Logic and Theory of Discrete Systems
RWTH AACHEN UNIVERSITY

Prof. Dr. M. Grohe

E. Fluck, N. Runde

Algorithmic Foundations of Data Science

Mock Exam, whenever you want

Last Name: ____________________

First Name: ____________________

Student ID: ____________________

Study Program: ____________________

Remarks

- Write your name and student ID on **every** sheet of paper you use.
- Write your solution in the space provided on the problem sheets. If you need more paper, ask us. Do **not** use your **own paper** and hand in all paper you obtained (even scratch paper).
- You can answer either in German or in English but please do not mix languages.
- Only use **black** or **blue** pens. Do **not** use pencils.
- If you provide multiple solutions to a question, the worst of those counts.
- You have **120 minutes** to work on the exam. With **50 points** you have passed this exam.

I hereby declare that I have read the above guidelines and that I am healthy enough to take the exam.

____

(Signature)

Do not write below this line.

	1	2	3	4	5
Points	/ 19	/ 24	/ 18	/ 17	/ 22
Sign.					

$\sum$	/ 100

Problem 1 (Questions on Learning Problems) 5+5+4+5 = 19 points

a) Briefly explain the concepts of supervised and unsupervised learning. What is the difference? For both concepts give a learning algorithm that fits to this concept.

b) Define the entropy of a probability distribution $\mathcal{P}$ on a finite sample space Ω. What is the intuitive meaning of the entropy?

c) The PAC-learning framework allows to derive upper bounds on the number of training examples required for a learning problem. Which assumptions does the framework make on the training data?

d) In the lecture we have discussed two methods for Dimension Reduction of high-dimensional data. Which ones?

Problem 2 (Eigenvalues and Eigenvectors) 6+8+4+6 = 24 points

a) Consider the following matrix

$$M = \begin{pmatrix} 2 & -1 & 3 & 1 & -1 \\ 1 & -2 & -3 & 1 & 1 \end{pmatrix}.$$

What are the eigenvectors for all non-zero eigenvalues of $M^T M$?

b) Describe the Power Iteration Algorithm. What does it compute? What are the requirements on the input for the algorithm to work?

c) Give a square matrix A and an initial vector x_0 such that the Power Iteration algorithm does not converge on A with initial vector x_0. Explain why it does not converge.

d) Suppose we are given ℓ points $x_1, \ldots, x_\ell \in \mathbb{R}^n$ with similarity measure $s\colon \mathbb{R}^n \times \mathbb{R}^n \to [0,1]$. We want to cluster the points into k clusters using Spectral Clustering. Give pseudocode for the Spectral Clustering Method that realizes this task. Briefly explain the steps of your algorithm.

Problem 3 (Markov Chains) 4+3+7+4 = 18 points

a) Suppose your are on a summer holiday and choose your daily activity only based on the activity of the previous day. There are only three possible activities: going to the beach **(B)**, hiking (H), or a trip to the nearby city **(C)**.

- If you went to the beach yesterday, then today you go to the beach again **(B)** with probability 0.4, you choose hiking **(H)** with probability 0.4 and the city **(C)** with probability 0.2.
- If you went hiking yesterday, then today you go to the beach **(B)** with probability 0.6, you choose hiking **(H)** with probability 0.1 and the city **(C)** with probability 0.3.
- If you went to the city yesterday, then today you go to the beach **(B)** with probability 0.4, you choose hiking **(H)** with probability 0.5 and the city **(C)** with probability 0.1.

Model your activities by a Markov chain. Give the corresponding transition matrix.

b) If you are hiking today, what is the probability to be on the beach on the day after tomorrow?

c) What fraction of the time do you spent in the city in the long run? Towards this end, compute the stationary distribution of the chain.

d) A Markov chain is said to be *symmetric* if its transition matrix is symmetric. What is the stationary distribution of a connected symmetric chain? Prove your answer.

Problem 4 (Map-Reduce Algorithms) 4+4+9 = 17 points

Let G be the internet graph and suppose for simplicity the vertices (i.e. the websites) are numbered from 1 to n. In this exercise we consider MapRecude algorithms where the initial input of the first phase are tuples of the form $(siteA, linksA)$ where $siteA$ is a website and $linksA$ is a list of websites $siteB$ such that there is a link from $siteA$ to $siteB$.

a) Consider the following MapReduce algorithm for computing the set of all key-value pairs $(\{siteA, siteB\}, comAB)$ where $siteA$ and $siteB$ are websites and $comAB \neq \emptyset$ are the common links of the two websites.

Map on input $(siteA, linksA)$ emit $(\{siteA, siteB\}, linksA)$ for all sites $siteB$

Reduce on input $(\{siteA, siteB\}, val)$ emit $(\{siteA, siteB\}, comAB)$ if $comAB \neq \emptyset$ where $comAB$ is the intersection of all sets $links \in val$.

Determine the communication cost of the algorithm by counting the number of input key-value pairs for both phases (in the worst-case).

b) Give a MapReduce algorithm that outputs all pairs $(siteA, inLinksA)$ where $siteA$ is a website and $inLinksA$ is a list of all incoming links, i.e. websites $siteB$ such that there is a link from $siteB$ to $siteA$.

c) Let d be the maximum degree of the internet graph (counting incoming and outgoing edges). It is reasonable to assume that d is much smaller than n, the total number of websites. Use part b) to give an improved algorithm for the problem from part a) taking the maximum degree d into account. Again, analyse the communication cost. Compare it to your results from part a).

Problem 5 (Streaming Algorithms) 4+6+4+3+5 = 22 points

a) Consider the stream $a = 1, 4, 4, 1, 5, 3, 8, 2, 2, 1, 5$ over the universe $\mathbb{U} = \{1, \ldots, 10\}$. Compute the p-th frequency moment for $p = 0, 1, 2$, i.e. give $F_0(a)$, $F_1(a)$ and $F_2(a)$.

b) Apply the Tug-Of-War estimator to the stream of part a) using the hash $h\colon \mathbb{U} \to \{-1, 1\}$ given below. Give the result as well as some some intermediate steps showing how the estimator is computed from a.

u	1	2	3	4	5	6	7	8	9	10
$h(u)$	1	1	-1	1	-1	1	1	1	-1	-1

c) What are the requirements on the family of hash functions for the Tug-Of-War estimator? What guarantee on the output does the Tug-Of-War estimator give?

d) The Tug-Of-War estimator itself typically does not give very good results. What is the problem with the Tug-Of-War estimator?

e) In the lecture we have discussed an extension of the Tug-Of-War estimator which avoids the problems discussed in part d). How does this extension work? What additional guarantee can we get for the extended version?

Empty Page for Notes