1

Profs. Nicolas Flammarion and Martin Jaggi
Machine Learning – CS-433 - IC
20.01.2023 from 15h15 to 18h15 in STCC
Duration : 180 minutes

Student One

SCIPER : **111111**

Do not turn the page before the start of the exam. This document is double-sided, has 24 pages, the last ones are possibly blank. Do not unstaple.

- This is a closed book exam. No electronic devices of any kind.
- Place on your desk: your student ID, writing utensils, one double-sided A4 page cheat sheet if you have one; place all other personal items below your desk.
- You each have a different exam.
- This exam has many questions. We do *not* expect you to solve all of them even for the best grade
- Only answers in this booklet count. No extra loose answer sheets. You can use the last two pages as scrap paper.
- For the **multiple choice** questions, we give :
    +2 points if your answer is correct,
    0 points for incorrect or no answer
- For the **true/false** questions, we give :
    +1.5 points if your answer is correct,
    0 points for incorrect or no answer
- Use a **black or dark blue ballpen** and clearly erase with **correction fluid** if necessary.
- If a question turns out to be wrong or ambiguous, we may decide to nullify it.

For your examination, preferably print documents compiled from auto-multiple-choice.

First part: multiple choice questions

For each question, mark the box corresponding to the correct answer. Each question has **exactly one** correct answer.

Cost functions

Question 1 Assume we have N training samples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$ where for each sample $i \in \{1, \ldots, N\}$ we have that $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$. We want to classify the dataset using the exponential loss $L(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^{N} \exp(-y_i \mathbf{x}_i^\top \mathbf{w})$ for $\mathbf{w} \in \mathbb{R}^d$. Which of the following statements is **true**:

☐ The loss function L is non-convex in $\mathbf{w}$.

☐ There exists a vector $\mathbf{w}^\star$ such that $L(\mathbf{w}^\star) = 0$.

☐ If I find a vector $\mathbf{w}^\star$ such that $L(\mathbf{w}^\star) < 1/N$, then $\mathbf{w}^*$ linearly separates my dataset.

☐ None of the statements are true.

☐ This corresponds to doing logistic regression as seen in class.

Question 2 Assume we have N training samples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$ where for each sample $i \in \{1, \ldots, N\}$ we have that $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. For $\lambda \geq 0$, we consider the following loss:

$$L_\lambda(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^{N} (y_i - \mathbf{x}_i^\top \mathbf{w})^2 + \lambda \|\mathbf{w}\|_2,$$

and let $C_\lambda = \min_{\mathbf{w} \in \mathbb{R}^d} L_\lambda(\mathbf{w})$ denote the optimal loss value. Which of the following statements is **true**:

☐ C_λ is a non-increasing function of λ.

☐ For $\lambda = 0$, the loss L_0 is convex and has a unique minimizer.

☐ None of the statements are true.

☐ C_λ is a non-decreasing function of λ.

Optimization

Question 3 Consider the logistic regression loss $L : \mathbb{R}^d \to \mathbb{R}$ for a binary classification task with data $(\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \{0, 1\}$ for $i \in \{1, \ldots N\}$:

$$L(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^{N} \left(\log\left(1 + e^{\mathbf{x}_i^\top \mathbf{w}}\right) - y_i \mathbf{x}_i^\top \mathbf{w} \right).$$

Which of the following is a gradient of the loss L?

☐ $\nabla L(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^{N} \left(\mathbf{x}_i \frac{e^{\mathbf{x}_i^\top \mathbf{w}}}{1 + e^{\mathbf{x}_i^\top \mathbf{w}}} - y_i \mathbf{x}_i^\top \mathbf{w} \right)$

☐ $\nabla L(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i \left(y_i - \frac{e^{\mathbf{x}_i^\top \mathbf{w}}}{1 + e^{\mathbf{x}_i^\top \mathbf{w}}} \right)$

☐ $\nabla L(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{e^{\mathbf{x}_i^\top \mathbf{w}}}{1 + e^{\mathbf{x}_i^\top \mathbf{w}}} - y_i \mathbf{x}_i \right)$

☐ $\nabla L(\mathbf{w}) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i \left(\frac{1}{1 + e^{-\mathbf{x}_i^\top \mathbf{w}}} - y_i \right)$

Question 4 Consider the loss function $L : \mathbb{R}^d \to \mathbb{R}$, $L(\mathbf{w}) = \frac{\beta}{2}\|\mathbf{w}\|^2$, where $\beta > 0$ is a constant. We run gradient descent on L with a stepsize $\gamma > 0$ starting from some $\mathbf{w}_0 \neq 0$. Which of the statements below is true?

- [ ] Gradient descent converges in two steps for $\gamma = \frac{1}{\beta}$ (i.e., $\mathbf{w}_2$ is the **first** iterate attaining the global minimum of L).
- [ ] Gradient descent with stepsize $\gamma = \frac{2}{\beta}$ produces iterates that diverge to infinity ($\|\mathbf{w}_t\| \to \infty$ as $t \to \infty$).
- [ ] Gradient descent converges to the global minimum for any stepsize $\gamma > 0$.
- [ ] Gradient descent converges to the global minimum for any stepsize in the interval $\gamma \in \left(0, \frac{2}{\beta}\right)$.

LASSO/Ridge regression

Question 5 LASSO is a linear regression with L1 regularization, it is equivalent to imposing on the weights a :

- [ ] Laplace prior.
- [ ] Logistic prior.
- [ ] Gaussian prior.
- [ ] None of the other options.

Question 6 How does the bias-variance decomposition of a ridge regression estimator compare with that of the ordinary least-squares estimator in general?

- [ ] Ridge has a smaller bias, and smaller variance.
- [ ] Ridge has a larger bias, and larger variance.
- [ ] Ridge has a larger bias, and smaller variance.
- [ ] Ridge has a smaller bias, and larger variance.

Logistic Regression

Our task is to classify whether an animal is a dog (class 0) or a cat (class 1) based on the following features:

- x_1: height
- x_2: length of whiskers
- x_3: thickness of fur

We perform standard normal scaling on the training features so that they have a mean of zero and standard deviation of 1. We have trained a Logistic Regression model to determine the probability that the animal is a cat, $p(1|\mathbf{x}, \mathbf{w})$.

Question 7   Our classifier learns that cats have a lower height and longer whiskers than dogs, while the thickness of fur is not relevant to the classification outcome. Which of the following is true about the weights **w** learned by the classifier?

- [ ] $w_1 < w_2 < w_3$
- [ ] $w_2 < w_3 < w_1$
- [ ] $w_3 < w_2 < w_1$
- [ ] $w_1 < w_3 < w_2$
- [ ] $w_2 < w_1 < w_3$
- [ ] $w_3 < w_1 < w_2$

Question 8   Let's imagine that every pet owner in Europe sends us their dog and cat data to train our model, so we have lots of training data. Which of the following is true about the possible optimizers used to train our model? We assume that the optimal hyperparameters are used for each optimizer.

- [ ] Gradient descent takes fewer steps to converge, while Newton steps are more computationally efficient.
- [ ] Overall, gradient descent and Newton's method are equal in terms of computational complexity.
- [ ] Gradient descent steps are more computationally efficient, while Newton's method takes fewer steps to converge.
- [ ] Gradient descent and Newton's method take similar number of iterations to converge.

SVM

Question 9   Recall that the hard-margin SVM problem corresponds to:

$$\min_{\substack{\mathbf{w} \in \mathbb{R}^d, \\ \forall i:\ y_i \mathbf{w}^\top \mathbf{x}_i \geq 1}} \|\mathbf{w}\|_2.$$

Now consider the 2-dimensional classification dataset corresponding to the 3 following datapoints: $\mathbf{x}_1 = (-1, 2)$, $\mathbf{x}_2 = (1, 2)$, $\mathbf{x}_3 = (0, -2)$ and $y_1 = y_2 = 1$, $y_3 = -1$. Which of the following statements is **true**:

- [ ] Our dataset is not linearly separable and hence it does not make sense to consider the hard-margin problem.
- [ ] The unique vector which solves the hard-margin problem for our dataset is $\mathbf{w}^\star = (0, 1)$.
- [ ] None of the other statements are true.
- [ ] There exists a unique $\mathbf{w}^\star$ which linearly separates our dataset.

Generative Models

Question 10

Which of the following is true for Generative Adversarial Networks (GANs) but not Diffusion Models:

- [ ] They gradually remove noise from inputs using a Markov chain.
- [ ] They use a simple L2 loss function.
- [ ] They can generate new data from input noise.
- [ ] They use two different models during training.

PCA

Question 11

Which of the following transformations to the data matrix $\mathbf{X}$ will affect the principal components obtained through PCA?

- [ ] Adding a constant value to all elements of $\mathbf{X}$.
- [ ] None of the other options.
- [ ] Multiplying one of the features of $\mathbf{X}$ by a constant.
- [ ] Adding an extra feature that is constant across all data points.

Gaussian Mixture Models

Question 12   We apply a Gaussian Mixture Model made of K *isotropic* Gaussians (invariant to rotation around its center) to N vectors of dimension D. What is the number of *free* parameters of this model?

- [ ] $NKD + N$
- [ ] $2NKD$
- [ ] $NKD + NKD^2$
- [ ] $NKD + NKD^2 + N$
- [ ] $KD + 2K - 1 + N - 1$
- [ ] $2KD + N - 1$
- [ ] NKD
- [ ] $2NKD + N$
- [ ] $KD + KD^2 - 1$
- [ ] $KD + K - 1$
- [ ] $KD + K + N$
- [ ] $2KD - 1$
- [ ] $KD + 2K - 1$ ✖

Question 13   We define a simplified Gaussian Mixture Model consisting of 2 equally likely Gaussians, i.e. $K = 2$ and $\pi_1 = \pi_2 = 0.5$, and covariance matrices of the form $\boldsymbol{\Sigma}_i = \sigma_i \mathbf{I}_{D \times D}$ for $i \in \{1, 2\}$ with $\mathbf{I}_{D \times D}$ the identity matrix of size D. The dataset consists of only 2 points $\mathbf{x}_1$ and $\mathbf{x}_2$ that are distinct ($\mathbf{x}_1 \neq \mathbf{x}_2$). We initialize the model at some finite $\boldsymbol{\mu}_1^{(0)}, \boldsymbol{\mu}_2^{(0)}$ and $\sigma_1^{(0)}, \sigma_2^{(0)}$. We fit the model by the EM method on these parameters (keeping π_1 and π_2 fixed to 0.5). After $T \to \infty$ steps, select the true statement among the following:

- [ ] $\sigma_1^{(T)}$ and $\sigma_2^{(T)}$ diverge to ∞ for some but not all the initializations.
- [ ] $\sigma_1^{(T)}$ and $\sigma_2^{(T)}$ diverge to ∞ for any initializations.
- [ ] $\sigma_1^{(T)}$ and $\sigma_2^{(T)}$ converge to 0 for any initializations.
- [ ] $\sigma_1^{(T)}$ and $\sigma_2^{(T)}$ converge to 0 for some but not all the initializations.
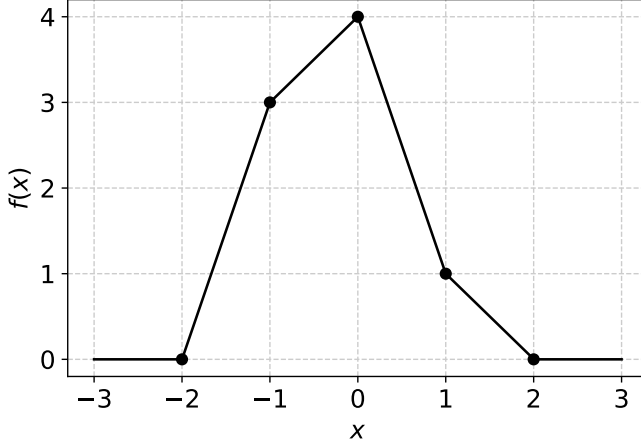
Neural Networks

Question 14 A neural network with a single hidden layer and a scalar input/output is parameterized as:

$$f(x) = \mathbf{a}^\top \mathrm{ReLU}(x\mathbf{1} + \mathbf{b})$$

where $\mathbf{a}, \mathbf{b}, \mathbf{1} \in \mathbb{R}^n$ and $x \in \mathbb{R}$. What values of $\mathbf{a}, \mathbf{b}$ result in the graph shown in the figure below?

- [ ] $\mathbf{a} = [1, 2, -4, -2, 3]^\top, \mathbf{b} = [-2, -1, 0, 1, 2]^\top$
- [ ] $\mathbf{a} = [1, 2, -4, -2, 3]^\top, \mathbf{b} = [2, 1, 0, -1, -2]^\top$
- [ ] $\mathbf{a} = [3, 1, -3, -1, 0]^\top, \mathbf{b} = [2, 1, 0, -1, -2]^\top$
- [ ] $\mathbf{a} = [3, 1, -3, -1, 0]^\top, \mathbf{b} = [-2, -1, 0, 1, 2]^\top$

Question 15 Let $\mathbf{Y} = \mathbf{XW} + \mathbf{b}$, where $\mathbf{X}, \mathbf{W}, \mathbf{Y} \in \mathbb{R}^{k \times k}$ and $\mathbf{b} \in R^{1 \times k}$, represent a linear layer of width k operating on a batch of k inputs where the addition is broadcasted as in Numpy or PyTorch. The network is trained with respect to a loss function $L(\mathbf{Y})$ that only depends on $\mathbf{W}$ and $\mathbf{b}$ through $\mathbf{Y}$. Given $\delta_{\mathbf{Y}} = \frac{\partial L}{\partial \mathbf{Y}}$, how can we compute $\delta_{\mathbf{W}} = \frac{\partial L}{\partial \mathbf{W}}$ and $\delta_{\mathbf{b}} = \frac{\partial L}{\partial \mathbf{b}}$? Let $\mathbf{1}_{1,k} = [1, 1, \ldots, 1]$ with shape $1 \times k$.

- [ ] $\delta_{\mathbf{W}} = \mathbf{X}\delta_{\mathbf{Y}}, \quad \delta_{\mathbf{b}} = \mathbf{1}_{1,k}\delta_{\mathbf{Y}}^\top$
- [ ] $\delta_{\mathbf{W}} = \mathbf{X}^\top\delta_{\mathbf{Y}}, \quad \delta_{\mathbf{b}} = \mathbf{1}_{1,k}\delta_{\mathbf{Y}}$
- [ ] $\delta_{\mathbf{W}} = \delta_{\mathbf{Y}}\mathbf{X}, \quad \delta_{\mathbf{b}} = \mathbf{1}_{1,k}\delta_{\mathbf{Y}}^\top$
- [ ] $\delta_{\mathbf{W}} = \delta_{\mathbf{Y}}\mathbf{X}^\top, \quad \delta_{\mathbf{b}} = \mathbf{1}_{1,k}\delta_{\mathbf{Y}}^\top$
- [ ] $\delta_{\mathbf{W}} = \mathbf{X}\delta_{\mathbf{Y}}, \quad \delta_{\mathbf{b}} = \mathbf{1}_{1,k}\delta_{\mathbf{Y}}$
- [ ] $\delta_{\mathbf{W}} = \delta_{\mathbf{Y}}\mathbf{X}^\top, \quad \delta_{\mathbf{b}} = \mathbf{1}_{1,k}\delta_{\mathbf{Y}}$
- [ ] $\delta_{\mathbf{W}} = \mathbf{X}^\top\delta_{\mathbf{Y}}, \quad \delta_{\mathbf{b}} = \mathbf{1}_{1,k}\delta_{\mathbf{Y}}^\top$
- [ ] $\delta_{\mathbf{W}} = \delta_{\mathbf{Y}}\mathbf{X}, \quad \delta_{\mathbf{b}} = \mathbf{1}_{1,k}\delta_{\mathbf{Y}}$

Question 16 A neural network has been trained for multi-class classification using cross-entropy but has not necessarily achieved a global or local minimum on the training set. The output of the neural network is $\mathbf{z} = [z_1, \ldots, z_d]^\top$ obtained from the penultimate values $\mathbf{x} = [x_1, \ldots, x_d]^\top$ via softmax $z_k = \frac{\exp(x_k)}{\sum_i \exp(x_i)}$ that can be interpreted as a probability distribution over the d possible classes. The cross-entropy is given by $H(\mathbf{y}, \mathbf{z}) = -\sum_{i=1}^d y_i \ln z_i$ where $\mathbf{y}$ is one-hot encoded meaning the entity corresponding to the true class is 1 and other entities are 0.

We now modify the neural network, either by scaling $\mathbf{x} \mapsto \alpha\mathbf{x}$ where $\alpha \in \mathbb{R}_{>0}$ or through a shift $\mathbf{x} \mapsto \mathbf{x} + b\mathbf{1}$ where $b \in \mathbb{R}$. The modified $\mathbf{x}$ values are fed into the softmax to obtain the final output and the network parameters are otherwise unchanged. How do these transformations affect the training accuracy of the network?

- [ ] Both transformations sometimes increase and sometimes decrease the accuracy.
- [ ] Both transformations decrease the accuracy in some cases (but never increase it).
- [ ] Neither transformation affects the accuracy.
- [ ] One transformation has no effect, the other sometimes increases and sometimes decreases the accuracy.
- [ ] One transformation has no effect, the other one decreases the accuracy in some cases (but never increases it).

Question 17 Consider a 2D convolutional layer with a kernel size of 3×3, 10 input channels and 10 output channels. The input to the convolution is a tensor $\mathbf{X} \in \mathbb{R}^{N \times C \times H \times W}$ with batch size $N = 1$, number of channels $C = 10$, height $H = 10$ and width $W = 10$. We use a zero padding of 1 on each side and obtain an output $\mathbf{Y}$ of the same shape as $\mathbf{X}$.

If we flatten the input and output into vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^{CHW}$ respectively, we can compute $\mathbf{y} = \mathbf{A}\mathbf{x}$ for a matrix $\mathbf{A}$ that depends on the original convolutional weights. What is the sparsity ζ of the $\mathbf{A}$, assuming the original convolutional weights contain no zeros? Note that the sparsity of a matrix is defined as the fraction (or percentage) of its elements that are zeros.

- [ ] $91.5\% < \zeta \le 93.0\%$
- [ ] $90.0\% < \zeta \le 91.5\%$
- [ ] $\zeta > 96.0\%$
- [ ] $94.5\% < \zeta \le 96.0\%$
- [ ] $93.0\% < \zeta \le 94.5\%$
- [ ] $\zeta \le 90.0\%$

Adversarial ML

Question 18 Let $\mathbf{x}, \mathbf{w}, \boldsymbol{\delta} \in \mathbb{R}^d$, $y \in \{-1, 1\}$, and $\varepsilon \in \mathbb{R}_{>0}$ be an arbitrary positive value. Which of the following is NOT true in general:

- [ ] $\mathrm{argmax}_{\|\boldsymbol{\delta}\|_2 \le \varepsilon} \log_2(1 + \exp(-y\mathbf{w}^\top(\mathbf{x} + \boldsymbol{\delta}))) = \mathrm{argmax}_{\|\boldsymbol{\delta}\|_2 \le \varepsilon} \mathbf{1}_{y\mathbf{w}^\top(\mathbf{x}+\boldsymbol{\delta}) \le 0}$
- [ ] $\mathrm{argmax}_{\|\boldsymbol{\delta}\|_2 \le \varepsilon} \log_2(1 + \exp(-y\mathbf{w}^\top(\mathbf{x} + \boldsymbol{\delta}))) = \mathrm{argmax}_{\|\boldsymbol{\delta}\|_2 \le \varepsilon} 1 - \tanh(y\mathbf{w}^\top(\mathbf{x} + \boldsymbol{\delta}))$
  where $\tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$
- [ ] $\mathrm{argmax}_{\|\boldsymbol{\delta}\|_2 \le \varepsilon} \log_2(1 + \exp(-y\mathbf{w}^\top(\mathbf{x} + \boldsymbol{\delta}))) = \mathrm{argmax}_{\|\boldsymbol{\delta}\|_2 \le \varepsilon} \exp(-y\mathbf{w}^\top(\mathbf{x} + \boldsymbol{\delta}))$
- [ ] $\mathrm{argmax}_{\|\boldsymbol{\delta}\|_2 \le \varepsilon} \log_2(1 + \exp(-y\mathbf{w}^\top(\mathbf{x} + \boldsymbol{\delta}))) = \mathrm{argmin}_{\|\boldsymbol{\delta}\|_2 \le \varepsilon} y\mathbf{w}^\top(\mathbf{x} + \boldsymbol{\delta})$

Question 19 Which statement about *black-box* adversarial attacks is true:

☐ They require access to the gradients of the model being attacked.

☐ They cannot be implemented via gradient-free (e.g., grid search or random search) optimization methods.

☐ They are highly specific and cannot be transferred from a model which is similar to the one being attacked.

☐ They can be implemented using gradient approximation via a finite difference formula.

Generalized Linear Models

Question 20 Which of the following probability distributions are members of the exponential family:

☐ Uniform distribution over $[0, \eta], \eta > 0$: $p(y|\eta) = \frac{1}{\eta} 1_{y \in [0, \eta]}$.

☐ Cauchy distribution: $p(y|y_0, \gamma) = \frac{1}{\pi \gamma [1 + (\frac{y - y_0}{\gamma})^2]}$.

☐ Poisson distribution: $p(y|\mu) = \frac{e^{-y}}{y!} \mu^y$.

Nearest Neighbor Classifiers

Question 21 Which of the following statements is true about nearest neighbor classifiers:

☐ None of the other answers.

☐ Nearest neighbors can be slow to find in high-dimensional spaces.

☐ Nearest neighbor classifiers can only work with the Euclidean distance.

☐ Nearest neighbor classifiers do not need to store the training data.

Recommender systems and text representation

Question 22 Consider a movie recommendation system which minimizes the following objective

$$\frac{1}{2} \sum_{(d,n) \in \Omega} [x_{dn} - (\mathbf{W}\mathbf{Z}^\top)_{dn}]^2 + \frac{\lambda_w}{2} \|\mathbf{W}\|_{\text{Frob}}^2 + \frac{\lambda_z}{2} \|\mathbf{Z}\|_{\text{Frob}}^2$$

where $\mathbf{W} \in \mathbb{R}^{D \times K}$ and $\mathbf{Z} \in \mathbb{R}^{N \times K}$. Suppose movies are divided into genre A and genre B (i.e., $\mathbf{W}_A \in \mathbb{R}^{D_A \times K}, \mathbf{W}_B \in \mathbb{R}^{D_B \times K}, \mathbf{W} = [\mathbf{W}_A; \mathbf{W}_B]$, with $D_A + D_B = D$) and users are divided into group 1 and group 2 (i.e., $\mathbf{Z}_1 \in \mathbb{R}^{N_1 \times K}, \mathbf{Z}_2 \in \mathbb{R}^{N_2 \times K}, \mathbf{Z} = [\mathbf{Z}_1; \mathbf{Z}_2]$, with $N_1 + N_2 = N$). In addition, group 1 users only rate genre A movies while group 2 users only rate genre B movies. Then instead of training a large recommendation system with $(\mathbf{W}, \mathbf{Z})$, one may train two smaller recommendation systems with parameters $(\mathbf{W}_A, \mathbf{Z}_1)$ and $(\mathbf{W}_B, \mathbf{Z}_2)$ separately. If SGD is used to solve the minimization problems and all conditions remain the same (e.g., hyperparameters, sampling order, initialization, etc), then which of the following statements is true about the two training methods?

☐ Feature vectors obtained in both cases can be either same or different, depending on the sparsity of rating matrix.

☐ Feature vectors obtained in both cases remain the same.

☐ Feature vectors obtained in both cases are different.

☐ Feature vectors obtained in both cases can be either same or different, depending on if ratings in two groups and genres are evenly distributed.

Question 23 Which of the following statements is **incorrect**?

☐ Word2Vec is based on the idea that a word's meaning is given by the words that frequently appear close-by.

☐ The order of words is ignored in FastText.

☐ Word vectors can be obtained through unsupervised training.

☐ Skip-gram predicts the center word from the bag of context words.

Second part: true/false questions

For each question, mark the box (without erasing) TRUE if the statement is **always true** and the box FALSE if it is **not always true** (i.e., it is sometimes false).

Question 24   (Subgradients) Consider the function $f : \mathbb{R} \to \mathbb{R}$, $f(x) = |x - 2023|$. A subgradient of f at $x = 2023$ exists **and** is unique.

☐ TRUE   ☐ FALSE

Question 25   (Ridge Regression and overfitting) In ridge regression, a large regularization parameter λ causes overfitting whereas a small regularization parameter causes underfitting.

☐ TRUE   ☐ FALSE

Question 26   (Logistic Regression) The loss function used in logistic regression equally penalizes positive and negative deviations from the correct class label.

☐ TRUE   ☐ FALSE

Question 27   (Bias-Variance) A model which has a high bias necessarily has a low variance.

☐ TRUE   ☐ FALSE

Question 28   (Kernels) *Recall that we say that a kernel $K : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is valid if there exists $k \in \mathbb{N}$ and $\Phi : \mathbb{R} \to \mathbb{R}^k$ such that for all $(x, x') \in \mathbb{R} \times \mathbb{R}$, $K(x, x') = \Phi(x)^\top \Phi(x')$.*

The kernel $K(x, x') = \cos(x + x')$ is a valid kernel.

☐ TRUE   ☐ FALSE

Question 29   (GANs) Generative Adversarial Networks use the generator and discriminator models during training but only the discriminator for data synthesis.
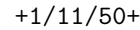
☐ TRUE   ☐ FALSE

Question 30   (GMM) Fitting a Gaussian Mixture Model with a single Gaussian ($K = 1$) will converge after one step of Expectation-Maximization.

☐ TRUE   ☐ FALSE

Question 31   (NN) Consider two fully connected networks, A and B, with a constant width for all layers, inputs and outputs. Network A has depth $3L$ and width H, network B has depth L and width $2H$. Everything else is identical for the two networks and both L and H are large. In this case, performing a single iteration of backpropagation requires fewer scalar multiplications for network A than for network B.

☐ TRUE   ☐ FALSE

Question 32 (CNN) The output of a 2D convolutional layer with filter size $S \times S$, where $S \geq 1$, always has a larger receptive field than the input to the convolution.

☐ TRUE ☐ FALSE

Question 33 (Adversarial Robustness) The primal formulation of the soft-margin SVM is NOT equivalent to ℓ_2 adversarial training for a linear model trained with the hinge loss ($\ell(z) = \max\{0, 1 - z\}$).

☐ TRUE ☐ FALSE

Question 34 (kNN) Nearest neighbor classifiers cannot be used for regression because they rely on majority voting, which is not suited for continuous labels.

☐ TRUE ☐ FALSE

Question 35 (Text representation) Antonyms, such as "hot" and "cold", have very different context words.

☐ TRUE ☐ FALSE

Third part, open questions

Answer in the space provided! Your answer must be justified with all steps. Leave the check-boxes empty, they are used for the grading.

1 Empirical risk minimization for classification

Given a joint data distribution $\mathcal{D}$ on $\mathcal{X} \times \{-1, 1\}$ and n independent and identically distributed observations from $\mathcal{D}$, the goal of the classification task is to learn a classifier $f : \mathcal{X} \rightarrow \{-1, 1\}$ with minimum true risk $\mathcal{L}(f) = \mathbb{E}_{(X,Y)\sim\mathcal{D}}[\mathbb{1}_{f(X)\neq Y}]$ where $\mathbb{1}_C = \begin{cases} 1 & \text{if } C \text{ is true} \\ 0 & \text{otherwise} \end{cases}$. We denote by $\mathcal{D}_X$ the marginal law (probability distribution) of X, and $\mathcal{D}_{Y|X}$ the conditional law of Y given X.

Question 36: (*1 point.*) Give the two reasons seen in the course which explain that minimizing the true risk with the $0 - 1$ loss over the set of classifiers $f : \mathcal{X} \rightarrow \{-1, 1\}$ is problematic.

☐ 0 ☐ 1

Thus, instead of directly learning a classifier $f : \mathcal{X} \to \{-1, 1\}$, we learn instead a function $g : \mathcal{X} \to \mathbb{R}$ and use the function $f(x) = \text{sgn}(g(x))$ to make our prediction, where

$$
\text{sgn}(a) = \begin{cases} 1 & \text{if } a > 0 \\ 0 & \text{if } a = 0 \\ -1 & \text{if } a < 0. \end{cases} \tag{1}
$$

The risk of the function $f = \text{sgn} \circ g$ is still denoted by $\mathcal{L}(g)$ and is equal to:

$$
\mathcal{L}(g) = \mathbb{E}_{(X,Y) \sim \mathcal{D}}[\mathbb{1}_{\text{sgn}(g(X)) \neq Y}] = \mathbb{E}_{(X,Y) \sim \mathcal{D}}[\mathbb{1}_{Y g(X) \leq 0}]. \tag{2}
$$

Instead of minimizing the true risks $\mathcal{L}$, we replace the 0-1 loss $\ell_{0-1}(u) = \mathbb{1}_{u \leq 0}$ by a convex surrogate ϕ and consider minimizing the true ϕ-risk defined as

$$
\mathcal{L}_\phi = \mathbb{E}_{(X,Y) \sim \mathcal{D}}[\phi(Y g(X))]. \tag{3}
$$

The convex surrogates we will consider are:

- the quadratic loss $\phi(u) = \frac{1}{2}(u - 1)^2$

- the logistic loss $\phi(u) = 2 \log(1 + e^{-u})$

- the hinge loss $\phi(u) = \max(1 - u, 0) = (1 - u)_+$ (def : for any $x \in \mathbb{R}$ we have $x_+ = \max(0, x)$).

Many convex surrogates ϕ can be used to upper bound the 0-1 loss $\phi_{0-1}(u) = \mathbb{1}_{u \leq 0}$ if they are properly rescaled. However, using this property alone to justify the good performance of a convex surrogate can be misleading.

Let us denote by $\eta(\mathbf{x}) = \mathbb{P}(Y = 1 | X = \mathbf{x}) \in [0, 1]$.

Question 37: (*1 point.*) Derive a formula for $\mathbb{E}[Y | X = \mathbf{x}]$ as a function of $\eta(\mathbf{x})$.

☐ 0 ☐ 1
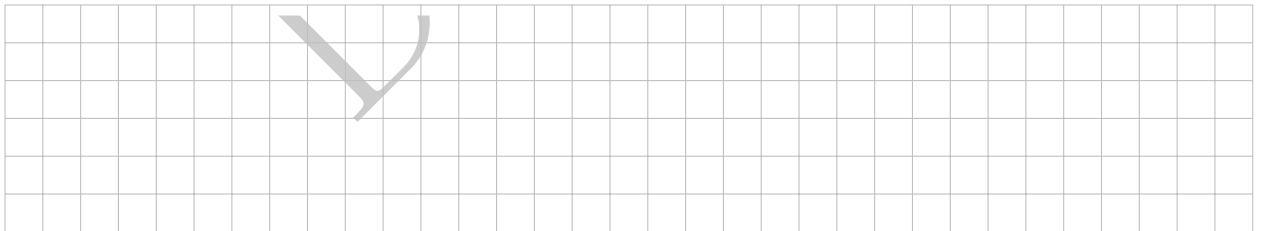
We have seen in the course that a classifier which minimizes the true risk $\mathcal{L}(f)$ over classifiers f is given by

$$
f^\star(x) \in \underset{y \in \{-1,1\}}{\text{argmax}} \ \mathbb{P}(Y = y | X = \mathbf{x}).
$$

Such a function is called a Bayes classifier.

Question 38: (*2 points.*) Show that the function $f^\star(\mathbf{x}) = \text{sgn}(\mathbb{E}[Y | X = \mathbf{x}])$ is a Bayes classifier.
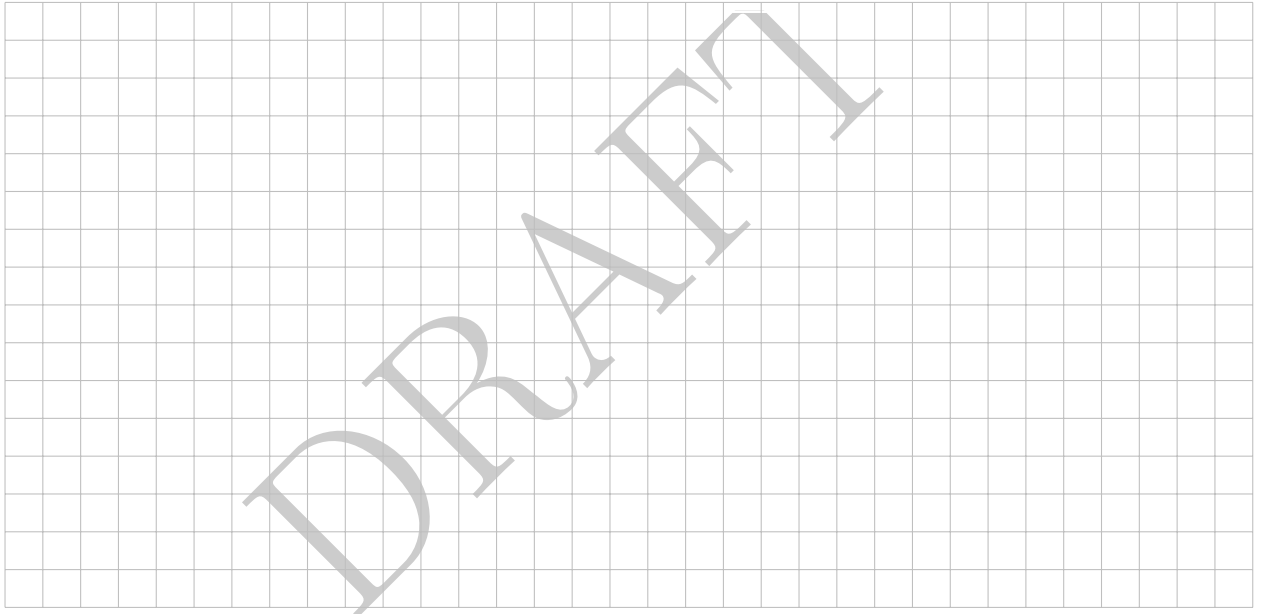
☐ 0 ☐ 1 ☐ 2

Question 39: (*2 points.*) Show that $\mathcal{L}^\star = \mathcal{L}(f^\star) = \mathbb{E}[\min(\eta(\mathbf{x}), 1 - \eta(\mathbf{x}))]$.

Hint: Perform a case distinction depending on the sign of $\eta - 1/2$.

☐ 0 ☐ 1 ☐ 2

To control the quality of a classifier f we define the *excess risk* of f as $\mathcal{L}(f) - \mathcal{L}^\star$.

1.1 Classification calibration and conditional ϕ-risk

The purpose of this first exercise part is to ensure that the predictions produced by minimizing the true ϕ-risk are optimal. As for the $0 - 1$ loss, it can be shown that the true ϕ-risk is minimized at a predictor $g^\star : \mathcal{X} \to \mathbb{R}$ satisfying for all $\mathbf{x} \in \mathcal{X}$:

$$g^\star(\mathbf{x}) \in \arg\min_{z \in \mathbb{R}} \mathbb{E}[\phi(zY)|X = \mathbf{x}].$$

Thus the function $g^\star$ that minimizes the ϕ-risk can be determined by looking at each $\mathbf{x}$ separately.
We will start with the square loss.

Square loss.

Question 40: (*2 points.*) Give a formula of the function $g^\star : \mathcal{X} \to \mathbb{R}$ which minimizes the true ϕ-risk, as a function of $\eta(\mathbf{x})$.
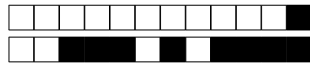
☐ 0 ☐ 1 ☐ 2

Question 41: (*2 points.*) Show that the classifier $f(\mathbf{x}) = \mathrm{sgn}(g^\star(\mathbf{x}))$ leads to the optimal prediction of the Bayes classifier.

☐ 0 ☐ 1 ☐ 2

We can therefore conclude in the population case that using the square loss for binary classification is leading to the optimal prediction.

Other losses. To study the impact of using the ϕ-risk for other losses, we first define the conditional ϕ-risk for a predictor $g : \mathcal{X} \to \mathbb{R}$ as

$$\mathbb{E}[\phi(Yg(\mathbf{x}))|X = \mathbf{x}] = \eta(\mathbf{x})\phi(g(\mathbf{x})) + \big(1 - \eta(\mathbf{x})\big)\phi(-g(\mathbf{x})),$$

which we denote $C_{\eta(\mathbf{x})}(g(x))$, with the function $C_\eta(\alpha) = \eta\phi(\alpha) + (1 - \eta)\phi(-\alpha)$ for $\alpha \in \mathbb{R}$.

A convex surrogate should at least guarantee that in the population case, where all $\mathbf{x}$ are independent, the optimal $g(\mathbf{x})$ produced by minimizing the conditional ϕ-risk results in the same prediction as the Bayes predictor.

We say that a function ϕ is *classification-calibrated* if it satisfies the following two statements:

$$\eta > 1/2 \iff \underset{\alpha \in \mathbb{R}}{\operatorname{argmin}}\, C_\eta(\alpha) \subseteq \mathbb{R}_+^* \text{ , and} \tag{4}$$

$$\eta < 1/2 \iff \underset{\alpha \in \mathbb{R}}{\operatorname{argmin}}\, C_\eta(\alpha) \subseteq \mathbb{R}_-^* \text{ .} \tag{5}$$

here the notation means $\mathbb{R}_+^* = \{x \in \mathbb{R} : x > 0\}$ and $\mathbb{R}_-^* = \{x \in \mathbb{R} : x < 0\}$, so the reals excluding zero.

Question 42: (*2 points.*) Show that when ϕ is classification calibrated, then minimizing the ϕ-risk exactly leads to the same prediction as the Bayes predictor f^*.

☐ 0 ☐ 1 ☐ 2
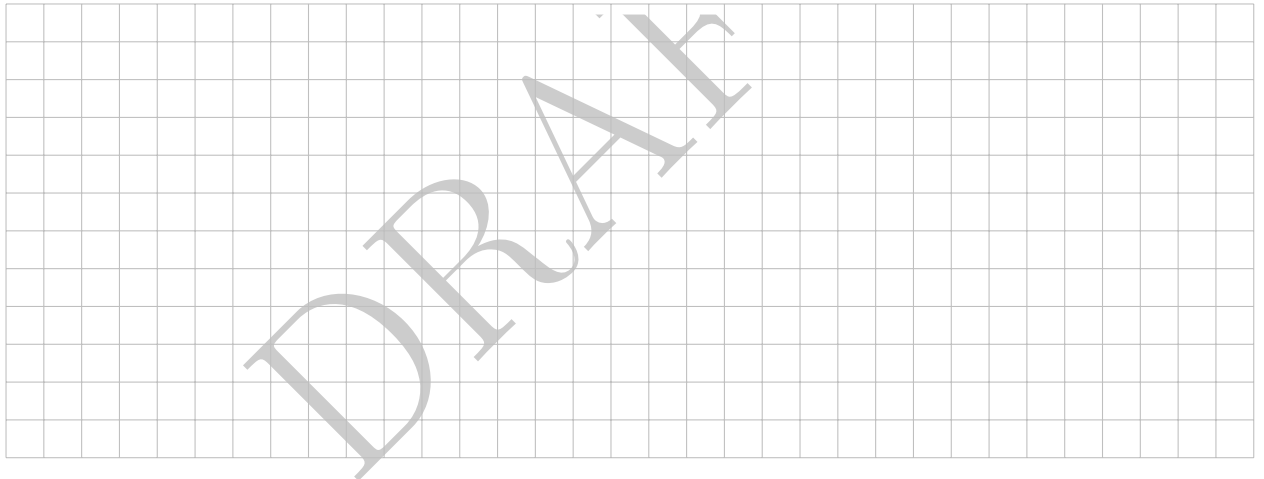
When the function ϕ is convex, calibration can be shown to be equivalent to a simple condition.

Question 43: (*2 points.*) Let $\phi : \mathbb{R} \to \mathbb{R}$ convex and differentiable in 0. Show the following equivalences:

$$\operatorname*{argmin}_{\alpha \in \mathbb{R}} C_\eta(\alpha) \subseteq \mathbb{R}_+^* \iff (2\eta - 1)\phi'(0) < 0 \ , \ \text{and}$$
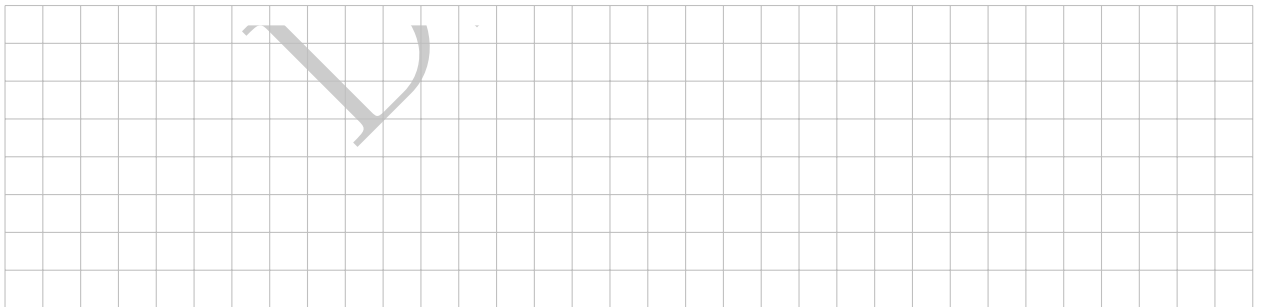
$$\operatorname*{argmin}_{\alpha \in \mathbb{R}} C_\eta(\alpha) \subseteq \mathbb{R}_-^* \iff (2\eta - 1)\phi'(0) > 0.$$

☐ 0 ☐ 1 ☐ 2
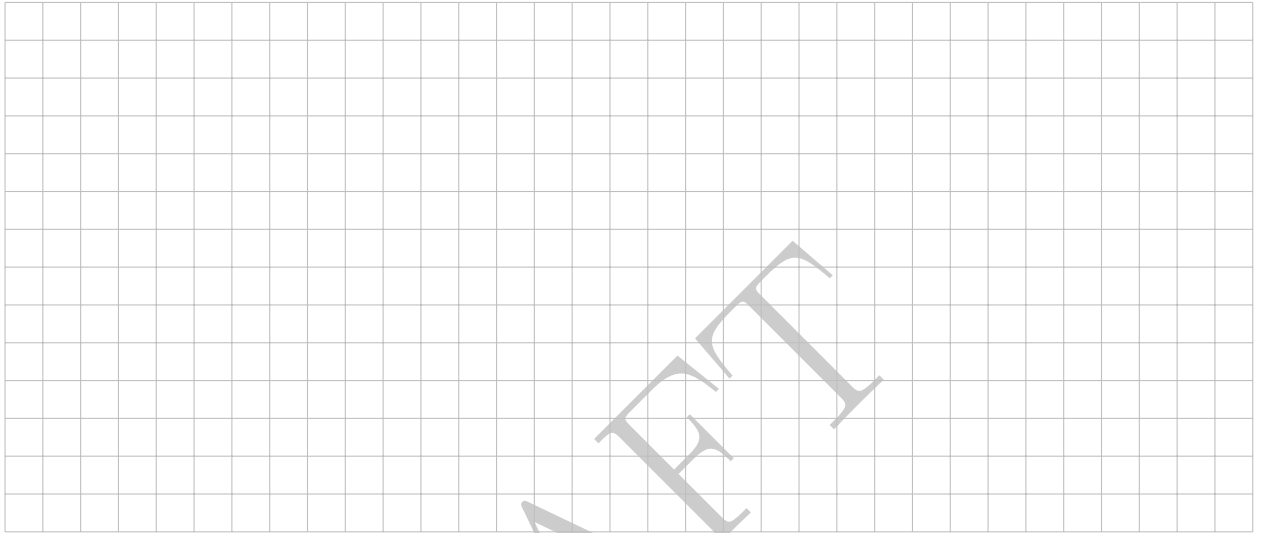
Question 44: (*1 point.*) Let $\phi : \mathbb{R} \to \mathbb{R}$ be convex, and differentiable at 0. Show that the surrogate function ϕ is classification-calibrated if and only if $\phi'(0) < 0$.

☐ 0 ☐ 1

Note that it is in fact possible to show the stronger statement that ϕ is classification-calibrated if and only if ϕ is differentiable at 0 and $\phi'(0) < 0$.

Question 45: (*1 point.*) Is the hinge-loss classification-calibrated?

☐ 0 ☐ 1
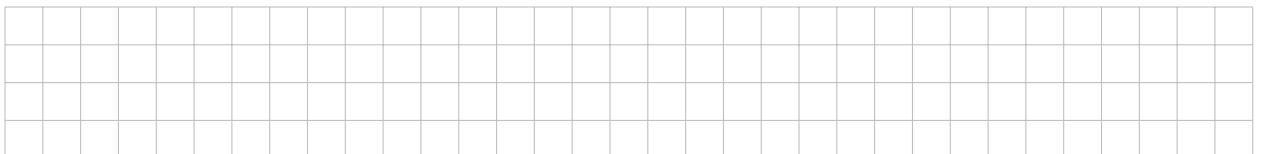
We now assume that ϕ is classification calibrated and convex.

1.2 Relation between the classification risk and the ϕ-risk

We have shown in the previous exercise that for any $\mathbf{x} \in \mathcal{X}$, minimizing $C_{\eta(\mathbf{x})}(g(\mathbf{x}))$ with respect to $g(\mathbf{x})$ leads to an optimal prediction using $\mathrm{sgn}(g(\mathbf{x}))$. We now want to show that controlling the excess ϕ-risk enables to control the classification excess risk. To that end, we will show the existence of an increasing function $H : \mathbb{R}_+ \to \mathbb{R}_+$ such that:

$$\mathcal{L}(g) - \mathcal{L}^\star \leq H\big(\mathcal{L}_\phi(g) - \mathcal{L}_\phi^\star\big).$$

Such a function H is called a calibration function.

Question 46: (*2 points.*) For any function $g : \mathcal{X} \to \mathbb{R}$, and for a Bayes predictor $g^\star : \mathcal{X} \to \mathbb{R}$ (i.e., such that $\mathrm{sgn} \circ g^\star$ is a Bayes classifier), show that

$$\mathcal{L}(g) - \mathcal{L}^\star = \mathbb{E}[\mathbb{1}_{g(X)g^\star(X)<0}|2\eta(X) - 1|].$$

☐ 0 ☐ 1 ☐ 2
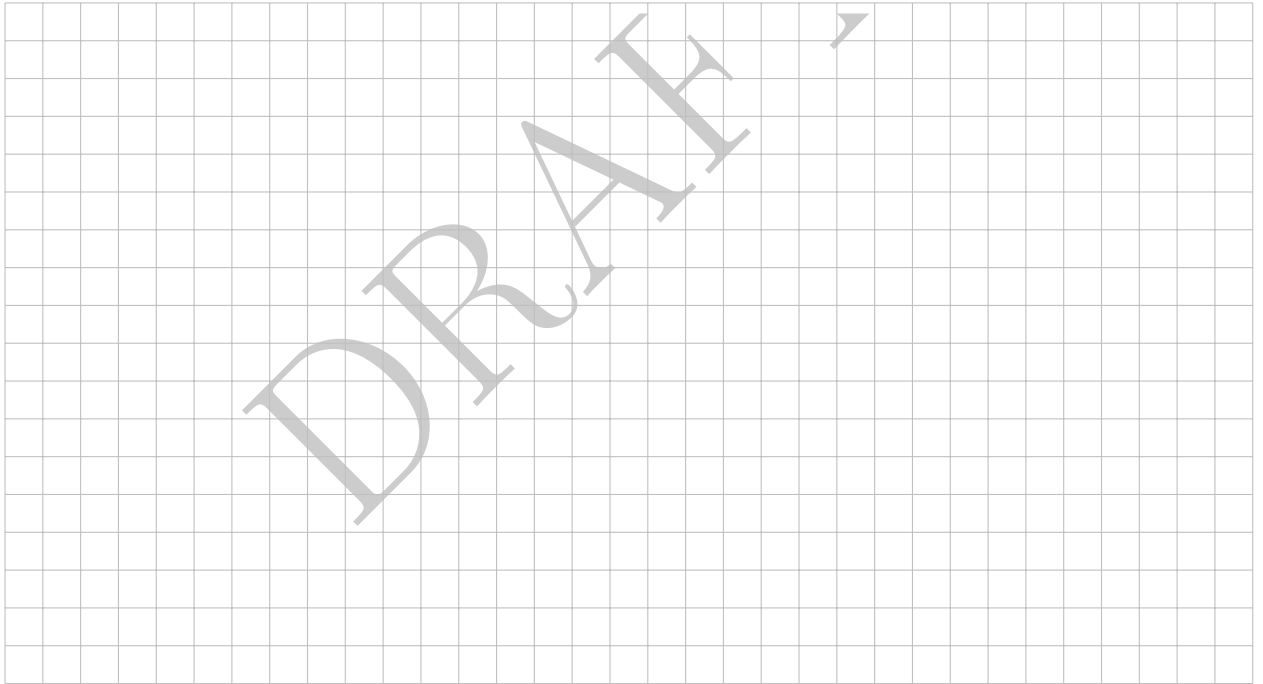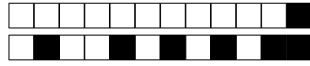
Question 47: (*2 points.*) Let $b : \mathbb{R} \to \mathbb{R}$ a function that preserves the sign, i.e., $b(\mathbb{R}_+^*) \subseteq \mathbb{R}_+^*$ and $b(\mathbb{R}_-^*) \subseteq \mathbb{R}_-^*$. Show that

$$\mathcal{L}(g) - \mathcal{L}^\star \leq \mathbb{E}[|2\eta(X) - 1 - b(g(X))|]$$

☐ 0 ☐ 1 ☐ 2
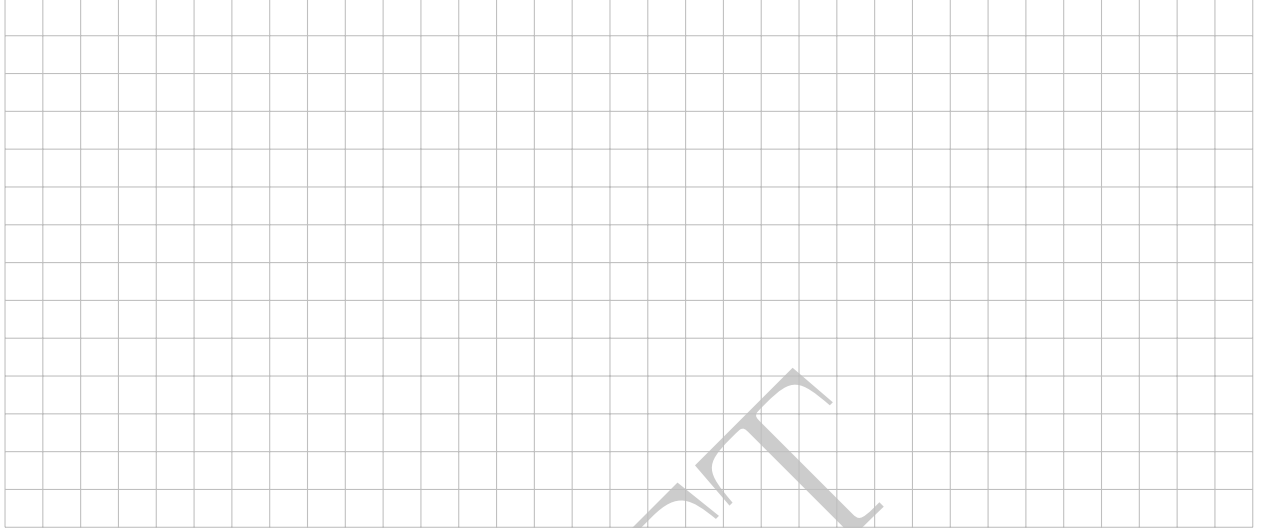
For $\mathbf{x} \in \mathcal{X}$, the excess conditional ϕ-risk can be rewritten as:

$$\eta(\mathbf{x})\phi(g(\mathbf{x})) + (1 - \eta(\mathbf{x}))\phi(-g(\mathbf{x})) - \inf_{\alpha \in \mathbb{R}} \{\eta(\mathbf{x})\phi(\alpha) + (1 - \eta(\mathbf{x}))\phi(-\alpha)\}. \tag{6}$$

We will now relate this quantity to $\mathbb{1}_{g(\mathbf{x})g^\star(\mathbf{x}) < 0} |2\eta(\mathbf{x}) - 1|$

Hinge loss.

Question 48: (*2 points.*) Compute the minimizer and the minimum of the conditional ϕ-risk for the hinge loss

$$\min_{\alpha \in \mathbb{R}} \{\eta(\mathbf{x})(1 - \alpha)_+ + (1 - \eta(\mathbf{x}))(1 + \alpha)_+\}. \tag{7}$$
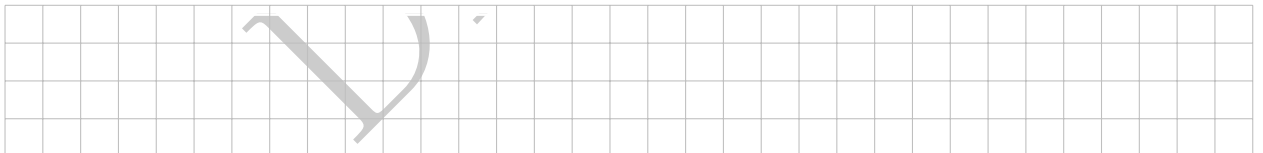
☐ 0 ☐ 1 ☐ 2

Question 49: (*3 points.*) Show that the excess conditional ϕ-risk is greater than the excess risk.

☐ 0 ☐ 1 ☐ 2 ☐ 3
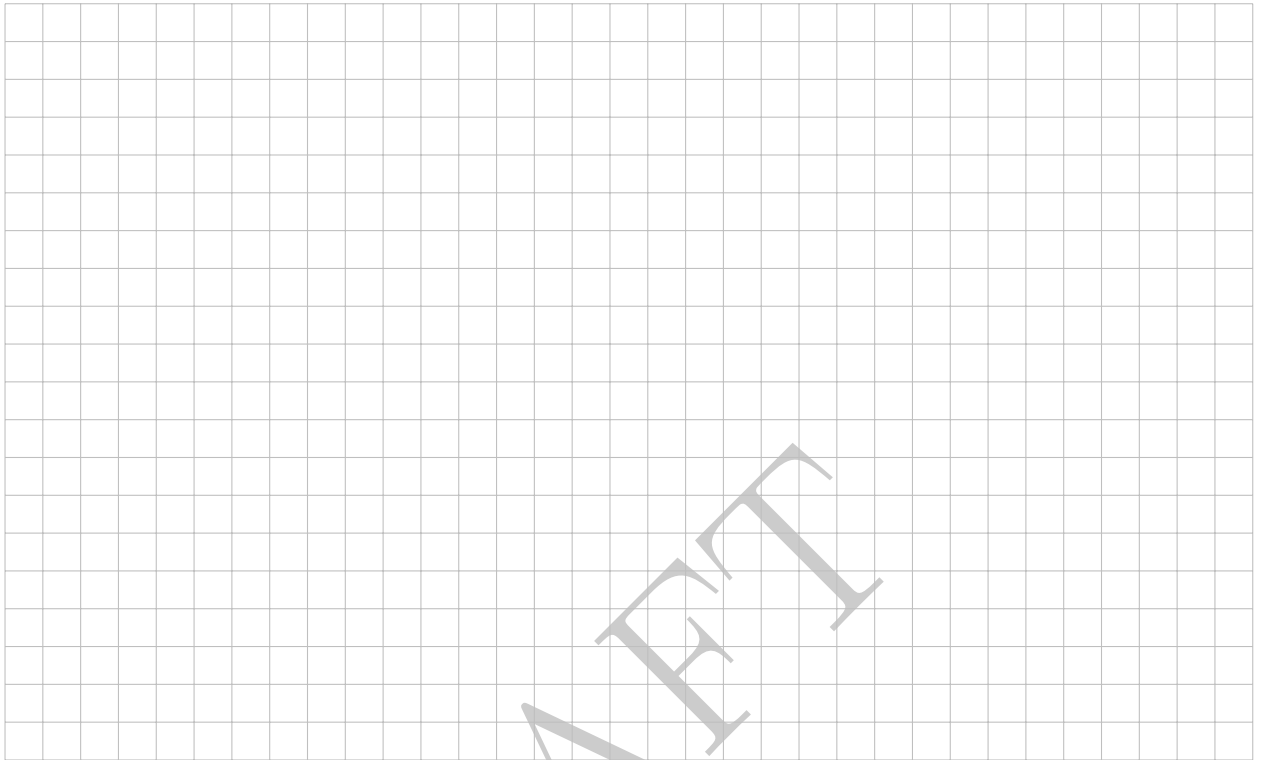
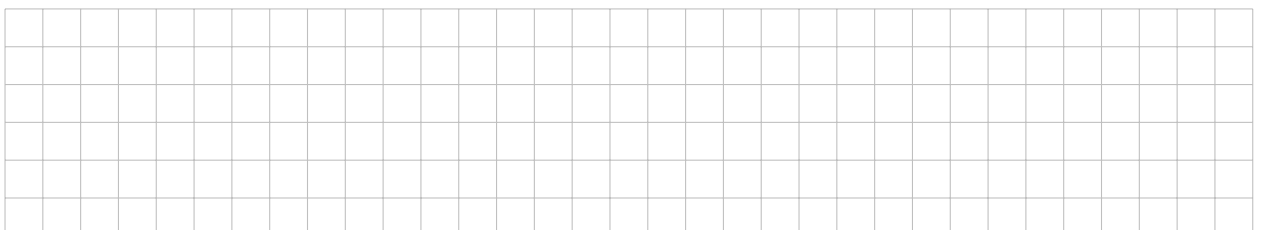**Question 50:** (*1 point.*) What is the calibration function for the Hinge loss?

☐ 0 ☐ 1

Quadratic and logistic loss. We consider smooth loss functions of the form (up to additive constants) $\phi(v) = a(v) - v$.

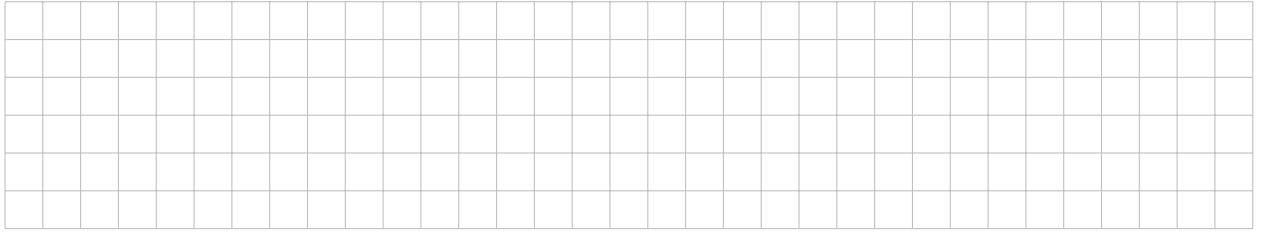**Question 51:** (*1 point.*) Compute the function a corresponding to the quadratic loss.

☐ 0 ☐ 1

Question 52: (*1 point.*) Compute the function a corresponding to the logistic loss.
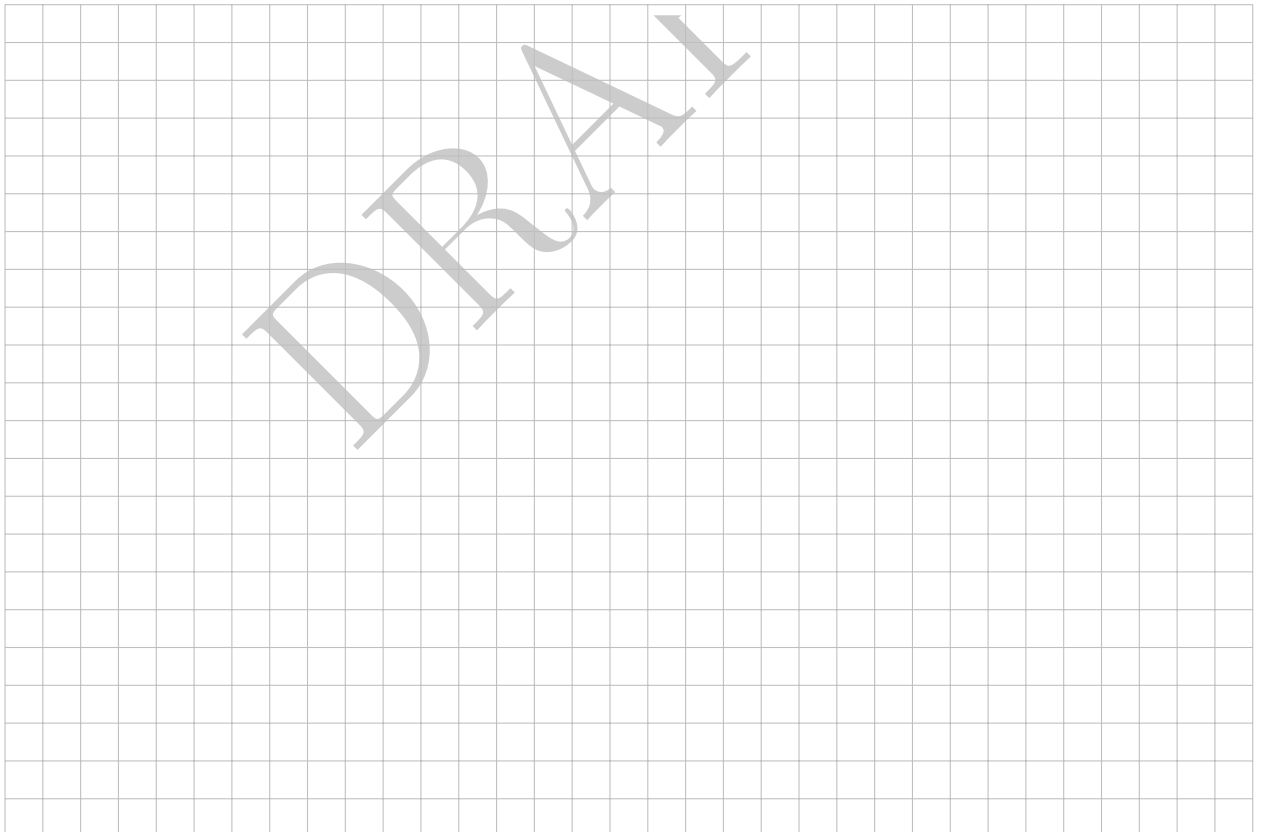
$\square_0$ $\square_1$

We assume that the function a is even, $a(0) = 0$, twice differentiable with $|a''(v)| \leq \beta$ for all $v \in \mathbb{R}$ (we call such function β-smooth).

Bonus Question 53: (*3 points.*) Show that for such function a we have:

$$a(v) - \alpha v - \inf_{w \in \mathbb{R}} \{a(w) - \alpha w\} \geq \frac{1}{2\beta} |\alpha - a'(v)|^2. \tag{8}$$

$\square_0$ $\square_1$ $\square_2$ $\square_3$

Question 54: (*2 points.*) Show that the excess conditional ϕ-risk is lower bounded by $\frac{1}{2\beta}|2\eta(\mathbf{x}) - 1 - a'(g(\mathbf{x}))|^2$.
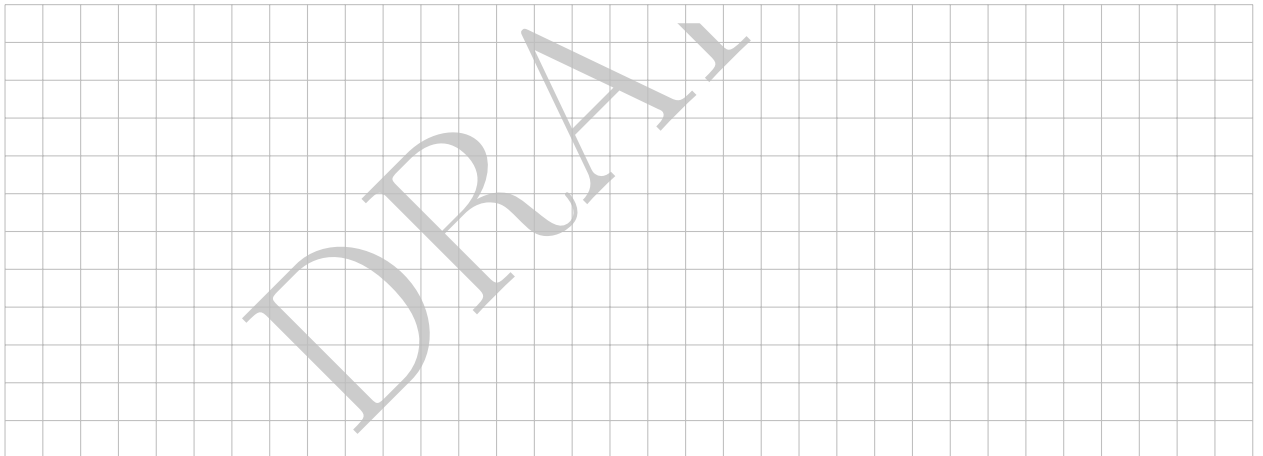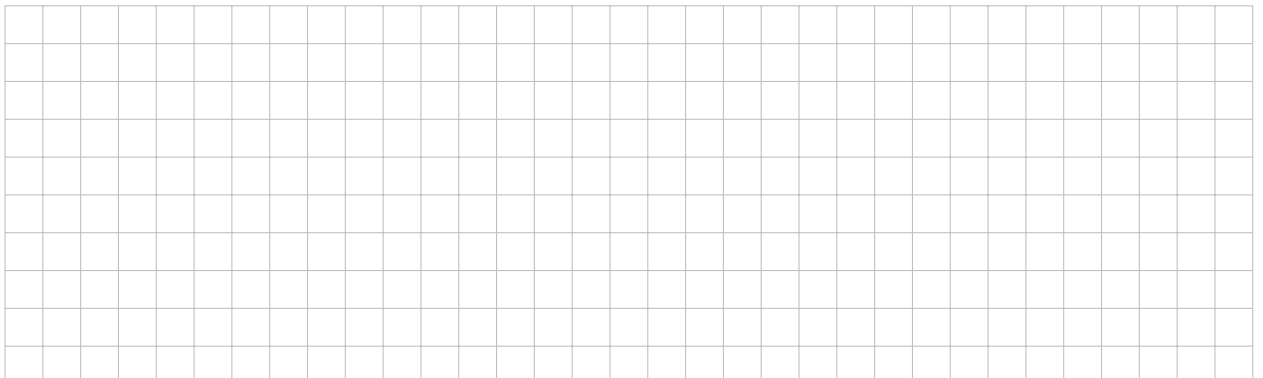
☐ 0 ☐ 1 ☐ 2

Question 55: (*2 points.*) Show that

$$\mathcal{L}_\phi(g) - \mathcal{L}_\phi^\star \geq \frac{1}{2\beta}(\mathcal{L}(g) - \mathcal{L}^\star)^2 \tag{9}$$
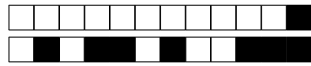
☐ 0 ☐ 1 ☐ 2

Question 56: (*1 point.*) What are the calibration functions for the quadratic loss and for the logistic loss?

☐ 0 ☐ 1