EECS 553

Machine Learning (ECE)

Fall 2024

Course Information and Policies

- Questions?

Unconstrained Optimization

Unconstrained Optimization

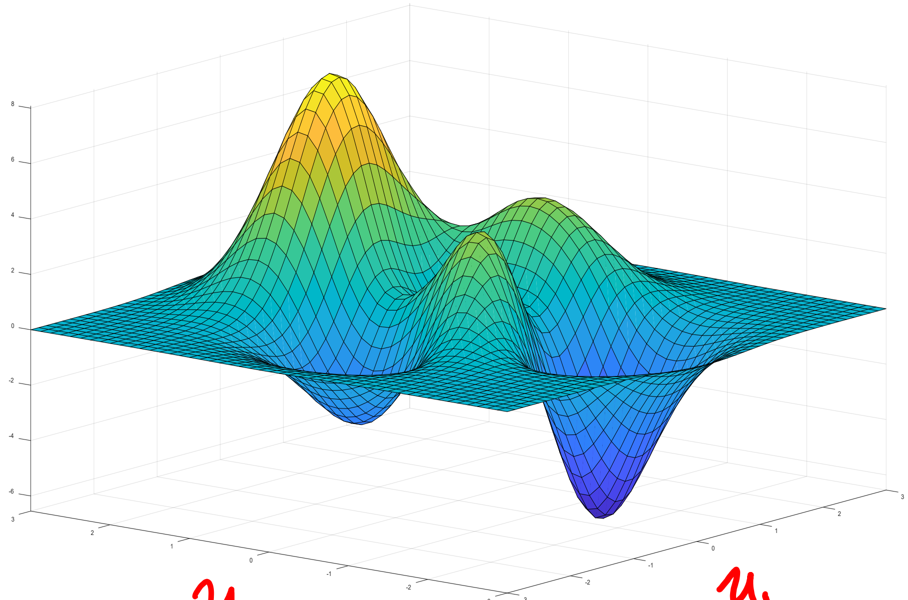
An *unconstrained optimization problem* has the form

$$\min_{\boldsymbol{u} \in \mathbb{R}^d} f(\boldsymbol{u})$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is called the *objective function.*

$d = 2$

$f(u)$



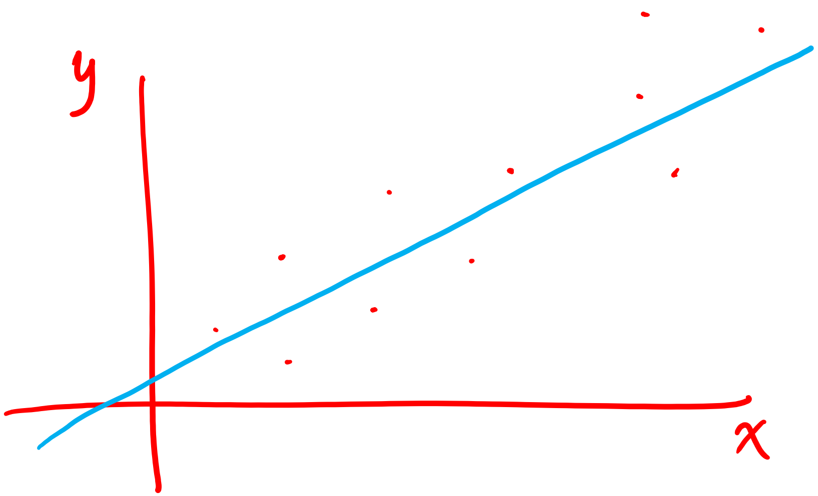
$u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \in \mathbb{R}^2$

u_2 u_1

Motivation

- Many machine learning methods are derived as the minimizer or maximizer of a certain objective function.
- Example: least squares linear regression

$(x_1, y_1), \ldots, (x_n, y_n)$ $\quad x_i \in \mathbb{R}, \; y_i \in \mathbb{R}$

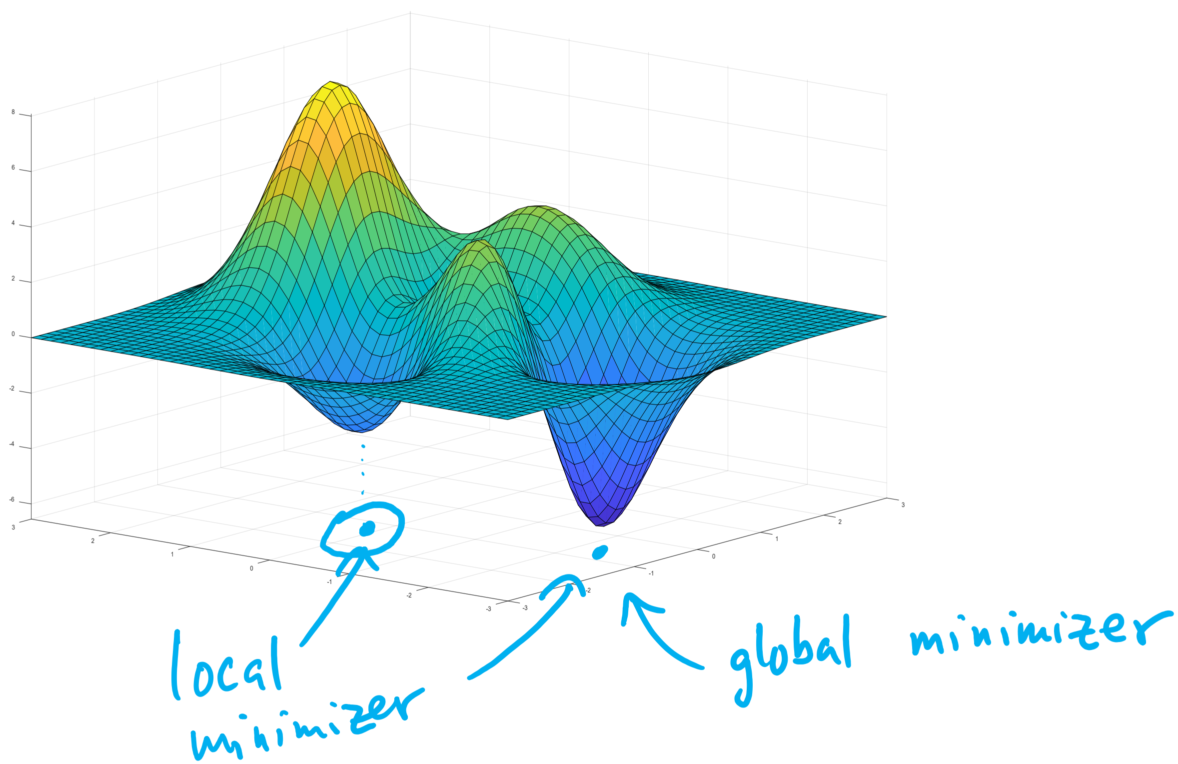


$$\min_{a,b} \underbrace{\sum_{i=1}^{n} \left(y_i - [a x_i + b]\right)^2}_{f(w)}$$

$$w = \begin{bmatrix} a \\ b \end{bmatrix}$$

Local and Global Minimizers

- A point $\boldsymbol{u}^* \in \mathbb{R}^d$ is called a *local minimizer* if $\exists r > 0$ such that $f(\boldsymbol{u}^*) \leq f(\boldsymbol{u})\ \forall \boldsymbol{u}$ satisfying $\|\boldsymbol{u} - \boldsymbol{u}^*\| < r$.
- $\boldsymbol{u}^*$ is called a *global minimizer* if $f(\boldsymbol{u}^*) \leq f(\boldsymbol{u})\ \forall \boldsymbol{u} \in \mathbb{R}^d$.

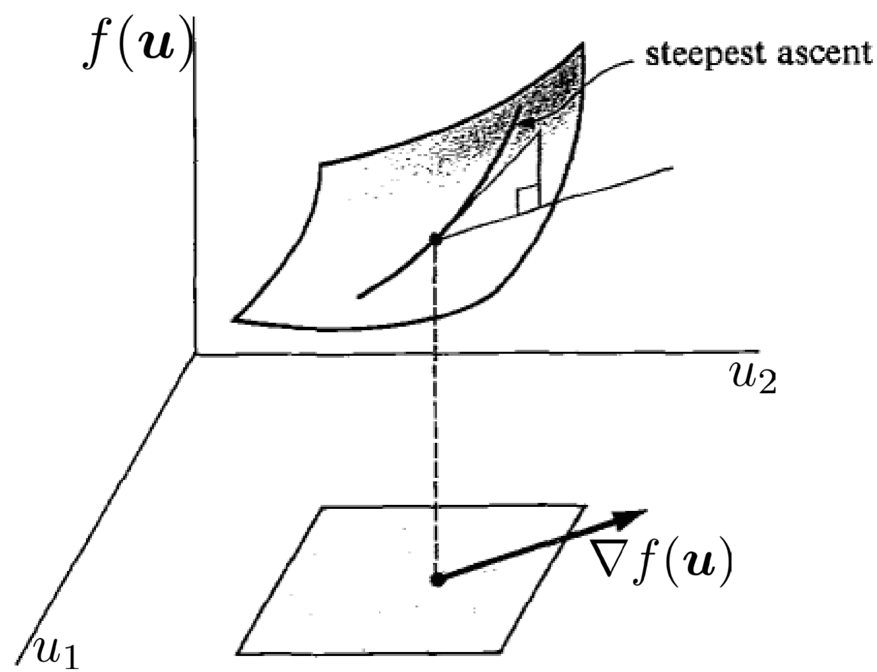


Gradient

- Given a function $f : \mathbb{R}^d \to \mathbb{R}$, the *gradient* f at $\boldsymbol{u} = [u_1 \ \cdots \ u_d]^T \in \mathbb{R}^d$ is defined by

$$\nabla f(\boldsymbol{u}) := \begin{bmatrix} \frac{\partial f(\boldsymbol{u})}{\partial u_1} \\ \vdots \\ \frac{\partial f(\boldsymbol{u})}{\partial u_d} \end{bmatrix}$$

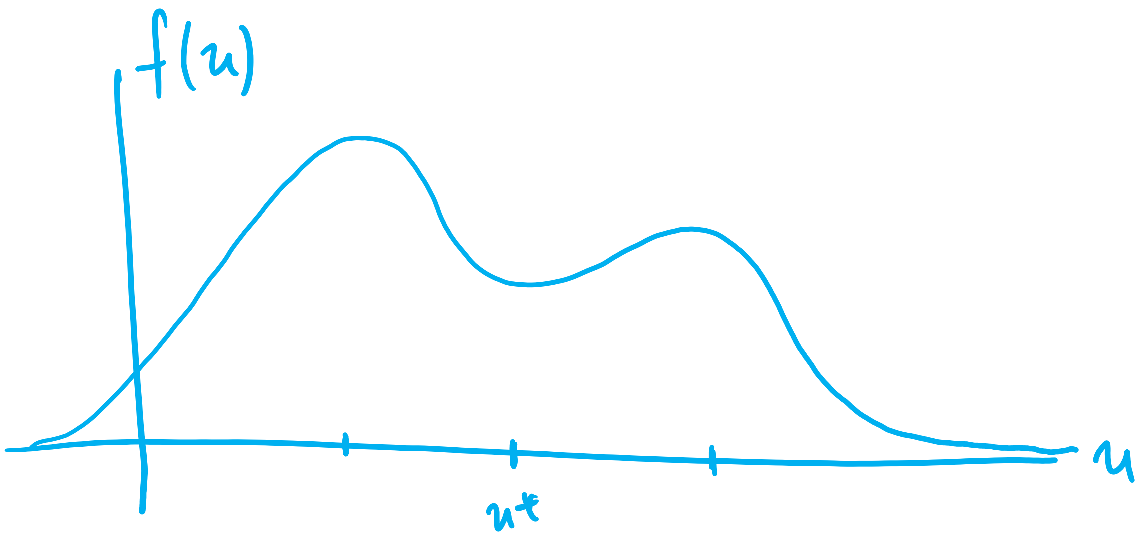
- The gradient gives the direction of *steepest ascent*.



First Order Necessary Condition

- If f is differentiable and $\boldsymbol{u}^*$ is a local minimizer of f, then

$$\nabla f(\boldsymbol{u}^*) = \mathbf{0}.$$



- Note that $\nabla f(\boldsymbol{u}^*) = \mathbf{0}$ is necessary, but *not* sufficient for $\boldsymbol{u}^*$ to be a local minimizer.

- If $\nabla f(\boldsymbol{u}) = \mathbf{0}$ for some $\boldsymbol{u}$, then $\boldsymbol{u}$ is said to be a *critical point* or *stationary point* of f.

Hessian

- The *Hessian* of f at $\boldsymbol{u}$ is the $d \times d$ matrix

$$\nabla^2 f(\boldsymbol{u}) := \begin{bmatrix} \frac{\partial^2 f(\boldsymbol{u})}{\partial u_1^2} & \cdots & \frac{\partial^2 f(\boldsymbol{u})}{\partial u_1 \partial u_d} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\boldsymbol{u})}{\partial u_d \partial u_1} & \cdots & \frac{\partial^2 f(\boldsymbol{u})}{\partial u_d^2} \end{bmatrix}$$

- If f is twice continuously differentiable, then $\nabla^2 f(\boldsymbol{u})$ is a symmetric matrix $\forall \boldsymbol{u}$.

Positive (Semi-)Definite Matrices

Let $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ be a square matrix. We say

- $\boldsymbol{A}$ is *positive semi-definite* if $\boldsymbol{z}^T \boldsymbol{A} \boldsymbol{z} \geq 0$ for all $\boldsymbol{z} \in \mathbb{R}^d$
- $\boldsymbol{A}$ is *positive definite* if $\boldsymbol{z}^T \boldsymbol{A} \boldsymbol{z} > 0$ for all $\boldsymbol{z} \neq \boldsymbol{0}$.

Clearly if $\boldsymbol{A}$ is PD, it is also PSD.

Properties: If $\boldsymbol{A}$ is a symmetric matrix, then

- $\boldsymbol{A}$ is *positive semi-definite* iff all eigenvalues of $\boldsymbol{A}$ are nonnegative
- $\boldsymbol{A}$ is *positive definite* iff all eigenvalues of $\boldsymbol{A}$ are positive

Spectral Thm: If A is symmetric, then

$$A = U \Lambda U^T$$

where $U^T U = U U^T = I$, $\Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_d \end{bmatrix}$

Second Order Necessary Condition

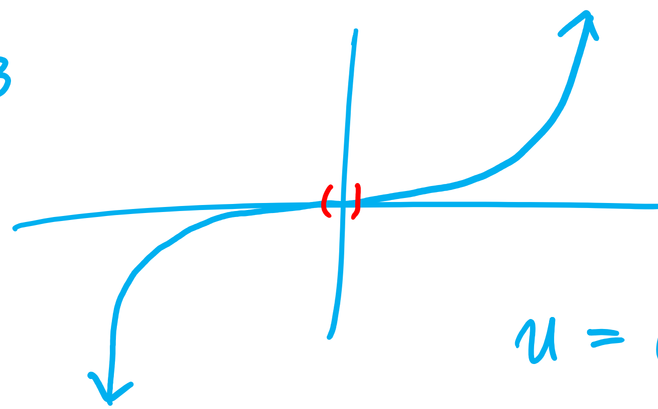
- If f is twice continuously differentiable and $\boldsymbol{u}^*$ is a local min, then $\nabla^2 f(\boldsymbol{u}^*)$ is positive semi-definite, i.e.,

$$\boldsymbol{z}^T \nabla^2 f(\boldsymbol{u}^*) \boldsymbol{z} \geq 0 \quad \forall \boldsymbol{z} \in \mathbb{R}^d.$$

- This generalizes the result from single-variable calculus that the second derivative is nonnegative at a local min.

- Give an example of a function f and a critical point $\boldsymbol{u}$ such that $\nabla^2 f(\boldsymbol{u})$ is PSD but $\boldsymbol{u}$ is not a local minimizer

$d=1$ $\qquad f(u) = u^3$



$u = 0$

Example

Notation:

$$\boldsymbol{u} = \begin{bmatrix} u \\ v \end{bmatrix}$$

Consider the function

$$f(u, v) = u^2 + 4uv - v^2 - 8u - 6v + 10$$

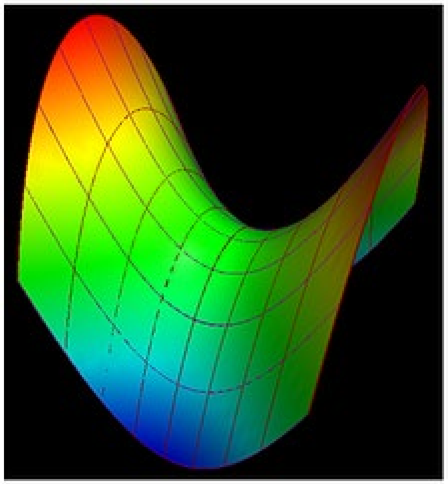
1. Determine $\nabla f(u, v)$ $= \begin{bmatrix} \partial f / \partial u \\ \partial f / \partial v \end{bmatrix} = \begin{bmatrix} 2u + 4v - 8 \\ 4u - 2v - 6 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$
2. Determine $\nabla^2 f(u, v)$
3. Determine a critical point $\boldsymbol{u}^*$
4. Is $\boldsymbol{u}^*$ a local min, a local max, or neither?

$\Downarrow$

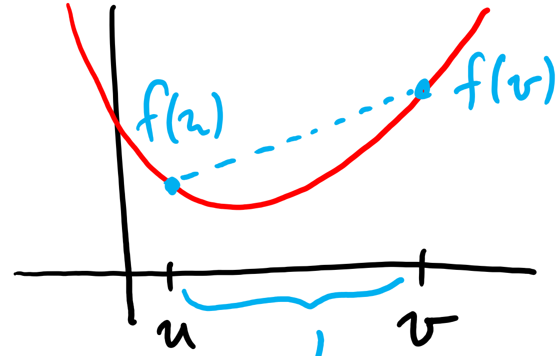
$u^* = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$

$$\nabla^2 f(u) = \begin{bmatrix} 2 & 4 \\ 4 & -2 \end{bmatrix} \xrightarrow{\text{evals}} \pm 4.47$$

Example, Continued



Convexity



- We say that f is *convex* if

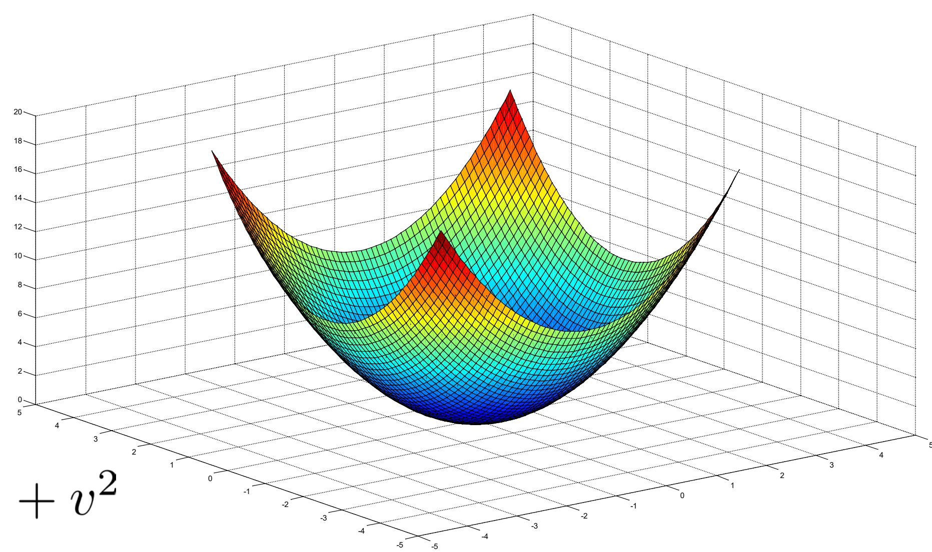
$$f(t\boldsymbol{u} + (1-t)\boldsymbol{v}) \leq tf(\boldsymbol{u}) + (1-t)f(\boldsymbol{v})$$

$\forall \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$ and $t \in [0,1]$.

- We say f is *strictly convex* if

$$f(t\boldsymbol{u} + (1-t)\boldsymbol{v}) < tf(\boldsymbol{u}) + (1-t)f(\boldsymbol{v})$$

$\forall \boldsymbol{u} \neq \boldsymbol{v}$ and $t \in (0,1)$.

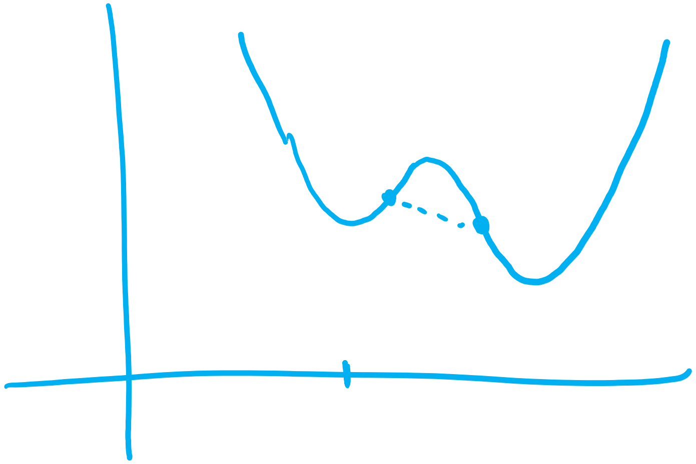


$$f(u,v) = u^2 + v^2$$

Convex Functions are Nice

Properties of convex functions

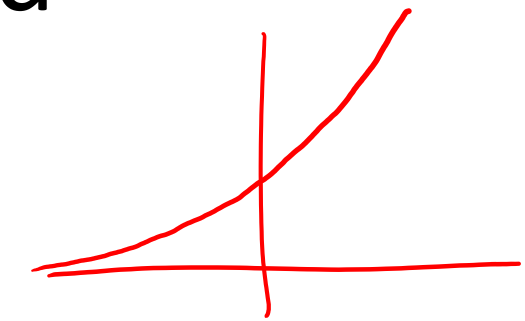
1. If f is convex, then every local min is a global min (see lecture notes).
2. If f is strictly convex, then f has at most one global min (exercise).



Raise Your Hand

Give an example of a function $f : \mathbb{R} \to \mathbb{R}$ that is

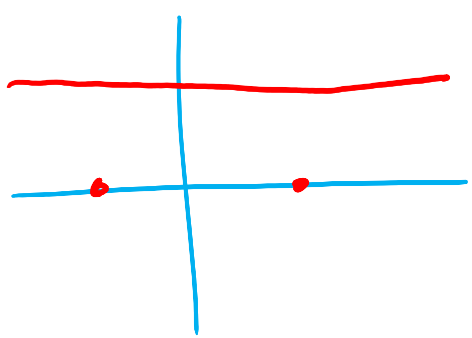
- convex but not strictly convex
- convex and has more than one global minimizer
- strictly convex, but has no global minimizer

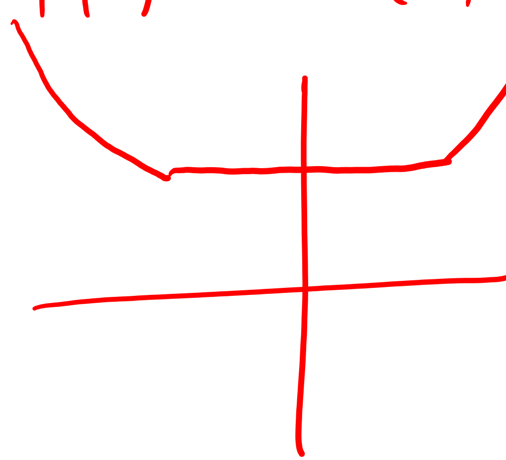


$f(u) = e^u$

$f(u) = 1$

$f(u) = \max(1, u^2)$





Poll

True or false: The product of convex functions is necessarily convex.

(A) True

(B) False ✓

$f(x) = x^2$

$g(x) = -1$

$(f \cdot g)(x) = f(x)g(x) = -x^2$

Poll

True or false: The composition of convex functions is necessarily convex.

(A) True

(B) False ✓

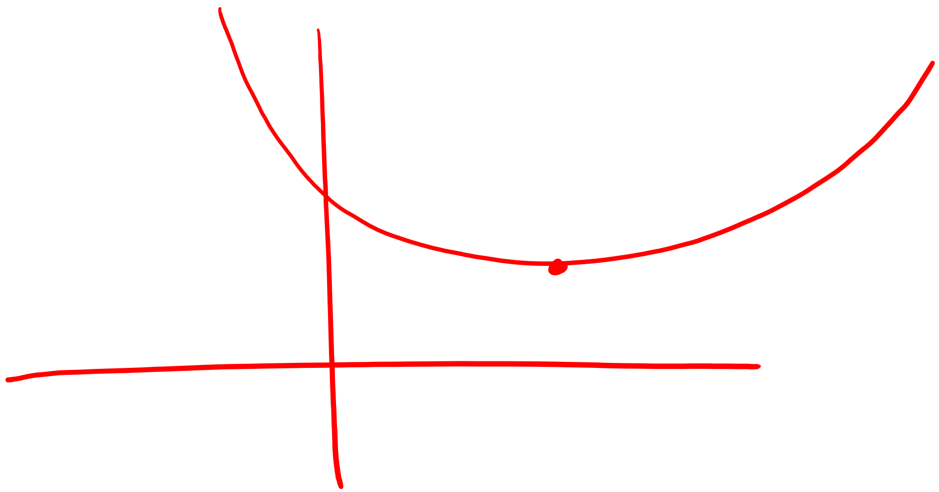
$$f(u) = -u$$

$$g(u) = u^2$$

$$f(g(u)) = -u^2 \quad \text{not convex}$$

1^{st} Order Condition for Local Min, Revisited

- For convex functions, $\nabla f(\boldsymbol{u}^*) = \mathbf{0}$ is both necessary *and* sufficient for $\boldsymbol{u}^*$ to be a local min.



Second Order Characterizations of Convexity

- f is convex $\Longleftrightarrow \nabla^2 f(\boldsymbol{u})$ is positive semidefinite $\forall \boldsymbol{u} \in \mathbb{R}^d$
- f is strictly convex $\Longleftarrow \nabla^2 f(\boldsymbol{u})$ is positive definite $\forall \boldsymbol{u} \in \mathbb{R}^d$

Exercise

Numerically determine a critical point of

$$f(u,v) = u^2 + 2uv + 3v^2 + 4u + 5v + 6$$

and also determine if it is a local/global min or max. *Note:* If you don't have immediate access to Python/Matlab/etc., you can also use Wolfram Alpha for many calculations like eigenvalue decompositions

$$\nabla f(u) = \begin{bmatrix} \partial f/\partial u \\ \partial f/\partial v \end{bmatrix} = \begin{bmatrix} 2u + 2v + 4 \\ 2u + 6v + 5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \Rightarrow u^* = \begin{bmatrix} -1.75 \\ -.25 \end{bmatrix}$$

$$\nabla^2 f(u) = \begin{bmatrix} 2 & 2 \\ 2 & 6 \end{bmatrix} \xrightarrow{\text{e-vals}} 4 \pm 2\sqrt{2} > 0$$

Exercise Solution

- Since the Hessian is PD $\forall u$, f is strictly convex
- Since f is convex, the critical pt u^* is a local min.
- Since f is convex, every local min. is a global min.
- Since f is strictly convex, u^* is the unique global min.

Summary

- The gradient and Hessian allow us to state necessary and sufficient conditions for local and global optimality in unconstrained optimization problems
- Convex objections function make an unconstrained optimization problem easier to understand (and, as we will see, to solve)
- Next time: Begin supervised learning, apply today's material