MA 581 Notes: Mathematics of Data Science

Instructor: Dmitriy Drusvyatskiy
Scribe: Mars Gao

October 25, 2022

1 Introduction

How does one optimally extract information from data $S_n = z_1, ..., z_n \sim^{i.i.d.} \mathcal{P}$

1.1 Complexity

There are two sources to understand and measure complexity.

1. Statistical complexity: samples
2. Computational complexity: flops, gradient evaluations, optimization, computer science

Question: How does everything work under high dimensional settings?

Example 1.1. Mean estimation and Shrinkage
Suppose you get to observe $S_n x_1, ..., x_n \sim \mathcal{N}(\mu, \Sigma)$. Your goal is to estimate μ.
One solution is just to compute the mean that

$$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

But in what sense $\bar{x}_n$ is a good estimation? A: Mean squared error defined as

$$\mathbb{E}_{P_n}||\bar{x}_n - \mu||_2^2 = \frac{tr(\Sigma)}{n}$$

Is there a better estimator?
Simple answer: NO! Because the sample mean is minimax-optimal that

$$\inf_{\hat{x}_n}\sup_{\mu}\mathbb{E}_{S_n\sim\mathcal{N}(\mu,\Sigma)}||\hat{x}_n - \mu||_2^2 \geq c\frac{tr(\Sigma)}{n}$$

But a more complicated answer is "yes".
Suppose for simplicity $\Sigma = I$.
Consider bias-variance decomposition that

$$\mathbb{E}||\hat{x}_n - \mu||_2^2 = \mathbb{E}||\hat{x}_n - \mathbb{E}\hat{x}_n||_2^2 + ||\mathbb{E}\hat{x}_n - \mu||_2^2$$

However, in high dimensions, it pays to trade bias for variance!!

Definition 1.2. $\hat{x}_n$ strictly dominates $\tilde{x}_n$ if

$$\mathbb{E}||\hat{x}_n - \mu||^2 \leq \mathbb{E}||\tilde{x}_n - \mu||^2, \ \forall\mu$$

and there exists μ_0 s.t.

$$\mathbb{E}||\hat{x}_n - \mu_0|| < \mathbb{E}||\tilde{x}_n - \mu_0||^2.$$

Then $\tilde{x}_n$ is called inadmissable.

Theorem 1.3. *$\bar{x}_n$ is inadmissable if and only if $d \geq 3$.*

To show this Theorem, let's define the famous James-Stein skrinkage estimator that

$$x_n^{JS} = \left(1 - \frac{\sigma^2(d-2)}{n||\bar{x}||^2}\right)\bar{x}_n$$

The intuition behind is that in high dimensions, the ball has much larger volumn given radius $\sigma\sqrt{d}$. Therefore, it pays to shrink x to reduce the variance. In high-dimension, it pays a lot to achieve unbiasedness.

Proof. We compute the MSE of JS estimator that

$$\begin{aligned}\mathbb{E}||x_n^{JS} - \mu||_2^2 &= \frac{\sigma^2 d}{n} - \frac{\sigma^2}{n}(d-2)^2\mathbb{E}\left[\frac{\sigma^2/n}{||\bar{x}_n||^2}\right] \\ &\le \frac{\sigma^2 d}{n} - \frac{\sigma^2(d-2)^2}{n(d-2+\frac{n}{\sigma^2}||\mu||^2)}\end{aligned}$$

□

Example 1.4. Compressed sensing

Suppose we get to observe

$$y = Ax_{\#},$$

where $A \in \mathbb{R}^{m\times d}$ is a Gaussian random matrix and $x_{\#} \in \mathbb{R}^d$ has at most s nonzero entries.
Our goal is to recover $x_{\#}$.
From convex optimization, we can do in the following way that

$$\begin{aligned}&\min_x ||x||_1 \\ &Ax = y\end{aligned}$$

As soon as $m < s\log\left(\frac{d}{s}\right)$, with high probability, $x_{\#}$ is the unique solution.
A geometric reason is that $x_{\#}$ solves the optimization problem if and only if

$$ker(A) \cap \{v : ||x_{\#} + v|| \le ||x_{\#}||_1\} = \{0\}$$

Q: What is the probability that a random subspace intersects a convex cone trivially?

2 Basic Probability

Definition 2.1. Expectation and variance. Let X be a random variable on probability space. The expectation

$$\mathbb{E}[X]$$

Conditional expectation,

$$\mathbb{E}[X|Y]$$

and Variance

$$Var(X) = \mathbb{E}(X - \mathbb{E}X)^2 = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

Definition 2.2. Moment generating function is defined as

$$m_X(t) = \mathbb{E}[e^{tX}], \quad t \in \mathbb{R}.$$

Definition 2.3. Denote the L^p norm as

$$||X||_p = (\mathbb{E}[|X^p|])^{1/p}$$

Definition 2.4. Banach space is

$$L^p = \{X : ||X||_p < \infty\}$$

Remark 2.5. L^2 is a Hilbert space.

We denote

$$\langle X, Y\rangle_2 = \mathbb{E}[XY], \qquad ||X||_2 = \sqrt{\langle X, X\rangle} = \sqrt{\mathbb{E}[X^2]}$$

The covariance

$$\begin{aligned}cov(X,Y) &= \mathbb{E}\left([X - \mathbb{E}[X]][Y - \mathbb{E}[Y]]\right) \\ &= \langle X - \mathbb{E}[X], Y - \mathbb{E}[Y]\rangle\end{aligned}$$

2.1 Important Distributions

1. Uniform distribution
2. Gaussian distribution
3. Rademacher distribution
$$p(x=1) = p(x=-1) = \frac{1}{2}$$
4. Bernoulli(p)
5. Poisson λ

2.2 A few basic facts

Definition 2.6. A family $(X_1, ..., X_k)$ is independent if

$$P[X_i \in E_i, \forall i = 1, ..., k] = \prod_{i=1}^{k} P[X_i \in E_i]$$

Remark 2.7. [Linearlity of expectation]

$$\mathbb{E}[\sum c_i X_i] = \sum_{i=1}^{k} \mathbb{E} X_i$$

Remark 2.8. [Linearlity of variance] If $X_1, ..., X_k$ are pairwise independent, then

$$Var(\sum_{i=1}^{k} X_i) = \sum_{i=1}^{k} Var(X_i)$$

Remark 2.9. [Tower rule]

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$$

Lemma 2.10. *[Markov inequality] For any non-negative X and $t > 0$, we have*

$$\mathbb{P}[X \geq t] \leq \frac{\mathbb{E}X}{t}$$

Proof. We see

$$\begin{aligned} \mathbb{E}X &= \mathbb{E}X\mathbf{1}_{\{x \geq t\}} + \mathbb{E}X\mathbf{1}_{\{x<t\}} \\ &\geq t\mathbb{E}_{\{x \geq t\}} \\ &= t\mathbb{P}[X \geq t] \end{aligned}$$

□

3 Concentration Inequalities

3.1 Chernoff Bound

Let $X_1, ..., X_n$ be r.v.'s with $\mathbb{E}X = 0$. The question is: how big is $|\sum X_i|$ typically?

In general, this quantity can be $\mathcal{O}(n)$. But if $X_1, ..., X_n$ are pairwise independen, then using Chebyshev gives us

$$P\left(|\sum X_i| \geq t\right) \leq \frac{\sum Var(X_i)}{t^2}$$

So,

$$P\left(\left|\sum X_i\right| \geq \lambda\sqrt{\sum Varr(X_i)}\right) \leq \frac{1}{\lambda^2}$$

Therefore, with high probability,

$$\left|\sum X_i\right| = \mathcal{O}(\sqrt{n}),$$

if $Var(X_i) = \sigma^2$.
Question:
When ca we expect to replace $\frac{1}{\lambda^2}$ by $e^{-\lambda}$ or $e^{-\lambda^2}$?

Example 3.1. [Motivating example] Consider if we wish to control that

$$P\left[\sup_{i\in I} X_i \geq t\right] \leq \sum_{i\in I} P\left[X_i \geq t\right]$$

If $|I|$ is huge, need $P\left[X_i \geq t\right]$
E.g. the control of $\sup_{x\in X}\left|\mathbb{E}_z f(x,z) - \frac{1}{n}\sum f(x,z_i)\right|$ which is an empirical process.

The Chernoff method is described in the following.
Let X be r.v. with $\mu = \mathbb{E}X < \infty$. Then, for all $\lambda \geq 0$, we have

$$\begin{aligned} P\left[X-\mu \geq t\right] &= P\left[e^{\lambda(X-\mu)} \geq e^{\lambda t}\right] \\ By\ Markov &\leq \frac{\mathbb{E}e^{\lambda(X-\mu)}}{e^{\lambda t}} \end{aligned}$$

This derives that

$$\begin{aligned} \log P\left[X-\mu \geq t\right] &\leq \inf_{\lambda\geq 0}\left\{\log \mathbb{E}e^{\lambda(X-\mu)} - \lambda t\right\} \\ &= -\sup_{\lambda\geq 0}\left\{\lambda t - \log \mathbb{E}e^{\lambda(X-\mu)}\right\} \end{aligned}$$

Define any function $\varphi : \mathbb{R} \to \mathbb{R}\cup\{+\infty\}$, the Fenchle conjugate is defined as

$$\varphi^*(t) = \sup_{\lambda}\left\{\lambda t - \psi(\lambda)\right\}$$

Let's look at the main example

$$\psi_X(\lambda) = \log \mathbb{E}e^{\lambda(X-\mu)}$$

For all $\lambda \in \mathbb{R}$, observe from Jensen

$$\psi_X(\lambda) = \log \mathbb{E}e^{\lambda(X-\mu)} \geq \mathbb{E}\log e^{\lambda(X-\mu)} = 0$$

So when $\lambda < 0$ and $t > 0$, we have

$$\lambda t - \psi(\lambda) \leq 0 = 0 - \psi(0)$$

Therefore, for $t \geq 0$, the equality holds.

$$\psi_X^*(t) = \sup_{\lambda\geq 0}\left\{t\lambda - \psi(\lambda)\right\}$$

We arrive at the Chernoff bound that

$$P\left[X-\mu \geq t\right] \leq \exp\left(-\psi_X^*(t)\right)$$

where $\psi_X(\lambda) = \log\left(\mathbb{E}e^{\lambda(X-\mu)}\right)$.

Example 3.2. Let $X \sim \mathcal{N}(\mu,\sigma^2)$. Then,

$$\mathbb{E}e^{\lambda(X-\mu)} = e^{\frac{\sigma^2\lambda^2}{2}}$$

Then,

$$\psi_X^*(t) = \sup_{\lambda} \lambda t - \frac{\sigma^2\lambda^2}{2} = \frac{t^2}{2\sigma^2}$$

Therefore,

$$P\left[X \geq \mu + t\right] \leq \exp\left(-t^2/2\sigma^2\right), \quad \forall t > 0$$

3.2 Sub-Gaussian Random variable

Definition 3.3. [Sub-Gaussian variable] Define X with mean μ is sub-Gaussian with parameter $\sigma > 0$ if

$$\mathbb{E}e^{\lambda(X-\mu)} \leq e^{\frac{\sigma^2\lambda^2}{2}}, \quad \forall \lambda \in \mathbb{R}.$$

If X is sub-gaussian, so is $-X$. We have the tail bound that

$$P\left[|X-\mu| \geq t\sigma\right] \leq 2e^{-t^2/2}$$

Lemma 3.4. *[Bounded random variable] Suppose X is supported on $[a,b]$. Then X is $\frac{b-a}{2}$ sub-Gaussian.*

Proof. Set $y = X - \mu$ and define

$$f(\lambda) = \log\left(\mathbb{E}\exp(\lambda y)\right)$$

Then,

$$f'(\lambda) = \frac{\mathbb{E}y\exp(\lambda y)}{\mathbb{E}\exp(\lambda y)}$$

$$f''(\lambda) = \frac{\mathbb{E}y^2\exp(\lambda y)}{\mathbb{E}\exp(\lambda y)} - \left[\frac{\mathbb{E}y\exp(\lambda y)}{\mathbb{E}\exp(\lambda y)}\right]^2$$

Define a measure $dm = \frac{\exp(\lambda y)dy}{\mathbb{E}\exp(\lambda y)}$
Then,

$$\begin{aligned} f''(\lambda) &= Var_m(y) \\ &= \inf_t \left[(y-t)^2\right] \\ &\leq \mathbb{E}\left[(y - \frac{a+b}{2})^2\right] \\ &= \frac{(b-a)^2}{4} \end{aligned}$$

Finally, using Tylor's theorem, we know

$$f(\lambda) = f(0) + f'(0)\lambda + \frac{1}{2}f''(\tilde{\lambda})\lambda^2$$

We could further know that

$$f(\lambda) \leq 0 + 0 + \frac{1}{2}\frac{(b-a)^2}{4}\lambda^2$$

□

Lemma 3.5. *[Sum rule] Suppose X_i are independent σ_i-sub-Gaussian, then*

$$\sum X_i \text{ is } \sqrt{\sum \sigma_i^2}\text{-sub-Gaussian}$$

From here, we have the corollary which is the famour Hoeffding inequality.

Corollary 3.6. *[Hoeffding]. Suppose $X_1, ..., X_n$ are independent with $\mathbb{E}X_i = \mu_i$ and these X_i's are σ_i-sub-Gaussian. Then*

$$P\left[\sum(X_i - \mu_i) \geq t||\sigma||_2\right] \leq \exp\left\{-\frac{t^2}{2}\right\}$$

Additionally, if $\mu_i = \mu$, $\sigma_i = \sigma$, then

$$P\left[\sum(X_i - \mu) \geq t\sigma\sqrt{n}\right] \leq \exp\left\{-\frac{t^2}{2}\right\}$$

It turns out the indepence in Hoeffding can be weakened to martingale difference sequences.

Theorem 3.7. *[Azuma] Let $X_1, ..., X_n$ be r.v.'s with*

$$\mathbb{E}\left(X_i|X_{i-1},...,X_1\right) = \mathbb{E}\left(X_i|X_{i-1}\right)$$

and

$$\mathbb{E}\left(\exp(\lambda X_i)|X_{i-1},...,X_1\right) \leq e^{\sigma_i^2\lambda^2/2}$$

Then, $\sum X_i$ is $||\sigma||_2$-subGaussian.

Proof. Set $S_n = \sum X_i$. Then

$$\begin{aligned}\mathbb{E}\exp\left(\lambda S_n\right) &= \mathbb{E}\left[\exp(\lambda S_{n-1})\mathbb{E}\left[\exp(\lambda X_n)|X_1,...,X_{n-1}\right]\right] \\ &\leq e^{\sigma_n^2\lambda^2/2}\mathbb{E}\exp(\lambda S_{n-1}) \\ &\leq e^{||\sigma||_2^2\lambda^2/2}\end{aligned}$$

□

3.3 Sub-exponential random variable

Example 3.8. Let $z \sim \mathcal{N}(0,1)$. Let's compute

$$\begin{aligned}\mathbb{E}\left[e^{\lambda(Z^2-1)}\right] &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{\lambda(x^2-1)}e^{-x^2/2}dx \\ &= \begin{cases}\frac{e^{-\lambda}}{\sqrt{1-2\lambda}} & if\ \lambda \leq \frac{1}{2} \\ +\infty & if\ \lambda > \frac{1}{2}\end{cases}\end{aligned}$$

Definition 3.9. [Sub-exponential] Define X with mean μ is sub-exponential with parameters (ν, α) if

$$\mathbb{E}\left[e^{\lambda(X-\mu)}\right] \leq e^{\nu^2\lambda^2/2}, \quad \forall|\lambda| \leq \frac{1}{\alpha}.$$

Back to the example 3.8, we see that

$$\mathbb{E}\left[e^{\lambda(z^2-1)}\right] \leq \frac{e^{-\lambda}}{\sqrt{1-2\lambda}} \leq e^{4\lambda^2/2}, \quad |\lambda| < \frac{1}{4}$$

So, z^2 is $(2,4)$-subexponential.

Theorem 3.10. *[Sub-exponential tail bound] Let X be subexponential with (ν, α). Then*

$$P\left[X-\mu \geq t\right] \leq \begin{cases} e^{-t^2/2\nu^2} & ,if\ |t| \leq \nu^2/\alpha \\ e^{-t/2\alpha} & ,otherwise\end{cases}$$

Proof. Back to Chernoff.

$$\log P\left[X-\mu \geq t\right] \leq -\psi_X^*(t)$$

where $\psi_X(\lambda) = \log \mathbb{E}e^{\lambda(X-\mu)}$. This quantity, we have

$$\begin{aligned}\psi_X(\lambda) &= \log \mathbb{E}e^{\lambda(X-\mu)} \\ &= \begin{cases}\nu^2\lambda^2/2 & ,if\ |\lambda| \leq 1/\alpha \\ +\infty & ,otherwise\end{cases}\end{aligned}$$

□

Theorem 3.11. *[Bernstein] Let X be subexponential with parameter (ν, α) and mean μ. Then*

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

Lemma 3.12. *[Sum rule] X_i are (ν_1, α_i)-subexponential, then*

$$\sum X_i\ is\ \left(||\sigma||_2, ||\alpha||_\infty\right)\text{-}subExponential$$

Theorem 3.13. *[Bernstein for summation] Let X_i are (ν_1, α_i)-subexponential with mean $\mu_i = \mathbb{E}X_i$*

$$P\left[\left|\sum (X_i - \mu_i)\right| \geq t\right] \leq 2\exp\left[-\frac{1}{2}\left(\frac{t^2}{||\nu||_2^2} \wedge \frac{t}{||\alpha||_\infty}\right)\right].$$

Theorem 3.14. *[Improved Bernstein for bounded RVs] Suppose $|X - \mu| \leq b$, $\mathbb{E}(X-\mu)^2 = \sigma^2$. Then,*

$$\mathbb{E}e^{\lambda(X-\mu)} \leq \exp\left(\frac{\lambda^2\sigma^2}{2(1-b|\lambda|)}\right),\ \forall|\lambda| > \frac{1}{b}.$$

Therefore,

$$P\left[|X-\mu| \geq t\right] \leq 2\exp\left(-\frac{t^2}{2(\sigma^2 + bt)}\right)$$

Proof. Using Taylor expnsion:

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

Follow from Chernoff, by setting $\lambda = \frac{t}{bt+\sigma^2} \in [0, \frac{1}{b}]$ ☐

This is superior to Hoeffding when $\sigma \ll b$.

3.4 Application: Dimensionality Reduction

Given $u_1, u_2, ..., u_m \in \mathbb{R}^d$ with $m \ll d$, can we map $u_1, .u_2, ..., u_m$ to a lower dimensional space with low distortion?

Theorem 3.15. *[Johnson-Lindenstrauss] Fix $\epsilon, \delta \in (0,1)$, a set $U \subseteq \mathbb{R}^d$ of m points and a number $n > \frac{16\log(\frac{m^2}{\sigma})}{\epsilon^2}$. Let $X \in \mathbb{R}^{n\times d}$ consist of i.i.d. $\mathcal{N}(0,1)$ entries. Then with probability $1-\delta$, the map $f(u) = \frac{1}{\sqrt{n}}Xu$ satisfies*

$$1 - \epsilon \leq \frac{||f(u) - f(v)||_2^2}{||u-v||_2^2} \leq 1 + \epsilon, \quad \forall u, v \in U$$

Proof. Observe that

$$\frac{||Xu||_2^2}{||u||_2^2} = \sum_{i=1}^{n} \frac{\left\langle X_i, \frac{u}{||u||}\right\rangle^2}{i.i.d.\ \mathcal{N}(0,1)}$$

This gives rise to

$$\frac{||Xu||_2^2}{||u||_2^2}\ is\ (2\sqrt{n}, n) - subExponential$$

Using Bernstein,

$$\begin{aligned}
P\left[\left|\frac{||Xu||_2^2}{n||u||_2^2} - 1\right| > \epsilon\right] &\leq 2\exp\left[-\left(\frac{n\epsilon^2}{8} \wedge \frac{n\epsilon}{8}\right)\right] \\
&= 2\exp\left[-\left(\frac{n\epsilon^2}{8}\right)\right]
\end{aligned}$$

So for any i, j, we have

$$P\left[\frac{||f(u_i-u_j)||_2^2}{||u_i-u_j||_2^2}\notin[1-\epsilon,1+\epsilon]\right]\le 2e^{-n\epsilon^2/8}$$

Take the union bound over $\binom{m}{2}$ pairs, we have

$$2\binom{m}{2}e^{-n\epsilon^2/8}\le m^2e^{-n\epsilon^2/8}=\delta.$$

□

Question 3.16. *What if $m=\infty$ but U only has a few "degree of freedom"? Next, we will look at concentration of $f(x_1,...,x_n)$ where f is a "well-behaved" function and $x_1,...,x_n$ are independent r.v's.*

Bounded differences inequality (McDiarmid) So far, we have focused on n concentration of the average $\frac{1}{n}\sum_{i=1}^n X_i$.

Remark 3.17. [Useful insight] As long as $f(x_1,...,x_n)$ depends weakly on individual x_i, the concentration holds!

Theorem 3.18. *[McDiarmid] Suppose that $f:X^n\to\mathbb{R}$ has the bounded difference property that $\exists L_1,L_2,...,L_n$ such that*

$$|f(x_1,...,x_k,...,x_n)-f(x_1,...,x_k',...x_n)|\le L_k,\quad\forall x,x'\in X^n.$$

Then, for independent rv's $X=(x_1,...,x_n)$, we have

$$P\left[|f(X)-\mathbb{E}f(X)|>t\right]\le 2e^{-\frac{2t^2}{||L||_2^2}}$$

Proof. We will use the martingale method.

Define

$$y_0=\mathbb{E}f(X)\ and\ y_i=\mathbb{E}[f(X)|x_1,...,x_i]$$

We observe that

$$y_i=y_0+\sum_{j=0}^{i-1}(y_{j+1}-y_j)=y_0+\sum_{j=1}^{i}D_j$$

Further, we see

$$\begin{aligned}\mathbb{E}\left[y_i|x_1,...,x_{i-1}\right]&=\mathbb{E}\left[\mathbb{E}\left[f(X)|x_1,...,x_i\right]|x_1,...,x_{i-1}\right]\\&=\mathbb{E}\left[f(X)|x_1,...,x_{i-1}\right]\\&=y_{i-1}\end{aligned}$$

Therefore, we know that

$$\mathbb{E}\left[y_i-y_{i-1}|x_1,...,x_{i-1}\right]=\mathbb{E}\left[D_{j+1}|x_1,...,x_{i-1}\right]=0$$

Then, we can compute that

$$\begin{aligned}\mathbb{E}\left[e^{\lambda(f(x)-\mathbb{E}[f(x)])}\right]&=\mathbb{E}\left[e^{\lambda(y_n-y_0)}\right]\\&=\mathbb{E}\left[e^{\lambda\sum_{j=1}^n D_j}\right]\\&=\mathbb{E}\left[e^{\lambda(y_{n-1}-y_0)}e^{\lambda D_n}\right]\\&=\mathbb{E}\left[e^{\lambda(y_{n-1}-y_0)}\mathbb{E}\left[e^{\lambda D_n}|x_1,x_2,...,x_{n-1}\right]\right]\end{aligned}$$

Let $x'\neq x$ be another random sample from x_i that $x_i'\sim^{iid}x_i$. Then,

$$\begin{aligned}\mathbb{E}\left[e^{\lambda D_i}|x_1,...,x_{i-1}\right]&=\mathbb{E}\left[e^{\lambda(y_i-y_{i-1})}|x_1,...,x_{i-1}\right]\\&=\mathbb{E}\left[e^{\lambda\mathbb{E}\left[f(X)-f(X')|x_1,,,.x_i\right]}|x_1,...,x_{i-1}\right]\\(Jensen)&\le\mathbb{E}\left[e^{\lambda(f(X)-f(X'))}|x_1,...,x_{i-1}\right]\\&\le e^{\frac{\lambda^2L_1^2}{8}}\end{aligned}$$

Therefore, in total, we see

$$\mathbb{E}\left[e^{\lambda(f(X)-\mathbb{E}f(X))}\right] \leq e^{\frac{\lambda^2}{8}||L||_2^2}$$

Then, apply Chernoff, we have

$$P\left[|f(X)-\mathbb{E}f(X)|>t\right] \leq 2e^{-\frac{2t^2}{||L||_2^2}}$$

□

3.5 Lipschitz transformation of Gaussians

Theorem 3.19. *Let $X_1, ..., X_n \sim^{iid} \mathcal{N}(0,1)$ and let $F : \mathbb{R}^n \to \mathbb{R}$ be L-Lipschitz:*

$$|F(x)-F(y)| \leq L||x-y||_2, \quad \forall x, y \in \mathbb{R}^n$$

Then,

$$F(X)-\mathbb{E}F(X) \text{ is } \frac{\pi L}{\sqrt{2}} - \text{subGaussian}$$

To show the theorem above, we need the following exercise.

Exercise 3.20. Suppose that (X, Y) are jointly normal. Then, X and Y are independent iff

$$\mathbb{E}\left[XY\right] = \mathbb{E}X\mathbb{E}Y$$

Proof. We can assume WLOG:
$L = 1$, $\mathbb{E}F(X_1, ..., X_n) = 0$, F is C'-smooth (otherwise approximate). Let Y be an independent realization of X. Then

$$\begin{aligned}\mathbb{E}\exp\left(\lambda F(X)\right) &= \mathbb{E}\exp\left(\lambda F(X)\right)\cdot 1 \\ &\leq \mathbb{E}\exp\left(\lambda F(X)\right)\mathbb{E}\exp\left(-\lambda F(Y)\right) \\ &= \mathbb{E}\exp\left(\lambda\left(F(X)-F(Y)\right)\right)\end{aligned}$$

We write $F(X)-F(Y)$ that

$$F(X)-F(Y) = \int_0^{\pi/2} \left(F\circ\gamma\right)'(\theta)d\theta$$

where $\gamma(\theta) = Y\cos(\theta) + X\sin(\theta)$.
Note here that

$$\dot{\gamma}(\theta) = -Y\sin(\theta) + X\cos(\theta)$$

So, $(\gamma(\theta), (\theta))$ jointly normal with $Cor\left(\gamma(\theta), \dot{\gamma}(\theta)\right) = 0$. □

The proof is a little bit beyond my understanding so I will understand it later.
Recall if $X_1, ..., X_n$ are independent σ-subGaussian with $\mathbb{E}X_i = \mu$. The Hoeffding implies that $\hat{x} = \frac{1}{n}\sum x_i$ satisfies

$$P\left[|\hat{x}-\mu| \leq t\right] \geq 1-2\exp\left(-\frac{nt^2}{2\sigma^2}\right)$$

or equivalently

$$P\left[|\hat{x}-\mu| \leq \sqrt{\frac{2\sigma^2\log(2/\rho)}{n}}\right] \geq 1-\rho.$$

Can we achieve similar guarantee without subGaussian assumption with a different estimator $\hat{x}$?

Theorem 3.21. *[Mediam of means] Consider $X \in \mathbb{R}$ with $\mathbb{E}X = \mu$ and $Var(X) = \sigma^2$. Let $X_1, ..., X_n$ be i.i.d. realizations of X subdivide into $k = 18\log\left(\frac{1}{\rho}\right)$ bins and form the empirical means $\hat{x}_j$ for $j = 1, ..., k$. Then $\hat{x} = median(\hat{x}_1, ..., \hat{x}_k)$ satisfies*

$$P\left[|\hat{x}-\mu| \leq \sqrt{\frac{54\sigma^2\log(1/\rho)}{n}}\right] \geq 1-\rho$$

Proof. By Chebyshev,

$$P\left[|\hat{x}_i - \mu| \geq \sqrt{\frac{3\sigma^2 k}{n}}\right] \leq \frac{\sigma^2 k/n}{3\sigma^2 k/n} = \frac{1}{3}, \quad \forall i$$

By Hoeffding

$$P\left[\frac{1}{k}\sum_{i=1}^{k} \mathbf{1}\left\{|\hat{x}_i - \mu| \geq \sqrt{\frac{3\sigma^2 k}{n}}\right\} > \frac{1}{2}\right] \geq 1 - \exp\left(-\frac{k}{18}\right).$$

We can know that

$$|\hat{x} - \mu| \leq \sqrt{\frac{3\sigma^2 k}{2n}}$$

In this case, $\hat{x}$ depends on the confidence level ρ. □

4 Random vectors in High Dimensions

- Concentration of the norm
- Isotropy
- Similarity of Normal and Spherical
- Sub-Gaussian and Sub-Exponential random vectors.

Two main results we'll prove in this chapter.

- Sub-Gaussian vectors are concentrated around a sphere.
- Two independent isotropic subGaussian random vectors are nearly orthogonal in high dimensions.

We will next investigate the behavior of random vectors in high dimensions!!!

Concentration of the norm Let $X = (X_1, ..., X_d) \in \mathbb{R}^d$ have independent σ-subGaussian coordinates with

$$\mathbb{E}X_i = 0 \textit{ and } \mathbb{E}X_i^2 = 1$$

What should we expect for

$$||X||_2^2 \textit{ and } ||X||_2$$

Lemma 4.1. *Suppose y is σ-subGaussian. Then y^2 is $(\sigma, 4\sigma^2)$ subexponential.*

Proof. [Sketch]
Step 1: Estimate $\mathbb{E}\left[|y|^r\right] \leq r2^{r/2}\sigma^r\Gamma\left(\frac{r}{2}\right)$ using $\mathbb{E}\left[|y|^r\right] = \int_0^\infty P\left[|y| > t^{1/r}\right] dr$.
Step 2: Use Taylor expansion that

$$\begin{aligned}\mathbb{E}\left[e^{\lambda\left(y^2 - \mathbb{E}y^2\right)}\right] &\leq 1 + \sum_{r=2}^{\infty} \lambda^r 2^{r+1}\sigma^{2r} \\ &\leq 1 + \frac{8\lambda^2\sigma^4}{1 - 2\lambda\sigma^2} \\ &\leq \exp\left(...\right)\end{aligned}$$

□

Corollary 4.2. *Let $X = (X_1, ..., X_d) \in \mathbb{R}^d$ have independent σ-subGaussian coordinates with*

$$\mathbb{E}X_i = 0 \textit{ and } \mathbb{E}X_i^2 = 1$$

Then $P\left[|||X||_2^2 - d| \geq td\right] \leq 2\exp\left(-\frac{d}{4\sigma^2}\left(t \wedge t^2\right)\right)$ which is just

$$P\left[|||X||_2 - \sqrt{d}| \geq t\sqrt{d}\right] \leq 2\exp\left(-\frac{dt^2}{4\sigma^2}\right)$$

Proof. We see $||X||_2^2$ that

$$||X||_2^2 = \sum_{i=1}^{d} X_i^2$$

This is the sum of $d$random Chi-square samples. We see that (i) $\mathbb{E}||X||_2^2 = d$ and (ii) $||X||_2^2$ is $\left(\sigma\sqrt{d}, 4\sigma^2\right)$ subexponential.

Using Bernstein, we see that

$$P\left[\left|\frac{1}{d}||X||_2^2 - 1\right| \geq t\right] \leq 2\exp\left[-\frac{d}{4\sigma^2}\left(t \wedge t^2\right)\right]$$

Observe that $\forall z \geq 0$, we have

$$|z-1| \geq t \rightarrow |z^2 - 1| \geq \min(t, t^2)$$

So

$$\begin{aligned} P\left[\left|\frac{1}{\sqrt{d}}||X||_2 - 1\right| \geq t\right] &\leq P\left[\left|\frac{1}{d}||X||_2^2 - 1\right| \geq t^2 \wedge t\right] \\ &\leq 2\exp\left(-\frac{dt^2}{4\sigma^2}\right) \end{aligned}$$

□

4.1 Isotropic vectors

Recall for $X \in \mathbb{R}^d$, covariance

$$cov(X) = \mathbb{E}\left[(X-\mu)(X-\mu)^T\right]$$

where $\mu = \mathbb{E}[X]$.

Definition 4.3. A random vector $X \in \mathbb{R}^d$ with $\mathbb{E}X = 0$ is isotropic if

$$\Sigma(X) = \mathbb{E}\left[XX^T\right] = I_d$$

Remark 4.4. If $\Sigma = \Sigma(X)$ is invertible, then $z := \Sigma^{-1/2}(X - \mu)$ is isotropic.

Lemma 4.5. *X is isotropic iff*

$$\mathbb{E}\langle X, y\rangle^2 = ||y||_2^2, \quad \forall y \in \mathbb{R}^d.$$

Proof. X is isotropic

$$\begin{aligned} &iff\ \mathbb{E}XX^T = I_d \\ &iff\ y^T\mathbb{E}XX^Ty = y^Ty \\ &iff\ \mathbb{E}y^TXX^Ty = ||y||_2^2 \\ &iff\ \mathbb{E}\langle X, y\rangle^2 = ||y||_2^2 \end{aligned}$$

□

Thus, if $\mathbb{E}X = 0$, then X is isotropic iff marginal $\left\langle X, \frac{y}{||y||}\right\rangle$ has unit variance $\forall y \in \mathbb{R}^d$.

Lemma 4.6. *Let $X \in \mathbb{R}^d$ be isotropic. Then $\mathbb{E}||X||_2^2 = d$. Moreover, if X and y are two independent isotropic vectors, then*

$$\mathbb{E}\langle X, y\rangle^2 = d$$

Proof. First,

$$||X||_2^2 = X^TX = tr\left(XX^T\right)$$

Therefore,

$$\mathbb{E}||X||_2^2 = tr(I_d) = d$$

Nest,

$$\begin{aligned} \mathbb{E}\langle X, y\rangle^2 &= \mathbb{E}_y\left[\mathbb{E}_X\langle X, y\rangle^2 |y\right] \\ &= \mathbb{E}_y\left[||y||_2^2\right] \\ &= d \end{aligned}$$

□

Let X and Y be independent and isotropic. Then we see $||X|| \sim \sqrt{d}, ||y|| \sim \sqrt{d}$ and $\left\langle \frac{X}{||X||}, \frac{y}{||y||}\right\rangle \sim \frac{1}{\sqrt{d}}$. Can be rigorous by assuming light tails.

4.1.1 Examples of isotropic Random Variables

1. Spherical uniform RV. $X \sim Unif\left(\sqrt{d}S^{d-1}\right)$.
2. Symmertic Bernoulli: $X \sim Unif\left(\{-1,1\}^d\right)$
3. Any vector $X = (X_1, ..., X_d)$ where X_i are independent, zero mean, unit variance.
4. Coordinate $Unif\left(\left\{\sqrt{d}e_i\right\}_{i=1}^d\right)$
5. Gaussian $g = (g_1, ..., g_d) \sim \mathcal{N}(0, I_d)$. Recall this means g_i are i.i.d. $\mathcal{N}(0,1)$.
   The density of the Gaussian is
   $$p(x) = \prod_{i=1}^d p_i(x) = \prod_{i=1}^d \frac{1}{\sqrt{2\pi}} e^{-\frac{x_i^2}{2}} = \frac{1}{(2\pi)^{n/2}} e^{-\frac{||x||^2}{2}}.$$
   After applying a random rotation matrix, the standard multivariate Gaussian is still standard multivarriate Gaussian.

Exercise 4.7. Let $g \sim \mathcal{N}(0, I_d)$. Then $r := ||g||_2$ and $\theta = \frac{g}{||g||_2}$ are independent random variables and $\theta \sim Unif(S^{d-1})$.

Definition 4.8. X is $\mathbb{R}^d$ is σ-subGaussian if $\langle X, u\rangle$ is σ-subGaussian $\forall u \in S^{d-1}$.

Example 4.9. Let $X = (X_1, ..., X_d)$ be RV with independent σ-subGaussian X_i. Then X is σ-subGaussian.

1. $\mathcal{N}(0, I_d)$ is 1-subGaussian.
2. $Unif\left(\{-1,1\}^d\right)$ is 1-subGaussian.
3. $Unif\left(\left\{\sqrt{d}e_i\right\}_{i=1}^d\right)$ is σ-subGaussian with $\sigma \asymp \sqrt{\frac{d}{\log d}}$
4. $Unif\left(\sqrt{d}S^{d-1}\right)$ is c-subGaussian for a constant c.

5 Introduction to Statistical Inference