EECS 245 Fall 2025

Math for ML

Lecture 18: Gradient Descent

→ Read: 4.2

Agenda

ch. 4.1

- Use gradients to minimize

$$R_{sq}(\vec{w}) = \frac{1}{n} \|\vec{y} - X\vec{w}\|^2$$

- Use gradients to minimize functions that we can't minimize by hand

→ gradient descent

ch. 4.2

Announcements

- HW 8 due tomorrow
- class suggestions for next semester are on Ed
- want to record a 30s video to help advertise the class? let me know!
- no live lecture on Tuesday: videos will be posted by 3PM on Tuesday

Gradient of a vector-to-scalar function

$f : \mathbb{R}^d \to \mathbb{R}$ is

$$\nabla f(\vec{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_d} \end{bmatrix}_{d \times 1}$$

$\nabla f(\vec{x})$ points in the direction of steepest ascent

$$\vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}$$

$$f(\vec{x}) = x_1^2 + x_1 \cos(x_2)$$

$$\nabla f(\vec{x}) = \begin{bmatrix} 2x_1 + \cos(x_2) \\ -x_1 \sin(x_2) \end{bmatrix}$$

"Big 3"

① $f(\vec{x}) = \vec{a}^T\vec{x} \Rightarrow \nabla f(\vec{x}) = \vec{a}$ dot product

② $f(\vec{x}) = \|\vec{x}\|^2 = \vec{x}^T\vec{x} \Rightarrow \nabla f(\vec{x}) = 2\vec{x}$ norm^2

③ $\underbrace{f(\vec{x}) = \vec{x}^T A\vec{x}}_{\text{quadratic form}} \Rightarrow \nabla f(\vec{x}) = (A+A^T)\vec{x}$

if $A = A^T$ (A symmetric)

$\Rightarrow \; = 2A\vec{x}$

$$R_{sq}(\vec{w}) = \frac{1}{n}\|\vec{y} - X\vec{w}\|^2$$

key: $\|\vec{v}\|^2 = \vec{v}^T\vec{v}$

$$= \frac{1}{n}(\vec{y} - X\vec{w})^T(\vec{y} - X\vec{w})$$

$$= \frac{1}{n}\left(\vec{y}^T - (X\vec{w})^T\right)(\vec{y} - X\vec{w})$$

QF!

$$= \frac{1}{n}\left(\vec{y}^T\vec{y} - \vec{y}^T(X\vec{w}) - (X\vec{w})^T\vec{y} + \overbrace{(X\vec{w})^T(X\vec{w})}^{\vec{w}^T X^T X\vec{w}}\right)$$

same! both dot product of $\vec{y}$ and $X\vec{w}$

$$= \frac{1}{n}\left(\vec{y}^T\vec{y} - 2\underbrace{\vec{y}^T X}_{=\left((\vec{y}^T X)^T\right)^T = (X^T\vec{y})^T}\vec{w} + \vec{w}^T X^T X\vec{w}\right)$$

$$R_{sq}(\vec{w}) = \frac{1}{n}\left(\vec{y}^T\vec{y} - \underbrace{2\left(X^T\vec{y}\right)^T\vec{w}}_{\text{rule ①}} + \underbrace{\vec{w}^T X^T X \vec{w}}_{\text{rule ③}} \right)$$

$$\nabla R_{sq}(\vec{w}) = \frac{1}{n}\left(0 - 2X^T\vec{y} + 2\underbrace{X^T X}_{\text{symmetric!}}\vec{w} \right)$$

$$= \frac{2}{n}\left(X^T X\vec{w} - X^T\vec{y} \right)$$

$$\nabla R_{sq}(\vec{w}) = \vec{0} \implies \frac{2}{n}\left(X^T X\vec{w} - X^T\vec{y}\right) = \vec{0}$$

$$\implies X^T X\vec{w} = X^T\vec{y}$$

Ch. 4.2

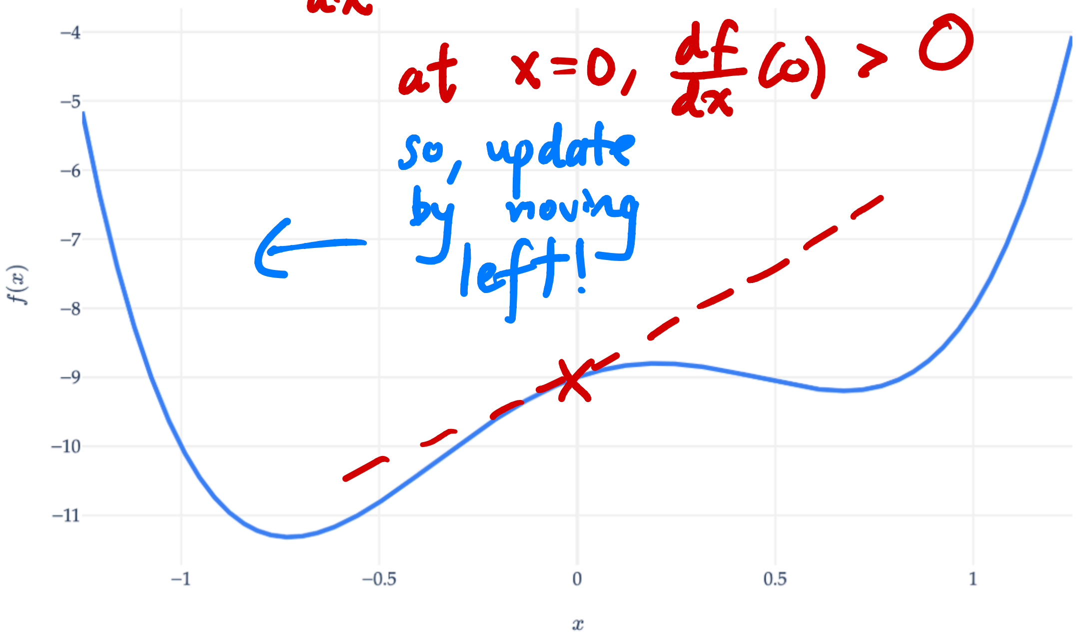
3:50

$$f(x) = 5x^4 - x^3 - 5x^2 + 2x - 9$$

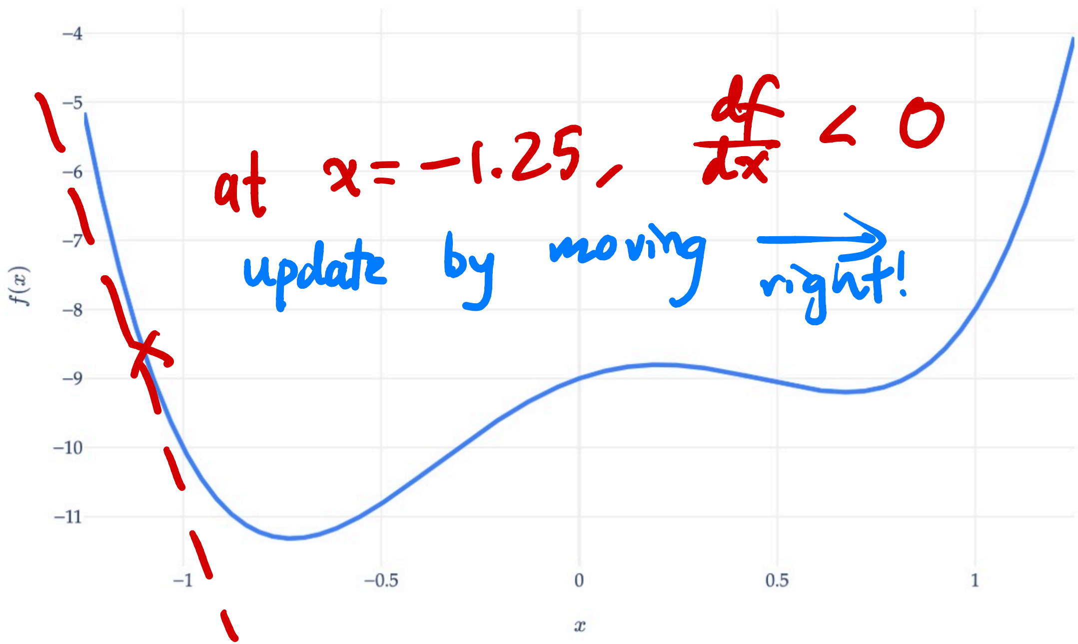
$$\frac{df}{dx} = 20x^3 - 3x^2 - 10x + 2$$

at $x=0$, $\frac{df}{dx}(0) > 0$

so, update by moving left!



$$f(x) = 5x^4 - x^3 - 5x^2 + 2x - 9$$



Gradient descent

to minimize (global) $f: \mathbb{R}^d \to \mathbb{R}$

- Choose an initial guess, $\vec{x}^{(0)}$
- Choose a learning rate/step size, $\alpha > 0$

- update our guesses using:

timestep t

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

timestep $t+1$

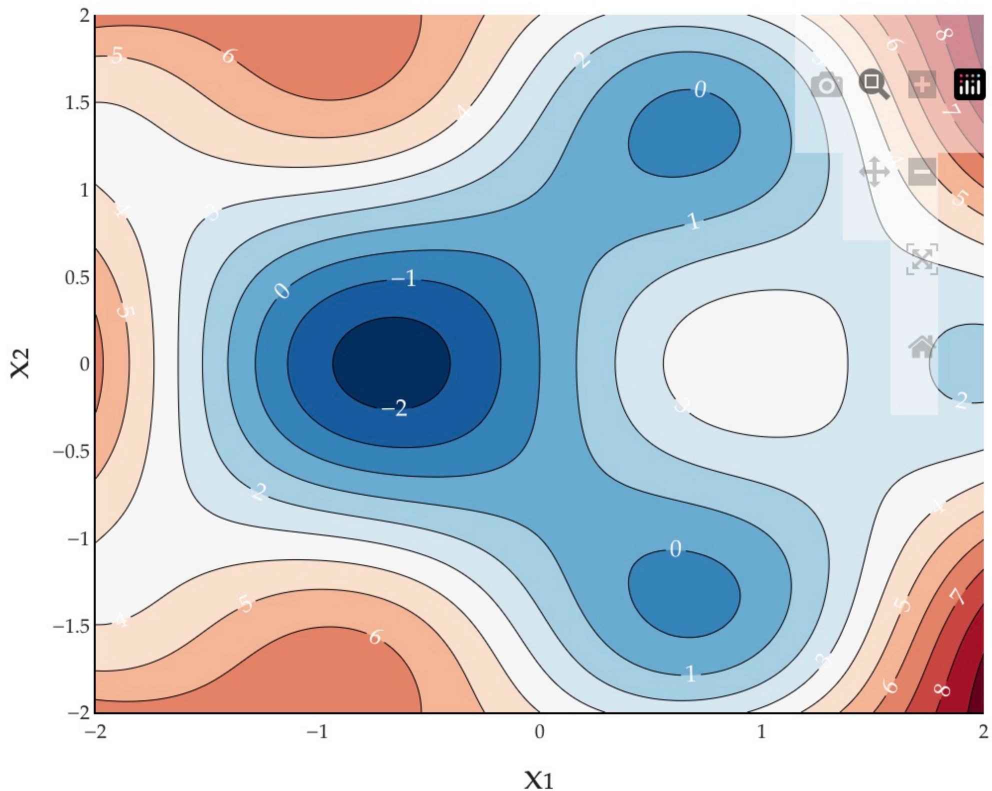
terminate when

$$\|\nabla f(\vec{x}^{(t)})\| < 0.0001$$

$$R_{sq}(\vec{w}) = \frac{1}{n} \|\vec{y} - X\vec{w}\|^2$$

$$= \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \vec{w} \cdot \text{Aug}(\vec{x}_i)\right)^2$$

$$R_{sq} : \mathbb{R}^{d+1} \to \mathbb{R}$$


X1
X2