Lecture 1

Aug / 27 / 23

Instructor: Mateo Diaz (mateodd@jhu.edu)

Office Hours: Mondays 4:00 - 5:30 pm
Wyman S429

TAs: Wyman S425

- Pedro Izquierdo (pizquie1@jh.edu)
  OH: Th 10:00 - 11:30 am
- Daniel Lopez (jlopezc1@jh.edu)
  OH: Tue 10:00 - 11:30 am
- Thabo Samakhona (tsamakh1@jhu.edu)
  OH: Wed 10:30 - 11:15 am

Resources

- Canvas
- Website (mateodd25.github.io/nonlinear)
- Piazza ← Ask questions here before emailing me
- Gradescope ← All submissions.

Agenda

- Syllabus
- Motivation
- Overview
- Background Review

Syllabus

Four components:

Some theory
Some code
(Python please)

- Homework (5 - 6)
- Midterm Takehome (Oct 11 - 17)
- Final Takehome (Dec 13 - 15)

Might change

- Participation
  - Engaging in class, OH, Piazza.

Grading System

Let C_H, C_M, C_F, C_P denote your normalize grades $(0 - 100)$.

Let H, M, F be variable weights for each component.

Your grade will be the optimal value of

First piece of motivation

$$\begin{aligned}
\max \quad & C_H \cdot H + C_M \cdot M + C_F \cdot F \\
& + C_p \cdot \underbrace{(100 - H - M - F)}_{= P} \\
\text{s.t.} \quad & (H, M, F) \in \mathbb{R}^3 \\
& 90 \leq H + M + F \leq 100 \\
& H, M \geq 15 \\
& F \geq M \\
& 50 \leq M + F \leq 80
\end{aligned}$$

Motivation

We want to solve

$$\min_{x \in C} f(x)$$

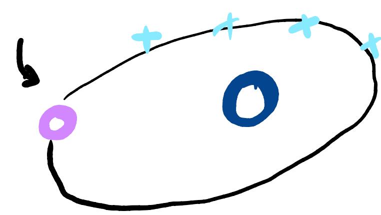
In this class we will focused in the unconstrained setting $C = \mathbb{R}^d$.

Example 1 (Least-Squares)

Gauss was interested in predicting the position of Ceres (planetoid)

From 22 observations made by Joseph Piazzi:

where next?



$(x_1, y_1), \ldots, (x_{22}, y_{22})$.

Gauss assumed that the data was close to an ellipse:

$$\alpha x^2 + \beta y^2 + \gamma xy = 1.$$

To find α, β, γ he formulated

$$\min_{\alpha, \beta, \gamma} \sum_{i=1}^{22} (\alpha x_i^2 + \beta y_i^2 + \gamma x_i y_i - 1)^2$$

Gauss solved this problem and obtained meaningful predictions (after a 100 hours of computation).

This is an instance of a least-squares problem

$$\min_{\bar{w} \in \mathbb{R}^d} \|\underline{A}\bar{w} - \bar{b}\|^2 = \sum_{i=1}^{n} (\bar{a}_i^T \bar{w} - b_i)^2$$

$$a_i = \begin{bmatrix} x_i^2 \\ y_i^2 \\ x_i y_i \end{bmatrix} \quad b_i = 1, \quad \bar{\beta} = \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix}$$

Example 2: Data fitting in general

Learning problem

Data: $(\bar{x}_1, y_1), \dots, (\bar{x}_n, y_n)$

Goal: Find function $f(x_i) \approx y_i$.

Approach

loss function measuring similarity

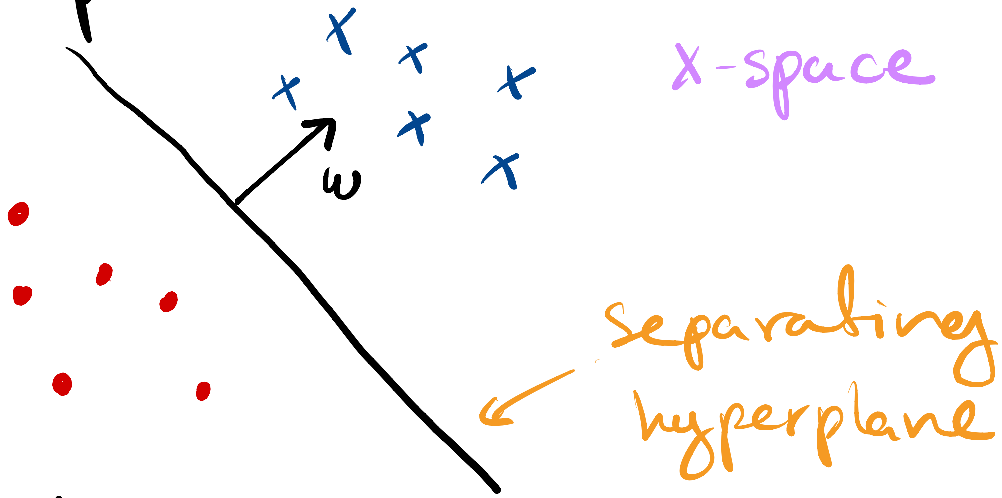
$$\min_{\bar{w}} \sum_{i=1}^{n} \ell(f_{\bar{w}}(\bar{x}_i), y_i)$$

Parametrized function, before $f_{\bar{w}}(a_i) = \bar{a}_i^T \bar{w}$

Imagine we want to predict whether a patient has COVID from observations

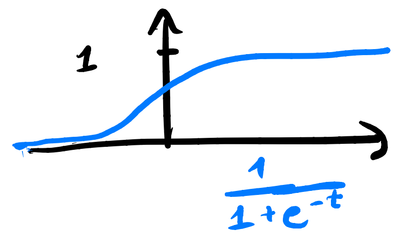
$$x_i = \begin{bmatrix} \text{age} \\ \text{heart rate} \\ \text{blood preassure} \\ \text{temperature} \\ \vdots \end{bmatrix} \qquad y = \begin{cases} 1 & \text{if COVID} \\ 0 & \text{otherwise} \end{cases}$$

We might imagine that data is linearly separable



Consider

$$f_\omega(x) = \frac{1}{1+e^{-\omega^T x}}$$



With logistic regression we minimize

$$\min_{w} \sum_i y_i \ln f_w(x_i) + (1-y_i)\ln(1-f_w(x_i))$$

In general we can use more complicated parametrizations

$$f_W(x) = W_L \circ \dots \circ \sigma_2 \circ W_2 \circ \sigma_1 \circ W_1 \bar{x}$$

Nonlinear functions (Relu)

Parameters

This is what gives rise to neural networks.

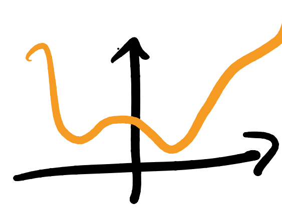
Overview

Geometry $\begin{cases} \text{Optimality conditions} \\ \text{Basic convex analysis.} \end{cases}$

First-order methods

When we can only use $x \mapsto \nabla f(x)$

- For smooth functions



- For convex functions

- For nonsmooth functions

- For stochastic funtions

$$f(x) = \mathbb{E}_z \, F(x, z)$$

Second-order methods

When we have access to

$$x \longmapsto (\nabla f(x), \nabla^2 f(x))$$

- Newton's method
- Quasi-Newton methods
- Trust region

Time permitting

- Linear Programming
- Conjugate Gradient
- Composite optimization.