MA 581 Notes: Mathematics of Data Science

September 28, 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?