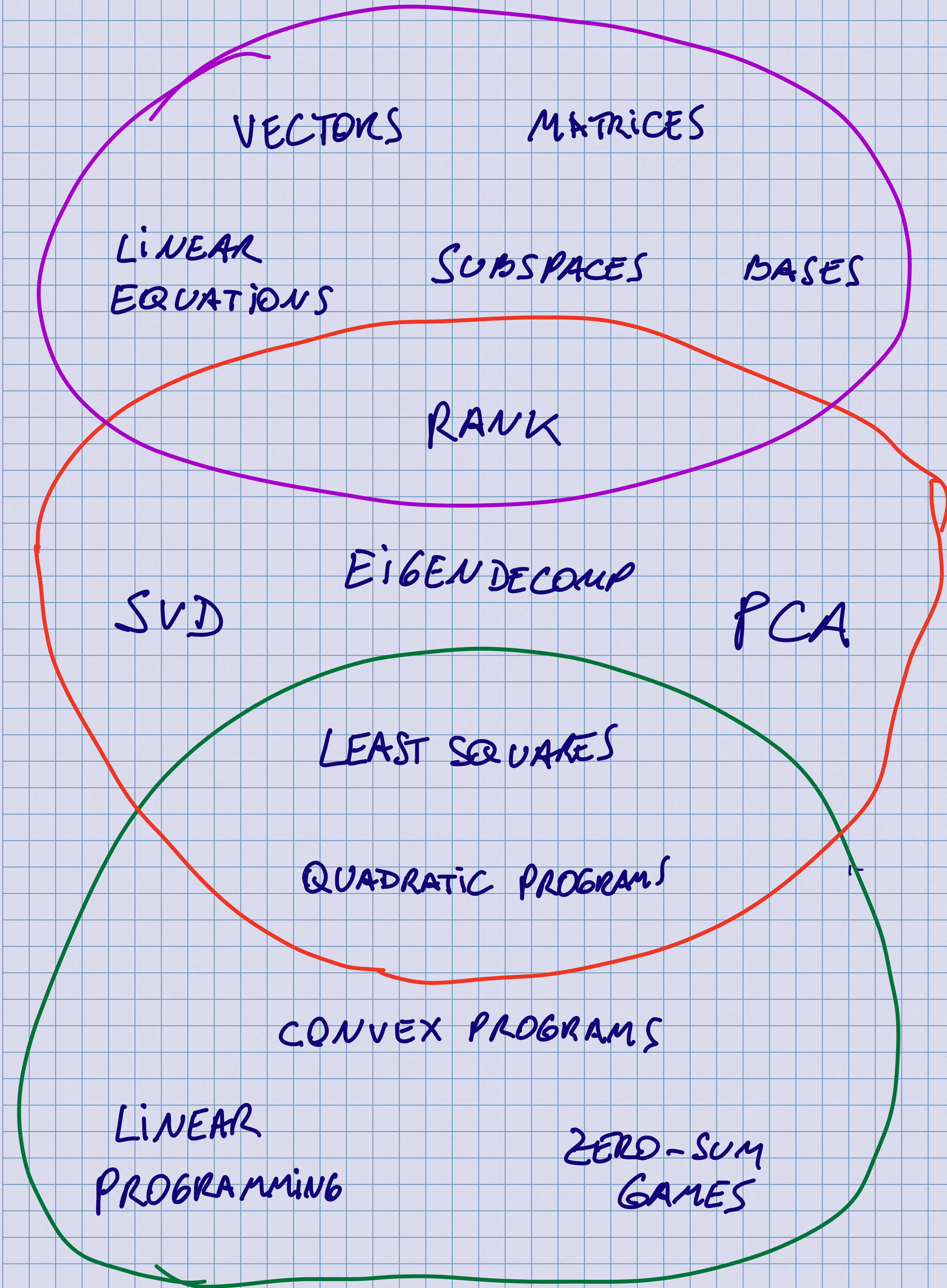
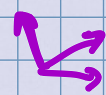
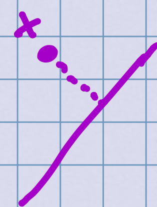
Lecture 35

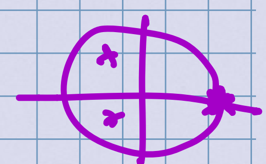
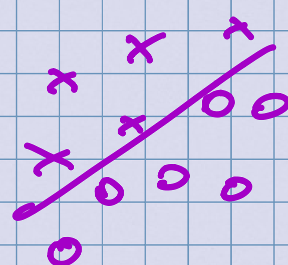
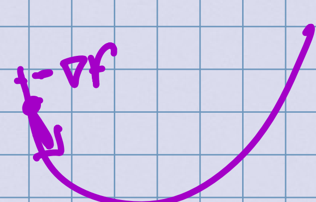
FINAL REVIEW (I)

- FINAL EXAM ON FRI 12/18
- PRACTICE EXAM WILL BE POSTED SOON


VECTORS
MATRICES
LINEAR EQUATIONS
SUBSPACES
BASES
RANK
EIGENDECOMP
SVD
PCA
LEAST SQUARES
QUADRATIC PROGRAMS
CONVEX PROGRAMS
LINEAR PROGRAMMING
ZERO-SUM GAMES

ALGORITHMS !

- GAUSSIAN ELIMINATION — $Ax = b$
- GRAM-SCHMIDT  $\{v_1, \dots, v_n\}$
- LEAST SQUARES / PROJECTIONS 

- SVD / PCA 
- POWER METHOD $v_k = A^k v_0$ 
- PERCEPTRON 
- GRADIENT DESCENT 

- SGD $\min_x \sum_i f_i(x)$

- YOU SHOULD HAVE A GOOD SENSE OF WHAT ALL THESE TERMS MEAN, AND HOW THEY RELATE TO EACH OTHER

- WHICH ALGORITHMS TO USE WHEN, AND WHY.

- UNDERSTAND WHAT DOES NOT MAKE SENSE, E.G.

  ✗ divide two vectors u/v, $u,v \in \mathbb{R}^n$

  ✗ eigenvalues of a nonsquare matrix

  ✗ gradient of a matrix

  ✗ quadratic form x^TQx with Q rectangular

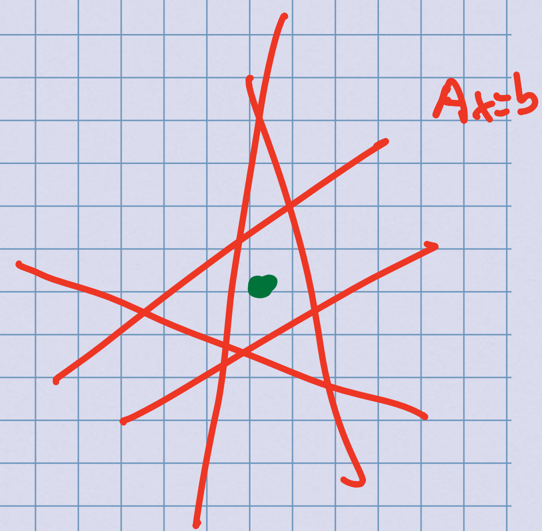
LEAST SQUARES:

Two Flavors

OVERDETERMINED

$$Ax = b$$

has no solution

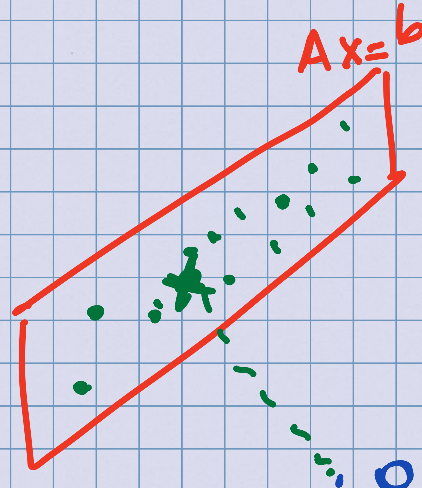


Define approx. least sq sol by

$$\min_{x} \|Ax-b\|^2$$

UNDERDETERMINED

$Ax=b$ has infinitely many solutions

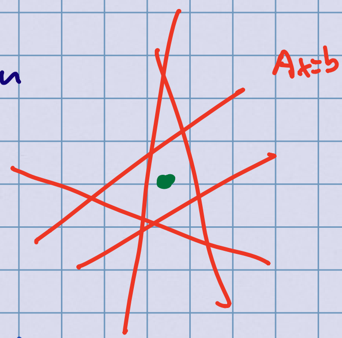


Define minimum norm (LS) solution by

$$\min \|x\|^2$$
$$\text{s.t. } Ax=b$$

OVERDETERMINED $Ax = b$

has <u>no</u> solution

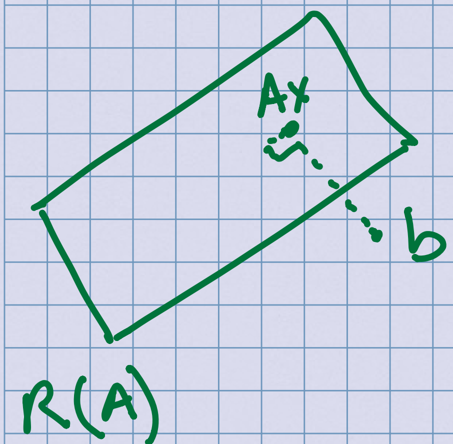


<u>Define</u> approx. least sq sol by

$$\min_x \|\underbrace{Ax}_{}-b\|^2$$

$\|y-b\|^2 \qquad y \in R(A)$

How to solve it?



- PROJECTION FORMULA

(project b onto the range of A)

- SVD $\quad (A = U\Sigma V^T)$

$$\|Ax-b\|^2 = \|U\Sigma V^T x - b\|^2$$
$$= \|\Sigma \underbrace{V^T x}_{} - \underbrace{U^T b}_{}\|^2$$
$$= \|\Sigma \tilde{x} - \tilde{b}\|^2$$

↖ easy to solve.

$\|Ux\| = \|x\|$

(U orthogonal)

- OPTIMIZATION

$$\min_x \frac{1}{2}(Ax-b)^T(Ax-b) \qquad \frac{1}{2}\|Ax-b\|^2$$

(optimality conditions)

$\nabla F(x) = 0$

$\hookrightarrow \quad A^T(Ax-b) = 0 \qquad$ (normal eqs)

$A^TA\,x = A^Tb$

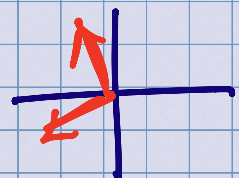
$$\Rightarrow \quad \boxed{x = (A^TA)^{-1}A^Tb}$$

(ASIDE:)

COOL LINEAR ALGEBRA FACT # 731

How many orthogonal unit-norm vectors can I find in $\mathbb{R}^n$?

E.g. in $\mathbb{R}^2$



2 vectors

$\begin{pmatrix}1\\0\end{pmatrix}, \begin{pmatrix}0\\1\end{pmatrix}$ (or any rotations)

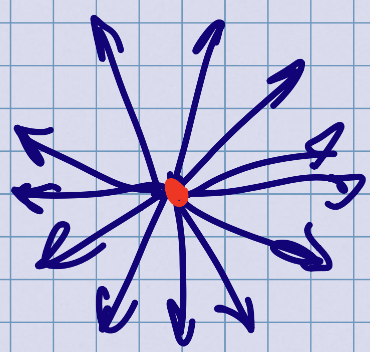
in $\mathbb{R}^n$ → at most $\underline{\underline{n}}$ orthogonal vectors.

(why? orthogonal ⇒ LI)

But, what if I allow them to be "almost orthogonal"?

SURPRISE ?

In high dimensions, a LOT more.



all nearly orthogonal to each other

For instance, in $\mathbb{R}^{\overbrace{1000000}^{10^6}}$, can find

$\boxed{1.17 \times 10^{33}}$ vectors with

$$89^\circ \leq \text{angle}(v_i, v_j) \leq 91^\circ$$

(compare against $n = 10^6$ (!))

QUADRATIC PROGRAMS

convex quadratic function of x

$$\min_x \quad q(x)$$

$$\text{s.t.} \quad Ax = b$$

linear equations (subspace)

$$\begin{cases} \|Ax-b\|^2 \\ \|x\|^2 \\ x^TQx \\ \frac{1}{2}x^TQx + b^Tx \\ \dots \end{cases}$$

(matrix Q is symmetric)

$$q(x) = (Ax-b)^T(Ax-b)$$
$$= x^T A^T A x + \dots$$

$Q = A^TA$

Q psd

q(x) must be convex, i.e.,

matrix Q is PSD $(\lambda_i \geq 0)$

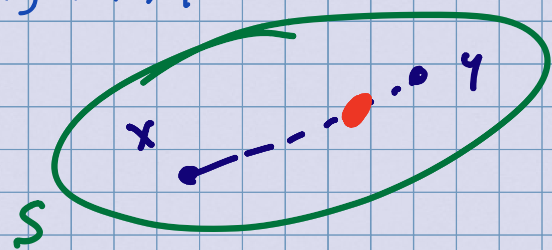
CONVEX PROGRAMS

$$\boxed{\begin{aligned} &\min_x \quad f(x) \\ &g_i(x) \leq 0 \end{aligned}}$$

$f(x)$, $g_i(x)$ are CONVEX functions

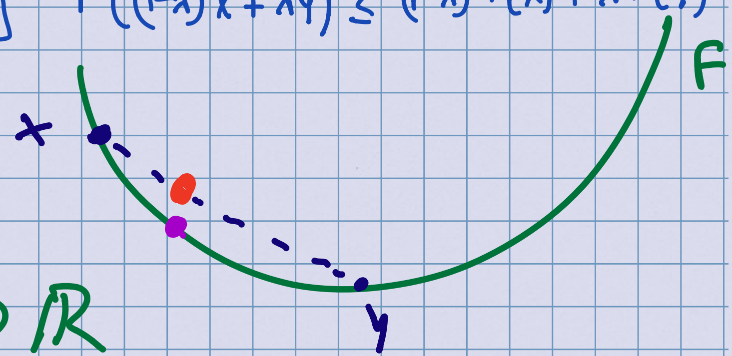
→ convex sets

$$(1-\lambda)x + \lambda y \in S \quad \forall \lambda \in [0,1]$$



convex functions

$$\forall \lambda \in [0,1] \quad f\big((1-\lambda)x + \lambda y\big) \leq (1-\lambda) f(x) + \lambda f(y)$$



$$f : \mathbb{R}^n \to \mathbb{R}$$

IF CONVEX: LOCAL MINIMUM $\Longrightarrow$ GLOBAL MINIMUM

x^* GLOBAL OPTIMUM $\iff$ $\nabla f(x^*) = 0$

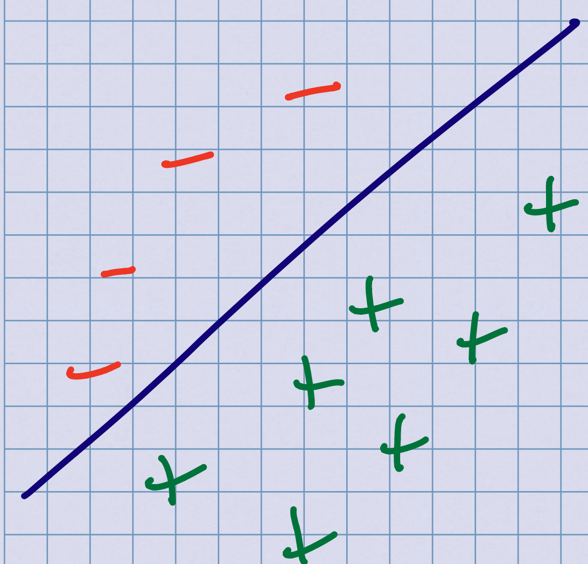
NONNEGATIVE LEAST SQUARES

$$\min \quad \|Ax-b\|^2 \qquad (NNLS)$$
$$\text{s.t.} \quad x \geq 0$$

- **NOT** A LEAST SQUARES PROBLEM
  (of the type we have seen)

CLASSIFICATION : (SVM)

$c^Tx+b=0$



$$\left\{\begin{array}{l}\min \quad \|c\|^2 \\ \underset{\text{label}}{y}\left(c^T\underset{\text{point}}{x}+b\right) \geq -1\end{array}\right.$$

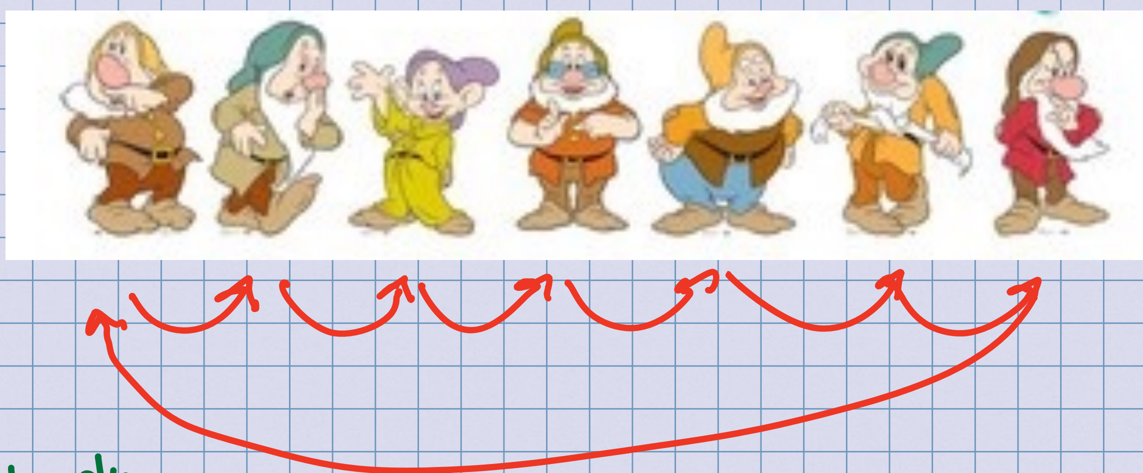
$\{+1,-1\}$

CONVEX PROBLEMS $\Rightarrow$ GOOD ALGORITHMS

SNOW WHITE : (PIAZZA @420)

Snow White distributed 399 liters of milk among the seven dwarfs. The first dwarf then distributed the contents of his pail evenly to the pails of other six dwarfs. Then the second did the same, and so on. After the seventh dwarf distributed the contents of his pail evenly to the other six dwarfs, it was found that each dwarf had exactly as much milk in his pail as at the start.

a) What was the initial distribution of the milk?
b) Generalize to N dwarfs.



$$T_i = \begin{array}{c} \text{i-th column} \\ \left[\begin{array}{ccccccc} 1 & & & 1/6 & & & \\ & 1 & & 1/6 & & & \\ & & 1 & 1/6 & & & \\ & & & 0 & & & \\ & & & 1/6 & 1 & & \\ & & & 1/6 & & 1 & \\ & & & \vdots & & & 1 \\ & & & 1/6 & & & \end{array}\right] \end{array} \qquad = \underbrace{T_7 \cdot \ldots \cdot T_2 \, T_1}$$

julia helps!

GRADIENT DESCENT

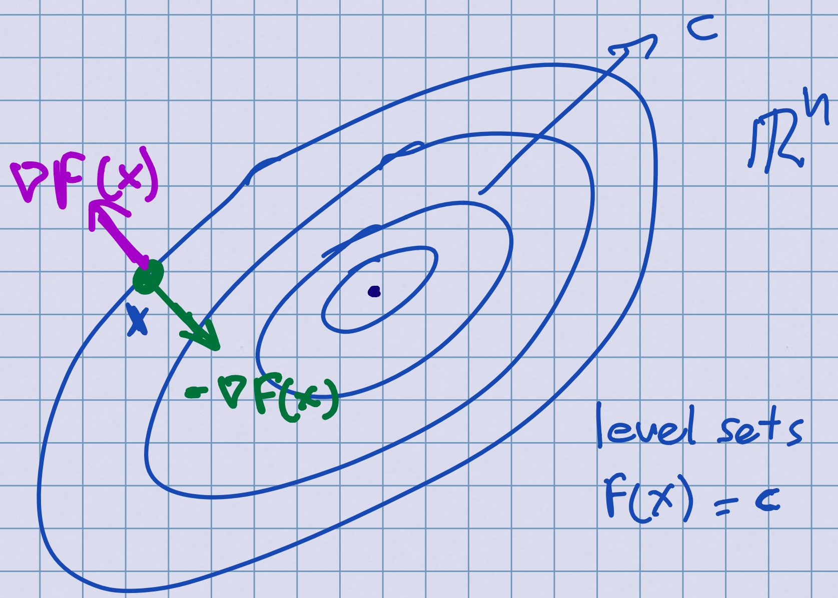
$$\boxed{x_{k+1} = x_k - \gamma \underbrace{\nabla f(x_k)}}$$

next point

current point

stepsize

gradient at current point



GRADIENT DESCENT (convex functions)

- CONVERGES FOR A RANGE OF STEPSIZES

$$0 < \gamma < \frac{2}{L}$$

$0 < m \leq \lambda_i(x) \leq L$

eigs of Hessian

- For strictly convex functions (i.e, $m > 0$) rate depends on condition number $\frac{L}{m}$

- Accelerated methods (e.g., heavy ball) can achieve improvements

- For large, data-oriented problems, SGD is a good option

NEXT TIME :

OVERVIEW OF SOME RESEARCH DIRECTIONS

+

OTHER COURSES AT MIT