Introduction to Machine Learning Exam
Questions Pack

January 27, 2022

Time limit: 120 minutes

Instructions. This pack contains all questions for the final exam. It contains the questions only. Please use the accompanying answer sheet to provide your answers by blackening out the corresponding squares. As the exam will be graded by a computer, please **make sure to do blacken out the whole square and do not use ticks or crosses**. *During* the exam you can use a **pencil** to fill out the squares as well as an **eraser** to edit your answers. *After* the exam is over, we will collect the questions pack and provide you with additional time to blacken out the squares on the answer sheet with a **black pen**. *Nothing* written on 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 all answers are clearly marked by filling the squares out completely. 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, you will encounter the following question types.

- **Multiple Choice questions with a single answer.**

  Multiple Choice questions have **exactly one** correct choice. Depending on the difficulty of the question **2, 3, or 4** points are awarded if answered correctly, and **zero points** are awarded if answered wrong or not attempted.

- **True Or False questions.**

  Each True Or False questions has a value of **1 point** if answered correctly, **0 points** if answered wrong or not attempted.

Not all questions need to be answered correctly to achieve the best grade. There are **no negative grades** so you are incentivized to attempt all questions.

1 Regression and Classification

We are given a dataset consisting of n labeled training points $\mathcal{D} = \{(x_1, y_1), ..., (x_n, y_n)\}$, where $x_i \in \mathbb{R}^d$ are the feature vectors and $y_i \in \mathbb{R}$ are the labels. Here samples are generated independently from a distribution $p(x, y)$ for which the following holds:

$$y = {w^*}^\top x + \epsilon \text{ where } \epsilon \sim \mathcal{N}(0, \sigma^2).$$

The true underlying parameters $w^* \in \mathbb{R}^d$ are unknown. The rows of the *design matrix* $X \in \mathbb{R}^{n \times d}$ are the feature vectors $x_i \in \mathbb{R}^d$. The label vector is denoted by $y = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$. **In all of Section 1, we assume X is full rank i.e.,** $\text{rank}(X) = \min(n, d)$**.**
Recall from lecture that the *empirical risk* is defined as follows:

$$\hat{R}_{\mathcal{D}}(w) = \sum_{i=1}^{n} (w^\top x_i - y_i)^2 = \|y - Xw\|_2^2. \tag{1}$$

The goal is to to find $w \in \mathbb{R}^d$ that minimizes the empirical risk.

1.1 Ordinary Least Squares

Let X have a singular value decomposition $X = U\Lambda^{\frac{1}{2}}V^\top$. Here $U \in \mathbb{R}^{n \times n}$ and $V \in \mathbb{R}^{d \times d}$ are orthogonal, and $\Lambda^{\frac{1}{2}} \in \mathbb{R}^{n \times d}$ has the singular values $\sigma_i > 0$ on its diagonal and is zero elsewhere, i.e., $\sigma_i = \Lambda_{i,i}^{\frac{1}{2}}$ or equivalently $\sigma_i^2 = \Lambda_{i,i}$.

Question 1 [reg1] *(4 points)* What is the estimator $\widehat{w}$ you obtain by minimizing the empirical risk in Equation (1), i.e.,

$$\widehat{w} \triangleq \arg\min_w \hat{R}_{\mathcal{D}}(w),$$

in terms of w^* , V, Λ, and $\tilde{\epsilon} \triangleq U^\top \epsilon$?
Hint: Since U and V are orthogonal, it holds that, $U^\top U = UU^\top = I_{n \times n}, VV^\top = V^\top V = I_{d \times d}$.

- [A] $V\Lambda^{-1}V^\top U\Lambda^{\frac{1}{2}}V^\top(w^* + \tilde{\epsilon})$
- [B] $V\Lambda^{-\top}U^\top(w^* + \tilde{\epsilon})$
- [C] $V\Lambda^{-\frac{1}{2}}V^\top w^* + V\Lambda^{-\frac{1}{2}}\tilde{\epsilon}$
- [D] $V\Lambda^{-\frac{1}{2}}V^\top w^* + V\Lambda^{-1}\tilde{\epsilon}$
- [■] $w^* + V\Lambda^{-\frac{\top}{2}}\tilde{\epsilon}$
- [F] $w^* + V\Lambda^{-\top}\tilde{\epsilon}$
- [G] $w^* + V\Lambda^{\top}\tilde{\epsilon}$
- [H] $w^* + V\Lambda^{\frac{\top}{2}}\tilde{\epsilon}$

Question 2 [reg2] *(3 points)* Assume that the feature vectors of our training set are centered, i.e., $\sum_{i=1}^n x_i = 0$. Compute the following:
(*i.*) The empirical covariance matrix of our training data-points: $\Sigma \triangleq \frac{1}{n}\sum_{i=1}^n x_i x_i^\top$.
(*ii.*) The covariance matrix of the random vector $\tilde{\epsilon} \triangleq U^\top \epsilon$.

- [■] (*i.*) $\frac{1}{n}V\Lambda^{\frac{1}{2}\top}\Lambda^{\frac{1}{2}}V^\top$(*ii.*)$\sigma^2 I_{n \times n}$
- [B] (*i.*) $\frac{1}{n}U\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}\top}U^\top$(*ii.*)$\sigma^2 I_{n \times n}$
- [C] (*i.*) $\frac{1}{n}V\Lambda^{\frac{1}{2}\top}\Lambda^{\frac{1}{2}}V^\top$ (*ii.*) U
- [D] (*i.*) $\frac{1}{n}U\Lambda^{\frac{1}{2}}\Lambda^{\frac{1}{2}\top}U^\top$ (*ii.*) U

For the following statements, decide if they are True or False.

Question 3 [reg3] *(1 point)* When $n \geq d$, the empirical risk $\hat{R}_{\mathcal{D}}$, has a unique minimizer.

[■] True [B] False

Question 4 [reg4] *(1 point)* A local minimizer for the empirical risk $\hat{R}_{\mathcal{D}}$, is also a global minimizer.

■ True B False

Question 5 [reg5] *(1 point)* When $n \leq d$, there always exists w such that $Xw = y$.

■ True B False

Question 6 [reg6] *(3 points)* We would like to minimize the empirical risk $\hat{R}_{\mathcal{D}}$ using gradient descent. What is the update formula?

A $w_{t+1} = w_t + \eta_t(X^\top X w_t - 2X^\top y)$

■ $w_{t+1} = w_t - \eta_t(2X^\top X w_t - 2X^\top y)$

C $w_{t+1} = w_t + \eta_t(2X w_t - 2XX^\top y)$

D $w_{t+1} = w_t - \eta_t(X w_t - 2XX^\top y)$

E $w_{t+1} = w_t + \eta_t(2y_i - 2w_t^\top x_i)x_i$, for some randomly chosen $i \in \{1, 2, \ldots, n\}$

F $w_{t+1} = w_t - \eta_t(2y_i - 2w_t^\top x_i)x_i$, for some randomly chosen $i \in \{1, 2, \ldots, n\}$

G $w_{t+1} = w_t + \eta_t(y_i - 2w_t^\top x_i)x_i$, for some randomly chosen $i \in \{1, 2, \ldots, n\}$

H $w_{t+1} = w_t - \eta_t(2y_i - w_t^\top x_i)x_i$, for some randomly chosen $i \in \{1, 2, \ldots, n\}$

1.2 Ridge Regression

To avoid overfitting to the data, we add a regularization term to the empirical risk and minimize the following objective

$$l_{\mathcal{D}}(w) \triangleq \hat{R}_{\mathcal{D}}(w) + \lambda \|w\|_2^2 = \sum_{i=1}^{n} (w^\top x_i - y_i)^2 + \lambda \|w\|_2^2, \quad \lambda > 0. \tag{2}$$

The minimizer of Equation (2) is denoted by $\widehat{w}_\lambda \in \mathbb{R}^d$.

Question 7 [reg7] *(2 points)* Assume $n > d$. The minimizer $\widehat{w}_\lambda$ of Equation (2) in closed form is given by

- [A] $\widehat{w}_\lambda = \left(X^\top X + \lambda I\right)^{-1} X y.$
- [■] $\widehat{w}_\lambda = \left(X^\top X + \lambda I\right)^{-1} X^\top y.$
- [C] $\widehat{w}_\lambda = \left(X X^\top + \lambda I\right)^{-1} X y.$
- [D] $\widehat{w}_\lambda = \left(X X^\top + \lambda I\right)^{-1} X^\top y.$
- [E] there is no closed form solution.

Remember that for fixed $x \in \mathbb{R}^d$ the *bias-variance tradeoff* can be written as follows:

$$\mathbb{E}_{\mathcal{D},\epsilon}[(y - \widehat{w}_\lambda^\top x)^2] = (\mathbb{E}_{\mathcal{D}}[\widehat{w}_\lambda^\top x - w^{*\top} x])^2 + \mathrm{Var}_{\mathcal{D}}[\widehat{w}_\lambda^\top x] + \sigma^2.$$

The first term is called the **bias term**, the second term is called the **variance** term and the third term is the **irreducible noise**.

Question 8 [reg8] *(1 point)* A bigger λ (Equation 2) reduces the bias term in the bias variance trade-off.

[A] True [■] False

Question 9 [reg9] *(1 point)* A smaller λ (Equation 2) increases the variance in the bias-variance trade-off.

[■] True [B] False

Question 10 [reg10] *(1 point)* Smaller λ (Equation 2) prevents overfitting to the training data.

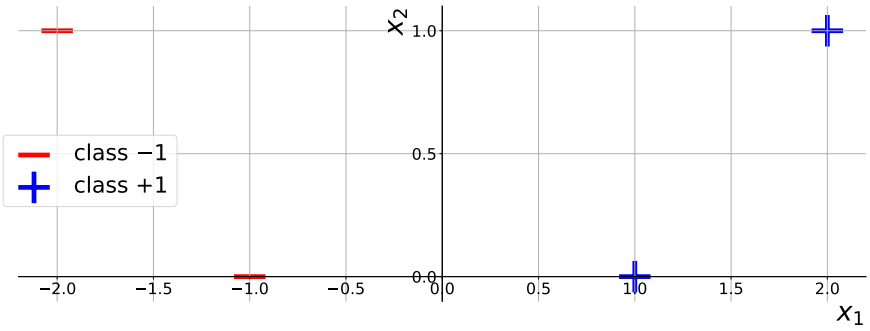
[A] True [■] False

Question 11 [reg11] *(1 point)* The population risk $\mathbb{E}_{\mathcal{D},\epsilon}[(y - \widehat{w}_\lambda^\top x)^2]$ is constant with respect to λ (Equation 2):

[A] True [■] False

1.3 Classification

Question 12 [classf1] *(3 points)* Consider linear classification with weights $w = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$. Predictions take the form $y_{pred} = \text{sign}(w^\top x)$. Consider the dataset $\{(\begin{pmatrix} -1 \\ 0 \end{pmatrix}, -1), (\begin{pmatrix} -2 \\ 1 \end{pmatrix}, -1), (\begin{pmatrix} 1 \\ 0 \end{pmatrix}, +1), (\begin{pmatrix} 2 \\ 1 \end{pmatrix}, +1)\}$, where the first element in each data-point $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ is the feature vector and the second element is its class label $y = \pm 1$. The points are represented in the Figure below.



The solution (normalized such that $\|w\|_2 = 1$) that classifies all points correctly and achieves the maximum margin is given by (Recall that the *margin* is defined as the minimum distance between all of the data-points and the decision boundary of the classifier):

A $w = \frac{1}{\sqrt{2}} \begin{pmatrix} -1 \\ 1 \end{pmatrix}$ with margin 2

B $w = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}$ with margin 2

C $w = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ with margin 2

D $w = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ with margin 1

■ $w = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ with margin 1

F $w = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ with margin 1

G $w = \frac{1}{\sqrt{5}} \begin{pmatrix} 2 \\ 1 \end{pmatrix}$ with margin 5

H $w = \frac{1}{\sqrt{5}} \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ with margin 5

Question 13 [classf2] *(2 points)* Remember that the zero-one loss is given by

$$l_{0-1}(z) = \begin{cases} 0 & z \geq 0 \\ 1 & z < 0 \end{cases}.$$

Which property is shared between the following "surrogate" loss functions?

hinge	$\ell_{\text{hinge}}(z) = \max\{0, 1 - z\}$
squared	$2\ell_{\text{sq}}(z) = (1 - z)^2$
logistic	$\ell_{\text{logistic}}(z)/\ln(2) = \ln(1 + e^{-z})/\ln(2)$
exponential	$\ell_{\text{exp}}(z) = e^{-z}$

■ Each one is an upper bound for the 0-1 loss.

B Each one is a lower bound for the 0-1 loss.

C Each one is differentiable on its whole domain.

D They are equally robust to outliers.

2 Kernels

Question 14 [kernel1] *(2 points)* Consider the feature map $\Phi : \mathbb{R} \to \mathbb{R}^3$ defined as $\Phi(x) = \left(x, x^2, e^x\right)^T$. Find the kernel $k(x, y)$ associated with Φ.

- [A] $x + x^2 + e^x$
- [B] $xy + e^{x+y}$
- [C] $x + y + x^2 + y^2 + e^{x+y}$
- [D] $x^2 + y^2 + xy + e^{x+y}$
- [■] $xy + (xy)^2 + e^{x+y}$
- [F] $x + y + (xy)^2 + e^{xy}$
- [G] $(xy + 1)^2 + e^{xy}$
- [H] $xy + (xy)^2$

Question 15 [kernel2] *(4 points)* Let $x, x' \in \mathbb{R}^3$ and $k(x, x') = (x^\top x' + 1)^2$. What is the minimal dimensionality of a feature map $\phi(x)$, such that $k(x, x') = \phi(x)^\top \phi(x')$?

[A] 6 [B] 9 [■] 10 [D] 12 [E] 13 [F] 15 [G] 16 [H] 27

Question 16 [kernel3] *(1 point)* Is the following statement True or False?
For every valid kernel $k(x, x')$, $k(x, x') \geq 0$ for all x and x'.

[A] True [■] False

For each of the following functions k, decide if it is a valid kernel function (True) or not (False).

Question 17 [kernel4] *(1 point)* For $x, x' \in \mathbb{R}^+$, define $k(x, x') = \frac{\max(x,x')}{\min(x,x')}$.

[A] True [■] False

Question 18 [kernel5] *(1 point)* For $x, x \in \mathbb{R}^d$, define $k(x, x') = (x^\top x' + 1)^3 + e^{(x^\top x')}$.

[■] True [B] False

3 Dimension Reduction with PCA

In (linear) principal component analysis (PCA), we map the data points $x_i \in \mathbb{R}^d, i = 1, \ldots, n$, to $z_i \in \mathbb{R}^k$, $k \ll d$, by solving the following optimization problem:

$$C_* = \frac{1}{n} \min_{\substack{W \in \mathbb{R}^{d \times k}, W^\top W = I \\ z_1, \ldots, z_n \in \mathbb{R}^k}} \sum_{i=1}^{n} \|W z_i - x_i\|_2^2 . \tag{3}$$

We denote by $W_*, z_1^*, \ldots, z_n^*$ the optimal solution of Equation (3). For all questions in section 3, assume the data points are centered i.e., $\sum x_i = 0$. Therefore, the empirical covariance of the data is as follows: $\Sigma_x = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^\top$

Question 19 [pca1] *(2 points)* What is the empirical covariance Σ_z^* of the latent variables z_i^*?

A $\Sigma_z^* = \Sigma_x W_*$ B $\Sigma_z^* = W_* \Sigma_x W_*^\top$ C $\Sigma_z^* = W_*^\top \Sigma_x$ ■ $\Sigma_z^* = W_*^\top \Sigma_x W_*$

Question 20 [pca2] *(2 points)* We obtain new data by first sampling $s \sim \mathcal{N}(0, \Sigma_z^*)$ in the latent space and then projecting the obtained sample to the original space $x_{\text{new}} = W_* s$. What is the distribution of x_{new}?

■ $\mathcal{N}\left(0, W_* \Sigma_z^* W_*^\top\right)$ B $\mathcal{N}(0, \Sigma_x)$ C $\mathcal{N}\left(0, W_*^\top \Sigma_z^* W_*\right)$ D $\mathcal{N}\left(0, W_* W_*^\top \Sigma_x\right)$

Question 21 [pca3] *(2 points)* In Figure 1 we plot 1000 points. We apply PCA to those 1000 points. What is the direction of the principal eigenvector?

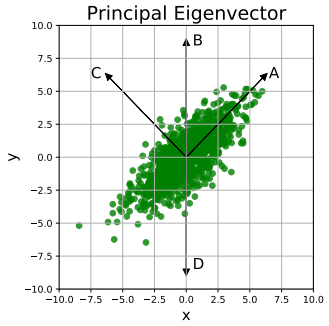


Figure 1: Direction of principal eigenvector.

■ A B B C C D D

Question 22 [pca4] *(1 point)* Figure 2 shows the Swiss roll dataset. All data points in this dataset lie on a 2D plane that has been wrapped around an axis. Linear (non-kernelized) PCA with $k = 2$ can explain all sources of variance in this dataset (C^* in Equation 3 is equal to 0).

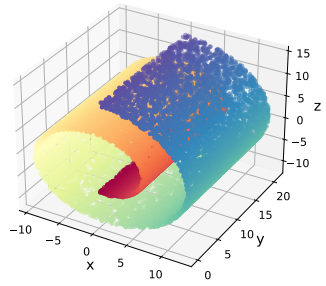


Figure 2: The Swiss roll dataset.

A True ■ False

Question 23 [pca5] *(1 point)* Both the PCA and the k-means problems can be formulated as

$$\min_{W, z_1, \ldots, z_n} \sum_{i=1}^{n} \|W z_i - x_i\|_2^2 ,$$

albeit, with different constraints on the matrix W and vectors $z_1, \ldots, z_n$.

■ True B False

Question 24 [pca6] *(1 point)* If we use the Gaussian kernel for kernel PCA, we implicitly perform PCA on an infinite-dimensional feature space.

■ True B False

Question 25 [pca7] *(1 point)* If we use neural network autoencoders with nonlinear activation functions, we seek to compress the data with a nonlinear map.

■ True B False

4 Neural Networks

4.1 Linear Neural Networks

This subsection is regarding linear networks. For input $x \in \mathbb{R}^{d_0}$, a deep linear network $F : \mathbb{R}^{d_0} \to \mathbb{R}$ of depth K will output $F(x) = W_K W_{K-1} ... W_1 x$, where each W_j is a matrix of appropriate dimension. We aim to train F to minimize the mean squared error loss on predicting real-valued scalar labels y. The loss is specified by

$$\ell(F) = \frac{1}{n} \sum_i (y_i - F(x_i))^2,$$

where i ranges over the dataset.

Question 26 [nn1] *(1 point)* With depth $K = 1$, we recover linear regression (with no bias term).

■ True B False

Question 27 [nn2] *(1 point)* For $K = 2$ there is a unique pair of matrices W_1, W_2 that minimizes ℓ.

A True ■ False

Question 28 [nn3] *(1 point)* Networks with increasing depth K allow one to model more complex relationships between x and y.

A True ■ False

Question 29 [nn4] *(1 point)* W_K must be a row vector.

■ True B False

Explanation: For 2), we can see that the pair of matrices is not unique by taking invertible matrix A and considering the pair $W_2 A^{-1}$ and AW_1, which will achieve the same loss as W_2, W_1. For 3), this is false since for all K, the function class remains the same as linear regression (it always only contains functions of the form $w \cdot x$). 4) is true since this is necessary to achieve scalar output.

Question 30 [nn5] *(3 points)* You plan to train this model with stochastic gradient descent and batch size 1. In each batch, you minimize $\tilde{\ell}_x(F) = (y - F(x))^2$, for a fixed data point x. For simplicity, suppose $K = 3$ and W_3 is a scalar.

Then, $\partial \tilde{\ell}_x / \partial W_3$ is equal to

■ $-2(y - F(x))(W_2 W_1 x)$.
B $2(y - F(x))(W_2 W_1 x)$.
C $(y - F(x))(W_2 W_1 x)$.
D $-(y - F(x))(W_2 W_1 x)$.
E $-2(y - F(x))$.
F $2(y - F(x))$.
G $(y - F(x))$.
H $-(y - F(x))$.

Explanation: Application of the chain rule.

Question 31 [nn6] *(2 points)* Again consider the loss calculated on a fixed data point $x \in \mathbb{R}^{d_0}$,

$$\tilde{\ell}_x(F) = (y - F(x))^2.$$

We use backpropagation to compute $\partial\tilde{\ell}_x/\partial W_1$. Suppose that $z_1(x) = W_1 x \in \mathbb{R}^{d_1}$. From previous steps in the backpropagation algorithm you know that $\partial\tilde{\ell}_x/\partial z_1(x) = a \in \mathbb{R}^{d_1}$. Please assume a and x are column vectors, i.e., $a \in \mathbb{R}^{d_1 \times 1}$ and $x \in \mathbb{R}^{d_0 \times 1}$.

Compute $\partial\tilde{\ell}_x/\partial W_1 \in \mathbb{R}^{d_1 \times d_0}$.

A $2ax^\top$	■ $ax^\top$	E $2a(W_1x)^\top$	G $a(W_1x)^\top$
B $-2ax^\top$	D $-ax^\top$	F $-2a(W_1x)^\top$	H $-a(W_1x)^\top$

Explanation: $\frac{\partial\tilde{L}_x}{\partial W_1} = \left(\frac{\partial\tilde{L}_x}{\partial z_1(x)}\right)\left(\frac{\partial z_1(x)}{\partial W_1}\right) = (a)(x^\top)$ Chain rule for backpropagation.

Question 32 [nn7] *(2 points)* Which of the following describes one iteration of a stochastic gradient descent update (still batch size 1 with single data point x) on W_1 with step size α?

■ $W_1 \leftarrow W_1 - \alpha\frac{\partial\tilde{\ell}_x}{\partial W_1}$	C $W_1 \leftarrow W_1 - \alpha\frac{\partial z_1(x)}{\partial W_1}$
B $W_1 \leftarrow W_1 + \alpha\frac{\partial\tilde{\ell}_x}{\partial W_1}$	D $W_1 \leftarrow W_1 + \alpha\frac{\partial z_1(x)}{\partial W_1}$

Explanation: Substituting into the definition of gradient descent.

4.2 Training Neural Networks

All questions in this subsection are independent from the previous subsection and are regarding training neural networks.

Question 33 [nn8] *(1 point)* Increasing the minibatch size in stochastic gradient descent (SGD) lowers the variance of gradient estimates (assuming data points in the mini-batch are selected independently).

■ True B False

Question 34 [nn9] *(1 point)* For a minibatch size of 1, SGD is guaranteed to decrease the loss in every iteration.

A True ■ False

Question 35 [nn10] *(1 point)* There exists a fixed learning rate $\eta > 0$ such that SGD with momentum is guaranteed to converge to a global minimum of the empirical risk for any architecture of a neural network and any dataset.

A True ■ False

Question 36 [nn11] *(1 point)* The cross entropy loss is designed for regression tasks, where the goal is to predict arbitrary real-valued labels.

A True ■ False

5 Decision Theory

In the following questions, assume that data is generated from some known probabilistic model $P(x, y)$. In both questions, we use the shorthand $p(x) = P(y = +1 \mid x)$.

Question 37 [decthe1] *(3 points)* Assume that we want to train a classifier $y = f(x)$ where labels y take values $y \in \{1, -1\}$. We extend the action (label) space and allow the classifier to *abstain* i.e., refrain from making a prediction. This extends the label space to $y \in \{+1, -1, r\}$. In order to make sure the classifier does not always abstain, we introduce a cost $c > 0$ for an abstention. The resulting 0-1 loss with abstention is given by:

$$\ell(f(x), y) = \mathbb{1}_{f(x)\neq y}\mathbb{1}_{f(x)\neq r} + c\mathbb{1}_{f(x)=r}.$$

An (Bayes) optimal classifier is one that minimizes the expected loss (risk) under the known conditional distribution. For a given input x, for which range of c should the optimal classifier abstain from predicting $+1$ or -1?

A $c < \max\{p(x), 1 - p(x)\}$

B $c > \min\{p(x), 1 - p(x)\}$

■ $c < \min\{p(x), 1 - p(x)\}$

D $c > 1 - \min\{p(x), 1 - p(x)\}$

E $c > 1 - p(x)$

F $c < p(x)$

Question 38 [decthe2] *(2 points)* We want to use regression with the quantile loss to estimate the current price y of our house given features x, defined as

$$\ell(f(x), y) = \tau \max(y - f(x), 0) + (1 - \tau) \max(f(x) - y, 0).$$

Here, $\tau \in (0, 1)$ is a parameter that balances overestimation and underestimation errors.
As we have enough time to sell the house, overestimation errors of the predictor are less critical than underestimation errors. Which of the asymmetric loss functions in Figure 3 would you use for the estimation of the current price of your house?

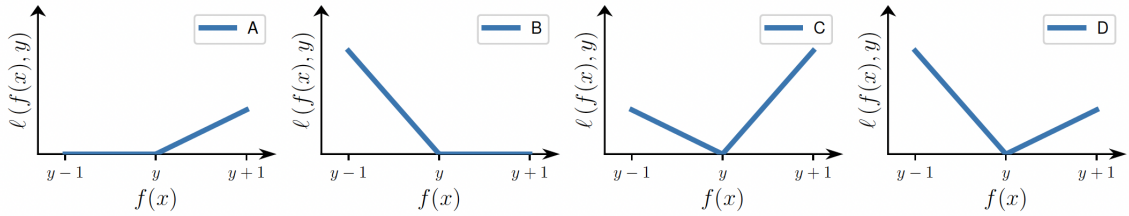


Figure 3: Different quantile loss functions.

☐ A A ☐ B B ☐ C C ■ D

6 Expectation Maximization Algorithm

In this question, we use the (soft) expectation maximization (EM) algorithm to compute a maximum likelihood estimator (MLE) for the average lifetime of light bulbs. We assume the lifetime of a light bulb is exponentially distributed with unknown mean $\theta > 0$, i.e., its cumulative distribution function is given by $F(x) = \left(1 - e^{-\frac{x}{\theta}}\right) \mathbb{1}_{\{x \geq 0\}}$.

We test $N + M$ independent light bulbs in two independent experiments. In the first experiment, we test the first N light bulbs. Let $Y = (Y_1, \ldots, Y_N)$, where each random variable Y_i represents the exact lifetime of light bulb i. In the second experiment we test the remaining M bulbs, but we only check the light bulbs at some fixed time $t > 0$ and record for each bulb whether it is still working or not.
Let $X = (X_1, \ldots, X_M)$, where the random variable $X_j = 1$ if the bulb j from the second experiment was still working at time t and 0 if it already expired. We denote by $Z = (Z_1, \ldots, Z_M)$ the unobserved lifetime of the light bulbs from the second experiment.

Question 39 [ema1] *(2 points)* What is the log-likelihood $\log p(X, Y, Z|\theta)$?

■ $\log p(X, Y, Z|\theta) = -(N + M) \log \theta - \frac{1}{\theta} \sum_{i=1}^{N} Y_i - \frac{1}{\theta} \sum_{j=1}^{M} Z_j$

B $\log p(X, Y, Z|\theta) = -(N + M) \log \theta - \theta \sum_{i=1}^{N} Y_i - \theta \sum_{j=1}^{M} Z_j$

C $\log p(X, Y, Z|\theta) = -N \log \theta - \theta \sum_{i=1}^{N} Y_i - \theta \sum_{j=1}^{M} Z_j$

D $\log p(X, Y, Z|\theta) = -M \log \theta - \frac{1}{\theta} \sum_{i=1}^{N} Y_i - \frac{1}{\theta} \sum_{j=1}^{M} Z_j$

Question 40 [ema2] *(3 points)* What is $E_1(\theta') \triangleq \mathbb{E}[Z_j|X_j = 1, \theta']$?

■ $\theta' + t$ B $\frac{1}{\theta'} + t$ C $t\theta' + t$ D $\frac{t}{\theta'} + t$

Question 41 [ema3] *(2 points)* What is $E_0(\theta') \triangleq \mathbb{E}[Z_j|X_j = 0, \theta']$?

■ $\theta' - \frac{te^{\frac{-t}{\theta'}}}{1-e^{\frac{-t}{\theta'}}}$ B $\theta' - \frac{2te^{\frac{-t}{\theta'}}}{1-e^{\frac{-t}{\theta'}}}$ C $\theta' - \frac{1-e^{\frac{-t}{\theta'}}}{te^{\frac{-t}{\theta'}}}$ D $\frac{1}{\theta'} - \frac{te^{\frac{-t}{\theta'}}}{1-e^{\frac{-t}{\theta'}}}$

Question 42 [ema4] *(2 points)* We define the expected complete data log-likelihood $Q(\theta, \theta')$ to be

$$Q(\theta, \theta') \triangleq \mathbb{E}_Z \left[\log p(X, Y, Z \mid \theta)|X, Y, \theta'\right]$$

and

$$k \triangleq \sum_{j=1}^{M} \mathbb{1}_{\{X_j = 1\}}$$

to be the number of light bulbs still working at time t in the second experiment. What is $Q(\theta, \theta')$?

■ $Q(\theta, \theta') = -(N + M) \log \theta - \frac{1}{\theta} \sum_{i=1}^{N} y_i - \frac{k}{\theta} E_1(\theta') - \frac{M-k}{\theta} E_0(\theta')$

B $Q(\theta, \theta') = -(N + M) \log \theta' - \frac{1}{\theta'} \sum_{i=1}^{N} y_i - \frac{k}{\theta'} E_1(\theta) - \frac{M-k}{\theta'} E_0(\theta)$

C $Q(\theta, \theta') = -(N + M) \log \theta - \theta \sum_{i=1}^{N} y_i - \theta k E_1(\theta') - \theta(M - k) E_0(\theta')$

D $Q(\theta, \theta') = -(N + M) \log \theta' - \theta' \sum_{i=1}^{N} y_i - \theta' k E_1(\theta) - \theta'(M - k) E_0(\theta)$

Question 43 [gmm1] *(1 point)* The MLE objective for Gaussian mixture models (GMM) is non-convex with respect to the cluster's means, covariances, and weights when we have strictly more than one Gaussian in the mixture.

■ True B False

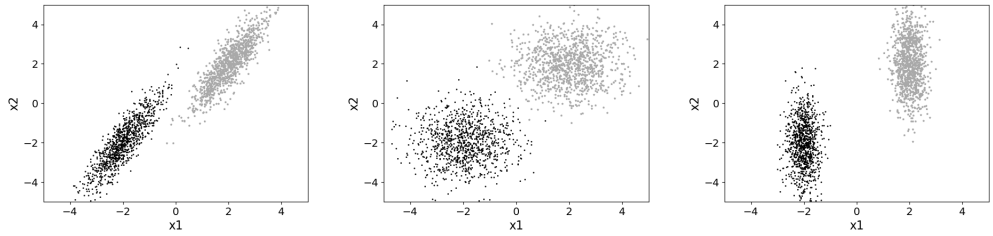
Question 44 [gmm2] *(1 point)* An EM algorithm can also be used to fit GMMs in the semi-supervised setting, where some data points are labeled and some are unlabeled.

■ True B False

Question 45 [gmm3] *(1 point)* We fit a GMM to a dataset utilizing the (soft) EM algorithm. We compute the log-likelihood of the data after each iteration. During this process the log-likelihood of the data never decreases.

■ True B False

Question 46 [gmm4] *(2 points)* You get 2D scatter plots of 3 different sets of data points (A, B, C respectively, see Figure below). You decide to cluster them with GMMs. You could model the covariance matrices of the two clusters as spherical, unrestricted, and diagonal. For datasets A, B, and C, assign the most appropriate covariance matrix.



A B C

A A: spherical, B: unrestricted, C: diagonal
B A: spherical, B: diagonal, C: unrestricted
■ A: unrestricted, B: spherical, C: diagonal
D A: diagonal, B: spherical, C: unrestricted
E A: unrestricted, B: diagonal, C: spherical
F A: diagonal, B: unrestricted, C: spherical

Explanation: B is spherical since the covariance matrix is the identity, whilst C is diagonal since, within clusters, there is no correlation between x1 and x2.

7 Generative Adversarial Networks

You train a generative adversarial network (GAN) with neural network discriminator D and neural network generator G. Let $z \sim \mathcal{N}(0, I)$ represent the random Gaussian (normal) noise input for G. Here, I is the identity matrix. The objective during training is given by

$$\min_G \max_D \mathbb{E}_{x \sim p_{\text{data}}}[\log D(x)] + \mathbb{E}_z[\log(1 - D(G(z)))],$$

where p_{data} is the data-generating distribution.

Question 47 [gan1] *(2 points)* Consider a fixed data point x with probability density $p_{\text{data}}(x)$. Suppose the probability density of x under the (not necessarily optimal) trained generator is $p_G(x)$. Moreover, assume that the trained discriminator D^* is the *optimal* discriminator for G, based on the loss above. That is:

$$D^* = \arg\max_D \mathbb{E}_{x \sim p_{\text{data}}}[\log D(x)] + \mathbb{E}_z[\log(1 - D(G(z)))],$$

For the data point x, what is $D^*(x)$?

A $D(x) = \frac{p_G(x)}{p_G(x) + p_{\text{data}}(x)}$

■ $D(x) = \frac{p_{\text{data}}(x)}{p_G(x) + p_{\text{data}}(x)}$

C 0

D 1

E Not enough information

Explanation: The second point is a sufficient condition. The third is not sufficient since, even at convergence, D may not be the optimal discriminator given G.

GANs can be used for the task of learning a *generative model* of data. However, GANs are not the only generative models we have seen in the course. Indicate whether each of the following models is generative or discriminative.

Question 48 [gan2] *(1 point)* Support Vector Machines.

A Generative Model ■ Discriminative Model

Question 49 [gan3] *(1 point)* Gaussian Mixture Models.

■ Generative Model B Discriminative Model

Question 50 [gan4] *(1 point)* Decision Trees.

A Generative Model ■ Discriminative Model

Answer Sheet of the Introduction to Machine Learning 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: A B C D ■ F G H
Question 2: ■ B C D
Question 3: ■ B
Question 4: ■ B
Question 5: ■ B
Question 6: A ■ C D E F G H
Question 7: A ■ C D E
Question 8: A ■
Question 9: ■ B
Question 10: A ■
Question 11: A ■
Question 12: A B C D ■ F G H
Question 13: ■ B C D
Question 14: A B C D ■ F G H
Question 15: A B ■ D E F G H
Question 16: A ■
Question 17: A ■
Question 18: ■ B
Question 19: A B C ■
Question 20: ■ B C D
Question 21: ■ B C D
Question 22: A ■
Question 23: ■ B
Question 24: ■ B
Question 25: ■ B
Question 26: ■ B
Question 27: A ■
Question 28: A ■
Question 29: ■ B
Question 30: ■ B C D E F G H
Question 31: A B ■ D E F G H
Question 32: ■ B C D
Question 33: ■ B
Question 34: A ■
Question 35: A ■
Question 36: A ■
Question 37: A B ■ D E F
Question 38: A B C ■
Question 39: ■ B C D
Question 40: ■ B C D
Question 41: ■ B C D
Question 42: ■ B C D
Question 43: ■ B
Question 44: ■ B
Question 45: ■ B
Question 46: A B ■ D E F
Question 47: A ■ C D E
Question 48: A ■
Question 49: ■ B
Question 50: A ■