• This is a closed book, closed notes exam. Switch off your cell phone and do not communicate with anyone other than an exam proctor.
• Start writing when instructed. Stop writing when your time is up.
• Remember that your work is graded on the quality of your writing and explanation as well as the validity of the mathematics.

(1) Recall that \( I_d \) is the \( d \times d \) identity matrix. State whether the following statement is true or false, and briefly justify your answer.

If \( C \) is a \( d \times d \) symmetric positive semi-definite matrix, then \( I_d + C \) is also a symmetric positive semi-definite matrix.

Solution: Let \( x \) be any \( d \times 1 \) vector. By linearity of matrix multiplication, we have that:

\[
x^\top (I_d + C)x = x^\top I_d x + x^\top C x
\]

Recall that from class, for any \( d \times d \) matrix \( A \), \( x^\top Ax = \sum_{i=1}^{d} \sum_{j=1}^{d} A_{ij} x_i x_j \). As \((I_d)_{ij} = 1 \) when \( i = j \) and 0 otherwise, we have: \( x^\top I_d x = \sum_{i=1}^{d} \sum_{j=1}^{d} (I_d)_{ij} x_i x_j = \sum_{i=1}^{d} x_i^2 \geq 0 \). Also, since \( C \) is a positive semi-definite matrix, \( x^\top C x \geq 0 \). Therefore, \( x^\top (I_d + C)x \geq 0 \) for any \( x \), which implies that \( I_d + C \) is a positive semidefinite matrix.

Additionally, since both \( C \) and \( I_d \) are symmetric matrices, so is their sum \( I_d + C \). Thus, \( I_d + C \) is symmetric positive semidefinite.

(2) Consider the following loss function on vectors \( w \in \mathbb{R}^3 \): \( L(w) = \exp(w_1) + \exp(w_2) + \exp(w_3) \)

(a) What is \( \nabla L(w) \)?

Solution: Taking the partial derivative with respect to an arbitrary \( w_k \):

\[
\frac{\partial L(w)}{\partial w_k} = e^{w_k}
\]

Therefore, the gradient must be:

\[
\nabla L(w) = \begin{bmatrix} e^{w_1} \\ e^{w_2} \\ e^{w_3} \end{bmatrix}
\]

(b) Suppose we use gradient descent to minimize this function, and that the current estimate is \( w = (0, 0, 0) \). If the step size is \( \eta \), what is the next estimate?

Solution: Evaluating the gradient at \( w = (0, 0, 0) \):

\[
\nabla L(w) = \begin{bmatrix} e^0 \\ e^0 \\ e^0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}
\]
Plugging this into the gradient descent update rule:

\[ w_{new} = w + \eta_t \nabla L(w) \]

\[
= \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} - \eta_t \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}
= \begin{bmatrix} w_1 - \eta_t \\ w_2 - \eta_t \\ w_3 - \eta_t \end{bmatrix}
= \begin{bmatrix} -\eta_t \\ -\eta_t \\ -\eta_t \end{bmatrix}
\]

(c) Does \( L(w) \) have a minimum? Briefly justify your answer.

**Solution:** No. \( L(w) \) is differentiable, but there is no point at which the gradient becomes zero. However, there is a well-defined infimum (greatest lower bound) of zero if we take

\[ w = \begin{bmatrix} -\infty \\ -\infty \\ -\infty \end{bmatrix} \]
A logistic regression model given by parameters $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ is fit to a data set of points $x \in \mathbb{R}^d$ with binary labels $y \in \{-1, 1\}$. Write down a precise expression for the set of points $x$ that satisfy each of the equations below. You may write your answer in terms of $w$. Hint: Remember, the equation for the conditional probability of $y = 1$ under a binary logistic regression model is:

$$\Pr(y = 1|x) = \frac{1}{1 + \exp(-w \cdot x)}$$

(a) The set of points, $x$, such that $\Pr(y = 1|x) = 1/2$.

**Solution:** Rearranging the given equation and taking the log of both sides:

$$\exp(-w \cdot x) = \frac{1}{\Pr(y = 1|x)} - 1$$

$$w \cdot x = -\log\left(\frac{1}{\Pr(y = 1|x)} - 1\right)$$

we obtain the following equation of a line:

$$w \cdot x + \log\left(\frac{1}{\Pr(y = 1|x)} - 1\right) = 0$$

Plugging in $\Pr(y = 1|x) = 1/2$:

$$w \cdot x + \log(2 - 1) = 0$$

$$w \cdot x = 0$$

Therefore, the set of points $x$ satisfying this equation is the hyperplane $w \cdot x = 0$.

(b) $\Pr(y = 1|x) = 3/4$

**Solution:** Plugging in $\Pr(y = 1|x) = 3/4$:

$$w \cdot x + \log\left(\frac{4}{3} - 1\right) = 0$$

$$w \cdot x + \log(1/3) = 0$$

Therefore, the set of points $x$ satisfying this equation is the hyperplane $w \cdot x + \log(1/3) = 0$.

(c) $\Pr(y = 1|x) = 1/4$

**Solution:** Plugging in $\Pr(y = 1|x) = 1/4$:

$$w \cdot x + \log(4 - 1) = 0$$

$$w \cdot x + \log(3) = 0$$

Therefore, the set of points $x$ satisfying this equation is the hyperplane $w \cdot x + \log(3) = 0$. Notice that all of the above hyperplanes are parallel to one another!

(4) For each of the following functions of one variable, say whether it is convex.

(a) [1 pt] Is $f(x) = x^2 + x + 1$ convex? Circle your answer: **Yes** or **No**

**Solution:** Yes. Taking the first derivative:

$$f'(x) = 2x + 1$$
Taking the second derivative:

\[ f''(x) = 2 \]

which is always nonnegative.

(b) [1 pt] Is \( f(x) = \exp(x) \) convex? Circle you answer: Yes or No

**Solution:** Yes. Taking the first derivative:

\[ f'(x) = e^x \]

Taking the second derivative:

\[ f''(x) = e^x \]

which is always nonnegative.

(c) [1 pt] Is \( f(x) = \exp(x^2 + x + 1) \) convex? Circle you answer: Yes or No

**Solution:** Yes. Taking the first derivative (applying the chain rule):

\[ f'(x) = e^{x^2+x+1}(2x+1) \]

Taking the second derivative (applying the product rule):

\[ f''(x) = 2e^{x^2+x+1} + e^{x^2+x+1}(2x+1)^2 \]

which is always nonnegative.

(d) [1 pt] Is \( f(x) = \exp(x) + x^2 + x + 1 \) convex? Circle you answer: Yes or No

**Solution:** Yes. Taking the first derivative:

\[ f'(x) = e^x + 2x + 1 \]

Taking the second derivative:

\[ f''(x) = e^x + 2 \]

which is always nonnegative.

(5) Is the matrix \( M = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} \) positive semidefinite? Justify your answer.

**Solution:** Yes. \( M \) is positive semidefinite if and only if \( x^T M x \geq 0 \) for all \( x \). Expanding:

\[
x^T M x = x^T \begin{bmatrix} x_1 + x_2 + x_3 \\ x_1 + x_2 + x_3 \\ x_1 + x_2 + x_3 \end{bmatrix} \\
= x_1(x_1 + x_2 + x_3) + x_2(x_1 + x_2 + x_3) + x_3(x_1 + x_2 + x_3) \\
= (x_1 + x_2 + x_3)^2 \\
\geq 0
\]
Consider the following binary classification dataset:

\[(x_1, x_2, x_3, y) : \]
1. \((0, 0, 1, +1)\)
2. \((0, 1, 1, -1)\)
3. \((1, 0, 1, -1)\)
4. \((1, 1, 1, +1)\)

(a) [4pts] Suppose we initialize the weight vector for a linear classifier to \(w = (0, 0, 0)\) and then proceed to train using the perceptron algorithm on this dataset (we will use no additional bias term). What are the weights after one pass through the data in the order listed above?

**Solution:** Let \(w_i\) represent \(w\) after the \(i\)th iteration of learning. We are given that \(w_0 = (0, 0, 0)\).

Calculating the \(w_i\) in turn (let \(x_i\) below represent the feature vector for the \(i\)th training point):

\[
\begin{align*}
  w_1 &= w_0 + x_1 & (y_1 w_0^T x_1 = 0) \\
  &= (0, 0, 1) \\
  w_2 &= w_1 - x_2 & (y_2 w_1^T x_2 = -1) \\
  &= (0, -1, 0) \\
  w_3 &= w_2 - x_3 & (y_3 w_2^T x_3 = 0) \\
  &= (-1, -1, -1) \\
  w_4 &= w_3 + x_4 & (y_4 w_3^T x_4 = -3) \\
  &= (0, 0, 0)
\end{align*}
\]

Notice that we ended up doing a whole lot of nothing! Because we ended up where we started, this same cycle will repeat on the next iteration (and the next one, \textit{ad infinitum}).

(b) [2pts] Is this dataset linearly separable? [Hint: Plot the dataset.]

**Solution:** No. Let’s take “linearly separable” to mean separable given the constraints of our model (i.e. no bias). Because the last feature in each training point is 1, this problem is equivalent to fitting a perceptron with bias term to the set of points:

\[
\begin{align*}
  x_1 &= (0, 0), \quad y_1 = 1, \\
  x_2 &= (0, 1), \quad y_1 = -1, \\
  x_3 &= (1, 0), \quad y_1 = -1, \quad \text{and} \quad x_4 = (1, 1), \quad y_1 = 1.
\end{align*}
\]

If you plot these points on a two-dimensional plane with their labels, you will find that there is no line (with bias) which separates them.

You can equivalently plot the original points in a three-dimensional plane with their labels, and find that there is no plane through the origin that separates them. Even if you take “linearly separable” to mean with any plane (with bias), you can make the same argument.

(c) [2pts] Will the perceptron algorithm eventually converge on this dataset?

**Solution:** No. As we saw experimentally in part (a), the algorithm enters an infinite loop. To back this up theoretically, we saw in part (b) that there is no separating plane through the origin between the points of each class. Therefore, the perceptron algorithm will not converge.