, j and with all entries independent and supported on $[-1, 1]$. Let

$$\sigma_1(X) = \max_{\|v\|_2 = 1} \|Xv\|_1 = \max_{\|U\|_2 = 1, \|v\|_2 = 1} U^T X v.$$

Can show the triangle inequality holds so

$$|\sigma_1(A) - \sigma_1(B)| \leq \sigma_1(A - B).$$

Can also show by Cauchy-Schwarz that

$$\sigma_1(A)^2 \leq \sum_{\substack{1 \leq i \leq n \\ 1 \leq j \leq d}} A_{ij}^2 = \|A\|_F^2.$$

Thus

$$|\sigma_1(A) - \sigma_1(B)| \leq \|A - B\|_F.$$

Therefore σ_1 is Frobenius-1-Lipschitz. This means (assuming derivatives exist) $\|\nabla \sigma_1(X)\| \leq 1$. So using the convex Poincaré inequality, $\mathrm{Var}(\sigma_1(X)) \leq 4$.

Theorem (Gaussian Poincaré inequality)**.** *Let $X_1, \dots, X_n$ be iid $\mathcal{N}(0,1)$. Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable. Then Var$(f(X)) \leq \mathbb{E}[\|\nabla f(X)\|^2]$.*

Proof. It is enough to show the $n = 1$ case. Indeed if the $n = 1$ case is true we have

$$\mathrm{Var}(f(X_1, \dots, X_n)) \leq \sum_{i=1}^{n} \mathbb{E}[\mathrm{Var}^{(i)}(Z)]$$

by Efron-Stein. Also

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

by the $n = 1$ case, so we get the general case.

Now we prove the $n = 1$ case. Let $X_1, \dots, X_n$ be iid (Rademacher) symmetric Ber$(1/2)$ (i.e takes values ± 1 with probabilities $1/2$). Define $S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i$ so $S_n \xrightarrow{d} \mathcal{N}(0,1)$ by the CLT. Then

$$\begin{aligned}\mathrm{Var}(f(S_n)) &\leq \sum_{i=1}^{n} \mathbb{E}[\mathrm{Var}^{(i)}(f(S_n))] \\ &= \sum_{i=1}^{n} \mathbb{E}\left[\frac{1}{4}\left(f(S_n - \frac{X_i}{\sqrt{n}} + \frac{1}{\sqrt{n}}) - f(S_n - \frac{X_i}{\sqrt{n}} - \frac{1}{\sqrt{n}})\right)^2\right].\end{aligned}$$

For the rest of the proof we assume f is twice-continuously differentiable on a bounded domain. Then

$$f(S_n - \frac{X_i}{\sqrt{n}} + \frac{1}{\sqrt{n}}) = f(S_n) + f'(S_n)\frac{1 - X_i}{\sqrt{n}} + f''(\theta_1)\frac{(1 - X_i)^2}{2n}$$

$$f(S_n - \frac{X_i}{\sqrt{n}} - \frac{1}{\sqrt{n}}) = f(S_n) - f'(S_n)\frac{1 + X_i}{\sqrt{n}} + f''(\theta_2)\frac{(1 + X_i)^2}{2n}$$

so

$$|f(S_n - \frac{X_i}{\sqrt{n}} + \frac{1}{\sqrt{n}}) - f(S_n - \frac{X_i}{\sqrt{n}} - \frac{1}{\sqrt{n}})| \leq |f'(S_n)|\frac{2}{\sqrt{n}} + \|f''\|_\infty \frac{2}{n}.$$

Hence

$$\begin{aligned}&|f(S_n - \frac{X_i}{\sqrt{n}} + \frac{1}{\sqrt{n}}) - f(S_n - \frac{X_i}{\sqrt{n}} - \frac{1}{\sqrt{n}})|^2 \\ &\leq |f'(S_n)|^2 \frac{4}{n} + \frac{4\|f''\|_\infty^2}{n^2} + \frac{8|f'(S_n)|\|f''\|_\infty}{n^{3/2}}\end{aligned}$$

and summing over $\{1, \dots, n\}$ we get

$$\mathrm{Var}(f(S_n)) \leq \mathbb{E}[f'(S_n)^2] + \frac{\|f''\|_\infty^2}{n} + \frac{8\mathbb{E}[|f'(S_n)|]\|f''\|_\infty}{n^{1/2}}$$

so taking $n \to \infty$ gives the result. $\square$

Entropy

Definition. For a random variable taking values on a discrete set $\mathcal{X}$ with PMF P_X, the *Shannon entropy* is defined as $H(X) = H(P_X) = \mathbb{E}[-\log P_X(X)]$.

Definition. Given two probability measures P, Q on a discrete set $\mathcal{X}$, define the *relative entropy* or *Kullback-Leibler divergence* $D(Q\|P) = \sum_x q(x) \log \frac{q(x)}{p(x)}$ where p, q are the PMF's of P, Q respectively.

Some basic properties of relative entropy are

1. $D(Q\|P) \geq 0$ with equality iff $Q = P$;
2. $D(Q\|P)$ is jointly convex, i.e

   $$D(\lambda Q_1 + (1-\lambda)Q_2 \| \lambda P_1 + (1-\lambda)P_2) \leq \lambda D(Q_1\|P_1) + (1-\lambda)D(Q_2\|P_2).$$

Suppose $|\mathcal{X}| < \infty$, then

$$D(Q\|U) = \log|X| - H(Q)$$

where $U \sim \text{Uniform}(\mathcal{X})$.

Definition. We define the *conditional entropy* $H(Y|X)$ by

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

Note $H(Y|X) \leq H(Y)$ by concavity of H together with Jensen. We define the *joint entropy* $H(X,Y)$ by

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y) = \mathbb{E}[-\log P_{X,Y}(X,Y)].$$

Theorem (Chain rule). $H(X_1, \ldots, X_n) = \sum_{i=1}^n H(X_i|X_{1:i-1})$.

Proof. We have

$$\begin{aligned} H(X_1, \ldots, X_n) &= \mathbb{E}[-\log P_{X_{1:n}}(X_{1:n})] \\ &= \mathbb{E}\left[-\log \prod_{i=1}^n P_{X_i|X_{1:i-1}}(X_i|X_{1:i-1})\right] \\ &= \sum_{i=1}^n H(X_i|X_{1:i-1}). \end{aligned}$$

□

Theorem (Chain rule for KL-divergence)**.** *Let P, Q be measures on $\mathcal{X}^n$. Then*

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

Proof. We have

$$\begin{aligned}
D(Q\|P) &= \sum_{x_{1:n}} q(x_{1:n}) \log \frac{q(x_{1:n})}{p(x_{1:n})} \\
&= \mathbb{E}_Q \left[\log \frac{q(X_{1:n})}{p(X_{1:n})}\right] \\
&= \mathbb{E}_Q \left[\log \prod_{i=1}^{n} \frac{q(X_i|X_{1:i-1})}{p(X_i|X_{1:i-1})}\right] \\
&= \sum_{i=1}^{n} \mathbb{E}_Q \left[\log \frac{q(X_i|X_{1:i-1})}{p(X_i|X_{1:i-1})}\right].
\end{aligned}$$

Note that

$$\begin{aligned}
\sum_{i=1}^{n} \mathbb{E}_Q \left[\log \frac{q(X_i|X_{1:i-1})}{p(X_i|X_{1:i-1})}\right] &= \sum_{x_{1:i}} q(x_{1:i}) \log \frac{q(x_i|x_{1:i-1})}{p(x_i|x_{1:i-1})} \\
&= \sum_{x_{1:i-1}} q(x_{1:i-1}) \left[\sum_{x_i} q(x_i|x_{i-1}) \log \frac{q(x_i|x_{1:i-1})}{p(x_i|x_{1:i-1})}\right] \\
&= \mathbb{E}_{Q_{X_{1:i-1}}}[D(Q_{X_i|X_{1:i-1}}\|P_{X_i|X_{1:i-1}})] \\
&:= D(Q_{X_i|X_{1:i-1}}\|P_{X_i|X_{1:i-1}}|Q_{X_{1:i-1}}).
\end{aligned}$$

Hence

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

□

Usually we'll have $P = P_1 \otimes P_2 \otimes \ldots \otimes P_n$, which simplifies this expression. If $Q = Q_1 \otimes Q_2 \otimes \ldots \otimes Q_n$ then it simplifies further to

$$D(Q\|P) = \sum_{i=1}^{n} D(Q_i\|P_i).$$

Theorem (Han's inequality for Shannon entropy)**.** *We have*

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

Example. Let $X_{1:n}$ be sampled iid from the uniform distribution on $A \subseteq \mathbb{Z}^n$. Then $H(X_{1:n}) = \log |A|$. Then Han's inequality implies

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

which is called the Loomis-Whitney inequality.