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

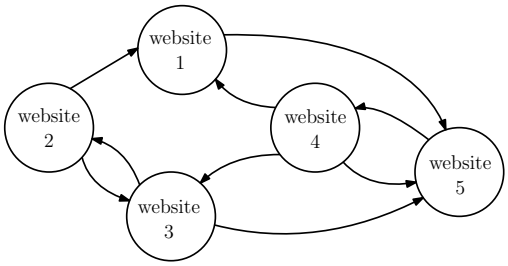
Name/Sciper:

Instructions:

- Duration of the exam: 3 hours, 26. 01. 2024 from 09h15 to 12h15. Rooms CM1105, CM1120.
- Material allowed: 2 pages (i.e. one sheet recto-verso or 2 one-sided sheets) of personal notes. Pen and paper.
- Problems can be solved in any order.
- Write your full name on **each** additional sheet of paper you hand in. You can also use the last page if you need more space.
- Total number of points is 75.

1 PageRank & Graphs [8 points]

1. (2 points) Given the following network of websites, write their adjacency matrix $A \in \{0,1\}^{5\times5}$, where $A_{ij} = 1$ if website j links towards website i (arrow from j to i in figure) and zero otherwise. The i is the row index and j the column index of the matrix.



2. (1 point) Which is the website with the highest out-degree? Which is the website with the highest in-degree?

3. (1 point) When is a matrix column stochastic?

4. (1 point) Write the column stochastic version of the following adjacency matrix A where the entries are normalized by the out-degree of the corresponding website.

$$A = \begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \end{bmatrix}$$

5. (2 points) Recall that the Google Matrix in PageRank was defined as

$$G = (1 - \epsilon)S + \epsilon I \tag{1}$$

where I is the matrix with all entries equal to $1/n$, $\epsilon \in (0, 1)$. Explain why we add the second term of the sum and need the weighting by $1 - \epsilon$ and ϵ.

6. (1 point) You run the page rank algorithm on the subset of Wikipedia of all websites including the keyword physics, and you retrieve the list of page rank values for the websites. Is the website which is the most relevant according to page rank the one with the smallest or largest value?

2 SVD [8 points]

1. (2 points) Consider a matrix $X \in \mathbb{C}^{n \times d}$. Define the Singular Value Decomposition (SVD) of X.

2. (2 points) Consider a matrix $X \in \mathbb{C}^{n \times d}$. Define what are left and right singular vectors. How are they related to the SVD of X?

3. (2 points) Consider a matrix $X \in \mathbb{C}^{n \times d}$. What is its best rank-$k$ approximation under the mean-squared error norm (a.k.a. Frobenius norm)?

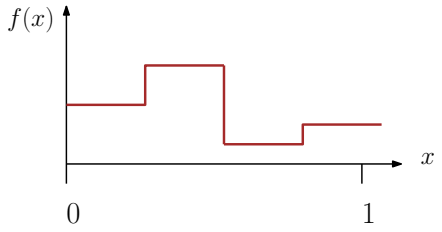
4. (2 points) Suppose that the matrix $X \in \mathbb{R}^{n \times d}$ represents a dataset of points. The dataset contains n points, each of which is d-dimensional. Call $X = U\Sigma V^*$ the SVD of X. What is the interpretation of the first k right singular vectors $v_1, \ldots v_k$ for $k \leq d$?

3 Gradient Descent [4 points]

1. (1 point) Write the equation for one step of the gradient descent algorithm that aims to minimize a function $L(w)$ over $w \in \mathbb{R}^d$.

2. (2 points) Consider the learning rate in the gradient descent algorithm. Describe one inconvenience of taking a very small learning rate. Describe one inconvenience of taking a very large learning rate.

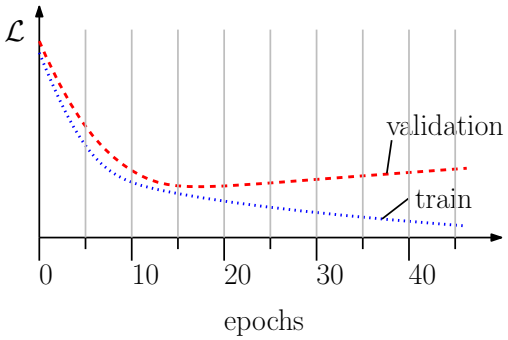
3. (1 point) Why is it not a good idea to use the following function as loss function for gradient-based optimization?



4 Training, validation, and test error [6 points]

1. (2 points) When we talked about overfitting we described the use of the validation, training and test sets. It was important that the samples in the dataset are split at random in order to create these three sets. What could go wrong if we took the first 1/3 of the samples into the training set, the second third into the validation set and the last third into the test set?

2. (1 point) You are training a classification model to classify cats and dogs using gradient descent, and you are observing the following loss curves for the training and validation set over time:



Is the model overfitting at epoch 45? Explain your answer.

3. (2 points) Following the previous question, you have saved the model parameters for every epoch during training. From these saved parameters, you want to choose the one that is best at classifying cats and dogs. Explain what early stopping is, and choose the epoch in the plot (to an accuracy of 5 epochs), which you should use for prediction according to early stopping.

4. (1 point) The training and validation accuracy for the model you chose in the previous part are 98% and 95%, respectively. Should you use one of these numbers to describe how well your model generalizes to new, previously unseen images?

5 Bayes formula [4 points]

1. (2 points) Urn A contains five balls: one black, two white, one green and one pink; urn B contains five hundred balls: two hundred black, one hundred white, 50 yellow, 40 cyan, 30 sienna, 25 green, 25 silver, 20 gold, and 10 purple. [One fifth of A's balls are black; two-fifths of B's are black.] 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 urn is urn A?

2. (2 points) The inhabitants of an island tell the truth one third of the time. They lie with probability 2/3. On an occasion, one of them (called Alice) tells you a statement. You ask another of them (called Bob) 'Did Alice tell the truth?' and Bob answers 'Yes'. What is the probability that Alice told the truth?

6 Uncertainty Propagation and Probability [7 points]

1. (3 points) Provide a proof of the Chebyshev inequality that states the following: Let $\rho(x)$ be the p.d.f. of a random variable X with finite mean and variance. Then

$$\text{Proba}\left(|X - \mathbb{E}(X)| \geq l\sigma_X\right) \leq \frac{1}{l^2} \text{ with } l \in \mathbb{R}, l > 0 \tag{4}$$

where σ_X is the standard deviation of X. Hint: You are allowed to use the Markov inequality that states that $\forall a > 0$ we have that $\text{Proba}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}$.

2. We discussed two formulas for the propagation of error that are often used in experimental physics.

$$\sigma_G = \sum_{i=1}^{k} \left| \frac{\partial G(X_1, ..., X_k)}{\partial X_i}\bigg|_{X_j = \mathbb{E}(X_j)\,\forall j} \right| \sigma_{X_i} \tag{5}$$

and

$$\sigma_G^2 = \sum_{i=1}^{k} \left[\frac{\partial G(X_1, ..., X_k)}{\partial X_i}\bigg|_{X_j = \mathbb{E}(X_j)\,\forall j} \right]^2 \sigma_{X_i}^2 \tag{6}$$

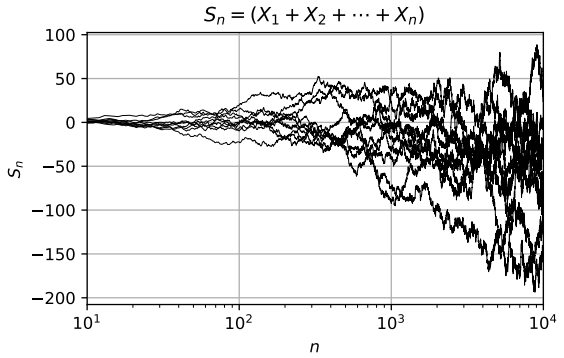
   (i) (1 point) What is the main assumption that needs to be true for both of them?

(ii) (2 points) What is the main assumption to check when deciding which of the two formulas to use?

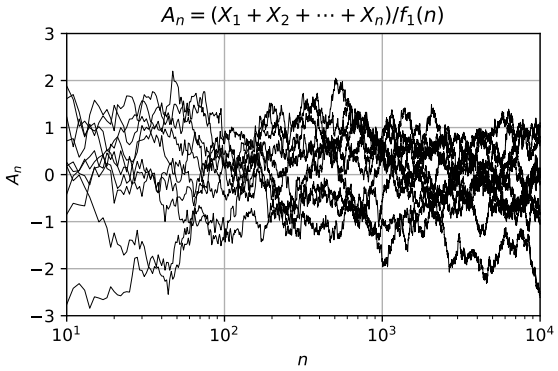
(iii) (1 point) Which of the two formulas gives smaller σ_G, especially when k is large?

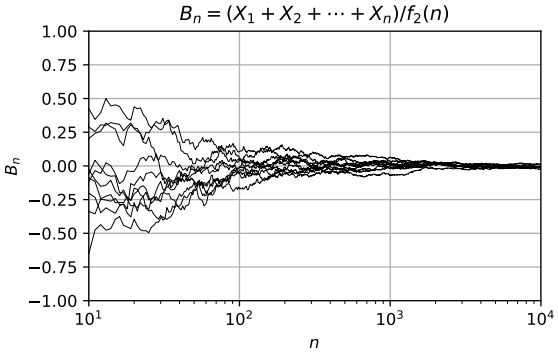
7 Central Limit Theorem and Law of Large Numbers [6 points]

We sample $X_i \sim \mathcal{N}(0,1)$ i.i.d. from the normal distribution. We define $S_n = X_1 + X_2 + \cdot + X_n$. This is a random walk, as you can see in the figure below.



Your friend shows you two other plots, where they visualized the convergence of the Law of Large Numbers and the Central Limit Theorem as n grows to infinity. They remember that they plotted the scaled versions of S_n, namely $A_n = S_n/f_1(n)$ and $B_n = S_n/f_2(n)$ for each of the laws, but they forgot what the function $f_1(n) : \mathbb{R} \to \mathbb{R}$ and $f_2(n) : \mathbb{R} \to \mathbb{R}$ were.





1. (4 points) Explain for each plot if the convergence behaviour shown is linked to the law of large numbers or the central limit theorem. For your explanation define the concrete $f_1(n)$ and $f_2(n)$ your friend must have used as a function of n.

2. (2 points) The central limit theorem states that the sum of many random variables is distributed as a Gaussian. What are the two assumptions on the random variables for this to hold.

8 Maximum likelihood for Laplace distribution [6 points]

Consider a random variable taken from the Laplace distribution

$$\rho(x) = \frac{1}{2b} e^{-\frac{|x-\mu|}{b}} \tag{7}$$

where $\mu, b \in \mathbb{R}$, $b > 0$.

1. (3 points) Consider that one observed n independent samples from this distribution x_i, $i = 1, \ldots, n$. Use the maximum likelihood method to estimate the constant μ.

2. (3 points) Consider that one observed n independent samples from this distribution x_i, $i = 1, \ldots, n$. Use the maximum likelihood method to estimate the constant b.

9 Predicting success or failure [12 points]

We will consider the following model for predicting whether a student μ will pass an exam or not. Consider we have a database of n previous students, and for each of them, we collected information about how long they studied, how many lectures they attended in person, what their grades were in previous years in other lectures, how many hours they slept before the exam etc. We gathered all this information in a d-dimensional vector $\vec{X}_\mu$. A matrix $X \in \mathbb{R}^{n\times d}$ then gathers all this information from all the n past students.

We also know which students passed the exam, we denote $y_\mu = +1$, and which of them failed, denoted as $y_\mu = -1$. All together the data we observe is the matrix $X \in \mathbb{R}^{n\times d}$ and the vector $y \in \mathbb{R}^n$.

We further assume that there exists a vector of parameters $\vec{w}^* \in \mathbb{R}^d$ (that is not known to us) such that the probability of a given student μ passes the exam ($y_\mu = 1$) or not ($y_\mu = -1$) is given by

$$P_{\text{exam}}(y_\mu|\vec{w}^*, \vec{X}_\mu) = \frac{\exp\left(y_\mu \sum_{i=1}^d X_{\mu i} w_i^*\right)}{\exp\left(\sum_{i=1}^d X_{\mu i} w_i^*\right) + \exp\left(-\sum_{i=1}^d X_{\mu i} w_i^*\right)} \,. \tag{14}$$

Different students passing or not are considered independent random variables, all conditional to the same vector $\vec{w}^*$.

1. (2 points) Sketch the probability that the student μ passes the exam as a function of the parameter $z_\mu = y_\mu \sum_{i=1}^d X_{\mu i} w_i^*$.

2. (2 points) Write the quantity that needs to be maximized for the maximum likelihood estimation of the parameters w.

3. (2 points) Maximizing the likelihood is equivalent to minimization of a loss, along the same lines as what we did in the lecture when we described the probabilistic derivations of the least-squares loss. Write the loss function we obtain in the present case (Hint: it should be a sum over the students.).

4. (2 points) Write here the loss function used in logistic regression that we covered in the lecture.

5. (2 points) Show that the loss function of the exam-passing problem above is a special case of the logistic regression.

6. (1 point) Which algorithm would you use to minimize this loss function? (Just give the name.)

7. (1 point) Once the minimizer $\hat{\vec{w}}$ of the loss is obtained, one can predict whether a new student for whom you have the data $\vec{X}_{\text{new}} \in \mathbb{R}^d$ will pass the exam or not. Write the corresponding predictor, i.e. a function from $\vec{X}_{\text{new}} \to \pm 1$.

10 Monte Carlo Markov Chains [6 points]

1. (3 points) Consider a Markov chain on a state space X defined by the transition probability $p(a \to b)$ of going from state $a \in X$ to state $b \in X$ at each time step. Consider a probability distribution on the states space $\pi(a)$, and suppose that it satisfies the condition

$$\pi(a)p(a \to b) = \pi(b)p(b \to a) \, \forall a, b \in X \, . \tag{17}$$

What is this condition called? Then, prove that π is a stationary distribution for the Markov chain, i.e.

$$\sum_a \pi(a)p(a \to b) = \pi(b) \, . \tag{18}$$

2. (3 points) Which of the following algorithms would you use to sample points uniformly in the d-dimensional disk $D_d = \{x \in \mathbb{R}^d \text{ such that } ||x|| \leq 1\}$?

   (a) Direct sampling: generate points uniformly in $[-1, 1]^d$, and reject all samples that fall utside D_d.

   (b) MCMC: random walk inside the disk, and every time a step would make you exit from the disk, do not accept the step (but still keep as a sample the point where you are).

Motivate briefly, distinguishing the case of low dimension, e.g. $d \lessapprox 10$, and large dimension, e.g. $d \gtrapprox 10$.

11 Sampling with Monte Carlo Markov Chains - triangles [8 points]

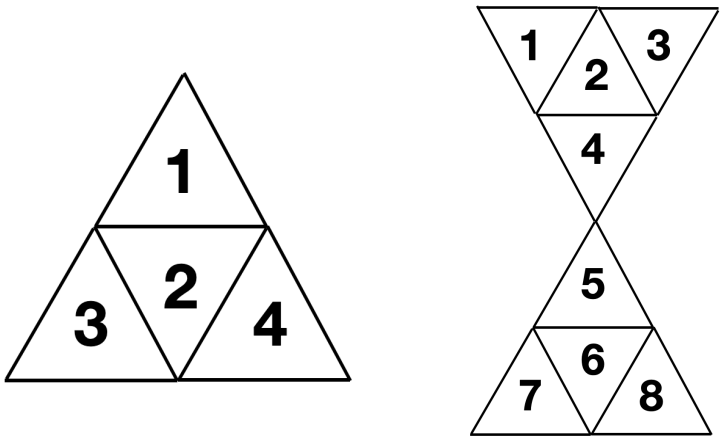


Figure 1: Triangle grids.

1. (2 points) Consider the triangle grid in Fig. 1 left hand side. The aim is to sample the cells 1,2,3,4 uniformly using a Markov Chain. It is given that the transition probability $p(2 \to 1) = p(2 \to 4) = p(2 \to 3) = 1/3$. 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 triangle grid in Fig. 1 left hand side. The aim is to sample the cells 1,2,3,4 uniformly using a Markov Chain. It is given that the transition probability $p(2 \to 1) = p(2 \to 4) = p(2 \to 3) = 1/3$, $p(1 \to 4) = p(4 \to 3) = p(1 \to 3) = 1/2$. 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 triangle grid in Fig. 1 right hand side. The aim is to sample cells 1,2,3,4,5,6,7,8 uniformly using a Markov Chain. It is given that the transition probability can only be non-zero for neighbours i.e. only between $(1, 2)$, $(3, 2)$, $(2, 4)$, $(5, 6)$, $(6, 7)$, $(6, 8)$. 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 triangle grid in Fig. 1 right hand side. The aim is to sample cells 1,2,3,4,5,6,7,8 uniformly using a Markov Chain. It is given that the transition probabilities $p(2 \to 1) \geq 1/4$, $p(2 \to 3) \geq 1/4$, $p(2 \to 4) \geq 1/4$. 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.