Notes for High-Dimensional Probability*

Vishwak Srinivasan

Contents

*Vershynin [2019]

1 Preliminaries

1.1 Example on approximate Caratheodory's Theorem

First, we begin by discussing Caratheodory's Theorem:

Theorem 1.1.1 (Caratheodory's Theorem)**.** *Consider a convex set $S \subseteq \mathbb{R}^p$. Any point $x \in S$ can be represented as a convex combination of at most $p+1$ distinct points from S.*

Remark. This result is a popular result in convex analysis, and is tight. The tight lower bound is achieved by a simplex in p dimensions, which corresponds to $p+1$ vertices.

Now, we seek an approximation of the above theorem like so: given k points $\{x_i\}_{i=1}^k \subset S$, is it possible to approximate a point $x \in S$? We answer this in the affirmative below:

Theorem 1.1.2 (Approx. Caratheodory's Theorem)**.** *Given $x \in S \subseteq \mathbb{R}^p$, where S is convex, there exists a set of k points $\{x_i\}_{i=1}^k \in S$, such that the following holds:*

$$\left\| x - \frac{1}{k}\sum_{i=1}^{k} x_i \right\|_2 \leq \frac{\text{diam}(S)}{\sqrt{k}}$$

where $\text{diam}(S) = \sup_{s,t \in S} \|s - t\|_2$.

Proof. By the fact that $x \in S$, we know that we can write x as a convex combination of a subset $\{z_i\}_{i=1}^m$ that satisfy $\text{CONV}(\{z_i\}_{i=1}^m) = S$, where $m \leq p+1$. Let the coefficients be $\{\lambda_i\}_{i=1}^m$ where $\sum_{i=1}^m \lambda_i = 1$ and $\lambda_i \geq 0$ for all $i \in [m]$.

Consider a random variable Z that takes m different values from the set $\{z_i\}_{i=1}^m$ with probability λ_i. Note that $\mathbb{E}[Z] = x$, since $\mathbb{E}[Z] = \sum_{i=1}^m \Pr(Z = z_i) z_i = \sum_{i=1}^m \lambda_i z_i = x$.

We know that for any $x \in \mathbb{R}^p$ and independent random variables $\{Z_i\}_{i=1}^k$ that satisfy $\mathbb{E}[Z_i] = x$ for all $i \in [k]$:

$$\begin{aligned}
\mathbb{E}\left[\left\| x - \frac{1}{k}\sum_{i=1}^{k} Z_i \right\|_2^2\right] &= \mathbb{E}\left[\left\| \frac{1}{k}\sum_{i=1}^{k} (x - Z_i) \right\|_2^2\right] \\
&= \frac{1}{k^2}\mathbb{E}\left[\left\| \sum_{i=1}^{k} (x - Z_i) \right\|_2^2\right] \\
&\overset{(i)}{=} \frac{1}{k^2}\sum_{i=1}^{k} \mathbb{E}\left[\|x - Z_i\|_2^2\right] \\
&\overset{(ii)}{\leq} \frac{1}{k^2}\sum_{i=1}^{k} \text{diam}(S)^2 = \frac{\text{diam}(S)^2}{k}
\end{aligned}$$

Step (i) holds true due to Lemma 1.1.1. Step (ii) follows from the fact that $Z_i, x \in S$ which implies that $\|Z_i - x\|_2 \leq \text{diam}(S)$ followed by the fact that $\mathbb{E}[c] = c$ for constant c.

Therefore, there exists a realization of $\{Z_i\}_{i=1}^k$, that satisfies:

$$\left\| x - \frac{1}{k}\sum_{i=1}^{k} Z_i \right\|_2 \leq \frac{\text{diam}(S)}{\sqrt{k}}$$

□

Remark. First note the dimension independence in the result. Secondly, in the special case where S consists of elements with bounded norms i.e., $\|x\|_2 \leq B$ for all $x \in S$, the diameter of the set is bounded by $2B$ by an application of the triangle inequality. Finally, note that if we have $k \to \infty$ samples from the set, then our approximation is going to be perfect.

The method used to prove Theorem 1.1.1 is called Maurey's Empirical Method.

1.1.1 Auxiliary Lemmata

Lemma 1.1.1. *Let $\{X_i\}_{i=1}^k$ be a set of independent zero-mean random variables. The following holds true:*

$$\mathbb{E}\left[\left\|\sum_{i=1}^{k} X_i\right\|_2^2\right] = \sum_{i=1}^{k} \mathbb{E}\left[\|X_i\|_2^2\right]$$

Proof. First note that:

$$\begin{aligned}\left\|\sum_{i=1}^{k} X_i\right\|_2^2 &= \left\langle \sum_{i=1}^{k} X_i, \sum_{j=1}^{k} X_j \right\rangle \\ &= \sum_{i=1}^{k}\sum_{j=1}^{k} X_i^T X_j \\ &= \sum_{i=1}^{k} \|X_i\|_2^2 + 2\sum_{\substack{i,j=1 \\ i\neq j}}^{k} X_i^T X_j\end{aligned}$$

Taking expectations on both sides:

$$\begin{aligned}\mathbb{E}\left[\left\|\sum_{i=1}^{k} X_i\right\|_2^2\right] &= \mathbb{E}\left[\sum_{i=1}^{k} \|X_i\|_2^2\right] + 2\mathbb{E}\left[\sum_{\substack{i,j=1 \\ i\neq j}}^{k} X_i^T X_j\right] \\ &= \sum_{i=1}^{k} \mathbb{E}\left[\|X_i\|_2^2\right] + 2\sum_{\substack{i,j=1 \\ i\neq j}}^{k} \mathbb{E}\left[X_i^T X_j\right]\end{aligned}$$

Since X_is are independent, $\mathbb{E}\left[X_i^T X_j\right] = \mathbb{E}\left[X_i\right]^T \mathbb{E}\left[X_j\right] = 0$, and this completes the proof. □

Lemma 1.1.2. *For all integers $m \in [1, n]$, we have the following series of inequalities:*

$$\left(\frac{n}{m}\right)^m \leq \binom{n}{m} \leq \sum_{k=0}^{m} \binom{n}{m} \leq \left(\frac{en}{m}\right)^m$$

Proof. First inequality:

$$\binom{n}{m} m^m = \frac{n!}{(n-m)! \cdot m!} m^m \geq \frac{n!}{(n-m)!} \geq n^m \Rightarrow \binom{n}{m} \geq \left(\frac{n}{m}\right)^m$$

Second inequality:

$$\binom{n}{m} \leq \binom{n}{m} + \sum_{k=0}^{m-1} \binom{n}{k} = \sum_{k=0}^{m} \binom{n}{k}$$

Third inequality:

$$\left(\frac{m}{n}\right)^m \sum_{k=0}^{m} \binom{n}{k} \leq \sum_{k=0}^{m} \binom{n}{k} \left(\frac{m}{n}\right)^k \leq \sum_{k=0}^{n} \binom{n}{k} \left(\frac{m}{n}\right)^k = \left(1+\frac{m}{n}\right)^n \leq e^m \Rightarrow \sum_{k=0}^{m} \binom{n}{k} \leq \left(\frac{en}{m}\right)^m$$

□

1.2 Quantities and Inequalities associated with RVs

- Expectation: $\mathbb{E}[X]$
- Variance: $\text{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$
- MGF: $M_X(t) = \mathbb{E}[e^{tX}]$, $t \in \mathbb{R}$
- p^{th} moment: $\mathbb{E}[X^p]$ and p^{th} absolute moment: $\mathbb{E}[|X|^p]$
- L^p norm: $\|X\|_{L^p} = \sqrt[p]{\mathbb{E}[|X|^p]}$
- L^∞ norm: $\|X\|_{L^\infty} = \text{ess sup}\,|X|$, where $\text{ess sup}\,|X|$ denotes the supremum over all set with measure not 0. Also note that: $\text{ess sup}\,|X| \leq \sup|X|$.
- Covariance: $\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$
- CDF: $F_X(t) = \Pr(X \leq t)$, $t \in \mathbb{R}$

For a convex function f and any random variable X, we have by *Jensen's inequality* that:

$$f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$$

Consequently, for a concave function f and any random variable X, we have:

$$f(\mathbb{E}[X]) \geq \mathbb{E}[f(X)]$$

As a special case, consider $f(x) : x^{q/p}$ where $q > p$. Note that f is convex. Therefore:

$$\left(\mathbb{E}\left[|X|^p\right]\right)^{q/p} \leq \mathbb{E}\left[|X|^q\right] \Rightarrow \|X\|_{L^p} \leq \|X\|_{L^q}$$

Another inequality is the *Cauchy-Schwarz inequality*, which states that for any two RVs X and Y:

$$\mathbb{E}[|XY|] \leq \sqrt{\mathbb{E}[X^2]}\sqrt{\mathbb{E}[Y^2]} = ||X||_{L^2}||Y||_{L^2}$$

We also have *Holder's inequality* which generalizes *Cauchy-Schwarz* to dual norms as:

$$\mathbb{E}[|XY|] \leq ||X||_{L^p}||Y||_{L^q} \qquad ; \qquad \frac{1}{p} + \frac{1}{q} = 1$$

The following lemma characterizes the expectation as a quantity involving only tails:

Lemma 1.2.1. *Consider a non-negative random variable X. The expectation of this random variable can be written as:*

$$\mathbb{E}[X] = \int_0^\infty \Pr(X > t)dt$$

Proof. For any $x \geq 0$, we have that:

$$x = \int_0^\infty \mathbf{1}_{\{t<x\}} dt = \int_0^x 1 dt + \int_x^\infty 0 dt$$

Therefore:

$$\begin{aligned} X = \int_0^\infty \mathbf{1}_{\{t<X\}} dt \Rightarrow \mathbb{E}[X] &= \mathbb{E}\left[\int_0^\infty \mathbf{1}_{\{t<X\}} dt\right] \\ &= \int_0^\infty \int_{-\infty}^\infty \mathbf{1}_{\{t<x\}} \Pr(X = x) dx dt \\ &= \int_0^\infty \int_t^\infty \Pr(X = x) dx dt \\ &= \int_0^\infty \Pr(X > t) dt \end{aligned}$$

□

A simple generalization for real-valued random variables from the proof of Lemma 1.2.1 is as follows:

Corollary 1.2.1. *Consider a real valued random variable X. The expectation of this random variable can be written as:*

$$\mathbb{E}[X] = \int_0^\infty \Pr(X > t) dt - \int_{-\infty}^0 \Pr(X < t) dt$$

An application of Lemma 1.2.1 is to use it to bound the p^{th} absolute moments via tails:

Corollary 1.2.2. *For any random variable X:*

$$\mathbb{E}\left[|X|^p\right] = \int_0^\infty p t^{p-1} \Pr(|X| > t) dt$$

Classical inequalities: Markov and Chebyshev's:

Lemma 1.2.2 (Markov's Inequality)**.** *Consider a non-negative random variable X. Then the tails of X can be bounded as:*

$$\Pr(X > t) \leq \frac{\mathbb{E}[X]}{t}$$

Proof. Note that:

$$\mathbb{E}[X] = \mathbb{E}[X \cdot \mathbf{1}_{\{X>t\}}] + \mathbb{E}[X \cdot \mathbf{1}_{\{X \leq t\}}] \geq \mathbb{E}[X \cdot \mathbf{1}_{\{X>t\}}] \geq t\mathbb{E}[\mathbf{1}_{\{X>t\}}] = t \Pr(X > t) \Rightarrow \Pr(X > t) \leq \frac{\mathbb{E}[X]}{t}$$

□

Corollary 1.2.3 (Chebyshev's Inequality)**.** *Consider a random variable X. Then the probability of deviation from the expectation of X can be bounded as:*

$$\Pr(|X - \mathbb{E}[X]| > t) \leq \frac{\mathrm{Var}(X)}{t^2}$$

Proof. Take $Y = |X - \mathbb{E}[X]|$ as the random variable as apply Markov's inequality:

$$\Pr(Y > t) = \Pr(Y^2 > t^2) \leq \frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{t^2}$$

□

Remark. Note that one can achieve better dependence on t by using higher moments - provided they exist:

$$\Pr(Y > t) = \Pr(Y^{2k} > t^{2k}) \leq \frac{\mathbb{E}[(X - \mathbb{E}[X])^{2k}]}{t^{2k}}$$

1.3 Basic Limit Theorems

Theorem 1.3.1 (Strong Law of Large Numbers)**.** *Let $\{X_i\}_{i=1}^n$ be a sequence of identically and independently distributed random variables with mean μ. The quantity $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ satisfies:*

$$\bar{X}_n \xrightarrow{a.s.} \mu$$

as $n \to \infty$.

Here $\xrightarrow{a.s.}$ denotes *almost sure convergence*, which is:

$$\Pr\left(\lim_{n\to\infty} \bar{X}_n = \mu\right) = 1$$

There is a *weak law of large numbers*, which can be derived from Chebyshev's Inequality, for distributions with bounded variance. It is stated below:

Corollary 1.3.1 (Weak Law of Large Numbers)**.** *Let $\{X_i\}_{i=1}^n$ be a sequence of identically and independently distributed random variables with mean μ and variance $\sigma^2 < \infty$. The quantity $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ satisfies:*

$$\bar{X}_n \xrightarrow{p} \mu$$

where $\xrightarrow{p}$ denotes convergence in probability*, which is;*

$$\forall \epsilon > 0, \qquad \lim_{n\to\infty} \Pr\left(|\bar{X}_n - \mu| > \epsilon\right) = 0$$

Proof. First note $\mathbb{E}\left[\bar{X}_n\right] = \mu$, and hence $\mathrm{Var}(\bar{X}_n) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left(X_i\right) = \frac{\sigma^2}{n}$.

By Chebyshev's inequality, for any $\epsilon > 0$:

$$\Pr(|\bar{X}_n - \mu| > \epsilon) \leq \frac{\sigma^2}{n\epsilon} \Rightarrow \lim_{n\to\infty} \Pr(|\bar{X}_n - \mu| > \epsilon) = 0 \ \ (\because \text{Sandwich theorem})$$

□

Remark. This weak result is *weak* because $\xrightarrow{a.s.}$ implies $\xrightarrow{p.}$.

Next, we state a result that gives the asymptotic distribution of $\bar{X}_n$.

Theorem 1.3.2 (Central Limit Theorem)**.** *Let $\{X_i\}_{i=1}^n$ be a sequence of identically and independently distributed random variables with mean μ and variance $\sigma^2 < \infty$. Define $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then:*

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} \mathcal{N}(0,1) \qquad \text{as } n \to \infty$$

While this result states that the deviation between the sample mean and population mean is 0 in the limit, we can give some non asymptotic guarantees on the deviation as follows:

Lemma 1.3.1. *Let $\{X_i\}_{i=1}^n$ be a sequence of identically and independently distributed random variables with mean μ and variance $\sigma^2 < \infty$. We have that:*

$$\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right|\right] = O\left(\frac{1}{\sqrt{n}}\right)$$

Proof. By Jensen's inequality:

$$\mathbb{E}\left[|Z|\right] \leq \sqrt{\mathbb{E}\left[Z^2\right]}$$

(Note that this also follows from the fact that $\|Z\|_{L^1} \leq \|Z\|_{L^2}$)

Therefore:

$$\begin{aligned}
\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right|\right] &\leq \sqrt{\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right)^2\right]} \\
&= \sqrt{\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} (X_i - \mu)\right)^2\right]} \\
&= \sqrt{\frac{1}{n^2}\mathbb{E}\left[\left(\sum_{i=1}^{n} (X_i - \mu)\right)^2\right]} \\
&\overset{(i)}{=} \sqrt{\frac{1}{n^2}\sum_{i=1}^{n} \mathbb{E}\left[(X_i - \mu)^2\right]} \\
&= \sqrt{\frac{\sigma^2}{n}} = O\left(\frac{1}{\sqrt{n}}\right)
\end{aligned}$$

where Step (i) follows from Lemma 1.1.1 for 1-D random variables. □

A special case of the Central Limit Theorem is to provide approximate distributions for binomial distributions. Recall that the binomial distribution $\text{Bin}(n, p)$ is the sum of n independent Bernoulli distribution with parameter p. Therefore, we get that:

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} = \frac{n\bar{X}_n - n\mu}{\sigma\sqrt{n}} = \frac{B_{n,p} - np}{\sqrt{n}\sqrt{p(1-p)}} \xrightarrow{d} \mathcal{N}(0, 1) \text{ as } n \to \infty$$

where $X_i \sim \text{Ber}(p), i \in [n]$ and $B_{n,p} \sim \text{Bin}(n, p)$. This means that $B_{n,p} \xrightarrow{d} \mathcal{N}(np, np(1-p))$ as $n \to \infty$.

However, there is a better limit theorem in the regime where $p \to \infty, n \to \infty$ and $np = \lambda > 0$. This is the Poisson Limit Theorem:

Theorem 1.3.3 (Poisson Limit Theorem)**.** *Consider $\{X_i\}_{i=1}^n$ to be n independent Bernoulli variables with parameters p_i. Then, for $n \to \infty$, $\max_{i \in [n]} p_i \to 0$ and $\sum_{i=1}^{n} p_i = \lambda > 0$, we have that:*

$$\sum_{i=1}^{n} X_i \xrightarrow{d} \text{Poi}(\lambda)$$

Remark. In the special case when all p_is are equal, we obtain the same result with $n \to \infty$, $p \to 0$ and $np = \lambda > 0$ as described informally earlier.

2 Concentration inequalities

2.1 Basic Gaussian Inequalities

Lemma 2.1.1 (Mill's inequalities)**.** *Let $g \sim \mathcal{N}(0,1)$. We have the following lower and upper bounds for the tail $\Pr(g > t)$, $t > 0$ as follows:*

$$\left(\frac{1}{t} - \frac{1}{t^3}\right)\frac{1}{\sqrt{2\pi}}e^{-t^2/2} \leq \Pr(g > t) \leq \frac{1}{t}\cdot\frac{1}{\sqrt{2\pi}}e^{-t^2/2}$$

Proof. First, the upper bound:

$$\begin{aligned}
\Pr(g > t) &= \frac{1}{\sqrt{2\pi}}\int_t^\infty e^{-x^2/2}dx \\
&= \frac{1}{t}\cdot\frac{1}{\sqrt{2\pi}}\int_t^\infty te^{-x^2/2}dx \\
&\leq \frac{1}{t}\cdot\frac{1}{\sqrt{2\pi}}\int_t^\infty xe^{-x^2/2}dx \\
&= \frac{1}{t}\cdot\frac{1}{\sqrt{2\pi}}\int_{t^2/2}^\infty ye^{-y}dy \qquad \left(\because y = \frac{x^2}{2}\right) \\
&= \frac{1}{t}\cdot\frac{1}{\sqrt{2\pi}}e^{-t^2/2}
\end{aligned}$$

Second, the lower bound:

$$\begin{aligned}
\Pr(g > t) &= \frac{1}{\sqrt{2\pi}}\int_t^\infty e^{-x^2/2}dx \\
&\geq \frac{1}{\sqrt{2\pi}}\int_t^\infty \left(1 - \frac{3}{x^4}\right)e^{-x^2/2}dx \qquad \left(\because 1 - \frac{3}{x^4} \leq 1 \ \forall \ x > 0\right) \\
&\geq \left(\frac{1}{t} - \frac{1}{t^3}\right)\frac{1}{\sqrt{2\pi}}e^{-t^2/2}
\end{aligned}$$

An alternative proof for the lower bound can be obtained as follows. First note that for $\phi(z) = \frac{1}{\sqrt{2\pi}}e^{-z^2/2}$, we have (see [Wainwright, 2019]):

$$\phi'(z) = \frac{1}{\sqrt{2\pi}}\cdot -ze^{-z^2/2} = -z\phi(z)$$

Therefore:

$$\begin{aligned}
\Pr(g > t) &= \int_t^\infty \phi(x)dx \\
&= \int_t^\infty -\frac{\phi'(x)}{x}dx \\
&= -\frac{1}{x}\phi(x)\Big|_t^\infty - \int_t^\infty \frac{1}{x^2}\phi(x)dx \qquad \left(\because \text{int. by parts with } f(x) = -\frac{1}{x}, g(x) = \phi'(x)\right) \\
&= \frac{\phi(t)}{t} + \int_t^\infty \frac{1}{x^3}\phi'(x)dx \\
&= \frac{\phi(t)}{t} + \frac{\phi(x)}{x^3}\Big|_t^\infty + \underbrace{\int_t^\infty \frac{3}{x^4}\phi(x)dx}_{\geq 0} \qquad \left(\because \text{int. by parts with } f(x) = -\frac{1}{x^3}, g(x) = \phi'(x)\right) \\
&\geq \frac{\phi(t)}{t} - \frac{\phi(t)}{t^3}
\end{aligned}$$

□

Remark. Note that one can get tail bounds for $\mathcal{N}(0, \sigma^2)$ by simply reparameterising the integrals as:

$$\left(\frac{\sigma}{t} - \frac{\sigma}{t^3}\right) \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \leq \Pr(g > t) \leq \frac{\sigma}{t} \cdot \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$$

The Central Limit Theorem (Theorem 1.3.2) states that averages tend in distribution to a Gaussian. But what can be said about the distribution function itself? The following theorem gives this result:

Theorem 2.1.1 (Berry-Esseen CLT)**.** *Let $\{X_i\}_{i=1}^n$ be a sequence of identically and independently distributed random variables with mean μ and variance $\sigma^2 < \infty$. Define $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\bar{Z}_n = \sqrt{n}\frac{\bar{X}_n - \mu}{\sigma}$. Then:*

$$\left|\Pr(\bar{Z}_n > t) - \Pr(g > t))\right| \leq \frac{\rho}{\sqrt{n}}$$

where $\rho = \frac{\mathbb{E}\left[|X_i - \mu|^3\right]}{\sigma^3}$, $i \in [n]$ and $g \sim \mathcal{N}(0, 1)$.

Remark. This theorem basically states that the error of approximation scales as $O\left(\frac{1}{\sqrt{n}}\right)$, which is bad, since we can't always leverage the normal approximation from the central limit theorem always.

2.1.1 Auxiliary Lemmata

Lemma 2.1.2. *Let $g \sim \mathcal{N}(0, 1)$. For $t \geq 1$, we have that:*

$$\mathbb{E}\left[g^2 \mathbf{1}_{\{g>t\}}\right] = \frac{t}{\sqrt{2\pi}} e^{-t^2/2} + \Pr(g > t) \leq \left(t + \frac{1}{t}\right) \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$$

Proof.

$$\begin{aligned}
\mathbb{E}\left[g^2 \mathbf{1}_{\{g>t\}}\right] &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 \mathbf{1}_{\{x>t\}} e^{-x^2/2} dx \\
&= \frac{1}{\sqrt{2\pi}} \int_{t}^{\infty} x^2 e^{-x^2/2} dx \\
&= \frac{1}{\sqrt{2\pi}} \left(\left(x \cdot e^{-x}\right)\Big|_{t^2/2}^{\infty} + \int_{t}^{\infty} e^{-x^2/2} dx \right) \qquad \left(\because \text{int. by parts with } f(x) = x, g(x) = xe^{-x^2/2}\right) \\
&= \frac{t}{\sqrt{2\pi}} e^{-t^2/2} + \Pr(g > t) \\
&\leq \frac{t}{\sqrt{2\pi}} e^{-t^2/2} + \frac{1}{t} \cdot \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \\
&= \left(t + \frac{1}{t}\right) \frac{1}{\sqrt{2\pi}} e^{-t^2/2}
\end{aligned}$$

□

2.2 Hoeffding's Inequality

First, we describe a simple bounded random variable namely the Rademacher random variable. X with support $\{-1, +1\}$ is said to be distributed w.r.t. a Rademacher distribution if:

$$\Pr(X = +1) = \Pr(X = -1) = \frac{1}{2}$$

Let's discuss about the basic quantities attributed to a random variable. Note that $\mathbb{E}[X] = 0$ and $\text{Var}[X] = \frac{1}{2}$. Also the moment generating function is:

$$M_X(t) = \mathbb{E}\left[e^{tX}\right] = \frac{1}{2}e^t + \frac{1}{2}e^{-t} = \cosh(t)$$

We have the first concentration inequality for these random variables.

Theorem 2.2.1. *Let $\{X_i\}_{i=1}^n$ be a sequence of n i.i.d. Rademacher random variables. Then, for any $a \in \mathbb{R}^n$ we have:*

$$\Pr\left(\sum_{i=1}^n a_i X_i > t\right) \leq \exp\left(-\frac{t^2}{2||a||_2^2}\right)$$

Proof. Let's begin with Markov's inequality:

$$\begin{aligned}
\Pr\left(\sum_{i=1}^n a_i X_i > t\right) &= \Pr\left(\exp\left(s \cdot \sum_{i=1}^n a_i X_i\right) > \exp(s \cdot t)\right) \qquad (s > 0) \\
&\overset{(i)}{\leq} \frac{\mathbb{E}\left[\exp\left(s \cdot \sum_{i=1}^n a_i X_i\right)\right]}{\exp(st)} \\
&= \frac{1}{\exp(st)} \mathbb{E}\left[\prod_{i=1}^n \exp(s \cdot a_i X_i)\right] \\
&\overset{(ii)}{=} \frac{1}{\exp(st)} \prod_{i=1}^n \mathbb{E}\left[\exp(s \cdot a_i X_i)\right] \\
&\overset{(iii)}{=} \frac{1}{\exp(st)} \prod_{i=1}^n \cosh(s \cdot a_i) \\
&\overset{(iv)}{\leq} \frac{1}{\exp(st)} \prod_{i=1}^n \exp\left(\frac{s^2 a_i^2}{2}\right) \\
&= \frac{1}{\exp(st)} \exp\left(\frac{s^2 ||a||_2^2}{2}\right)
\end{aligned}$$

Step (i) uses Markov's inequality over $Y = \exp\left(s \cdot \sum_{i=1}^n a_i X_i\right)$. Step (ii) uses the fact that X_is are independent, and Step (iii) uses the MGF derived earlier. Step (iv) can be attributed to Lemma 2.2.1. Therefore, we have shown for any $s > 0$ that:

$$\Pr\left(\sum_{i=1}^n a_i X_i > t\right) \leq \exp\left(s^2 \frac{||a||_2^2}{2} - st\right)$$

To obtain the tighest / smallest upper bound, we minimize the RHS. The minimizer of the quadratic: $s^2 \frac{||a||_2^2}{2} - st$ w.r.t s is $\frac{t}{||a||_2^2}$, resulting in:

$$\Pr\left(\sum_{i=1}^n a_i X_i > t\right) \leq \exp\left(-\frac{t^2}{2||a||_2^2}\right)$$

□

Remark. Note that the first step generally holds for any invertible function f. We used such a trick in proving Chebyshev's inequality. Also note that since $\mathbb{E}[X_i] = 0$ for all $i \in [n]$, we didn't have an explicit mean term.

Also, in the case where $a_i = \frac{1}{n}$ i.e., we care about the concentration of the sample average to the mean, we get:

$$\Pr\left(\bar{X}_n > t\right) \leq \exp\left(-\frac{nt^2}{2}\right)$$

Generally, high probability guarantees are given by setting the RHS to a confidence parameter δ, like so:

$$\bar{X}_n \leq \sqrt{\frac{2}{n}\log\left(\frac{1}{\delta}\right)} \qquad \text{w. p. } \geq 1 - \delta$$

Fix the error to ϵ. Smaller the δ, the more confident the estimate, but requires more samples.

There is a two sided version to this inequality as well, and the stated below:

Corollary 2.2.1. *Let $\{X_i\}_{i=1}^n$ be a sequence of n i.i.d. Rademacher random variables. Then, for any $a \in \mathbb{R}^n$ we have:*

$$\Pr\left(\left|\sum_{i=1}^n a_i X_i\right| > t\right) \leq 2\exp\left(-\frac{t^2}{2||a||_2^2}\right)$$

Proof. Simple probability states that if for two events A and B such that $A \Rightarrow B$, then $\Pr(A) \leq \Pr(B)$. Therefore:

$$\left|\sum_{i=1}^n a_i X_i\right| > t \Rightarrow \sum_{i=1}^n a_i X_i > t \vee \sum_{i=1}^n a_i X_i < -t$$

$$\Rightarrow \Pr\left(\left|\sum_{i=1}^n a_i X_i\right| > t\right) \leq \Pr\left(\sum_{i=1}^n a_i X_i > t \vee \sum_{i=1}^n a_i X_i < -t\right) \leq \Pr\left(\sum_{i=1}^n a_i X_i > t\right) + \Pr\left(\sum_{i=1}^n a_i X_i < -t\right)$$

$$\Rightarrow \Pr\left(\left|\sum_{i=1}^n a_i X_i\right| > t\right) \leq 2\exp\left(-\frac{t^2}{2||a||_2^2}\right)$$

□

An interesting application of this inequality is as follows: say that you have a set of objects of two kinds, equally probably occuring. If you want to provide high probability guarantees on the number of objects of one kind, then you could use it as follows. For each object, assign $Z_i = 1$ if the object is of the kind you want and 0 otherwise. Now, $S_n = \sum_{i=1}^n Z_i$ is the number of objects of the kind you desire. However, Z_i is supported on $\{0, 1\}$, which means you would have to reparameterize. Note that:

$$\Pr(S_n > t) = \Pr(2S - n > 2t - n) = \Pr\left(\sum_{i=1}^n (2Z_i - 1) > 2t - n\right) \leq \exp\left(-\frac{(2t-n)^2}{2n}\right)$$

So, if you anticipated having at least $\frac{3}{4}n$ objects of the kind you want, then this would happen with probability at most $e^{-n/8}$.

Now the derivation above has given an idea of how to obtain bounds for *bounded random variables* - of which the Rademacher random variable is an example. Below, we have a generalized theorem:

Theorem 2.2.2. *Let $\{X_i\}_{i=1}^n$ be a sequence of n independent bounded random variables i.e., every X_i satisfies $X_i \in [l_i, u_i]$. Then, we have that:*

$$\Pr\left(\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) > t\right) \leq \exp\left(-\frac{2t^2}{||u-l||_2^2}\right)$$

Proof. The standard series of steps used in the proof of Theorem 2.2.1 yields:

$$\begin{aligned}
\Pr\left(\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) > t\right) &= \Pr\left(\exp\left(s \cdot \sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])\right) > \exp(st)\right) \qquad (s > 0) \\
&\leq \frac{1}{\exp(st)}\mathbb{E}\left[\exp\left(s \cdot \sum_{i=1}^{n}(X_i - \mathbb{E}[X_i])\right)\right] \\
&\leq \frac{1}{\exp(st)}\prod_{i=1}^{n}\mathbb{E}\left[\exp(s \cdot (X_i - \mathbb{E}[X_i]))\right]
\end{aligned}$$

To bound $\mathbb{E}\left[\exp(s \cdot (X_i - \mathbb{E}[X_i]))\right]$: let $Z_i = X_i - \mathbb{E}[X_i]$ for $i \in [n]$. Therefore, we focus on the quantity $\mathbb{E}\left[\exp(s \cdot Z_i)\right]$. Consider Z_i' to be an independent copy of Z_i[1].

$$\mathbb{E}_{Z_i}\left[\exp(s \cdot Z_i)\right] = \mathbb{E}_{Z_i}\left[\exp(s \cdot (Z_i - \mathbb{E}_{Z_i'}[Z_i']))\right] = \mathbb{E}_{Z_i}\left[\exp(s \cdot (\mathbb{E}_{Z_i'}[Z_i - Z_i']))\right] \leq \mathbb{E}_{Z_i, Z_i'}\left[\exp\left(s \cdot (Z_i - Z_i')\right)\right]$$

Now, we have by symmetricity:

$$\mathbb{E}_{Z_i, Z_i'}\left[\exp\left(s \cdot (Z_i - Z_i')\right)\right] = \mathbb{E}_{\epsilon \sim \text{Rad}}\left[\mathbb{E}_{Z_i, Z_i'}\left[\exp\left(s\epsilon(Z_i - Z_i')\right)\right]\right] = \mathbb{E}_{Z_i, Z_i'}\left[\mathbb{E}_{\epsilon \sim \text{Rad}}\left[\exp\left(\epsilon \cdot s(Z_i - Z_i')\right)\right]\right]$$

We showed via Lemma 2.2.1 that:

$$\mathbb{E}_{\epsilon \sim \text{Rad}}\left[\exp\left(\epsilon \cdot s(Z_i - Z_i')\right)\right] \leq \exp\left(\frac{s^2(Z_i - Z_i')^2}{2}\right) \leq \exp\left(\frac{s^2(u_i - l_i)^2}{2}\right)$$

and therefore:

$$\mathbb{E}_{Z_i}\left[\exp(s \cdot Z_i)\right] \leq \exp\left(\frac{s^2(u_i - l_i)^2}{2}\right)$$

Therefore, we obtain:

$$\Pr\left(\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) > t\right) \leq \frac{1}{\exp(st)}\exp\left(\frac{s^2||u-l||_2^2}{2}\right)$$

From the proof of Theorem 2.2.1, with $a = u - l$, we get:

$$\Pr\left(\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) > t\right) \leq \exp\left(-\frac{t^2}{2||u-l||_2^2}\right)$$

Note the sub-optimal constant. Hoeffding showed that one can obtain a better bound on the moments as:

$$\mathbb{E}_{Z_i}\left[\exp(s \cdot Z_i)\right] \leq \exp\left(\frac{s^2(u_i - l_i)^2}{8}\right)$$

which leads to:

$$\Pr\left(\sum_{i=1}^{n}(X_i - \mathbb{E}[X_i]) > t\right) \leq \exp\left(-\frac{2t^2}{||u-l||_2^2}\right)$$

□

[1]The proof is adapted from [Wainwright, 2019, Example 2.3]

Remark. The introduction and use of the Rademacher random variable is known as the *symmetrization trick*. The result of Hoeffding that gives the improved bound on the MGF is classically know as *Hoeffding's Lemma*.

The basic idea is to obtain bounds on the MGF. We will see later about classes of distributions that have a bounded MGF, and therefore allow for the application of this theorem despite being unbounded.

With these results, let's discuss some guarantees for robust mean estimation. Given n samples from a distribution with bounded variance σ^2 and mean μ, the mean $\bar{X}_n$ satisfies:

$$\Pr\left(|\bar{X}_n - \mu| > \epsilon\right) \leq \frac{\mathrm{Var}(\bar{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$$

Therefore, we require $n = O\left(\frac{\sigma^2}{\epsilon^2}\right)$ samples to obtain an estimate with high probability, say at least 0.75.

Now, given k sets of n samples, can we do any better w.r.t. dependence on success probability? This is tantamount to being given $\tilde{n} = n \cdot k$ samples, and splitting them into k subsets of size n each. The answer is yes, and can be obtained by considering the median of these k means.

To see this, consider k Bernoulli random variables $\{Z_i\}_{i=1}^k$, which take value 1 if $|\bar{X}_{i,n} - \mu| > \epsilon$ and 0 otherwise. Since this is a bounded variable, we can apply the result from Theorem 2.2.2 and the fact that $\mathbb{E}[Z_i] = \Pr(Z_i = 1) \leq \frac{1}{4}$ to get:

$$\Pr\left(\sum_{i=1}^{k} Z_i > \frac{k}{2}\right) \leq \Pr\left(\sum_{i=1}^{k} (Z_i - \mathbb{E}[Z_i]) > \frac{k}{4}\right) \leq \exp\left(-\frac{k}{8}\right)$$

Therefore, if $k = \log\left(\frac{8}{\delta}\right)$, we get a mean estimate which with probability atleast $1 - \delta$, provides an ϵ-error estimate.

2.2.1 Auxiliary Lemmata

Lemma 2.2.1. *For any x, we have that* $\cosh(x) \leq \exp\left(\frac{x^2}{2}\right)$.

Proof. Note that $\cosh(x) = \frac{1}{2}e^x + \frac{1}{2}e^{-x}$

The Maclaurin series for e^x thus gives:

$$\begin{aligned}
\cosh(x) &= \frac{1}{2}\sum_{k=0}^{\infty}\frac{x^{2k}}{(2k)!} + \frac{1}{2}\sum_{k=0}^{\infty}\frac{x^{2k+1}}{(2k+1)!} + \frac{1}{2}\sum_{k=0}^{\infty}\frac{x^{2k}}{(2k)!} - \frac{1}{2}\sum_{k=0}^{\infty}\frac{x^{2k+1}}{(2k+1)!} \\
&= \sum_{k=0}^{\infty}\frac{x^{2k}}{(2k)!} \\
&\overset{(i)}{\leq} \sum_{k=0}^{\infty}\frac{x^{2k}}{2^k k!} = \sum_{k=0}^{\infty}\frac{\left(x^2/2\right)^k}{k!} = \exp\left(\frac{x^2}{2}\right)
\end{aligned}$$

where in Step (i) we have used $(2k)! \geq 2^k \cdot k!$. □

Lemma 2.2.2. *Let X is a non-negative random variable whose density function is uniformly bounded by* 1. *Then:*

$$M_X(-t) \leq \frac{1}{t} \qquad t > 0$$

Proof. By definition:

$$M_X(-t) = \mathbb{E}[e^{-tX}] = \int_0^\infty p(x)e^{-tx}dx \leq \int_0^\infty e^{-tx}dx = \frac{1}{t}$$

□

2.3 Chernoff's Inequality

We discussed earlier about the Poisson Limit Theorem 1.3.3. While Hoeffding's inequality gives us good concentration bounds for bounded random variables such as Bernoulli, we can hope for better bounds by leveraging the fact that for small parameters, the distribution of sum of Bernoulli's could be closer to a Poisson as compared to a Gaussian.

Theorem 2.3.1. *Let $\{X_i\}_{i=1}^n$ be a sequence of n independent Bernoulli random variables with parameters $\{p_i\}_{i=1}^n$. Define the sum $S_n = \sum_{i=1}^n X_i$ and mean $\mu = \mathbb{E}[S_n]$. For any $t > \mu$, we have:*

$$\Pr(S_n > t) \leq e^{-\mu}\left(\frac{e\mu}{t}\right)^t$$

Proof. Again, following the usual steps give us:

$$\Pr(S_n > t) \leq \frac{1}{\exp(st)}\prod_{i=1}^n \mathbb{E}\left[\exp(s \cdot X_i)\right] \qquad (s > 0)$$

Now, since X_i is a Bernoulli random variable, we get:

$$\mathbb{E}\left[\exp(s \cdot X_i)\right] = \exp(s)p_i + (1 - p_i) = 1 + (\exp(s) - 1)p_i \leq \exp((\exp(s) - 1)p_i)$$

and hence:

$$\Pr(S_n > t) \leq \frac{1}{\exp(st)}\exp(\exp(s) - 1)\mu)$$

To obtain the tightest upper bound, we minimize for $s > 0$. The minimizer is $s = \log\left(\frac{t}{\mu}\right)$, and the minimum is:

$$\Pr(S_n > t) \leq \left(\frac{\mu}{t}\right)^t e^t e^{-\mu}$$

□

We can also obtain a bound on the other side of the tail like so.

Corollary 2.3.1. *Let $\{X_i\}_{i=1}^n$ be a sequence of n independent Bernoulli random variables with parameters $\{p_i\}_{i=1}^n$. Define the sum $S_n = \sum_{i=1}^n X_i$ and mean $\mu = \mathbb{E}[S_n]$. For any $t > \mu$, we have:*

$$\Pr(S_n \leq t) \leq e^{-\mu}\left(\frac{e\mu}{t}\right)^t$$

Proof. We repeat the proof of Theorem 2.3.1 using the fact that $\Pr(S_n \leq t) = \Pr(-S_n \geq -t)$.

$$\Pr(-S_n \geq -t) \leq \frac{1}{\exp(-st)}\prod_{i=1}^n \mathbb{E}\left[\exp(-s \cdot X_i)\right] \qquad (s > 0)$$

Now, since X_i is a Bernoulli random variable, we get:

$$\mathbb{E}\left[\exp(-s \cdot X_i)\right] = \exp(-s)p_i + (1-p_i) = 1 + (\exp(-s)-1)p_i \leq \exp((\exp(-s)-1)p_i)$$
$$\Rightarrow \Pr(-S_n \geq -t) \leq \exp(st)\exp((\exp(-s)-1)\mu)$$

To obtain the tighest upper bound, we minimize for $s > 0$. The minimizer is $s = \log\left(\frac{\mu}{t}\right)$, and the minimum is:

$$\Pr(S_n \leq t) \leq \left(\frac{\mu}{t}\right)^t e^t e^{-\mu}$$

□

An interesting consequence of the Chernoff bound when $\mu = \sum_{i=1}^{n} p_i = \lambda$, but when $n \to \infty$ is that, we retrieve tail bounds for the Poisson random variable:

Corollary 2.3.2. *Consider the setting of Theorem 2.3.1. Then, for $n \to \infty$ and $\mu = \lambda$, we have:*

$$\Pr(X \geq t) \leq \left(\frac{e\lambda}{t}\right)^t e^{-\lambda}$$

where $X \sim \text{Poi}(\lambda)$.

Proof. Note by the Poisson Limit Theorem 1.3.3, we have that:

$$\lim_{n\to\infty} \Pr(S_n \geq t) = \Pr(X \geq t)$$

and this completes the proof. □

A popular version of Chernoff's inequality is to give bounds on the event of deviation from the mean μ. We state it as follows:

Corollary 2.3.3. *Let $\{X_i\}_{i=1}^n$ be a sequence of n independent Bernoulli random variables with parameters $\{p_i\}_{i=1}^n$. Define the sum $S_n = \sum_{i=1}^{n} X_i$ and mean $\mu = \mathbb{E}[S_n]$. For any $\delta \in (0, 1]$, we have:*

$$\Pr\left(|S_n - \mu| \leq \delta\mu\right) \leq 2\exp\left(-c\mu\delta^2\right)$$

where $c > 0$ is a universal constant.

Proof. To prove this bound, we will transform it to the form we say in Theorem 2.3.1.

$$|S_n - \mu'| \leq \delta\mu' \Rightarrow S_n - \mu' > \delta\mu' \vee S_n - \mu < -\delta\mu'$$

Now, we will bound the probability of the event $S_n - \mu' > \delta\mu'$. For any $\mu' \geq \mu$:

$$\begin{aligned}
\Pr\left(S_n - \mu' > \delta\mu'\right) &= \Pr\left(S_n > \mu'(\delta+1)\right) \\
&\leq \exp(-s\mu'(1+\delta))\exp((\exp(s)-1)\mu) \qquad (s > 0) \\
&\leq \exp((\exp(s)-1)\mu - s\mu'(1+\delta))
\end{aligned}$$

Set $s = \log(1+\delta)$ to get:

$$\Pr\left(S_n - \mu' > \delta\mu'\right) \leq \exp(\mu\delta - \mu'(1+\delta)\log(1+\delta)) \leq \exp(\mu'\delta - \mu'(1+\delta)\log(1+\delta)) \leq \exp\left(-\frac{\mu'\delta^2}{3}\right)$$

where the last step uses Lemma 2.3.1.

Therefore, for any $\mu' \geq \mu$, we have:

$$\Pr\left(S_n > \mu'(1+\delta)\right) \leq \exp\left(-\frac{\mu'\delta^2}{3}\right)$$

Now let's bound the probability of the event $S_n - \mu' < -\delta\mu'$. For any $\mu' \leq \mu$:

$$\begin{aligned}\Pr\left(S_n - \mu' \leq -\delta\mu'\right) &= \Pr\left(-S_n \geq \mu'(\delta-1)\right) \\ &\leq \exp(s\mu'(1-\delta))\exp((\exp(-s)-1)\mu) \qquad (s>0) \\ &\leq \exp((\exp(-s)-1)\mu + s\mu'(1-\delta))\end{aligned}$$

Set $s = -\log(1-\delta)$ to get:

$$\Pr\left(S_n - \mu' \leq -\delta\mu'\right) \leq \exp(-\delta\mu - \mu'(1-\delta)\log(1-\delta)) \leq \exp(-\delta\mu' - \mu'(1-\delta)\log(1-\delta)) \leq \exp\left(-\frac{\mu'\delta^2}{3}\right)$$

Therefore, for any $\mu' \leq \mu$, we have:

$$\Pr\left(S_n - \mu' \leq -\delta\mu'\right) \leq \exp\left(-\frac{\mu'\delta^2}{3}\right)$$

Now, one can apply a union bound for $\mu' = \mu$, to get:

$$\Pr\left(|S_n - \mu| > \delta\mu\right) \leq 2\exp\left(-\frac{\mu\delta^2}{3}\right)$$

□

2.3.1 Auxiliary Lemmata

Lemma 2.3.1. *For $x \in [0,1]$, we have the following inequality:*

$$\log(1+x) \geq \frac{2x}{2+x}$$

and consequently:

$$x - (1+x)\log(1+x) \leq \frac{-x^2}{2+x} \leq \frac{-x^2}{3}$$

Proof. We know that:

$$\log(1+x) = \sum_{k=1}^{\infty} \frac{x^k(-1)^{k-1}}{k} \qquad \frac{2x}{2+x} = \sum_{k=1}^{\infty} \frac{x^k(-1)^{k-1}}{2^{k-1}}$$

Therefore:

$$\log(1+x) - \frac{2x}{2+x} = \sum_{k=1}^{\infty} x^k(-1)^{k-1}\left(\frac{1}{k} - \frac{2}{2^k}\right)$$

We know that since $x \in [0,1]$:

$$\frac{x^k}{k} - \frac{x^{k+1}}{k+1} \geq \frac{x^k}{k(k+1)}$$

and

$$\frac{x^k}{2^{k-1}} - \frac{x^{k+1}}{2^k} \geq \frac{x^k}{2^{k+1}}$$

Due to these facts, and using $k(k+1) \leq 2^{k+1}$, we get $\log(1+x) - \frac{2x}{2+x} \geq 0$. As a consequence:

$$x - (1+x)\log(1+x) \leq x\left(\frac{-x}{2+x}\right) = \frac{-x^2}{2+x} \leq -\frac{x^2}{3}$$

□

2.4 Sub-Gaussian Random Variables

Earlier, we saw Hoeffding's inequality applied to Rademacher random variables. In general, can we say anything about the random variables that satisfy:

$$\Pr\left(\left|\sum_{i=1}^{n} a_i X_i\right| > t\right) \leq 2\exp\left(-c\frac{t^2}{\|a\|_2^2}\right)$$

In the special case that $n = 1$ and $a_i = 1$, we have:

$$\Pr\left(|X_i| > t\right) \leq 2\exp(-ct^2)$$

Such random variables are called *sub-Gaussian random variables*. From Theorem 2.2.2, we know that bounded random variables are sub-Gaussian with an appropriate constant $c > 0$. Incidentally, Gaussian random variables also fall in this class (hence the name). The following lemma proves this:

Lemma 2.4.1. *Let $X \sim \mathcal{N}(0,1)$. Then for any $t > 0$:*

$$\Pr(|X| > t) \leq 2\exp\left(-\frac{t^2}{2}\right)$$

Proof. First, let's consider $\Pr(X > t)$.

$$\Pr(X > t) = \Pr\left(e^{sX} > e^{st}\right) \leq \frac{\mathbb{E}\left[e^{sX}\right]}{e^{st}} = e^{\frac{s^2}{2} - st}$$

Minimizing the upper bound, gives $s = t$ and consequently:

$$\Pr(X > t) \leq \exp\left(-\frac{t^2}{2}\right)$$

Next, let's consider $\Pr(X < -t)$. Note that $\Pr(X < -t) = \Pr(-X > t) = \Pr(X > t)$ since X is symmetric about 0, and we have bounded this earlier. Combining them we get:

$$\Pr(|X| > t) \leq 2\Pr(X > t) \leq 2e^{-\frac{t^2}{2}}$$

□

Now, let's also compute moments of a standard normal random variable:

Lemma 2.4.2. *Let $X \sim \mathcal{N}(0,1)$. Then for any $p \geq 1$, we have:*

$$||X||_{L^p} = \sqrt{2}\left(\frac{\Gamma\left((1+p)/2\right)}{\Gamma(1/2)}\right)^{1/p}$$

Proof.

$$\mathbb{E}\left[|X|^p\right] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}|x|^p e^{-x^2/2}dx = \frac{2}{\sqrt{2\pi}}\int_0^{\infty}x^p e^{-x^2/2}dx \overset{(i)}{=} \frac{\sqrt{2}^p}{\sqrt{\pi}}\int_0^{\infty}t^{\frac{p+1}{2}-1}e^{-t}dt = \sqrt{2}^p\frac{\Gamma\left((p+1)/2\right)}{\sqrt{\pi}}$$

where Step (i) uses the change of variable $t = \frac{x^2}{2}$

By the definition of the L^p norm, we have:

$$\|X\|_{L^p} = \left(\mathbb{E}[|X|^p]\right)^{1/p} = \sqrt{2}\left(\frac{\Gamma\left((1+p)/2\right)}{\Gamma(1/2)}\right)^{1/p}$$

Note that:

$$\Gamma\left((1+p)/2\right) = \underbrace{\frac{p-1}{2}\frac{p-3}{2}\cdots\frac{1}{2}}_{\frac{p}{2}\text{ terms}}\cdots\Gamma(1/2) = O\left(\left(\frac{p}{2}\right)^{p/2}\right) \Rightarrow \left(\frac{\Gamma\left((1+p)/2\right)}{\Gamma(1/2)}\right)^{1/p} = O(\sqrt{p})$$

which means that $\|X\|_{L^p} = O(\sqrt{p})$. □

Now, let's look at some properties of sub-Gaussian random variables. Recall that a sub-Gaussian random variable X is one that satisfies:

$$\Pr(|X| > t) \leq 2\exp\left(-\frac{t^2}{K_1^2}\right) \tag{1}$$

for some $K_1 > 0$, and $t \geq 0$. We will see that the following properties are equivalent:

Lemma 2.4.3. *Let X be a sub-Gaussian random variable as defined in Eqn 1. The following statement are equivalent:*

(i) *X is sub-Gaussian*

(ii) $\|X\|_{L^p} \leq K_2\sqrt{p}$ *for all* $p \geq 1$

(iii) $\mathbb{E}[\exp(\lambda^2 X^2)] \leq \exp(K_3^2\lambda^2)$ *for all* $|\lambda| \leq \frac{1}{K_3}$

(iv) $\mathbb{E}\left[\exp\left(\frac{X^2}{K_4^2}\right)\right] \leq 2$

(v) $\mathbb{E}\left[\exp(\lambda X)\right] \leq \exp(K_5^2\lambda^2)$ *for all* $\lambda \in \mathbb{R}$ *(if X is zero mean sub-Gaussian)*

Proof. **(i)** $\Rightarrow$ **(ii)** First, consider $\mathbb{E}[|X|^p]$. From Corollary 1.2.2, we have:

$$\begin{aligned}
\mathbb{E}[|X|^p] &= \int_0^{\infty} pt^{p-1}\Pr(|X| > t)dt \\
&\leq \int_0^{\infty} 2pt^{p-1}\exp\left(-\frac{t^2}{K_1^2}\right)dt \\
&= p\int_0^{\infty} K_1^p u^{\frac{p}{2}-1}\exp(-u)du \\
&= pK_1^p\Gamma(p/2) \\
&\leq 3K_1^p p\left(\frac{p}{2}\right)^{p/2}
\end{aligned}$$

This implies: $\|X\|_{L^p} \leq 3K_1\sqrt{p}$, and therefore for $K_2 \leq 3K_1$, we get obtain the proof.

(ii) ⇒ (iii) By the series expansion:

$$\begin{aligned}\mathbb{E}\left[\exp(\lambda^2 X^2)\right] &= \mathbb{E}\left[\sum_{k=0}^{\infty}\frac{(\lambda^2X^2)^k}{k!}\right]\\ &= 1+\sum_{k=1}^{\infty}\frac{\lambda^{2k}\mathbb{E}[X^{2k}]}{k!}\\ &\le 1+\sum_{k=1}^{\infty}\frac{\lambda^{2k}(2K_1^{2k}k\Gamma(k))}{k!}\\ &= 1+\sum_{k=1}^{\infty}\lambda^{2k}2K_1^{2k}\\ &\le \sum_{k=0}^{\infty}2^{2k}\lambda^{2k}K_1^{2k}\end{aligned}$$

The above series converges if $2^2\lambda^2K_1^2<1\Rightarrow|\lambda|<\frac{1}{2K_1}$. If this condition is satisfied:

$$\mathbb{E}\left[\exp(\lambda^2X^2)\right]\le\frac{1}{1-4\lambda^2K_1^2}\overset{(i)}{\le}e^{8\lambda^2K_1^2}$$

where Step (i) is due to Lemma 2.4.4 where $4\lambda^2K_1^2\le\frac{1}{2}\Rightarrow|\lambda|\le\frac{1}{2\sqrt{2}K_1}$. This yields (iii) with $K_3=2\sqrt{2}K_1$.

(iii) ⇒ (iv) Choose $\lambda=\frac{1}{\sqrt{2}K_3}$ in (iii) to get:

$$\mathbb{E}\left[\exp\left(\frac{X^2}{2K_3^2}\right)\right]\le\exp(1/2)\le 2$$

(iv) ⇒ (i)

$$\Pr\left(\frac{|X|}{K_4}>t\right)=\Pr\left(\frac{X^2}{K_4^2}>t^2\right)=\mathbb{E}\left[\exp\left(\frac{X^2}{K_4^2}\right)\right]\exp(-t^2)\le 2\exp(-t^2)$$

With $t'=K_4t$, we get:

$$\Pr(|X|>t')\le 2\exp\left(-\frac{t^2}{K_4^2}\right)$$

(ii) ⇒ (v) From the series expansion:

$$\begin{aligned}\mathbb{E}[\exp(\lambda X)] &= \sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\mathbb{E}[X^k]\\ &\le 1+\sum_{k=2}^{\infty}\frac{\lambda^k}{k!}\mathbb{E}[|X|^k]\\ &\overset{(i)}{\le} 1+\sum_{k=1}^{\infty}\frac{2^k\lambda^{2k}}{(2k)!}\mathbb{E}[|X|^{2k}]\\ &\le 1+\sum_{k=1}^{\infty}\frac{2^k\lambda^{2k}}{(2k)!}K_1^{2k}\underbrace{k\Gamma(k)}_{k!}\\ &\le 1+\sum_{k=1}^{\infty}\lambda^{2k}K_1^{2k}\frac{2^kk!}{(2k)!}\\ &\overset{(ii)}{\le}\sum_{k=0}^{\infty}\frac{\lambda^{2k}(K_1)^{2k}}{k!}=\exp(\lambda^2K_1^2)\end{aligned}$$

where in Step (i) we have bounded odd moments using even moments (Lemma 2.4.5), and in Step (ii) we have used $\frac{2^k k!}{(2k)!} \leq \frac{1}{k!}$ (which follows from $\left(\frac{2k}{k}\right)^k \leq \binom{2k}{k}$).

(v) $\Rightarrow$ (i)

$$\Pr(X > t) = \Pr(\exp(\lambda X) > \exp(\lambda t)) \leq \frac{\mathbb{E}[\lambda X]}{e^{\lambda t}} \leq \mathbb{E}\left[\lambda^2 K_1^2 - \lambda t\right]$$

Minimizing the upper bound over $\lambda > 0$ gives:

$$\Pr(X > t) \leq \exp\left(-\frac{t^2}{2K_1^2}\right)$$

Analogously, we can obtain:

$$\Pr(X < -t) \leq \exp\left(-\frac{t^2}{2K_1^2}\right)$$

Combining these via the union bound gives:

$$\Pr(|X| > t) \leq 2\exp\left(-\frac{t^2}{2K_1^2}\right)$$

□

Remark. Note that (ii) states that sub-Gaussian random variables have bounded moments. It is also worth noting that when $\mathbb{E}[X] \neq 0$, we can still bound the moment generating function of X as follows:

Corollary 2.4.1. *Let X be a non-zero mean sub-Gaussian random variable satisfying $\Pr(|X| > t) \leq 2\exp\left(-\frac{t^2}{K_1^2}\right)$. Then we have that:*

$$\mathbb{E}[\exp(\lambda X)] \leq 2\exp(\lambda^2 K_1^2)$$

Proof. The proof mostly follows from the proof of **(ii)** $\Rightarrow$ **(v)**, except that:

$$\mathbb{E}[\exp(\lambda X)] \leq 1 + \mathbb{E}[\lambda X] + \sum_{k=1}^{\infty} \frac{2^k \lambda^{2k}}{(2k)!}\mathbb{E}[|X|^{2k}] \leq 1 + \frac{1}{2} + \frac{\lambda^2}{2}\mathbb{E}[|X|^2] + \sum_{k=1}^{\infty} \frac{2^k \lambda^{2k}}{(2k)!}\mathbb{E}[|X|^{2k}] \leq 2\sum_{k=0}^{\infty} \frac{2^k \lambda^{2k}}{(2k)!}\mathbb{E}[|X|^{2k}]$$

We know that $\sum_{k=0}^{\infty} \frac{2^k \lambda^{2k}}{(2k)!}\mathbb{E}[|X|^{2k}] \leq \exp(\lambda^2 K_1^2)$, and this completes the proof. □

Therefore, except for a constant factor, the bound on the MGF is the same for non-centered sub-Gaussian random variables.

2.4.1 Auxiliary Lemmata

Lemma 2.4.4. *For $x \in [0, 1/2]$, we have that $\frac{1}{1-x} \leq e^{2x}$.*

Proof. Both functions are monotonically increasing in $[0, 1/2]$ and satisfy: $f(0) = g(0)$ and $f(1/2) \leq g(1/2)$, with no crossings i.e., $\nexists x \in (0, 1/2]$ such that $f(x) = g(x)$. This means that $f(x) \leq g(x)$ for all $x \in [0, 1/2]$. □

Lemma 2.4.5. *For a random variable X, we have for $k \geq 0$:*

$$\mathbb{E}[|\lambda X|^{2k+1}] \leq \frac{1}{2}\mathbb{E}[|\lambda X|^{2k}] + \frac{1}{2}\mathbb{E}[|\lambda X|^{2k+2}]$$

As a consequence:

$$\mathbb{E}[\exp(\lambda X)] \leq 1 + \mathbb{E}[X] + \sum_{k=1}^{\infty} \frac{2^k \lambda^{2k}\mathbb{E}[|X|^{2k}]}{(2k)!}$$

Proof. By Cauchy-Schwarz and AM-GM inequality, we have:

$$\mathbb{E}[|\lambda X|^{2k+1}] \leq \sqrt{\mathbb{E}[|\lambda X|^{2k}]\mathbb{E}[|\lambda X|^{2k+2}]} \leq \frac{1}{2}\mathbb{E}[|\lambda X|^{2k}] + \frac{1}{2}\mathbb{E}[|\lambda X|^{2k+2}]$$

Now, by the series expansion:

$$\begin{aligned}
\mathbb{E}[\exp(\lambda X)] &\leq 1 + \mathbb{E}[\lambda X] + \sum_{k=2}^{\infty} \frac{\lambda^k \mathbb{E}[X^k]}{k!} \\
&\leq 1 + \mathbb{E}[X] + \sum_{k=2}^{\infty} \frac{\lambda^k \mathbb{E}[|X|^k]}{k!} \\
&\leq 1 + \mathbb{E}[X] + \sum_{k=1}^{\infty} \mathbb{E}[|X|^{2k}] \left(\frac{\lambda^{2k}}{(2k)!} + \frac{1}{2}\left(\frac{\lambda^{2k}}{(2k-1)!} + \frac{\lambda^{2k}}{(2k+1)!} \right) \right) \\
&\leq 1 + \mathbb{E}[X] + \sum_{k=1}^{\infty} \mathbb{E}[|X|^{2k}] \frac{\lambda^{2k} 2^k}{(2k)!}
\end{aligned}$$

□

References

Roman Vershynin. *High-dimensional probability*. 2019.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.