Introduction
oooo

Methods
ooooo

Convexity
oooooooooooooooooo

Optimization for Data Science

Axel Böhm

October 2, 2021

Introduction
oooo

Methods
ooooo

Convexity
oooooooooooooooo

1 Introduction

2 Methods

3 Convexity

Introduction
○○○○

Methods
○○○○○

Convexity
○○○○○○○○○○○○○○○○

Course organization

- ◇ Lectures (contribution counts)
- ◇ hands on sessions on some Thursdays
- ◇ a small weekly problem set
- ◇ Project (prices for most creative, best presentation, cleanest code, etc.)
- ◇ oral exam

Find everything on github (please contribute with pull requests: typos, etc.)

- ◇ Quick introductory round?

What is Optimization

Given a function f which represents some cost/regret/loss (or gain/profit/utility) we aim to find the argument/decision associated with the smallest cost (or largest profit).

$$\min_{x \in C} f(x)$$

◇ variables, parameters, candidate solutions x

◇ objective function f (typically real-valued)

◇ typically: technical assumptions on f

◇ constrained set $C \subset \mathbb{R}^d$

◇ convexity / differentiability

Applications of optimization

- ◇ Economics
  - ▶ Microeconomics: Agents maximizing utility
  - ▶ Game theory and equilibria
- ◇ Statistics
  - ▶ maximum likelihood
- ◇ Physics
  - ▶ soap bubble is a sphere because it minimizes surface tension
- ◇ Chemistry
  - ▶ Protein folding
- ◇ Inverse problems
  - ▶ imaging, denoising, deblurring

Introduction
○○●○

Methods
○○○○○

Convexity
○○○○○○○○○○○○○○○○

Optimization for ML

$$\min_{\beta_1,\beta_0} \sum_{i=1}^{n} (\beta_1 x_i + \beta_0 - y_i)^2$$

For data points (x_i, y_i).
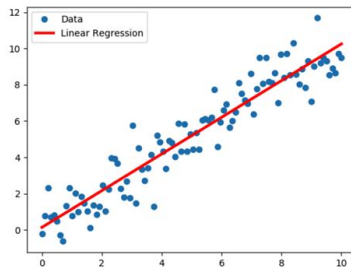
Optimization for ML

◇ Mathematical modeling
▶ defining & modeling the problem
▶ finding a good metric / what is success
▶ accuracy vs. solvability trade-off

◇ Computational optimization
▶ running an (appropriate) optimization algorithm

◇ theory vs. practice
▶ libraries available, but algorithms treated as "black box" by practitioners
▶ we will try and understand why and how they work

Introduction
○○○○

Methods
●○○○○

Convexity
○○○○○○○○○○○○○○○○

Optimization Algorithms

*Simplicity* rules in the large scale setting.

Main approaches:

- ◇ First order methods: **gradient descent**
- ◇ Stochastic gradient descent (SGD)
- ◇ Coordinate descent

History

- ◇ 1847: Cauchy proposes gradient descent
- ◇ 1950s: Linear programming, operations research, soon followed by nonlinear
- ◇ 1980s: general convergence theory
- ◇ 2005-today: large scale optimization, SGD, distributed optimization

Introduction
oooo

Methods
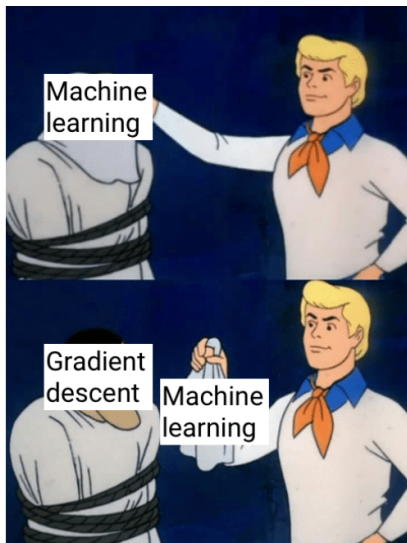○●○○○

Convexity
○○○○○○○○○○○○○○○○○

Example: Coordinate descent
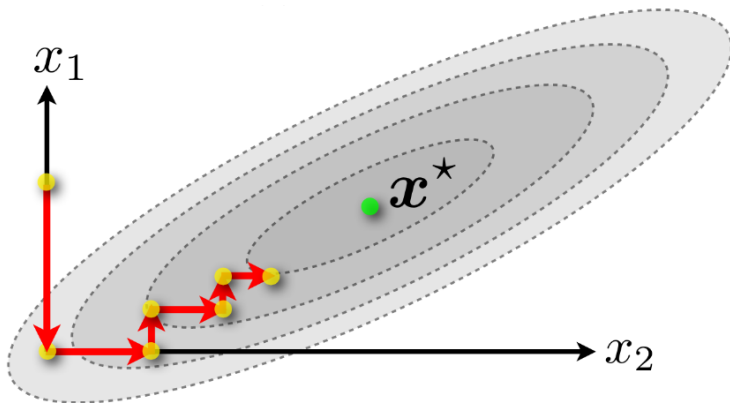
Strategy: Minimize along one coordinate at a time, while keeping the others fixed.

Introduction
0000

Methods
00●00

Convexity
0000000000000000

Example: Gradient descent
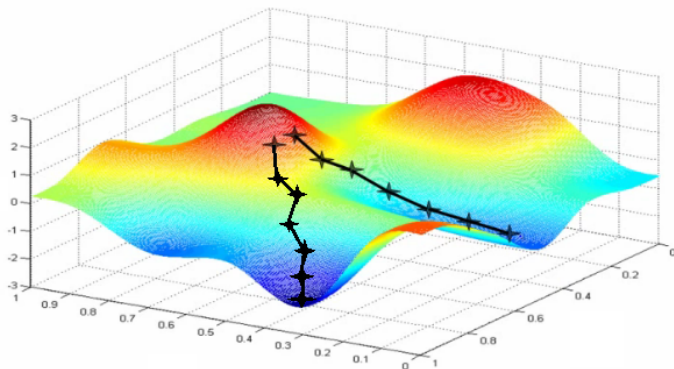
Strategy: Follow the direction of (local) **steepest descent**.

Introduction
○○○○

Methods
○○○●○

Convexity
○○○○○○○○○○○○○○○○

Machine learning behind the scenes

Introduction
oooo

Methods
ooooo●

Convexity
ooooooooooooooooo

Optimization in other settings

◇ Second order
- ▶ if high precision in solution is required
- ▶ too **expensive** in high dimensions

◇ Zeroth order
- ▶ no gradient or functional representation available
- ▶ only function values
- ▶ for simulation, hyperparameters, black box models

◇ constrained problems

◇ discrete optimization
- ▶ involving graphs, traveling salesman
- ▶ scheduling

Introduction
oooo

Methods
ooooo

Convexity
●oooooooooooooooo

Convex sets

A set C is convex if the line segment between any two points remains inside C, i.e. for any $x, y \in C$ and $\lambda \in [0, 1]$.

$$\lambda \mathbf{x} + (1 - \lambda)\mathbf{y} \in C.$$
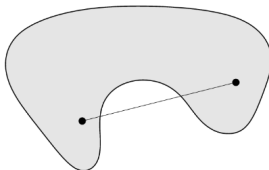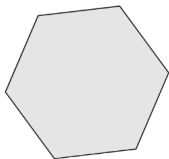
*Figure 2.2 from S. Boyd, L. Vandenberghe

Which of these sets are convex?

Introduction
oooo

Methods
ooooo

Convexity
○●○○○○○○○○○○○○○○

Properties of convex sets

- ◇ intersection remains convex
- ◇ can separated by a hyperplane
- ◇ projections onto them are unique

Introduction
oooo
Methods
ooooo
Convexity
ooo●oooooooooooooo

Convex functions

We call a function $f \to \mathbb{R} \cup \{+\infty\}$ convex if the function values lie below the line segment between $(x, f(x))$ and $(y, f(y))$

$$f(\lambda \mathbf{x} + (1 - \lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1 - \lambda)f(\mathbf{y}).$$

$(y, f(y))$

$(x, f(x))$

*Figure 3.1 from S. Boyd, L. Vandenberghe

Sometimes we will call $\{x : f(x) < +\infty\}$ the domain of f.

Introduction
0000

Methods
00000

Convexity
0000000000000000

Convex optimization

Are of the form

$$\min_x f(x) \quad \text{such that } x \in C$$

where both

$\diamond$ f is a convex function

$\diamond$ C is a convex set

Why?

$\diamond$ Every local minimum is a global minimum

$\diamond$ Not all problems are convex but can be used as approximate model

Introduction
oooo

Methods
ooooo

Convexity
oooo●oooooooooooo

Examples of convex functions

◇ linear: $f(x) = a^T x$

◇ affine: $f(x) = a^T x + b$

◇ exponential: $f(x) = e^{\alpha x}$

◇ norms, $\| \cdot \|_1, \| \cdot \|_2, \| \cdot \|_\infty$ show this

◇ composition of linear and convex:
  for example $f(x) = \|Ax - b\|^2$

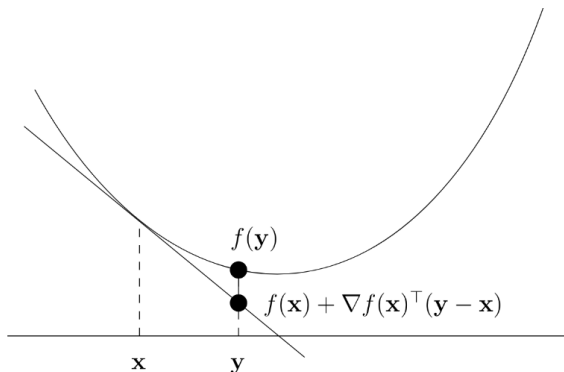Introduction
oooo

Methods
ooooo

Convexity
oooooo●oooooooooo

Differentiable function

Derivative at a point is the best linear approximation of the function at this point.
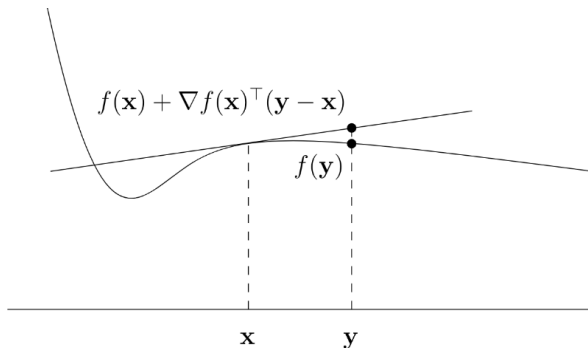
Graph of $f(x) + \nabla f(x)^T(y - x)$ is a tangent hyperplane to the graph of f at $(x, f(x))$
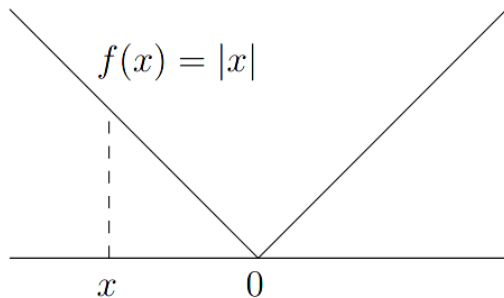
First-order characterization of convexity

If f is differentiable, then

f is convex *if and only if:* $f(y) \geq f(x) + \nabla f(x)^T (y - x)$

Introduction
oooo

Methods
ooooo

Convexity
oooooooo●oooooooo

Nonsmooth functions

do in fact play a role in practice

⋄ ReLu, Hinge loss, norms

⋄ can induce sparsity in the solution

⋄ appear as the maximum over a family of functions (max pooling, or min-max)

Second-order characterization of convexity

If f is twice differentiable then it is convex **if and only if** its
Hessian $\nabla^2 f(x) \mathbb{R}^d$, given by

$$\nabla^2 f(x)_{ij} := \frac{\partial^2 f}{\partial x_i \partial x_j}$$

is positive semidefinite, i.e.

$$\nabla^2 f(x) \succeq 0$$

A matrix M is *positive semidefinite* if $x^T M x \geq 0$ for all x.
Also used in algorithm like *Newtons* method.

Introduction
0000

Methods
00000

Convexity
00000000000●000000

Examples

$\diamond$ quadratic function: $f(x) = \frac{1}{2}x^T Q x + c^T x$, then

$$\nabla^2 f(x) = Q$$

and f is convex iff $Q \succeq 0$.

$\diamond$ least squares objective: $f(x) = \|Ax - b\|^2$, then

$$\nabla^2 f(x) = A^T A$$

is always convex for any A.

Introduction
oooo

Methods
ooooo

Convexity
ooooooooooo●ooooo

Local minima are global

Definition

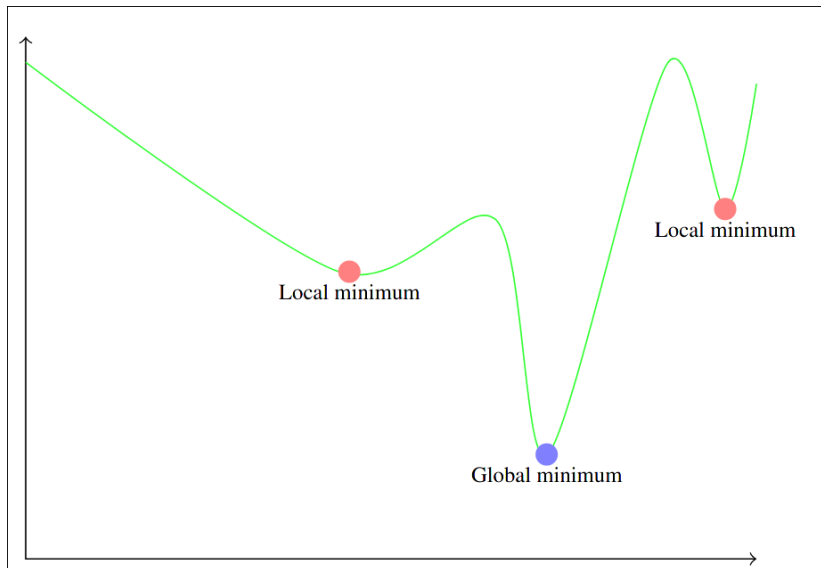A local minimum of f is a point $\bar{x}$ such that there exists $\epsilon > 0$

$$f(\bar{x}) \leq f(y) \quad \forall y : \text{s.t.} \|\bar{x} - y\| \leq \epsilon$$

Lemma

Let x^* be local minimum of a convex function f then x^* is a global minimum.

Prove this!

Introduction
oooo

Methods
ooooo

Convexity
ooooooooooooo●oooo

Local vs. global minima

Introduction
oooo

Methods
ooooo

Convexity
oooooooooooo●ooo

Critical points are global minima

Definition

We call a point $\bar{x}$ critical or stationary if $\nabla f(\bar{x}) = 0$.

Lemma

If $\bar{x}$ is a stationary point of the **convex** function f, then $\bar{x}$ is a global minimizer of f.

Prove this and give a geometric intuition in words using the first order characterization of convexity

Introduction
oooo

Methods
ooooo

Convexity
ooooooooooooo●oo

Strong convexity

Definition

We call f strongly convex if there exist $\mu > 0$ such that

$$f - \frac{\mu}{2}\|\cdot\|^2 \quad \text{is convex.}$$

Equivalently:

◇ can be lower bounded by a quadratic

$$f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2 \le f(y)$$

◇ Hessian is pos. def. everywhere

$$\nabla^2 f(x) \succ 0.$$

Introduction
oooo

Methods
ooooo

Convexity
ooooooooooooooo●o

Constrained minimization
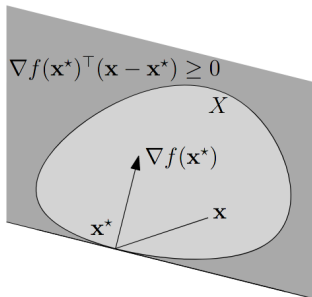
Definition

x^* is a minimizer of f over C if

$$f(x^*) \leq f(x), \forall x \in C$$

Lemma

x^* is a minimizer of f over C if and only if

$$\langle \nabla f(x^*), x - x^* \rangle \geq 0 \quad \forall x \in C$$

$\nabla f(\mathbf{x}^\star)^\top(\mathbf{x} - \mathbf{x}^\star) \geq 0$

X

$\nabla f(\mathbf{x}^\star)$

$\mathbf{x}^\star$

$\mathbf{x}$

Introduction
0000

Methods
00000

Convexity
00000000000000●

Proof

x^* is a minimizer of f over C *if and only if*

$$\langle \nabla f(x^*), x - x^* \rangle \geq 0 \quad \forall x \in C$$

Assume that $f(x^*) \leq f(x)$ forall $x \in C$ then $\forall t \in [0, 1]$

$$
\begin{aligned}
0 &\leq f(x^* + t(x - x^*)) - f(x^*) \\
0 &\leq \lim_{t \to 0} \frac{f(x^* + t(x - x^*)) - f(x^*)}{t} \\
&= \langle \nabla f(x^*), x - x^* \rangle.
\end{aligned}
$$

where the last equality follows from the chain rule.