CORRECTED

Probabilistic Artificial Intelligence
Questions Pack

January 22, 2025

Time limit: 120 minutes

Instructions. This pack contains all questions for the final exam. It contains the questions only. You must provide your answers to the questions on the accompanying **answer sheet** by blackening out the corresponding squares. Please carefully read the following instructions regarding the answer sheet:

- During the exam, you may use pencil and eraser to mark and edit your answers on the answer sheet. After the time is up, we will collect the questions pack. Make sure that you marked all your answers by then, at least with a pencil. After collecting the questions pack, we provide you with additional **10** minutes of time to blacken out the squares on the answer sheet with a **black pen**.
- As the answer sheet will be graded by a computer, please **make sure to blacken out the whole square. Do not use ticks or crosses**.
- Nothing written on the pages of the question pack will be collected or marked. **Only the separate answer sheet with the filled squares will be marked**.
- Please make sure that your answer sheet is clean and do not write anything on the answer sheet except the squares you blackened out. We reserve the right to classify answers as wrong without further consideration if the sheet is filled out ambiguously.

Collaboration on the exam is strictly forbidden. You are allowed a summary of *two* A4 pages and a simple, non-programmable calculator. The use of any other helping material or collaboration will lead to being excluded from the exam and subjected to disciplinary measures by the ETH Zurich disciplinary committee.

Question Types In this exam, questions award **1, 2, 3, or 4** points if answered correctly, depending on the difficulty of the question. You will encounter the following question types:

- **True or False.**
- **Multiple choice questions with a *single* correct answer.**

  These multiple choice questions have **exactly one** correct choice.
- **Multiple choice questions with *multiple* correct answers.**

  These multiple choice questions are marked with a "♣" and have **one or more** correct choices (at least one choice is correct). You will receive full points if you have chosen **all** the correct choices and **none** of the wrong choices. You will get zero points otherwise.

There are no negative points in this exam, so **0 points** are awarded if a question is answered wrongly or not attempted. The sections and questions are not ordered by difficulty. Thus, if you find a question too difficult, it may make sense to get back to it in the end.

Questions in the same section might be related. In addition, some questions require information from previous parts. Hence, if you intend to skip some questions, make sure that you read all notes and details between the questions.

The notation $\mathcal{N}(\mu, \sigma^2)$ denotes the Gaussian random distribution with mean μ and variance σ^2. For example, $\mathcal{N}(3, 4)$ designates the Gaussian distribution with mean 3 and standard deviation 2.

Good luck!

1 Probability and Regression (11 points)

Question 1 *(1 point)*
For two jointly Gaussian random variables X, Y, the variance of the conditional random variable $X|Y = y$ is always less than or equal to the marginal variance of X.

■ True [B] False

Question 2 *(1 point)*
Consider random variables X, Y, Z, where X and Y are conditionally independent given Z, i.e. $X \perp Y|Z$. Then for all x, y, z it holds that $P(X = x, Y = y|Z = z) = P(X = x|Z = z) \cdot P(Y = y|Z = z)$

■ True [B] False

Question 3 *(1 point)*
For n binary random variables $X_1, \ldots, X_n$, we need $\mathcal{O}(n^2)$ parameters in the worst case to specify the joint distribution $P(X_1, \ldots, X_n)$.

[A] True ■ False

Question 4 *(1 point)*
Let $(X, Y)^\top \sim \mathcal{N}((\mu_X, \mu_Y)^\top, \Sigma)$ with Σ being a diagonal matrix. It is possible that X and Y are not independent.

[A] True ■ False

Question 5 *(2 points)*
Let X and Y be i.i.d. Gaussian random variables distributed according to $\mathcal{N}(0, 1)$. Suppose $Z_1 = X + Y$ and $Z_2 = X - Y$. The random variables Z_1 and Z_2 are independent.

■ True [B] False

1.1 Bayesian Linear Regression (5 points)

We want to predict a target variable $y \in \mathbb{R}$ as a linear function of an input feature vector $\boldsymbol{x} \in \mathbb{R}^2$. We consider a Bayesian linear regression model of the form

$$y = \boldsymbol{x}^\top \boldsymbol{\theta} + \epsilon,$$

where $\boldsymbol{\theta}$ represents the model parameters, and ϵ is i.i.d. Gaussian noise, with a mean of zero and a variance of 4. We choose the prior $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}_\theta = \mathbf{0}, \boldsymbol{\Sigma}_\theta = \boldsymbol{I})$, where $\boldsymbol{I}$ is the identity matrix and $\mathbf{0} \in \mathbb{R}^2$ is the null vector.

Question 6 *(2 points)* What is the distribution of the predicted value y_* for a new input $\boldsymbol{x}_* = \begin{pmatrix} 3 \\ 2 \end{pmatrix}$? No data has been observed yet.

- A $p(y_* \mid \boldsymbol{x}_*) = \mathcal{N}(0, 4)$
- B $p(y_* \mid \boldsymbol{x}_*) = \mathcal{N}(0, 5)$
- C $p(y_* \mid \boldsymbol{x}_*) = \mathcal{N}(0, 13)$
- ■ $p(y_* \mid \boldsymbol{x}_*) = \mathcal{N}(0, 17)$

Question 7 *(3 points)* We observe a dataset $\mathcal{D} = (\boldsymbol{X}, \boldsymbol{y})$ which consists of three data points

$$\boldsymbol{X} = \begin{pmatrix} 3 & 0 \\ 1 & 0 \\ 0 & 2 \end{pmatrix}, \quad \boldsymbol{y} = \begin{pmatrix} 5 \\ 3 \\ 4 \end{pmatrix}.$$

Assuming the same prior distribution on the model parameters as above, what is the posterior distribution of θ conditioned on this data?

- ■ $p(\boldsymbol{\theta}|\mathcal{D}) = \mathcal{N}(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}})$ with $\bar{\boldsymbol{\mu}} = \begin{pmatrix} \frac{9}{7} \\ 1 \end{pmatrix}$ and $\bar{\boldsymbol{\Sigma}} = \begin{pmatrix} \frac{2}{7} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}$
- B $p(\boldsymbol{\theta}|\mathcal{D}) = \mathcal{N}(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}})$ with $\bar{\boldsymbol{\mu}} = \begin{pmatrix} \frac{9}{7} \\ 1 \end{pmatrix}$ and $\bar{\boldsymbol{\Sigma}} = \begin{pmatrix} \frac{1}{11} & 0 \\ 0 & \frac{1}{5} \end{pmatrix}$
- C $p(\boldsymbol{\theta}|\mathcal{D}) = \mathcal{N}(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}})$ with $\bar{\boldsymbol{\mu}} = \begin{pmatrix} \frac{18}{11} \\ \frac{8}{5} \end{pmatrix}$ and $\bar{\boldsymbol{\Sigma}} = \begin{pmatrix} \frac{2}{7} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}$
- D $p(\boldsymbol{\theta}|\mathcal{D}) = \mathcal{N}(\bar{\boldsymbol{\mu}}, \bar{\boldsymbol{\Sigma}})$ with $\bar{\boldsymbol{\mu}} = \begin{pmatrix} \frac{18}{11} \\ \frac{8}{5} \end{pmatrix}$ and $\bar{\boldsymbol{\Sigma}} = \begin{pmatrix} \frac{1}{11} & 0 \\ 0 & \frac{1}{5} \end{pmatrix}$

2 Gaussian Processes (12 points)

Question 8 *(1 point)* The polynomial function $k : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ with $k(x, x') = (1 + xx')^m$ is not a valid kernel for odd m because it can take negative values.

[A] True ■ False

Question 9 *(1 point)*
The mean function $\mu : \mathbb{R}^n \to \mathbb{R}$ of a Gaussian process has to be non-negative: $\mu(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^n$.

[A] True ■ False

Question 10 *(1 point)* Every stationary kernel $k : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ satisfies $k(\boldsymbol{x}, \boldsymbol{x}') = \tilde{k}(\|\boldsymbol{x} - \boldsymbol{x}'\|_2)$ for some function $\tilde{k} : \mathbb{R} \to \mathbb{R}$.

[A] True ■ False

Question 11 ♣ *(4 points)*
Consider an undirected square graph, i.e. the graph $G = (V, E)$ with vertices $V = \{1, 2, 3, 4\}$ and edges $E = \{\{1, 2\}, \{2, 3\}, \{3, 4\}, \{4, 1\}\}$, as depicted below:

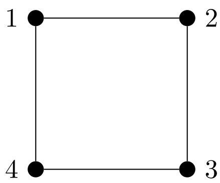


Figure 1: Corresponding to Question 11.

Define the kernel $k_l : V \times V \to \mathbb{R}$ on the vertices of G similarly to the standard squared exponential kernel, by $k_l(x, x') = \exp(-\frac{d(x,x')^2}{2l^2})$, where $d(x, x')$ is the length of the shortest path on G between x and x' and $l \neq 0$. The distance of a vertex from itself, i.e. $d(x, x)$, is zero. **Mark all that apply.**

- ■ For all permissible values of l, k_l is a symmetric function, i.e. $k_l(x, x') = k_l(x', x)$.
- [B] k_l for $l = 1$ is a valid kernel, i.e. a symmetric positive semidefinite function.
- [C] For all permissible values of l, k_l is a valid kernel, i.e. a symmetric positive semidefinite function.

Hint: $(1, -1, 1, -1)^\top$ is an eigenvector of the matrix $\mathbf{K} = \begin{pmatrix} k_l(1,1) & k_l(1,2) & k_l(1,3) & k_l(1,4) \\ k_l(2,1) & k_l(2,2) & k_l(2,3) & k_l(2,4) \\ k_l(3,1) & k_l(3,2) & k_l(3,3) & k_l(3,4) \\ k_l(4,1) & k_l(4,2) & k_l(4,3) & k_l(4,4) \end{pmatrix}$.

2.1 Matching Kernels to Samples from the Posterior (5 points)

In Figure 2, we show samples from posterior Gaussian processes with different kernels fitted to the same dataset, with the same fixed value of observation noise. The kernels $k_i : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ considered are:

- $k_1(x, x') = \exp\left(-\frac{(x-x')^2}{2l^2}\right)$ with $l = 0.1$
- $k_2(x, x') = \exp\left(-\frac{(x-x')^2}{2l^2}\right)$ with $l = 0.4$

- $k_3(x, x') = \exp\left(-\frac{|x-x'|}{l}\right)$ with $l = 1$
- $k_4(x, x') = \exp\left(-\frac{|x-x'|}{l}\right)$ with $l = 10$
- $k_5(x, x') = (1 + xx')^m$ with $m = 2$
- $k_6(x, x') = (1 + xx')^m$ with $m = 10$

Match each of the plots in Figure 2 to the corresponding kernel. Each kernel corresponds to a single plot.

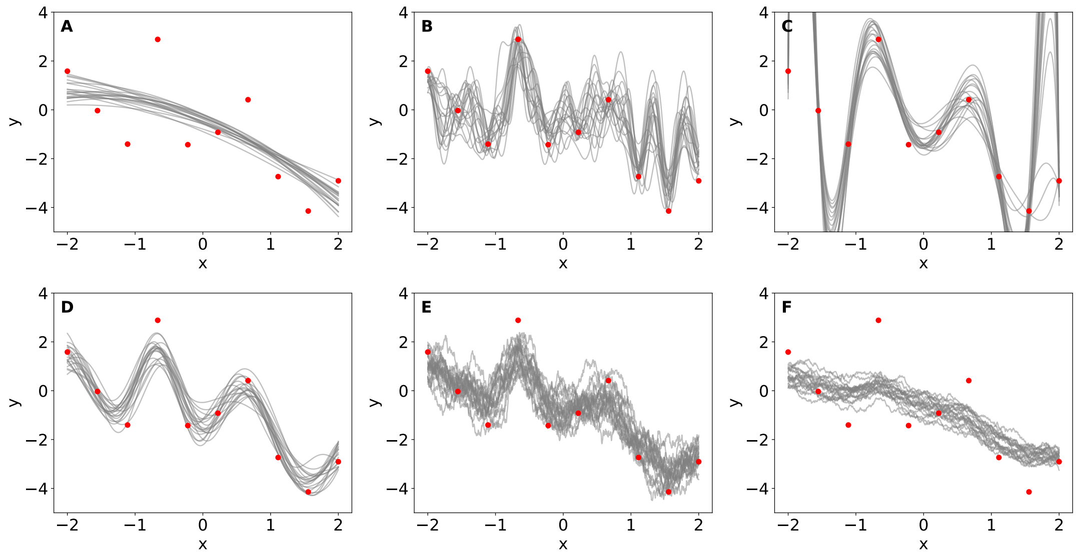


Figure 2: Samples of posterior Gaussian processes for different kernels k. The red dots represent the data.

Question 12 *(1 point)*
Which plot corresponds to k_1, the squared exponential kernel with a smaller length scale?

[A] A ■ B [C] C [D] D [E] E [F] F

Question 13 *(1 point)*
Which plot corresponds to k_2, the squared exponential kernel with a larger length scale?

[A] A [B] B [C] C ■ D [E] E [F] F

Question 14 *(1 point)*
Which plot corresponds to k_3, the exponential kernel with a smaller length scale.

☐ A ☐ B ☐ C ☐ D ■ E ☐ F

Question 15 *(1 point)*
Which plot corresponds to k_4, the exponential kernel with a larger length scale.

☐ A ☐ B ☐ C ☐ D ☐ E ■ F

Question 16 *(1 point)*
Which plot corresponds to k_5, the quadratic polynomial kernel.

■ A ☐ B ☐ C ☐ D ☐ E ☐ F

3 Variational Inference and MCMC (12 Points)

Question 17 *(1 point)* The covariance matrix in the Laplace approximation for Bayesian logistic regression depends on the normalization constant of the true posterior.

[A] True ■ False

Question 18 *(1 point)* Let $X_1, X_2, \ldots$ be an ergodic Markov chain over a finite state space. Let T_0 be the burn-in time. Then, for a function f defined on the state space of the chain, the following holds

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} f(X_i) = \lim_{N\to\infty} \frac{1}{N - T_0} \sum_{i=T_0+1}^{N} f(X_i).$$

■ True [B] False

Question 19 *(2 points)* In the context of variational inference, which one of the following statements about the Evidence Lower Bound (ELBO) is true?

[A] Maximizing the ELBO is equivalent to minimizing the Kullback–Leibler divergence of the variational distribution from the likelihood.

[B] The ELBO provides an upper bound on the log marginal likelihood of the data.

■ Maximizing the ELBO is equivalent to minimizing the Kullback–Leibler divergence of the variational distribution from the posterior distribution.

[D] The ELBO is always positive.

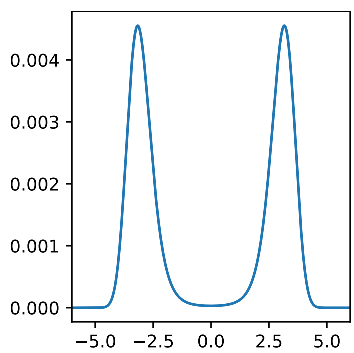


Figure 3: Related to Question 20.

Question 20 *(2 points)* Consider the "double-well" potential $f(x) = 0.05x^4 - x^2$. We use the Metropolis-adjusted Langevin Algorithm (MALA) to sample from the distribution $\pi(x) \propto e^{-f(x)}$ shown in Figure 3. Which of the following is true?

[A] MALA with the step-size $\tau = 0.01$ does not converge to π.

■ MALA with the step-size $\tau = 5$ converges to the correct distribution π.

3.1 Sampling from the Cube (6 Points)

Let $n \geq 4$ and $S = \{(x_1, \dots, x_n) \mid x_i \in \{0,1\}, \ \sum_{i=1}^n x_i \leq 2\} \subset \{0,1\}^n$ be the set of all n-bit strings with at most two 1-s. Suppose we want to sample from the probability distribution p defined as

$$p(x_1, \dots, x_n) = \begin{cases} \frac{1}{Z} \exp(\sum_{i=1}^n x_i w_i) & \text{if } (x_1, \dots, x_n) \in S \\ 0 & \text{otherwise,} \end{cases}$$

where $w_i \in \mathbb{R}$ are fixed weights and Z is a normalization constant that makes p a probability distribution.

Question 21 *(1 point)* Suppose the random vector $\mathbf{X}$ is distributed according to p. Denote the i-th component of the vector $\mathbf{X}$ by X_i. The random variables $X_1, \dots, X_n$ are independent.

[A] True ■ False

Consider the following process for sampling from p: Starting from an arbitrary $\boldsymbol{x}^{(0)} \in S$, at each iteration $t = 1, 2, \dots$ we select an index $i \in \{1, \dots, n\}$ uniformly at random, and

- if $\boldsymbol{x}_i^{(t-1)} = 1$, set $\boldsymbol{y}$ to be the same as $\boldsymbol{x}^{(t-1)}$ with the ith index flipped to 0 with probability A_i.
- if $\boldsymbol{x}_i^{(t-1)} = 0$, set $\boldsymbol{y}$ to be the same as $\boldsymbol{x}^{(t-1)}$ with the ith index flipped to 1 with probability B_i.
- if $\boldsymbol{y} \in S$, set $\boldsymbol{x}^{(t)} = \boldsymbol{y}$, otherwise, set $\boldsymbol{x}^{(t)} = \boldsymbol{x}^{(t-1)}$.

Here, $0 < A_i \leq 1$ and $0 < B_i \leq 1$ are parameters that will be defined later for all $i \in \{1, \dots, n\}$.

Question 22 *(2 points)* The process above describes an ergodic Markov chain.

■ True [B] False

Question 23 *(3 points)* Suppose the Markov chain above satisfies the detailed balance equation for p. Which of the following are values that A_i and B_i could take, such that the chain admits p as its stationary distribution for all possible values of $w_i \in \mathbb{R}$?

[A] $A_i = B_i = \frac{1}{2}$

■ $A_i = \dfrac{1}{1 + e^{w_i}}$ and $B_i = \dfrac{e^{w_i}}{1 + e^{w_i}}$

[C] $A_i = e^{-w_i}$ and $B_i = 1$

[D] $A_i = 1$ and $B_i = e^{w_i}$

[E] $A_i = \dfrac{1}{1 + w_i}$ and $B_i = \dfrac{w_i}{1 + w_i}$

4 Bayesian Deep Learning and Active Learning (10 points)

Question 24 *(2 points)*
Consider a Bayesian optimization problem of maximizing an unknown function on a discrete domain $\{x_1, x_2, x_3, x_4\}$. At each round we select a point x from this domain and observe the value $f(x) + \epsilon$ where ϵ is Gaussian noise and f is an unknown function. Suppose at some round t, the posterior distribution of the vector $\big(f(x_1), f(x_2), f(x_3), f(x_4)\big)^T$ is a multivariate normal distribution $\mathcal{N}(\mu_t, \Sigma_t)$ where

$$\mu_t = \begin{pmatrix} 1 \\ 2 \\ 0 \\ -1 \end{pmatrix}, \Sigma_t = \begin{pmatrix} 9 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 4 & 1 \\ 0 & 0 & 1 & 16 \end{pmatrix}.$$

Which input would we choose at this round if we use the upper confidence bound algorithm with $\beta_t = 1$?
Hint: Recall that the upper confidence bound algorithm selects the input x with maximum upper confidence bound: $\mu_t(x) + \beta_t \sigma_t(x)$ where $\mu_t(x)$ and $\sigma_t(x)$ are the posterior mean and posterior marginal standard deviation of $f(x)$ respectively.

■ x_1 $\boxed{\text{B}}$ x_2 $\boxed{\text{C}}$ x_3 $\boxed{\text{D}}$ x_4

4.1 Properties of Information Gain (2 points)

Recall that for random variables X, Y, the information gain between X and Y is given by

$$I(X;Y) = H(X) - H(X \mid Y)$$

where $H(X)$ is the entropy of X and $H(X \mid Y)$ is the conditional entropy of X given the observation of Y. Now given a new random variable Z, we similarly define the information gain between X and (Y, Z) as

$$I(X;Y,Z) = H(X) - H(X \mid Y, Z).$$

Question 25 *(1 point)*
It holds that $I(X;Y) \geq 0$.

■ True $\boxed{\text{B}}$ False

Question 26 *(1 point)*
It holds that $I(X;Y,Z) - I(X;Y) \geq 0$.

■ True $\boxed{\text{B}}$ False

4.2 Bayesian Deep Learning (6 points)

We are modeling a binary classification problem with a 2-layer Bayesian neural network where the probability that the binary label y of input $\boldsymbol{x}$ is equal to one is

$$p(y = 1 \mid \boldsymbol{w}, \boldsymbol{W}, \boldsymbol{x}) = \sigma(\boldsymbol{w}^\top \sigma(\boldsymbol{W}\boldsymbol{x}))$$

where σ is the element-wise applied sigmoid activation function defined as $\sigma(z) = (1 + \exp(-z))^{-1}$. The network depends on two *unknown* parameters $\boldsymbol{w} \in \mathbb{R}^2$ and $\boldsymbol{W} \in \mathbb{R}^{2\times 3}$. We use the following priors:

$$\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, 0.5 \cdot \boldsymbol{I}), \qquad \boldsymbol{W}_{ij} \sim \mathcal{N}(0, 0.5)$$

where all entries $\boldsymbol{W}_{ij}$ and $\boldsymbol{w}$ are independent of each other and $\boldsymbol{I}$ is the 2-dimensional identity matrix.

Question 27 *(3 points)*
Given the i.i.d. dataset $\{\boldsymbol{x}_i, y_i\}_{i=1}^{100}$, which of the following optimization problems finds the maximum a posteriori (MAP) estimate of the model parameters $\boldsymbol{w}$ and $\boldsymbol{W}$?
Note: Recall that the Frobenius matrix norm is defined via $\|\boldsymbol{A}\|_F^2 = \sum_{i,j} \boldsymbol{A}_{ij}^2$, while $\|\cdot\|_2$ denotes the spectral norm (the largest singular value of a matrix and for vectors the Euclidean norm).

- [A] $\arg\min_{\boldsymbol{w},\boldsymbol{W}} \|\boldsymbol{w}\|_2^2 + \|\boldsymbol{W}\|_F^2 - \frac{1}{100}\sum_{i=1}^{100} y_i \log(\sigma(\boldsymbol{w}^\top\sigma(\boldsymbol{W}\boldsymbol{x}_i))) + (1-y_i)\log(1-\sigma(\boldsymbol{w}^\top\sigma(\boldsymbol{W}\boldsymbol{x}_i)))$
- ■ $\arg\min_{\boldsymbol{w},\boldsymbol{W}} \|\boldsymbol{w}\|_2^2 + \|\boldsymbol{W}\|_F^2 - \sum_{i=1}^{100} y_i \log(\sigma(\boldsymbol{w}^\top\sigma(\boldsymbol{W}\boldsymbol{x}_i))) + (1-y_i)\log(1-\sigma(\boldsymbol{w}^\top\sigma(\boldsymbol{W}\boldsymbol{x}_i)))$
- [C] $\arg\min_{\boldsymbol{w},\boldsymbol{W}} \|\boldsymbol{w}\|_2^2 + \|\boldsymbol{W}\|_F^2 - \frac{1}{100}\sum_{i=1}^{100} y_i \log(\sigma(\boldsymbol{w}^\top\sigma(\boldsymbol{W}\boldsymbol{x}_i)))$
- [D] $\arg\min_{\boldsymbol{w},\boldsymbol{W}} \|\boldsymbol{w}\|_2^2 + \|\boldsymbol{W}\|_F^2 - \sum_{i=1}^{100} y_i \log(\sigma(\boldsymbol{w}^\top\sigma(\boldsymbol{W}\boldsymbol{x}_i)))$
- [E] $\arg\min_{\boldsymbol{w},\boldsymbol{W}} \|\boldsymbol{w}\|_2^2 + \|\boldsymbol{W}\|_2^2 - \sum_{i=1}^{100} y_i \log(\sigma(\boldsymbol{w}^\top\sigma(\boldsymbol{W}\boldsymbol{x}_i))) + (1-y_i)\log(1-\sigma(\boldsymbol{w}^\top\sigma(\boldsymbol{W}\boldsymbol{x}_i)))$

Question 28 *(3 points)*
We choose to approximate the posterior of $\boldsymbol{w}$ by a Laplace approximation $\boldsymbol{w} \sim \mathcal{N}(\hat{\boldsymbol{w}}, \boldsymbol{\Lambda}^{-1})$ with MAP estimate $\hat{\boldsymbol{w}}$ and precision matrix $\boldsymbol{\Lambda}$. Which of the following represents the correct precision matrix?
Note: $\ell(y_i \mid \boldsymbol{w}, \boldsymbol{W}, \boldsymbol{x}_i)$ denotes the negative log-likelihood of data point $(\boldsymbol{x}_i, y_i)$ and $\boldsymbol{H}_{\boldsymbol{w}} f(\boldsymbol{w})|_{\boldsymbol{w}=\hat{\boldsymbol{w}}}$ denotes the Hessian of a function f with respect to $\boldsymbol{w}$ evaluated at $\hat{\boldsymbol{w}}$.

- [A] $\boldsymbol{\Lambda} = 2\boldsymbol{I} + \frac{1}{100}\sum_{i=1}^{100} \boldsymbol{H}_{\boldsymbol{w}}\, \ell(y_i \mid \boldsymbol{w}, \boldsymbol{W}, \boldsymbol{x}_i)|_{\boldsymbol{w}=\hat{\boldsymbol{w}}}$
- [B] $\boldsymbol{\Lambda} = \frac{1}{100}\sum_{i=1}^{100} \boldsymbol{H}_{\boldsymbol{w}}\, \ell(y_i \mid \boldsymbol{w}, \boldsymbol{W}, \boldsymbol{x}_i)|_{\boldsymbol{w}=\hat{\boldsymbol{w}}}$
- ■ $\boldsymbol{\Lambda} = 2\boldsymbol{I} + \sum_{i=1}^{100} \boldsymbol{H}_{\boldsymbol{w}}\, \ell(y_i \mid \boldsymbol{w}, \boldsymbol{W}, \boldsymbol{x}_i)|_{\boldsymbol{w}=\hat{\boldsymbol{w}}}$
- [D] $\boldsymbol{\Lambda} = \sum_{i=1}^{100} \boldsymbol{H}_{\boldsymbol{w}}\, \ell(y_i \mid \boldsymbol{w}, \boldsymbol{W}, \boldsymbol{x}_i)|_{\boldsymbol{w}=\hat{\boldsymbol{w}}}$
- [E] $\boldsymbol{\Lambda} = -2\boldsymbol{I} - \sum_{i=1}^{100} \boldsymbol{H}_{\boldsymbol{w}}\, \ell(y_i \mid \boldsymbol{w}, \boldsymbol{W}, \boldsymbol{x}_i)|_{\boldsymbol{w}=\hat{\boldsymbol{w}}}$
- [F] $\boldsymbol{\Lambda} = -\sum_{i=1}^{100} \boldsymbol{H}_{\boldsymbol{w}}\, \ell(y_i \mid \boldsymbol{w}, \boldsymbol{W}, \boldsymbol{x}_i)|_{\boldsymbol{w}=\hat{\boldsymbol{w}}}$

5 MDPs and Reinforcement Learning (15 points)

Question 29 ♣ *(2 points)*
Which of the following algorithms, as taught in the course, are off-policy RL algorithms? **Mark all that apply.**

- ■ Deep Q-Networks (DQN)
- ■ Deep Deterministic Policy Gradients (DDPG)
- ■ Soft Actor Critic (SAC)
- E REINFORCE
- C PPO (Proximal Policy Optimization)
- F Advantage Actor Critic (A2C)

Question 30 ♣ *(2 points)* Which of the following algorithms are based on the "optimism in the face of uncertainty" principle? **Mark all that apply.**

■ R-Max   B PETS   C DDPG   ■ H-UCRL

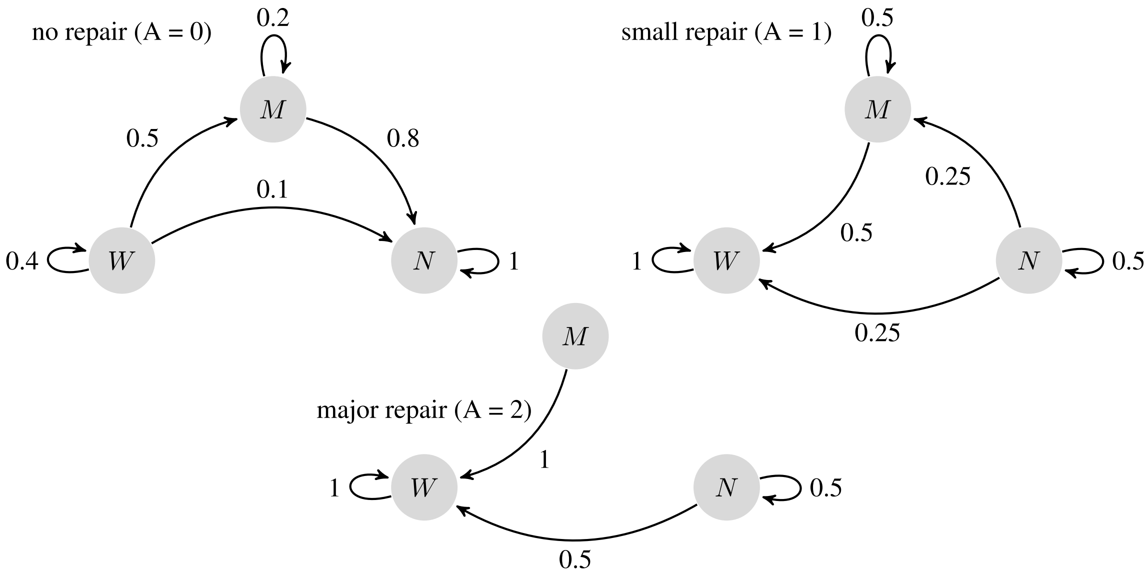
Question 31 ♣ *(3 points)* Consider an RL problem with reward $r(x, a)$, where $x \in \mathbb{R}, a \in \mathbb{R}$, are the state and action, respectively. We roll out a parametric policy $\pi_\theta(a|x)$ for T steps and denote with $\tau \sim \pi_\theta$ the resulting trajectory detailed as $\tau = \{x_0, a_0, x_1, a_1, \dots, x_T, a_T\}$. The expected sum of rewards under π_θ is defined as $J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[\sum_{t=0}^T r(x_t, a_t)]$. Which of the following is equal to $\nabla J(\theta)$, the gradient of this objective with respect to θ? Suppose $b \leq 0$ is a constant and **mark all that apply.**

- A $\mathbb{E}_{\tau\sim\pi_\theta}[\sum_{t=0}^T(\sum_{t'=0}^T r(x_{t'}, a_{t'}))\nabla\pi_\theta(a_t|x_t)]$
- ■ $\mathbb{E}_{\tau\sim\pi_\theta}[\sum_{t=0}^T(\sum_{t'=0}^T r(x_{t'}, a_{t'}))\nabla\log\pi_\theta(a_t|x_t)]$
- C $\nabla\mathbb{E}_{\tau\sim\pi_\theta}[\sum_{t=0}^T(\sum_{t'=0}^T r(x_{t'}, a_{t'}))\log\pi_\theta(a_t|x_t)]$
- D $\mathbb{E}_{\tau\sim\pi_\theta}[\sum_{t=0}^T(\sum_{t'=0}^T r(x_{t'}, a_{t'}) - b)\nabla\pi_\theta(a_t|x_t)]$
- ■ $\mathbb{E}_{\tau\sim\pi_\theta}[\sum_{t=0}^T(\sum_{t'=0}^T r(x_{t'}, a_{t'}) - b)\nabla\log\pi_\theta(a_t|x_t)]$
- F $\nabla\mathbb{E}_{\tau\sim\pi_\theta}[\sum_{t=0}^T(\sum_{t'=0}^T r(x_{t'}, a_{t'}) - b)\log\pi_\theta(a_t|x_t)]$
- ■ $\mathbb{E}_{\tau\sim\pi_\theta}[\sum_{t=0}^T(\sum_{t'=t}^T r(x_{t'}, a_{t'}))\nabla\log\pi_\theta(a_t|x_t)]$
- H $\mathbb{E}_{\tau\sim\pi_\theta}[\sum_{t=0}^T(\sum_{t'=0}^{t-1} r(x_{t'}, a_{t'}))\nabla\log\pi_\theta(a_t|x_t)]$

5.1 Markov Decision Processes (8 points)

Consider a machine that can exist in one of three states: *working* (W), *maintenance needed* (M), and *not working* (N). Our goal is to minimize the time that the machine spends in the *not working* state while avoiding unnecessary repair costs.

We develop a policy for repairing the machine at times $t = 1, 2, 3, \ldots$, given the state of the machine at this time. Three options are available: no repairs ($A = 0$), minor repairs ($A = 1$), or a major repair ($A = 2$). The three diagrams below show the transition probabilities between states when we play each of the respective actions $A = 0, 1$ or 2. The action corresponding to each diagram is written next to it.



Question 32 *(1 points)* Suppose we design a policy that performs a major repair at $t = 2$, but does no repairs at $t = 1, 3, 4, \ldots$, regardless of the state at that time. Is this policy a stationary policy?

[A] True ■ False

Question 33 *(3 points)* Assume that at $t = 1$ the machine is in working condition. What is the probability that the machine is in the *maintenance needed* (M) state once the above policy is applied for 3 rounds?

[A] 0.145 [B] 0.38 [C] 0.5 ■ 0.475

We model our objective of having a working machine at low repairs cost via the reward function

$$R(s, a) = R_1(s) + R_2(a)$$

where s is the state, a is the action and

$$R_1(s) = \begin{cases} -6 & s = N \\ 0 & s \in \{W, M\} \end{cases}, \qquad R_2(a) = \begin{cases} -5 & a = 2 \\ -2.5 & a = 1 \\ 0 & a = 0 \end{cases}$$

Question 34 *(4 points)* Assume a discount factor $\gamma = 0.5$. Assume that the optimal value function that we obtained by value iteration is $V^*(W) = -2, V^*(M) = -4, V^*(N) = -12$. Given this, which of the following is a corresponding optimal policy? Note that the function A^* : state $\rightarrow$ action, assigns optimal actions to each state.

■ $A^*(W) = 0, A^*(M) = 1, A^*(N) = 0$

B $A^*(W) = 0, A^*(M) = 1, A^*(N) = 1$

C $A^*(W) = 0, A^*(M) = 0, A^*(N) = 2$

D $A^*(W) = 0, A^*(M) = 1, A^*(N) = 2$

E $A^*(W) = 0, A^*(M) = 0, A^*(N) = 1$

CORRECTED

Answer Sheet of the Probabilistic Artificial Intelligence 2024/25 Exam

0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4
5	5	5	5	5	5	5	5
6	6	6	6	6	6	6	6
7	7	7	7	7	7	7	7
8	8	8	8	8	8	8	8
9	9	9	9	9	9	9	9

⟵ Please encode your student number on the left, and write your first and last names below.

Firstname and Lastname:

. . . .

Question 1: ■ B
Question 2: ■ B
Question 3: A ■
Question 4: A ■
Question 5: ■ B
Question 6: A B C ■
Question 7: ■ B C D
Question 8: A ■
Question 9: A ■
Question 10: A ■
Question 11: ■ B C
Question 12: A ■ C D E F
Question 13: A B C ■ E F
Question 14: A B C D ■ F
Question 15: A B C D E ■
Question 16: ■ B C D E F
Question 17: A ■
Question 18: ■ B
Question 19: A B ■ D
Question 20: A ■
Question 21: A ■
Question 22: ■ B
Question 23: A ■ C D E
Question 24: ■ B C D
Question 25: ■ B
Question 26: ■ B
Question 27: A ■ C D E
Question 28: A B ■ D E F
Question 29: ■ ■ C ■ E F
Question 30: ■ B C ■
Question 31: A ■ C D ■ F ■ H
Question 32: A ■
Question 33: A B C ■
Question 34: ■ B C D E