Convex Optimization for Statistics and Machine Learning

Part I: Basic Convex Analysis

Ryan Tibshirani

Depts. of Statistics & Machine Learning
Carnegie Mellon University

`http://www.stat.cmu.edu/~ryantibs/talks/cuso-part1-2019.pdf`

Optimization in Statistics and Machine Learning

Optimization underlies almost everything we do in Statistics and Machine Learning. In many settings, you learn how to:

translate into $P \ : \ \min_{x \in D} \ f(x)$

Conceptual idea *Optimization problem*

Examples of this? Examples of the contrary?

This course: how to solve P, and why this is a good skill to have

Motivation: why do we bother?

Presumably, other people have already figured out how to solve

$$P \ : \ \min_{x \in D} \ f(x)$$

So why bother? Many reasons. Here's three:

1. Different algorithms can perform better or worse for different problems P (sometimes drastically so)
2. Studying P through an optimization lens can actually give you a deeper understanding of the task/procedure at hand
3. Knowledge of optimization can actually help you create a new problem P that is even more interesting/useful

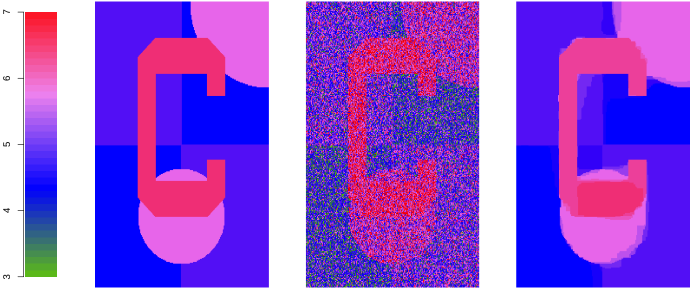
Optimization moves quickly as a field. But there is still much room for progress, especially its intersection with ML and Stats

Example: algorithms for the 2d fused lasso

The 2d fused lasso or 2d total variation denoising problem:

$$\min_{\theta} \ \frac{1}{2}\sum_{i=1}^{n}(y_i-\theta_i)^2+\lambda\sum_{(i,j)\in E}|\theta_i-\theta_j|$$

This fits a piecewise constant function over an image, given data y_i, $i=1,\ldots,n$ at pixels. Here $\lambda \geq 0$ is a tuning parameter

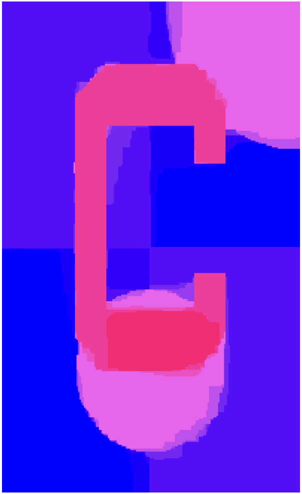


True image

Data

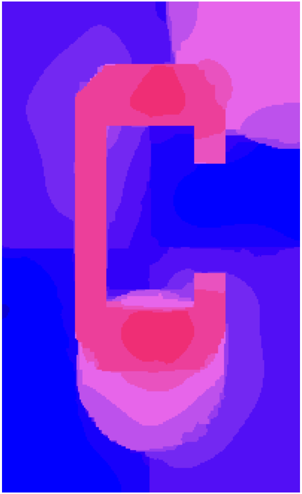
Solution

Our problem: $$\min_{\theta} \frac{1}{2}\sum_{i=1}^{n}(y_i-\theta_i)^2 + \lambda \sum_{(i,j)\in E} |\theta_i - \theta_j|$$



Specialized ADMM, 20 iterations

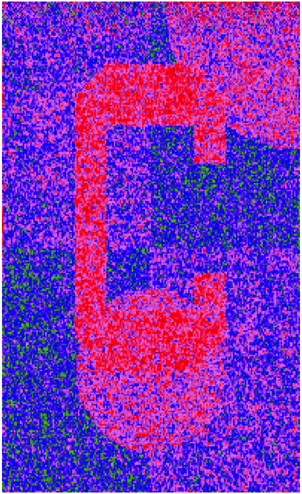
Our problem: $$\min_{\theta} \frac{1}{2}\sum_{i=1}^{n}(y_i-\theta_i)^2+\lambda\sum_{(i,j)\in E}|\theta_i-\theta_j|$$



Specialized ADMM, 20 iterations

Proximal gradient descent, 1000 iterations

Our problem: $$\min_{\theta} \frac{1}{2}\sum_{i=1}^{n}(y_i - \theta_i)^2 + \lambda \sum_{(i,j)\in E} |\theta_i - \theta_j|$$

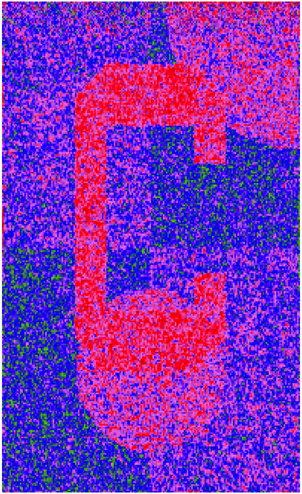


Specialized ADMM, 20 iterations

Proximal gradient descent, 1000 iterations

Coordinate descent, 10K cycles

Our problem: $$\min_{\theta} \frac{1}{2}\sum_{i=1}^{n}(y_i - \theta_i)^2 + \lambda \sum_{(i,j)\in E} |\theta_i - \theta_j|$$



Specialized ADMM, 20 iterations

Proximal gradient descent, 1000 iterations

Coordinate descent, 10K cycles

(Last two from the dual)

What's the message here?

So what's the right conclusion here?

Is the alternating direction method of multipliers (ADMM) method simply a better method than proximal gradient descent, coordinate descent? ... No

In fact, different algorithms will perform better or worse in different situations. We'll learn details later

In the 2d fused lasso problem:

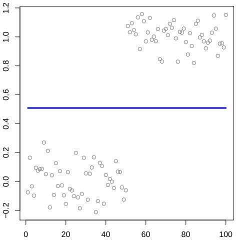
- Special ADMM: fast (structured subproblems)
- Proximal gradient: slow (poor conditioning)
- Coordinate descent: slow (large active set)

Example: changepoints in the 1d fused lasso

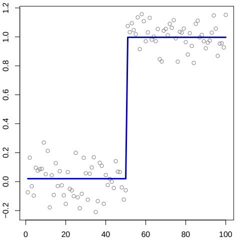
The 1d fused lasso or 1d total variation denoising problem:

$$\min_{\theta} \ \frac{1}{2}\sum_{i=1}^{n}(y_i-\theta_i)^2+\lambda\sum_{i=1}^{n-1}|\theta_i-\theta_{i+1}|$$

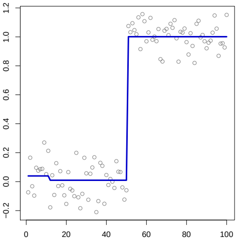
Again here $\lambda \geq 0$ is a tuning parameter. As λ decreases, we see more changepoints in the solution $\hat{\theta}$



$\lambda = 25$

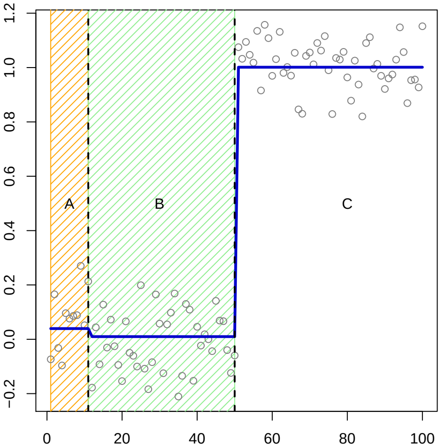


$\lambda = 0.62$



$\lambda = 0.41$

Let's look at the solution at $\lambda = 0.41$ a little more closely

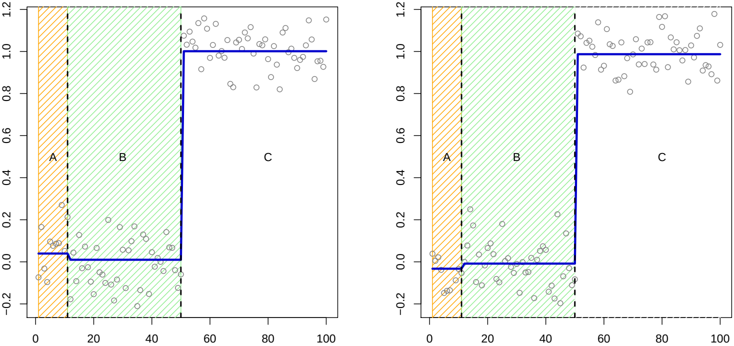


How can we test the significance of detected changepoints? Say, at location 11?

Classically (z-test): average the data in region A minus the average in B, compare this to what we expect if the signal was flat

But this is incorrect, because location 11 was selected based on the data, so of course the difference in averages looks high!

What we want: compare our observed difference to that in proper null data, where the signal was flat and we happen to select same location 11 (and 50)

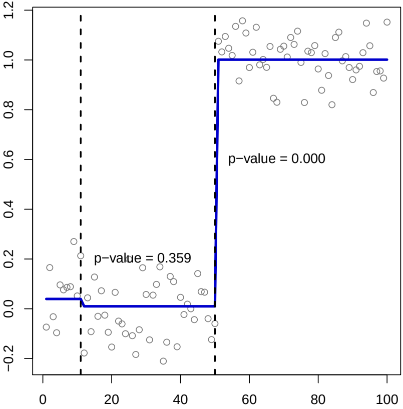


Observed data
Test stat ≈ 0.088

Null data
Test stat ≈ 0.072

But it took 1222 simulated data sets to get one null data set!

The role of optimization: if we understand the fused lasso, i.e., the way it selects changepoints (stems from KKT conditions), then we can come up with a null distribution without simulation

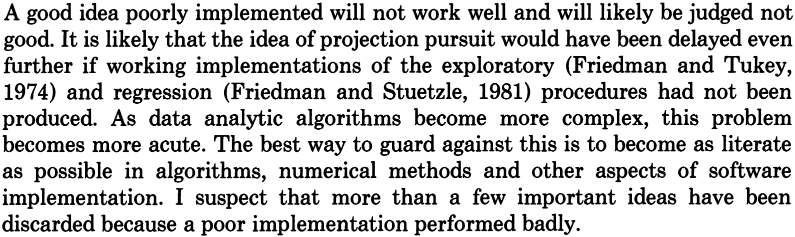


We can use this to efficiently conduct significance tests[1]

[1] Hyun et al. 2018, "Exact post-selection inference for the generalized lasso path"

Widsom from Friedman (1985)

From Jerry Friedman's discussion of Peter Huber's 1985 projection pursuit paper, in Annals of Statistics:



A good idea poorly implemented will not work well and will likely be judged not good. It is likely that the idea of projection pursuit would have been delayed even further if working implementations of the exploratory (Friedman and Tukey, 1974) and regression (Friedman and Stuetzle, 1981) procedures had not been produced. As data analytic algorithms become more complex, this problem becomes more acute. The best way to guard against this is to become as literate as possible in algorithms, numerical methods and other aspects of software implementation. I suspect that more than a few important ideas have been discarded because a poor implementation performed badly.

Arguably, less true today due to the advent of disciplined convex programming? But it still rings true in large part ...

Central concept: convexity

Historically, linear programs were the focus in optimization

Initially, it was thought that the important distinction was between linear and nonlinear optimization problems. But some nonlinear problems turned out to be much harder than others ...

Now it is widely recognized that the right distinction is between convex and nonconvex problems

My two favorite textbooks:

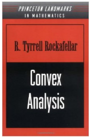
Boyd and Vandenberghe (2004)

Stephen Boyd and Lieven Vandenberghe

Convex Optimization

and

Rockafellar (1970)



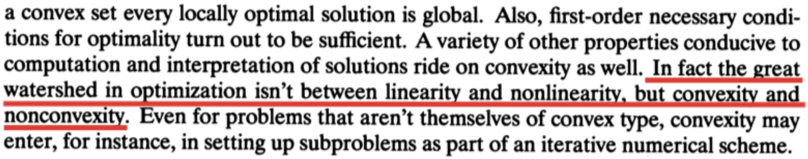
PRINCETON LANDMARKS IN MATHEMATICS

R. Tyrrell Rockafellar

Convex Analysis

Wisdom from Rockafellar (1993)

From Terry Rockafellar's 1993 SIAM Review survey paper:



a convex set every locally optimal solution is global. Also, first-order necessary conditions for optimality turn out to be sufficient. A variety of other properties conducive to computation and interpretation of solutions ride on convexity as well. In fact the great watershed in optimization isn't between linearity and nonlinearity, but convexity and nonconvexity. Even for problems that aren't themselves of convex type, convexity may enter, for instance, in setting up subproblems as part of an iterative numerical scheme.

Credit to Nemirovski, Yudin, Nesterov, others for formalizing this

This view was dominant both within the optimization community and in many application domains for many decades (... currently being challenged by successes of neural networks?)

Prerequisites?

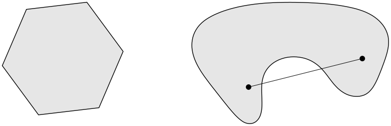
I will assume working knowledge of/proficiency with:

- Real analysis, calculus, linear algebra
- Core problems in Machine Learning and Statistics
- Data structures, computational complexity
- Formal mathematical thinking

Convex sets and functions

Convex set: $C \subseteq \mathbb{R}^n$ such that

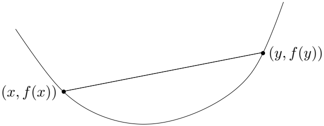
$$x, y \in C \implies tx + (1-t)y \in C \text{ for all } 0 \le t \le 1$$



Convex function: $f : \mathbb{R}^n \to \mathbb{R}$ such that $\mathrm{dom}(f) \subseteq \mathbb{R}^n$ convex, and

$$f(tx + (1-t)y) \le tf(x) + (1-t)f(y) \text{ for all } 0 \le t \le 1$$

and all $x, y \in \mathrm{dom}(f)$



Convex optimization problems

Optimization problem:

$$\begin{aligned} \min_{x \in D} \quad & f(x) \\ \text{subject to} \quad & g_i(x) \le 0,\ i = 1, \dots, m \\ & h_j(x) = 0,\ j = 1, \dots, r \end{aligned}$$

Here $D = \mathrm{dom}(f) \cap \bigcap_{i=1}^m \mathrm{dom}(g_i) \cap \bigcap_{j=1}^p \mathrm{dom}(h_j)$, common domain of all the functions

This is a convex optimization problem provided the functions f and $g_i, i = 1, \dots, m$ are convex, and $h_j, j = 1, \dots, p$ are affine:

$$h_j(x) = a_j^T x + b_j, \quad j = 1, \dots, p$$

Local minima are global minima

For convex optimization problems, local minima are global minima

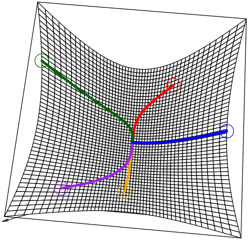
Formally, if x is feasible—$x \in D$, and satisfies all constraints—and minimizes f in a local neighborhood,

$$f(x) \leq f(y) \text{ for all feasible } y, \; \|x - y\|_2 \leq \rho,$$

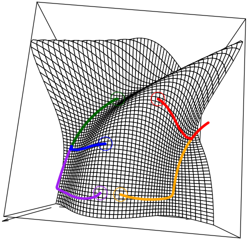
then

$$f(x) \leq f(y) \text{ for all feasible } y$$

This is a very useful fact and will save us a lot of trouble!



Convex



Nonconvex

Outline for Part I

- Part A. Convex sets and functions
- Part B. Basics of optimization
- Part C. Canonical problem forms

Part I: Basic convex analysis

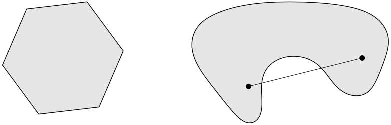
A. Convex sets and functions

Convex sets

Convex set: $C \subseteq \mathbb{R}^n$ such that

$$x, y \in C \implies tx + (1-t)y \in C \text{ for all } 0 \le t \le 1$$

In words, line segment joining any two elements lies entirely in set



Convex combination of $x_1, \ldots, x_k \in \mathbb{R}^n$: any linear combination

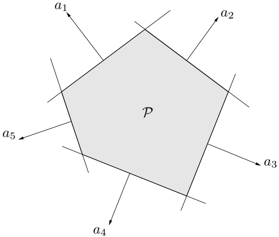
$$\theta_1 x_1 + \cdots + \theta_k x_k$$

with $\theta_i \ge 0$, $i = 1, \ldots, k$, and $\sum_{i=1}^k \theta_i = 1$. Convex hull of a set C, $\mathrm{conv}(C)$, is all convex combinations of elements. Always convex

Examples of convex sets

- Trivial ones: empty set, point, line
- Norm ball: $\{x : \|x\| \leq r\}$, for given norm $\|\cdot\|$, radius r
- Hyperplane: $\{x : a^T x = b\}$, for given a, b
- Halfspace: $\{x : a^T x \leq b\}$
- Affine space: $\{x : Ax = b\}$, for given A, b

- Polyhedron: $\{x : Ax \leq b\}$, where inequality $\leq$ is interpreted componentwise. Note: the set $\{x : Ax \leq b, Cx = d\}$ is also a polyhedron (why?)



- Simplex: special case of polyhedra, given by $\mathrm{conv}\{x_0, \ldots, x_k\}$, where these points are affinely independent. The canonical example is the probability simplex,

$$\mathrm{conv}\{e_1, \ldots, e_n\} = \{w : w \geq 0,\ 1^T w = 1\}$$

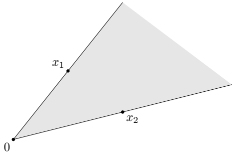
Cones

Cone: $C \subseteq \mathbb{R}^n$ such that

$$x \in C \implies tx \in C \text{ for all } t \geq 0$$

Convex cone: cone that is also convex, i.e.,

$$x_1, x_2 \in C \implies t_1 x_1 + t_2 x_2 \in C \text{ for all } t_1, t_2 \geq 0$$



Conic combination of $x_1, \ldots, x_k \in \mathbb{R}^n$: any linear combination

$$\theta_1 x_1 + \cdots + \theta_k x_k$$

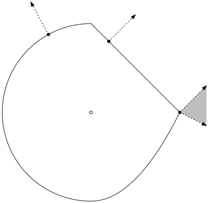
with $\theta_i \geq 0$, $i = 1, \ldots, k$. Conic hull collects all conic combinations

Examples of convex cones

- Norm cone: $\{(x,t) : \|x\| \leq t\}$, for a norm $\|\cdot\|$. Under the ℓ_2 norm $\|\cdot\|_2$, called second-order cone

- Normal cone: given any set C and point $x \in C$, we can define

$$\mathcal{N}_C(x) = \{g : g^T x \geq g^T y, \text{ for all } y \in C\}$$

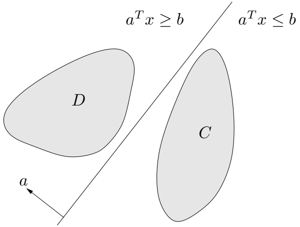


This is always a convex cone, regardless of C

- Positive semidefinite cone: $\mathbb{S}^n_+ = \{X \in \mathbb{S}^n : X \succeq 0\}$, where $X \succeq 0$ means that X is positive semidefinite (and $\mathbb{S}^n$ is the set of $n \times n$ symmetric matrices)

Key properties of convex sets

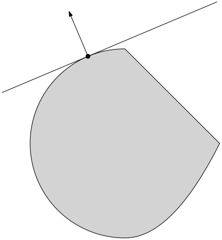
- Separating hyperplane theorem: two disjoint convex sets have a separating between hyperplane them



Formally: if C, D are nonempty convex sets with $C \cap D = \emptyset$, then there exists a, b such that

$$C \subseteq \{x : a^T x \le b\}$$
$$D \subseteq \{x : a^T x \ge b\}$$

- Supporting hyperplane theorem: a boundary point of a convex set has a supporting hyperplane passing through it



Formally: if C is a nonempty convex set, and $x_0 \in \mathrm{bd}(C)$, then there exists a such that

$$C \subseteq \{x : a^T x \leq a^T x_0\}$$

Both of the above theorems (separating and supporting hyperplane theorems) have partial converses; see Section 2.5 of BV

Operations preserving convexity

- Intersection: the intersection of convex sets is convex
- Scaling and translation: if C is convex, then

  $$aC + b = \{ax + b : x \in C\}$$

  is convex for any a, b
- Affine images and preimages: if $f(x) = Ax + b$ and C is convex then

  $$f(C) = \{f(x) : x \in C\}$$

  is convex, and if D is convex then

  $$f^{-1}(D) = \{x : f(x) \in D\}$$

  is convex

Example: linear matrix inequality solution set

Given $A_1, \ldots, A_k, B \in \mathbb{S}^n$, a linear matrix inequality is of the form

$$x_1 A_1 + x_2 A_2 + \cdots + x_k A_k \preceq B$$

for a variable $x \in \mathbb{R}^k$. Let's prove the set C of points x that satisfy the above inequality is convex

Approach 1: directly verify that $x, y \in C \Rightarrow tx + (1-t)y \in C$. This follows by checking that, for any v,

$$v^T \Big(B - \sum_{i=1}^k (t x_i + (1-t) y_i) A_i \Big) v \geq 0$$

Approach 2: let $f : \mathbb{R}^k \to \mathbb{S}^n$, $f(x) = B - \sum_{i=1}^k x_i A_i$. Note that $C = f^{-1}(\mathbb{S}^n_+)$, affine preimage of convex set

More operations preserving convexity

- Perspective images and preimages: the perspective function is $P : \mathbb{R}^n \times \mathbb{R}_{++} \to \mathbb{R}^n$ (where $\mathbb{R}_{++}$ denotes positive reals),

$$P(x, z) = x/z$$

for $z > 0$. If $C \subseteq \mathrm{dom}(P)$ is convex then so is $P(C)$, and if D is convex then so is $P^{-1}(D)$

- Linear-fractional images and preimages: the perspective map composed with an affine function,

$$f(x) = \frac{Ax + b}{c^T x + d}$$

is called a linear-fractional function, defined on $c^T x + d > 0$. If $C \subseteq \mathrm{dom}(f)$ is convex then so if $f(C)$, and if D is convex then so is $f^{-1}(D)$

Example: conditional probability set

Let U, V be random variables over $\{1, \ldots, n\}$ and $\{1, \ldots, m\}$. Let $C \subseteq \mathbb{R}^{nm}$ be a set of joint distributions for U, V, i.e., each $p \in C$ defines joint probabilities

$$p_{ij} = \mathbb{P}(U = i, V = j)$$

Let $D \subseteq \mathbb{R}^{nm}$ contain corresponding conditional distributions, i.e., each $q \in D$ defines

$$q_{ij} = \mathbb{P}(U = i | V = j)$$

Assume C is convex. Let's prove that D is convex. Write

$$D = \left\{ q \in \mathbb{R}^{nm} : q_{ij} = \frac{p_{ij}}{\sum_{k=1}^{n} p_{kj}}, \text{ for some } p \in C \right\} = f(C)$$

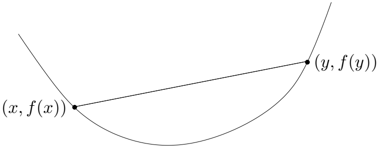
where f is a linear-fractional function, hence D is convex

Convex functions

Convex function: $f : \mathbb{R}^n \to \mathbb{R}$ such that $\mathrm{dom}(f) \subseteq \mathbb{R}^n$ convex, and

$$f(tx + (1-t)y) \le tf(x) + (1-t)f(y) \quad \text{for } 0 \le t \le 1$$

and all $x, y \in \mathrm{dom}(f)$



In words, function lies below the line segment joining $f(x), f(y)$

Concave function: opposite inequality above, so that

$$f \text{ concave} \iff -f \text{ convex}$$

Important modifiers:

- Strictly convex: $f\big(tx + (1-t)y\big) < tf(x) + (1-t)f(y)$ for $x \neq y$ and $0 < t < 1$. In words, f is convex and has greater curvature than a linear function

- Strongly convex with parameter $m > 0$: $f - \frac{m}{2}\|x\|_2^2$ is convex. In words, f is at least as convex as a quadratic function

Note: strongly convex ⇒ strictly convex ⇒ convex

(Analogously for concave functions)

Examples of convex functions

- Univariate functions:
  - Exponential function: e^{ax} is convex for any a over $\mathbb{R}$
  - Power function: x^a is convex for $a \geq 1$ or $a \leq 0$ over $\mathbb{R}_+$ (nonnegative reals)
  - Power function: x^a is concave for $0 \leq a \leq 1$ over $\mathbb{R}_+$
  - Logarithmic function: $\log x$ is concave over $\mathbb{R}_{++}$

- Affine function: $a^T x + b$ is both convex and concave

- Quadratic function: $\frac{1}{2} x^T Q x + b^T x + c$ is convex provided that $Q \succeq 0$ (positive semidefinite)

- Least squares loss: $\|y - Ax\|_2^2$ is always convex (since $A^T A$ is always positive semidefinite)

- Norm: $\|x\|$ is convex for any norm; e.g., ℓ_p norms,

$$\|x\|_p = \left(\sum_{i=1}^n x_i^p\right)^{1/p} \text{ for } p \geq 1, \quad \|x\|_\infty = \max_{i=1,\ldots,n} |x_i|$$

and also operator (spectral) and trace (nuclear) norms,

$$\|X\|_{\text{op}} = \sigma_1(X), \quad \|X\|_{\text{tr}} = \sum_{i=1}^r \sigma_r(X)$$

where $\sigma_1(X) \geq \ldots \geq \sigma_r(X) \geq 0$ are the singular values of the matrix X

- Indicator function: if C is convex, then its indicator function

  $$I_C(x) = \begin{cases} 0 & x \in C \\ \infty & x \notin C \end{cases}$$

  is convex

- Support function: for any set C (convex or not), its support function

  $$I_C^*(x) = \max_{y \in C} \; x^T y$$

  is convex

- Max function: $f(x) = \max\{x_1, \ldots, x_n\}$ is convex

Key properties of convex functions

- A function is convex if and only if its restriction to any line is convex
- Epigraph characterization: a function f is convex if and only if its epigraph

$$\mathrm{epi}(f) = \{(x, t) \in \mathrm{dom}(f) \times \mathbb{R} : f(x) \leq t\}$$

is a convex set

- Convex sublevel sets: if f is convex, then its sublevel sets

$$\{x \in \mathrm{dom}(f) : f(x) \leq t\}$$

are convex, for all $t \in \mathbb{R}$. The converse is not true

- First-order characterization: if f is differentiable, then f is convex if and only if $\mathrm{dom}(f)$ is convex, and

  $$f(y) \geq f(x) + \nabla f(x)^T (y - x)$$

  for all $x, y \in \mathrm{dom}(f)$. Therefore for a differentiable convex function $\nabla f(x) = 0 \iff x$ minimizes f

- Second-order characterization: if f is twice differentiable, then f is convex if and only if $\mathrm{dom}(f)$ is convex, and $\nabla^2 f(x) \succeq 0$ for all $x \in \mathrm{dom}(f)$

- Jensen's inequality: if f is convex, and X is a random variable supported on $\mathrm{dom}(f)$, then $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$

Operations preserving convexity

- Nonnegative linear combination: $f_1, \ldots, f_m$ convex implies $a_1 f_1 + \cdots + a_m f_m$ convex for any $a_1, \ldots, a_m \geq 0$

- Pointwise maximization: if f_s is convex for any $s \in S$, then $f(x) = \max_{s \in S} \; f_s(x)$ is convex. Note that the set S here (number of functions f_s) can be infinite

- Partial minimization: if $g(x, y)$ is convex in x, y, and C is convex, then $f(x) = \min_{y \in C} \; g(x, y)$ is convex

More operations preserving convexity

- Affine composition: if f is convex, then $g(x) = f(Ax + b)$ is convex

- General composition: suppose $f = h \circ g$, where $g : \mathbb{R}^n \to \mathbb{R}$, $h : \mathbb{R} \to \mathbb{R}$, $f : \mathbb{R}^n \to \mathbb{R}$. Then:
  - f is convex if h is convex and nondecreasing, g is convex
  - f is convex if h is convex and nonincreasing, g is concave
  - f is concave if h is concave and nondecreasing, g concave
  - f is concave if h is concave and nonincreasing, g convex

  How to remember these? Think of the chain rule when $n = 1$:

$$f''(x) = h''(g(x))g'(x)^2 + h'(g(x))g''(x)$$

- Vector composition: suppose that

$$f(x) = h\big(g(x)\big) = h\big(g_1(x), \ldots, g_k(x)\big)$$

where $g : \mathbb{R}^n \to \mathbb{R}^k$, $h : \mathbb{R}^k \to \mathbb{R}$, $f : \mathbb{R}^n \to \mathbb{R}$. Then:

  - f is convex if h is convex and nondecreasing in each argument, g is convex
  - f is convex if h is convex and nonincreasing in each argument, g is concave
  - f is concave if h is concave and nondecreasing in each argument, g is concave
  - f is concave if h is concave and nonincreasing in each argument, g is convex

Example: log-sum-exp function

Log-sum-exp function: $g(x) = \log(\sum_{i=1}^k e^{a_i^T x + b_i})$, for fixed a_i, b_i, $i = 1, \ldots, k$. Often called "soft max", as it smoothly approximates $\max_{i=1,\ldots k} \, (a_i^T x + b_i)$

How to show convexity? First, note it suffices to prove convexity of $f(x) = \log(\sum_{i=1}^n e^{x_i})$ (affine composition rule)

Now use second-order characterization. Calculate

$$\begin{aligned}
\nabla_i f(x) &= \frac{e^{x_i}}{\sum_{\ell=1}^n e^{x_\ell}} \\
\nabla_{ij}^2 f(x) &= \frac{e^{x_i}}{\sum_{\ell=1}^n e^{x_\ell}} 1\{i=j\} - \frac{e^{x_i} e^{x_j}}{(\sum_{\ell=1}^n e^{x_\ell})^2}
\end{aligned}$$

Write $\nabla^2 f(x) = \mathrm{diag}(z) - zz^T$, where $z_i = e^{x_i}/(\sum_{\ell=1}^n e^{x_\ell})$. This matrix is diagonally dominant, hence positive semidefinite

Part I: Basic convex analysis

B. Basics of optimization

Optimization terminology

Reminder: a convex optimization problem (or program) is

$$\begin{aligned} \min_{x \in D} \quad & f(x) \\ \text{subject to} \quad & g_i(x) \leq 0,\ i = 1, \ldots, m \\ & Ax = b \end{aligned}$$

where f and g_i, $i = 1, \ldots, m$ are all convex, and the optimization domain is $D = \mathrm{dom}(f) \cap \bigcap_{i=1}^m \mathrm{dom}(g_i)$ (often we do not write D)

- f is called criterion or objective function
- g_i is called inequality constraint function
- If $x \in D$, $g_i(x) \leq 0$, $i = 1, \ldots, m$, and $Ax = b$ then x is called a feasible point
- The minimum of $f(x)$ over all feasible points x is called the optimal value, written $f^\star$

- If x is feasible and $f(x) = f^\star$, then x is called optimal; also called a solution, or a minimizer[2]
- If x is feasible and $f(x) \leq f^\star + \epsilon$, then x is called ϵ-suboptimal
- If x is feasible and $g_i(x) = 0$, then we say g_i is active at x
- Convex minimization can be reposed as concave maximization

$$
\begin{array}{lll}
\min\limits_{x} & f(x) \\
\text{subject to} & g_i(x) \leq 0, \\
& \quad i = 1, \ldots, m \\
& Ax = b
\end{array}
\quad \Longleftrightarrow \quad
\begin{array}{lll}
\max\limits_{x} & -f(x) \\
\text{subject to} & g_i(x) \leq 0, \\
& \quad i = 1, \ldots, m \\
& Ax = b
\end{array}
$$

Both are called convex optimization problems

[2] Note: a convex optimization problem need not have solutions, i.e., need not attain its minimum, but we will not be careful about this

Solution set

Let X_{opt} be the set of all solutions of convex problem, written

$$\begin{array}{rlll} X_{\mathsf{opt}} & = & \text{argmin} & f(x) \\ & & \text{subject to} & g_i(x) \leq 0,\ i = 1, \ldots, m \\ & & & Ax = b \end{array}$$

Key property 1: X_{opt} is a convex set

Proof: use definitions. If x, y are solutions, then for $0 \leq t \leq 1$,

- $g_i(tx + (1-t)y) \leq tg_i(x) + (1-t)g_i(y) \leq 0$
- $A(tx + (1-t)y) = tAx + (1-t)Ay = b$
- $f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) = f^\star$

Therefore $tx + (1-t)y$ is also a solution

Key property 2: if f is strictly convex, then solution is unique, i.e., X_{opt} contains one element

Example: lasso

Given $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, consider the lasso problem:

$$\begin{aligned} \min_{\beta} \quad & \|y - X\beta\|_2^2 \\ \text{subject to} \quad & \|\beta\|_1 \le s \end{aligned}$$

Is this convex? What is the criterion function? The inequality and equality constraints? Feasible set? Is the solution unique, when:

- $n \ge p$ and X has full column rank?
- $p > n$ ("high-dimensional" case)?

How do our answers change if we changed criterion to Huber loss:

$$\sum_{i=1}^{n} \rho(y_i - x_i^T \beta), \quad \rho(z) = \begin{cases} \frac{1}{2} z^2 & |z| \le \delta \\ \delta |z| - \frac{1}{2}\delta^2 & \text{else} \end{cases} \quad ?$$

Example: support vector machines

Given $y \in \{-1, 1\}^n$, $X \in \mathbb{R}^{n \times p}$ with rows $x_1, \ldots x_n$, consider the support vector machine or SVM problem:

$$\begin{aligned} \min_{\beta, \beta_0, \xi} \quad & \frac{1}{2}\|\beta\|_2^2 + C \sum_{i=1}^{n} \xi_i \\ \text{subject to} \quad & \xi_i \geq 0, \ i = 1, \ldots, n \\ & y_i(x_i^T \beta + \beta_0) \geq 1 - \xi_i, \ i = 1, \ldots, n \end{aligned}$$

Is this convex? What is the criterion, constraints, feasible set? Is the solution (β, β_0, ξ) unique? What if changed the criterion to

$$\frac{1}{2}\|\beta\|_2^2 + \frac{1}{2}\beta_0^2 + C \sum_{i=1}^{n} \xi_i^{1.01} ?$$

For original criterion, what about β component, at the solution?

Rewriting constraints

The optimization problem

$$\begin{aligned} \min_x \quad & f(x) \\ \text{subject to} \quad & g_i(x) \le 0,\ i = 1, \ldots, m \\ & Ax = b \end{aligned}$$

can be rewritten as

$$\min_x \ f(x) \ \ \text{subject to} \ \ x \in C$$

where $C = \{x : g_i(x) \le 0,\ i = 1, \ldots, m,\ Ax = b\}$, the feasible set. Hence the latter formulation is completely general

With I_C the indicator of C, we can write this in unconstrained form

$$\min_x \ f(x) + I_C(x)$$

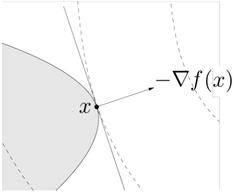
First-order optimality condition

For a convex problem

$$\min_x \ f(x) \ \text{ subject to } \ x \in C$$

and differentiable f, a feasible point x is optimal if and only if

$$\nabla f(x)^T (y - x) \geq 0 \quad \text{for all } y \in C$$



This is called the first-order condition for optimality

In words: all feasible directions from x are aligned with gradient $\nabla f(x)$

Important special case: if $C = \mathbb{R}^n$ (unconstrained optimization), then optimality condition reduces to familiar $\nabla f(x) = 0$

Example: quadratic minimization

Consider minimizing the quadratic function

$$f(x) = \frac{1}{2}x^T Q x + b^T x + c$$

where $Q \succeq 0$. The first-order condition says that solution satisfies

$$\nabla f(x) = Qx + b = 0$$

- if $Q \succ 0$, then there is a unique solution $x = -Q^{-1}b$
- if Q is singular and $b \notin \operatorname{col}(Q)$, then there is no solution (i.e., $\min_x \; f(x) = -\infty$)
- if Q is singular and $b \in \operatorname{col}(Q)$, then there are infinitely many solutions

  $$x = -Q^{+}b + z, \quad z \in \operatorname{null}(Q)$$

  where Q^{+} is the pseudoinverse of Q

Example: equality-constrained minimization

Consider the equality-constrained convex problem:

$$\min_x \; f(x) \;\; \text{subject to} \;\; Ax = b$$

with f differentiable. Let's prove Lagrange multiplier optimality condition

$$\nabla f(x) + A^T u = 0 \quad \text{for some } u$$

According to first-order optimality, solution x satisfies $Ax = b$ and

$$\nabla f(x)^T (y - x) \geq 0 \quad \text{for all } y \text{ such that } Ay = b$$

This is equivalent to

$$\nabla f(x)^T v = 0 \quad \text{for all } v \in \mathrm{null}(A)$$

Result follows because $\mathrm{null}(A)^\perp = \mathrm{row}(A)$

Example: projection onto a convex set

Consider projection onto convex set C:

$$\min_x \ \|a-x\|_2^2 \ \text{ subject to } \ x \in C$$

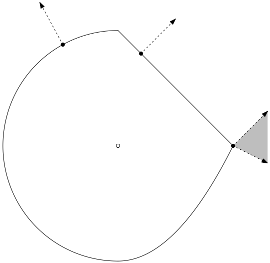
First-order optimality condition says that the solution x satisfies

$$\nabla f(x)^T(y-x) = (x-a)^T(y-x) \geq 0 \quad \text{for all } y \in C$$

Equivalently, this says that

$$a - x \in \mathcal{N}_C(x)$$

where recall $\mathcal{N}_C(x)$ is the normal cone to C at x



Partial optimization

Reminder: $g(x) = \min_{y \in C} \; f(x, y)$ is convex in x, provided that f is convex in (x, y) and C is a convex set

Therefore we can always partially optimize a convex problem and retain convexity

E.g., if we decompose $x = (x_1, x_2) \in \mathbb{R}^{n_1+n_2}$, then

$$\begin{array}{ll} \displaystyle\min_{x_1, x_2} & f(x_1, x_2) \\ \text{subject to} & g_1(x_1) \leq 0 \\ & g_2(x_2) \leq 0 \end{array} \quad \Longleftrightarrow \quad \begin{array}{ll} \displaystyle\min_{x_1} & \tilde{f}(x_1) \\ \text{subject to} & g_1(x_1) \leq 0 \end{array}$$

where $\tilde{f}(x_1) = \min\{f(x_1, x_2) : g_2(x_2) \leq 0\}$. The right problem is convex if the left problem is

Example: hinge form of SVMs

Recall the SVM problem

$$\begin{aligned} \min_{\beta,\beta_0,\xi} \quad & \frac{1}{2}\|\beta\|_2^2 + C\sum_{i=1}^n \xi_i \\ \text{subject to} \quad & \xi_i \geq 0,\ y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i,\ i = 1,\ldots,n \end{aligned}$$

Rewrite the constraints as $\xi_i \geq \max\{0, 1 - y_i(x_i^T\beta + \beta_0)\}$. Indeed we can argue that we have $=$ at solution

Therefore plugging in for optimal ξ gives the hinge form of SVMs:

$$\min_{\beta,\beta_0} \ \frac{1}{2}\|\beta\|_2^2 + C\sum_{i=1}^n \big[1 - y_i(x_i^T\beta + \beta_0)\big]_+$$

where $a_+ = \max\{0, a\}$ is called the hinge function

Transformations and change of variables

If $h : \mathbb{R} \to \mathbb{R}$ is a monotone increasing transformation, then

$$\begin{aligned} &\min_x \; f(x) \;\; \text{subject to} \;\; x \in C \\ \iff &\min_x \; h(f(x)) \;\; \text{subject to} \;\; x \in C \end{aligned}$$

Similarly, inequality or equality constraints can be transformed and yield equivalent optimization problems. Can use this to reveal the "hidden convexity" of a problem

If $\phi : \mathbb{R}^n \to \mathbb{R}^m$ is one-to-one, and its image covers feasible set C, then we can change variables in an optimization problem:

$$\begin{aligned} &\min_x \; f(x) \;\; \text{subject to} \;\; x \in C \\ \iff &\min_y \; f(\phi(y)) \;\; \text{subject to} \;\; \phi(y) \in C \end{aligned}$$

Example: geometric programming

A monomial is a function $f : \mathbb{R}^n_{++} \to \mathbb{R}$ of the form

$$f(x) = \gamma x_1^{a_1} x_2^{a_2} \cdots x_n^{a_n}$$

for $\gamma > 0$, $a_1, \ldots, a_n \in \mathbb{R}$. A posynomial is a sum of monomials,

$$f(x) = \sum_{k=1}^{p} \gamma_k x_1^{a_{k1}} x_2^{a_{k2}} \cdots x_n^{a_{kn}}$$

A geometric program is of the form

$$\begin{aligned} \min_x \quad & f(x) \\ \text{subject to} \quad & g_i(x) \le 1,\ i = 1, \ldots, m \\ & h_j(x) = 1,\ j = 1, \ldots, r \end{aligned}$$

where f, g_i, $i = 1, \ldots, m$ are posynomials and h_j, $j = 1, \ldots, r$ are monomials. This is nonconvex

Given $f(x) = \gamma x_1^{a_1} x_2^{a_2} \cdots x_n^{a_n}$, let $y_i = \log x_i$ and rewrite this as

$$\gamma (e^{y_1})^{a_1} (e^{y_2})^{a_2} \cdots (e^{y_n})^{a_n} = e^{a^T y + b}$$

for $b = \log \gamma$. Also, a posynomial can be written as $\sum_{k=1}^{p} e^{a_k^T y + b_k}$. With this variable substitution, and after taking logs, a geometric program is equivalent to

$$\begin{aligned} \min_{x} \quad & \log \left(\sum_{k=1}^{p_0} e^{a_{0k}^T y + b_{0k}} \right) \\ \text{subject to} \quad & \log \left(\sum_{k=1}^{p_i} e^{a_{ik}^T y + b_{ik}} \right) \le 0,\ i = 1, \ldots, m \\ & c_j^T y + d_j = 0,\ j = 1, \ldots, r \end{aligned}$$

This is convex, recalling the convexity of soft max functions

Eliminating equality constraints

Important special case of change of variables: eliminating equality constraints. Given the problem

$$\begin{aligned} \min_x \quad & f(x) \\ \text{subject to} \quad & g_i(x) \le 0,\ i = 1, \ldots, m \\ & Ax = b \end{aligned}$$

we can always express any feasible point as $x = My + x_0$, where $Ax_0 = b$ and $\mathrm{col}(M) = \mathrm{null}(A)$. Hence the above is equivalent to

$$\begin{aligned} \min_y \quad & f(My + x_0) \\ \text{subject to} \quad & g_i(My + x_0) \le 0,\ i = 1, \ldots, m \end{aligned}$$

Note: this is fully general but not always a good idea (practically)

Introducing slack variables

Essentially opposite to eliminating equality contraints: introducing slack variables. Given the problem

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{subject to} \quad & g_i(x) \leq 0, \ i = 1, \ldots, m \\ & Ax = b \end{aligned}$$

we can transform the inequality constraints via

$$\begin{aligned} \min_{x,s} \quad & f(x) \\ \text{subject to} \quad & s_i \geq 0, \ i = 1, \ldots, m \\ & g_i(x) + s_i = 0, \ i = 1, \ldots, m \\ & Ax = b \end{aligned}$$

Note: this is no longer convex unless g_i, $i = 1, \ldots, n$ are affine

Convex relaxations

Given an optimization problem

$$\min_x \ f(x) \ \text{ subject to } \ x \in C$$

we can always take an enlarged constraint set $\tilde{C} \supseteq C$ and consider

$$\min_x \ f(x) \ \text{ subject to } \ x \in \tilde{C}$$

This is called a relaxation and its optimal value is always smaller or equal to that of the original problem

Important special case: relaxing nonaffine equality constraints, i.e.,

$$h_j(x) = 0, \ j = 1, \ldots, r$$

where h_j, $j = 1, \ldots, r$ are convex but nonaffine, are replaced with

$$h_j(x) \leq 0, \ j = 1, \ldots, r$$

Example: principal components analysis

Given $X \in \mathbb{R}^{n \times p}$, consider the low rank approximation problem:

$$\min_{R} \ \|X - R\|_F^2 \ \text{ subject to } \ \mathrm{rank}(R) = k$$

Here $\|A\|_F^2 = \sum_{i=1}^n \sum_{j=1}^p A_{ij}^2$, the entrywise squared ℓ_2 norm, and $\mathrm{rank}(A)$ denotes the rank of A

Also called principal components analysis or PCA problem. Given $X = UDV^T$, singular value decomposition or SVD, the solution is

$$R = U_k D_k V_k^T$$

where U_k, V_k are the first k columns of U, V and D_k is the first k diagonal elements of D. That is, R is reconstruction of X from its first k principal components

The PCA problem is not convex. Let's recast it. First rewrite as

$$\begin{aligned} &\min_{Z \in \mathbb{S}^p} \; \|X - XZ\|_F^2 \;\; \text{subject to} \;\; \mathrm{rank}(Z) = k, \;\; Z \text{ is a projection} \\ \iff &\max_{Z \in \mathbb{S}^p} \; \mathrm{tr}(SZ) \;\; \text{subject to} \;\; \mathrm{rank}(Z) = k, \;\; Z \text{ is a projection} \end{aligned}$$

where $S = X^T X$. Hence constraint set is the nonconvex set

$$C = \Big\{ Z \in \mathbb{S}^p : \lambda_i(Z) \in \{0,1\}, \; i = 1, \ldots, p, \; \mathrm{tr}(Z) = k \}$$

where $\lambda_i(Z)$, $i = 1, \ldots, n$ are the eigenvalues of Z. Solution in this formulation is

$$Z = V_k V_k^T$$

where V_k gives first k columns of V

Now consider relaxing constraint set to $\mathcal{F}_k = \mathrm{conv}(C)$, its convex hull. Note

$$\begin{aligned}\mathcal{F}_k &= \{Z \in \mathbb{S}^p : \lambda_i(Z) \in [0,1],\ i = 1,\ldots,p,\ \mathrm{tr}(Z) = k\} \\ &= \{Z \in \mathbb{S}^p : 0 \preceq Z \preceq I,\ \mathrm{tr}(Z) = k\}\end{aligned}$$

This set is called the Fantope of order k. It is convex. Hence, the linear maximization over the Fantope, namely

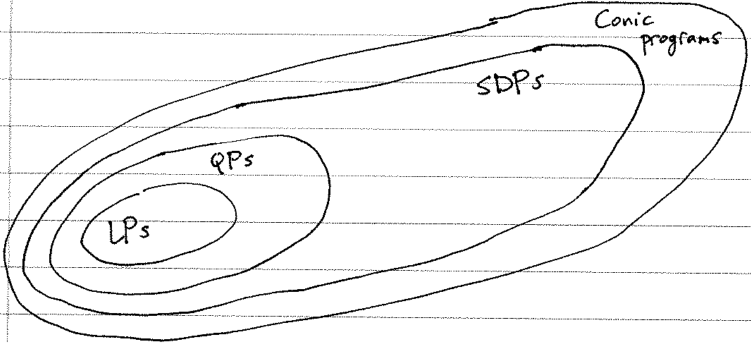
$$\max_{Z \in \mathcal{F}_k} \ \mathrm{tr}(SZ)$$

is a convex problem. Remarkably, this is equivalent to the original nonconvex PCA problem (admits the same solution)!

(Famous result: Fan (1949), "On a theorem of Weyl conerning eigenvalues of linear transformations")

Part I: Basic convex analysis

C. Canonical problem forms


Conic programs
SDPs
QPs
LPs

Linear program

A linear program or LP is an optimization problem of the form

$$\begin{aligned} \min_x \quad & c^T x \\ \text{subject to} \quad & Dx \le d \\ & Ax = b \end{aligned}$$

Observe that this is always a convex optimization problem

- First introduced by Kantorovich in the late 1930s and Dantzig in the 1940s
- Dantzig's simplex algorithm gives a direct (noniterative) solver for LPs (later in the course we'll see interior point methods)
- Fundamental problem in convex optimization. Many diverse applications, rich history

Example: diet problem

Find cheapest combination of foods that satisfies some nutritional requirements (useful for graduate students!)

$$\begin{aligned} \min_{x} \quad & c^T x \\ \text{subject to} \quad & Dx \geq d \\ & x \geq 0 \end{aligned}$$

Interpretation:

- c_j : per-unit cost of food j
- d_i : minimum required intake of nutrient i
- D_{ij} : content of nutrient i per unit of food j
- x_j : units of food j in the diet

Example: basis pursuit

Given $y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times p}$, where $p > n$. Suppose that we seek the sparsest solution to underdetermined linear system $X\beta = y$:

$$\begin{aligned} \min_{\beta} \quad & \|\beta\|_0 \\ \text{subject to} \quad & X\beta = y \end{aligned}$$

where recall $\|\beta\|_0 = \sum_{j=1}^p 1\{\beta_j \neq 0\}$, the ℓ_0 "norm"

The ℓ_1 approximation, often called basis pursuit:

$$\begin{aligned} \min_{\beta} \quad & \|\beta\|_1 \\ \text{subject to} \quad & X\beta = y \end{aligned}$$

This can be reformulated as a linear program (check this!)

Example: Dantzig selector

Modification of previous problem, where we allow for $X\beta \approx y$ (we don't require exact equality), the Dantzig selector:[3]

$$\begin{array}{ll} \displaystyle\min_{\beta} & \|\beta\|_1 \\ \text{subject to} & \|X^T(y - X\beta)\|_\infty \le \lambda \end{array}$$

Here $\lambda \geq 0$ is a tuning parameter

Again, this can be reformulated as a linear program (check this!)

[3] Candes and Tao (2007), "The Dantzig selector: statistical estimation when p is much larger than n"

Standard form

A linear program is said to be in standard form when it is written as

$$\begin{aligned} \min_x \quad & c^T x \\ \text{subject to} \quad & Ax = b \\ & x \geq 0 \end{aligned}$$

Any linear program can be rewritten in standard form (check this!)

Convex quadratic program

A convex quadratic program or QP is an optimization problem of the form

$$\begin{aligned} \min_{x} \quad & c^T x + \frac{1}{2} x^T Q x \\ \text{subject to} \quad & Dx \le d \\ & Ax = b \end{aligned}$$

where $Q \succeq 0$, i.e., positive semidefinite

Note that this problem is not convex when $Q \not\succeq 0$

From now on, when we say quadratic program or QP, we implicitly assume that $Q \succeq 0$ (so the problem is convex)

Example: portfolio optimization

Construct a financial portfolio, trading off performance and risk:

$$\begin{array}{ll} \max\limits_{x} & \mu^T x - \frac{\gamma}{2} x^T Q x \\ \text{subject to} & 1^T x = 1 \\ & x \geq 0 \end{array}$$

Interpretation:

- μ : expected assets' returns
- Q : covariance matrix of assets' returns
- γ : risk aversion
- x : portfolio holdings (percentages)

Example: support vector machines

Given $y \in \{-1, 1\}^n$, $X \in \mathbb{R}^{n \times p}$ having rows $x_1, \ldots, x_n$, recall the support vector machine or SVM problem:

$$\begin{aligned} \min_{\beta, \beta_0, \xi} \quad & \frac{1}{2}\|\beta\|_2^2 + C \sum_{i=1}^n \xi_i \\ \text{subject to} \quad & \xi_i \geq 0, \ i = 1, \ldots, n \\ & y_i(x_i^T \beta + \beta_0) \geq 1 - \xi_i, \ i = 1, \ldots, n \end{aligned}$$

This is a quadratic program

Example: lasso

Given $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, recall the lasso problem:

$$\begin{aligned} \min_{\beta} \quad & \|y - X\beta\|_2^2 \\ \text{subject to} \quad & \|\beta\|_1 \le s \end{aligned}$$

Here $s \ge 0$ is a tuning parameter. Indeed, this can be reformulated as a quadratic program (check this!)

Alternative parametrization (called Lagrange, or penalized form):

$$\min_{\beta} \; \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

Now $\lambda \ge 0$ is a tuning parameter. And again, this can be rewritten as a quadratic program (check this!)

Standard form

A quadratic program is in standard form if it is written as

$$\begin{aligned} \min_x \quad & c^T x + \frac{1}{2} x^T Q x \\ \text{subject to} \quad & Ax = b \\ & x \geq 0 \end{aligned}$$

Any quadratic program can be rewritten in standard form

Motivation for semidefinite programs

Consider linear programming again:

$$\begin{aligned} \min_{x} \quad & c^T x \\ \text{subject to} \quad & Dx \le d \\ & Ax = b \end{aligned}$$

Can generalize by changing $\le$ to different (partial) order. Recall:

- $\mathbb{S}^n$ is space of $n \times n$ symmetric matrices
- $\mathbb{S}^n_+$ is the space of positive semidefinite matrices, i.e.,

$$\mathbb{S}^n_+ = \{X \in \mathbb{S}^n : u^T X u \ge 0 \text{ for all } u \in \mathbb{R}^n\}$$

- $\mathbb{S}^n_{++}$ is the space of positive definite matrices, i.e.,

$$\mathbb{S}^n_{++} = \left\{X \in \mathbb{S}^n : u^T X u > 0 \text{ for all } u \in \mathbb{R}^n \setminus \{0\}\right\}$$

Semidefinite program

A semidefinite program or SDP is an optimization problem of the form

$$\begin{array}{ll} \min\limits_{x} & c^T x \\ \text{subject to} & x_1 F_1 + \cdots + x_n F_n \preceq F_0 \\ & Ax = b \end{array}$$

Here $F_j \in \mathbb{S}^d$, for $j = 0, 1, \ldots n$, and $A \in \mathbb{R}^{m \times n}$, $c \in \mathbb{R}^n$, $b \in \mathbb{R}^m$. Observe that this is always a convex optimization problem

Also, any linear program is a semidefinite program (check this!)

Standard form

A semidefinite program is in standard form if it is written as

$$\begin{aligned} \min_{X} \quad & C \bullet X \\ \text{subject to} \quad & A_i \bullet X = b_i,\ i = 1, \ldots, m \\ & X \succeq 0 \end{aligned}$$

where $X \bullet Y = \mathrm{tr}(XY)$. Any semidefinite program can be written in standard form (for a challenge, check this!)

Example: theta function

Let $G = (N, E)$ be an undirected graph, $N = \{1, \ldots, n\}$, and

- $\omega(G)$: clique number of G
- $\chi(G)$: chromatic number of G

The Lovasz theta function:[4]

$$\begin{array}{rcll} \vartheta(G) & = & \max\limits_{X} & 11^T \bullet X \\ & & \text{subject to} & I \bullet X = 1 \\ & & & X_{ij} = 0, \ (i,j) \notin E \\ & & & X \succeq 0 \end{array}$$

The Lovasz sandwich theorem: $\omega(G) \leq \vartheta(\bar{G}) \leq \chi(G)$, where $\bar{G}$ is the complement graph of G

[4] Lovasz (1979), "On the Shannon capacity of a graph"

Example: trace norm minimization

Let $A : \mathbb{R}^{m\times n} \to \mathbb{R}^p$ be a linear map,

$$A(X) = \begin{pmatrix} A_1 \bullet X \\ \dots \\ A_p \bullet X \end{pmatrix}$$

for $A_1, \dots, A_p \in \mathbb{R}^{m\times n}$ (and where $A_i \bullet X = \mathrm{tr}(A_i^T X)$). Finding lowest-rank solution to an underdetermined system, nonconvex:

$$\begin{array}{ll} \displaystyle\min_X & \mathrm{rank}(X) \\ \text{subject to} & A(X) = b \end{array}$$

Trace norm approximation:

$$\begin{array}{ll} \displaystyle\min_X & \|X\|_{\mathrm{tr}} \\ \text{subject to} & A(X) = b \end{array}$$

This is indeed an SDP (but harder to show, requires duality ...)

Conic program

A conic program is an optimization problem of the form:

$$\begin{aligned} \min_{x} \quad & c^T x \\ \text{subject to} \quad & Ax = b \\ & D(x) + d \in K \end{aligned}$$

Here:

- $c, x \in \mathbb{R}^n$, and $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$
- $D : \mathbb{R}^n \to Y$ is a linear map, $d \in Y$, for Euclidean space Y
- $K \subseteq Y$ is a closed convex cone

Both LPs and SDPs are special cases of conic programming. For LPs, $K = \mathbb{R}^n_+$; for SDPs, $K = \mathbb{S}^n_+$

Hey, what about QPs?

Lastly, our old friend QPs "sneak" into the hierarchy. It's not easy to directly show that every QP is an SDP

But it turns out QPs are a special type of conic program called a second-order cone program (SOCP):

$$\begin{aligned} \min_x \quad & c^T x \\ \text{subject to} \quad & \|D_i x + d_i\|_2 \leq e_i^T x + f_i, \ i = 1, \ldots, p \\ & Ax = b \end{aligned}$$

In fact, every SOCP is an SDP. This gives the (extended) hierachy

$$\text{LPs} \subseteq \text{QPs} \subseteq \text{SOCPs} \subseteq \text{SDPs} \subseteq \text{Conic programs}$$

completing the picture we saw at the start

References and further reading

Part A:

- S. Boyd and L. Vandenberghe (2004), “Convex optimization”, Chapters 2 and 3
- J.P. Hiriart-Urruty and C. Lemarechal (1993), “Fundamentals of convex analysis”, Chapters A and B
- R. T. Rockafellar (1970), “Convex analysis”, Chapters 1–10,

Part B:

- S. Boyd and L. Vandenberghe (2004), “Convex optimization”, Chapter 4
- O. Guler (2010), “Foundations of optimization”, Chapter 4

Part C:

- S. Boyd and L. Vandenberghe (2004), “Convex optimization,” Chapter 4
- D. Bertsimas and J. Tsitsiklis (1997), “Introduction to linear optimization,” Chapters 1, 2
- A. Nemirovski and A. Ben-Tal (2001), “Lectures on modern convex optimization,” Chapters 1–4