Introduction to Machine Learning
Final Exam (Questions Pack)

August 15, 2020

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 appropriate squares. *Nothing* written on pages of the question pack will be collected or marked and you may take this questions pack home if you wish to. **Only the separate answer sheet will be marked**. Collaboration on the exam is strictly forbidden. You are allowed a summary of 2 A4 pages and a simple, non-programmable calculator and a dictionary. The use of any other helping material will lead to being excluded from the exam and subjected to disciplinary measures by the ETH disciplinary committee.
Please make sure that your answer sheet is clean and all answers are clearly marked. You are allowed to use a pencil so that you can easily correct wrong answers. We reserve the right to classify answers as wrong without further consideration if the sheet is filled out ambiguously.

Question Types In this exam, you will encounter the following question types.

- **Multiple choice multiple answer questions (marked with ♣):** These multiple choice questions are all marked with the ♣ symbol at the top. They have **at least one** correct choice and **may contain multiple** correct choices. In some cases, all choices may be correct. **Three points** are awarded if all correct choices were identified and no incorrect choice was made and **zero points** are awarded otherwise.
- **Multiple choice single answer question (no mark):** Multiple choice questions that do not have the ♣ symbol have **exactly one** correct choice. **Three points** are awarded if the correct choice is identified, **negative one point** is awarded if a wrong choice is identified and **zero points** are given if not attempted.
- **True/False questions:** Each True/False questions has a value of **one point** if answered correctly, **negative one point** point if answered wrong and **zero** points if not attempted.

Not all questions need to be answered correctly to achieve the best grade.

1 Regression

1.1 Multiple choice questions

The following are multiple choice questions about ordinary least squares (OLS), ridge and L1-regularised (Lasso) regression and feature selection.

Question 1 ♣ For OLS regression, which of the following will never increase the least squares loss? *Hint: The bias term is the component of the weight vector associated with a constant feature for all samples.*

- [A] Setting the bias term to zero or not fitting a bias term.
- [■] Augmenting the set of features used for the regression.
- [C] Projecting all samples onto a lower dimensional feature space with PCA before performing regression on the projected samples.
- [■] Subtracting the empirical mean from the data before performing regression on the centered samples.

Question 2 In general, how do the bias and variance properties of the ridge regression estimator compare to those of the ordinary least squares (OLS) estimator?

- [■] The ridge regression estimator has larger bias and smaller variance than the OLS estimator.
- [B] The ridge regression estimator has larger bias and larger variance than the OLS estimator.
- [C] The ridge regression estimator has smaller bias and smaller variance than the OLS estimator.
- [D] The ridge regression estimator has smaller bias and larger variance than the OLS estimator.

Explanation: The ridge regression estimator has larger bias and smaller variance than the OLS estimator.

Question 3 ♣ Which of the following statements about ridge regression (with regularisation term $\lambda\|\boldsymbol{w}\|_2^2$ and $\lambda > 0$) are true?

- [■] The norm of the optimal weight vector is a monotonically decreasing function of λ.
- [B] As λ increases, model complexity decreases resulting in smaller bias and larger variance of the model.
- [C] The regularisation term can be interpreted as a Laplacian prior on the weight vector.
- [■] The objective function has a unique optimiser.

Question 4 For a ridge regression task, which of the sketches in Figure 1 below is the most likely to describe the training and test loss as a function of λ best (at least qualitatively)?

[■] A. [B] B. [C] C. [D] D.

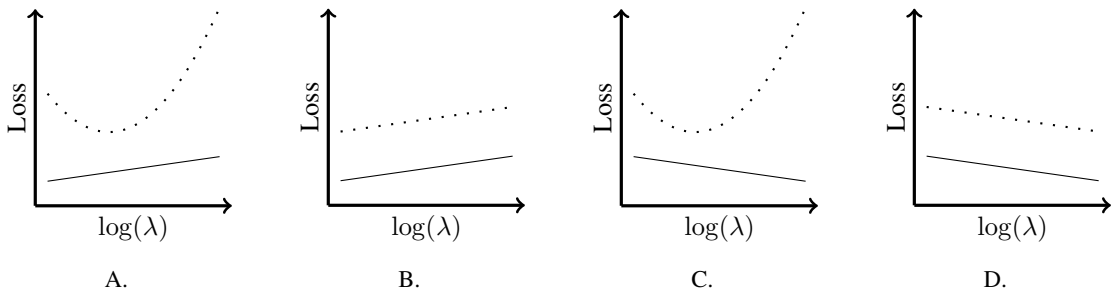

Figure 1: For a ridge regression task, Figures A. - D. are sketches to qualitatively describe the training (solid line) and test loss (loosely dotted line) respectively as a function of λ.

Question 5 ♣ Which of the following statements about L1-regularised (Lasso) regression are true?

A The optimiser of the objective function can be computed in closed form.

■ Lasso regression selects a subset of the input features.

C Standardising the data (centering and scaling variance of each feature to a value of one) will not change the optimal value of the objective function.

D Greedy forward selection always selects a model with fewer features than Lasso regression.

Question 6 ♣ Which of the following statements are true about greedy forward selection with d features for regression?

A $\mathcal{O}(2^d)$ models must be trained for greedy forward selection.

B It greedily adds features to reduce the training loss.

■ Greedy forward selection is faster than backward selection if only few features are relevant.

D It always finds the subset of features with the lowest validation loss.

1.2 Linear Regression and Line Search

Consider the following regression setting: You are given a design matrix $X \in \mathbb{R}^{n \times d}$ and a target vector $y \in \mathbb{R}^n$ for a dataset with n samples and features of dimension d. For this question, assume that $n > d$ unless otherwise stated. For this dataset, a linear regression model minimises the following objective,

$$L(\boldsymbol{w}) = \|\boldsymbol{y} - X\boldsymbol{w}\|^2, \tag{1}$$

where $\boldsymbol{w} \in \mathbb{R}^d$ and $\|\cdot\|$ is the Euclidean norm.

Question 7 Which of the following expressions gives the minimiser of equation 1 in closed form?

A $\hat{\boldsymbol{w}} = \left(X^\top X\right)^{-1} Xy$

■ $\hat{\boldsymbol{w}} = \left(X^\top X\right)^{-1} X^T y$

C $\hat{\boldsymbol{w}} = \left(XX^\top\right)^{-1} Xy$

D $\hat{\boldsymbol{w}} = \left(XX^\top\right)^{-1} X^T y$

Explanation: Basic bookwork.

Question 8 What is the computational complexity of computing the closed form solution?

A $\Theta(n^3)$ B $\Theta(n^2 d)$ ■ $\Theta(nd^2)$ D $\Theta(d^3)$

Question 9 For large n and d, instead of computing the closed form solution, you may consider minimising equation 1 using methods based on gradient descent. Such methods typically perform iterative updates of the following form

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta_t \nabla L(\boldsymbol{w})|_{\{\boldsymbol{w}=\boldsymbol{w}_t\}} \tag{2}$$

where $\boldsymbol{w}_t$ and $\boldsymbol{w}_{t+1}$ are the solutions at iteration t and $t+1$ respectively and η_t is the step size at iteration t. What is the correct expression for $\nabla L(\boldsymbol{w})$?

A $X^\top X\boldsymbol{w} - 2X^\top \boldsymbol{y}$
B $2X\boldsymbol{w} - 2XX^\top \boldsymbol{y}$
C $2X\boldsymbol{w} - XX^\top \boldsymbol{y}$
D $2X^\top X\boldsymbol{w} - X^\top \boldsymbol{y}$
E $X^\top X\boldsymbol{w} - X^\top \boldsymbol{y}$
F $X\boldsymbol{w} - 2XX^\top \boldsymbol{y}$
G $X\boldsymbol{w} - XX^\top \boldsymbol{y}$
■ $2X^\top X\boldsymbol{w} - 2X^\top \boldsymbol{y}$

Explanation: Basic bookwork.

Question 10 What is the computational complexity of computing $\nabla L(\boldsymbol{w})$ at a specific $\boldsymbol{w}_t$?

■ $\Theta(dn)$
C $\Theta(d^2n)$
E $\Theta(dn\log(d))$
G $\Theta(d^2n\log(d))$
B $\Theta(dn^2)$
D $\Theta(nd^2+n^2d)$
F $\Theta(dn\log(n))$
H $\Theta(dn^2\log(n))$

Question 11 It is possible to improve the gradient descent schedule by choosing an optimal step size. For example, we can use a line search and choose η_t to optimise the following objective,

$$\eta_t^* = \arg\min_{\eta\in\mathbb{R}} L\left(\boldsymbol{w_t} - \eta J\right) \tag{3}$$

where $J = \nabla_{\boldsymbol{w}} L|_{\{\boldsymbol{w}=\boldsymbol{w_t}\}}$. What is the optimal value η_t^*?

A $\frac{(\boldsymbol{y}-X\boldsymbol{w})^\top XJ}{\|XJ\|_2^2}$
E $\frac{X^\top(\boldsymbol{y}-X\boldsymbol{w})J}{\|XJ\|_2^2}$
B $\frac{(\boldsymbol{y}-X\boldsymbol{w})^\top XJ}{\|\boldsymbol{y}-X\boldsymbol{w}\|_2^2}$
F $\frac{X^\top(\boldsymbol{y}-X\boldsymbol{w})J}{\|\boldsymbol{y}-X\boldsymbol{w}\|_2^2}$
■ $\frac{(X\boldsymbol{w}-\boldsymbol{y})^\top XJ}{\|XJ\|_2^2}$
G $\frac{X^\top(X\boldsymbol{w}-\boldsymbol{y})J}{\|XJ\|_2^2}$
D $\frac{(X\boldsymbol{w}-\boldsymbol{y})^\top XJ}{\|\boldsymbol{y}-X\boldsymbol{w}\|_2^2}$
H $\frac{X^\top(X\boldsymbol{w}-\boldsymbol{y})J}{\|\boldsymbol{y}-X\boldsymbol{w}\|_2^2}$

Question 12 What is the computational complexity of performing one step of line search (including the computation of J)?

■ $\Theta(dn)$
C $\Theta(d^2n)$
E $\Theta(dn\log(d))$
G $\Theta(d^2n\log(d))$
B $\Theta(dn^2)$
D $\Theta(nd^2+n^2d)$
F $\Theta(dn\log(n))$
H $\Theta(dn^2\log(n))$

Explanation: Computing J takes time $\Theta(d\cdot n)$. Computing $\frac{(X\boldsymbol{w}-\boldsymbol{y})^\top XJ}{\|XJ\|_2^2}$ is dominated by $(X\boldsymbol{w})$ which also takes time $\theta(d\cdot n)$

2 Kernels and Support Vector Machines

2.1 Kernel Functions

For each of the following functions k, decide if is a valid kernel function (True) or not (False). *Hint: For the questions involving set operations, try to represent a subset as a binary vector.*

Question 13 Let $x, x' \in \mathbb{R}$ and $k(x, x') = x^2 + (x')^3 + 1$

A True ■ False

Question 14 Let $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^n$ and $k(\boldsymbol{x}, \boldsymbol{x}') = (\boldsymbol{x}^\top \boldsymbol{x}' - 1)^2$

A True ■ False

Question 15 Let $A, B \subseteq \Omega$ for some finite set Ω and $k(A, B) = |A \cap B|^2$.

■ True B False

Question 16 Let $A, B \subseteq \Omega$ for some finite set Ω and $k(A, B) = 2|A \cap B| - |A| - |B|$.

A True ■ False

Question 17 Let $A, B \subseteq \Omega$ for some finite set Ω and $k(A, B) = \exp\left(\frac{1}{2} \cdot |A \cap B|\right)$.

■ True B False

2.2 Support Vector Machines

Consider the following binary classification task: You are given a dataset $\{(\boldsymbol{x}_i, y_i)\}_{i \in \mathcal{D}}$ with $\boldsymbol{x}_i \in \mathbb{R}^2$ and $y_i \in \{-1, +1\}$ for all $i \in \mathcal{D}$. This dataset is displayed in Figure 2.
You are asked to fit a linear support vector machine with parameters $\boldsymbol{w} \in \mathbb{R}^2$ and $b \in \mathbb{R}$ to minimise the following objective,

$$L(\boldsymbol{w}, b, \mathcal{D}) = \frac{1}{2}\|\boldsymbol{w}\|_2^2 + \sum_{i \in \mathcal{D}} \max\left\{0, 1 - y_i\left(\boldsymbol{w}^\top \boldsymbol{x}_i + b\right)\right\}, \tag{4}$$

where $\|\cdot\|$ is the Euclidean norm. The objective in Equation (4) could be minimised using stochastic gradient descent with a constant step size η to perform iterative updates of the following form:

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \nabla_{\boldsymbol{w}} L(\boldsymbol{w}, b, \mathcal{B}_t)|_{\boldsymbol{w}=\boldsymbol{w}_t, b=b_t} \tag{5}$$

$$b_{t+1} = b_t - \eta \nabla_b L(\boldsymbol{w}, b, \mathcal{B}_t)|_{\boldsymbol{w}=\boldsymbol{w}_t, b=b_t} \tag{6}$$

where $\boldsymbol{w}_t$ and b_t are the solutions at iteration t and $\mathcal{B}_t$ is a batch (subset) of the dataset at iteration t. For the remainder of this question, define $\frac{d}{dz}\max\{0, z\}\,|_{z=0} = 0$.

Question 18 Let $\mathcal{B}_t = \{A, B\}$, $\boldsymbol{w}_t = (1, 0)^\top$ and $b_t = -8$. What is $\nabla_{\boldsymbol{w}} L(\boldsymbol{w}, b, \mathcal{B}_t)|_{\boldsymbol{w}=\boldsymbol{w}_t, b=b_t}$?

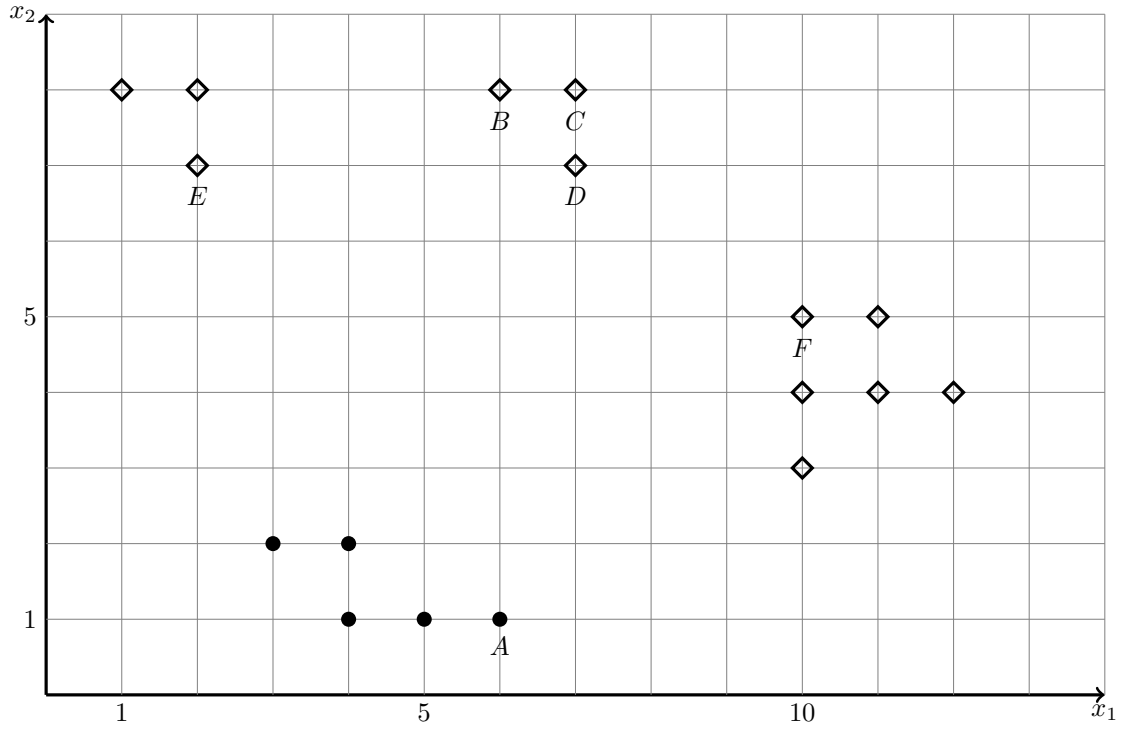

Figure 2: Binary classification dataset $\mathcal{D}$ in $\mathbb{R}^2$ with $y = -1$ as symbol ● and $y = +1$ as symbol ◇.

A $(-10, -9)^\top$	C $(-5, -1)^\top$	E $(1, 0)^\top$	G $(6, 1)^\top$
■ $(-5, -8)^\top$	D $(1, -7)^\top$	F $(1, 7)^\top$	H $(6, 8)^\top$

Explanation: For point A with $y = -1$, the margin is -2 and hence the Hinge loss is saturated. For point B the margin is also -2, but $y = +1$. The gradient of the Hinge loss (if not saturated) is $-yx$. Hence, the correct answer is given by $\boldsymbol{w} - y_B \boldsymbol{x_B} = (-5, -8)^\top$.

Question 19 ♣ For the optimal SVM classifier $(\boldsymbol{w}^*, b^*)$ that minimises the objective in Equation (4), which of the following labelled samples in the dataset (see again Figure 2) are support vectors? *Hint: There may be other samples (not among the choices below) that are support vectors.*

■ A. B B. C C. D D.

Explanation: B, C, D cannot be support vectors since we are fitting a linear classifier and any reasonable decision boundary would result in other samples from the same class as B, C, D being closer to the decision boundary. It is immediate that A must be a support vector from the same reasoning.

Question 20 You learn that the data points B, C and D in Figure 2 were incorrectly labelled and that they actually belong to the class $y = -1$ (filled circle symbol). You observe that the data is not anymore linearly separable in the input space. Therefore, you consider fitting a linear decision boundary in a higher dimensional space with feature map $\phi(\boldsymbol{x}) \in \mathbb{R}^d$ for some $d \in \mathbb{N}, d > 2$ instead. For this purpose, you set the bias $b = 0$ and recall that the optimal weight vector $\hat{\boldsymbol{w}} \in \mathbb{R}^d$ for the SVM classifier will be given by:

$$\hat{\boldsymbol{w}} = \sum_{i \in \mathcal{D}} a_i y_i \phi(\boldsymbol{x}_i) \tag{7}$$

where $\alpha_i \in \mathbb{R}$ for $i \in \mathcal{D}$ are some trained parameters. You decide to use the following kernel,

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \frac{1}{100}(\boldsymbol{x}_i^\top \boldsymbol{x}_j + 1)^2. \tag{8}$$

What is the dimension of the feature vector $\phi(\boldsymbol{x}) \in \mathbb{R}^d$ corresponding to the kernel given in Equation (8)?

A $d = 2$	C $d = 4$	■ $d = 6$	G $d = 8$
B $d = 3$	D $d = 5$	F $d = 7$	H $d = 9$

Explanation: The feature map is $\phi(\boldsymbol{x}) = \left[x_1^2, x_2^2, x_1x_2, \sqrt{2}x_1, \sqrt{2}x_2, 1\right]^\top$ and hence $d = 6$.

Question 21 Continuing from the previous question (with data points B, C and D belonging to class $y = -1$) , assume the set of support vectors corresponding to the optimal SVM classifier $\boldsymbol{w}^*$ with the kernel in Equation 8 is $\mathcal{S} = \{A, B, E, F\}$ with $\alpha_A = 2$, $\alpha_B = 2$, $\alpha_E = 3$, $\alpha_F = 1$. Define the signed margin at a point x as

$$\pi(\boldsymbol{x}) = \hat{\boldsymbol{w}}^\top \phi(\boldsymbol{x}). \tag{9}$$

For the test point $\boldsymbol{x_T} = (5, 5)^\top$, using $k(\boldsymbol{x_A}, \boldsymbol{x_T}) = 12.96$, $k(\boldsymbol{x_B}, \boldsymbol{x_T}) = 50.41$, $k(\boldsymbol{x_E}, \boldsymbol{x_T}) = 21.16$ and $k(\boldsymbol{x_F}, \boldsymbol{x_T}) = 57.76$, compute its signed margin $\pi(\boldsymbol{x_T})$.

A $\pi(\boldsymbol{x_T}) = -25.92$	E $\pi(\boldsymbol{x_T}) = -1.00$
B $\pi(\boldsymbol{x_T}) = -12.96$	F $\pi(\boldsymbol{x_T}) = 0.00$
■ $\pi(\boldsymbol{x_T}) = -5.50$	G $\pi(\boldsymbol{x_T}) = 20.42$
D $\pi(\boldsymbol{x_T}) = -2.50$	H $\pi(\boldsymbol{x_T}) = 31.10$

Explanation: The margin is $\pi(\boldsymbol{x}) = \sum_{i \in \mathcal{S}} a_i y_i k(\boldsymbol{x_i}, \boldsymbol{x})$. We have $k(\boldsymbol{x_A}, \boldsymbol{x_T}) = 1296$. Hence, we have $0.02 \times -1296 + 0.02 \times -5041 + 0.03 \times 2116 + 0.01 \times 5776 = -5.5$.

3 Classification and Probabilistic Modelling

3.1 Classification

Question 22 ♣ For a binary classification problem, let $f(\boldsymbol{x}) : \mathcal{X} \to \mathbb{R}$ be a predictor for which a classifier can be defined as

$$\hat{y}_\tau(\boldsymbol{x}) = \begin{cases} 1 & \text{if } f(\boldsymbol{x}) > \tau, \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

for any decision threshold $\tau \in \mathbb{R}$. Which of the following statements about the ROC curve of f is true?

- A The ROC curve is increasing and concave for any f.
- ■ The ROC curve plots the true positive rate (y-axis) against the false positive rate (x-axis) for classifiers obtained from f at different thresholds τ.
- ■ The area under the ROC curve associated with a random predictor $f(\boldsymbol{x}) \overset{\text{d}}{=} \text{Uniform}(0, 1)$, is one half (in expectation), even if the dataset is imbalanced. *Hint: The notation $f(\boldsymbol{x}) \overset{\text{d}}{=} Uniform(0, 1)$ means for any $\boldsymbol{x}$ a number is sampled uniformly at random from the interval* $[0, 1]$.
- ■ The area under the ROC curve associated with a perfect predictor f is 1.

Explanation: The ROC curve is increasing but not concave. Indeed, true positive rate is on the y-axis. You may not always choose the decision threshold such that the area under the ROC curve is maximized. For example, if you have an unbalanced dataset or care about false positives more than false negatives.

Question 23

	Positive Condition	Negative Condition
Positive Prediction	10	30
Negative Prediction	10	50

The table above shows the confusion matrix for a binary classifier. What is the F1 score of this binary classifier given the results in the table?

A $\frac{1}{4}$ ■ $\frac{1}{3}$ C $\frac{1}{2}$ D $\frac{3}{5}$

Explanation: The precision is given by 0.25. The recall is given by 0.5. Therefore, the F1 score is $2 \times \frac{0.25 \times 0.5}{0.25 + 0.5} = \frac{1}{3}$.

Question 24 Consider a (binary) logistic regression model $P(Y = y|\boldsymbol{x}, \boldsymbol{w}) = \frac{1}{1+\exp(-y\boldsymbol{w}^\top \boldsymbol{x})}$ parameterized by $\boldsymbol{w}$ and define a classifier with the rule $P(Y = 1|\boldsymbol{x}, \boldsymbol{w}) > 0.9$ as follows

$$\hat{y}_{\boldsymbol{w}}(\boldsymbol{x}) = \begin{cases} +1 & \text{if } P(Y = 1|\boldsymbol{x}, \boldsymbol{w}) > 0.9 \\ -1 & \text{otherwise.} \end{cases} \tag{11}$$

Which of the following expressions does the rule $P(Y = 1|\boldsymbol{x}, \boldsymbol{w}) > 0.9$ simplify to?

- A $\boldsymbol{w}^\top \boldsymbol{x} > 0.9$
- B $\boldsymbol{w}^\top \boldsymbol{x} > \ln(0.9)$
- C $\boldsymbol{w}^\top \boldsymbol{x} > -1/9$
- ■ $\boldsymbol{w}^\top \boldsymbol{x} > -\ln(1/9)$

Question 25 ♣ Consider a multi-class logistic regression model with K classes. Recall the model maintains parameters $\boldsymbol{w}_k$ for each class $k \in \{1, \ldots, K\}$. You can temper the output probabilities introducing a temperature parameter $\tau > 0$. For test point $\boldsymbol{x}$, the model then predicts probability for class j as

$$P_\tau(Y = k|\boldsymbol{x}) = \frac{\exp(\boldsymbol{w}_k^\top \boldsymbol{x}/\tau)}{\sum_{i=1}^K \exp(\boldsymbol{w}_i^\top \boldsymbol{x}/\tau)}$$

Let $\hat{y}_\tau = \arg\max_k P_\tau(Y = k \mid \boldsymbol{x})$ be the predicted class label at temperature τ. Which of the following statements about tempered multi-class logistic regression are true?

■ For all $\tau > 0$, the predicted class label $\hat{y}_\tau$ is the same.

B If $\tau \neq 1$, the predicted class probabilities no longer outputs valid class probabilities, i.e., $\sum_{k=1}^{K} P_\tau(Y = k|\boldsymbol{x}) \neq 1$.

C In general, as the temperature increases without bound ($\tau \to \infty$), the predicted class probability converges to one, such that $P_\tau(Y = \hat{y}_\tau \mid \boldsymbol{x}) \to 1$.

D In general, as the temperature decreases to zero ($\tau \to 0^+$), the class distribution converges to a uniform distribution and $P_\tau(Y = \hat{y}_\tau \mid \boldsymbol{x}) \to \frac{1}{K}$.

Explanation: Softmax is a monotonic transformation and always projects onto the simplex. The asymptotic statements are in the wrong order.

3.2 Regularization and Probabilistic Modeling

Consider the following Bayesian regression model for fitting a quadratic function,

$$\begin{aligned} y &= \left(\theta_0 + \theta_1 x + \theta_2 x^2\right) + \varepsilon \\ \varepsilon &\sim \text{Normal}(0, \sigma^2) \\ \theta_i &\sim \text{Laplace}\,(0, s) \;\text{ for all } i \in \{0, 1, 2\} \end{aligned}$$

where σ and s are assumed to be known. You are given a dataset $\{(x_i, y_i)\}_{i=1}^n$ generated from this model.

Question 26 ♣ Which of the following optimisation problems solves correctly for the MAP (maximum-a-posteriori) estimate?

■ $\arg\max_\theta p(\theta) \prod_{i=1}^n p_\theta(y_i|x_i)$

B $\arg\max_\theta \prod_{i=1}^n p(\theta) p_\theta(y_i|x_i)$

C $\arg\min_\theta \log p(\theta) + \sum_{i=1}^n p_\theta(y_i|x_i)$

■ $\arg\min_\theta - \log p(\theta) - \sum_{i=1}^n \log p_\theta(y_i|x_i)$

Explanation: The MAP estimate maximizes $P\left(\theta|\{(x_i, y_i)\}_{i=1}^N\right) \propto P(\theta)P(\{(x_i, y_i)\}_{i=1}^N|\theta)$. Hence, the MAP estimate for the given dataset maximizes $p(\theta) \prod_{i=1}^N p_\theta(y_i|x_i)$ or equivalently minimizes the negative log of this expression given by $-\log p(\theta) - \sum_{i=1}^N p_\theta(y_i|x_i)$.

Question 27 For the given model, MAP estimation may be written as a regularized least squares optimization problem in the following form

$$\arg\min_\theta \sum_{i=1}^{n} (y_i - h_\theta(x_i))^2 + \lambda C(\theta) \tag{12}$$

where $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2$. Recall that the density of the Laplace$(0, s)$ distribution is given by

$$p(\theta) = \frac{1}{2s} \exp\left(-\frac{|\theta|}{s}\right).$$

What are the correct expressions for λ and C in Equation (12)?

A $\lambda = \frac{2}{s\sigma^2}$ and $C(\theta) = -\sum_{i=0}^2 |\theta_i|$

B $\lambda = \frac{1}{s\sigma^2}$ and $C(\theta) = \sum_{i=0}^2 |\theta_i|$

■ $\lambda = \frac{2\sigma^2}{s}$ and $C(\theta) = \sum_{i=0}^2 |\theta_i|$

D $\lambda = \frac{\sigma^2}{s}$ and $C(\theta) = -\sum_{i=0}^2 |\theta_i|$

Explanation: Using the correct definition of the MAP problem in negative log form and simplifying gives $\lambda = \frac{2\sigma^2}{s}$ and $C(\theta) = \sum_{i=0}^2 |\theta_i|$.

4 Dimension reduction and clustering

4.1 Multiple choice questions

The following are multiple choice questions with at least one correct choice and possibly multiple correct choices about clustering and dimension reduction.

Question 28 ♣ Which of the following statements are true about k-means clustering?

- ■ It seeks cluster centres and assignments to minimise the within-cluster sum of squares.
- ■ It is appropriate if the underlying clusters are separable, spherical and approximately of same size.
- C For fixed assignments of sample points to cluster centres, computing the optimal cluster centres is a non-convex optimisation problem.
- ■ k-means clustering can be kernelised.

Question 29 ♣ Which of the following statements are true about Lloyd's algorithm for k-means clustering?

- ■ It cannot cycle; i.e. it does never return to a particular solution after having previously changed to a different solution.
- B It always terminates with the globally optimal solution.
- C The number of iterations until convergence is guaranteed to be polynomial in the number of cluster centres and data points.
- ■ Using specialised initialisation schemes (e.g. k-means++) can improve the quality of solutions found by the algorithm and reduce its runtime.

Explanation: k-means++ does lead to better solutions and reduced runtime in practice. The algorithm cannot cycle, since Lloyd's algorithm is guaranteed to monotonically decrease the average squared distance. It does not terminate to a globally optimal solution, k-means clustering is NP-hard and in the worst-case the algorithm may take super-polynomial time to converge.

Question 30 ♣ In k-means, how can the number of cluster centers k be selected?

- ■ By using a heuristic like the elbow method that identifies the diminishing returns from increasing k.
- ■ By using an information criterion that regularises solutions to favour simpler models with lower k.
- C By using a validation set to select the best k on the held-out data.
- D By using an algorithm like Lloyd's algorithm that automatically selects k during runtime.

Explanation: The elbow method and using an information criterion are well-known techniques for picking k. Lloyd's algorithm requires k to be specified by the user. A validation set cannot be used, because the validation criterion typically improves in k.

Question 31 ♣ Which of the following is true about principal component analysis (PCA)?

- A PCA is a supervised learning algorithm.
- B PCA is a method for non-linear dimension reduction.
- C If the underlying data distribution is a Gaussian distribution with diagonal covariance matrix, then PCA is equivalent to k-means clustering.
- ■ PCA can be kernelised.

Question 32 ♣ Which of the following is true about the first principal component found by PCA?

■ It is orthogonal to all other principal components found by PCA.

B It is the direction that minimises the variance of the projected data.

C Scaling some of the features with a factor $c > 1$ does not change the first principal component if the data is centred.

■ It corresponds to a line that minimises the sum of squares of the distances of the sample points from that line.

Question 33 ♣ Assume you are given data $\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^n$ with each $\boldsymbol{x_i} \sim \mathcal{N}(\mathbf{0}, C)$ and $C = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$. Which of the following statements is true for performing PCA on $\mathcal{D}$?

A The expected value of the first principal component is $\mathbb{E}[\boldsymbol{w}_1] = \left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right)^\top$.

B The variance associated with the first and second principal component is the same in expectation.

■ Assuming instead $\boldsymbol{x_i} \sim \mathcal{N}(\mathbf{0}, C')$ with $C' = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$ does not change the principal components in expectation.

■ The first two principal components are sufficient to perfectly reconstruct the data.

Explanation: The dataset is clearly in $\mathbb{R}^2$, there are only two principal components, so these are sufficient for perfect reconstruction. The first principal component is not $\left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right)^\top$ but $\left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)^\top$ instead. Making the positive correlation weaker does not change the directions of the principal components.

5 Neural Networks

5.1 Multiple choice questions

The following are multiple choice questions with at least one correct choice and possibly multiple correct choices about neural networks.

Question 34 ♣ Which of the following statements about backpropagation used for computing gradients when training neural networks are true?

■ It can be applied to compute gradients for neural networks for unsupervised learning.

B On GPU, its computation can be easily parallelised over the different layers.

■ It is based on the chain rule for differentiation.

D Its running time grows quadratically in the number of parameters in a feedforward network.

Explanation: The running time grows linearly in the number of parameters. Backpropagation is a dynamic programme. It is easily parallelized over the batch dimension, but the gradients for different layers are computed sequentially. Hence, it is not easily parallelized, even though this is an active area of research. Backpropagation can be used for unsupervised learning (e.g. auto-encoders).

Question 35 ♣ Which of the following statements about the vanishing gradient problem in neural networks are true?

A A neural network that suffers from vanishing gradients for one training example will suffer from vanishing gradients for all training examples.

■ Neural networks with ReLU activations are typically less susceptible to suffer from vanishing gradients than those with sigmoid activations.

C Shallow networks do not suffer from vanishing gradients regardless of their weight initialisation.

■ Batch Normalization can sometimes alleviate the vanishing gradient problem.

Explanation: You can easily design two training examples, only one of them will induce vanishing gradients. Similarly, you could choose very large weights (at initialization or because your learning rate is too high) in a shallow network to produce vanishing gradients. ReLU units have been shown to suffer less from vanishing gradients. Intuitively, they have non-zero gradients for a larger domain, i.e. the positive real line. They do not completely solve the problem (dead ReLU units). Batch Normalization has been observed to alleviate vanishing gradients, because it rescales the input units to a hidden layer (and the gradients).

Question 36 ♣ Which of the following statements about nonlinear activation functions in neural networks are true?

A On GPUs, they can speed up the gradient calculation in backpropagation as compared to linear units.

■ They help to learn nonlinear decision boundaries.

C They are often only applied to the output units.

D They are everywhere differentiable.

Explanation: They do not speed up back-propagation, instead they introduce an additional operation that needs to be performed. Typically, they are applied to the hidden units to learn nonlinear decision boundaries. ReLUs are not differentiable at zero.

Question 37 ♣ When using neural networks, which of the following methods will typically result in a lower training error?

■ Adding an additional hidden layer with a non-linear activation function to the network.

B Reducing the learning rate when training for a fixed number of iterations with stochastic gradient descent.

C Training with weight decay to regularize the L2 norm of the weights of the network.

D Increasing the batch size when training for a fixed number of epochs with stochastic gradient descent.

Explanation: An additional hidden layer increases the capacity of the network. Weight decay reduces the model complexity and thus does not result in lower training error generally. When training with a larger batch size for a fixed number of epochs, fewer weight updates are performed. As with a smaller learning rate, the model may not converge as a result. Training error is thus not generally lower.

Question 38 ♣ Which of the following methods for training neural networks are generally believed to mitigate overfitting?

■ Early Stopping.
■ Dropout.
■ Weight decay.
■ Batch Normalization.

Explanation: All methods are believed to mitigate overfitting (Neural Networks III). In particular, for batch normalization also see Ioffe and Szegdy (2015), batch normalization "[...] also acts as a regularizer, in some cases eliminating the need for Dropout".

Question 39 ♣ Which of the following statements about convolutional neural networks (CNNs) for image analysis are true?

A They do not require non-linear activations to learn non-linear decision boundaries.

B They can only be used in shallow neural networks.

■ Pooling layers reduce the spatial resolution of the image.

D They cannot be used for unsupervised learning.

Explanation: Deep convolutional neural networks are a thing. Convolutional networks can be used for unsupervised learning and pooling layers reduce spatial resolution. They definitely require non-linear activation functions as the convolutional operation itself is a linear transformation.

Question 40 ♣ Which of the following statements about Generative Adversarial Networks (GANs) are true?

■ GANs are a modular neural network architecture comprised of a generator and a discriminator.

■ Training a GAN requires finding a saddle point of the objective function rather than a local optimum.

C In practice, training a GAN is easy if the generator and discriminator have enough capacity.

D GANs can be evaluated by computing their log-likelihood for held-out samples.

Explanation: GANs are indeed a modular neural network architecture. The equilibrium of the minimax game they pose is a saddle point. Training GANs is notoriously difficult and subject to much research. GANs are difficult to evaluate and compare against each other, because they do not fit a likelihood function.

5.2 Convolutional Networks in 3D

The two-dimensional convolutional layers seen in class for image processing can be generalized to three dimensions (3D). Input data in 3D arises naturally in spatial modelling and medical imaging. Let I be such a 3D image of shape $W \times H \times D \times C$, where W, H and D are width, height and depth of the image respectively, and C is the number of channels. See Figure 3.
The following questions consider 3D convolutional layers. For all questions, assume that spatial dimensions $W = H = D = 30$ and $C = 3$ channels for I and ignore the bias term, i.e. only consider the parameters of the weight matrix.

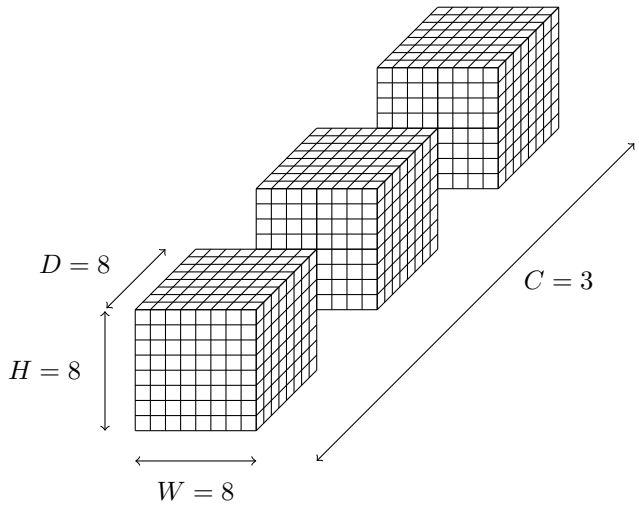

Figure 3: An example of a 3D image I with $W = H = D = 8$ and $C = 3$ for illustration only. The figure is not to scale. Note that for the questions you are given $W = H = D = 30$ and $C = 3$.

Question 41 How many parameters does the 3D convolutional layer have when using a kernel size of 5 in all dimensions and 100 filters?

A 375	C $12,500$	■ $37,500$	G $2,700,000$
B $7,500$	D $25,000$	F $81,000$	H $8,100,000$

Explanation: $100 \times 3 \times 5 \times 5 \times 5 = 37,500$

Question 42 For the same convolutional layer as above, what is the number of elements in the output when using stride 3 and padding 1 in all dimensions?

A 40,000	C 60,000	E 80,000	■ 100,000
B 50,000	D 70,000	F 90,000	H 110,000

Explanation:

$$x = \frac{30 + 2 - (5 - 1) - 1}{3} + 1 = 10$$
$$100 \times x \times x \times x = 100,000$$

Question 43 Assume the image was preprocessed and projected to $2D$ such that $D = 1$. How many parameters does the 2D convolutional layer have when using a kernel size of 5 and 100 filters?

A 375	C $12,500$	E $37,500$	G $2,700,000$
■ $7,500$	D $25,000$	F $81,000$	H $8,100,000$

Explanation:

$$100 \times 3 \times 5 \times 5 = 7,500$$

Question 44 In comparison, how many parameters does a fully connected layer have that produces 100 elements in the output from processing I (again with $W = H = D = 30$ and $C = 3$)?

A 375	C $12,500$	E $37,500$	G $2,700,000$
B $7,500$	D $25,000$	F $81,000$	■ $8,100,000$

Explanation:

$$100 \times 3 \times 30 \times 30 \times 30 = 81,000,000$$

6 Expectation Maximisation

6.1 True/ False questions

The following statements are about using the Expectation Maximisation (EM) algorithm to learn the parameters and latent (probabilistic) assignments of a Gaussian Mixture Model (GMM). For each of the statements, decide if it is true or false.

Question 45 The EM algorithm only converges to a local maximum or a saddle point of the objective function when using careful initialization.

A True ■ False

Question 46 Every iteration of the EM algorithm increases the marginal likelihood (of the data).

■ True B False

Question 47 Instead of the EM algorithm, it is possible to adapt gradient descent for learning the parameters of the GMM and its latent assignments.

■ True B False

Question 48 The step size of the EM algorithm may be tuned via random search or cross-validation.

A True ■ False

Explanation: The EM algorithm does not have a step size. We can indeed use a gradient-based method to optimise the marginal likelihood with respect to the GMM parameters and the latent assignments. The EM algorithm always converges irrespective of its initialisation and it increases the marginal likelihood function on every iteration.

6.2 Diagnostics with Expectation Maximization

In response to a new disease, you decided to develop your own diagnostic test. You convinced four friends Anton (A), Bella (B), Charlie (C) and Dora (D) to volunteer for initial trials. You tested each friend for the disease independently five times and obtained the test results $\mathcal{D}$ in Table 1.

Subject	Test Results
Anton	$1, 0, 0, 1, 1$
Bella	$0, 1, 0, 0, 0$
Charlie	$0, 0, 1, 1, 1$
Dora	$1, 1, 1, 1, 1$

Table 1: The test results $\mathcal{D}$ of five independent trials for your four friends, where 0 or 1 indicate a negative or positive test outcome respectively. For example, on the second and third trial your test indicated Anton has not contracted the disease.

Given the results in Table 1, you would like to estimate the true positive rate $\mu_1 := \mathbb{P}(X = 1 \mid Z = 1)$ and false positive rate $\mu_0 := \mathbb{P}(X = 1 \mid Z = 0)$ of your test. Here, Z is a binary **unobserved** variable that is 1 if a subject suffers from the disease, and 0 otherwise. X is a binary **observed** variable that is 1 if your test is positive, and 0 otherwise. In addition, you are interested in estimating the probability that any of your friends has contracted the disease, $\pi_i = \mathbb{P}(Z_i = 1 \mid \mathcal{D})$ for $i \in \{A, B, C, D\}$.

You realize you can estimate μ_0, μ_1 and π_i for $i \in \{A, B, C, D\}$ jointly using the (soft) expectation maximization (EM) algorithm. For this purpose, you treat Z as the latent variable and μ_0 and μ_1 as the parameters of the likelihood function. You also assume $\mathbb{P}(Z_i = 1) = 0.2$ for each friend $i \in \{A, B, C, D\}$ independently, because it is believed that every fifth person suffers from the new disease.

Question 49 *E-step.* Assuming you obtained estimates $\hat{\mu}_0 = 0.2$ and $\hat{\mu}_1 = 0.8$ from the previous iteration, carry out a single expectation step of the soft EM algorithm, to compute $\hat{\pi}_A$. What is $\hat{\pi}_A$? *Hint: The answer may be rounded to three decimal places to the closest choice below.*

A 0.020	C 0.250	■ 0.500	G 0.750
B 0.200	D 0.256	F 0.512	H 0.800

Explanation: Consider the odds-ratio for Anton $\frac{0.2 \times 0.8^3 \times 0.2^2}{0.8 \times 0.2^3 \times 0.8^2} = 1$. Hence, $\hat{\pi}_A = 0.5$.

Question 50 *M-step.* Assuming you obtained estimates $\hat{\pi}_A = 0.5$, $\hat{\pi}_B = 0.2$, $\hat{\pi}_C = 0.5$, $\hat{\pi}_D = 0.8$ from the previous iteration, carry out a single maximization step of the soft EM algorithm to compute $\hat{\mu}_1$. What is $\hat{\mu}_1$?

A 0.20	C 0.40	E 0.52	■ 0.72
B 0.28	D 0.48	F 0.60	H 0.80

Explanation: $\hat{\mu}_1 = \frac{0.5 \times 3 + 0.2 \times 1 + 0.5 \times 3 + 0.8 \times 5}{(0.5 + 0.2 + 0.5 + 0.8) \times 5} = 0.72$

Question 51 Using the soft EM algorithm and initializing $\hat{\mu}_0 = 0.5$ and $\hat{\mu}_1 = 0.5$, what solution do you converge to for $\hat{\pi}_i$ for $i \in \{A, B, C, D\}$?

■ $\hat{\pi}_A = 0.20, \hat{\pi}_B = 0.20, \hat{\pi}_C = 0.20, \hat{\pi}_D = 0.20$
B $\hat{\pi}_A = 0.48, \hat{\pi}_B = 0.16, \hat{\pi}_C = 0.48, \hat{\pi}_D = 0.80$
C $\hat{\pi}_A = 0.60, \hat{\pi}_B = 0.16, \hat{\pi}_C = 0.60, \hat{\pi}_D = 0.80$
D $\hat{\pi}_A = 0.48, \hat{\pi}_B = 0.20, \hat{\pi}_C = 0.48, \hat{\pi}_D = 0.80$
E $\hat{\pi}_A = 0.48, \hat{\pi}_B = 0.16, \hat{\pi}_C = 0.48, \hat{\pi}_D = 1.00$
F $\hat{\pi}_A = 0.48, \hat{\pi}_B = 0.20, \hat{\pi}_C = 0.48, \hat{\pi}_D = 1.00$
G $\hat{\pi}_A = 0.60, \hat{\pi}_B = 0.16, \hat{\pi}_C = 0.60, \hat{\pi}_D = 1.00$
H $\hat{\pi}_A = 0.60, \hat{\pi}_B = 0.20, \hat{\pi}_C = 0.60, \hat{\pi}_D = 0.80$

Explanation: Initializing both classes with the same likelihood parameters produces a probabilistic assignment that is equal to the prior. Since this assignment is uniform over the data for the two classes, the likelihood parameters remain coupled and thus $\hat{\pi}_A = \hat{\pi}_B = \hat{\pi}_C = \hat{\pi}_D = 0.2$.

Question 52 Using the soft EM algorithm and initializing $\hat{\mu}_0 = 0.5$ and $\hat{\mu}_1 = 0.5$, what solution do you converge to for $\hat{\mu}_0$ and $\hat{\mu}_1$?

A $\hat{\mu}_0 = 0.1, \hat{\mu}_1 = 0.1$	E $\hat{\mu}_0 = 0.5, \hat{\mu}_1 = 0.5$
B $\hat{\mu}_0 = 0.2, \hat{\mu}_1 = 0.2$	■ $\hat{\mu}_0 = 0.6, \hat{\mu}_1 = 0.6$
C $\hat{\mu}_0 = 0.3, \hat{\mu}_1 = 0.3$	G $\hat{\mu}_0 = 0.7, \hat{\mu}_1 = 0.7$
D $\hat{\mu}_0 = 0.4, \hat{\mu}_1 = 0.4$	H $\hat{\mu}_0 = 0.8, \hat{\mu}_1 = 0.8$

Explanation: Continuing from the the previous part or otherwise, $\hat{\mu}_0 = \hat{\mu}_1 = 0.6$

Intro to ML - Final Exam (Answer Sheet)

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 ■ C ■
Question 2: ■ B C D
Question 3: ■ B C ■
Question 4: ■ B C D
Question 5: A ■ C D
Question 6: A B ■ D
Question 7: A ■ C D
Question 8: A B ■ D
Question 9: A B C D E F G ■
Question 10: ■ B C D E F G H
Question 11: A B ■ D E F G H
Question 12: ■ B C D E F G H
Question 13: A ■
Question 14: A ■
Question 15: ■ B
Question 16: A ■
Question 17: ■ B
Question 18: A ■ C D E F G H
Question 19: ■ B C D
Question 20: A B C D ■ F G H
Question 21: A B ■ D E F G H
Question 22: A ■ ■ ■
Question 23: A ■ C D
Question 24: A B C ■
Question 25: ■ B C D
Question 26: ■ B C ■
Question 27: A B ■ D
Question 28: ■ ■ C ■
Question 29: ■ B C ■
Question 30: ■ ■ C D
Question 31: A B C ■
Question 32: ■ B C ■
Question 33: A B ■ ■
Question 34: ■ B ■ D
Question 35: A ■ C ■
Question 36: A ■ C D
Question 37: ■ B C D
Question 38: ■ ■ ■ ■
Question 39: A B ■ D
Question 40: ■ ■ C D
Question 41: A B C D ■ F G H
Question 42: A B C D E F ■ H
Question 43: A ■ C D E F G H
Question 44: A B C D E F G ■
Question 45: A ■
Question 46: ■ B
Question 47: ■ B
Question 48: A ■
Question 49: A B C D ■ F G H
Question 50: A B C D E F ■ H
Question 51: ■ B C D E F G H
Question 52: A B C D E ■ G H