ECE 595: Machine Learning I

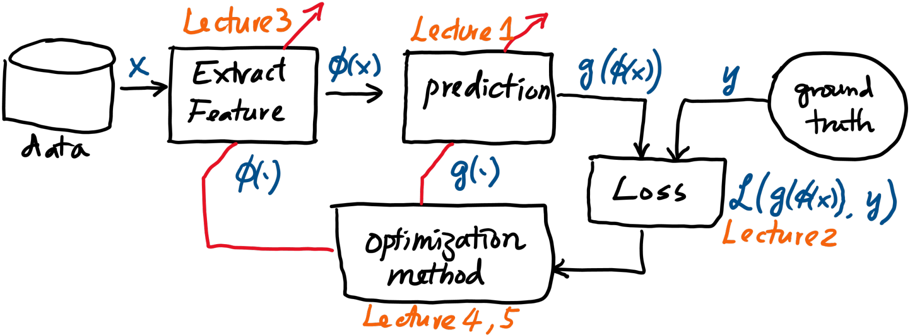
Lecture 4.1: Intro to Optimization - Unconstrained Optimization

Spring 2020

Stanley Chan

School of Electrical and Computer Engineering Purdue University

Outline


Lecture3
Lecture 1
data
x
Extract Feature
ϕ(x)
prediction
g(ϕ(x))
y
ground truth
ϕ(·)
g(·)
Loss
L(g(ϕ(x)), y)
Lecture2
optimization method
Lecture 4, 5

Outline

Mathematical Background

- Lecture 4: Intro to Optimization
- Lecture 5: Gradient Descent

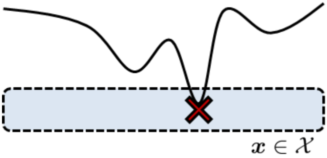
Lecture 4: Intro to Optimization

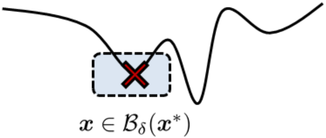
- Unconstrained Optimization
  - First Order Optimality
  - Second Order Optimality
- Convexity
  - What is convexity?
  - Convex optimization
- Constrained Optimization
  - Lagrangian
  - Examples

Unconstrained Optimization

$$\underset{\boldsymbol{x} \in \mathcal{X}}{\text{minimize}} \quad f(\boldsymbol{x})$$

- $\boldsymbol{x}^* \in \mathcal{X}$ is a **global minimizer** if
  - $f(\boldsymbol{x}^*) \leq f(\boldsymbol{x})$ for any $\boldsymbol{x} \in \mathcal{X}$
- $\boldsymbol{x}^* \in \mathcal{X}$ is a **local minimizer** if
  - $f(\boldsymbol{x}^*) \leq f(\boldsymbol{x})$, for any $\boldsymbol{x}$ in a neighborhood $\mathcal{B}_\delta(\boldsymbol{x}^*)$
  - $\mathcal{B}_\delta(\boldsymbol{x}^*) = \{\boldsymbol{x} \mid \|\boldsymbol{x} - \boldsymbol{x}^*\|_2 \leq \delta\}$





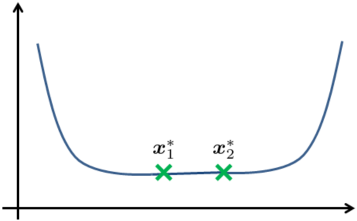
Uniqueness of Global Minimizer

If $\boldsymbol{x}^*$ is global minimizer, then

- Objective value $f(\boldsymbol{x}^*)$ is unique
- Solution $\boldsymbol{x}^*$ is not necessarily unique

Therefore:

- Suppose $f(\boldsymbol{x}) = g(\boldsymbol{x}) + \lambda\|\boldsymbol{x}\|_1$ for some convex g.
- "$\underset{\boldsymbol{x}}{\text{minimize}}\ f(\boldsymbol{x})$" has a global optimal $f(\boldsymbol{x}^*)$.
- But there could be multiple $\boldsymbol{x}^*$'s.
- Some $\boldsymbol{x}^*$ maybe better, but not in the sense of $f(\boldsymbol{x})$.



First and Second Order Optimality

$$\underbrace{\nabla f(\boldsymbol{x}^*) = \mathbf{0}}_{\text{First order condition}} \quad \text{and} \quad \underbrace{\nabla^2 f(\boldsymbol{x}^*) \succeq 0}_{\text{Second order condition}} .$$

Necessary Condition:

If $\boldsymbol{x}^*$ is a global (or local) minimizer, then

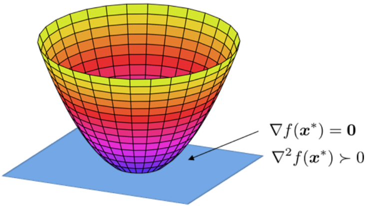
- $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$.
- $\nabla^2 f(\boldsymbol{x}^*) \succeq 0$.

Sufficient Condition:

If $\boldsymbol{x}^*$ satisfies

- $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$.
- $\nabla^2 f(\boldsymbol{x}^*) \succ 0$.

then $\boldsymbol{x}^*$ is a global (or local) minimizer.



Why? First Order

- Why is $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$ necessary?
- Suppose $\boldsymbol{x}^*$ is the minimizer.
- Pick any direction $\boldsymbol{d}$, and any step size ϵ. Then

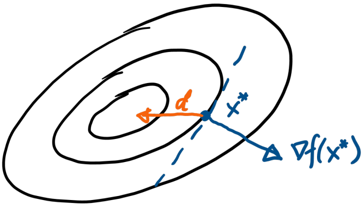
$$f(\boldsymbol{x}^* + \epsilon \boldsymbol{d}) = f(\boldsymbol{x}^*) + \epsilon \nabla f(\boldsymbol{x}^*)^T \boldsymbol{d} + \cancel{\mathcal{O}(\epsilon^2)}.$$

- Rearranging the terms yields

$$\underbrace{\lim_{\epsilon \to 0} \left\{ \frac{f(\boldsymbol{x}^* + \epsilon \boldsymbol{d}) - f(\boldsymbol{x}^*)}{\epsilon} \right\}}_{\geq 0,\ \forall \boldsymbol{d}} = \nabla f(\boldsymbol{x}^*)^T \boldsymbol{d}.$$

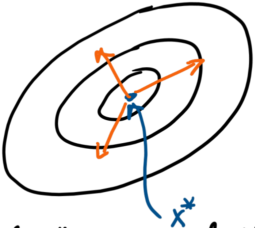
- So $\nabla f(\boldsymbol{x}^*)^T \boldsymbol{d} \geq 0$ for all $\boldsymbol{d}$. True only when $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$.

First Order Condition Illustrated



if x^* is not optimal, then

$\nabla f(x^*)^T d \le 0$, for **some** d



if x^* is optimal, then

$\nabla f(x^*)^T d = 0$, for **all** d

Why? Second Order

Do third order approximation:

$$f(\boldsymbol{x}^* + \epsilon \boldsymbol{d}) = f(\boldsymbol{x}^*) + \underbrace{\epsilon \nabla f(\boldsymbol{x}^*)^T \boldsymbol{d}}_{=0} + \frac{\epsilon^2}{2} \boldsymbol{d}^T \nabla^2 f(\boldsymbol{x}^*) \boldsymbol{d} + \frac{\epsilon^3}{6} \mathcal{O}(\|\boldsymbol{d}\|^3)$$

Therefore,

$$\frac{1}{\epsilon^2}\Big[f(\boldsymbol{x}^* + \epsilon \boldsymbol{d}) - f(\boldsymbol{x}^*)\Big] = \frac{1}{2} \boldsymbol{d}^T \nabla^2 f(\boldsymbol{x}^*) \boldsymbol{d} + \left[\frac{\epsilon}{6} \mathcal{O}(\|\boldsymbol{d}\|^3)\right]$$

$$\lim_{\epsilon \to 0} \frac{1}{\epsilon^2} \underbrace{\Big[f(\boldsymbol{x}^* + \epsilon \boldsymbol{d}) - f(\boldsymbol{x}^*)\Big]}_{\geq 0} = \frac{1}{2} \boldsymbol{d}^T \nabla^2 f(\boldsymbol{x}^*) \boldsymbol{d} + \lim_{\epsilon \to 0} \left[\frac{\epsilon}{6} \mathcal{O}(\|\boldsymbol{d}\|^3)\right],$$

Hence,

$$\frac{1}{2} \boldsymbol{d}^T \nabla^2 f(\boldsymbol{x}^*) \boldsymbol{d} \geq 0, \qquad \forall \boldsymbol{d}.$$

$\Rightarrow$ positive semi-definite!

Second Order Condition Illustrated