

Esolution

Place student sticker here

Note:

- During the attendance check a sticker containing a unique code will be put on this exam.
- This code contains a unique number that associates this exam with your registration number.
- This number is printed both next to the code and to the signature field in the attendance check list.

Machine Learning for Graphs and Sequential Data

Exam: IN2323 / Endterm

Date: Wednesday 5th August, 2020

Examiner: Prof. Dr. Stephan Günnemann

Time: 11:30 – 12:45

Working instructions

- This exam consists of **14 pages** with a total of **10 problems**.
Please make sure now that you received a complete copy of the exam.
- The total amount of achievable credits in this exam is 43 credits.
- Detaching pages from the exam is prohibited.
- Allowed resources:
 - all materials that you will use on your own (lecture slides, calculator etc.)
 - **not allowed are any forms of collaboration between examinees and plagiarism**
- You have to sign the code of conduct.
- Make sure that the **QR codes are visible** on every uploaded page. Otherwise, we cannot grade your exam.
- Only write on the provided sheets, **submitting your own additional sheets is not possible**.
- Last two pages can be used as scratch paper.
- All sheets (including scratch paper) have to be submitted to the upload queue. Missing pages will be considered empty.
- **Only use a black or blue color (no red or green)!**
- Write your answers only in the provided solution boxes or the scratch paper.
- **For problems that say "Justify your answer" you only get points if you provide a valid explanation.**
- **For problems that say "Prove" you only get points if you provide a valid mathematical proof.**
- If a problem does not say "Justify your answer" or "Prove" it's sufficient to only provide the correct answer.
- Exam duration - 75 minutes.

Left room from _____ to _____ / Early submission at _____

Problem 1 Normalizing Flows (4 credits)

We consider two transformations $f_1(\mathbf{z}) = \begin{bmatrix} z_1 \\ z_2^{1/3} \end{bmatrix}$ and $f_2(\mathbf{z}) = \begin{bmatrix} z_1(|z_2| + 1) \\ z_2 \end{bmatrix}$ from \mathbb{R}^2 to \mathbb{R}^2 .

The respective inverse transformation are $f_1^{-1}(\mathbf{x}) = \begin{bmatrix} x_1 \\ x_2^3 \end{bmatrix}$ and $f_2^{-1}(\mathbf{x}) = \begin{bmatrix} \frac{x_1}{|x_2|+1} \\ x_2 \end{bmatrix}$.

The respective Jacobians are

$$J_{f_1} = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{3}z_2^{-\frac{2}{3}} \end{bmatrix} \quad J_{f_2} = \begin{bmatrix} |z_2| + 1 & \text{sign}(z_2)z_1 \\ 0 & 1 \end{bmatrix}$$

$$J_{f_1^{-1}} = \begin{bmatrix} 1 & 0 \\ 0 & 3x_2^2 \end{bmatrix} \quad J_{f_2^{-1}} = \begin{bmatrix} \frac{1}{|x_2|+1} & \frac{-\text{sign}(x_2)x_1}{(|x_2|+1)^2} \\ 0 & 1 \end{bmatrix}$$

We assume a Gaussian base distribution $p_1(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$. We observed one point $\mathbf{x}_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$.

We propose to stack the transformations f_1, f_2 to transform the base distribution p_1 in the distribution p_2 with normalizing flows. Compute the likelihood for \mathbf{x} under the transformed distribution p_2 if the order of transformations is f_1 followed by f_2 .

Hint: You might use the density of the unit variate Gaussian $p = \mathcal{N}(0, 1)$ at the following points: $p(1/2) = 0.3521$, $p(1/3) = 0.3774$, $p(1/9) = 0.3965$, $p(5) = 1.4867e^{-06}$, $p(8) = 5.0523e^{-15}$, $p(10) = 7.6946e^{-23}$

We consider $f_1 \cdot f_2$ and compute:

$$f_2^{-1}(\mathbf{x}) = \begin{bmatrix} \frac{1}{3} \\ 2 \end{bmatrix}, f_1^{-1}(f_2^{-1}(\mathbf{x})) = \begin{bmatrix} \frac{1}{8} \\ \frac{1}{8} \end{bmatrix}$$

By using the determinant of the Jacobians, we apply the change of variable formula:

$$\begin{aligned} p_2(\mathbf{x}) &= p_1\left(\begin{bmatrix} \frac{1}{8} \\ \frac{1}{8} \end{bmatrix}\right) \times \frac{1}{|2|+1} \times (3 \times 2^2) \\ &= p_1\left(\begin{bmatrix} \frac{1}{8} \\ \frac{1}{8} \end{bmatrix}\right) \times 4 \\ &= 7.6266e^{-15} \end{aligned}$$

Version 2:

We consider $f_2 \cdot f_1$ and compute:

$$f_1^{-1}(\mathbf{x}) = \begin{bmatrix} \frac{1}{8} \\ \frac{1}{8} \end{bmatrix}, f_2^{-1}(f_1^{-1}(\mathbf{x})) = \begin{bmatrix} \frac{1}{9} \\ \frac{1}{8} \end{bmatrix}$$

By using the determinant of the Jacobians, we apply the change of variable formula:

$$\begin{aligned} p_2(\mathbf{x}) &= p_1\left(\begin{bmatrix} \frac{1}{9} \\ \frac{1}{8} \end{bmatrix}\right) \times (3 \times 2^2) \times \frac{1}{|8|+1} \\ &= p_1\left(\begin{bmatrix} \frac{1}{9} \\ \frac{1}{8} \end{bmatrix}\right) \times \frac{4}{3} \\ &= 2.6709e^{-15} \end{aligned}$$

Problem 2 Variational Inference (5 credits)

We are performing variational inference in some latent variable model $p_\theta(x, z)$ using the following family of variational distributions $\mathcal{Q}_1 = \{\mathcal{N}(z|\phi, 1) : \phi \in \mathbb{R}\}$.

a) Assume that the variational distribution $q \in \mathcal{Q}_1$ is fixed, and we are trying to maximize the ELBO w.r.t. θ using gradient ascent. Is it necessary to use the reparametrization trick in this case? If yes, explain how to do it for our family of distributions \mathcal{Q}_1 ; if not, provide a justification.

0
1
2
3

No, we don't need to use the reparametrization trick when optimizing only w.r.t. θ (i.e., when q is fixed). We can approximate the gradient of the ELBO w.r.t. θ simply using Monte Carlo, and the reparametrization trick is not necessary (see slide 75 of Lecture 2).

b) Consider another family of distributions $\mathcal{Q}_2 = \{\mathcal{N}(z|0, s^2) : s \in (0, \infty)\}$. Which of the following statements is true? Justify your answer.

0
1
2

1. $\max_{\theta, q \in \mathcal{Q}_1} \text{ELBO}(\theta, q) < \max_{\theta, q \in \mathcal{Q}_2} \text{ELBO}(\theta, q)$
2. $\max_{\theta, q \in \mathcal{Q}_1} \text{ELBO}(\theta, q) = \max_{\theta, q \in \mathcal{Q}_2} \text{ELBO}(\theta, q)$
3. $\max_{\theta, q \in \mathcal{Q}_1} \text{ELBO}(\theta, q) > \max_{\theta, q \in \mathcal{Q}_2} \text{ELBO}(\theta, q)$
4. It's impossible to tell without additional information.

It's impossible to tell without additional information about $p_\theta(x, z)$.

For example, if the true posterior is $\mathcal{N}(z|\mu, 1)$ for some $\mu \in \mathbb{R}$, then (3) will hold. As another example, if the true posterior is $\mathcal{N}(z|0, \sigma^2)$ for some $\sigma \in (0, \infty)$, then (1) will hold. This means that it's impossible to tell which one is the case without additional information.

Problem 3 Robustness of Machine Learning Models (6 credits)

Suppose we have trained a binary classifier $f : \mathbb{R}^d \rightarrow \{0, 1\}$ and want to certify its robustness via randomized smoothing. Therefore, the *smoothed classifier* $g_{\sigma^2}(\mathbf{x}) = \mathbb{E}[\mathbb{I}[f(\mathbf{x} + \varepsilon) = 1]]$, where $\varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$.

Fact: $\Phi^{-1}(g_{\sigma^2}(\mathbf{x}))$ is $1/\sigma$ -**Lipschitz** w.r.t. \mathbf{x} and the L_2 norm, where $\Phi(z)$ denotes the cumulative distribution function (CDF) of the standard normal distribution.

- 0 ☐
1 ☐
2 ☐
3 ☐
4 ☐
- a) **Using the above fact about the Lipschitz-continuity** of $\Phi^{-1}(g_{\sigma^2}(\mathbf{x}))$, show that the largest certifiable L_2 radius r around a sample \mathbf{x} is identical to the result shown in the lecture. More precisely, show that

$$r = \sigma \Phi^{-1}(g_{\sigma^2}(\mathbf{x})).$$

Hint: You may assume we can evaluate $g_{\sigma^2}(\mathbf{x})$ in closed form and you may use the following results: $\lim_{z \rightarrow 0} \Phi^{-1}(z) = -\infty$, $\Phi^{-1}(0.5) = 0$, $\lim_{z \rightarrow 1} \Phi^{-1}(z) = \infty$.

$$\begin{aligned} \|\Phi^{-1}(g_{\sigma^2}(\mathbf{x})) - \Phi^{-1}(0.5)\|_2 &\leq \frac{1}{\sigma} \|\mathbf{x} - \tilde{\mathbf{x}}\|_2 \\ \Phi^{-1}(g_{\sigma^2}(\mathbf{x})) &\leq \frac{1}{\sigma} \|\mathbf{x} - \tilde{\mathbf{x}}\|_2 \\ \|\mathbf{x} - \tilde{\mathbf{x}}\|_2 &\geq \sigma \Phi^{-1}(g_{\sigma^2}(\mathbf{x})) \end{aligned}$$

- 0 ☐
1 ☐
2 ☐
- b) A fellow student has a promising idea: by letting $\sigma \rightarrow \infty$ we can make the Lipschitz constant of the smoothed classifier arbitrarily small, leading to arbitrarily large certifiable radii, i.e. a very robust model. Is this a good idea? Why or why not?

No, since $\sigma \rightarrow \infty$ means that the signal-to-noise ratio of the smoothed examples approaches zero. Similar, also valid arguments:

- For $\sigma \rightarrow \infty$ we sample from very large regions of the input space, for which a large fraction of samples belongs to other classes \Rightarrow we don't get certificates.
- $\sigma \rightarrow \infty$ means we introduce so much noise that we lose accuracy / performance of the smoothed classifier.

Problem 4 Markov Property (3 credits)

We consider the following sequences of random variables U_0, U_1, \dots, U_t .

- a) $U_t = \begin{bmatrix} X_t \\ Z_t \end{bmatrix}$ where X_t are observed variables and Z_t are latent variables of an Hidden Markov Model. Does the sequence of variables U_t fulfill the Markov property i.e. $P(U_t|U_{t-1}) = P(U_t|U_{t-1}, \dots, U_0)$? Justify your answer.

0
1

Yes, since X_t and Z_t are conditionally independent $X_{t-2}, Z_{t-2}, \dots, X_0, Z_0$ of given Z_{t-1} we can write:

$$P(X_t, Z_t | X_{t-1}, Z_{t-1}, \dots, X_0, Z_0) = P(X_t, Z_t | X_{t-1}, Z_{t-1})$$

- b) We consider an AR(p) process X_t . Under what condition on p and k does the sequence of variables $U_t = [X_{t-1}, \dots, X_{t-k}]$ fulfill the Markov property i.e. $P(U_t|U_{t-1}) = P(U_t|U_{t-1}, \dots, U_0)$? Justify your answer.

0
1

We consider an AR(p) process X_t and compute:

$$P(X_{t-1}, \dots, X_{t-k} | X_{t-2}, \dots, X_0) = P(X_{t-1} | X_{t-1}, \dots, X_{t-p-1})$$

This quantity is equal to $P(X_{t-1}, \dots, X_{t-k} | X_{t-2}, \dots, X_{t-k-1}) = P(X_{t-1} | X_{t-2}, \dots, X_{t-k-1})$ iff $p \leq k$.

- c) We consider a recurrent neural network which produces X_t . Does the sequence of variables $U_t = X_t$ fulfill the Markov property i.e. $P(U_t|U_{t-1}) = P(U_t|U_{t-1}, \dots, U_0)$? Justify your answer.

0
1

No, both variables X_t and X_{t-2} depend on the hidden state at time $t-2$ given X_{t-1} .

Problem 5 Markov Chain (3 credits)

0 ☐
1 ☐
2 ☐
3 ☐

We consider a Markov chain X_t in $\{1, C\}$ with parameters π, \mathbf{A} . We assume we observed the sequence $S_k = [\underbrace{v_0, \dots, v_0}_{k \text{ times}}, \underbrace{v_1, \dots, v_1}_{k \text{ times}}, \dots, \underbrace{v_T, \dots, v_T}_{k \text{ times}}]$ where each value is observed k times. The parameter k can be seen as a discretization parameter of the time space.

Compute the likelihood of the sequence under the parameters π, \mathbf{A} i.e. $P_{\pi, \mathbf{A}}(S_k)$. What happens to this quantity if you increase the discretization parameter from k to $k' > k$ but keep the same model parameter π, \mathbf{A} ?

We compute the likelihood:

$$P_{\pi, \mathbf{A}}(S_k) = \pi_{v_0} \times \prod_{t=0}^{T-1} A_{v_t, v_{t+1}} \times \prod_{t=0}^T A_{v_t, v_t}^{k-1}.$$

Since the parameters $A_{i,j} < 1$ the likelihood decreases when k increases.

Problem 6 Temporal Point Process (6 credits)

Consider an inhomogeneous Poisson process (IPP) on the interval $[0, 4]$ with the intensity function

$$\lambda(t) = \begin{cases} a & \text{if } t \in [0, 3] \\ b & \text{if } t \in (3, 4] \end{cases}$$

where $a > 0$, $b > 0$ are some positive parameters.

a) Assume that you observed a sequence of events $\{0.2, 1.0, 1.5, 2.9, 3.1, 3.8\}$ generated by the above IPP. What is the maximum likelihood estimate of the parameters a and b ?

The log-likelihood of a TPP realization $\{t_1, \dots, t_N\}$ is

$$\log p(\{t_1, \dots, t_N\}) = \sum_{i=1}^N \log \lambda(t_i) - \int_0^T \lambda(t) dt$$

Plugging in our values $\{0.2, 1.0, 1.5, 2.9, 3.1, 3.8\}$, we obtain

$$= 4 \log a + 2 \log b - 3a - b$$

To maximize w.r.t. a and b , we can compute the derivatives, set them to zero and solve for a and b .

$$\frac{\partial}{\partial a} \log p(\{t_1, \dots, t_N\}) = \frac{4}{a} - 3 \stackrel{!}{=} 0$$

$$\Rightarrow a^* = \frac{4}{3}$$

$$\frac{\partial}{\partial b} \log p(\{t_1, \dots, t_N\}) = \frac{2}{b} - 1 \stackrel{!}{=} 0$$

$$\Rightarrow b^* = 2$$

b) Assume that $a = 1$ and $b = 5$. What is the expected number of events generated by the IPP in this case?

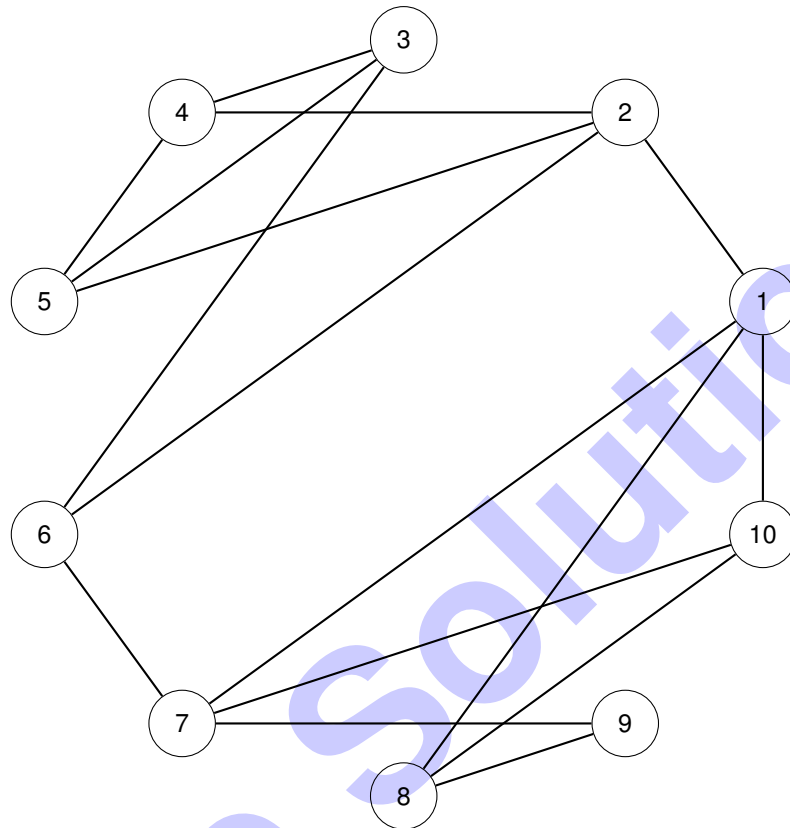
We know from the properties of Poisson process that the expected number of events equals to

$$\begin{aligned} \int_0^T \lambda(t) dt &= \int_0^4 \lambda(t) dt \\ &= \int_0^3 1 dt + \int_3^4 5 dt \\ &= 3 + 5 \\ &= 8 \end{aligned}$$

Problem 7 Clustering with the Planted Partition Model (4 credits)

0 ☐
1 ☐
2 ☐
3 ☐
4 ☐

The following graph has been generated from a planted partition model with in-community edge probability p and between-community edge probability q .



Assuming $p < q$, find the maximum likelihood community assignments under a PPM.

Give your solution as two sets of node labels making up the two discovered communities. Justify your answer.

The likelihood of graph under a planted partition model is proportional to a $\frac{q(1-p)}{p(1-q)}$ to the power of the induced cut size of the partitioning. If $p < q$, this fraction is greater than 1 and the likelihood is maximized by the maximum cut size. Therefore, the two clusters are made up of the nodes $\{2, 3, 7, 8\}$ and $\{1, 4, 5, 6, 9, 10\}$.

Problem 8 PageRank in a Wheel (6 credits)

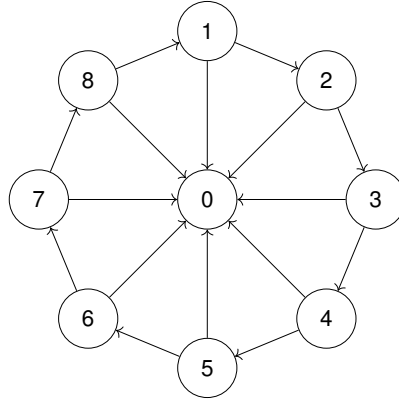


Figure 8.1: Example of a directed wheel graph with $n + 1 = 9$ nodes

Consider a directed graph of size $n + 1$ with a cycle of n nodes and an additional central node that every other node connects to (see figure). So we have a graph with node set $\mathcal{V} = \{0, 1, \dots, n\}$ and edge set

$$\mathcal{E} = \{(i, i + 1) \mid i \in \{1, \dots, n - 1\}\} \cup \{(n, 1)\} \cup \{(i, 0) \mid i \in \{1, \dots, n\}\}.$$

We want to compute the PageRank scores with a link-follow probability of β (a teleport probability of $1 - \beta$) and some arbitrary teleport vector π , $\sum_{i=0}^n \pi_i = 1$. Note that we index π from 0 to n .

We define the predecessor function pa as the index of the predecessor of a node in the directed cycle, i.e.

$$\text{pa}(1) = n \quad \text{and} \quad \text{pa}(i) = i - 1 \quad \forall i \in \{2, \dots, n\}.$$

You can write $\text{pa}^k(i)$ for the k -th predecessor of node i , i.e. $\text{pa}^3(i) = \text{pa}(\text{pa}(\text{pa}(i)))$ and $\text{pa}^0(i) = i$.

a) Set up the PageRank equations for all nodes in scalar form, i.e. each r_i separately instead of matrix form.

The nodes 1 through n have exactly 2 outgoing edges, so their degree is 2 and their PageRank equations are

$$r_i = \beta \cdot \frac{r_{\text{pa}(i)}}{2} + (1 - \beta)\pi_i \quad \forall i \in \{1, \dots, n\}$$

The central node has incoming edges from all other nodes and therefore the following PageRank equation.

$$r_0 = \beta \cdot \left(\sum_{i=1}^n \frac{r_i}{2} \right) + (1 - \beta)\pi_0$$

b) Why is this graph problematic for PageRank without random teleportation ($\beta = 1$)?

The central node has no outgoing edges and is a dead end. A random walker would be stuck there and any PageRank would be “lost”.

c) Show that the PageRank for node $i \in \{1, \dots, n\}$ in the outer cycle is given by

$$r_i = \frac{(1 - \beta)}{1 - \left(\frac{\beta}{2}\right)^n} \sum_{j=0}^{n-1} \left(\frac{\beta}{2}\right)^j \pi_{pa^j(i)}.$$

In the formula of some cycle node i , we can plug in the formula for the predecessor nodes repeatedly and get

$$\begin{aligned} r_i &= \frac{\beta}{2} r_{pa(i)} + (1 - \beta) \pi_i \\ &= \frac{\beta}{2} \left(\frac{\beta}{2} r_{pa(pa(i))} + (1 - \beta) \pi_{pa(i)} \right) + (1 - \beta) \pi_i = \left(\frac{\beta}{2}\right)^2 r_{pa(pa(i))} + \frac{\beta}{2} (1 - \beta) \pi_{pa(i)} + (1 - \beta) \pi_i \\ &= \left(\frac{\beta}{2}\right)^2 \left(\frac{\beta}{2} r_{pa(pa(pa(i)))} + (1 - \beta) \pi_{pa(pa(i))} \right) + \frac{\beta}{2} (1 - \beta) \pi_{pa(i)} + (1 - \beta) \pi_i \\ &= \left(\frac{\beta}{2}\right)^3 r_{pa(pa(pa(i)))} + \left(\frac{\beta}{2}\right)^2 (1 - \beta) \pi_{pa(pa(i))} + \frac{\beta}{2} (1 - \beta) \pi_{pa(i)} + (1 - \beta) \pi_i \\ &= \dots \end{aligned}$$

Repeating this process for k steps expresses the rank of node i in terms of its k -th predecessor.

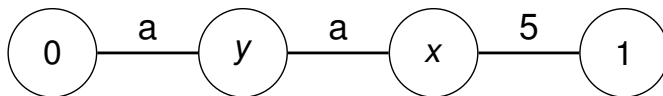
$$r_i = \left(\frac{\beta}{2}\right)^k r_{pa^k(i)} + (1 - \beta) \sum_{j=0}^{k-1} \left(\frac{\beta}{2}\right)^j \pi_{pa^j(i)}$$

But node i is its own n -th predecessor, so we get

$$r_i = \left(\frac{\beta}{2}\right)^n r_i + (1 - \beta) \sum_{j=0}^{n-1} \left(\frac{\beta}{2}\right)^j \pi_{pa^j(i)} \Leftrightarrow r_i = \frac{(1 - \beta)}{1 - \left(\frac{\beta}{2}\right)^n} \sum_{j=0}^{n-1} \left(\frac{\beta}{2}\right)^j \pi_{pa^j(i)}$$

Problem 9 Label Propagation (4 credits)

Consider the following graph



<input type="checkbox"/>	0
<input type="checkbox"/>	1
<input type="checkbox"/>	2
<input type="checkbox"/>	3
<input type="checkbox"/>	4

The nodes labeled 0 and 1 are observed and from class 0 and 1, respectively. One edge has a fixed weight, the other two have a variable edge weight of $a \geq 0$. The two center nodes are unobserved and we call their labels x and y .

We want to predict classes for the two center nodes that minimize the Label Propagation objective exactly,

$$\frac{1}{2} \sum_{ij} w_{ij} (y_i - y_j)^2$$

where W is the weighted adjacency matrix and y_i, y_j are the labels of the nodes.

Find the set of all possible edge weights a that guarantee that node x is assigned to class 0. Justify your answer.

If we instantiate the objective for this graph (disregarding constant factors), we get

$$5(x - 1)^2 + a * (x - y)^2 + a(y - 0)^2 = (5 + a)x^2 - 10x + 5 - 2axy + 2ay^2.$$

Since x and y are restricted to be either 0 or 1, we can just look at all cases.

$$\begin{aligned} x = 0, y = 0 &\Rightarrow 5 & x = 0, y = 1 &\Rightarrow 5 + 2a \\ x = 1, y = 0 &\Rightarrow a & x = 1, y = 1 &\Rightarrow a \end{aligned}$$

If $x = 0$, setting $y = 1$ always increases the cost, so the assignment $x = 0, y = 1$ will never be made. In the case of $x = 1$, the value of y is irrelevant. In conclusion, we assign $x = 0$ if $5 < a$ and $(5, \infty)$ is the set we are looking for.

Alternatively: In Label Propagation with two classes, the exact solution is given by the minimum cut between the sets of labeled nodes. For two unlabeled nodes, there are 4 possible cuts corresponding to the 4 possible assignments of x and y above. By the same reasoning as above, we arrive at the same solution.

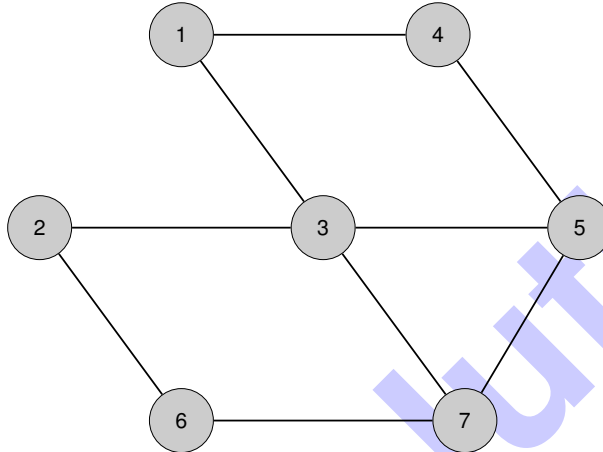
Problem 10 Adversarial Attacks on Graph Neural Networks (2 credits)

Suppose you are given the following two-layer graph neural network.

$$f(\mathbf{A}, \mathbf{X}) = \mathbf{Z} = \text{Softmax}(\hat{\mathbf{A}} \text{ReLU}(\hat{\mathbf{A}} \mathbf{X} \mathbf{W}_1) \mathbf{W}_2)$$

$\mathbf{X} \in \mathbb{R}^{N \times D}$ are the node features, \mathbf{Z} are the node predictions, \mathbf{W}_x are weight matrices of appropriate dimensions and $\hat{\mathbf{A}} = \tilde{\mathbf{D}}^{-\frac{1}{2}} \tilde{\mathbf{A}} \tilde{\mathbf{D}}^{-\frac{1}{2}}$ is the propagation matrix as defined for GCNs. Here, $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$, where \mathbf{A} is the adjacency matrix and \mathbf{I} is the identity matrix, and $\tilde{\mathbf{D}}$ is a diagonal matrix of node degrees $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$.

The model was trained for the task of semi-supervised node classification, and we want to predict a class c for node 6 in the following graph \mathbf{A} :



- 0 ☐ a) An adversary with complete knowledge about the graph \mathbf{A} and the trained model $f(\mathbf{A}, \mathbf{X})$ may delete one edge to perturb the prediction for node 6. Deleting which of the following edges would lead to a greater change to the prediction for node 6? Justify your answer.

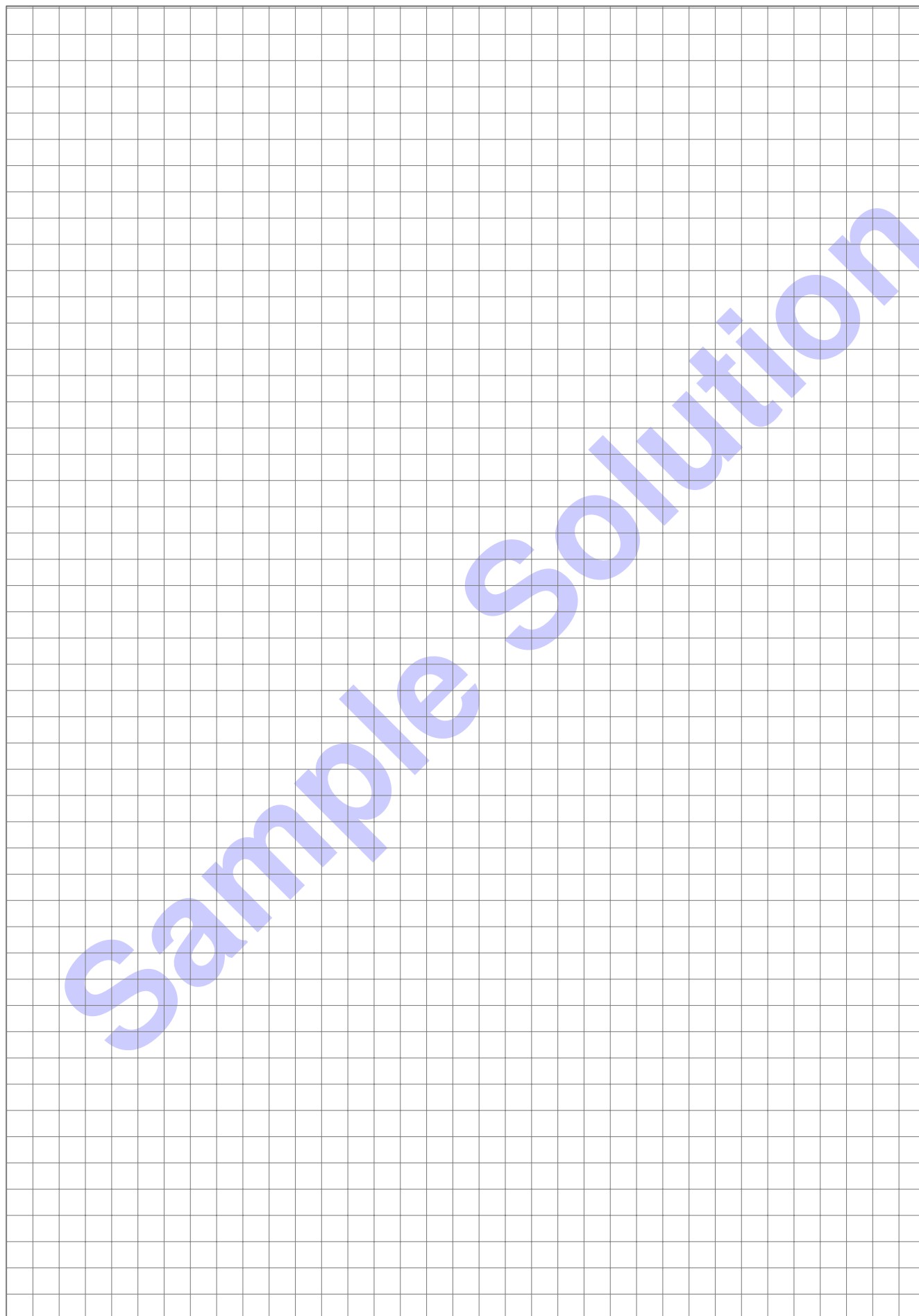
1. The edge connecting node 5 and 7
2. The edge connecting node 3 and 5
3. There is not enough information to determine which deletion leads to a greater change.

The edge connecting node 5 and 7. The model is a two-layer GCN with a propagation depth of 2. The edge connecting nodes 3 and 5 is more than 2 hops away from node 6, so it has no influence.

- 0 ☐ b) Assume we instead have a Personalized Propagation of Neural Predictions (PPNP) model instead of the two-layer GCN. How does this affect your choice? Justify your answer.

The PPNP model is the limit of infinite propagation steps with teleportation, so any node can influence any other and in this general setting there is not enough information to decide which deletion leads to a greater change.

Additional space for solutions—clearly mark the (sub)problem your answers are related to and strike out invalid solutions.

A large grid of graph paper for solutions, with a diagonal watermark reading "Sample Solution". The grid is composed of small squares, and the watermark is written in a large, light blue, sans-serif font, oriented diagonally from the bottom-left to the top-right.

Sample Solution