

Lecture 28

Convexity, optimality, and

Gradient descent.

* MIDTERM

^{this}
FRIDAY 13 TH

* EXAM LOGISTICS

→ SIGN UP SLOTS

→ NEED ALTERNATIVES?

(EMAIL US)

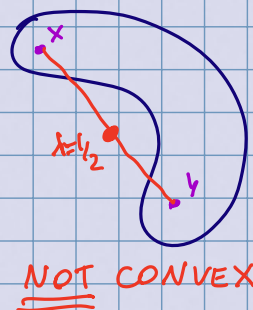
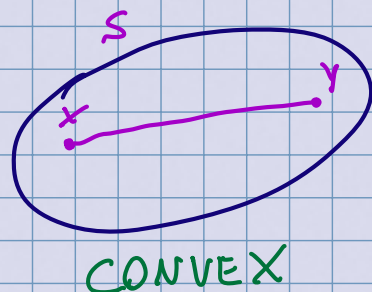
CONVEX SETS

(in $\mathbb{R}^n$,
or any real vector
space)

A set S is convex if

$$x, y \in S \Rightarrow \lambda x + (1-\lambda)y \in S$$

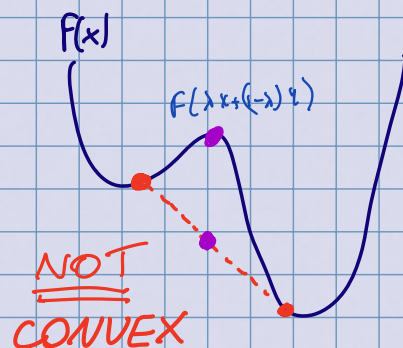
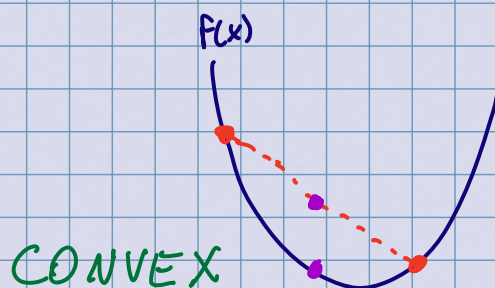
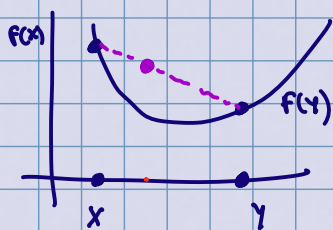
$$\forall \lambda \in [0, 1]$$



CONVEX FUNCTIONS

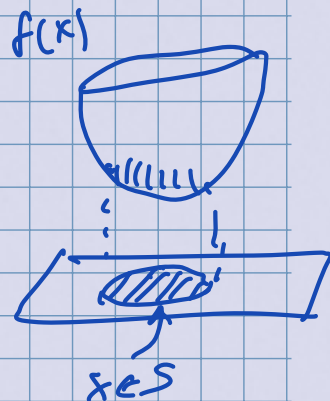
A function f is convex if

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$$



CONVEX OPTIMIZATION

PROBLEMS



$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & F(x) \\ \text{s.t.} \quad & g_i(x) \leq 0 \quad i=1 \dots m \end{aligned}$$

where f and g_i are convex functions

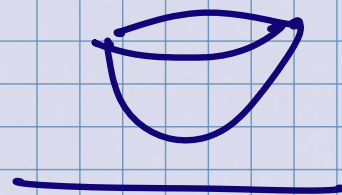
- The feasible set $\{x \in \mathbb{R}^n : g_i(x) \leq 0\}$ is a convex set.
- Important!
 - minimization
 - form of constraints

$$\begin{aligned} \min \\ g_i \leq 0 \end{aligned}$$

EXAMPLES:

● (CONVEX) QUADRATIC PROGRAMMING

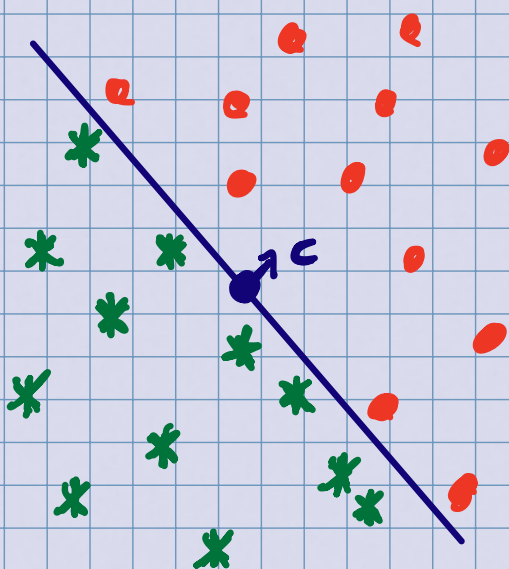
$$\min_x \quad \frac{1}{2} x^T A x - b^T x$$



where A is positive semidefinite

● LINEAR CLASSIFICATION

RECALL
(PERCEPTRON)



$$\min_c \quad \|c\|^2$$

$$c^T p \leq -1$$

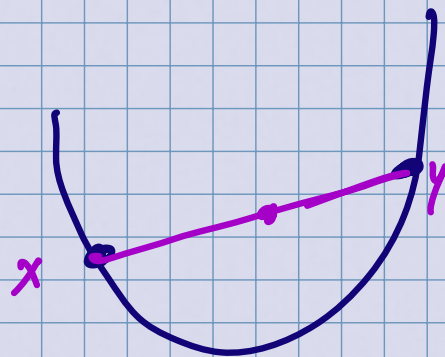
$$c^T p \geq 1$$

$$\forall p \in \text{green asterisks}$$

$$\forall p \in \text{red circles}$$

HOW TO RECOGNIZE A CONVEX FUNCTION?

- BY DEFINITION



- OFTEN, BY CONSTRUCTION RULES
(SUMS, MAXIMUM, ETC.)

e.g. $\max \left(x^2, e^x, 3x + (x-3)^4 \right)$

"DISCIPLINED CONVEX PROGRAMMING"
(JULIA'S CONVEX.JL)

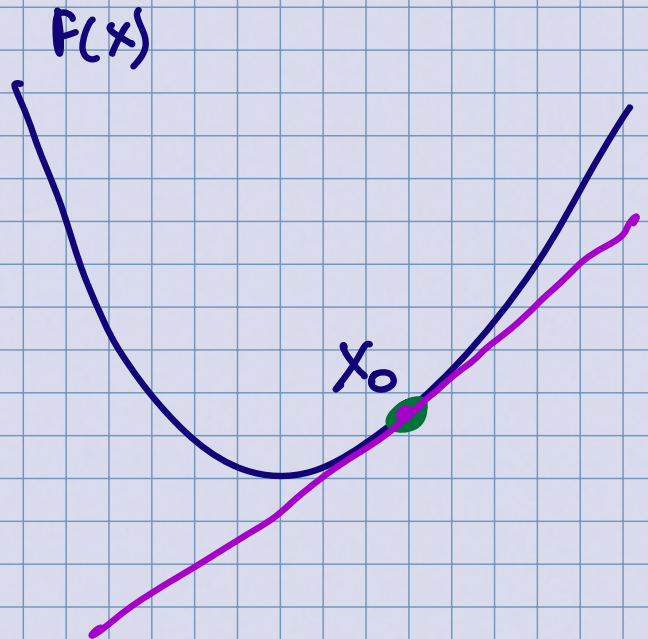
- ALTERNATIVELY, CHARACTERIZATIONS.
BASED ON DERIVATIVES

"FIRST ORDER" CHARACTERIZATION

- Assume $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is convex
and differentiable

LINEAR APPROXIMATION

$$f(x_0) + f'(x_0)(x - x_0)$$



f is convex



Function

linear (tangent)
approx.

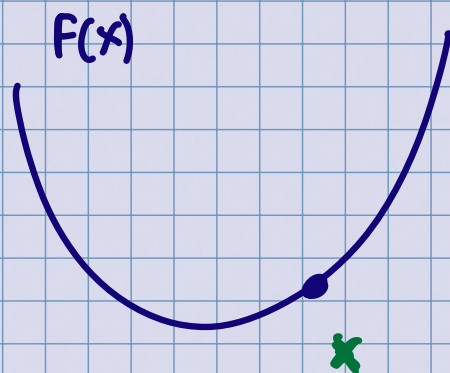
$$f(x) \geq f(x_0) + \nabla f(x_0)^T (x - x_0) \quad \forall \underline{x}, \underline{x_0}$$

"SECOND ORDER" CHARACTERIZATION

- Assume $F: \mathbb{R}^n \rightarrow \mathbb{R}$ is convex
and twice differentiable

"positive curvature"

Hessian $\nabla^2 F(x)$
(matrix of second derivatives)

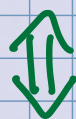


$$\nabla^2 F(x) = \begin{bmatrix} \frac{\partial^2 F}{\partial x_1^2} & \frac{\partial^2 F}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 F}{\partial x_1 \partial x_n} \\ \frac{\partial^2 F}{\partial x_1 \partial x_2} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \frac{\partial^2 F}{\partial x_n^2} \end{bmatrix}$$



SYMMETRIC

F is convex



$\nabla^2 F(x_0)$ is positive semidefinite $\forall x_0$

Example:

"log-sum-exp"

"softmax"

$$f(x, y) = \log(e^x + e^y)$$

psd

$$\nabla^2 f(x, y) = \frac{e^{x+y}}{(e^x + e^y)^2} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

positive semidefinite

"Negative entropy" is CONVEX.

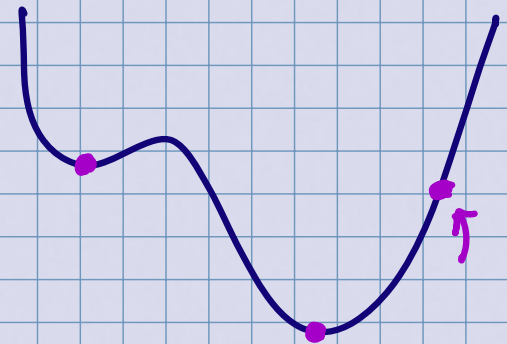
$$F(x_1, \dots, x_n) = \sum_{i=1}^n x_i \log x_i \quad (\text{on } \underline{\mathbb{R}}_+^n)$$

$$\nabla^2 F(x_1, \dots, x_n) = \begin{bmatrix} 1/x_1 & & 0 \\ & \ddots & \\ 0 & & 1/x_n \end{bmatrix} \rightarrow \text{positive definite (if } x_i > 0)$$

OPTIMALITY CONDITION :

$$\min_{x \in \mathbb{R}^n} f(x)$$

- What condition(s) must an optimal point satisfy?



NECESSARY :
CONDITION
(for local optimality).

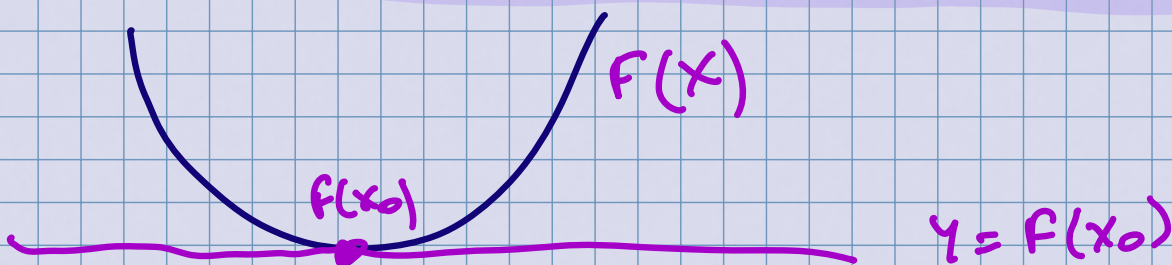
$$\nabla f(x_0) = 0$$

function

linear approx

$$f(x) \geq f(x_0) + \nabla f(x_0)^T (x - x_0)$$

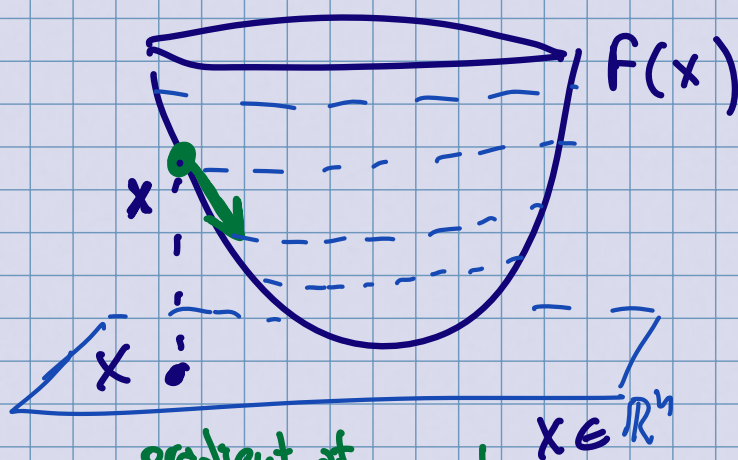
IN THE CONVEX CASE, $\nabla f = 0$ IS ALSO
SUFFICIENT FOR GLOBAL OPTIMALITY!



GRADIENT METHOD.

Therefore, to (globally!) optimize a convex function, want to find a point where the gradient vanishes.

Idea:



next
iterate

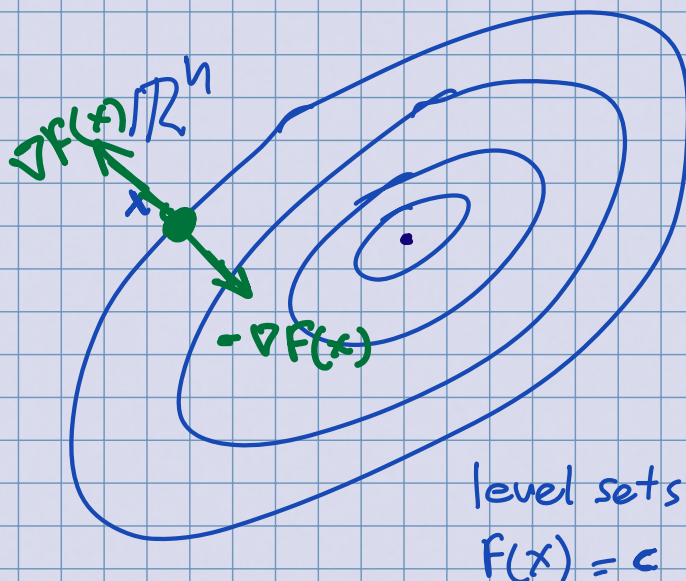
current iterate.

gradient at current
point

$$X_{k+1} = X_k - \gamma \nabla F(X_k)$$

stepsize

"go downhill,
along the direction
of the ^{minus} gradient"



QUESTIONS



- DOES THIS WORK?
- HOW QUICKLY DOES IT CONVERGE?
(if it does)
- HOW TO CHOOSE STEPSIZE α ?
(does it matter?)
- IS THIS METHOD GOOD?
ARE THERE BETTER ONES?

What happens for quadratic functions?

$$x_{k+1} = x_k - \gamma \nabla f(x_k)$$

$$F(x) = \frac{1}{2} x^T A x$$

(without loss of generality)

A is pd
(so f is ^{strictly} convex).

$$\nabla f(x) = \underline{Ax}$$

eigenvalues

$$\underline{\lambda_1} \geq \lambda_2 \geq \dots \lambda_n > \underline{0}$$

$$\Rightarrow x_{k+1} = x_k - \gamma \nabla f(x_k) = x_k - \gamma A x_k$$

$$x_{k+1} = (I - \gamma A) x_k$$

When does this converge?

$$x_{k+1} = M x_k$$

convergence if

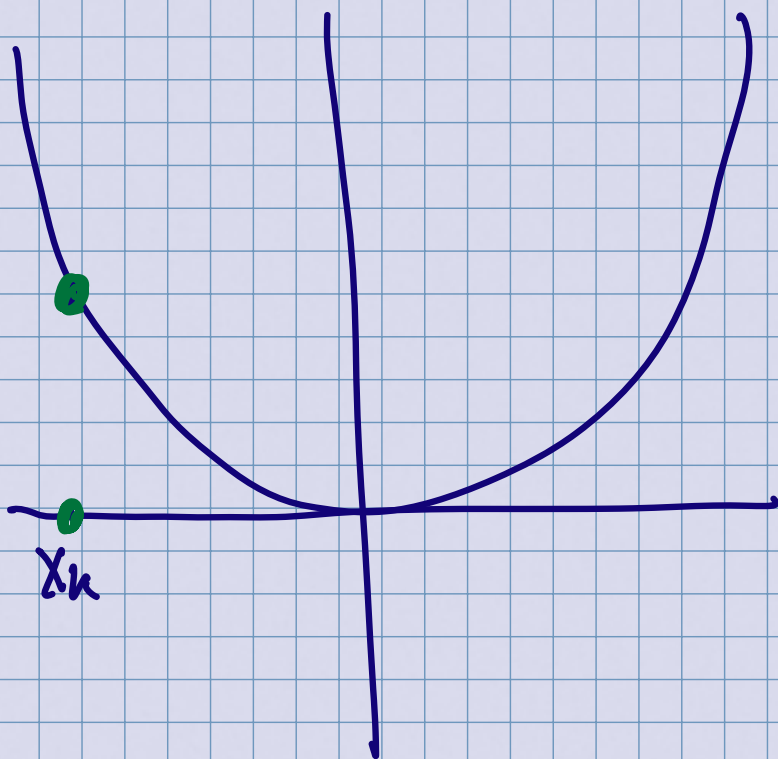
$$|\text{eig}(M)| < 1$$

$$M = I - \gamma A$$

$$\text{eig}(M) = 1 - \gamma \lambda_i$$

NOT EVERY STEPSIZE WORKS!

E.g., in the UNIVARIATE CASE $A=1$



$$F(x) = \frac{1}{2} x^2$$

$$x_{k+1} = x_k - \gamma \nabla F(x_k)$$

$$x_{k+1} = (1 - \gamma) x_k$$

$$\gamma = -1?$$



$$\gamma = 0?$$



$$\gamma = 2?$$



$$\gamma = 1?$$



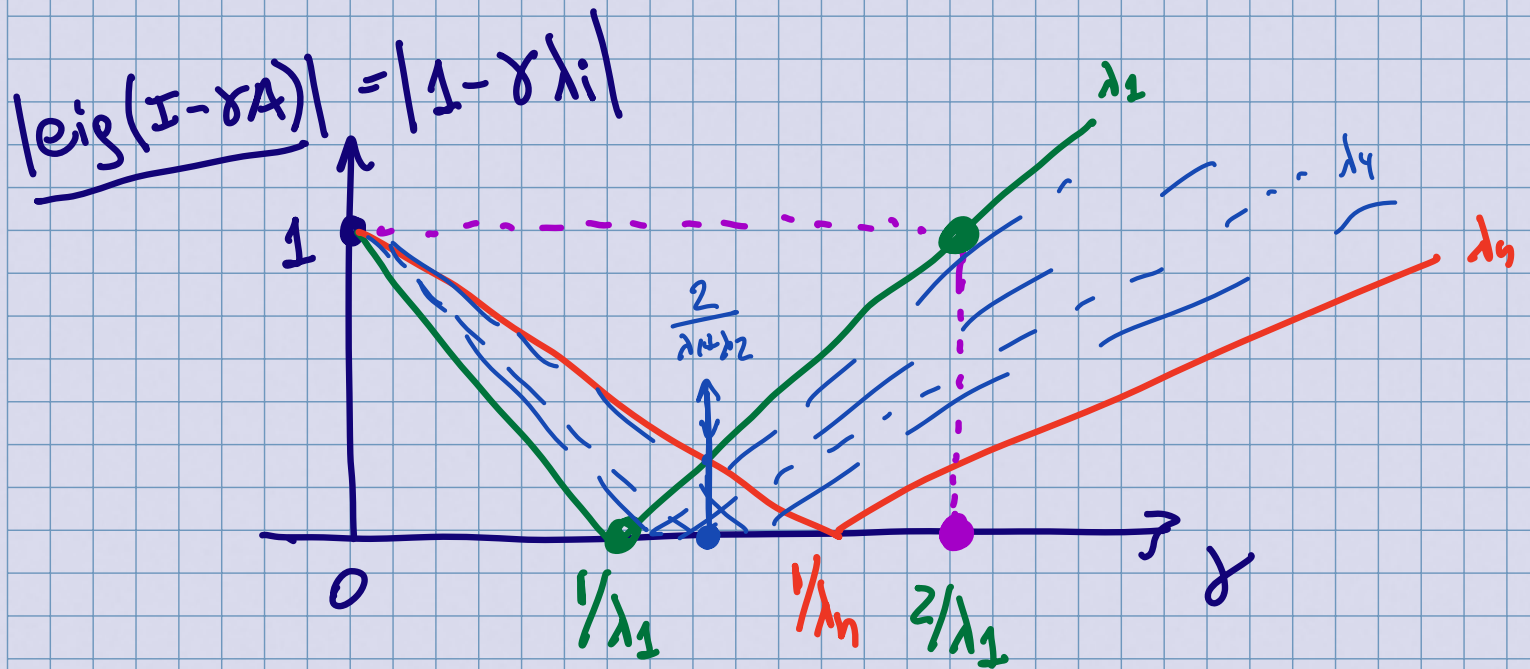
$$\gamma = 3?$$



$$\underline{\text{eig}(I - \gamma A)} = \underline{1 - \gamma \lambda_i}$$

$\lambda_i \text{ eig}(A)$

What is the "best" stepsize?



$\Rightarrow$ if $0 < \gamma < 2/\lambda_1$ $\leftarrow$
The gradient method converges.

If we measure convergence rate by

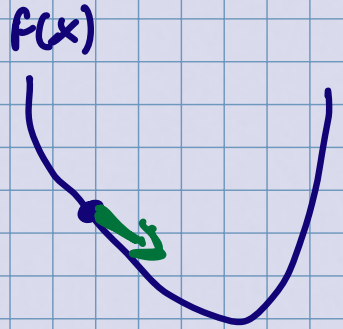
$$|\text{eig}(\mathbf{I} - \gamma \mathbf{A})| = |1 - \gamma \lambda_i|, \text{ the}$$

"best" stepsize is

$$\frac{2}{\lambda_1 + \lambda_n}$$

NEXT TIME

- What happens in the general (convex) case?



- Better methods?

