Internet Analytics (COM-308): Final Exam

June 29, 2024

Duration: **2h45**.

Total points: **100**.

Number of pages: **20**.

Allowed documents: **class notes, lab handouts, homeworks, your own code**.

There should in general be enough room below every question for intermediate calculations and your answer. However, you are allowed to use additional sheets of paper; please **write your name on every sheet**, number them, and staple them to this document before handing in.

The use of **mobile phones, tablets, laptop computers**, and other communication devices is **prohibited**.

Last name:
First name:
SCIPER number:
Signature:

Please leave blank.

1	2	3	4	5	6	7	Total
30	12	14	14	10	10	10	100

Question 1: Multiple Choice Questions (30 points)

(30 pts) All questions have a single answer. Check the correct one. Grading:

- Correct answer: +2 points;
- Wrong answer: −1 point;
- No answer or "I don't know": 0 point.

1. Here is a basic implementation of the power method to estimate the PageRank scores using python and numpy:

```
def power_method(probas, G_matrix, S):
    for iteration in range(S):
        probas = probas @ G_matrix
    return probas
```

G_matrix is the google matrix, which has been created using a teleportation probablility $0 < 1 - \theta < 1$.
We note N the number of nodes of the graph, M the number of edges of the graph and S the number of iterations of the algorithm.
What is the smallest upper bound we can give on the number of operations of this function (additions and multiplications)?

■ $O(SN^2)$
□ $O(SNM)$
□ $O(SM)$
□ I don't know

2. We propose the following algorithm to reduce the number of operations without changing the result:

```
def power_method2(probas, graph_transitions, theta, S):
    nb_nodes = len(probas[0])
    for iteration in range(S):
        probas = probas @ graph_transitions * theta
        [.......................................................]
    return probas
```

Which line of code is missing (marked by [...])?

■ `probas += (1 - theta) * np.ones((1, nb_nodes)) / nb_nodes`
□ `probas = (1 - theta) * probas @ graph_transitions`
□ `probas = (1 - theta) * probas + np.ones((1, nb_nodes)) @ graph_transitions`
□ I don't know

3. We use the same notation as 1.
What is the smallest upper bound we can give on the number of operations of the new function `power_method2()`? (we assume that adding 0 or multiplying by 0 does not count as an operation)

□ $O(SN\log(M))$
■ $O(SM)$
□ $O(SN)$
□ I don't know

4. We consider a random graph G whose edges are all independent. G has N nodes split in two communities A and B. 50% of the nodes are in A and 50% are in B. For every pair of nodes in A, the probablility of an edge is $P_{AA} = 0.1$, for every pair in B, $P_{BB} = 0.2$ and for every pair of nodes in different communities, the edge probability is $P_{AB} = 0.001$.
When N goes to infinity, which of the following is more likely to happen?

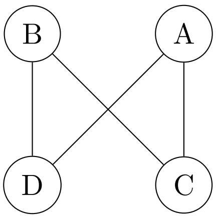
   □ G will have a giant component and some other connected components.
   ■ G will be connected.
   □ G will not have any giant component.
   □ I don't know

5. On the same graph G as the previous question, someone is trying to estimate the size of A and B by performing a random walk on the graph. The random walk starts from a random node. For each node visited, she records the cluster of the node (A or B).
After a very large number of steps, she tries to estimate the size of each cluster by naively comparing the number of nodes of each cluster she visited.

   Which cluster will she estimate to be bigger?

   □ A
   ■ B
   □ It depends on the starting point of the random walk.
   □ I don't know

6. Consider the undirected graph $G = (V, E)$ with vertices $V = \{A, B, C, D\}$ below. What is the number of Bayesian networks compatible with G (same edges as G but all oriented) where the following two conditional independences hold: $A \perp B \mid D$ and $C \perp D \mid A, B$?



   □ 0
   ■ 3
   □ 4
   □ I don't know

7. Here is a simple implementation of the Latent Dirichlet Allocation (LDA) model using RDD operations in spark.

   - **RDD_corpus**: An RDD of documents, where each document is represented as a tuple of (docID, list of terms). Example:

```
RDD_corpus = sc.parallelize([
    (1, ['word1', 'word2', 'word3']),
    (2, ['word2', 'word4', 'word5'])
])
```

   - **num_topics**: Number of topics to be extracted from the documents.
   - **max_iterations**: Number of iterations for the LDA algorithm to converge.

```
# Step 1: Initialize random topic assignments
def initialize_topics(doc):
    docID, terms = doc
    return [(docID, term, randint(0, num_topics - 1)) for term in terms]

topic_assignments = RDD_corpus.flatMap(initialize_topics)

# Step 2: Iterate to optimize the topic assignments
for iteration in range(max_iterations):
    # Calculate topic probabilities and assign new topics
    def sample_topics((docID, term, topic)):
        # Placeholder for actual topic sampling logic
        return (docID, term, randint(0, num_topics - 1))

    topic_assignments = topic_assignments.map(sample_topics)

# Step 3: Aggregate topic assignments into final model structures
A = topic_assignments.map(lambda x: (x[0], x[2])).countByValue()
B = topic_assignments.map(lambda x: (x[1], x[2])).countByValue()
```

What are A and B:

■ doc topic distribution, term topic distribution

□ number of docs, number of topics

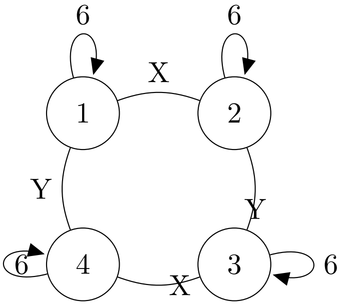
□ number of topics, number of docs

□ I don't know

8. As seen in class, the Louvain method generates a sequence of increasingly smaller, weighted graphs containing self-loops. The general expression for modularity in this case is

$$Q = \frac{1}{2m} \sum_{c_i \in C} \sum_{u,v \in c_i} \left(w_{uv} - \frac{d_u d_v}{2m} \right), \tag{1}$$

where w_{uv} is the weight of the edge between node u and v (w_{uu} for a self-loop), m is the sum of the weights of all links in the network, d_u and d_v are the sum of the weights of the links incident to node u and v respectively.
Which combination of X and Y will cause the graph to have only two communities in the next iteration?



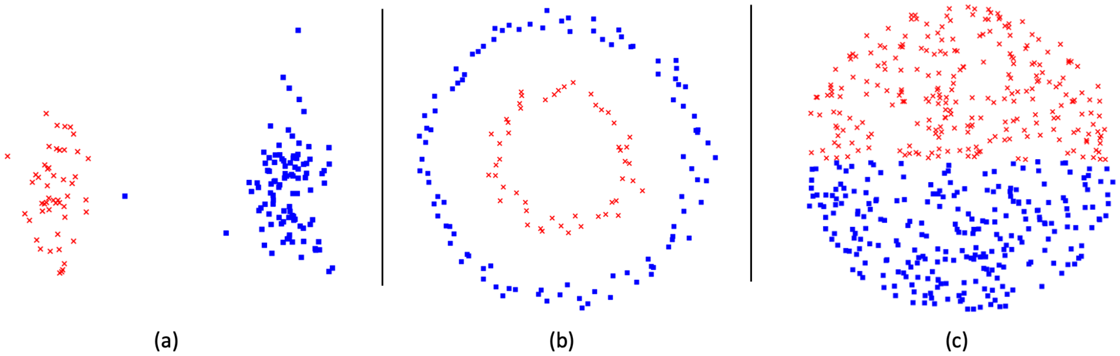
□ $X = 1, Y = 1$

□ $X = 0.5, Y = 1$

■ $X = 1, Y = 2$

□ I don't know

9. We run the clustering methods K-means and Gaussian Mixture Models (GMM) on two-dimensional data points, with both methods using $K = 2$ clusters. In each figure below, points assigned to cluster 1 are represented by red crosses, while points assigned to cluster 2 are represented by blue squares. For each of the following figures, determine which clustering methods could have led to the cluster assignments. Note that more than one method could be correct for each figure, or none of them.
For GMM, a point is assigned to the cluster with maximum posterior.



Which of the following statements is true?

□ Only GMM can produce (c)

□ Only K-means can produce (c)

■ Both GMM and K-means can produce (c)

□ I don't know

10. As in the previous question, which of the following statements is true?

□ K-means can produce (a)

■ GMM can produce (a) and (b)

□ None of the above statements are true

□ I don't know

11. Which of the following statements is true?

□ The generative model of probabilistic Latent Semantic Analysis (pLSA) is:
For each document d in the corpus: (a) sample a topic z from a prior $p(z)$; (b) sample a document d according to $p(d|z)$; (c) sample each word w in d according to $p(w|z)$

□ If PCA is performed on an uncorrelated data set (meaning that the covariance matrix is diagonal), the variances along the different principal directions are all equal

■ If A is a real symmetric matrix with non-negative eigenvalues, then the eigenvalues and singular values of A coincide

□ I don't know

12. We classify a message $M = (w_1, w_2, w_3)$ consisting of three words as ham (G) or spam (B), using a Naive Bayes classifier. Suppose the prior for both classes is positive: $P(B) > 0$, $P(G) > 0$, and the class posterior for message M is positive as well: $P(G|M) > 0$, $P(B|M) > 0$.
Now we create a new, longer message M', by repeating M multiple times: $M' = (w_1, w_2, w_3, w_1, w_2, \ldots, w_3)$. Which of the following inequalities guarantees that for M' sufficiently long, we classify M' as ham (i.e., $P(G|M') > P(B|M')$)?

□ $P(G) > P(B)$

■ $P(M|G) > P(M|B)$

□ $P(G)P(M|G) > P(B)P(M|B)$

□ I don't know

13. Which of the following assertions about the Latent Dirichlet Allocation (LDA) text model is false?

□ Conditional on the topic, the word is independent of the document.

■ Increasing the α parameter of the Dirichlet prior tends to lead to sparser topic distributions for documents.

□ It is possible to generate words for a new document not seen during training.

□ I don't know

14. Consider a SIR epidemic model with one percent of the population initially infected on a population of $N = 10'000$ people. We assume that each person meets on average 120 people every year, and patients stay infected for an average of one week.
Among the following options, what is the smallest percentage of the population that needs to be vaccinated in order to prevent the epidemic to already be a large-scale epidemic? (we assume the vaccine fully prevents transmission)

□ 1/2

■ 2/3

□ 3/4

□ I don't know

15. In relation to Skip-gram, for a specific word-context pair (w, c), which of the following is the partial derivative of the objective function $L_{wc} = \log \sigma_s(c, w)$ with respect to the context vector v_c? ($u_w \in \mathbb{R}^d$ and $v_c \in \mathbb{R}^d$ are the vectors of the word and context and $\sigma_s(c, w) = \frac{e^{u_w^T v_c}}{\sum_{c' \in V} e^{u_w^T v_{c'}}}$)

□ u_w

■ $u_w(1 - \sigma_s(c, w))$

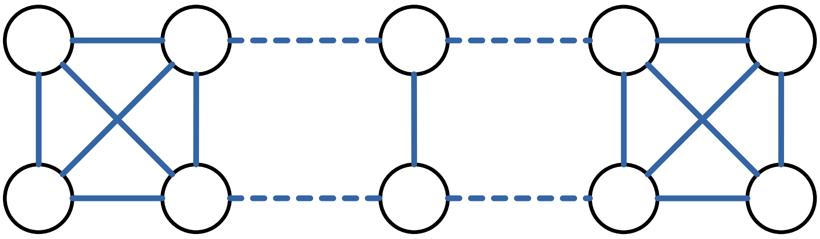
□ $u_w(1 - \frac{1}{\sum_{c' \in V} e^{u_w^T v_{c'}}})$

□ I don't know

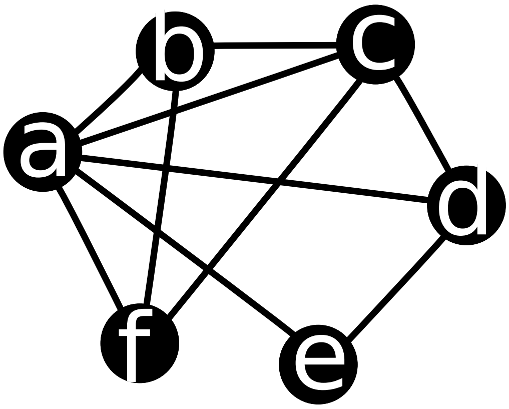
Question 2: Network Structure (12 points)

1. (4 pts) Find the largest possible set of strong links such that the strong triadic closure (STC) property holds for all nodes of the following graph. Mark the strong edges in the figure.

   Solution:



Consider the following graph. We are interested in its transitivity.



2. (2 pts) What is the clustering coefficient c_d of node d?

   Solution:

$$\frac{2}{3} \tag{2}$$

3. (4 pts) What is the weighted average clustering coefficient (transitivity) of the entire graph of the previous question?

   Solution:

$$\frac{5+3+4+2+1+3}{\binom{5}{2}+\binom{3}{2}+\binom{4}{2}+\binom{3}{2}+\binom{2}{2}+\binom{3}{2}} = \frac{3\times 6}{26} = \frac{9}{13} \tag{3}$$

   The six triangles are abc, abf, bcf, acd, ade, and acf.

4. (2 pts) Is this graph strongly clustered relative to an equivalent-density $G(n, p)$?

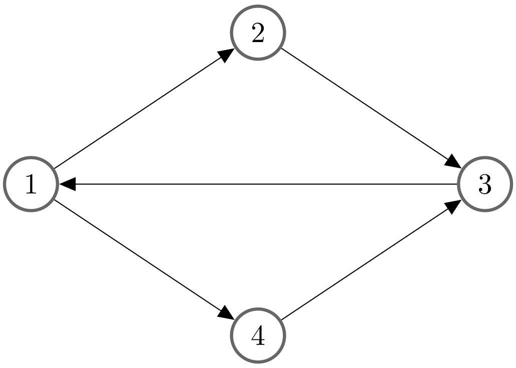
Solution:
We have $n = 6$ and $m = 10$, so the equivalent

$$p = \frac{m}{\binom{n}{2}} = 2/3. \tag{4}$$

The actual clustering coefficient $c(G) = 9/13$ is slightly larger than $2/3$, so we would consider this graph clustered.

Question 3: PageRank (14 points)

We perform PageRank on the following graph.



1. (1 pt) For $\theta = 1$, which nodes are dangling nodes and what are the connected components?

   Solution:
   For $\theta = 1$ (and for any other value of θ), any node can be reached from any node. Consequently, there are no dangling nodes, and only one connected component (the whole graph).

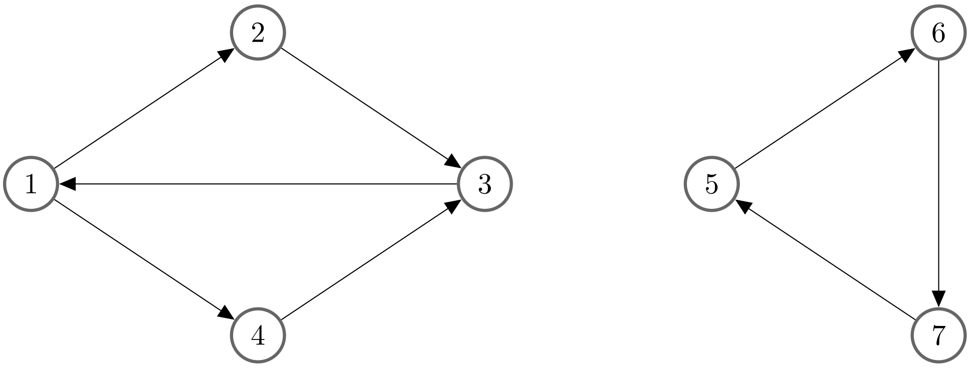
2. (2 pts) For which values of θ is this Markov chain ergodic? When it is not, justify why not.

   Solution:
   To be ergodic, the Markov chain has to be irreductible and aperiodic.
   For $\theta < 1$, the teleportation adds an edge between each pair of nodes (in both directions). This ensures both irreducibility and aperiodicity, so the Markov chain is ergodic.
   For $\theta = 1$, there is no teleportation. In that case, all nodes have a period of 3, which means the Markov chain is not ergodic.

3. (3 pts) Approximate the PageRank scores for $\theta = 0.999$.

   Solution:
   $\theta = 0.999$ means that the probability of teleportation is very small (0.001). In this case, the PageRank score of a node is very close to the frequency at which the random walk visits this node when $\theta = 1$.
   For $\theta = 1$, the frequencies of visit is $1/3$ for nodes 1 and 3, and $1/6$ for nodes 2 and 4. These values are good approximations of the PageRank scores for $\theta = 0.999$.

   Now we add nodes 5, 6, 7 to the graph as shown on the image below.



4. (1 pt) For $\theta = 1$, which nodes are dangling nodes and what are the connected components?

Solution:
Every node has at least one edge going from it, so there are no dangling nodes.
There are 2 connected components : $(1, 2, 3, 4)$ and $(5, 6, 7)$.

5. (4 pts) Approximate the PageRank scores for $\theta = 0$ and for $\theta = 0.999$.

   Solution:
   For $\theta = 0$, the probability of teleportation is 1, so the edges are irrelevant. At each step of the random walk, we walk to any of the nodes with the same probability 1/7. So the PageRank scores of all nodes are 1/7.

   For any $\theta < 1$, the random walk will spend 4/7 of the time in the component $(1, 2, 3, 4)$ and 3/7 of the time in the component $(5, 6, 7)$. In particular, this is the case for $\theta = 0.999$.
   Inside each component, this time will be split almost as if $\theta = 1$, so $\frac{1}{3}, \frac{1}{3}, \frac{1}{3}$ for component $(5, 6, 7)$, and we can reuse the results of question 3. for $(1, 2, 3, 4)$.
   This gives us PageRank scores $PR(1) = PR(3) = \frac{4}{7}\frac{1}{3} = \frac{4}{21}$, $PR(2) = PR(4) = \frac{4}{7}\frac{1}{6} = \frac{2}{21}$ and $PR(5) = PR(6) = PR(7) = \frac{3}{7}\frac{1}{3} = \frac{3}{21} = \frac{1}{7}$.

6. (3 pts) Rank the nodes by increasing PageRank scores if $\theta = 0.001$ (specify if there are equalities).

   Solution:
   For $\theta = 0.001$, the probability of teleportation is very high. This means the PageRank scores are almost all equal (to $\frac{1}{7}$).
   However, on average once every 1000 steps, the random does not teleport. This is what makes a difference between the scores. So we have to look at the probability that each node is reached after taking a random edge from a node selected uniformely at random.
   Let us define, for all nodes i, the probability to reach that node after only one step without teleportation (the previous step is teleportation). For node i, this is

   $$P_1(i) = \sum_{j \in parent(i)} P_0(j) \frac{1}{outdegree_j} = \sum_{j \in parent(i)} \frac{1}{7} \frac{1}{outdegree_j}$$

   where $P_0(i)$ is the probability of being art node i after a teleportation step.

   As we only care about the rank, we can ignore the factor $\frac{1}{7}$, which is the same for all nodes.
   This gives us $P_1(1) = P_1(5) = P_1(6) = P_1(7) = 1$, $P_1(2) = P_1(4) = \frac{1}{2}$ and $P_1(3) = 2$.

   We see that we have a lot of equalities. Nodes 2 and 4 are completely symetrical in the graph, so it is clear that their rank should be the same. The same is true for nodes 5, 6, 7. However, this is less clear for node 1 and 5.
   If we consider the -one in a million- case of having no teleportation two steps in a row, we can break this tie. Let $P_2(i)$ be the probability of reaching i after two non-teleportation steps. Then we have :

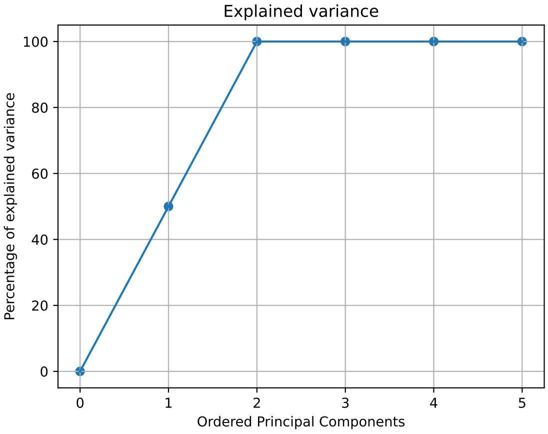
   $$P_2(i) = \sum_{j \in parent(i)} P_1(j) \frac{1}{outdegree_j}$$

   Using this formula for nodes 1 and 5 : $P_2(1) = P_1(3) = 2 > 1 = P_1(7) = P_2(5)$, so we have $P_2(1) > P_2(5)$.

   So the final rank is $2 = 4 < 5 = 6 = 7 < 1 < 3$.

Question 4: PCA (14 points)

You have a dataset of N data points $x_1, ..., x_N \in \mathbb{R}^5$. You rigorously follow the steps of PCA and you plot the cumulative variance explained by the principal components.
This is the plot you obtain:



1. (1 pt) From this plot, what can you deduce about the dimensionality of the data?

   Solution:
   We see that the variance is fully explained by the first 2 principal components. This means that the data is actually 2-dimensionnal (more precisely that the points are all in the same 2-dimensionnal affine subspace of $\mathbb{R}^5$).

2. (4 pts) Is it possible to observe this plot if $N = 1$, $N = 2$, $N = 3$, $N = 4$? Justify your answers.

   Solution:
   A dataset of 1 point has no variance, and a dataset of 2 points can only have variance along 1 direction (2 points are always aligned).
   Consequently, $N = 1$ and $N = 2$ are not possible.
   In order to have the two first components explaining each 50% of the variance (as shown on the plot), there must be at least 2 directions giving exactly the exact same (maximal) explained variance. This happens if the dataset has a rotational invariance.
   In particular, it works for $N = 3$ with an equilateral triangle, and for $N = 4$ for a square.

3. (3 pts) Let N_0 be the smallest N you found for the previous question. Give an example of a set of point(s) $x_1, ..., x_{N_0} \in \mathbb{R}^5$ that explain the plot (give their coordinates).

   Solution:
   The answer should be the coordinates of 3 points forming an equilateral triangle (centered on 0 or not).
   Using the Pythagorean theorem, we can find that if the triangle has edges of length a, then its height is $\frac{\sqrt{3}}{2}a$.
   So some possible answers are $(x_1 = (0, 0, 0, 0, 0), x_2 = (a, 0, 0, 0, 0), x_3 = (\frac{r}{2}a, \frac{\sqrt{3}}{2}a, 0, 0, 0))$ for all $a \in \mathbb{R}^*$.

4. (3 pts) Compute the covariance matrix for these points. What are the eigenvalues of the covariance matrix?

Solution:

To compute the covariance matrix, the data must be centered first. By taking the same example as above, with $a = 1$, and focusing on the 2 first dimensions (the 3 others are only zeros), we find that the center of the triangle is $(\frac{1}{2}, \frac{\sqrt{3}}{6})$.

This means that the coordinates of the centered dataset are $((-\frac{1}{2}, -\frac{\sqrt{3}}{6}), (\frac{1}{2}, -\frac{\sqrt{3}}{6}), (0, 2\frac{\sqrt{3}}{6}))$.

So the matrix in dimension 2 is :

$$\begin{pmatrix} -\frac{1}{2} & -\frac{\sqrt{3}}{6} \\ \frac{1}{2} & -\frac{\sqrt{3}}{6} \\ 0 & 2\frac{\sqrt{3}}{6} \end{pmatrix}$$

So the covariance matrix in dimension 2 is :

$$\frac{1}{3}\begin{pmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}$$

So the full covariance matrix in dimension 5 is :

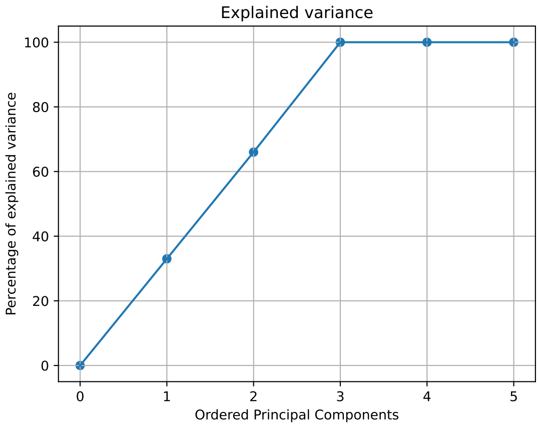
$$\begin{pmatrix} \frac{1}{6} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{6} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

The eigenvalues of the covariance matrix are $(\frac{1}{6}, \frac{1}{6}, 0, 0, 0)$ for this example.

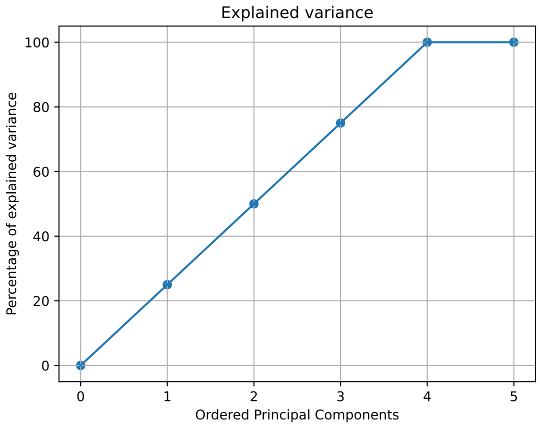
Depending on the dataset we chose (the size of the equilateral triangle), the eigenvalues can be different, but two are nonzero and equal, and the 3 others are 0. This result is expected because it is necessary and sufficient in order to get the plot.

5. (3 pts) You don't change your dataset, but you add a new point to it. It now contains $N_0 + 1$ points $x_1, ..., x_{N_0+1} \in \mathbb{R}^5$.

Can the new plot be plot (a)? Can it be plot (b)? Justify your answers.



(a)



(b)

Solution:
By adding only one point, the dataset can "gain" at most one dimension. So plot (b) is not possible because it corresponds to a dataset where the points lie in a 4 dimensional subspace, whereas the original data is in 2-dimensional subspace.
By adding one point to the equilateral triangle, it is possible to obtain a regular tetrahedron. The regular thetraedron having rotational invariances in 3 dimensions, it has the same amount of covariance explained by its 3 principal components.
So plot (a) is possible.

Question 5: Collaborative Filtering (10 points)

Here are two different datasets (a) and (b) of the ratings given by 3 users (1, 2, 3) about 4 items (A, B, C, D). The X represents a missing score.

	A	B	C	D
1	1	2	3	-1
2	0	1	2	-2
3	3	4	5	X

(a)

	A	B	C	D
1	1	2	3	-1
2	0	1	2	-2
3	0	4	5	X

(b)

1. (1 pt) Call r_{ui} the score given by user u to item i. Call $(b_u), u \in \{1, 2, 3\}$ the user biases and $(b_i), i \in \{A, B, C, D\}$ the item biases. One of the two datasets follows the model $r_{ui} = b_u + b_i$, the other one follows the model $r_{ui} = b_u + b_i + q_u p_i$, where q_u qnd p_i are respectively a user parameter and an item parameter.
   Identify which dataset follows which model. Justify your answer.

   Solution:
   The easiest way to identify which matrix follows which model is to look at the rank. The first column of matrix (b) has 2 zeros. This implies that row 1 cannot be generated by a linear combination of the 2 other rows. In addition, rows 2 and 3 are not aligned vectors. So rows 1, 2 and 3 are independant, which means the rank of the matrix is 3. Consequently, dataset (b) cannot be following the model with only biases.

   For the rest of the exercise, we focus on the model $r_{ui} = b_u + b_i$.

2. (1 pt) Predict the score that user 3 would give to item D in the dataset you identified to follow the model $r_{ui} = b_u + b_i$.

   Solution:
   In dataset (a), we can see that the scores of user 3 are 3 more than the scores of user 2. This means that $b_3 = b_2 + 3$. Consequently, the score given by user 3 to item D will be $r_{3D} = r_{2D} + 3 = 1$.

3. (2 pts) Suppose we have a single user u and a single item i, and the corresponding rating r_{ui}. We are trying to find bias parameters in the usual way, which amounts to the optimization problem $(b_u^*, b_i^*) = \text{argmin}_{(b_u,b_i)}(r_{ui} - b_u - b_i)^2$.
Give (all) the solution(s) to this minimization problem.

Solution:
There is no constraint on the values of b_u and b_i, so we can solve the problem such that $(r_{ui} - b_u - b_i)^2 = 0$, which is equivalent to $b_u + b_i = r_{ui}$. So the solutions are $(b_u, b_i) \in \{(r_{ui} + c, r_{ui} - c), \forall c \in \mathbb{R}\}$.

4. (3 pts) Modify the minimization problem above by adding a quadratic regularizer to the parameters.
Give (all) the solution(s) to the modified minimization problem.

Solution:
Adding a quadratic regularization to the parameters yields the following optimization problem :

$$(b_u^*, b_i^*) = \text{argmin}_{(b_u,b_i)}(r_{ui} - b_u - b_i)^2 + b_u^2 + b_i^2$$

Because of the regularisation, the problem now has a unique solution. As the problem is symmetrical for b_u and b_i, we have $b_u^* = b_i^*$. Consequently, we can solve

$$(b_u^*, b_i^*) = \text{argmin}_{(b_u,b_i)}(r_{ui} - b_u - b_i)^2 + b_u^2 + b_i^2, \quad b_u = b_i$$

So

$$b_u^* = b_i^* = \text{argmin}_b(r_{ui} - 2b)^2 + 2b^2$$

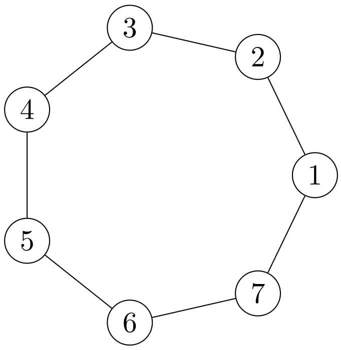
Let us note $f(x, b) = (x - 2b)^2 + 2b^2$. Then we have $f(x, b) = x^2 - 4xb + 4b^2 + 2b^2 = x^2 - 4xb + 6b^2$.
This is a second degree polynom in b, so the derivative (in b) is zero at only one point.
Because the dominant coefficient is positiv (6), this point is the unique minimum of $f(x, .)$.
So we solve $0 = \frac{d}{db}(f(x, b)) = -4x + 12b$.
Which gives $b = \frac{1}{3}x$.
So we have $(b_u^*, b_i^*) = (\frac{1}{3}r_{ui}, \frac{1}{3}r_{ui})$

5. (3 pts) We now return to the full problem with multiple u and i :
$\min_{b_u,b_i} \sum_{u,i}(r_{ui} - b_u - b_i)^2$
Under which condition does this problem have a unique solution?

Solution:
The solution is explained in Exercise 2 of Homework 4.
In the general case, the condition is the invertibility of the matrix $A^\top A$.
In our case, this matrix is not invertible. An easy way to see that is to realize that if you have a valid solution to the problem, for any constant $c \in \mathbb{R}$, you can remove c from all the user biases and add it to the item biases, which produces another solution to the problem (like in question 3.).

Question 6: Random Walk and Conductance (10 points)

A particle performs a random walk on a cycle graph with nodes numbered from 1 to n, denoted as C_n. For example C_7 is as follows:



1. (2 pts) For which n is this random walk ergodic? Justify your answer.

   Solution:
   A random walk on a cycle graph C_n is ergodic if it is both irreducible and aperiodic.

   - **Irreducibility**: A graph is irreducible if there is a path between any two nodes. In a cycle graph C_n, it is always possible to reach any node from any other node, as the nodes form a closed loop. Thus, the random walk on C_n is irreducible for any $n \geq 3$.
   - **Aperiodicity**: A Markov chain is aperiodic if there is no fixed number of steps k (greater than 1) such that all returns to a particular state occur in multiples of k. In C_n, if n is even, the period of the graph is 2, as the only way to return to the starting node in an odd number of steps is to move back and forth along the same edge, which is not possible in a cycle. If n is odd, the period of the graph is 1, because it is possible to return to any starting node in any number of steps.

   Therefore, the random walk on C_n is ergodic if and only if n is odd.

   For the rest of the exercise, consider all n such that this ergodicity condition is satisfied.

2. (3 pts) Let $\lambda_1(n)$ and $\lambda_2(n)$ be the first and second largest eigenvalues of the transition matrix of the random walk C_n. Determine $\lambda_1(n)$ and $\lambda_2(n)$ as n goes to infinity?

   Solution:

   Because it is ergodic, the transition matrix P has a stationary distribution with a maximum eigenvalue of 1.

   To determine $\lambda_2(n)$, we compute the conductance $\Phi(C_n)$ of C_n for odd n. Conductance is given by:

$$\Phi(C_n) = \min_{S \subset V, 0 < |S| \leq \frac{n}{2}} \frac{|\partial S|}{\min(|S|, |V \setminus S|)}$$

   For C_n, the smallest cut is obtained by taking any consecutive set of $\lceil \frac{n}{2} \rceil$ nodes. The boundary size $|\partial S|$ for this set is 2. Therefore,

$$\Phi(C_n) = \frac{4n}{n^2 - 1}$$

Following the Cheeger inequality, the second largest eigenvalue $\lambda_2(n)$ is related to the conductance by:

$$1 - 2\Phi(C_n) \leq \lambda_2(n) \leq 1 - \frac{\Phi(C_n)^2}{2}$$

Substituting $\Phi(C_n) = \frac{4n}{n^2-1}$, we get:

$$1 - \frac{8n}{n^2-1} \leq \lambda_2(n) \leq 1 - \frac{(\frac{4n}{n^2-1}^2)}{2}$$

Therefore, as n goes to infinity:

$$\lambda_2(n) \approx 1$$

Thus,

$$\lambda_1(n) = 1, \quad \lambda_2(n) \approx 1 \quad \text{as} \quad n \to \infty$$

3. (3 pts) For a fixed n, add an edge to C_n that increases the conductance the most. Where would you add this edge? Call this new graph C'_n and compute its conductance $\Phi_{C'_n}$. Compute the ratio $\frac{\Phi_{C'_n}}{\Phi_{C_n}}$ asymptotically when n goes to infinity.

Solution:

To maximize the conductance, we should add an edge that connects two nodes with the largest shortest path distance in C_n. For C_n, the longest shortest path is $\lfloor n/2 \rfloor$. Therefore, we add an edge between nodes 1 and $\lfloor n/2 \rfloor + 1$.

For C_n, $\Phi(C_n) = \frac{4n}{n^2-1}$.

For C'_n, the edge doesn't change the the boundary size $|\partial S|$ for some subsets S. Specifically, the set S that includes half of the nodes will now have an additional edge in the cut set, so:

$$\Phi(C'_n) = \frac{2}{\frac{2(n+1)(n+3)(n-1)}{2(n+1)2(n+1)}}$$

Therefore, the ratio is:

$$\frac{\Phi_{C'_n}}{\Phi_{C_n}} = \frac{\frac{4(n+1)}{(n+3)(n-1)}}{\frac{4n}{n^2-1}} =^{n\to\infty} 1$$

4. (2 pts) For the graph C_n, compute the expected number of steps for a particle starting at node 1 to visit node $\frac{n+1}{2}$ for the first time. Show your detailed calculations (Hint: consider this expected number of steps for every node and write a system of equations based on them).

Solution:

Let $E(i)$ be the expected number of steps to reach node $\frac{n+1}{2}$ from node i. We set up the following system of equations for the expected hitting times:

$$\begin{cases} E\left(\frac{n+1}{2}\right) = 0 \\ E(i) = 1 + \frac{1}{2}\left(E(i-1) + E(i+1)\right) \quad \text{for } i \neq \frac{n+1}{2} \end{cases}$$

This simplifies to:

$$E(i) = 1+\frac{1}{2}\left(E(i-1) + E(i+1)\right) \Rightarrow 2E(i) = 2+E(i-1)+E(i+1) \Rightarrow E(i) = 1+\frac{E(i-1)+E(i+1)}{2}$$

Solving this for all i from 1 to $\frac{n-1}{2}$ and using symmetry, we find:

$$E(1) = \sum_{k=1}^{\frac{n-1}{2}} 2k = \frac{(n-1)(n+1)}{4} = \frac{n^2-1}{4}$$

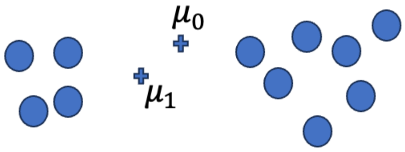
Thus, the expected number of steps for a particle starting at node 1 to visit node $\frac{n+1}{2}$ for the first time is:

$$E(1) = \frac{n^2 - 1}{4}$$

Question 7: Clustering (10 points)

This question concerns training Gaussian Mixture Models (GMMs). Throughout, we will assume there are two Gaussian components in the GMM. We will use μ_0, μ_1, σ_0^2, and σ_1^2 to define the means and variances of these two components, and will use π_0 and $(1 - \pi_0)$ to denote the mixture proportions of the two Gaussians (i.e., $p(x) = \pi_0 N(\mu_0, \sigma_0^2 I) + (1 - \pi_0) N(\mu_1, \sigma_1^2 I)$). We will also use θ to refer to the entire collection of parameters $\langle \mu_0, \mu_1, \sigma_0^2, \sigma_1^2, \pi_0 \rangle$ defining the mixture model $p(x)$.

1. (2 pts) Consider applying EM to train a Gaussian Mixture Model (GMM) to cluster the data below into two clusters. The ''+" points indicate the current means μ_0 and μ_1 of the two Gaussian mixture components after the k^{th} iteration of EM. Draw on the figure the directions in which μ_0 and μ_1 will move during the next M-step. Justify your answer.



   Solution:
   μ_0 moves to the right, and μ_1 moves to the left.

2. (1 pt) Is it possible that the marginal likelihood of the data (objective function), $\prod_j P(x_j|\theta)$, decreases on the next EM iteration? Justify your answer.

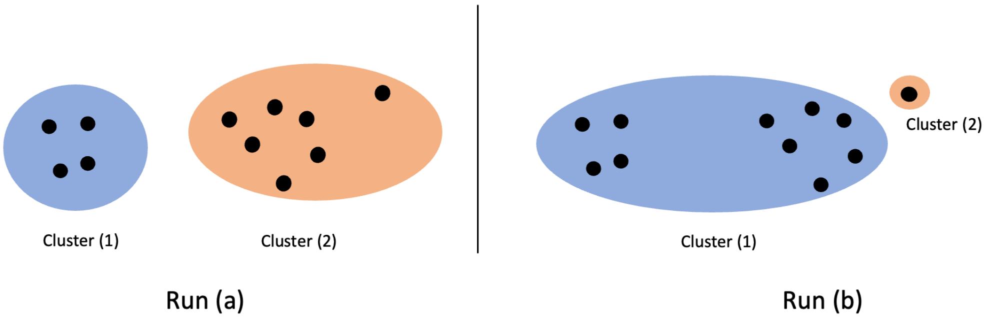
   Solution:
   No, Each iteration of the EM algorithm increases likelihood of the data, unless you happen to be exactly at a local optimum.

3. (2 pts) Consider the set of training data below. After running the Gaussian Mixture Model (GMM) twice on this data, with different initializations, we obtain the following results for the determinants of the covariance matrices of the probability density functions (pdf) for each cluster (we have two clusters, $K = 2$):

   - For run **a**: $\det(\Sigma_1) = 5$ and $\det(\Sigma_2) = 10$
   - For run **b**: $\det(\Sigma_1) = 15$ and $\det(\Sigma_2) = 0.00001$

   Here, Σ_1 and Σ_2 represent the covariance matrices for the Gaussian distributions of clusters 1 and 2, respectively.

   Which run, **a** or **b**, has the higher likelihood value? Provide a detailed justification for your answer.



Solution:
Run **b** has the higher likelihood value. This is because the very small determinant of Σ_2 indicates that the corresponding Gaussian component is very sharply peaked, which suggests that it fits tightly around the point because of a bad initialization, leading to a higher likelihood. The determinant of the covariance matrix is inversely related to the sharpness of the peak of the Gaussian distribution: the smaller the determinant, the sharper the peak, and thus the higher the likelihood of the data points near the peak.

4. (2 pts) Consider a training dataset consisting of partially labeled samples. For the first m samples $x_1, x_2, \dots, x_m$, we have corresponding z labels (which can be 0 or 1). The remaining samples, $x_{m+1} \dots x_{m+n}$, are unlabeled with unknown z values. Given this setup, which of the likelihood functions should be maximized to exploit both labeled and unlabeled data in GMM clustering? Justify your answer.

   (a) $(\prod_{i=1}^{m} P(x_i|z_i, \theta))(\prod_{i=m+1}^{m+n} P(x_i|\theta))$

   (b) $(\prod_{i=1}^{m} P(x_i, z_i|\theta))(\prod_{i=m+1}^{m+n} P(x_i|\theta))$

   (c) $(\prod_{i=1}^{m} P(z_i|\theta)P(x_i|\theta))(\prod_{i=m+1}^{m+n} P(x_i|\theta))$

   Solution:
   The correct answer is (b): $(\prod_{i=1}^{m} P(x_i, z_i|\theta))(\prod_{i=m+1}^{m+n} P(x_i|\theta))$. This is because for the labeled data, we need to consider the joint probability of the samples and the labels to fully explain the information provided by the data. For the unlabeled data, we only consider the probability of the samples itself.

5. (3 pts) How do you need to change the E-step and M-steps of the EM algorithm to fit the above objective function?

   Solution:
   E: Same as before for $x_{m+1} \dots x_{m+n}$, and for $i \leq m$, $\gamma_{ij} = \delta_{j,z(i)}$.

   M: Same as before.

6. (0 pts, but a deep sense of satisfaction if you are right) Who wins the Euro 2024 championship?

   Solution:
   Based on the current team performances, the popular prediction is Spain. However, surprises are always possible in football tournaments:)