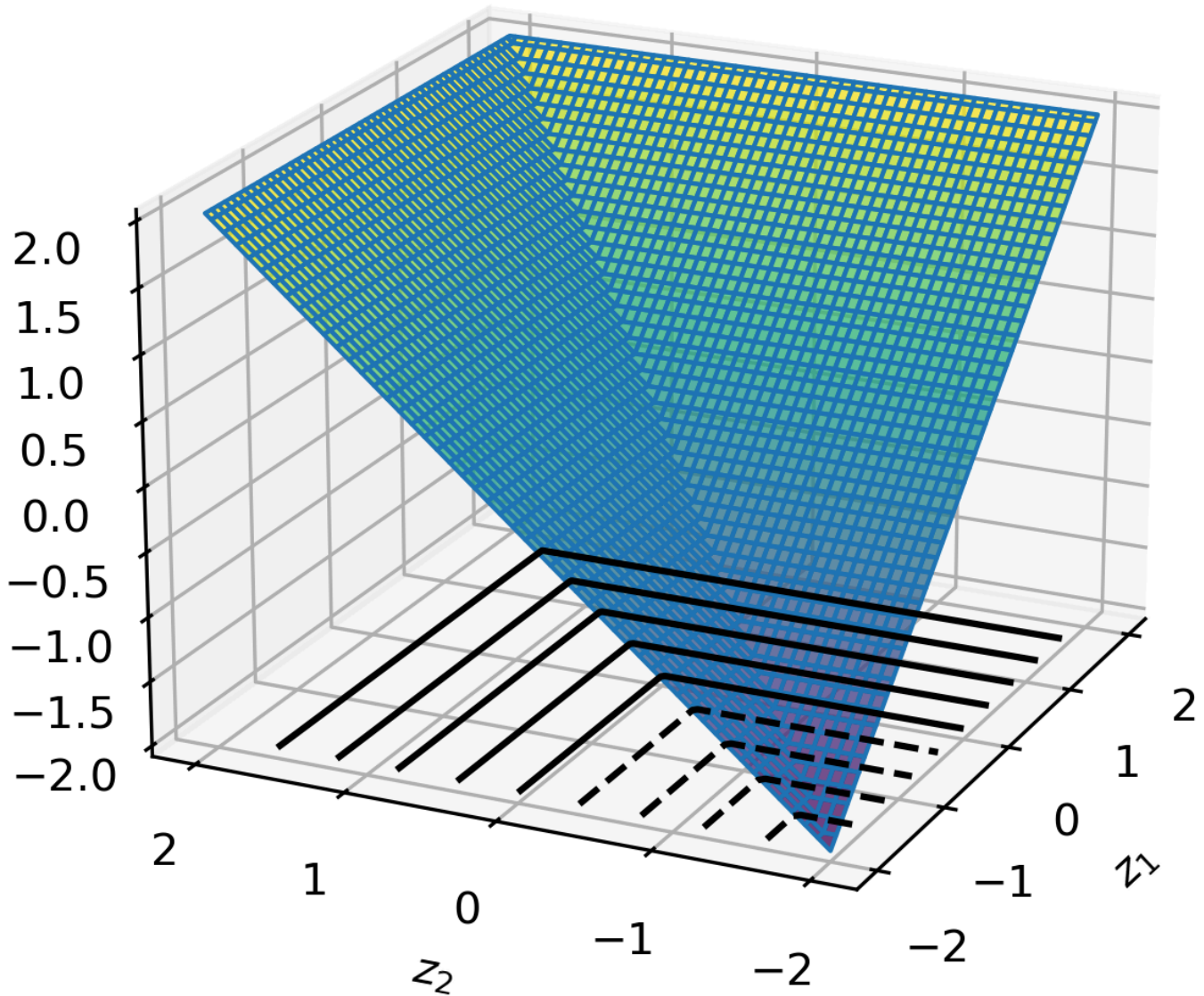


Consider the function $f(z_1, z_2) = \max(z_1, z_2)$. For convenience, this function is plotted below.



Which of the below are valid subgradients of the function at the point $(1, 1)$? Check all that apply.

- $(0, 1)^T$
- $(1, 0)^T$
- $(-0.5, 1)^T$
- $(0.5, 1)^T$
- $(0.5, 0.5)^T$
- $(1, 1)^T$

We are testing

$$f_s(\vec{z}) = f(\vec{z}^{(0)}) + \vec{s} \cdot (\vec{z} - \vec{z}^{(0)}) \leq f(\vec{z})$$

for all \vec{z} .

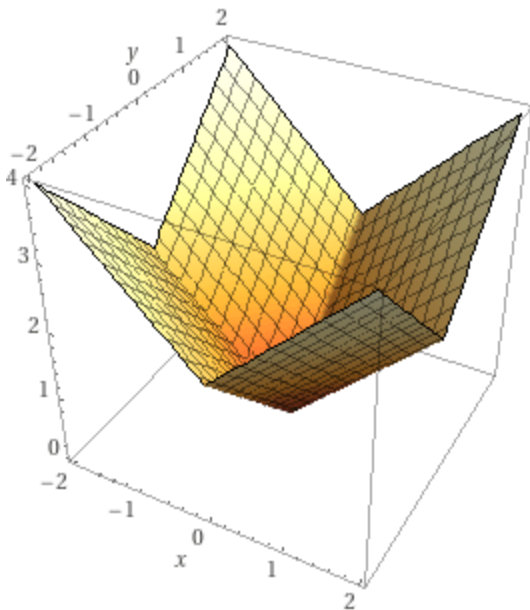
For $\vec{s} = (0, 1)^T$, $f_s(\vec{z}) = z_2 \leq \max(z_1, z_2)$ for all \vec{z} , so it is a valid subgradient. By similar logic, $\vec{s} = (1, 0)^T$ is also a valid subgradient.

Conversely, $\vec{s} = (-0.5, 1)^T$ is not a valid subgradient because $f_s(\vec{z}) = 1 + -0.5(z_1 - 1) + 1(z_2 - 1) = -0.5z_1 + z_2 + 0.5$ is not necessarily $\leq \max(z_1, z_2)$ for all \vec{z} , e.g. when $\vec{z} = (0, 0)^T$.

$\vec{s} = (0.5, 1)^T$ and $\vec{s} = (1, 1)^T$ are also not valid subgradients, e.g. when $\vec{z} = (5, 4)^T$.

$\vec{s} = (0.5, 0.5)^T$ is a valid subgradient by Problem 2c in Discussion 3.

Consider the function $f(x, y) = |x| + |y|$, shown below for convenience.



Which of the below are subgradients of f at the point $(0, 1)$? Check all that apply.

- $(-1, -1)^T$
- $(-1, 0)^T$
- $(-1, 1)^T$
- $(0, -1)^T$
- $(0, 0)^T$
- $(0, 1)^T$
- $(1, -1)^T$
- $(1, 0)^T$
- $(1, 1)^T$

Observe that at the point $(0, 1)$, the gradient of f is undefined for x , but is equal to 1 for y .

Moreover, the slopes to the left and right of $x = 0$ are -1 and 1 respectively. Thus, the subgradient at $(0, 1)$ is any vector from $(-1, 1)^T$ to $(1, 1)^T$.

Which of the below are subgradients of f at the point $(-2, 0.5)$? Check all that apply.

- $(-1, -1)^T$
- $(-1, 0)^T$
- $(-1, 1)^T$
- $(0, -1)^T$
- $(0, 0)^T$
- $(0, 1)^T$
- $(1, -1)^T$
- $(1, 0)^T$
- $(1, 1)^T$

Observe that at the point $(-2, 0.5)$, the gradient of f is in fact well-defined; in particular, since $x = -2$ is less than 0 and $y = 0.5$ is greater than 0, the subgradient is precisely the gradient, $(-1, 1)^T$.

Which of the below are subgradients of f at the point $(-2, 0)$? Check all that apply.

- $(-1, -1)^T$
- $(-1, 0)^T$
- $(-1, 1)^T$
- $(0, -1)^T$
- $(0, 0)^T$

$$[] (0, 1)^T$$

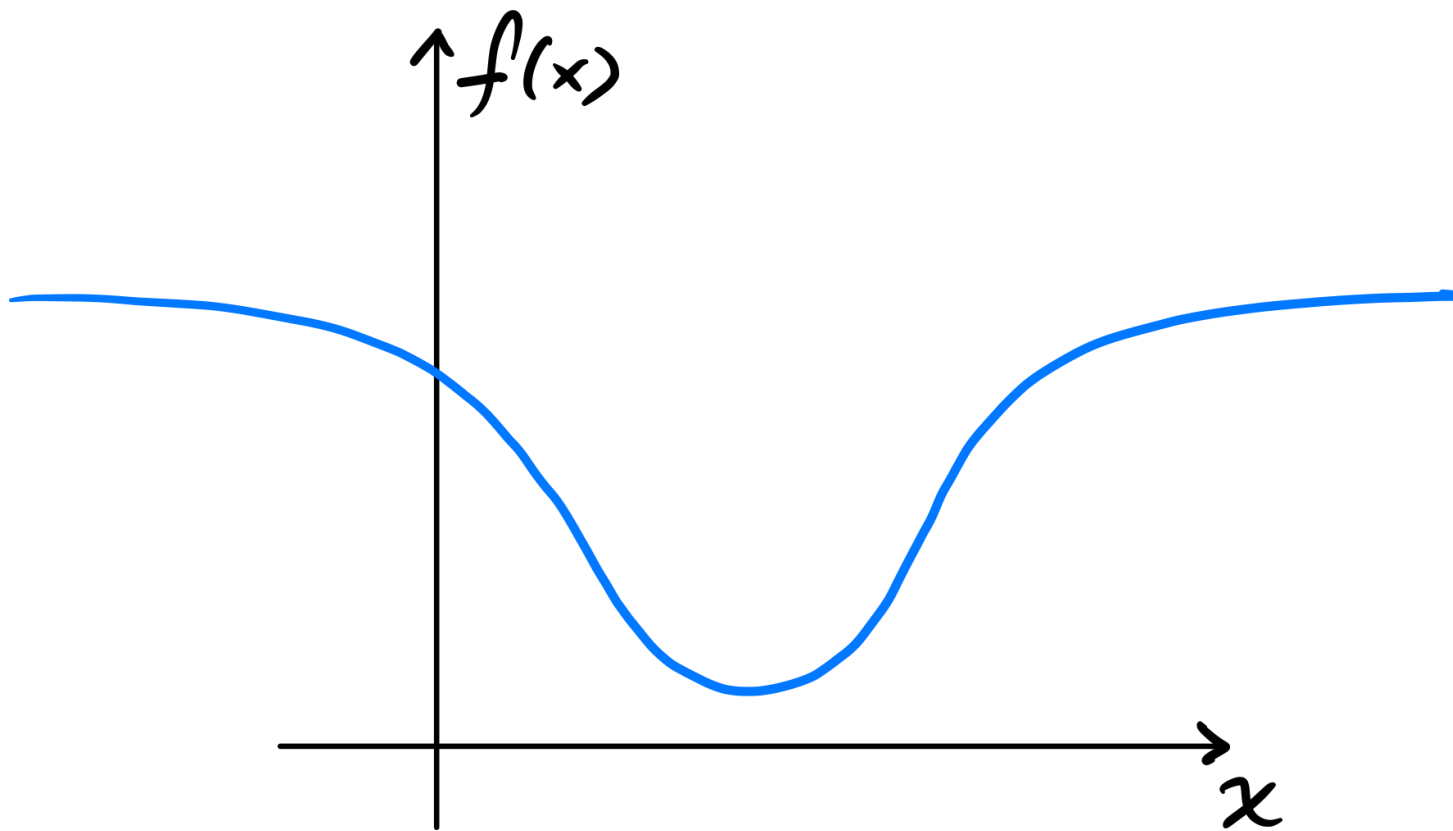
$$[] (1, -1)^T$$

$$[] (1, 0)^T$$

$$[] (1, 1)^T$$

By similar logic to Q1.1, the gradient is not defined for $y = 0$, but is equal to -1 for $x = -2$. Thus, the subgradient at $(-2, 0)$ is any vector from $(-1, -1)^T$ to $(-1, 1)^T$.

Is the function shown below convex or non-convex?



Convex

Non-convex

Recall that one geometric interpretation/property of convexity is that for every a, b the line segment between $(a, f(a))$ and $(b, f(b))$ does not go below the graph of f . This does not hold for the above function, since one can draw a line segment to the left or right of the bowl that goes below the graph.

Define $f(x) = \sum_{i=1}^k \alpha_i e^{\beta_i(x-\mu)}$, where k is a positive integer, μ is a real number, and α_i (for $i \in \{1, 2, \dots, k\}$) is a positive real number, and β_i (for $i \in \{1, 2, \dots, k\}$) is a (possibly-negative) real

number.

Is f convex (no matter how α_i, β_i, μ are chosen) or is it possibly non-convex?

Convex

Non-convex

Recall a property of convexity: sums of convex functions are convex. This leaves us to show that each summand is convex, as α_i is a *positive* real number.

One way to prove convexity for this 1-dimensional function is to show that the second derivative is ≥ 0 . Suppose we define $f_i(x)$ s.t. $f(x) = \sum_{i=1}^k f_i(x)$. Then,

$$\begin{aligned}f_i(x) &= \alpha_i e^{\beta_i(x-\mu)} \\ \frac{d}{dx} f_i(x) &= \alpha_i \beta_i e^{\beta_i(x-\mu)} \\ \frac{d^2}{dx^2} f_i(x) &= \alpha_i \beta_i^2 e^{\beta_i(x-\mu)}\end{aligned}$$

Since $\alpha_i > 0$, $\beta_i^2 \geq 0$, and $e^x > 0$ for any real number x , the second derivative is indeed non-negative. Thus, $f_i(x)$ is convex, and their sum, $f(x)$ is also convex.

Recall the 0-1 loss from lecture. Define $R_{01}(\vec{w})$ to be the risk of a linear predictor $H(\vec{x}) = \vec{w} \cdot \text{Aug}(\vec{x})$ on a data set. Is R_{01} convex or non-convex as a function of \vec{w} ?

Convex

Non-convex

Recall that $R_{0,1}$ is flat almost everywhere, i.e. its graph is a piecewise constant function. So, one can draw a line segment between two points on different flat segments and obtain a line segment below the graph. Thus, $R_{0,1}$ is not convex. Although we have a linear predictor, we have a non-convex loss function, so the risk is also non-convex.

Suppose $f(x_1, x_2)$ is a convex function. Define the new function $g(x) = f(x, 0)$. True or False: $g(x)$ must be convex.

- True
 False

Since f is convex, it must be convex for all points in its domain, in particular for points of the form $(x, 0)$ where $x_1 = x$ and $x_2 = 0$. So, $g(x)$ must be convex.

Let $f_1(x)$ be a convex function and let $f_2(x)$ be a non-convex function. Define the new function $f(x) = f_1(x) + f_2(x)$. True or False: $f(x)$ must be non-convex.

- True
 False

Let $f_1(x) = 4x^2$ and $f_2(x) = -x^2$ s.t. we satisfy the assumptions (f_1 is convex and f_2 is non-convex). Then, $f(x) = f_1(x) + f_2(x) = 3x^2$ is convex.

Let H be a linear classifier and suppose $\vec{x}^{(1)}$, $\vec{x}^{(2)}$, and $\vec{x}^{(3)}$ are three points. Suppose

- $H(\vec{x}^{(1)}) = 1$
- $H(\vec{x}^{(2)}) = -3$
- $H(\vec{x}^{(3)}) = 2$

Which point is furthest away from the decision boundary?

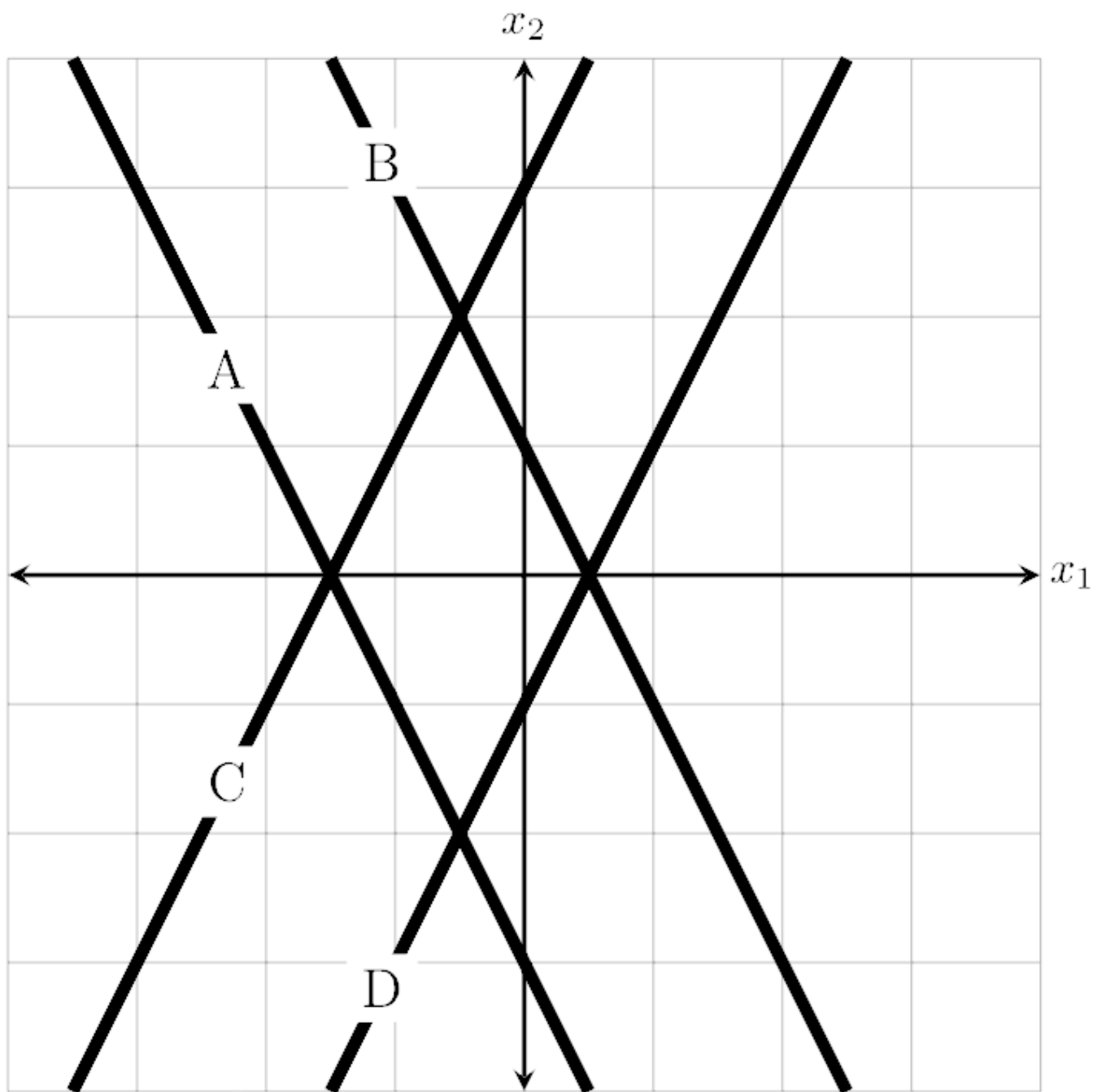
- $\vec{x}^{(1)}$
 $\vec{x}^{(2)}$
 $\vec{x}^{(3)}$

Since the decision boundary is at $H(\vec{x}) = 0$, and the magnitude of $H(\vec{x})$ is proportional to the distance from the decision boundary, $\vec{x}^{(2)}$ is the furthest as it has the highest magnitude.

Suppose a linear prediction function $H(\vec{x}) = w_0 + w_1x_1 + w_2x_2$ is trained to perform classification, and the weight vector is found to be $\vec{w} = (1, -2, 1)^T$.

The figure below shows four possible decision boundaries: A , B , C , and D Which of them is the decision boundary of the prediction function H ?

You may assume that each grid square is 1 unit on each side, and the axes are drawn to show you where the origin is.



- A
- B
- C
- D

There are many ways to approach this problem; here is one of them:

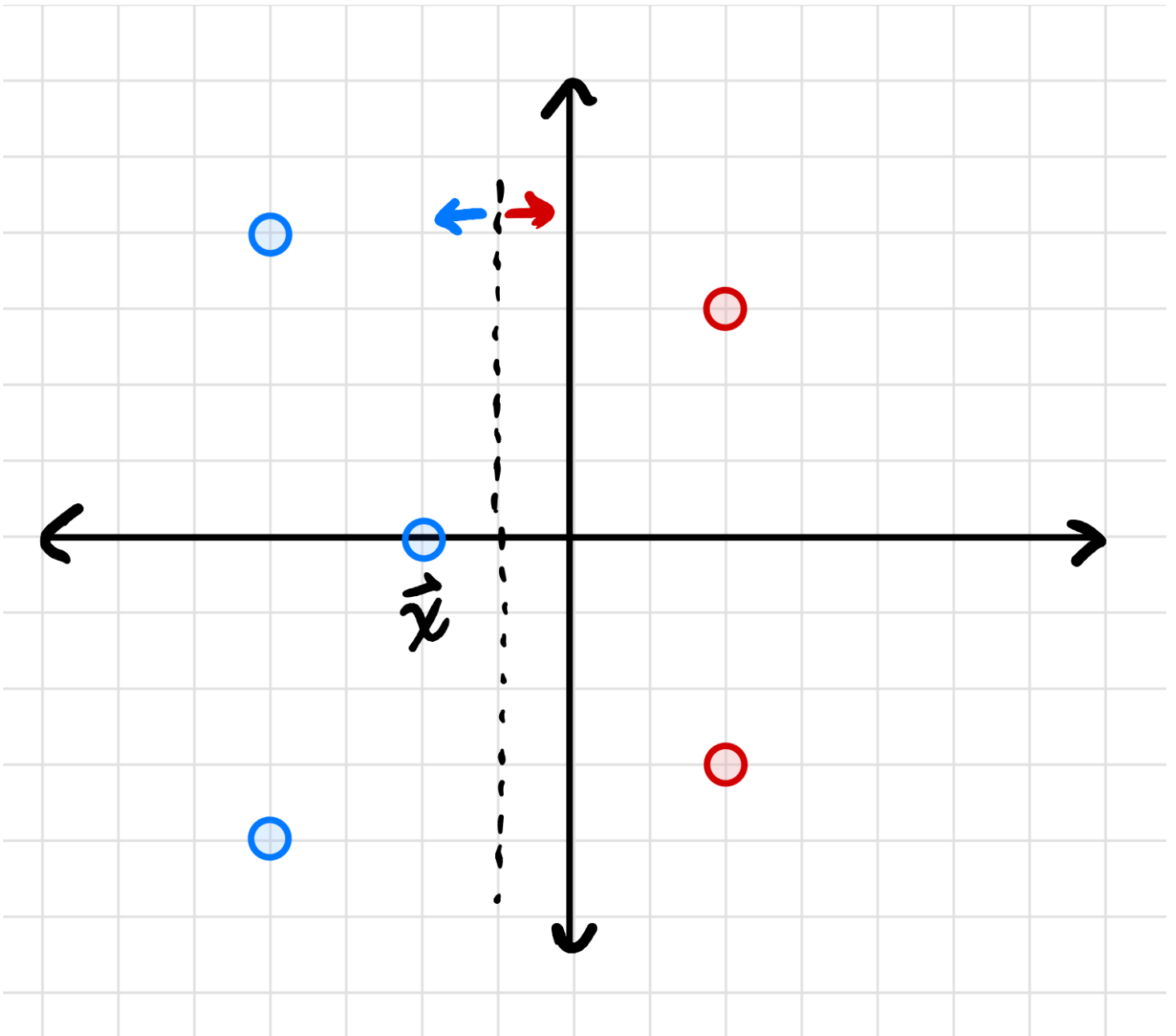
Recall from lecture that $\vec{w}' = (w_1, \dots, w_d)$ is orthogonal to the decision boundary. For this problem, $\vec{w}' = (-2, 1)$, so this rules out options A and B because we're looking for a decision boundary orthogonal to a vector pointing west-northwest.

To choose between C and D , we can leverage the fact that the prediction at the decision boundary must be 0, i.e., $H(\vec{x}) = 0$. Let us choose $\vec{x} = (0, -1)$ since it is on D .

$$\text{Aug}(\vec{x}) \cdot \vec{w} = (1, 0, -1)^T \cdot (1, -2, 1)^T = 0.$$

Thus, the decision boundary must be D .

Consider the image below:



The blue points have label +1, and the red points have label -1. Suppose H is a linear prediction function; the dashed line in the image shows H 's decision boundary.

Suppose that when H is applied to the point labeled \vec{x} in the above image, $H(\vec{x}) = 1$.

What is the mean square loss of this prediction function, H ?

The key fact in solving problems like this is that, since H is a *linear* prediction function, $H(\vec{z})$ is *proportional* to the distance between \vec{z} and the decision boundary.

In this case, $H(\vec{x}) = 1$ for the \vec{x} shown in the picture. Notice that \vec{x} is 1 unit away from the decision boundary. Therefore, $H(\vec{x}) = 3$ for any point on the same side of the boundary that is three units

away. On the other side of the boundary, the output of H is negative. Both of the red points are 3 units from the boundary, and so $H(\vec{x}) = -3$ on these points.

To compute the mean square loss, we compute $H(\vec{x}^{(i)} - y_i)$ for each data point, and average. For the blue points, $y_i = 1$, and for the red points $y_i = -1$. So the total square loss is $(3 - 1)^2 + (3 - 1)^2 + (1 - 1)^2 + (-3 - -1)^2 + (-3 - -1)^2 = 4 + 4 + 0 + 4 + 4 = 16$. Thus the mean square loss is $16/5 = 3.2$.

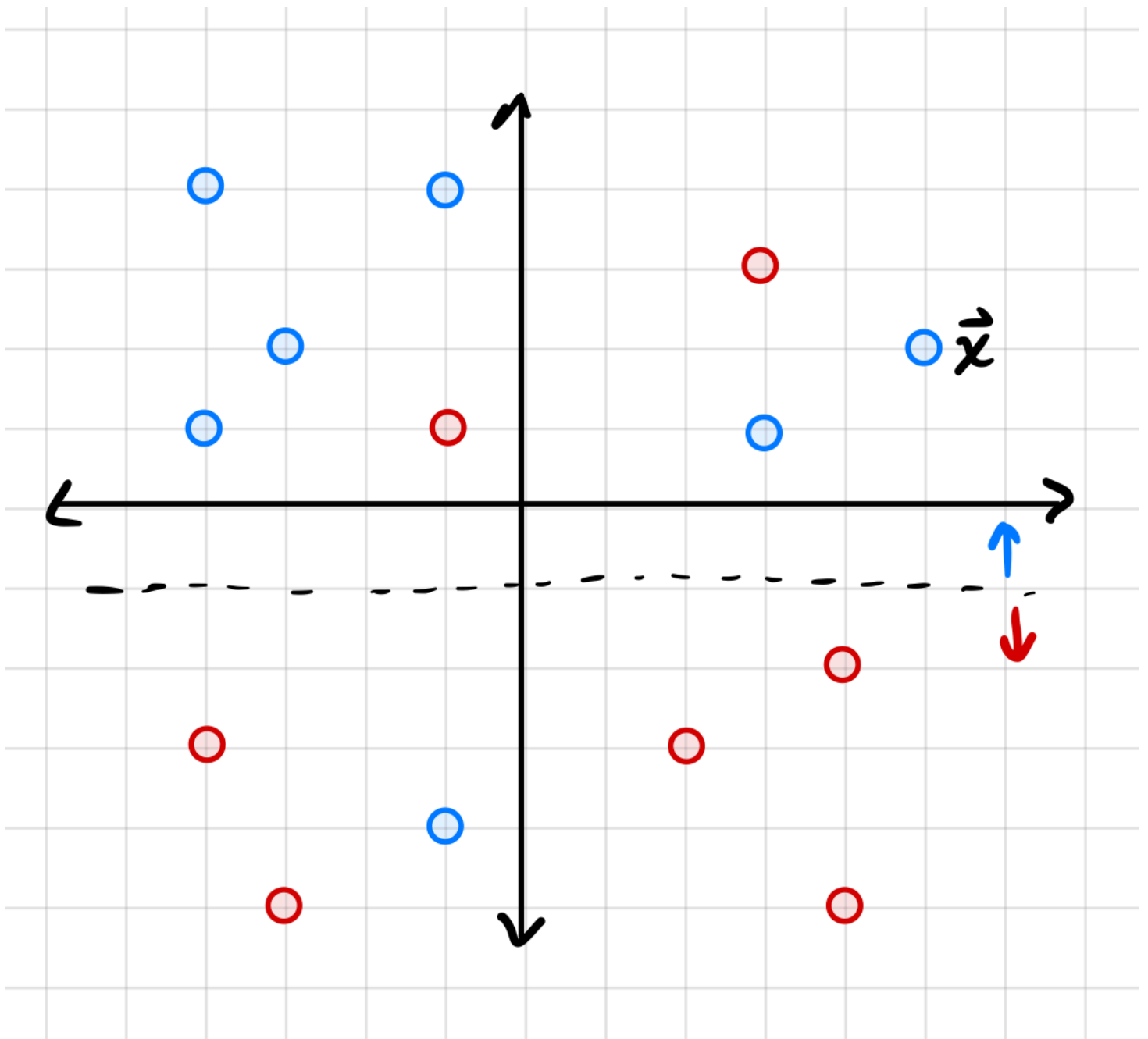
Notice that the mean square loss is not zero, even though every point is correctly classified!

Consider again the dataset above. True or false: there exists a linear classifier $H(\vec{x}; \vec{w}) = w_0 + w_1x_1 + w_2x_2$ which has a mean square loss of zero on this data.

- True
- False

In order for H to be a linear classifier with a mean square loss of 0, the data would need to be arranged in a way such that each data point was correctly classified and only 1 unit away from the decision boundary (so that the loss at each point would be equal to 0). No matter how we tilt/reposition the decision boundary, this will never be true; thus, there does not exist such a linear classifier.

The picture below shows the decision boundary of a linear classifier, H , as a dashed line at $y = -1$. The classifier predicts that points above the line are blue, and points below are red. Three points out of the 14 shown have been misclassified.



Assume that when H is applied to the point labeled \vec{x} , $H(\vec{x}) = 6$.

What is the expected perceptron loss of H ? That is, what is the risk with respect to the perceptron loss?

—

As in Problem 5, the key fact is that H is linear, and so $H(\vec{z})$ is proportional to the distance between \vec{z} and the decision boundary. In this case, since $H(\vec{x}) = 6$ for the point \vec{x} marked above, and \vec{x} is 3 units away from the boundary, this tells us that $H(\vec{z}) = 2$ if \vec{z} is 1 unit away (and on the same side of the boundary).

To compute the expected loss, we could compute the loss on each data point. However, since we are working with the perceptron loss, we know that the loss is zero on any data point that has been correctly classified. Therefore, we only need to worry about red points that are on the blue side of the boundary, or blue points that are on the red side. There are three such points.

One of the misclassified red points is 2 units away from the boundary. At this point, the output of the prediction function H will be 4. The perceptron loss assigns a loss of $|H(\vec{z})| = |4| = 4$ to this point. The other misclassified red point is 4 units away from the boundary, and H is 8 at this point. Therefore, the loss for this point is 8.

The misclassified blue point is 3 units from the boundary, but on the red side, and so the output of H at this point is -6. The loss for this prediction is $|-6| = 6$.

The *total* perceptron loss is therefore $4 + 8 + 6 = 18$. The *mean* perceptron loss is then $18/14 \approx 1.28$ (be careful to divide by the total number of points, not just how many were misclassified!).