Gradient Descent

Natalie Parde
UIC CS 421

Finding Optimal Weights

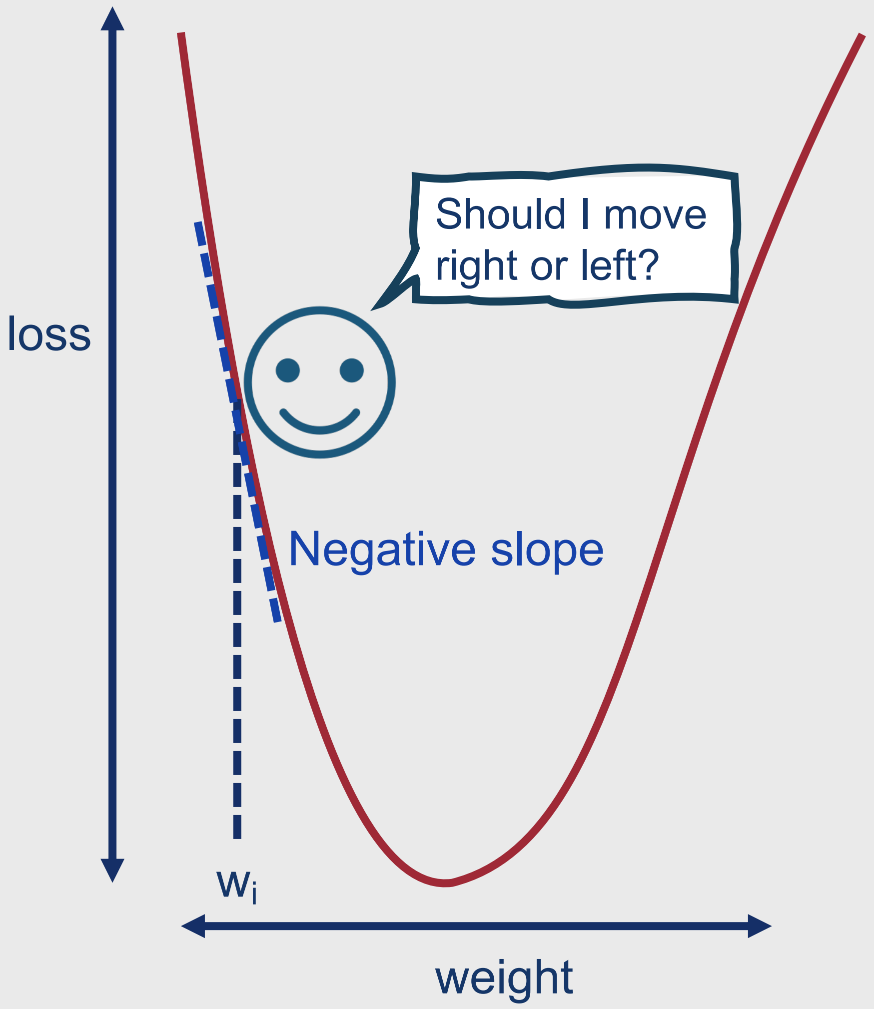
- Goal: Minimize the loss function defined for the model
  - $\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \frac{1}{m} \sum_{i=1}^{m} L_{CE}(y^{(i)}, x^{(i)}; \theta)$
- For logistic regression, $\theta = w, b$
- One way to do this is by using **gradient descent**

Gradient Descent

- Finds the minimum of a function by:
  - Figuring out the direction (in the space of θ) the function's slope
  - Moving in the opposite direction
- For logistic regression, loss functions are **convex**
  - Only one minimum
  - Gradient descent starting at any point is guaranteed to find it

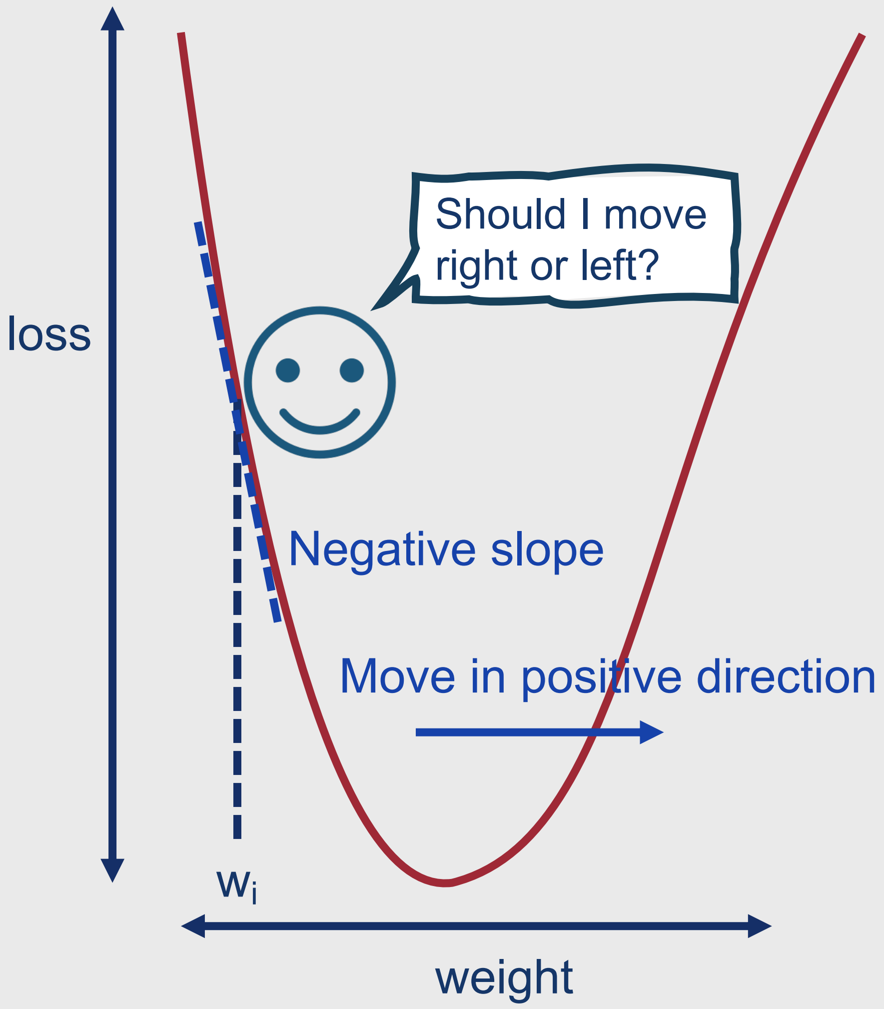
Gradient Descent

- Finds the minimum of a function by:
  - Figuring out the direction (in the space of θ) the function's slope
  - Moving in the opposite direction
- For logistic regression, loss functions are **convex**
  - Only one minimum
  - Gradient descent starting at any point is guaranteed to find it



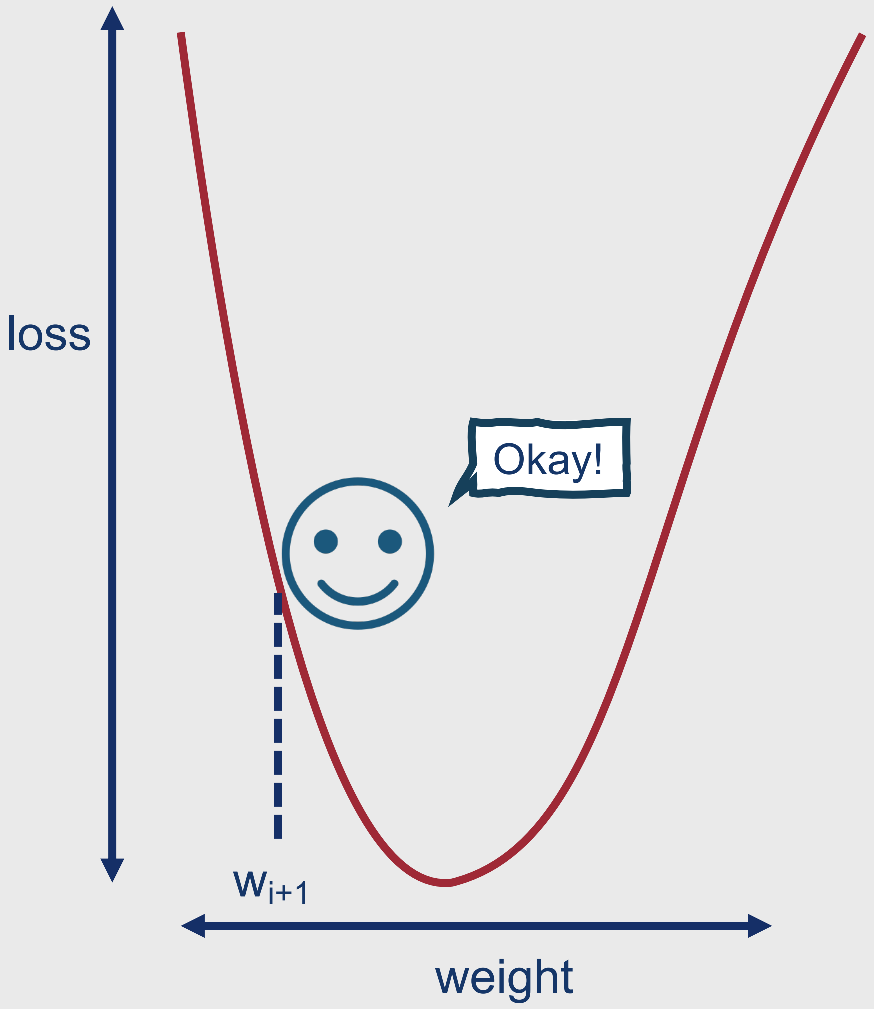
Gradient Descent

- Finds the minimum of a function by:
  - Figuring out the direction (in the space of θ) the function's slope
  - Moving in the opposite direction
- For logistic regression, loss functions are **convex**
  - Only one minimum
  - Gradient descent starting at any point is guaranteed to find it



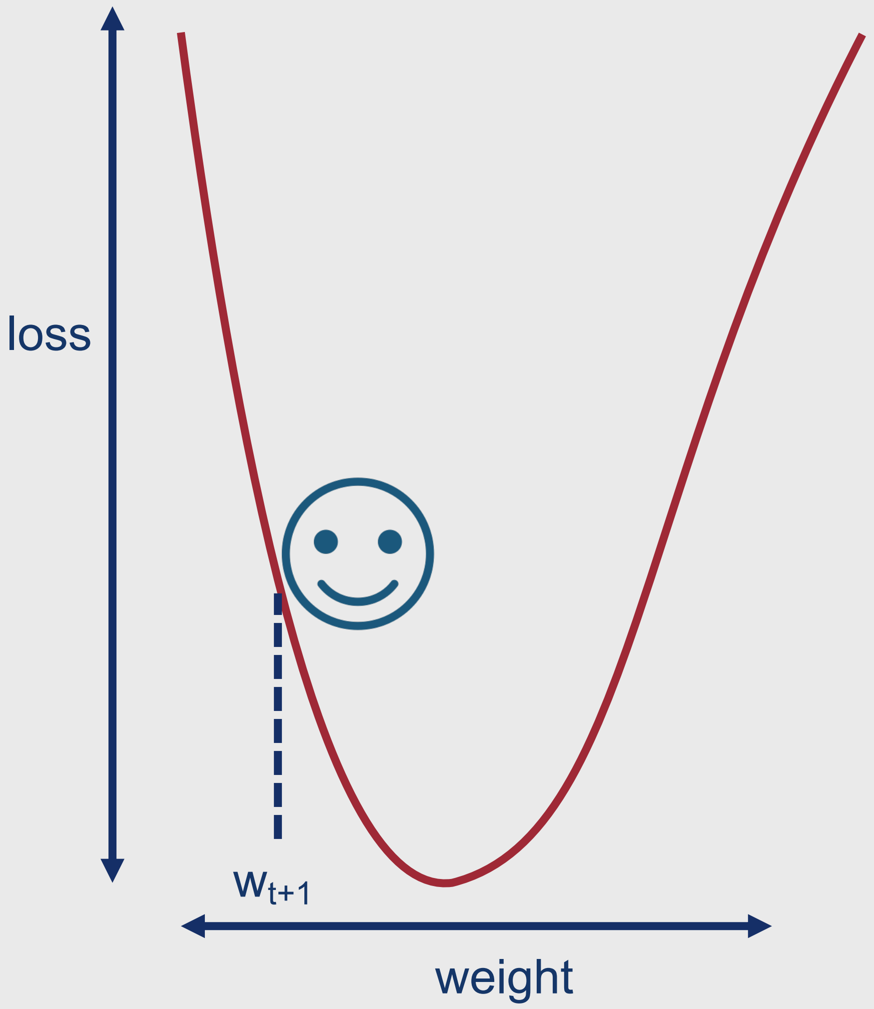
Gradient Descent

- Finds the minimum of a function by:
  - Figuring out the direction (in the space of θ) the function's slope
  - Moving in the opposite direction
- For logistic regression, loss functions are **convex**
  - Only one minimum
  - Gradient descent starting at any point is guaranteed to find it



Gradient Descent

- How much do we move?
  - Value of the slope
    - $\frac{d}{dw} f(x; w)$
  - Weighted by a learning rate η
- Faster learning rate → move *w* more on each step
- So, the change to a weight at time t is actually:
  - $w^{t+1} = w^t - \eta \frac{d}{dw} f(x; w)$



Remember, in actual logistic regression, there are weights for each feature.

- The gradient is then a vector of the slopes of each dimension:
  - $\nabla_\theta L(f(x;\theta),y) = \begin{bmatrix} \frac{d}{dw_1} L(f(x;\theta),y) \\ \cdots \\ \frac{d}{dw_n} L(f(x;\theta),y) \end{bmatrix}$
- This in turn means that the final equation for updating θ is:
  - $\theta_{t+1} = \theta_t - \eta \nabla L(f(x;\theta),y)$

The Gradient for Logistic Regression

- Recall our cross-entropy loss function:
  - $loss(y_i, \hat{y_i}) = -\sum_{c=1}^{|C|} y \log \hat{y} = -\sum_{c=1}^{|C|} y \log \sigma(\boldsymbol{w} \cdot \boldsymbol{x} + b)$
- The derivative for this function is:
  - $\frac{dL_{CE}(w,b)}{dw_j} = [\sigma(\boldsymbol{w} \cdot \boldsymbol{x} + b) - y]x_j$

Difference between true and estimated *y*

Corresponding input observation

Stochastic Gradient Descent Algorithm

```
θ←0  # initialize weights to 0
repeat until convergence:
      For each training instance (x^(i),y^(i)) in random order:
            # What is our gradient, given our current parameters?
            g ← ∇_θ L(f(x^(i); θ), y^(i))

            θ ← θ − ηg  # What are our updated parameters?
      return θ
```

Example: Gradient Descent (Single Step)

I'm just thrilled that I have five final exams on the same day. 🙄

Sarcastic

Feature	Weight	Value
Contains 🙄	0	1
Contains 😊	0	0
Contains "I'm"	0	1

Example: Gradient Descent (Single Step)

I'm just thrilled that I have five final exams on the same day. 🙄

Sarcastic

Feature	Weight	Value
Contains 🙄	0	1
Contains 😊	0	0
Contains "I'm"	0	1

Bias (b) = 0
Learning rate (η) = 0.1

$$\theta^{t+1} = \theta^{t} - \eta \nabla_{\theta} L\left(f\left(x^{(i)}; \theta\right), y^{(i)}\right)$$

Example: Gradient Descent (Single Step)

I'm just thrilled that I have five final exams on the same day. 🙄 ← Sarcastic

Feature	Weight	Value
Contains 🙄	0	1
Contains 😊	0	0
Contains "I'm"	0	1

Bias (b) = 0
Learning rate (η) = 0.1

$$\theta^{t+1} = \theta^{t} - \eta \nabla_{\theta} L\left(f\left(x^{(i)}; \theta\right), y^{(i)}\right)$$

$$\nabla_{\theta} L\left(f\left(x^{(i)}; \theta\right), y^{(i)}\right) = \begin{bmatrix} \frac{dL_{CE}(w,b)}{dw_1} \\ \frac{dL_{CE}(w,b)}{dw_2} \\ \frac{dL_{CE}(w,b)}{dw_3} \\ \frac{dL_{CE}(w,b)}{db} \end{bmatrix} = \begin{bmatrix} (\sigma(w \cdot x + b) - y)x_1 \\ (\sigma(w \cdot x + b) - y)x_2 \\ (\sigma(w \cdot x + b) - y)x_3 \\ \sigma(w \cdot x + b) - y \end{bmatrix} = \begin{bmatrix} (\sigma(0) - 1)x_1 \\ (\sigma(0) - 1)x_2 \\ (\sigma(0) - 1)x_3 \\ \sigma(0) - 1 \end{bmatrix} = \begin{bmatrix} (0.5 - 1)x_1 \\ (0.5 - 1)x_2 \\ (0.5 - 1)x_3 \\ (0.5 - 1) \end{bmatrix} = \begin{bmatrix} -0.5 * 1 \\ -0.5 * 0 \\ -0.5 * 1 \\ -0.5 \end{bmatrix} = \begin{bmatrix} -0.5 \\ 0 \\ -0.5 \\ -0.5 \end{bmatrix}$$

Example: Gradient Descent (Single Step)

I'm just thrilled that I have five final exams on the same day. 🙄

Sarcastic

Feature	Weight	Value
Contains 🙄	0	1
Contains 😊	0	0
Contains "I'm"	0	1

Bias (b) = 0
Learning rate (η) = 0.1

$$\theta^{t+1} = \theta^t - \eta \nabla_\theta L\left(f\left(x^{(i)};\theta\right), y^{(i)}\right)$$

$$\nabla_\theta L\left(f\left(x^{(i)};\theta\right), y^{(i)}\right) = \begin{bmatrix} -0.5 \\ 0 \\ -0.5 \\ -0.5 \end{bmatrix}$$

$$\theta^{t+1} = \theta^t - \eta \nabla_\theta L\left(f\left(x^{(i)};\theta\right), y^{(i)}\right) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} - \eta \begin{bmatrix} -0.5 \\ 0 \\ -0.5 \\ -0.5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} - 0.1 \begin{bmatrix} -0.5 \\ 0 \\ -0.5 \\ -0.5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} - \begin{bmatrix} -0.05 \\ 0 \\ -0.05 \\ -0.05 \end{bmatrix} = \begin{bmatrix} 0.05 \\ 0 \\ 0.05 \\ 0.05 \end{bmatrix}$$

Example: Gradient Descent (Single Step)

I'm just thrilled that I have five final exams on the same day. 🙄 ← Sarcastic

Feature	Weight	Value
Contains 🙄	0	1
Contains 😊	0	0
Contains "I'm"	0	1

Bias (b) = 0
Learning rate (η) = 0.1

$$\theta^{t+1} = \theta^{t} - \eta \nabla_{\theta} L\left(f\left(x^{(i)}; \theta\right), y^{(i)}\right)$$

$$\nabla_{\theta} L\left(f\left(x^{(i)}; \theta\right), y^{(i)}\right) = \begin{bmatrix} -0.5 \\ 0 \\ -0.5 \\ -0.5 \end{bmatrix}$$

$$\theta^{t+1} = \theta^{t} - \eta \nabla_{\theta} L\left(f\left(x^{(i)}; \theta\right), y^{(i)}\right) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} - \eta \begin{bmatrix} -0.5 \\ 0 \\ -0.5 \\ -0.5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} - 0.1 \begin{bmatrix} -0.5 \\ 0 \\ -0.5 \\ -0.5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} - \begin{bmatrix} -0.05 \\ 0 \\ -0.05 \\ -0.05 \end{bmatrix} = \begin{bmatrix} 0.05 \\ 0 \\ 0.05 \\ 0.05 \end{bmatrix}$$