# DSC 140A

## Probabilistic Modeling & Machine Learning

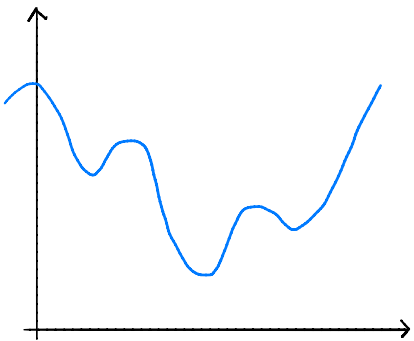
Lecture 5 | Part 1

## Convexity

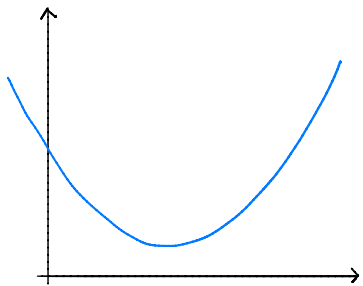
# Question

- When is gradient descent guaranteed to work?

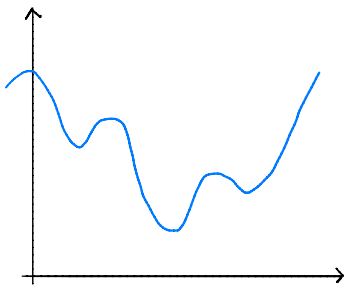
# Not here...



# Convex Functions



Convex



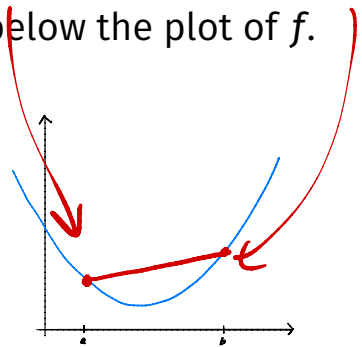
Non-convex

# 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 $f$.

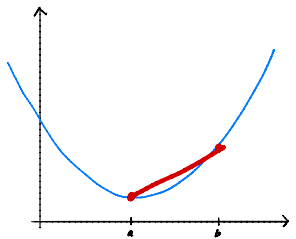


# Convexity: Definition

- $f$ 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 $f$.

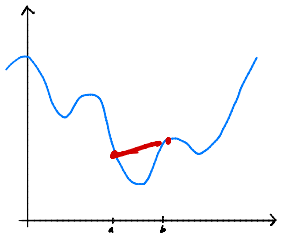


# Convexity: Definition

- $f$ 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 $f$.

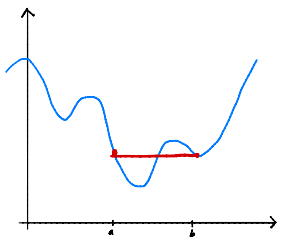


# Convexity: Definition

- $f$ 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 $f$.



# Other Terms

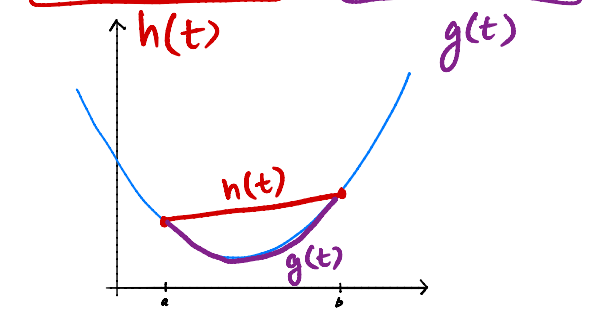
- If a function is not convex, it is **non-convex**.
- **Strictly convex**: the line lies strictly above curve.
- **Concave**: the line lines on or below curve.



# Convexity: Formal Definition

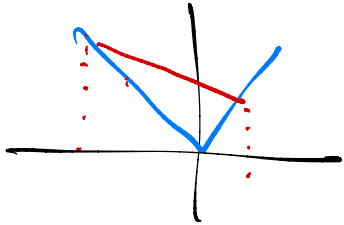
- A function $f : \mathbb{R} \to \mathbb{R}$ is **convex** if for every choice of $a, b \in \mathbb{R}$ and $t \in [0, 1]$:

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



# Example

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



Fact 1: $|u| + |v| \geq |u+v|$ (triangle ineq.)

Fact 2: $|uv| = |u||v|$

# Example

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

$(1-t)f(a) + tf(b) = (1-t)|a| + t|b|$ → since $t \in [0,1]$

$= |1-t| \cdot |a| + |t||b|$ → by Fact #2.

$= |(1-t)a| + |tb|$ — by the triangle ineq.

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

$= f((1-t)a + tb)$

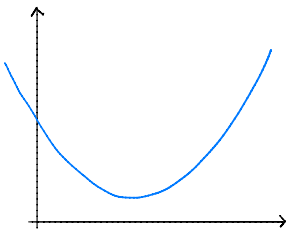
# Another View: Second Derivatives

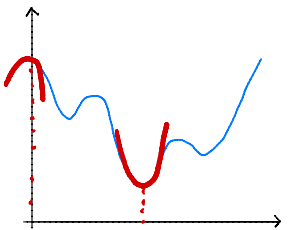
- If $\frac{d^2 f}{dx^2}(x) \geq 0$ for all $x$, then $f$ is convex.
- Example: $f(x) = x^4$ is convex.

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

$f''(x) = 12x^2$

- **Warning!** Only works if $f$ is twice differentiable!



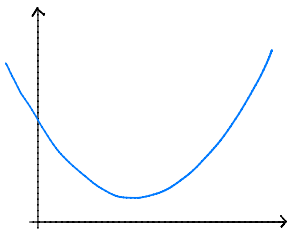


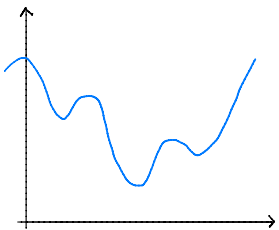
# Another View: Second Derivatives

- "Best" straight line at $x_0$:
  - $h_1(z) = f'(x_0) \cdot z + b$
- "Best" parabola at $x_0$:
  - At $x_0$, $f$ looks likes $h_2(z) = \frac{1}{2} f''(x_0) \cdot z^2 + f'(x_0)z + c$
  - Possibilities: upward-facing, downward-facing.

# Convexity and Parabolas

- Convex if for **every** $x_0$, parabola is upward-facing.
  - That is, $f''(x_0) \geq 0$.





# Proving Convexity Using Properties

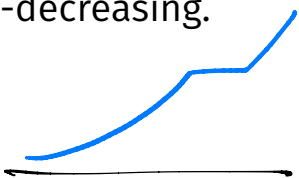
Suppose that $f(x)$ and $g(x)$ are convex. Then:

- $w_1 f(x) + w_2 g(x)$ is convex, provided $w_1, w_2 \geq 0$
  - Example: $3x^2 + |x|$ is convex

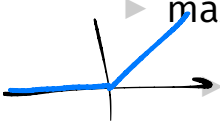
$\sum f_i(x)$

- $g(f(x))$ is convex, provided $g$ is non-decreasing.
  - Example: $e^{x^2}$ is convex

$g(z) = e^z$ $f(x) = x^2$



- $\max\{f(x), g(x)\}$ is convex



  - Example: $\begin{cases} 0, & x < 0 \\ x, & x \geq 0 \end{cases}$ is convex

# Convexity and Gradient Descent

- Convex functions are (relatively) easy to optimize.
- **Theorem**: if $f(x)$ is convex and "not too steep"[1] then (stochastic) (sub)gradient descent converges to a **global optimum** of $f$ *provided* that the step size is small enough[2].

[1] Technically, $c$-Lipschitz

[2] step size related to steepness, should decrease like $1/\sqrt{t}$, where $t$ is step number

# 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.

# DSC 140A

## Probabilistic Modeling & Machine Learning

Lecture 5 | Part 2

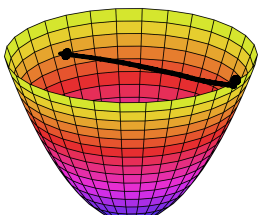
## Convexity in Many Dimensions

# Convexity: Definition

- $f(\vec{x})$ is **convex** if for **every** $\vec{a}, \vec{b}$ the line segment between

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

does not go below the plot of $f$.



# Convexity: Formal Definition

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

$$(1 - t)f(\vec{a}) + tf(\vec{b}) \geq f((1 - t)\vec{a} + t\vec{b}).$$

# The Second Derivative Test

- For 1-d functions, convex if second derivative $\geq 0$.
- For 2-d functions, convex if ???

$a + bx + cx^2$

$a + bx + cy + dx^2 + ey^2 + fxy$

# Second Derivatives in 2-d

- In 2-d, there are 4 second derivatives of $f(\vec{x})$:
  - $\frac{\partial f^2}{\partial x_1^2}, \frac{\partial f^2}{\partial x_2^2}, \frac{\partial f^2}{\partial x_1 x_2}, \frac{\partial f^2}{\partial x_2 x_1}$

# Convexity in 2-d

- "Best" quadratic function approximating $f$ at $\vec{x}$:

$$h_2(z_1, z_2) = az_1^2 + bz_2^2 + cz_1z_2 + \ldots$$
$$= \frac{1}{2}\frac{\partial f^2}{\partial x_1^2}(\vec{x}) \cdot z_1 + \frac{1}{2}\frac{\partial f^2}{\partial x_2^2}(\vec{x}) \cdot z_2 + \frac{\partial f^2}{\partial x_1 x_2}(\vec{x}) \cdot z_1 z_2 + \ldots$$

- $a, b, c$ determine rough shape. Possibilities:
  - Upward-facing bowl.
  - Downward-facing bowl.
  - "Saddle"

# Convexity in 2-d

- Convex if at any $\vec{x}$, for any $z_1, z_2$:

$$\frac{1}{2}\frac{\partial f^2}{\partial x_1^2}(\vec{x}) \cdot z_1 + \frac{1}{2}\frac{\partial f^2}{\partial x_2^2}(\vec{x}) \cdot z_2 + \frac{\partial f^2}{\partial x_1 x_2}(\vec{x}) \cdot z_1 z_2 \geq 0$$

$f(x_1, x_2)$

# The Hessian Matrix

- Create the **Hessian** matrix of second derivatives:

$$H(\vec{x}) = \begin{pmatrix} \frac{\partial f^2}{\partial x_1^2}(\vec{x}) & \frac{\partial f^2}{\partial x_1 x_2}(\vec{x}) \\ \frac{\partial f^2}{\partial x_2 x_1}(\vec{x}) & \frac{\partial f^2}{\partial x_2^2}(\vec{x}) \end{pmatrix}$$

# In General

- If $f : \mathbb{R}^d \to \mathbb{R}$, the **Hessian** at $\vec{x}$ is:

$$H(\vec{x}) = \begin{pmatrix} \frac{\partial f^2}{\partial x_1^2}(\vec{x}) & \frac{\partial f^2}{\partial x_1 x_2}(\vec{x}) & \cdots & \frac{\partial f^2}{\partial x_1 x_d}(\vec{x}) \\ \frac{\partial f^2}{\partial x_2 x_1}(\vec{x}) & \frac{\partial f^2}{\partial x_2^2}(\vec{x}) & \cdots & \frac{\partial f^2}{\partial x_2 x_d}(\vec{x}) \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial f^2}{\partial x_d x_1}(\vec{x}) & \frac{\partial f^2}{\partial x_d^2}(\vec{x}) & \cdots & \frac{\partial f^2}{\partial x_d^2}(\vec{x}) \end{pmatrix}$$

# Observations

- $H$ is square.
- $H$ is symmetric.

# Convexity in 2-d

- Convex if at any $\vec{x}$, for any $z_1, z_2$:

$$\frac{1}{2}\frac{\partial f^2}{\partial x_1^2}(\vec{x}) \cdot z_1 + \frac{1}{2}\frac{\partial f^2}{\partial x_2^2}(\vec{x}) \cdot z_2 + \frac{\partial f^2}{\partial x_1 x_2}(\vec{x}) \cdot z_1 z_2 \geq 0$$

- Equivalently, convex if for any $\vec{x}$ and any $\vec{z}$:

$$\vec{z}^T H(\vec{x})\, \vec{z} \geq 0$$

# Positive Semi-Definite

- A square, $d \times d$ symmetric matrix $X$ is **positive semi-definite** (PSD) if for any $\vec{u}$:

$$\vec{u}^T X \vec{u} \geq 0$$

# The Second Derivative Test

- A function $f : \mathbb{R}^d \to \mathbb{R}$ is **convex** if for any $\vec{x} \in \mathbb{R}^d$, the Hessian matrix $H(\vec{x})$ is positive semi-definite.

# But wait...

- How can we tell if a matrix is positive semi-definite?

PSD if all of its eigenvalues are $\geq 0$

# Example

$$M = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

det(m) = prod of its eigenvalues

$$\begin{aligned} \det(m) &= (1)(1) - (1)(1) \\ &= 1 - 1 \\ &= 0 \end{aligned}$$

# Example

$$M = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$$

Not PSD

$$\det(M) = (1)(1) - (2)(2)$$
$$= 1 - 4$$
$$= -3$$

# Example

$\frac{d^2}{dx\,dy}$

Is $f(x, y) = x^2 + 4xy + y^2$ convex?

No

$\frac{df}{dx} = 2x + 4y$

$\frac{df}{dy} = 4x + 2y$

1) Hessian:

$$\begin{pmatrix} 2 & 4 \\ 4 & 2 \end{pmatrix}$$

2) det = −12, not convex

# Proving Convexity with Properties

Suppose that $f(\vec{x})$ and $g(\vec{x})$ are convex. Then:

- $w_1 f(\vec{x}) + w_2 g(\vec{x})$ is convex, provided $w_1, w_2 \geq 0$
- $\max\{f(\vec{x}), g(\vec{x})\}$ is convex

# Another Property: Affine Composition

- Suppose that $f(x)$ is convex. Let $A$ be a matrix, and $\vec{x}$ and $\vec{b}$ be vectors. Then

$$g(\vec{x}) = f(A\vec{x} + \vec{b})$$

  is convex as a function of $\vec{x}$.

- Useful!

# Proving Convexity Properties

- Show that the definition holds.
- Remember: a function $f : \mathbb{R}^d \to \mathbb{R}$ is **convex** if for every choice of $\vec{a}, \vec{b} \in \mathbb{R}^d$ and $t \in [0, 1]$:

$$(1 - t)f(\vec{a}) + tf(\vec{b}) \geq f((1 - t)\vec{a} + t\vec{b}).$$

$(1-t)f(\vec{a}) + tf(\vec{b}) \geq f((1-t)\vec{a}+t\vec{b})$

# Example

$h(\vec{x})$

- Prove that $w_1 f(\vec{x}) + w_2 g(\vec{x})$ is convex, provided that $w_1, w_2 \geq 0$ and $f$ and $g$ are convex.

$$(1-t)h(\vec{a}) + th(\vec{b}) = (1-t)\left[w_1 f(\vec{a}) + w_2 g(\vec{a})\right] + t\left[w_1 f(\vec{b}) + w_2 g(\vec{b})\right]$$

$$= w_1\left[(1-t)f(\vec{a}) + tf(\vec{b})\right] + w_2\left[(1-t)g(\vec{a}) + tg(\vec{b})\right]$$

by convexity of $f$ $\geq$ $\qquad$ $\geq$

$$\geq w_1 f((1-t)\vec{a}+t\vec{b}) + w_2 g((1-t)\vec{a}+t\vec{b})$$

$$= h((1-t)\vec{a} + t\vec{b})$$

# DSC 140A

## Probabilistic Modeling & Machine Learning

Lecture 5 | Part 3

## Convex Loss Functions

# Convexity and Gradient Descent

- Convex functions are (relatively) easy to optimize.
- **Theorem**: if $f(x)$ is convex and "not too steep"[3] then (stochastic) (sub)gradient descent converges to a **global optimum** of $f$ *provided* that the step size is small enough[4].

[3] Technically, $c$-Lipschitz

[4] step size related to steepness, should decrease like $1/\sqrt{t}$, where $t$ is step number

$R = \frac{1}{n} \sum_i \mathcal{L}(\vec{x}^{(i)})$

# Convex Loss

- **Recall:** sums of convex functions are convex.
- **Implication:** if loss function is convex as a function of $\vec{w}$, so is the risk.
- Convex losses are **nice**.

# Example

- Recall the square loss:

$$L(H(\vec{x}, \vec{w}), y) = (\vec{x} \cdot \vec{w} - y)^2$$

- Is this convex as a function of $\vec{w}$?

# Mean Squared Error

- The square loss is a convex function of $\vec{w}$.
- We had an explicit solution for the best $\vec{w}$:

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

- But we could also have used gradient descent.

# Perceptron Loss



- The perceptron loss is:

$$L_{\text{tron}}(H(\vec{x};\vec{w}),y) = \begin{cases} 0, & \text{sign}(\vec{w}\cdot\vec{x}) = \text{sign}(y) \\ |\vec{w}\cdot\vec{x}|, & \text{sign}(\vec{w}\cdot\vec{x}) \neq \text{sign}(y) \end{cases}$$

- Is it convex as a function of $\vec{w}$?

# Summary

- 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.