Gradient Descent

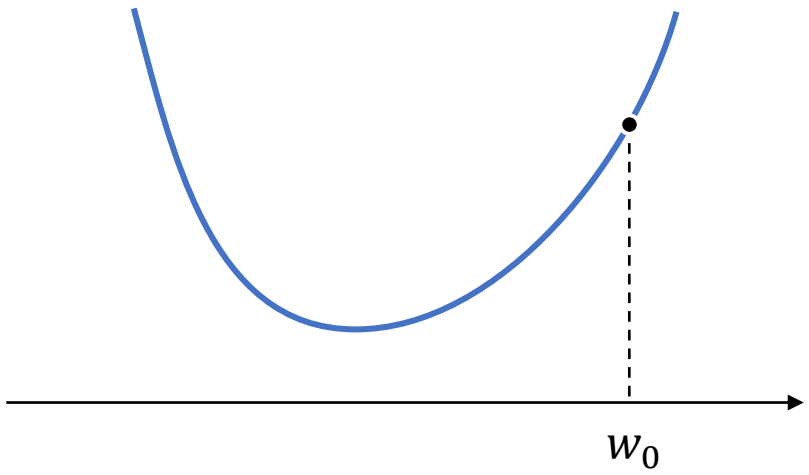
Hao-Wen Dong

Material based on Intro to Machine Learning (CSE 251A), Fall 2021

Gradient descent

- Pseudocode:
  - Choose an initial weight vector w_0 and learning rate η
  - Repeat until convergence:

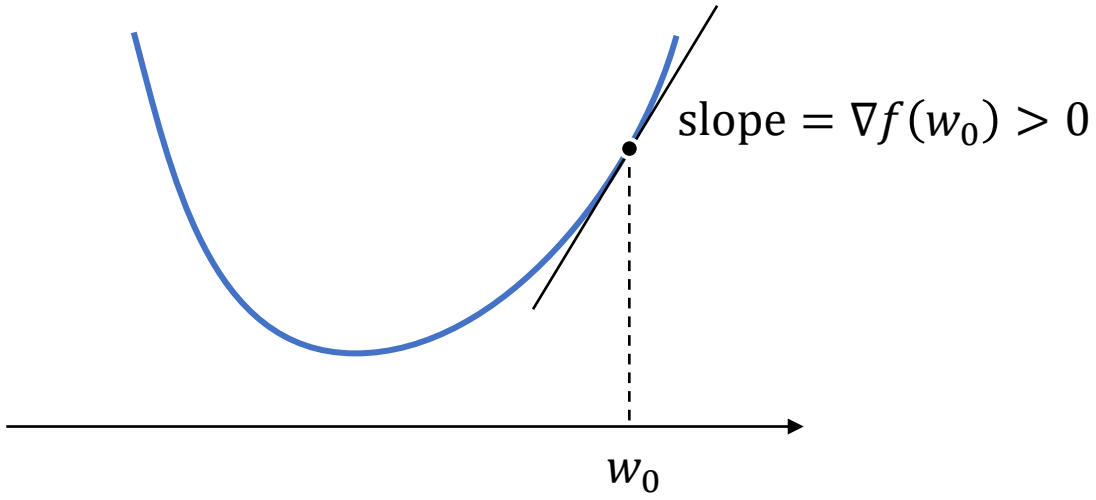
$$w_{t+1} = w_t - \eta \nabla f(w_t)$$



Gradient descent

- Pseudocode:
  - Choose an initial weight vector w_0 and learning rate η
  - Repeat until convergence:

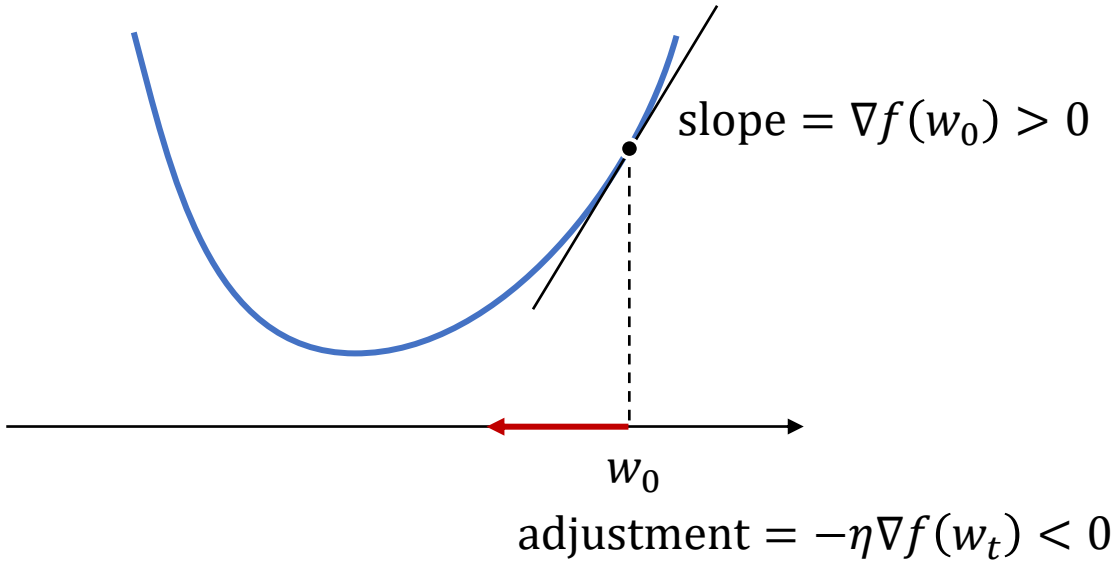
$$w_{t+1} = w_t - \eta \nabla f(w_t)$$



Gradient descent

- Pseudocode:
  - Choose an initial weight vector w_0 and learning rate η
  - Repeat until convergence:

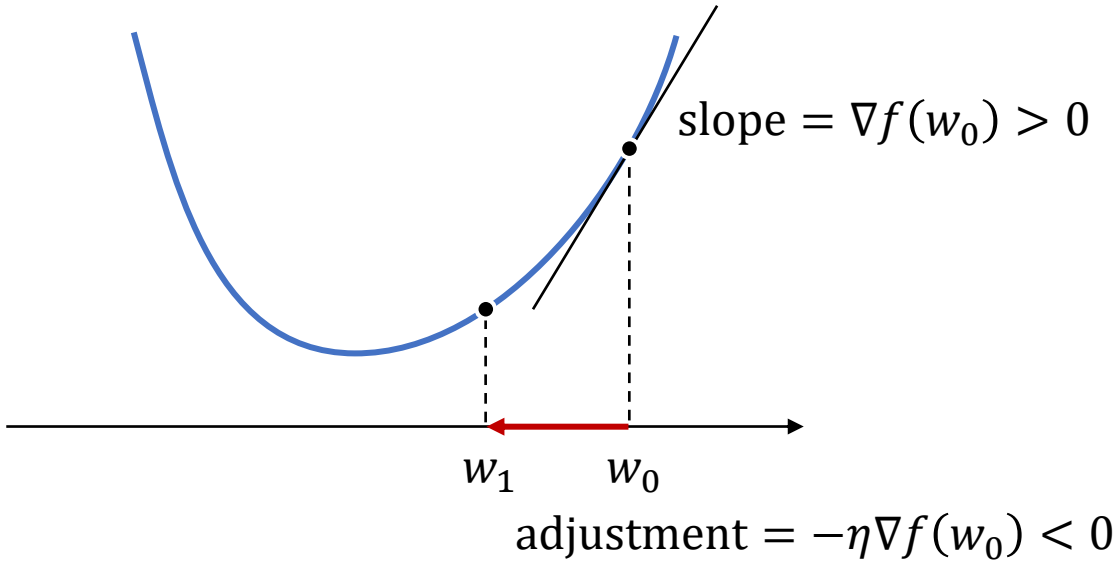
$$w_{t+1} = w_t - \eta \nabla f(w_t)$$



Gradient descent

- Pseudocode:
  - Choose an initial weight vector w_0 and learning rate η
  - Repeat until convergence:

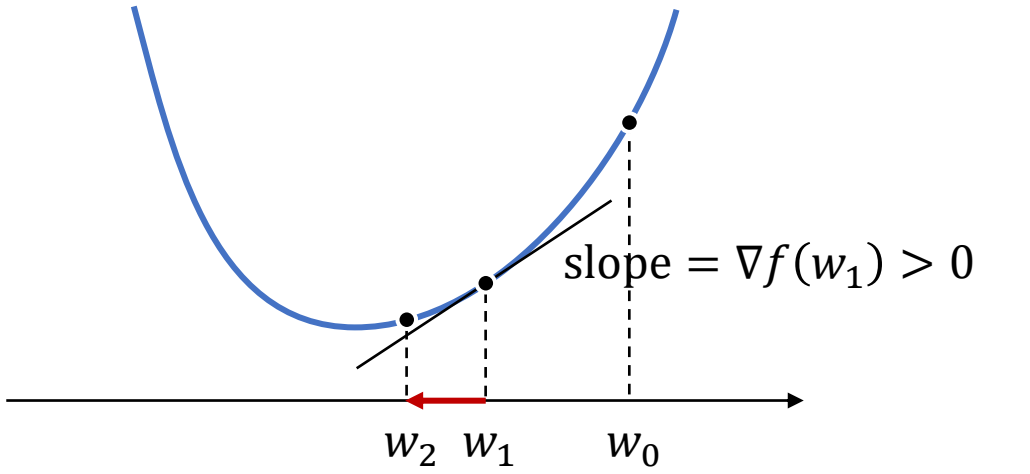
$$w_{t+1} = w_t - \eta \nabla f(w_t)$$



Gradient descent

- Pseudocode:
  - Choose an initial weight vector w_0 and learning rate η
  - Repeat until convergence:

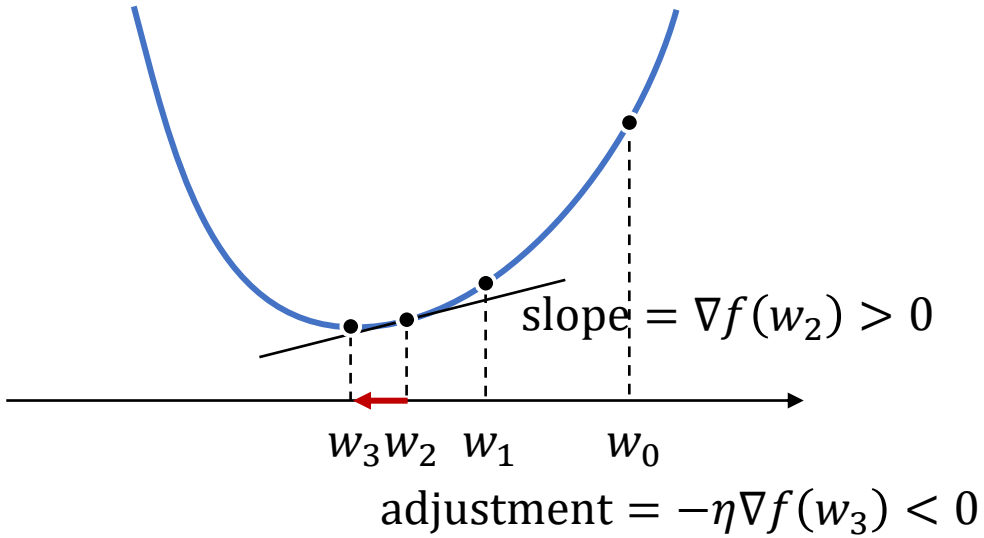
$$w_{t+1} = w_t - \eta \nabla f(w_t)$$



Gradient descent

- Pseudocode:
  - Choose an initial weight vector w_0 and learning rate η
  - Repeat until convergence:

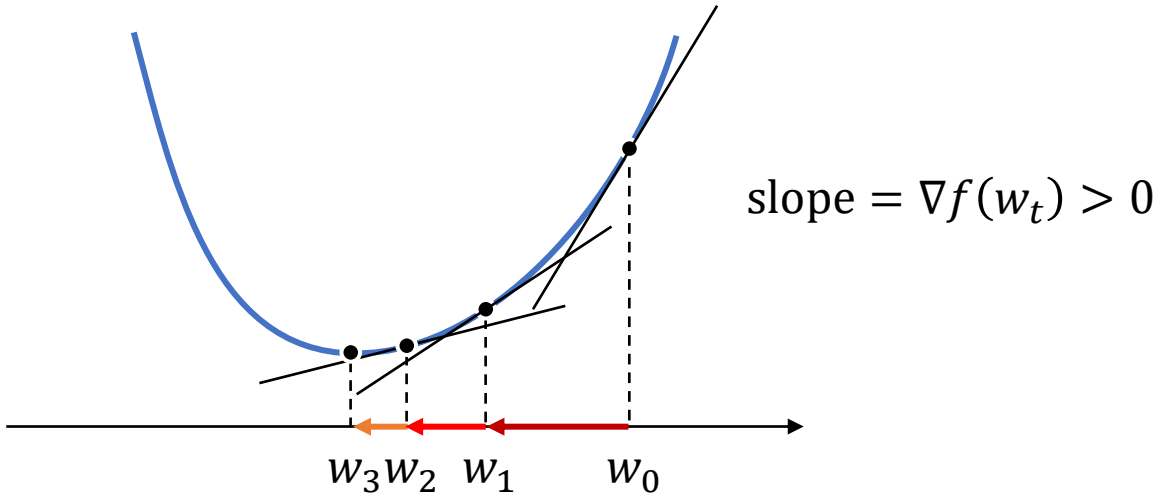
$$w_{t+1} = w_t - \eta \nabla f(w_t)$$



Gradient descent

- Pseudocode:
  - Choose an initial weight vector w_0 and learning rate η
  - Repeat until convergence:

$$w_{t+1} = w_t - \eta \nabla f(w_t)$$



$$\text{slope} = \nabla f(w_t) > 0$$

$$\text{adjustment} = -\eta \nabla f(w_t) < 0$$

Stochastic gradient descent

- Pseudocode:
  - Choose an initial weight vector w_0 and learning rate η
  - Repeat until convergence:
    - Randomly pick a sample (x, y)
    - Update the weight

$$w_{t+1} = w_t - \eta \nabla g(w_t, x, y)$$

Stochastic gradient descent

- Pseudocode:
  - Choose an initial weight vector w_0 and learning rate η
  - Repeat until convergence:
    - Randomly pick a sample (x, y)
    - Update the weight

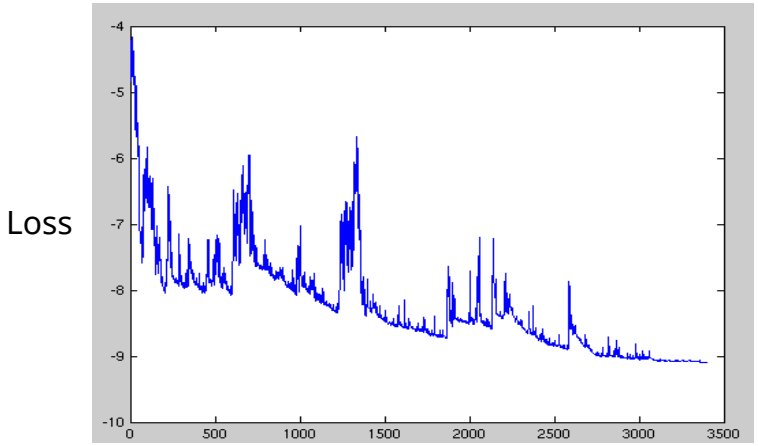
$$w_{t+1} = w_t - \eta \nabla g(w_t, x, y)$$

- Assuming $f(w_t) = \sum_{i=1}^{n} g(w_t, x_i, y_i)$
  - Total loss is the sum of sample loss
  - Holds for many ML problems

Stochastic gradient descent

- Pseudocode:
  - Choose an initial weight vector w_0 and learning rate η
  - Repeat until convergence:
    - Randomly pick a sample (x, y)
    - Update the weight

$$w_{t+1} = w_t - \eta \nabla g(w_t, x, y)$$



(By Joe pharos at the English-language Wikipedia, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=42498187)

Mini-batch stochastic gradient descent

- Pseudocode:
  - Choose an initial weight vector w_0 and learning rate η
  - Repeat until convergence:
    - Randomly pick a batch of samples $\{(x_1, y_1), (x_2, y_2), \dots\}$
    - Update the weight

$$w_{t+1} = w_t - \eta \sum_{i=1}^{n} \nabla g(w_t, x_i, y_i)$$

- Provide better estimate of the true gradient
  - Trade off between stability and speed

Comparisons

- Gradient descent (batch gradient descent): $\text{batch size} = N$
- Stochastic gradient descent: $\text{batch size} = 1$
- Mini-batch gradient descent: $1 < \text{batch size} < N$