Contents

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

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

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

Theorem 0.1 (Central Limit Theorem) Let $X_1, ..., X_n$ be IID RVs with mean $\mathbb{E}[X_1] = \mu$. Let $\text{Var}(X_1) = \sigma^2 < \infty$. Then $\frac{1}{\sigma\sqrt{n}} \sum_{i=1}^n (X_i - \mu) \underset{D}{\to} N(0, 1)$, 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) $\mathbb{P}(|X - \mu| \geq \varepsilon) \leq \frac{\text{Var}(X)}{\varepsilon^2}$ for all $\varepsilon > 0$.

Corollary 0.3 $\mathbb{P}\left(\left|\sum_{i=1}^n (X_i) - \frac{n}{2}\right| \geq t\right) \leq \frac{\text{Var}\left(\sum_{i=1}^n X_i\right)}{t^2} = n\frac{\sigma^2}{t^2} \leq \delta$ 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 $X = \sum_{i=1}^n X_i$, $X_i \sim \text{Geom}\left(\frac{i}{n}\right)$. $\mathbb{E}[X] = \sum_i \mathbb{E}[X_i] \approx n \log n$ (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

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

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. $\square$

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. $\square$

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. $\square$

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

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

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

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

□

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

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

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

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

The set of all such 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 $\psi''_Y(\lambda) \leq (b-a)^2/4$ and $Y \in \mathcal{G}((b-a)^2/4)$.

Proof (Hints).

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

$\square$

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

$$P_\lambda(y) = \frac{e^{\lambda y}P(y)}{\sum_z P(z)e^{\lambda z}},$$

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

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

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 $\left|Y - \frac{b-a}{2}\right|^2 \leq (b-a)^2/4$. So we have

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

Finally, using a Taylor expansion at 0, we obtain

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

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

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

Theorem 1.20 (Hoeffding's Inequality) Let $X_1, ..., X_n$ 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 $\psi_{X_i}(\lambda) = \log\left(1 + \frac{\sigma_i^2}{c^2}(e^{\lambda c} - \lambda c - 1)\right) \le \frac{\sigma_i^2}{c^2}(e^{\lambda c} - \lambda c - 1)$. So by additivity of ψ, we have

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

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

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

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

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

So we are done by the Chernoff Bound. □

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

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

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

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

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 $X_{i:j} = (X_i, ..., X_j)$.

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

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

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

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 $\Delta_i = E_i\left(Z - E^{(i)} Z\right)$.

□

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

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

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 $\Delta_i = E_i Z - E_{i-1} Z = E_i\left(Z - E^{(i)} Z\right)$, 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,

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

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

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

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

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

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

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 $\mathbb{E}[Z] = \mathbb{E}[Z_i']$,

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

Thus we have

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

Finally, recall that $\mathrm{Var}(X) = \inf_a \mathbb{E}\left[(X-a)^2\right]$, with equality if $a = \mathbb{E}[X]$. So $\mathrm{Var}^{(i)}(Z) = \inf_{Z_i} E^{(i)}\left((Z - Z_i)^2\right)$, with equality if $Z_i = E^{(i)}Z$. Taking expectations and summing completes the proof. $\square$