Deep Learning Unveiled: Theory, Mathematics, and Programming.

Session 2: Multi-variable calculus for deep learning and the Gradient Descent Algorithm.

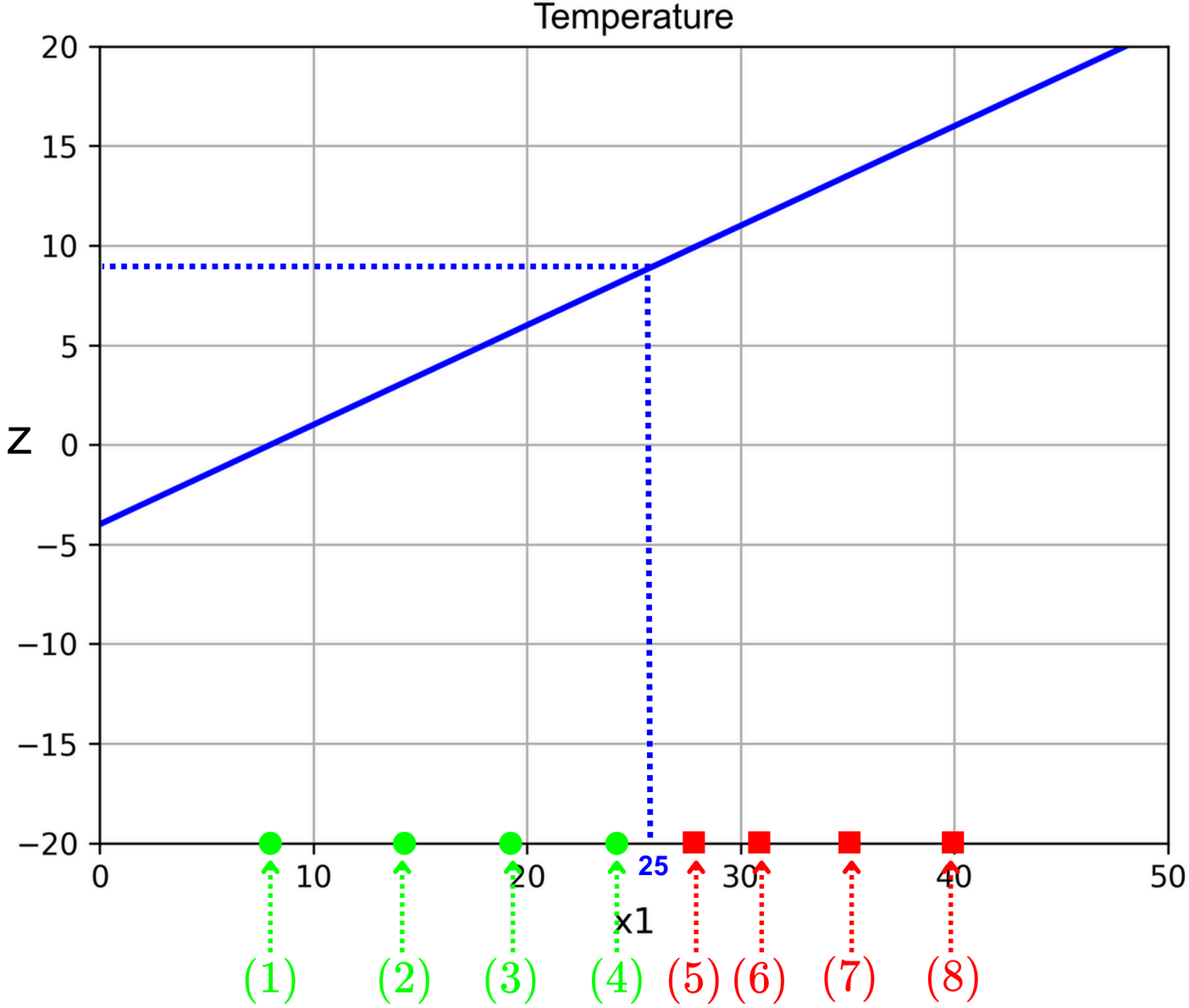
Likelihood

The likelihood of our model is given by:

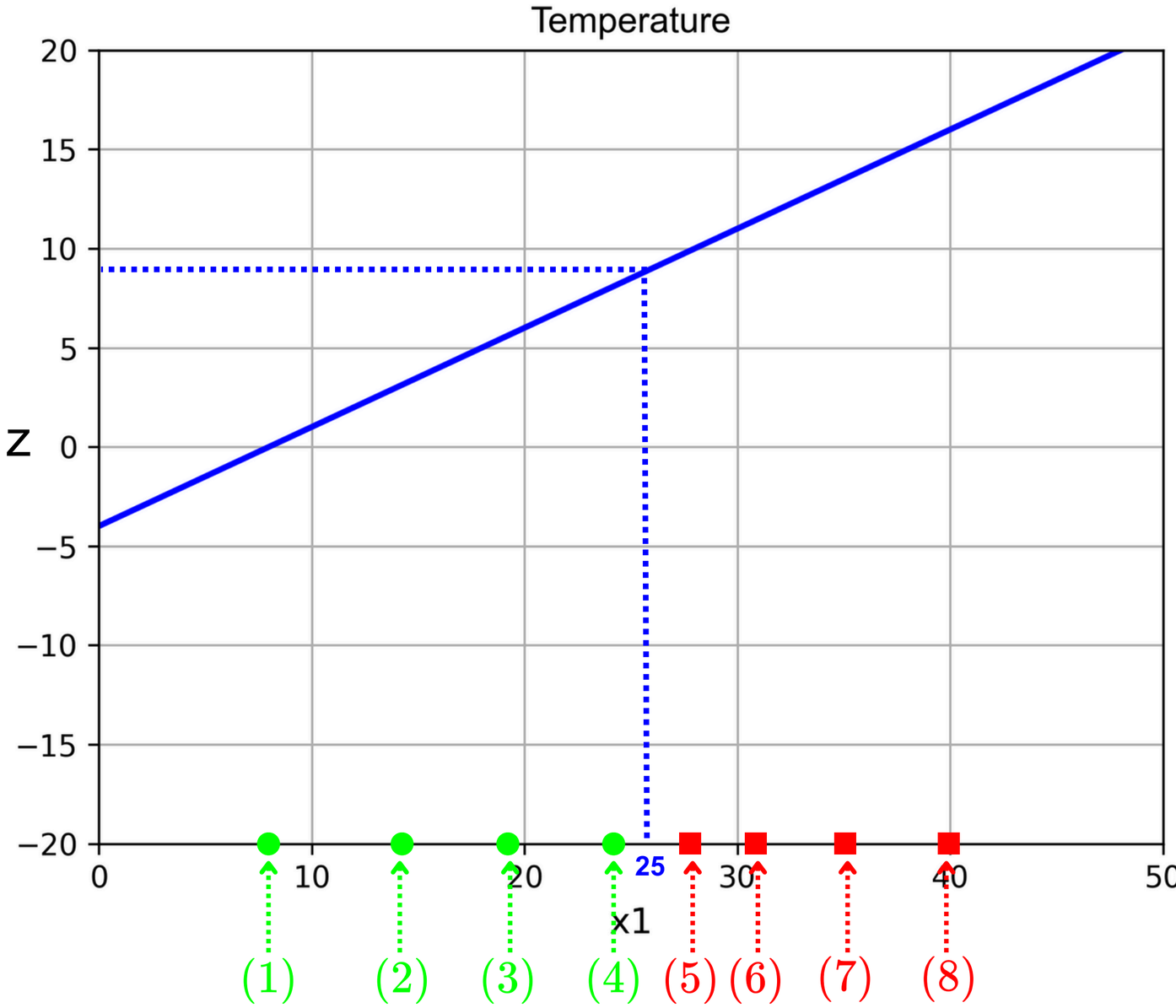
$$L = \prod_{i=1}^{m} P(Y = y_i)$$

$$L = \prod_{i=1}^{m} a_i^{y_i}(1 - a_i)^{1-y_i}$$

Let's give some Id (i) for our examples.



ID	x_1	y
1	8	0
2	15	0
3	19	0
4	24	0
5	27	1
6	31	1
7	35	1
8	40	1



$$L = \prod_{i=1}^{m} a_i^{y_i} (1 - a_i)^{1-y_i}$$

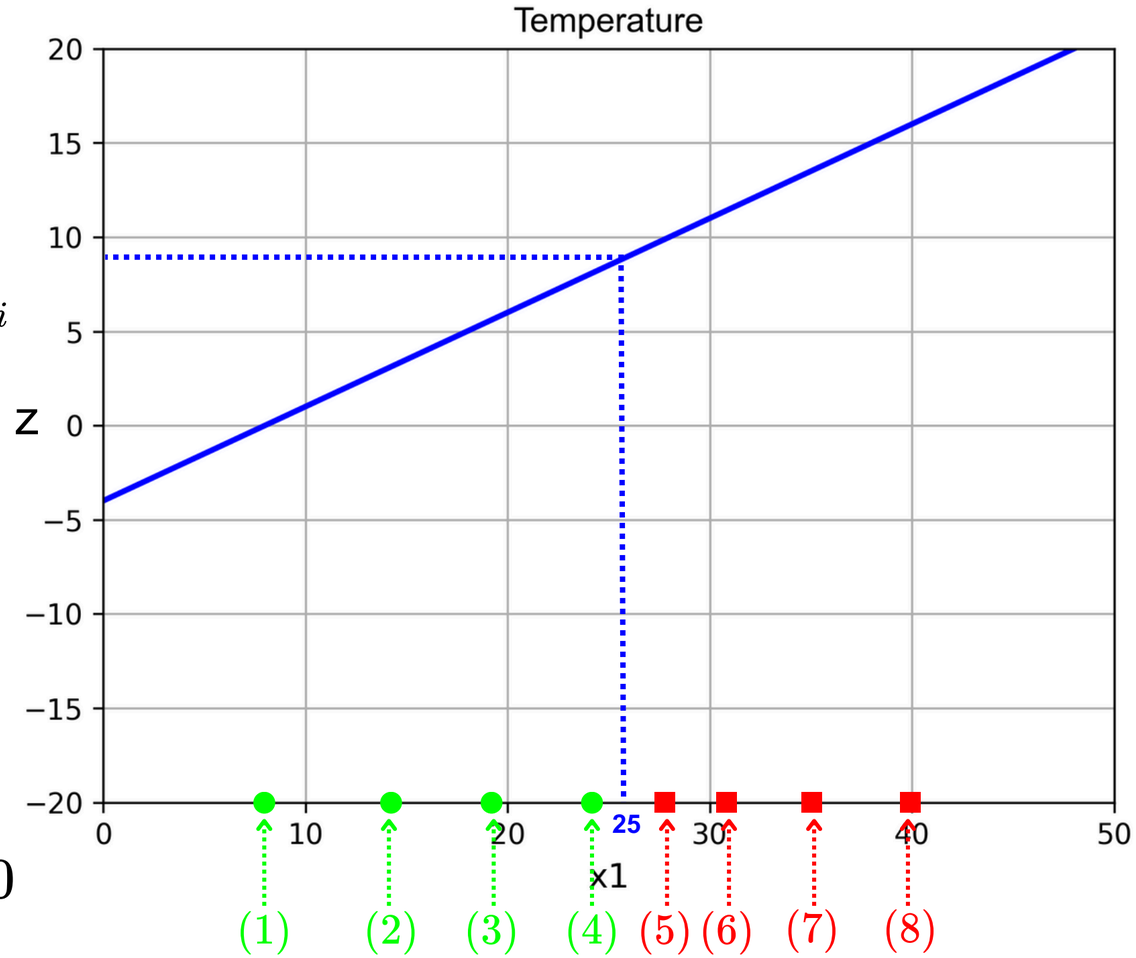
$$P(Y = y_i) = a(z_i)^{y_i} \times (1 - a(z_i))^{1-y_i}$$

example (1):

$$\begin{cases} x_1 = 8 \\ y = 0 \end{cases}$$

$$z_1 = z\left(x_1^{(1)}\right) = (0.5)x_1^{(1)} - 4 = (0.5) \times 8 - 4 = 0$$

$$z(x_1) = (0.5)x_1 - 4$$



$$P(Y=0) = a(z_1)^0 \times (1 - a(z_1))^{1-0}$$

$$z_1 = z\left(x_1^{(1)}\right) = (0.5)x_1^{(1)} - 4 = (0.5) \times 8 - 4 = 0$$

$$a(z_1) = \frac{1}{1+e^{-z_1}} = \frac{1}{1+e^0} = \frac{1}{2}$$

$$P(Y=0) = a(z_1) = 1 - \frac{1}{2} = \frac{1}{2}$$

example (1):

$$\begin{cases} x_1 = 8 \\ y = 0 \end{cases}$$

$$\boxed{z(x_1) = (0.5)x_1 - 4}$$

i	$x_1^{(i)}$	y_i	z_i	a_i	$P(Y = y_i)$
1	8	0	0	0.50000	0.50000
2	15	0	3.5	0.97068	0.02931
3	19	0	5.5	0.99592	0.00407
4	24	0	8	0.99966	0.00033
5	27	1	9.5	0.99992	0.99992
6	31	1	11.5	0.99998	0.99998
7	35	1	13.5	0.99999	0.99999
8	40	1	16	0.99999	0.99999

Likelihood

Calculating the initial likelihood of our model:

$$L = \prod_{i=1}^{m} P(Y = y_i) = P(Y = y_1) \times P(Y = y_2) \times \ldots \times P(Y = y_8)$$

$$= 0.5 \times 0.02931 \times 0.00407 \times \ldots \times 0.99999$$

$$= 2.0002713231 \times 10^{-8}$$

On the other hand, the likelihood for the other model $z(x_1) = x_1 - 25$

i	$x_1^{(i)}$	y_i	z_i	a_i	$P(Y = y_i)$
1	8	0	-17	0.00000	0.99999
2	15	0	-10	0.00004	0.99995
3	19	0	-6	0.00247	0.99752
4	24	0	-1	0.26894	0.73105
5	27	1	2	0.99992	0.88079
6	31	1	6	0.99998	0.99752
7	35	1	10	0.99999	0.99995
8	40	1	15	0.99999	0.99999

Likelihood

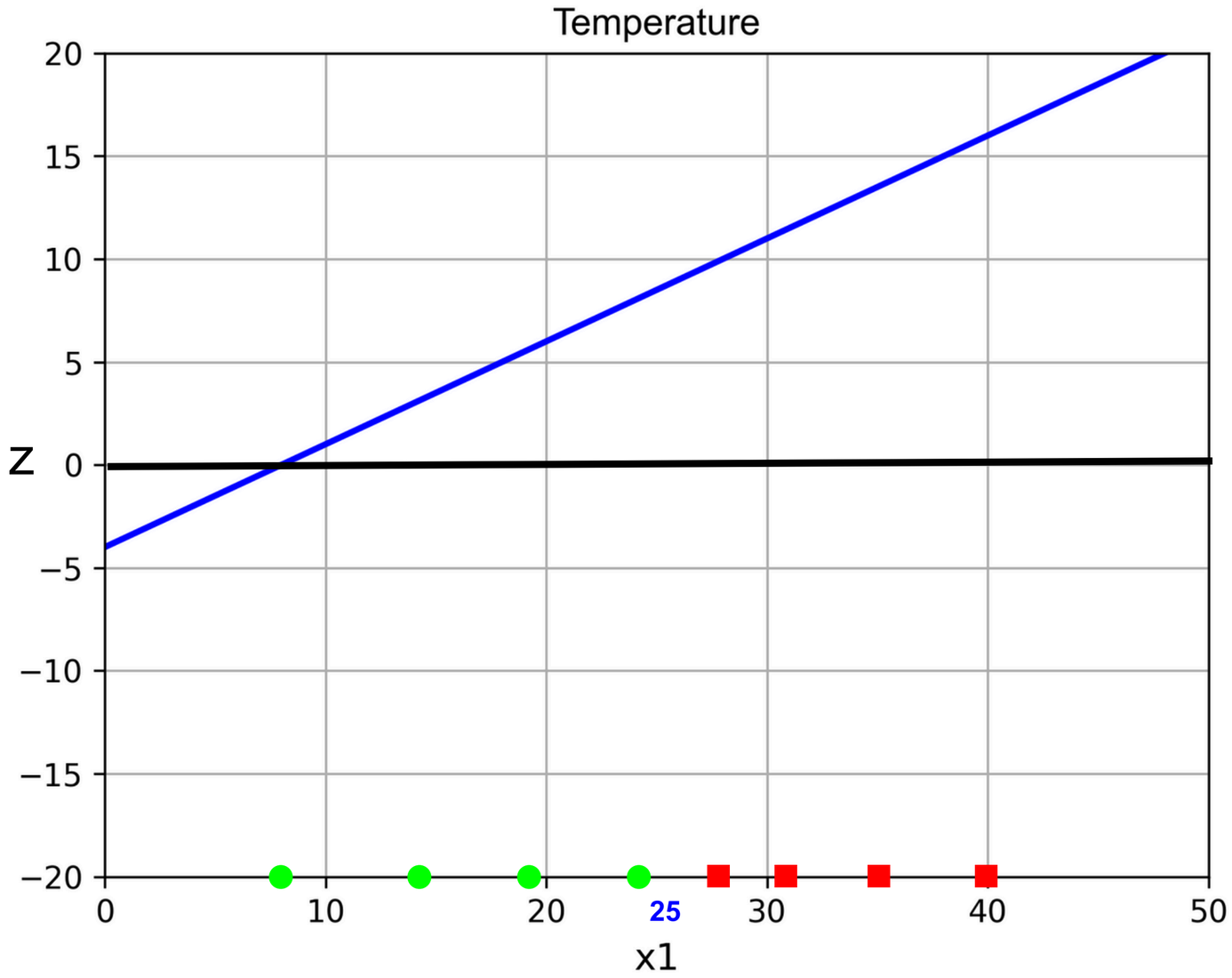
$$L = 0.640675$$

$$z(x_1) = x_1 - 25$$

We notice that the likelihood for this model is significantly higher than our model's.

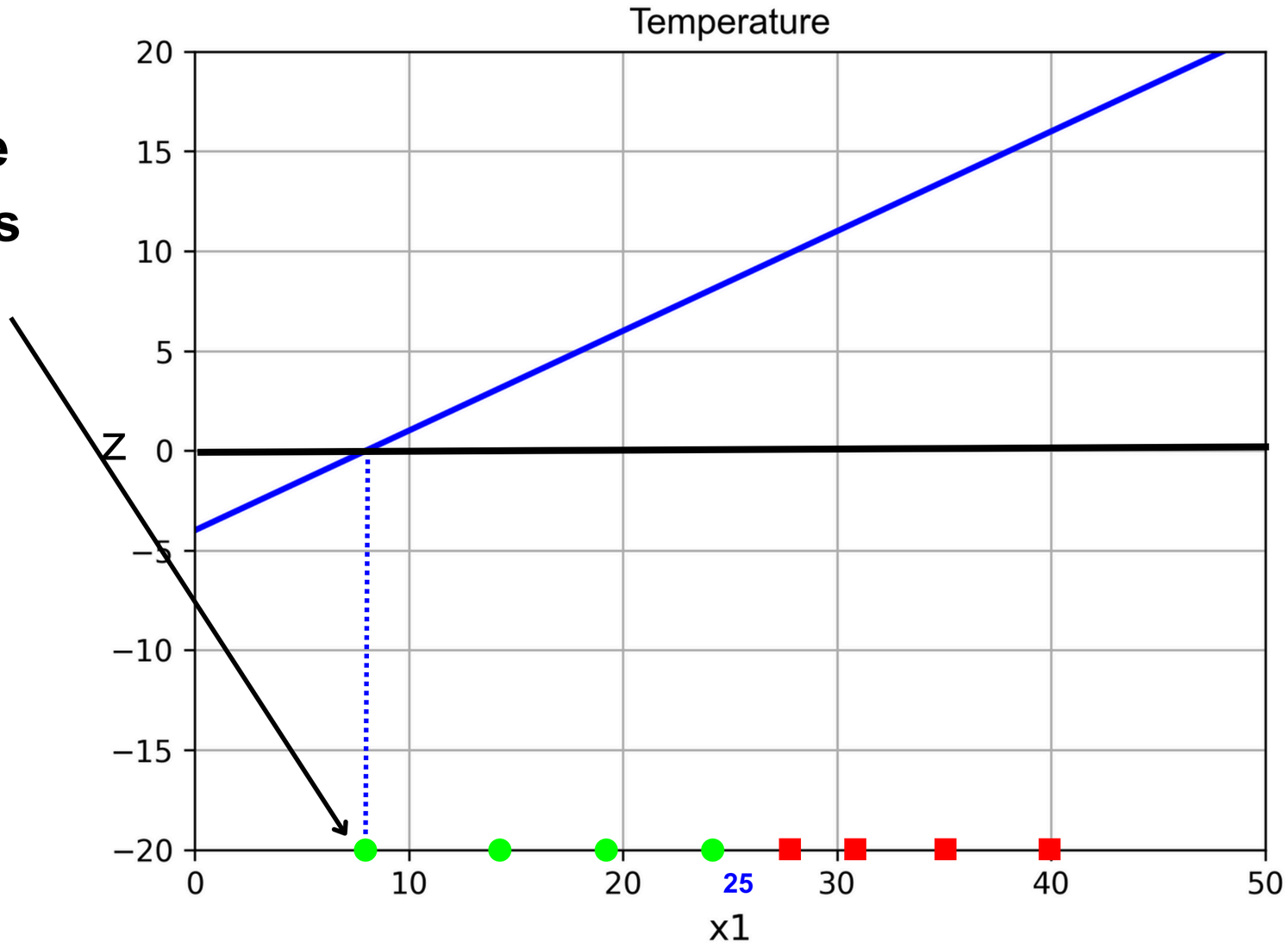
The reason behind this is:

$$z(x_1) = (0.5)x_1 - 4$$



$$z(x_1) = (0.5)x_1 - 4$$

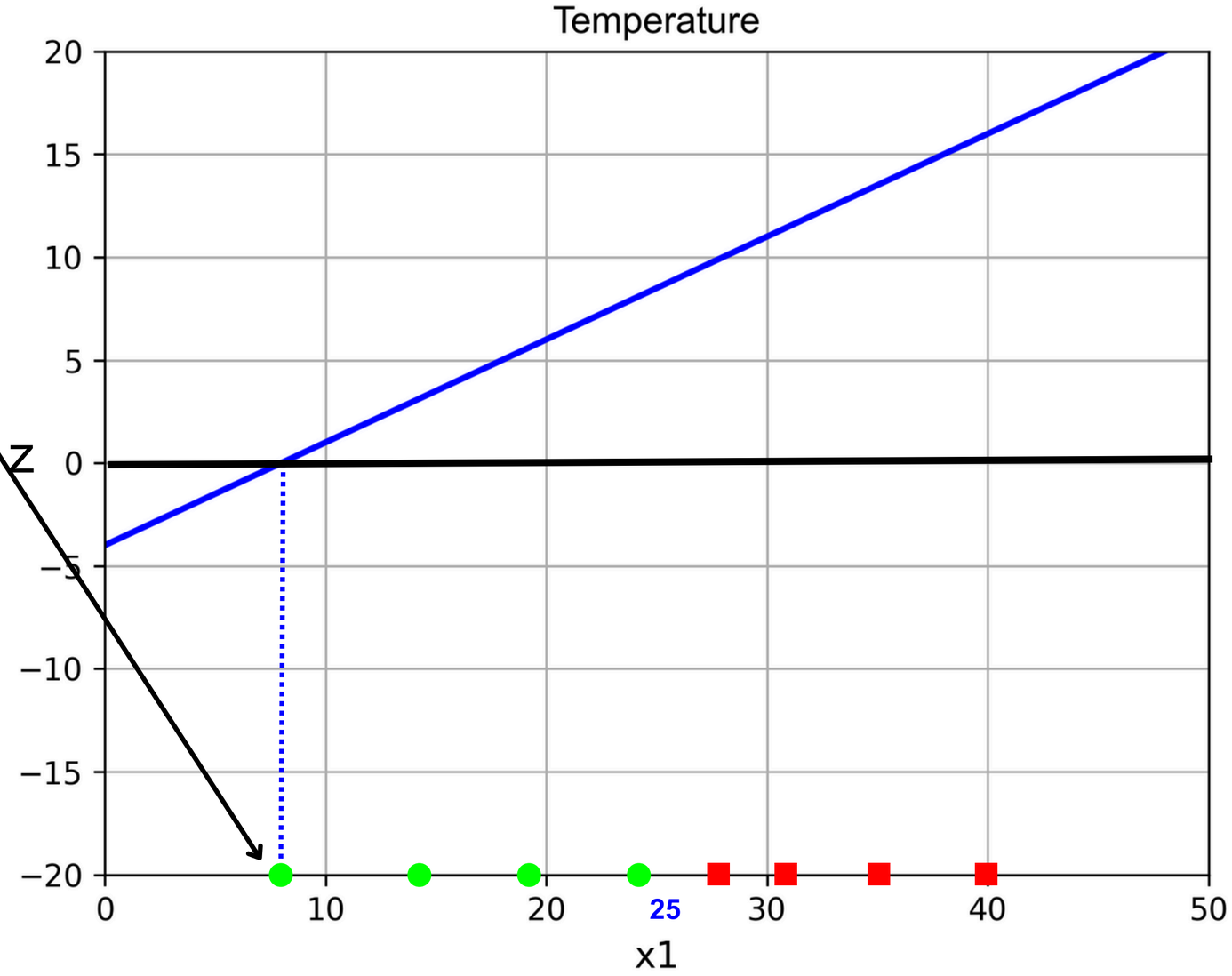
look at where the function is nul!



$$z(x_1) = (0.5)x_1 - 4$$

look at where the function is nul!

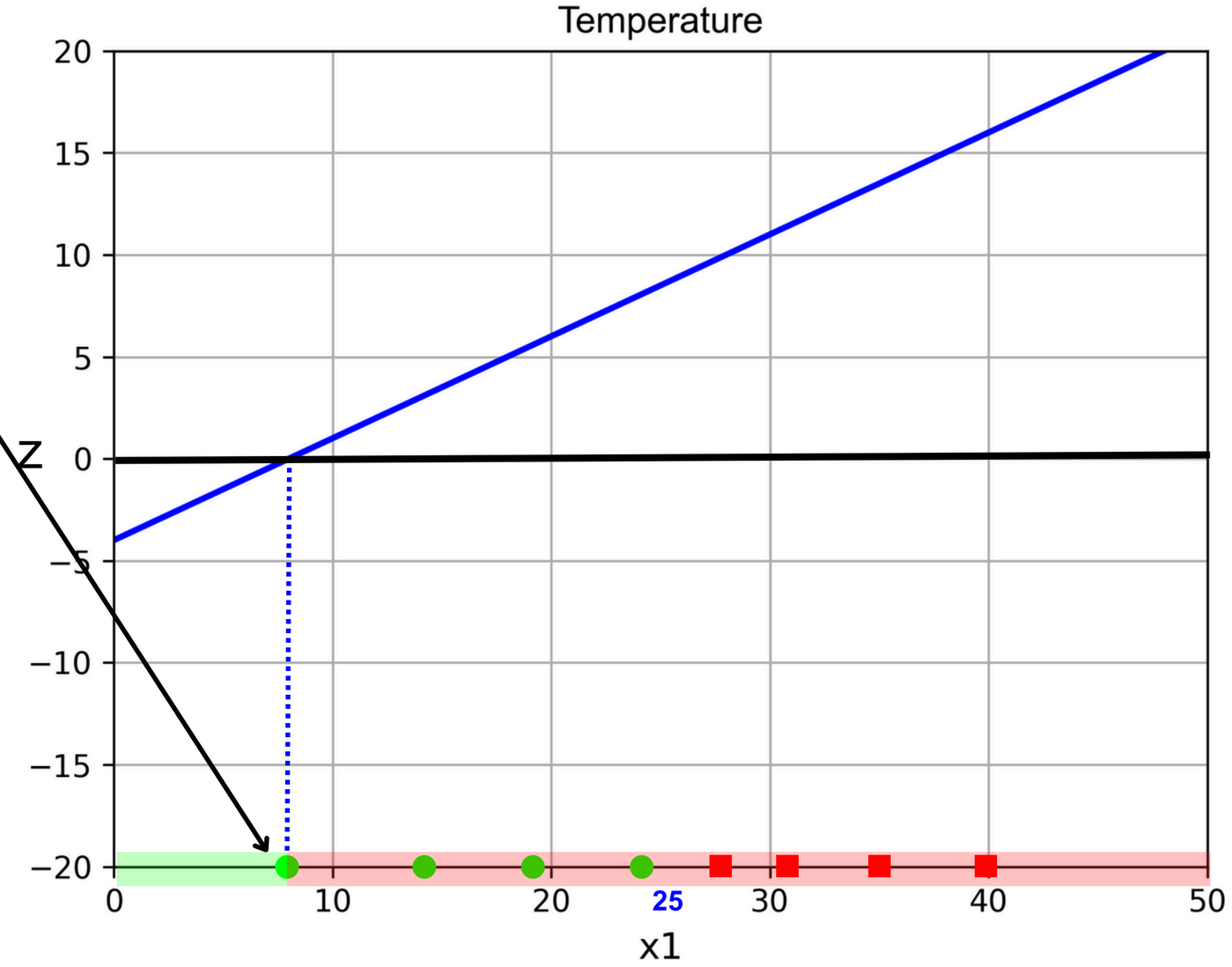
It means that any example with $x_1 \geqslant 8$ is considered as belonging to class 1



$$z(x_1) = (0.5)x_1 - 4$$

look at where the function is nul!

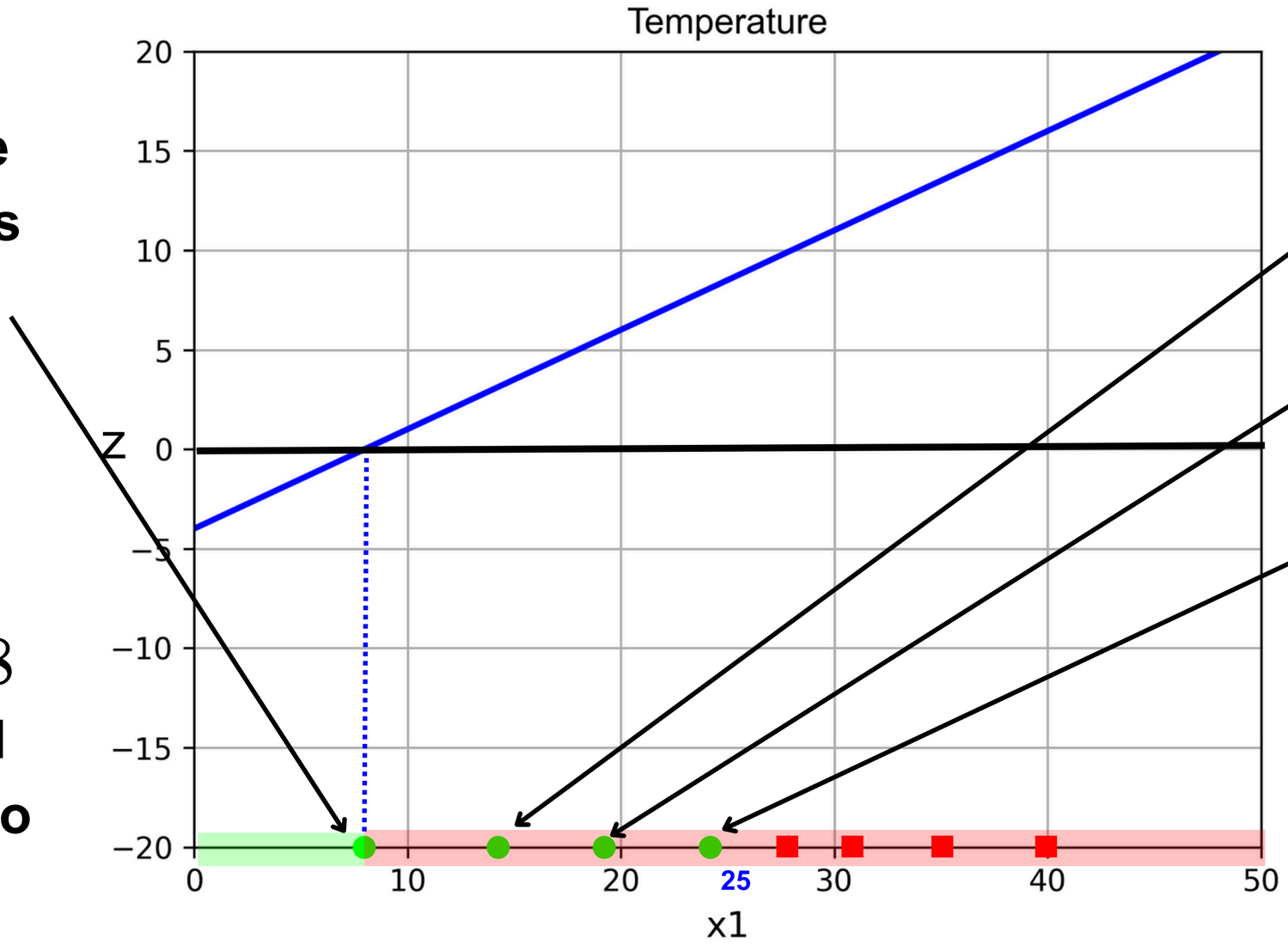
It means that any example with $x_1 \geqslant 8$ is considered as belonging to class 1



$$z(x_1) = (0.5)x_1 - 4$$

look at where the function is nul!

It means that any example with $x_1 \geqslant 8$ is considered as belonging to class 1



$z_2 = 3.5$
$P(Y = 0) = 0.02931$

$z_3 = 5.5$
$P(Y = 0) = 0.00407$

$z_4 = 8$
$P(Y = 0) = 0.00033$

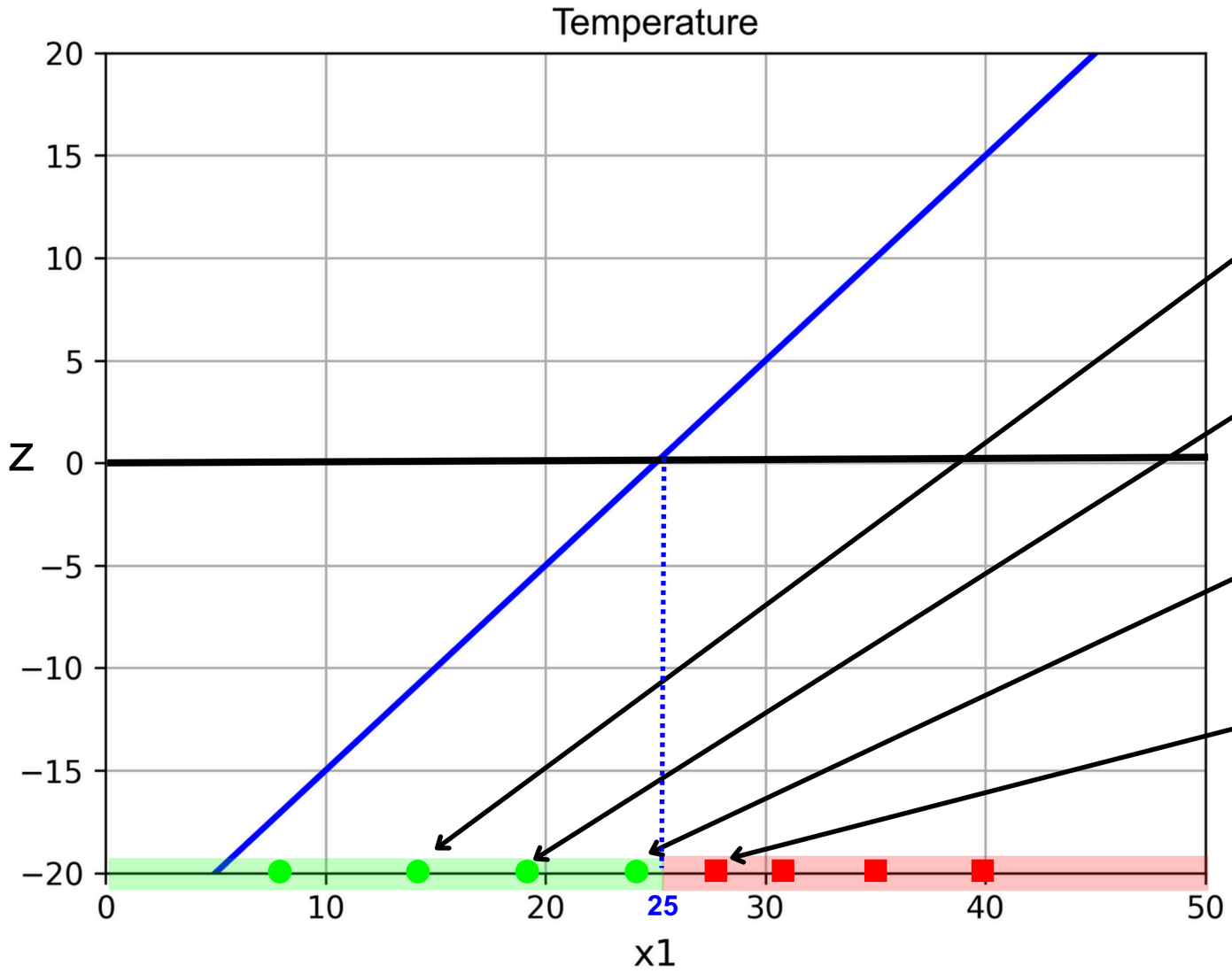
Likelihood

These incorrect values are very small, causing the likelihood product to diminish significantly.

$$0.0293 \times 0.0040 \times 0.0003 = 3.516 \times 10^{-8} = 0.00000003516$$

$$L = 2.0002713231 \times 10^{-8}$$

$$z(x_1) = x_1 - 25$$



$z_2 = -10$

$P(Y = 0) = 0.99995$

$z_3 = -6$

$P(Y = 0) = 0.99752$

$z_4 = -1$

$P(Y = 0) = 0.73105$

$z_5 = 2$

$P(Y = 1) = 0.88079$

Likelihood

We'll actually use this initial likelihood to train our model, adjusting the weight and bias to improve its accuracy.

Our goal is to have the highest likelihood possible.

Log Loss function

To train our model, we will not directly use the current Likelihood of our model but instead we will use the negative log of L.

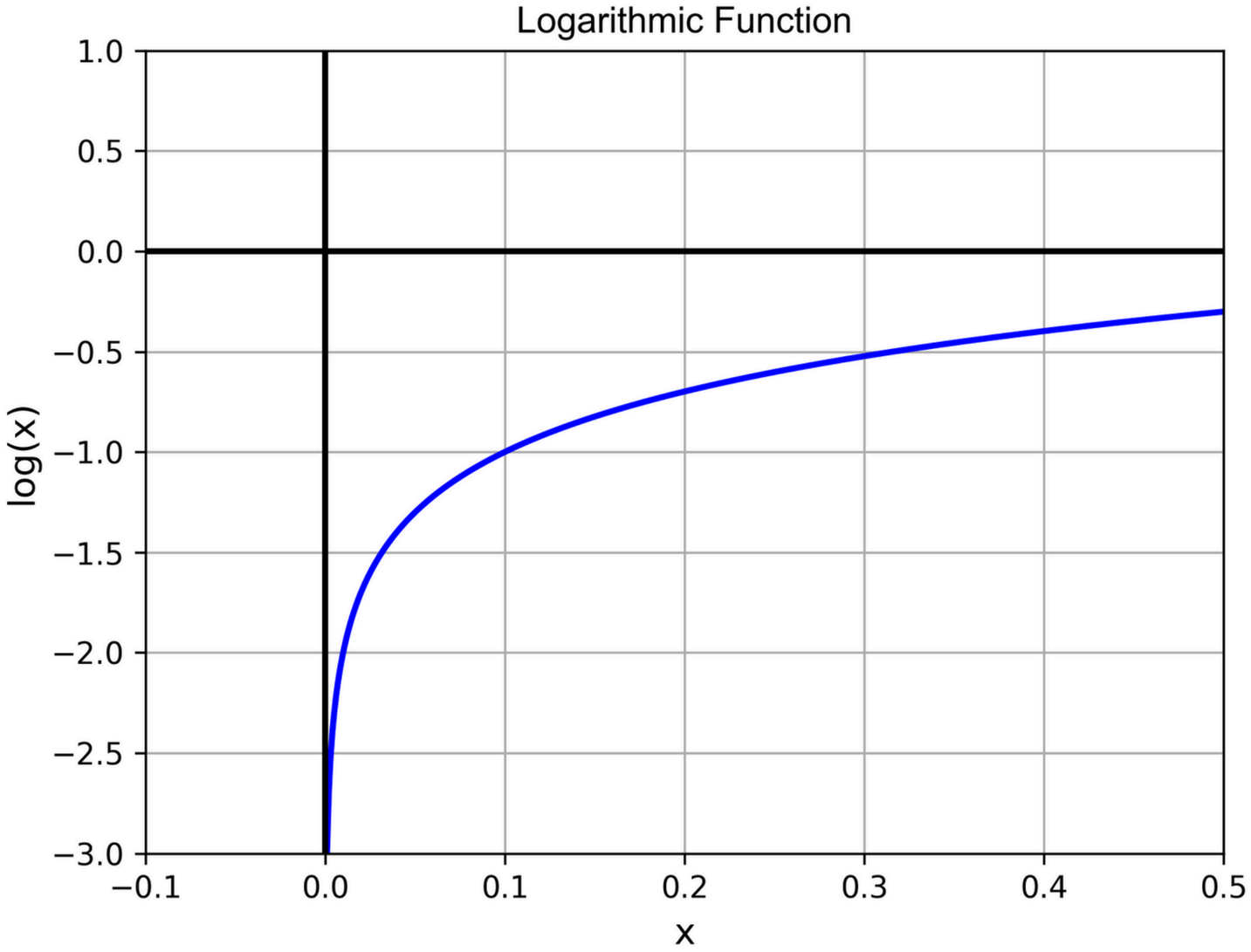
Because the Likelihood is strictly between 0 and 1

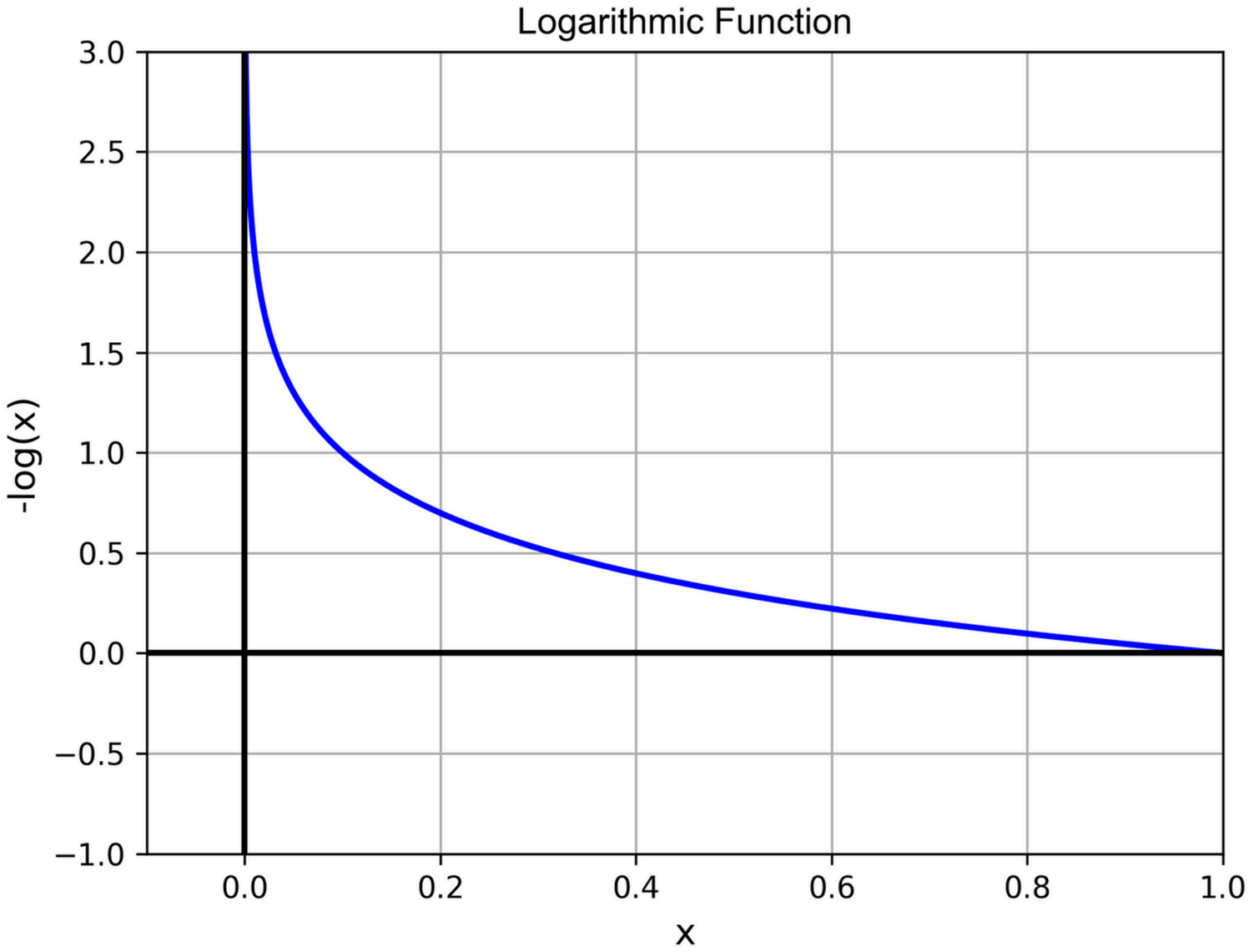
Log Loss function

To train our model, we will not directly use the Likelihood of our model but instead we will use the negative log of L.

Because the Likelihood is strictly between 0 and 1 $0 < L < 1$

And the logarithmic function provides a well-defined scale for values between 0 and 1, capturing small variations effectively.


Logarithmic Function
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
−3.0
log(x)
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
x


Logarithmic Function
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
-log(x)
0.0
0.2
0.4
0.6
0.8
1.0
x

Log Loss function

The negative logarithm transforms likelihood values from a range between 0 and 1 to a scale from 0 to $+\infty$, allowing us to more clearly interpret the likelihood of our model.

The -log(L) is called the Cost Function. (some use Loss/Cost interchangeably)

$$-log\left(L\right) = -log\left(\prod_{i=1}^{m}\left(a_i^{y_i}\left(1-a_i\right)^{1-y_i}\right)\right)$$

Log Loss function

We generally divide the Cost by the number of examples in the dataset, to normalize it relative to the size of the dataset.

The final Cost expression is:

$$-\frac{1}{m}log\left(L\right) = -\frac{1}{m}log\left(\prod_{i=1}^{m}\left(a_i^{y_i}\left(1-a_i\right)^{1-y_i}\right)\right)$$

Log Loss function

A more simplified expression:

$$L = -\frac{1}{m}\sum_{i=1}^{m}(y_i.log(a_i) + (1-y_i).log(1-a_i))$$

Log Loss function

Let’s calculate the current Cost of our model

$$L = -\frac{1}{8} log\left(2.0002 \times 10^{-8}\right) = 0.96236$$

Let’s calculate the Cost of the descent perceptron $z\left(x_1\right) = x_1 - 25$

$$L = -\frac{1}{8} log\left(0.640675\right) = 0.02417$$

Gradient descent

Now that we've compared the costs of our current model and a more accurate one, we can see that our model's cost is too high. Our goal is to minimize this cost effectively.

To achieve this, we'll use an algorithm called Gradient Descent, one of the most widely used optimization techniques in machine learning and deep learning.

Gradient descent

The Gradient Descent algorithm uses the Log Loss function to adjust the model's weights and biases, improving its performance. To fully understand this process, we first need to explore multivariable functions.

Functions of several variables

Up until now, we have only met functions of single variables. From now on we will meet functions such as $z = f(x, y)$ and $w = f(x, y, z)$ which are functions of two variables and three variables respectively.

The domain of $z = f(x, y)$ is the set of all points (x, y) which f is defined.

Functions of several variables

For example, let f be

$$f(x,y,z) = x^2 + xy + y^2 - \sqrt{z}$$

We will find its value at the point (1,3,4). We get

$$f(1,3,4) = 1^2 + (1)(3) + 3^2 - \sqrt{4} = 11$$

Functions of several variables

$$f(x, y, z) = x^2 + xy + y^2 - \sqrt{z}$$

To find its domain, we notice that the first three terms are defined for all real numbers; However, the last term $\sqrt{z}$ is only defined if $z \geqslant 0$ Hence the domain is everything on or above the z-axis.

$$D_f = \mathbb{R} \times \mathbb{R} \times \mathbb{R}_{x \geqslant 0}$$

Graphs of functions of two variables

A graph of a function of two variables is the graph of the function

$$z = f(x, y)$$

For example, let's sketch the graph of

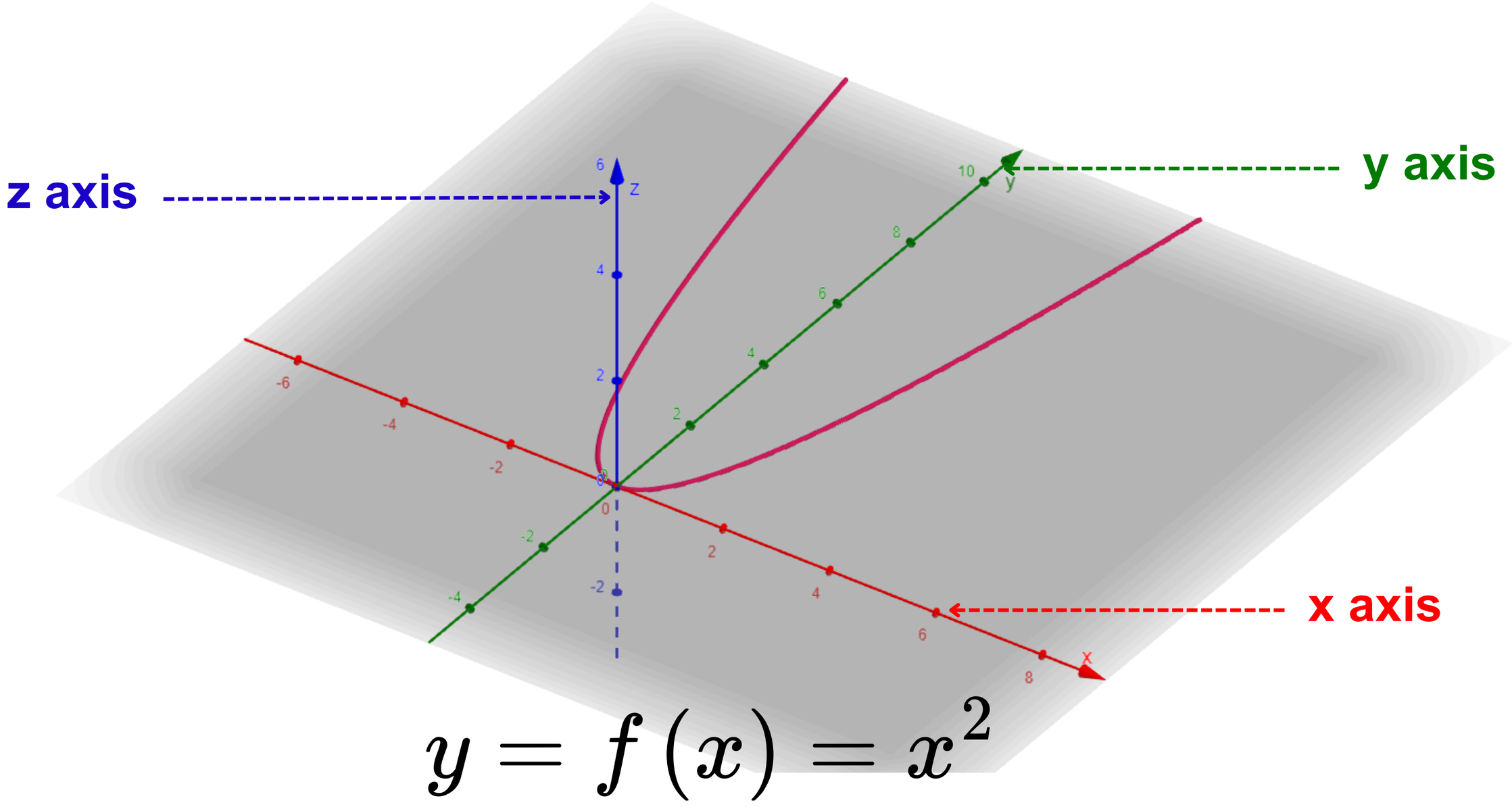
$$z = x^2 + y$$

Graphs of functions of two variables

$$z = x^2 + y$$

We begin by sketching

$$y = f(x) = x^2$$



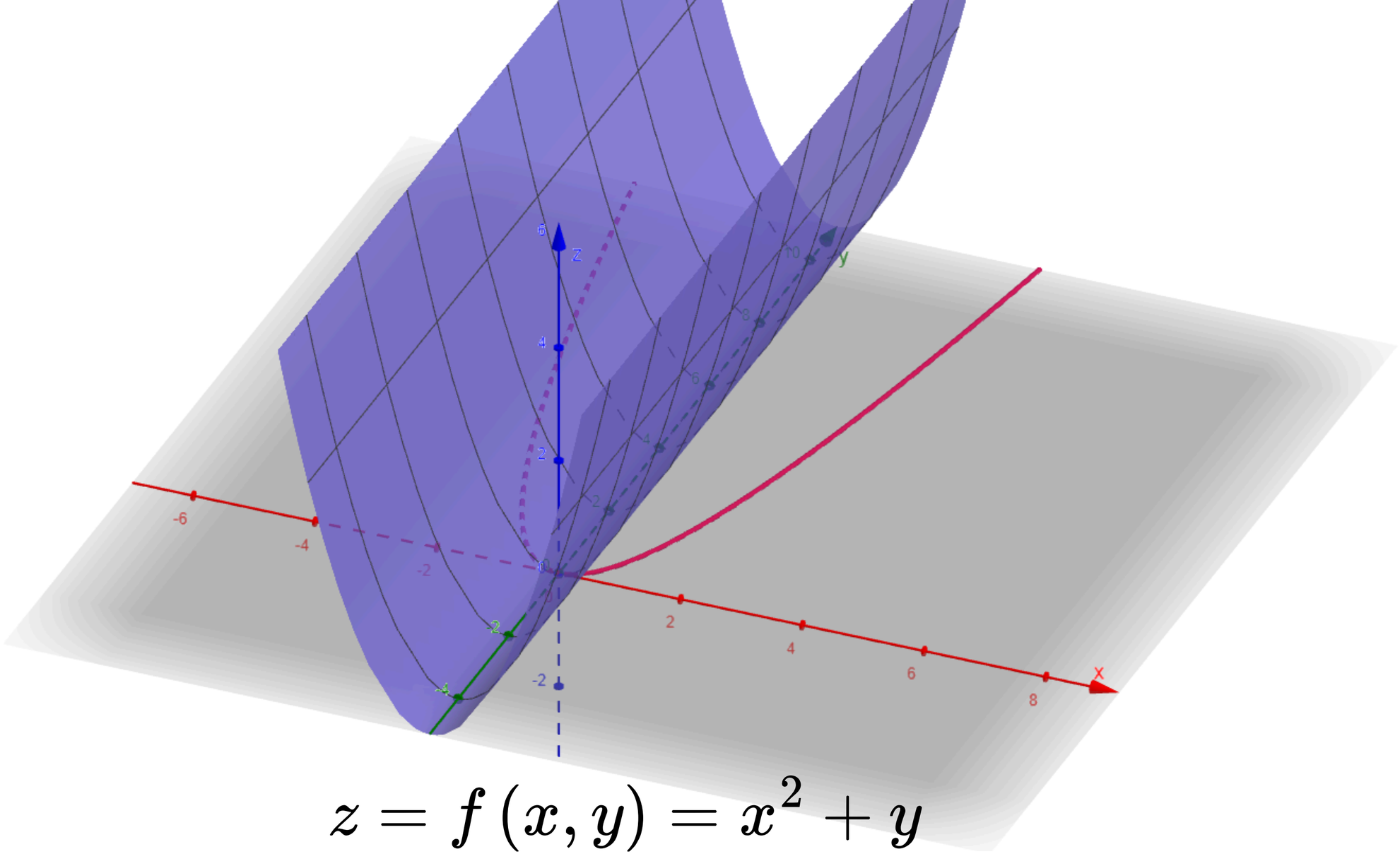
$$y = f(x) = x^2$$

Graphs of functions of two variables

Now, what do you think will happen if we add the y values to x^2 and sketch the graph?

We will have the variable z that depends on both x and y

$$z = f(x, y) = x^2 + y$$



$$z = f(x, y) = x^2 + y$$

Graphs of functions of two variables

LIVE DEMO

Limits and continuity in functions of two variables

As always, we will first review limits in functions of a single variable

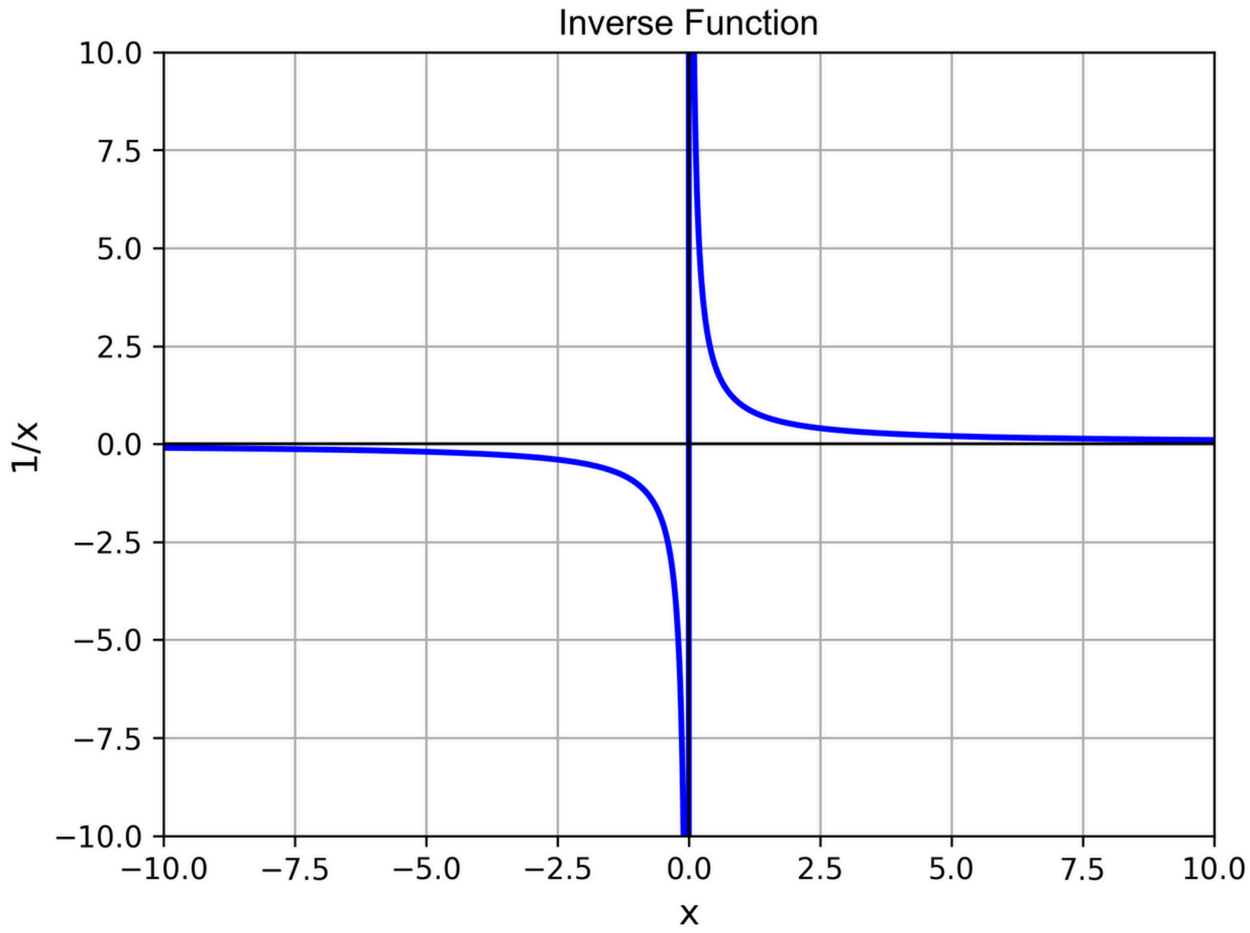
The limit of a function is a fundamental concept in calculus and analysis concerning the behavior of that function near a particular input which may or may not be in the domain of the function. They are used to define integrals, derivatives, and continuity.

Limits and continuity in functions of two variables

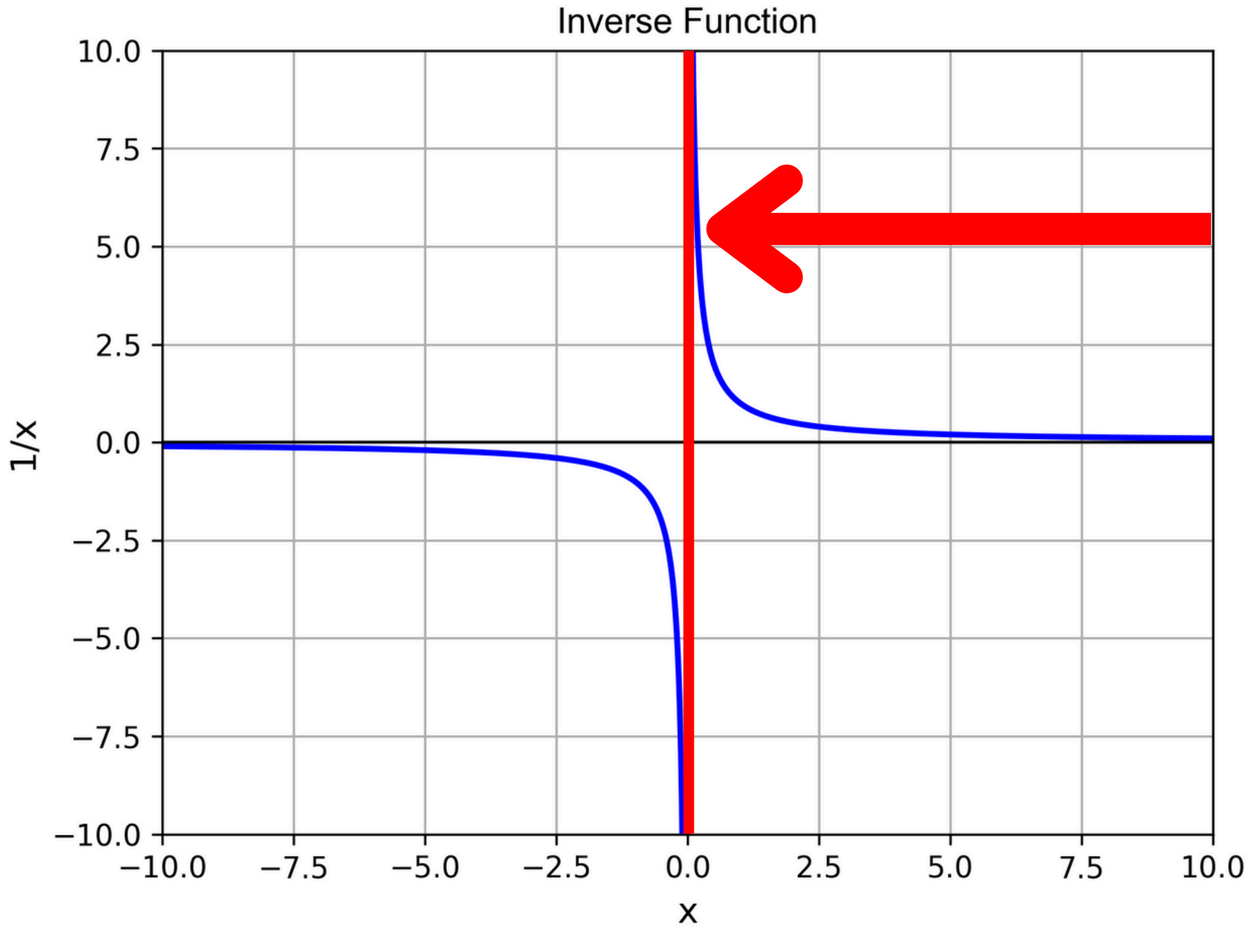
A limit describes the behavior of a function as its input approaches a particular value. Think of it as observing how the output of a function changes as we get closer to a specific point on the x-axis.

Limits and continuity in functions of two variables

$$\lim_{x \to 0^+} \frac{1}{x}$$

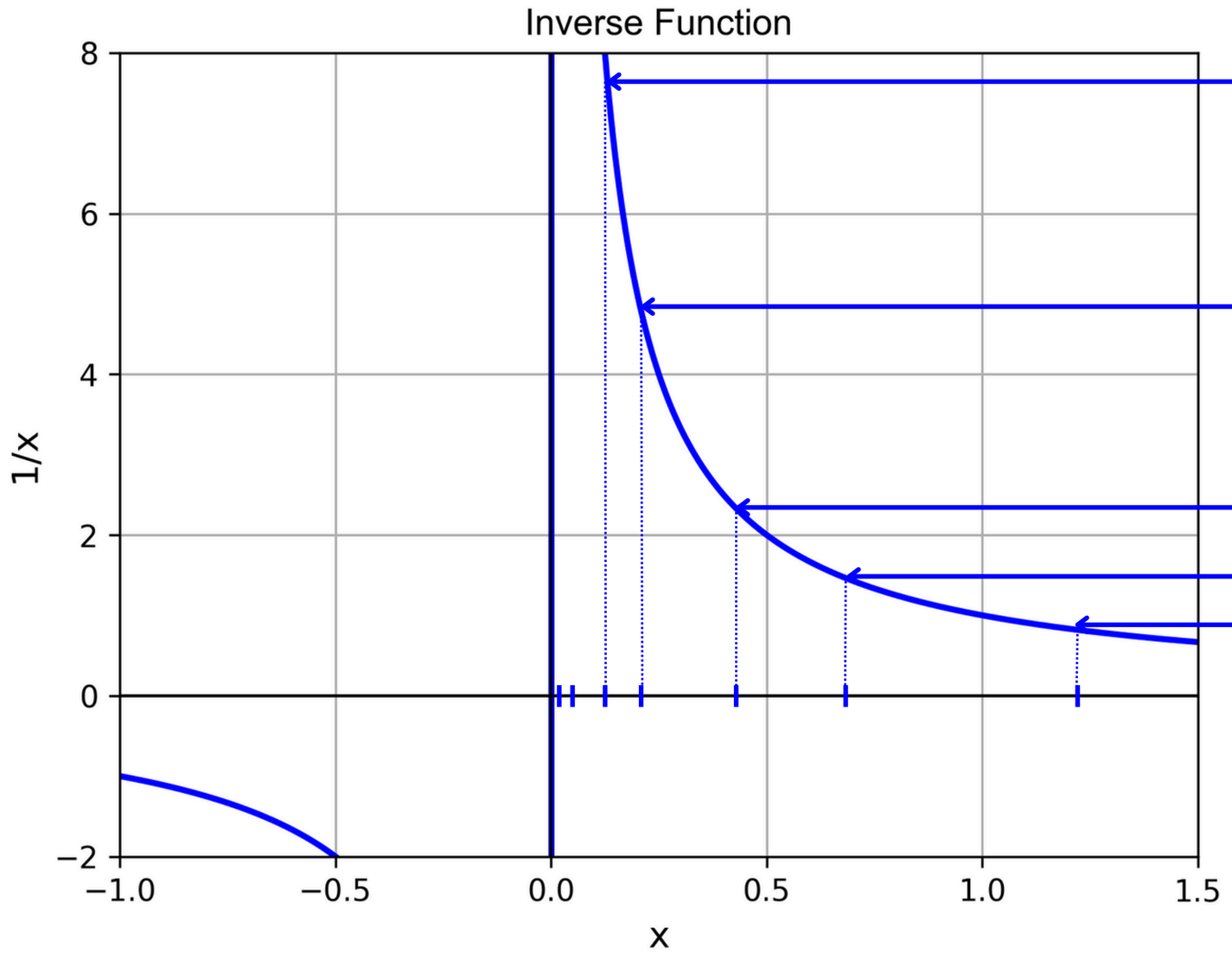


$$\lim_{x \to 0^+} \frac{1}{x}$$



Right-hand limit, it is is the value the function approaches when the variable approaches its limit from the right.

There is also a left-hand limit.



$x = 0.2, f(x) = 5$

$x = 0.25, f(x) = 4$

$x = 0.45, f(x) = 2.22$

$x = 0.7, f(x) = 1.42$

$x = 1.25, f(x) = 0.8$

$x = 0.1, f(x) = 10$

$x = 0.0001, f(x) = 10000$

Limits and continuity in functions of two variables

Finding some limits (LIVE DEMO) (INTUITIVE)

$$\lim_{x \to 0^+} \frac{1}{x} = +\infty \qquad \lim_{x \to +\infty} \frac{1}{x} = 0^+ \qquad \lim_{x \to -\infty} x^2 = +\infty$$

$$\lim_{x \to 0^-} \frac{1}{x} = -\infty \qquad \lim_{x \to -\infty} \frac{1}{x} = 0^- \qquad \lim_{x \to +\infty} x^2 = +\infty$$

Limits and continuity in functions of two variables

Actually, the definition of of limits in calculus are as follows:

Limit of a function at infinity:

We say that f tends to l at $+\infty$ if, for x large enough, $f\left(x\right)$ is as close to l as we want. Precisely:

$$\forall \varepsilon > 0, \exists A \in R, \forall x \geq A, \left| f(x) - l \right| < \varepsilon$$

Limits and continuity in functions of two variables

Limit of a function at a point

except that this time x can approach as close as we want to a real number a
We therefore suppose that we have a function f defined on an interval I of R and that a is an element of l, or a bound of l. We say that f tends to l at a if:

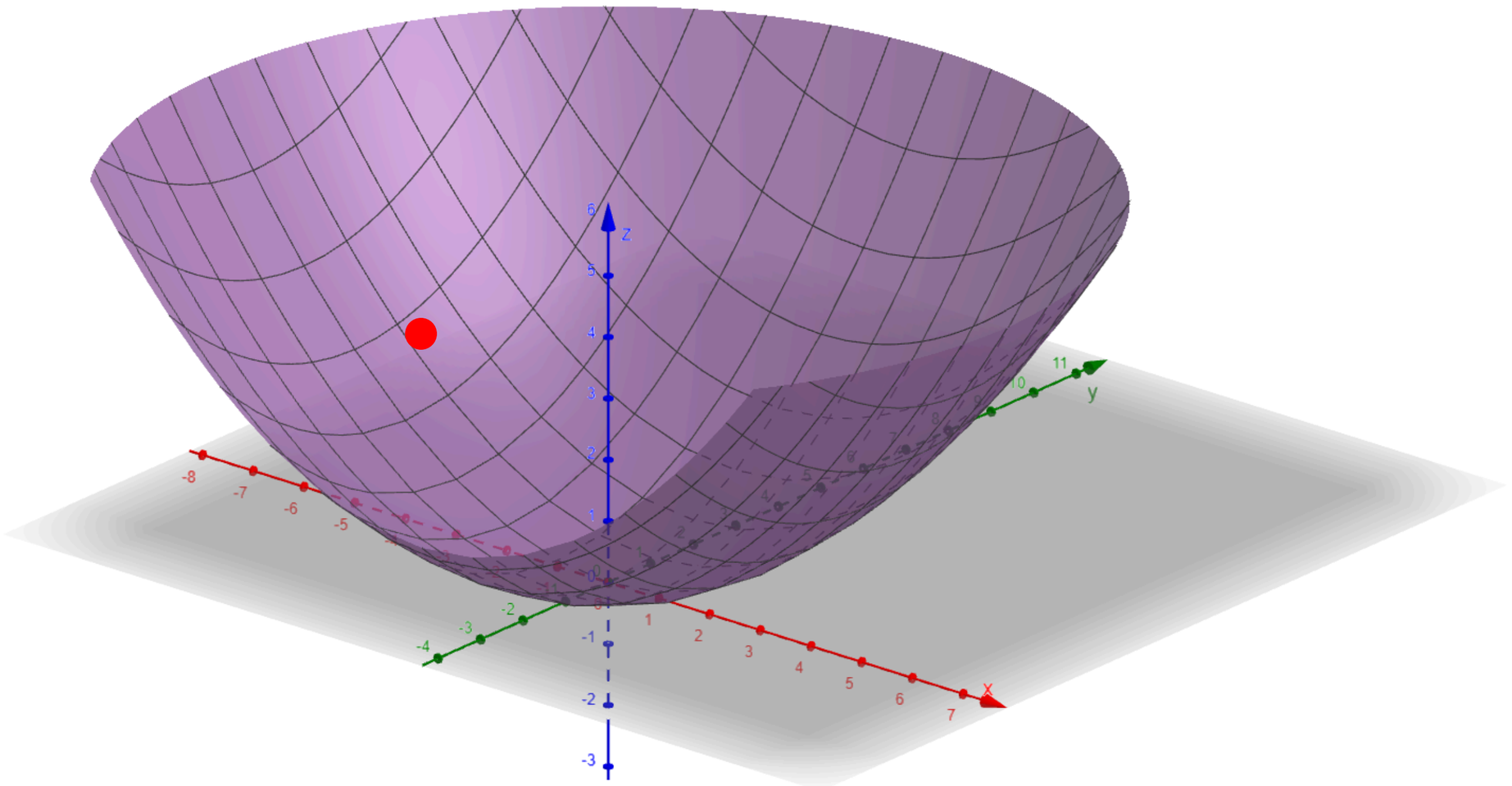
$$\forall \varepsilon > 0, \exists \delta > 0, \forall x \in I \cap (a-\delta, a+\delta), |f(x) - l| < \varepsilon$$

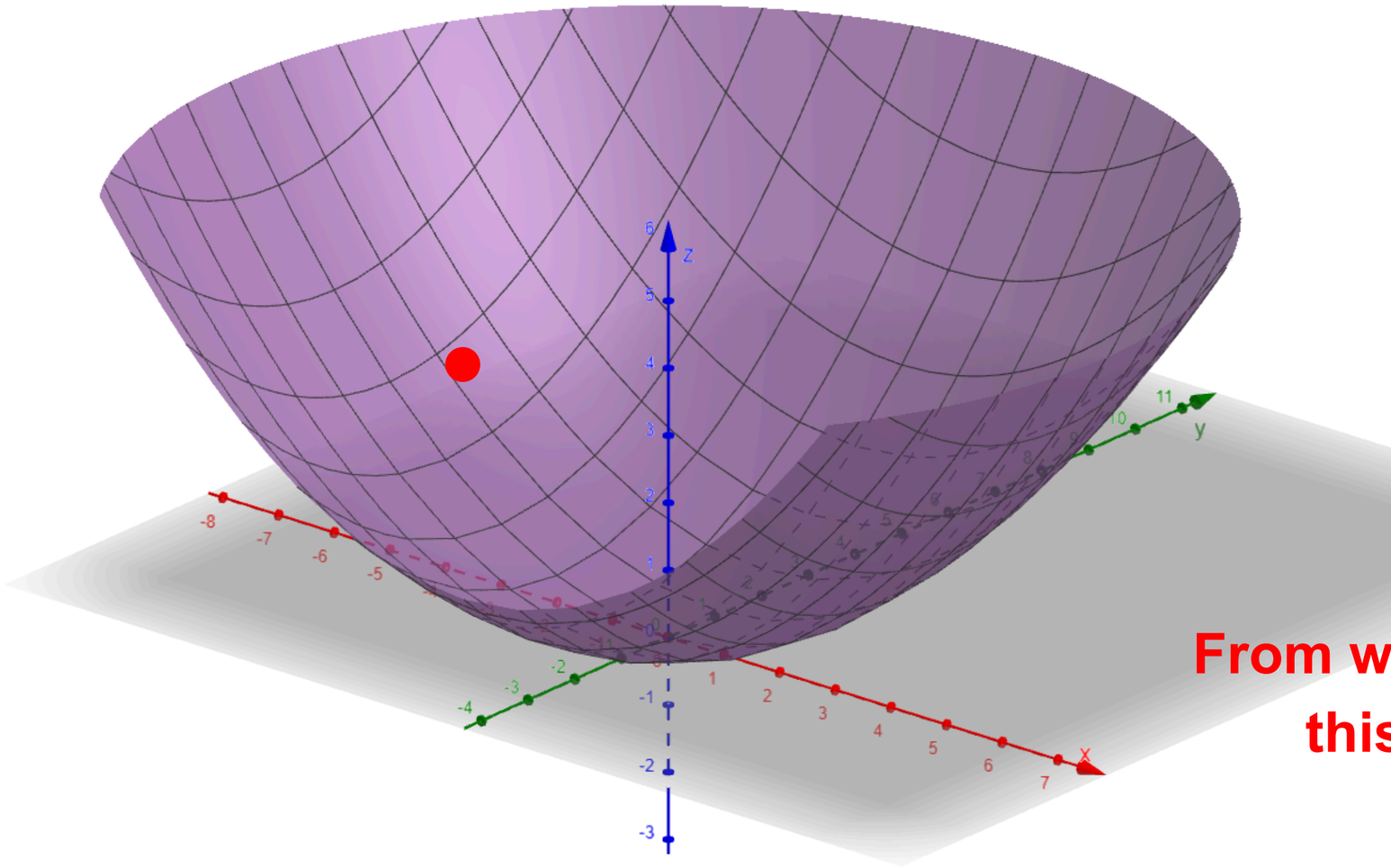
Limits and continuity in functions of two variables

With functions of a single variable, we can take the limit from either above or below. These denoted:

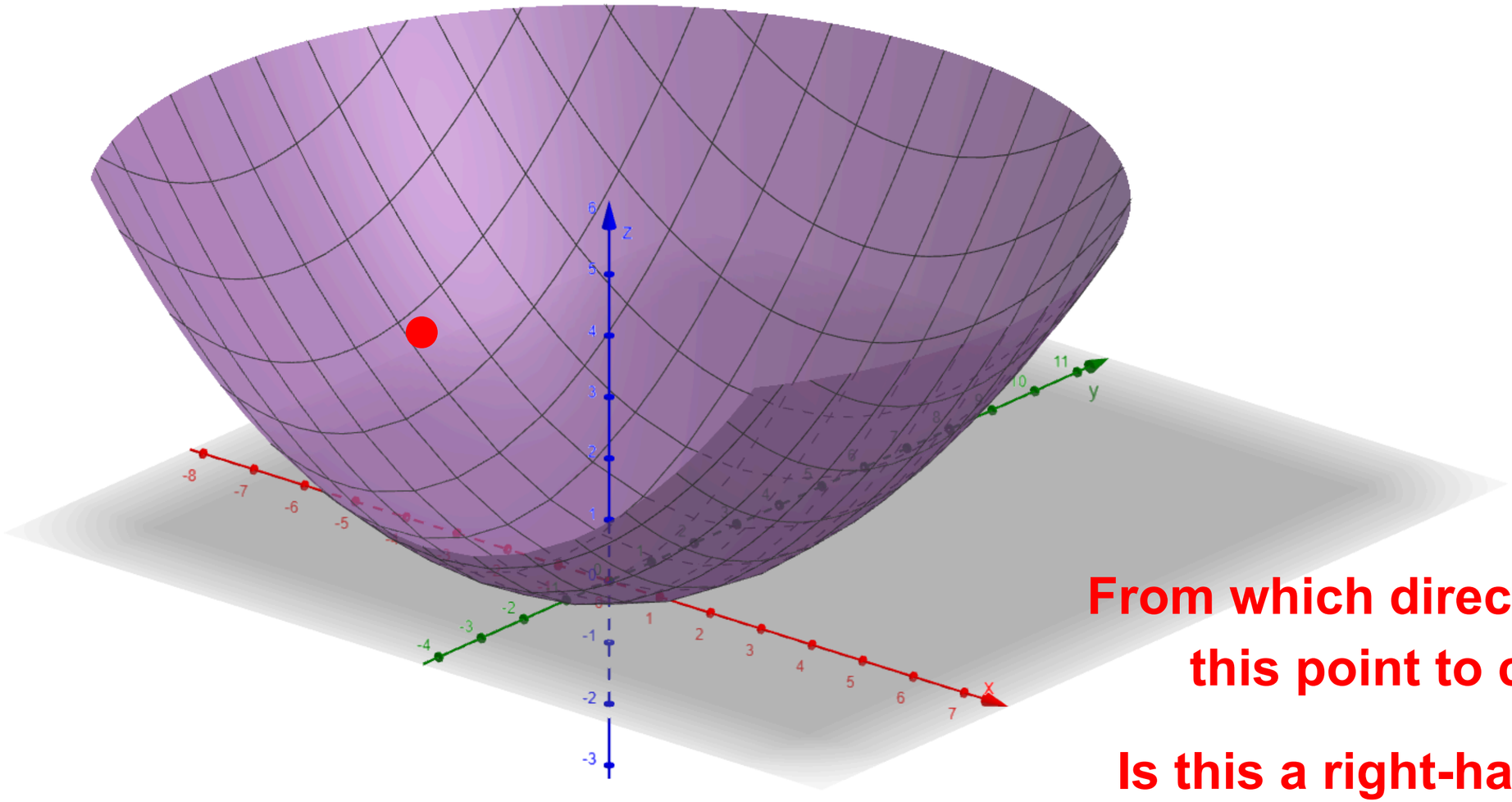
$$\lim_{x \to x_0^+} f(x), \lim_{x \to x_0^-} f(x)$$

When we have two or three variables, there are infinitely many ways to approach a point (think of a point on the plane) and as a result we must take a limit along a curve.


z
6
5
4
3
2
1
0
-1
-2
-3
y
11
10
x
-8
-7
-6
-5
-4
-3
-2
1
2
3
4
5
6
7

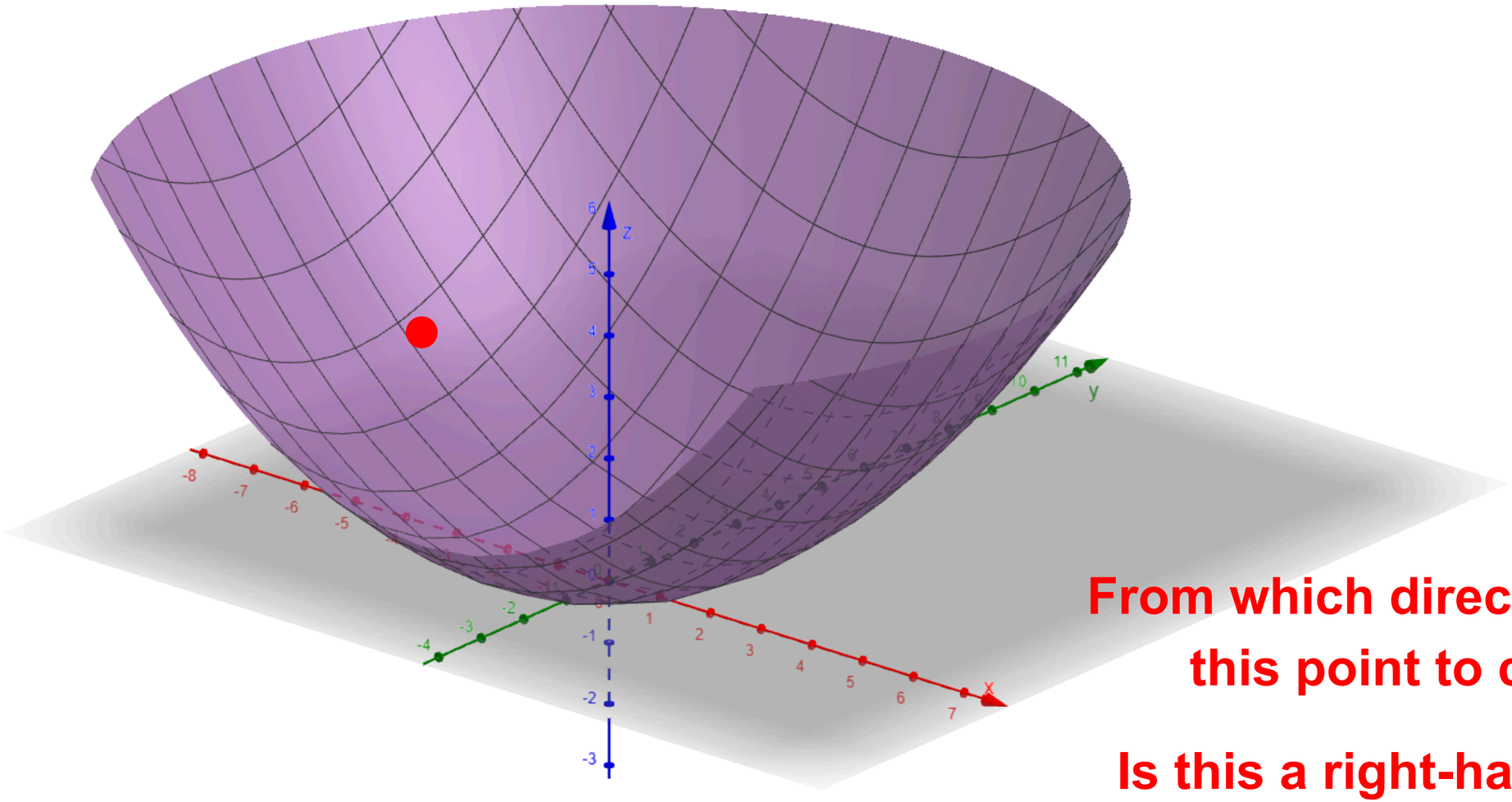


From which direction should I approach this point to calculate the limit?



From which direction should I approach this point to calculate the limit?

Is this a right-hand side limit? or left?



From which direction should I approach this point to calculate the limit?

Is this a right-hand side limit? or left?

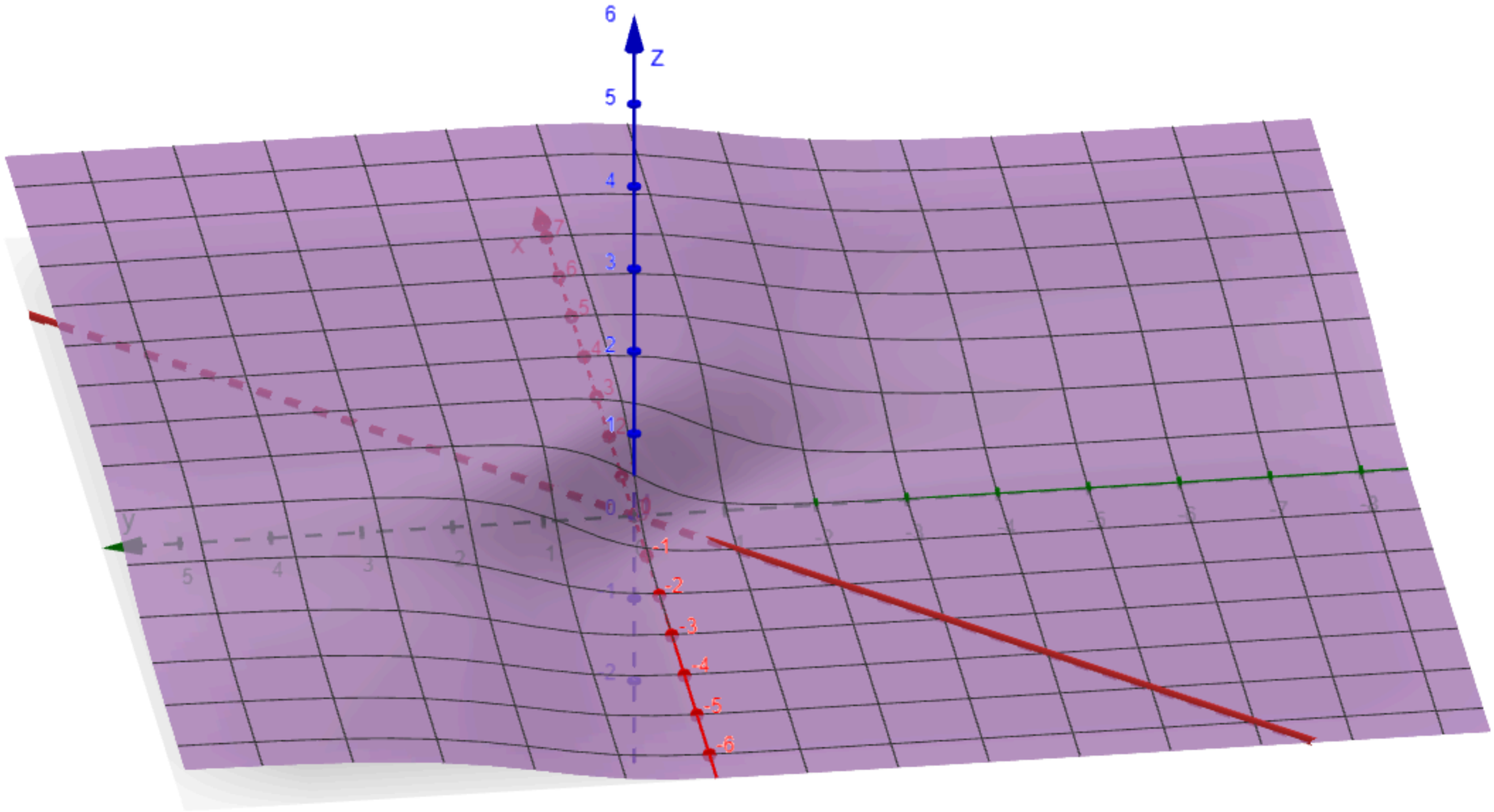
There are infinite direction

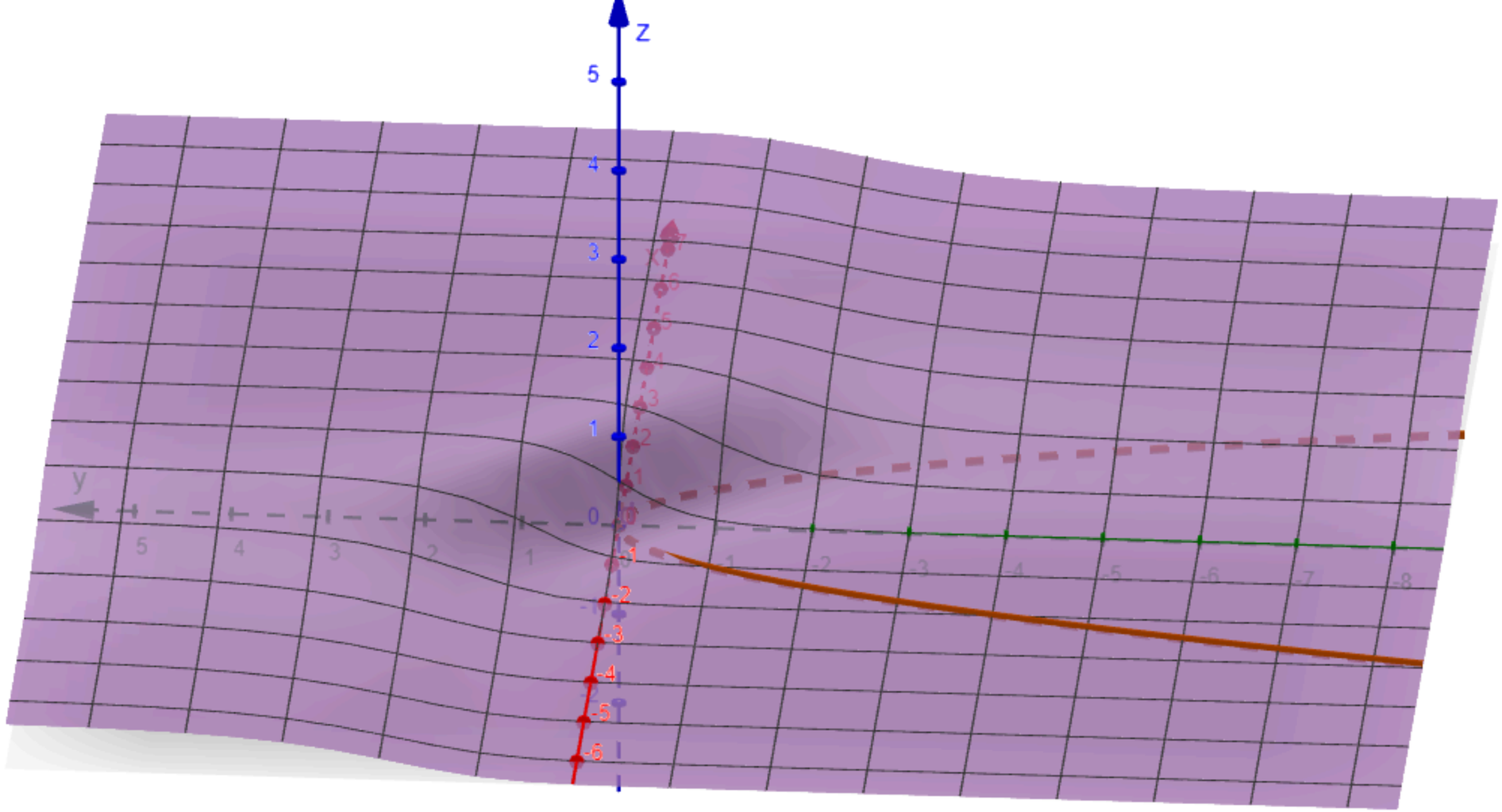
Limits and continuity in functions of two variables

(LIVE DEMO)

Limits and continuity in functions of two variables

So to calculate the limit, we MUST choose a curve to define a direction along which we take the limit.


z
6
5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
x
y


z
5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
y
5
4
3
2
1
-1
-2
-3
-4
-5
-6
-7
-8

Limits and continuity in functions of two variables

For a two variable function case, we take a curve C such that the point $(x_{0,}y_0)$ is on it, if the curve is parameterised by t, we will have:

$$x = x\,(t),\;\; y = y\,(t)$$

This "mechanism" allows us to vary two variables, x and y, <u>simultaneously</u> by expressing each as a function of a single variable, t
This approach can be helpful when calculating limits in multiple dimensions.

Limits and continuity in functions of two variables

(LIVE DEMO)

Limits and continuity in functions of two variables

so basically

$$x_0 = x\left(t_0\right), y_0 = y\left(t_0\right)$$

And the limits along the curve are defined as

$$\lim_{(x,y)\to(x_0,y_0)} f\left(x,y\right) = \lim_{t\to t_0} f\left(x\left(t\right), y\left(t\right)\right)$$

Limits and continuity in functions of two variables

Example: Find the limit of the function

$$f(x,y) = -\frac{xy}{x^2+y^2}$$

at $(0,0)$ **along the parabola** $y = x^2$

The line $y = x^2$ **can be parameterised by** $x = t$, $y = t^2$

Therefore, we substitue these values into f and take the limit $t \to 0$

Limits and continuity in functions of two variables

$$f(x,y) = -\frac{xy}{x^2+y^2}$$

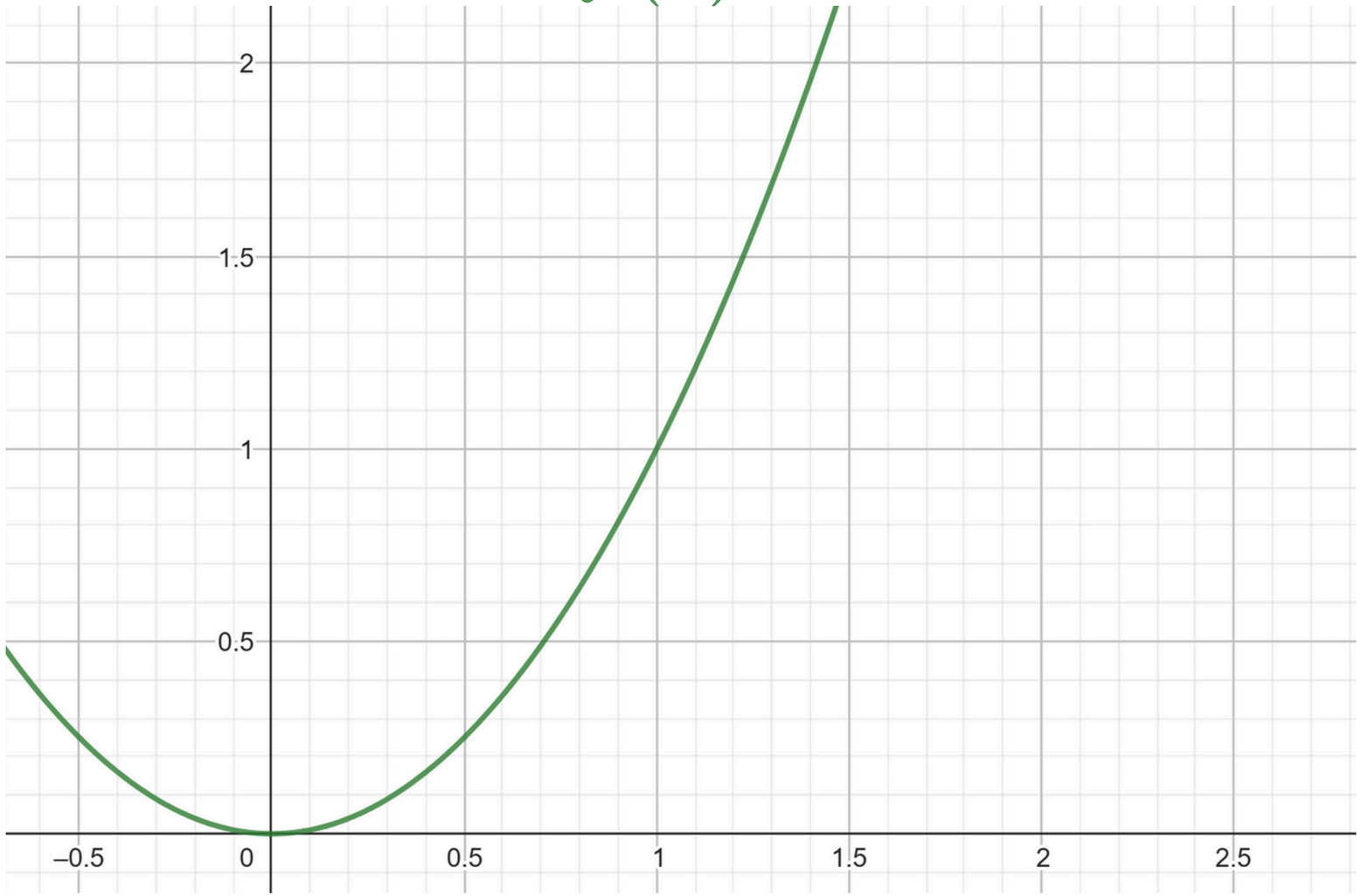
$$\lim_{t\to 0} f\left(t,t^2\right) = \lim_{t\to 0} -\frac{(t)\left(t^2\right)}{t^2+\left(t^2\right)^2}$$

$$= \lim_{t\to 0} -\frac{t^3}{t^2+t^4} = 0$$

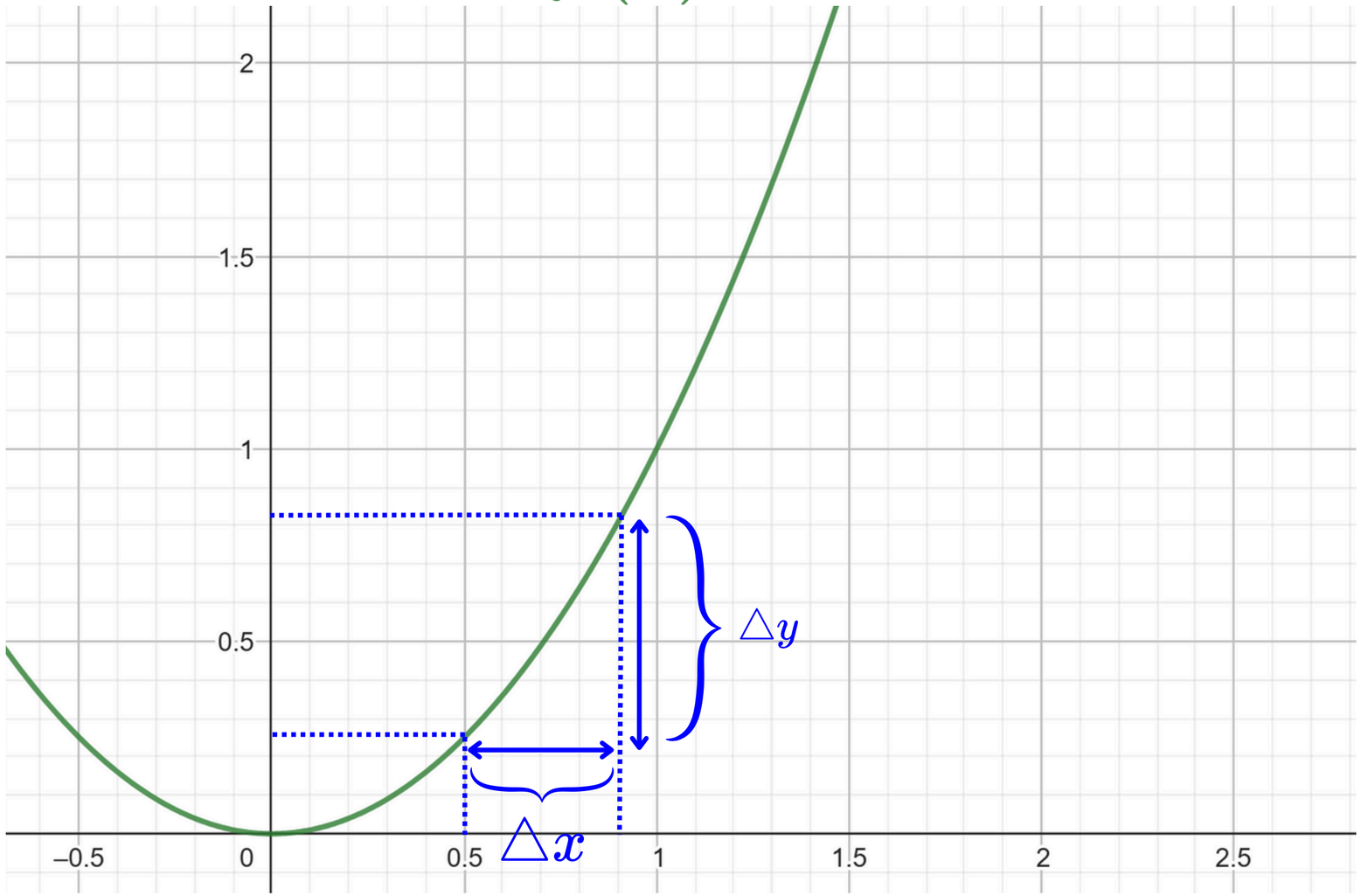
Derivatives

In mathematics, the derivative is a fundamental tool that quantifies the sensitivity of change of a function's output with respect to its input. The derivative of a function of a single variable at a chosen input value, when it exists, is the slope of the tangent line to the graph of the function at that point

$$f(x) = x^2$$



$$f(x) = x^2$$

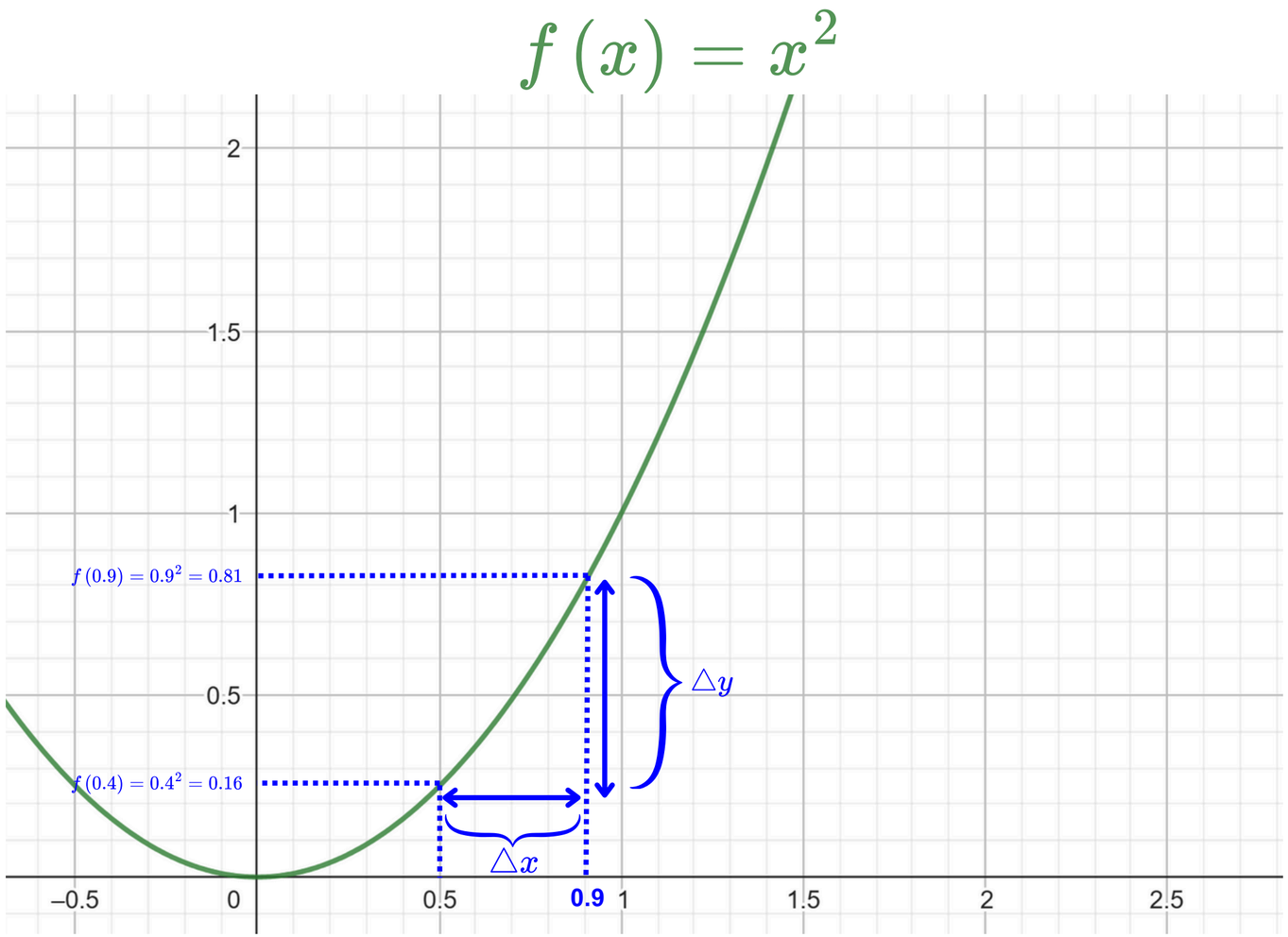


$$\triangle x = 0.9 - 0.5 = 0.4$$

$$\triangle y = f\left(0.9\right) - f\left(0.5\right) = 0.56$$

$\triangle x$ denotes the change on the x-axis

$\triangle y$ denotes the <u>exact</u> change on the y-axis

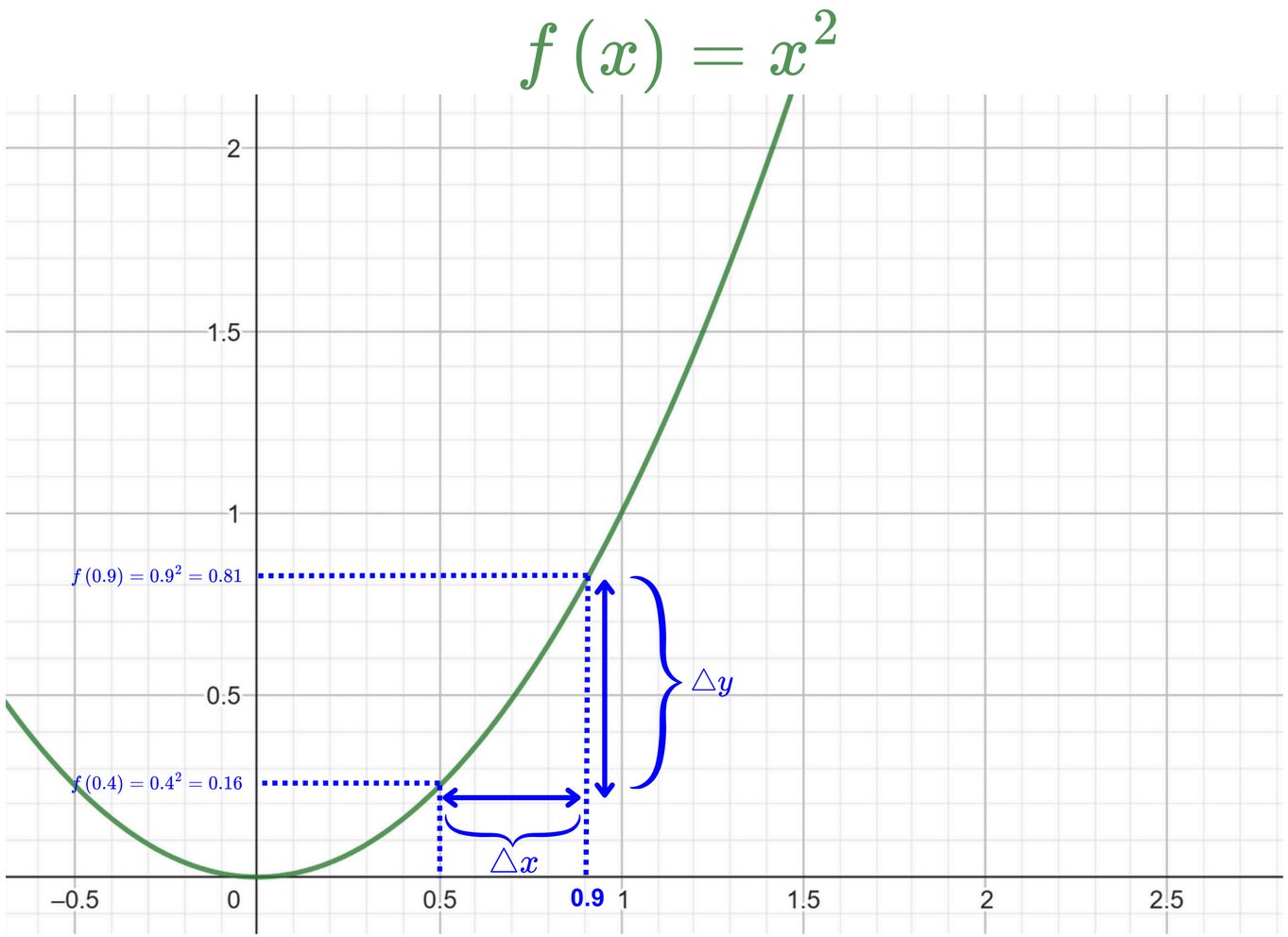


$$\triangle x = 0.9 - 0.5 = 0.4$$

$$\triangle y = f\left(0.9\right) - f\left(0.5\right) = 0.56$$

$\triangle x$ denotes the change on the x-axis

$\triangle y$ denotes the <u>exact</u> change on the y-axis



Derivatives

Derivative of a function f(x) signifies the rate of change of the function f(x) with respect to x at a point 'a' lying in its domain.

For a function to be differentiable at any point x = a in its domain, it must be continuous at that particular point.

Derivatives

To better understand this,

Let’s try to calculate the derivative of the function $f(x) = \sqrt[2]{x}$

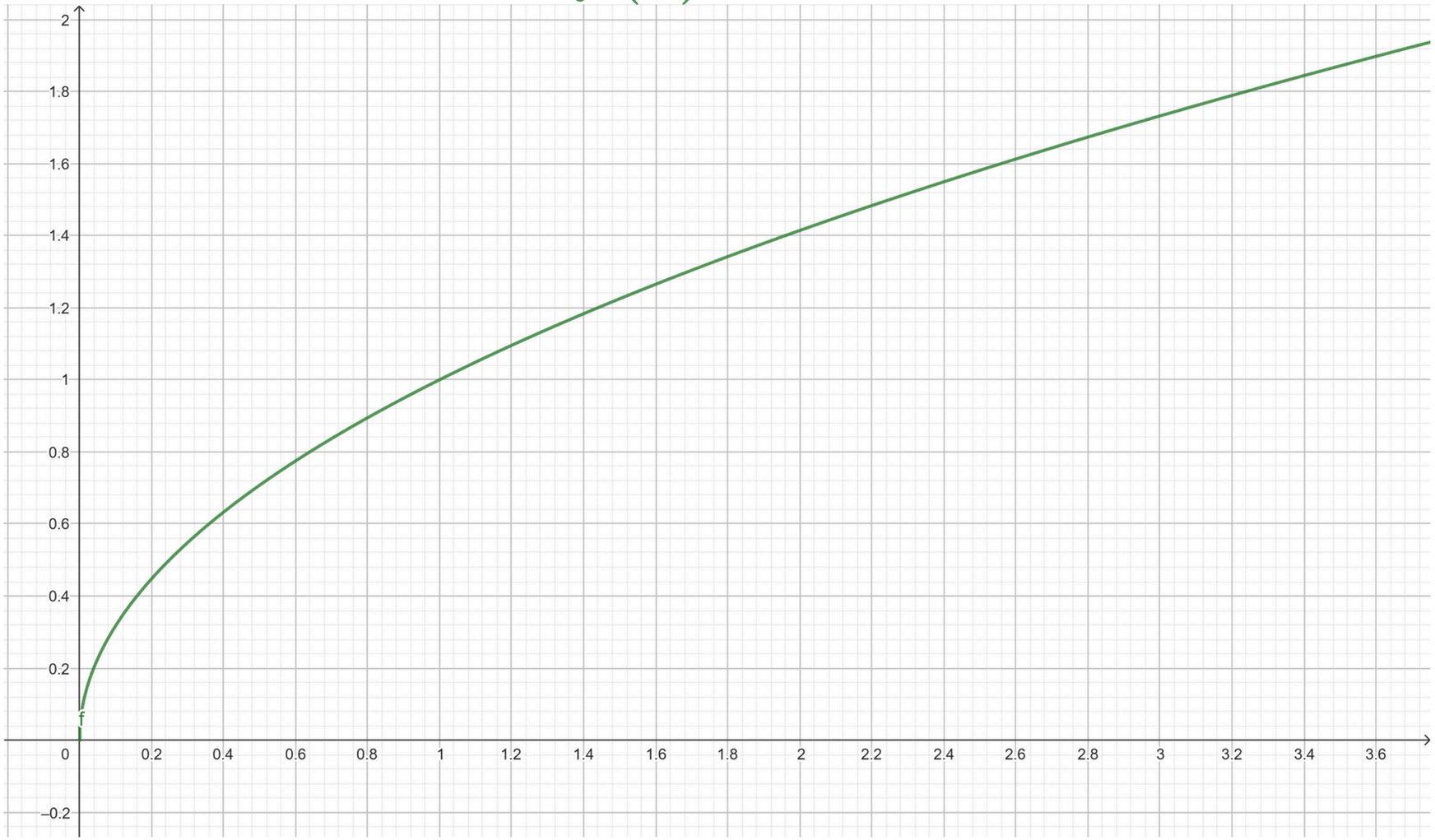
at the point a = 4

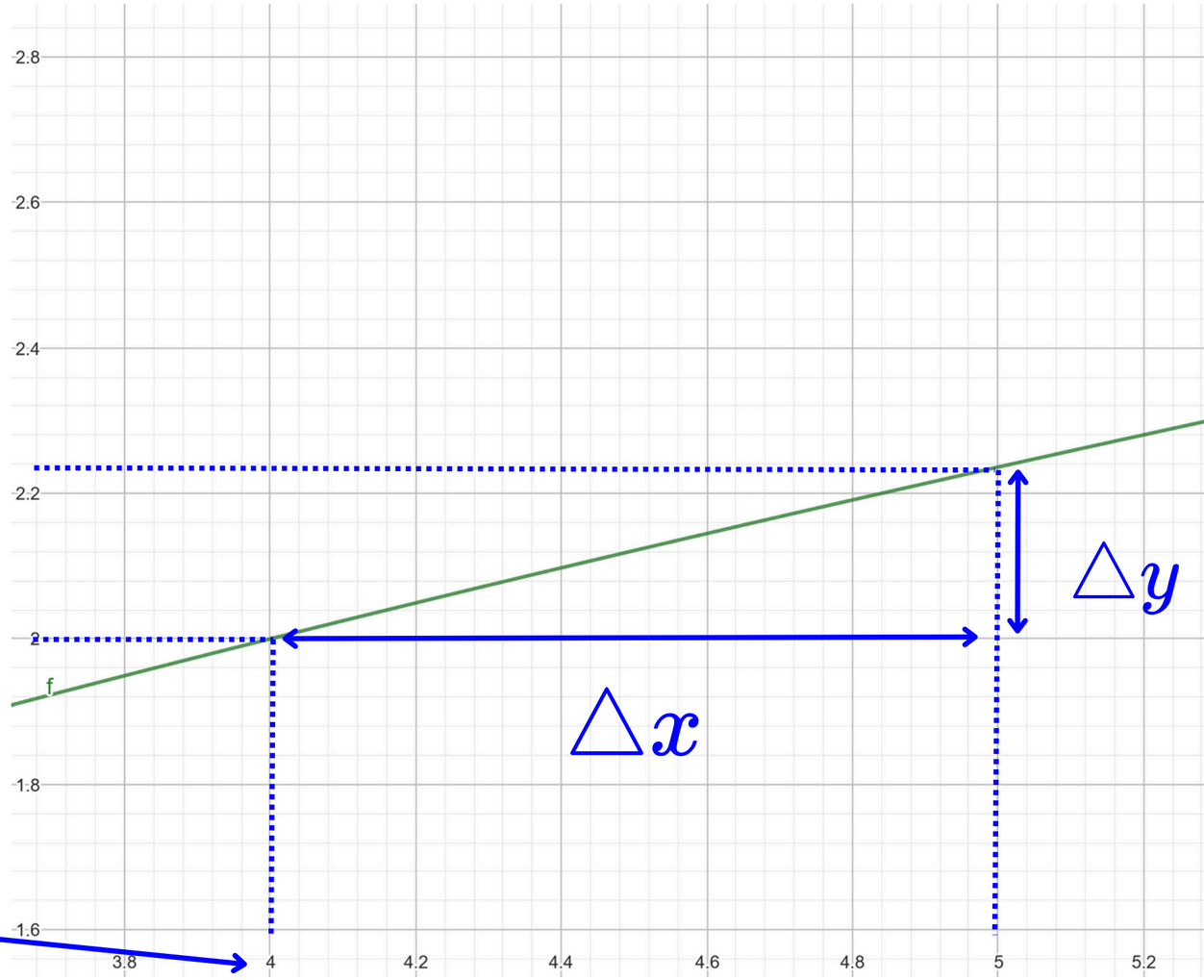
Derivatives

By derivative, we mean the rate of change of the function at the point a = 4.

By rate of change of the function at the point a = 4, we simply mean the change on y-axis $\triangle y$ if the change on the x-axis $\triangle x$ is so small.

$$f(x) = \sqrt[2]{x}$$

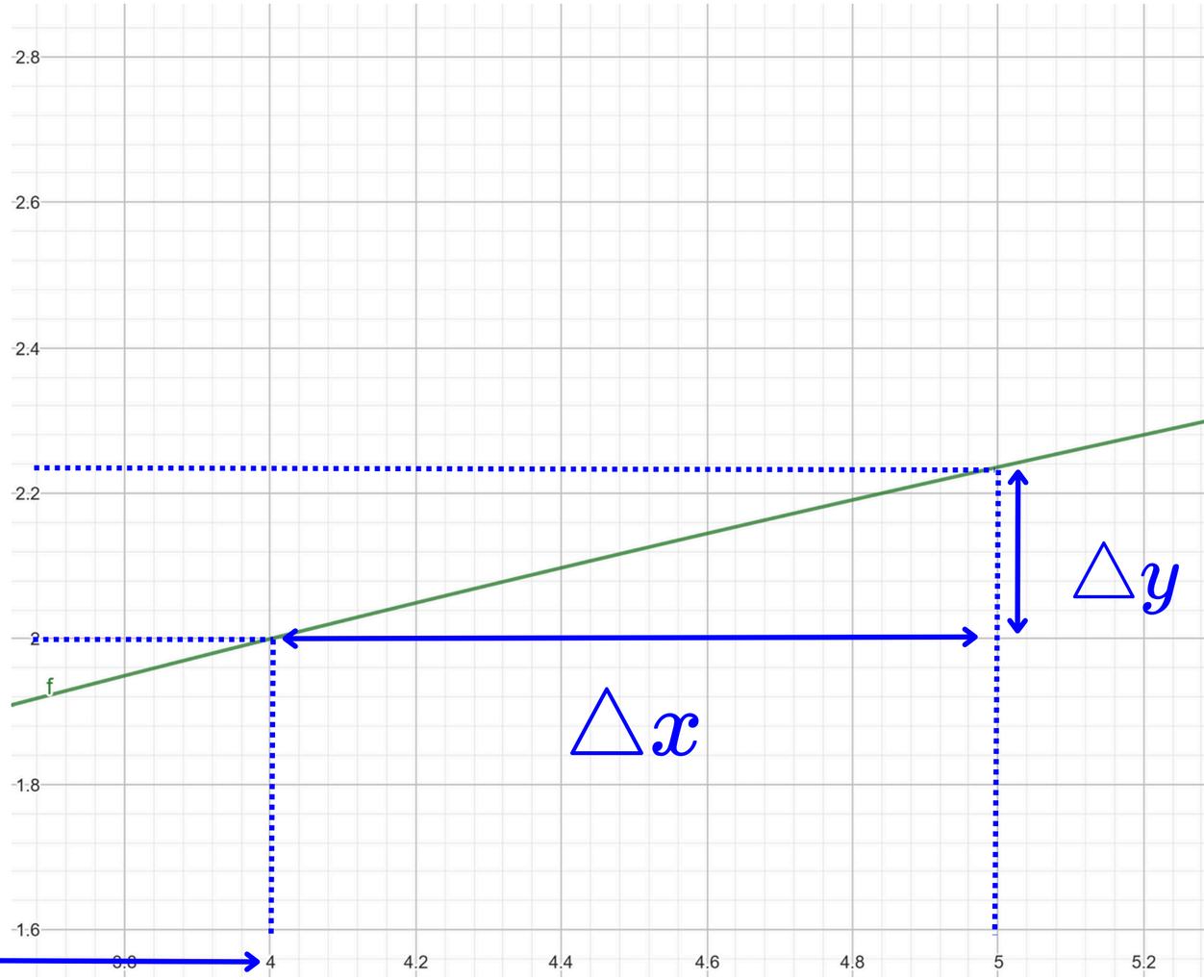

2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
−0.2
f
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
2.2
2.4
2.6
2.8
3
3.2
3.4
3.6



We start at the point a = 4

In this case, the we chose a change in x-axis equal to 1.

$$\triangle x = 5 - 4 = 1$$



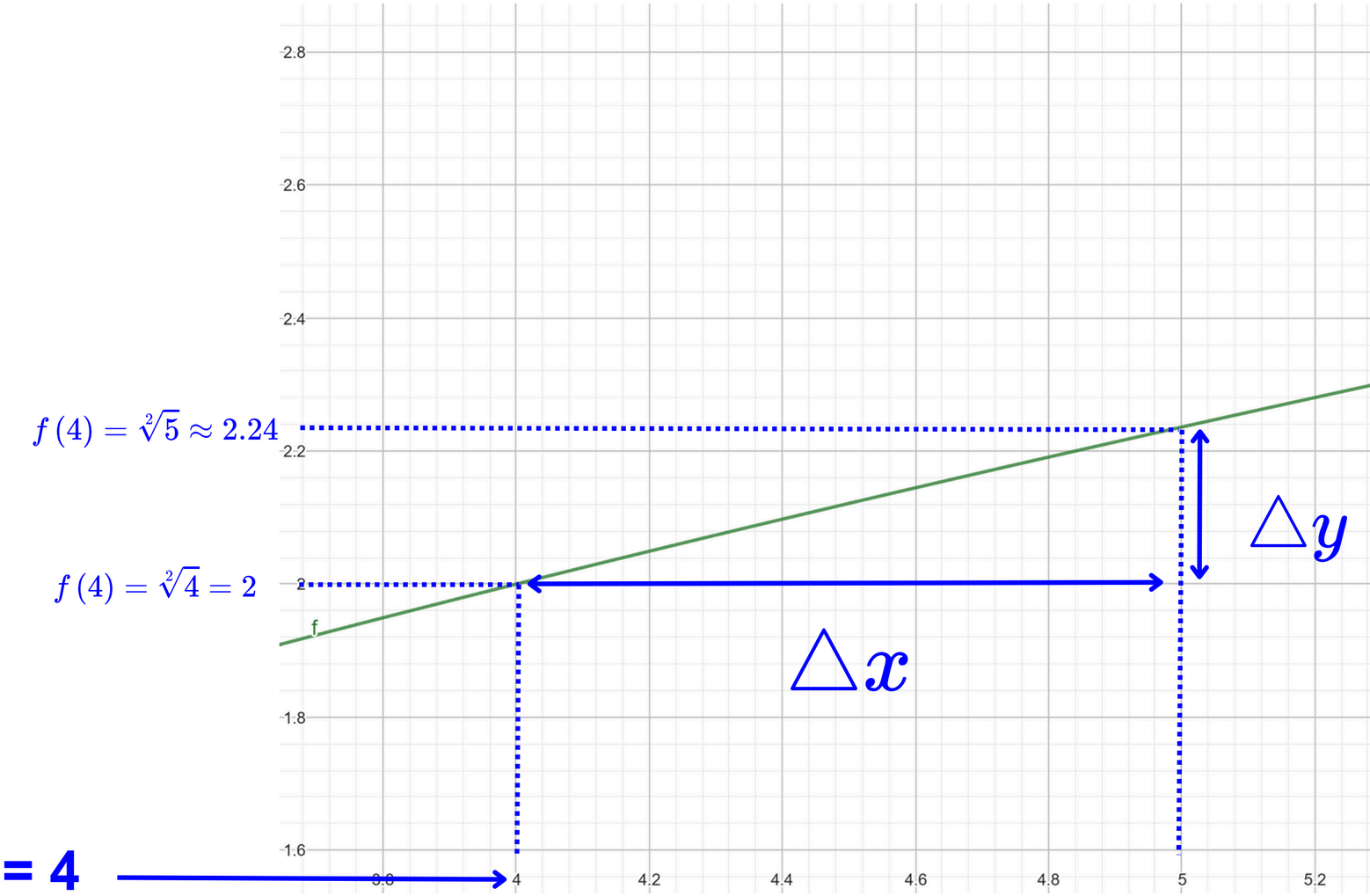
This is the point a = 4

In this case, the we chose a change on x-axis equal to 1.

$$\triangle x = 5 - 4 = 1$$

Which resulted in a change on y-axis equal to approximately 0.24.

$$\triangle y = \sqrt[2]{5} - \sqrt[2]{4} \approx 0.24$$



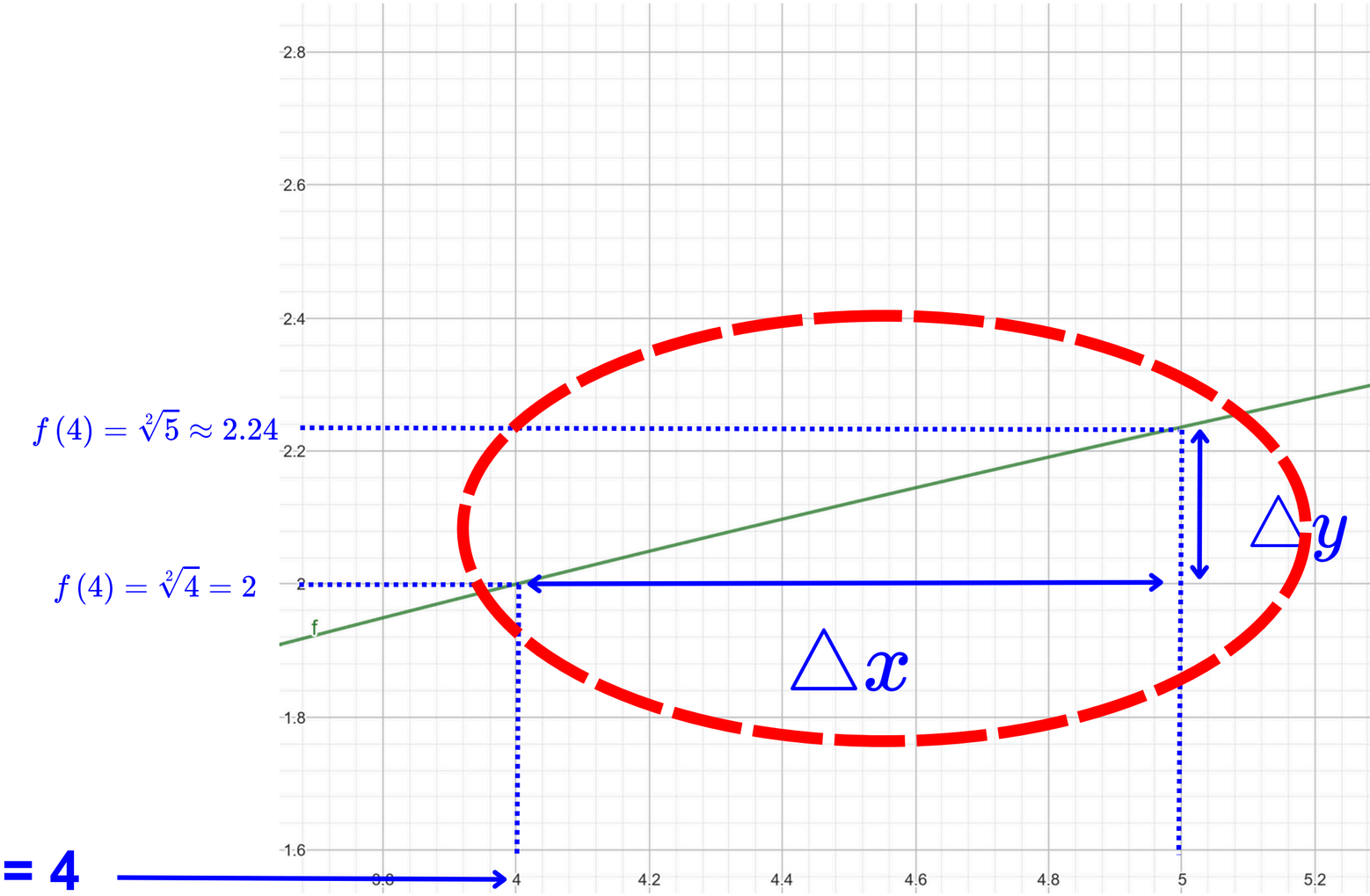
This is the point a = 4

In this case, the we chose a change on x-axis equal to 1.

$$\triangle x = 5 - 4 = 1$$

Which resulted in a change on y-axis equal to approximately 0.24.

$$\triangle y = \sqrt[2]{5} - \sqrt[2]{4} \approx 0.24$$

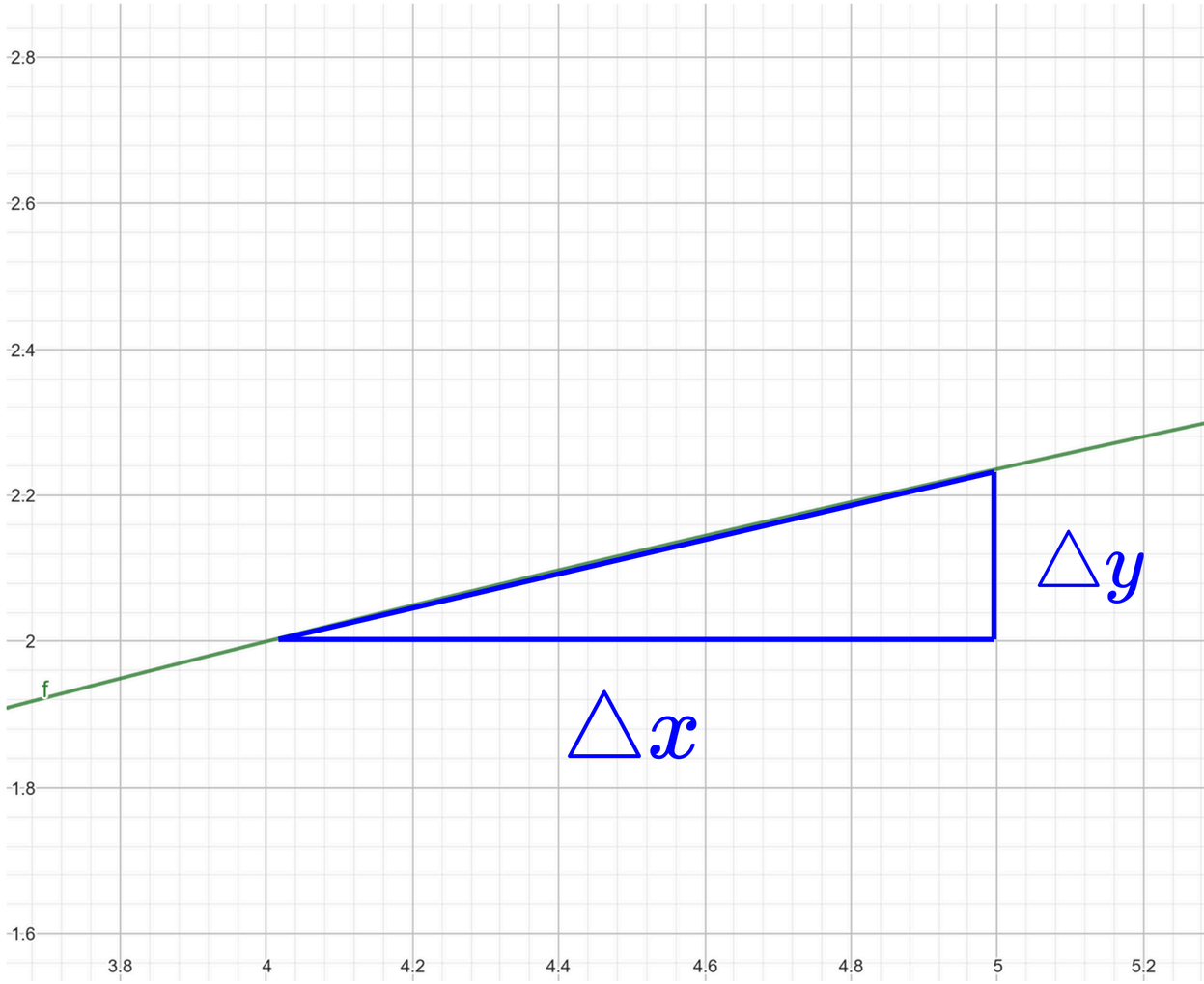


This is the point a = 4

$$\triangle y \approx 0.24$$

Here, with $\triangle x = 1$ the rate of change of the function is:

$$\frac{\triangle y}{\triangle x} \approx \frac{0.24}{1} \approx 0.24$$

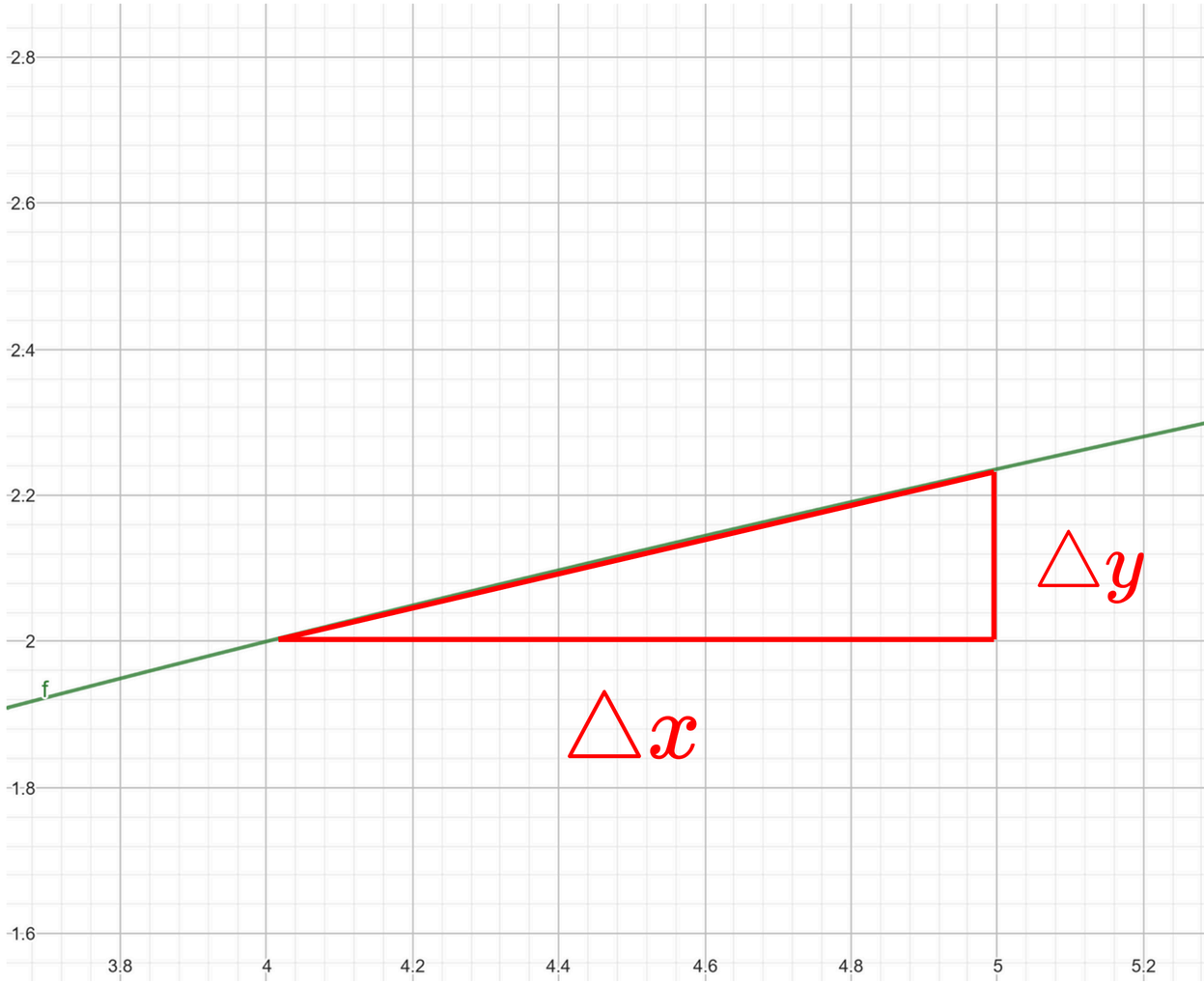


$$\triangle y \approx 0.24$$

Here, with $\triangle x = 1$ the rate of change of the function is:

$$\frac{\triangle y}{\triangle x} \approx \frac{0.24}{1} \approx 0.24$$

Our function increased by 0.24



Derivatives

Okay, $\triangle x = 1$ is a small change on the x-axis, but to calculate the rate of change of the function, $\triangle x$ should be a lot smaller that this.

$\triangle x$ should be consistent and helpful regardless of how the function behaves.

Derivatives

For example, the cosine function oscillates significantly over small intervals, so if $\triangle x$ is as large as 1, it won't accurately capture the true rate of change at a specific point. By making $\triangle x$ very small, we get a much finer, more precise measure of how the cosine function changes locally.

Derivatives

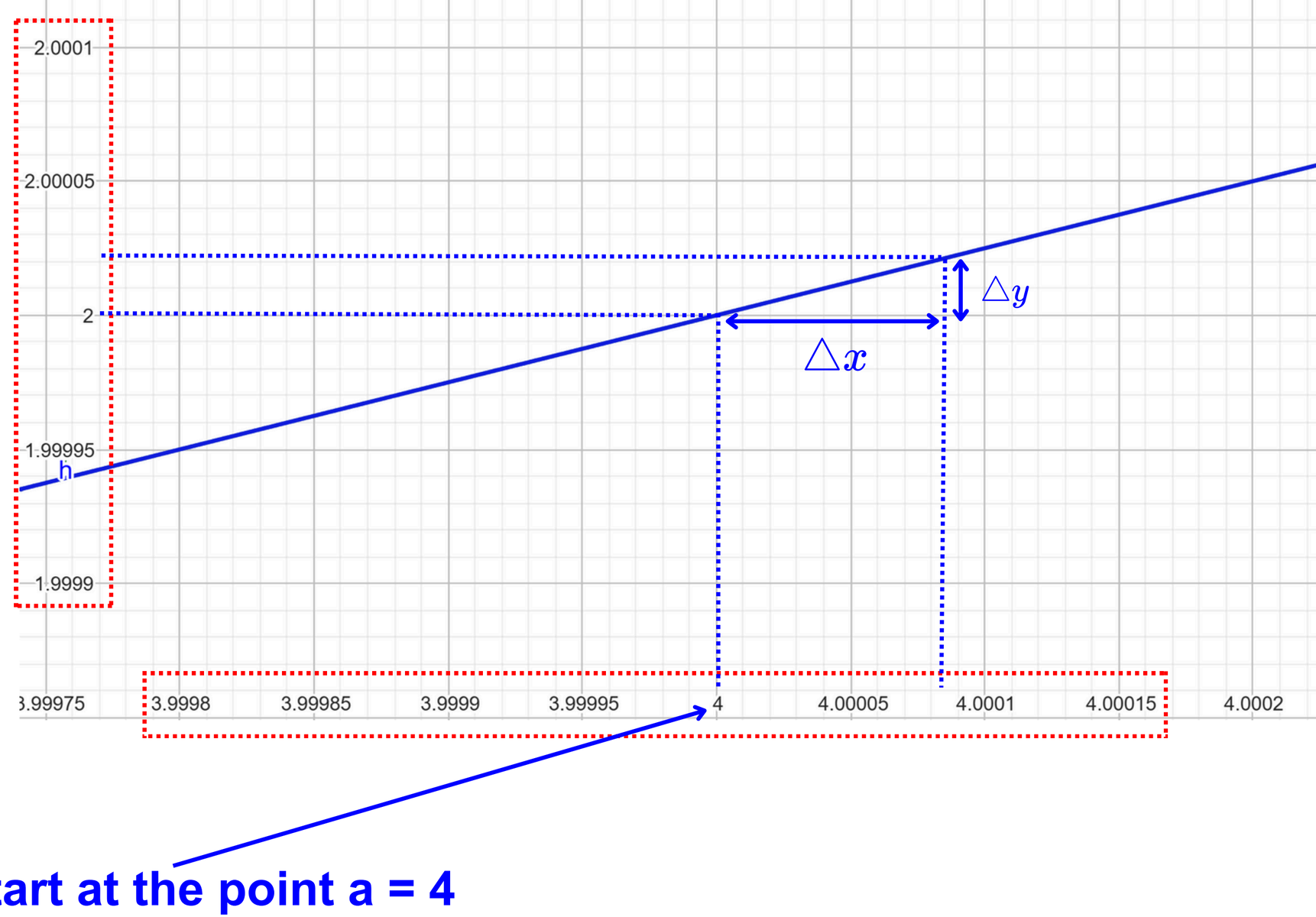
We must take an infinitely small interval, $\triangle x$, approaching zero. By applying limits, we can precisely evaluate this rate of change, capturing the behavior of the function as $\triangle x$ becomes exceedingly small.

In this case, the we chose an even smaller change on x-axis.

$$\triangle x \approx 0.00008$$

Which resulted in a change on y-axis

$$\triangle y \approx 0.000021$$

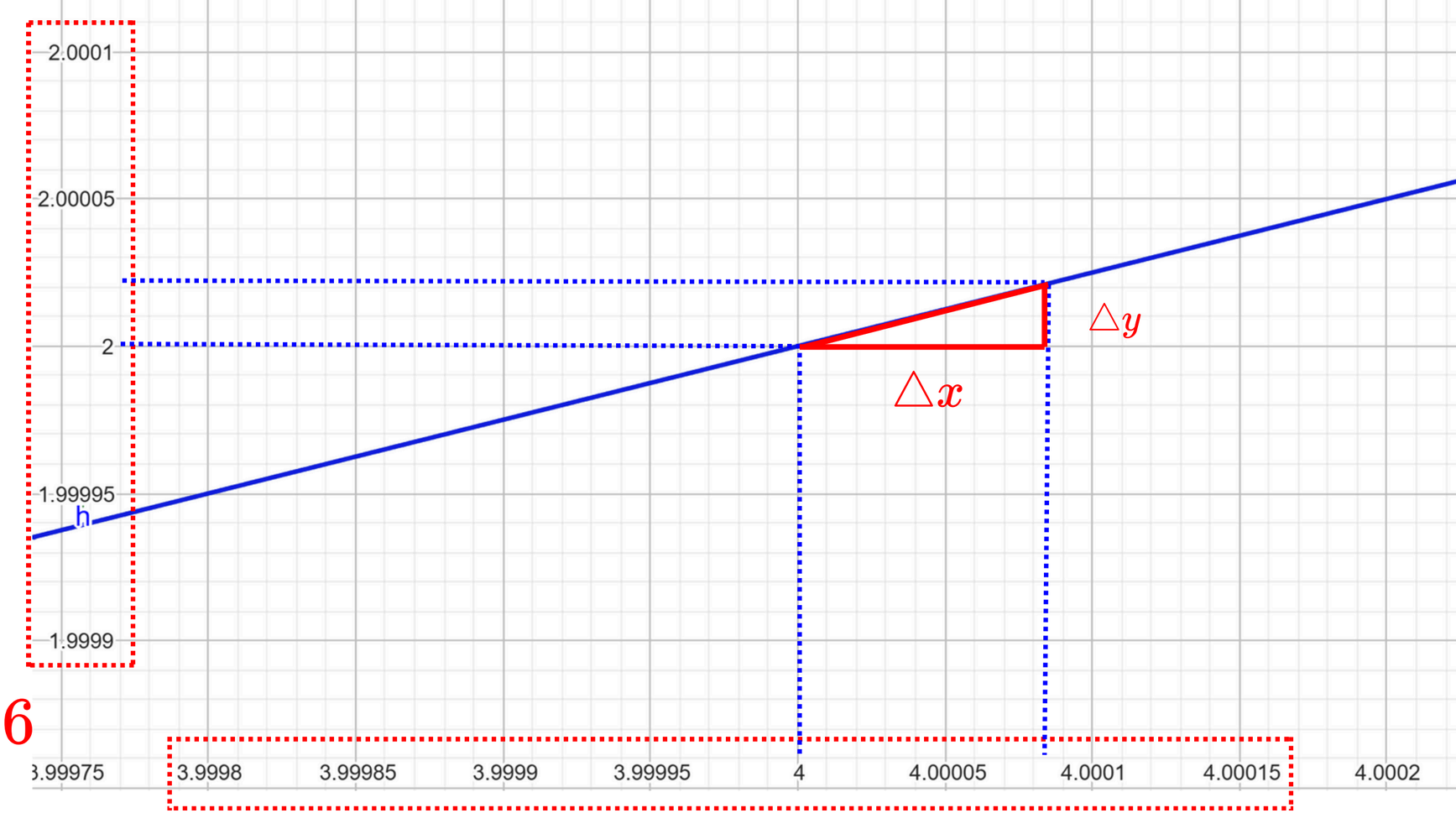


We start at the point a = 4

$$\triangle x \approx 0.00008$$

$$\triangle y \approx 0.000021$$

$$\frac{\triangle y}{\triangle x} \approx \frac{0.000021}{0.00008} \approx 0.26$$



With change in x-axis equal to 0.00008 our function increased by 0.26

Derivatives

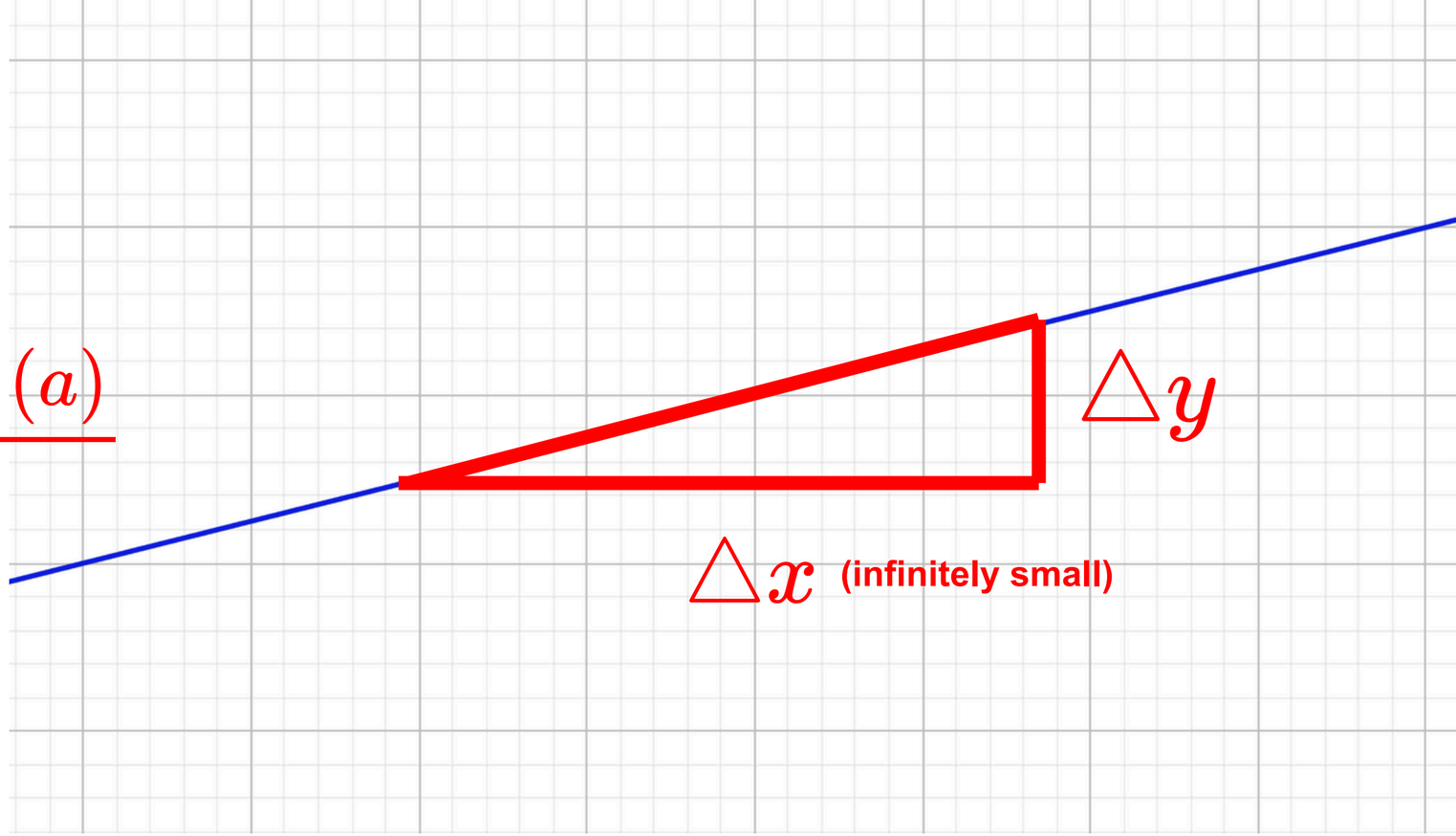
As we mentioned before, to measure the exact rate of change of the function, $\triangle x$ should be so small that we actually take $\triangle x \to 0$

Derivatives

and

$$\textbf{The rate of change} = \lim_{\triangle x \to 0} \frac{f(a + \triangle x) - f(a)}{\triangle x}$$

$$\frac{\triangle y}{\triangle x} = \lim_{\triangle x \to 0} \frac{f(a + \triangle x) - f(a)}{\triangle x}$$

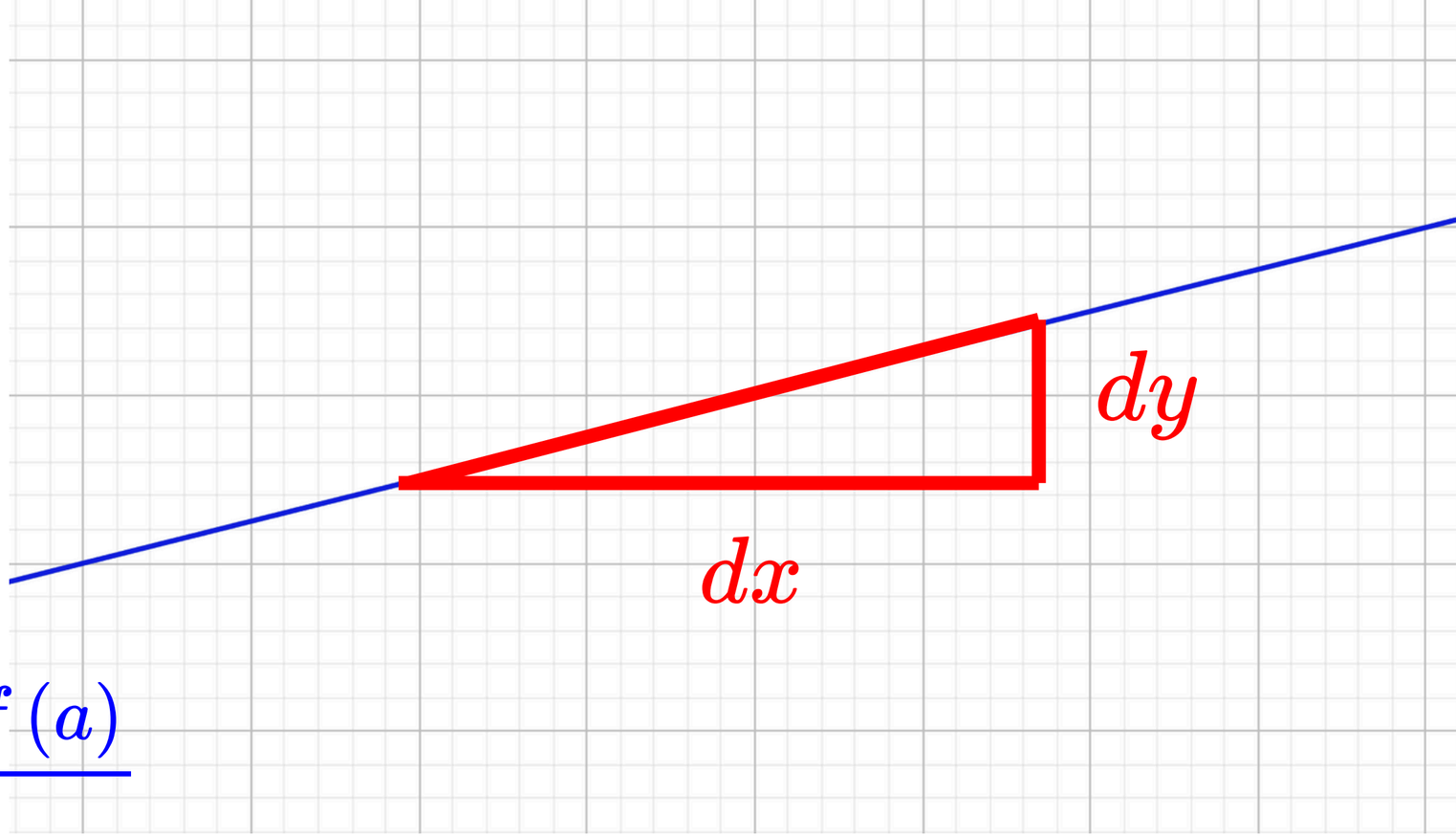


The increase in our function as $\triangle x \to 0$ **is** $\frac{\triangle y}{\triangle x} = f'(a)$

Note: when $\triangle x \to 0$

$\triangle x$ becomes dx

$\triangle y$ becomes dy



$$\frac{dy}{dx} = \lim_{\triangle x \to 0} \frac{f(a + \triangle x) - f(a)}{\triangle x}$$

The increase in our function as $\triangle x \to 0$ **is** $\frac{\triangle y}{\triangle x} = \frac{dy}{dx} = f'(a)$

Derivatives

If f is a real-valued function and ‘a’ is any point in its domain for which f is defined then f(x) is said to be differentiable at the point x=a if the derivative f'(a) exists at every point in its domain. It is given by

$$f'(a) = \lim_{h \to 0} \frac{f(a+h) - f(a)}{h}$$

Derivatives

$$f'(a) = \lim_{h \to 0} \frac{f(a+h) - f(a)}{h}$$

$$f'(4) = \lim_{h \to 0} \frac{\sqrt[2]{4+h} - \sqrt[2]{4}}{h} = \frac{1}{4}$$

Derivatives

In general:

$$f'(x) = \lim_{h \to 0} \frac{\sqrt[2]{x+h} - \sqrt[2]{x}}{h} = \frac{1}{2\sqrt[2]{x}}$$

Partial Derivatives

Partial derivatives are a way to measure how a multivariable function changes with respect to one variable at a time, while keeping the other variables constant. Imagine a function $f(x, y)$ that depends on both x and y the partial derivative with respect to x, written as $\frac{\partial f}{\partial x}$ tells us how f changes as x changes, with y held fixed.

Partial Derivatives

Take, for instance, $z = f(x, y)$. Let us imagine that we can fix y at some value, say $y = y_0$. Then, the derivative at $f(x, y_0)$ in x is

$$\frac{d}{dx} f(x, y_0) .$$

Partial Derivatives

In other words, treat y as a constant. Similarly, we could fix x and take a derivative in y. We define the partial derivatives as follows

$$f_x(x,y) = \lim_{\Delta x \to 0} \frac{f(x+\Delta x, y) - f(x,y)}{\Delta x},$$

$$f_y(x,y) = \lim_{\Delta y \to 0} \frac{f(x, y+\Delta y) - f(x,y)}{\Delta y}.$$

Partial Derivatives

The notation $f_x(x, y)$ means the partial derivative of $z = f(x, y)$ with respect to x. Other notations are

$$\frac{\partial f}{\partial x}, \quad \frac{\partial z}{\partial x}.$$

Often, we will want to find the partial derivative at a given point, say (x_0, y_0). To do this, find the partial derivative and then substitute the values of the point (x_0, y_0). This will be denoted

$$\left.\frac{\partial f}{\partial y}\right|_{x=x_0, y=y_0}, \quad \left.\frac{\partial f}{\partial y}\right|_{(x_0, y_0)}, \quad \frac{\partial f}{\partial y}(x_0, y_0).$$

Partial Derivatives

Let us look at an example.

Example: Let $z = x^2 \sin y$, and find $\frac{\partial z}{\partial x}\Big|_{(\pi,\pi)}$ and $\frac{\partial z}{\partial y}\Big|_{(\pi,\pi)}$.

Solution:

$$\frac{\partial z}{\partial x} = 2x \sin y$$

$$\Rightarrow \frac{\partial z}{\partial x}\bigg|_{(\pi,\pi)} = 2\pi \sin \pi = 0\,.$$

Similarly,

$$\frac{\partial z}{\partial y} = x^2 \cos y$$

$$\Rightarrow \quad \frac{\partial z}{\partial x}\bigg|_{(\pi,\pi)} = \pi^2 \cos \pi = -\pi^2\,.$$

Partial Derivatives

Slope:

You probably recall that the slope of a function is given by its derivative, slope $= \frac{df}{dx}$. If a function has three variables, i.e three independent directions, it has three slopes. Therefore the function $f(x, y, z)$ has slope $\frac{df}{dx}$ in the x-direction, slope $\frac{df}{dy}$ in the y-direction, and slope $\frac{df}{dz}$ in the z-direction. We will revisit this later when we discuss gradient.

Partial Derivatives

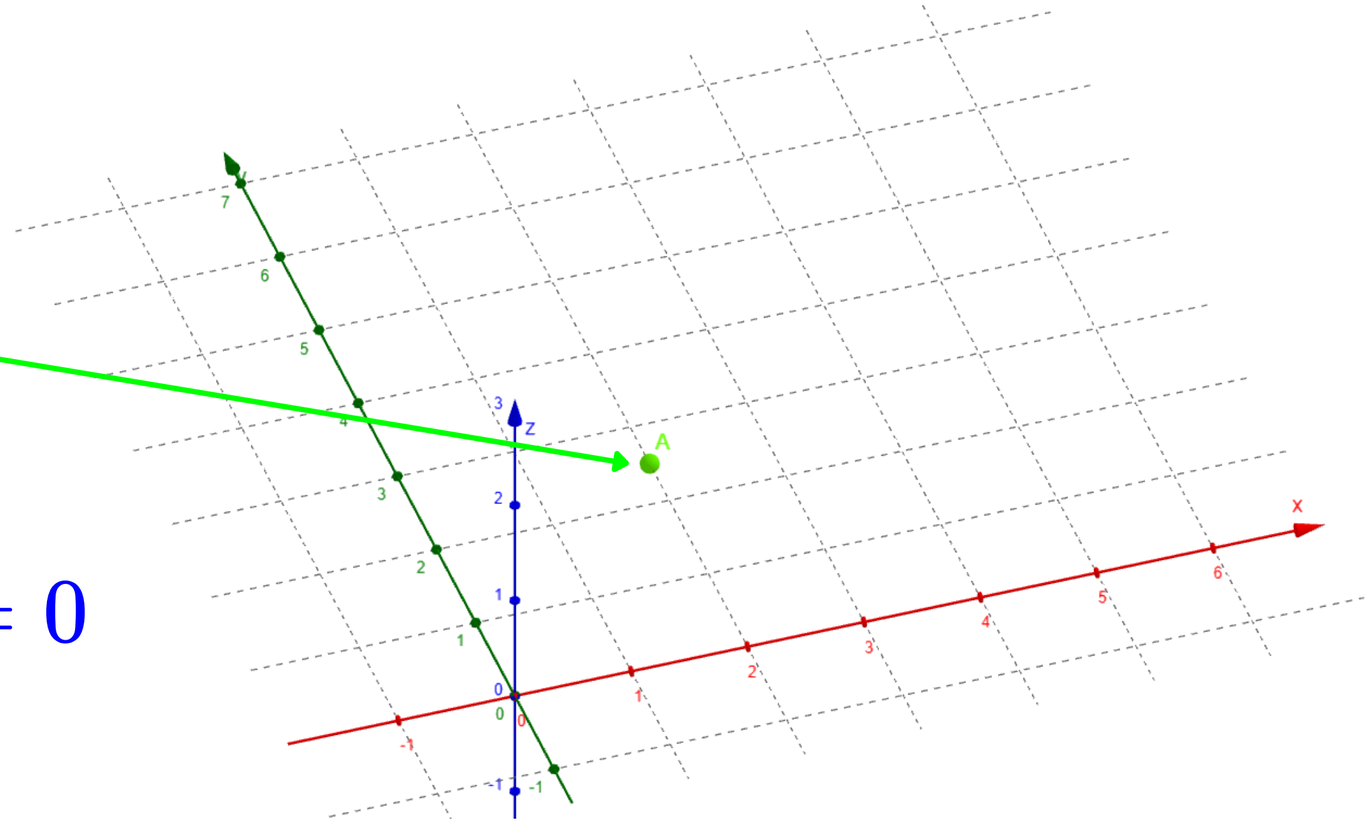
We define the point A on the plane:

$$A = (2, 2.5, 0)$$

$x = 2$

$y = 2.5$

$z = 0$



Partial Derivatives

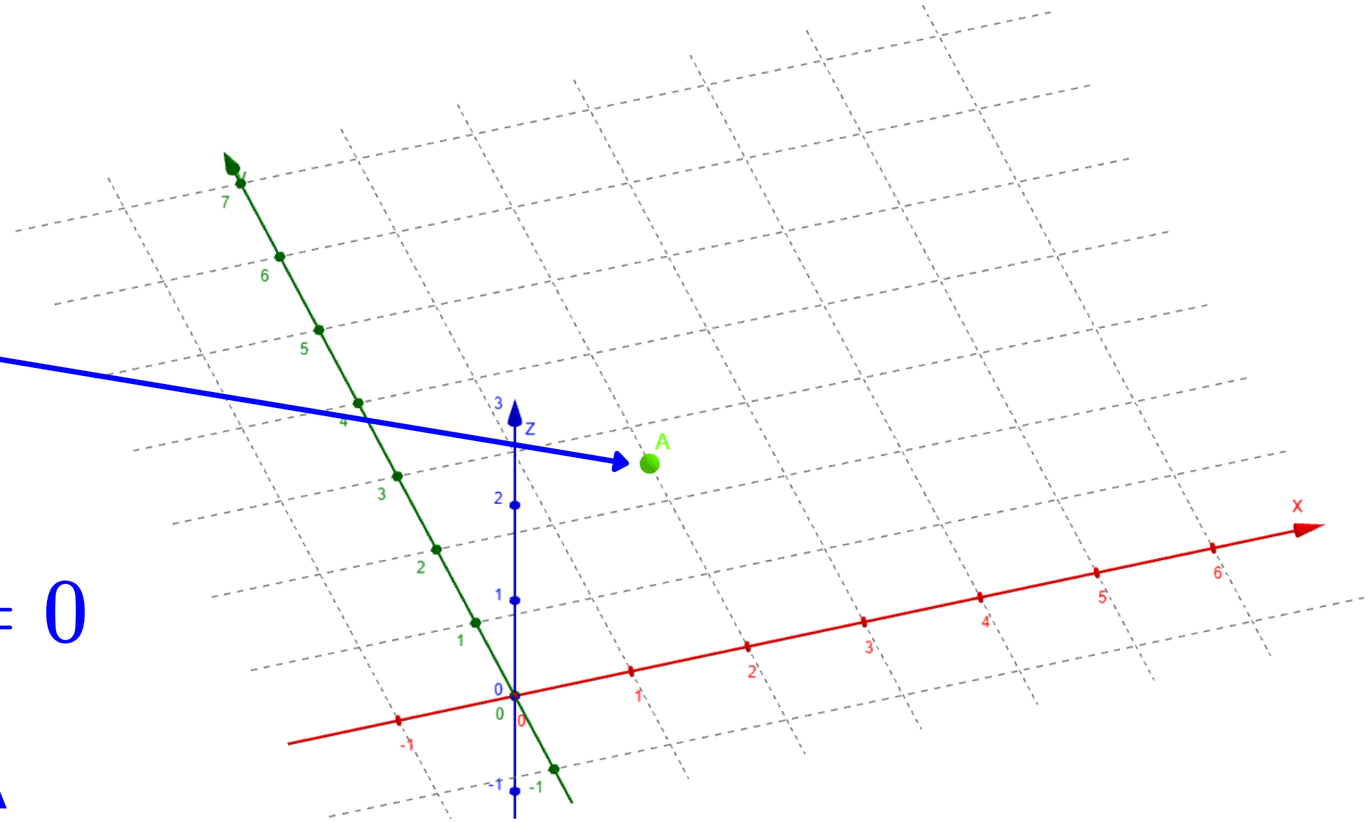
We define the point A on the plane:

$$A = (2, 2.5, 0)$$

$x = 2$

$y = 2.5$

$z = 0$



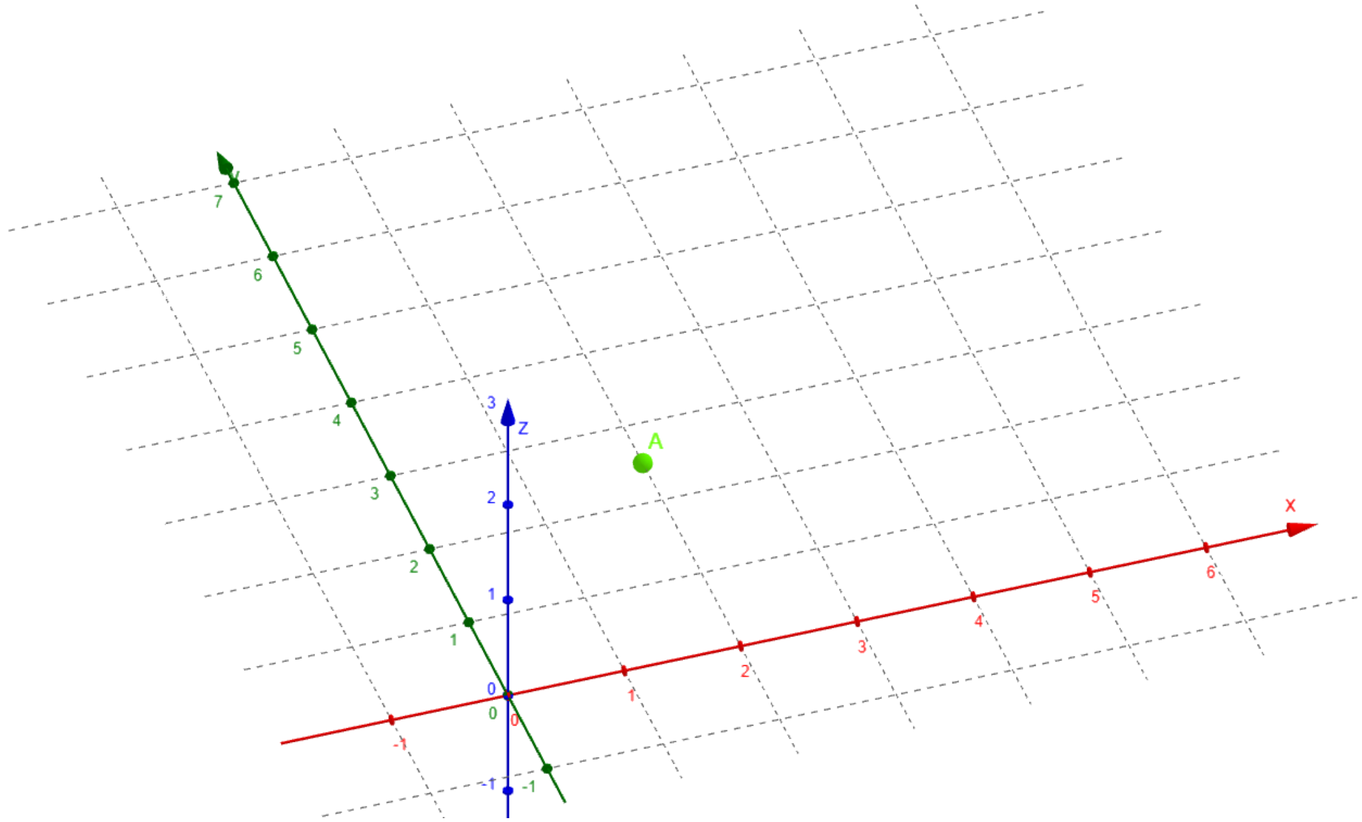
Coordinates of point A

Partial Derivatives

$$A = (2, 2.5, 0)$$

We also define our two-variable function:

$$f\,(x, y) = sin\,(x) \times cos\,(y)$$

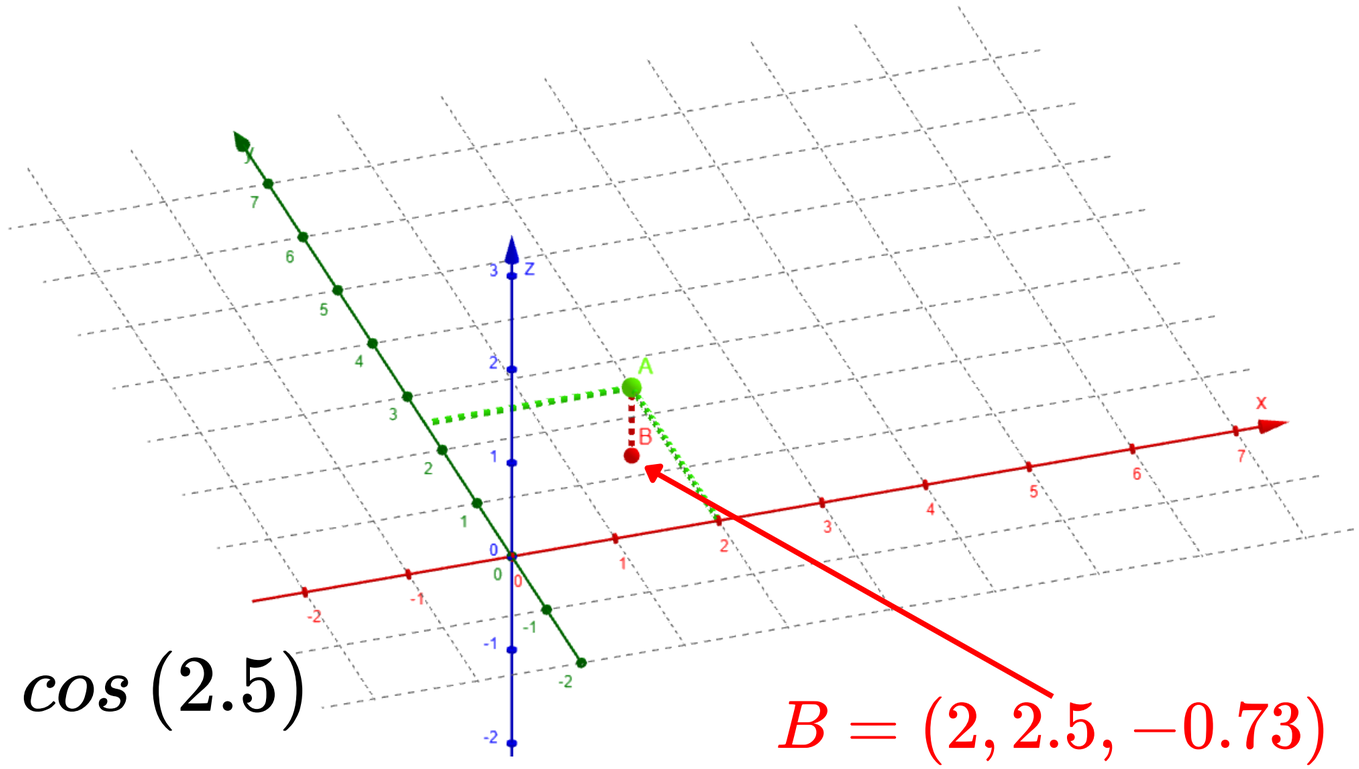


Partial Derivatives

$$A_0 = (2, 2.5)$$

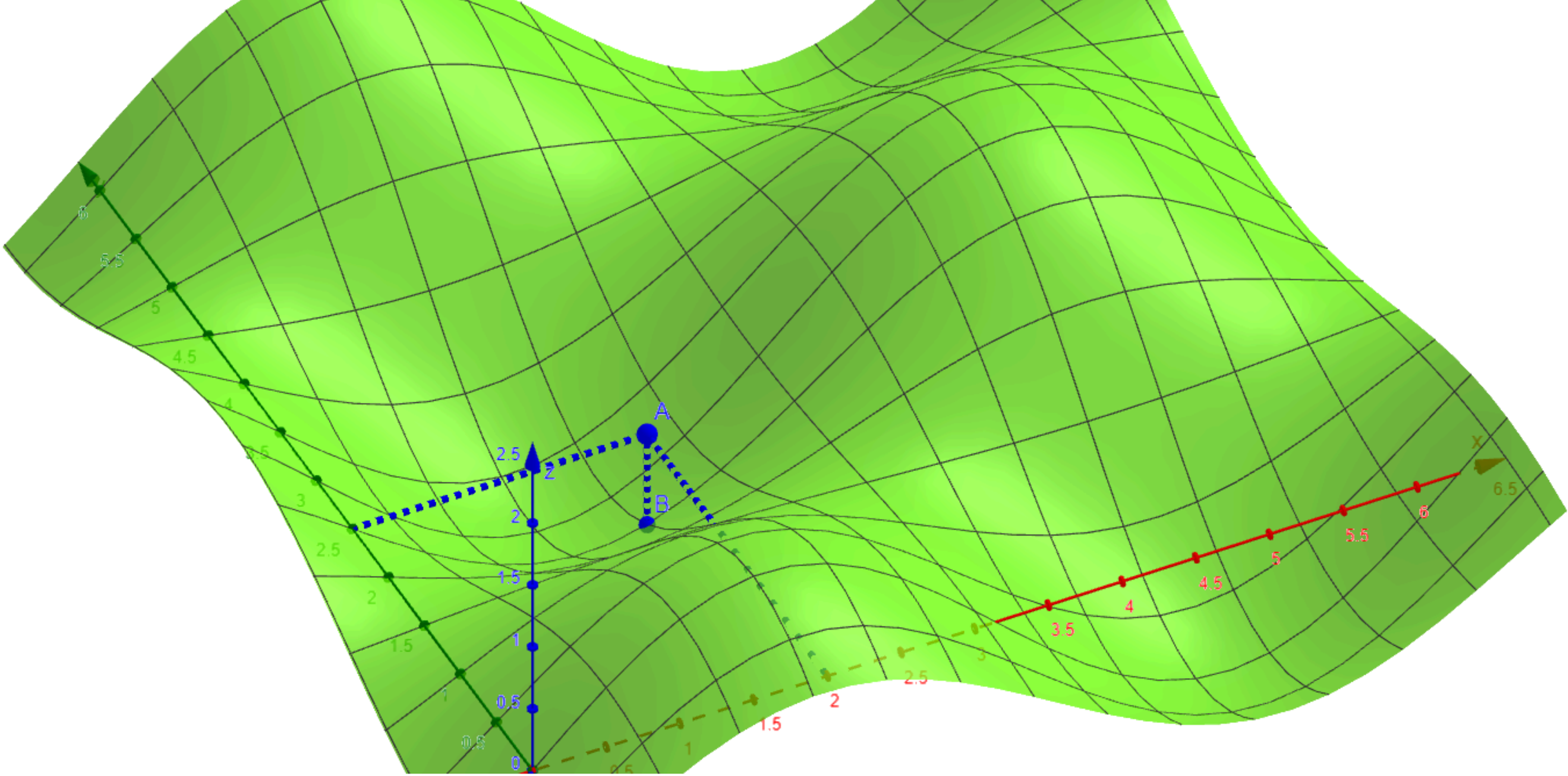
$$f\left(x, y\right) = sin\left(x\right) \times cos\left(y\right)$$

We calculate the image of A with the function f:



$$f\left(A_0\right) = f\left(2, 2.5\right) = sin\left(2\right) \times cos\left(2.5\right)$$

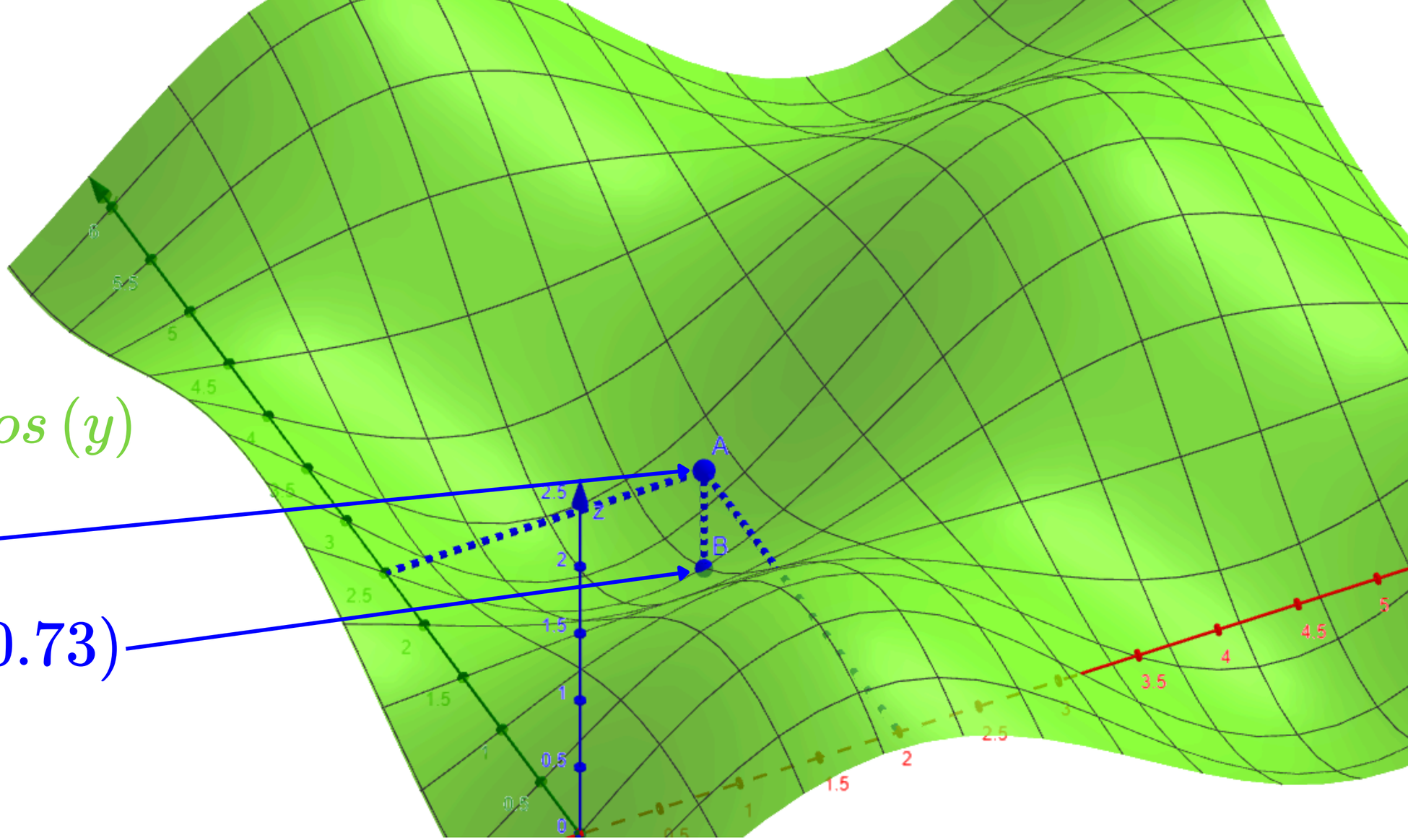
$$f\left(A_0\right) = z \approx -0.73$$

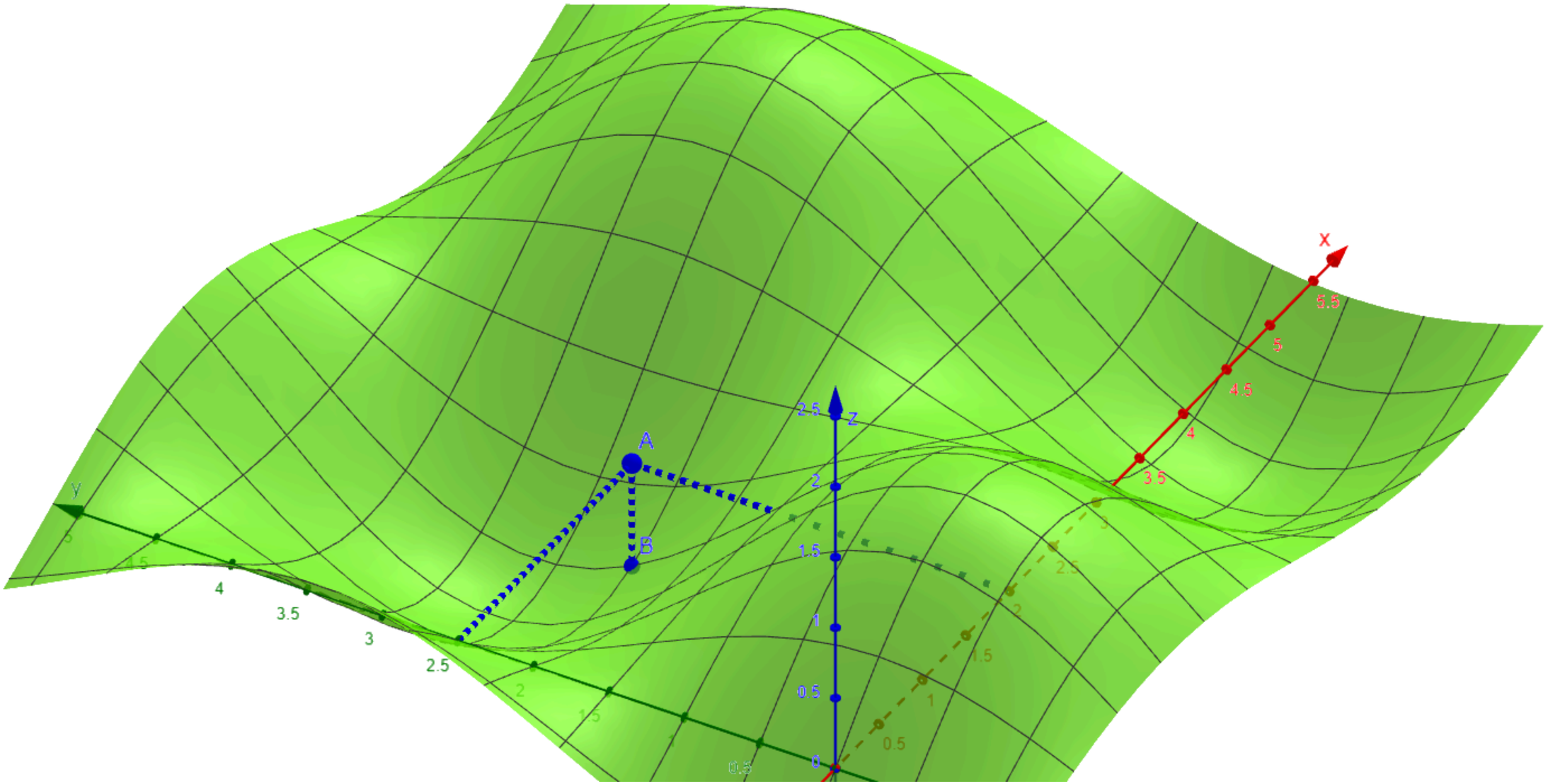

A
B
X
6
5.5
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
6.5

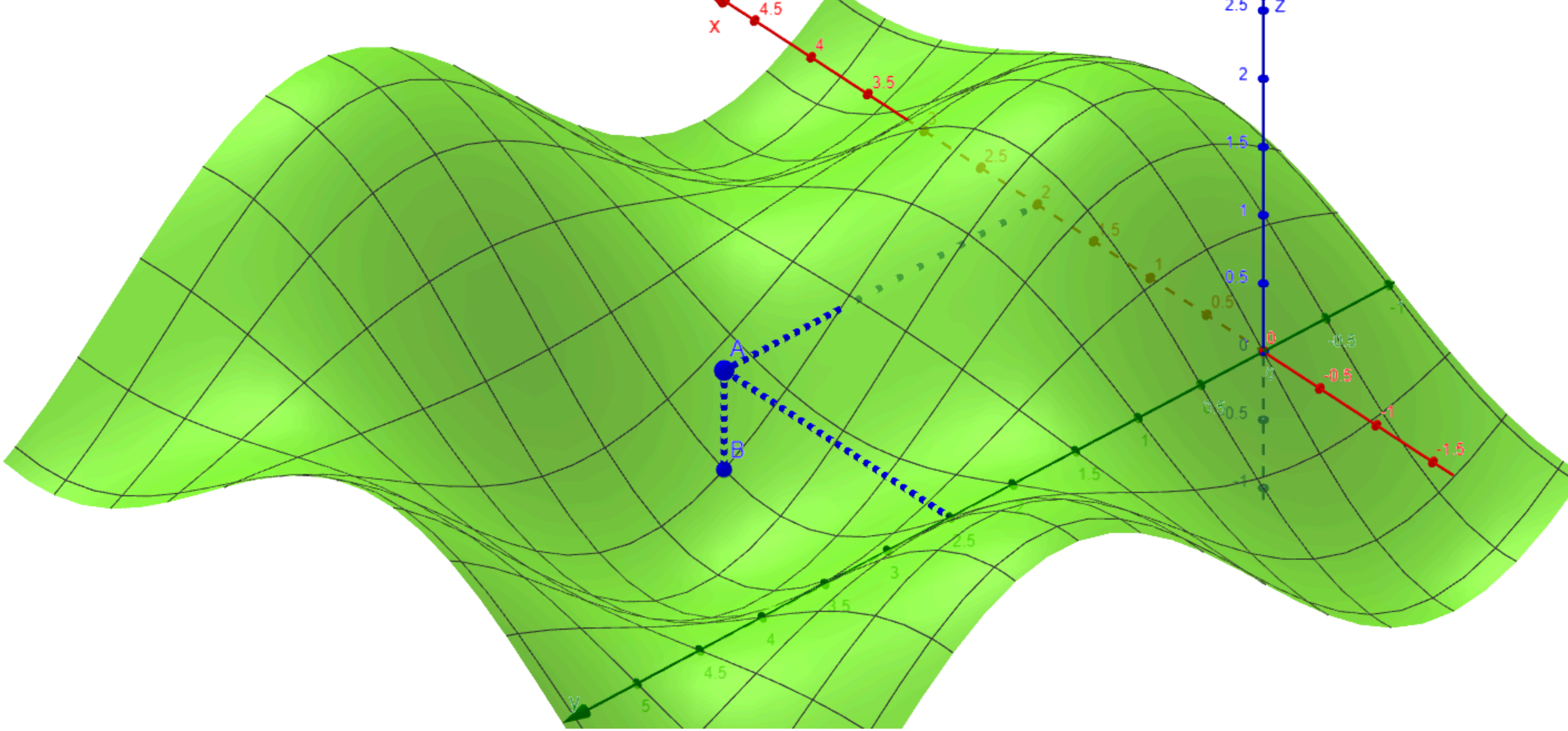
$$z = sin\,(x) \times cos\,(y)$$

$$A = (2, 2.5, 0)$$

$$B = (2, 2.5, -0.73)$$



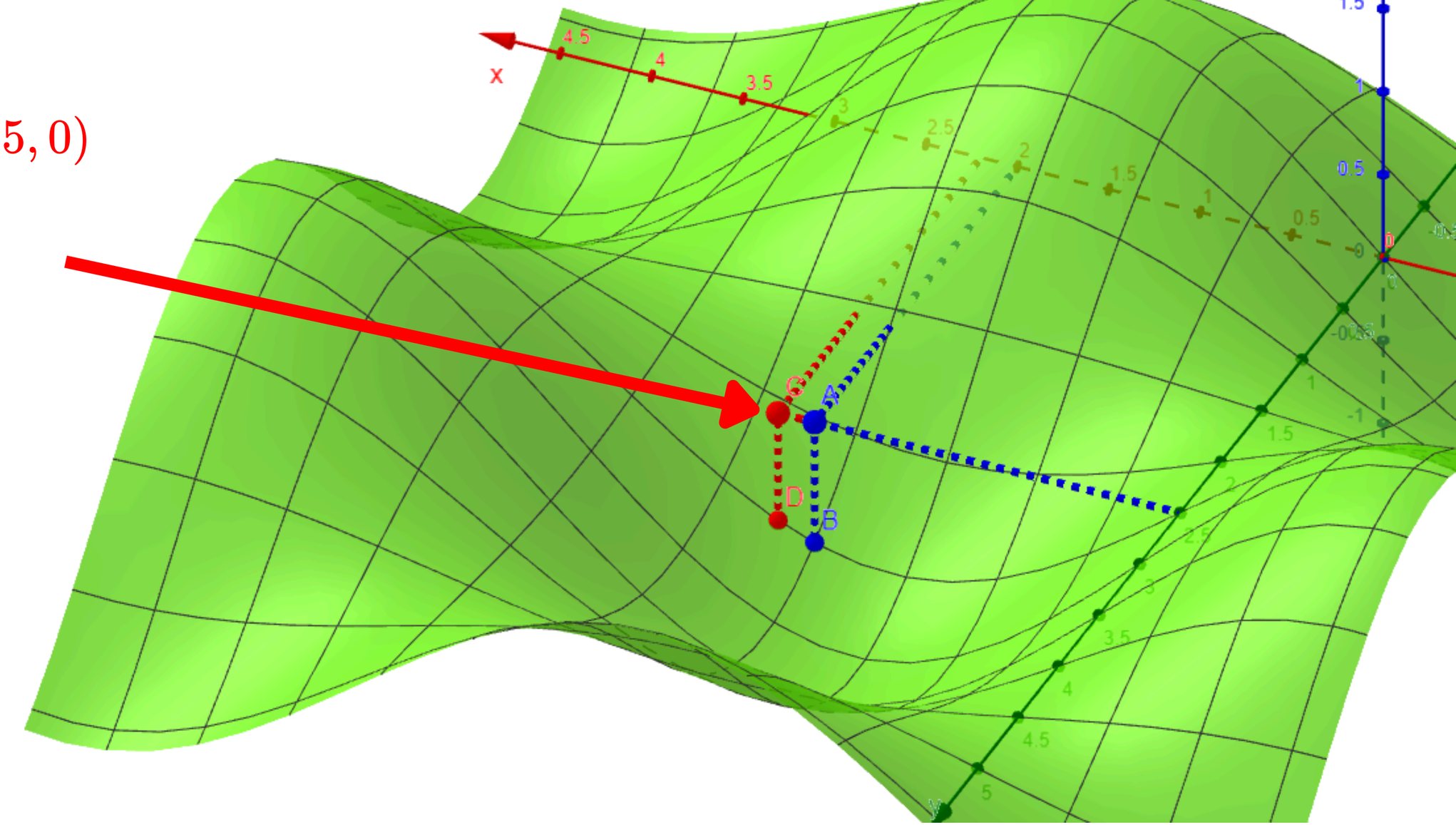

x
5.5
5
4.5
4
3.5
z
2.5
2
1.5
1
0.5
0
A
B
y
4
3.5
3
2.5
2
1.5
1
0.5


x
4.5
4
3.5
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
z
A
B
y
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5

$$C = (2 + \triangle x, 2.5, 0)$$

$$\triangle x = 0.2$$

$$C = (2.2, 2.5, 0)$$



$$C = (2 + \triangle x, 2.5, 0)$$

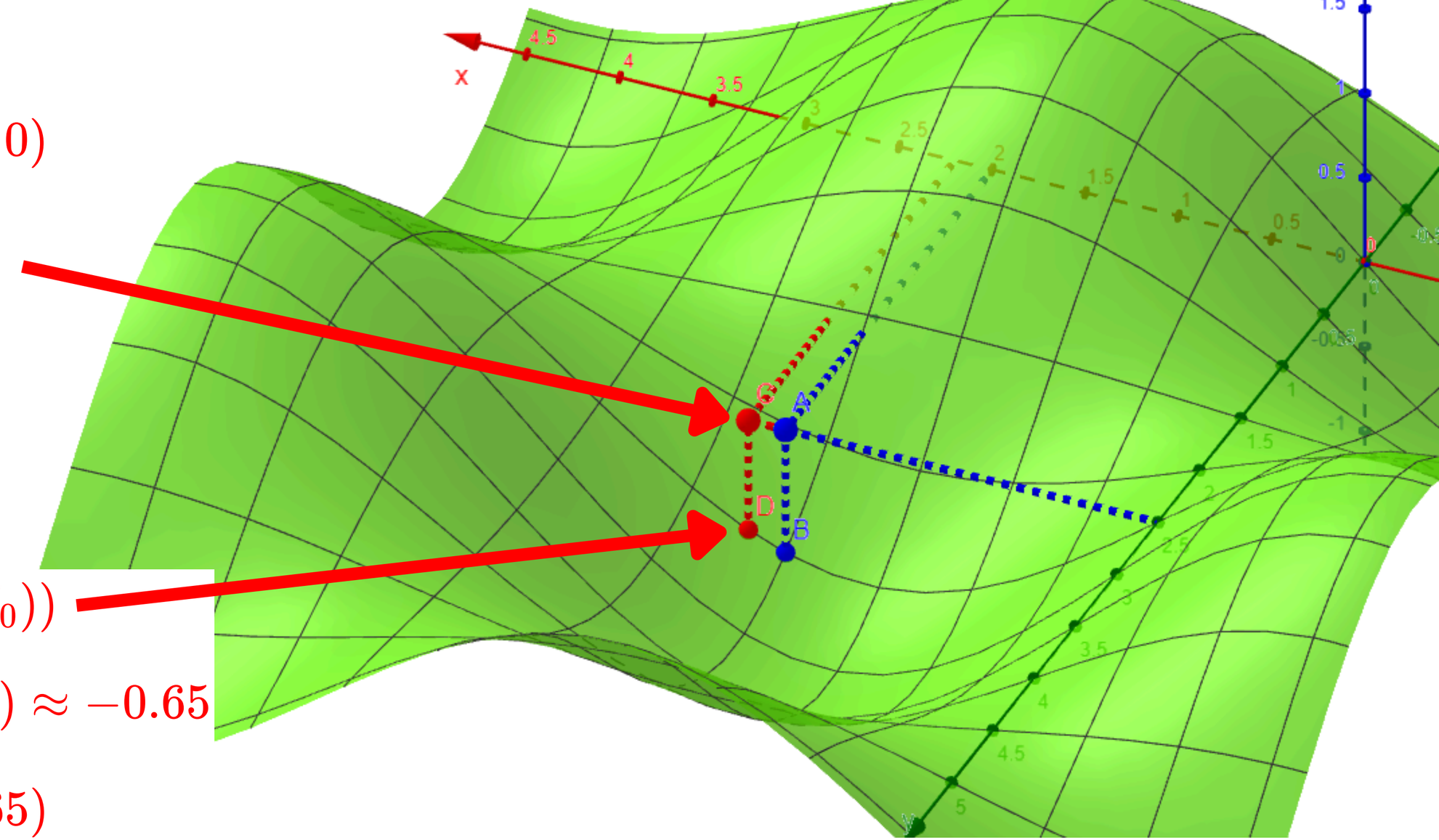
$$\triangle x = 0.2$$

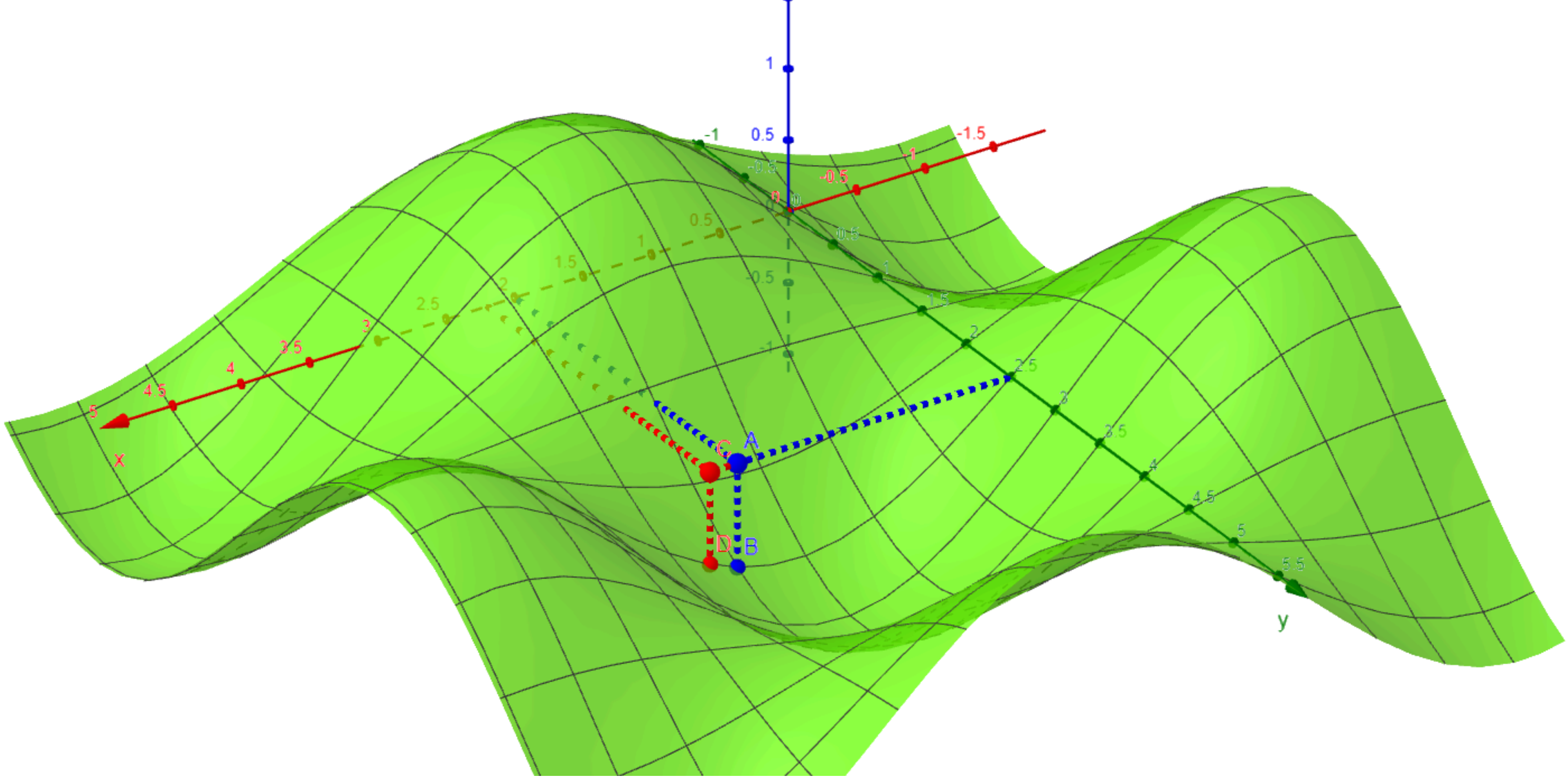
$$C = (2.2, 2.5, 0)$$

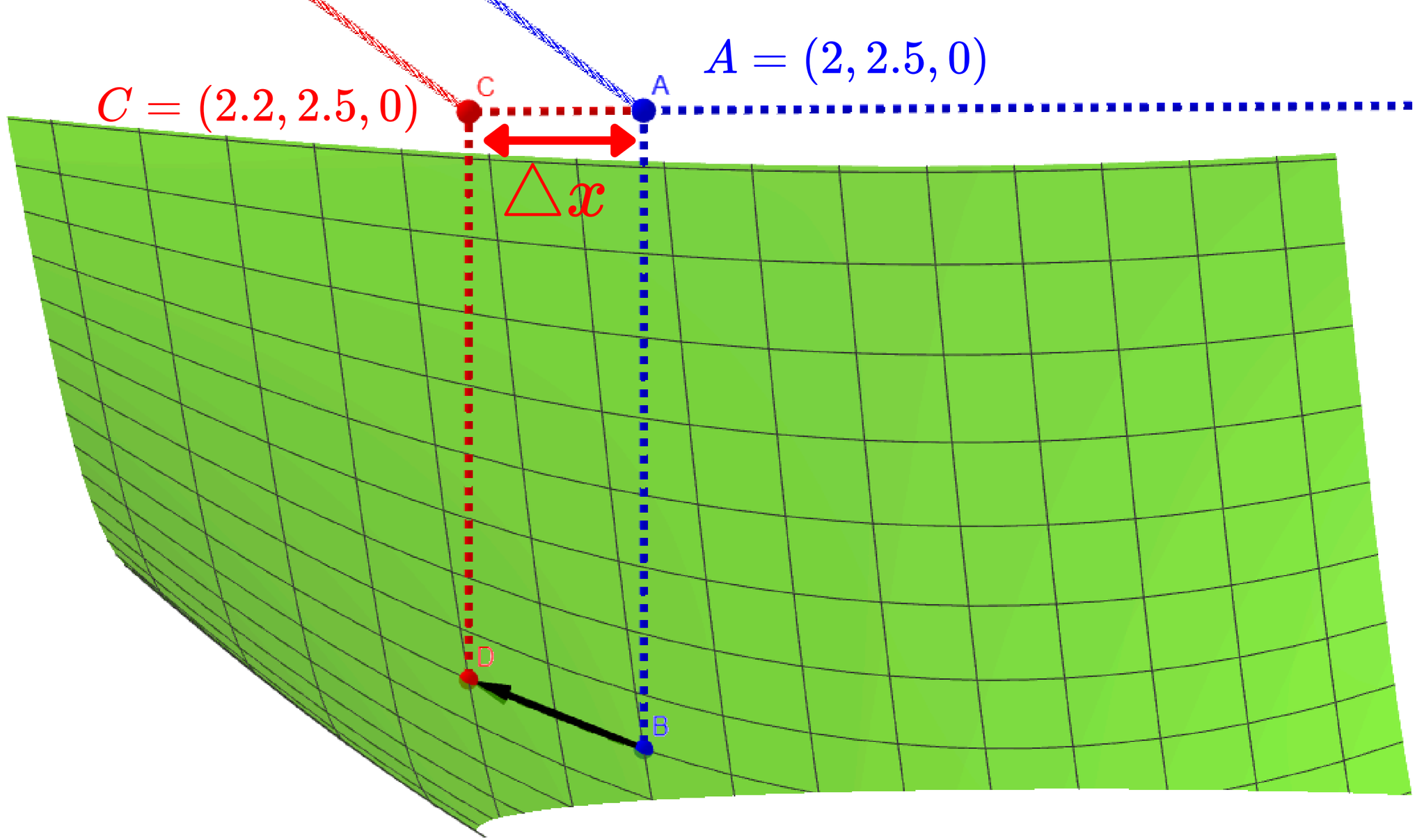
$$D = (2.2, 2.5, f(C_0))$$

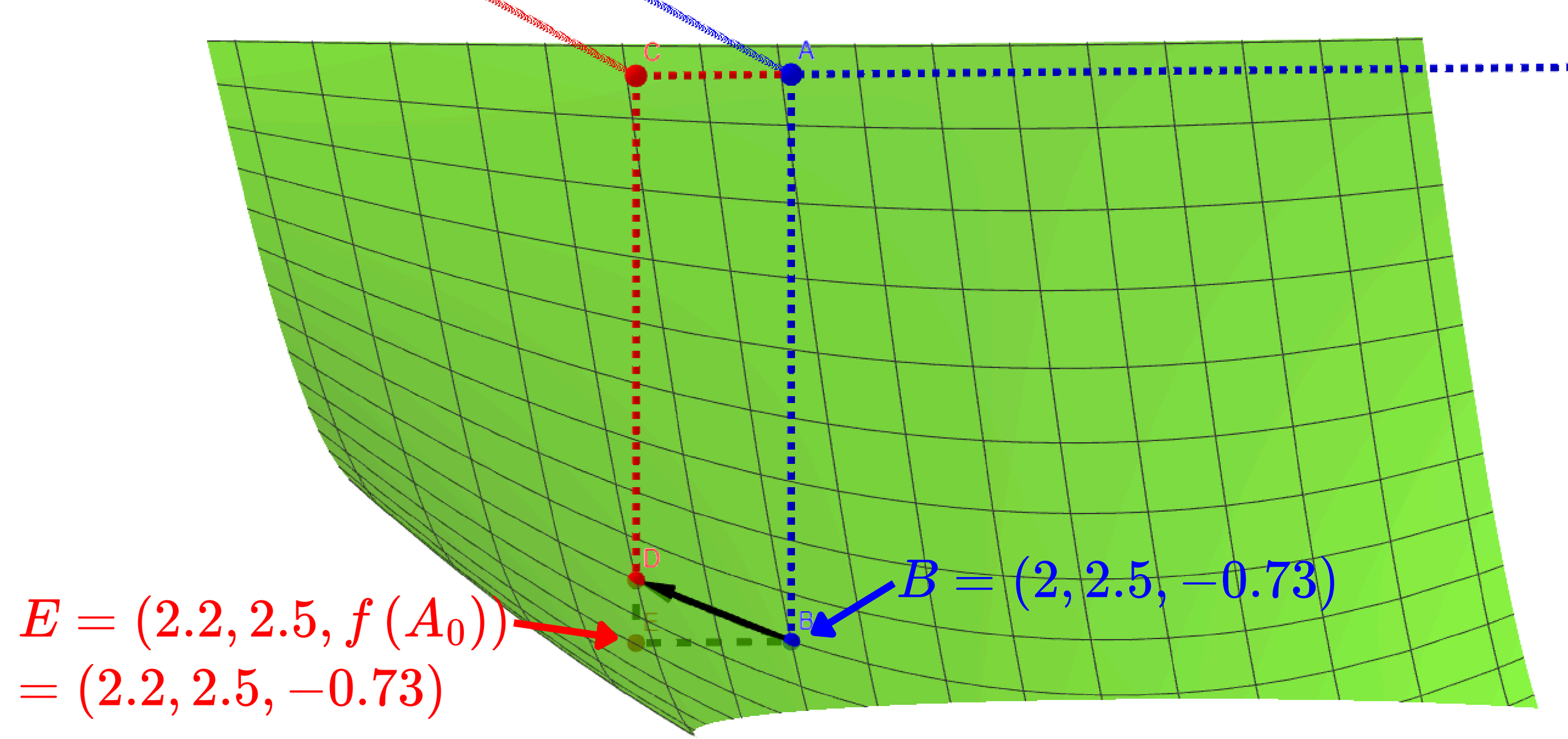
$$f(C_0) = f(2.2, 2.5) \approx -0.65$$

$$D = (2.2, 2.5, -0.65)$$



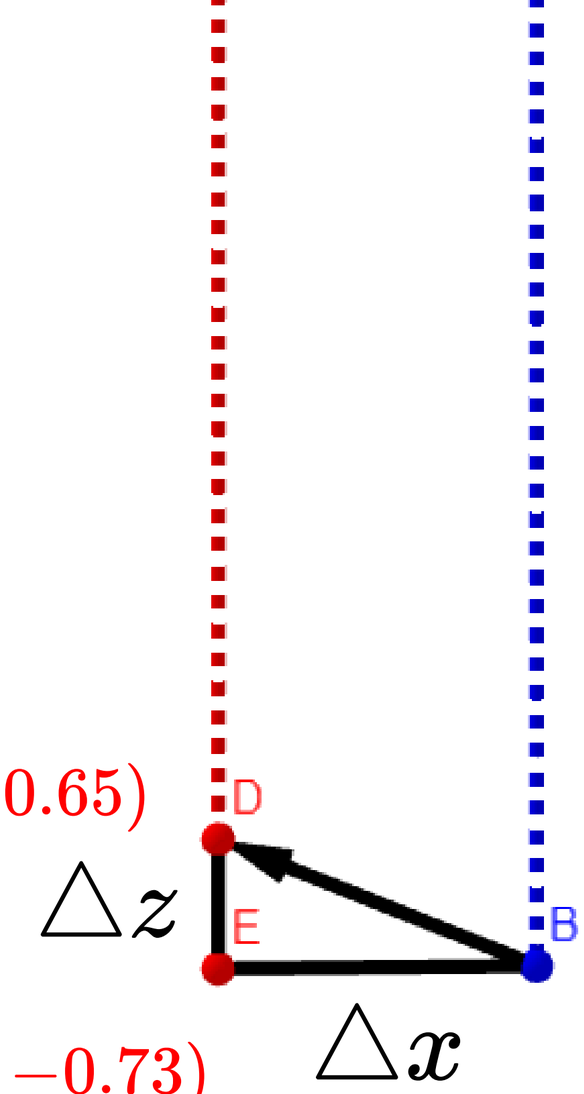

1
0.5
-1
-1.5
-0.5
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
5.5
x
y
A
B
C
D


$A = (2, 2.5, 0)$
$C = (2.2, 2.5, 0)$
C
A
$\triangle x$
D
B


C
A
D
E
B
$B = (2, 2.5, -0.73)$
$E = (2.2, 2.5, f(A_0))$
$= (2.2, 2.5, -0.73)$

$C = (2.2, 2.5, 0)$

$D = (2.2, 2.5, -0.65)$



$B = (2, 2.5, -0.73)$

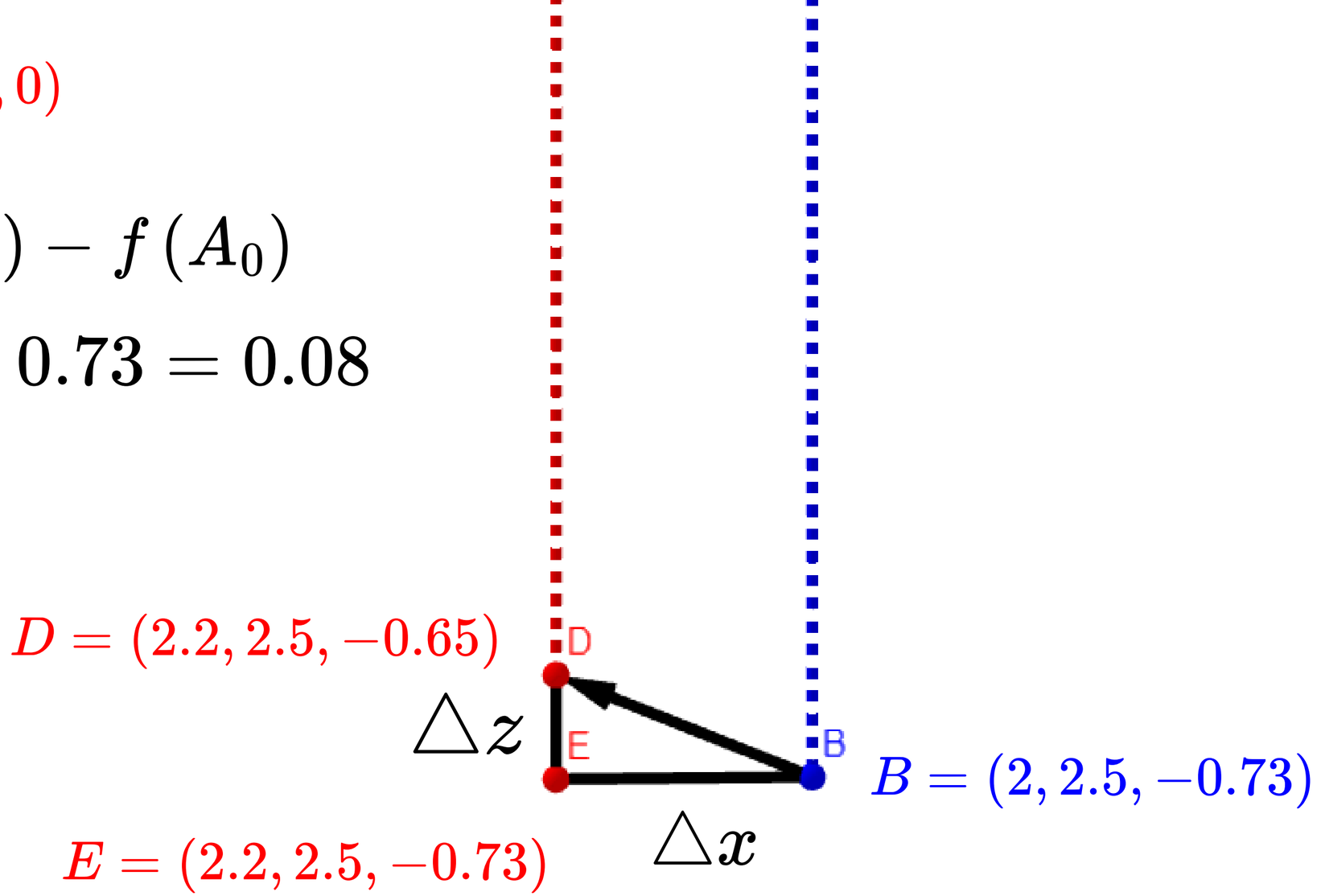
$E = (2.2, 2.5, -0.73)$

$C = (2.2, 2.5, 0)$

$$\triangle z = f(C_0) - f(A_0)$$

$$= -0.65 + 0.73 = 0.08$$

$$\triangle x = 0.2$$



$D = (2.2, 2.5, -0.65)$

$B = (2, 2.5, -0.73)$

$E = (2.2, 2.5, -0.73)$

$C = (2.2, 2.5, 0)$

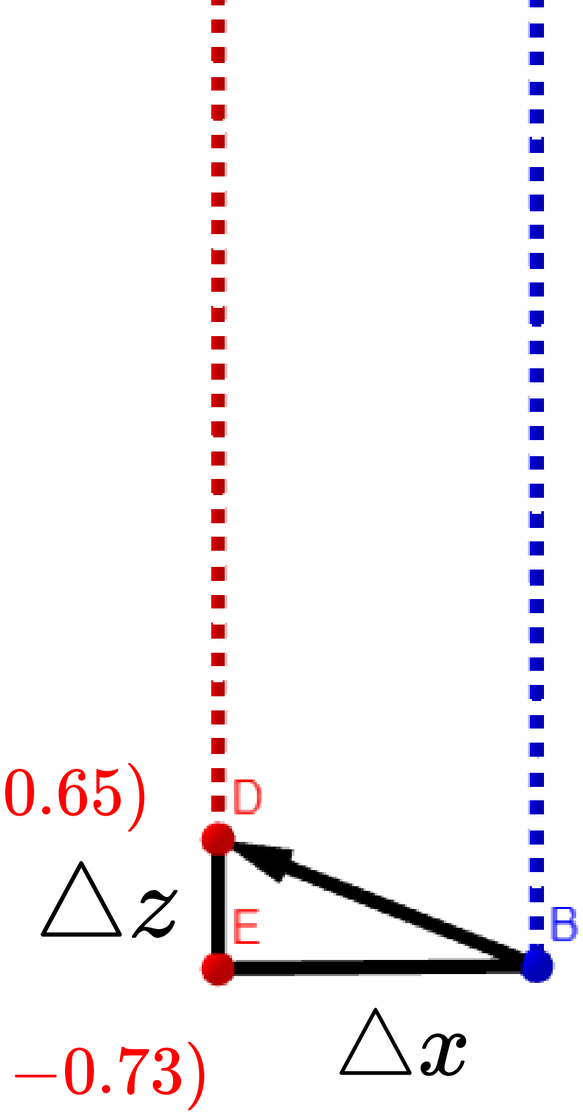
$$\triangle z = f(C_0) - f(A_0)$$

$$= -0.65 + 0.73 = 0.08$$

$$\triangle x = 0.2$$

Rate of change of f at (2,2.5) in x direction when $\triangle x = 0.2$ is :

$$\frac{\triangle z}{\triangle x} = 0.4$$



$D = (2.2, 2.5, -0.65)$

$B = (2, 2.5, -0.73)$

$E = (2.2, 2.5, -0.73)$

Rate of change of f at (2,2.5) when $\triangle x = 0.2$ is :

$$\frac{\triangle z}{\triangle x} = 0.4$$

But we did say that to calculate the accurate rate of change, $\triangle x \to 0$ and in this case,

$$\frac{\triangle z}{\triangle x} = \frac{\partial f}{\partial x} = f_x\left(x, y\right) = \lim_{\triangle x \to 0} \frac{f\left(x + \triangle x, y\right) - f\left(x, y\right)}{\triangle x}$$

Partial Derivatives

$$f(x,y) = sin(x) \times cos(y)$$

$$\frac{\partial f}{\partial x} = f_x(x,y) = cos(x) \times cos(y)$$

Therefore, the exact rate of change in the x direction at (2,2.5) is:

$$f_x(2,2.5) = cos(2) \times cos(2.5) \approx 0.33$$

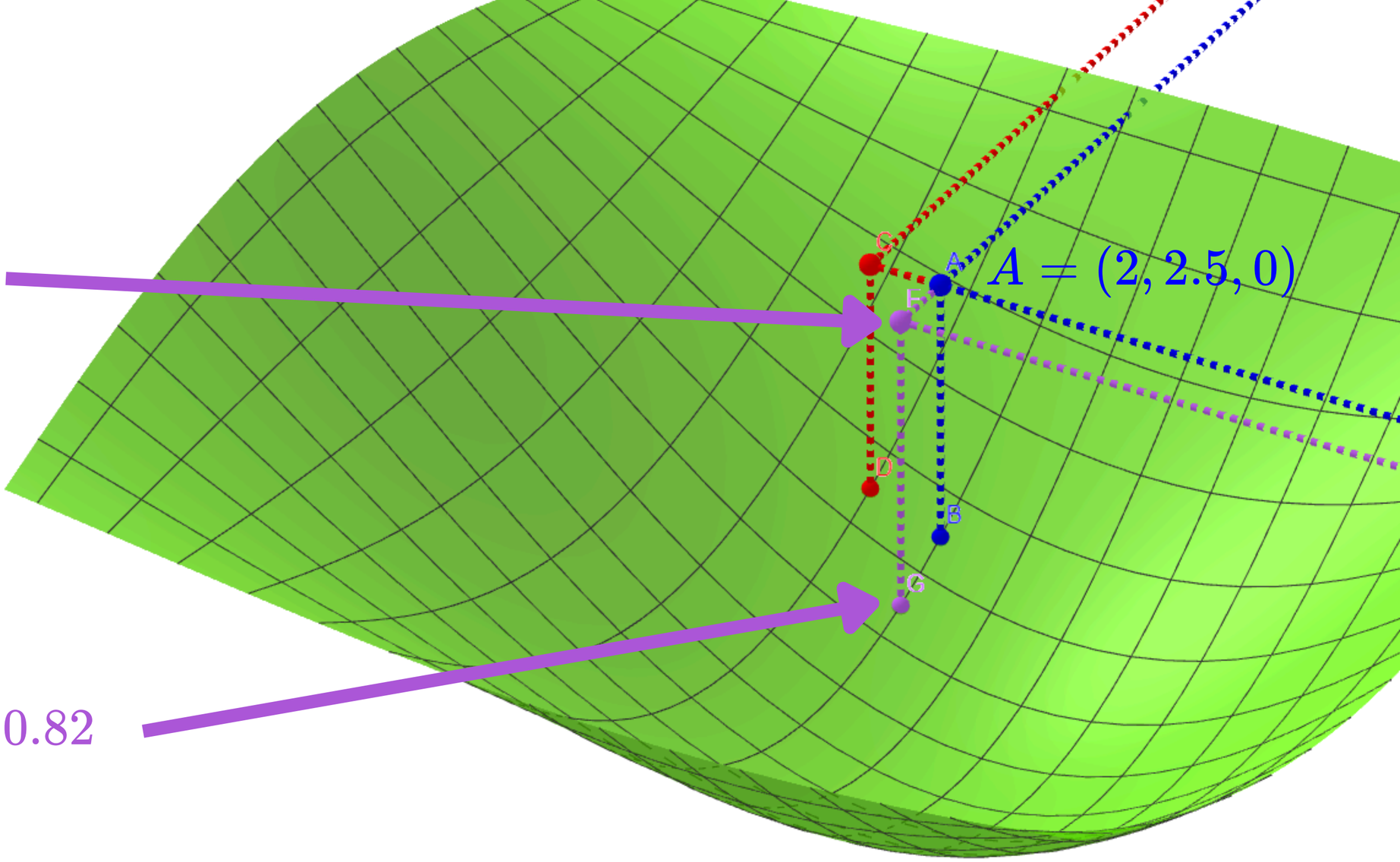
Partial Derivatives

Now let's calculate the rate of change of f in y direction at the same point (2,2.5):

$$F = (2, 2.5 + \triangle y, 0)$$

$$\triangle y = 0.2$$

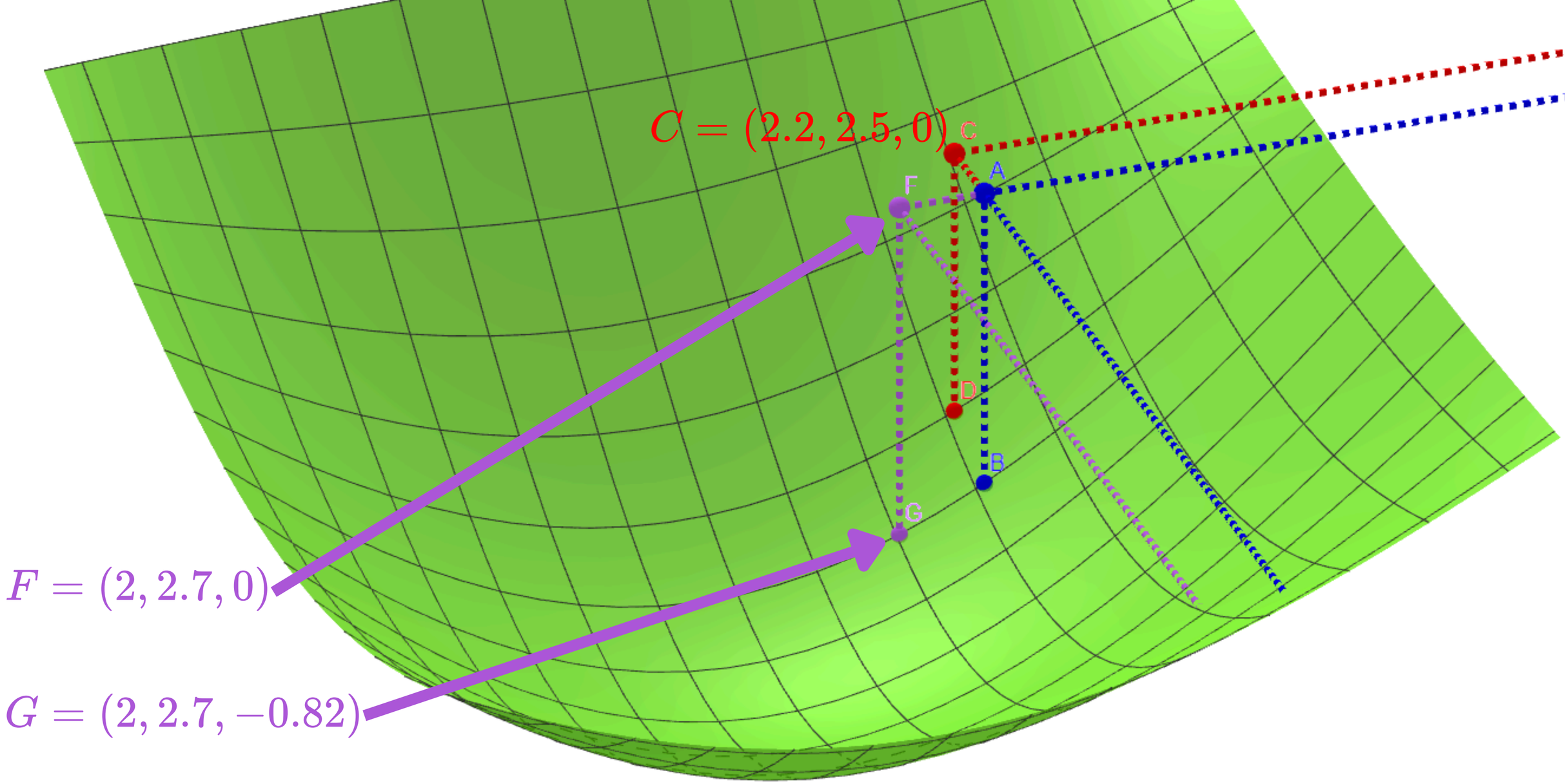
$$F = (2, 2.7, 0)$$

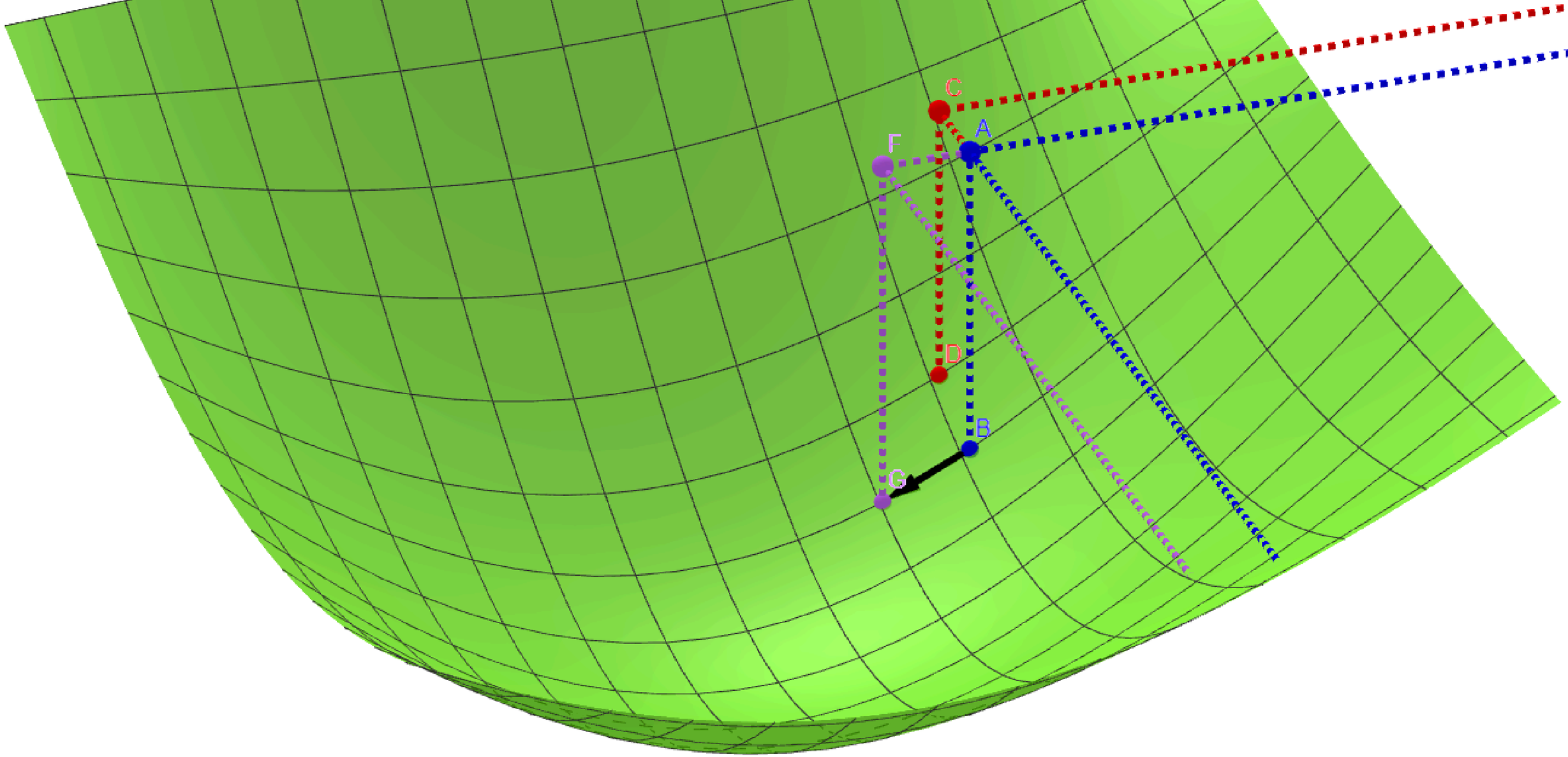


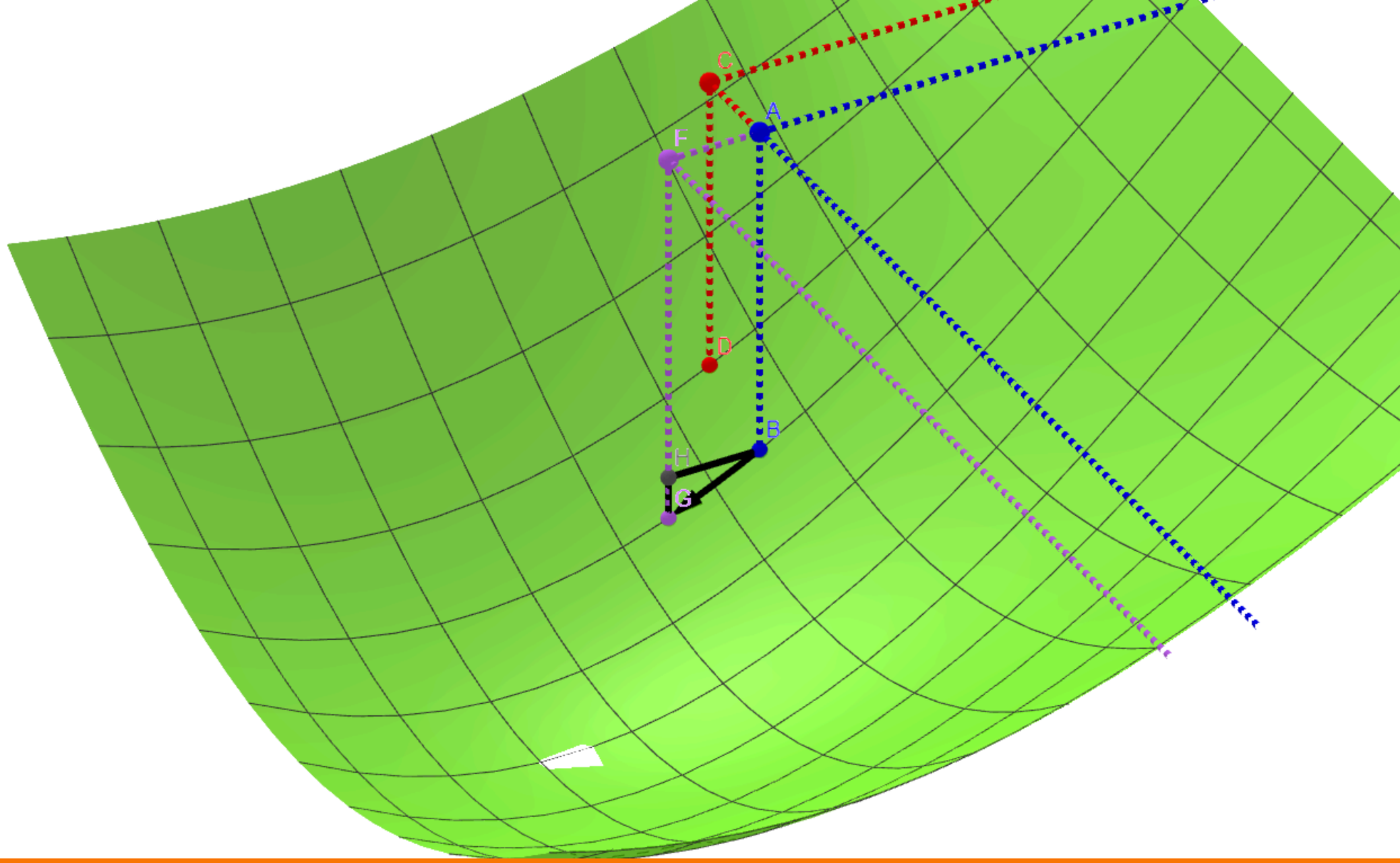
$$G = (2, 2.7, f(F_0))$$

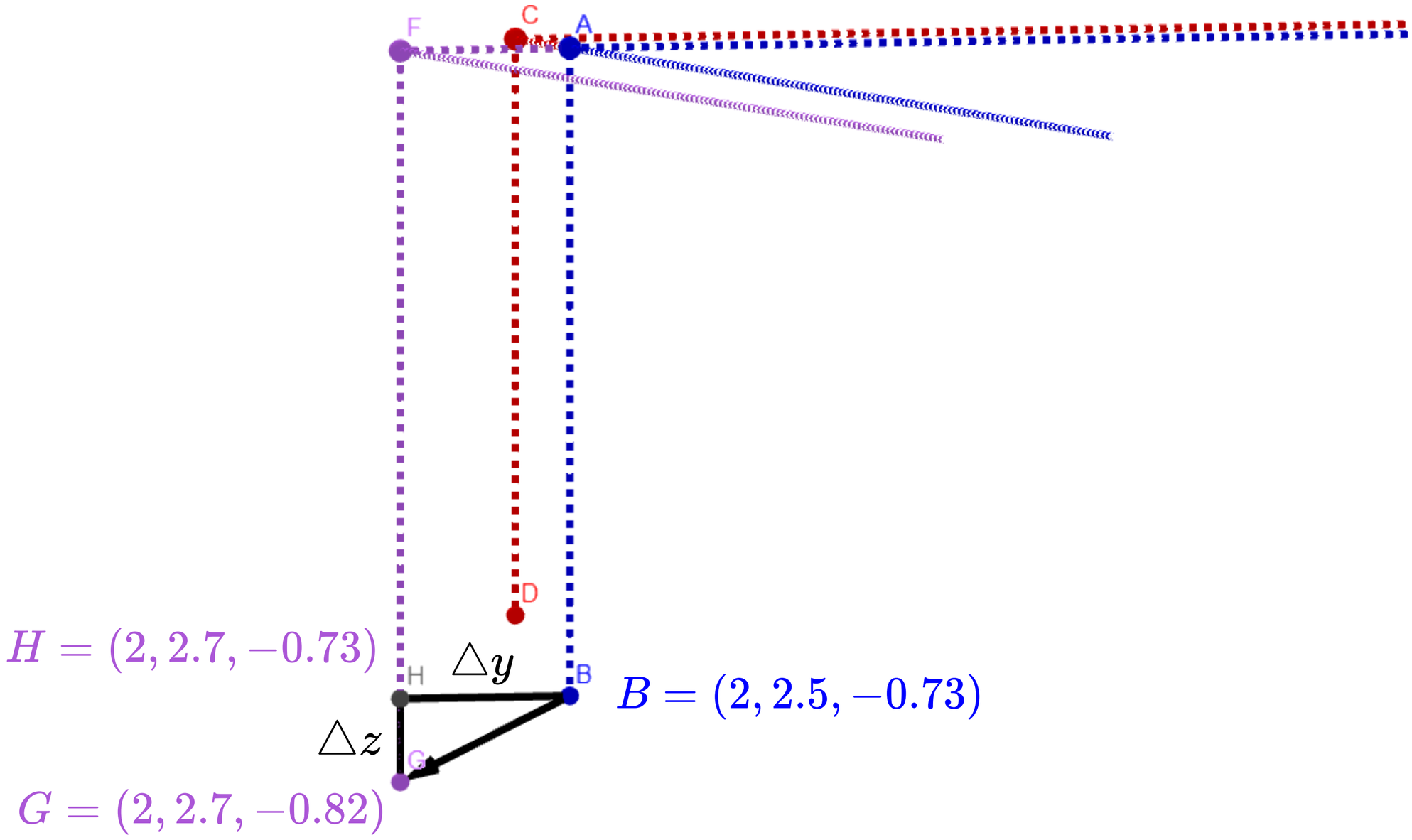
$$f(F_0) = f(2, 2.7) \approx -0.82$$

$$G = (2, 2.7, -0.82)$$


$C = (2.2, 2.5, 0)$
C
A
F
D
B
G
$F = (2, 2.7, 0)$
$G = (2, 2.7, -0.82)$


C
A
F
D
B
G

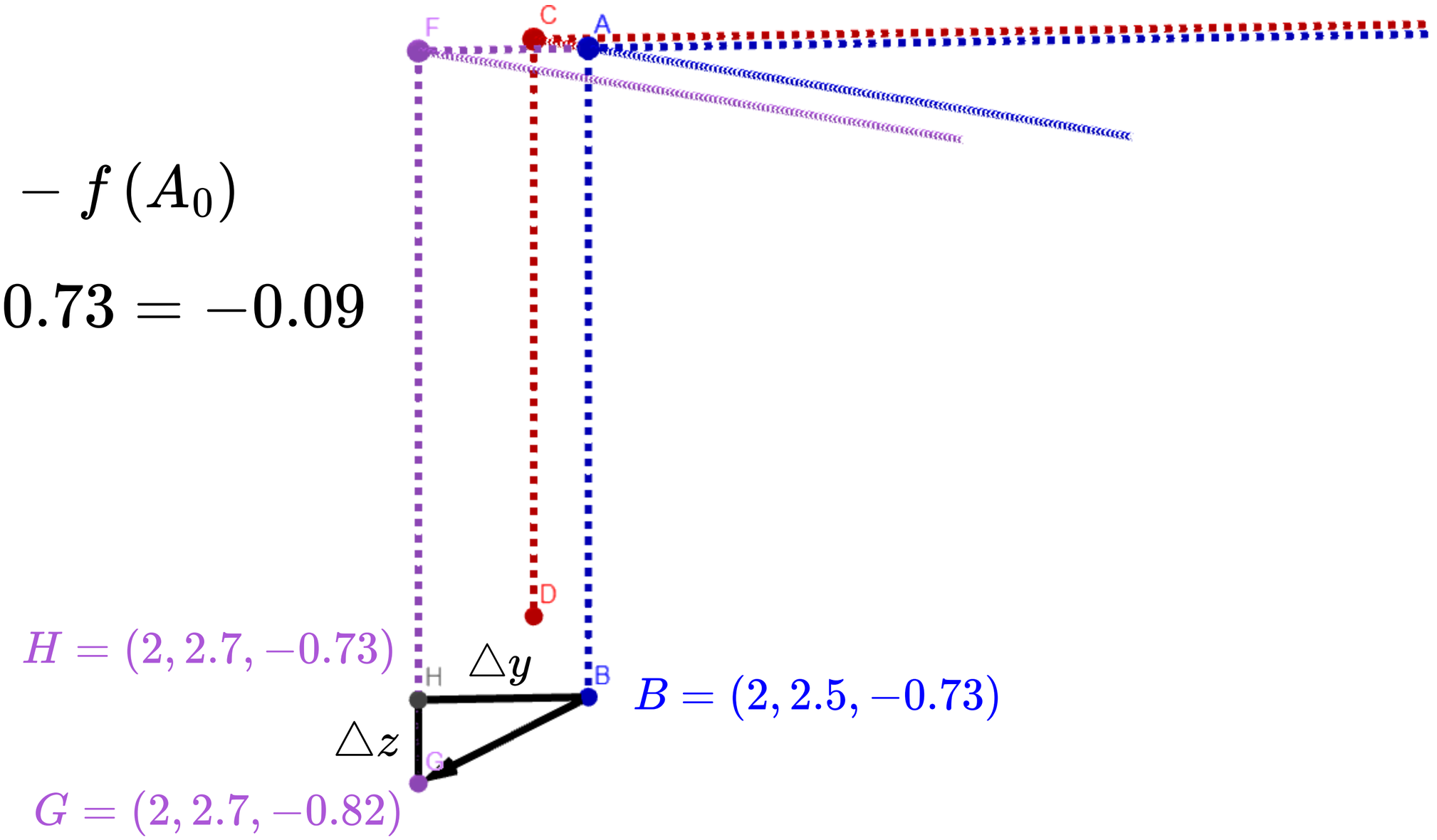

C
A
F
D
B
H
G


F
C
A
D
H
B
G
$H = (2, 2.7, -0.73)$
$\triangle y$
$B = (2, 2.5, -0.73)$
$\triangle z$
$G = (2, 2.7, -0.82)$

$$\triangle z = f(F_0) - f(A_0)$$

$$= -0.82 + 0.73 = -0.09$$

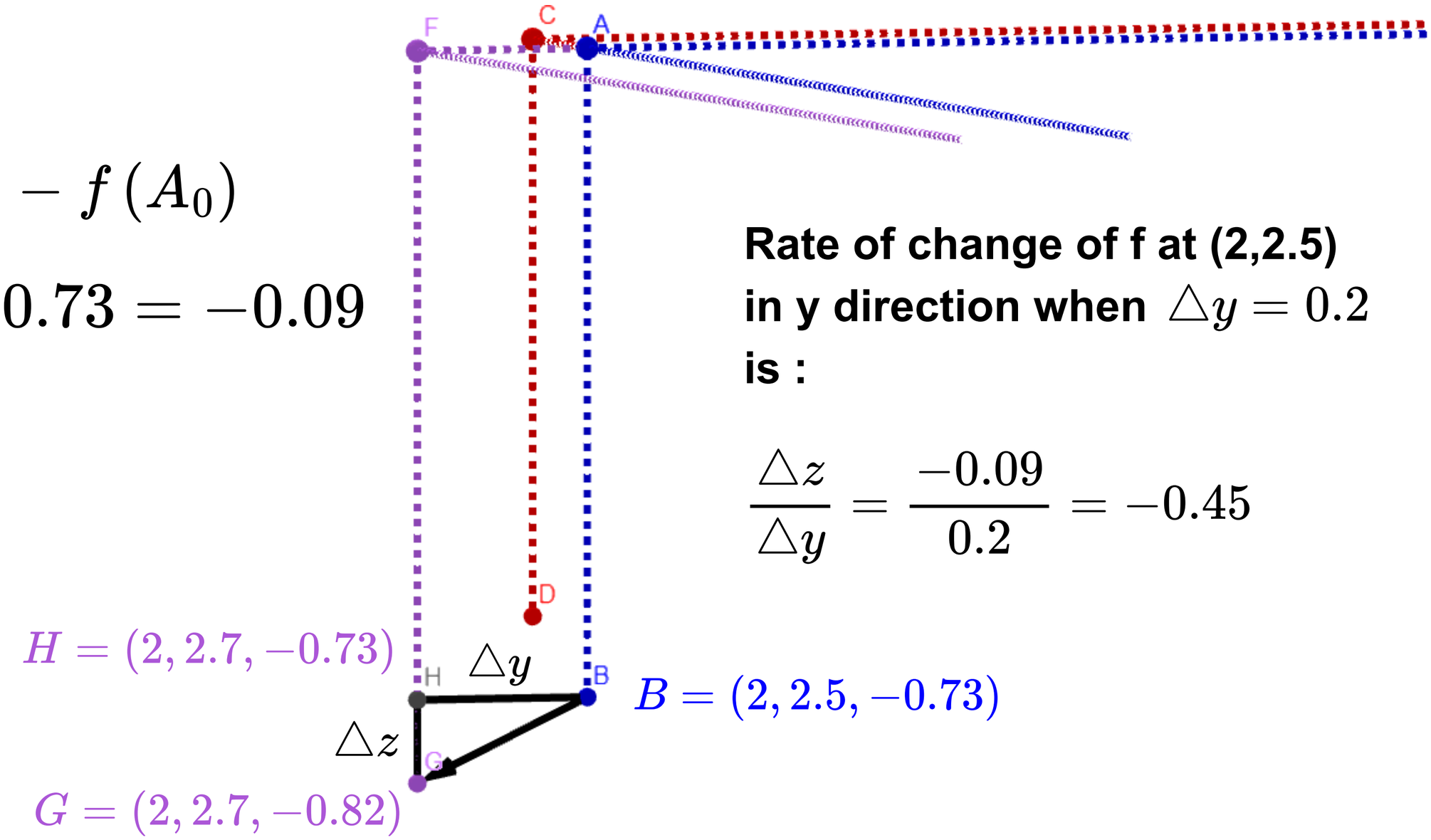
$$\triangle y = 0.2$$



$$\triangle z = f(F_0) - f(A_0)$$

$$= -0.82 + 0.73 = -0.09$$

$$\triangle y = 0.2$$



Rate of change of f at (2,2.5) in y direction when $\triangle y = 0.2$ is :

$$\frac{\triangle z}{\triangle y} = \frac{-0.09}{0.2} = -0.45$$

Partial Derivatives

To calculate the accurate rate of change, $\triangle y \to 0$ and in this case,

$$\frac{\triangle z}{\triangle y} = \frac{\partial f}{\partial y} = f_y(x,y) = \lim_{\triangle y \to 0} \frac{f(x, y+\triangle y) - f(x,y)}{\triangle y}$$

Partial Derivatives

$$f(x,y) = sin(x) \times cos(y)$$

$$\frac{\partial f}{\partial y} = f_y(x,y) = sin(x) \times (-sin(y))$$

Therefore, the exact rate of change in the y direction at (2,2.5) is:

$$f_y(2,2.5) = sin(2) \times (-sin(2.5)) \approx -0.54$$

Partial Derivatives

Directional derivatives

If we consider a function at a given point $f(x, y, z)$, there are obviously many different directions in which we could move away from the initial point. In general, any linear combination which is a unit vector $(a^2 + b^2 + c^2 = 1)$

$$\mathbf{u} = a\mathbf{i} + b\mathbf{j} + c\mathbf{k}\,,$$

For a two variable function case, the unit vector is given by

$$\mathbf{u} = a\mathbf{i} + b\mathbf{j}$$

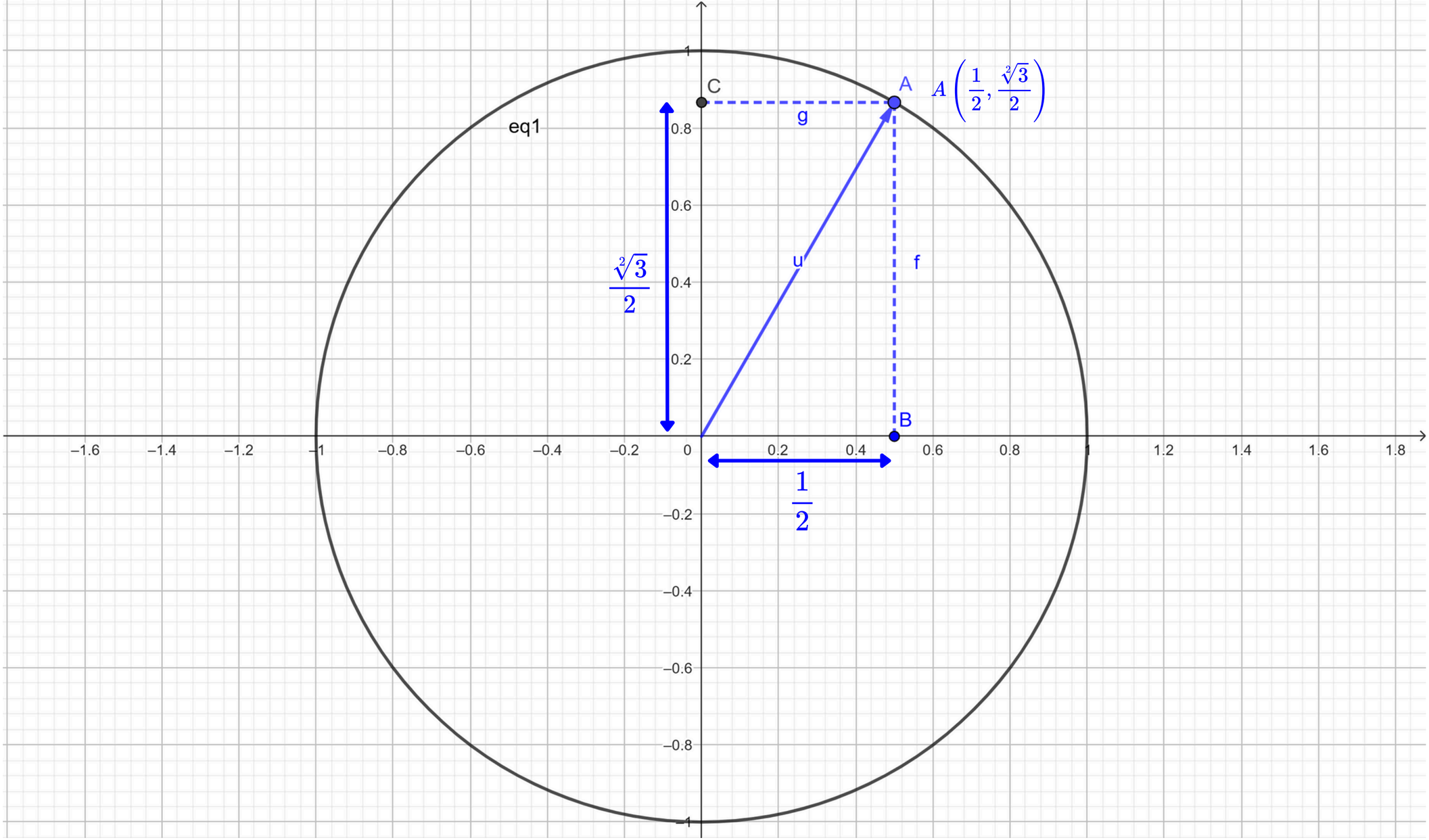
with $(a^2 + b^2 = 1)$

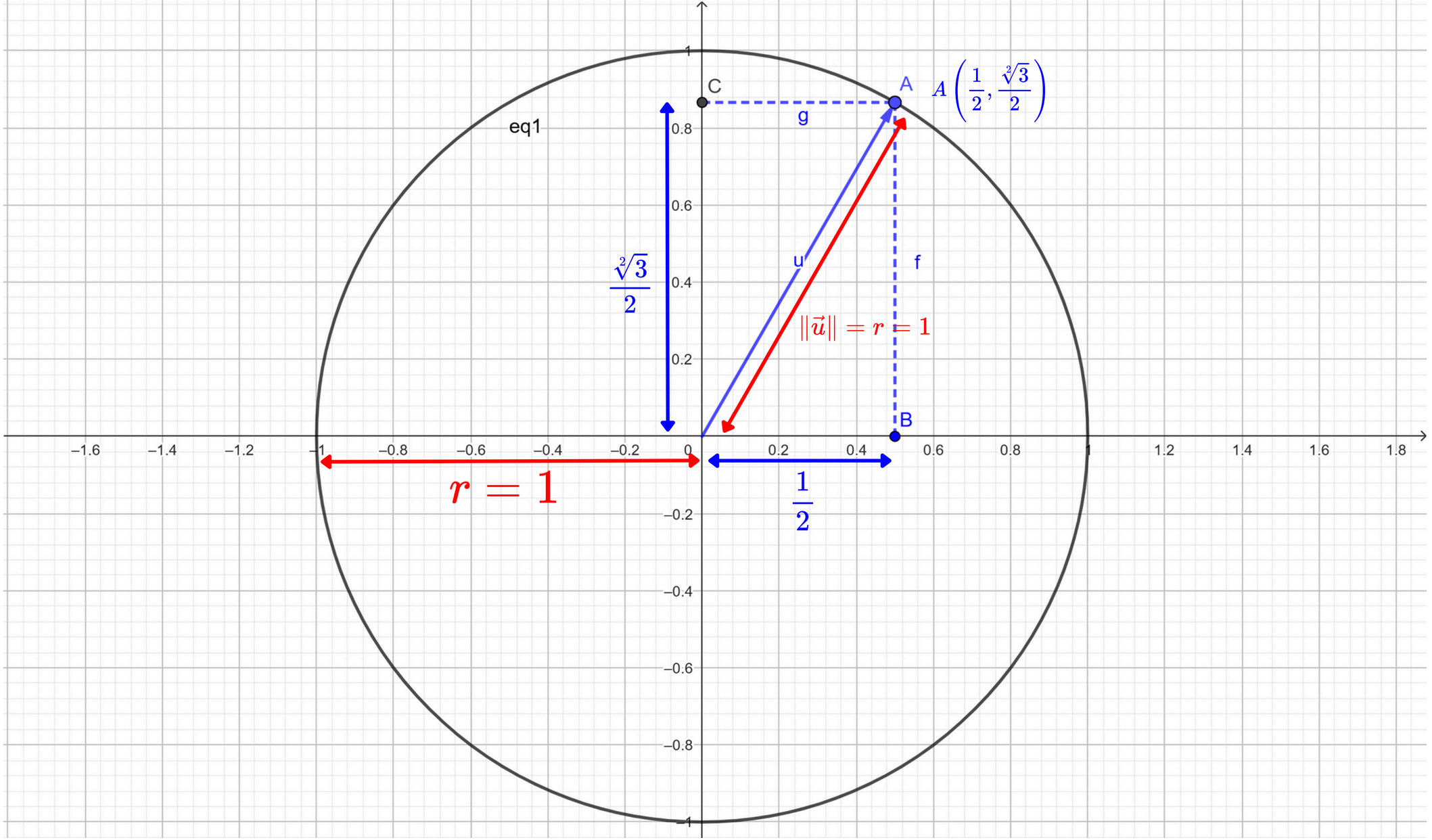
Partial Derivatives

Directional derivatives

$$\mathrm{u} = a\mathrm{i} + b\mathrm{j}$$

$(a^2 + b^2 = 1)$, because the length of u must be equal to 1 (Pythagorean theorem)


eq1
C
A
$A\left(\frac{1}{2}, \frac{\sqrt[2]{3}}{2}\right)$
g
u
f
B
$\frac{\sqrt[2]{3}}{2}$
$\frac{1}{2}$
1
0.8
0.6
0.4
0.2
−0.2
−0.4
−0.6
−0.8
−1
−1.6
−1.4
−1.2
−1
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8


eq1
C
A
$A\left(\frac{1}{2}, \frac{\sqrt[2]{3}}{2}\right)$
g
f
u
$\frac{\sqrt[2]{3}}{2}$
$||\vec{u}|| = r = 1$
B
$r = 1$
$\frac{1}{2}$

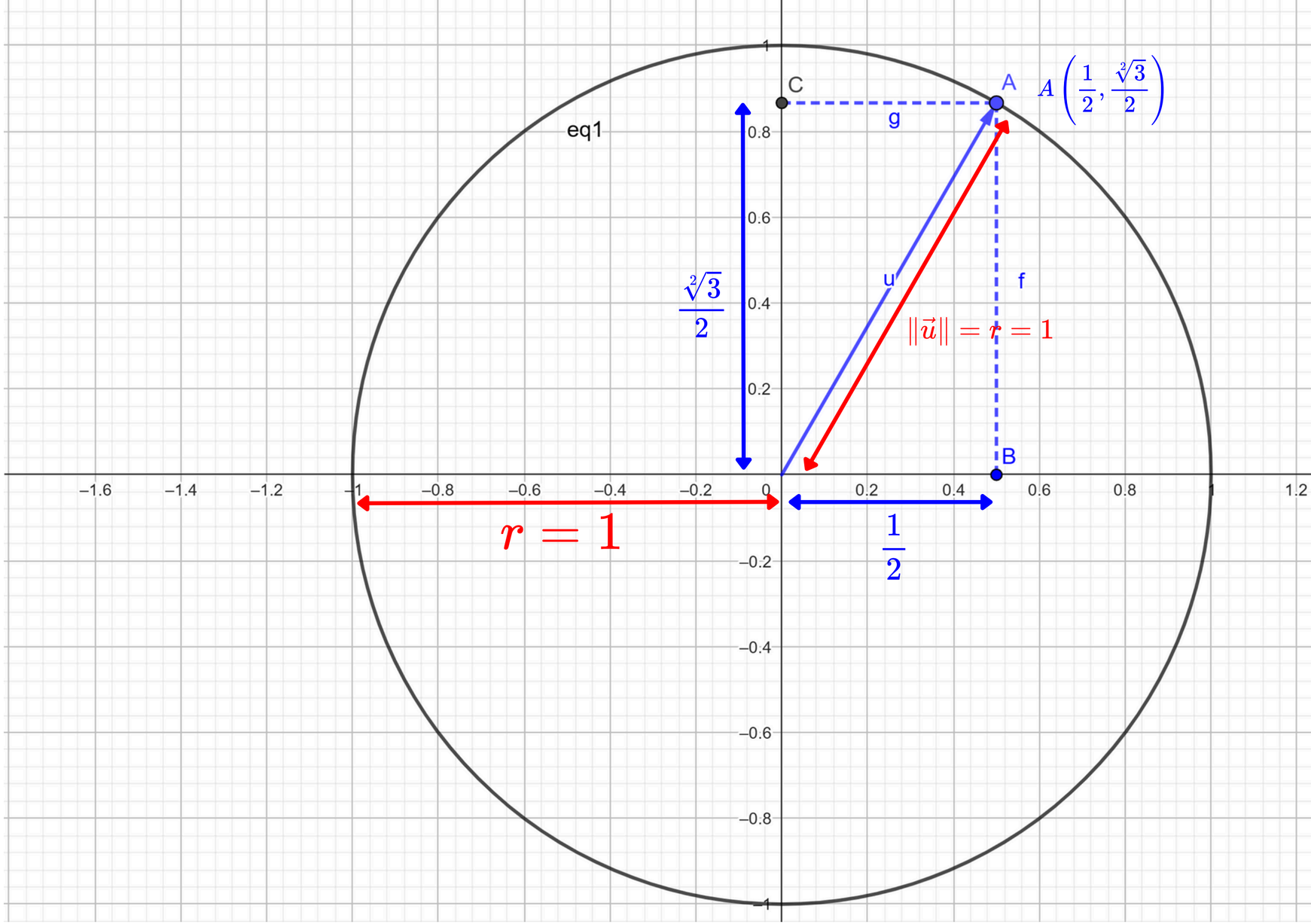
We got $\mathrm{u} = \frac{1}{2}\mathrm{i} + \frac{\sqrt{3}}{2}\mathrm{j}$.

$$a = \frac{1}{2}$$

$$b = \frac{\sqrt[2]{3}}{2}$$

Notice that

$$\left(\frac{1}{2}\right)^2 + \left(\frac{\sqrt[2]{3}}{2}\right)^2 = 1$$



Partial Derivatives

Directional derivatives

If we consider a function at a given point $f(x, y, z)$, there are obviously many different directions in which we could move away from the initial point. In general, any linear combination which is a unit vector $(a^2 + b^2 + c^2 = 1)$

$$\mathbf{u} = a\mathbf{i} + b\mathbf{j} + c\mathbf{k}\,,$$

defines a direction, if we fix the origin to be (x_0, y_0, z_0). In terms of the arc length parameter s, we express subsequent motion away from (x_0, y_0, z_0) through the equations

$$x = x_0 + a\ s\,, \quad y = y_0 + b\ s\,, \quad z = z_0 + c\ s\,.$$

Partial Derivatives

When we take s to 0, we recover the initial point. Then, differentiation with respect to s will give the slope in the direction of $\mathbf{u}$ when we set $s \to 0$. In other words, we use s to test how a small change affects the function f at (x_0, y_0, z_0). If we didn't set $s \to 0$ at the end, we would not find the derivative at (x_0, y_0, z_0), but at a point an arc length s away in the relevant directions. As a result, we define the **directional derivative of f in the direction of u** to be

$$\begin{aligned} D_{\mathbf{u}} f(x_0, y_0, z_0) &= \frac{d}{ds}[f(x_0 + a\ s, y_0 + b\ s, z_0 + c\ s)]|_{s=0} \\ &= f_x(x_0, y_0, z_0)a + f_y(x_0, y_0, z_0)b + f_z(x_0, y_0, z_0)c\,. \end{aligned} \tag{44}$$

This can be regarded as the slope of the surface $w = f(x, y, z)$ in the direction $\mathbf{u}$.

Partial Derivatives

Let’s apply this in our previous example

In the previous times, we were calculating the rate of change of our function in x-direction and y-direction, now we are going to calculate the rate of change in u direction.

We can choose any direction $\mathrm{u} = a\mathrm{i} + b\mathrm{j}$**. provided** $a^2 + b^2 = 1$

Partial Derivatives

We put $\mathrm{u} = \frac{1}{2}\mathrm{i} + \frac{\sqrt{3}}{2}\mathrm{j}.$

$a = \frac{1}{2} \qquad b = \frac{\sqrt[2]{3}}{2}$

$\left(\frac{1}{2}\right)^2 + \left(\frac{\sqrt[2]{3}}{2}\right)^2 = 1$ **(VERIFIED)**

Partial Derivatives

$$\mathrm{u} = \frac{1}{2}\mathrm{i} + \frac{\sqrt{3}}{2}\mathrm{j}.$$

As usual, we will calculate the rate of change of the function at $A_0\,(2, 2.5)$ with change in x and change in y depending on our initial vector u.

Partial Derivatives

$$\mathrm{u} = \frac{1}{2}\mathrm{i} + \frac{\sqrt{3}}{2}\mathrm{j}.$$

As usual, we will calculate the rate of change of the function at $A_0\,(2, 2.5)$ with change in x and change in y depending on our initial vector u.

$$\triangle x = a \times s = \frac{1}{2} \times 0.2$$

$$\triangle y = b \times s = \frac{\sqrt[2]{3}}{2} \times 0.2$$

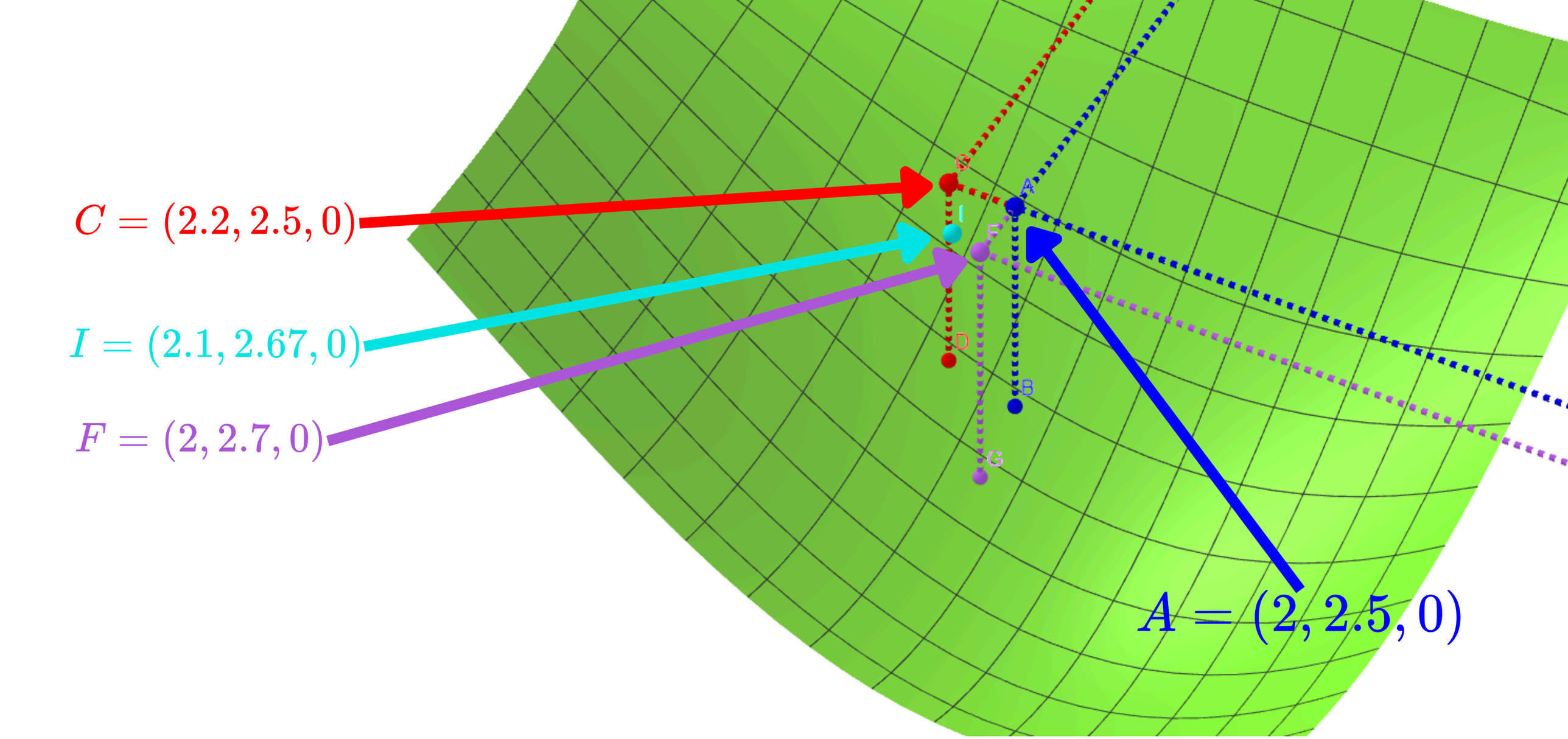
Partial Derivatives

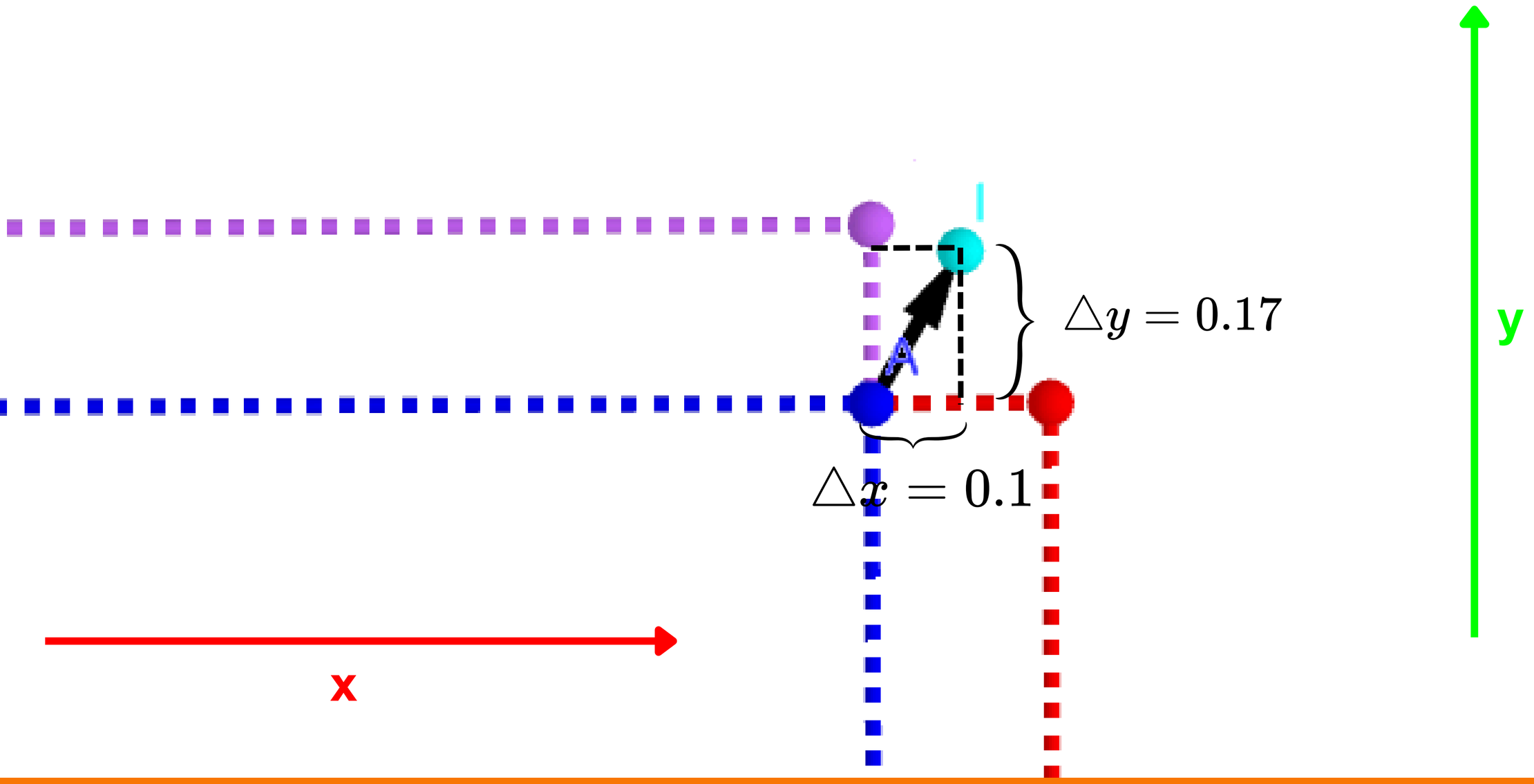
$$\triangle x = a \times s = \frac{1}{2} \times 0.2 = 0.1$$

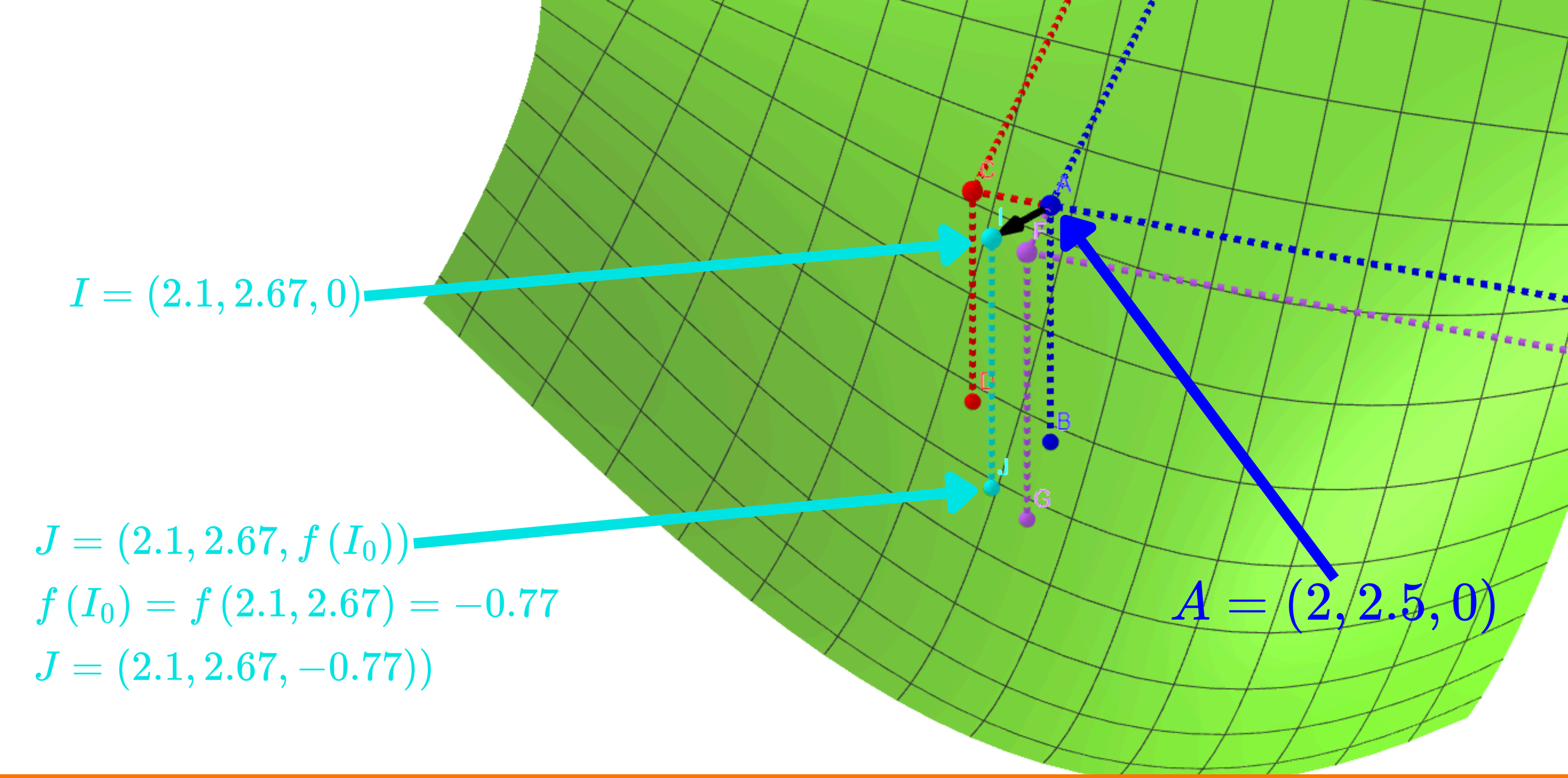
$$\triangle y = b \times s = \frac{\sqrt[2]{3}}{2} \times 0.2 = 0.17$$

$$I_0 = (2 + \triangle x, 2.5 + \triangle y)$$

$$= (2.1, 2.67)$$


$C = (2.2, 2.5, 0)$
$I = (2.1, 2.67, 0)$
$F = (2, 2.7, 0)$
$A = (2, 2.5, 0)$
C
A
I
F
D
B
G


$\triangle y = 0.17$
$\triangle x = 0.1$
A
y
x



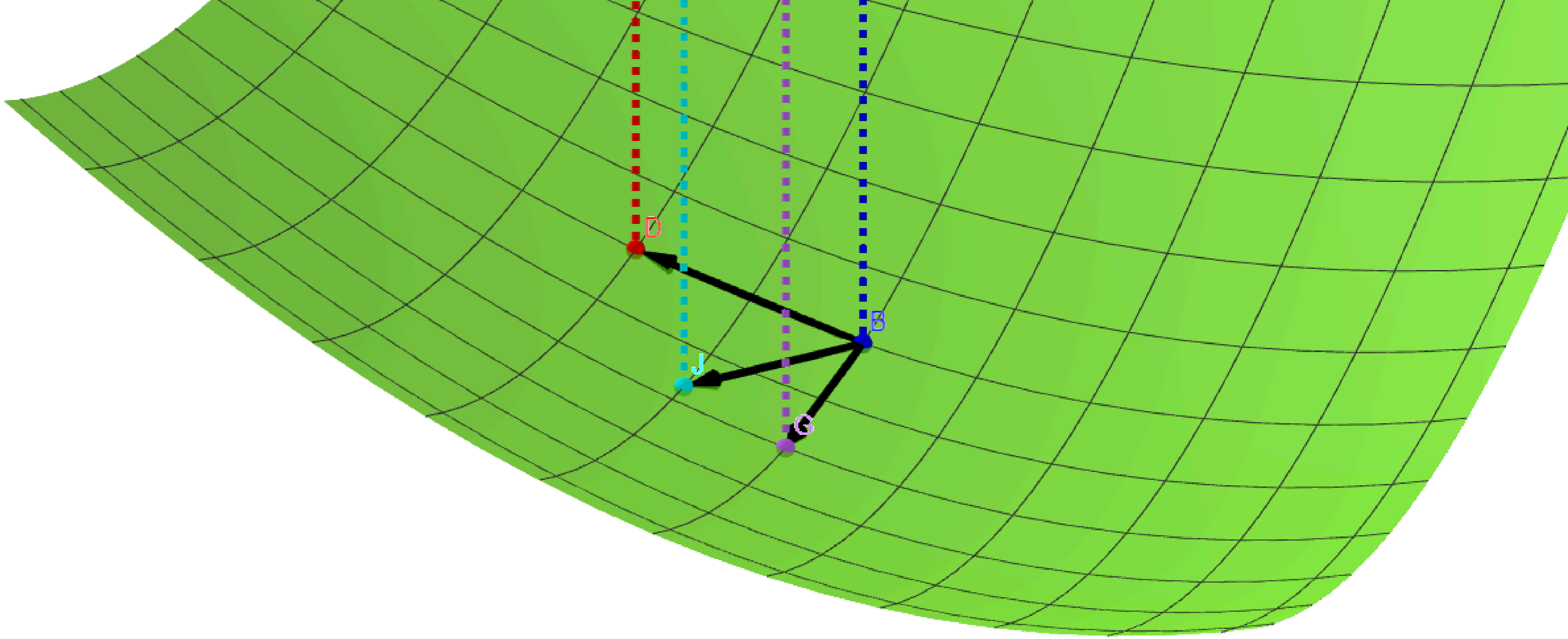
$I = (2.1, 2.67, 0)$

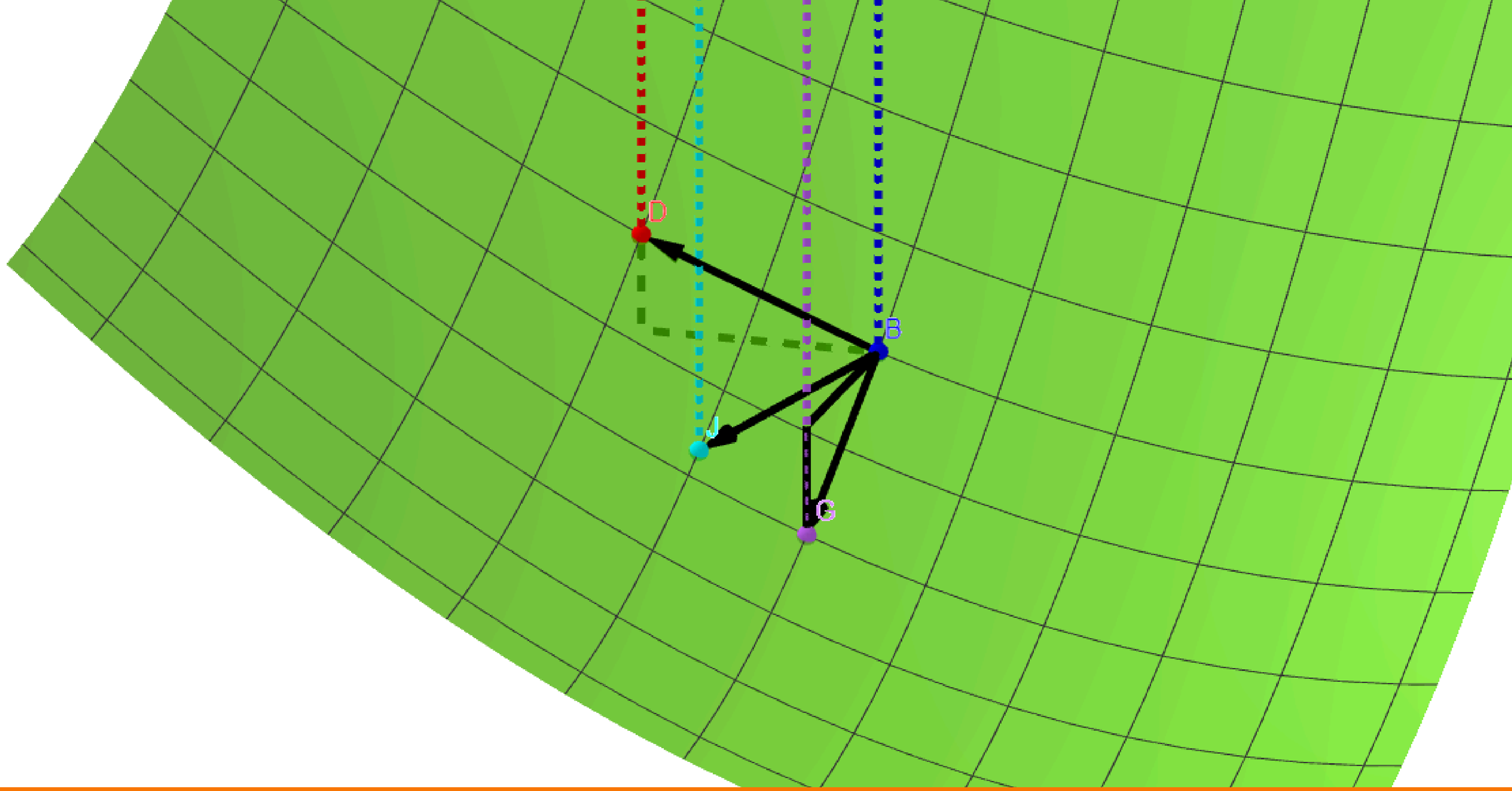
$J = (2.1, 2.67, f(I_0))$

$f(I_0) = f(2.1, 2.67) = -0.77$

$J = (2.1, 2.67, -0.77))$

$A = (2, 2.5, 0)$


D
B
J
G


D
B
J
G

Partial Derivatives: The Gradient

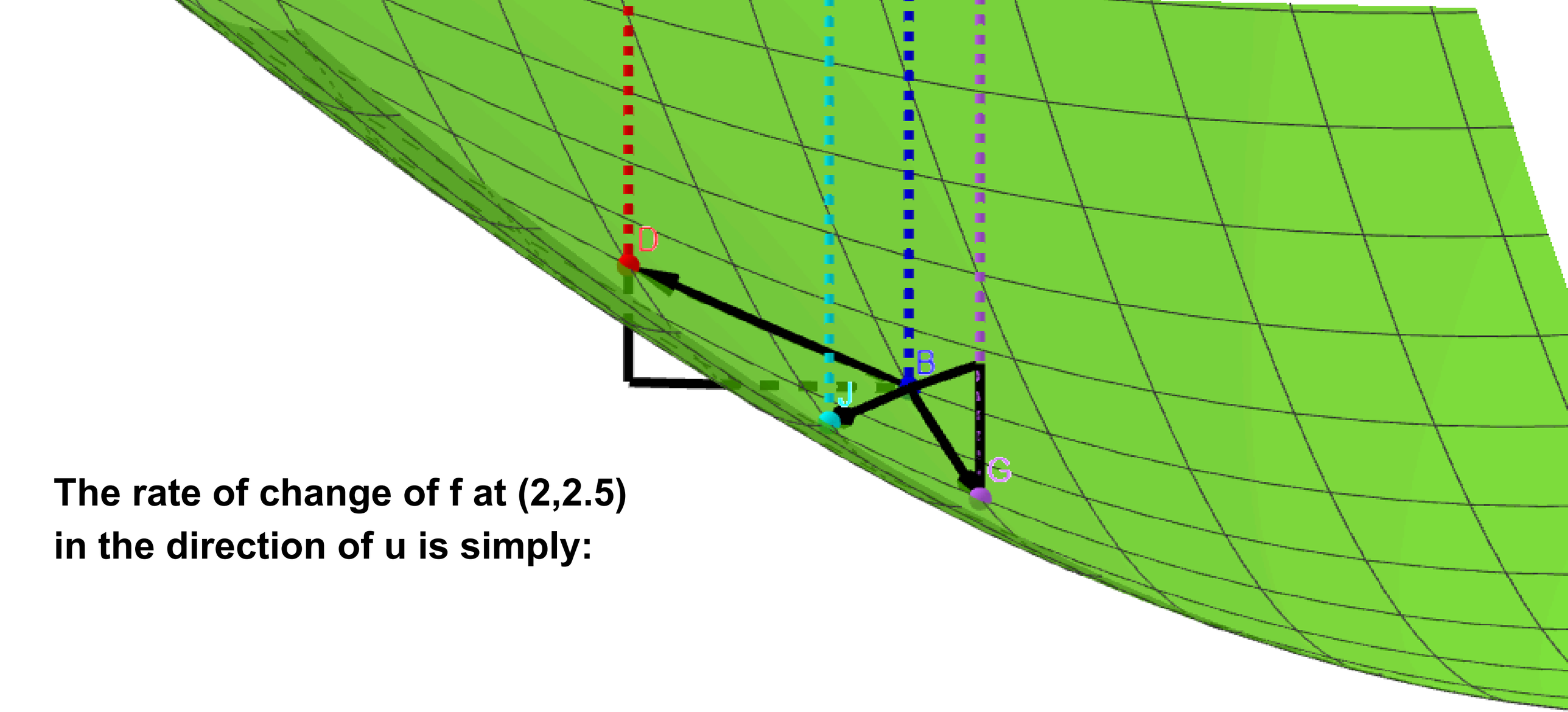
Calculating directional derivative is made easier using the **gradient**. It is denoted by ∇, which is called "nabla", but generally read as "del" and is given by

$$\nabla f(x, y, z) = f_x(x, y, z)\mathbf{i} + f_y(x, y, z)\mathbf{j} + f_z(x, y, z)\mathbf{k}\,.$$

Using this, we see that we can use it to express directional derivatives as

$$D_{\mathbf{u}} f(x, y, z) = \nabla f(x, y, z) \cdot \mathbf{u}\,.$$

This is why it is called a gradient, because it can give the slope in any direction if the dot product with a unit vector is taken. Properties of the gradient are:



The rate of change of f at (2,2.5)
in the direction of u is simply:

$$D_u f(x,y) = \bigtriangledown f(x,y) \cdot \mathbf{u}$$

$$= (f_x(x,y)\mathbf{i} + f_y(x,y)\mathbf{j}) \cdot \mathbf{u}$$

$$= (f_x(x,y)\mathbf{i} + f_y(x,y)\mathbf{j}) \cdot (\frac{1}{2}\mathbf{i} + \frac{\sqrt{3}}{2}\mathbf{j})$$

$$= \frac{1}{2} f_x(x,y) + \frac{\sqrt{3}}{2} f_y(x,y)$$

$$= \frac{1}{2} f_x(x, y) + \frac{\sqrt{3}}{2} f_y(x, y)$$

$$= \frac{1}{2} \cos(x) \cos(y) - \frac{\sqrt{3}}{2} \sin(x) \sin(y)$$

At (2,2.5) in u direction, the rate of change of the function is:

$$D_u f(2, 2.5) = \frac{1}{2} \cos(2) \cos(2.5) - \frac{\sqrt{3}}{2} \sin(2) \sin(2.5) = -0.3$$

Gradient Descent

So far, we've built a strong foundation by covering everything from limits to partial derivatives. With these tools in hand, we're ready to dive deeper into the gradient descent algorithm and explore how it truly works.

Gradient Descent: Back to our model

$$z\left(x_1\right) = (0.5)x_1 - 4$$

We found a very low likelihood, which corresponded to a very high loss.

$$L = \prod_{i=1}^{m} P(Y = y_i) = P\left(Y = y_1\right) \times P\left(Y = y_2\right) \times \ldots \times P\left(Y = y_8\right)$$

$$= 0.5 \times 0.02931 \times 0.00407 \times \ldots \times 0.99999$$

$$= 2.0002713231 \times 10^{-8}$$

Gradient Descent: Back to our model

$$Loss = -\frac{1}{m} log\left(L\right) = -\frac{1}{m} log\left(\prod_{i=1}^{m}\left(a_i^{y_i}(1-a_i)^{1-y_i}\right)\right)$$

After simplifying:

$$Loss = -\frac{1}{m}\sum_{i=1}^{m}\left(y_i.\,log\left(a_i\right) + (1-y_i).\,log\left(1-a_i\right)\right)$$

Gradient Descent: Back to our model

$$z(x_1) = (0.5)x_1 - 4 \longrightarrow Loss = 0.96236$$

Gradient Descent: Back to our model

This Loss function depends on w_1 and b

$$Loss = -\frac{1}{m}\sum_{i=1}^{m}(y_i.log(a_i) + (1-y_i).log(1-a_i))$$

Gradient Descent: Back to our model

$$Loss = -\frac{1}{m}\sum_{i=1}^{m}\left(y_i.\,log\left(a_i\right) + \left(1-y_i\right).\,log\left(1-a_i\right)\right)$$

and since $a\left(x\right) = \frac{1}{1+e^{-x}}$

$$Loss = -\frac{1}{m}\sum_{i=1}^{m}\left(y_i.\,log\left(\frac{1}{1+e^{-z_i}}\right) + \left(1-y_i\right).\,log\left(1-\frac{1}{1+e^{-z_i}}\right)\right)$$

Gradient Descent: Back to our model

$$Loss = -\frac{1}{m}\sum_{i=1}^{m}\left(y_i.log\left(\frac{1}{1+e^{-z_i}}\right)+(1-y_i).log\left(1-\frac{1}{1+e^{-z_i}}\right)\right)$$

and since $z\left(x_1\right)=(0.5)x_1-4$

$$Loss = -\frac{1}{m}\sum_{i=1}^{m}\left(y_i.log\left(\frac{1}{1+e^{-(0.5)x_1-4}}\right)+(1-y_i).log\left(1-\frac{1}{1+e^{-(0.5)x_1-4}}\right)\right)$$

Gradient Descent: Back to our model

$$Loss = -\frac{1}{m}\sum_{i=1}^{m}\left(y_i.log\left(\frac{1}{1+e^{-\left(w_1x_1^{(i)}+b\right)}}\right)+(1-y_i).log\left(1-\frac{1}{1+e^{-\left(w_1x_1^{(i)}+b\right)}}\right)\right)$$

The Loss function (In our case) is just a two-variable function where:

w_1 **is the first variable** (x)

b **is the second variable** (y)

Gradient Descent: Back to our model

$$Loss = -\frac{1}{m}\sum_{i=1}^{m}\left(y_i.log\left(\frac{1}{1+e^{-\left(w_1x_1^{(i)}+b\right)}}\right)+(1-y_i).log\left(1-\frac{1}{1+e^{-\left(w_1x_1^{(i)}+b\right)}}\right)\right)$$

The Loss function (In our case) is just a two-variable function where:

w_1 **is the first variable** (x)

b **is the second variable** (y)

Note: here we're only variying the weight and bias, $x_1^{(i)}$ **and** y_i **are related to the dataset examples and they never change when we train our model.**

Gradient Descent: Back to our model

$$Loss = -\frac{1}{m}\sum_{i=1}^{m}\left(y_i.log\left(\frac{1}{1+e^{-\left(w_1x_1^{(i)}+b\right)}}\right)+(1-y_i).log\left(1-\frac{1}{1+e^{-\left(w_1x_1^{(i)}+b\right)}}\right)\right)$$

$$z(x_1) = (0.5)x_1 - 4 \longrightarrow Loss = 0.96236$$

$$w_1 = 0.5$$

$$b = -4$$

Gradient Descent: Back to our model

$$Loss = -\frac{1}{m}\sum_{i=1}^{m}\left(y_i.log\left(\frac{1}{1+e^{-\left(w_1x_1^{(i)}+b\right)}}\right)+(1-y_i).log\left(1-\frac{1}{1+e^{-\left(w_1x_1^{(i)}+b\right)}}\right)\right)$$

$$z(x_1) = x_1 - 25 \longrightarrow Loss = 0.02417$$

$$w_1 = 1$$

$$b = -25$$

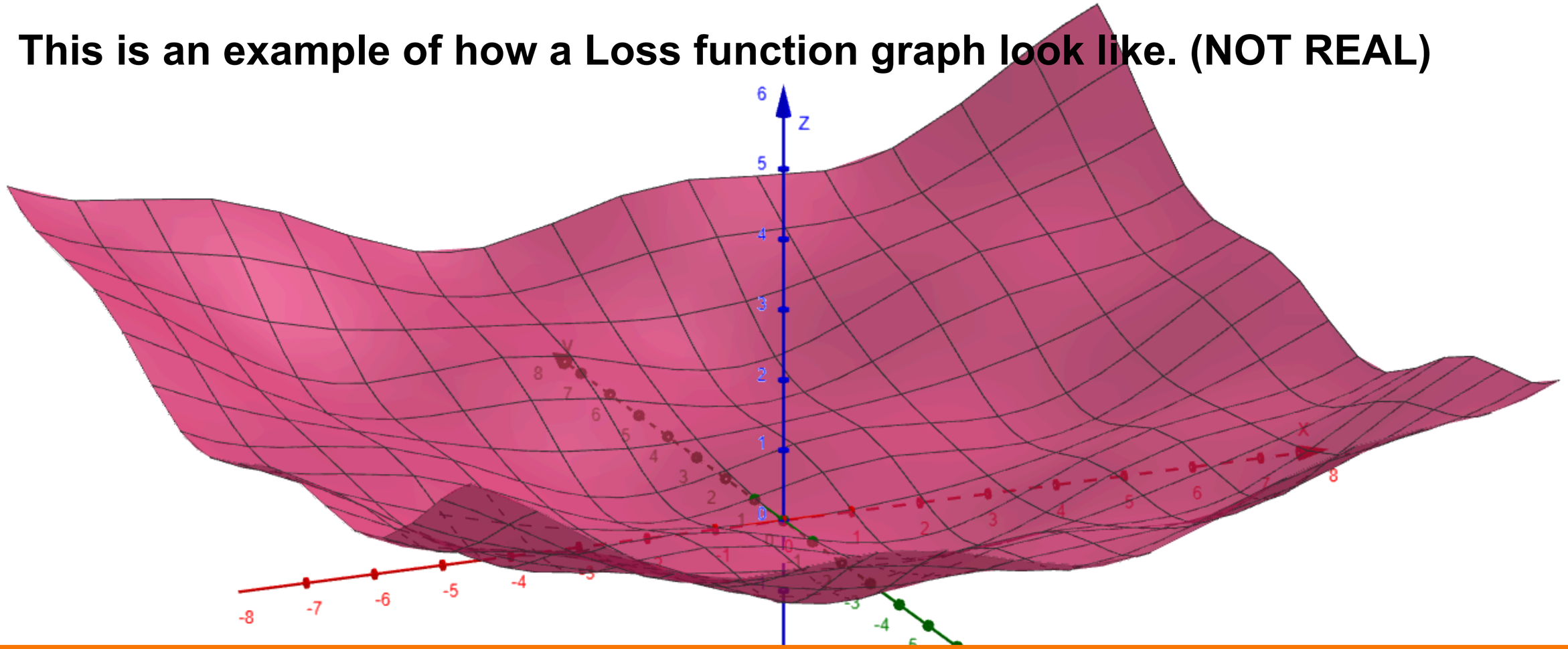
Gradient Descent: Back to our model

The Loss function also has a graph

$$Loss = -\frac{1}{m}\sum_{i=1}^{m}\left(y_i.log\left(\frac{1}{1+e^{-\left(w_1x_1^{(i)}+b\right)}}\right)+(1-y_i).log\left(1-\frac{1}{1+e^{-\left(w_1x_1^{(i)}+b\right)}}\right)\right)$$

Gradient Descent: Back to our model

This is an example of how a Loss function graph look like. (NOT REAL)



Gradient Descent: Back to our model

We will begin with random parameters w_1 and b and apply the Gradient Descent Algorithm to iteratively refine these parameters and minimize the loss.

In each iteration of the algorithm, we will increase or decrease the value of w_1 and b.

Gradient Descent: Back to our model

$$w_{t+1} = w_t - \alpha \frac{\partial L}{\partial w_t}$$

$$b_{t+1} = b_t - \alpha \frac{\partial L}{\partial b_t}$$

$$\begin{cases} w_t \text{ denotes the weight } w \text{ at instant } t \\ L \text{ represents the } Loss \end{cases}$$

Gradient Descent: Back to our model

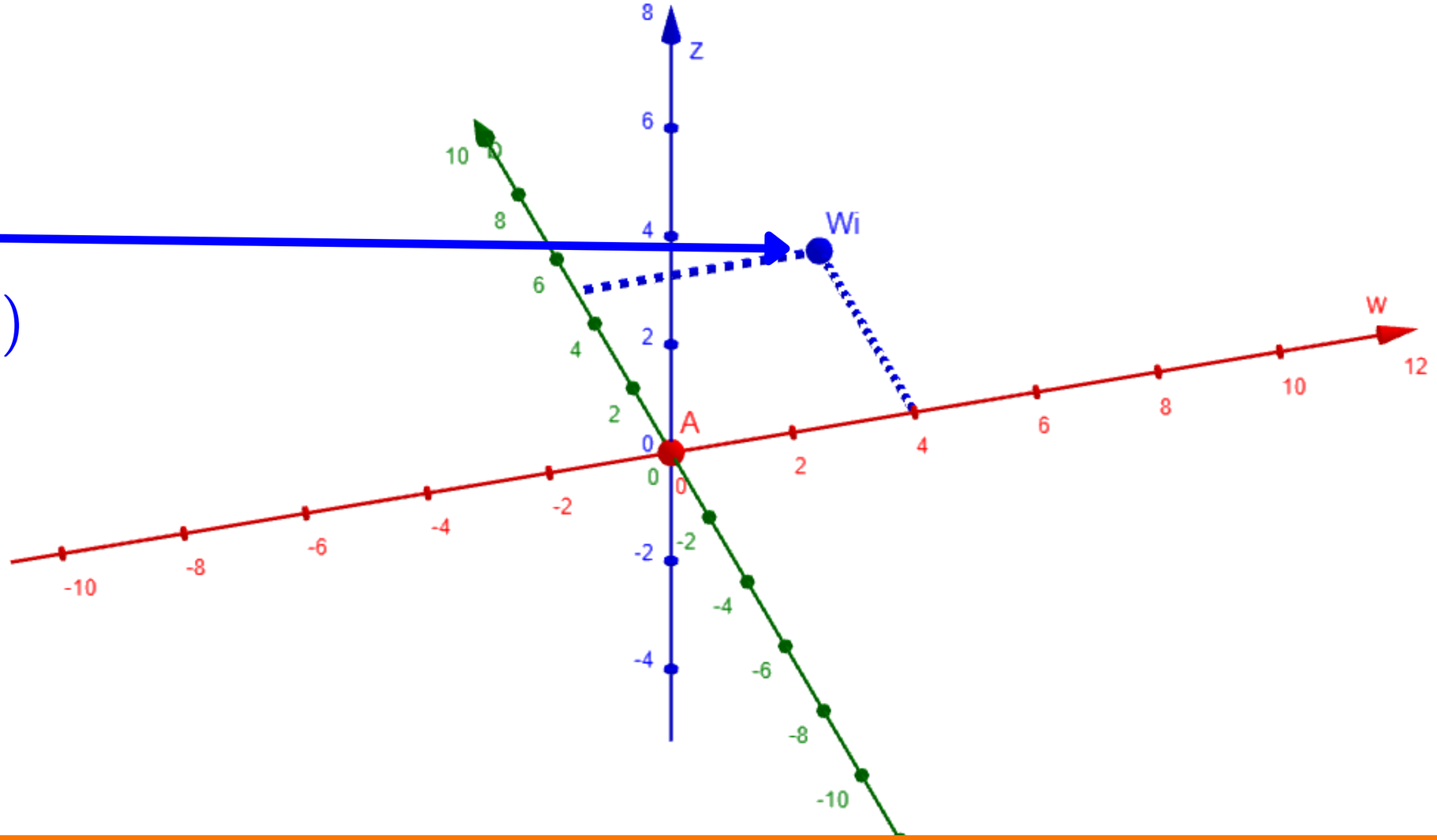
We start with random parameters: $w_1 = 4$ and $b = 5$.

Our initial $Loss$ is 1.38

Note: The example we're using is hypothetical and intended solely for explanatory purposes; in a real scenario, the Loss would not exceed 1.

Gradient Descent: Back to our model

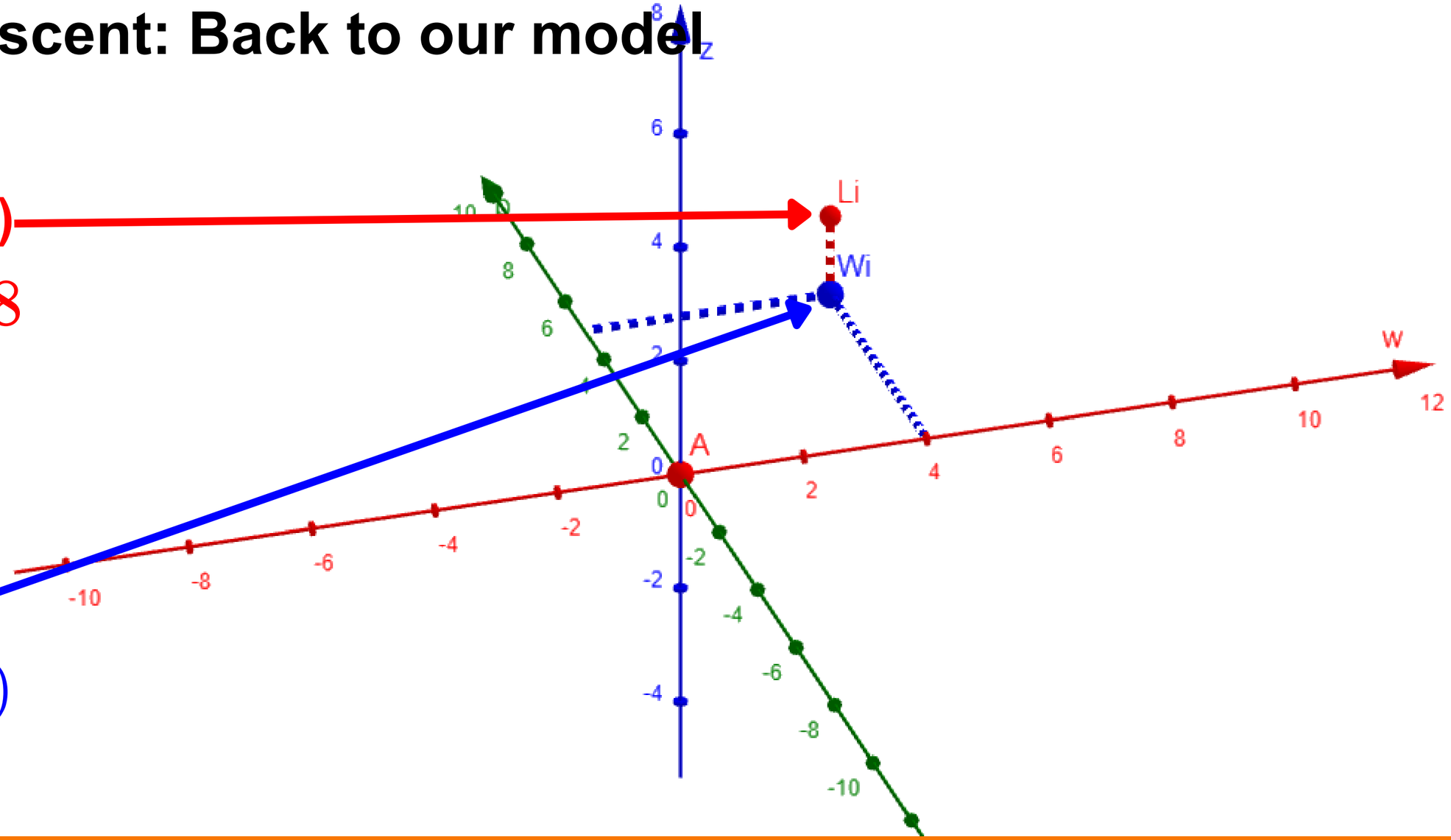
$$\begin{cases} \textbf{Wi(4,5,0)} \\ w_0 = (4, 5) \end{cases}$$



Gradient Descent: Back to our model

Li(4,5,1.38)

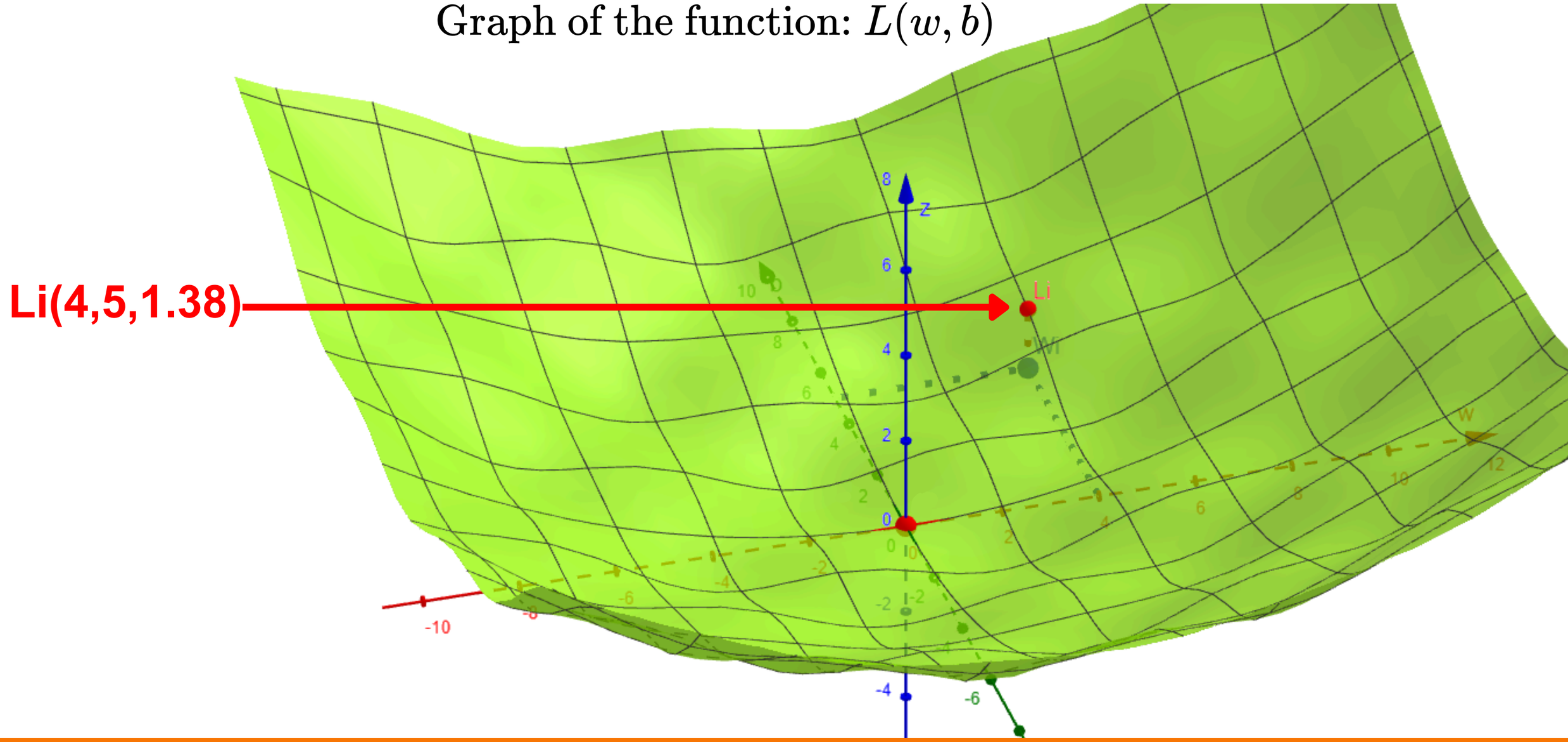
$L_0 = 1.38$

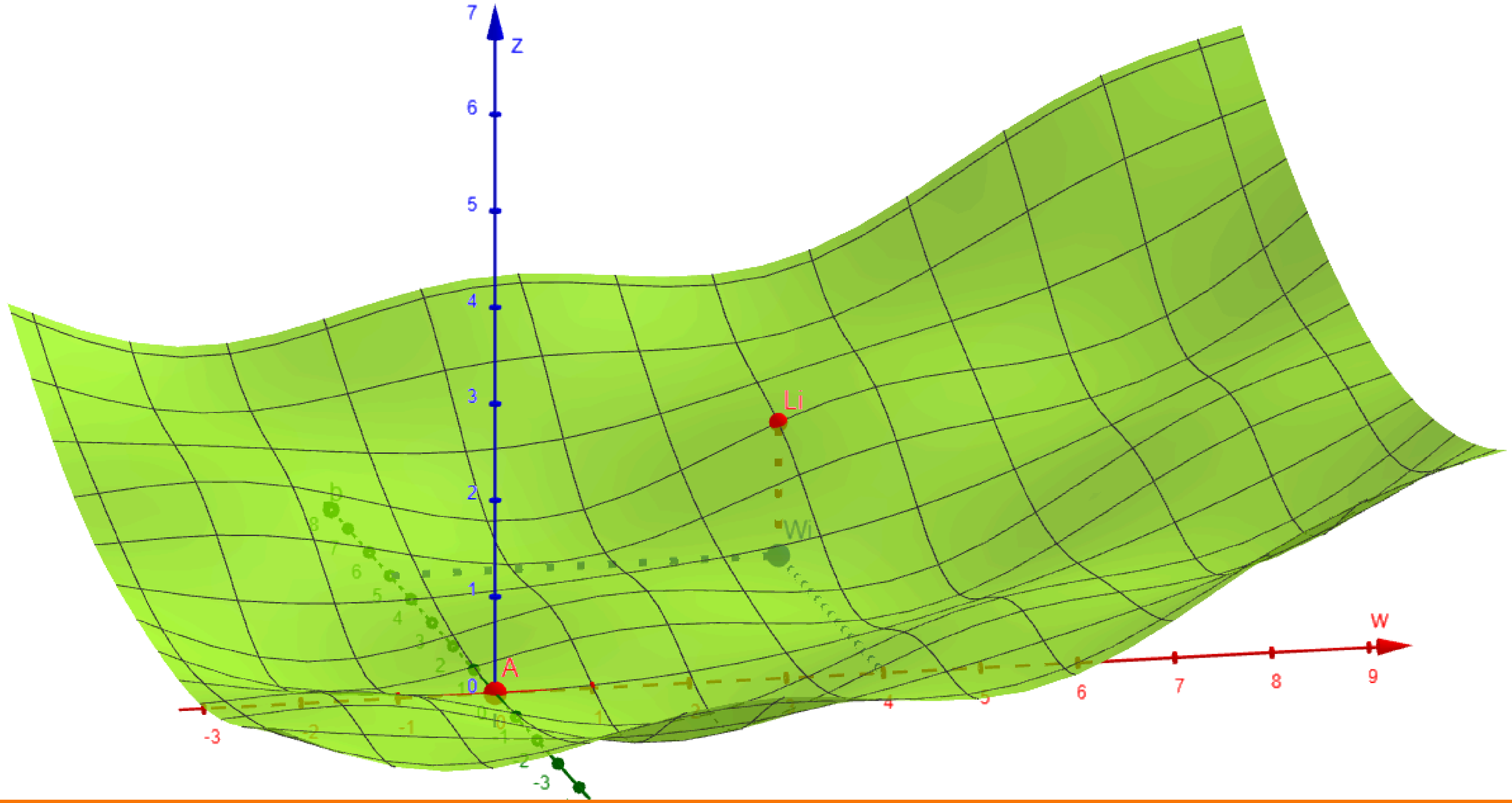


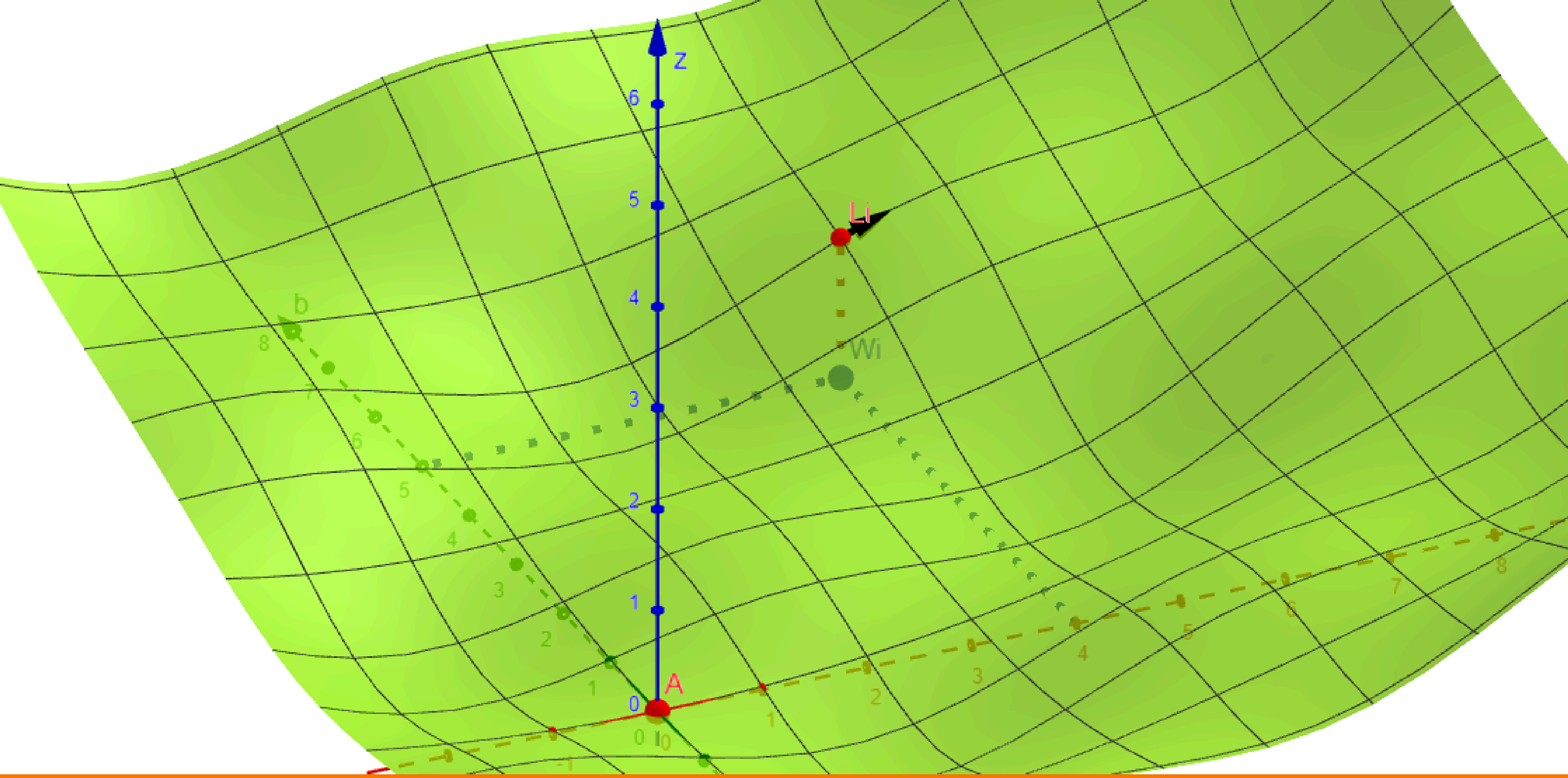
Wi(4,5,0)

$w_0 = (4, 5)$

Graph of the function: $L(w, b)$




z
7
6
5
4
3
2
1
0
Li
Wi
A
w
-3
-1
1
2
3
4
6
7
8
9


z
6
5
4
3
2
1
0
A
Li
Wi
b
8
7
6
5
4
3
2
1
0
-1
0
1
2
3
4
5
6
7
8

$$\frac{\partial L}{\partial w} = L_w(w, b)$$

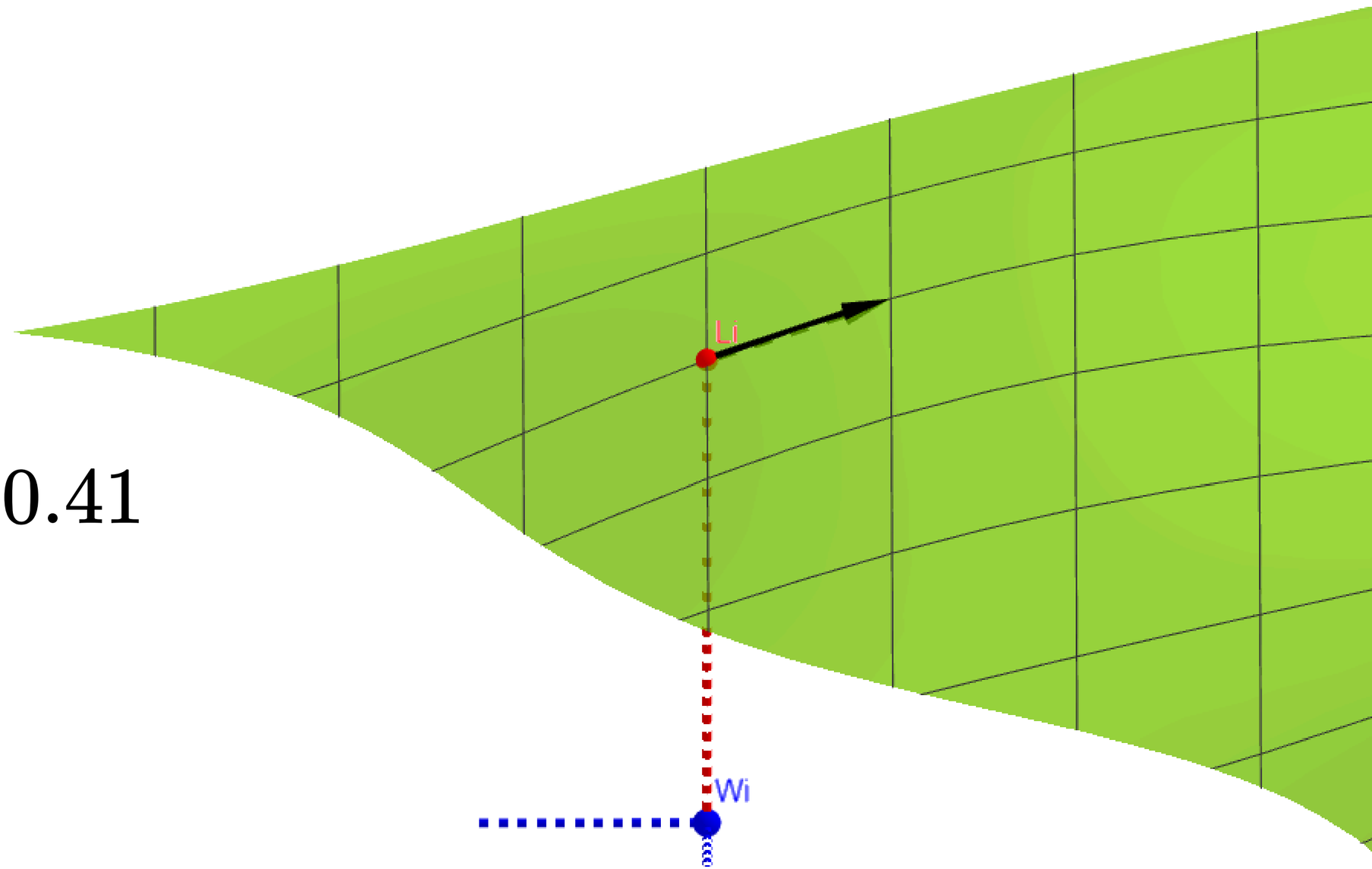
$$\frac{\partial L}{\partial b} = L_b(w, b)$$

To start applying our algorithm,

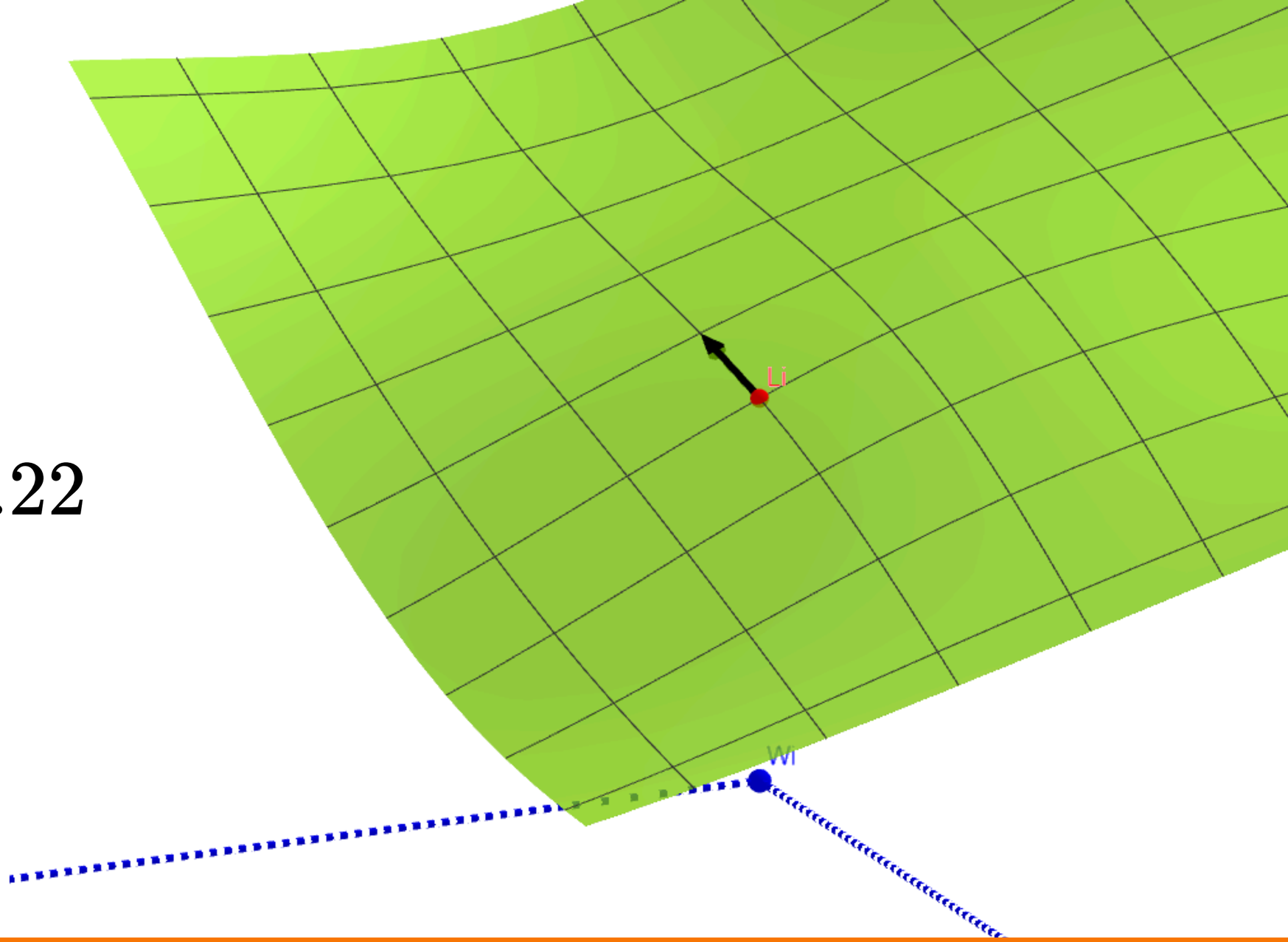
We need to calculate $L_w(w_0, b_0)$

and $L_b(w_0, b_0)$

$$L_w(w_0, b_0) = 0.41$$



$$L_b(w_0, b_0) = 0.22$$



Gradient Descent:

After calculating the initial partial derivatives,

$$L_w(w_0, b_0) = 0.41$$

$$L_b(w_0, b_0) = 0.22$$

We need to update the weigh and bias via the instructions:

$$w_1 = w_0 - \alpha \times L_w(w_0, b_0)$$

Gradient Descent:

$$w_1 = w_0 - \alpha \times L_w(w_0, b_0)$$

$$b_1 = b_0 - \alpha \times L_b(w_0, b_0)$$

α Is known as the learning rate, that specifies the step size at each iteration of the gradient descent algorithm. It is a hyper-parameter.

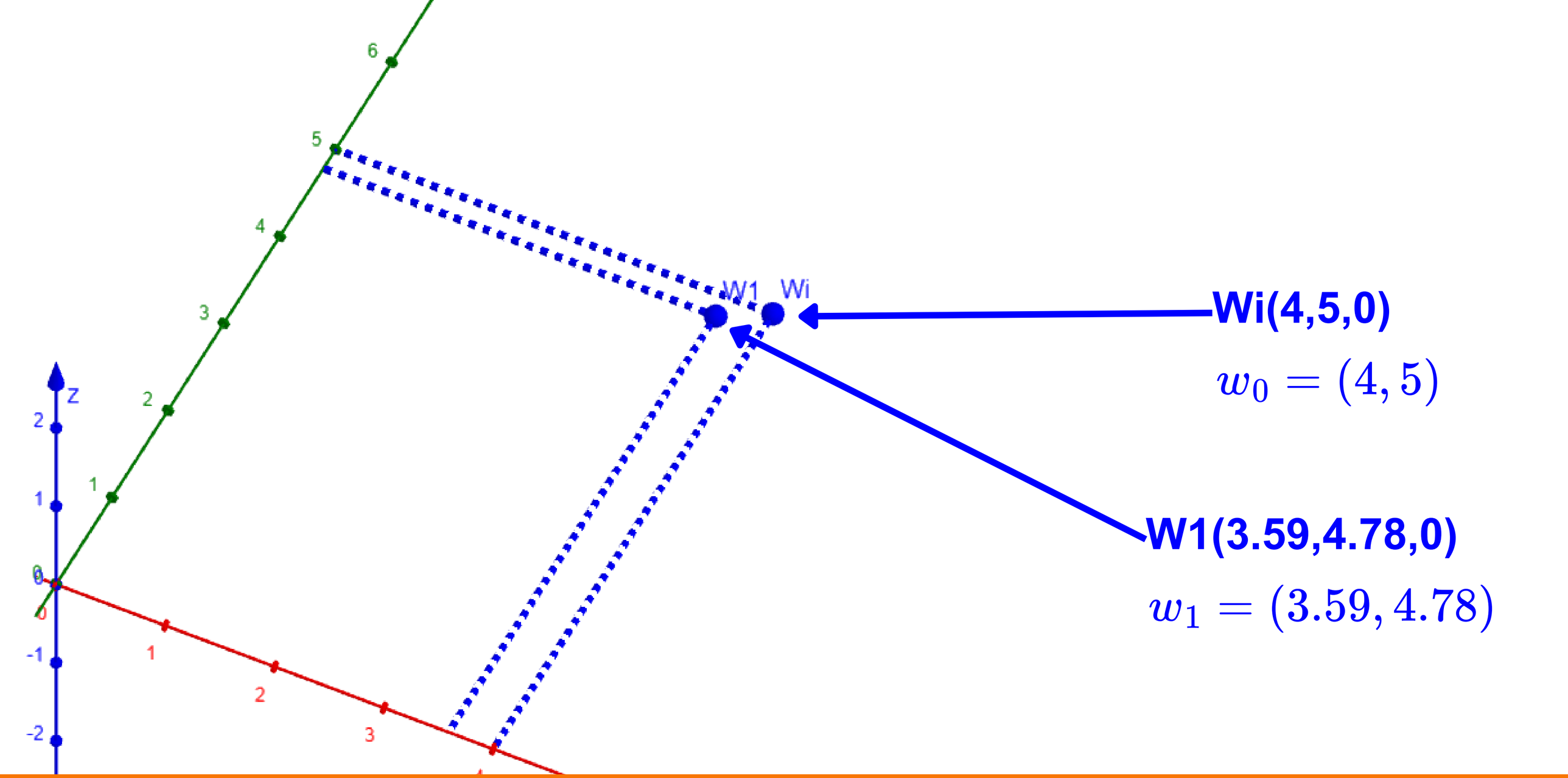
We usually set α equal to small values,

In our case, we're going to choose $\alpha = 1$ for simplifying purposes.

Gradient Descent:

$$w_1 = w_0 - \alpha \times L_w(w_0, b_0)$$
$$= 4 - 0.41 = 3.59$$

$$b_1 = b_0 - \alpha \times L_b(w_0, b_0)$$
$$= 4 - 0.22 = 4.78$$

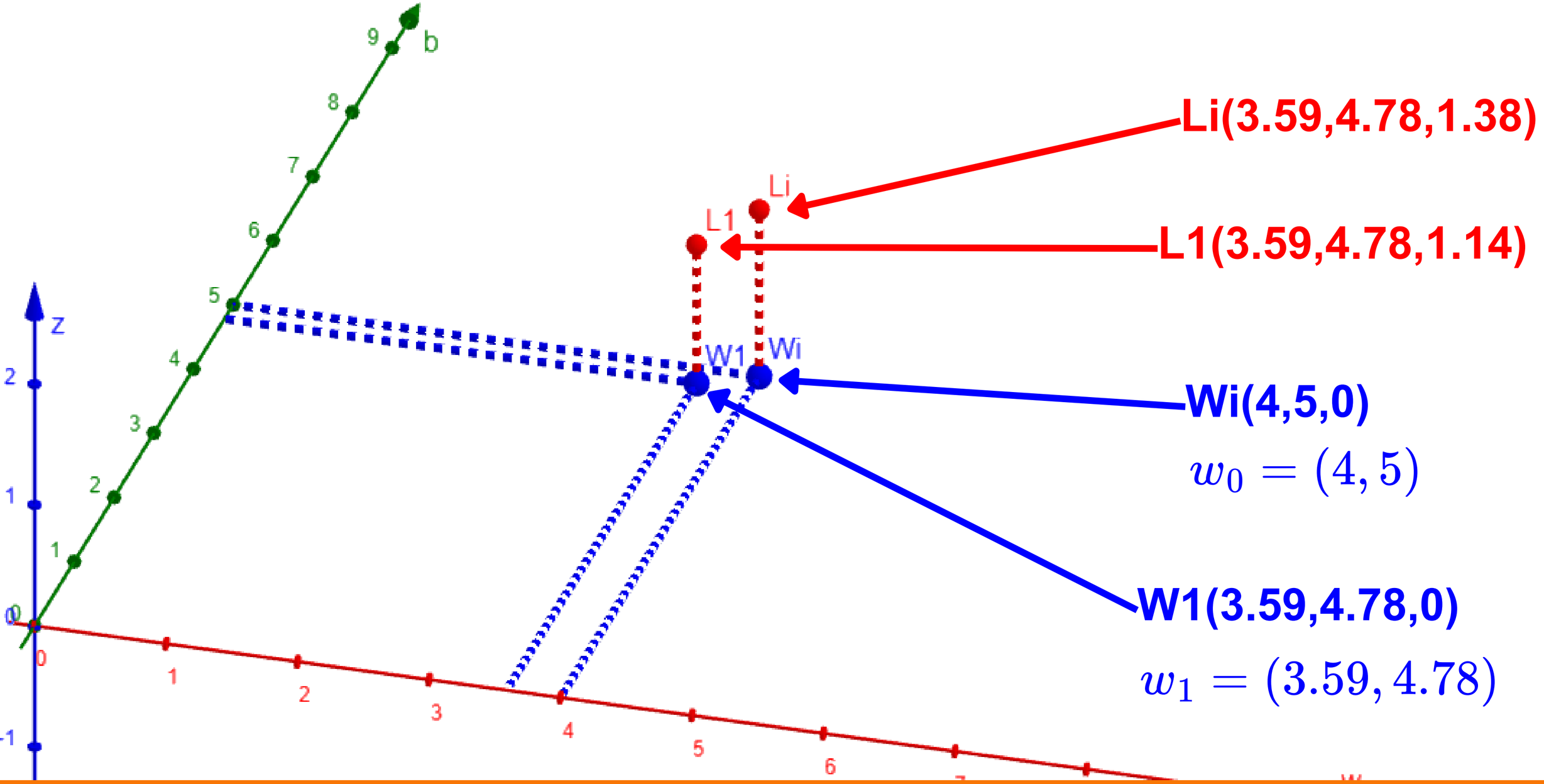


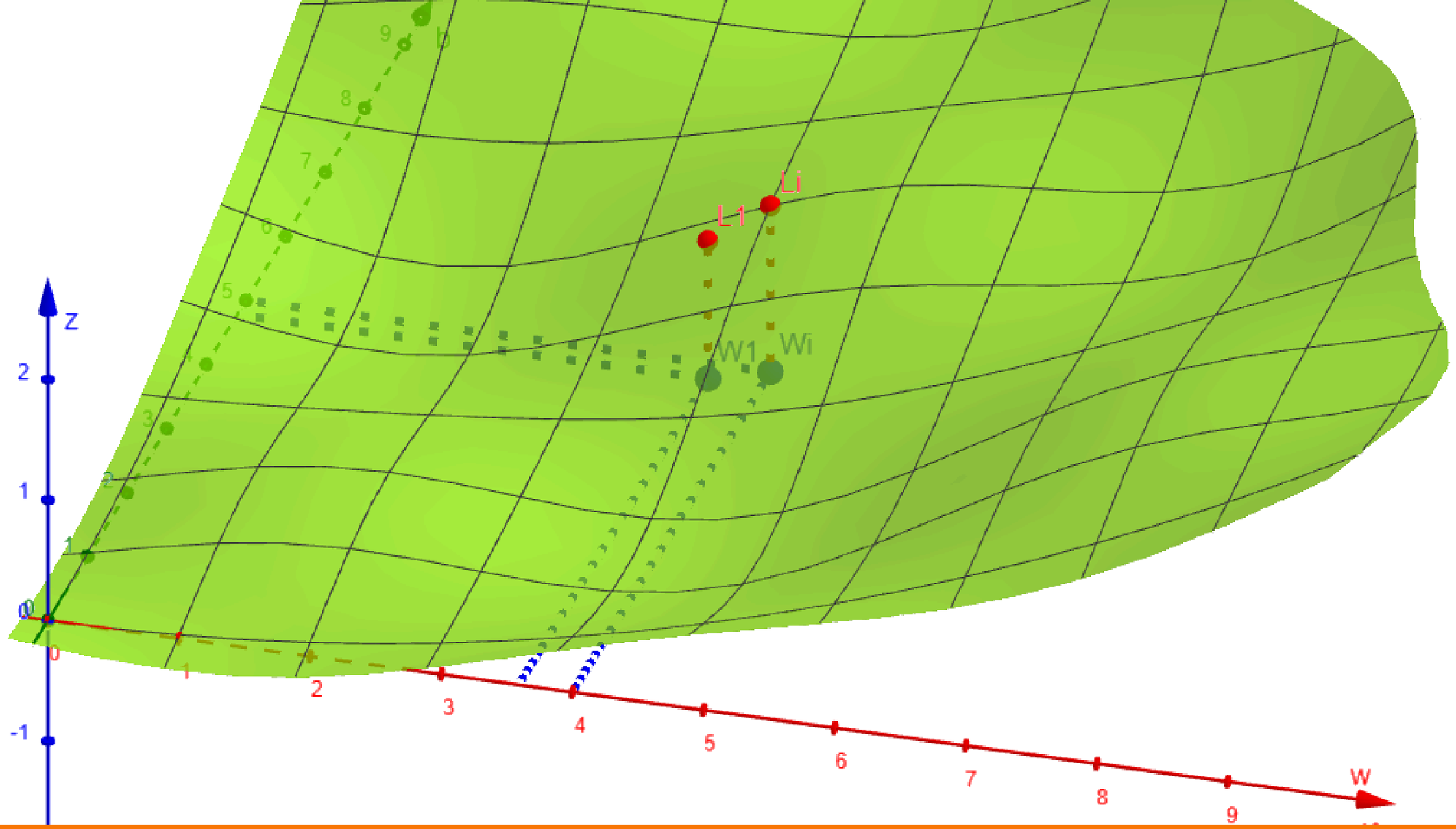
Wi(4,5,0)

$$w_0 = (4, 5)$$

W1(3.59,4.78,0)

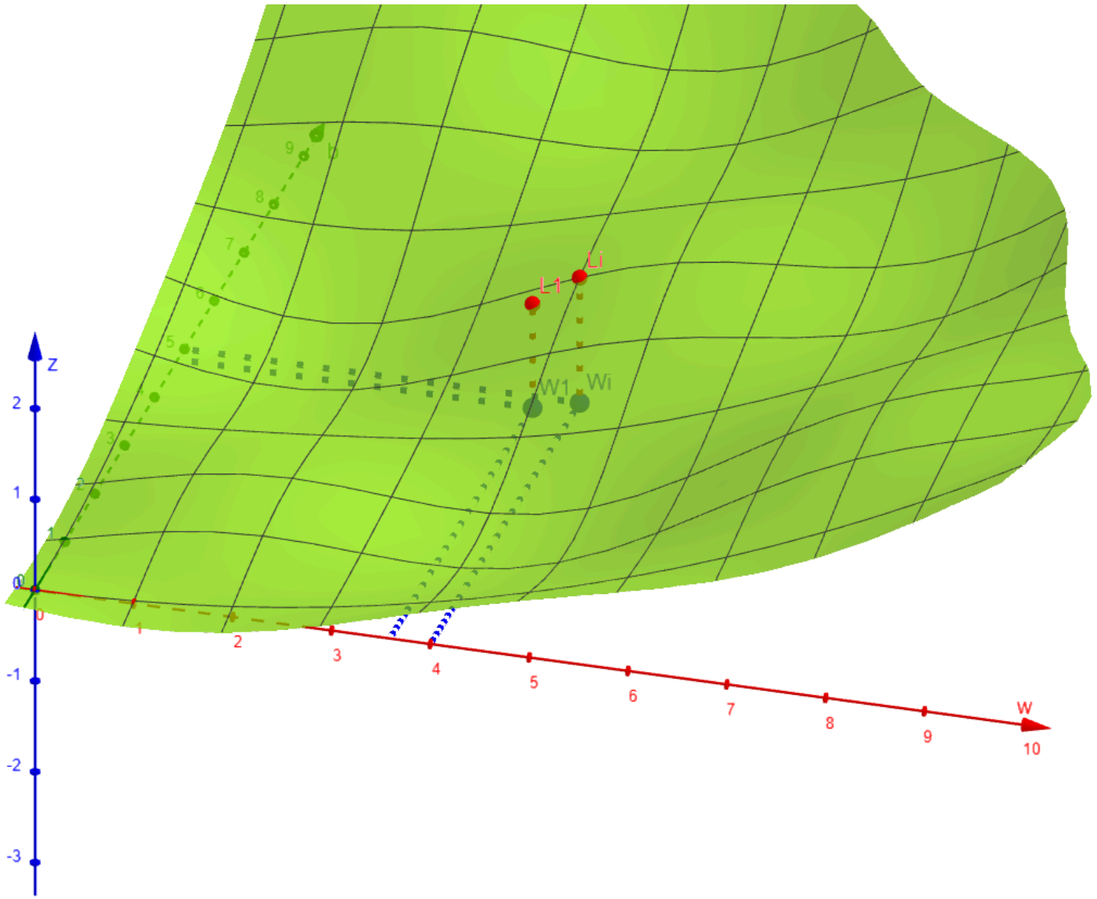
$$w_1 = (3.59, 4.78)$$

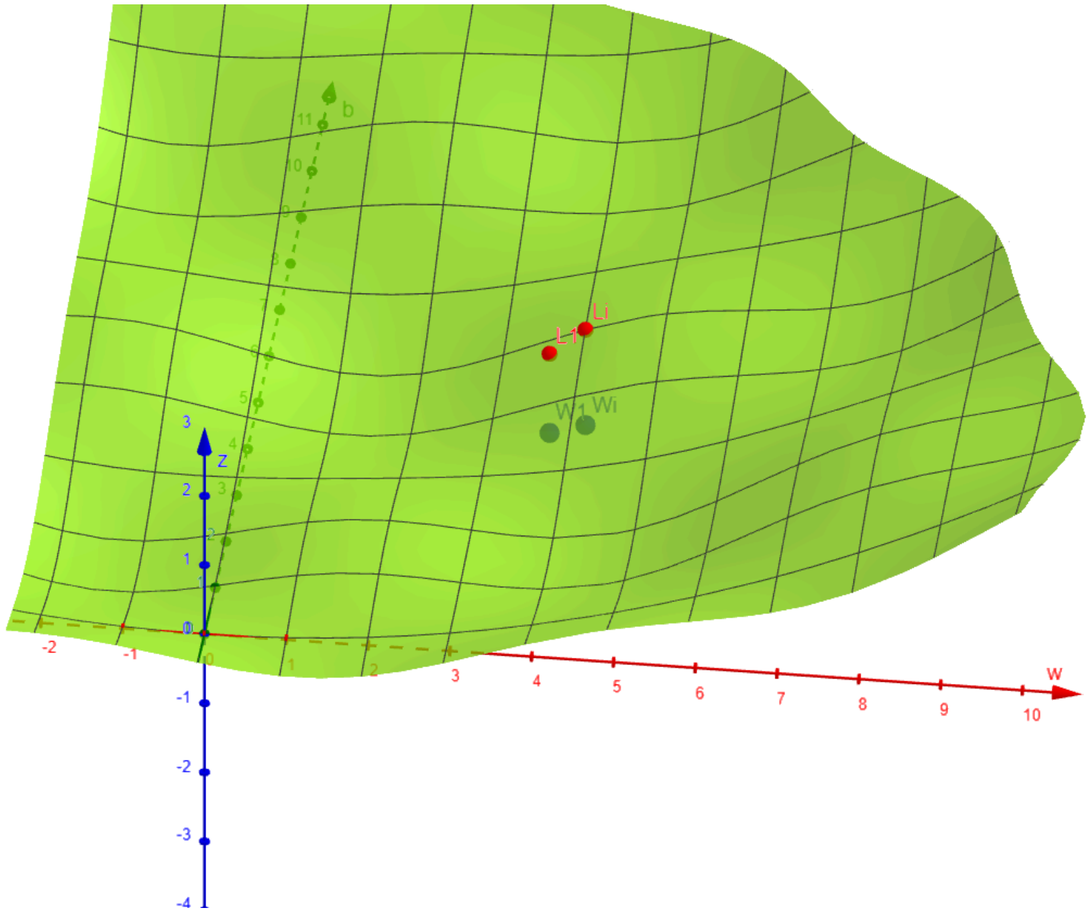

b
z
Li
L1
W1
Wi
Li(3.59,4.78,1.38)
L1(3.59,4.78,1.14)
Wi(4,5,0)
$w_0 = (4, 5)$
W1(3.59,4.78,0)
$w_1 = (3.59, 4.78)$

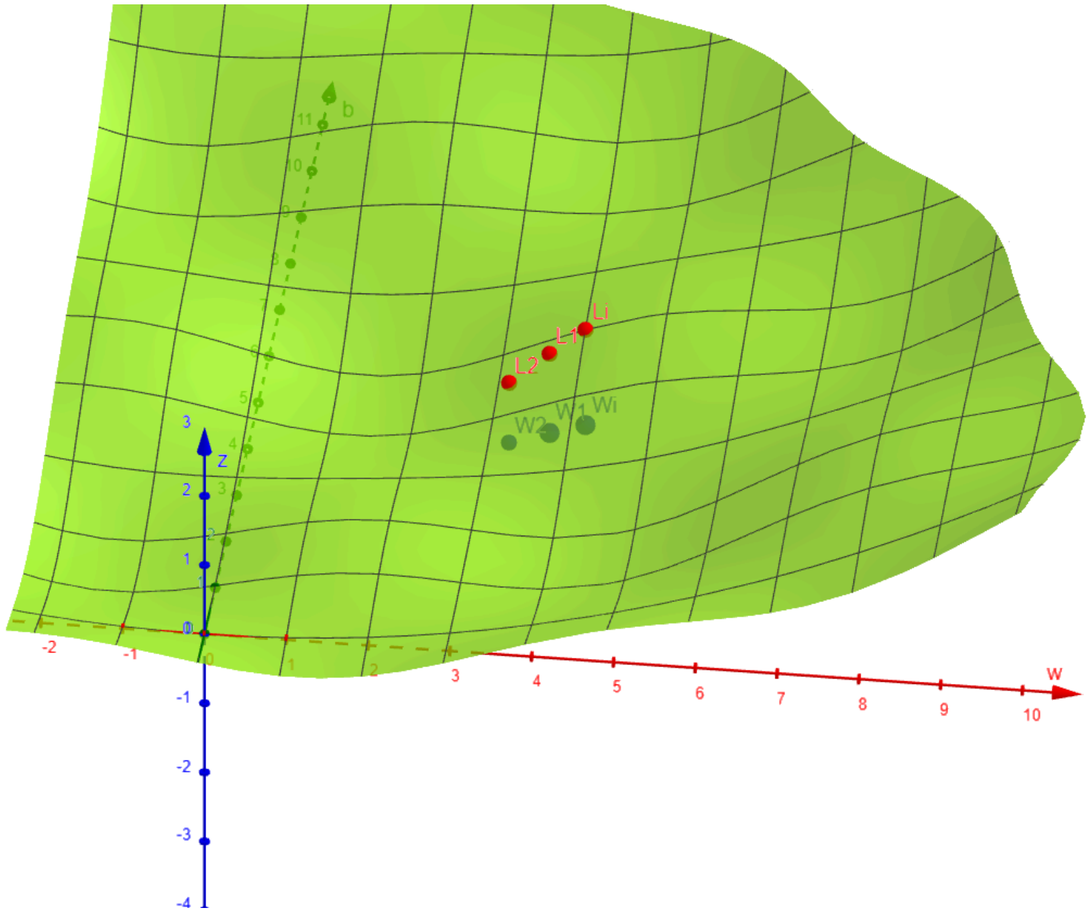

Z
2
1
0
-1
0
1
2
3
4
5
6
7
8
9
W
L1
Li
W1
Wi

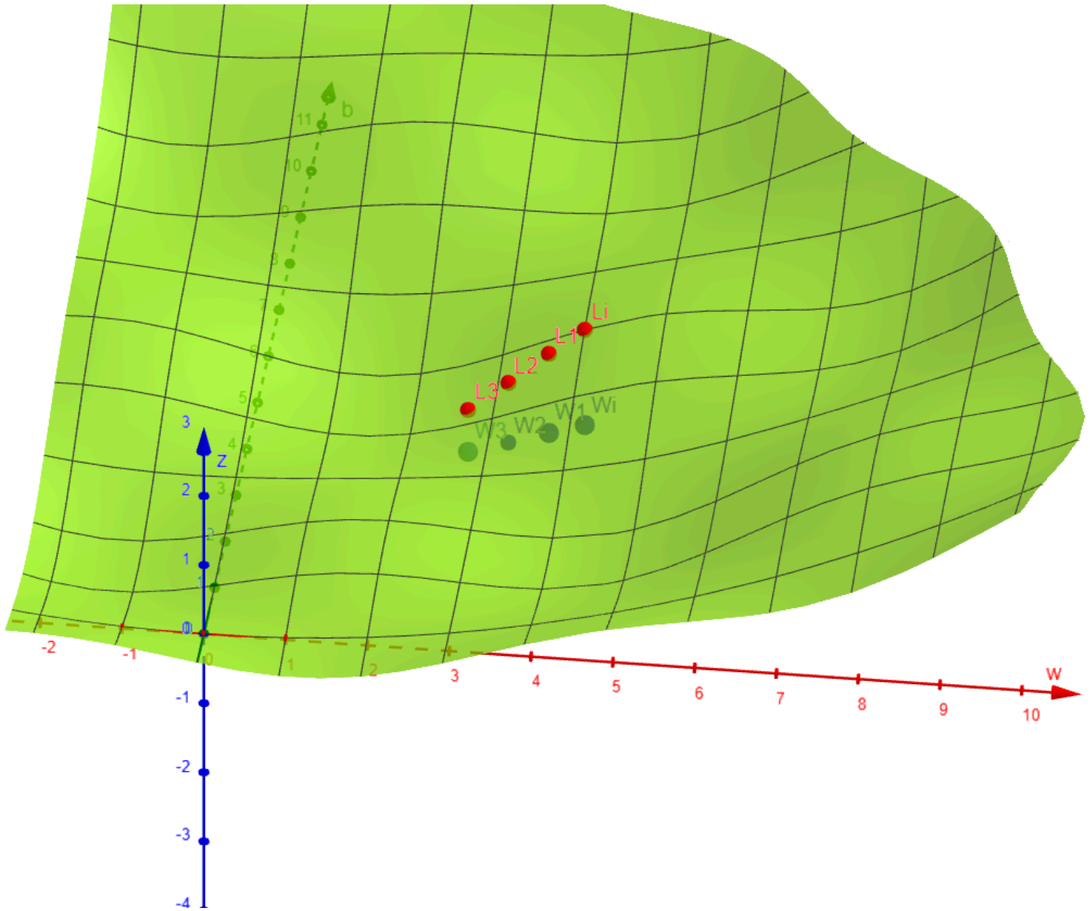
Gradient Descent:

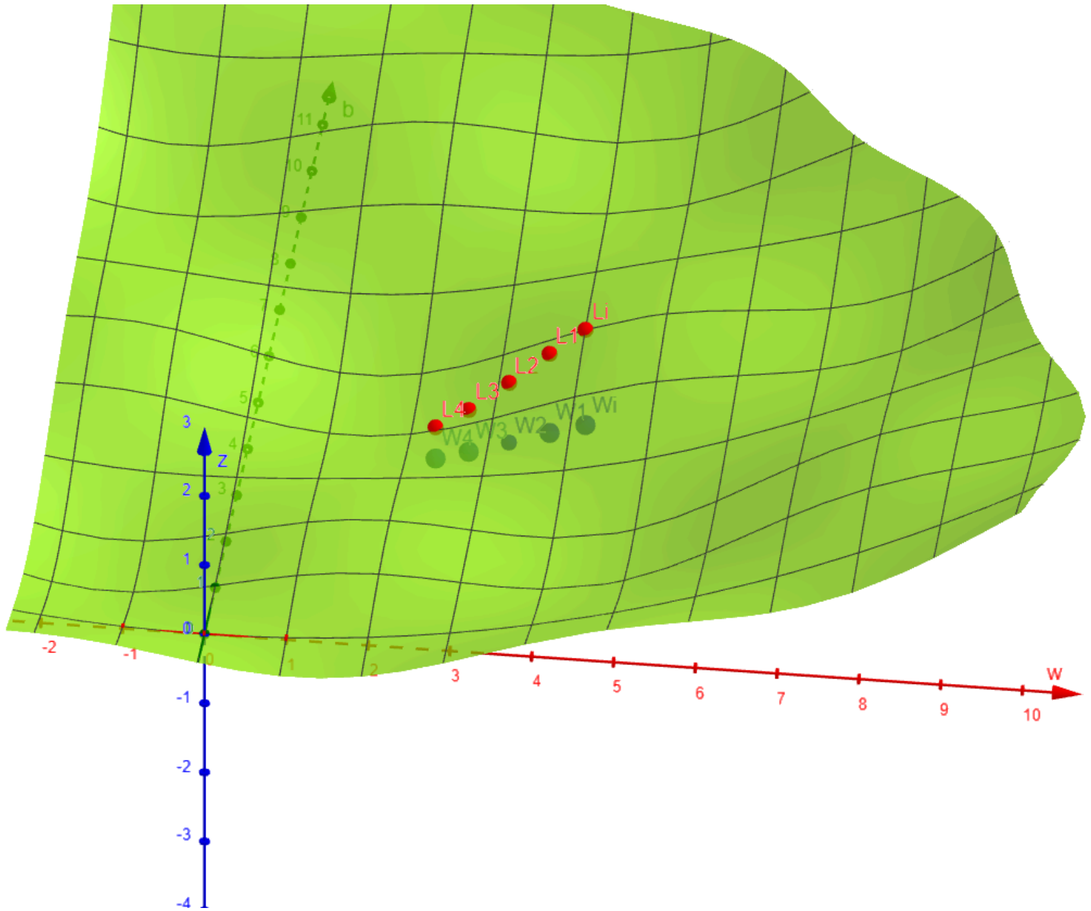
As we continue to iterate and adjust our parameters, we gradually approach the lowest point on the loss graph, known as the Minimum Loss. At this stage, the weight and bias values have been optimized, meaning they now yield the best possible fit for our model.

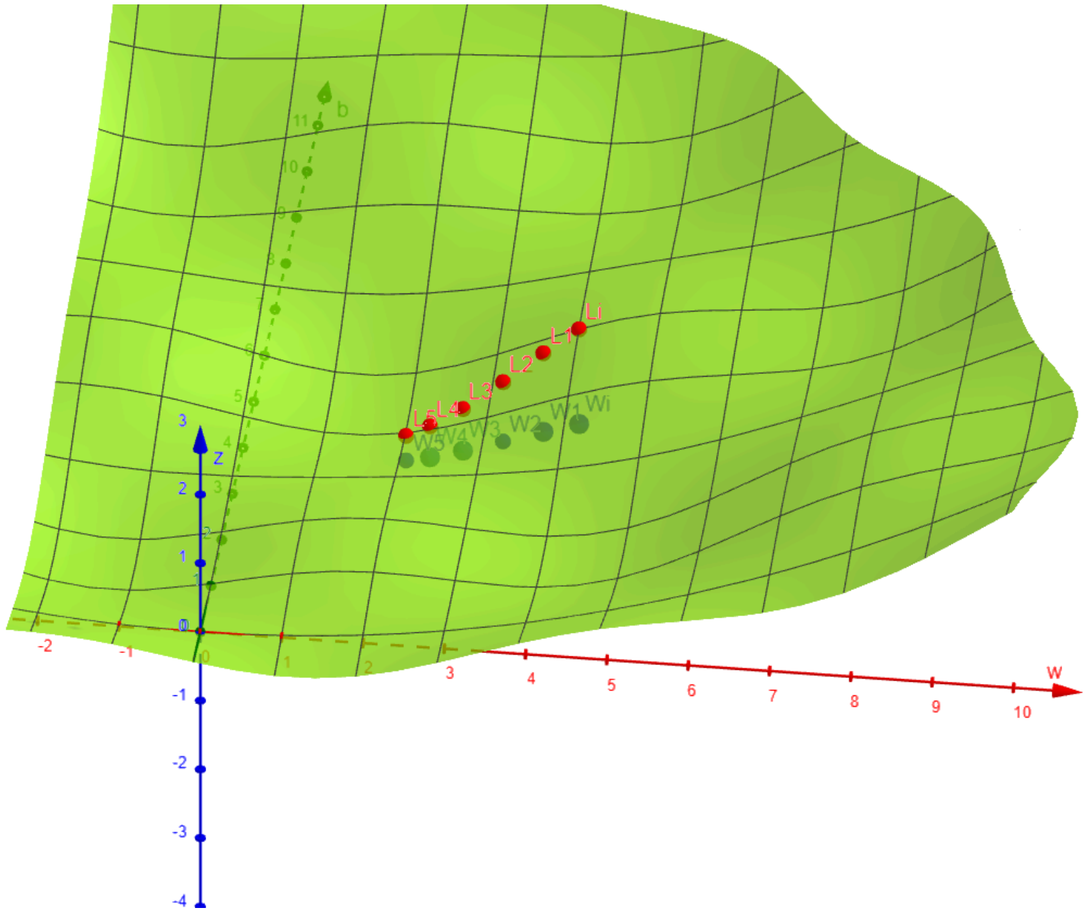

Z
2
1
0
-1
-2
-3
L1
Li
W1
Wi
W
10

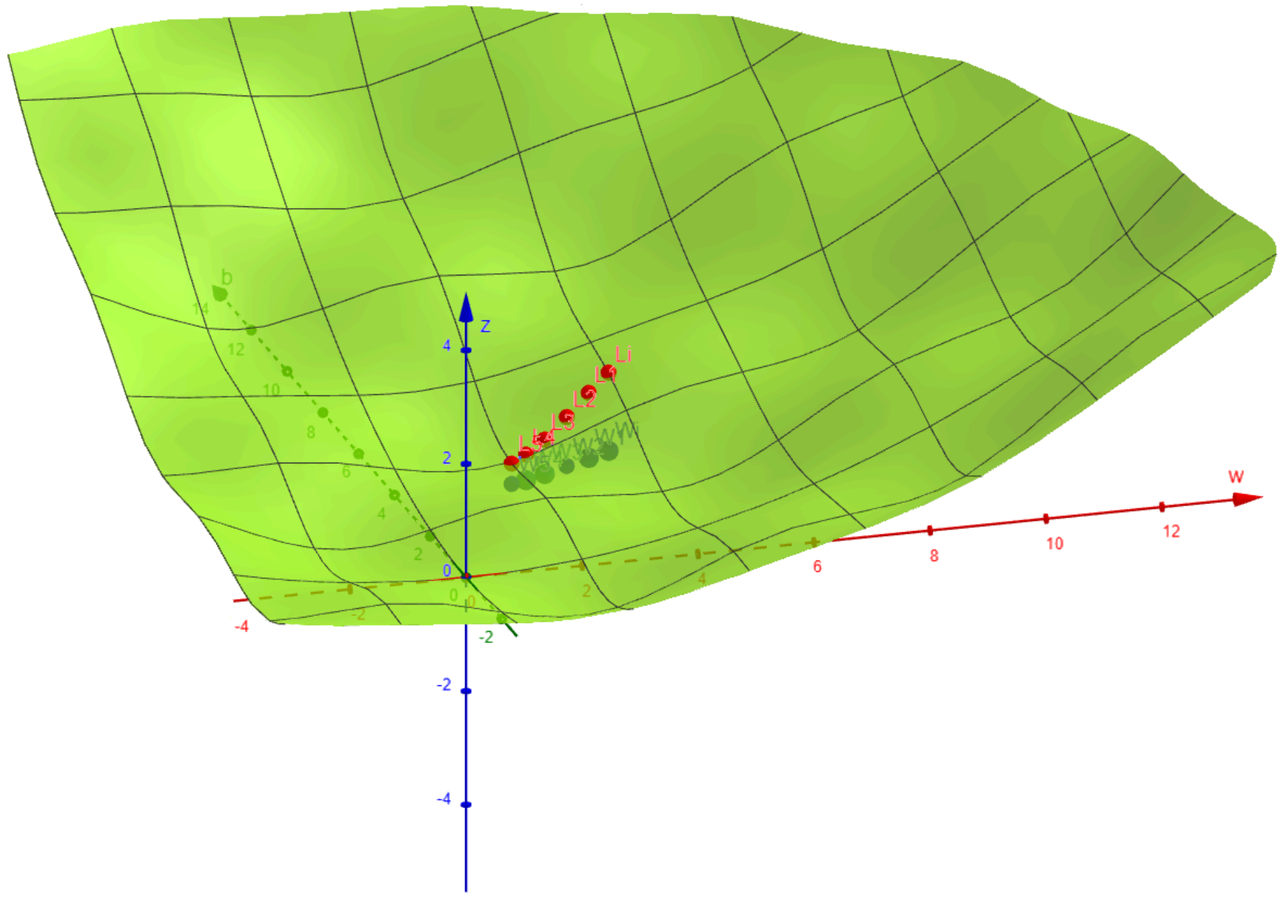

b
Li
L-1
W1
Wi
z
w
-2
-1
0
1
2
3
4
5
6
7
8
9
10
-4
-3
-2
-1
0
1
2
3


b
L1
L1
L2
W2
W1
Wi
z
w


b
Li
L1
L2
L3
W1
Wi
W2
W3
z
W
-2
-1
0
1
2
3
4
5
6
7
8
9
10
-1
-2
-3
-4

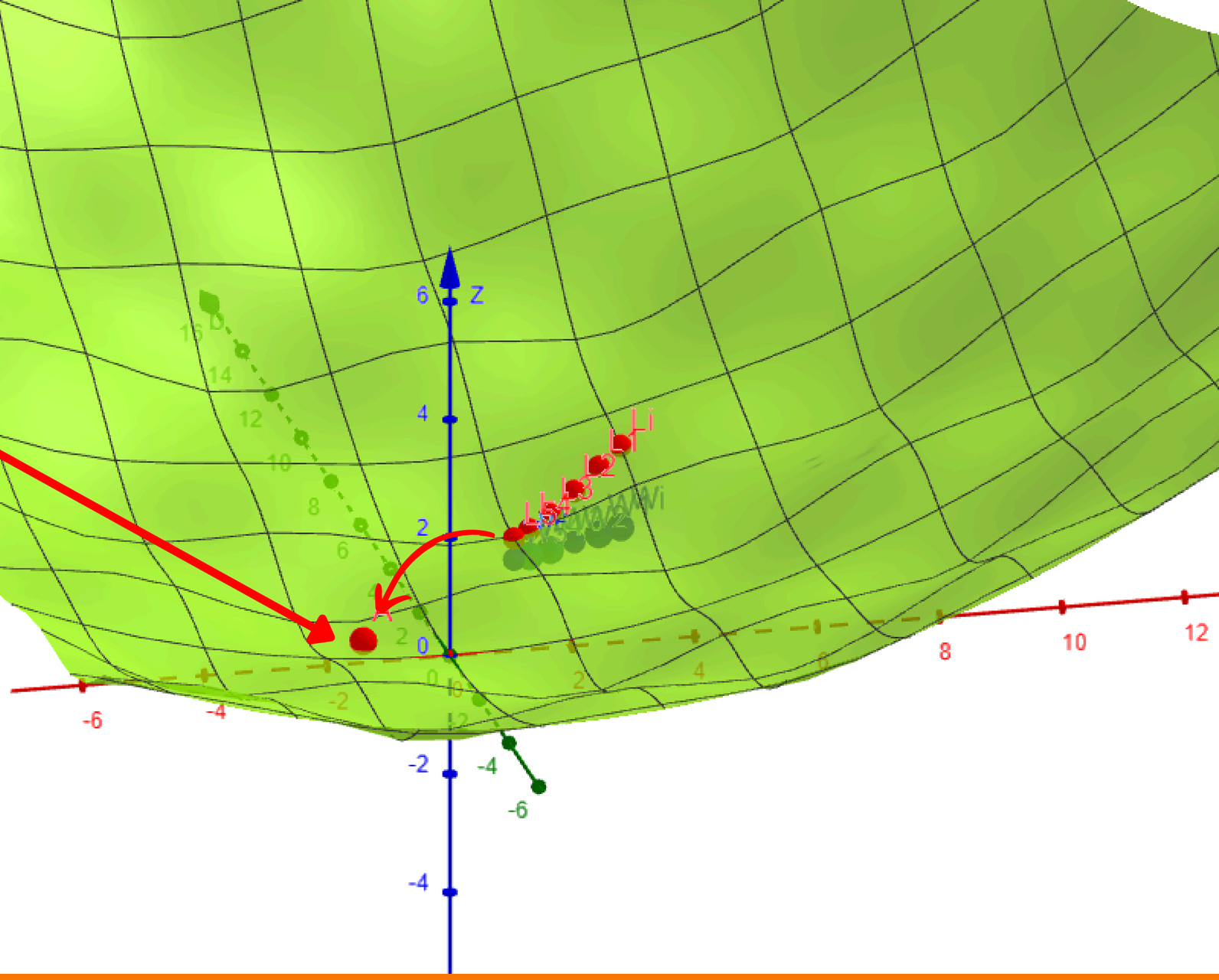

b
Li
L1
L2
L3
L4
Wi
W1
W2
W3
W4
z
w


b
Li
L1
L2
L3
L4
L5
W1
W2
W3
W4
W5
Wi
Z
W
-2
-1
1
2
3
4
5
6
7
8
9
10
-1
-2
-3
-4


z
w
b
Li
4
2
0
-2
-4
6
8
10
12

Global Minimum

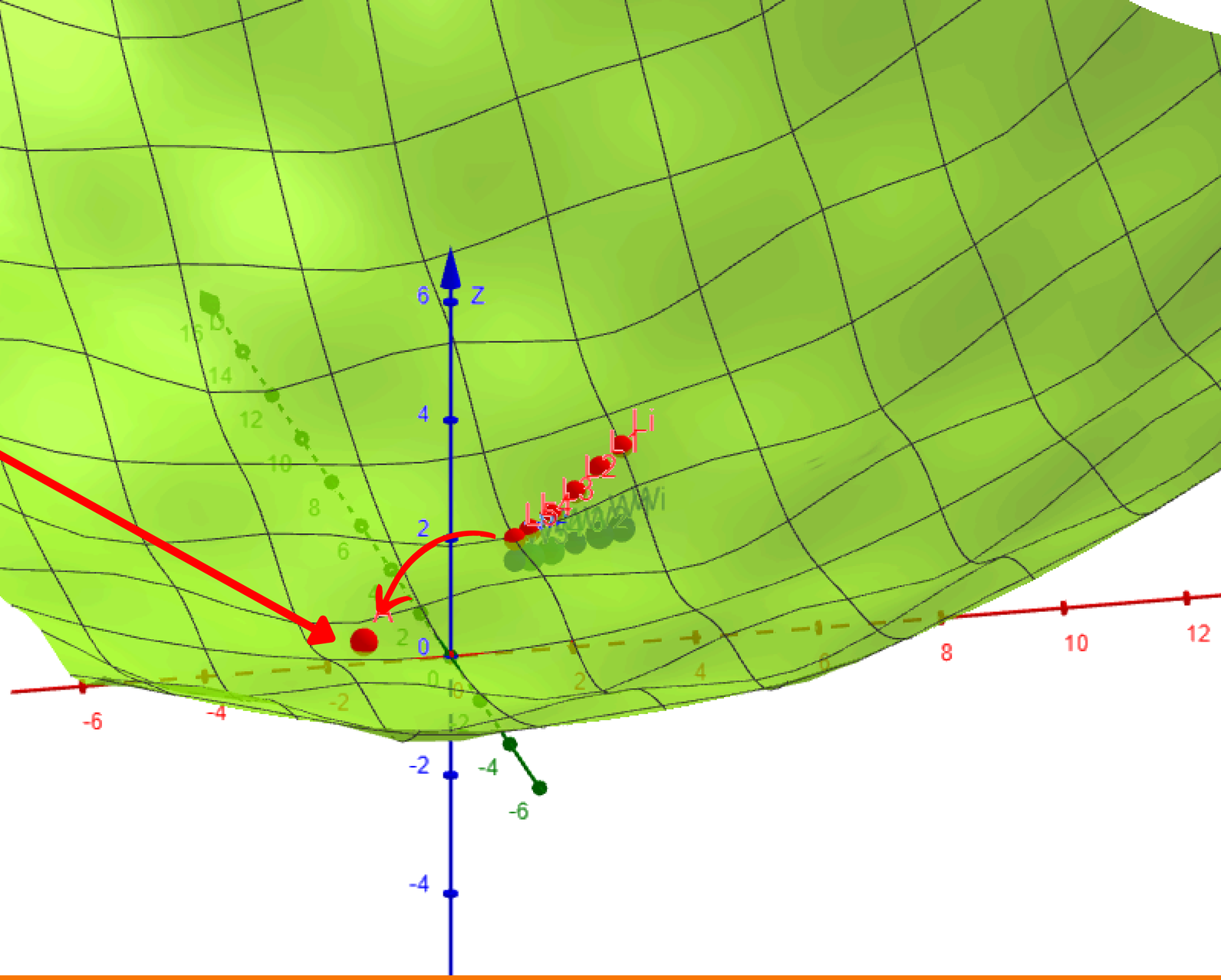
(In this case)



Global Minimum

(In this case)

The coordinates of A are the optimal parameters for our model.

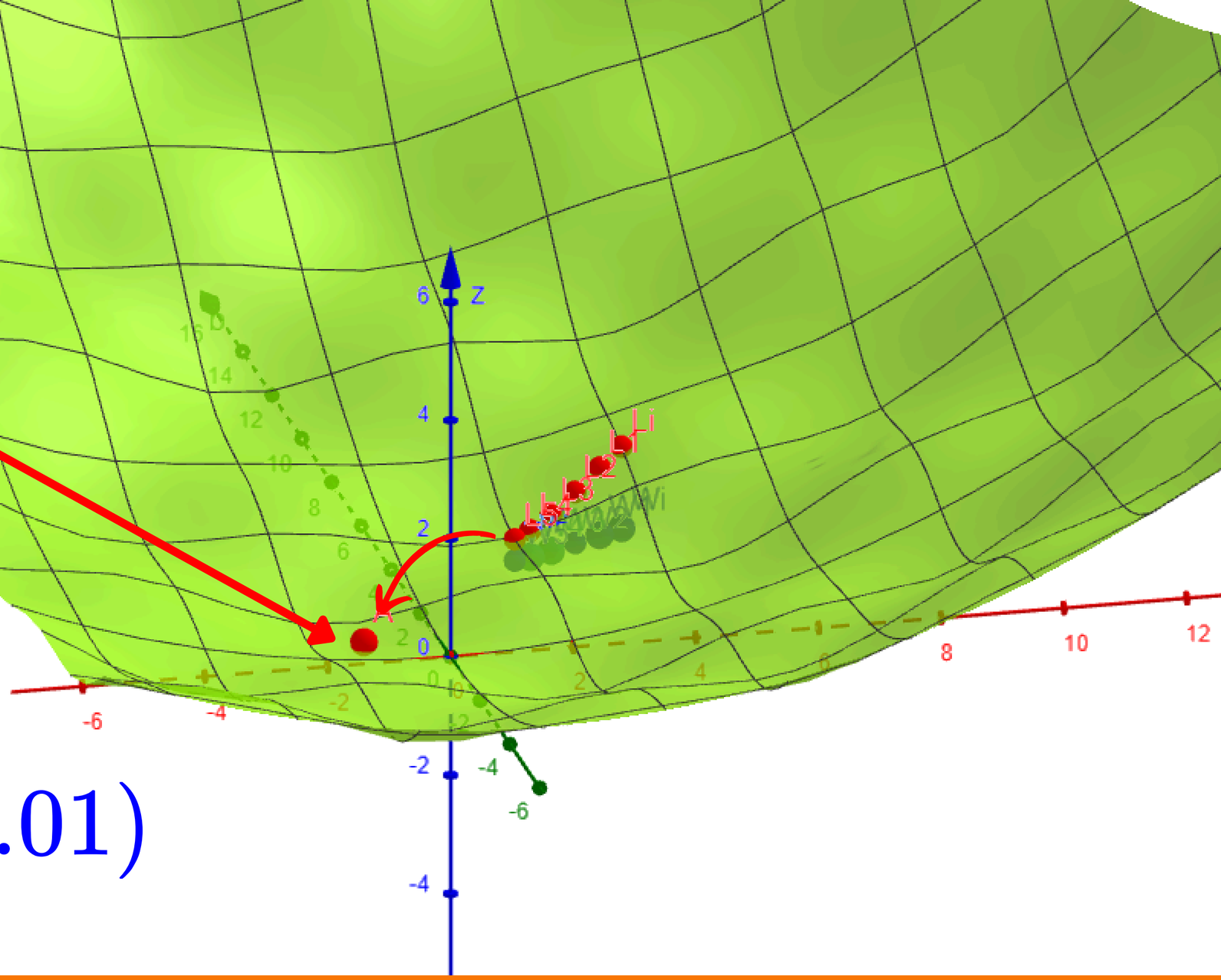


Global Minimum

(In this case)

The coordinates of A are the optimal parameters for our model.

$$A\left(-1.18, 0.94, 0.01\right)$$



Gradient Descent

$$A\left(-1.18, 0.94, 0.01\right)$$

So after training our model, we end up with:

$$w_f = -1.18$$

$$b_f = 0.94$$

$$Loss = 0.01$$