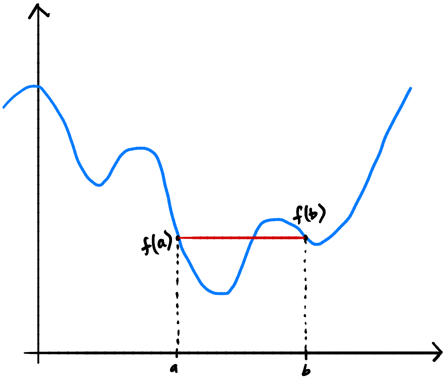


DSC 40A

Lecture 05

Learning via Optimization, pt V

Last Week: Empirical Risk Minimization

- To learn, pick a **loss function** L and minimize the **empirical risk**:

$$R(h) = \frac{1}{n}\sum_{i=1}^{n} L(h, y_i)$$

- Absolute loss: $L_{\text{abs}}(h, y) = |h - y|$ (gives the **median**)
- Square loss: $L_{\text{sq}}(h, y) = (h - y)^2$ (gives the **mean**)

Last Week: The UCSD Loss

- We defined the "UCSD Loss":

$$L_{\text{ucsd}}(h, y) = 1 - e^{-(h-y)^2/\sigma^2}$$

- Goal: minimize the "UCSD Risk",

$$R_{\text{ucsd}}(h, y) = \frac{1}{n} \sum_{i=1}^{n} \left[1 - e^{-(h-y_i)^2/\sigma^2}\right]$$

- We tried taking a derivative and solving, but we couldn't solve for h.

Last Week: Gradient Descent

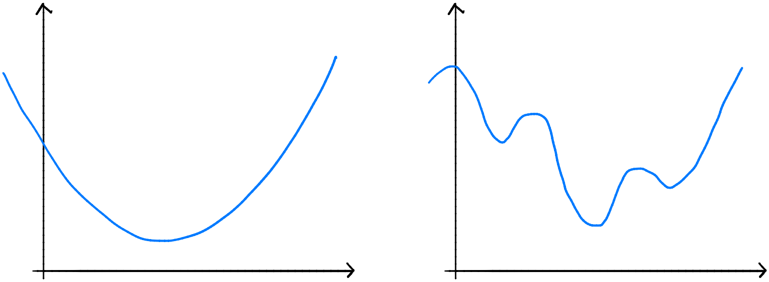
- Pick α to be a positive number. It is the **learning rate**.
- Pick a starting prediction, h_0.
- On step i, perform update $h_i = h_{i-1} - \alpha \cdot \frac{dR}{dh}(h_{i-1})$
- Repeat until convergence (when h doesn't change much).

Demo notebook on DataHub

Today

When is gradient descent guaranteed to work?

Convex Functions



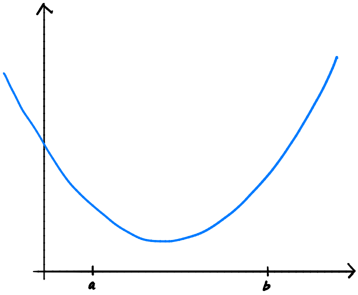
Convex Non-convex

Convexity: Definition

- f is **convex** if for **every** a, b the line segment between

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

does not go below the plot of f.

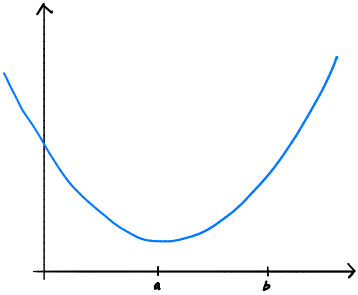


Convexity: Definition

- f is **convex** if for **every** a, b the line segment between

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

does not go below the plot of f.

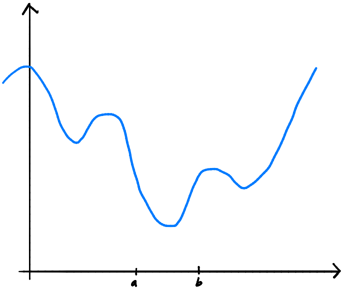


Convexity: Definition

- f is **convex** if for **every** a, b the line segment between

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

does not go below the plot of f.

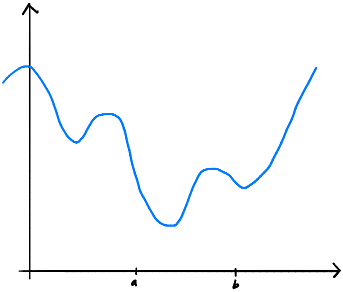


Convexity: Definition

- f is **convex** if for **every** a, b the line segment between

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

does not go below the plot of f.



Deriving a More Useful/Formal Definition

- Walk from a at time $t = 0$ to b at time $t = 1$.
- Let $\text{height}_f(t)$ be height of f at time t.
- Let $\text{height}_{\text{line}}(t)$ be height of line segment at time t.
- If f is convex, then for every $t \in [0, 1]$:

$$\text{height}_{\text{line}}(t) \geq \text{height}_f(t)$$

Position at time t

- Let $x(t)$ be horizontal position at time t.
- At time $t = 0$, we're at a, so $x(0) = a$.
- At time $t = 1$, we're at b, so $x(1) = b$.
- This formula works:

$$x(t) =$$

$$=$$

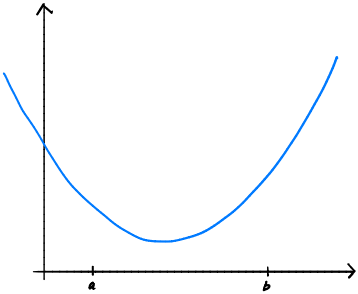
Height of f at time t

- We want a formula for $\text{height}_f(t)$

- Remember $x(t) = (1 - t)a + bt$. So:

$$\text{height}_f(t) =$$

$$=$$



Height of line segment at time t

- We want a formula for $\text{height}_{\text{line}}(t)$
- It is a linear function: $\text{height}_{\text{line}}(t) = w_1 t + w_0$
- We know $\text{height}_{\text{line}}(0) = f(a)$ and $\text{height}_{\text{line}}(1) = f(b)$.

Height of line segment at time t

- We want a formula for $\text{height}_{\text{line}}(t)$
- It is a linear function: $\text{height}_{\text{line}}(t) = w_1 t + w_0$
- We know $\text{height}_{\text{line}}(0) = f(a)$ and $\text{height}_{\text{line}}(1) = f(b)$.

Discussion Question

What is the formula for $\text{height}_{\text{line}}(t)$?

a) $at + (1 - b)t$
b) $(1 - t)f(a) + tf(b)$
c) $(a \cdot f(t) + b \cdot f(t))/2$
d) $t[f(b) - f(a)]$

Height of line segment at time t

$\text{height}_{\text{line}}(t) = w_1 t + w_0$

$\text{height}_{\text{line}}(0) = f(a)$ $\qquad$ $\text{height}_{\text{line}}(1) = f(b)$

Convexity: Formal Definition

$$\text{height}_{\text{line}}(t) \geq \text{height}_f(t)$$

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

Convexity: Formal Definition

$$\text{height}_{\text{line}}(t) \geq \text{height}_f(t)$$

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

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

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

Convexity: Formal Definition

$$\text{height}_{\text{line}}(t) \geq \text{height}_f(t)$$
$$(1-t)f(a) + tf(b) \geq f((1-t)a + tb)$$

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

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

- A function f is **nonconvex** if it is not convex.

Discussion Question

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

a) Yes.
b) No.
c) Maybe.

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

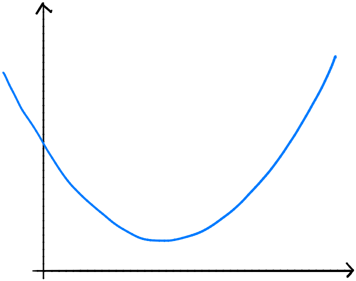
Hint: remember triangle inequality, $|\alpha + \beta| \leq |\alpha| + |\beta|$.

Proving Convexity: Second Derivative Test

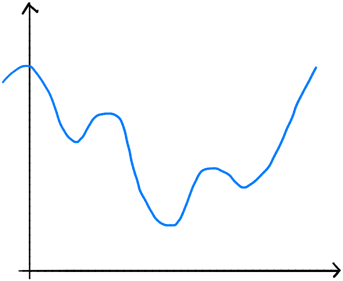
- If $\frac{d^2 f}{dx^2}(x) \geq 0$ for all x, then f is convex.

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

- Only works if f is twice differentiable!



Convex



Non-convex

Proving Convexity: Using Properties

Suppose that $f(x)$ and $g(x)$ are convex. Then:

- $w_1 f(x) + w_2 g(x)$ is convex, provided $w_1, w_2 \geq 0$
  - Example: $3x^2 + |x|$ is convex
- $g(f(x))$ is convex, provided g is non-decreasing.
  - Example: e^{x^2} is convex
- $\max\{f(x), g(x)\}$ is convex
  - Example: $\begin{cases} 0, & \mathrm{x} < 0 \\ x, & \mathrm{x} \geq 0 \end{cases}$ is convex (max of 0 and x)

Convex Losses

- If $L(h, y)$ is a convex function (when y is fixed) then

$$R(h) = \frac{1}{n} \sum_{i=1}^{n} L(h, y_i)$$

  is convex.

- Proof: sums of convex functions are convex.

Convexity and Gradient Descent

- Convex functions are (relatively) easy to optimize.
- **Theorem**: if $R(h)$ is convex and differentiable[1] then gradient descent converges to a **global optimum** of R *provided* that the step size is small enough[2].

---

[1] and it's derivative is not too wild
[2] step size related to steepness.

Convexity and Gradient Descent

- Convex functions are (relatively) easy to optimize.
- **Theorem**: if $R(h)$ is convex and differentiable[1] then gradient descent converges to a **global optimum** of R *provided* that the step size is small enough[2].
- We can even modify GD to work with convex, non-differentiable functions.

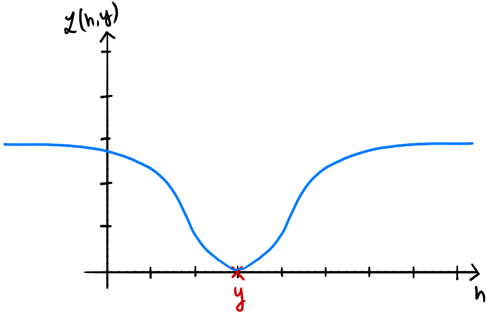
[1] and it's derivative is not too wild
[2] step size related to steepness.

Nonconvexity and Gradient Descent

- Nonconvex functions are (relatively) hard to optimize.
- Gradient descent can still be useful.
- But not guaranteed to converge to a global minimum.

Convexity of Losses

- Is $L_{\text{sq}}(h, y) = (h - y)^2$ convex? Yes or No.
- Is $L_{\text{abs}}(h, y) = |h - y|$ convex? Yes or No.
- Is $L_{\text{ucsd}}(h, y)$ convex? Yes or No.



Convexity of UCSD Risk

- A function can be convex in a region.
- If σ is large, $R_{\text{ucsd}}(h)$ is convex in a big region around data.
- If σ is small, $R_{\text{ucsd}}(h)$ is convex in only small regions.

Status Update

- We learned what it means for a function to be **convex**.
- Convex functions are (relatively) **easy** to optimize with gradient descent.
- We like **convex loss functions**, like the square loss and absolute loss.

What’s Left?

- We’ve been predicting salary without using any information about the individual.
- Making predictions using some information.