CS5284 : Graph Machine Learning

Administrative (Week 7)

Semester 1 2024/25

Xavier Bresson

https://x.com/xbresson

Department of Computer Science
National University of Singapore (NUS)

Midterm

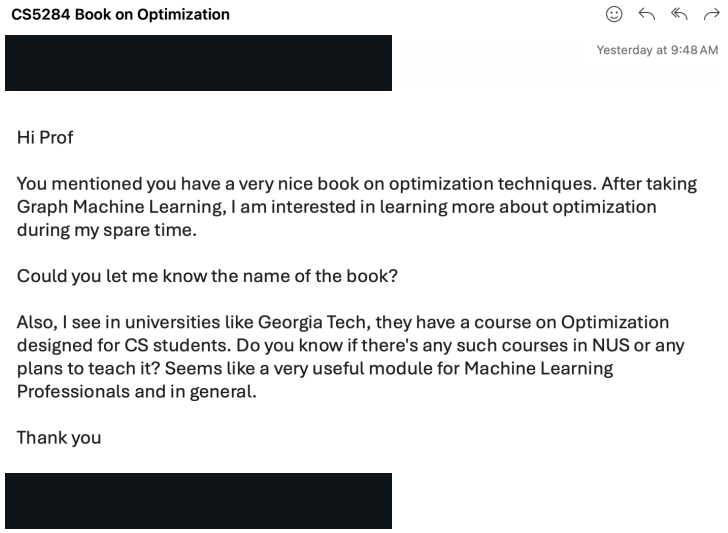
Midterm

- Grade release : Next Monday Oct 7^{th} noon (expected).
- Next week's tutorial : TAs will present and discuss the midterm solution.
- We will *not* share the electronic versions of the notebook solution and the MRQ.
- Additionally, you are not allowed to take photos of the solution during the tutorial.

Student question

Math books for machine learning

- Linear algebra for ML : Linear Algebra by Lieven Vandenberghe and Stephen Boyd: https://web.stanford.edu/~boyd/vmls/vmls.pdf
- Optimization for ML : Convex Optimization by Lieven Vandenberghe and Stephen Boyd: https://web.stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf
- Theory of ML : Elements of Statistical Learning, Data Mining, Inference, and Prediction by Jerome Friedman, Robert Tibshirani, Trevor Hastie, https://www.sas.upenn.edu/~fdiebold/NoHesitations/BookAdvanced.pdf


CS5284 Book on Optimization

[redacted]

Yesterday at 9:48 AM

Hi Prof

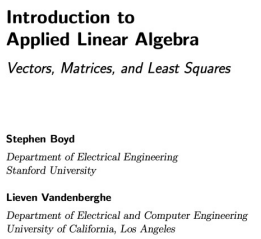
You mentioned you have a very nice book on optimization techniques. After taking Graph Machine Learning, I am interested in learning more about optimization during my spare time.

Could you let me know the name of the book?

Also, I see in universities like Georgia Tech, they have a course on Optimization designed for CS students. Do you know if there's any such courses in NUS or any plans to teach it? Seems like a very useful module for Machine Learning Professionals and in general.

Thank you

[redacted]


Introduction to Applied Linear Algebra

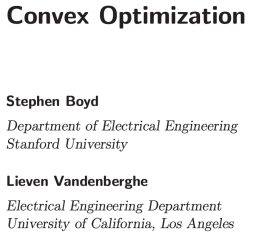
Vectors, Matrices, and Least Squares

Stephen Boyd

Department of Electrical Engineering
Stanford University

Lieven Vandenberghe

Department of Electrical and Computer Engineering
University of California, Los Angeles

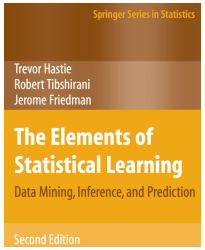

Convex Optimization

Stephen Boyd

Department of Electrical Engineering
Stanford University

Lieven Vandenberghe

Electrical Engineering Department
University of California, Los Angeles



In-class questions

In-lecture question [Answer]

- Assuming the data distribution is not centered at the origin and we do not center the data points, what impact does this absence of centering have on the PCA results?
- In Slack #lectures
  - Identify the question and Reply in thread with a short response
- Answer : Not centering the data in PCA will distort the result. It will still attempt to capture the directions of maximum variance in the data but the offset from the origin will not capture meaningful relationships between variables.

In-lecture question [Answer]

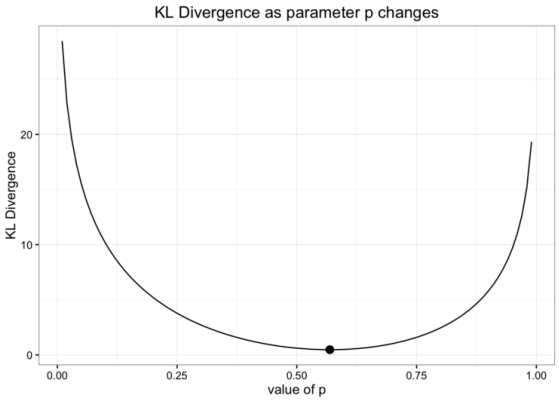
- What is the simplest technique for dimensionality reduction given a set of distances between data points, where $D_{ij} = \text{dist}(x_i, x_j)$? Hint: Think about PCA technique.
- In Slack #lectures
  - Identify the question and Reply in thread with a short response
- Answer : When the data points x_i are available, the simplest dimensionality reduction technique is PCA, performing EVD on the covariance matrix $C=X^TX$.

  For cases where only the distance matrix D is available, a direct extension of PCA is to apply EVD to D, as D can be interpreted as a Gram matrix $D_{ij} = x_i.x_j = C_{ij}$. To generalize further, a positive-definite function of D can be considered to form a standard kernel matrix, to which EVD can also be applied. This approach leads to a natural extension of PCA known as kernel PCA[1].

[1] Scholkopf, Smola, Muller, Nonlinear Component Analysis as a Kernel Eigenvalue Problem, 1998 (11,000 citations)

In-lecture question [Answer]

- The Kullback-Leibler divergence KL(p,q) is a continuous and convex function with respect to p. Could we potentially use a faster optimization technique to solve Y instead of relying on standard gradient descent? Additionally, is the primary reason for using PCA as initialization to accelerate the optimization process?
- In Slack #lectures
  - Identify the question and Reply in thread with a short response
- Answer : Actually, the minimization problem is non-convex with respect to Y due to the non-linear relationship in q(Y), which necessitates the use of gradient descent (GD). Since the problem is non-convex, GD is not guaranteed to find a global minimum. In practice, PCA is commonly used as an initialization strategy to help avoid poor local minima and improve convergence.



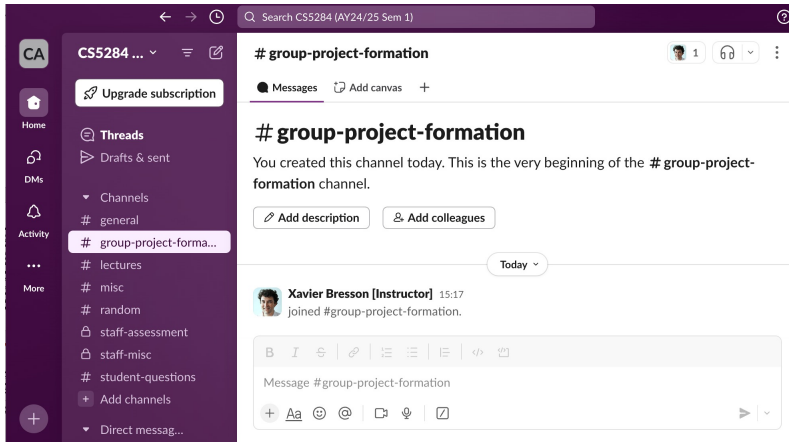
Project

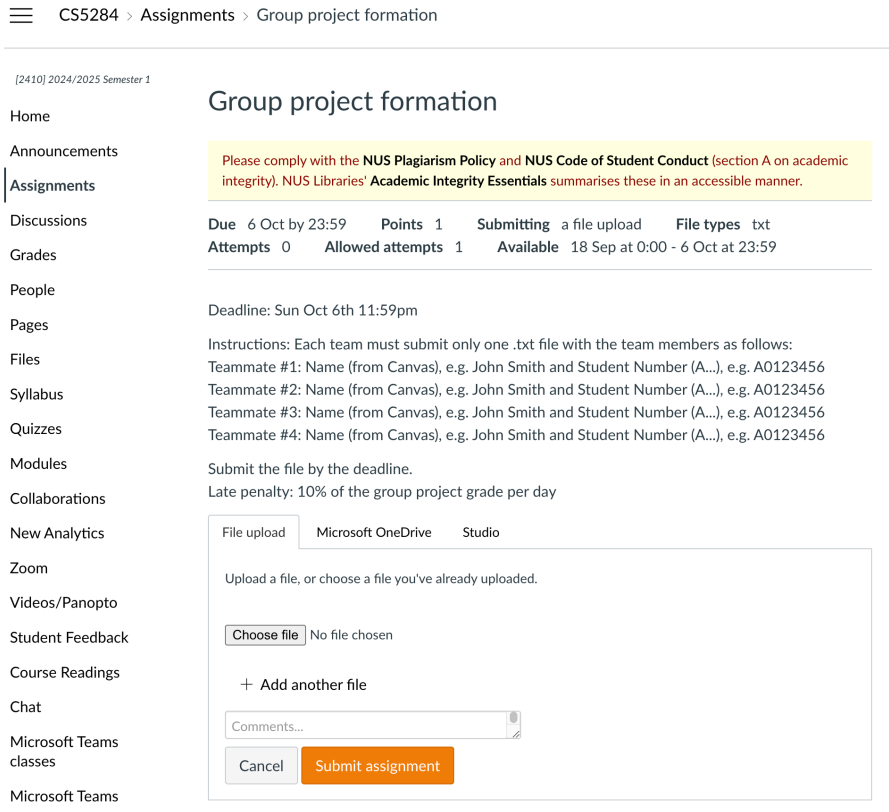
Group project formation

- Your team
  - You can select your teammates (each group may have 2-5 members), but you will have to make an agreement/contract to distribute and contribute equally to the tasks (for minimizing future conflict).
  - Contribute equally does not mean contributing like an expert in coding, maths, engineering, presentation, etc (some people are beginners) -- it means that the effort and attitude to make the project successful must be at the same level than others.
  - Each group, i.e. each teammate, will receive the same grade.
    - Choose your group wisely!
    - Not only people you know, i.e. your friends, but people willing to work on the project (good friend $\neq$ good teammate).
  - Submission deadline : Sun Oct 6th 11:59pm (Week 7)
  - Penalty : 10% of the group grade per late day

Group project formation

- Canvas>Assignments>Group project formation
- Looking for teammates : Use e.g. slack channel “group-project-formation”, accessible as follows:
  - Click "Add channels">Browse channels>group-project-formation>join





Group project

- Group project
  - Group project counts for 35% of the total course mark.
  - Deadline : Fri Nov 22nd 11:59pm (Week 14)
  - Penalty : 10% of the group grade per late day
  - Deliveries : Python notebook, report, presentation slides and video recording

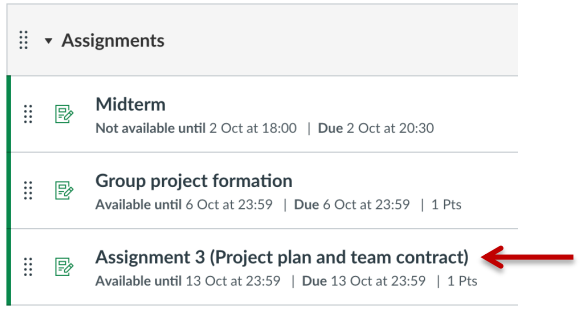
- Do not start working on the project 1-2 weeks before the deadline!

- If any question about the group project, please contact your group TA, who will be allocated to your group ID after the group formation submission.
  - Note that Group TA $\neq$ Tutorial TA

Project plan and team contract

- Project plan and team contract
  - Write a clear and concise one-page description of the project.
    - Strictly one-page limit. If > one-page limit then project plan grade will be a 0.
    - Exception : References (if any) can be provided as extra pages.
    - Any style and format can be used, e.g. single/double columns, etc.
  - Possible template of project plan
    - Project motivation, description, proposed solution, project milestones.
  - Team contract/agreement
    - Add an additional page which describes the tasks assigned to each team member.
    - Each teammate must contribute equally to the project.
    - Each teammate must sign the contract.
    - If the signed contract is not submitted with the project plan, then the project grade will be a 0.

Project plan and team contract



- Project plan and team contract
  - Submit the project plan and team contract in Canvas:
    - Upload one .pdf file into Canvas > Assignment > Assignment 3 (Project plan and team contract)
    - Submit one single project plan and team contract per team.
    - File name must be "project_plan_contract_groupIDXX.pdf", (for example project_plan_contract_groupID31.pdf) (see later slide for groupIDXX).
    - Deadline : Sun Oct 13th 11:59pm (Week 8)
    - Penalty : 10% of the group grade per late day
    - Project plan and contract do not bring any point to the project grade
    - The TA allocated to your group will provide a feedback in Canvas by Fri Oct 18th (Week 9)

Group ID

- After the deadline of group formation, your team will be assigned an ID number, i.e. IDXX.
  - For example, team ID27: John Smith and Joe Doe
  - Your team ID number will be available at Canvas > Home > W08 > list_ID_project_groups.pdf
    - Please, check and use your group ID for any future communication and submission.

TA allocated to Groups

- TA allocated to your group will be as follows :
  - Groups ID XX to XX(included) : Liu Nian
  - Groups ID XX to XX(included) : Fu Guoji
  - Groups ID XX to XX(included) : Wang Jiaming
- Reminder : Group TA ≠ Tutorial TA
- If you have any question about the project, please ask the TA in charge of your group.

Project philosophy

- This project focuses on
  - The understanding of the fundamental concepts of graph machine learning techniques,
  - The practical skills required to develop a data analysis project.
- It is not about learning to use GitHub codes.
- It is not about winning a Kaggle competition.
- It is not about three lines of Keras' code to run machine learning techniques.
- It is not about running long experiments with the best possible GPUs.
  - Google Colab, Google Cloud, and your computer/laptop are enough.
- It is not about getting 90% of accuracy.
- It is about how to design from scratch, debug, understand and train learning algorithms.
- It is about to understand why it works and why it does not.

Project philosophy

- This project focuses on :
  - Theoretical knowledge received in this module.
  - Practical skills with data acquisition, exploration, exploitation, analysis.
  - Teamwork with management of tasks.
  - Concise and clear communication with written report and oral presentation.

Project goals

- Project goals
  - Download or prepare a dataset(s)
    - This dataset(s) can be novel or not.
  - Implement graph machine learning techniques on this dataset(s)
    - Use simple model(s) as baseline.
  - Propose improvement(s)
    - Motivation, description, equation, implementation, result, discussion.
  - Demonstrate initiatives
    - Develop own scrapper, dynamic visualization, discover new data insights, etc.
  - Deliveries
    - Python notebook for code demo
    - Project report (it can be merged with the notebook)
    - Video presentation and slides.

Dataset(s)

- Dataset(s) can be collected from an existing repository
  - UCI : https://archive.ics.uci.edu/datasets
  - Kaggle : https://www.kaggle.com/datasets
  - Paperswithcode : https://paperswithcode.com/datasets
  - GitHub : https://github.com/topics/dataset
  - DGL : https://docs.dgl.ai/en/2.2.x/api/python/dgl.data.html
  - PyG : https://pytorch-geometric.readthedocs.io/en/2.5.2/modules/datasets.html

- Dataset(s) can be new, i.e.
  - Scrap using an API, e.g. Twitter API or Meta API
  - Collect data with your hand-crafted scrapper

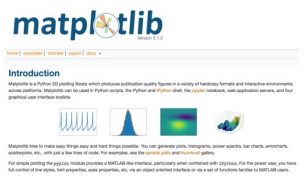
Project development

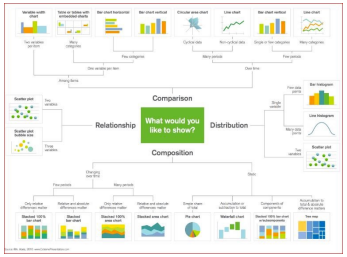
- Step 1: Identify a data analysis problem that ca be solved with graph machine learning.
  - You may use your own field of expertise or your personal interests.
  - The problem is neither too easy nor too difficult!

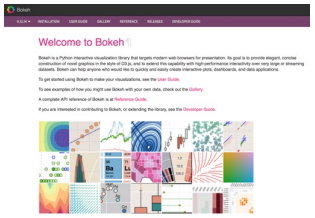
- Step 2: Dataset collection
  - Use existing dataset(s)
  - Develop new dataset(s)

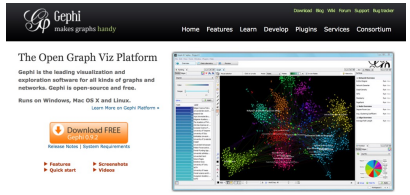
Project development

- Step 3: Data exploration (analyze your data, get insights)
  - Use statistics
  - Use visualization libraries, for example
    - Matplotlib: https://matplotlib.org
    - Bokeh: https://bokeh.pydata.org
    - Graphlab: https://gephi.org







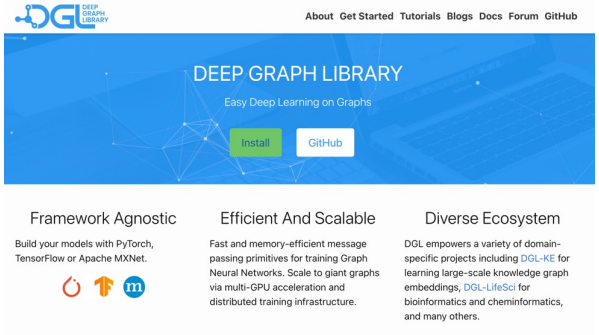


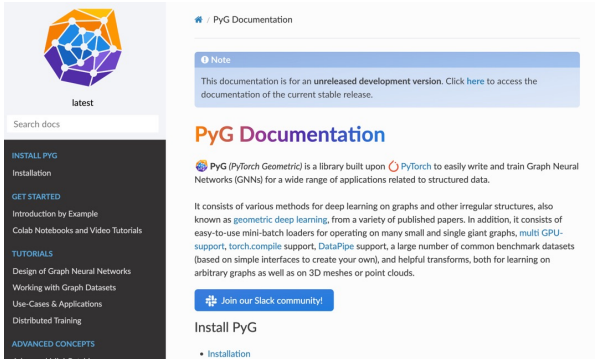
- Step 4: Pre-processing
  - Data cleaning (missing features)
  - Data normalization (unbalanced scaling)
  - Important and consuming step to prepare data as clean as possible for analysis

Project development

- Step 5: Data analysis with graph machine learning
  - Apply machine learning to solve your data problem :
    - Regression, classification, etc
  - Compare different models

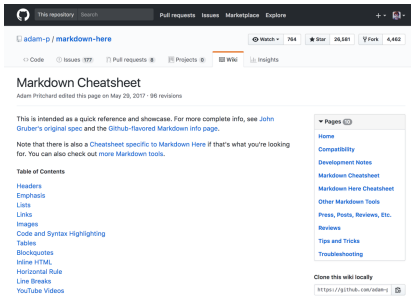
- Step 6: Numerical results
  - Analysis, interpretation, conclusion





Project development

- Step 7: Report
  - Standard approach
    - Word/latex report
  - Modern approach
    - Use Python Notebook and Markdown: https://github.com/adam-p/markdown-here/wiki/Markdown-Cheatsheet
    - Future of scientific reports:
      - Code + description + analysis merged into a single document.
      - Code is reproducible, transferable to a new dataset, can be extended with new ideas.
  - You are free to select the mode of report.



Project development

- Step 8: Video presentation
  - The project presentation must present concisely the project:
    - Project motivation and description, data acquisition, data exploration, pre-processing, proposed deep learning solutions, analysis of results, future development.
  - Each teammate must present her/his contribution to the project.
    - You will receive a project grade 0 if you do not present your contribution.
  - Use slides (one slide is 1-2min).
  - The length of the presentation is maximum 10min.
    - The time is strict, no more than 10min.
      - You will receive a project grade 0 if you video is beyond 11min.
    - Each member has 3-4min if your group size is 3, and 2min/member for a group size of 5.
  - Convince us you understood what you did !

Team communication

- Conflicts arise from few and lack of communication between teammates.
- It is strongly recommended the team meet once a week (even 5 minutes) to discuss the project status, the progress and the challenges faced.

Weekly monitoring

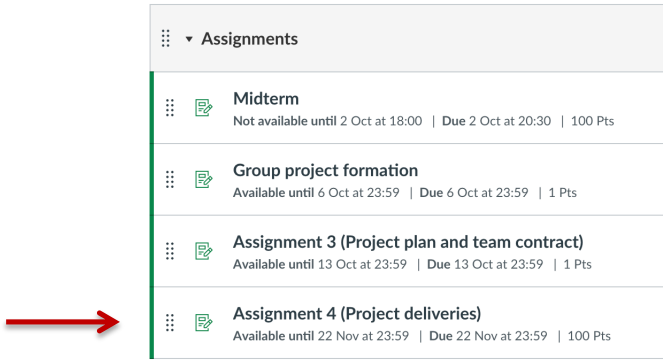
- Work weekly on the project.
  - Do not wait for the last weeks to start working on the project.
- We will monitor the project progress :
  - Each team must send a short update (s.a. one paragraph) of the week's progress to the TA allocated to your group. Deadline : Every Friday by 6pm from week 8 to week 13.
  - There will be at least one Zoom meeting scheduled between you and the TA :
    - Ideally by Week 11, but multiple Zoom meetings (i.e. before/after Week 11) can be organized, as needed.
    - You are responsible for scheduling one Zoom meeting with the TA, e.g. in Week 11.
  - Weekly updates and one zoom meeting count for 10pts of the project grade.

Marking scheme

- Project plan & team contract do not count.
- Weekly updates and one zoom meeting count for 10pts.
- Steps 1-8 count for 65pts.
- Anything that demonstrates initiatives will receive up to 25pts additional points.

Project submission

- Submit notebook, report, presentation slides and video recording :
  - Canvas > Assignment > Assignment 4 (Project deliveries)
  - Create a .zip file with your notebook, report, presentation slides and video recording.
    - Use the format “project_groupID.zip” (for example project_group31.zip).
  - Note that the maximum upload file size is 500MB.
  - Contact your allocated TA if your submission is larger than 500MB.



Deliveries and deadlines

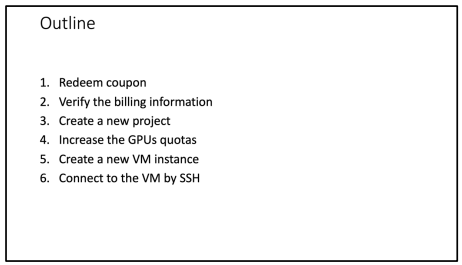
- Week 7 : Group formation, deadline: Sun Oct 6th 11:59pm
- Week 8 : Project proposal and team contract, deadline: Sun Oct 13th 11:59pm
- Week 14 :
  - A working/reproducible python notebook
  - Project report (it can be merged with the notebook)
  - Presentation slides and video recording (with e.g. Zoom)
  - Deadline : Fri Nov 22nd 11:59pm
- Penalty : 10% of the group grade per late day

GPU

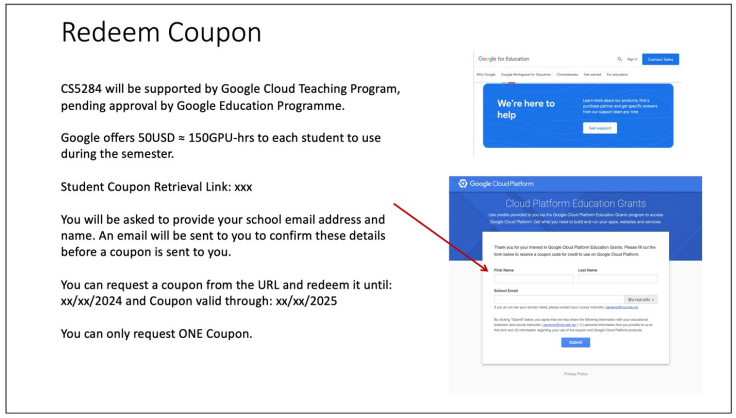
- The project should not require extensive experimentation with top-tier GPUs.
- Google Colab (free GPU with limitations) and your computer should suffice.
- Additionally, the course benefits from support through the Google Cloud Teaching Program.
- If your project requires GPU resources, use the Google Cloud platform and 150 hrs of free GPU.
- Refer to the lecture "admin_week07_google_cloud_gpu.pdf" for instructions on setting up GPUs.


Google Cloud Guide

Mario Michelessa, Xavier Bresson AY2023/24


Outline

1. Redeem coupon
2. Verify the billing information
3. Create a new project
4. Increase the GPUs quotas
5. Create a new VM instance
6. Connect to the VM by SSH


Redeem Coupon

CS5284 will be supported by Google Cloud Teaching Program, pending approval by Google Education Programme.

Google offers 50USD ≈ 150GPU-hrs to each student to use during the semester.

Student Coupon Retrieval Link: xxx

You will be asked to provide your school email address and name. An email will be sent to you to confirm these details before a coupon is sent to you.

You can request a coupon from the URL and redeem it until: xx/xx/2024 and Coupon valid through: xx/xx/2025

You can only request ONE Coupon.

Teaching assistants

- Teaching assistants are available to support the development of your project as best as possible.
- When the group ID are announced, then
  - Ask the TA in charge of your group for any questions.
  - Do not hesitate to communicate with them to clarify anything.

Questions?