18.656 Notes

Lecturer:

Andrew Liu

Spring 2024

Last updated on Tuesday 27th February, 2024.

Contents

1 February 6, 2024

First day

2 February 8, 2024

2.1 Tail Bounds

Some important tail bounds that we'll use in this class.

Theorem 2.1 (Markov's Inequality)
For nonnegative real-valued r.v. X,

$$\mathbb{P}[X \geq t] \leq \frac{\mathbb{E}[X]}{t}, \quad t > 0.$$

Theorem 2.2 (Chebyshev's Inequality)
For any real-valued r.v. X with mean μ,

$$\mathbb{P}[|X - \mu| \geq t] \leq \frac{\sigma^2}{t^2}.$$

Some useful applications of Markov's inequality:

- higher moments:

$$\mathbb{P}[|X - \mu| \geq t] = \mathbb{P}[|X - \mu|^p \geq t^p] \leq \min_{p \geq 1} \frac{\mathbb{E}[|X - \mu|]^p}{t^p}$$

- exponentiated r.v.s:

$$\mathbb{P}[X - \mu \geq t] = \mathbb{P}[e^{\lambda(X-\mu)} \geq e^{\lambda t}] \leq \inf_{\lambda > 0} e^{-t\lambda} \mathbb{E}[e^{\lambda(X-\mu)}].$$

The second expression shows us that deducing tail bounds for means is intimately related to better understanding **moment generating functions** (MGFs), i.e.,

$$\mathrm{MGF}_X(\lambda) = \mathbb{E}[e^{\lambda X}].$$

2.2 Sub-Gaussian Random Variables

Definition 2.3

A random variable X with mean $\mu = \mathbb{E}[X]$ is **σ-sub-Gaussian** if $\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\lambda^2\sigma^2/2}$ for all $\lambda \in \mathbb{R}$.

We can show that this holds when $X \sim \mathcal{N}(\mu, \sigma^2)$, hence motivating the definition, by directly deriving the MGF for X.

$$\begin{aligned}
\mathrm{MGF}_X(\lambda) = \mathbb{E}[e^{\lambda X}] &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\lambda x} e^{-\frac{1}{2}\left(\frac{(x-\mu)^2}{\sigma^2}\right)} \mathrm{d}x \\
&= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2\sigma^2}\left(x^2 - 2\mu x + \mu^2 - 2\sigma^2\lambda x\right)\right] \mathrm{d}x \\
&= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left[-\frac{1}{2\sigma^2}\left(x - (\mu + \sigma^2\lambda)\right)^2 - 2\mu\sigma^2\lambda - \sigma^4\lambda^2\right] \mathrm{d}x \\
&= e^{\mu\lambda + \sigma^2\lambda^2/2}.
\end{aligned}$$

The key is the quadratic exponential tail decay; this is generally what we use to characterize Gaussian / sub-Gaussian behavior.

Claim 2.4 (Bounded r.v.s are sub-Gaussian)

Given r.v. $X \in [a, b]$, $\mathbb{E}[X] = \mu$. Then, X is sub-Gaussian with $\sigma = (b - a)$.

It turns out that we can also show $\sigma = (b-a)/2$, but the technique used to show the weaker result is more interesting.

Proof. Let $\tilde{X}$ be i.i.d. to X. Then,

$$\begin{aligned}
\mathbb{E}_X[e^{\lambda(X-\mu)}] &= \mathbb{E}_X[e^{\lambda(X - \mathbb{E}_{\tilde{X}}[\tilde{X}])}] \\
&\leq \mathbb{E}_{X,\tilde{X}}[e^{\lambda(X-\tilde{X})}],
\end{aligned}$$

by Jensen's inequality. Since $X, \tilde{X}$ are i.i.d., $(X - \tilde{X})$ has a distribution symmetric around 0. Now, we also have that

$$(X - \tilde{X}) \overset{\text{dist}}{=} \varepsilon(X - \tilde{X}),$$

where $\varepsilon \in \{\pm 1\}$ with equal probability (also called a **Rademacher** random variable).

Therefore,

$$\begin{aligned}\mathbb{E}_{\tilde{X}}[e^{\lambda(X-\mu)}] &\leq \mathbb{E}_{\tilde{X},X}[\mathbb{E}_{\varepsilon}[e^{\lambda\varepsilon(X-\tilde{X})}]] \\ &\leq \mathbb{E}_{X,\tilde{X}}[e^{\lambda^2(X-\tilde{X})^2/2}],\end{aligned}$$

since ε is 1-sub-gaussian. Finally, since X is bounded,

$$\mathbb{E}_{X,\tilde{X}}[e^{\lambda^2(X-\tilde{X})^2/2}] \leq e^{\lambda^2(b-a)^2/2}.$$

□

Definition 2.5 (Addition property of Gaussians)
Given $X_1 \sim \mathcal{N}(0,\sigma_1^2)$ and $X_2 \sim \mathcal{N}(0,\sigma_2^2)$,

$$X_1 + X_2 \sim \mathcal{N}(0,\sigma_1^2+\sigma_2^2).$$

Claim 2.6 (Addition property of sub-Gaussians)
Given $X_i \sim \sigma_i$-sub-Gaussian, $i \in \{1,2\}$, then X_1+X_2 is $\sqrt{\sigma_1^2+\sigma_2^2}$-sub-Gaussian.

Proof.

$$\begin{aligned}\mathbb{E}_{X_1,X_2}[e^{\lambda(X_1+X_2)}] &= \mathbb{E}_{X_1,X_2}[e^{\lambda X_1}e^{\lambda X_2}] \\ &= \mathbb{E}_{X_1}[e^{\lambda X_1}]\mathbb{E}_{X_2}[e^{\lambda X_2}] \\ &\leq e^{\lambda^2\sigma_1^2/2}e^{\lambda^2\sigma_2^2/2} = e^{\lambda^2(\sigma_1^2+\sigma_2^2)/2}.\end{aligned}$$

□

A consequence of this fact is that given $X_i \sim \sigma$-sub-Gaussian, i.i.d. with zero-mean, then

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i \sim \sigma\text{-sub-Gaussian},$$

or equivalently

$$\frac{1}{n}\sum_{i=1}^{n} X_i \sim \frac{\sigma}{\sqrt{n}}\text{-sub-Gaussian}.$$

Example 2.7 (Survey sampling)
Consider a race where two candidates A and B are up for election, and a survey is distributed to the population for their preferred candidate.

Say that we sample $i = 1, \dots, n$ who give responses $X_i = 1$ if A or 0 if B. Let μ^* be the theoretical fraction of people who will vote A. Let $\hat{\mu}$ be our estimator for μ^*:

$$\hat{\mu} = \sum_{i=1}^{n} X_i.$$

During a typical analysis, we would like to construct a confidence interval $\hat{\mathcal{I}}$, and know the probability that the true proportion falls outside of this confidence interval. More formally, given δ and $\hat{\mathcal{I}}$, we would like to know the n at which we could say

$$\mathbb{P}[\hat{\mathcal{I}} \ni \mu^*] \geq 1 - \delta.$$

For example, with $\delta = 0.02$, interval width of 0.03, we would require $n \approx 10000$ to make this guarantee.

We can model $X_i \sim \textsc{Bern}(\mu^*)$. Since $X_i \in [0, 1]$, our earlier result shows that X_i is 1/2-sub-Gaussian. Using additivity and i.i.d., our sample mean $\hat{\mu}$ is $1/(2\sqrt{n})$-sub-Gaussian. Thus,

$$\mathbb{E}[e^{\lambda(\hat{\mu}-\mu^*)}] \leq e^{\lambda^2/2 \cdot 1/(4n)} = e^{\lambda^2/(8n)}.$$

Using Chernoff,

$$\mathbb{P}[|\hat{\mu} - \mu^*| \geq s] \leq 2e^{2ns^2},$$

for some $s > 0$. So, for any fixed δ, if our interval has width $s \geq \sqrt{\log(2/\delta)/(2n)}$, then we could say that the probability that our true mean lies outside of the interval is less than δ.

2.3 Hoeffding

Lemma 2.8 (Hoeffding's Lemma)
For any zero-mean r.v. X with values in $[a, b]$, the MGF satisfies

$$\mathbb{E}[e^{\lambda X}] \leq e^{\lambda^2 (b-a)^2/8}.$$

for all λ. In other words, bounded random variables are sub-Gaussian with parameter $\sigma = (b-a)/2$.

We'll present two proofs. The first proof is taken from these lecture notes. The second proof is a more slick argument.

Proof. Since e^{sx} is convex,

$$e^{sX} \leq \frac{b-X}{b-a} e^{sa} + \frac{X-a}{b-a} e^{sb},$$

so

$$\mathbb{E}[e^{sX}] \leq \frac{b-\mathbb{E}[X]}{b-a} e^{sa} + \frac{\mathbb{E}[X]-a}{b-a} e^{sb} = \frac{b}{b-a} e^{sa} - \frac{a}{b-a} e^{sb}.$$

Make the substitution $p = -a/(b-a)$ so that the above expression simplifies to

$$(1-p+pe^{s(b-a)})e^{-sp(b-a)},$$

and again substitute $u = s(b-a)$ so that it further simplifies to

$$\varphi(u) := (1-p+pe^{u})e^{pu}.$$

Now we can bound $\varphi(u)$. Taking derivatives,

$$\varphi'(u) := -p + \frac{pe^u}{1-p+pe^u}$$
$$\varphi''(u) := \frac{p(1-p)e^u}{(1-p+pe^u)^2}.$$

By Taylor's theorem (see here), we have for some $z \in [0, u]$

$$\varphi(u) = \phi(0) + u\phi'(0) + \frac{1}{2}u^2\phi''(z) \leq \phi(0) + u\phi'(0) + \sup_z \frac{1}{2}u^2\phi''(z),$$

so substituting in the expressions from above

$$\varphi(u) \le \sup_z \frac{1}{2} u^2 \varphi''(z).$$

Bashing critical points eventually gives the upper bound 1/4, from which we get

$$\mathbb{E}[e^{sX}] \le e^{\varphi(u)} \le e^{u^2/8} \le e^{s^2(b-a)^2/8},$$

as desired. □

Now the cleaner argument:

Proof. Let $\psi(\lambda) = \log \mathbb{E}[e^{\lambda X}]$. We have

$$\mathbb{E}[e^{\lambda X}] = \mathbb{E}[e^{\psi(\lambda)}],$$

so it suffices to bound $\psi(\lambda)$. Similar to the previous proof, we use Taylor's theorem:

$$\psi(\lambda) \le 1 + \lambda \psi'(0) + \sup_{z \in (0,\lambda)} \frac{\lambda^2}{2} \psi''(z).$$

We have

$$\psi'(\lambda) = \frac{\mathbb{E}[Xe^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} \quad \text{and} \quad \psi''(\lambda) = \frac{\mathbb{E}[X^2 e^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} - \left(\frac{\mathbb{E}[Xe^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} \right)^2.$$

Notice that if we sample from a new distribution with density

$$f_Y(y) = \frac{e^{\lambda y} f_X(y)}{\mathbb{E}[e^{\lambda X}]},$$

then $\psi''(\lambda) = \mathrm{Var}[Y]$. Since variance is maximized when points are clustered at the endpoints, this is bounded by $(b-a)^2/4$, which gives

$$\psi(\lambda) \le \lambda^2 \frac{(b-a)^2}{8},$$

finishing the proof. □

Theorem 2.9 (Hoeffding's Inequality)
Let $X_1, \ldots, X_n$ be r.v. with $\mathbb{E}[X_i] = \mu_i$, $a_i \leq X_i \leq b_i$, and independent. Then,

$$\mathbb{P}\left[\sum_{i=1}^{n}(X_i - \mu_i) \geq t\right] \leq e^{-2t^2/(\sum_i (b_i - a_i)^2)}.$$

Proof. Define $Z_i = X_i - \mathbb{E}[X_i]$, so that $\mathbb{E}[Z_i] = 0$. By Chernoff, for any $s > 0$, we have

$$\mathbb{P}[\sum_i Z_i \geq t] = \mathbb{P}[\exp\left(s\sum_i Z_i\right) \geq e^{st}] \leq \frac{\mathbb{E}[\prod_{i=1}^{m} e^{sZ_i}]}{e^{st}}.$$

Since Z_i are independent, we can move the expectation inside of the product. Applying the Hoeffding Lemma then gives

$$\frac{\mathbb{E}[\prod_i e^{sZ_i}]}{e^{st}} = \frac{\prod_i \mathbb{E}[e^{sZ_i}]}{e^{st}} \leq \exp\left(-st + \frac{s^2}{8}\sum_i (b_i - a_i)^2.\right)$$

Substituting carefully chosen $s = \frac{4t}{\sum_i (b_i - a_i)^2}$ gives the bound that we want. □

This statement is a bit specialized. A more general statement in terms of sub-Gaussian coefficients is as follows:

Theorem 2.10 (Hoeffding's Inequality)
Let $X_1, \ldots, X_n$ be r.v. with $\mathbb{E}[X_i] = \mu_i$, and X_i sub-Gaussian with parameter σ_i. Then,

$$\mathbb{P}\left[\sum_{i=1}^{n}(X_i - \mu_i) \geq t\right] \leq e^{-t^2/(2\sum_i \sigma_i^2)}.$$

We already showed that bounded r.v.s have parameters $\sigma = (b-a)/2$, so we can plug this into the above to arrive at the more specific version of the inequality.

Example 2.11
$X \sim \mathcal{N}(0,1)$ is 1-sub-Gaussian. Let $Y = X^2$ is not sub-Gaussian.

First compute the MGF:

$$M_X(\lambda) = \mathbb{E}[e^{\lambda X^2}] = \frac{1}{\sqrt{2\pi}} \int e^{\lambda z^2} e^{-(z-1)^2/2} \mathrm{d}z = \frac{1}{\sqrt{1-\lambda}}$$

for $\lambda \in (0,1)$. Despite not being sub-Gaussian, it is close, which we will show in more detail next time.

Example 2.12

Consider an n-dimensional Gaussian, $X = (X_1, \ldots, X_n)$, where each $X_i \sim \mathcal{N}(0,1)$.

Then $\mathbb{E}[\|X\|_2^2/n] = \mathbb{E}[\sum X_i^2/n] = 1$, it turns out that

$$\mathbb{P}\left[\left|\frac{\|X\|^2}{n^2} - 1\right| \geq \delta\right] \leq 2e^{-cn\delta^2}$$

holds for all $\delta \in (0,1)$. This looks very similar to the sub-gaussian tail bound from earlier, but will only hold for small delta. For larger delta, the tail bound becomes linear in δ, i.e., exponential. This is example is just to provide some intuition on tail bounds. We will show this in more detail in the next lecture.

3 February 15, 2024

Class was cancelled on Tuesday due to snow.

3.1 Exponential Random Variables

Here we discuss another interesting and useful class of random variables. Consider the chi-square random variable χ^2, which follows distribution $X = Z^2$ where $Z \sim \mathcal{N}(0,1)$. Note that $\mathbb{E}[X] = \mathbb{E}[Z^2] = \mathrm{Var}[Z] = 1$, so its zero-mean moment generating function is given by

$$M_X(\lambda) = \mathbb{E}[e^{\lambda(Z^2-1)}] = \frac{1}{\sqrt{2\pi}} \int e^{\lambda(z^2-1)} e^{-z^2/2} \mathrm{d}z.$$

Force $\lambda < 1/2$ so that this is finite. Then,

$$M_X(\lambda) = \frac{1}{e^{\lambda}\sqrt{1-2\lambda}}.$$

over the domain $0 < \lambda < 1/2$. We can show that this is sub-Gaussian for bounded $|\lambda|$ in the following way. Take Taylor expansions:

$$e^{-\lambda} = 1 - \lambda + \frac{\lambda^2}{2} + o(\lambda^2)$$
$$\frac{1}{\sqrt{1-2\lambda}} = 1 + \lambda + \frac{3}{2}\lambda^2 + o(\lambda^2),$$

so

$$\frac{e^{-\lambda}}{\sqrt{1-2\lambda}} = (1 - \lambda + \frac{\lambda^2}{2} + o(\lambda^2))(1 + \lambda + \frac{3}{2}\lambda^2 + o(\lambda^2)) = 1 + \lambda^2 + o(\lambda^2).$$

The sub-Gaussian bound is

$$e^{\lambda^2\sigma^2/2} = 1 + \frac{\lambda^2\sigma^2}{2} + o(\lambda^2),$$

so we can choose a suitable σ to have some region around 0 where the MGF is sub-Gaussian. By graphing we can see that $\sigma = 2$ works for all $|\lambda| < 1/4$, so in this range Z^2 is sub-Gaussian.

Definition 3.1

Random variable X with mean $\mu = \mathbb{E}[X]$ is **sub-exponential** if there are non-negative parameters (s^2, α) such that

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{s^2\lambda^2/2}$$

for all $|\lambda| < 1/\alpha$.

The naming is motivated by the linear exponential tail bound outside of this special region.

Lemma 3.2

Given r.v. X is (ν^2, α)-sub-exponential,

$$\mathbb{P}[X - \mu \geq t] \leq \begin{cases} e^{-t^2/(2\nu^2)}, & 0 \leq t \leq \nu^2/\alpha \\ e^{-t/(2\alpha)}, & t \geq \nu^2/\alpha. \end{cases}$$

This is equivalent to writing

$$\mathbb{P}[X - \mu \geq t] \leq \exp\left(-\min\left\{\frac{t^2}{2\nu^2}, \frac{t}{2\alpha}\right\}\right).$$

Proof. By Markov, for some $\lambda \in [0, 1/\alpha)$,

$$\mathbb{P}[X - \mu \geq t] = \mathbb{P}[e^{\lambda(X-\mu)} \leq e^{\lambda t}] \leq \inf_{\lambda \in [0,1/\alpha)} e^{-t\lambda} \mathbb{E} e^{\lambda(X-\mu)} \leq \inf_{\lambda \in [0,1/\alpha)} e^{-t\lambda} e^{\nu^2\lambda^2/2}.$$

By taking a derivative, the critical point lands at $\lambda = t/\nu^2$. If this is between $t/\nu^2 \in [0, 1/\alpha)$, then our bound is

$$e^{\nu^2/2 \cdot t^2/\nu^4 - t^2/\nu^2} = e^{-t^2/(2\nu^2)},$$

which is the first case. Else, the critical point is larger than $1/\alpha$, so we take the endpoint $\lambda = 1/\alpha$ and get the bound

$$e^{\nu^2/2 \cdot 1/\alpha^2 - t/\alpha} \leq e^{-t/(2\alpha)},$$

where this last inequality uses $\nu^2 \leq \alpha t$. □

3.2 Randomized dimension reduction

Example 3.3

Consider a dataset of N points $\{u^1, \ldots, u^N\}$ where $u^j \in \mathbb{R}^d$.

Storing the entire dataset gets very expensive very quickly when N, d are large. Is there a lower-dimensional representation of this dataset that is still useful? We

would like to accomplish this while preserving pairwise distances:

$$\|u_i - u_j\|_2^2 \quad \forall i \neq j$$

This is useful for estimating clustering algorithms, densities (computing neighborhoods of points), and more. A more formal representation of this problem:

$$\mathcal{F} : \mathbb{R}^d \to \mathbb{R}^m$$

We call m the **sketch dimension**, or the **embedding dimension**. The goal is for us to find a "useful" representation where $m \ll d$, which, using the distance metric as our notion of usefulness, we can bound w.r.t. a new parameter ε:

$$1 - \varepsilon \leq \frac{\|\mathcal{F}(u_i) - \mathcal{F}(u_j)\|_2^2}{\|u_i - u_j\|_2^2} \leq 1 + \varepsilon.$$

We'll solve this from the perspective of a fixed ϵ, so that our goal is to minimize m while preserving some notion of distance. We'll also introduce another parameter δ, so that we want

$$\frac{\|\mathcal{F}(u_i) - \mathcal{F}(u_j)\|_2^2}{\|u_i - u_j\|_2^2} \in [1 - \varepsilon, 1 + \varepsilon]$$

w.p. $1 - \delta$.

3.3 Bernstein's condition

We begin with some motivation for Bernstein's condition:

Example 3.4 (Motivating Bernstein's condition)
Given $X_i \sim \textsc{Bern}(p)$ for $i \in \{1, \ldots, n\}$. Say we have $\mathbb{E}[X_i] = p$, $\text{Var}[X_i] = p(1-p)$, and $|X_i| \leq b := 1$.

Then, it turns out that

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_i X_i - p\right| \geq t\right) \leq 2e^{-nt^2/(2p(1-p)+2t)}.$$

The behavior of this bound is interesting. When t is small, we get a sub-Gaussian tail bound. If t is large, then it converges to $e^{-nt/2}$, which is only sub-Exponential.

Definition 3.5 (Bernstein's condition)
Given random variable with parameters $\mu = \mathbb{E}[X]$, $\sigma^2 = \mathrm{Var}[X]$, we say that it satisfies Bernstein's condition with parameter b if

$$|\mathbb{E}[(X-\mu)^k]| \le \frac{1}{2} k! \sigma^2 b^{k-2}$$

for $k = 2, 3, \ldots$.

Example 3.6
Bounded random variables satisfy the Bernstein condition. Let's say $|X - \mu| \le B$. We can show that this satisfies the condition with $b = B$ in a pretty strong sense; we won't need the extra factor of $k!$ on the RHS.

Proof.

$$\begin{aligned} |\mathbb{E}[(X-\mu)^k]| &\le \mathbb{E}|X-\mu|^2 |X-\mu|^{k-2} \\ &\le \sigma^2 B^{k-2} \ll \frac{1}{2} k! \sigma^2 B^{k-2}. \end{aligned}$$

□

Note that this works with $b = B/3$ as well, because the bound is so loose.
Next, we show that the Bernstein condition also implies a nice bound on MGFs.

Lemma 3.7 (Bernstein's Inequality)
For r.v. X satisfying the Bernstein condition with parameter $b > 0$,

$$\mathbb{E}[e^{\lambda(X-\mu)}] \le e^{(\lambda^2\sigma^2/2)/(1-b|\lambda|)},$$

for all $|\lambda| b < 1$.

Proof.

$$\begin{aligned}\mathbb{E}[e^{\lambda(X-\mu)}] &= 1 + \lambda\mathbb{E}[X-\mu] + \frac{\lambda^2}{2} + \sum_{k\geq 3}\frac{\lambda^k\mathbb{E}[(X-\mu)^k]}{k!}\\ &\leq 1 + \frac{\lambda^2}{2}\sigma^2 + \sum_{k\geq 3}\frac{|\lambda|^k k!/2\cdot\sigma^2 b^{k-2}}{k!}\\ &\leq 1 + \frac{\lambda^2}{2}\sigma^2 + \frac{\lambda^2}{2}\sigma^2\sum_{k\geq 3}|\lambda|^{k-2}b^{k-2}\\ &\leq 1 + \frac{\lambda^2}{2}\sigma^2\left(\sum_{k\geq 0}|\lambda|^k b^k\right)\\ &= 1 + \frac{\lambda^2\sigma^2/2}{1-b|\lambda|},\end{aligned}$$

as long as $|\lambda|b < 1$. Now we are done by the fact that $1 + a \leq e^a$. □

This shows that **random variables satisfying the Bernstein condition are (σ^2, b)-sub-Exponential**. Using the bound we established earlier on sub-Exponential random variables, we can choose $\lambda = t/(bt+\sigma^2) \in [0, 1/b)$ to establish tail bound

$$\mathbb{P}[|X-\mu| \geq t] \leq 2e^{-(t^2/2)/(\sigma^2+tb)}.$$

Let's see what happens when we try to use the strictest bound possible.

Theorem 3.8

Suppose it holds for some v, b that

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{(\lambda^2 v^2/2)/(1-b\lambda)}$$

for all $\lambda \in (0, 1/b)$. Then,

$$\mathbb{P}(X-\mu \geq t) \leq \frac{v^2}{b^2}\left(\sqrt{1+\frac{2bt}{v^2}} - \frac{bt}{v^2} - 1\right),$$

and equivalently

$$\mathbb{P}[X-\mu \geq \sqrt{2v^2 t} + bt] \leq e^{-t}.$$

Proof. By Markov,

$$\mathbb{E}[X-\mu \geq t] = \mathbb{E}[e^{\lambda(X-\mu)} \leq e^{\lambda t}] \leq \inf_{\lambda \in (0,1/b)} e^{-\lambda t}\mathbb{E}[e^{\lambda(X-\mu)}] \leq \inf_{\lambda \in (0,1/b)} e^{-\lambda t} e^{(\lambda^2 v^2/2)/(1-b\lambda)}.$$

So we'll have to optimize

$$-\lambda t + \frac{\lambda^2 v^2}{2(1-b\lambda)}.$$

Taking the derivative gives

$$-t + \frac{\lambda v^2 - \lambda^2 v^2 b/2}{(1-b\lambda)^2},$$

so we want

$$\lambda^2\left(-\frac{v^2 b}{2} - b^2 t\right) + \lambda(v^2 + 2bt) - t = 0.$$

Here, we can make the convenient substitution $s = v^2 + 2bt$, whence

$$\lambda^2\left(-\frac{sb}{2}\right) + \lambda s - t = 0,$$

and solving the quadratic gives

$$\lambda = \frac{-s + \sqrt{s}\sqrt{s-2bt}}{-sb} = \frac{1}{b}(1-p),$$

for $p := v/\sqrt{s} = \sqrt{v^2/(v^2+2bt)}$. Now, plugging back into the original expression gives

$$\frac{1/b^2 \cdot (1-p)^2 \cdot v^2/2}{1 - b\cdot(1/b)\cdot(1-p)} - \frac{t}{b}(1-p) = \frac{v^2}{b^2}(1-p)\left(\frac{(1-p)}{2p} - \frac{bt}{v^2}\right).$$

We make one final substitution $x := bt/v^2$, so that $p = \sqrt{1/(1+2x)}$, which gives

$$\frac{v^2}{b^2}\left(\frac{\sqrt{1+2x}-1}{\sqrt{1+2x}}\right)\left(\frac{\sqrt{1+2x}-1-2x}{2}\right) = \frac{v^2}{b^2}(\sqrt{1+2x}-1-x).$$

This is the same as the first expression that we are trying to prove, so we're done with the first part.

To show that the second part is the same, we have to invert this expression. We have

$$\sqrt{1+2x} - x - 1 = -\frac{b^2}{v^2}t,$$

where simplifying a bit gives

$$x^2 + x \cdot \frac{-2b^2}{v^2}t + \left(-\frac{b^2}{v^2}t + 1\right)^2 - 1 = 0$$

and thus

$$\left(x - \frac{b^2}{v^2}t\right)^2 = \frac{2b^2}{v^2}t \implies x = \frac{b^2}{v^2}t + \sqrt{\frac{2b^2}{v^2}t}.$$

Finally, since $x = bt^{-1}/v^2$, we have

$$t^{-1} = bt + \sqrt{2v^2 t},$$

as desired. □

3.4 Johnson-Lindenstrauss Lemma

We have the tools we need to tackle the problem of dimensionality reduction. First consider the problem we introduced last lecture:

Example 3.9
Consider an n-dimensional Gaussian, $X = (X_1, \ldots, X_n)$ where each $X_i \sim \mathcal{N}(0,1)$.

We showed earlier that $\|X_i\|^2$ is $(2^2, 4)$-sub-Exponential, hence $\|X\|$ is $(4n, 4)$-sub-Exponential. Therefore,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\|X_i^2\| - 1\right| \geq t\right) = \mathbb{P}[|\|X\| - n| \geq nt] \leq 2e^{-nt^2/8}$$

for $t \in (0,1)$.

In the dimensionality reduction setup, we had map $\mathcal{F} : \mathbb{R}^d \to \mathbb{R}^m$. We analyze the specific mapping

$$\mathcal{F}(u) = \frac{Fu}{\sqrt{m}}$$

where $F \in \mathbb{R}^{m \times d}$ and $F_{i,j} \sim \mathcal{N}(0,1)$. Consider some $v \in \mathbb{R}^d$ with $\|v\| = 1$. Then,

$\langle F_i, v\rangle \sim \mathcal{N}(0,1)$, so $\sum_i \langle F_i, v\rangle = \|Fv\|^2 \sim \chi^2_m$. Thus, by our above result,

$$\mathbb{P}\left[\left|\frac{1}{m}\|Fv\|^2 - 1\right| \geq t\right] \leq 2e^{-mt^2/8}.$$

Thus, for arbitrary $u \in \mathbb{R}^d$,

$$\mathbb{P}\left[\frac{\|Fu\|^2}{\|u\|^2} \notin [1-t, 1+t]\right] \leq 2e^{-mt^2/8}.$$

The same logic holds for pairwise distances, i.e.,

$$\mathbb{P}\left[\frac{\|Fu_i - Fu_j\|^2}{\|u_i - u_j\|^2} \notin [1-t, 1+t]\right] \leq 2e^{-mt^2/8},$$

so union bounding gives

$$\mathbb{P}\left[\frac{\|Fu_i - Fu_j\|^2}{\|u_i - u_j\|^2} \notin [1-t, 1+t]\right] \leq 2e^{-mt^2/8}\binom{N}{2}$$

over all pairs of distinct $i, j \in [N]$. Thus, as long as

$$m > \frac{8}{t^2}\ln\frac{N(N-1)}{\delta},$$

there is a guarantee of no pairwise distances escaping ratio $[1-t, 1+t]$ w.p. $(1-\delta)$.

4 February 27, 2024

Outline for today:

- Martingales and Azuma-Hoeffding
- Some examples
- Rademacher and Gaussian complexities
- Gauss-Lipschitz functions

4.1 Recap

Let

$$D_i = \mathbb{E}[f(X_1)^n | X_1, \dots, X_i] - \mathbb{E}[f(X_1^n) | X_1, \dots, X_{i-1}].$$

By telescoping,

$$f(X_1^n) - \mathbb{E}[f(X_1^n)] = \sum_{i=1}^{n} D_i,$$

and the Martingale difference property says that

$$\mathbb{E}[D_i | X_1, \dots, X_{i-1}] = 0.$$

4.2 Idk

Claim 4.1

Say that we have

$$\mathbb{E}[e^{\lambda D_i} | X_1, \dots, X_{i-1}] \le e^{\lambda^2 \nu_i^2/2} \quad \forall |\lambda| < \frac{1}{\alpha_i}, i \in [n].$$

Then, we have

$$\mathbb{E}[e^{\lambda \sum_{i=1}^n D_i}] \le e^{\lambda^2/2 \cdot (\sum_{i=1}^n \nu_i^2)} \quad \forall |\lambda| \le \frac{1}{\max_{i \in [n]} \alpha_i}.$$

Proof.

$$\begin{aligned}
\mathbb{E}[e^{\lambda \sum_{i=1}^n D_i}] &= \mathbb{E}[e^{\lambda \sum_{i=1}^{n-1} D_i} e^{\lambda D_n}] \\
&= \mathbb{E}_{X_{[1 \dots n-1]}}[\mathbb{E}_{X_n}[e^{\lambda \sum_{i=1}^{n-1} D_i} e^{\lambda D_n} | X_1, \dots, X_{n-1}]] \\
&= \mathbb{E}_{X_{[1 \dots n-1]}}[e^{\lambda \sum_{i=1}^{n-1} D_i} \mathbb{E}_{X_n}[e^{\lambda D_n} | X_1, \dots, X_{n-1}]].
\end{aligned}$$

By assumption, this quantity is bounded above by

$$\mathbb{E}_{X_1^{n-1}}[e^{\lambda \sum_{i=1}^{n-1}}] e^{\lambda^2 \nu_n^2/2},$$

for all $|\lambda| \le \frac{1}{\alpha_n}$. If we keep applying this inductively, we have the result. □

Theorem 4.2 (Azuma-Hoeffding Bound)
Consider a special case, when $D_i \in [a_i, b_i]$ almost surely. Then,

$$\mathbb{P}\left(\sum_{i=1}^{n} D_i \geq t\right) \leq e^{-2t^2/(\sum_{i=1}^{n}(b_i-a_i)^2)} \quad \forall t > 0.$$

Suppose f satisfies bounded-differences:

$$|f(x_1,\ldots,x_{i-1},x_i,\ldots,x_n) - f(x_1,\ldots,x_{i-1},\tilde{x_i},\ldots,x_n)| \leq c_i \quad \forall x_i, \tilde{x_i}.$$

Then,

$$|D_i| = |\mathbb{E}[f(X_1^n)|X_1,\ldots,X_i] - \mathbb{E}[f(X_1^n)|X_1,\ldots,X_{i-1}]| \leq c_i$$

since expected values are just linear combinations of functions **I'm confused?**. We can apply the Azuma-Hoeffding bound with $a_i = -c_i$ and $b_i = c_i$ to get the following claim.

Claim 4.3
If f satisfies $(c_i)_{i=1}^n$-bounded-differences,

$$\mathbb{P}[f(X_1^n) - \mathbb{E}[f(X_1^n)] \geq t] \leq e^{\frac{-2t^2}{4\sum_{i=1}^{n} c_i^2}} = e^{\frac{-t^2}{2\sum_{i=1}^{n} c_i^2}} \quad \forall t > 0.$$

look at this example later

Example 4.4
See book for concentration of clique number in a random graph model. Something about Vertex martingales, Edge martingales.

4.3 Noise Complexities

Suppose we observe some data

$$Y = \Theta^* + W,$$

where $\Theta^* \in \mathcal{C} \subseteq \mathbb{R}^n$ is an unknown object. We want to be able to recover Θ^* given noise $W \in \mathbb{R}^n$. Say that the noise components are i.i.d. satisfying $\mathbb{E}[w_i] = 0$ and

$\text{Var}[w_i] = \sigma^2$. Our goal is to recover a least-squares estimate

$$\hat{\Theta} = \arg\min_{\Theta \in \mathcal{C}} \frac{1}{n} \|Y - \Theta\|_2^2.$$

Now let's say we have another fresh sample

$$\tilde{Y} = \Theta^* + \tilde{W},$$

which could represent a holdout dataset or something similar. We are interested in an empirical risk function defined on this holdout data:

$$L(\theta) = \frac{1}{n} \mathbb{E}[\|\tilde{Y} - \theta\|_2^2].$$

With some math, **??**, we can show that

$$L(\theta) = \frac{1}{n} \|\theta - \theta^*\| + \sigma^2.$$

We can measure the quality of our procedure with the following quantity:

$$L_n(\hat{\theta}) - L(\hat{\theta}).$$

If this quantity is too large, then we are overfitting, because the empirical risk generated by our new samples is large compared to expected risk. We can expand this quantity further:

$$\begin{aligned} L_n(\hat{\theta}) &= \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\theta}_i)^2 \\ &= \frac{1}{n} \sum (\theta_i^* - \hat{\theta}_i + w_i)^2 \\ &= \frac{1}{n} \sum (\theta_i^* - \hat{\theta}_i)^2 + \frac{1}{n} \sum_{i=1}^{n} w_i^2 - \frac{2}{n} \sum_{i=1}^{n} w_i (\hat{\theta} - \theta^*). \end{aligned}$$

When we subtract $L(\hat{\theta})$,

$$L_n(\hat{\theta}) - L(\hat{\theta}) = \left(\frac{1}{n} \sum_{i=1}^{n} w_i^2 - \sigma^2 \right) - \frac{2}{n} \sum_{i=1}^{n} w_i (\hat{\theta} - \theta^*).$$

The first part of this expression (in parenthesis) is an i.i.d. zero-mean tail bound, which we know how to bound effectively. The worst case of over-fitting is when we have

$$\hat{\theta} = Y = \theta^* + W,$$

in which case the last term becomes very large:

$$-\frac{1}{n}\sum_{i=1}^{n} w_i(\hat{\theta}_i - \theta_i^*) = -\frac{1}{n}\sum_{i=1}^{n} w_i^2.$$

The quantity $L(\hat{\theta})$ is called that **population risk**, while $L_n(\hat{\theta})$ is called the **empirical risk**.

$$L(\hat{\theta}) \le L_n(\hat{\theta}) + \left|\frac{1}{n}\sum_{i=1}^{n} w_i^2 - \sigma^2\right| + \sup_{\theta\in\mathcal{C}}\left|\frac{1}{n}\sum_{i=1}^{n} w_i(\theta_i - \theta_i^*)^2\right|$$

The rightmost term is known as the **noise complexity** of the set

$$a = \{\theta - \theta^* | \theta \in \mathcal{C}\}.$$

4.4 Rademacher and Gaussian complexities

In general noise complexity, the only guarantees that we made about the errors were that they were zero-main with variance σ^2. We could focus on more specific distributios of error to obtain more specific complexities.

Let $a \subseteq \mathbb{R}^n$. Then define

$$Z(a) = \sup_{a\in a}\left|\frac{1}{n}\sum_{i=1}^{n} a_i\varepsilon_i\right|,$$

where $\varepsilon_i \in \{\pm 1\}$ are i.i.d. Rademacher. The inner sum can also be written more succinctly as the inner product $\langle a, \varepsilon\rangle$.

$$R(a) = \mathbb{E}_\varepsilon[Z_n(a)].$$

We could replace the inner ε with any type of random variable, like Gaussians, which would give a Gaussian complexity.

Example 4.5

L_p balls.

Define our noise set

$$a = \mathbb{B}_p(1) = \left\{ a \middle| \left(\sum_{j=1}^{n} |a_j|^p \right)^{1/p} \leq 1 \right\}.$$

Then

$$R(\mathbb{B}_1(1)) = \sup_{\|a\|_1 \leq 1} \frac{1}{n} \langle a, \varepsilon \rangle \leq \frac{1}{n} \sup_{\|a\|_1 \leq 1} \|a\|_1 \|\varepsilon\|_\infty = \frac{1}{n},$$

by Holder's inequality. We also have

$$R(\mathbb{B}_2(1)) = \sup_{\|a\|_2 \leq 1} \frac{1}{n} \langle a, \varepsilon \rangle \leq \frac{1}{n} \sup_{\|a\|_2 \leq 1} \|a\|_2 \|\varepsilon\|_2 = \frac{1}{\sqrt{n}},$$

by Cauchy Schwarz. It is left as an exercise to show that

$$R(\mathbb{B}_\infty(1)) = 1.$$

look at this later