CS457/557 Computational Intelligence and Machine Learning

Central Washington University

Gradient Descent - Definition

- Gradient descent is an optimization algorithm which is commonly-used to train machine learning models and neural networks.
- It trains machine learning models by minimizing errors between predicted and actual results - **minimize the cost function**.
- Training data helps these models learn over time, and the cost function within **gradient descent** specifically acts as a barometer, gauging its accuracy with each iteration of parameter updates.
- Until the function is close to or equal to zero, the model will continue to adjust its parameters to yield the smallest possible error.

Gradient Descent

- Method to find local optima of differentiable a function f
  - Intuition: gradient tells us direction of greatest increase, negative gradient gives us direction of greatest decrease
    - Take steps in directions that reduce the function value
  - Definition of derivative guarantees that if we take a small enough step in the direction of the negative gradient, the function will decrease in value
    - How small is small enough?

Gradient Descent

Gradient Descent Algorithm:

- Pick an initial point x_0
- Iterate until convergence

$$x_{t+1} = x_t - \alpha_t \nabla f(x_t)$$

where α_t is the t^{th} step size (sometimes called learning rate)

Gradient Descent

Gradient Descent Algorithm:

- Pick an initial point x_0
- Iterate until convergence

$$x_{t+1} = x_t - \alpha_t \nabla f(x_t)$$

where α_t is the t^{th} step size (sometimes called learning rate)

When do we stop?

Gradient Descent

Gradient Descent Algorithm:

- Pick an initial point x_0

- Iterate until convergence

$$x_{t+1} = x_t - \alpha_t \nabla f(x_t)$$

where α_t is the t^{th} step size (sometimes called learning rate)

Possible Stopping Criteria: iterate until
$\|\nabla f(x_t)\| \leq \epsilon$ for some $\epsilon > 0$

How small should ϵ be?

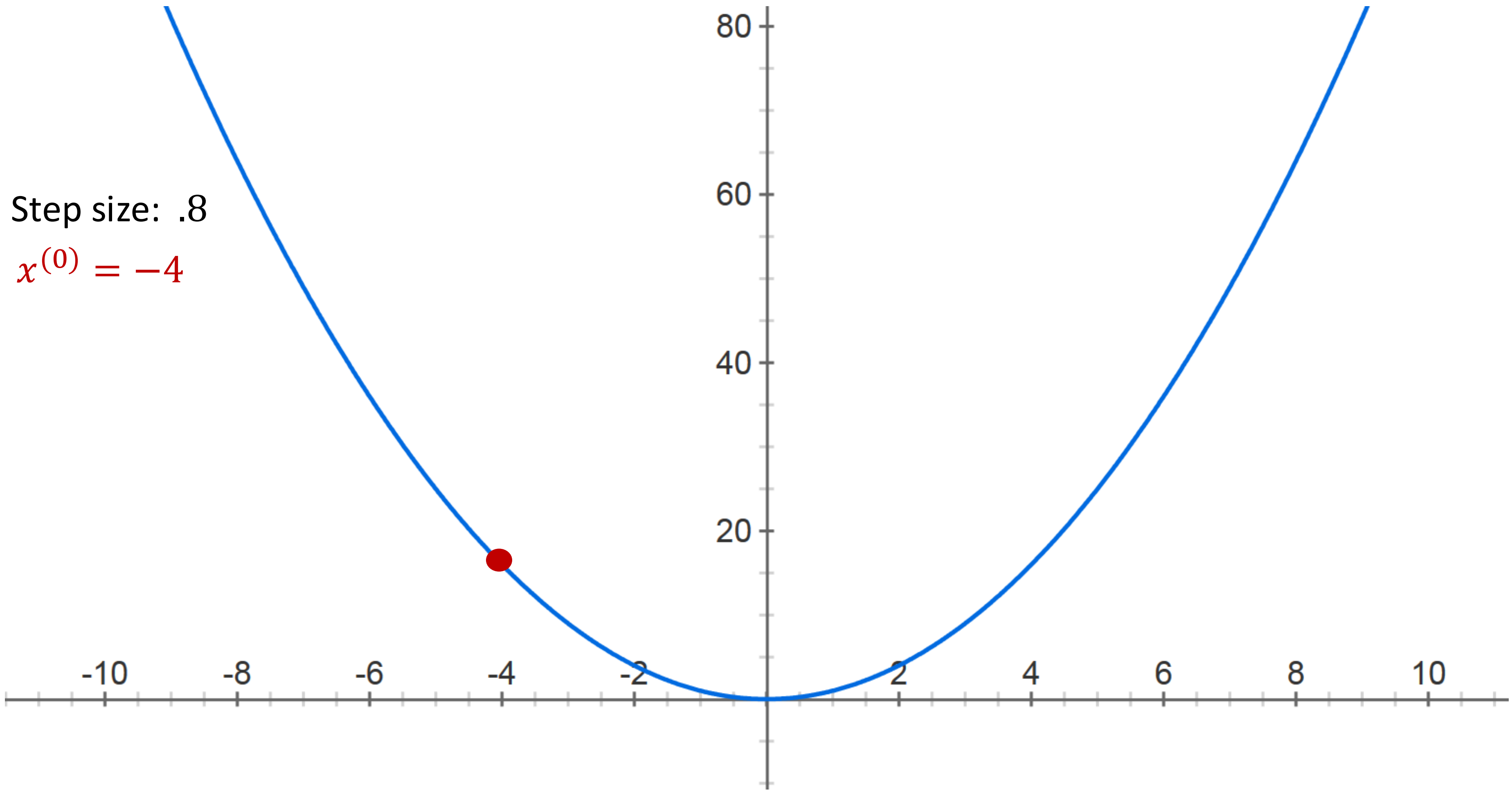
Gradient Descent

Parabola equation

$f(x) = x^2$

Step size: .8

$x^{(0)} = -4$



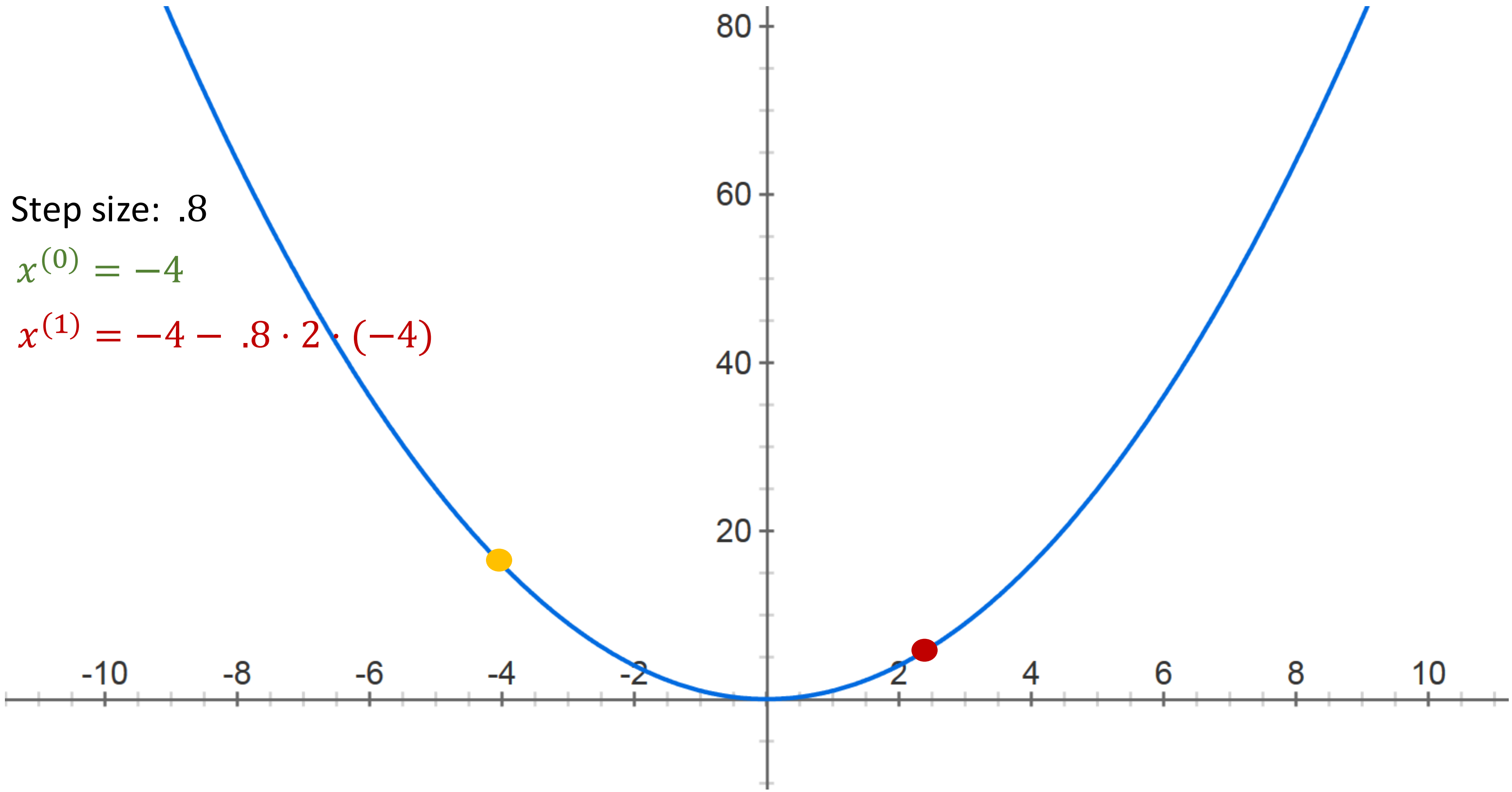
Gradient Descent

$f(x) = x^2$ $\partial x = 2\,x$

Step size: .8

$x^{(0)} = -4$

$x^{(1)} = -4 - .8 \cdot 2 \cdot (-4)$



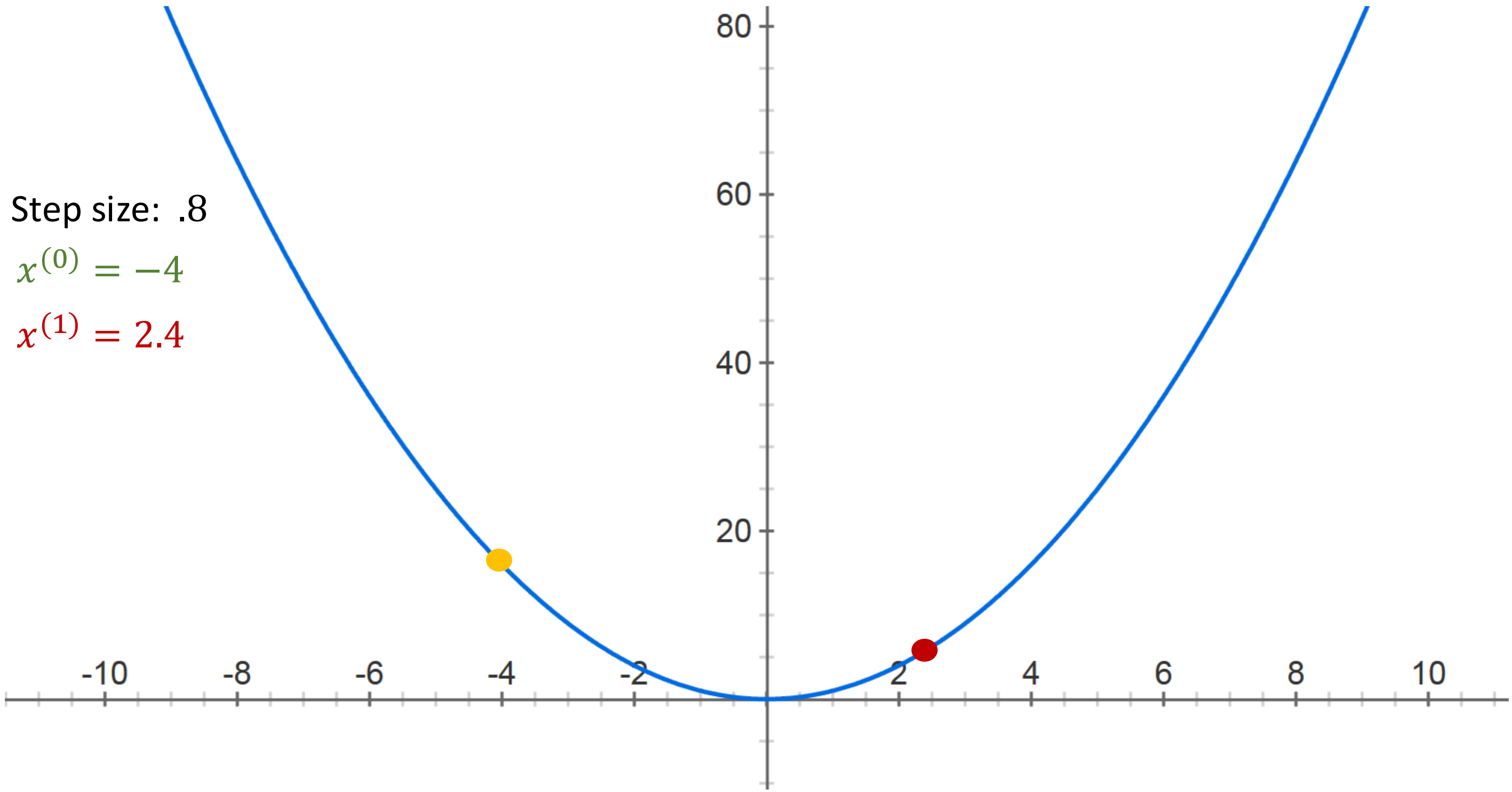
Gradient Descent

$f(x) = x^2$

Step size: .8

$x^{(0)} = -4$

$x^{(1)} = 2.4$



Gradient Descent

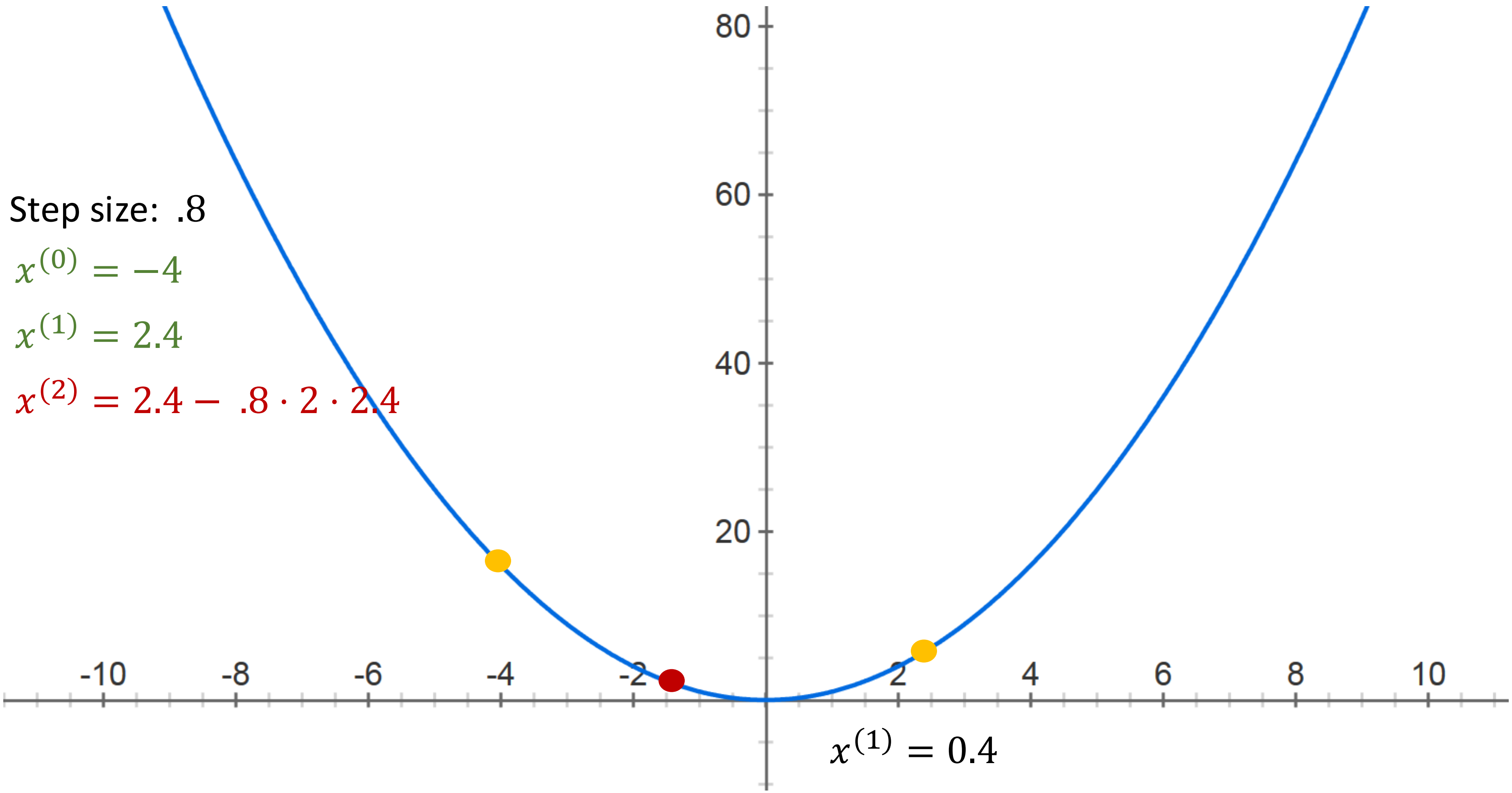
$f(x) = x^2$ $\qquad$ $\partial x = 2\,x$

Step size: .8

$x^{(0)} = -4$

$x^{(1)} = 2.4$

$x^{(2)} = 2.4 - \ .8 \cdot 2 \cdot 2.4$



$x^{(1)} = 0.4$

Gradient Descent

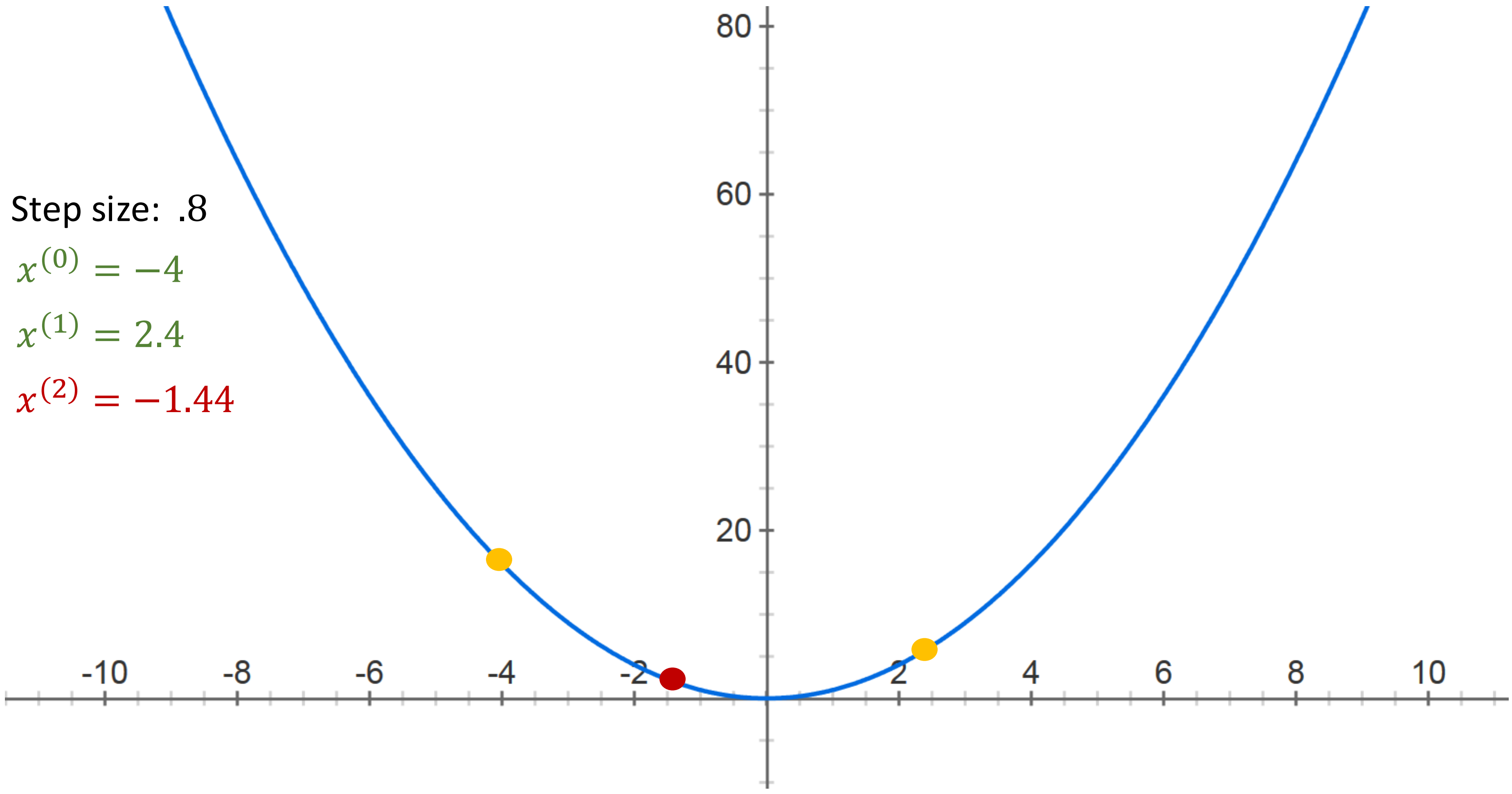
$f(x) = x^2$

Step size: .8

$x^{(0)} = -4$

$x^{(1)} = 2.4$

$x^{(2)} = -1.44$



Gradient Descent

$f(x) = x^2$

Step size: .8

$x^{(0)} = -4$

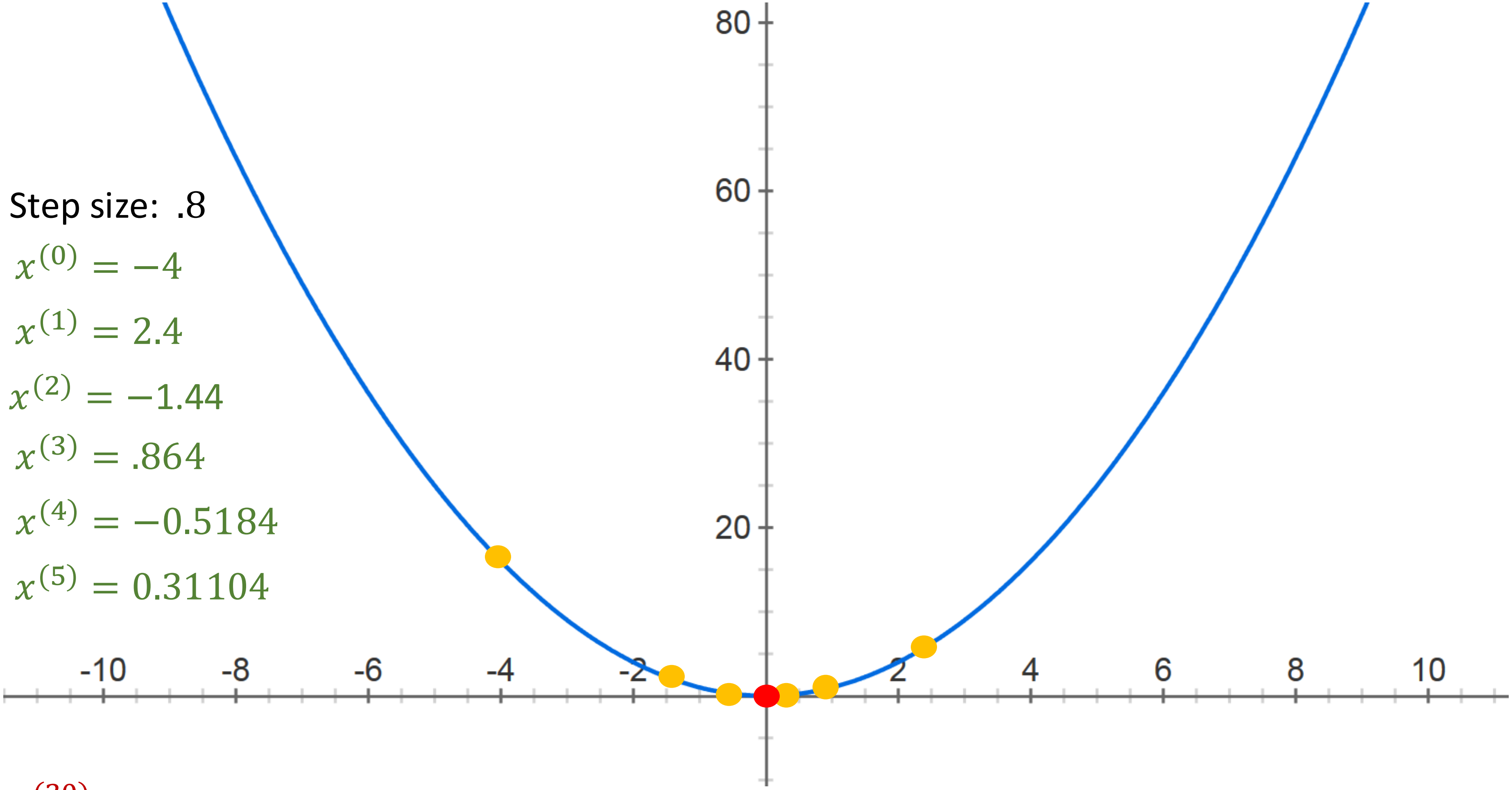
$x^{(1)} = 2.4$

$x^{(2)} = -1.44$

$x^{(3)} = .864$

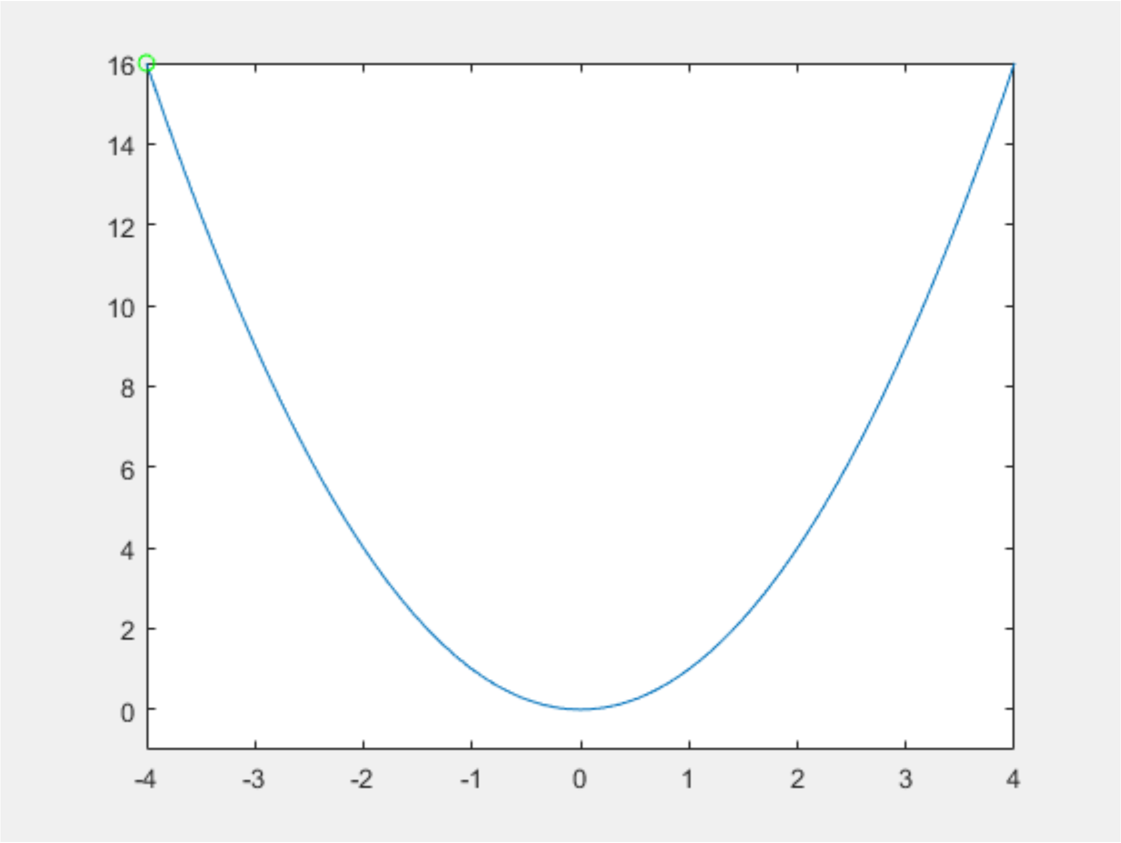
$x^{(4)} = -0.5184$

$x^{(5)} = 0.31104$

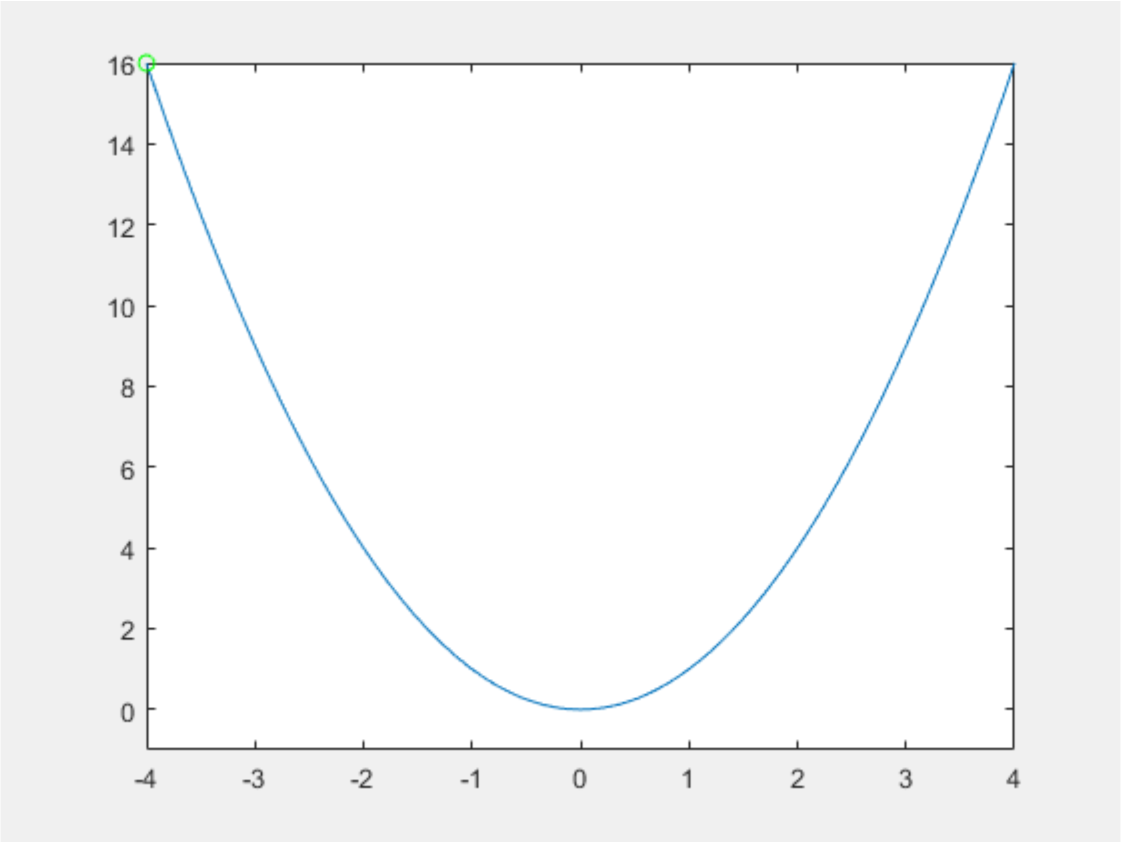


$x^{(30)} = -8.84296e - 07$

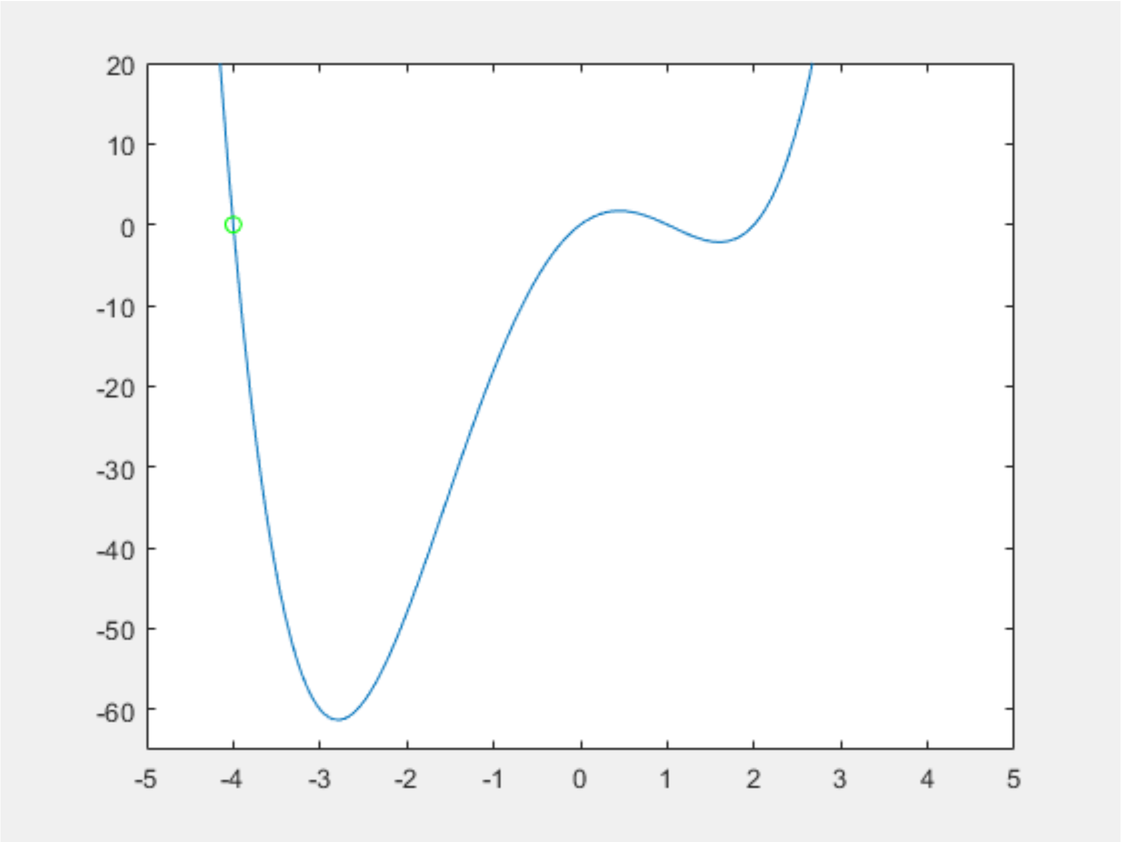
Gradient Descent



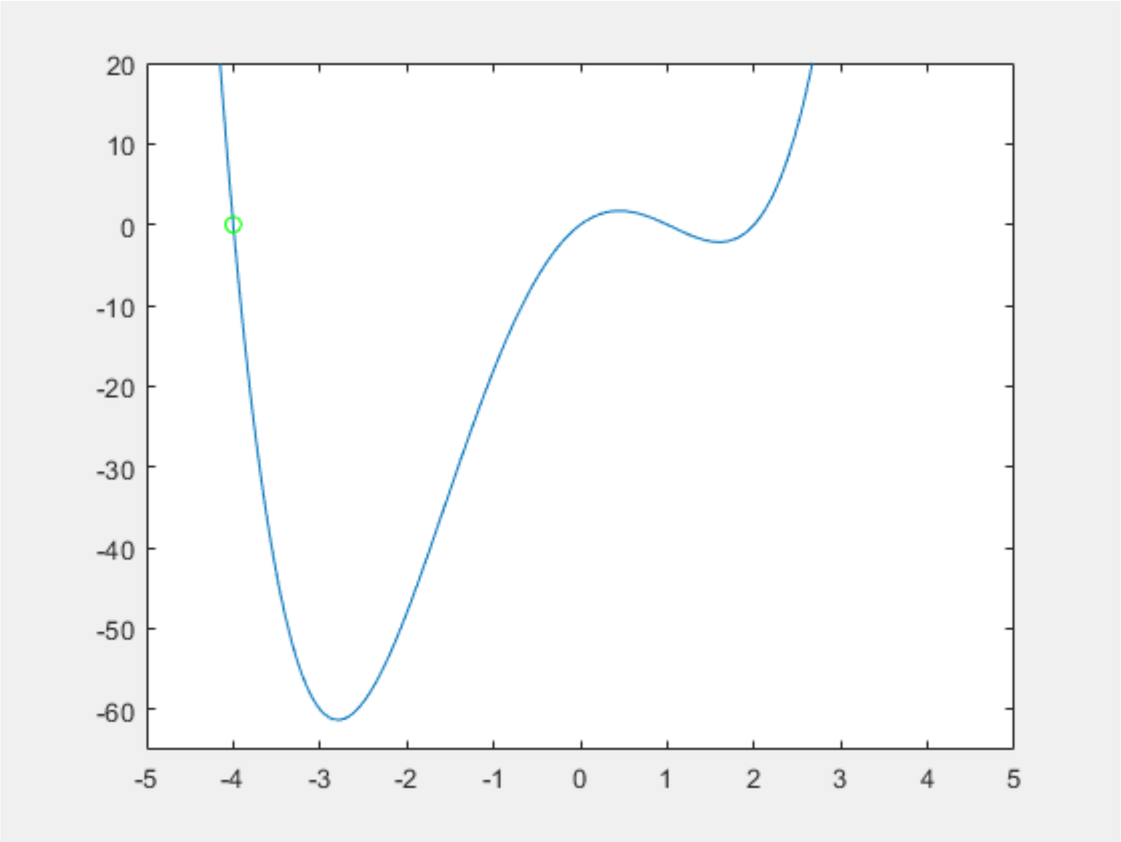
Gradient Descent



Gradient Descent



Gradient Descent

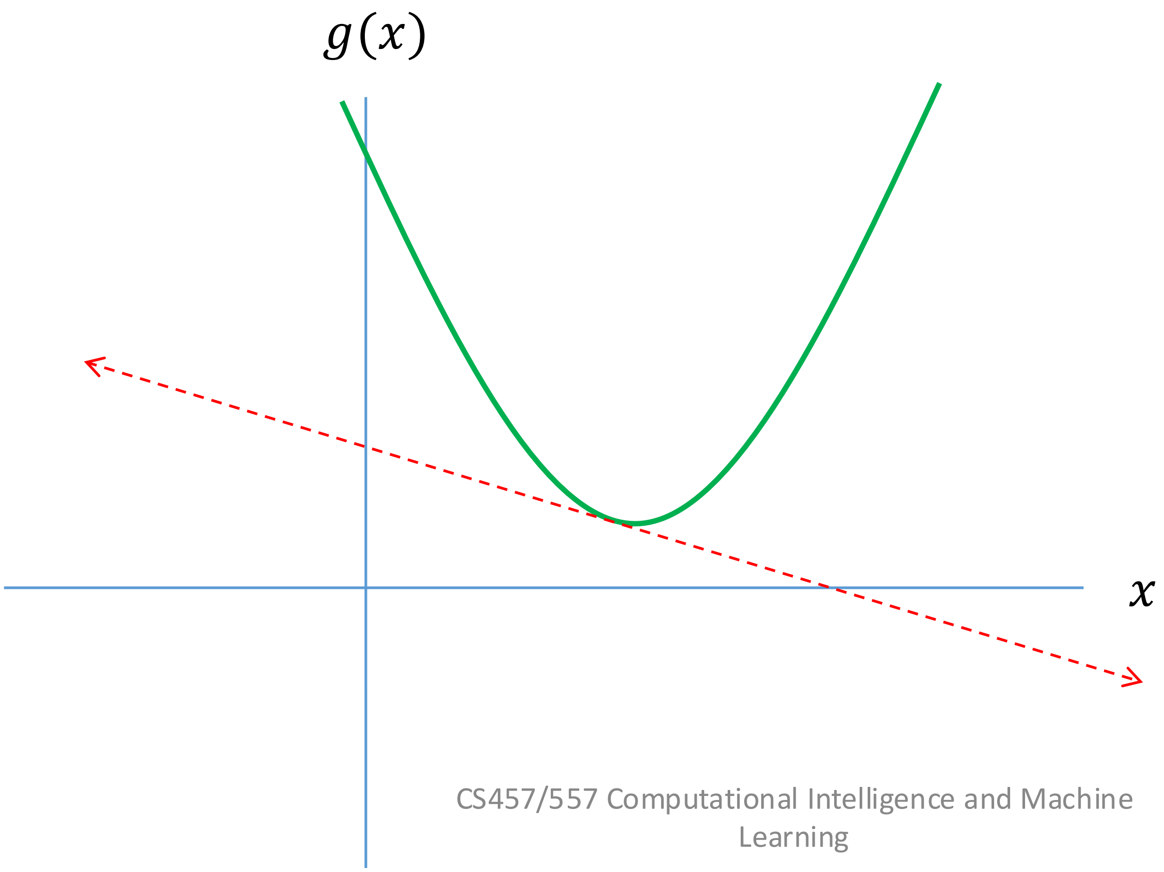


Gradient Descent: Convex Functions

- For convex functions, local optima are always global optima (this follows from the definition of convexity)
  - If gradient descent converges to a critical point, then the result is a global minimizer
- Not all convex functions are differentiable, can we still apply gradient descent?

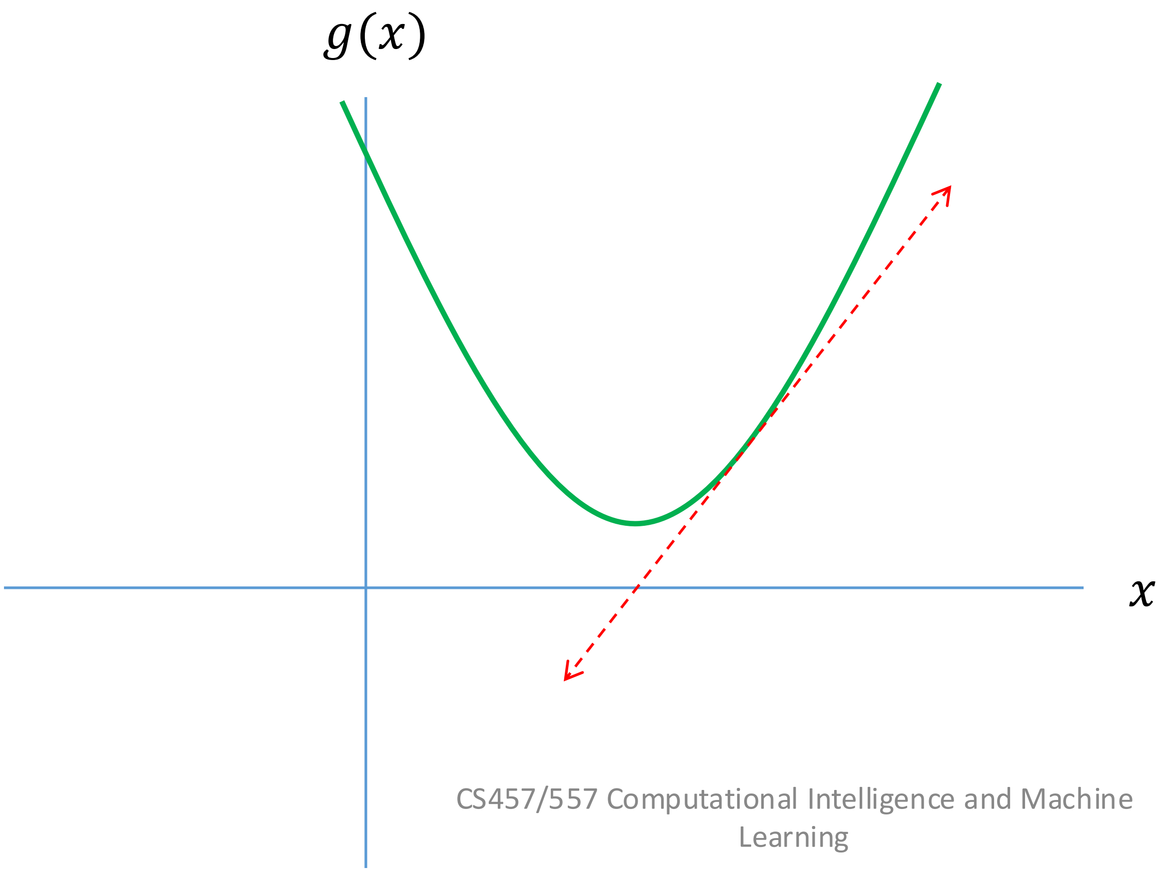
Gradients of Convex Functions

- For a differentiable convex function $g(x)$ its gradients yield **linear underestimators**



Gradients of Convex Functions

- For a differentiable convex function $g(x)$ its gradients yield **linear underestimators**



Gradients of Convex Functions

- For a differentiable convex function $g(x)$ its gradients yield **linear underestimators**: zero gradient corresponds to a global optimum