Computational Data Analysis

Machine Learning

Yao Xie, Ph.D.

Associate Professor

Harold R. and Mary Anne Nash Early Career Professor

H. Milton Stewart School of Industrial and Systems Engineering

Basic Optimization

Outline

- Concept
- Convex/concave function
- First-order condition
- Second-order condition
- Langragian and dual function
- KKT conditions

Motivation

We have seen quite a few optimization problems in machine learning already

Clustering:

$$\min_{c,\pi} \frac{1}{m} \sum_{i=1}^{m} \left\| x^i - c^{\pi(i)} \right\|^2$$

PCA:

$$\max_{w:\|w\|\leq 1} \frac{1}{m} \sum_{i=1}^{m} \left(w^\top x^i - w^\top \mu \right)^2$$

MLE:

$$\theta = argmax_\theta \log P(\mathcal{D}|\theta) = argmax_\theta \log \prod_{i=1}^{m} P(x^i|\theta)$$

Motivation

Optimization is foundational to ML

- Modeling (together with statistics and geometric models)
  - Clustering, PCA, MLE, regression, support vector machine, deep learning, ….
- Solutions methods
  - Heuristics (e.g., k-means)
  - Special methods (e.g., PCA, EM)
  - General methods (e.g., gradient descent, SGD)

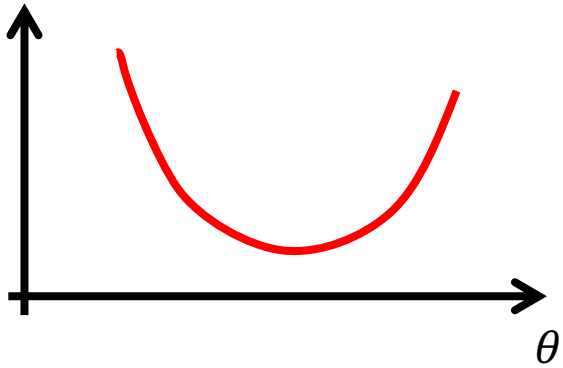
Three components

- Decision variables
- Constraints
- Objective function

Main watershed in optimization

- Convex/Nonconvex optimization (different strategy for solving them)

Convex vs. non-convex



Convex problem

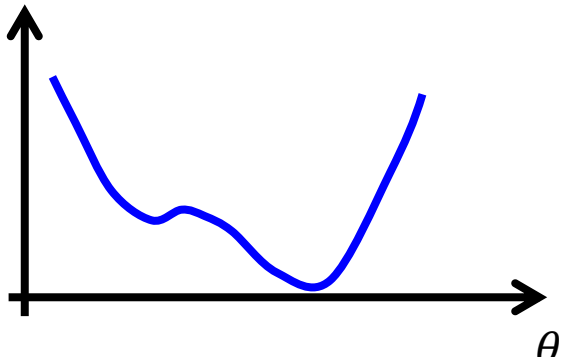
- e.g. linear regression
- Can be solved efficiently (find global optimal solution) in polynomial time
- Gradient descent (or many acceleration methods e.g., Newton method) can converge to global solution

Non-convex problem

- e.g. clustering, maximum likelihood for GMM
- Cannot find global solution in polynomial time

(NP-hard)

- Use heuristics to find local optimal solution



Optimization

- Definition: An optimization problem is specified by

$$\begin{aligned} &\text{minimize} && f_0(x) \\ &\text{subject to} && f_i(x) \leq 0,\ i = 1, \ldots, m \\ &&& h_i(x) = 0,\ i = 1, \ldots, p \end{aligned}$$

- A convex optimization problem has the following requirements
  - The objective function $f_0(x)$ must be convex
  - The inequality constraint functions $f_i(x)$ must be convex
  - The equality constraint functions $h_i(x)$ must be affine

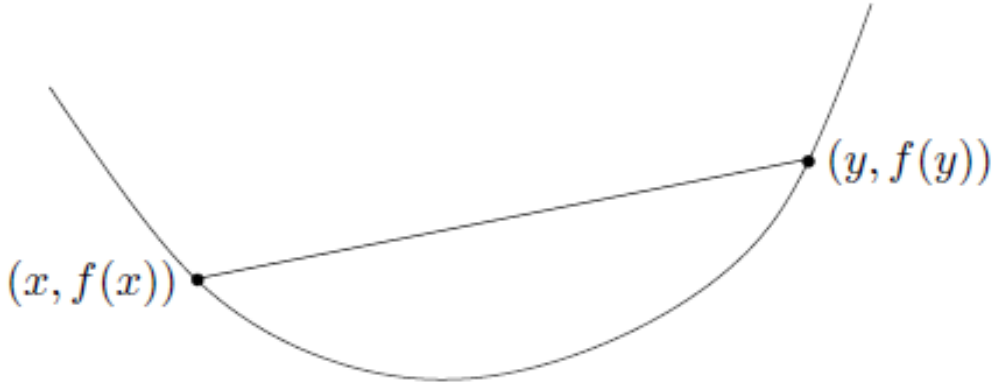
- Eg. support vector machines (SVM), logistic regression, maximum likelihood, ridge regression, …

Convex Functions

- Definition: A function $f: R^n \rightarrow R$ is convex if the domain $\mathbf{dom} f$ is a convex set and if for all $x, y \in \mathbf{dom} f$, and $0 \leq \theta \leq 1$, we have

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

- Geometrically, the line segment between $(x, f(x))$ and $(y, f(y))$ lies above the graph of f

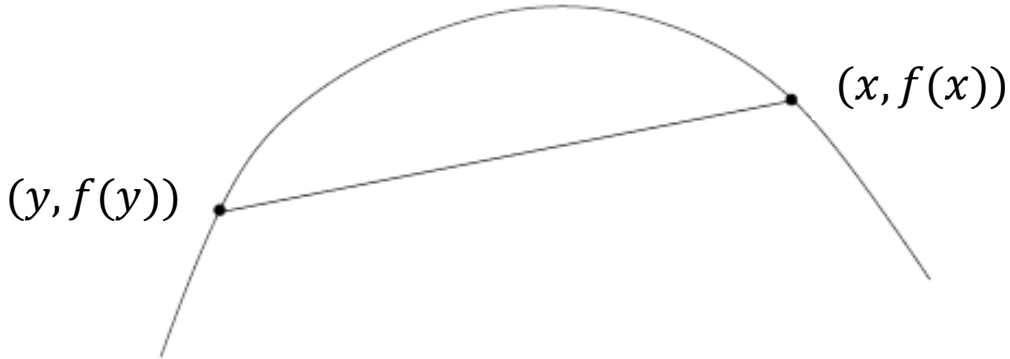


Concave Functions

- Definition: A function $f: R^n \rightarrow R$ is concave if the domain $\mathbf{dom} f$ is a convex set and if for all $x, y \in \mathbf{dom} f$, and $0 \leq \theta \leq 1$, we have

$$f(\theta x + (1-\theta)y) \geq \theta f(x) + (1-\theta) f(y)$$

- Geometrically, the line segment between $(x, f(x))$ and $(y, f(y))$ lies below the graph of f



Examples

- Exponential: e^{ax} for every $a \in R$
- Powers: x^a is convex on R_{++} when $a \geq 1$ or $a \leq 0$; concave (i.e., $-f$ is convex) for $0 \leq a \leq 1$
- Powers of absolute vaue: $|x|^p$ for $p \geq 1$
- Logarithm: $\log x$ is concave on R_{++}
- Negative entropy: $x \log x$ is convex
- Norms: All norms are convex (nonnegative; homogeneous; triangular inequality)
- Max function: $f(x) = \max\{x_1, \ldots, x_n\}$ is convex
- Log-determinant: $f(X) = \log \det X$ is convex for all positive definite matrices

Used in EM

Used in multivariate Gaussian fit

Operations that Preserve Convexity

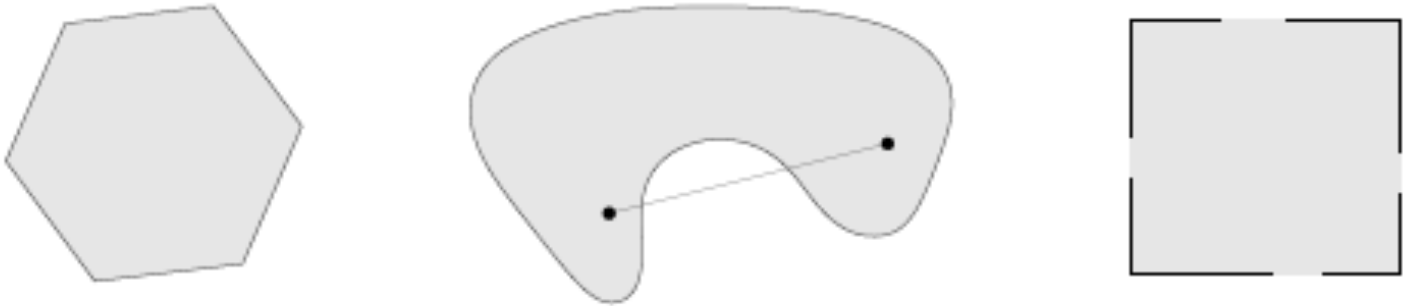
- Nonnegative weighted sums: If $f_1, \dots, f_m$ are convex, and $w_1, \dots, w_m \geq 0$, then
$$f = w_1 f_1 + \cdots + w_m f_m$$
is convex

- Composition with an affine mapping: suppose f is convex, then
$$g(x) = f(Ax + b)$$
with $\mathbf{dom} g = \{x | Ax + b \in \mathbf{dom} f\}$ is convex

- Pointwise maximum and supremum: If f_1 and f_2 are convex, then $f(x) = \max\{f_1, f_2\}$ is also convex. It easily extends to multiple functions.

Convex Set

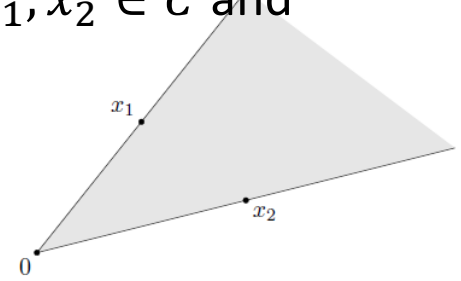
- Definition: A set A is convex, if for every $0 \leq \alpha \leq 1$ it satisfies
  - $\forall x,\ y \in A \rightarrow \alpha x + (1-\alpha)y \in A$
- The line segment between any two points is also in the set.
- Examples of convex and non-convex sets



Common Convex Set

- Cones: A set C is a convex cone, if for any $x_1, x_2 \in C$ and $\theta_1, \theta_2 \geq 0$, we have

$$\theta_1 x_1 + \theta_2 x_2 \in C$$

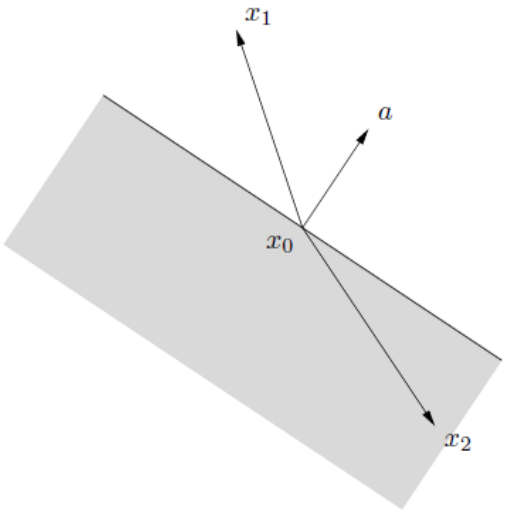


- Hyperplanes and halfspaces:
  A set is hyperplane if

$$\{x | a^\top (x - x_0) = 0, a \neq 0\}$$

  A halfspace is

$$\{x | a^\top (x - x_0) \leq 0, a \neq 0\}$$



Common Convex Set

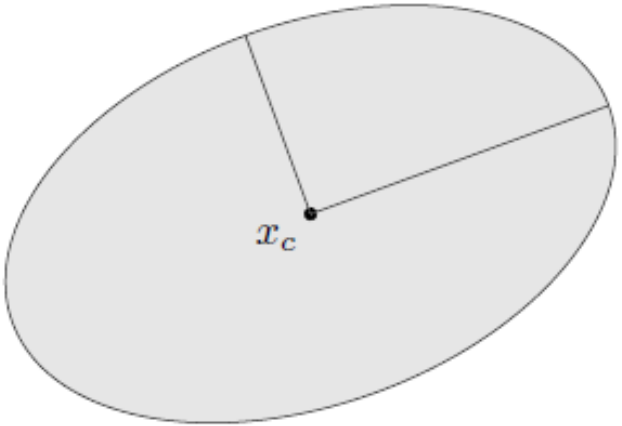
- Euclidean balls: A Euclidean ball has the form

$$B(x_c, r) = \{x | \|x - x_c\|_2 \leq r\}$$

- Ellipsoids:

$$E = \{x | (x - x_c)^\top P^{-1} (x - x_c) \leq 1\}$$

  - The eigen-vectors and eigen-values determine the direction and shape of the semi-axes



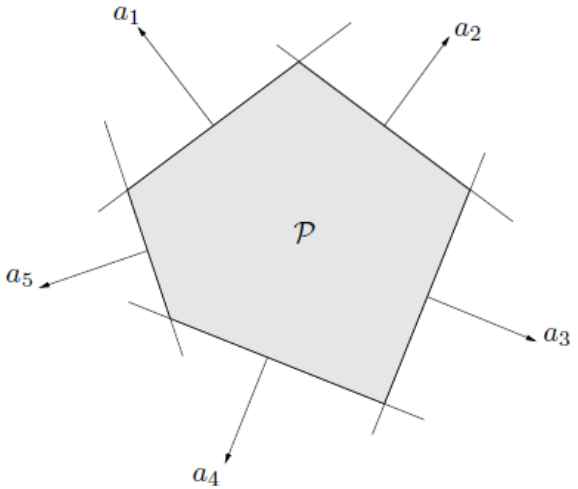
Used in support vector novelty detection

Common Convex Set

- Polyhedra: Intersection of a *finite* set of halfspaces/hyperplanes

$$P = \{x | a_j^\top x \leq b_j, j = 1, \ldots, m, c_j^\top x = d_j, j = 1, \ldots, p\}$$

- It is defined by as the solution set of a finite number of linear equalities and inequalities



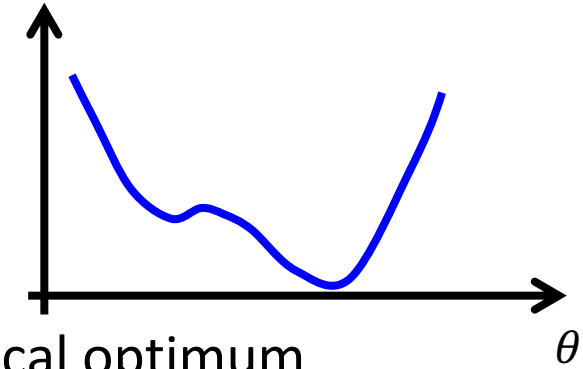
Used in SVM

Global vs. local optimum

- Global optimum: a point x^* in the feasible set is a global optimum iff

$$f_0(x^*) \leq f_0(x)$$

for all x in the feasible set

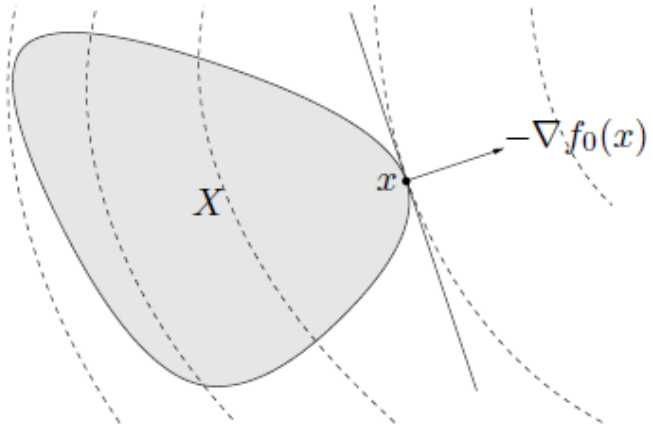


- Local optimum: a point x^* in the feasible set is a local optimum iff there exists $r > 0$, such that for all $x \in \{x | \|x - x^*\| \leq r\}$ and also in the feasible set, we have $f_0(x^*) \leq f_0(x)$

- For convex optimization problem, any local optimum is also a global optimum

First order optimality condition

- Let X denotes the feasible set, then x is optimal iff
$$\nabla f_0(x)^\top (y - x) \geq 0 \text{ for all } y \in X$$
- For an unconstrained problem, the condition becomes
$$\nabla f_0(x) = 0$$
- Geometrically, if $\nabla f_0(x) \neq 0$, it means $-\nabla f_0(x)$ is orthogonal to the feasible set at x



Constrained vs. unconstrained

For an unconstrained problem, the condition becomes

$$\nabla f_0(x) = 0$$

For constrained problem, we need to use the Lagrangian

$$L(x, \mu, \lambda) = f_0(x) + \sum_{i=1}^{p} \mu_i h_i(x) + \sum_{i=1}^{m} \lambda_i f_i(x)$$

$$\text{s.t.} \quad \lambda_i \geq 0$$

to transform it into an unconstrained problem

It is a lower bound of $f_0(x)$ for all $x \in X$, since $h_i(x)=0$, $f_i(x) \leq 0$ and $\lambda_i \geq 0$

$$L(x, \mu, \lambda) \leq f_0(x) \text{ for all } x \in X$$

Lagrange dual function

- The Lagrange dual function is

$$g(\mu, \lambda) = \inf_x L(x, \mu, \lambda)$$

- It is a lower bound for the optimal value

$$g(\mu, \lambda) = \inf_x L(x, \mu, \lambda) \leq L(x^*, \mu, \lambda) \leq f_0(x^*)$$

- We want to maximize the lower bound to make it tight

$$g(\mu^*, \lambda^*) = \max g(\mu, \lambda)$$

Primal and Dual problems

- Primal problem

$$\begin{aligned} &\text{minimize} && f_0(x) \\ &\text{subject to} && f_i(x) \leq 0, \ i = 1, \dots, m \\ &&& h_i(x) = 0, \ i = 1, \dots, p \end{aligned}$$

- Dual problem

$$\begin{aligned} &\text{maximize} && g(\mu, \lambda) \\ &\text{subject to} && \lambda_i \geq 0, \ i = 1, \dots, m \end{aligned}$$

- Strong duality (for convex problems, with mild condition)

$$g(\mu^*, \lambda^*) = f_0(x^*)$$

- Slater's condition: There exists an x inside the relative interior of the domain X such that, $f_i(x) < 0, \ i = 1, \dots, m$

KKT Optimality conditions

The following list of conditions for an optimal triplet (x^*, μ^*, λ^*), are called KKT conditions

- $\nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^{p} \mu_i^* \nabla h_i(x^*) = 0$
- $\lambda_i^* f_i(x^*) = 0$ (complementarity condition)
- $f_i(x^*) \leq 0$
- $h_i(x^*) = 0$
- $\lambda_i^* \geq 0$


SUMMARY