Data Analytics and Machine Learning Group
Department of Informatics
Technical University of Munich

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

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

The respective Jacobians are

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

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

0 1 2 3 4

We assume a Gaussian base distribution $p_1(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$. We observed one point $\boldsymbol{x}_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$.
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 $\boldsymbol{x}$ 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:* $p(1/2) = 0.3521$, $p(1/3) = 0.3774$, $p(1/9) = 0.3965$, $p(5) = 1.4867e^{-06}$, $p(8) = 5.0523e^{-15}$, $p(10) = 7.6946e^{-23}$

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

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.

Problem 3 Robustness of Machine Learning Models (6 credits)

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

Fact: $\Phi^{-1}(g_{\sigma^2}(\boldsymbol{x}))$ is $1/\sigma$-**Lipschitz** w.r.t. $\boldsymbol{x}$ 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 $\Phi^{-1}(g_{\sigma^2}(\boldsymbol{x}))$, show that the largest certifiable L_2 radius r around a sample $\boldsymbol{x}$ is identical to the result shown in the lecture. More precisely, show that

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

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

0 1 2

b) A fellow student has a promising idea: by letting $\sigma \to \infty$ 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?

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. $P(U_t|U_{t-1}) = P(U_t|U_{t-1}, ..., U_0)$? Justify your answer.

0
1

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

0
1

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. $P(U_t|U_{t-1}) = P(U_t|U_{t-1}, ..., U_0)$? Justify your answer.

0
1

Problem 5 Markov Chain (3 credits)

0 1 2 3

We consider a Markov chain X_t in $\{1, C\}$ with parameters $\pi, \boldsymbol{A}$. 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 $\pi, \boldsymbol{A}$ i.e. $P_{\pi, \boldsymbol{A}}(S_k)$. What happens to this quantity if you increase the discretization parameter from k to $k' > k$ but keep the same model parameter $\pi, \boldsymbol{A}$?

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?

0
1
2
3
4

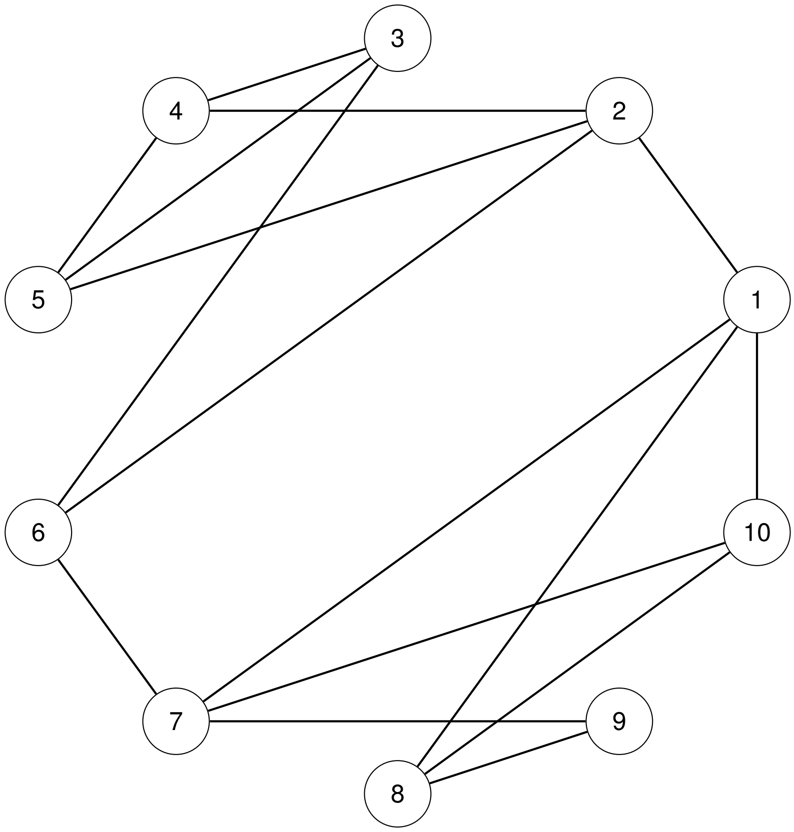
b) Assume that $a = 1$ and $b = 5$. What is the expected number of events generated by the IPP in this case?

0
1
2

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.

Problem 8 PageRank in a Wheel (6 credits)

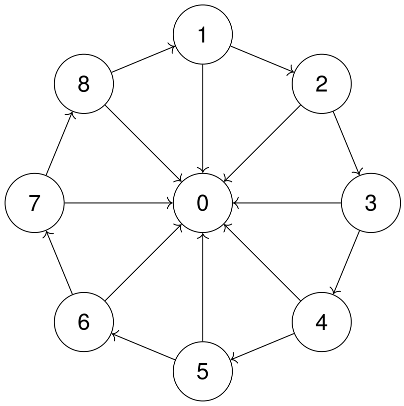


Figure 8.1: Example of a directed wheel graph with $n + 1 = 9$ nodes

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.

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

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

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

0
1
2

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

0
1

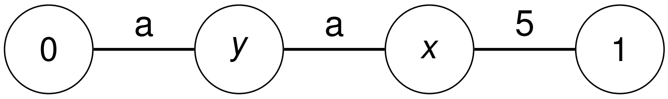
0
1
2
3

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

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

Problem 9 **Label Propagation (4 credits)**

Consider the following graph



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,

$$\frac{1}{2} \sum_{ij} W_{ij} \left(y_i - y_j \right)^2$$

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

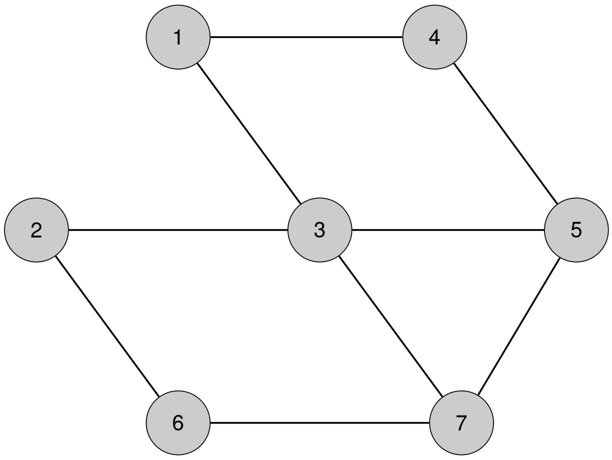
Find the set of all possible edge weights a that guarantee that node x is assigned to class 0. Justify your answer.

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

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

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

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



0 1 a) An adversary with complete knowledge about the graph $\boldsymbol{A}$ and the trained model $f(\boldsymbol{A}, \boldsymbol{X})$ 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.

0 1 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.

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