Introduction: Optimization for Data Science Course Outline

TANN Chantara

Department of Applied Mathematics and Statistics
Institute of Technology of Cambodia

October 7, 2022

Table of Contents

Prerequisites

Required:

- Calculus and linear algebra
  Familiarity with basic matrix manipulations and concepts from calculus (e.g., differentiability, gradient, continuity)
- Pleasure with mathematically rigorous statements
  Sufficient mathematical maturity regarding proof techniques (direct proof, proof by contradiction, composition)
- Basic knowledge on probability
  Familiarity with the following concepts: event, random variable, probability density function, cumulative distribution function, conditional probability, independence, expected value, variance, etc.

Aims of the Course

During this course, students should

- learn how to formalize decision problems in machine learning and statistics as mathematical optimization models
- build a good understanding of convex optimization problems
- see how to formulate scalable and accurate implementations of the most important optimization algorithms for machine learning
- be able to characterize trade-offs between time and accuracy, for machine learning methods
- understand how to assess/evaluate the most important algorithms, function classes, and algorithm convergence guarantees

This course provides foundations for advanced topics in ML, e.g., Deep learning, Reinforcement learning, Statistical learning theory

General Remarks

This course is split into two main themes

1. Convex optimization models are used in:
    - classification methods of patients in health care;
    - internet search engines to rank online documents;
    - supply chain management to predict inventory levels;
    - online advertisement to optimally place adds;
    - etc. etc.
2. Gradient descent and its variants are used to:
    - solve convex optimization problems in high dimensions;
    - train neural networks;
    - compute optimal policies for reinforcement learning problems;
    - dynamically price airline tickets;
    - etc. etc.

Data science crucially relies on mathematical optimization and its corresponding solution methods

Recommended Books

- Stephen Boyd and Lieven Vandenberghe, Convex Optimization, Cambridge University Press, 2009
  - extremely well written and comprehensive; the book and some supplementary material can be downloaded for free
- D.P. Bertsekas, Convex Optimization Theory, Athena Scientific, 2009.
  - theoretical perspective and introduction to convex optimization; excellent reference book for research; the book can be downloaded for free
- Y. Nesterov, Introductory Lectures on Convex Optimization, Springer, 2004.
  - Excellent and complete treatment of gradient descent methods for solving convex optimization probems

Course Outline

Part I: Convex optimization

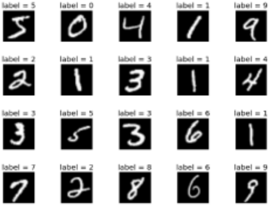
- Optimization problems
- Convex sets, convex functions, convex optimization problems
- Lagrangian duality
- Optimality conditions
- Optimization in Statistics and Machine Learning

Part II: Algorithms

- Gradient descent
- Projected gradient descent
- Stochastic gradient descent
- Subgradient method
- Outlook

Example 1: Handwritten digit recognition

- Goal: recognize handwritten decimal digits $0, 1, \ldots, 9$
- Set $\mathcal{P} \subset \mathbb{R}^{784}$ of grayscale images (28×28 pixels)
- Image (feature) $x \in \mathcal{P} \to$ digit (label) $d(x) \in \{0, 1, \ldots, 9\}$



- Predict new digit as $y = Wx \in \mathbb{R}^{10}$, y_j probability of digit being j
- Conversion to actual probabilities $z_j = z_j(y) = e^{y_j} / \sum_{k=0}^{9} e^{y_k}$
- Find the "best" matrix $W \in \mathbb{R}^{10 \times 784}$

Example 1: Handwritten digit recognition

- Find the "best" matrix $W \in \mathbb{R}^{10\times 784}$

$$\min_{W\in\mathbb{R}^{10\times 784}} \ell(W)$$

- Loss function

$$\ell(W) = -\sum_{x\in\mathcal{P}} \log\left(z_{d(x)}(Wx)\right) = \sum_{x\in\mathcal{P}} \left(\log\left(\sum_{k=0}^{9} e^{(Wx)_k}\right) - (Wx)_{d(x)}\right)$$

  - loss function "punishes" images for which the correct digit j has low probability z_j

How do we solve $\min_{W\in\mathbb{R}^{10\times 784}} \ell(W)$?

Example 2: Master's admission

Goal: Predict performance of MSc applicants based on application documents

- Features: $x^{(1)}$ =GPA, $x^{(2)}$ =TOEFL
- Label: y =MSc grade

GPA	TOEFL	MSc grade
3.52	100	3.08
3.66	109	2.67
3.76	113	2.20
3.74	100	2.33
3.93	100	1.48
3.88	115	1.56
3.77	115	1.96
3.66	107	2.27
3.87	106	1.97
3.84	107	1.94

Master's admission - center data

- Linear regression model

$$\text{MSc grade} = w_0 + w_1\text{GPA} + w_2\text{TOEFL}$$

- Center the data

$$x_j^{(i)} \leftarrow x_j^{(i)} - \frac{1}{n}\sum_{k=1}^{n} x_k^{(i)},\ i = 1, 2, \qquad y_j \leftarrow y_j - \frac{1}{n}\sum_{k=1}^{n} y_k$$

GPA	TOEFL	MSc grade
-0.24	-7.2	0.93
-0.10	1.8	0.52
-0.01	5.8	0.05
-0.02	-7.2	0.18
0.17	-7.2	-0.67
0.12	7.8	-0.59
0.01	7.8	-0.19
-0.10	-0.2	0.12
0.11	-1.2	-0.18
0.08	-0.2	-0.21

Master's admission - rescale data

- Linear regression model

$$\text{MSc grade} = w_1 \text{GPA} + w_2 \text{TOEFL}$$

- Rescale the data

$$x_j^{(i)} \leftarrow x_j^{(i)} \sqrt{n / \sum_{k=1}^{n} (x_k^{(i)})^2}, i = 1, 2$$

GPA	TOEFL	MSc grade
-2.06	-1.28	0.93
-0.87	0.32	0.52
-0.03	1.03	0.05
-0.19	-1.28	0.18
1.41	-1.28	-0.67
0.99	1.39	-0.59
0.06	1.39	-0.19
-0.87	-0.04	0.12
0.90	-0.21	-0.18
0.65	-0.04	-0.21

Master's admission - linear regression

Linear regression. Find the optimal parameter $w^\star$ as the solution to

$$\min_{w_1, w_2 \in \mathbb{R}} \sum_{k=1}^{10} (w_1 x_k^{(1)} + w_2 x_k^{(2)} - y_k)^2$$

- Predict the MSc grade of a new applicant with rescaled, normalized features $(\bar{x}^{(1)}, \bar{x}^{(2)})$ as $\bar{y} = w_1^\star \bar{x}^{(1)} + w_2^\star \bar{x}^{(2)}$
- In the example, we get $w^\star = (-0.92, -0.09)$, which implies that the first input (GPA) has a much higher influence on the output (MSc grade) than the second input (TOEFL)

Questions

- Should we set $w_2^\star = 0$ for better predictions? Which loss function leads to best prediction? $\Rightarrow$ Course on statistical learning theory
- How to compute $w^\star$? $\Rightarrow$ this course

Example 3: Designing portfolios in finance

An investor can put his money in n different assets

- Proportion of asset i in portfolio is x_i
- Each asset i has expected return p_i
- Desired return is $\mu > 0$
- Risk is measured by the variance of the portfolio, where Σ is the known covariance matrix of the portfolio

Goal: Find portfolio with minimal risk that provides return μ

Markowitz portfolio model.

$$\begin{cases} \min\limits_{x \in \mathbb{R}^n} & x^\top \Sigma x \\ \text{s.t.} & p^\top x = \mu \\ & \sum_{i=1}^n x_i = 1, x_i \geq 0 \ \forall i \end{cases}$$

Harry Markowitz
Nobelprize 1952