MA 581 Notes: Mathematics of Data Science

October 11, 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

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

□

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

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

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*

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

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

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

□