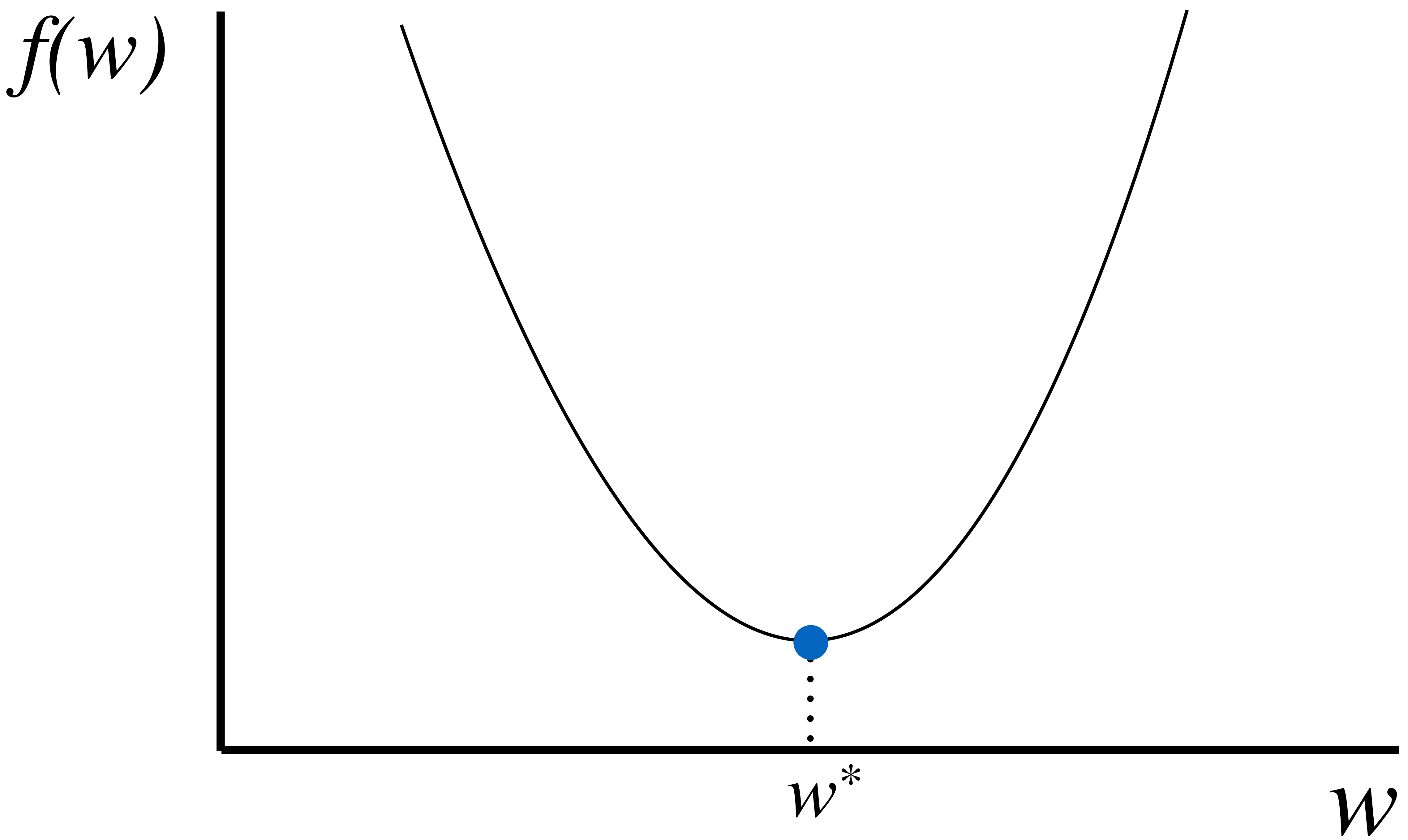
Linear Regression Optimization

Goal: Find $\mathbf{w}^*$ that minimizes

$$f(\mathbf{w}) = ||\mathbf{X}\mathbf{w} - \mathbf{y}||_2^2$$

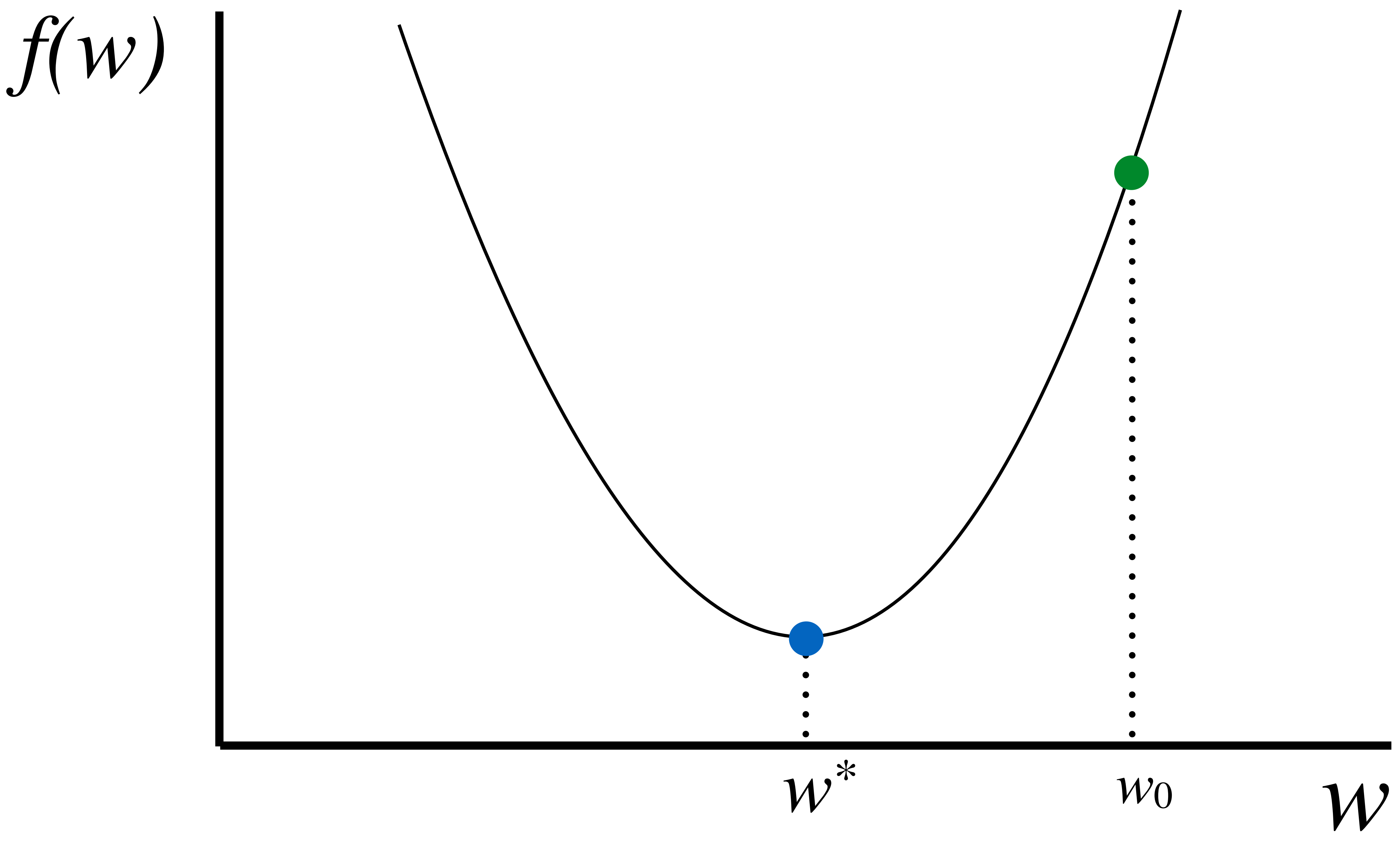
- Closed form solution exists
- Gradient Descent is iterative
(Intuition: go downhill!)



Scalar objective: $f(w) = ||w\mathbf{x} - \mathbf{y}||_2^2 = \sum_{j=1}^{n} (wx^{(j)} - y^{(j)})^2$

Gradient Descent

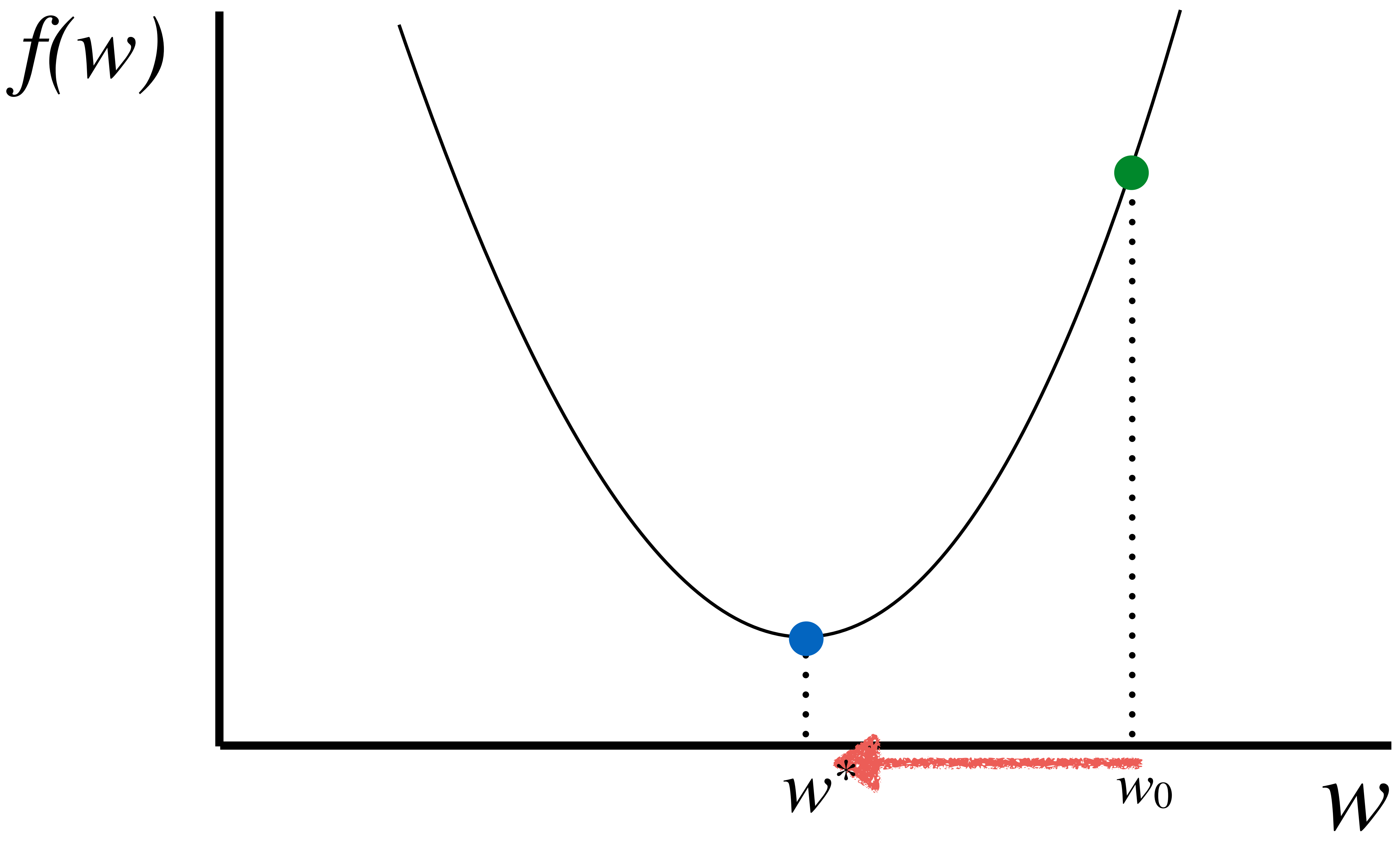
Start at a random point



Gradient Descent

Start at a random point

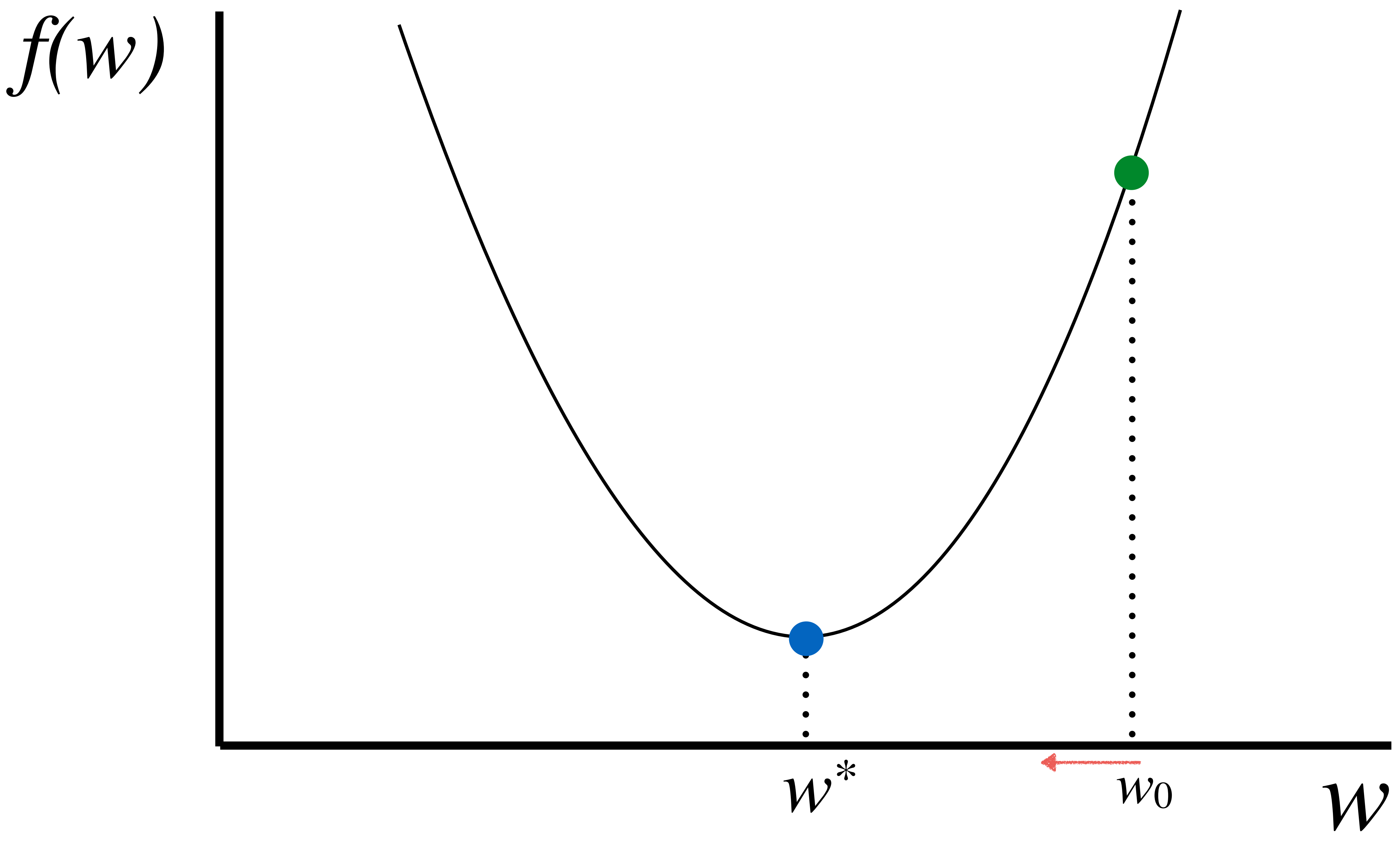
Determine a descent direction



Gradient Descent

Start at a random point

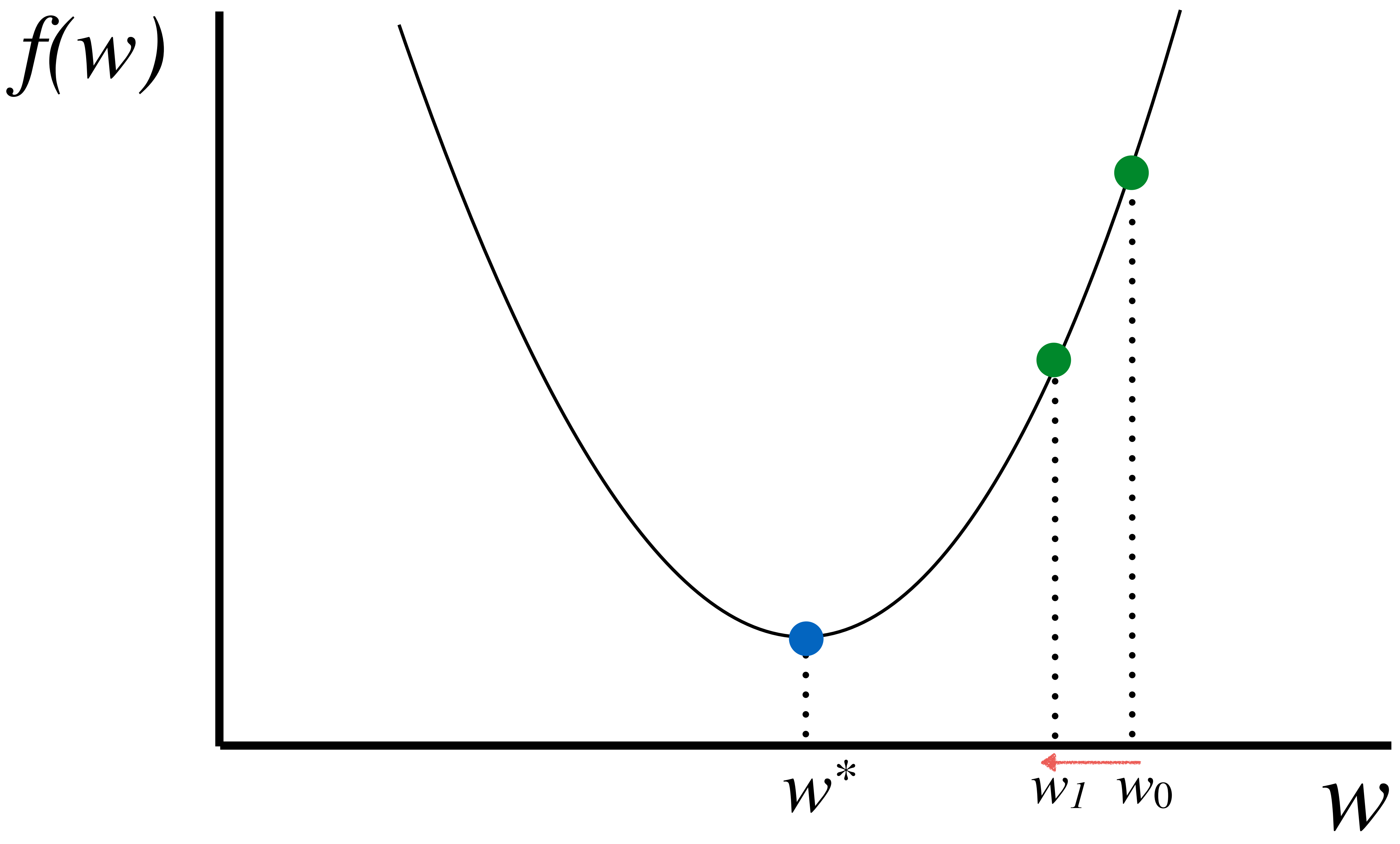
Determine a descent direction
Choose a step size



Gradient Descent

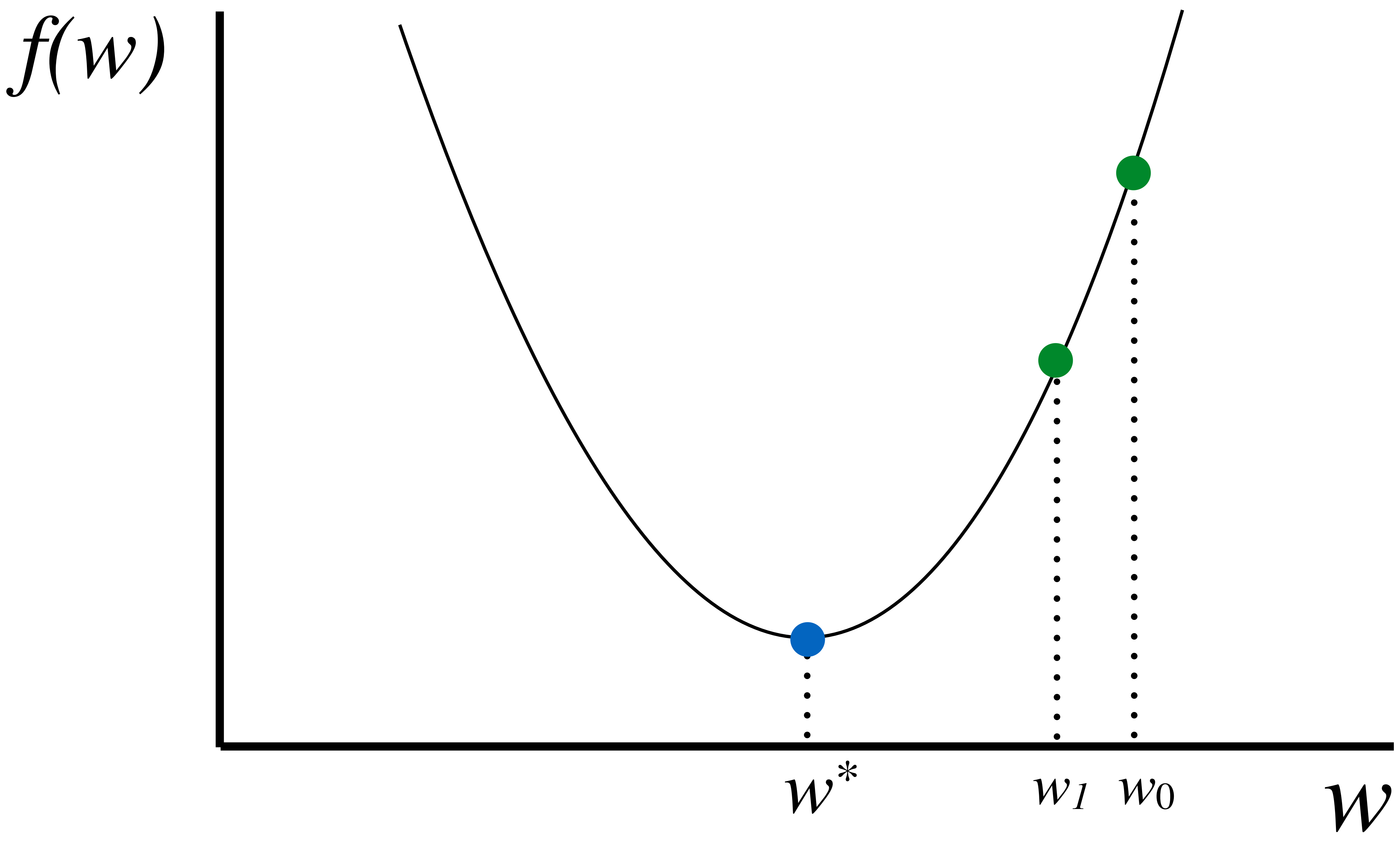
Start at a random point

Determine a descent direction
Choose a step size
Update



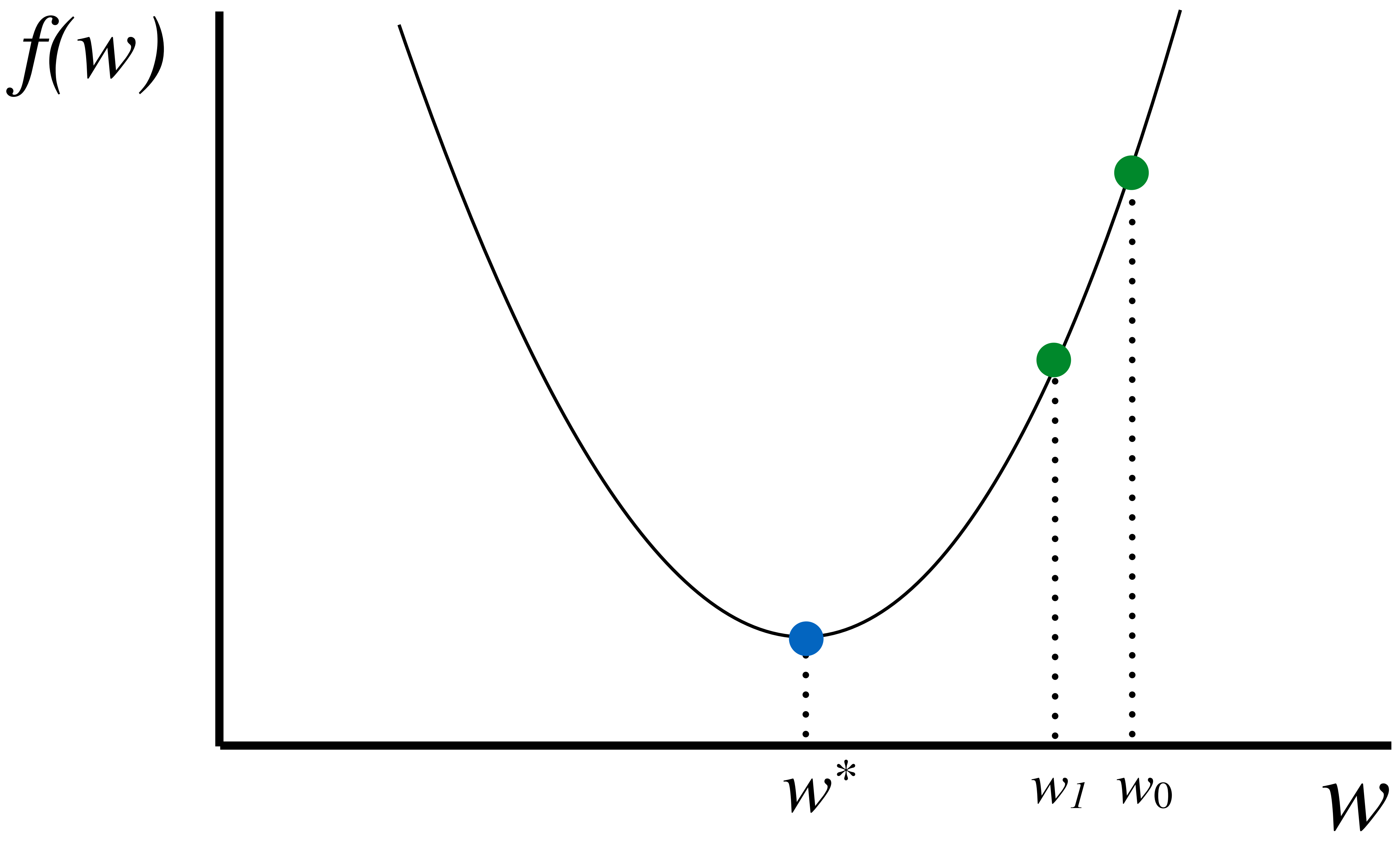
Gradient Descent

Start at a random point
Repeat
    Determine a descent direction
    Choose a step size
    Update
Until stopping criterion is satisfied



Gradient Descent

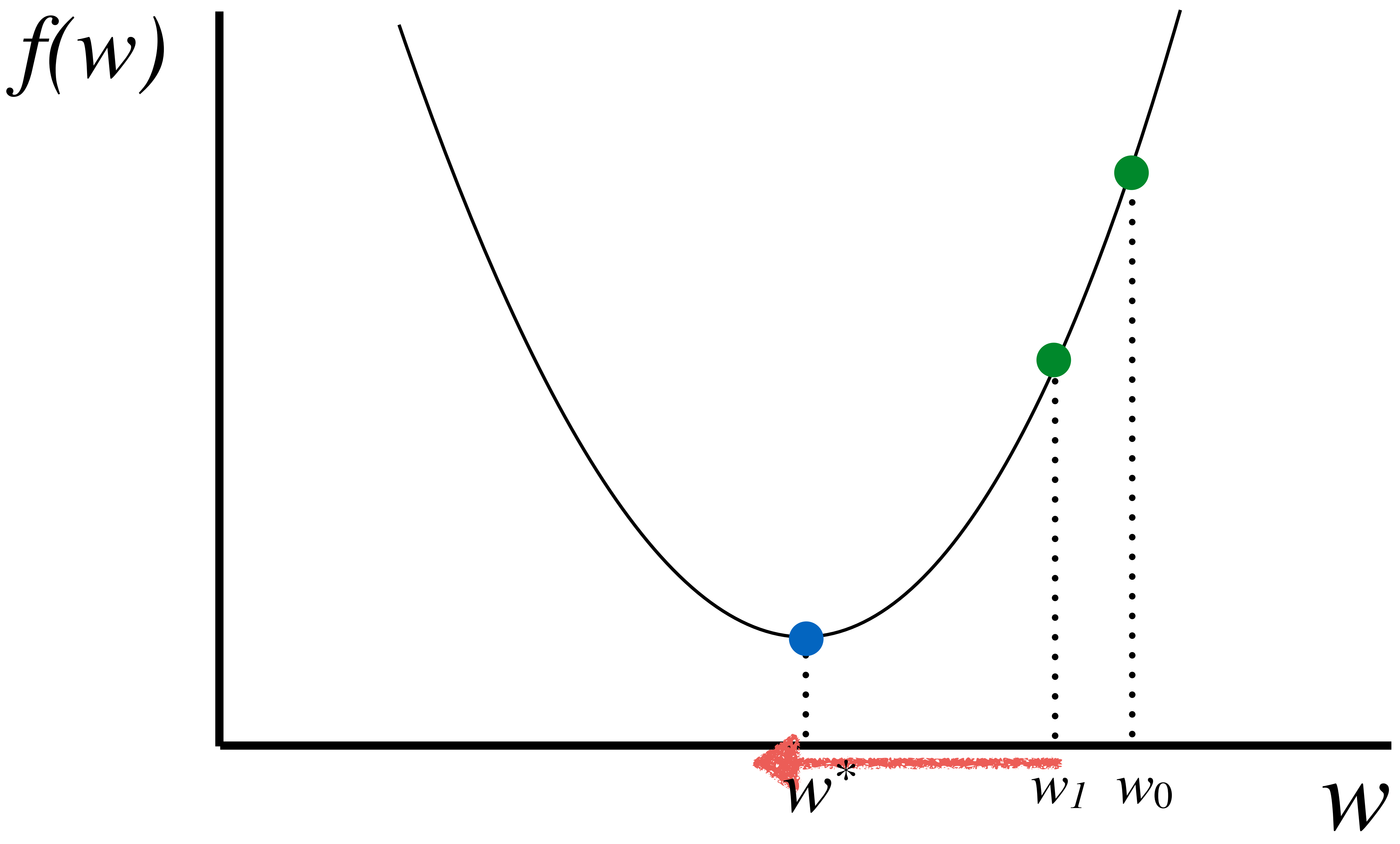
Start at a random point
Repeat
    Determine a descent direction
    Choose a step size
    Update
Until stopping criterion is satisfied



Gradient Descent

Start at a random point
Repeat
- Determine a descent direction
- Choose a step size
- Update

Until stopping criterion is satisfied



Gradient Descent

Start at a random point

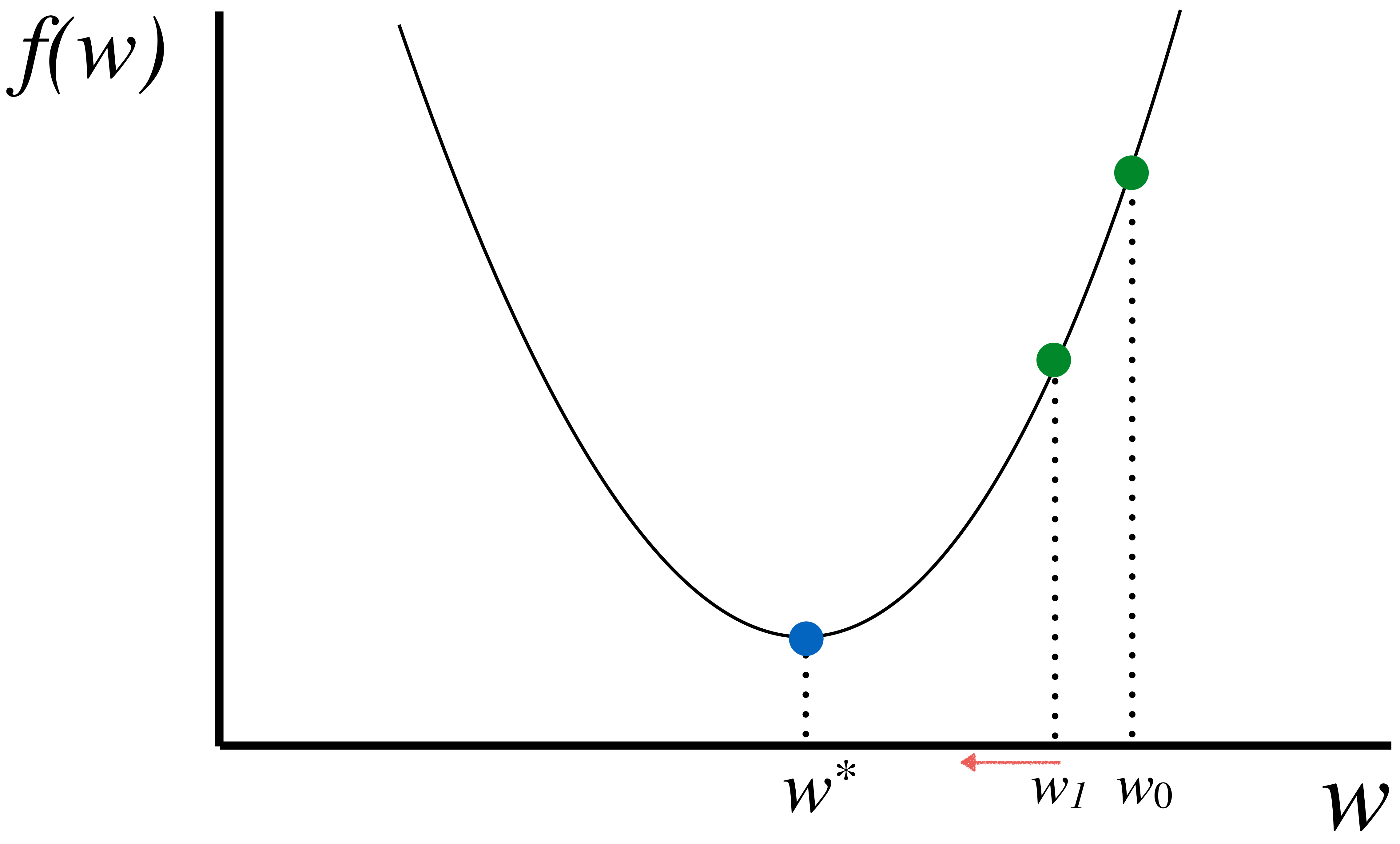
Repeat

Determine a descent direction

Choose a step size

Update

Until stopping criterion is satisfied



Gradient Descent

Start at a random point

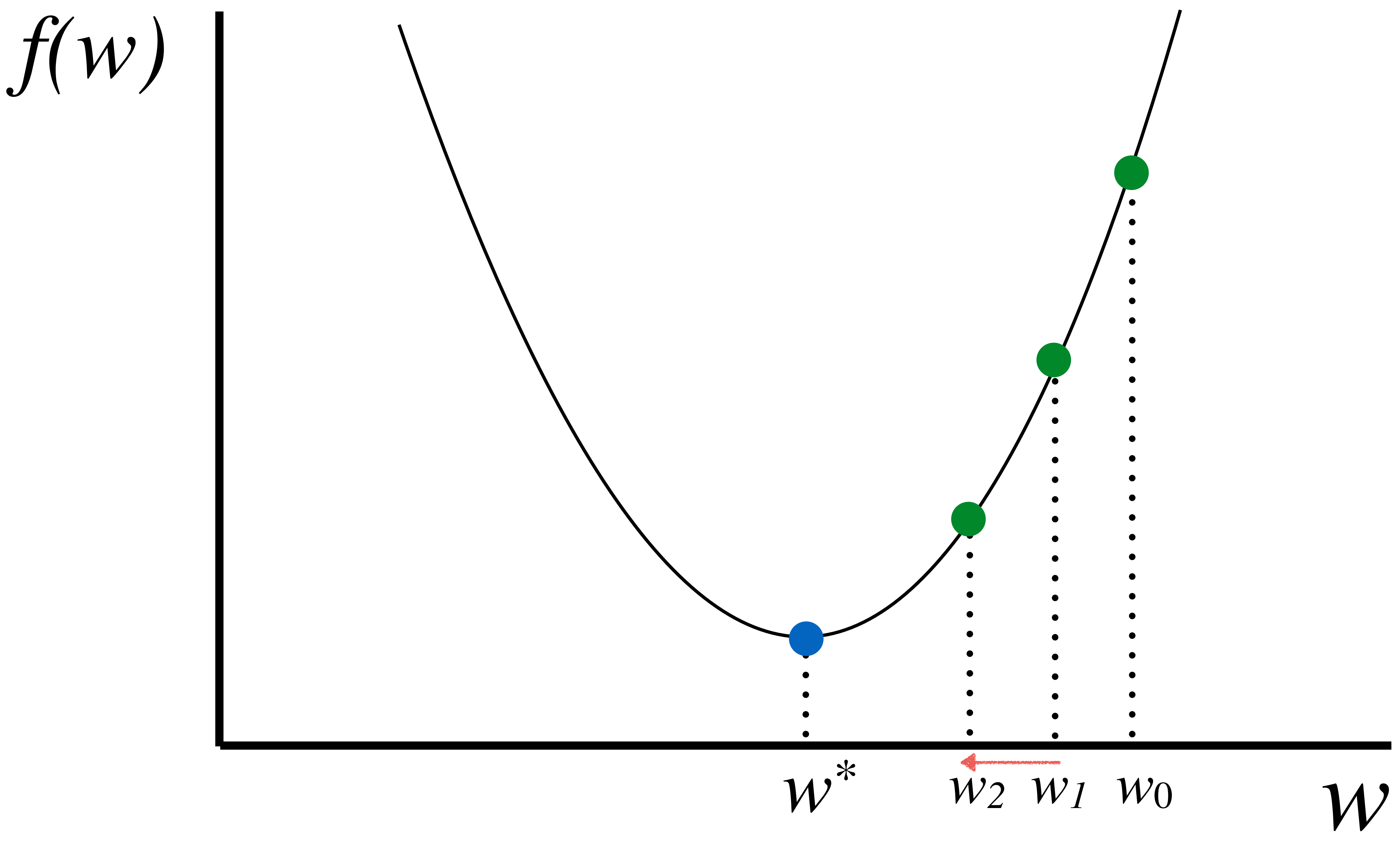
Repeat

Determine a descent direction

Choose a step size

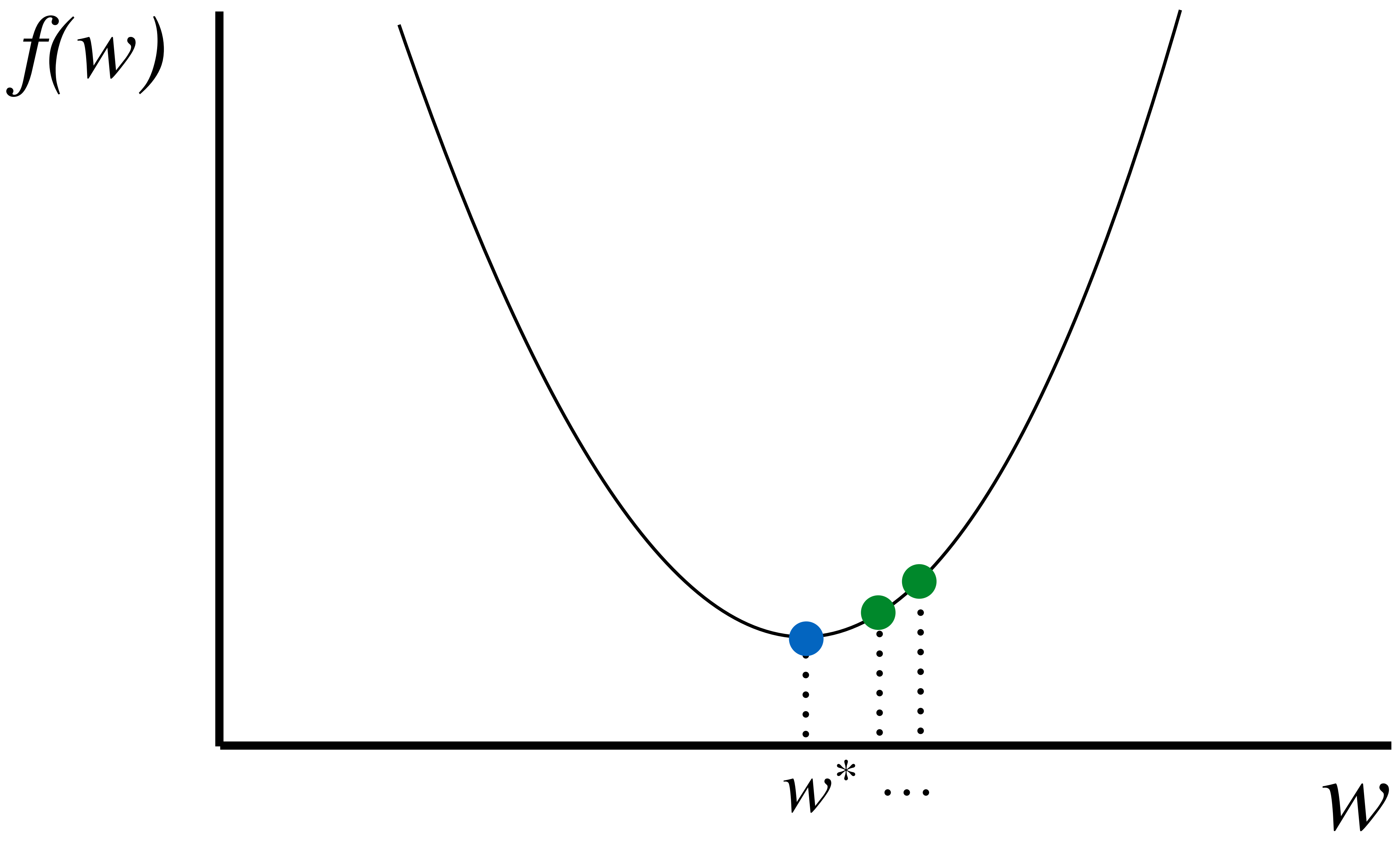
Update

Until stopping criterion is satisfied



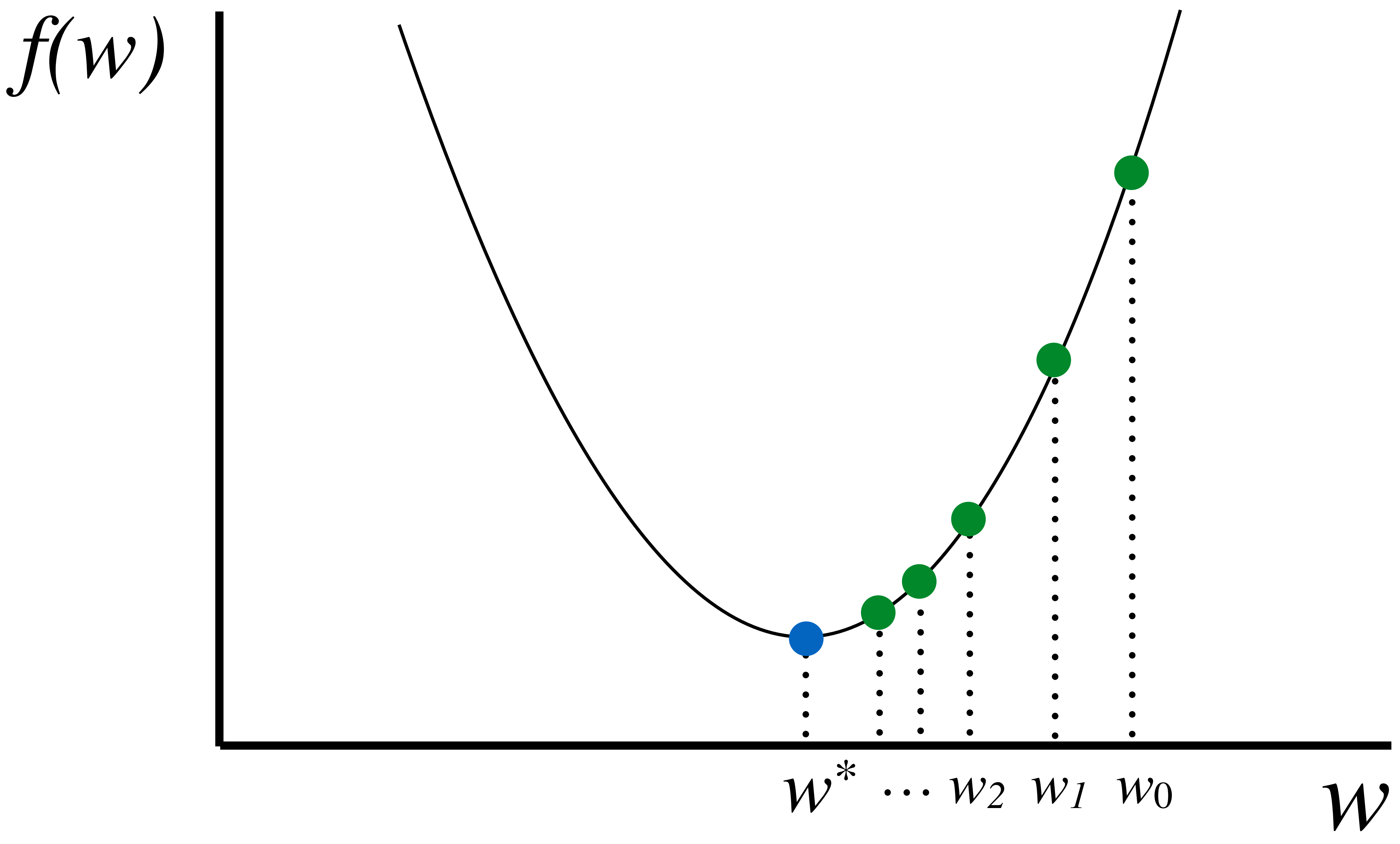
Gradient Descent

Start at a random point
Repeat
    Determine a descent direction
    Choose a step size
    Update
Until stopping criterion is satisfied

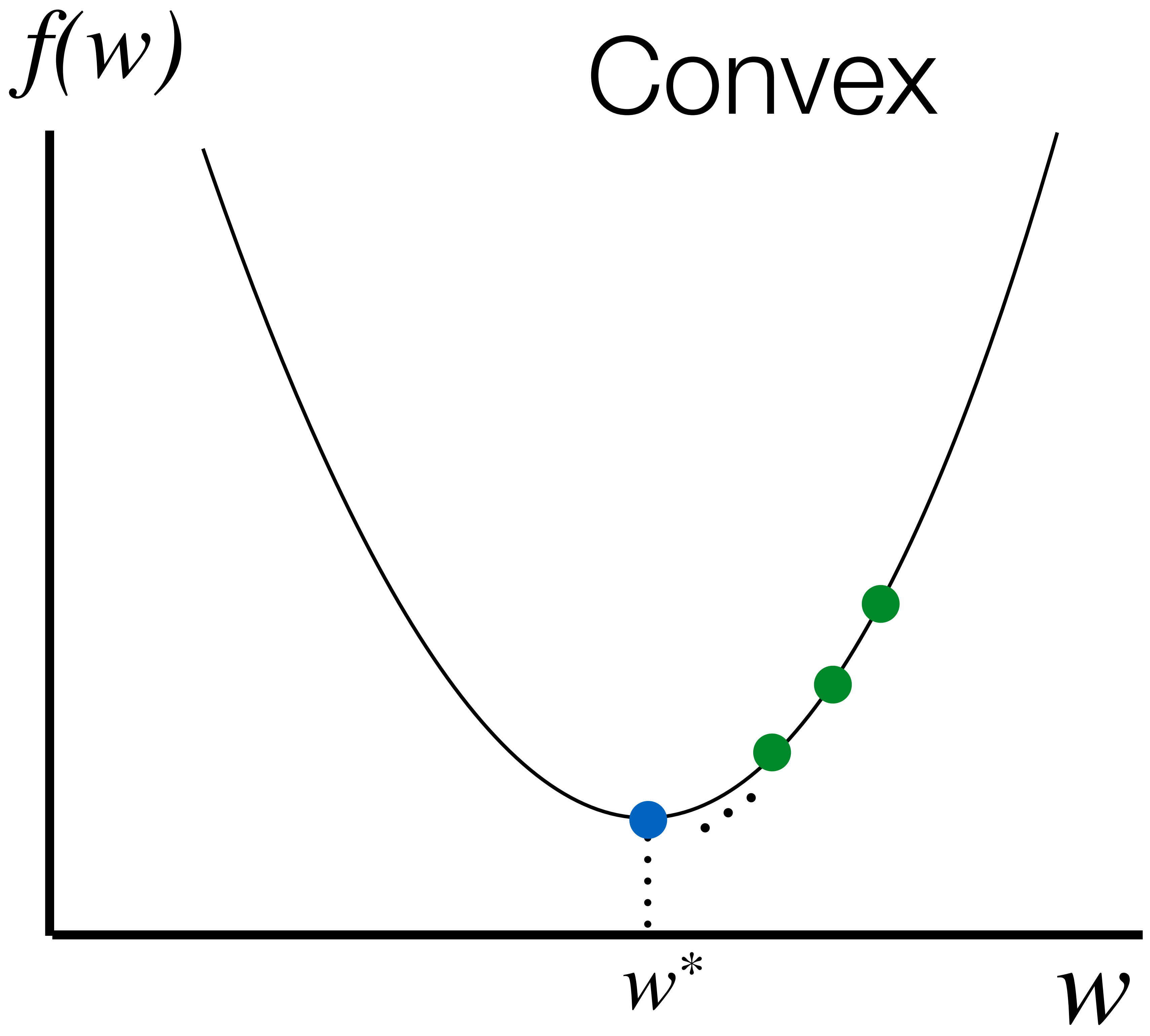


Gradient Descent

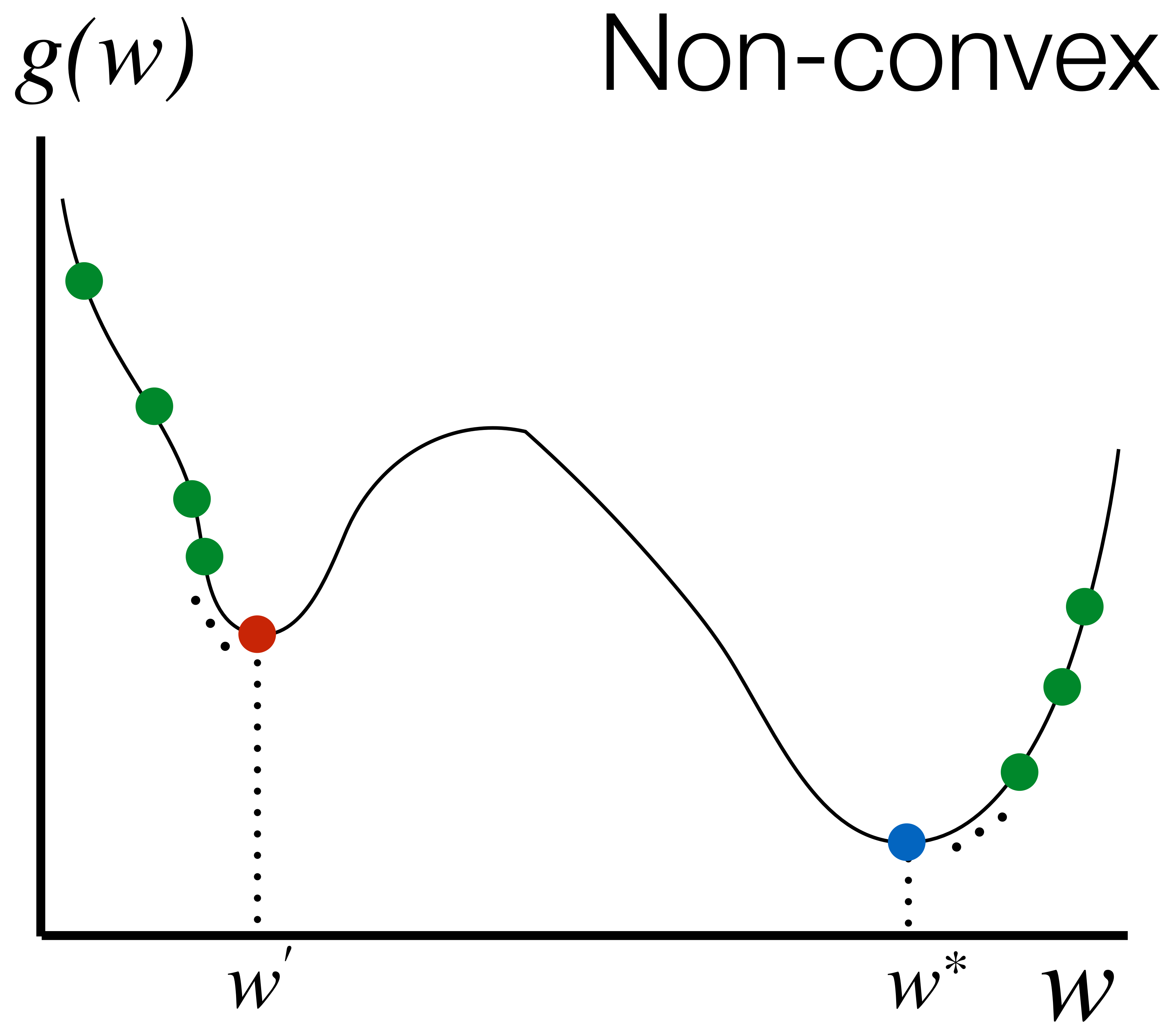
Start at a random point
Repeat
    Determine a descent direction
    Choose a step size
    Update
Until stopping criterion is satisfied



Where Will We Converge?



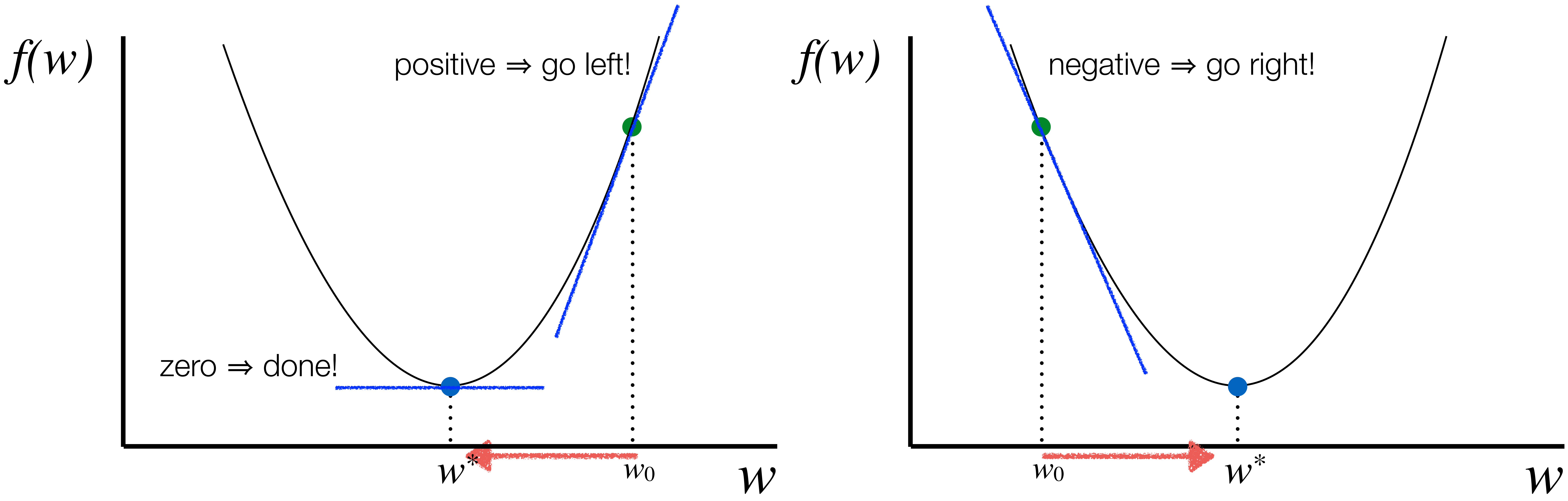
Any local minimum is a global minimum



Multiple local minima may exist

Least Squares, Ridge Regression and Logistic Regression are all convex!

Choosing Descent Direction (1D)



We can only move in two directions

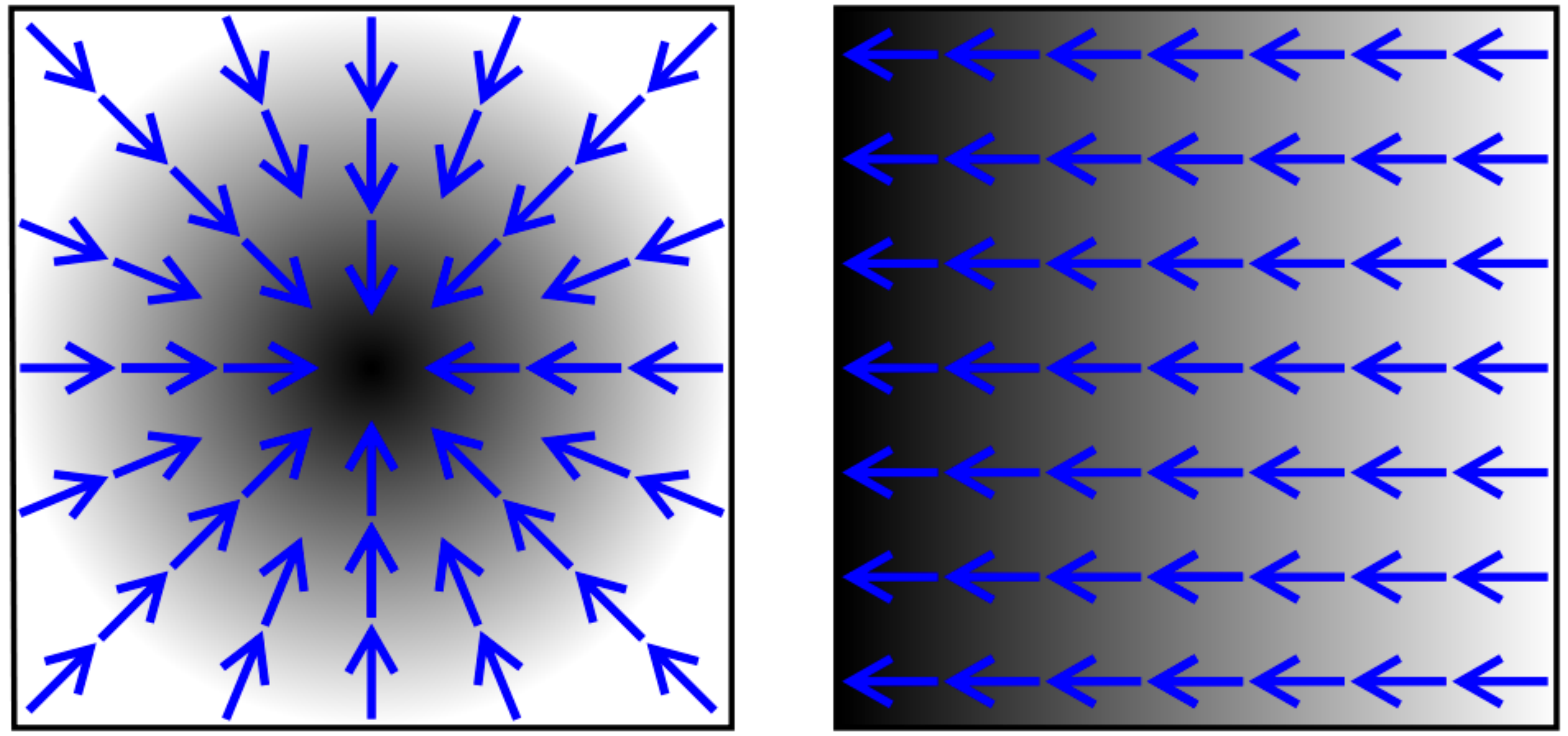
Negative slope is direction of descent!

Step Size

Update Rule: $w_{i+1} = w_i - \alpha_i \frac{df}{dw}(w_i)$

Negative Slope

Choosing Descent Direction



"Gradient2" by Sarang. Licensed under CC BY-SA 2.5 via Wikimedia Commons
http://commons.wikimedia.org/wiki/File:Gradient2.svg#/media/File:Gradient2.svg

2D Example:

- Function values are in black/white and black represents higher values
- Arrows are gradients

We can move anywhere in $\mathbb{R}^d$

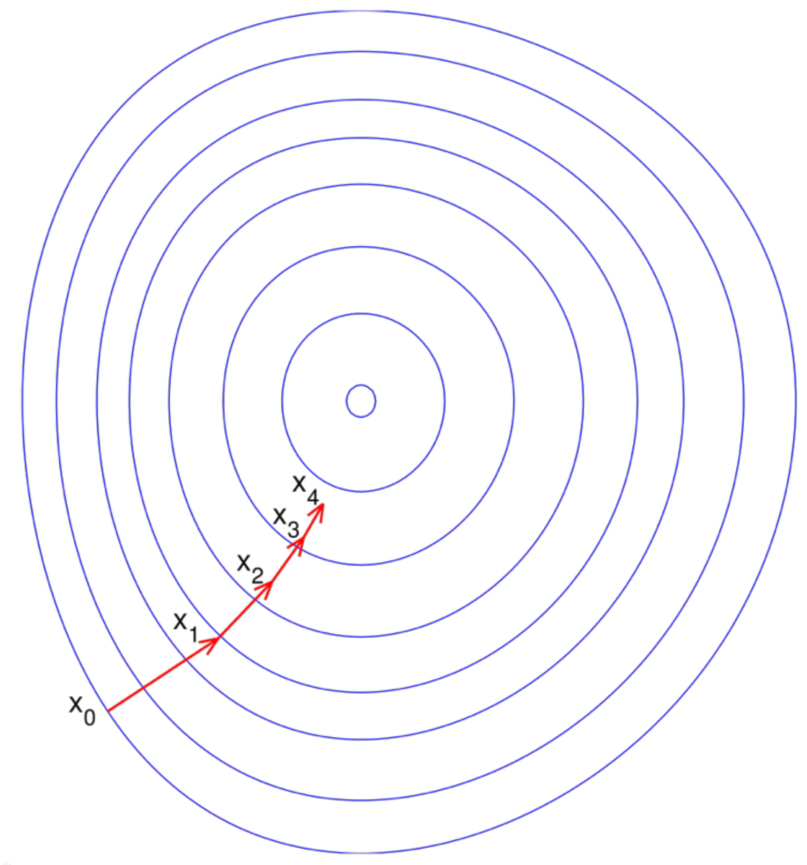
Negative gradient is direction of *steepest* descent!

Step Size

Update Rule: $\mathbf{w}_{i+1} = \mathbf{w}_i - \alpha_i \nabla f(\mathbf{w}_i)$

Negative Slope

Gradient Descent for Least Squares



Update Rule: $w_{i+1} = w_i - \alpha_i \frac{df}{dw}(w_i)$

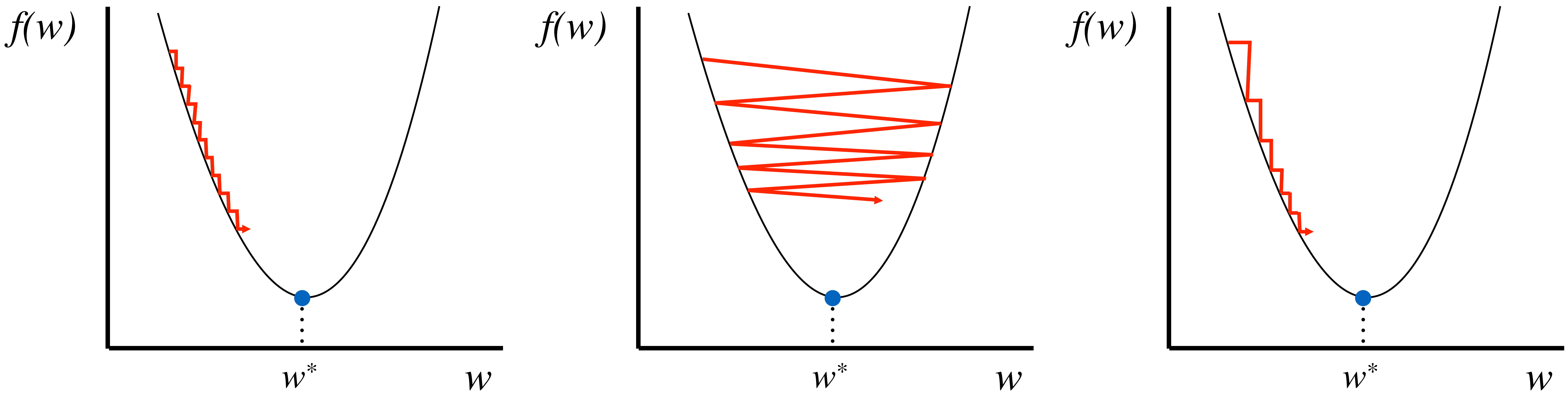
Scalar objective: $f(w) = ||w\mathbf{x} - \mathbf{y}||_2^2 = \sum_{j=1}^{n} (wx^{(j)} - y^{(j)})^2$

Derivative: $\frac{df}{dw}(w) = 2\sum_{j=1}^{n} (wx^{(j)} - y^{(j)})x^{(j)}$
(chain rule)

Scalar Update: $w_{i+1} = w_i - \alpha_i \sum_{j=1}^{n} (w_i x^{(j)} - y^{(j)})x^{(j)}$
(2 absorbed in α)

Vector Update: $\mathbf{w}_{i+1} = \mathbf{w}_i - \alpha_i \sum_{j=1}^{n} (\mathbf{w}_i^\top \mathbf{x}^{(j)} - y^{(j)})\mathbf{x}^{(j)}$

Choosing Step Size



Too small: converge very slowly

Too big: overshoot and even diverge

Reduce size over time

Theoretical convergence results for various step sizes

A common step size is $\alpha_i = \dfrac{\alpha}{n\sqrt{i}}$

Constant

Iteration #

Training Points

Gradient Descent Summary

Pros:

- Easily parallelized
- Cheap at each iteration
- Stochastic variants can make things even cheaper

Cons:

- Slow convergence (especially compared with closed-form)
- **Requires communication across nodes!**