Let $\vec{x}^{(1)}, \ldots, \vec{x}^{(n)}$ be vectors in $\mathbb{R}^d$ and let $\vec{w}$ be a vector in $\mathbb{R}^d$. Consider the function:

$f(\vec{w}) = \frac{1}{n} \sum_{i=1}^{n} ||\vec{w} - \vec{x}^{(i)}||^2$

Which of the below is the gradient of f with respect to $\vec{w}$?

( ) $\frac{2}{n} \sum_{i=1}^{n} \vec{x}^{(i)}$
( ) $\frac{2}{n} \sum_{i=1}^{n} \vec{w}$
(x) $\frac{2}{n} \sum_{i=1}^{n} (\vec{w} - \vec{x}^{(i)})$
( ) $\vec{x}^{(i)}$

One approach: Similarly to how we derived $\frac{d}{d\vec{w}} ||\vec{w}||^2 = 2\vec{w}$ we can apply this to our function and invoke chain rule to obtain, $\frac{d}{d\vec{w}} f(\vec{w}) = \frac{d}{d\vec{w}} \frac{1}{n} \sum_{i=1}^{n} ||\vec{w} - \vec{x}^{(i)}||^2 = \frac{2}{n} \sum_{i=1}^{n} (\vec{w} - \vec{x}^{(i)})$.

Another approach: Similarly to how we derived the normal equations, we can expand the squared 2-norm as follows, $f(\vec{w}) = \frac{1}{n} \sum_{i=1}^{n} ||\vec{w} - \vec{x}^{(i)}||^2 = \frac{1}{n} \sum_{i=1}^{n} (\vec{w} - \vec{x}^{(i)})^T (\vec{w} - \vec{x}^{(i)}) = \frac{1}{n} \sum_{i=1}^{n} (||\vec{w}||^2 - 2{\vec{x}^{(i)}}^T \vec{w} + ||\vec{x}^{(i)}||^2)$.
Then, we take the gradient and arrive at the same conclusion, $\frac{d}{d\vec{w}} f(\vec{w}) = \frac{1}{n} \sum_{i=1}^{n} (2\vec{w} - 2\vec{x}^{(i)}) = \frac{2}{n} \sum_{i=1}^{n} (\vec{w} - \vec{x}^{(i)})$.

Which of the below is a minimizer of $f(\vec{w})$?

(x) $\vec{w}^* = \frac{1}{n} \sum_{i=1}^{n} \vec{x}^{(i)}$
( ) $\vec{w}^* = \frac{1}{n} \sum_{i=1}^{n} \vec{w}$
( ) $\vec{w}^* = \frac{1}{n} \vec{x}^{(i)}$
( ) $\vec{w}^* = \vec{x}^{(i)}$

To find the minimizer of $f(\vec{x})$, we can take its derivative, set it equal to 0, and solve. We have already found the gradient of f with respect to $\vec{w}$ in the previous problem, so we solve as follows,

$$0 = \frac{2}{n} \sum_{i=1}^{n} (\vec{w}^* - \vec{x}^{(i)}) \Leftrightarrow \vec{w}^* = \frac{1}{n} \sum_{i=1}^{n} \vec{x}^{(i)}.$$

This checks out: if you recall from DSC 40A, we know that the mean minimizes mean squared error.

Consider the function $f(z_1, z_2) = z_1^4(z_1^2 + z_2^2)$. Suppose that a single iteration of gradient descent is run on this function with a starting location of $(1, 1)^T$ and a learning rate of $\eta = 1/10$. What will be the z_2 coordinate after one iteration?

Recall our gradient descent formula, $\vec{z}^{(t+1)} = \vec{z}^{(t)} - \eta \nabla f(\vec{z}^{(t)})$.

We can find the gradient to be $\nabla f(\vec{z}) = (6z_1^5 + 4z_1^3 z_2^2, 2z_1^4 z_2)^T$ and evaluate it at our initial point, $\vec{z}^{(0)} = (1, 1)^T$

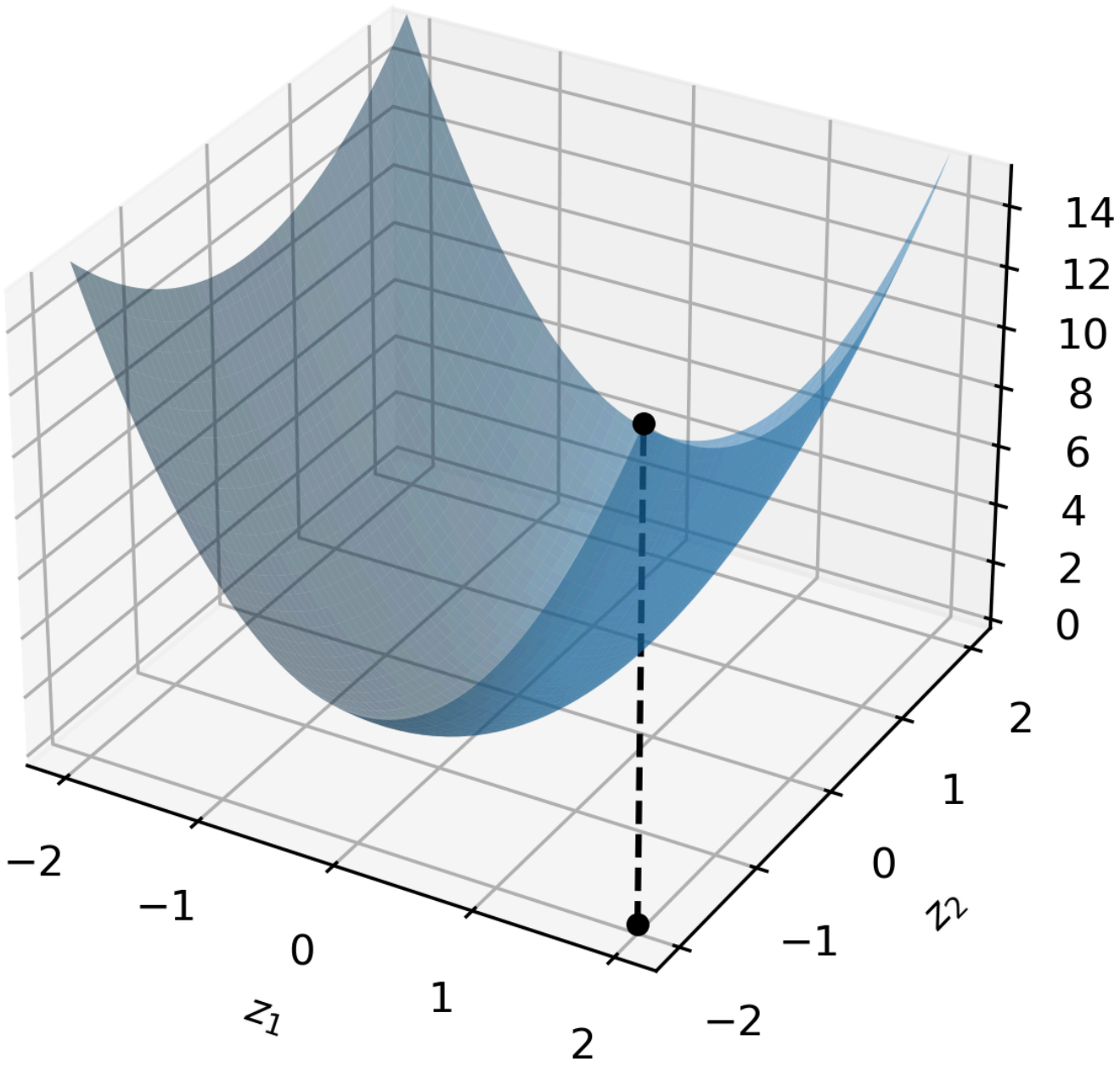
$$\vec{z}^{(1)} = \vec{z}^{(0)} - \eta \nabla f(\vec{z}^{(0)}) = \begin{bmatrix} 1 \\ 1 \end{bmatrix} - \frac{1}{10} \begin{bmatrix} 10 \\ 2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0.8 \end{bmatrix}.$$

Using the same learning rate, what will be the z_2 coordinate after the *second* iteration?

We repeat the same process to find $\vec{z}^{(2)}$,

$$\vec{z}^{(2)} = \vec{z}^{(1)} - \eta \nabla f(\vec{z}^{(1)}) = \begin{bmatrix} 0 \\ 0.8 \end{bmatrix} - \frac{1}{10} \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0.8 \end{bmatrix}.$$

Consider the function $f(z_1, z_2)$ whose plot is shown below:

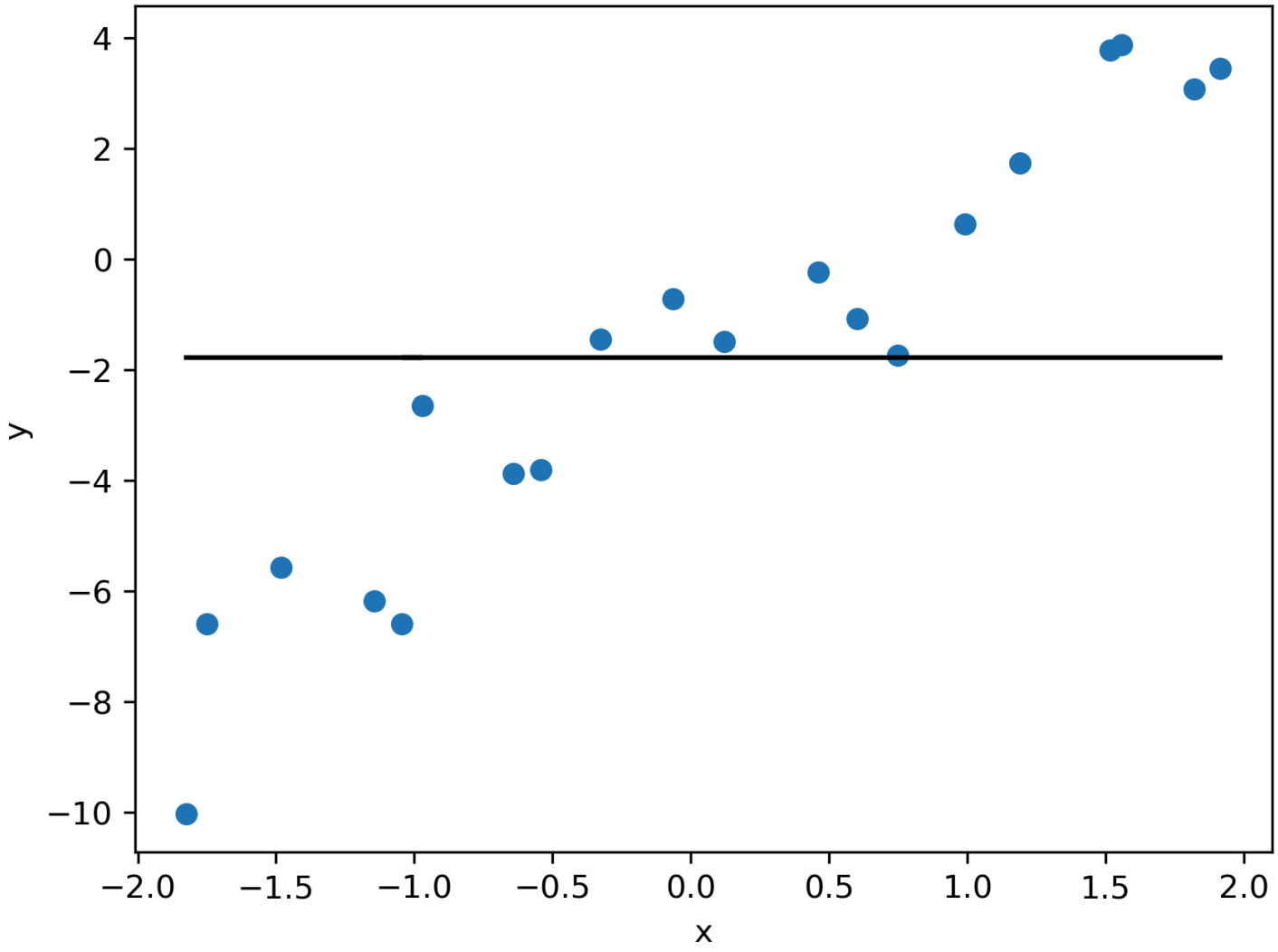


Marked on the plot is the point (2, -2). Which of the below show the signs of each entry of the gradient vector at this point?

( ) (+, +)
( ) (-, -)
(x) (+, -)
( ) (-, +)

We can see that at the point $(2, -2)$, the graph is increasing in the z_1 direction, so its slope is positive with respect to z_1. On the other hand, the graph is decreasing at $(-2, 2)$ in the z_2 direction, so its slope is negative with respect to z_2.

Consider the regression data set $(x_1, y_1), \ldots, (x_n, y_n)$ shown below. Also shown as a black line is the linear prediction function $H(x) = w_0 + w_1 x$ with $w_1 = 0$ and $w_0 = \frac{1}{n}\sum_{i=1}^{n} y_i$. That is, w_0 is set to be equal to the mean of the y_i values.



Let $R(\vec{w})$ be the empirical risk with respect to this data and the square loss.

Which of the following is true about $\partial R / \partial w_0(w_0, w_1)$?

( ) It is positive
( ) It is negative
(x) It is zero

The following question is similar to Problem 3 in Discussion 2. For the first component, we want to ask ourselves, in order to increase the risk, should we increase of decrease w_0? We can see from the graph that the line of best will have an intercept around its current value, -2. Since we are already at the optimal solution, we would not want to change w_0.

One can also prove this mathematically,

$$\frac{\partial R}{\partial w_0}(w_0, w_1) = \frac{\partial}{\partial w_0}\frac{1}{n}\sum_{i=1}^{n}(w_0 + w_1 x^{(i)} - y^{(i)})^2 = 2w_0 - \frac{2}{n}\sum_{i=1}^{n} y^{(i)} = 0.$$

Which of the following is true about $\partial R / \partial w_1(w_0, w_1)$?

( ) It is positive
(x) It is negative
( ) It is zero

For the second component, we want to ask ourselves, in order to increase the risk, should we increase or decrease w_1? We can see from the graph that the line of best will have a positive slope. Currently, $w_1 = 0$, so if we want to increase the risk, we want to make w_1 negative (and more negative).

Let $(\vec{x}^{(1)}, y_1), \ldots, (\vec{x}^{(n)}, y_n)$ be a data set of n points in $\mathbb{R}^d \times \mathbb{R}$. Let $\vec{w}$ be a vector in $\mathbb{R}^d$. Let $R(\vec{w})$ be the mean squared error of a linear prediction function $H(\vec{x}) = \vec{w} \cdot \text{Aug}(\vec{x})$ on this data.

True or False: if $R(\vec{w})$ is exactly zero, then it must be the case that $H(\vec{x})$ makes no errors on the data set (every prediction is exactly equal to the corresponding y_i).

(x) True
( ) False

True or False: if the gradient $\frac{dR}{d\vec{w}}(\vec{w}) = \vec{0}$, then it must be the case that $H(\vec{x})$ makes no errors on the data set.

( ) True
(x) False

Consider the regression data set shown below consisting of points $(x_1, y_1), \ldots, (x_n, y_n)$.

$\vec{x}$	y
(1, 1)	1
(2, 0)	5
(-1, 1)	-3

$\vec{x}$	y
(0, 0)	0

Consider fitting a linear model $H(\vec{x}) = w_0 + w_1 x_1 + w_2 x_2$ to this data set using gradient descent to minimize the empirical risk with respect to the square loss (also known as the mean squared error).

What type of object is the gradient of the empirical risk evaluated at $\vec{w}^{(0)} = (2, 1, -2)$?

( ) A scalar
( ) A vector in $\mathbb{R}^2$
(x) A vector in $\mathbb{R}^3$
( ) A vector in $\mathbb{R}^4$

The empirical risk is a function $R_{sq} : \mathbb{R}^3 \to \mathbb{R}$ that takes in a vector $\vec{w}$ of length 3. Thus, the gradient is a vector in $\mathbb{R}^3$.

What is the value of the first entry of the gradient of the empirical risk evaluated at $\vec{w}^{(0)} = (2, 1, -2)^T$? That is, what is the value of the entry corresponding to $\partial R / \partial w_0$? Give your answer as a decimal number.

___

Recall from lecture that the gradient of the empirical risk is

$$\frac{dR}{d\vec{w}}(\vec{w}) = \frac{2}{n} \sum_{i=1}^{n} (\text{Aug}(\vec{x}^{(i)}) \cdot \vec{w} - y_i)(\text{Aug}(\vec{x}^{(i)})).$$

After substituting the given data and taking the partial derivative with respect to w_0, we obtain

$$\frac{\partial R}{\partial w_0}(\vec{w}) = \frac{1}{2}(4w_0 + 2w_1 + 2w_2 - 3)$$

, which evaluated at $\vec{w}^{(0)} = (2, 1, -2)^T$ is 1.5.

Alternatively, one can compute the entire gradient evaluated at $\vec{w}^{(0)} = (2, 1, -2)^T$, which is $(1.5, -2, 1)^T$.

After one step of gradient descent with a learning rate of $\eta = 1$, which of the below is the current weight vector, $\vec{w}^{(1)}$?

( ) $(3, 1, -1)^T$
(x) $(1/2, 3, -3)^T$
( ) $(3/2, 2, -2)^T$
( ) $(2, 1, -2)^T$
( ) $(1, 0, -2)^T$

The iterative step in gradient descent for this scenario is

$$\vec{w}^{(1)} = \vec{w}^{(0)} - \eta \frac{dR}{d\vec{w}}(\vec{w}^{(0)}).$$

From the previous problem, we found that the gradient evaluated at $\vec{w}^{(0)}$, so $\vec{w}^{(1)} = (2, 1, -2)^T - (1.5, -2, 1)^T = (0.5, 3, -3)^T$.

Consider again the regression data set shown below consisting of points $(x_1, y_1), \ldots, (x_n, y_n)$.

$\vec{x}$	y
(1, 1)	1
(2, 0)	5
(-1, 1)	-3
(0, 0)	0

Consider fitting a linear model $H(\vec{x}) = w_0 + w_1 x_1 + w_2 x_2$ to this data set, this time using *stochastic* gradient descent to minimize the empirical risk with respect to the square loss (also known as the mean squared error).

Suppose that you use a batch size of 2, and that on the first step of SGD the first and last points are used in the batch. If you use an initial weight vector of $\vec{w}^{(0)} = (2, 1, -2)^T$, and a learning rate of $\eta = 1$, what will be the weight vector $w^{(1)}$ after the first step of SGD?

( ) $(3, 1, -1)^T$
( ) $(2, 1, -2)^T$
( ) $(1, 0, -2)^T$

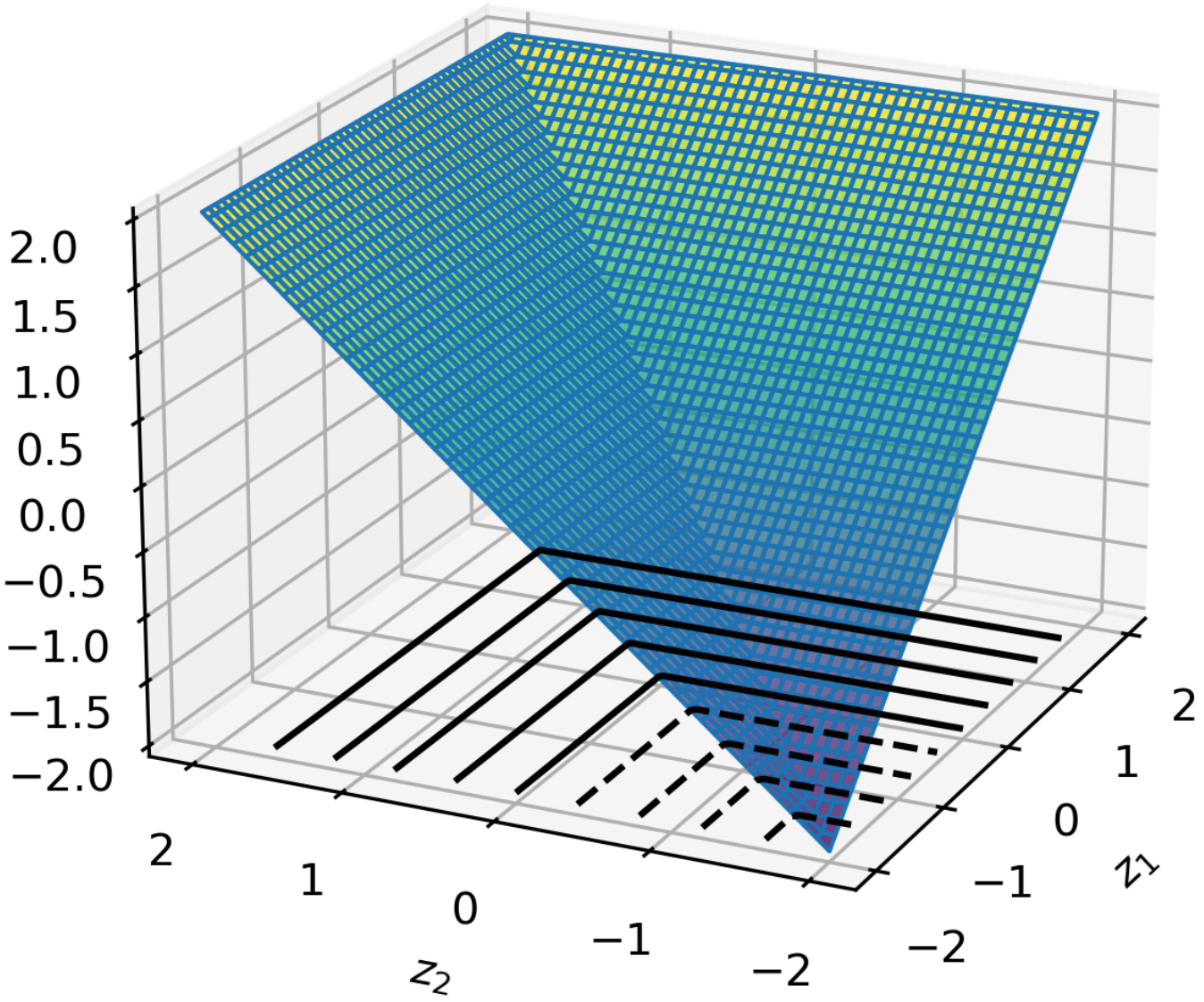
( ) $(0, -1, -2)^T$
(x) $(0, 1, -2)^T$

Since we are performing SGD, the update step will look something like this,

$$\vec{w}^{(1)} = \vec{w}^{(0)} - \eta \cdot \frac{2}{2}\left[(\text{Aug}(\vec{x}^{(1)}) \cdot \vec{w}^{(0)} - y_1)(\text{Aug}(\vec{x}^{(1)})) + (\text{Aug}(\vec{x}^{(4)}) \cdot \vec{w}^{(0)} - y_4)(\text{Aug}(\vec{x}^{(4)}))\right]$$

which evaluates to $\vec{w}^{(1)} = (2, 1, -2)^T - (2, 0, 0)^T = (0, 1, -2)^T$.

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



What is the gradient of the function at the point $(1, -1)$?

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

We can see from the graph that at the point $(1, -1)$, the graph is increasing in the z_1 direction, so this rules out all choices except for $(1, -1)^T$ and $(1, 0)^T$. Then, we see that the graph is constant in the z_2

direction, so this leaves $(1, 0)^T$ as the gradient.

We can still do this problem without the graph. To find the slope in the z_1 direction, we can fix $z_2 = -1$ and see what happens when we change z_1, say, from 1 to 2. $f(1, -1) = 1$ and $f(2, -1) = 2$, so we can infer that the slope in the z_1 direction is 1.

By similar logic, we can fix $z_1 = 1$ and change z_2 from -1 to -2. $f(1, -1) = f(1, -2) = 1$, so we can infer that the slope in the z_2 direction is 0.