• 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) Consider the following binary classification dataset consisting of four training examples, each with two features, $x_1$ and $x_2$:

$$(x_1, x_2, y):$$
1. $(-1, -1, +1)$
2. $(-1, -1, +1)$
3. $(+1, +1, -1)$
4. $(+1, +1, -1)$

(a) (2 pts) Can this dataset be perfectly classified using a linear classifier with no bias term? Yes or No

Solution: Yes or No. Without bias term, we are fitting a line through the origin to separate the positive points from the negative points. One such separating line is $x_2 = -x_1$.

(b) (2 pts) Will the perceptron algorithm with no bias term converge on this dataset? Yes or No

Solution: Yes or No. As shown above, there is a separating hyperplane through the origin between the points, so the perceptron algorithm is guaranteed to converge.

(c) (2 pts) Suppose you train a binary logistic regression classifier on this dataset with no bias term, using no regularization term, and a step size small enough that gradient descent always decreases the loss after each update. Let $w_t$ be the model parameters after $t$ gradient updates.

Is $P_{w_{40}}(y = +1|x = (-1, -1)) < P_{w_{80}}(y = +1|x = (-1, -1))$? Yes or No or Can’t say

Solution: Yes or No or Can’t say. The negative log-likelihood is:

$$L(w) = -2 \log(P(+1|(-1, -1))) - 2 \log(P(-1|(+1, +1)))$$
$$= -2 \log(P(+1|(-1, -1))) - 2 \log(P(+1|(-1, -1)))$$
$$= -4 \log(P(+1|(-1, -1)))$$

where the second equality follows because

$$P(-1|(+1, +1)) = \frac{1}{1 + e^{-w \cdot (1, 1)}} = \frac{1}{1 + e^{w \cdot (-1, -1)}} = P(+1|(-1, -1))$$
Alternatively, in the two-class multiclass case:
\[
P(-1|(+1,+1)) = \frac{e^{w_2(1,1)}}{e^{w_1(1,1)} + e^{w_2(1,1)}}
\]
\[
= \frac{1}{e^{(w_1-w_2)(1,1)} + 1}
\]
\[
= \frac{e^{-w_1(1,1)}}{e^{-w_1(1,1)} + e^{-w_2(1,1)}}
\]
\[
= P(+1|(-1,-1))
\]

Therefore, as the negative log likelihood of the data decreases (implied by problem assumption), we must be increasing \(P(+1|(-1,-1))\).

(2) For each of the following functions, say whether it is convex.

(a) (1 pt) For \(x \in \mathbb{R}\), is \(f(x) = -x^2\) convex? Circle you answer: \textbf{Yes} or \textbf{No}

\textbf{Solution: Yes} or \textbf{No}. Taking the first derivative:
\[
f'(x) = -2x
\]
Taking the second derivative:
\[
f''(x) = -2
\]
which is always nonpositive, so the function is strictly concave, not convex.

(b) (1 pt) For \(x \in \mathbb{R}\), is \(f(x) = 0\) convex? Circle you answer: \textbf{Yes} or \textbf{No}

\textbf{Solution: Yes} or \textbf{No}. Taking the first derivative:
\[
f'(x) = 0
\]
Taking the second derivative:
\[
f''(x) = 0
\]
which is always nonnegative. In fact, because it is also always nonpositive, the function is also concave.

(c) (2 pts) For \(x \in \mathbb{R}^d\), is \(f(x) = x^\top x\) convex? Circle you answer: \textbf{Yes} or \textbf{No}

\textbf{Solution: Yes} or \textbf{No}. Note that \(f(x) = \sum_{i=1}^{d} x_i^2\). Taking the first derivative:
\[
\nabla f(x) = 2x
\]
Taking the second derivative:
\[
\nabla^2 f(x) = 2I_{d \times d}
\]
which is positive semidefinite and also positive definite, so the function is strictly convex.

(d) (2 pts) For \(x \in \mathbb{R}^d\), is \(f(x) = -x^\top x\) convex? Circle you answer: \textbf{Yes} or \textbf{No}

\textbf{Solution: Yes} or \textbf{No}. A negated strictly convex function is strictly concave.
(3) Consider the following loss function on vectors \( w \in \mathbb{R}^d \): \( L(w) = w^\top w + 5 \)

(a) (2 pts) What is \( \nabla L(w) \)?

**Solution:** Note that \( L(w) = \sum_{i=1}^{d} w_i^2 + 5 \). Taking the first derivative:
\[
\nabla L(w) = 2w
\]

(b) (2 pts) Suppose we use gradient descent to *minimize* this function, and that the current estimate is \( w = (0, 0, \ldots, 0)^\top \). If the step size is \( \eta \), what is the next estimate?

**Solution:** \((0, 0, \cdots 0)^\top\). The gradient at this point is:
\[
\nabla L(w) = 2w = (0, 0, \cdots, 0)^\top
\]
so we are at a local minimum, and \( w \) will not change. Therefore the next estimate is \( w_{next} = w - \eta \nabla L(w) = (0, 0, \cdots, 0)^\top \).

(c) (2 pts) What is the argmin of \( L(w) \)? Briefly justify your answer.

**Solution:** \((0, 0, \cdots 0)^\top\). We can solve for the hessian of \( L(w) \):
\[
\nabla L(w) = 2I_{d\times d}
\]
which is positive semidefinite, so the function is convex. This requires that any local minimum is also a global minimum. As we found a local minimum in the previous part at \( w = (0, 0, \cdots, 0)^\top \), this must also be the global minimum of the function.