Introduction

Tuesday, January 11, 2022 9:24 AM

SEEM 5380: https://www.se.cuhk.edu.hk/~manchoso

What is this course about?

Example: Linear Regression

Samples: $z_i = (x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$; $i = 1, \dots, n$

parameters: $\theta \in \mathbb{R}^d$

model: $y = X\theta^* + \varepsilon$; $X \in \mathbb{R}^{n \times d}$, $y \in \mathbb{R}^n$

$\theta^* \in \mathbb{R}^d$: ground truth (unknown) ; $\varepsilon \in \mathbb{R}^n$: additive noise

- Classical setting: $n >> d$ (overdetermined)

If $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, then the MLE of θ^* is

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \underbrace{\|y - X\theta\|_2^2}_{\mathcal{L}(\theta)} \quad \text{— (*)}$$

Issues:

1) Optimization: How to find $\hat{\theta}$?

(*) is convex; solvable in polynomial time

Lightweight methods?

e.g.: Gradient descent

$$\theta^{k+1} \leftarrow \theta^k - \alpha_k \nabla \mathcal{L}(\theta^k), \quad \alpha_k > 0 \text{ step size}$$

$$\nabla \mathcal{L}(\theta^k) = 2X^T(X\theta^k - y)$$

We care about efficiency. How fast does

$\mathcal{L}(\theta^*) - \mathcal{L}(\hat{\theta})$ decay? — (△)

gap to optimality

or $\|\theta^* - \hat{\theta}\|_2$ decay? — (△△)

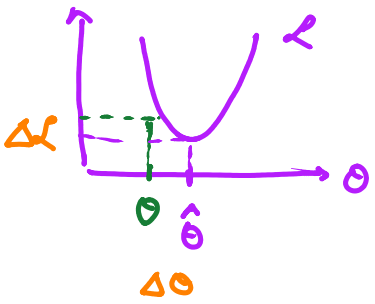
distance to optimal solution

2) Statistics: Is (△) or (△△) enough?

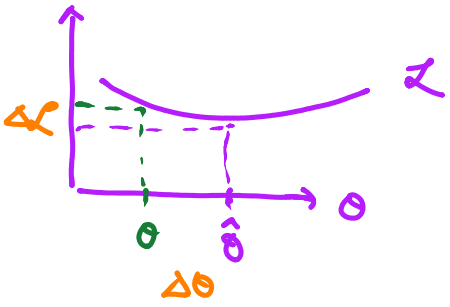
What is the estimation error?

Typically measured by $\|\hat{\theta} - \theta^*\|_2$ or other norms of interest.

Given $\mathcal{L}(\theta^*) - \mathcal{L}(\hat{\theta})$ is small, do we necessarily have $\|\hat{\theta} - \theta^*\|_2$ small?



high curvature



low curvature

⇒ need to understand the curvature of the loss function.

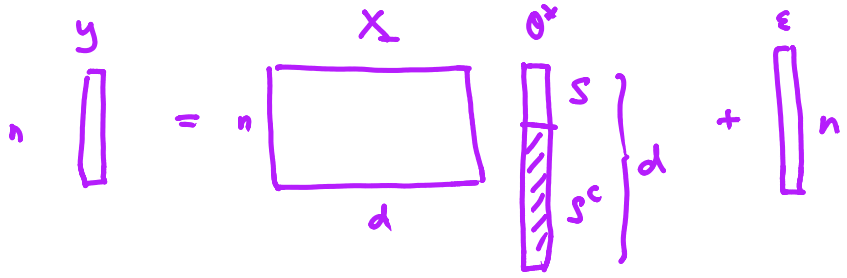
- contemporary setting: $d \gg n$ (underdetermined)

  - Impossible to recover θ^* without further assumptions.

- Typically assume θ^* has some "low-dimensional" structure

Continuing with linear regression:

$$y = X\theta^* + \varepsilon \quad ; \quad \theta^* \text{ is sparse}$$

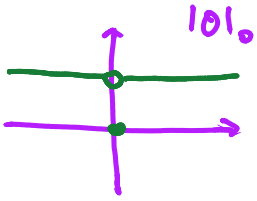


Ideally,

(constrained version) $$\hat{\theta} = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \|y - X\theta\|_2^2 \; : \; \|\theta\|_0 \le R \right\}$$

data fidelity; sparsity; sparsity threshold

$\|\theta\|_0$ = # of non-zero entries in θ.



or

(regularized version) $$\hat{\theta} = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \left\{ \|y - X\theta\|_2^2 + \lambda \|\theta\|_0 \right\}$$

Abstract Estimation Problem

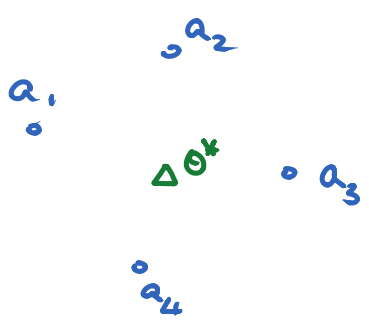
- measurable map F

$$(X, \theta^*, \varepsilon) \longmapsto F(X, \theta^*, \varepsilon)$$

- Given X, $F(X, \theta^*, \varepsilon)$, recover θ^*.
  - Dimensions of X, θ^* and nature of noise ε affect the formulation of the loss function

Example: Source Localization

Example: Source Localization



θ^* : source (unknown position)

a_i : sensor (known position)

measurements

$$d_i = \| \theta^* - a_i \|_2 + \varepsilon_i \; ; \; i = 1, \dots, m$$

Estimate θ^* by

$$\hat{\theta} \in \underset{\theta \in \mathbb{R}^2}{\operatorname{argmin}} \sum_{i=1}^{m} \left(\| \theta - a_i \|_2 - d_i \right)^2$$