GO CLASSES

Gradient Descent

Gradient Descent is one method for solving an optimization problem.

$f(x) = x^2$

$\min f(x)$

we already know how to find minima / maxima.

Why We Need Gradient Descent

From calculus, we know that the minimum of f must lie at a point where its derivative vanishes, i.e. $\frac{\partial f(\theta^*)}{\partial \theta} = 0$.

- Sometimes, we can solve this equation analytically for θ.
- Mostly, we are not so lucky and must resort to iterative methods.

Recall the Gradient:

$$\nabla_\theta f = \left(\frac{\partial f(\theta)}{\partial \theta_1}, \frac{\partial f(\theta)}{\partial \theta_2}, \ldots, \frac{\partial f(\theta)}{\partial \theta_n}\right)$$

$$f(x) = x^2 \cdot \log_e(x) - x$$

$$f'(x) = 2x \cdot \log_e(x) + x - 1 = 0 \quad \Rightarrow \quad x = \ldots\ldots$$

$\Rightarrow \quad 2x \ln x + x = 1$

$x =$

Not able to find closed form.

Closed Form Solutions are not always available

- Often it is not possible to simply solve $\nabla_X f(X) = 0$
  - The function to minimize/maximize may have an intractable form
- In these situations, iterative solutions are used
  - Begin with a "guess" for the optimal X and refine it iteratively until the correct value is obtained

https://deeplearning.cs.cmu.edu/F22/document/slides/lec4.learning.pdf

Solving the equation $f'(x) = 0$ for x provides an *analytical* solution for a critical point. Unfortunately, it is not always possible to compute such analytical solutions in closed form. It is often difficult to exactly solve the equation $f'(x) = 0$ because this derivative might itself be a complex function of x. In other words, a *closed form solution* (like the example above) typically does not exist. For example, consider the following function that needs to be minimized:

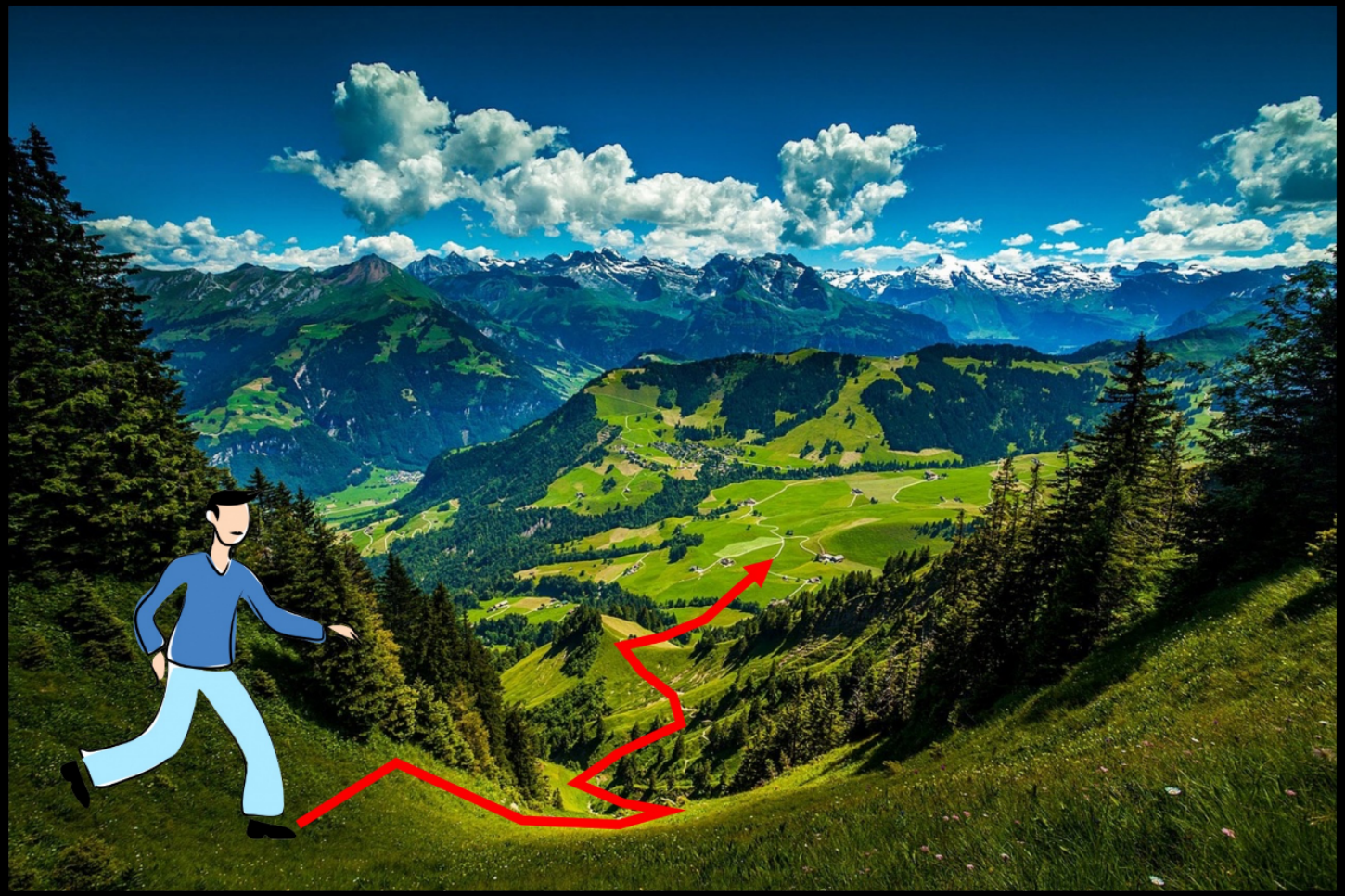
$$f(x) = x^2 \cdot \log_e(x) - x \tag{4.5}$$

Setting the first derivative of this function to 0 yields the following condition:

$$f'(x) = 2x \cdot \log_e(x) + x - 1 = 0$$

This equation is somewhat hard to solve, although iterative methods exist for solving it. By trial and error, one might get lucky and find out that $x = 1$ is indeed a solution to the first-order optimality condition because it satisfies $f'(1) = 2\log_e(1) + 1 - 1 = 0$. Furthermore, the second derivative $f''(x)$ can be shown to be positive at $x = 1$, and therefore this point is at least a local minimum. However, solving an equation like this numerically causes all types of numerical and computational challenges; these types of challenges increase when we move from univariate optimization to multivariate optimization.

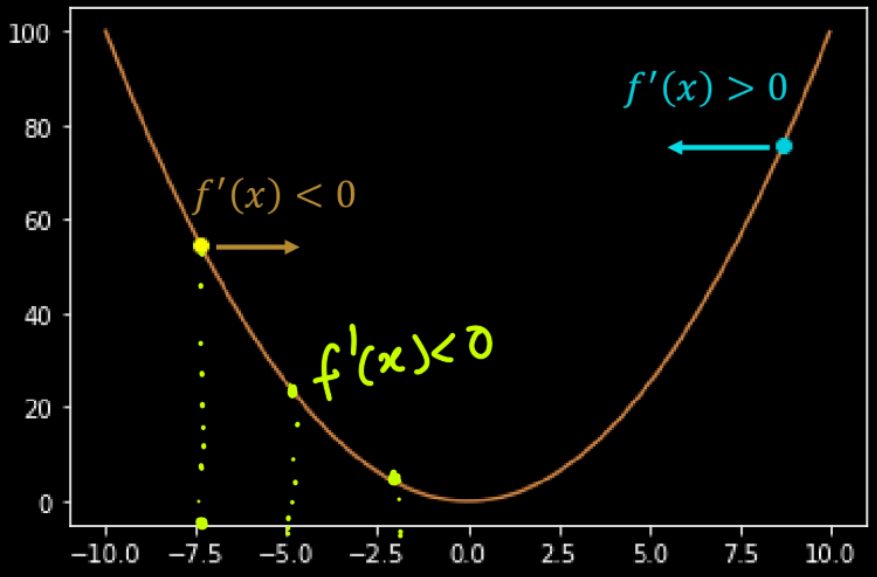
Calculus



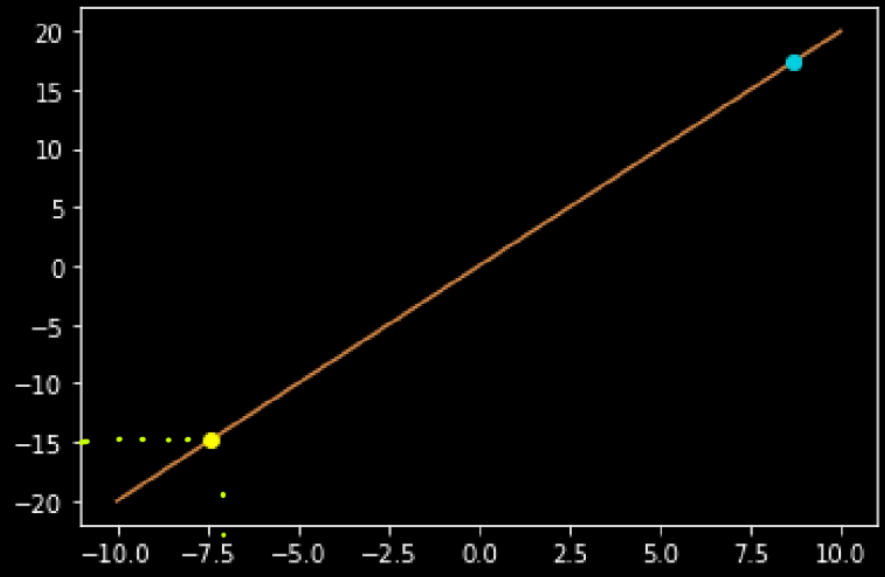
1D Gradient Descent

1D Gradient Descent

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



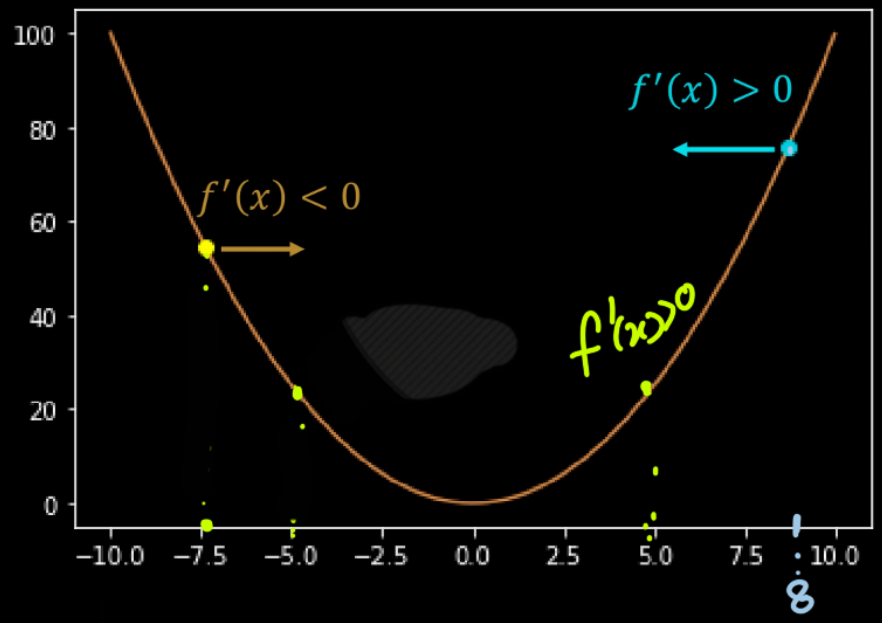
$$f'(x) = 2x$$



$-7.5 \rightsquigarrow -5.0 \rightsquigarrow -2 \rightsquigarrow -1 \rightsquigarrow 0$

1D Gradient Descent

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



$8 \longrightarrow 5 \longrightarrow 2 \longrightarrow 1 \longrightarrow 0$

$x_1 = -7.5$

$x_{new} = x_1 + 2.5$

$= -7.5 + 2.5$

$= -5.0$

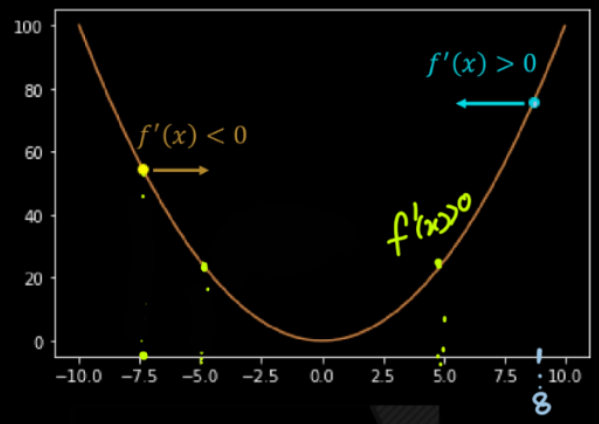
$x_1 = 8$

$x_{new} = x_1 - 2.5$

$= 5.5$

1D Gradient Descent

$f(x) = x^2$



$8 \rightarrow 5 \rightarrow 2 \rightarrow 1 \rightarrow 0$

$x_1 = -7.5$

$x_{new} = x_1 + 2.5$

$= -7.5 + 2.5$

$= -5.0$

$x_1 = 8$

$x_{new} = x_1 - 2.5$

$= 5.5$

Step size or learning rate

$x_{new} = x_1 - \alpha f'(x)$

$x_{new} = x_1 - \frac{1}{10}(-15)$

$= -7.5 + 1.5$

$-f'(x)$ ← direction of steepest decrease

$x_{new} = x_1 - \alpha f'(x)$

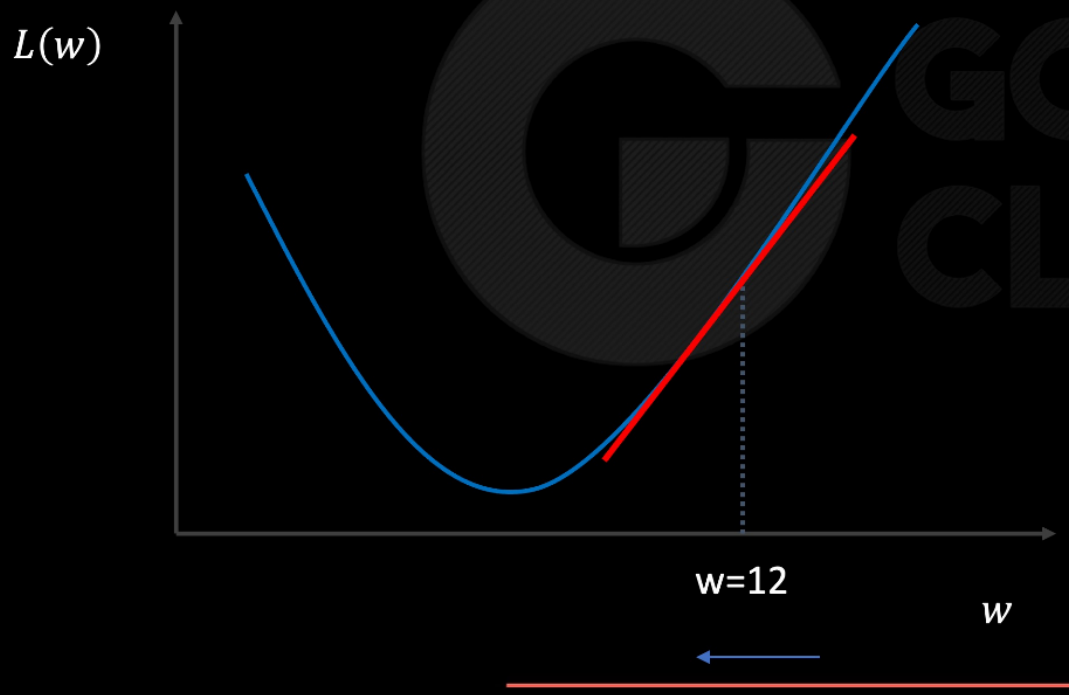
→ in which direction we should go (to decrease the function)

→ How much to go in that direction.

→ in which direction we should go (to decrease the function) $-f'(x)$

→ How much to go in that direction. α

Gradient Descent (GD) (idea)



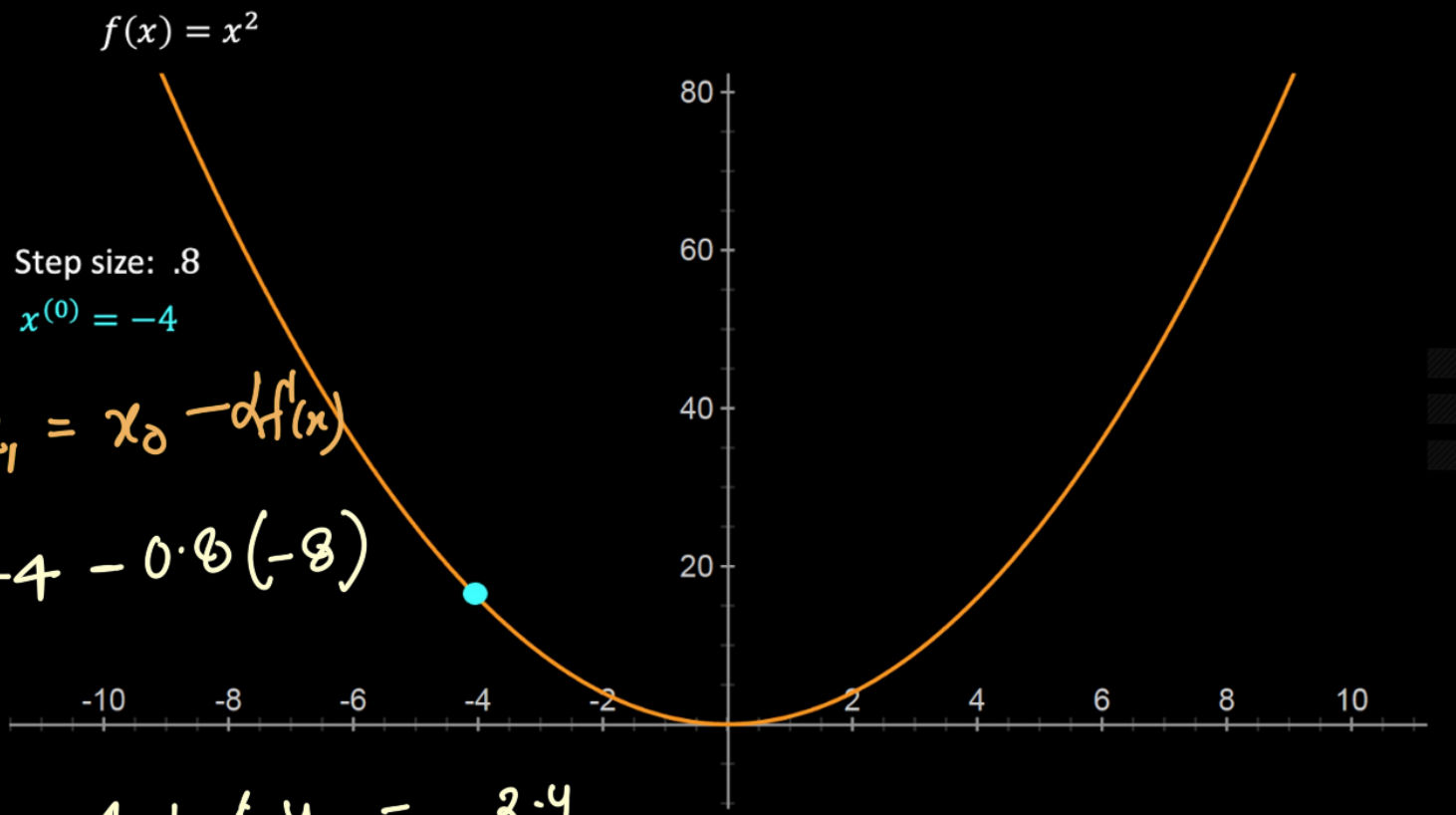
1. Start with a random value of w (e.g. w = 12)

2. Compute the gradient (derivative) of L(w) at point w = 12. (e.g. dL/dw = 6)

3. Recompute w as:

w = w – lambda * (dL / dw)

Gradient Descent

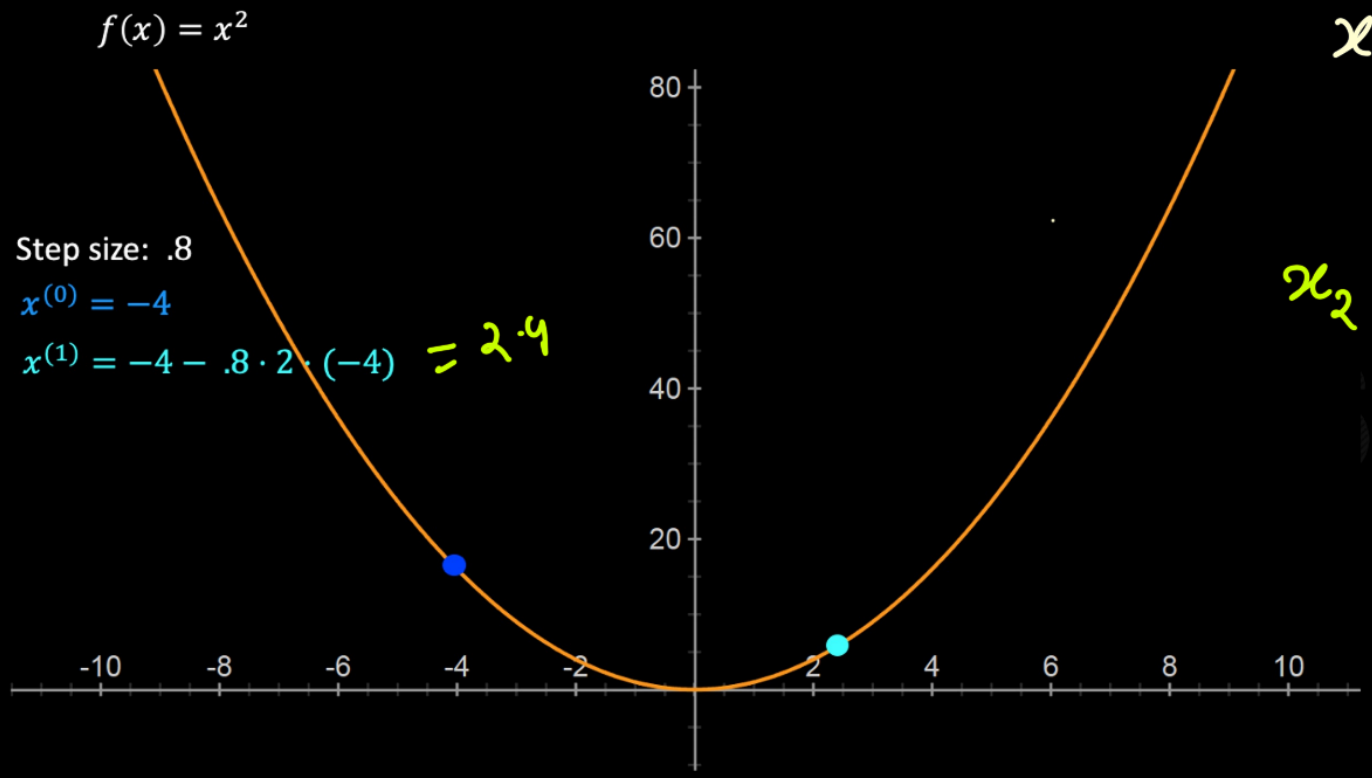


$x_1 = x_0 - \alpha f'(x)$

$-4 - 0.8(-8)$

$= -4 + 6.4 = 2.4$

Gradient Descent



$x_{new} = x_{old} - \alpha f'_{x_{old}}(x)$

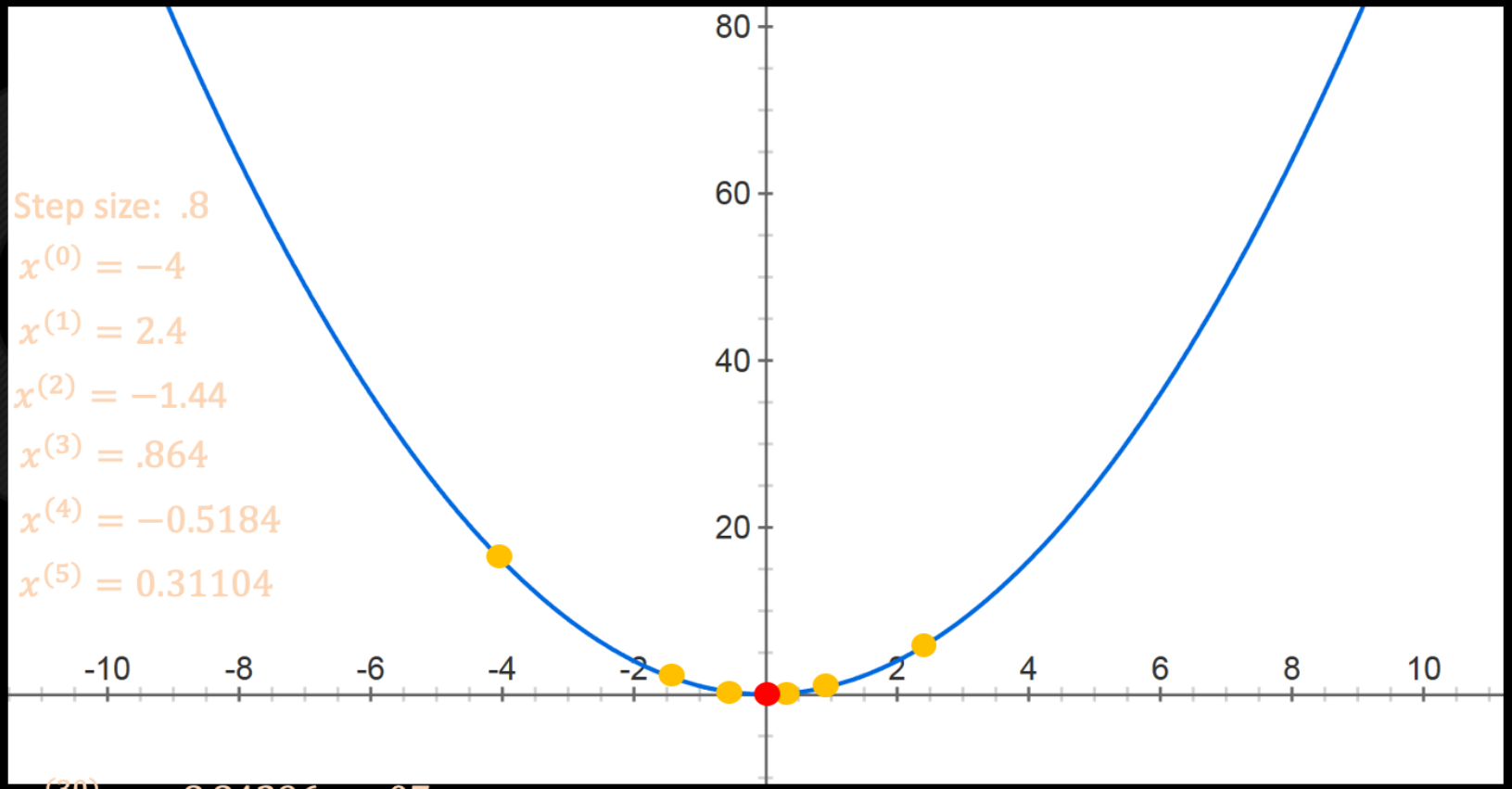
$x_2 = x_1 - 0.8\,(2 \times 2.4)$

$= 2.4 - 3.84$

$= -1.44$

Gradient Descent

$f(x) = x^2$



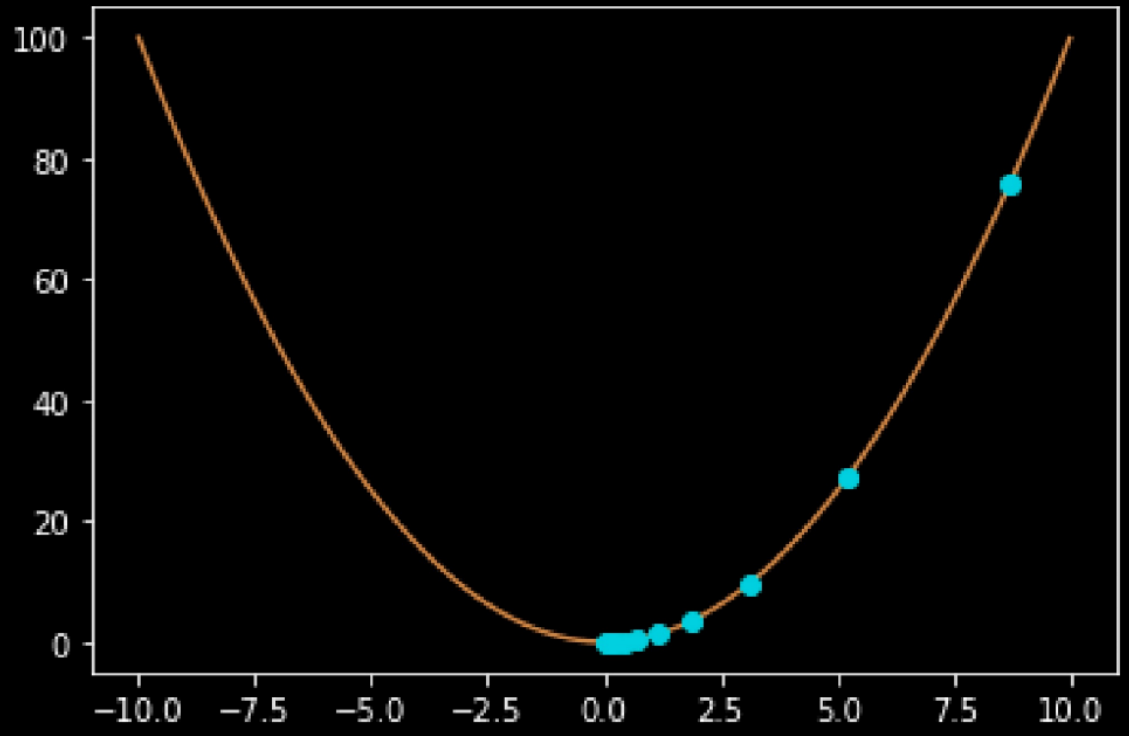
Calculus

$f(x) = x^2$

$f'(x) = 2x$

$x_0 = 8.7, \alpha = 0.2$

$x \leftarrow x - \alpha f'(x)$

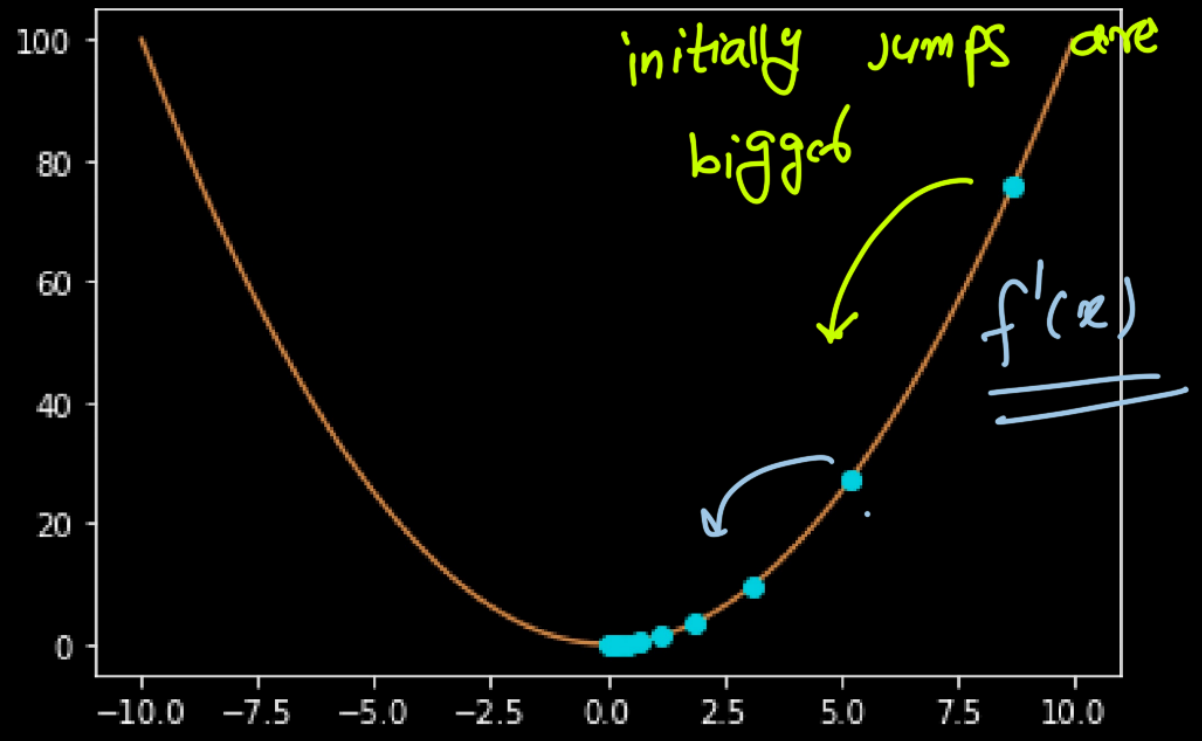


$f(x) = x^2$

$f'(x) = 2x$

$x_0 = 8.7, \alpha = 0.2$

$x \leftarrow x - \alpha f'(x)$

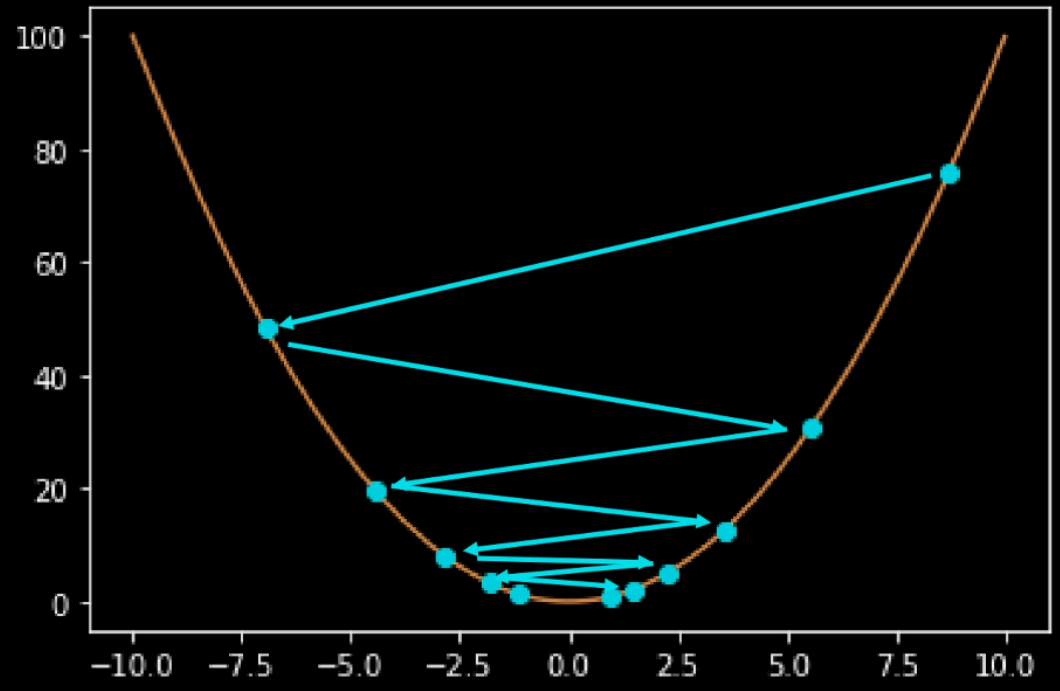


$f(x) = x^2$

$f'(x) = 2x$

$x_0 = 8.7, \alpha = 0.9$

$x \leftarrow x - \alpha f'(x)$

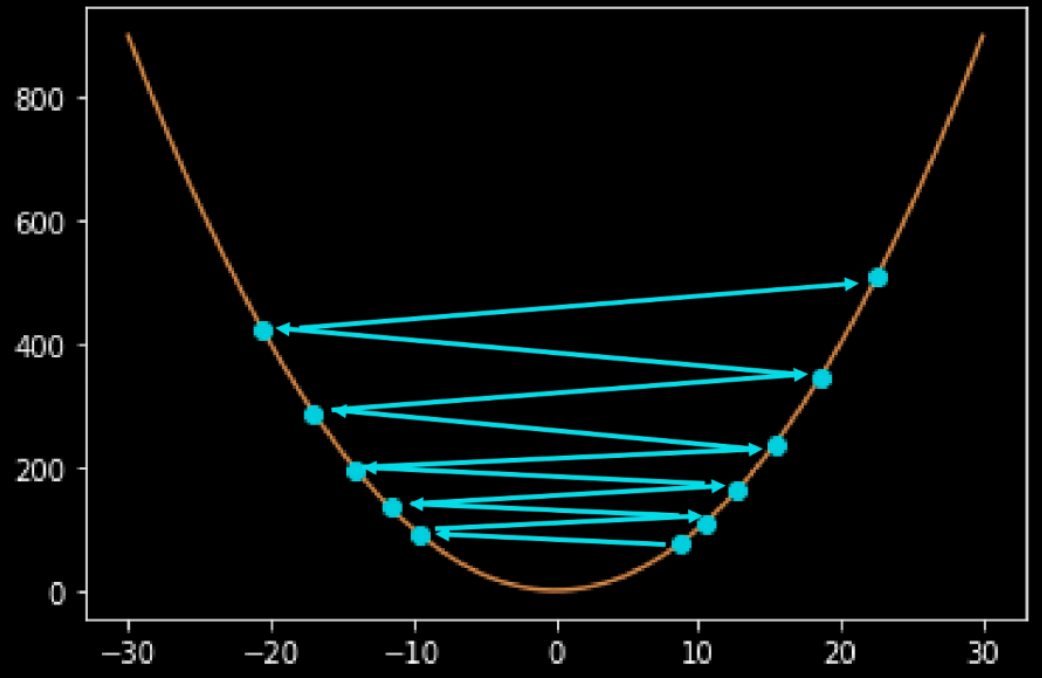


$f(x) = x^2$

$f'(x) = 2x$

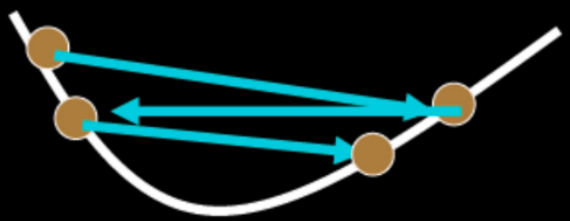
$x_0 = 8.7, \alpha = 1.05$

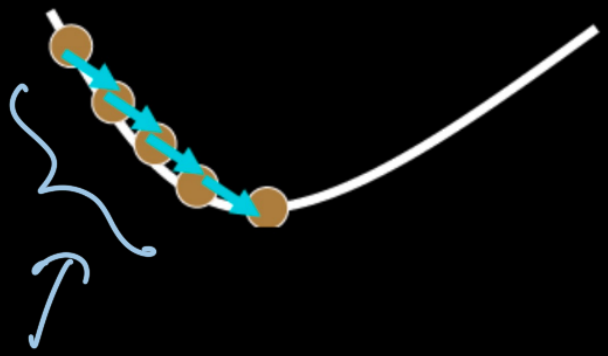
$x \leftarrow x - \alpha f'(x)$



Gradient descent may not even converge.

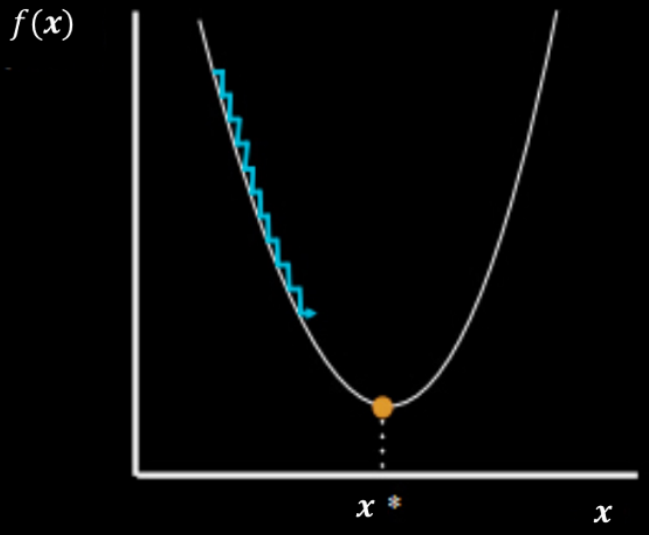
How to set step size?



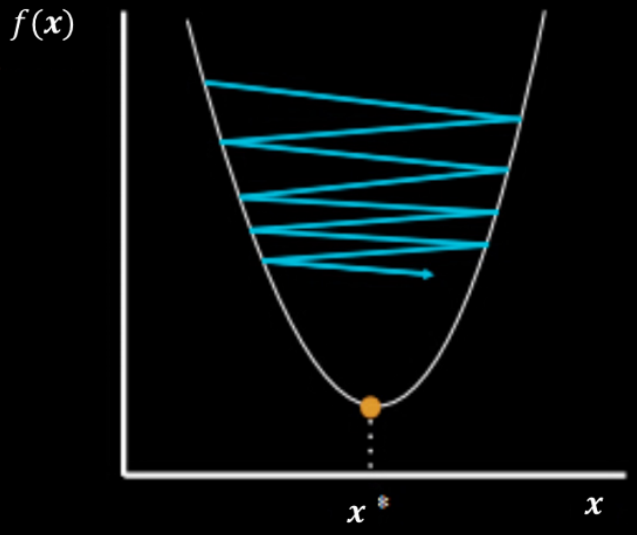


https://www.cs.tufts.edu/comp/135/2020f/slides/day06_gradient_descent.pdf

Choosing good step size matters!

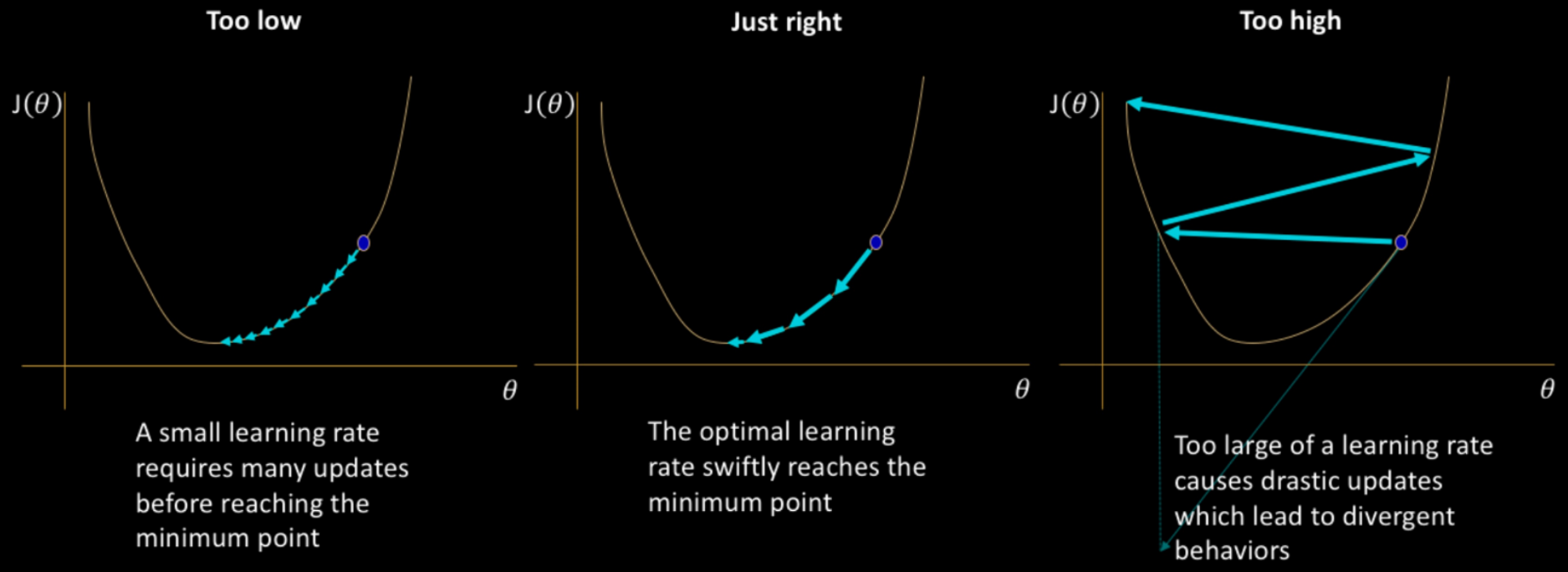


Too small: converge very slowly

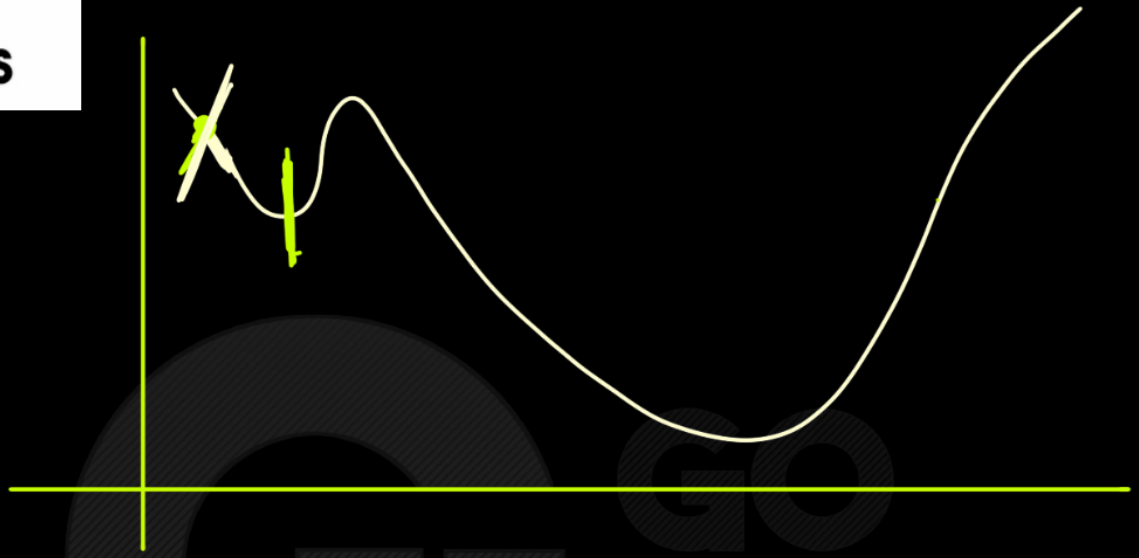


Too big: overshoot and even diverge

$\Rightarrow \alpha = 0.1, 0.01, 0.2$



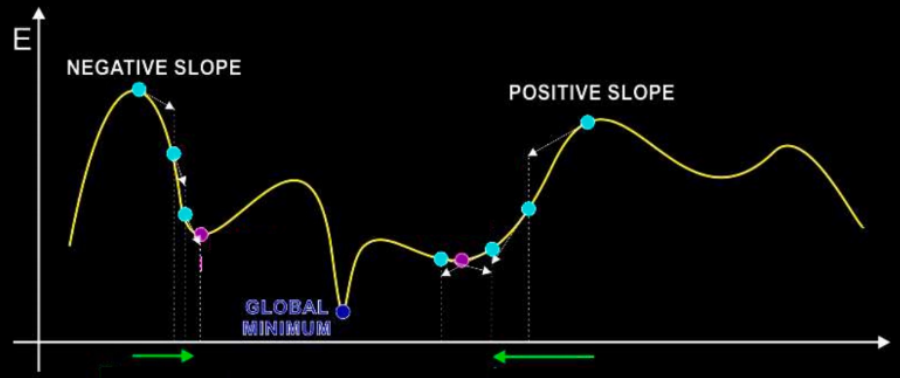
https://www.davidinouye.com/course/ece47300-spring-2024/lectures/gradient-descent.pdf



algorithm may find local minima, not the global minima.

↳ crucial factor on which it depends ⇒ Starting point.

The Approach of Gradient Descent



- Iterative solution:
  - Start at some point
  - Find direction in which to shift this point to decrease error
    - This can be found from the derivative of the function
      - A negative derivative → moving right decreases error
      - A positive derivative → moving left decreases error
  - Shift point in this direction

Gradient Descent

1. Initialize the parameters **w** to some guess (usually all zeros, or random values)
2. Update the parameters:
   $\mathbf{w} = \mathbf{w} - \eta \nabla L(\mathbf{w})$
3. Update the learning rate η
4. Repeat steps 2-3 until $\nabla L(\mathbf{w})$ is close to zero.

$x_{new} = x_{old} - \eta f'(x)$

Stopping criteria → of number of iterations reaches 30

2D Gradient Descent

or multivariate

Gradient Descent

The algorithm:

$$\boldsymbol{x}^{(t+1)} = \boldsymbol{x}^{(t)} - \alpha \nabla f(\boldsymbol{x}^{(t)}), \qquad t = 0, 1, 2, \ldots,$$

where α is called the **step size**.

$x_{new} = x_{old} - \alpha f'(x)$

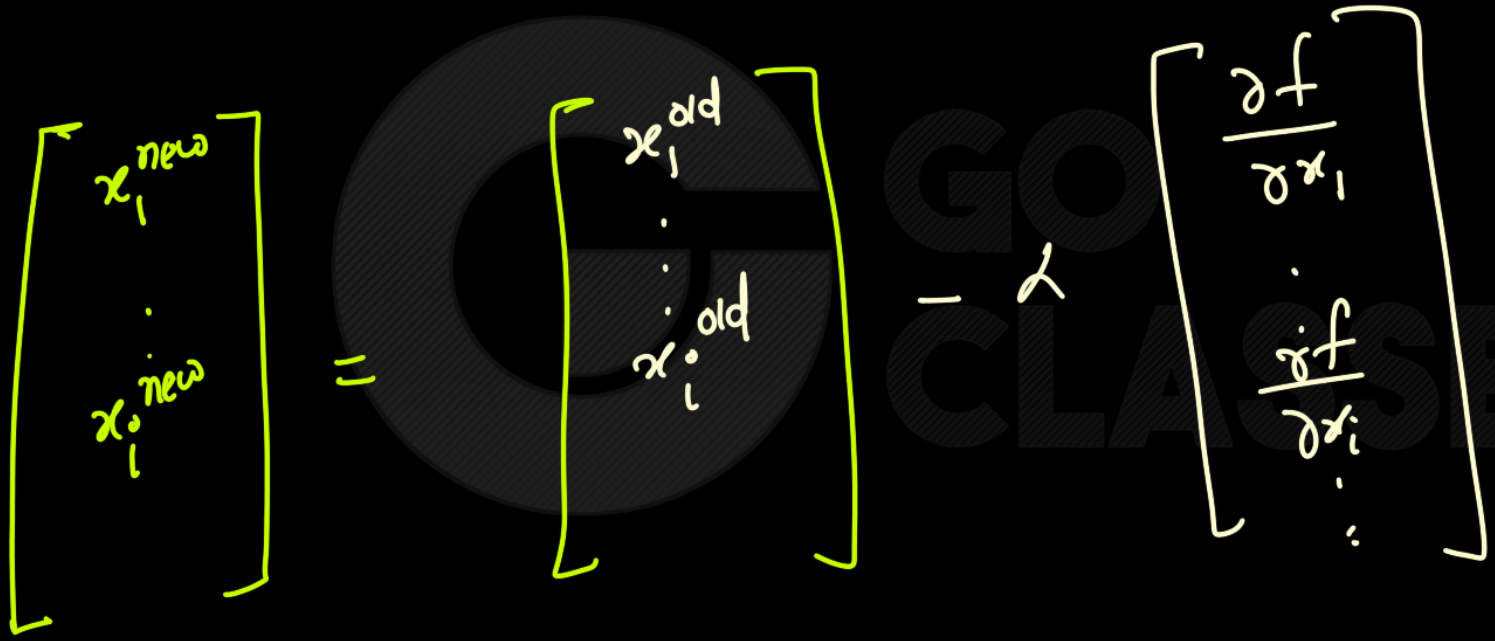
$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

$$x_i^{new} = x_i^{old} - \alpha \frac{\partial f}{\partial x_i}$$

https://engineering.purdue.edu/ChanGroup/ECE595/files/Lecture05_descent.pdf

GO CLASSES

$$x^{(t+1)} = x^{(t)} - \alpha \nabla f(x^{(t)})$$



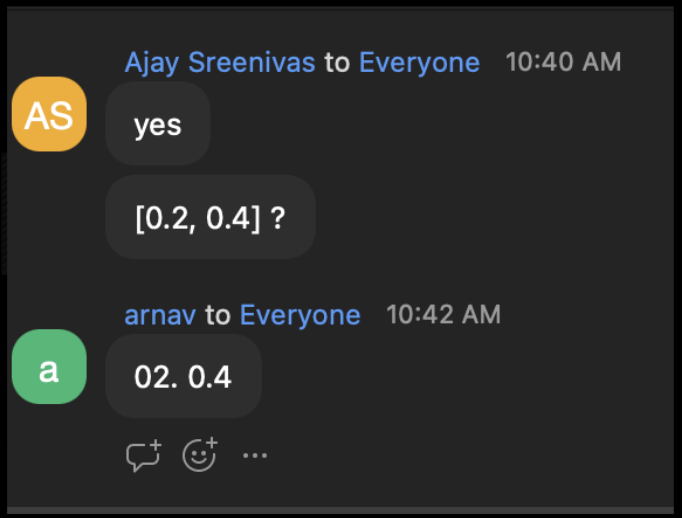
- Gradient descent is an iterative algorithm, which means we apply an update repeatedly until some criterion is met.
- We initialize the weights to something reasonable (e.g. all zeros) and repeatedly adjust them in the direction of steepest descent.

Question:

(2 points) Consider applying gradient descent with step size $\alpha = 0.1$ to find the $\mathbf{x}$ that minimizes the function $f(\mathbf{x}) = f\left((x^{(1)}, x^{(2)})\right) = \left(x^{(1)} - 1\right)^2 + \left(x^{(2)} - 2\right)^2$ starting from $\mathbf{x}_0 = (0, 0)$. Find the value $\mathbf{x}_1$ after one iteration.

$$f(x) = (x_1 - 1)^2 + (x_2 - 2)^2$$

$$x_0 = (0, 0) \qquad \alpha = 0.1$$



https://pages.stat.wisc.edu/~jgillett/451/exam/midtermSpring2023key.pdf

Question:

(2 points) Consider applying gradient descent with step size $\alpha = 0.1$ to find the $\mathbf{x}$ that minimizes the function $f(\mathbf{x}) = f\left((x^{(1)}, x^{(2)})\right) = \left(x^{(1)} - 1\right)^2 + \left(x^{(2)} - 2\right)^2$ starting from $\mathbf{x}_0 = (0, 0)$. Find the value $\mathbf{x}_1$ after one iteration.

$$f(x) = (x_1 - 1)^2 + (x_2 - 2)^2$$

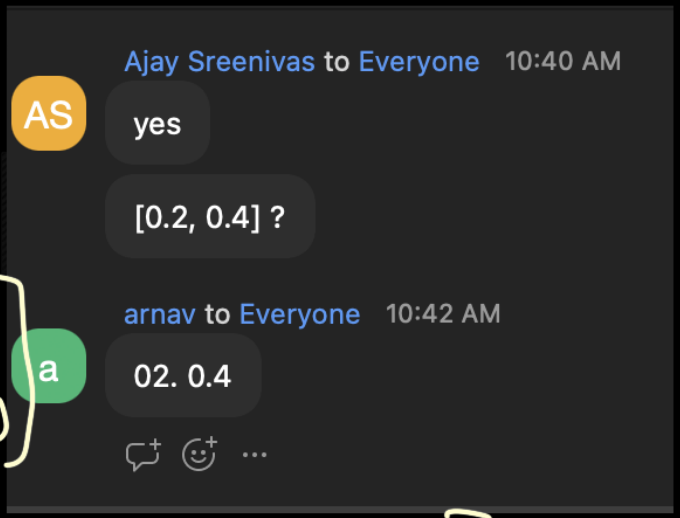
$$\nabla = \begin{bmatrix} 2(x_1 - 1) \\ 2(x_2 - 2) \end{bmatrix}$$

$$x_0 = (0, 0) \qquad \alpha = 0.1$$

$$x_1 = x_0 - \alpha \nabla f(x)$$

$$x_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} - 0.1 \begin{bmatrix} -2 \\ -4 \end{bmatrix}$$

$$= \begin{bmatrix} 0.2 \\ 0.4 \end{bmatrix}$$



(2 points) Consider applying gradient descent with step size $\alpha = 0.1$ to find the $\mathbf{x}$ that minimizes the function $f(\mathbf{x}) = f\left((x^{(1)}, x^{(2)})\right) = \left(x^{(1)} - 1\right)^2 + \left(x^{(2)} - 2\right)^2$ starting from $\mathbf{x}_0 = (0, 0)$. Find the value $\mathbf{x}_1$ after one iteration.

ANSWER:

$\nabla f(\mathbf{x}) = \left(2(x^{(1)} - 1), 2(x^{(2)} - 2)\right)$, which is $(-2, -4)$ at $\mathbf{x} = (0, 0)$.

Move to $\mathbf{x}_1 = \mathbf{x}_0 - \alpha \nabla f(\mathbf{x}_0) = (0, 0) - (-0.2, -0.4) = (0.2, 0.4)$.

Question:

(8 points) Mark each statement as true (T) or false (F).

____ Gradient descent can fail to converge on a convex function if step size α is such that we get stuck in a cycle, oscillating between two or several values.

____ For a non-convex function, gradient descent can fail to converge by descending without bound.

____ Gradient descent can fail to converge on a convex function if it gets stuck in a local minimum.

____ Gradient descent can fail to converge on a convex function if the step size $\alpha > 0$ is too small.

Question:

(8 points) Mark each statement as true (T) or false (F).

T ___ Gradient descent can fail to converge on a convex function if step size α is such that we get stuck in a cycle, oscillating between two or several values.

___ For a non-convex function, gradient descent can fail to converge by descending without bound.

F ___ Gradient descent can fail to converge on a convex function if it gets stuck in a local minimum.

F ___ Gradient descent can fail to converge on a convex function if the step size $\alpha > 0$ is too small.

(8 points) Mark each statement as true (T) or false (F).

____ Gradient descent can fail to converge on a convex function if step size α is such that we get stuck in a cycle, oscillating between two or several values. ANSWER: T

____ For a non-convex function, gradient descent can fail to converge by descending without bound. ANSWER: T

____ Gradient descent can fail to converge on a convex function if it gets stuck in a local minimum. ANSWER: F

____ Gradient descent can fail to converge on a convex function if the step size $\alpha > 0$ is too small. ANSWER: F

Question:

H.W.

Consider applying gradient descent with step size $\alpha = 0.5$ to find the $\mathbf{x}$ that minimizes the function $f(\mathbf{x}) = f\left((x^{(1)}, x^{(2)})\right) = \left(x^{(1)} - 3\right)^2 + \left(x^{(2)} - 4\right)^2$ starting from $\mathbf{x}_0 = (1, 2)$. Find the value $\mathbf{x}_1$ after one iteration.

https://pages.stat.wisc.edu/~jgillett/451/exam/midtermSummer2023key.pdf

Consider applying gradient descent with step size $\alpha = 0.5$ to find the $\mathbf{x}$ that minimizes the function $f(\mathbf{x}) = f\left((x^{(1)}, x^{(2)})\right) = \left(x^{(1)} - 3\right)^2 + \left(x^{(2)} - 4\right)^2$ starting from $\mathbf{x}_0 = (1, 2)$. Find the value $\mathbf{x}_1$ after one iteration.

ANSWER:

$\nabla f(\mathbf{x}) = \left(2(x^{(1)} - 3), 2(x^{(2)} - 4)\right)$, which is $(-4, -4)$ at $\mathbf{x} = (1, 2)$.

Move to $\mathbf{x}_1 = \mathbf{x}_0 - \alpha \nabla f(\mathbf{x}_0) = (1, 2) - 0.5(-4, -4) = (3, 4)$.

Question:

[3 Pts] Suppose we run gradient descent with a fixed learning rate of $\alpha = 0.1$ to minimize the 2D function $f(x, y) = 5 + x^2 + y^2 + 5xy$.

The gradient of this function is

$$\nabla_{x,y} f(x, y) = \begin{bmatrix} 2x + 5y \\ 2y + 5x \end{bmatrix} = \begin{bmatrix} 12 \\ 9 \end{bmatrix}$$

If our starting guess is $x^{(0)} = 1, y^{(0)} = 2$, what will be our next guess $x^{(1)}, y^{(1)}$?

$$\begin{bmatrix} 1 \\ 2 \end{bmatrix} - 0.1 \begin{bmatrix} 12 \\ 9 \end{bmatrix}$$

$x^{(1)} =$ ____ $y^{(1)} =$ ____ $= \begin{bmatrix} -0.2 \\ 1.1 \end{bmatrix}$

https://ds100.org/fa20/resources/assets/exams/fa19/fa19midterm2.pdf

$x^{(1)} =$ $y^{(1)} =$

Solution: The gradient is = [2*1 + 5*2, 2*2 + 5*1] = [12, 9] so next guess is [1, 2] - 0.1 * [12, 9] = -0.2, 1.1

Question:

Consider the following loss function on vectors $w \in \mathbb{R}^4$:

$$L(w) = w_1^2 + 2w_2^2 + w_3^2 - 2w_3w_4 + w_4^2 + 2w_1 - 4w_2 + 4.$$

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

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

(c) What is the minimum value of $L(w)$?

(d) Is there is a unique solution w at which this minimum is realized?

$$\nabla L(w) = (2w_1 + 2, 4w_2 - 4, 2w_3 - 2w_4, -2w_3 + 2w_4)$$

Question:

Consider the following loss function on vectors $w \in \mathbb{R}^4$:

$$L(w) = w_1^2 + 2w_2^2 + w_3^2 - 2w_3w_4 + w_4^2 + 2w_1 - 4w_2 + 4.$$

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

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

(c) What is the minimum value of $L(w)$?

(d) Is there is a unique solution w at which this minimum is realized?

$$\nabla L(w) = (2w_1 + 2, 4w_2 - 4, 2w_3 - 2w_4, -2w_3 + 2w_4)$$

$(2, -4, 6, 0)$

$$\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} - \eta \begin{bmatrix} 2 \\ -4 \\ 0 \\ 0 \end{bmatrix}$$

$$L(w) = w_1^2 + 2w_2^2 + w_3^2 - 2w_3w_4 + w_4^2 + 2w_1 - 4w_2 + 4$$

(a) The derivative is

$$\nabla L(w) = (2w_1 + 2, 4w_2 - 4, 2w_3 - 2w_4, -2w_3 + 2w_4)$$

(b) The derivative at $w = (0, 0, 0, 0)$ is $(2, -4, 0, 0)$. Thus the update at this point is:

$$w_{new} = w - \eta\nabla L(w) = (0, 0, 0, 0) - \eta(2, -4, 0, 0) = (-2\eta, 4\eta, 0, 0).$$

$\eta = 1/2$

(c) To find the minimum value of $L(w)$, we will equate $\nabla L(w)$ to zero:

- $2w_1 + 2 = 0 \implies w_1 = -1$
- $4w_2 - 4 = 0 \implies w_2 = 1$
- $2w_3 - 2w_4 = 0 \implies w_3 = w_4$

The function is minimized at any point of the form $(-1, 1, k, k)$.

(d) No, there is not a unique solution.

0, 0, 0, 0

↓

$-2\eta, 4\eta, 0, 0$

$-1, 2, 0, 0$

https://cseweb.ucsd.edu/classes/wi21/cse251A-a/HW/sol2.pdf

H.W.

Question:

Let

$f(x, y) = (x-1)^2(y-3)^2$

$P_0 = (x_0, y_0) = (4, 5)$

$\eta = 0.05$

$P_0 \rightarrow P_1 \rightarrow P_2$

Using the gradient descent algorithm find $P_2 = (x_2, y_2)$ and then calculate $f(P_2)$ to at least the nearest thousandth (at least 3 decimal places).

Answer

$\nabla f = \left\langle \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y} \right\rangle = \left\langle 2(x-1)(y-3)^2, 2(x-1)^2(y-3) \right\rangle$

$P_0 = (4,5)$

$P_1 = P_0 - \eta \nabla f(P_0)$

$= (4,5) - (.05)\nabla f(\langle 4,5, \rangle)$

$= (4,5) - (.05)\left\langle 2(4-1)(5-3)^2, 2(4-1)^2(5-3) \right\rangle$

$= (4,5) - (.05)\langle 24, 36 \rangle$

$= (4,5) - \langle 1.2, 1.8 \rangle$

$\boxed{P_1 = (2.8, 3.2)}$

$P_2 = P_1 - \eta \nabla f(P_1)$

$= (2.8, 3.2) - (.05)\nabla f(2.8, 3.2)$

$= (2.8, 3.2) - (.05)\left\langle 2(2.8-1)(3.2-3)^2, 2(2.8-1)^2(3.2-3) \right\rangle$

$= (2.8, 3.8 - (.05)\left\langle 2(1.8)(.2)^2, 2(1.8)^2(.2) \right\rangle$

$\boxed{P_2 = (2.7928, 3.152)}$

$f(P_2) = f(2.7928, 3.152)$

$= (2.7928-1)^2(3.152-3)^2$

$= \boxed{.0588}$ answer

Given:

$f(x,y) = (x-1)^2(y-3)^2$

$P_0 = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} = \begin{pmatrix} 4 \\ 5 \end{pmatrix}$

$\eta = 0.05$

The gradient of $f(x,y)$ is:

$\nabla f(x,y) = \begin{pmatrix} 2(x-1)(y-3)^2 \\ 2(x-1)^2(y-3) \end{pmatrix}$

Calculate $\nabla f(4,5)$:

$\nabla f(4,5) = \begin{pmatrix} 2 \cdot (4-1) \cdot (5-3)^2 \\ 2 \cdot (4-1)^2 \cdot (5-3) \end{pmatrix} = \begin{pmatrix} 24 \\ 36 \end{pmatrix}$

Now, update P_1:

$P_1 = P_0 - \eta \cdot \nabla f(4,5) = \begin{pmatrix} 4 \\ 5 \end{pmatrix} - 0.05 \cdot \begin{pmatrix} 24 \\ 36 \end{pmatrix} = \begin{pmatrix} 4 \\ 5 \end{pmatrix} - \begin{pmatrix} 1.2 \\ 1.8 \end{pmatrix} = \begin{pmatrix} 2.8 \\ 3.2 \end{pmatrix}$

Calculate $\nabla f(2.8, 3.2)$:

$\nabla f(2.8, 3.2) = \begin{pmatrix} 2 \cdot (2.8-1) \cdot (3.2-3)^2 \\ 2 \cdot (2.8-1)^2 \cdot (3.2-3) \end{pmatrix} = \begin{pmatrix} 0.072 \\ 0.648 \end{pmatrix}$

Update P_2:

$P_2 = P_1 - \eta \cdot \nabla f(2.8, 3.2) = \begin{pmatrix} 2.8 \\ 3.2 \end{pmatrix} - 0.05 \cdot \begin{pmatrix} 0.072 \\ 0.648 \end{pmatrix} = \begin{pmatrix} 2.8 \\ 3.2 \end{pmatrix} - \begin{pmatrix} 0.0036 \\ 0.0324 \end{pmatrix} = \begin{pmatrix} 2.7964 \\ 3.1676 \end{pmatrix}$

Finally, calculate $f(P_2)$:

$f(P_2) = (2.7964 - 1)^2 \cdot (3.1676 - 3)^2$

Question:

Consider the univariate function $f(x) = x^2$. This function has a unique minimum at $x^* = 0$. We're using gradient descent (GD) to find this minimum and at time t we arrive at the point $x_t = 2$. What is the step size that would bring us to x^* at time $t + 1$?

https://courses.cs.washington.edu/courses/cse446/23au/exams/pastexams/23sp-midterm-solutions.pdf

Consider the univariate function $f(x) = x^2$. This function has a unique minimum at $x^* = 0$. We're using gradient descent (GD) to find this minimum and at time t we arrive at the point $x_t = 2$. What is the step size that would bring us to x^* at time $t + 1$?

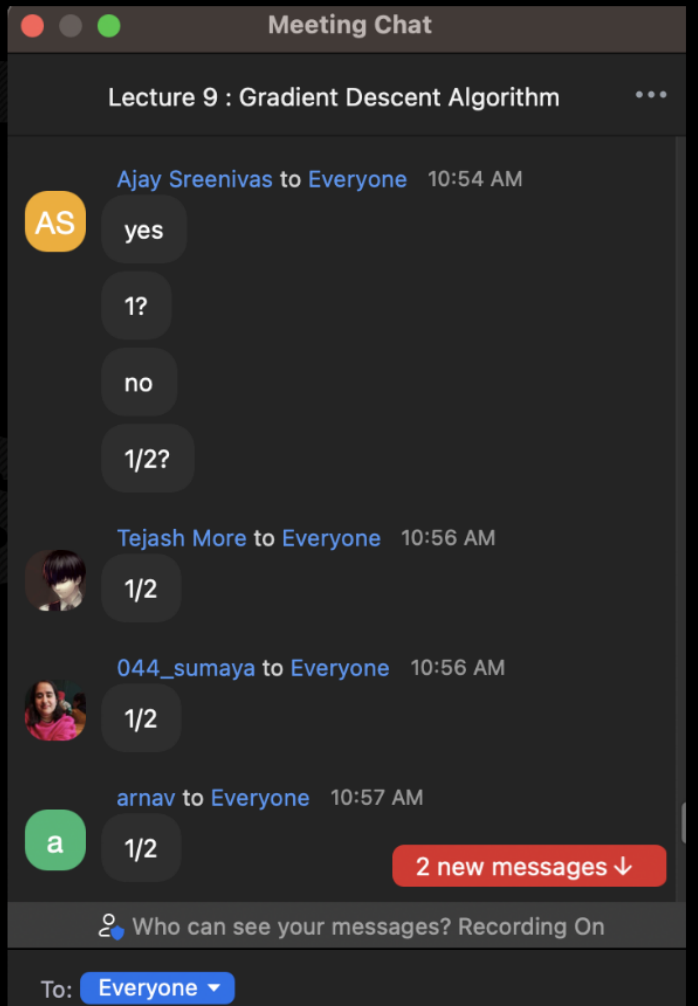
$x_t = 2$

$f'(x) = 2x$

$x_{t+1} = 2 - \alpha f'(x)$

$0 = 2 - \alpha \cdot (4)$

$\Rightarrow \alpha = 1/2$



Explanation: Using the definition of GD and the requirements in the problem statement we are looking for η such that:

$$x_* = x_t - \eta \nabla f(x_t) \iff 0 = 2 - \eta \nabla f(2) \iff \eta \cdot (2 \cdot 2) = 2 \iff \eta = \frac{1}{2}$$

Question:

(5 points) Consider the following loss function:

$$L(w) = \sum_{i=1}^{n} \log(1 + w^\top x_i)$$

$\nabla L_w = \sum_{i=1}^{n} \nabla L_i$

What is $\nabla L(w)$? Write down the update step for gradient descent.

$L_i = \log(1 + w^T x_i)$

$L(w) = \sum_{i=1}^{n} L_i$

$\nabla_w L_i \quad \frac{1}{1 + w^T x_i} \cdot x_i$

$$w_{new} = w_{old} - \eta \cdot \sum_{i=1}^{n} \nabla L_i$$

https://cseweb.ucsd.edu/classes/wi21/cse251A-a/HW/past_exams/wi20_quiz2_sol.pdf

(5 points) Consider the following loss function:

$$L(w) = \sum_{i=1}^{n} \log(1 + w^\top x_i)$$

What is $\nabla L(w)$? Write down the update step for gradient descent.

Let $L_i(w) = \log(1 + w^\top x_i)$. Then

$$\begin{aligned} \nabla L(w) &= \sum_{i=1}^{n} \nabla L_i(w) \\ &= \sum_{i=1}^{n} \frac{x_i}{1 + w^\top x_i}. \end{aligned}$$

The gradient descent update step for w is then

$$w_{t+1} = w_t - \eta_t \nabla L(w) = w_t - \eta_t \sum_{i=1}^{n} \frac{x_i}{1 + w_t^\top x_i},$$

where η_t is the learning rate at the t-th descent step.

Question: MSQ

[3 Pts] The learning rate can *potentially* affect which of the following? Select all that apply. Assume nothing about the function being minimized other than that its gradient exists. You may assume the learning rate is positive.

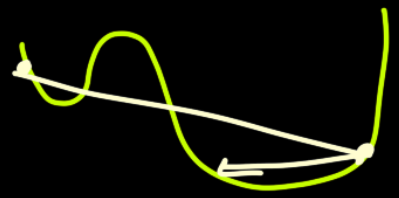
- ☐ The speed at which we converge to a minimum.
- ☐ Whether gradient descent converges.
- ☐ The direction in which the step is taken.
- ☐ Whether gradient descent converges to a local minimum or a global minimum.

https://ds100.org/fa20/resources/assets/exams/fa19/fa19midterm2.pdf

Question: MSQ

[3 Pts] The learning rate can *potentially* affect which of the following? Select all that apply. Assume nothing about the function being minimized other than that its gradient exists. You may assume the learning rate is positive.

- ☑ The speed at which we converge to a minimum.
- ☑ Whether gradient descent converges.
- ☐ The direction in which the step is taken.
- ☑ Whether gradient descent converges to a local minimum or a global minimum.



https://ds100.org/fa20/resources/assets/exams/fa19/fa19midterm2.pdf

[3 Pts] The learning rate can *potentially* affect which of the following? Select all that apply. Assume nothing about the function being minimized other than that its gradient exists. You may assume the learning rate is positive.

- ☐ **The speed at which we converge to a minimum.**
- ☐ **Whether gradient descent converges.**
- ☐ The direction in which the step is taken.
- ☐ **Whether gradient descent converges to a local minimum or a global minimum.**

Question:

35. For the dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$ and the loss function:

$$L(w) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sin\left(w_0 + w_1 x_i\right)\right)^2$$

H.W.

answer each of the following questions.

Which of the following is the gradient of the loss function with respect to $w = [w_0, w_1]$

- ◯ 0
- ◯ $\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sin\left(w_0 + w_1 x_i\right)\right)^2$
- ◯ $\frac{1}{n}\sum_{i=1}^{n}\sin\left(y_i - \cos\left(w_0 + w_1 x_i\right) w_1\right)$
- ◯ $-\frac{2}{n}\sum_{i=1}^{n}\left(y_i - \sin\left(w_0 + w_1 x_i\right)\right)\cos\left(w_0 + w_1 x_i\right)[1, x_i]$
- ◯ $\left[-\frac{2}{n}\sum_{i=1}^{n}\cos\left(w_0 + w_1 x_i\right), -\frac{2}{n}\sum_{i=1}^{n}\cos\left(w_0 + w_1 x_i\right) x_i\right]$

(a) [3 Pts] Which of the following is the gradient of the loss function with respect to $w = [w_0, w_1]$

- 0
- $\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sin\left(w_0 + w_1 x_i\right)\right)^2$
- $\frac{1}{n}\sum_{i=1}^{n}\sin\left(y_i - \cos\left(w_0 + w_1 x_i\right) w_1\right)$
- $-\frac{2}{n}\sum_{i=1}^{n}\left(y_i - \sin\left(w_0 + w_1 x_i\right)\right)\cos\left(w_0 + w_1 x_i\right)\left[1, x_i\right]$
- $\left[-\frac{2}{n}\sum_{i=1}^{n}\cos\left(w_0 + w_1 x_i\right), -\frac{2}{n}\sum_{i=1}^{n}\cos\left(w_0 + w_1 x_i\right) x_i\right]$

Question:

Which of the following could affect whether or not gradient descent converges to the global minimum?

(a) [1 Pt] Learning Rate

◯ True ◯ False

(b) [1 Pt] Initialization of parameters

◯ True ◯ False

(c) [1 Pt] The ordering of the data (ignoring issues related to numerical precision).

◯ True ◯ False

https://ds100.org/fa20/resources/assets/exams/sp20/sp20checkpoint.pdf

Question:

Which of the following could affect whether or not gradient descent converges to the global minimum?

(a) [1 Pt] Learning Rate ← learning rate can make it diverge

◯ True ◯ False

(b) [1 Pt] Initialization of parameters crucial thing

◯ True ◯ False

(c) [1 Pt] The ordering of the data (ignoring issues related to numerical precision).

◯ True ◯ False

https://ds100.org/fa20/resources/assets/exams/sp20/sp20checkpoint.pdf

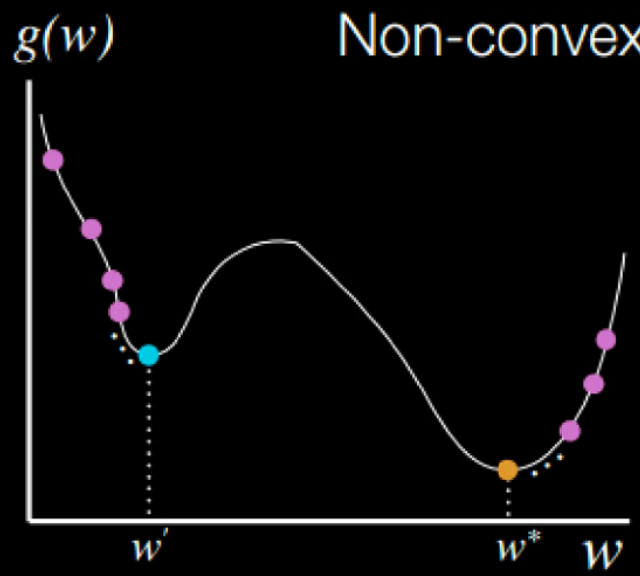
36.

(a) [1 Pt] Learning Rate
◯ **True** ◯ False

(b) [1 Pt] Initialization of parameters
◯ **True** ◯ False

(c) [1 Pt] The ordering of the data (igno
◯ True ◯ **False**

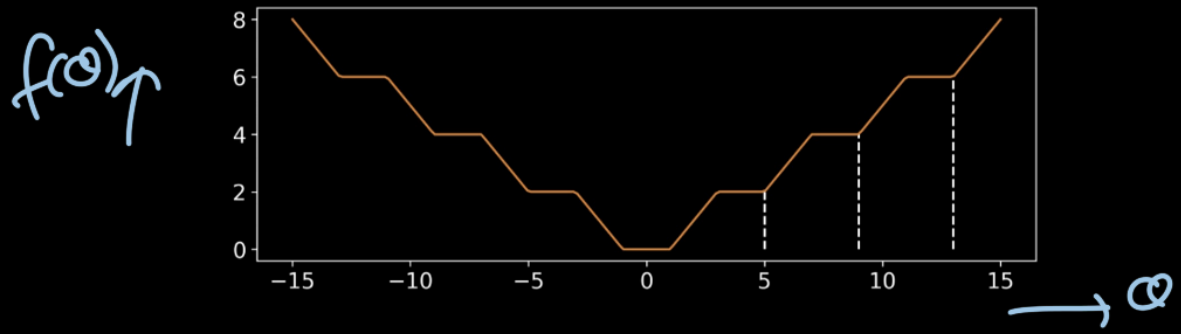
Solution: The initialization of parameters affects the starting parameters of gradient descent. If the loss function is not convex, the choice of starting parameters will affect whether the gradient descent algorithm converges to a local minimum or global minimum. The graph below shows a non-convex loss function with respect to parameter w.



If we initialize w to be very small, then gradient descent will converge to a local minimum (given a reasonable learning rate). If we initialize w to be very large, then gradient descent will converge to a global minimum (given a reasonable learning rate).

Question:

13. Consider the following function of $f(\theta)$, which alternates between completely flat regions and regions of absolute slope equal to 1. **There is only one correct answer for each part.**



(a) Assuming that θ starts in a flat region that is not a minimum and $\alpha > 0$, will the basic gradient descent algorithm terminate at a minimum?

$f'(\theta) = 0$

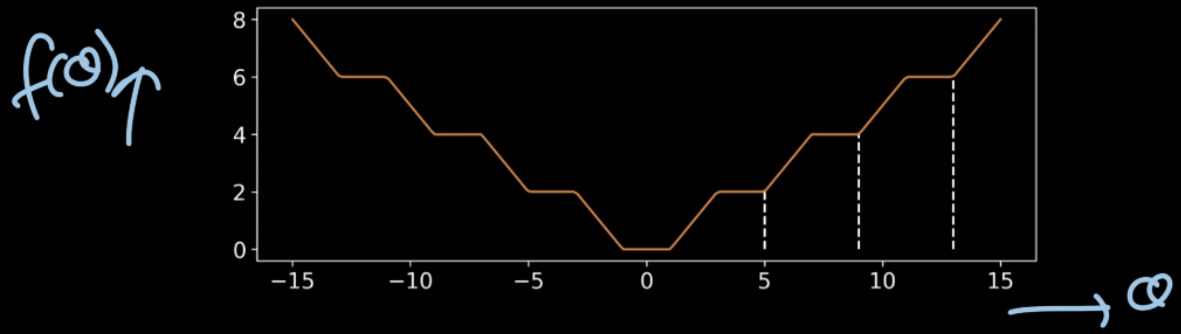
◯ A. Never ◯ B. Maybe ◯ C. Yes with enough iterations

(b) Assuming that θ starts in a sloped region and $\alpha > 0$, will the basic gradient descent algorithm find the minimum?

◯ A. Never ◯ B. Maybe ◯ C. Yes with enough iterations

Question:

13. Consider the following function of $f(\theta)$, which alternates between completely flat regions and regions of absolute slope equal to 1. **There is only one correct answer for each part.**



(a) Assuming that θ starts in a flat region that is not a minimum and $\alpha > 0$, will the basic gradient descent algorithm terminate at a minimum?

$f'(\theta) = 0$

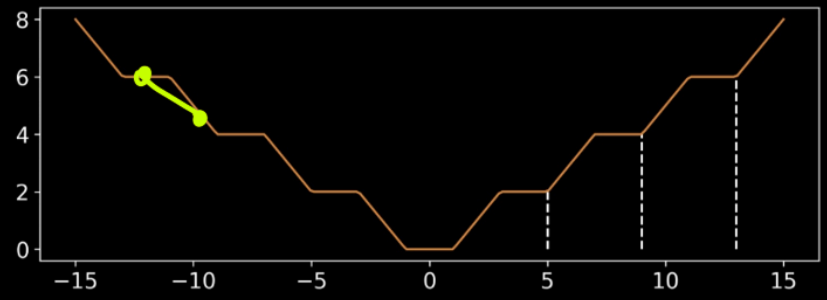
◯ A. Never ◯ B. Maybe ◯ C. Yes with enough iterations

(b) Assuming that θ starts in a sloped region and $\alpha > 0$, will the basic gradient descent algorithm find the minimum?

◯ A. Never ◯ B. Maybe ◯ C. Yes with enough iterations

Question:

13. Consider the following function of $f(\theta)$, which alternates between completely flat regions and regions of absolute slope equal to 1. **There is only one correct answer for each part.**



Is $f(\theta)$ convex?

- ◯ A. Yes
- ◯ B. No

https://ds100.org/fa20/resources/assets/exams/fa18/fa18midterm.pdf

(a) Assuming that θ starts in a flat region that is not a minimum and $\alpha > 0$, will the basic gradient descent algorithm terminate at a minimum? Note that the basic gradient descent algorithm is just the same as version with momentum on the previous page, but where $\gamma = 0$.

◯ **A. Never** ◯ B. Maybe ◯ C. Yes with enough iterations

(b) Assuming that θ starts in a sloped region and $\alpha > 0$, will the basic gradient descent algorithm find the minimum?

◯ A. Never ◯ **B. Maybe** ◯ C. Yes with enough iterations

Is $f(\theta)$ convex?

◯ A. Yes

◯ **B. No**

https://ds100.org/fa20/resources/assets/exams/fa18/fa18midtermsol.pdf

Question:

27. [10 Pts] Consider the following loss function based on data $x_1, \ldots, x_n$ with mean $\overline{x}$:

$$\ell(\beta) = \log \beta + \frac{\overline{x}}{\beta} + \frac{1}{n} \sum_{i=1}^{n} e^{-x_i/\beta}$$

Given an estimate $\beta^{(t)}$, write out the update $\beta^{(t+1)}$ after one iteration of gradient descent with step size α. Show your work in the box below.

Solution: The update is

$$\beta^{(t+1)} \leftarrow \beta^{(t)} - \alpha\, \ell'(\beta^{(t)}),$$

where

$$\ell'(\beta) = \frac{1}{\beta}\left(1 - \frac{\overline{x}}{\beta} + \frac{1}{n\beta}\sum_{i=1}^{n} x_i e^{-x_i/\beta}\right)$$
$$= \frac{1}{\beta} - \frac{\overline{x}}{\beta^2} + \frac{1}{n\beta^2}\sum_{i=1}^{n} x_i e^{-x_i/\beta}$$

Alternate notation:

$$\beta^{(t+1)} \leftarrow \beta^{(t)} - \alpha \left.\frac{\partial \ell}{\partial \beta}\right|_{\beta=\beta^{(t)}}$$

With everything substituted in:

$$\beta^{(t+1)} \leftarrow \beta^{(t)} - \alpha\left(\frac{1}{\beta^{(t)}} - \frac{\overline{x}}{{\beta^{(t)}}^2} + \frac{1}{n{\beta^{(t)}}^2}\sum_{i=1}^{n} x_i e^{-x_i/\beta^{(t)}}\right)$$

Question:

(m) [3 pts] Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous, smooth function whose derivative $f'(x)$ is also continuous. Suppose f has a unique global minimum $x^* \in (-\infty, \infty)$, and you are using **gradient descent** to find x^*. You fix some $x^{(0)} \in \mathbb{R}$ and $\epsilon > 0$, and run $x^{(t)} = x^{(t-1)} - \epsilon f'(x^{(t-1)})$ repeatedly. Which of the following statements are true?

- ○ Gradient descent is sure to converge, to *some* value, for any step size $\epsilon > 0$
- ○ If f has a local minimum x' different from the global one, i.e., $x' \neq x^*$, and $x^{(t)} = x'$ for some t, gradient descent will not converge to x^*
- ○ Assuming gradient descent converges, it converges to x^* if and only if f is convex
- ○ If, additionally, f is the objective function of logistic regression, and gradient descent converges, then it converges to x^*

Skip this option

https://people.eecs.berkeley.edu/~jrs/189/exam/mids19.pdf

(m) [3 pts] Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous, smooth function whose derivative $f'(x)$ is also continuous. Suppose f has a unique global minimum $x^* \in (-\infty, \infty)$, and you are using **gradient descent** to find x^*. You fix some $x^{(0)} \in \mathbb{R}$ and $\epsilon > 0$, and run $x^{(t)} = x^{(t-1)} - \epsilon f'(x^{(t-1)})$ repeatedly. Which of the following statements are true?

○ Gradient descent is sure to converge, to *some* value, for any step size $\epsilon > 0$

● If f has a local minimum x' different from the global one, i.e., $x' \neq x^*$, and $x^{(t)} = x'$ for some t, gradient descent will not converge to x^*

○ Assuming gradient descent converges, it converges to x^* if and only if f is convex

● If, additionally, f is the objective function of logistic regression, and gradient descent converges, then it converges to x^*

The top-left option is false because for a large enough step size, gradient descent may not converge. The bottom-left option is correct because $f'(x') = 0$, so gradient descent will never move from a local minimium. The top-right option is false because you could "accidentally" initialize GD at x^* even if f is non-convex. The bottom-right option is correct because the objective of logistic regression is convex.

Question:

Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous, smooth function whose derivative $f'(x)$ is also continuous. Suppose f has a unique global minimum $x^\star$, and you are using gradient descent to find $x^\star$. You fix some x_0 and $\eta > 0$, and apply $x_{t+1} = x_t - \eta f'(x_t)$ repeatedly. Which of the following statements is true?

(A) Gradient descent is sure to converge to some value, for any step size $\eta > 0$.

(B) Assuming gradient descent converges, it converges to $x^\star$ if and only if f is convex.

(C) If f has a local minimum x' different from the global one, and $x_t = x'$ for some t, gradient descent will not converge to $x^\star$.

https://courses.cs.washington.edu/courses/cse446/24sp/sections/05/practice_midterm_solutions.pdf

Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous, smooth function whose derivative $f'(x)$ is also continuous. Suppose f has a unique global minimum $x^\star$, and you are using gradient descent to find $x^\star$. You fix some x_0 and $\eta > 0$, and apply $x_{t+1} = x_t - \eta f'(x_t)$ repeatedly. Which of the following statements is true?

(A) Gradient descent is sure to converge to some value, for any step size $\eta > 0$.

(B) Assuming gradient descent converges, it converges to $x^\star$ if and only if f is convex.

(C) If f has a local minimum x' different from the global one, and $x_t = x'$ for some t, gradient descent will not converge to $x^\star$.

Solution:

The solution is (C). For a large enough step size, gradient descent may not converge. If $x_0 = x^\star$ then GD converges regardless of convexity. When $x_t = x'$, then $f'(x_t) = 0$, so it converges to x' instead of $x^\star$.

Question:

Gradient descent is sure to converge, to some value, for any step size greater than 0.

Solution: False. The step-size can be too large, which may cause the algorithm to diverge.

https://introml.mit.edu/_static/spring23/midterm/6_390__Midterm__Spring_2023_solutions.pdf

Question:

Suppose that we are given $f(x) = x^3 + x^2$ and learning rate $\alpha = 1/4$.

a) [5 Points] First of all, write down the updating rule for gradient descent in general and for this function.

b) [5 Points] If we start at $x_0 = -1$, should we go left or right? Can you verify mathematically? What is x_1? Can gradient descent converge? If so, where it might converge to, given appropriate step size?

c) [5 Points] If we start at $x_0 = 1$, should we go left or right? Can you verify mathematically? What is x_1? Can gradient descent converge? If so, where it might converge to, given appropriate step size?

d) [5 Points] Write down 1 condition to terminate the gradient descent algorithm (in general).

https://people.cs.uchicago.edu/~hytruongson/DSC40A_FA22/Midterm_Solution.pdf

a) [5 Points] First of all, write down the updating rule for gradient descent in general and for this function.

Solution: In general, the updating rule for gradient descent is:

$$x_{i+1} = x_i - \alpha \nabla f(x_i) = x_i - \alpha \frac{\partial f}{\partial x}(x_i),$$

where $\alpha \in \mathbb{R}_+$ is the learning rate or step size. For this function, since f is a single-variable function, we can write down the updating rule as:

$$x_{i+1} = x_i - \alpha \frac{df}{dx}(x_i) = x_i - \alpha f'(x_i).$$

We also have:

$$\frac{df}{dx} = f'(x) = 3x^2 + 2x,$$

thus the updating rule can be written down as:

$$x_{i+1} = x_i - \alpha(3x_i^2 + 2x_i) = -\frac{3}{4}x_i^2 + \frac{1}{2}x_i.$$

b) [5 Points] If we start at $x_0 = -1$, should we go left or right? Can you verify mathematically? What is x_1? Can gradient descent converge? If so, where it might converge to, given appropriate step size?

Solution: We have

$$f'(x_0) = f'(-1) = 3(-1)^2 + 2(-1) = 1 > 0,$$

so we go left, and

$$x_1 = x_0 - \alpha f'(x_0) = -1 - \frac{1}{4} = -\frac{5}{4}.$$

Intuitively, the gradient descent cannot converge in this case because

$$\lim_{x \to -\infty} f(x) = -\infty,$$

We need to find all local minimums and local maximums. First, we solve the equation $f'(x) = 0$ to find all critical points. We have:

$$f'(x) = 0 \Leftrightarrow 3x^2 + 2x = 0 \Leftrightarrow x = -\frac{2}{3} \quad \text{and} \quad x = 0.$$

Now, we consider the second-order derivative:

$$f''(x) = \frac{d^2 f}{dx^2} = 6x + 2.$$

We have $f''(x) = 0$ only when $x = -1/3$. Thus, for $x < -1/3$, $f''(x)$ is negative or the slope $f'(x)$ decreases; and for $x > -1/3$, $f''(x)$ is positive or the slope $f'(x)$ increases. Keep in mind that $-1 < -2/3 < -1/3 < 0 < 1$. Therefore, f has a local maximum at $x = -2/3$ and a local minimum at $x = 0$. If the gradient descent starts at $x_0 = -1$ and it always goes left then it will never meet the local minimum at $x = 0$, and it will go left infinitely. We say the gradient descent cannot converge, or is divergent.

c) [5 Points] If we start at $x_0 = 1$, should we go left or right? Can you verify mathematically? What is x_1? Can gradient descent converge? If so, where it might converge to, given appropriate step size?

Solution: We have

$$f'(x_0) = f'(-1) = 3 \cdot 1^2 + 2 \cdot 1 = 5 > 0,$$

so we go left, and

$$x_1 = x_0 - \alpha f'(x_0) = 1 - \frac{1}{4} \cdot 5 = -\frac{1}{4}.$$

From the previous part, function f has a local minimum at $x = 0$, so the gradient descent can converge (given appropriate step size) at this local minimum.

d) [5 Points] Write down 1 condition to terminate the gradient descent algorithm (in general).

Solution: There are several ways to terminate the gradient descent algorithm:

- If the change in the optimization objective is too small, i.e. $|f(x_i) - f(x_{i+1})| < \epsilon$ where ϵ is a small constant,
- If the gradient is close to zero or the norm of the gradient is very small, i.e. $\|\nabla f(x_i)\| < \lambda$ where λ is a small constant.

MSQ Question:

Which of the following are true about gradient descent? (select all statements that are true.)

- [ ] After each iteration, we modify the weight vector in the direction of the gradient.
- [ ] We have to choose a non-variable learning rate.
- [ ] After each iteration, we modify the weight vector in the direction of the negative gradient.
- [ ] In the gradient descent algorithm each update of the weight vector depends on all the training examples.

https://www.seas.upenn.edu/~cis5190/fall2020/assets/quiz/quiz1.pdf

Question 3

1 / 1 pts

Which of the following are true about gradient descent? (select all statements that are true.)

- [ ] After each iteration, we modify the weight vector in the direction of the gradient.
- [ ] We have to choose a non-variable learning rate.

Correct!

- [x] After each iteration, we modify the weight vector in the direction of the negative gradient.

Correct!

- [x] In the gradient descent algorithm each update of the weight vector depends on all the training examples.

Question:

Select all that apply: Consider the convex function $f(z) = z^2$. Let α be our learning rate in gradient descent.

For which values of α will $\lim_{t\to\infty} f(z^{(t)}) = 0$, assuming the initial value of z is $z^{(0)} = 1$ and $z^{(t)}$ is the value of z after the t-th iteration of gradient descent?

- ☐ $\alpha = 0$
- ☐ $\alpha = \frac{1}{2}$
- ☐ $\alpha = 1$
- ☐ $\alpha = 2$
- ☐ None of the above

Numerical answer: Give the range of all values for $\alpha \geq 0$ such that $\lim_{t\to\infty} f(z^{(t)}) = 0$, assuming the initial value of z is $z^{(0)} = 1$.

https://www.cs.cmu.edu/~mgormley/courses/10601/handouts/exam2_practice_solutions.pdf

Select all that apply: Consider the convex function $f(z) = z^2$. Let α be our learning rate in gradient descent.

For which values of α will $\lim_{t\to\infty} f(z^{(t)}) = 0$, assuming the initial value of z is $z^{(0)} = 1$ and $z^{(t)}$ is the value of z after the t-th iteration of gradient descent?

- ☐ $\alpha = 0$
- ☐ $\alpha = \frac{1}{2}$ $\qquad\qquad \alpha = \frac{1}{2}$
- ☐ $\alpha = 1$
- ☐ $\alpha = 2$
- ☐ None of the above

Numerical answer: Give the range of all values for $\alpha \geq 0$ such that $\lim_{t\to\infty} f(z^{(t)}) = 0$, assuming the initial value of z is $z^{(0)} = 1$.

$(0, 1)$.

Question:

Convexity is a desirable property in machine learning because it:

(a) guarantees gradient descent finds a global minimum in optimization problems for functions that have a global minimum

(b) helps to avoid the model overfitting

(c) speeds up model training

(d) reduces model complexity

Skip these Options

https://courses.cs.washington.edu/courses/cse446/23au/exams/pastexams/23sp-midterm-solutions.pdf

Convexity is a desirable property in machine learning because it:

(a) guarantees gradient descent finds a global minimum in optimization problems for functions that have a global minimum

(b) helps to avoid the model overfitting

(c) speeds up model training

(d) reduces model complexity

Correct answers: (a)

$$x_{new} = x_{old} - \alpha \nabla f(x)$$

GO CLASSES

Constrained optimisation