Lectures 15-16

Gradient Descent and Convexity

DSC 40A, Fall 2024

Midterm topics do not include:

* center & spread (questions in practice site about mean absolute deviation)

* gradient descent

HW4 solution will be released on Sunday

Lingering questions

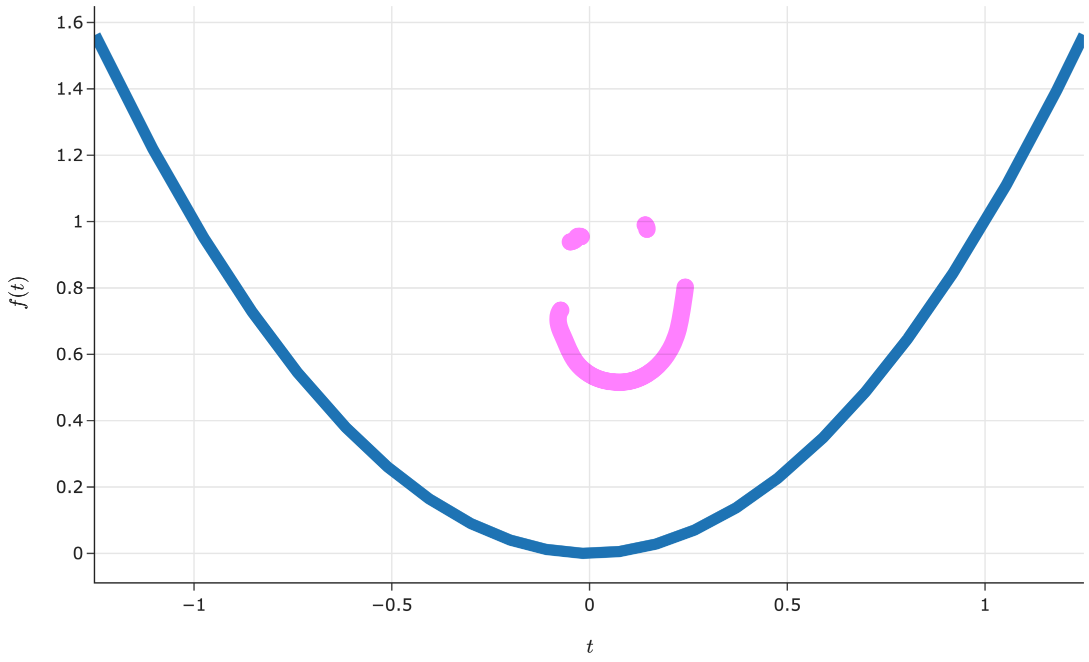
Now, we'll explore the following ideas:

- When is gradient descent *guaranteed* to converge to a global minimum?
    - What kinds of functions work well with gradient descent?
- How do I choose a step size?
- How do I use gradient descent to minimize functions of multiple variables, e.g.:

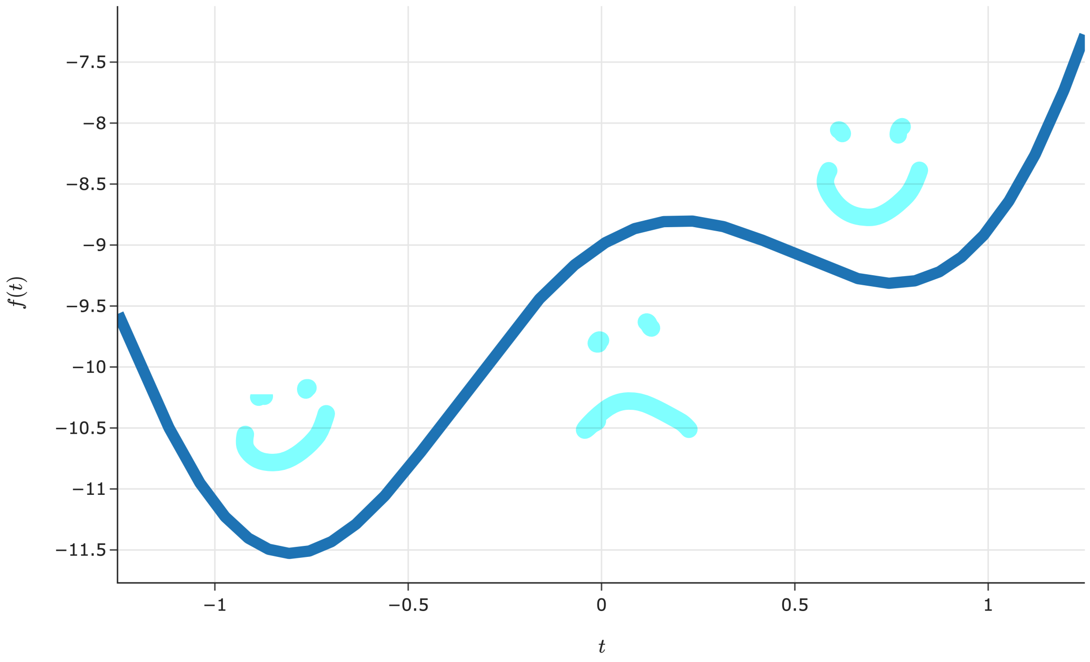
$$R_{\text{sq}}(w_0, w_1) = \frac{1}{n} \sum_{i=1}^{n} (y_i - (w_0 + w_1 x_i))^2$$

When is gradient descent guaranteed to work?

Convex functions



A **convex** function ✅



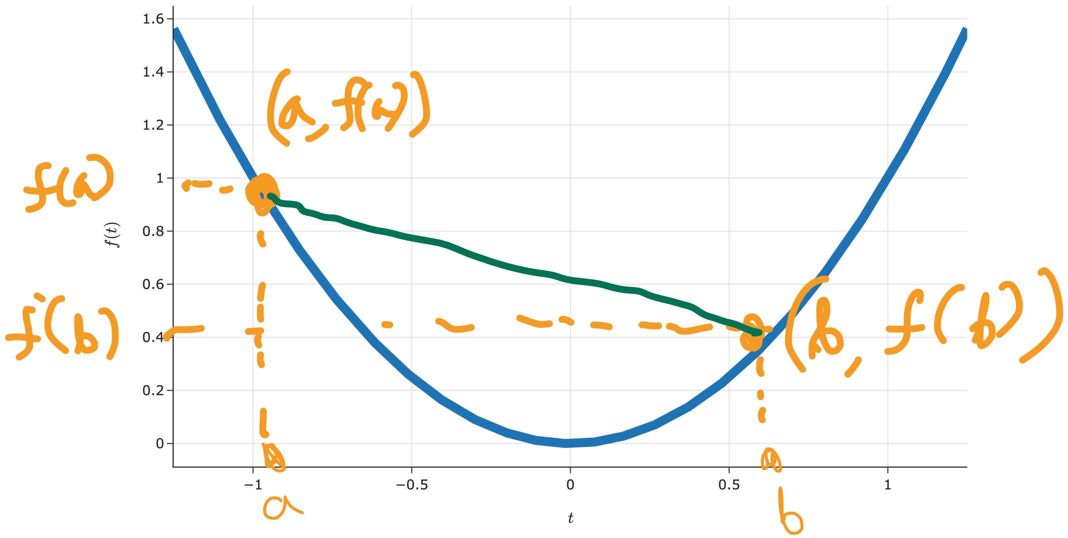
A **non-convex** function ❌

Convexity

- A function f is **convex** if, for **every** a, b in the domain of f, the line segment between:

$$(a, f(a)) \text{ and } (b, f(b))$$

does not go below the plot of f.



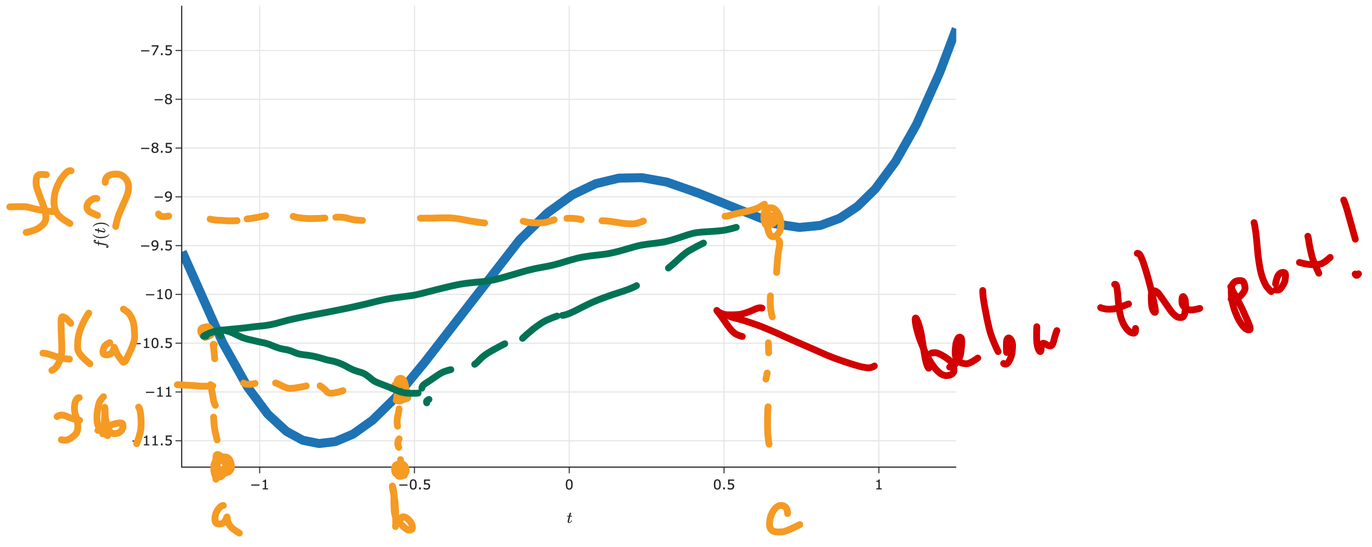
A **convex** function ✅

Convexity

- A function f is **convex** if, for **every** a, b in the domain of f, the line segment between:

$$(a, f(a)) \text{ and } (b, f(b))$$

does not go below the plot of f.



A **non-convex** function ✗

If $0 \le t \le 1$ what is $50 + 30t$

$(1-t)50 + 80t \quad 0 \le t \le 1$

Formal definition of convexity

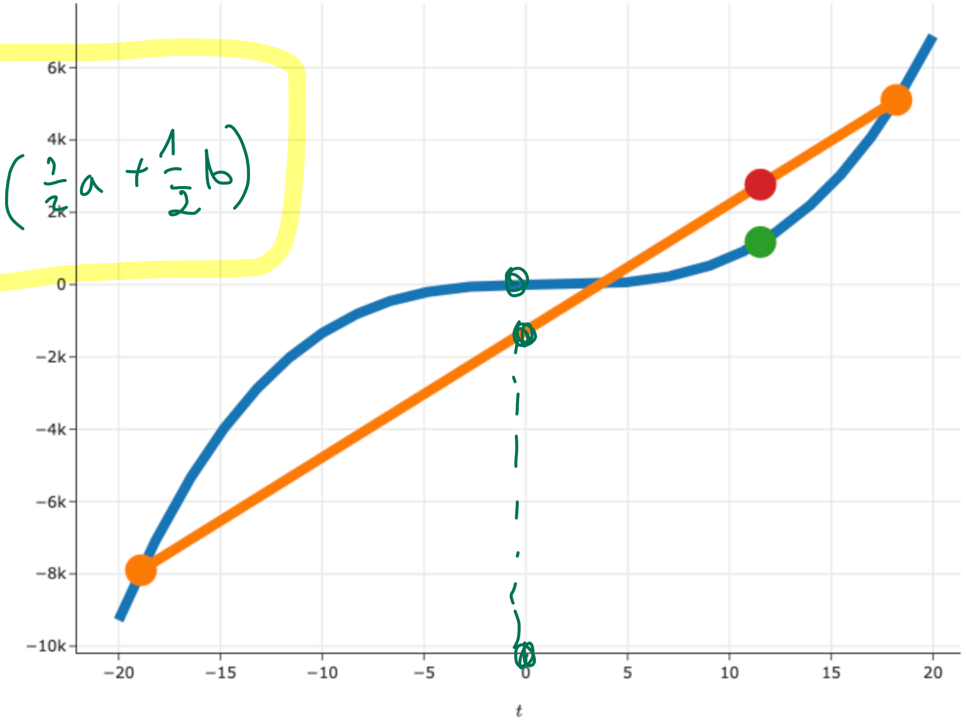
- A function $f : \mathbb{R} \to \mathbb{R}$ is **convex** if, for **every** a, b in the domain of f, and for every $t \in [0, 1]$:

plug in $t=0$: $f(a)$ plug in $t=1$: $f(b)$

Ex: $t = \frac{1}{2}$: $\frac{1}{2}f(a) + \frac{1}{2}f(b) \ge f(\frac{1}{2}a + \frac{1}{2}b)$

$$(1-t)f(a) + tf(b) \ge f((1-t)a + tb)$$

line between $f(a)$ and $f(b)$ function between $x=a$ and $x=b$



$\frac{1}{2}a + \frac{1}{2}b$

- A function is nonconvex if it is not convex.

- This is a formal way of restating the definition from the previous slide.

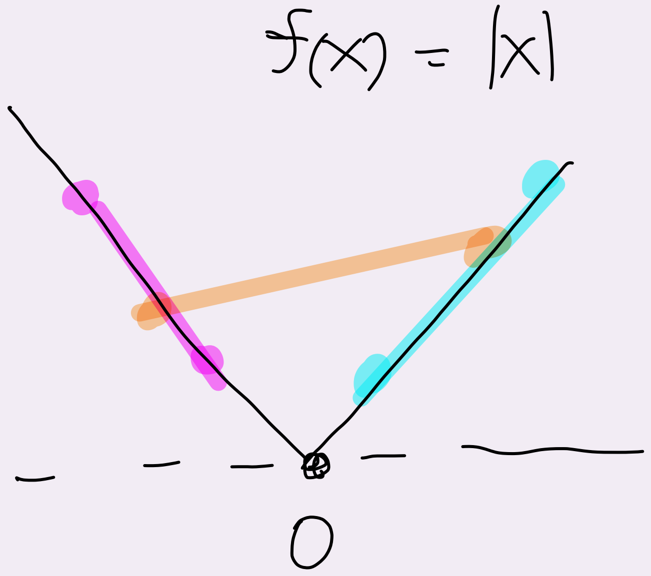
line $\ge$ function for $0 \le t \le 1$ $\iff$ convex

Question 🤔

Answer at q.dsc40a.com

Is $f(x) = |x|$ convex?

- A. Yes
- B. No
- C. Maybe



Example: Prove $f(x) = |x|$ is convex / nonconvex

Reminder: Traingle inequality: $|\alpha + \beta| \leq |\alpha| + |\beta|$

$(1-t)f(a) + t f(b) \geq f((1-t)a + t(b))$ for all $0 \leq t \leq 1$

$(1-t)|a| + t|b| \geq |(1-t)a + t(b)|$

line segment

$|(1-t)a + t(b)| \leq |(1-t)a| + |t b| \leq (1-t)|a| + t|b|$

function

alway non-negative for $0 \leq t \leq 1$

Question 🤔

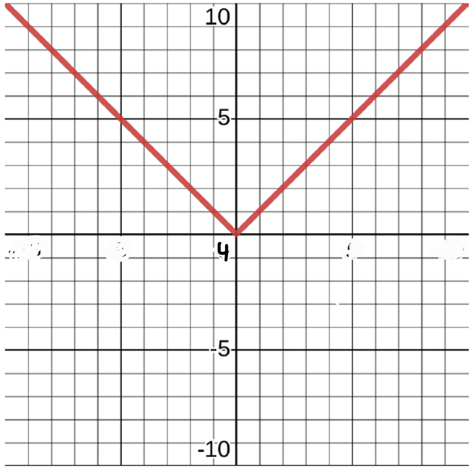
Answer at q.dsc40a.com

Which of these functions are **not** convex?

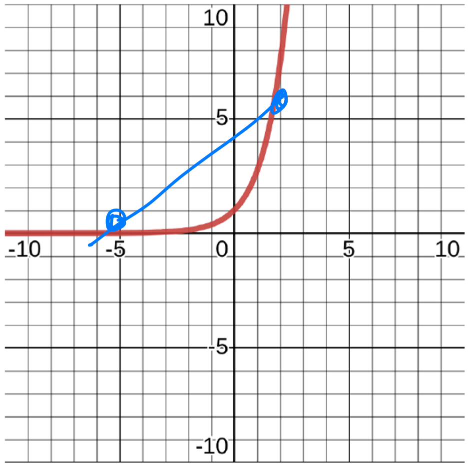
- A. $f(x) = |x - 4|$.
- B. $f(x) = e^x$.
- C. $f(x) = \sqrt{x - 1}$.
- D. $f(x) = (x - 3)^{24}$.
- E. More than one of the above are non-convex.

Convex vs. concave

convex



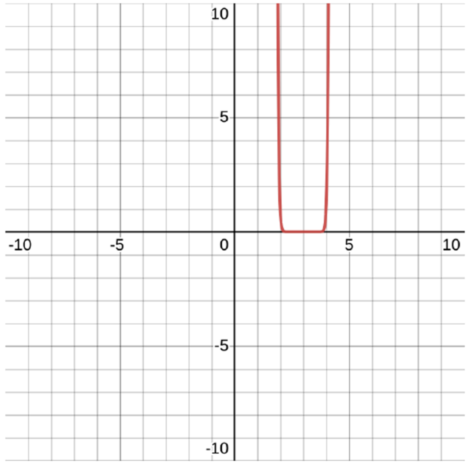
$f(x) = |x-4|$



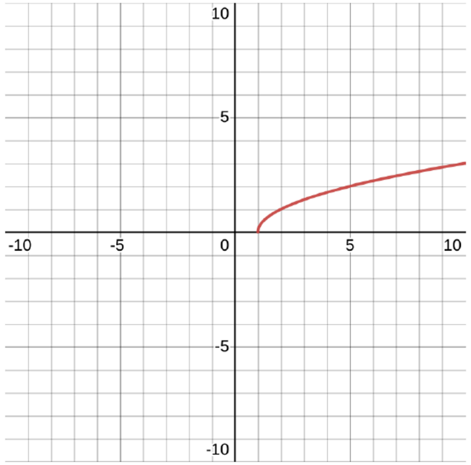
$f(x) = e^x$

convex

convex



$f(x) = (x-3)^{24}$

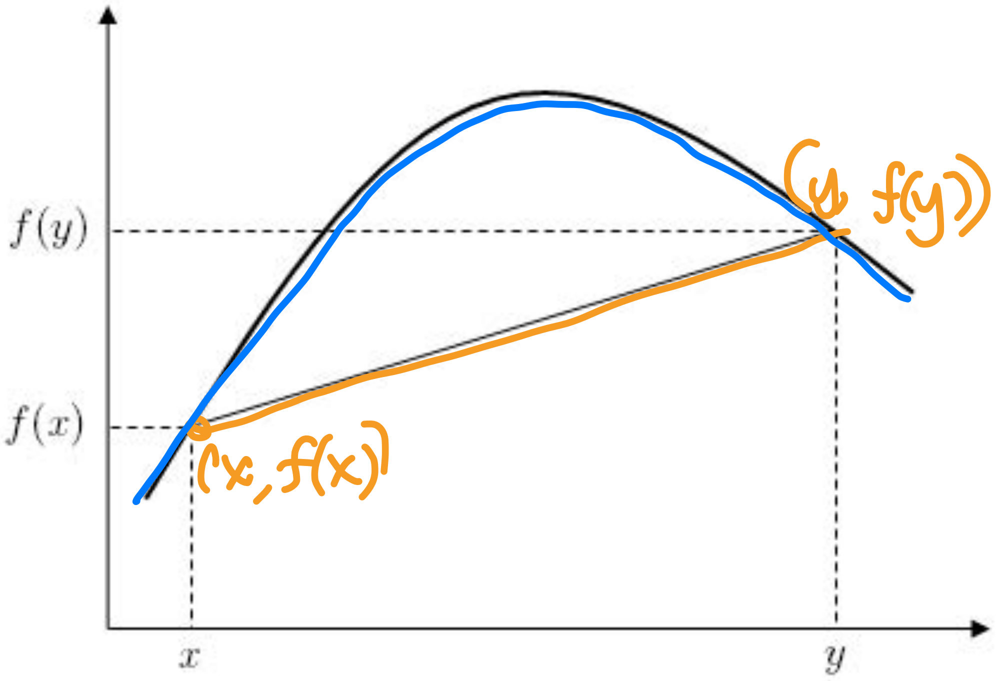


$f(x) = \sqrt{x-1}$

concav

Concave functions

- A **concave** function is the **negative** of a convex function.



Second derivative test for convexity

- If $f(t)$ is a function of a single variable and is **twice** differentiable, then $f(t)$ is
  - convex **if and only if**:

$\Leftrightarrow$

$$\frac{d^2 f}{dt^2}(t) \geq 0, \quad \forall t$$

→ for all t

  - concave **if and only if**:

$$\frac{d^2 f}{dt^2}(t) \leq 0, \quad \forall t$$

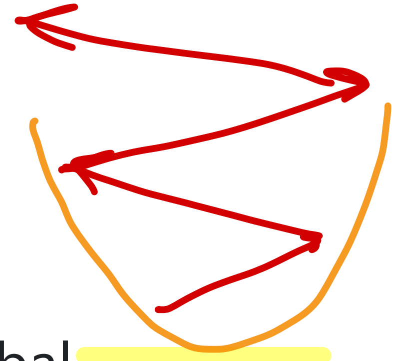
- Example: $f(x) = x^4$ is convex.

$f'(x) = 4x^3$

$f''(x) = 12x^2 \geq 0 \quad \forall x \Rightarrow$ convex

Why does convexity matter?

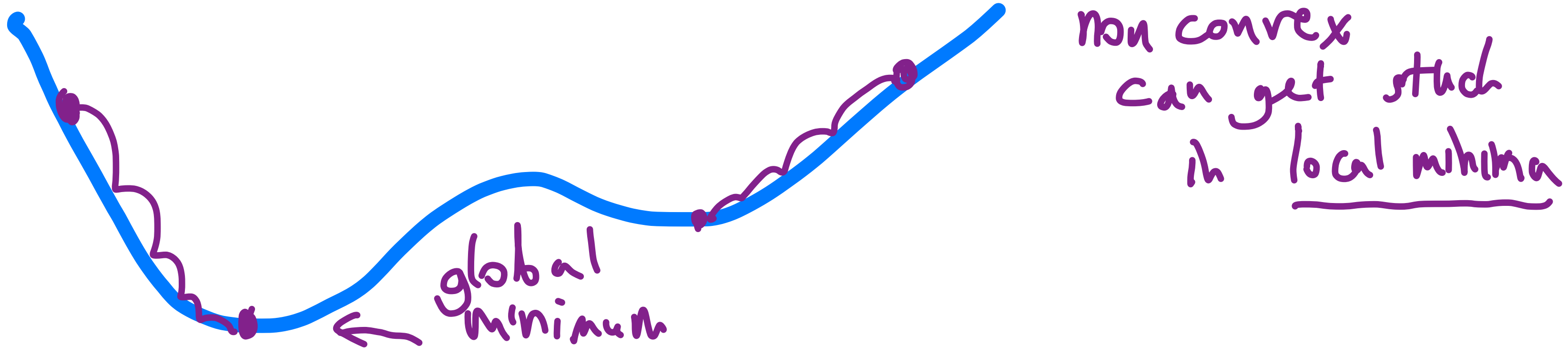
- Convex functions are (relatively) easy to minimize with gradient descent.
- **Theorem**: If $f(t)$ is convex and differentiable, then gradient descent converges to a **global minimum** of f, as long as the step size is small enough.



- **Why?**
  - Gradient descent converges when the derivative is 0.
  - For convex functions, the derivative is 0 only at one place – the global minimum.
  - In other words, if f is convex, gradient descent won't get "stuck" and terminate in places that aren't global minimums (local minimums, saddle points, etc.).

Nonconvex functions and gradient descent

- We say a function is **nonconvex** if it does not meet the criteria for convexity.
- Nonconvex functions are (relatively) difficult to minimize.
- Gradient descent **might** still work, but it's not guaranteed to find a global minimum.
    - We saw this at the start of the lecture, when trying to minimize $f(t) = 5t^4 - t^3 - 5t^2 + 2t - 9$.



Choosing a step size in practice

- In practice, choosing a step size involves a lot of trial-and-error.
- In this class, we've only touched on "constant" step sizes, i.e. where α is a constant.

$$t_{i+1} = t_i - \alpha \frac{df}{dt}(t_i)$$

- **Remember**: α is the "step size", but the amount that our guess for t changes is $\alpha \frac{df}{dt}(t_i)$, not just α.
- In future courses, you'll learn about "decaying" step sizes, where the value of α decreases as the number of iterations increases.
    - Intuition: take much bigger steps at the start, and smaller steps as you progress, as you're likely getting closer to the minimum.

More examples

Example: Huber loss and the constant model

- First, we learned about squared loss,
  $L_{\text{sq}}(y_i, H(x_i)) = (y_i - H(x_i))^2.$

  pro: differentiable, easy to minimize

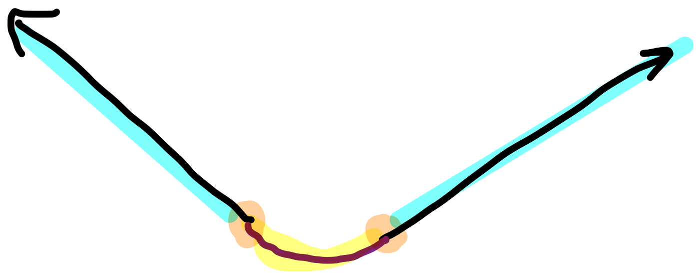
  con: sensitive to outliers

- Then, we learned about absolute loss,
  $L_{\text{abs}}(y_i, H(x_i)) = |y_i - H(x_i)|.$

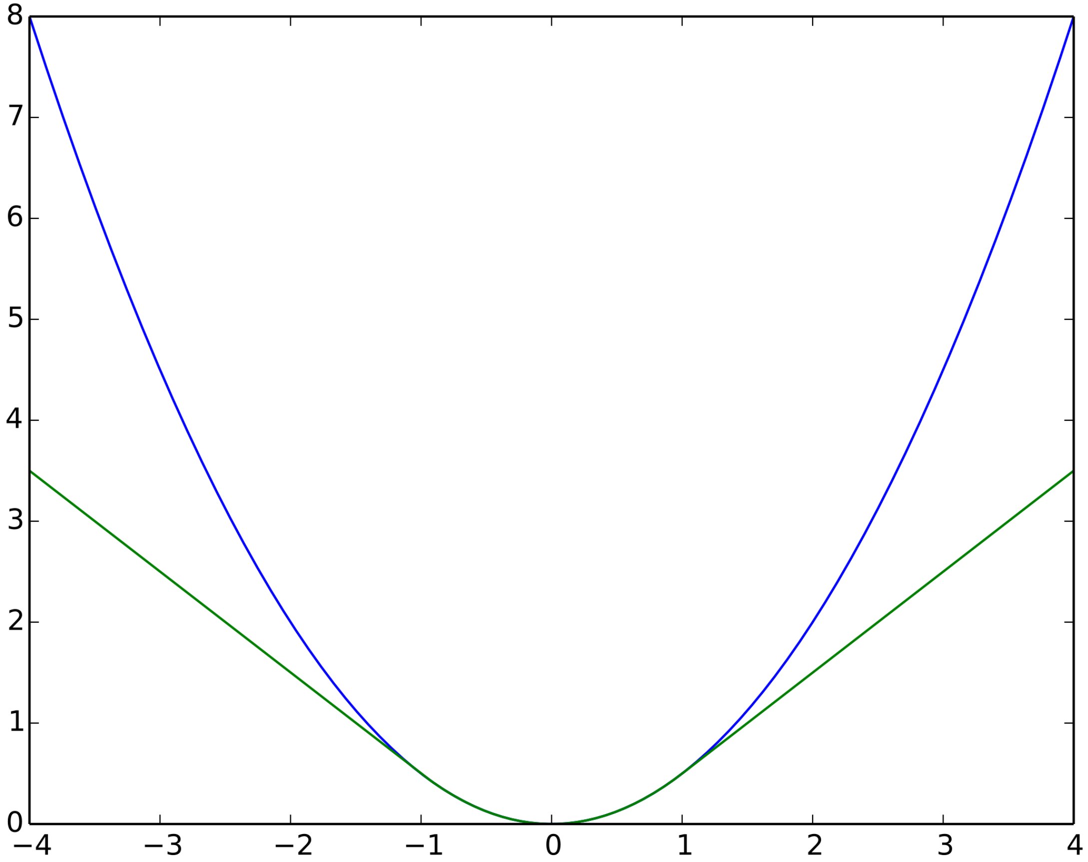
  pro: robust to outliers

  con: not differentiable, harder to minimize

- Let's look at a new loss function, **Huber loss**:



$$L_{\text{huber}}(y_i, H(x_i)) = \begin{cases} \frac{1}{2}(y_i - H(x_i))^2 & \text{if } |y_i - H(x_i)| \leq \delta \\ \delta \cdot \left(|y_i - H(x_i)| - \frac{1}{2}\delta\right) & \text{otherwise} \end{cases}$$



Squared loss in blue, **Huber** loss in green.

Note that both loss functions are convex!

Minimizing average Huber loss for the constant model

- For the constant model, $H(x) = h$:

$$L_{\text{huber}}(y_i, h) = \begin{cases} \frac{1}{2}(y_i - h)^2 & \text{if } |y_i - h| \leq \delta \\ \delta \cdot (|y_i - h| - \frac{1}{2}\delta) & \text{otherwise} \end{cases}$$

$$\implies \frac{\partial L}{\partial h}(h) = \begin{cases} -(y_i - h) & \text{if } |y_i - h| \leq \delta \\ -\delta \cdot \text{sign}(y_i - h) & \text{otherwise} \end{cases}$$

$y_i = h$

$\frac{\partial L}{\partial h}(h) = \begin{cases} 0 \\ -\delta \cdot 0 = 0 \end{cases}$

$\text{sign}(x) = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0 \end{cases}$

- So, the **derivative** of empirical risk is:

$$\frac{dR_{\text{huber}}}{dh}(h) = \frac{1}{n} \sum_{i=1}^{n} \begin{cases} -(y_i - h) & \text{if } |y_i - h| \leq \delta \\ -\delta \cdot \text{sign}(y_i - h) & \text{otherwise} \end{cases}$$

- It's **impossible** to set $\frac{dR_{\text{huber}}}{dh}(h) = 0$ and solve by hand: we need gradient descent!

Let's try this out in practice! Follow along in this notebook.

Minimizing functions of multiple variables

- Consider the function:

$$f(x_1, x_2) = (x_1 - 2)^2 + 2x_1 + (x_2 - 3)^2$$

- It has two **partial derivatives**: $\frac{\partial f}{\partial x_1}$ and $\frac{\partial f}{\partial x_2}$.

The gradient vector

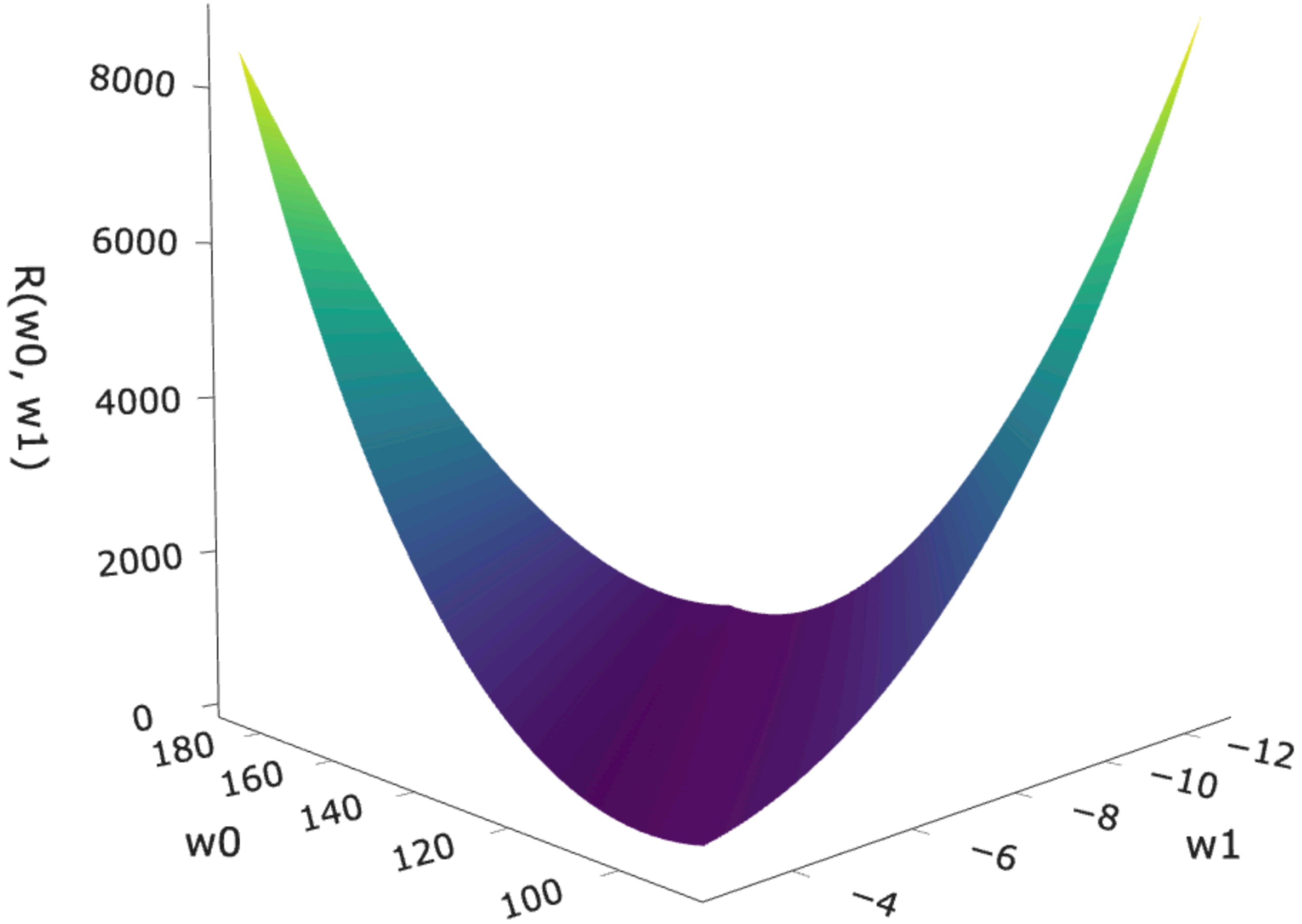
- If $f(\vec{x})$ is a function of multiple variables, then its **gradient**, $\nabla f(\vec{x})$, is a vector containing its partial derivatives.

- Example:

$$f(\vec{x}) = (x_1 - 2)^2 + 2x_1 + (x_2 - 3)^2$$
$$\nabla f(\vec{x}) = \begin{bmatrix} 2x_1 - 2 \\ 2x_2 - 6 \end{bmatrix}$$

- Example:

$$f(\vec{x}) = \vec{x}^T \vec{x}$$
$$\implies \nabla f(\vec{x}) =$$


8000
6000
4000
2000
0
R(w0, w1)
180
160
140
120
100
w0
−12
−10
−8
−6
−4
w1

Gradient descent for functions of multiple variables

- Example:

$$f(x_1, x_2) = (x_1 - 2)^2 + 2x_1 + (x_2 - 3)^2$$

$$\nabla f(\vec{x}) = \begin{bmatrix} 2x_1 - 2 \\ 2x_2 - 6 \end{bmatrix}$$

- The minimizer of f is a vector, $\vec{x}^* = \begin{bmatrix} x_1^* \\ x_2^* \end{bmatrix}$.
- We start with an initial guess, $\vec{x}^{(0)}$, and step size α, and update our guesses using:

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

Exercise

$$f(x_1, x_2) = (x_1 - 2)^2 + 2x_1 + (x_2 - 3)^2$$

$$\nabla f(\vec{x}) = \begin{bmatrix} 2x_1 - 2 \\ 2x_2 - 6 \end{bmatrix}$$

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

Given an initial guess of $\vec{x}^{(0)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and a step size of $\alpha = \frac{1}{3}$, perform **two** iterations of gradient descent. What is $\vec{x}^{(2)}$?

Example: Gradient descent for simple linear regression

- To find optimal model parameters for the model $H(x) = w_0 + w_1 x$ and squared loss, we minimized empirical risk:

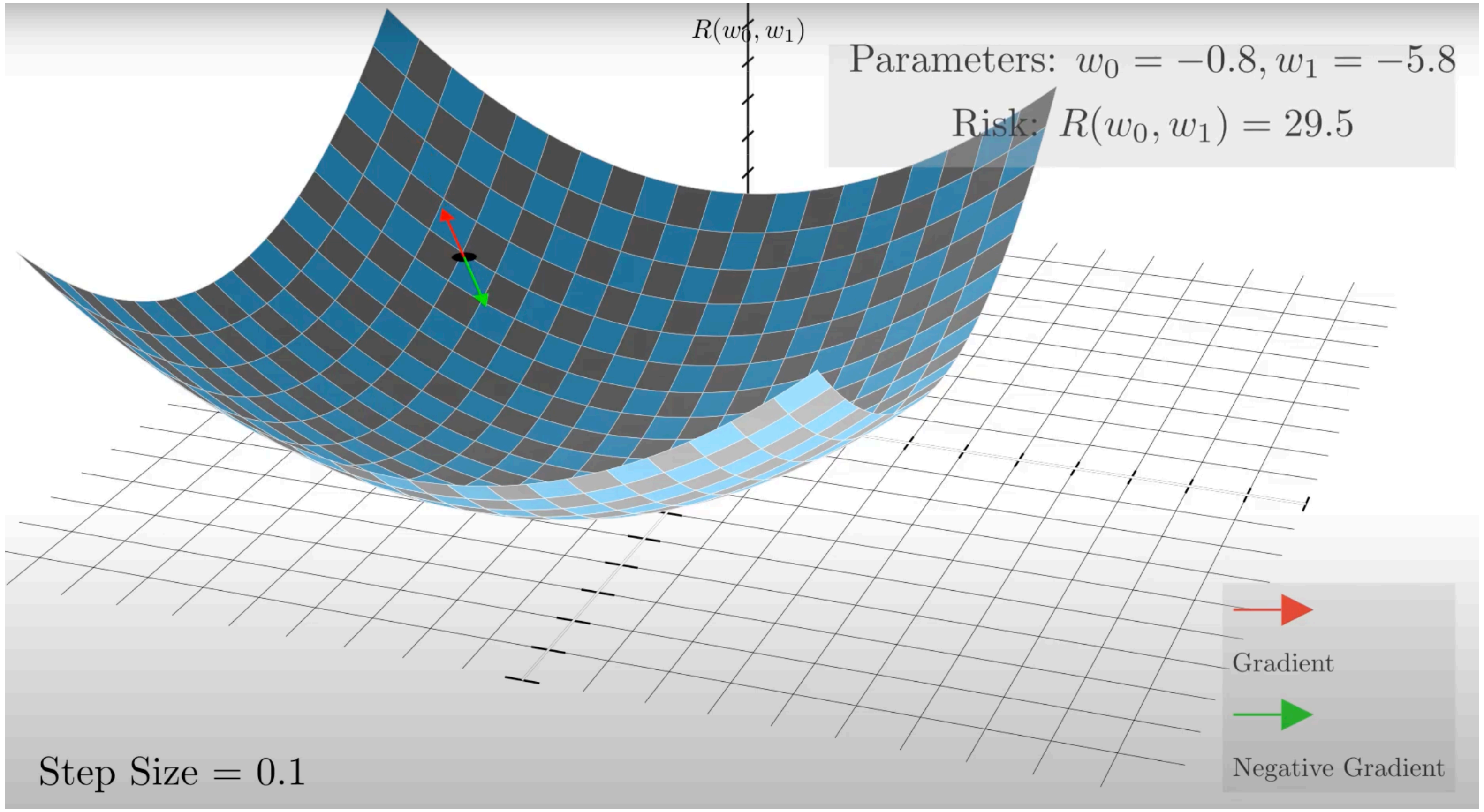
$$R_{\text{sq}}(w_0, w_1) = \frac{1}{n} \sum_{i=1}^{n} (y_i - (w_0 + w_1 x_i))^2$$

- This is a function of multiple variables, and is differentiable, so it has a gradient!

$$\nabla R(\vec{w}) = \begin{bmatrix} -\frac{2}{n} \sum_{i=1}^{n} (y_i - (w_0 + w_1 x_i)) \\ -\frac{2}{n} \sum_{i=1}^{n} (y_i - (w_0 + w_1 x_i)) x_i \end{bmatrix}$$

- **Key idea**: To find w_0^* and w_1^*, we *could* use gradient descent!

Gradient descent for simple linear regression, visualized



Let's watch 🎥 this animation that Jack made.

What's next?

- In Homework 5, you'll see a few questions involving today's material.
- After the midterm, we'll start talking about probability.