

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

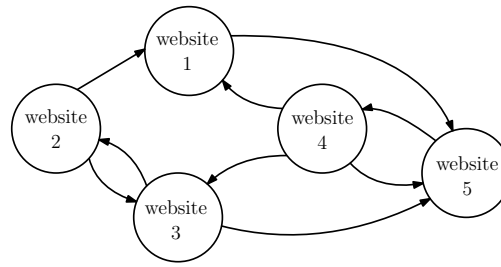
Name/Sciper:

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 $A \in \{0, 1\}^{5 \times 5}$, 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.



Solution:

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

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

Solution: Highest in-degree: Website 5
Highest out-degree: Website 4

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

Solution: A matrix is column-stochastic, if the entries of every column are ≥ 0 and sum to one.

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.

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

Solution:

$$S = \begin{bmatrix} 0 & 1 & 1/3 & 1/4 & 0 \\ 0 & 0 & 1/3 & 1/4 & 0 \\ 1/2 & 0 & 0 & 1/4 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1/2 & 0 & 1/3 & 1/4 & 0 \end{bmatrix}$$

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

$$G = (1 - \epsilon)S + \epsilon I \quad (1)$$

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 .

Solution: We would never visit websites that do not have incoming links. In terms of a random walk and MCMC context, not having ϵ would mean that the Random Surfer, which is a Markov chain, would not be ergodic and would hence not be guaranteed to converge to a stationary distribution. It would get stuck on connected components of the graph.

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?

Solution: The one with the 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 .

Solution: An SVD of X is a triple of matrices $U \in \mathbb{C}^{n \times n}$, $V \in \mathbb{C}^{d \times d}$ and $\Sigma \in \mathbb{C}^{n \times d}$ such that

- $X = U\Sigma V^*$.
- U and V are unitary, i.e. $UU^* = U^*U = \mathbb{I}$ and similarly for V .
- Σ is diagonal, and its diagonal elements are real and non-negative, and conventionally ordered in decreasing order.

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 ?

Solution: Two vectors $u \in \mathbb{C}^n$ and $v \in \mathbb{C}^d$ are a pair of left-right singular vectors of X if there exists a $\sigma \in \mathbb{C}$ such that

$$Av = \sigma u \quad \text{and} \quad A^*u = \sigma v. \quad (2)$$

Calling $X = U\Sigma V^*$ an SVD of X , we have that the pairs $\{u_a, v_a\}$, where u_a is the a -th row of U and similarly for v , are pairs of left-right singular vectors.

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)?

Solution: The best rank- k approximation of X under the mean-squared error norm is given by

$$\tilde{X} = U\Sigma_{\leq k}V^* \quad (3)$$

where $X = U\Sigma V$ is an SVD of X , and $\Sigma_{\leq k}$ is the matrix Σ in which we keep only the largest k singular values, and put to zero the others.

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, \dots, v_k for $k \leq d$?

Solution: The right singular vectors are the principal components of the dataset, and represent the k directions in d -dimensional space along which the datasets features have the largest variability.

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 $w \in \mathbb{R}^d$.

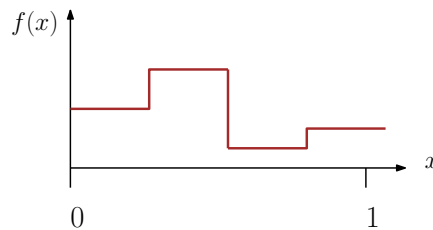
Solution:

$$\bar{w}_i^{t+1} = \bar{w}_i^t - \gamma \frac{\partial \mathcal{L}}{\partial w_i}(\bar{w}^t), \quad i = 1, \dots, d$$

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.

Solution: Slow convergence for small. Instability, jumping over optima or not converging at all for large one.

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



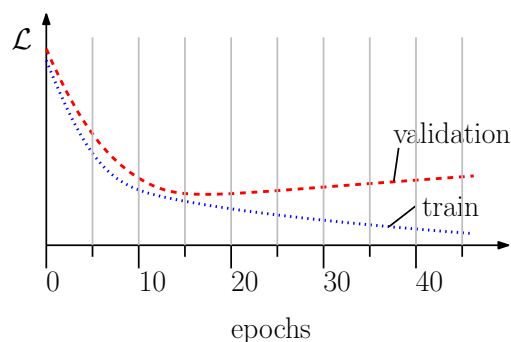
Solution: Its derivative is zero almost everywhere, so the algorithm will not move the model parameters anywhere.

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?

Solution: We would not get a proper estimation of the error. e.g. if the early samples were less noisy than the later ones or the other way round. Samples need to be iid so these 3 sets have comparable statistical properties.

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.

Solution: Yes, because the validation loss increases as the training loss decreases.

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.

Solution: Following the previous question, you should select the parameters from epoch 15, since this is the lowest point of the validation loss and the model is not yet overfitting.

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?

Solution: No, because we used the validation dataset in the process of the early stopping. We would need a separate test set.

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?

Solution:

$$p/(2 - p)$$

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?

Solution: $1/5$

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

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

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 $\text{Proba}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}$.

Solution: See course notes.

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

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

and

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

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

Solution: Linear approximation must be accurate in the range of the standard deviation of X .

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

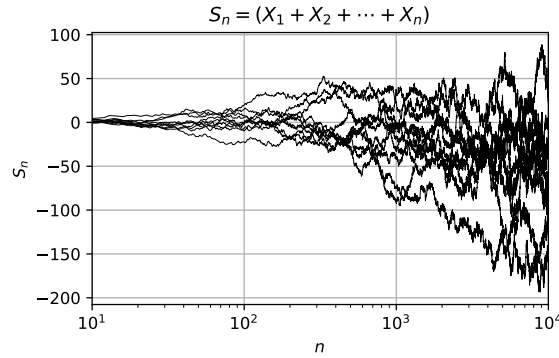
Solution: The 1st is an upper bound for correlated variables. The 2nd needs independent variables. Thus, the assumption to check is the independence of the random variables X_i .

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

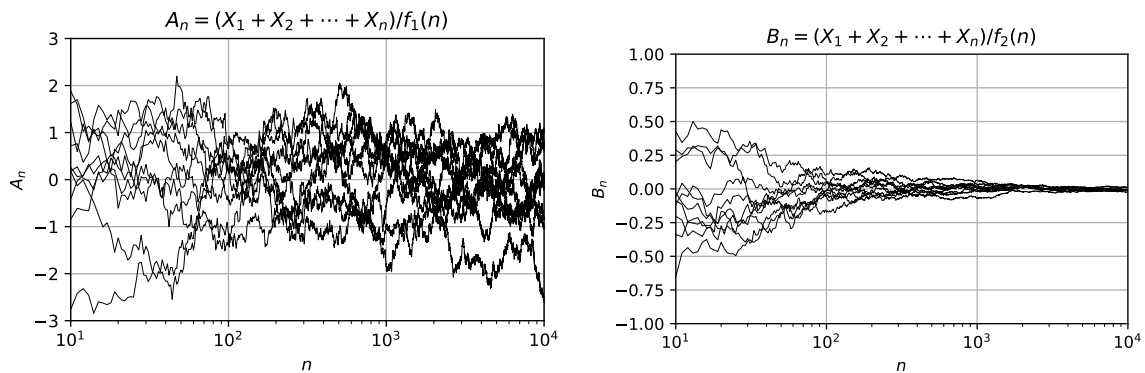
Solution: The 2nd, where variances are combined.

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 $S_n = X_1 + X_2 + \dots + X_n$. 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 $f_1(n) : \mathbb{R} \rightarrow \mathbb{R}$ and $f_2(n) : \mathbb{R} \rightarrow \mathbb{R}$ were.



- (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 .

Solution: On the left we see the central limit theorem, where $f_1(n) = \sqrt{n}$ and we see convergence to , as $n \rightarrow \infty$. On the right side we see the law of large numbers, where $f_2(n) = n$ and we see convergence to 0 (which is the expected value of X_i).GP: it could also be $n^{0.7}$ (for which lln holds), to make the power recognizable one should look at the slope in log-log scale

- (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.

Solution: Independence and finite second moment.

8 Maximum likelihood for Laplace distribution [6 points]

Consider a random variable taken from the Laplace distribution

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

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

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

Solution: The negative, normalised log-likelihood is

$$\ell(\mu, b) = \log(2b) + \frac{1}{bn} \sum_i |x_i - \mu| \quad (8)$$

Then

$$\hat{\mu} = \min_{\mu} \ell(\mu, b) \quad (9)$$

and

$$\partial_{\mu} \ell(\mu, b) = 0 \implies \sum_i \text{sign}(x_i - \mu) = 0 \quad (10)$$

i.e. $\hat{\mu}$ is the median of the samples.

Grading (Vittorio):

- 1 point for writing the likelihood and the optimisation problem in any correct form, max/min, log/non-log, normalised/non-normalised etc
 - -0.5 point if they commute abs with sum
 - -0.75 if likelihood is not summed over samples
- 1 point for deriving the expression of $\hat{\mu}$ (only 0.25 pt if they just say that they compute derivative and set to zero)
 - -0.5 points for wrong derivative or wrong maximisation in general
 - -0.5 for claiming that at $n = \infty$ the MLE is the average
- 0.5 point for recognising that $\hat{\mu}$ is the median
- 0.5 points for discussing if the MLE found is a max or min, and if it the global one
- -0.25 point for typos

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

Solution:

$$\hat{b} = \min_{\mu} \ell(\mu, b) \quad (11)$$

and

$$\partial_b \ell(\mu, b) = 0 \implies b = \frac{1}{n} \sum_i |x_i - \mu| \quad (12)$$

If we also optimise over μ , we get

$$b = \frac{1}{n} \sum_i |x_i - \hat{\mu}| \quad (13)$$

as the minimisation over μ is independent on b .

Grading (Vittorio):

- 1 point for writing the likelihood and the optimisation problem in any correct form, max/min, log/non-log, normalised/non-normalised etc Ok to say that it is the same as above
 - not penalised if they use same likelihood as above, even if wrong, and they stay consistent with this later
- 1 point for deriving the expression of \hat{b} (only 0.25 pt if they just say that they compute derivative and set to zero)
- 0.5 point for recognising that they can use $\mu = \hat{\mu}$

- *0.5 points for discussing second derivative or any other reasoning on whether it is a max min, global or not*
- *see previous point for other grading conventions*

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}_μ . 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 $y \in \mathbb{R}^n$.

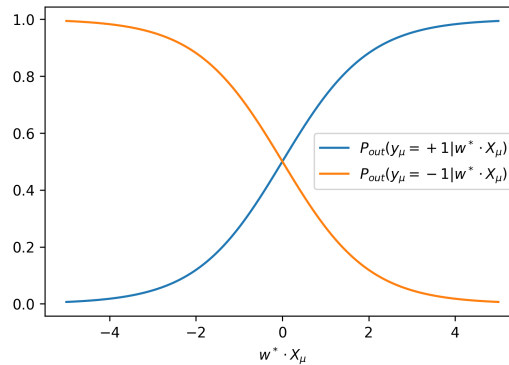
We further assume that there exists a vector of parameters $\vec{w}^* \in \mathbb{R}^d$ (that is not known to us) such that the probability of a given student μ passes the exam ($y_\mu = 1$) or not ($y_\mu = -1$) is given by

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

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

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

Solution: Probabilistic model $P_{\text{out}}(y_\mu | w^* \cdot X_\mu)$ leading to the logistic regression.



Grading (Vittorio):

- 0 points if plot gives probability < 0 or > 1 .
- 0.25 point for writing the probability as a function of z . If plot is perfect, award also these points
- 1.75 points for complete plot
 - -0.75 points if plot is only done for $z > 0$
 - -0.25 if plot is not strictly monotone
 - -0.5 for missing axis labels, y axis scale etc
 - -0.25 if wrong value in 0
 - 0 pts if wrong shape

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

Solution:

$$\prod_{\mu=1}^n P_{\text{exam}}(y_\mu | \vec{w}, \vec{X}_\mu)$$

Grading (Vittorio):

- 2 points for correct formula
- 0.5 points if \prod_μ missing
- 0 points if they just say that they need to maximise some undefined or not specified quantity ρ etc
- -1 point if they substitute sum over mu with power!

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.).

Solution:

$$L(\vec{w}) = \sum_{\mu=1}^n \log \left(1 + \exp \left(-2y_{\mu} \sum_{i=1}^d X_{\mu i} w_i \right) \right)$$

Grading (Vittorio):

- 1.75 points for any correct formula, if not logistic
- 2 points for logistic loss
- -1 point for correct but missing log
- -0.5 points if correct but with additional strange comments that are not correct
- -0.25 for typos
- -0.25 points if they max instead of min
- -0.25 points for writing the L2 loss and then ignoring it and giving the correct answer
- 0 points for L2
- 0 points if they rewrite point 2 without doing at least the log
- 0 points if they just transform product into sum
- 0 point if missing sum over mu from their reasoning

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

Solution:

$$L(\vec{w}) = \frac{1}{n} \sum_{\mu=1}^n \ell(y_{\mu}, z_{\mu}) + \frac{\lambda}{n} \sum_{i=1}^d w_i^2, \quad (15)$$

$$z_{\mu} = \sum_{i=1}^d X_{\mu i} w_i. \quad (16)$$

with $\ell(yz) = \log(1 + e^{-yz})$

Grading (Vittorio):

- 2 points for any correct formula
- 0.5 points if they state the per-sample loss without summing over mu or adding a regularisation
- 0.5 points if state generic loss minimisation but do not specify ell
- -0.25 if regularisation is missing
- 0 points for L2

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

Solution: Indeed $\lambda = 0$ and the constant 2 can be absorbed in the weights.

Grading (Vittorio):

- 2 points for correct answer
- 1 point if they kind of answer, mind do not address precisely the mapping
- 0.5 point if they were on the good road
- 0 if they do not address a comparison between the loss they find and another loss (the logistic)
- -0.25 missing regularisation matching. if pt 4 is given without lambda, remove 0 point for missing regularisation

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

Solution: Gradient Descent. *Grading (Vittorio):*

- 1 points for any numerical minimisation alg
- -0.5 if they say they can use least squares alternatively to numerical minimisation, and still mention GD
- if they have in mind L2 from before, saying that they minimize by least squares gives 0.25 pts

7. (1 point) Once the minimizer \hat{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 $\vec{X}_{\text{new}} \rightarrow \pm 1$.

Solution:

$$y_{\text{new}} = \text{sign}\left(\sum_{i=1}^d X_{\text{new } i} \hat{w}_i\right)$$

Grading (Vittorio):

- 1 points for correct predictor
- 0.5 for incorrect understanding of sign function
- 0 if they write L2

10 Monte Carlo Markov Chains [6 points]

1. (3 points) Consider a Markov chain on a state space X defined by the transition probability $p(a \rightarrow b)$ 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

$$\pi(a)p(a \rightarrow b) = \pi(b)p(b \rightarrow a) \forall a, b \in X. \quad (17)$$

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

$$\sum_a \pi(a)p(a \rightarrow b) = \pi(b). \quad (18)$$

Solution: The condition that π satisfied is called *detailed balance* condition.

To prove that π is stationary, we have

$$\sum_a \pi(a)p(a \rightarrow b) = \sum_a \pi(b)p(b \rightarrow a) = \pi(b) \sum_a p(b \rightarrow a) = \pi(b) \quad (19)$$

where we used first the detailed balance condition, and then that the transition probability is normalised.

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 outside 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. $d \lesssim 10$, and large dimension, e.g. $d \gtrsim 10$.

Solution: If the dimension is low the two algorithms perform similarly. There is no immediate criterion to choose one over the other.

If the dimension is large, then the direct sampling will perform poorly. In fact, the volume of the disk becomes exponentially small with respect to the volume of the cube as d increases, meaning that the overwhelming majority of points sampled while running direct sampling will be rejected. Thus, MCMC is to be preferred in large dimension.

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

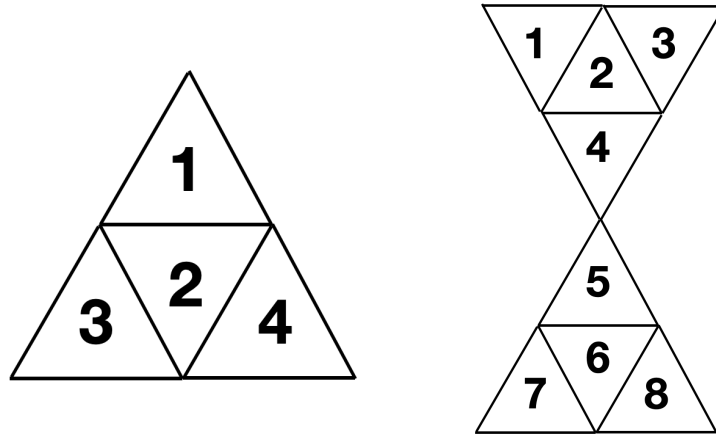


Figure 1: Triangle grids.

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 $p(2 \rightarrow 1) = p(2 \rightarrow 4) = p(2 \rightarrow 3) = 1/3$. 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.

Solution: $p(1 \rightarrow 2) = p(4 \rightarrow 2) = p(3 \rightarrow 2) = 1/3$, $p(1 \rightarrow 1) = p(3 \rightarrow 3) = p(4 \rightarrow 4) = 2/3$ all others zero. Other correct options are also accepted, e.g. those allowing transitions between 1,3; 1,4 or 3,4.

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 $p(2 \rightarrow 1) = p(2 \rightarrow 4) = p(2 \rightarrow 3) = 1/3$, $p(1 \rightarrow 4) = p(4 \rightarrow 3) = p(1 \rightarrow 3) = 1/2$. 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.

Solution: It does not exist. The DB implies $p(1 \rightarrow 2) = p(4 \rightarrow 2) = p(3 \rightarrow 2) = 1/3$, and also $p(4 \rightarrow 2) = p(3 \rightarrow 4) = p(3 \rightarrow 1) = 1/2$ this makes a "probability" to go out of the corner cells larger than 1.

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.

Solution: Not possible because such a Markov chain would not be ergodic.

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 $p(2 \rightarrow 1) \geq 1/4$, $p(2 \rightarrow 3) \geq 1/4$, $p(2 \rightarrow 4) \geq 1/4$. 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.

Solution: $p(2 \rightarrow 1) = p(1 \rightarrow 2) = p(2 \rightarrow 3) = p(3 \rightarrow 2) = p(4 \rightarrow 2) = p(2 \rightarrow 4) = p(6 \rightarrow 5) = p(5 \rightarrow 6) = p(6 \rightarrow 7) = p(7 \rightarrow 6) = p(6 \rightarrow 8) = p(8 \rightarrow 6) = 1/4 + 1/32$, $p(1 \rightarrow 1) = p(3 \rightarrow 3) = p(4 \rightarrow 4) = p(5 \rightarrow 5) = p(7 \rightarrow 7) = p(8 \rightarrow 8) = 1/2 + 1/32$, all others are $1/32$.