Optimization for Data Science

Axel Böhm

October 4, 2021

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

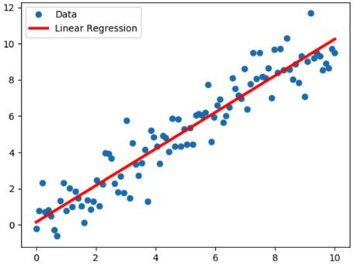
Applications of optimization

- 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

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).



◇ Loss functions express the discrepancy between the predictions of the model being trained and the actual problem instances

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

Optimization Algorithms

Simplicity rules in the large scale setting.

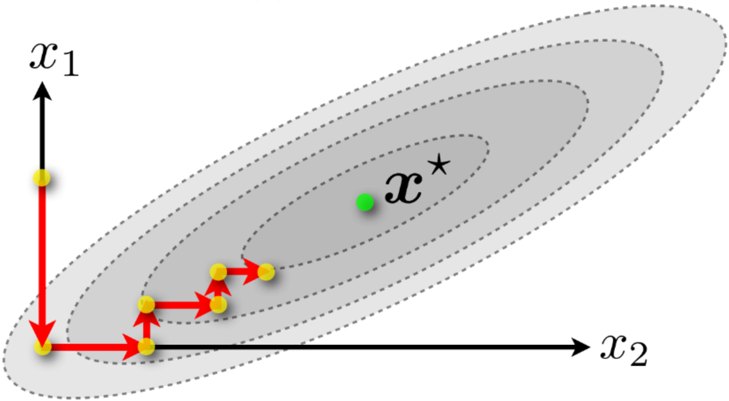
Main approaches:

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

History

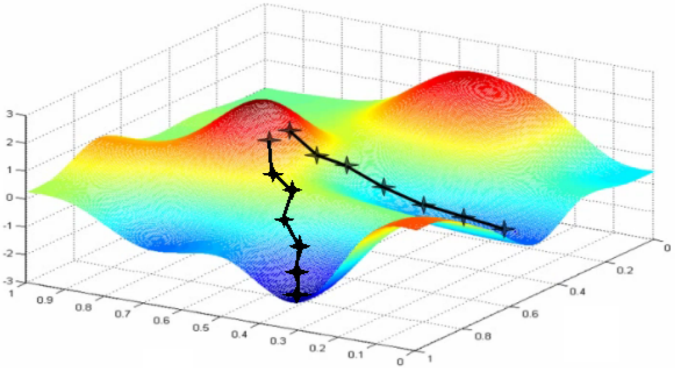
- 1847: Cauchy proposes gradient descent
- 1950s: Linear programming, operations research, soon followed by nonlinear
- 1980s: general convergence theory
- 2005-today: large scale optimization, SGD, distributed optimization

Example: Coordinate descent

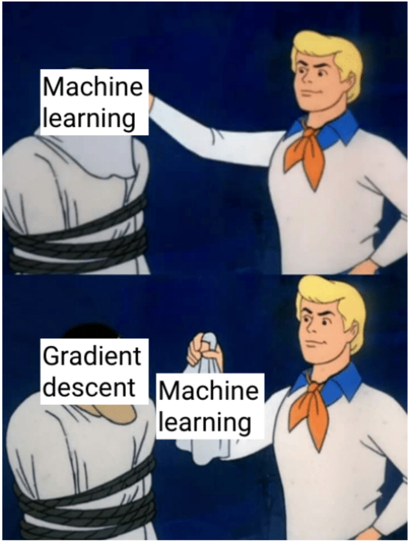


Strategy: Minimize along one coordinate at a time, while keeping the others fixed.

Example: Gradient descent



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



Machine learning behind the scenes

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

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

$$\lambda \mathbf{x} + (1 - \lambda)\mathbf{y} \in C.$$

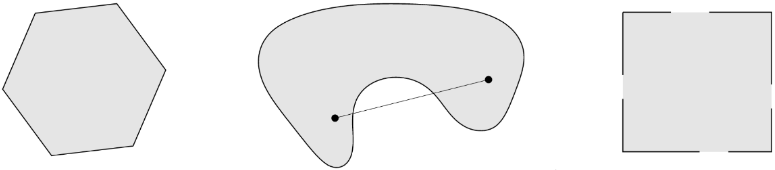


*Figure 2.2 from S. Boyd, L. Vandenberghe

Which of these sets are convex?

Properties of convex sets

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

$$P_C(x) := \arg\min_{y \in C} \|y - x\|$$

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 $(y, f(y))$, i.e./ for any $\lambda \in [0, 1]$

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

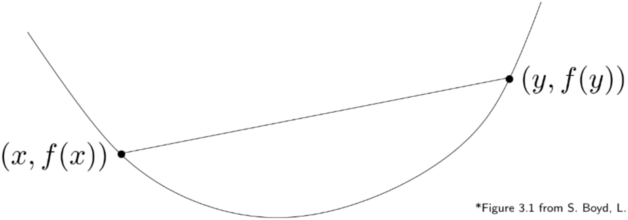


*Figure 3.1 from S. Boyd, L. Vandenberghe

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

Motivation: Convex optimization

Are of the form

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

where both

- f is a convex function
- C is a convex set

Why?

- *Every local minimum is a global minimum.*
- Not all problems are convex but can be used as approximate model.

Motivation: Provably (efficiently) solving convex problems

For convex optimization problems, basically all algorithms

- Coordinate Descent, (Stochastic) Gradient Descent, Proj. GD

converge provably to a global optimum including a

- **quantitative bound**.

Example Theorem

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex then the **convergence rate** is proportional to $1/k$, i.e.

$$f(x_k) - f(x^*) \leq \frac{c}{k}$$

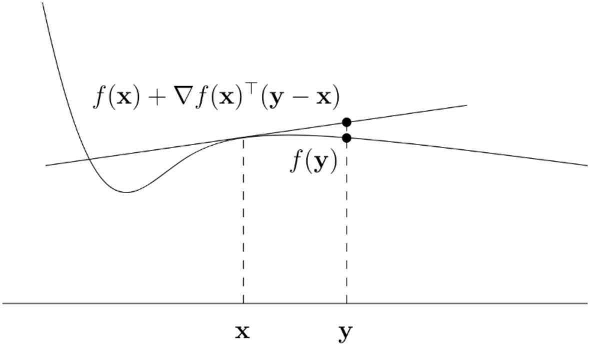
Explanation: The approximation error converges to zero and we know how many iterations are needed to achieve given target.

Examples of convex functions

- linear: $f(x) = a^T x$
- affine: $f(x) = a^T x + b$
- exponential: $f(x) = e^{\alpha x}$
- norms, $\|\cdot\|_1, \|\cdot\|_2, \|\cdot\|_\infty$ See ex. 1.(i)
- composition of linear and convex: See ex. 1.(ii)
  for example $f(x) = \|Ax - b\|^2$
- sum of two convex function $f + g$

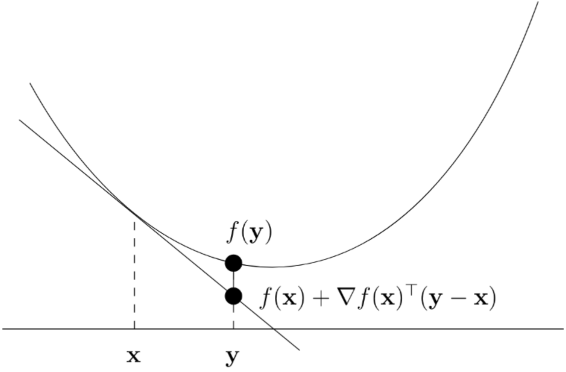
Differentiable function

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



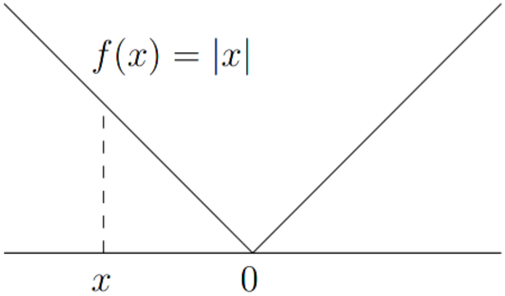
If f is differentiable, then

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

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* $\nabla^2 f(x) \in \mathbb{R}^{d\times d}$, given by

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

is positive semidefinite, i.e.

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

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

Examples

- quadratic function: $f(x) = \frac{1}{2}x^TQx + c^Tx$, then

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

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

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

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

is always convex for any A.

Local minima are global

Definition

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

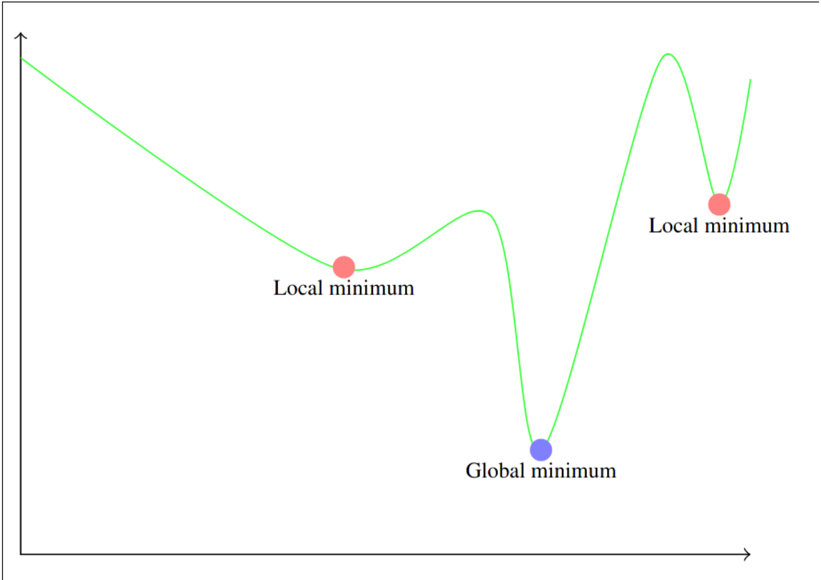
$$f(\bar{x}) \leq f(y) \quad \forall y : \text{s.t.} \|\bar{x} - y\| \leq \epsilon$$

Lemma

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

See ex 1.(iii)

Local vs. global minima



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.*

See ex. 1.(iv)

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

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

- Hessian is pos. def. everywhere

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

Constrained minimization

Definition

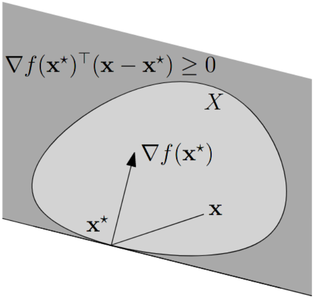
x^* is a minimizer of f over C if

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

Lemma

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

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



Lemma

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

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

"$\Leftarrow$" From the gradient inequality he deduce

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

"$\Rightarrow$" Assume that $f(x^*) \leq f(x)$ for all $x \in C$ 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. □

Lemma

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

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

"$\Leftarrow$" From the gradient inequality he deduce

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

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

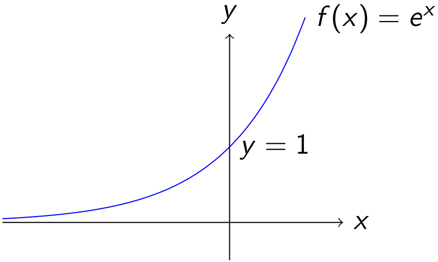
$$\begin{aligned} 0 &\leq f(x^* + t(x - x^*)) - f(x^*) \\ 0 &\leq \lim_{t \to 0} \frac{f(x^* + t(x - x^*)) - f(x^*)}{t} \\ &= \langle \nabla f(x^*), x - x^* \rangle. \end{aligned}$$

where the last equality follows from the chain rule. $\square$

Existence of a minimizer

In general a minimizer *does not need to exist*.

- can be unbounded from below (linear)
- bounded but infimum is not obtained



Typically we only consider problems where we assume a minimizer to exist (otherwise our model might be bad).

- if function is strongly convex a minimizer *always* exists.