University of Utah Spring 2024

MATH 2270-002 PSet 11 Specification

Instructor: Alp Uzman

Subject to Change; Last Updated: 2024-04-18 23:32:01-06:00

1 Background

This problem set focuses on Gerschgorin theory and PageRank. Gerschgorin theory is based on the following heuristic: "For a square matrix A, its diagonal entries approximate its eigenvalues. The smaller the off-diagonal entries, the better the approximation."

On the other hand, PageRank is a simple but then-revolutionary algorithm that lies (or at least, used to lie) at the heart of Google's search engine. Developed by Sergey Brin and Lawrence Page, and patented by Lawrence Page, PageRank offers a unique approach to ranking web pages by assessing their relative importance within the vast network of the internet.

PageRank Here is a summary of the theory and terminology for Pagerank. Let $\mathcal{N}$ be a network of webpages; mathematically this means that $\mathcal{N}$ is a directed graph with finitely many nodes. Say n is the number of webpages (aka vertices or nodes).

For a webpage i, denote by In(i) the collection of all those webpages that link to i and by Out(i) the collection of all those webpages that i links to. Then the **hyperlink matrix** $H = H(\mathcal{N})$ of the network $\mathcal{N}$ is defined by

$$H_{ij} = \begin{cases} \dfrac{1}{\#\,\mathrm{Out}(j)} & \text{, if } j \in \mathrm{In}(i) \\ 0 & \text{, otherwise} \end{cases} .$$

Here $\#\,\mathrm{Out}(j)$ is the number of links the webpage j has to other webpages. Then the **basic PageRank vector** $p_{basic} = p_{basic}(\mathcal{N}) \in \mathbb{R}^n$ of the network N is by definition a probability vector that is also a vector fixed by H (aka an eigenvector associated to the eigenvalue 1):

$$\boxed{H\,p_{basic} = p_{basic}}\,.$$

Since the hyperlink matrix H in practice is very large[1], instead of solving this system (say using row-reduction), one wants to be able to use the power method, which entails starting with an arbitrary probability vector p(0) and then setting up the linear dynamical system (with m as the discrete time)

$$p(m+1) = H\,p(m).$$

For large m, the hope is that p(m) approximates well the PageRank vector p. In order for this to work no matter what the structure of the network $\mathcal{N}$ is, one modifies the basic PageRank equation, ultimately leading to the patented version of PageRank. For the modification, one needs three more items:

1. A preprocessing scheme $\mathcal{S}$ for dangling nodes,
2. A number $0 \leq \alpha \leq 1$ as the **damping factor**,

[1] According to Siteefy, there are more than 190 million active (recently updated) webpages as of late February this year; thus the hyperlink matrix is very large. On the other hand many webpages do not have links between them, so many entries of H are zero, aka H is a sparse matrix.

3. A probability vector $B \in \mathbb{R}^n$ as the **bias (or personalization) vector**.

The **preprocessing scheme** $\mathcal{S}$ is applied to the hyperlink matrix H to transform it into a column-stochastic matrix $\mathcal{S}(H)$. Note that H need not be a column-stochastic matrix in general: if webpage j has no links to any other webpage, the jth column of H will be all zeros. While there are many prepreprocessing schemes for dangling nodes[2], let's assume that we will always use the "DISPERSE" preprocessing scheme that works like so: add a link from any dangling node to any other node (and do nothing if H is already column-stochastic). In terms of the hyperlink matrix, this corresponds to replacing any zero column with the vector

$$\frac{1}{n}\mathbb{1}, \qquad \mathbb{1} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}.$$

Recall that in problem set 8 too we had encountered the vector of ones $\mathbb{1} \in \mathbb{R}^n$.

With these three additional items, we may now define the **Google matrix** $G = G(\mathcal{N}; \mathcal{S}, \alpha, B)$ as the following column-stochastic matrix:

$$\boxed{G = G(\mathcal{N}; \mathcal{S}, \alpha, B) = \alpha \mathcal{S}(H(\mathcal{N})) + (1-\alpha) B \mathbb{1}^T}.$$

Note that the Google matrix is a linear interpolation between the matrix $\mathcal{S}(H)$ encoding the network structure and the rank one matrix $B\mathbb{1}^T$ that encodes specific "personalized" behavior. The larger the

[2]Two other important preprocessing schemes are "PRUNE", which involves simply pruning all dangling nodes as well as all dangling nodes that might have been created by pruning (this changes the dimensions of the hyperlink matrix), and "ABSORB", which entails adding a link from any dangling node to itself. Note that in general it might be beneficial to consider the preprocessing scheme $\mathcal{S}$ to be changing the network itself, as opposed to its associated hyperlink matrix.

damping factor α is, the more the network structure is emphasized, and because of this $(1-\alpha)$ is often called the **teleportation parameter**.

With this definition, we reach the true version of the PageRank vector. Whenever α is strictly less than 1 and each entry of B is positive, the **PageRank vector** $\mathsf{p} = \mathsf{p}(\mathcal{N}; \mathcal{S}, \alpha, \mathsf{B}) \in \mathbb{R}^{\mathsf{n}}$ is by definition the unique probability vector fixed by the Google matrix $\mathsf{G} = \mathsf{G}(\mathcal{N}; \mathcal{S}, \alpha, \mathsf{B})$:

$$\boxed{\mathsf{G}\,\mathsf{p} = \mathsf{p}}\,.$$

Suffice it to say, the modifications above are done so that indeed the eigenequation of the Google matrix G for the eigenvalue 1 has a unique solution that is a probability vector[3].

The entries of the PageRank vector p, when reordered from the largest to the smallest, rank the webpages in the network in question, and signify in which order the webpages ought to be presented to a user of a search engine. See the PageRank page under the Pages tab on Canvas for further materials. You may also find this account of stochastic matrices useful.

2 What to Submit

Submit your detailed solutions to each of the problems below. Though they may seem long, the additional text is meant to guide you by providing further context.

When documenting your solutions, be thorough. Your goal is not just to find the answer, but to create a clear, logical pathway to it that you or anyone else could follow in the future. It is likely that the textbook has the answers to some problems that are similar to some

[3]The linear algebraic theory that guarantees this ultimately is Perron-Frobenius theory. The conditions of $\alpha < 1$ and $\mathsf{B}_{\mathsf{i}} > 0$ are sufficient but not necessary for the uniqueness of the PageRank vector.

problems in this problem set; without further notice you may refer to these answers and reverse engineer them.

Make sure that each solution is properly enumerated and organized. Start the solution to each problem on a new page, and consider using headings or subheadings to structure your response clearly. This will not only aid in your thought process but also ensure that no part of your solution is overlooked during grading.

Some of the problems have "challenge" parts; you may attempt these parts if you feel comfortable with the rest of the problems. Challenge questions will not be graded, however you are more than welcome to reach out to staff to discuss them!

1. **For this problem, the use of a computational tool such as WolframAlpha is allowed and recommended.** Consider the following matrices:

   I:
$$\begin{pmatrix} 1 & -3/2 \\ 1/2 & -7/6 \end{pmatrix}$$

   II:
$$\begin{pmatrix} -1 & 1 & 1 \\ 2 & 2 & -1 \\ 0 & 3 & -4 \end{pmatrix}$$

   III:
$$\begin{pmatrix} 0 & 4 & 0 & 0 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/8 & -i & 1/8 & 0 & 0 \\ 0 & 0 & 0 & 1/4 & -2i & 1/4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 9/2 & 1/2 \\ 0 & 0 & 0 & 0 & 0 & 1/2 & -9/2 \end{pmatrix}$$

For each of these matrices, perform the following tasks:

(a) Sketch each Gerschgorin disk in the complex plane.

(b) Compute the eigenvalues and sketch the spectrum.

Challenge Let A be an $n \times n$ matrix with real entries. Denote by $r_i(A)$ the ith deleted absolute row sum of A:

$$r_i(A) = \sum_{j \neq i} |a_{ij}|;$$

thus $r_i(A)$ is precisely the radius of the ith Gerschgorin disk of A. Verify that if for $i \neq j$, $|a_{ii} - a_{jj}| \geq r_i(A) + r_j(A)$, then any eigenvalue of A is real.

2. In this problem your job is to study arbitrary networks of webpages, where the number of webpages is small.

   (a) List all networks with at most three websites.

   (b) For each such network, write down the hyperlink matrix.

   (c) Determine the rank of each hyperlink matrix. Can you find a relation between the rank of the hyperlink matrix and the associated network?

   (d) Classify each hyperlink matrix as nondiagonalizable, diagonalizable, orthogonally diagonalizable. Can you find a relation between diagonalizability of the hyperlink matrix and the associated network?

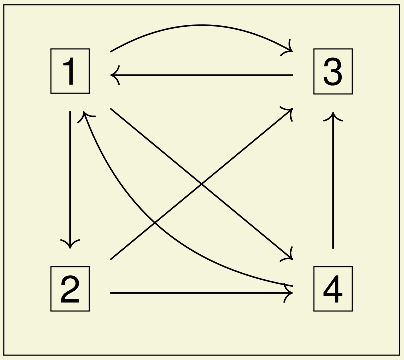
Challenge

(a) For n and arbitrary positive integer, how many distinct networks with exactly n webpages are there?

(b) For n and arbitrary positive integer, how many distinct networks with exactly n webpages are there up to a permutation of the names of the webpages?

3. Consider a network with no dangling nodes and k subnetworks that are disconnected from each other, for k a positive integer. Verify that the geometric multiplicity of the eigenvalue 1 of the associated hyperlink matrix is at least k.

4. Consider the following network consisting of four webpages:

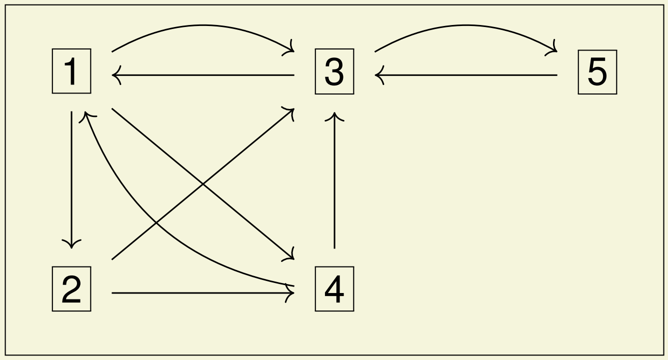


(a) Compute the associated hyperlink matrix.

(b) Set the damping factor α to be 1, so that the bias vector has no effect. Compute the PageRank vector.

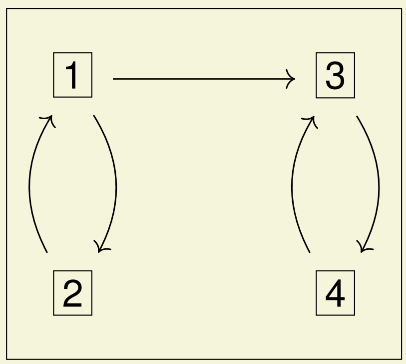
(c) Based on the entries of the PageRank vector, order the webpages according to their importance.

(d) Suppose the owners of webpage 3 are not happy with the ranking of their webpage. With the hopes of improving the ranking of their webpage, the owners thus create a new webpage, webpage 5, that only webpage 3 has a hyperlink to and that also only has a hyperlink to webpage 3, so that now the network looks like this:



Following the above steps for this new network, compute the PageRank vector. Does adding this new webpage improve the ranking of webpage 3?

5. Consider the following network with four webpages:



Perform the following tasks:

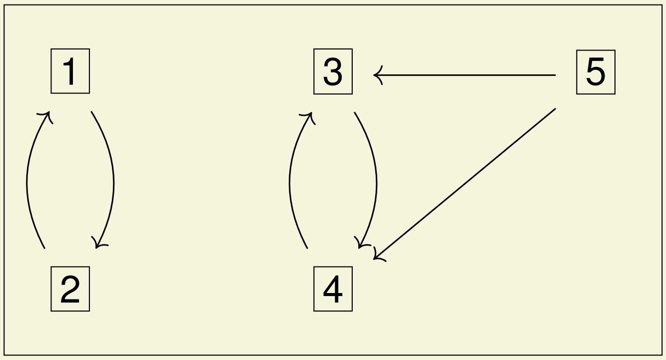
(a) Compute the hyperlink matrix associated to the network.

(b) Take as damping factor and bias vector

$$\alpha = 0.85, \qquad B = (1/2, 1/2, 0, 0),$$

respectively, and compute the PageRank vector. Based on your calculation, which webpage is the highest ranking?

6. Consider the following network consisting of five webpages:



Perform the following tasks:

(a) Compute the hyperlink matrix associated to the network.

(b) Find all eigenvectors of the hyperlink matrix associated to the eigenvalue 1.

(c) Take as bias vector

$$B = (1/5, 1/5, 1/5, 1/5, 1/5),$$

and find all damping factors α such that for the Google matrix G, the geometric multiplicity of the eigenvalue 1 is one.

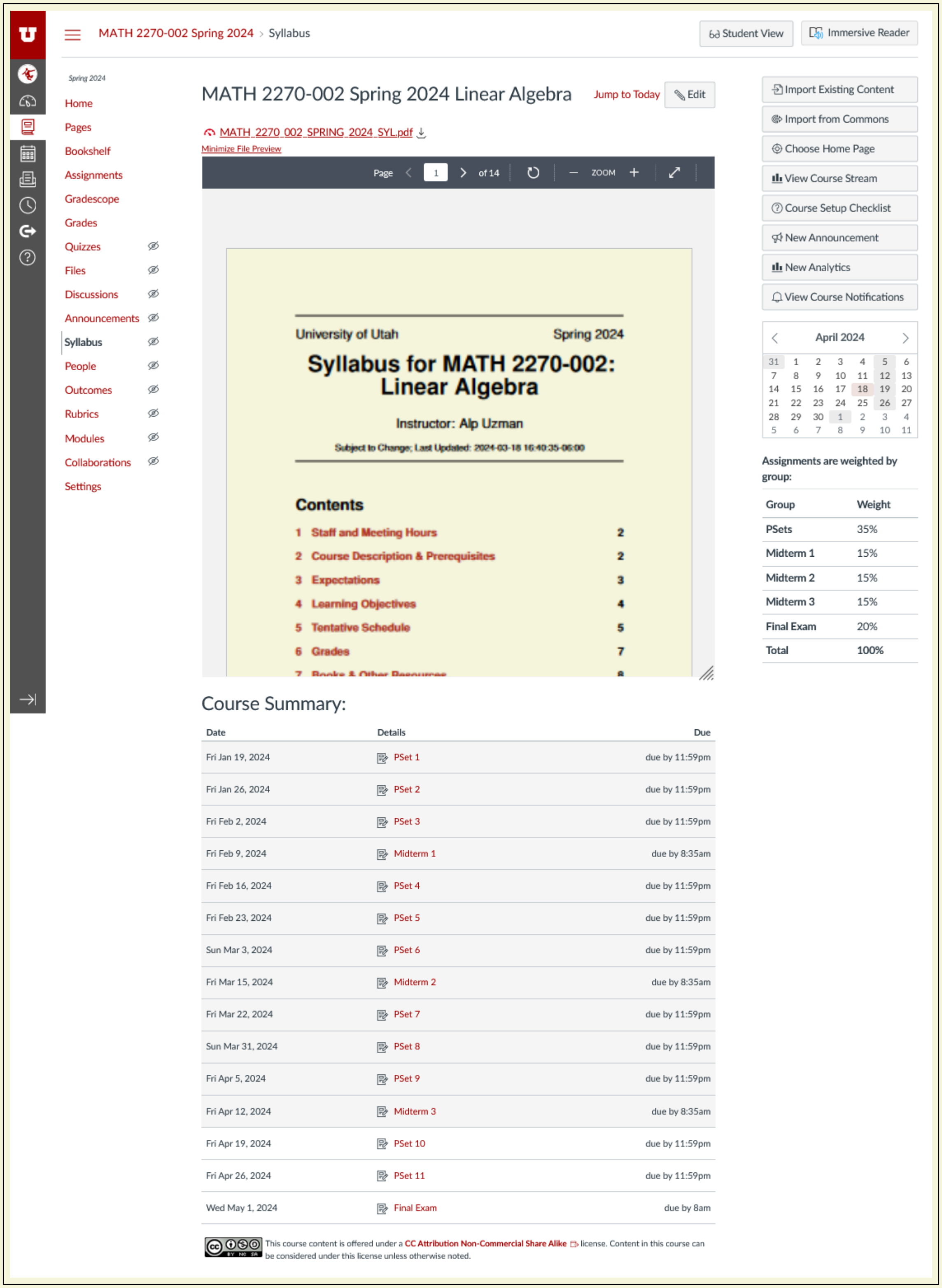
7. **For this problem, the use of a computational tool such as WolframAlpha is allowed and recommended.** In this problem your job is to find the most important webpage for MATH 2270-002 using the PageRank algorithm.

(a) Before you even start modeling the hyperlink structure, which webpage would you guess that is the highest ranking one for MATH 2270-002? That is to say, if you had to show a random student that wants to learn more about MATH 2270-002 exactly one webpage, which webpage would you choose? Your answer could be for instance

- The home/syllabus page, or
- The bookshelf page, where a digital copy of the textbook is stored, or
- The Desmos Grades page, or
- The page of this assignment, or
- ...

(b) Draw your network for MATH 2270-002. Recall that in such a network each webpage ought to be represented as a node, and if webpage j includes a link to webpage i, then this ought to be represented as an arrow from j to i in your network. Note that two webpages are distinct if their URL's are distinct. In your network you may represent all the links for MATH 2270-002 and draw the full network, or, applying a preliminary elimination and getting rid of webpages as "unimportant" you may represent only some of the webpages for MATH 2270-002. **Make sure that in your network there are at least 5 webpages and at least 20 links, and that the webpage of this assignment is definitely represented as a node in your network. For brevity it might also be useful to assign unique integers to each URL you decide to include in your network and record this assignment at the beginning of your solution to this problem.**

Here is an example as to how this could work. Say you decided to include the home/syllabus page in your network:



Group	Weight
PSets	35%
Midterm 1	15%
Midterm 2	15%
Midterm 3	15%
Final Exam	20%
Total	100%

Date	Details	Due
Fri Jan 19, 2024	PSet 1	due by 11:59pm
Fri Jan 26, 2024	PSet 2	due by 11:59pm
Fri Feb 2, 2024	PSet 3	due by 11:59pm
Fri Feb 9, 2024	Midterm 1	due by 8:35am
Fri Feb 16, 2024	PSet 4	due by 11:59pm
Fri Feb 23, 2024	PSet 5	due by 11:59pm
Sun Mar 3, 2024	PSet 6	due by 11:59pm
Fri Mar 15, 2024	Midterm 2	due by 8:35am
Fri Mar 22, 2024	PSet 7	due by 11:59pm
Sun Mar 31, 2024	PSet 8	due by 11:59pm
Fri Apr 5, 2024	PSet 9	due by 11:59pm
Fri Apr 12, 2024	Midterm 3	due by 8:35am
Fri Apr 19, 2024	PSet 10	due by 11:59pm
Fri Apr 26, 2024	PSet 11	due by 11:59pm
Wed May 1, 2024	Final Exam	due by 8am

Looking at this particular webpage, you will choose some number of links that seem to be relevant to MATH 2270-002. For instance on the left there is a menu that is associated to your Canvas account, with tabs "Account", "Dashboard", "Courses", ... These are not intrinsically related to MATH 2270-002, so you may as well ignore them for this assignment. Next, there is a table of tabs that link to other parts of this course; such as "Home", "Bookshelf", "Pages", ... You may decide to represent some or all of these URLs as separate nodes on your network. Then there is the embedded PDF file, which itself has many hyperlinks. You may decide to take into account these hyperlinks, or you may decide to ignore them. Further down, in the "Course Summary" section there are links to each assignment and so on. Ultimately how complex a network you will obtain depends on your implementation choices; for us all that matters is that you have a nice example of a network you can apply the PageRank algorithm to!

(c) Write the hyperlink matrix H associated to your network.

(d) Apply the "DISPERSE" scheme to H to ensure that you have a column-stochastic square matrix.

(e) Choose a damping factor α strictly less than 1 and a bias vector B, and write the associated Google matrix G.

(f) Compute the PageRank vector by either directly solving the eigenequation Gp = p or by employing the power method $p(m) = G^m p(0)$.

(g) Based on the PageRank equation you solved in the previous part, which webpage seems to be highest-ranking for MATH 2270-002, according to the network representation you came up with?

3 Generative AI and Computer Algebra Systems Regulations

This section applies if you decide to use either a generative AI tool or a computer algebra system for this problem set. If not, you may skip this section.

3.1 Providing Logs

If you use such tools, you are required to provide logs of your interactions. Here are some ways to submit them:

- If the tool generates a URL for the interaction (e.g. ChatGPT), list such URLs in the appropriate section of the form you will be filling as part of your submission.
- For tools without direct URL generation, use an appropriate external service to archive the session. An example of such a tool that might work is the Wayback Machine, see the documentation for the "Save Page Now" feature.
- If the tool allows PDF export of the interaction (e.g., Microsoft's Copilot), attach these PDFs to your Gradescope submission.

It is your responsibility to ensure that an archiving method is available for the tool you choose to use. If none of the archiving methods works, then that service is prohibited. If you use a service under the assumption that archiving is available, but it turns out not to be, you must report this in the submission form. Future assignments will be monitored accordingly.

3.2 Chat Guidelines; Prompt Engineering

For chatbot use, follow these guidelines:

- During the chat you may copy and paste parts of this specification document, as well as parts of the textbook or other sources.
- Directly asking the tool for complete problem solutions is prohibited.
- You are required to start any chat with a prompt that ideally would structure the chatbot's responses to you. This practice, known as prompt engineering, is your responsibility, and the staff will evaluate the reasonableness of your attempts. Here is an example of such a "guardrails" prompt that worked reasonably well in this context for ChatGPT as of October 2023:

  > Hello. I am working on a linear algebra problem as part of a university class. My instructor has permitted the use of ChatGPT, but only under specific guidelines to encourage independent critical thinking. Please assist me by asking probing questions, encouraging reflection, and providing general insights about the concepts involved. Do not offer direct hints, strategies, solutions, or step-by-step guidance. I seek to understand the underlying principles and want to develop my own approach to the problem. Your role is to facilitate my learning process without directly leading me to the answer. Thank you!

  Here are some guides regarding prompt engineering:

  - https://platform.openai.com/docs/guides/prompt-engineering
  - https://developers.google.com/machine-learning/resources/prompt-eng
  - https://www.ibm.com/topics/prompt-engineering
  - https://aws.amazon.com/what-is/prompt-engineering/

4 How to Submit

- **Step 1 of 2:** Submit the form at the following URL:

  https://forms.gle/p8k8mQ4awT2LMGaB9.

  You will receive a zero for this assignment if you skip this step, even if you submit your work on Gradescope on time.

- **Step 2 of 2:** Submit your work on Gradescope at the following URL:

  https://www.gradescope.com/courses/694951/assignments/3868166,

  see the Gradescope documentation for instructions.

5 When to Submit

This problem set is due on April 26, 2024 at 11:59 PM.

Late submissions, up to 24 hours after this deadline, will be accepted with a 10% penalty. Submissions more than 24 hours late will not be accepted unless you contact the course staff with a valid excuse before the 24-hour extension expires.