Convex Optimization: Fall 2019

Based on lectures by Ryan Tibshirani

Notes scribed by Ang Ji

April 2024

These notes are scribed by me, and it is difficult to avoids errors.

Course Information

Instructor: Ryan Tibshirani
Homepage: Machine Learning 10-725
Teaching: Carnegie Mellon University

Schedule

This course is divided into five parts:

(i) **Theory I: Fundamentals**

- Introduction
- Convexity I: Sets and Functions
- Convexity II: Optimization Basics
- Canonical Problem Forms

(ii) **Algorithms I: First-order methods**

(iii) **Theory II: Duality and optimality**

(iv) **Algorithms II: Second-order methods**

(v) **Advanced topics**

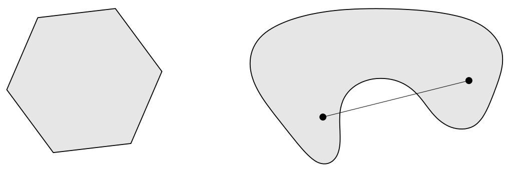
Contents

1 Convexity I: Sets and Functions

1.1 Convex sets

Definition (Convex set)**.** $C \subseteq \mathbb{R}^n$ such that

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



Definition (Convex combination)**.** For $x_1, \cdots, x_k \in \mathbb{R}^n$, any linear combination

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

with $\theta_i \geq 0, i = 1, \cdots, k$, and $\sum_{i=1}^k \theta_i = 1$.

Definition (Convex hull)**.** The convex hull of C, $\text{conv}(C)$, is all convex combinations of elements, and is always convex.

1.1.1 Example of convex sets:

(i) **Trivial ones:** empty set, point, line

(ii) **Norm ball:** $\{x : \|x\| \leq r\}$, for given norm $\|\cdot\|$, radius r

(iii) **Hyperplane:** $\{x : a^T x = b\}$, for given a, b

(iv) **Halfspace:** $\{x : a^T x \leq b\}$, for given a, b

(v) **Affine space:** $\{x : Ax = b\}$, for given A, b

(vi) **Polyhedron:** $\{x : Ax \leq b\}$, while inequality $\leq$ is interpreted componentwise. Note: the set $\{x : Ax \leq b, Cx = d\}$ is also a Polyhedron

(vii) **Simplex:** special case of polyhedra, given by $\text{conc}\{x_0, \cdots, x_k\}$, where these points are affinely independent. The canonical example is the **probability simplex**,

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

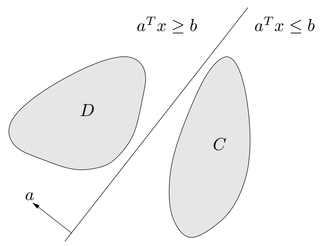
1.1.2 Key properties of convex sets:

(i) **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

$$C \subseteq \{x : a^T x \leq b\}$$
$$D \subseteq \{x : a^T x \geq b\}$$

(ii) **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 $x_0 \in \text{bd}(C)$, then there exists a such that

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



1.1.3 Operations preserving convexity

(i) **Intersection:** the intersection of convex sets is convex

(ii) **Scaling and translation:** if C is convex, then

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

is convex for any a, b

(iii) **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

(iv) **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 $C \subseteq \text{dom}(P)$ is convex then so is $P(C)$, and if D is convex then so is $P^{-1}(D)$

(v) **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 $C \subseteq \text{dom}(f)$ is convex then so if $f(C)$, and if D is convex then so is $f^{-1}(D)$

1.2 Cones

Definition (Cone)**.** $C \subseteq \mathbb{R}^n$ such that

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

Definition (Convex cone)**.** Cone that is also convex, i.e.,

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

Definition (Conic combination)**.** For $x_1, \cdots, x_k \in \mathbb{R}^n$, any linear combination

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

with $\theta_i \geq 0, i = 1, \cdots, k$.

Definition (Conic hull)**.** The conic hull of C, cone(C), is all conic combinations of elements.

1.2.1 Example of convex cones:

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

(ii) **Normal cone:** given any set C and point $x \in C$, we can define

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

This is always a convex cone, regardless of C

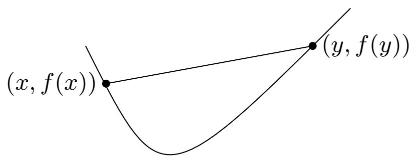
(iii) **Positive semidefinite cone:** $\mathbb{S}^n_+ = \{X \in \mathbb{S}^n : X \succeq 0\}$, where $X \succeq 0$ means that X is positive semidefinite (and $\mathbb{S}^n$ is the set of $n \times n$ symmetric matrices)

1.3 Convex function

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



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

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

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

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

Important modifiers:

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

(ii) **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

(iii) **Note:** strongly convex $\Rightarrow$ strictly convex $\Rightarrow$ convex (Analogously for concave functions)

1.3.1 Example of convex functions

(i) **Univariate functions:**

(a) Exponential function: e^{ax} is convex for any a over $\mathbb{R}$

(b) Power function: x^a is convex for $a \ge 1$ or $a \le 0$ over $\mathbb{R}_+$ (nonnegative reals)

(c) Power function: x^a is concave for $0 \le a \le 1$ over $\mathbb{R}_+$

(d) Logarithmic function: $\log(x)$ is concave over $\mathbb{R}_{++}$

(ii) **Affine function:** $a^T x + b$ is both convex and concave

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

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

(v) **Norm:** $\|x\|$ is convex for any norm; e.g. ℓ_p norms,

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

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

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

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

(vi) **Indicator function:** if C is convex, then its indicator function

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

(vii) **Support function:** for any set C(convex or not), its indicator function

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

is convex

(viii) **Max function:** $f(x) = \max\{x_1, \cdots, x_n\}$ is convex

1.3.2 Key properties of convex functions

(i) A function is convex if and only if its restriction to any line is convex

(ii) **Epigraph characterization:** a function f is convex if and only if its epigraph

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

is a convex set

(iii) **Convex sublevel sets:** if f is convex, then its sublevel sets

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

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

(iv) **First order characterization:** if f is differentiable, then f is convex if and only if dom(f) is convex, and

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

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

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

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

1.3.3 Operations preserving convexity

(i) **Nonnegative linear combination:** $f_1, \cdots, f_m$ convex implies $a_1 f_1 + \cdots + a_m f_m$ convex for any $a_1, \cdots, a_m \geq 0$

(ii) **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

(iii) **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

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

(v) **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 is concave
    f is concave if h is concave and nonincreasing, g is convex

(vi) **Vector composition:** suppose that

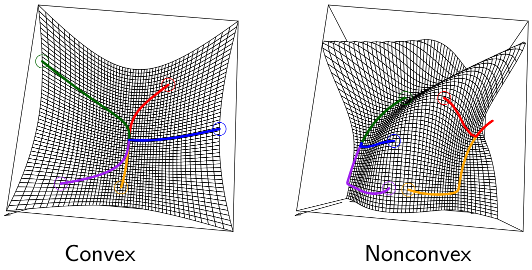
$$f(x) = h(g(x)) = h(g_1(x), \cdots, g_k(x))$$

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

2 Convexity II: Optimization Basics

2.1 Optimization terminology



Definition (Optimization problem).

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

here $D = \text{dom}(f) \cap \bigcap_{i=1}^{m} \text{dom}(g_i) \cap \bigcap_{j=1}^{r} \text{dom}(h_j)$, common domain of all functions.

Definition (Convex optimization problem).

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

where f and $g_i, i = 1, \cdots, m$ are all convex, and the optimization domain is $D = \text{dom}(f) \cap \bigcap_{i=1}^{m} \text{dom}(g_i)$.

For convex optimization problem, **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

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

Terminologies:

(i) f is called criterion or objective function

(ii) g_i is called inequality constraint function

(iii) If $x \in D, g_i(x) \leq 0, i = 1, \cdots, m$, and $Ax = b$ then x is called a feasible point

(iv) The minimum of $f(x)$ over all feasible points x is called the optimal value, written f^*

(v) If x is feasible and $f(x) = f^*$, then x is called optimal; also called a solution, or a minimizer

(vi) If x is feasible and $f(x) \leq f^* + \epsilon$, then x is called ϵ-suboptimal

(vii) If x is feasible and $g_i(x) = 0$, then we say g_i is active at x

(viii) Convex minimization can be reposed as concave maximization, i.e. $\min f(x) \Leftrightarrow \max -f(x)$, both are called convex optimization problems

(ix) the optimization problem can be rewritten as

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

where C is the feasible set. Hence the formulation is complete general. With I_C the indicator of C, it can also be written as the unconstrained form

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

Definition (Solution set)**.** Let X_{opt} be the set of all solutions of convex problem, written

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

Two key properties:

(i) X_{opt} is a convex set, and it can be proofed using definitions. If x, y are solutions, then for $0 \le t \le 1$, $t(x + (1-t)y)$ is also a solution

(ii) If f is strictly convex, then solution is unique, i.e., X_{opt} contains one element

2.2 First-order optimality condition

Definition (First-order optimality condition)**.** For a convex problem

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

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

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

this means all feasible directions from x are aligned with gradient $\nabla f(x)$

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

2.2.1 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

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

(i) if $Q \succeq 0$, then there is a unique solution $x = -Q^{-1} b$

(ii) if Q is singular and $b \notin \text{col}(Q)$, then there is no solution (i.e., $\min_x f(x) = -\infty$)

(iii) if Q is singular and $b \in \text{col}(Q)$, then there are infinitely many solutions

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

where Q^+ is the pseudoinverse of Q. Note: $\text{null}(Q) = \{z \in C^n : Qz = 0\}$

2.2.2 Example: equality-constrained minimization

Consider the equality-constrained convex problem

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

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

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

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

2.2.3 Example: projection onto a convex set

Consider projection onto convex set C

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

First-order optimality condition says that the solution x satisfies

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

Equivalent, this says that

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

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

2.3 Partial optimization

We can always partial optimize a convex problem and retain convexity, e.g.

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

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.

2.3.1 Excample: hinge form of SVMs

Recall the SVM problem

$$\begin{array}{ll} \min\limits_{\beta, \beta_0, \xi} & \frac{1}{2}\|\beta\|_2^2 + C \sum\limits_{i=1}^n \xi_i \\ \text{subject to} & \xi_i \geq 0, y_i \left(x_i^T \beta + \beta_0\right) \geq 1 - \xi_i, i = 1, \ldots, n \end{array}$$

Rewrite the constraints as $\xi_i \geq \max\{0, 1 - y_i \left(x_i^T \beta + \beta_0\right)\}$. Indeed we can argue that we have $=$ at solution. Therefor plugging in for optimal ξ gives the hinge form of SVMs:

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

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

2.4 Transformations and change of variables

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

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

Similarly, inequality or equality constraints can be transformed and yield equivalent optimization problem. This means we 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:

$$\min_x f(x) \quad \text{subject to } x \in C \iff \min_y f(\phi(y)) \quad \text{subject to } \phi(y) \in C$$

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

$$f(x) = \sum_{k=1}^{p} \gamma_k x_1^{a_k 1} x_2^{a_k 2} \cdots x_n^{a_k n}$$

A geometric program is of the form

$$\begin{array}{ll} \min\limits_x & f(x) \\ \text{subject to} & g_i(x) \leq 0, \ i = 1, \dots, m \\ & h_j(x) = 1, \ j = 1, \cdots, r \end{array}$$

where $f, g_i, i = 1, \cdots, m$ are posynomials and $h_j, j = 1, \cdots, r$ 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

$$\begin{array}{ll} \min\limits_x & \log(\sum\limits_{k=1}^{p_0} e^{a_{0k}^T y + b_{0k}}) \\ \text{subject to} & \log(\sum\limits_{k=1}^{p_i} e^{a_{ik}^T y + b_{ik}}) \leq 0, \ i = 1, \cdots, m \\ & c_j^T y + d_j = 0, \ j = 1, \cdots, r \end{array}$$

This is convex, recalling the convexity of soft max functions.

2.5 Eliminating equality constraints

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

$$\begin{array}{ll} \min\limits_x & f(x) \\ \text{subject to} & g_i(x) \leq 0, \ i = 1, \dots, m \\ & Ax = b \end{array}$$

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

$$\begin{array}{ll} \min\limits_{y} & f(My + x_0) \\ \text{subject to} & g_i(My + x_0) \leq 0,\ i = 1, \dots, m \end{array}$$

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

2.6 Introducing slack variables

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

$$\begin{array}{ll} \min\limits_{x} & f(x) \\ \text{subject to} & g_i(x) \leq 0,\ i = 1, \dots, m \\ & Ax = b \end{array}$$

we can transform the inequality constraints via

$$\begin{array}{ll} \min\limits_{x,s} & f(x) \\ \text{subject to} & s_i \geq 0,\ i = 1, \dots, m \\ & g_i(x) + s_i = 0,\ i = 1, \dots, m \\ & Ax = b \end{array}$$

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

2.7 Relaxing nonaffine equalities

Given an optimization problem

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

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

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

This is called a relaxation and its optimal value is always smaller or equal to that of the original problem. An important special case: relaxing nonaffine equality constraints, i.e.,

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

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

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

2.7.1 Example: maximum utility problem

The maximum utility problem models investment/consumption:

$$\begin{array}{ll} \min\limits_{x,b} & \sum_{t=1}^{T} \alpha_t u(x_t) \\ \text{subject to} & b_{t+1} = b_t + f(b_t) - x_t,\ t = 1, \dots, T \\ & 0 \leq x_t \leq b_t, t = 1, \cdots, T \end{array}$$

Here b_t is the budget and x_t is the amount consumed at time t; f is an investment return function, u utility function, both concave and increasing (Here, $f(0) = 0$, and $b_0 > 0$ is given).

2.7.2 Example: principal components analysis

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

$$\min_R \|X - R\|_F^2 \quad \text{subject to } \operatorname{rank}(R) = k$$

Here $\|A\|_F^2 = \sum_{i=1}^n \sum_{j=1}^p A_{ij}^2$, the entrywise squared ℓ_2 norm, and $\operatorname{rank}(A)$ denotes the rank of A. The problem is 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, but we can recast it. First rewrite as

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

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

$$C = \{Z \in \mathbb{S}^p : \lambda_i(Z) \in \{0, 1\}, i = 1, \cdots, p, \operatorname{tr}(Z) = k\}$$

where $\lambda_i(Z), i = 1, \cdots, n$ 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 $\mathcal{F}_k = \operatorname{conv}(V)$, its convex hull. Note

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

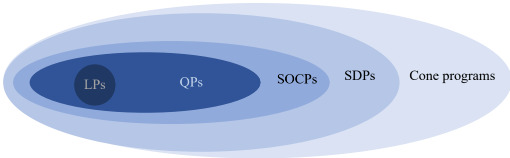
This set is called the Fantope of order k and it is convex. Hence, the linear maximization over the Fantope, namely

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

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

3 Canonical Problem Forms

The relationship among linear programs (LPs), quadratic programs (QPs), semidefinite programs (SDPs), second-order cone program (SOCPs) and cone programs is shown in the following figure.



3.1 Linear program

Definition (Linear program)**.** A linear program is an optimization problem of the form

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

and this is always a convex optimization problem.

The standard form of LP is

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

any LP can be rewritten in standard form.

3.1.1 Excample: basis pursuit

Given $y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times p}$, where $p > n$. Suppose that we seek the sparest solution to underdetermined linear system $X\beta = y$. If using ℓ_0 norm, it will be a nonconvex formulation. 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 problem can be reformulated as the LP form

$$\begin{aligned} \min_{\beta, z} \quad & 1^T z \\ \text{subject to} \quad & z \geq \beta \\ & z \geq -\beta \\ & X\beta = y \end{aligned}$$

3.1.2 Excample: Dantzig selector

Modification of previous problem, where we allow for $X\beta \approx y$, the Dantzig selector

$$\begin{aligned} \min_\beta \quad & \|\beta\|_1 \\ \text{subject to} \quad & |X^T(y - X\beta)\|_\infty \leq \lambda \end{aligned}$$

where $\lambda \geq 0$ is a tuning parameter. It can also be reformulated as a LP.

3.2 Convex quadratic program

Definition (Convex quadratic program)**.** A convex quadratic program is an optimization problem of the form

$$\begin{array}{ll} \min\limits_{x} & c^T x + \frac{1}{2} x^T Q x \\ \text{subject to} & Dx \leq d \\ & Ax = b \end{array}$$

where $Q \succeq 0$. Note that this problem is not convex when $Q \not\succeq 0$. From now on, when we say QP, we implicitly assume that $Q \succeq 0$.

The standard form of QP is

$$\begin{array}{ll} \min\limits_{x} & c^T x + \frac{1}{2} x^T Q x \\ \text{subject to} & x \geq 0 \\ & Ax = b \end{array}$$

any QP can be rewritten in standard form.

3.2.1 Example: support vector machines

Given $y \in \{-1, 1\}^n$, $X \in \mathbb{R}^{n \times p}$ having rows $x_1, \cdots, x_n$, recall the SVM problem

$$\begin{array}{ll} \min\limits_{\beta, \beta_0, \xi} & \frac{1}{2} \|\beta\|_2^2 + C \sum\limits_{i=1}^{n} \xi_i \\ \text{subject to} & \xi_i \geq 0, y_i \left(x_i^T \beta + \beta_0 \right) \geq 1 - \xi_i, i = 1, \ldots, n \end{array}$$

3.3 Semidefinite program

Definition (Semidefinite program)**.** A semidefinite program is an optimization problem of the form

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

where $F_j \in \mathbb{S}^d$, for $j = 0, 1, \cdots, n$, and $A \in \mathbb{R}^{m \times n}, c \in \mathbb{R}^n, b \in \mathbb{R}^m$.

The standard form of SDP is

$$\begin{array}{ll} \min\limits_{X} & C \bullet X \\ \text{subject to} & A_i \bullet X = b_i, i = 1, \cdots, m \\ & X \succeq 0 \end{array}$$

where $X \bullet Y = \text{tr}(XY)$ and any SDP can be rewritten in standard form.

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

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

$\mathbb{S}_{++}^n$ is the space of positive define matrices, i.e.,

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

3.3.1 Example: theta function

Let $G = (N, E)$ be an undirected graph, $N = \{1, \cdots, n\}$, $\omega(G)$ and $\chi(G)$ is the clique number and chromatic number of G, respectively. The Lovasz theta function is

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

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

3.3.2 Example: trace norm minimization

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

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

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

$$\begin{array}{ll} \min\limits_{X} & \text{rank}(X) \\ \text{subject to} & A(X) = b \end{array}$$

Trace norm approximation:

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

3.4 Conic program

Definition (Conic program)**.** A conic program is an optimization problem of the form

$$\begin{array}{ll} \min\limits_{x} & c^T x \\ \text{subject to} & D(x) + d \in K \\ & Ax = b \end{array}$$

where $c, x \in \mathbb{R}^n$, 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, $K = \mathbb{S}^n_+$.

Definition (Second-order cone program)**.** A second-order cone program or SOCP is an optimization problem of the form

$$\begin{array}{ll} \min\limits_{x} & c^T x \\ \text{subject to} & \|D_i(x) + d_i\|_2 \leq e_i^T x + f_i, i = 1, \cdots, p \\ & Ax = b \end{array}$$

This is indeed a cone program. Recall the second-order cone $Q = \{(x,t) : \|x\|_2 \leq t\}$, we have

$$\|D_i(x) + d_i\|_2 \leq e_i^T x + f_i \Longleftrightarrow (D_i(x) + d_i, e_i^T x + f_i) \in Q_i$$

for second-order cone Q_i of appropriate dimensions. Now take $K = Q_1 \times \cdots \times Q_p$.

Observe that every LP is an SOCP and every SOCP is an SDP. Turns out that

$$\|x\|_2 \leq t \Longleftrightarrow \begin{bmatrix} tI & x \\ x^T & t \end{bmatrix} \succeq 0$$

hence any SOCP constraint can be written as an SDP constraint. The above is a special case of the Schur complement theorem:

$$\begin{bmatrix} A & B \\ B^T & C \end{bmatrix} \succeq 0 \Longleftrightarrow A - BC^{-1}B^T \succeq 0$$

for A, C symmetric and $C \succ 0$.

In addition, QPs are SOCPs, which can be seen by rewriting a QP as

$$\begin{array}{ll} \min\limits_{x,t} & c^T x + t \\ \text{subject to} & Dx \leq d, \frac{1}{2}x^T Qx \leq t \\ & Ax = b \end{array}$$

now write $\frac{1}{2}x^T Qx \leq t \Leftrightarrow \left\|\left(\frac{1}{\sqrt{2}}Q^{1/2}x, \frac{1}{2}(1-t)\right)\right\|_2 \leq \frac{1}{2}(1+t)$. Thus we have established the hierachy.

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