Introduction to Optimization

CS 111
Winter 2023

Numerical Optimization

Minimize a real-valued function f of n real variables

$$\min_{x} f(x_0, x_1, \ldots, x_{n-1})$$

In vector notation, among $x \in \mathbb{R}^n$,

find $x^* = \operatorname{argmin} f(x)$.

Usually we're more interested in x^*

than in $p^* = \min f(x) = f(x^*)$.

$f(x) : \mathbb{R}^n \to \mathbb{R}$ is the "objective function"

or (in machine learning) the "loss function".

Examples

- Best location (x_0, x_1) for a new cell tower.

  $f(x_0, x_1) =$ weakest signal in neighborhood

- Best cross-section for an airplane wing.

  $f(x_0, x_1, \ldots, x_{1000}) = -\text{lift}$ (to maximize lift)

- Least-squares data fitting, $Ax \approx b$.

  $f(x) = \|Ax - b\|^2$

- Training deep neural nets: best weights "$x_0, \ldots, x_{10^{11}}$"

  $f(x) =$ "loss function"

  $$= \sum_{\text{training data}} (\text{DNN output} - \text{right answer})^2$$

Possible Additional Features

- Sometimes there are constraints on x:
  - $Ax=b$, $Ax\le b$, $x_i\in\{0,1\}$, etc.
  - Convex constraints, nonlinear constraints, etc.
- Sometimes the function f has special features:
  - convex
  - linear
  - smooth
  - integer-valued

IN CS111, WE WILL ONLY CONSIDER:

- Unconstrained minimization
  (x can be anything in $\mathbb{R}^n$)
- Convex functions f
  (every chord is above the function values)
- Smooth functions f
  (continuous, sometimes derivatives too)

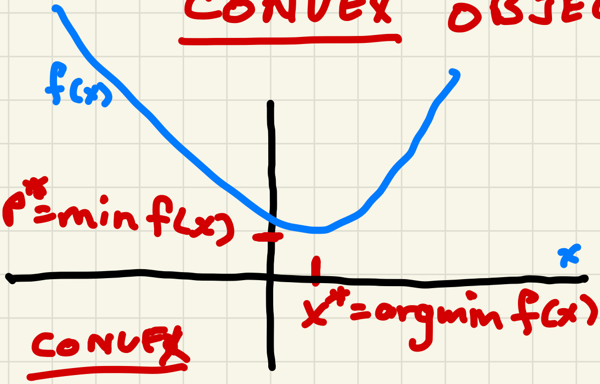
Wide Range of Problem Sizes & Algorithms!

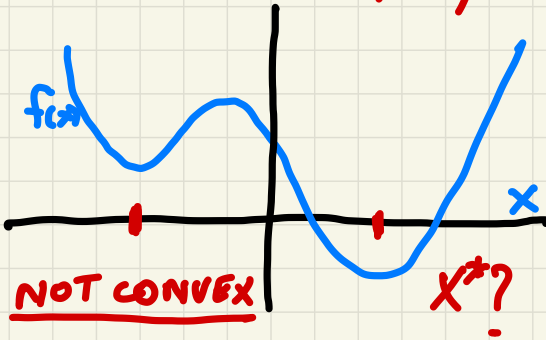
How many x's?

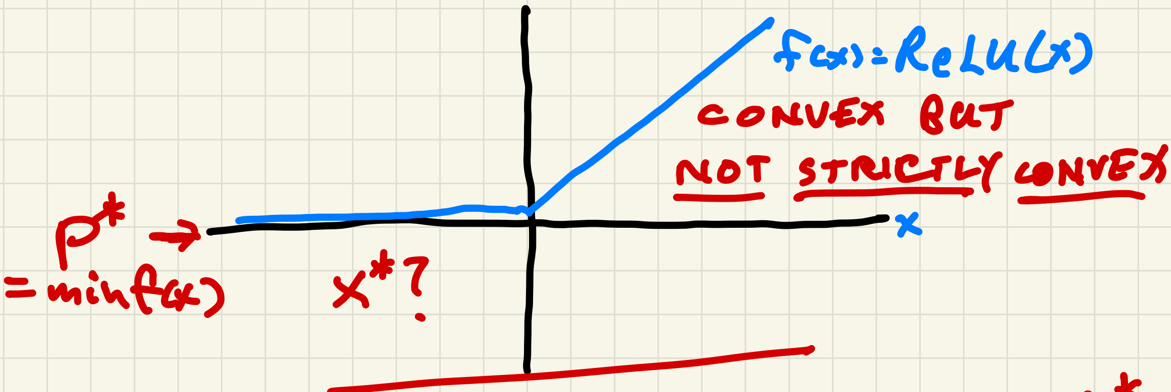
$1 - 100$	:	Newton's algorithm (dense)
$100 - 10^4$	:	Newton (with sparse matrices)
$10^4 - 10^6$	:	Quasi-Newton, e.g. "BFGS"
$10^6 - 10^8$	:	Gradient descent (with acceleration)
$10^8 - 10^{10++}$	:	Stochastic gradient descent (SGD)

(all very roughly speaking,
with exceptions & caveats)

CONVEX OBJECTIVE FUNCTION $f(x)$:







CONVEX $\Rightarrow$ UNIQUE MIN p^*

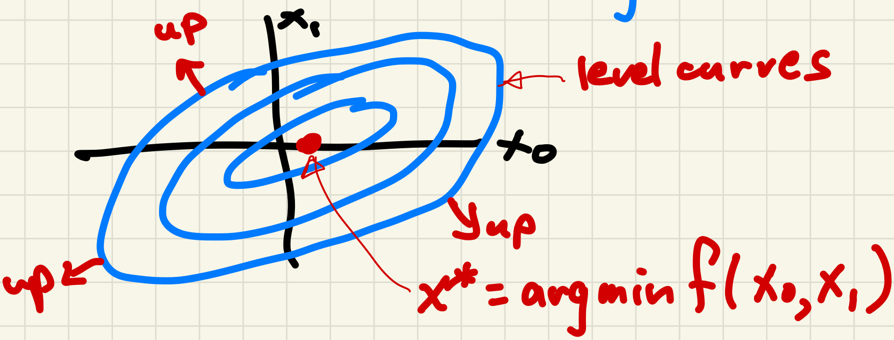
STRICTLY CONVEX $\Rightarrow$ UNIQUE ARGMIN x^*

<u>THM</u>: IF $f(x)$ and $g(x)$ are both convex, then so are:

$f(x) + g(x)$, $\max(f(x), g(x))$, $\alpha f(x)$ (for $\alpha \geq 0$).

What does "convex" mean in >1 dimension?

2-D: $f(x)$ is an "upward-pointing bowl":



n-D: EVERY CHORD OF f IS ABOVE THE FUNCTION VALUES.

DEFINITION:

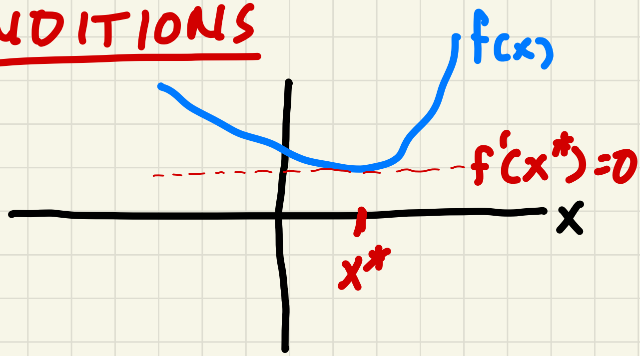
$f(x)$ is convex if for all $x, y \in \mathbb{R}^n$ and all α with $0 \le \alpha \le 1$,

$$f(\alpha x + (1-\alpha) y) \le \alpha f(x) + (1-\alpha) f(y).$$

("strictly convex": replace all $\le$ with $<$.)

OPTIMALITY CONDITIONS

In 1D, if f has a derivative then $f'(\text{argmin}) = 0$.



In n-D, $\dfrac{\partial f}{\partial x_i} = 0$ for $0 \le i < n$.

The first derivative of f is a vector in $\mathbb{R}^n$,

THE GRADIENT OF f:

$$\nabla f = \begin{bmatrix} \partial f / \partial x_0 \\ \partial f / \partial x_1 \\ \vdots \\ \partial f / \partial x_{n-1} \end{bmatrix}$$

$\nabla f(x)$ points uphill from x (length = steepness)

$-\nabla f$ points downhill (fall line)

∇f is perpendicular to the contour lines (level sets)

Example: Strang function:

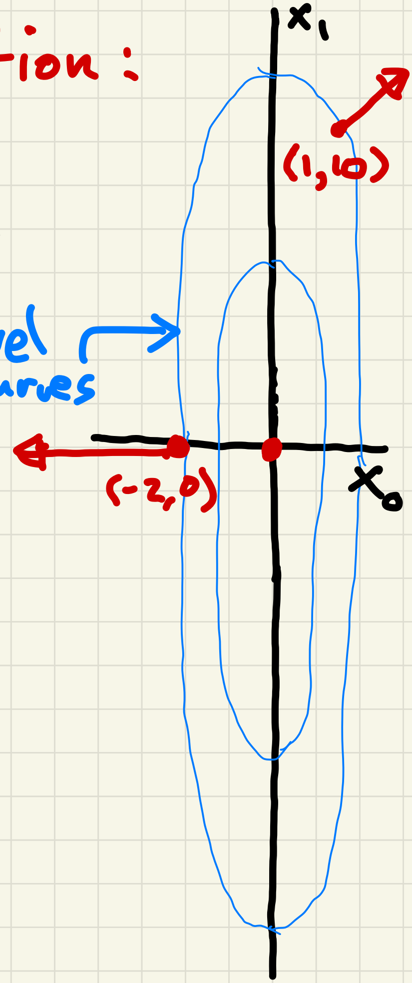
$$f(x) = (x_0^2 + \tfrac{1}{10}x_1^2)/2$$

$$\nabla f(x) = \begin{bmatrix} x_0 \\ x_1/10 \end{bmatrix}$$

At $x = \binom{1}{10}$, $\nabla f = \binom{1}{1}$

At $x = \binom{-2}{0}$, $\nabla f = \binom{-2}{0}$

At $x = x^* = \binom{0}{0}$, $\nabla f = \binom{0}{0}$



We are looking for x^* with $\nabla f(x^*) = 0$.

$$\left[\text{Usually we are more interested in } x^* = \operatorname{argmin} f(x) \text{ than in } p^* = \min f(x) = f(x^*)\right]$$