CHAPTER 1

Sub-Gaussian Random Variables

1.1 GAUSSIAN TAILS AND MGF

Recall that a random variable $X \in \mathbb{R}$ has Gaussian distribution iff it has a density p with respect to the Lebesgue measure on $\mathbb{R}$ given by

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big), \quad x \in \mathbb{R},$$

where $\mu = \mathbb{E}(X) \in \mathbb{R}$ and $\sigma^2 = \mathrm{var}(X) > 0$ are the *mean* and *variance* of X. We write $X \sim \mathcal{N}(\mu, \sigma^2)$. Note that $X = \sigma Z + \mu$ for $Z \sim \mathcal{N}(0,1)$ (called standard Gaussian) and where the equality holds in distribution. Clearly, this distribution has unbounded support but it is well known that it has *almost* bounded support in the following sense: $\mathbb{P}(|X-\mu| \le 3\sigma) \simeq 0.997$. This is due to the fast decay of the tails of p as $|x| \to \infty$ (see Figure 1.1). This decay can be quantified using the following proposition (Mills inequality).

Proposition 1.1. Let X be a Gaussian random variable with mean μ and variance σ^2 then for any $t > 0$, it holds

$$\mathbb{P}(X - \mu > t) \le \frac{1}{\sqrt{2\pi}} \frac{\mathrm{e}^{-\frac{t^2}{2\sigma^2}}}{t}.$$

By symmetry we also have

$$\mathbb{P}(X - \mu < -t) \le \frac{1}{\sqrt{2\pi}} \frac{\mathrm{e}^{-\frac{t^2}{2\sigma^2}}}{t}.$$

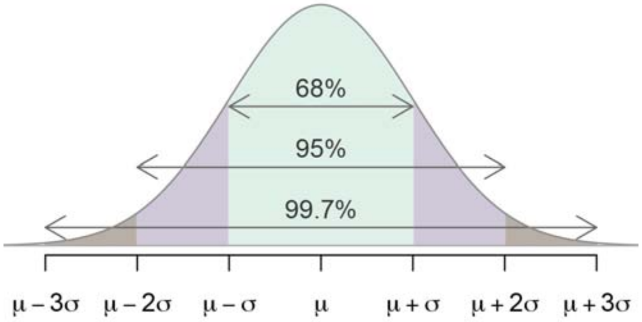

Figure 1.1. Probabilities of falling within 1, 2, and 3 standard deviations close to the mean in a Gaussian distribution. Source http://www.openintro.org/

and

$$\mathbb{P}(|X-\mu|>t) \le \sqrt{\frac{2}{\pi}}\frac{\mathrm{e}^{-\frac{t^2}{2\sigma^2}}}{t}\,.$$

Proof. Note that it is sufficient to prove the theorem for $\mu = 0$ and $\sigma^2 = 1$ by simple translation and rescaling. We get for $Z \sim \mathcal{N}(0,1)$,

$$\begin{aligned}
\mathbb{P}(Z>t) &= \frac{1}{\sqrt{2\pi}}\int_t^\infty \exp\Big(-\frac{x^2}{2}\Big)\mathrm{d}x \\
&\le \frac{1}{\sqrt{2\pi}}\int_t^\infty \frac{x}{t}\exp\Big(-\frac{x^2}{2}\Big)\mathrm{d}x \\
&= \frac{1}{t\sqrt{2\pi}}\int_t^\infty -\frac{\partial}{\partial x}\exp\Big(-\frac{x^2}{2}\Big)\mathrm{d}x \\
&= \frac{1}{t\sqrt{2\pi}}\exp(-t^2/2)\,.
\end{aligned}$$

The second inequality follows from symmetry and the last one using the union bound:

$$\mathbb{P}(|Z|>t) = \mathbb{P}(\{Z>t\}\cup\{Z<-t\}) \le \mathbb{P}(Z>t)+\mathbb{P}(Z<-t) = 2\mathbb{P}(Z>t)\,.$$

□

The fact that a Gaussian random variable Z has tails that decay to zero exponentially fast can also be seen in the *moment generating function* (MGF)

$$M\ :\ s \mapsto M(s) = \mathbb{E}[\exp(sZ)]\,.$$

Indeed in the case of a standard Gaussian random variable, we have

$$\begin{aligned} M(s) = \mathbb{E}[\exp(sZ)] &= \frac{1}{\sqrt{2\pi}} \int e^{sz} e^{-\frac{z^2}{2}} \mathrm{d}z \\ &= \frac{1}{\sqrt{2\pi}} \int e^{-\frac{(z-s)^2}{2} + \frac{s^2}{2}} \mathrm{d}z \\ &= e^{\frac{s^2}{2}} . \end{aligned}$$

It follows that if $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\mathbb{E}[\exp(sX)] = \exp\big(s\mu + \frac{\sigma^2 s^2}{2}\big)$.

1.2 SUB-GAUSSIAN RANDOM VARIABLES AND CHERNOFF BOUNDS

Definition and first properties

Gaussian tails are practical when controlling the tail of an average of independent random variables. Indeed, recall that if $X_1, \ldots, X_n$ are i.i.d $\mathcal{N}(\mu, \sigma^2)$, then $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i \sim \mathcal{N}(\mu, \sigma^2/n)$. Using Lemma 1.3 below for example, we get

$$\mathbb{P}(|\bar{X} - \mu| > t) \le 2\exp\big(-\frac{nt^2}{2\sigma^2}\big) .$$

Equating the right-hand side with some confidence level $\delta > 0$, we find that with probability at least[1] $1 - \delta$,

$$\mu \in \Big[\bar{X} - \sigma\sqrt{\frac{2\log(2/\delta)}{n}}, \bar{X} + \sigma\sqrt{\frac{2\log(2/\delta)}{n}}\Big] , \tag{1.1}$$

This is almost the confidence interval that you used in introductory statistics. The only difference is that we used an approximation for the Gaussian tail whereas statistical tables or software use a much more accurate computation. Figure 1.2 shows the ration of the width of the confidence interval to that of the confidence interval computer by the software R. It turns out that intervals of the same form can be also derived for non-Gaussian random variables as long as they have sub-Gaussian tails.

Definition 1.2. A random variable $X \in \mathbb{R}$ is said to be *sub-Gaussian* with variance proxy σ^2 if $\mathbb{E}[X] = 0$ and its moment generating function satisfies

$$\mathbb{E}[\exp(sX)] \le \exp\Big(\frac{\sigma^2 s^2}{2}\Big) , \quad \forall\, s \in \mathbb{R} . \tag{1.2}$$

In this case we write $X \sim \mathsf{subG}(\sigma^2)$. Note that $\mathsf{subG}(\sigma^2)$ denotes a class of distributions rather than a distribution. Therefore, we abuse notation when writing $X \sim \mathsf{subG}(\sigma^2)$.

More generally, we can talk about sub-Gaussian random vectors and matrices. A random vector $X \in \mathbb{R}^d$ is said to be *sub-Gaussian* with variance

[1] We will often commit the statement “at least” for brevity

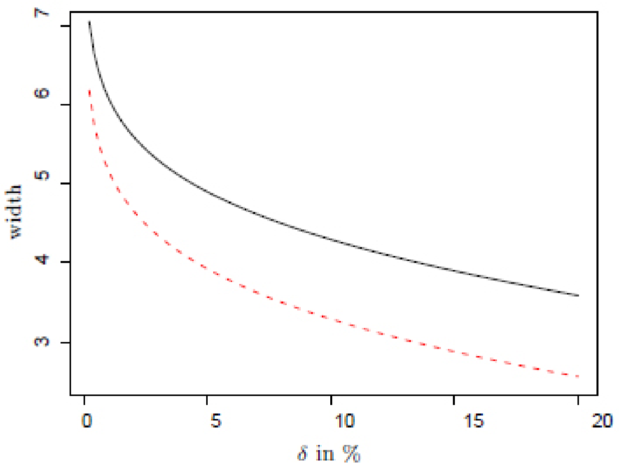

Figure 1.2. Width of confidence intervals from exact computation in R (red dashed) and (1.1) (solid black).

proxy σ^2 if $\mathbb{E}[X] = 0$ and $u^\top X$ is sub-Gaussian with variance proxy σ^2 for any unit vector $u \in \mathcal{S}^{d-1}$. In this case we write $X \sim \mathsf{subG}_d(\sigma^2)$. A random matrix $X \in \mathbb{R}^{d \times T}$ is said to be *sub-Gaussian* with variance proxy σ^2 if $\mathbb{E}[X] = 0$ and $u^\top X v$ is sub-Gaussian with variance proxy σ^2 for any unit vectors $u \in \mathcal{S}^{d-1}, v \in \mathcal{S}^{T-1}$. In this case we write $X \sim \mathsf{subG}_{d\times T}(\sigma^2)$.

This property can equivalently be expressed in terms of bounds on the tail of the random variable X.

Lemma 1.3. *Let $X \sim \mathsf{subG}(\sigma^2)$. Then for any $t > 0$, it holds*

$$\mathbb{P}[X > t] \le \exp\Big(-\frac{t^2}{2\sigma^2}\Big), \quad \text{and} \quad \mathbb{P}[X < -t] \le \exp\Big(-\frac{t^2}{2\sigma^2}\Big). \qquad (1.3)$$

Proof. Assume first that $X \sim \mathsf{subG}(\sigma^2)$. We will employ a very useful technique called **Chernoff bound** that allows to to translate a bound on the moment generating function into a tail bound. Using Markov's inequality, we have for any $s > 0$,

$$\mathbb{P}(X > t) \le \mathbb{P}\big(e^{sX} > e^{st}\big) \le \frac{\mathbb{E}\big[e^{sX}\big]}{e^{st}}.$$

Next we use the fact that X is sub-Gaussian to get

$$\mathbb{P}(X > t) \le e^{\frac{\sigma^2 s^2}{2} - st}.$$

The above inequality holds for any $s > 0$ so to make it the tightest possible, we minimize with respect to $s > 0$. Solving $\phi'(s) = 0$, where $\phi(s) = \frac{\sigma^2 s^2}{2} - st$, we find that $\inf_{s>0} \phi(s) = -\frac{t^2}{2\sigma^2}$. This proves the first part of (1.3). The second inequality in this equation follows in the same manner (recall that (1.2) holds for any $s \in \mathbb{R}$).

□

Moments

Recall that the absolute moments of $Z \sim \mathcal{N}(0, \sigma^2)$ are given by

$$\mathbb{E}[|Z|^k] = \frac{1}{\sqrt{\pi}}(2\sigma^2)^{k/2}\Gamma\big(\frac{k+1}{2}\big)$$

where $\Gamma(\cdot)$ denote the Gamma function defined by

$$\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}\mathrm{d}x\,, \quad t > 0\,.$$

The next lemma shows that the tail bounds of Lemma 1.3 are sufficient to show that the absolute moments of $X \sim \mathsf{subG}(\sigma^2)$ can be bounded by those of $Z \sim \mathcal{N}(0, \sigma^2)$ up to multiplicative constants.

Lemma 1.4. *Let X be a random variable such that*

$$\mathbb{P}[|X| > t] \le 2\exp\Big(-\frac{t^2}{2\sigma^2}\Big)\,,$$

then for any positive integer $k \ge 1$,

$$\mathbb{E}[|X|^k] \le (2\sigma^2)^{k/2}k\Gamma(k/2)\,.$$

In particular,

$$\big(\mathbb{E}[|X|^k]\big)^{1/k} \le \sigma e^{1/e}\sqrt{k}\,, \quad k \ge 2\,.$$

and $\mathbb{E}[|X|] \le \sigma\sqrt{2\pi}$.

Proof.

$$\begin{aligned}
\mathbb{E}[|X|^k] &= \int_0^\infty \mathbb{P}(|X|^k > t)\mathrm{d}t \\
&= \int_0^\infty \mathbb{P}(|X| > t^{1/k})\mathrm{d}t \\
&\le 2\int_0^\infty e^{-\frac{t^{2/k}}{2\sigma^2}}\mathrm{d}t \\
&= (2\sigma^2)^{k/2}k\int_0^\infty e^{-u}u^{k/2-1}\mathrm{d}u\,, \qquad\qquad u = \frac{t^{2/k}}{2\sigma^2} \\
&= (2\sigma^2)^{k/2}k\Gamma(k/2)
\end{aligned}$$

The second statement follows from $\Gamma(k/2) \le (k/2)^{k/2}$ and $k^{1/k} \le e^{1/e}$ for any $k \ge 2$. It yields

$$\big((2\sigma^2)^{k/2} k\Gamma(k/2)\big)^{1/k} \le k^{1/k}\sqrt{\frac{2\sigma^2 k}{2}} \le e^{1/e}\sigma\sqrt{k}\,.$$

Moreover, for $k = 1$, we have $\sqrt{2}\Gamma(1/2) = \sqrt{2\pi}$. □

Using moments, we can prove the following reciprocal to Lemma 1.3.

Lemma 1.5. *If (1.3) holds, then for any $s > 0$, it holds*

$$\mathbb{E}[\exp(sX)] \le e^{4\sigma^2 s^2}\,.$$

As a result, we will sometimes write $X \sim \mathsf{subG}(\sigma^2)$ when it satisfies (1.3).

Proof. We use the Taylor expansion of the exponential function as follows. Observe that by the dominated convergence theorem

$$\begin{aligned}
\mathbb{E}\big[e^{sX}\big] &\le 1 + \sum_{k=2}^{\infty} \frac{s^k \mathbb{E}[|X|^k]}{k!} \\
&\le 1 + \sum_{k=2}^{\infty} \frac{(2\sigma^2 s^2)^{k/2} k\Gamma(k/2)}{k!} \\
&= 1 + \sum_{k=1}^{\infty} \frac{(2\sigma^2 s^2)^{k} 2k\Gamma(k)}{(2k)!} + \sum_{k=1}^{\infty} \frac{(2\sigma^2 s^2)^{k+1/2} (2k+1)\Gamma(k+1/2)}{(2k+1)!} \\
&\le 1 + \big(2 + \sqrt{2\sigma^2 s^2}\big) \sum_{k=1}^{\infty} \frac{(2\sigma^2 s^2)^k k!}{(2k)!} \\
&\le 1 + \big(1 + \sqrt{\frac{\sigma^2 s^2}{2}}\big) \sum_{k=1}^{\infty} \frac{(2\sigma^2 s^2)^k}{k!} && 2(k!)^2 \le (2k)! \\
&= e^{2\sigma^2 s^2} + \sqrt{\frac{\sigma^2 s^2}{2}}(e^{2\sigma^2 s^2} - 1) \\
&\le e^{4\sigma^2 s^2}\,.
\end{aligned}$$

□

From the above Lemma, we see that sub-Gaussian random variables can be equivalently defined from their tail bounds and their moment generating functions, up to constants.

Sums of independent sub-Gaussian random variables

Recall that if $X_1, \dots, X_n$ are i.i.d $\mathcal{N}(0, \sigma^2)$, then for any $a \in \mathbb{R}^n$,

$$\sum_{i=1}^{n} a_i X_i \sim \mathcal{N}(0, |a|_2^2 \sigma^2).$$

If we only care about the tails, this property is preserved for sub-Gaussian random variables.

Theorem 1.6. *Let $X = (X_1, \dots, X_n)$ be a vector of independent sub-Gaussian random variables that have variance proxy σ^2. Then, the random vector X is sub-Gaussian with variance proxy σ^2.*

Proof. Let $u \in \mathcal{S}^{n-1}$ be a unit vector, then

$$\mathbb{E}[e^{su^\top X}] = \prod_{i=1}^n \mathbb{E}[e^{su_iX_i}] \le \prod_{i=1}^n e^{\frac{\sigma^2 s^2 u_i^2}{2}} = e^{\frac{\sigma^2 s^2 |u|_2^2}{2}} = e^{\frac{\sigma^2 s^2}{2}} .$$

□

Using a Chernoff bound, we immediately get the following corollary

Corollary 1.7. *Let $X_1, \dots, X_n$ be n independent random variables such that $X_i \sim \mathsf{subG}(\sigma^2)$. Then for any $a \in \mathbb{R}^n$, we have*

$$\mathbb{P}\Big[\sum_{i=1}^n a_i X_i > t\Big] \le \exp\Big(-\frac{t^2}{2\sigma^2|a|_2^2}\Big),$$

and

$$\mathbb{P}\Big[\sum_{i=1}^n a_i X_i < -t\Big] \le \exp\Big(-\frac{t^2}{2\sigma^2|a|_2^2}\Big)$$

Of special interest is the case where $a_i = 1/n$ for all i. Then, we get that the average $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$, satisfies

$$\mathbb{P}(\bar{X} > t) \le e^{-\frac{nt^2}{2\sigma^2}} \quad \text{and} \quad \mathbb{P}(\bar{X} < -t) \le e^{-\frac{nt^2}{2\sigma^2}}$$

just like for the Gaussian average.

Hoeffding's inequality

The class of subGaussian random variables is actually quite large. Indeed, Hoeffding's lemma below implies that all randdom variables that are bounded uniformly are actually subGaussian with a variance proxy that depends on the size of their support.

Lemma 1.8 (Hoeffding's lemma (1963))**.** *Let X be a random variable such that $\mathbb{E}(X) = 0$ and $X \in [a, b]$ almost surely. Then, for any $s \in \mathbb{R}$, it holds*

$$\mathbb{E}[e^{sX}] \le e^{\frac{s^2(b-a)^2}{8}} .$$

In particular, $X \sim \mathsf{subG}(\frac{(b-a)^2}{4})$.

Proof. Define $\psi(s) = \log \mathbb{E}[e^{sX}]$, and observe that and we can readily compute

$$\psi'(s) = \frac{\mathbb{E}[Xe^{sX}]}{\mathbb{E}[e^{sX}]}, \qquad \psi''(s) = \frac{\mathbb{E}[X^2 e^{sX}]}{\mathbb{E}[e^{sX}]} - \left[\frac{\mathbb{E}[Xe^{sX}]}{\mathbb{E}[e^{sX}]}\right]^2.$$

Thus $\psi''(s)$ can be interpreted as the variance of the random variable X under the probability measure $d\mathbb{Q} = \frac{e^{sX}}{\mathbb{E}[e^{sX}]} d\mathbb{P}$. But since $X \in [a,b]$ almost surely, we have, under any probability,

$$\mathrm{var}(X) = \mathrm{var}\big(X - \frac{a+b}{2}\big) \le \mathbb{E}\Big[\Big(X - \frac{a+b}{2}\Big)^2\Big] \le \frac{(b-a)^2}{4}.$$

The fundamental theorem of calculus yields

$$\psi(s) = \int_0^s \int_0^\mu \psi''(\rho)\, d\rho\, d\mu \le \frac{s^2(b-a)^2}{8}$$

using $\psi(0) = \log 1 = 0$ and $\psi'(0) = \mathbb{E}X = 0$. □

Using a Chernoff bound, we get the following (extremely useful) result.

Theorem 1.9 (Hoeffding's inequality)**.** *Let $X_1, \dots, X_n$ be n independent random variables such that almost surely,*

$$X_i \in [a_i, b_i], \qquad \forall\, i.$$

Let $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$, then for any $t > 0$,

$$\mathbb{P}(\bar{X} - \mathbb{E}(\bar{X}) > t) \le \exp\Big(-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\Big),$$

and

$$\mathbb{P}(\bar{X} - \mathbb{E}(\bar{X}) < -t) \le \exp\Big(-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\Big).$$

Note that Hoeffding's lemma is for *any* bounded random variables. For example, if one knows that X is a Rademacher random variable. Then

$$\mathbb{E}(e^{sX}) = \frac{e^s + e^{-s}}{2} = \cosh(s) \le e^{\frac{s^2}{2}}$$

Note that 2 is the best possible constant in the above approximation. For such variables $a = -1, b = 1$, $\mathbb{E}(X) = 0$ so Hoeffding's lemma yields

$$\mathbb{E}(e^{sX}) \le e^{\frac{s^2}{2}}.$$

Hoeffding's inequality is very general but there is a price to pay for this generality. Indeed, if the random variables have small variance, we would like to see it reflected in the exponential tail bound (like for the Gaussian case) but the variance does not appear in Hoeffding's inequality. We need a more refined inequality.

1.3 SUB-EXPONENTIAL RANDOM VARIABLES

What can we say when a centered random variable is not sub-Gaussian? A typical example is the double exponential (or Laplace) distribution with parameter 1, denoted by $\mathsf{Lap}(1)$. Let $X \sim \mathsf{Lap}(1)$ and observe that

$$\mathbb{P}(|X| > t) = e^{-t}, \quad t \geq 0\,.$$

In particular, the tails of this distribution do not decay as fast as the Gaussian ones (that decay as $e^{-t^2/2}$). Such tails are said to be *heavier* than Gaussian. This tail behavior is also captured by the moment generating function of X. Indeed, we have

$$\mathbb{E}\big[e^{sX}\big] = \frac{1}{1-s^2} \quad \text{if } |s| < 1\,,$$

and is not defined for $s \geq 1$. It turns out that a rather week condition on the moment generating function is enough to partially reproduce some of the bounds that we have proved for sub-Gaussian random variables. Observe that for $X \sim \mathsf{Lap}(1)$

$$\mathbb{E}\big[e^{sX}\big] \leq e^{2s^2} \quad \text{if } |s| < 1/2\,,$$

In particular, the Laplace distribution has its moment generating distribution that is bounded by that of a Gaussian in a neighborhood of 0 but does not even exist away from zero. It turns out that all distributions that have tails at least as heavy as that of a Laplace distribution satisfy such a property.

Lemma 1.10. *Let X be a centered random variable such that $\mathbb{P}(|X| > t) \leq 2e^{-2t/\lambda}$ for some $\lambda > 0$. Then, for any positive integer $k \geq 1$,*

$$\mathbb{E}[|X|^k] \leq \lambda^k k!\,.$$

Moreover,

$$\big(\mathbb{E}[|X|^k]\big)^{1/k} \leq 2\lambda k\,,$$

and the moment generating function of X satisfies

$$\mathbb{E}\big[e^{sX}\big] \leq e^{2s^2\lambda^2}\,, \qquad \forall |s| \leq \frac{1}{2\lambda}\,.$$

Proof.

$$\begin{aligned}
\mathbb{E}[|X|^k] &= \int_0^\infty \mathbb{P}(|X|^k > t)\mathrm{d}t \\
&= \int_0^\infty \mathbb{P}(|X| > t^{1/k})\mathrm{d}t \\
&\leq \int_0^\infty 2e^{-\frac{2t^{1/k}}{\lambda}}\mathrm{d}t \\
&= 2(\lambda/2)^k k \int_0^\infty e^{-u}u^{k-1}\mathrm{d}u\,, \qquad\qquad u = \frac{2t^{1/k}}{\lambda} \\
&\leq \lambda^k k\Gamma(k) = \lambda^k k!
\end{aligned}$$

The second statement follows from $\Gamma(k) \le k^k$ and $k^{1/k} \le e^{1/e} \le 2$ for any $k \ge 1$. It yields

$$\big(\lambda^k k\Gamma(k)\big)^{1/k} \le 2\lambda k\,.$$

To control the MGF of X, we use the Taylor expansion of the exponential function as follows. Observe that by the dominated convergence theorem, for any s such that $|s| \le 1/2\lambda$

$$\begin{aligned}
\mathbb{E}\big[e^{sX}\big] &\le 1 + \sum_{k=2}^{\infty} \frac{|s|^k \mathbb{E}[|X|^k]}{k!} \\
&\le 1 + \sum_{k=2}^{\infty} (|s|\lambda)^k \\
&= 1 + s^2\lambda^2 \sum_{k=0}^{\infty} (|s|\lambda)^k \\
&\le 1 + 2s^2\lambda^2 && |s| \le \frac{1}{2\lambda} \\
&\le e^{2s^2\lambda^2}
\end{aligned}$$

□

This leads to the following definition

Definition 1.11. A random variable X is said to be sub-exponential with parameter λ (denoted $X \sim \mathsf{subE}(\lambda)$) if $\mathbb{E}[X] = 0$ and its moment generating function satisfies

$$\mathbb{E}\big[e^{sX}\big] \le e^{s^2\lambda^2/2}\,, \qquad \forall |s| \le \frac{1}{\lambda}\,.$$

A simple and useful example of of a sub-exponential random variable is given in the next lemma.

Lemma 1.12. *Let $X \sim \mathsf{subG}(\sigma^2)$ then the random variable $Z = X^2 - \mathbb{E}[X^2]$ is sub-exponential: $Z \sim \mathsf{subE}(16\sigma^2)$.*

Proof. We have, by the dominated convergence theorem,

$$\begin{aligned}
\mathbb{E}[e^{sZ}] &= 1 + \sum_{k=2}^{\infty} \frac{s^k \mathbb{E}\big[X^2 - \mathbb{E}[X^2]\big]^k}{k!} \\
&\le 1 + \sum_{k=2}^{\infty} \frac{s^k 2^{k-1}\big(\mathbb{E}[X^{2k}] + (\mathbb{E}[X^2])^k\big)}{k!} && \text{(Jensen)} \\
&\le 1 + \sum_{k=2}^{\infty} \frac{s^k 4^k \mathbb{E}[X^{2k}]}{2(k!)} && \text{(Jensen again)} \\
&\le 1 + \sum_{k=2}^{\infty} \frac{s^k 4^k 2(2\sigma^2)^k k!}{2(k!)} && \text{(Lemma 1.4)} \\
&= 1 + (8s\sigma^2)^2 \sum_{k=0}^{\infty} (8s\sigma^2)^k \\
&= 1 + 128 s^2 \sigma^4 && \text{for} \quad |s| \le \frac{1}{16\sigma^2} \\
&\le e^{128 s^2 \sigma^4}\,.
\end{aligned}$$

□

Sub-exponential random variables also give rise to exponential deviation inequalities such as Corollary 1.7 (Chernoff bound) or Theorem 1.9 (Hoeffding's inequality) for weighted sums of independent sub-exponential random variables. The significant difference here is that the larger deviations are controlled in by a weaker bound.

Berstein's inequality

Theorem 1.13 (Bernstein's inequality)**.** *Let $X_1, \dots, X_n$ be independent random variables such that $\mathbb{E}(X_i) = 0$ and $X_i \sim \mathsf{subE}(\lambda)$. Define*

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i\,,$$

Then for any $t > 0$ we have

$$\mathbb{P}(\bar{X} > t) \vee \mathbb{P}(\bar{X} < -t) \le \exp\Big[-\frac{n}{2}\big(\frac{t^2}{\lambda^2} \wedge \frac{t}{\lambda}\big)\Big]\,.$$

Proof. Without loss of generality, assume that $\lambda = 1$ (we can always replace X_i by X_i/λ and t by t/λ. Next, using a Chernoff bound, we get for any $s > 0$

$$\mathbb{P}(\bar{X} > t) \le \prod_{i=1}^{n} \mathbb{E}\big[e^{sX_i}\big] e^{-snt}\,.$$

Next, if $|s| \leq 1$, then $\mathbb{E}\big[e^{sX_i}\big] \leq e^{s^2/2}$ by definition of sub-exponential distributions. It yields

$$\mathbb{P}(\bar{X} > t) \leq e^{\frac{ns^2}{2} - snt}$$

Choosing $s = 1 \wedge t$ yields

$$\mathbb{P}(\bar{X} > t) \leq e^{-\frac{n}{2}(t^2 \wedge t)}$$

We obtain the same bound for $\mathbb{P}(\bar{X} < -t)$ which concludes the proof. □

Note that usually, Bernstein's inequality refers to a slightly more precise result that is qualitatively the same as the one above: it exhibits a Gaussian tail $e^{-nt^2/(2\lambda^2)}$ and an exponential tail $e^{-nt/(2\lambda)}$. See for example Theorem 2.10 in [BLM13].

1.4 MAXIMAL INEQUALITIES

The exponential inequalities of the previous section are valid for linear combinations of independent random variables, and in particular, for the average $\bar{X}$. In many instances, we will be interested in controlling the *maximum* over the parameters of such linear combinations (this is because of empirical risk minimization). The purpose of this section is to present such results.

Maximum over a finite set

We begin by the simplest case possible: the maximum over a finite set.

Theorem 1.14. *Let $X_1, \ldots, X_N$ be N random variables such that $X_i \sim \mathsf{subG}(\sigma^2)$. Then*

$$\mathbb{E}[\max_{1 \leq i \leq N} X_i] \leq \sigma\sqrt{2\log(N)}\,, \qquad \textit{and} \qquad \mathbb{E}[\max_{1 \leq i \leq N} |X_i|] \leq \sigma\sqrt{2\log(2N)}$$

Moreover, for any $t > 0$,

$$\mathbb{P}\big(\max_{1 \leq i \leq N} X_i > t\big) \leq N e^{-\frac{t^2}{2\sigma^2}}\,, \qquad \textit{and} \qquad \mathbb{P}\big(\max_{1 \leq i \leq N} |X_i| > t\big) \leq 2N e^{-\frac{t^2}{2\sigma^2}}$$

Note that the random variables in this theorem need not be independent.

Proof. For any $s > 0$,

$$\begin{aligned}
\mathbb{E}[\max_{1\le i\le N} X_i] &= \frac{1}{s}\mathbb{E}\big[\log e^{s\max_{1\le i\le N} X_i}\big] \\
&\le \frac{1}{s}\log \mathbb{E}\big[e^{s\max_{1\le i\le N} X_i}\big] && \text{(by Jensen)} \\
&= \frac{1}{s}\log \mathbb{E}\big[\max_{1\le i\le N} e^{sX_i}\big] \\
&\le \frac{1}{s}\log \sum_{1\le i\le N} \mathbb{E}\big[e^{sX_i}\big] \\
&\le \frac{1}{s}\log \sum_{1\le i\le N} e^{\frac{\sigma^2 s^2}{2}} \\
&= \frac{\log N}{s} + \frac{\sigma^2 s}{2}
\end{aligned}$$

Taking $s = \sqrt{2(\log N)/\sigma^2}$ yields the first inequality in expectation.

The first inequality in probability is obtained by a simple union bound:

$$\begin{aligned}
\mathbb{P}\big(\max_{1\le i\le N} X_i > t\big) &= \mathbb{P}\Big(\bigcup_{1\le i\le N} \{X_i > t\}\Big) \\
&\le \sum_{1\le i\le N} \mathbb{P}(X_i > t) \\
&\le Ne^{-\frac{t^2}{2\sigma^2}},
\end{aligned}$$

where we used Lemma 1.3 in the last inequality.

The remaining two inequalities follow trivially by noting that

$$\max_{1\le i\le N} |X_i| = \max_{1\le i\le 2N} X_i\,,$$

where $X_{N+i} = -X_i$ for $i = 1, \dots, N$. □

Extending these results to a maximum over an infinite set may be impossible. For example, if one is given an infinite sequence of i.i.d $\mathcal{N}(0, \sigma^2)$ random variables $X_1, X_2, \dots,$, then for any $N \ge 1$, we have for any $t > 0$,

$$\mathbb{P}(\max_{1\le i\le N} X_i < t) = [\mathbb{P}(X_1 < t)]^N \to 0\,, \quad N \to \infty\,.$$

On the opposite side of the picture, if all the X_is are equal to the same random variable X, we have for any $t > 0$,

$$\mathbb{P}(\max_{1\le i\le N} X_i < t) = \mathbb{P}(X_1 < t) > 0 \quad \forall\, N \ge 1\,.$$

In the Gaussian case, lower bounds are also available. They illustrate the effect of the correlation between the X_is

Examples from statistics have structure and we encounter many examples where a maximum of random variables over an infinite set is in fact finite. This is due to the fact that the random variable that we are considering are not independent from each other. In the rest of this section, we review some of these examples.

Maximum over a convex polytope

We use the definition of a polytope from [Gru03]: a convex polytope P is a compact set with a finite number of vertices $\mathcal{V}(\mathsf{P})$ called extreme points. It satisfies $\mathsf{P} = \mathrm{conv}(\mathcal{V}(\mathsf{P}))$, where $\mathrm{conv}(\mathcal{V}(\mathsf{P}))$ denotes the convex hull of the vertices of P.

Let $X \in \mathbb{R}^d$ be a random vector and consider the (infinite) family of random variables

$$\mathcal{F} = \{\theta^\top X \, : \, \theta \in \mathsf{P}\},$$

where $\mathsf{P} \subset \mathbb{R}^d$ is a polytope with N vertices. While the family $\mathcal{F}$ is infinite, the maximum over $\mathcal{F}$ can be reduced to the a finite maximum using the following useful lemma.

Lemma 1.15. *Consider a linear form $x \mapsto c^\top x$, $x, c \in \mathbb{R}^d$. Then for any convex polytope $\mathsf{P} \subset \mathbb{R}^d$,*

$$\max_{x \in \mathsf{P}} c^\top x = \max_{x \in \mathcal{V}(\mathsf{P})} c^\top x$$

where $\mathcal{V}(\mathsf{P})$ denotes the set of vertices of P.

Proof. Assume that $\mathcal{V}(\mathsf{P}) = \{v_1, \dots, v_N\}$. For any $x \in \mathsf{P} = \mathrm{conv}(\mathcal{V}(\mathsf{P}))$, there exist nonnegative numbers $\lambda_1, \dots \lambda_N$ that sum up to 1 and such that $x = \lambda_1 v_1 + \cdots + \lambda_N v_N$. Thus

$$c^\top x = c^\top \Big(\sum_{i=1}^N \lambda_i v_i\Big) = \sum_{i=1}^N \lambda_i c^\top v_i \le \sum_{i=1}^N \lambda_i \max_{x \in \mathcal{V}(\mathsf{P})} c^\top x = \max_{x \in \mathcal{V}(\mathsf{P})} c^\top x \, .$$

It yields

$$\max_{x \in \mathsf{P}} c^\top x \le \max_{x \in \mathcal{V}(\mathsf{P})} c^\top x \le \max_{x \in \mathsf{P}} c^\top x$$

so the two quantities are equal. □

It immediately yields the following theorem

Theorem 1.16. *Let P be a polytope with N vertices $v^{(1)}, \dots, v^{(N)} \in \mathbb{R}^d$ and let $X \in \mathbb{R}^d$ be a random vector such that, $[v^{(i)}]^\top X, i = 1, \dots, N$ are sub-Gaussian random variables with variance proxy σ^2. Then*

$$\mathbb{E}[\max_{\theta \in \mathsf{P}} \theta^\top X] \le \sigma\sqrt{2\log(N)}\,, \qquad \textit{and} \qquad \mathbb{E}[\max_{\theta \in \mathsf{P}} |\theta^\top X|] \le \sigma\sqrt{2\log(2N)}\,.$$

Moreover, for any $t > 0$,

$$\mathbb{P}\big(\max_{\theta \in \mathsf{P}} \theta^\top X > t\big) \le N e^{-\frac{t^2}{2\sigma^2}}\,, \qquad \textit{and} \qquad \mathbb{P}\big(\max_{\theta \in \mathsf{P}} |\theta^\top X| > t\big) \le 2N e^{-\frac{t^2}{2\sigma^2}}$$

Of particular interests are polytopes that have a small number of vertices. A primary example is the ℓ_1 ball of $\mathbb{R}^d$ defined for any radius $R > 0$, by

$$\mathcal{B}_1 = \Big\{x \in \mathbb{R}^d \,:\, \sum_{i=1}^d |x_i| \le 1\}\,.$$

Indeed, it has exactly $2d$ vertices.

Maximum over the ℓ_2 ball

Recall that the unit ℓ_2 ball of $\mathbb{R}^d$ is defined by the set of vectors u that have Euclidean norm $|u|_2$ at most 1. Formally, it is defined by

$$\mathcal{B}_2 = \Big\{x \in \mathbb{R}^d \,:\, \sum_{i=1}^d x_i^2 \le 1\Big\}\,.$$

Clearly, this ball is not a polytope and yet, we can control the maximum of random variables indexed by $\mathcal{B}_2$. This is due to the fact that there exists a finite subset of $\mathcal{B}_2$ such that the maximum over this finite set is of the same order as the maximum over the entire ball.

Definition 1.17. Fix $K \subset \mathbb{R}^d$ and $\varepsilon > 0$. A set $\mathcal{N}$ is called an ε-net of K with respect to a distance $d(\cdot,\cdot)$ on $\mathbb{R}^d$, if $\mathcal{N} \subset K$ and for any $z \in K$, there exists $x \in \mathcal{N}$ such that $d(x, z) \le \varepsilon$.

Therefore, if $\mathcal{N}$ is an ε-net of K with respect to norm $\|\cdot\|$, then every point of K is at distance at most ε from a point in $\mathcal{N}$. Clearly, every compact set admits a finite ε-net. The following lemma gives an upper bound on the size of the smallest ε-net of $\mathcal{B}_2$.

Lemma 1.18. *Fix $\varepsilon \in (0, 1)$. Then the unit Euclidean ball $\mathcal{B}_2$ has an ε-net $\mathcal{N}$ with respect to the Euclidean distance of cardinality $|\mathcal{N}| \le (3/\varepsilon)^d$*

Proof. Consider the following iterative construction if the ε-net. Choose $x_1 = 0$. For any $i \ge 2$, take any x_i to be any $x \in \mathcal{B}_2$ such that $|x - x_j|_2 > \varepsilon$ for all $j < i$. If no such x exists, stop the procedure. Clearly, this will create an ε-net. We now control its size.

Observe that since $|x-y|_2 > \varepsilon$ for all $x, y \in \mathcal{N}$, the Euclidean balls centered at $x \in \mathcal{N}$ and with radius $\varepsilon/2$ are disjoint. Moreover,

$$\bigcup_{z \in \mathcal{N}} \{z + \frac{\varepsilon}{2}\mathcal{B}_2\} \subset (1 + \frac{\varepsilon}{2})\mathcal{B}_2$$

where $\{z + \varepsilon\mathcal{B}_2\} = \{z + \varepsilon x\,, x \in \mathcal{B}_2\}$. Thus, measuring volumes, we get

$$\mathrm{vol}\big((1 + \frac{\varepsilon}{2})\mathcal{B}_2\big) \ge \mathrm{vol}\Big(\bigcup_{z \in \mathcal{N}} \{z + \frac{\varepsilon}{2}\mathcal{B}_2\}\Big) = \sum_{z \in \mathcal{N}} \mathrm{vol}\big(\{z + \frac{\varepsilon}{2}\mathcal{B}_2\}\big)$$

This is equivalent to

$$(1+\frac{\varepsilon}{2})^d \ge |\mathcal{N}|(\frac{\varepsilon}{2})^d .$$

Therefore, we get the following bound

$$|\mathcal{N}| \le \Big(1+\frac{2}{\varepsilon}\Big)^d \le \Big(\frac{3}{\varepsilon}\Big)^d .$$

□

Theorem 1.19. *Let $X \in \mathbb{R}^d$ be a sub-Gaussian random vector with variance proxy σ^2. Then*

$$\mathbb{E}[\max_{\theta\in\mathcal{B}_2} \theta^\top X] = \mathbb{E}[\max_{\theta\in\mathcal{B}_2} |\theta^\top X|] \le 4\sigma\sqrt{d} .$$

Moreover, for any $\delta > 0$, with probability $1-\delta$, it holds

$$\max_{\theta\in\mathcal{B}_2} \theta^\top X = \max_{\theta\in\mathcal{B}_2} |\theta^\top X| \le 4\sigma\sqrt{d} + 2\sigma\sqrt{2\log(1/\delta)} .$$

Proof. Let $\mathcal{N}$ be a $1/2$-net of $\mathcal{B}_2$ with respect to the Euclidean norm that satisfies $|\mathcal{N}| \le 6^d$. Next, observe that for every $\theta \in \mathcal{B}_2$, there exists $z \in \mathcal{N}$ and x such that $|x|_2 \le 1/2$ and $\theta = z + x$. Therefore,

$$\max_{\theta\in\mathcal{B}_2} \theta^\top X \le \max_{z\in\mathcal{N}} z^\top X + \max_{x\in\frac{1}{2}\mathcal{B}_2} x^\top X$$

But

$$\max_{x\in\frac{1}{2}\mathcal{B}_2} x^\top X = \frac{1}{2}\max_{x\in\mathcal{B}_2} x^\top X$$

Therefore, using Theorem 1.14, we get

$$\mathbb{E}[\max_{\theta\in\mathcal{B}_2} \theta^\top X] \le 2\mathbb{E}[\max_{z\in\mathcal{N}} z^\top X] \le 2\sigma\sqrt{2\log(|\mathcal{N}|)} \le 2\sigma\sqrt{2(\log 6)d} \le 4\sigma\sqrt{d} .$$

The bound with high probability, follows because

$$\mathbb{P}\big(\max_{\theta\in\mathcal{B}_2} \theta^\top X > t\big) \le \mathbb{P}\big(2\max_{z\in\mathcal{N}} z^\top X > t\big) \le |\mathcal{N}|e^{-\frac{t^2}{8\sigma^2}} \le 6^d e^{-\frac{t^2}{8\sigma^2}} .$$

To conclude the proof, we find t such that

$$e^{-\frac{t^2}{8\sigma^2}+d\log(6)} \le \delta \ \Leftrightarrow\ t^2 \ge 8\log(6)\sigma^2 d + 8\sigma^2\log(1/\delta) .$$

Therefore, it is sufficient to take $t = \sqrt{8\log(6)}\sigma\sqrt{d} + 2\sigma\sqrt{2\log(1/\delta)}$. □

1.5 PROBLEM SET

Problem 1.1. Let $X_1, \ldots, X_n$ be independent random variables such that $\mathbb{E}(X_i) = 0$ and $X_i \sim \mathsf{subE}(\lambda)$. For any vector $a = (a_1, \ldots, a_n)^\top \in \mathbb{R}^n$, define the weighted sum

$$S(a) = \sum_{i=1}^n a_i X_i \,,$$

Show that for any $t > 0$ we have

$$\mathbb{P}(|S(a)| > t) \leq 2 \exp\left[-C\big(\frac{t^2}{\lambda^2 |a|_2^2} \wedge \frac{t}{\lambda |a|_\infty}\big)\right].$$

for some positive constant C.

Problem 1.2. A random variable X has χ_n^2 (chi-squared with n degrees of freedom) if it has the same distribution as $Z_1^2 + \ldots + Z_n^2$, where $Z_1, \ldots, Z_n$ are iid $\mathcal{N}(0,1)$.

(a) Let $Z \sim \mathcal{N}(0,1)$. Show that the moment generating function of $Y = Z^2 - 1$ satisfies

$$\phi(s) := E\big[e^{sY}\big] = \begin{cases} \dfrac{e^{-s}}{\sqrt{1-2s}} & \text{if } s < 1/2 \\ \infty & \text{otherwise} \end{cases}$$

(b) Show that for all $0 < s < 1/2$,

$$\phi(s) \leq \exp\Big(\frac{s^2}{1-2s}\Big).$$

(c) Conclude that

$$\mathbb{P}(Y > 2t + 2\sqrt{t}) \leq e^{-t}$$

`[Hint: you can use the convexity inequality` $\sqrt{1+u} \leq 1 + u/2$`].`

(d) Show that if $X \sim \chi_n^2$, then, with probability at least $1 - \delta$, it holds

$$X \leq n + 2\sqrt{n \log(1/\delta)} + 2\log(1/\delta)\,.$$

Problem 1.3. Let $X_1, X_2 \ldots$ be an infinite sequence of sub-Gaussian random variables with variance proxy $\sigma_i^2 = C(\log i)^{-1/2}$. Show that for C large enough, we get

$$\mathbb{E}\big[\max_{i \geq 2} X_i\big] < \infty\,.$$

Problem 1.4. Let $A = \{A_{i,j}\}_{\substack{1\le i\le n\\ 1\le j\le m}}$ be a random matrix such that its entries are iid sub-Gaussian random variables with variance proxy σ^2.

(a) Show that the matrix A is sub-Gaussian. What is its variance proxy?

(b) Let $\|A\|$ denote the operator norm of A defined by

$$\max_{x\in \mathbb{R}^m} \frac{|Ax|_2}{|x|_2}.$$

Show that there exits a constant $C > 0$ such that

$$\mathbb{E}\|A\| \le C(\sqrt{m} + \sqrt{n}).$$

Problem 1.5. Recall that for any $q \ge 1$, the ℓ_q norm of a vector $x \in \mathbb{R}^n$ is defined by

$$|x|_q = \Big(\sum_{i=1}^n |x_i|^q\Big)^{\frac{1}{q}}.$$

Let $X = (X_1, \dots, X_n)$ be a vector with independent entries such that X_i is sub-Gaussian with variance proxy σ^2 and $\mathbb{E}(X_i) = 0$.

(a) Show that for any $q \ge 2$, and any $x \in \mathbb{R}^d$,

$$|x|_2 \le |x|_q n^{\frac{1}{2}-\frac{1}{q}},$$

and prove that the above inequality cannot be improved

(b) Show that for for any $q > 1$,

$$\mathbb{E}|X|_q \le 4\sigma n^{\frac{1}{q}}\sqrt{q}$$

(c) Recover from this bound that

$$\mathbb{E}\max_{1\le i\le n} |X_i| \le 4e\sigma\sqrt{\log n}.$$

Problem 1.6. Let K be a compact subset of the unit sphere of $\mathbb{R}^p$ that admits an ε-net $\mathcal{N}_\varepsilon$ with respect to the Euclidean distance of $\mathbb{R}^p$ that satisfies $|\mathcal{N}_\varepsilon| \le (C/\varepsilon)^d$ for all $\varepsilon \in (0,1)$. Here $C \ge 1$ and $d \le p$ are positive constants. Let $X \sim \mathsf{subG}_p(\sigma^2)$ be a centered random vector.

Show that there exists positive constants c_1 and c_2 to be made explicit such that for any $\delta \in (0,1)$, it holds

$$\max_{\theta\in K} \theta^\top X \le c_1\sigma\sqrt{d\log(2p/d)} + c_2\sigma\sqrt{\log(1/\delta)}$$

with probability at least $1-\delta$. Comment on the result in light of Theorem 1.19 .

Problem 1.7. For any $K \subset \mathbb{R}^d$, distance d on $\mathbb{R}^d$ and $\varepsilon > 0$, the ε-covering number $C(\varepsilon)$ of K is the cardinality of the smallest ε-net of K. The ε-packing number $P(\varepsilon)$ of K is the cardinality of the largest set $\mathcal{P} \subset K$ such that $d(z, z') > \varepsilon$ for all $z, z' \in \mathcal{P}$, $z \neq z'$. Show that

$$C(2\varepsilon) \le P(2\varepsilon) \le C(\varepsilon)\,.$$

Problem 1.8. Let $X_1, \dots, X_n$ be n independent and random variables such that $\mathbb{E}[X_i] = \mu$ and $\mathrm{var}(X_i) \le \sigma^2$. Fix $\delta \in (0,1)$ and assume without loss of generality that n can be factored into $n = K \cdot G$ where $G = 8\log(1/\delta)$ is a positive integers.

For $g = 1, \dots, G$, let $\bar{X}_g$ denote the average over the gth group of k variables. Formally

$$\bar{X}_g = \frac{1}{k} \sum_{i=(g-1)k+1}^{gk} X_i\,.$$

1. Show that for any $g = 1, \dots, G$,

$$\mathbb{P}\big[\bar{X}_g - \mu > \frac{2\sigma}{\sqrt{k}}\big] \le \frac{1}{4}\,.$$

2. Let $\hat{\mu}$ be defined as the median of $\{\bar{X}_1, \dots, \bar{X}_G\}$. Show that

$$\mathbb{P}\big[\hat{\mu} - \mu > \frac{2\sigma}{\sqrt{k}}\big] \le \mathbb{P}\big[\mathcal{B} \ge \frac{G}{2}\big]\,,$$

where $\mathcal{B} \sim \mathsf{Bin}(G, 1/4)$.

3. Conclude that

$$\mathbb{P}\big[\hat{\mu} - \mu > 4\sigma\sqrt{\frac{2\log(1/\delta)}{n}}\big] \le \delta$$

4. Compare this result with 1.7 and Lemma 1.3. Can you conclude that $\hat{\mu} - \mu \sim \mathsf{subG}(\bar{\sigma}^2/n)$ for some $\bar{\sigma}^2$? Conclude.