SDSC6015
Stochastic Optimization for Machine Learning

0. Course Basics

Clint Chin Pang Ho
clint.ho@cityu-dg.edu.cn

Department of Data Science

Semester A 2025

Course Basics: Teaching Team

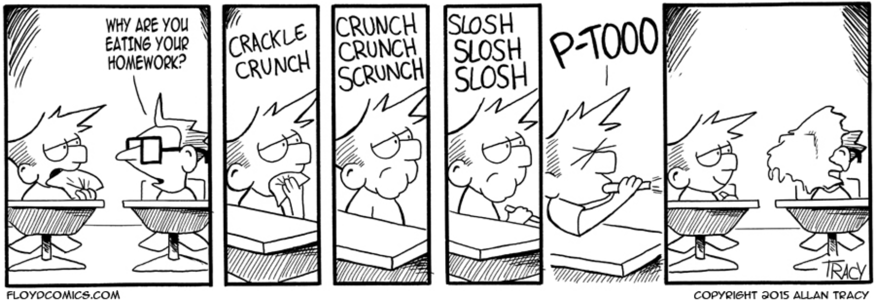
Instructor

- **Email:** clint.ho@cityu-dg.edu.cn (OR clint.ho@cityu.edu.hk) (aim to response within 2 working day)
- **Office hour:** By appointment

Our TAs

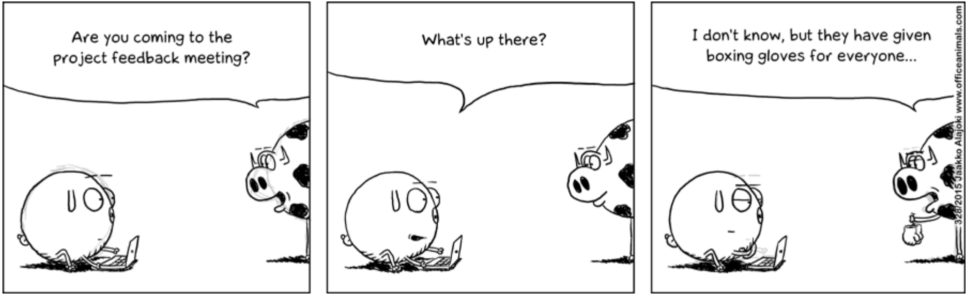
- **Zhuodong Yu,**
  Email: zhuodonyu2-c@my.cityu.edu.hk
- **Ling Dai,**
  Email: lingdai5-c@my.cityu.edu.hk
- **Wen Bai,**
  Email: wen.bai@my.cityu.edu.hk

Course Basics: Assessment



- **Two** assignments (each contributes 5% of the final grade).
- **One** midterm exam (contributes 20% of the final grade).
- **One** group project (contributes 30% of the final grade).
- **One** final exam (contributes 40% of the final grade).

Course Basics: Let me know your feedback!


Are you coming to the project feedback meeting?
What's up there?
I don't know, but they have given boxing gloves for everyone...
3/28/2015 Jaakko Alajoki www.officeanimals.com

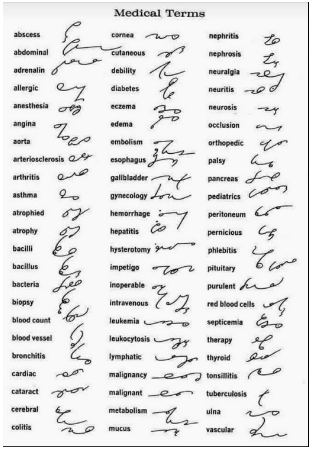
We need your feedback!

- This is the 2nd version of this course!
- This is a PG course for everyone!

Course Basics: About Submission

- Channel: Online (Canvas)
- File types: .pdf, .py, .mp4, and .txt
- **Submitting your code:** If you are using other formats (e.g., Jupyter Notebook), you can copy and paste your code to a txt file!
- **Late homework/project policy:** if you are late for t days ($t > 0$), the maximum of percentage is $(0.75)^t \times 100\%$.

Course Basics: About Submission



Medical Terms

abscess
abdominal
adrenalin
allergic
anesthesia
angina
aorta
arteriosclerosis
arthritis
asthma
atrophied
atrophy
bacilli
bacillus
bacteria
biopsy
blood count
blood vessel
bronchitis
cardiac
cataract
cerebral
colitis

cornea
cutaneous
debility
diabetes
eczema
edema
embolism
esophagus
gallbladder
gynecology
hemorrhage
hepatitis
hysterotomy
impetigo
inoperable
intravenous
leukemia
leukocytosis
lymphatic
malignancy
malignant
metabolism
mucus

nephritis
nephrosis
neuralgia
neuritis
neurosis
occlusion
orthopedic
palsy
pancreas
pediatrics
peritoneum
pernicious
phlebitis
pituitary
purulent
red blood cells
septicemia
therapy
thyroid
tonsillitis
tuberculosis
ulna
vascular

Figure: Handwriting: we are not in a medical school...

Course Basics: On GenAI (Reference)

- Students are allowed to used Generative Artificial Intelligence (GenAI) tools for non-exam/test assessment tasks (i.e., assignments and group project).

- The use of GenAI must be properly acknowledged. See `https://www.cityu.edu.hk/ah/Tutorial/citing`. Students should clearly indicate which part(s) of the material are generated by any GenAI tool(s).

- Students are entirely accountable for all materials submitted, regardless of whether they are generated using GenAI tools or not.

A few questions...

Overview

1. Course Basics
2. What is Optimization?
3. Course Intended Learning Outcomes
4. Some Machine Learning Models $\Rightarrow$ Optimization

What is Optimization?

Generic form

$$\begin{array}{ll} \underset{\boldsymbol{x}}{\text{minimize}} & f_0(\boldsymbol{x}) \\ \text{subject to} & \boldsymbol{x} \in \mathcal{X} \end{array}$$

- $\boldsymbol{x}$ is the decision (or optimization or design) variable
- f_0 is the objective function
- $\mathcal{X}$ is the feasible set, which is usually defined through constraints

$$\begin{aligned} f_i(\boldsymbol{x}) \leq 0 \quad & \forall i \in [m] \\ h_i(\boldsymbol{x}) = 0 \quad & \forall i \in [p] \end{aligned}$$

How about maximization?

Example: Portfolio Optimization

Challenge: Given \$100, how should we maximize our expected return by investing over n assets while controlling our risk?

Optimization problem:

- **decision variables:** the amounts invested in different assets
- **constraints:** the risk/variance is less than certain value, maxs/min. investment per asset
- **objective:** the expected return

No analytical solution!

Example: Flight Booking for Your Lecturers

Challenge: Plan a trip for the lecturers (different origins, same destination)

Name	Location
Matthias	LCY
Zijun	CHN
Xiang	ATH
Qi	BRL
Lishuai	BOS
Yu	LON

- 6 people, 10 flights from each location to destination 10^6 possibilities.
- **Combinatorial problem** – in practice intractable

Applications of Optimization I

Optimization & Engineering

- Engineering aims to transform scientific discoveries into practical devices (software, structures, machines, processes)
- Optimization is used to make the devices better:
  - Improve design
  - Cheaper
  - Lighter
  - ...

Optimization & Economics/Finance

- Optimize investment portfolios (e.g. Mean-Variance Optimization)
- Algorithmic Trading (e.g. Optimize trade execution)
- Calibrate statistical models

Applications of Optimization II

Optimization & Computing

- Machine Learning, e.g.
    - Neural Networks (unconstrained optimization)
    - Support Vector Machines (quadratic programming)
    - Reinforcement Learning (dynamic programming)
    - Markov Decision Processes (dynamic/linear programming)
    - Genetic/Evolutionary Algorithms (stochastic optimization)
- Computer Vision
- Many more...

Optimization is Everywhere!

Leonhard Euler:

"...nothing at all takes place in the universe in which some rule of maximum or minimum does not appear..."

L. Euler. *Methodus inveniendi lineas curvas maximi minimive proprietate gaudentes, sive Solutio problematis isoperimetrici latissimo sensu accepti* (1744).

It doesn't mean anything!

Most optimization problems are hard to solve!

Overview

1. Course Basics

2. What is Optimization?

3. **Course Intended Learning Outcomes**

4. Some Machine Learning Models ⇒ Optimization

Aims of SDSC6015

- Understand the methodologies and fundamental mathematical structures used in optimization.

- Explore the key principles and characteristics of stochastic approximation in the context of modern machine learning optimization problems.

- Study basic forms of stochastic optimization algorithms applied to different machine learning models.

- Develop and analyze practical algorithms for solving related optimization problems.

Topics for this Course (Tentative)

- Convexity
- Classifying Types of Optimization Problems
- Optimization & Optimality Conditions
- Duality
- Algorithm (Basic): Gradient Descent
- Algorithm (Constrained): Projected (Sub)Gradient Method
- Algorithm (Composite): Proximal Gradient Method
- Algorithm (Stochastic & Others): Stochastic Gradient Descent & ADMM & Coordinate & Incremental & Frank-Wolfe & Zero-Order

Reference Books

This course should be self-contained but below is a subset of good books for learning optimization:

- A. Ben-Tal and A. Nemirovski. *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications* (2001).
- D. P. Bertsekas. *Nonlinear Programming* (2016).
- D. P. Bertsekas. *Convex Optimization Algorithms* (2015).
- D. Bertsimas. *Machine Learning Under a Modern Optimization Lens* (2019).
- J. C. Duchi. *Introductory Lectures on Stochastic Optimization* (2016).
- G. Lan. *First-order and Stochastic Optimization Methods for Machine Learning* (2020).
- D. G. Luenberger and Y. Ye. *Linear and Nonlinear Programming* (2008).
- M. Jaggi and N. Flammarion. *CS-439: Optimization for Machine Learning* (2024).
- J. Nocedal and S. Wright. *Numerical Optimization* (2006).
- L. Vandenberghe and S. P. Boyd. *Convex Optimization* (2008).

Overview

1. Course Basics

2. What is Optimization?

3. Course Intended Learning Outcomes

4. **Some Machine Learning Models ⇒ Optimization**
   - 4.1 Least Squares Regression
   - 4.2 Logistics Regression
   - 4.3 More

Example 1: Haidilao or not?

Motivation

Wen needs to decide whether she should go to the restaurant “Haidilao” for lunch or not. She went to ask her friends Ling and Zhuodong, who had been to this restaurant. Both of them gave a rating of 3 on a scale between 1 and 5 for the service in this restaurant. Given these ratings, it is a bit difficult for Wen to decide if she should pay a visit to “Haidilao.” Fortunately, she has kept a table of Ling and Zhuodong’s ratings for some other restaurants, as well as her own ratings in the past, as shown below.

Restaurant	Ling’s rating	Zhuodong’s rating	Wen’s rating?
Meizhou Dongpo	1	5	2.5
Din Tai Fung	4.5	4	5
...	...	...	...
Haidilao	3	3	?

Who are Wen, Ling, and Zhuodong?

Example 1: Haidilao or not?

Notation

- "Input" Variables: $\boldsymbol{u} = (u_1, u_2) \in \mathcal{U} \subseteq \mathbb{R}^2$ (rating of Ling, rating of Zhuodong)
- "Output" Variables: $v \in \mathcal{V} \subseteq \mathbb{R}$ (rating of Wen)

Data

- Training Set/Dataset: $\left\{\boldsymbol{u}^{(i)}, v^{(i)}\right\}_{i=1}^{N}$
- For example, $\left(\boldsymbol{u}_1^{(i)}, v_1^{(i)}\right) = \left(u_1^{(1)}, u_2^{(1)}, v_1^{(1)}\right)$
  $= ($rating of Ling, rating of Zhuodong, rating of Wen$)$

Goal

Learn a function $h : \mathcal{U} \to \mathcal{V}$:D

Example 1: Haidilao or not?

Notes

1. This function h is usually called a *hypothesis* or *decision function*.
2. Machine learning tasks of these types are called *supervised learning*.
3. If $\mathcal{V}$ is continuous $\Rightarrow$ *regression*. If $\mathcal{V}$ is discrete $\Rightarrow$ *classification*.

Linear function

One simple idea is to approximate v by

$$\begin{aligned} h_{\boldsymbol{\theta}}(\boldsymbol{u}) &= \theta_0 + \theta_1 u_1 + \cdots + \theta_n u_n \\ &= \sum_{i=0}^{n} \theta_i u_i \quad = \quad \boldsymbol{\theta}^\top \boldsymbol{u} \end{aligned}$$

where we set $u_0 = 1$. Note that in our example $n = 2$.

Example 1: Haidilao or not?

Least square regression

To find the "best" parameters $\boldsymbol{\theta}$, least square regression optimizes

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^{n+1}} \left\{ f(\theta) := \sum_{i=1}^{N} \left(h_{\boldsymbol{\theta}}(\boldsymbol{u}^{(i)}) - v^{(i)} \right)^2 \right\}.$$

Here, $\epsilon^{(i)} = h_{\boldsymbol{\theta}}(\boldsymbol{u}^{(i)}) - v^{(i)} = \boldsymbol{\theta}^\top \boldsymbol{u}^{(i)} - v^{(i)}$ is the error associated with ith approximation.

Statistical properties

If $\epsilon^{(i)} \sim N(0, \sigma^2)$ and i.i.d., then the likelihood function w.r.t. $\boldsymbol{\theta}$ is

$$L(\boldsymbol{\theta}) := \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\boldsymbol{\theta}^\top \boldsymbol{u}^{(i)} - v^{(i)})^2}{2\sigma^2} \right)$$

Claim: The solution of least square regression maximizes the above likelihood function. How?

Example 1: Haidilao or not?

Statistical properties (cont.)

Note that maximizing $L(\boldsymbol{\theta})$ is same as maximizing $\log L(\boldsymbol{\theta})$. So, we have

$$\begin{aligned}\log L(\boldsymbol{\theta}) &= \log\left(\prod_{i=1}^{N}\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(\boldsymbol{\theta}^\top\boldsymbol{u}^{(i)}-v^{(i)})^2}{2\sigma^2}\right)\right)\\ &= \sum_{i=1}^{N}\log\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(\boldsymbol{\theta}^\top\boldsymbol{u}^{(i)}-v^{(i)})^2}{2\sigma^2}\right)\right)\\ &= N\log\left(\frac{1}{\sqrt{2\pi}\sigma}\right)-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(\boldsymbol{\theta}^\top\boldsymbol{u}^{(i)}-v^{(i)})^2.\end{aligned}$$

Therefore, maximizing $\log L(\boldsymbol{\theta})$ = least square regression.

Warning: The above claim is true only when $\epsilon^{(i)} \sim N(0, \sigma^2)$ and i.i.d.

Overview

1. Course Basics

2. What is Optimization?

3. Course Intended Learning Outcomes

4. Some Machine Learning Models $\Rightarrow$ Optimization

4.1 Least Squares Regression

4.2 Logistics Regression

4.3 More

Example 2: Haidilao or not?

We don’t need rating!

Suppose that Wen only cares about whether she will like the restaurant “Haidilao” or not!

Restaurant	Ling’s rating	Zhuodong’s rating	Wen’s likes?
Meizhou Dongpo	1	5	No
Din Tai Fung	4.5	4	Yes
$\cdots$	$\cdots$	$\cdots$	$\cdots$
Haidilao	3	3	?

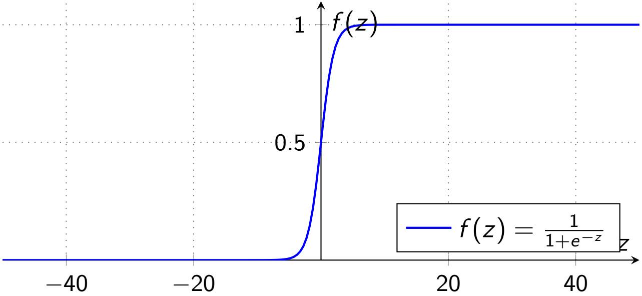
- “Input” Variables: $\boldsymbol{u} = (u_1, u_2) \in \mathcal{U} \subseteq \mathbb{R}^2$ (rating of Ling, rating of Zhuodong)
- “Output” Variables: $v \in \{0, 1\}$ (Wen likes or not)
- Training Set/Dataset: $\left\{\boldsymbol{u}^{(i)}, v^{(i)}\right\}_{i=1}^{N}$
- $v^{(i)} = 1$ means that Yuqi likes the i-*th* restaurant and $v^{(i)} = 0$ means that she dislikes the restaurant.

Example 2: Haidilao or not?

Logistic Regression

Recall that we aim to learn a function $h : \mathcal{U} \to \mathcal{V}$. However, the linear function $h_{\boldsymbol{\theta}}(\boldsymbol{u}) = \boldsymbol{\theta}^\top \boldsymbol{u}$ will lead to values that can be arbitrarily large or small. To force that we map to a value between 0 and 1, we make use of *logistic function* $g(z) = 1/(1 + \exp(-z))$, and we have

$$h_{\boldsymbol{\theta}}(\boldsymbol{u}) = g(\boldsymbol{\theta}^\top \boldsymbol{u}) = \frac{1}{1 + exp(-\boldsymbol{\theta}^\top \boldsymbol{u})}.$$



Example 2: Haidilao or not?

Statistical properties

We assume that $v^{(i)}$, $i = 1, \ldots, N$, are independent Bernoulli random variables with success probability (or mean) of $h_{\boldsymbol{\theta}}(\boldsymbol{u}^{(i)})$. Thus, their probability mass functions are given by

$$p\left(v^{(i)}|\boldsymbol{u}^{(i)};\boldsymbol{\theta}\right) = \ ???$$

Question

The probability mass function $p\left(v^{(i)}|\boldsymbol{u}^{(i)};\boldsymbol{\theta}\right)$ should be

- A : 0
- B : $\cos(\pi \cdot \theta)$
- C : $[h_{\boldsymbol{\theta}}(\boldsymbol{u}^{(i)})]^{v^{(i)}}[1 - h_{\boldsymbol{\theta}}(\boldsymbol{u}^{(i)})]^{1-v^{(i)}}, \quad v^{(i)} \in \{0, 1\}$
- D : All of the above

Answer: C

Example 2: Haidilao or not?

Statistical properties (cont.)

The the associated likelihood function $L(\boldsymbol{\theta})$ is defined as

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} \left\{ [h_{\boldsymbol{\theta}}(\boldsymbol{u}^{(i)})]^{v^{(i)}} [1 - h_{\boldsymbol{\theta}}(\boldsymbol{u}^{(i)})]^{1-v^{(i)}} \right\}.$$

Similar, we consider

$$\begin{aligned}
\log L(\boldsymbol{\theta}) &= \log \left(\prod_{i=1}^{N} \left\{ [h_{\boldsymbol{\theta}}(\boldsymbol{u}^{(i)})]^{v^{(i)}} [1 - h_{\boldsymbol{\theta}}(\boldsymbol{u}^{(i)})]^{1-v^{(i)}} \right\} \right) \\
&= \sum_{i=1}^{N} \left(v^{(i)} \cdot \log[h_{\boldsymbol{\theta}}(\boldsymbol{u}^{(i)})] + (1 - v^{(i)}) \cdot \log[1 - h_{\boldsymbol{\theta}}(\boldsymbol{u}^{(i)})] \right) \\
&= \sum_{i=1}^{N} \left(-\log \left[1 + \exp\left(-\boldsymbol{\theta}^\top \boldsymbol{u}^{(i)} \right) \right] - (1 - v^{(i)}) \cdot \boldsymbol{\theta}^\top \boldsymbol{u}^{(i)} \right).
\end{aligned}$$

Example 2: Haidilao or not?

Logistic regression

To find the "best" parameters $\boldsymbol{\theta}$, logistic regression optimizes

$$\max_{\boldsymbol{\theta}\in\mathbb{R}^{n+1}} \left\{ \sum_{i=1}^{N} -\log\left[1+\exp\left(-\boldsymbol{\theta}^\top \boldsymbol{u}^{(i)}\right)\right] - (1-v^{(i)})\cdot\boldsymbol{\theta}^\top \boldsymbol{u}^{(i)} \right\}.$$

- No explicit solution
- ⇒ optimization algorithms

Haidilao or not?

- Solve for $\boldsymbol{\theta}^\star$
- If $h_{\boldsymbol{\theta}^\star}((1,3,3)) > 0.5 \Rightarrow$ Yuqi will like the restaurant. [recall that $u_0 = 1$]

Overview

1. Course Basics

2. What is Optimization?

3. Course Intended Learning Outcomes

4. Some Machine Learning Models ⇒ Optimization
   - 4.1 Least Squares Regression
   - 4.2 Logistics Regression
   - 4.3 More

Regularization, Lasso, and Ridge Regression

Supervised learning

Many supervised machine learning models, including a few problems discussed before, can be written in the following form:

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \sum_{i=1}^{N} L\left(\boldsymbol{x}^\top \boldsymbol{u}_i, v_i\right) + \lambda \cdot r(\boldsymbol{x})$$

for some $\lambda \geq 0$, where $L(\cdot, \cdot)$ and $r(\cdot)$ are called the *loss* and *regularization* functions.

Examples

- Ordinary least square regression: $\boldsymbol{x} = \boldsymbol{\theta}$, $L(z, v) = (z - v)^2$, $r(\boldsymbol{x}) = 0$
- Support vector machines (SVM): $\boldsymbol{x} = (\boldsymbol{w}, b)$, $L(z, v) = \max\{0, 1 - vz\}$, $r(\boldsymbol{x}) = \|\boldsymbol{w}\|_2^2$
- Ridge regression: $\boldsymbol{x} = \boldsymbol{\theta}$, $L(z, v) = (z - v)^2$, $r(\boldsymbol{x}) = \|\boldsymbol{w}\|_2^2$ [shrinks coeff. for correlated data]
- Lasso regression: $\boldsymbol{x} = \boldsymbol{\theta}$, $L(z, v) = (z - v)^2$, $r(\boldsymbol{x}) = \|\boldsymbol{w}\|_1$ [sparse solution]

Population Risk Min. v.s. Empirical Risk Min.

Empirical Risk Minimization (ERM)

From our historical data, we optimize the average performance via

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^{n+1}} \sum_{i=1}^{N} \left(h_{\boldsymbol{\theta}}(\boldsymbol{u}^{(i)}) - v^{(i)}\right)^2 .$$

Population Risk Minimization (PRM)

Optimizing the true expected performance via

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^{n+1}} \mathbb{E}_{\mathbb{P}}\left[h_{\boldsymbol{\theta}}(\tilde{\boldsymbol{u}}) - \tilde{v}\right]^2, \qquad \text{where } (\tilde{\boldsymbol{u}}, \tilde{v}) \sim \mathbb{P}.$$

- These are examples of ERM and PRM. One could consider alternative models. (e.g., with regularization).
- For large N, ERM $\approx$ PRM. But how to solve ERM with large N?

Summary

- Many machine learning models are formulated as optimization problems.
- These problems exhibit specific structures and with unique computational challenges.
- This course covers:
    1. Basic optimization
    2. Classical algorithms
    3. Scalable algorithms for structured problems

To be continued...