Mock Exam for Data Science, PHYS-231, 2023/24

January 10, 2024

1 PageRank & Graphs [8 points]

1. (2 points) Recall the adjacency matrix we defined in the lecture for a directed graph with n nodes:

$$A_{ij} = \begin{cases} 1 & \text{if website } j \text{ links towards website } i \\ 0 & \text{otherwise} \end{cases} . \tag{1}$$

   a) How can you determine the in-degree and out-degree of a node i from its corresponding row and column in the adjacency matrix A?

   b) For what type of adjacency matrices is the use of a np.array matrix suboptimal, as a datastructure in your algorithm?

2. (4 points) How can you use a given adjacency matrix to find the number of directed triangles in a graph? The existence of a directed triangle on the internet means that starting on website i, a surfer can return to the starting website by clicking three links (e.g. ”Physics→Universe→Time→Physics” on wikipedia). Sketch the idea of an algorithm that would count these triangles, and explain why it works. Recall our definition of the matrix S which we used for Page Rank in Lecture 1:

$$S_{ij} = \begin{cases} \frac{A_{ij}}{d_j} & \text{if } d_j \geq 1 \\ \frac{1}{n} & \text{if } d_j = 0 \end{cases} , \tag{2}$$

with

$$d_j = \sum_{i=1}^{n} A_{ij} \, . \tag{3}$$

Hint: What conditions would the matrix A fullfill, if there existed a directed triangle between the nodes i, j and k?

3. (2 points) Prove that S is *column-wise stochastic*, i.e.

$$\sum_{i=1}^{n} S_{ij} = 1 \quad \text{for} \quad i = 1, \ldots, n\,.$$

2 Low-rank approximation [10 points]

We consider a $n \times d$ matrix A with real components. For $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^d$ we define the loss

$$L(u,v) = \frac{1}{2}||A - uv^T||_F^2 = \frac{1}{2}\sum_{i,j}(A_{ij} - u_i v_j)^2 \ . \tag{4}$$

1. (1 point) What is the rank of uv^T ? How do you interpret minimizing L ? (answer in one or two lines)

2. (2 points) Based on the course, describe what is the minimizer (u^*, v^*) of L? (aswer in two or three lines)

3. (2 points) Compute $\nabla_u L \in \mathbb{R}^n$, the differential (or gradient) of L with respect to u (you can start computing its derivative with respect to the scalar u_i for a i). Then compute also $\nabla_v L \in \mathbb{R}^d$.

4. (2 points) Assume $(u,v) \neq (0,0)$ is a minimizer of L, i.e. $\nabla_u L = 0$ and $\nabla_v L = 0$. What does it imply in terms of singular elements ? What is the corresponding singular value ?

5. (3 points) Assume (u,v) is a minimizer of L and (u',v') too. Assume they are othogonal, i.e. $u^T u' = v^T v' = 0$. Compute $L(u,v) - L(u',v')$ (you may use that, for a matrix M, $||M||_F^2 = \mathrm{tr}(MM^T)$). Which of these two minima $L(u,v)$ or $L(u',v')$ is lower? Comment on the global minimum of L.

3 Gradient Descent and Validation [5 points]

1. (2 points) Consider regression that has the goal of predicting labels on new datapoints. Why do we use a validation and a test dataset in the analysis?

2. (2 points) Gradient descent often requires selecting a learning rate. What is the learning rate, and what are the trade-offs in choosing a learning rate that is too small or too large?

3. (1 point) Why is it generally not a good idea to use the step function

$$\theta(x) = \begin{cases} 0 & \text{if } x > 0 \\ 1 & \text{if } else \end{cases} \tag{5}$$

as a loss for gradient based optimization?

Choose the correct answer and explain your reasoning.

4 Bayes formula [2 points]

1. (2 points) Urn A contains three balls: one black, and two white; urn B contains three balls: two black, and one white. One of the urns is selected at random, urn A with probability p, urn B with probability $1 - p$, and one ball is drawn. The ball is black. What is the probability that the selected urn is urn A?

5 Markov inequality [3 points]

1. (3 points) Markov inequality states that $\forall a > 0$ we have that $\text{Proba}(X \geq a) \leq \frac{\text{E}(X)}{a}$. Provide a proof.

6 Maximum likelihood estimators [7 points]

1. (2 points) Suppose that some data $\mathscr{D}$ (that we know) has been generated through a known probabilistic process $\mathscr{D} \sim \rho(\cdot|\theta_*)$ depending on some unknown parameters θ_*. Define the Maximum-Likelihood Estimator (MLE) for this problem.

2. (3 points) Suppose that the dataset $\mathscr{D}$ is composed of n observations $x_i \in \mathbb{R}$, each one generated independently through a known probabilistic process $x_i \sim \rho(\cdot|\theta_*)$ depending on some unknown parameters θ_*. As $n \to +\infty$, what properties does the MLE satisfy? You are free to assume that ρ is very well behaved, and satisfies any nice property you may want to impose.

3. (2 points) In the class, we derived the following expression for the weighted average of some measurements x_i and error bars σ_i

$$\hat{\theta} = \frac{\sum_i \frac{x_i}{\sigma_i^2}}{\sum_i \frac{1}{\sigma_i^2}} \,. \tag{11}$$

Write an example of a probabilistic process for sampling the measurements x_i, where their distribution is a function of some unknown parameter θ and the known error bars σ_i. For the process you write, the MLE for the unknown parameter θ should correspond exactly to the expression above.

7 Particle decay constant [12 points]

Consider an experimental system in which a radioactive source placed at the origin of the frame of reference emits unstable particles in the direction of the x axis with constant speed. We detect the radioactive decays with a detector of finite length, that can reveal decays only if they happen in the interval $[x_{\min}, x_{\max}]$.

The probability that the unstable particle decays at a distance X from the source in the interval $[x_{\min}, x_{\max}]$ is distributed according to a truncated exponential distribution, with probability density

$$\rho_{\lambda_*}(x) = \begin{cases} Z(\lambda_*)^{-1}\lambda_*^{-1}e^{-x/\lambda_*} & \text{for } x_{\min} < x < x_{\max} \\ 0 & \text{otherwise} \end{cases}, \tag{12}$$

where

$$Z(\lambda_*) = \lambda_*^{-1}\int_{x_{\min}}^{x_{\max}} dx\, e^{-x/\lambda_*} = e^{-x_{\min}/\lambda_*} - e^{-x_{\max}/\lambda_*}\,. \tag{13}$$

The unknown parameter $\lambda_* > 0$ is called the decay constant. We observe n independent decay events, at positions $\mathscr{D} = \{x_i\}_{i=1}^n$, and want to estimate λ_* from the data.

1. (1 points) Write the probability of observing the dataset $\mathscr{D}$ given a fixed value of λ.

2. (1 points) Write the maximisation problem defining the maximum-likelihood estimator for λ.

3. (3 points) Consider the case $x_{\min} = 0$ and $x_{\max} = +\infty$. Compute explicitly the maximum-likelihood estimator as a function of the decay positions $\{x_i\}_{i=1}^n$.

4. (1 points) Is it always the case (for example, whenever $x_{\min} \neq 0$ and $x_{\max} \neq +\infty$) that the MLE estimator equals the empirical average of the observations?

5. (2 points) Define what an unbiased estimator is, and check whether the MLE estimator for this problem (still in the case $x_{\text{min}} = 0$ and $x_{\text{max}} = +\infty$) is biased or not.

6. (2 points) Consider the general case $x_{\text{min}} \neq 0$ and $x_{\text{max}} \neq +\infty$, and let the number of measurements $n \to \infty$. In this case, it is not possible to compute explicitly the MLE. Can you nonetheless say whether it is biased or not as $n \to \infty$? Assume that any technical assumption you may need is satisfied in this case.

7. (2 points) Consider the MLE estimator you derived in point 3. How is it distributed when $n \to \infty$? How are its fluctuation around the mean distributed as $n \to \infty$?

8 Brownian motion and diffusion [10 points]

Consider a particle in 1 dimension with position $x \in \mathbb{R}$. The particle moves at each discrete time step following the equation

$$x(t+\tau) = x(t) + \Delta \tag{25}$$

where Δ is a random variable with probability density function $\phi(\Delta) : \mathbb{R} \to \mathbb{R}$.

1. (1 point) Let $\rho(x,t)$ be the probability density of x at time t. Write the discrete time evolution equation of $\rho(x,t)$, i.e. give an expression for $\rho(x, t+\tau)$.

2. (2 points) In the context of Brownian motion, give two standard assumptions about the random variable Δ.

3. (2 points) Using the two previous points, derive the equation

$$\rho(x,t+\tau) = \rho(x,t) + \frac{1}{2}\frac{\partial^2 \rho}{\partial x^2}(x,t)\mathrm{Var}(\Delta) + o(\Delta^2) \tag{27}$$

by expanding $\rho(x-\Delta)$ to second order.

4. (2 points) Expand $\rho(x, t+\tau)$ at first order in τ and deduce the diffusion equation as well as the diffusion coefficient.

5. (3 points) Apply the central limit theorem to $x(t = n\tau) = \sum_{i=1}^{n} \Delta_i$ with $\Delta_i \sim \phi(\Delta)$ i.i.d. with n large to find a solution to the diffusion equation.

9 Maximum entropy distribution, mean [5 points]

(5 points) Consider a non-negative random variable $x \in \mathbb{R}$. We observe its mean $\mu \geq 0$ from the data. Derive the probability distribution $P(x)$ that maximizes the entropy given the constraint that the mean is equal to the observed ones.

10 Sampling with Monte Carlo Markov Chains [8 points]

The goal in this exercise is to sample from the hexagonal grids below, using a random walk where the transitions are restricted to only adjacent cells (analogous to sampling from the 3x3 grid from the lecture).

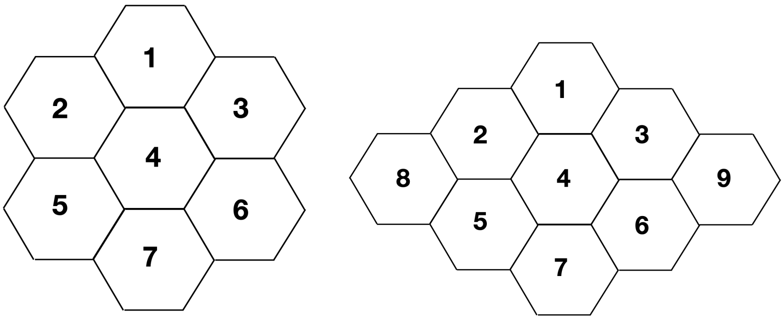


Figure 1: Hexagonal grids.

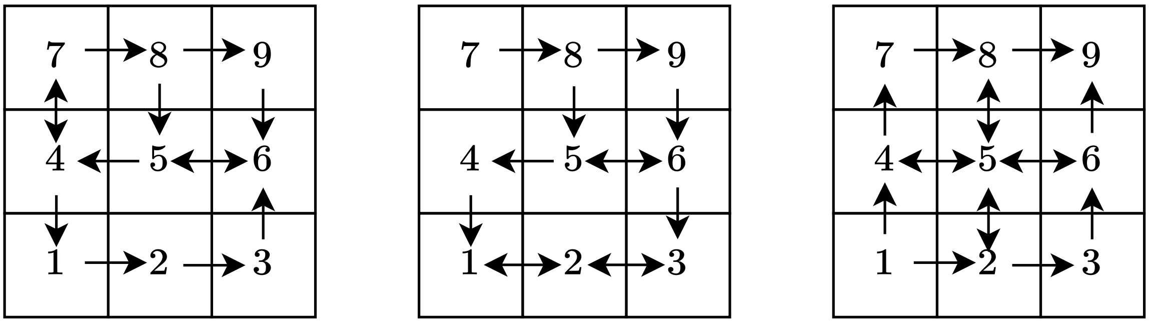
1. (2 points) Consider the hexagon grid in Fig. 1 left hand side. The aim is to sample the cells 1,2,3,4,5,6,7 uniformly using a Markov Chain. It is given that the transition probability $p(4 \to 1) = p(4 \to 2) = p(4 \to 3) = p(4 \to 5) = p(4 \to 6) = p(4 \to 7) = 1/6$. Write an example of a Markov chain that satisfies the detailed balance condition and samples uniformly the cells. If such a Markov Chain does not exist, explain why.

2. (2 points) Consider the hexagon grid in Fig. 1 left hand side. The aim is to sample the cells 1,2,3,4,5,6,7 uniformly using a Markov Chain. It is given that the transition probability $p(4 \to 1) = p(4 \to 2) = 1/6$, $p(4 \to 3) = p(4 \to 5) = 1/12$ $p(4 \to 6) = p(4 \to 7) = 1/6$. Write an example of a Markov chain that satisfies the detailed balance condition and samples uniformly the cells. If such a Markov Chain does not exist, explain why.

3. (2 points) Consider the the hexagon grid in Fig. 1 right hand side. The aim is to sample the cells 1,2,3,4,5,6,7,8,9 uniformly using a Markov Chain. It is given that the transition probability can only be non-zero for neighbours, i.e. cells that share an edge. It is also given that $p(4 \to 1) = p(4 \to 2) = p(4 \to 3) = p(4 \to 5) = p(4 \to 6) = p(4 \to 7) = 1/6$. Write an example of a Markov chain that satisfies the detailed balance condition and samples uniformly the cells. If such a Markov Chain does not exist, explain why.

4. (2 points) Consider the the hexagon grid in Fig. 1 right hand side. The aim is to sample the cells 1,2,3,4,5,6,7,8,9 uniformly using a Markov Chain. It is given that the transition probability can only be non-zero for neighbours, i.e. cells that share an edge. It is also given that $p(3 \to 9) = 0$ and $p(9 \to 6) = 0$. Write an example of a Markov chain that satisfies the detailed balance condition and samples uniformly the cells. If such a Markov Chain does not exist, explain why.

11 Markov Chains irreducibility [8 points]



The figure shows three Markov chains on the discrete space $\{1, 2, ..., 9\}$. An arrow between two cells means that the chain can move from one cell to the other in one time step. You can also assume that each state has a certain strictly positive probability to transition to itself (i.e. all transitions $x \to x$ are allowed).

1) (3 points) Which chains are irreducible and which ones aren't?

2) (3 points) For the chains that are not irreducible indicate the set of states in which we find the chain after running it for a long time.

3) (2 points) Explain why the left chain is not periodic