IE 539: Convex Optimization **KAIST, Fall 2023**
Lecture #5: convex optimization problems II September 11, 2023
Lecturer: Dabeen Lee

1 Outline

In this lecture, we consider

- Convex optimization applications,
- Classes of convex programming problems I (linear programming)

2 Convex optimization applications

2.1 Portfolio optimization

In the last lecture, we learned that the following formulation models a portfolio optimization problem.

$$\begin{aligned} \text{maximize} \quad & \mu^\top x - \gamma x^\top \Sigma x \\ \text{subject to} \quad & 1^\top x = 1, \\ & x \in C' \end{aligned}$$

where

- μ is the vector of expected returns of financial assets,
- Σ is the covariance matrix of the financial assets' random returns,
- C' is a convex constraint set which could be $C' = \mathbb{R}^d_+$ (long positions only) or $C' = \{x \in \mathbb{R}^d : \|x\|_1 \leq B\}$ (bounded leverage),
- $\gamma > 0$ is the risk aversion parameter.

Note again that the formulation above is a convex optimization problem. First, the feasible region is the intersection of a hyperplane

$$\{x \in \mathbb{R}^d : \ 1^\top x = 1\}$$

and a convex set C'. Therefore, the feasible region is a convex set. Moreover, it is known that any covariance matrix is positive semidefinite. This in turn implies that the objective function $\mu^\top x - \gamma x^\top \Sigma x$ is a concave function. Moreover, observe that

$$\begin{aligned} &\max\left\{\mu^\top x - \gamma x^\top \Sigma x : \ 1^\top x = 1, \ x \in C'\right\} \\ &= -\min\left\{-\mu^\top x + \gamma x^\top \Sigma x : \ 1^\top x = 1, \ x \in C'\right\}. \end{aligned}$$

As $\mu^\top x - \gamma x^\top \Sigma x$ is concave, it follows that the function $-\mu^\top x + \gamma x^\top \Sigma x$ is convex. Thus, the minimization problem is indeed a convex optimization problem. From now on, we may skip this process and simply say that concave maximization is equivalent to convex minimization.

2.2 Uncertainty quantification

Recall that given a portfolio x and the covariance matrix Σ of financial assets, the quadratic term

$$x^\top \Sigma x$$

models the risk of the portfolio x. However, in practice, it is difficult to predict the exact value of Σ. Instead, we estimate it and deduce an empirical estimation of it. Let us denote it as $\bar{\Sigma}$. Then

$$x^\top \bar{\Sigma} x$$

is an estimation of the actual risk term $x^\top \Sigma x$. Although $x^\top \bar{\Sigma} x$ may provide a proxy for the risk, underestimation of risk could result in a critical situation. Then the question is, given the empirical estimate of the risk term $x^\top \bar{\Sigma} x$, what would be the worst-case risk value under estimation noise?

How do we consider possible estimation noise? One common way to use some statistics tool such as

$$\Bigg\| \underbrace{\Sigma}_{\text{true covariance}} - \underbrace{\bar{\Sigma}}_{\text{empirical covariance}} \Bigg\|_{\text{nuc}} \leq \epsilon$$

which holds for some ϵ (with high probability) where ϵ depends on the number of data to obtaine the estimate $\bar{\Sigma}$. Here, $\|\cdot\|_{\text{nuc}}$ denotes what is called the *nuclear norm* over matrices. Then we may draw a ball around the empirical covariance matrix with respect to the nuclear norm. Formally,

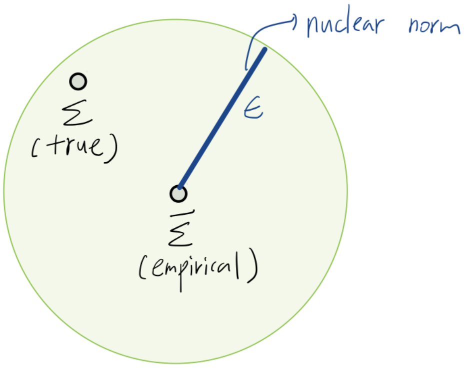

Figure 5.1: Nuclear-norm-ball around the empirical covariance matrix

we consider

$$\begin{aligned} &\left\{ A \in \mathbb{R}^{d\times d} : \ \|A - \bar{\Sigma}\|_{\text{nuc}} \leq \epsilon \right\} \\ &= \left\{ \bar{\Sigma} + S \in \mathbb{R}^{d\times d} : \ \|S\|_{\text{nuc}} \leq \epsilon \right\}. \end{aligned}$$

Then it follows from the above statistical result that the true covariance matrix Σ belongs to the ball (with high probability). Then now we consider

$$\begin{aligned} \text{maximize} \quad & x^\top (\bar{\Sigma} + S) x \\ \text{subject to} \quad & \bar{\Sigma} + S \succeq 0, \\ & \|S\|_{\text{nuc}} \leq \epsilon \end{aligned}$$

Note that $\Sigma = \bar{\Sigma} + S$ for some matrix S with $\|S\|_{\text{nuc}} \leq \epsilon$. Moreover, we know that Σ is positive semidefinite. Therefore,

$$x^\top \Sigma x \leq \max\left\{x^\top(\bar{\Sigma} + S)x : \ \bar{\Sigma} + S \succeq 0, \ \|S\|_{\text{nuc}} \leq \epsilon\right\}.$$

Hence, the optimum value provides an upper bound on the true risk value of portfolio x.

2.3 Support vector machine

Given n data $(x_1, y_1), \ldots, (x_n, y_n)$ where $y_i \in \{-1, 1\}$ are labels, we want to find a separating hyperplane

$$w^\top x = b$$

to classify data with $+1$ and data with -1. The goal is to find a separating hyperplane $w^\top x = b$ with the "gap" $(1/\|w\|_2)$ being maximized. Here, the gap means the Euclidean distance between two consecutive hyperplanes

$$\{x \in \mathbb{R}^d : \ w^\top x = b\}, \quad \{x \in \mathbb{R}^d : \ w^\top x = b + 1\}.$$

Then the problem can be formulated as

$$\begin{aligned} \text{minimize} \quad & \|w\|_2 \\ \text{subject to} \quad & y_i(w^\top x_i - b) \geq 0, \ i = 1, \ldots, n. \end{aligned}$$

If this problem is feasible, then $x \to \text{sign}(w^\top x_i - b)$ is a valid classifier for the data set.

What if the data set is not entirely separable? What if no hyperplane separates the data without an error? In such cases, we force separation via a penalty term, instead of imposing hard constraints. The number of misclassifications can be used as penalty. Namely,

$$\sum_{i=1}^n 1(y_i \neq \text{sign}(w^\top x_i - b)).$$

However, this is not convex. Instead, we apply the *hinge loss*[1], which is an upper bound on the number of misclassifications, given by

$$\sum_{i=1}^n \max\{0, \ 1 - y_i(w^\top x_i - b)\}.$$

Then we solve

$$\min_{w,b} \quad \lambda\|w\|_2^2 + \frac{1}{n}\sum_{i=1}^n \max\{0, \ 1 - y_i(w^\top x_i - b)\}$$

where λ determines the trade-off between the margin size and the penalty.

[1] Here, $\max\{0, a\}$ is called the hinge function.

2.4 LASSO (least absolute shrinkage and selection operator)

Based on n data points $(x_1, y_1), \ldots, (x_n, y_n)$, we want to find a linear rule

$$y = \beta^\top x$$

that best represents the relationship between x and y. The goal is to find β minimizing the "mean squared error", given by

$$\min_{\beta} \quad \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta^\top x_i)^2 \quad = \quad \min_{\beta} \quad \frac{1}{n}\|y - X\beta\|_2^2$$

where the rows of X are $x_1^\top, \ldots, x_n^\top$.

However, there are issues such as highly collinear covariates and overfitting. Motivated by this, LASSO is a regression method that achieves variable selection and regularization. The LASSO problem is to solve

$$\begin{aligned} \text{minimize} \quad & \frac{1}{n}\|y - X\beta\|_2^2 \\ \text{subject to} \quad & \|\beta\|_1 \leq t \end{aligned}$$

where t is a parameter determining the degree of regularization. Basically, the problem induces sparsity in β. The problem is often transformed into the *Lagrangian* form, given as follows.

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

where λ is set to control the degree of regularization.

2.5 Facility location

Given the locations of n households $x_1, \ldots, x_n \in \mathbb{R}^d$, we wish to build a hospital covering the households. A desired location would minimize the longest distance to a household. The problem can be formulated as

$$\min_{x} \quad \max_{i=1,\ldots,n} \|x - x_i\|.$$

3 Convex optimization hierarchy

On top of the applications we studied, there are many interesting "classes" of convex optimization problems. In this lecture, we consider them in this section, and specifically, we discuss

- Linear programming (LP),
- Quadratic programming (QP),
- Semidefinite programming (SDP),
- Conic programming,
- Second-order cone programming (SOCP).

In fact, these problems are closely related, and in fact, the problem classes form a hierarchy described in Figure 5.2

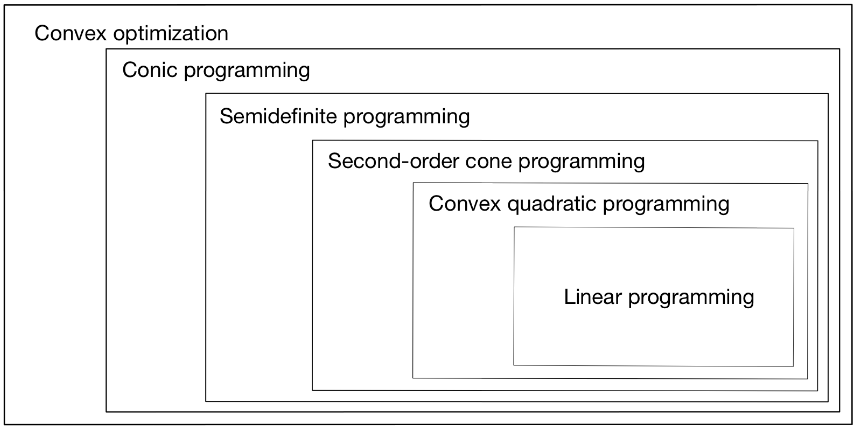

Figure 5.2: Hierarchy of classes of convex optimization problems

4 Linear programming

A linear program (LP) is an optimization problem of the following form.

$$\begin{aligned} \text{minimize} \quad & c^\top x \\ \text{subject to} \quad & Ax \geq b \end{aligned} \tag{P}$$

where $c^\top x$ is the linear objective function and $Ax \geq b$ is the system of linear constraints $a_1^\top x \geq b_1, \ldots, a_n^\top x \geq b_n$, i.e., $a_1^\top, \ldots, a_n^\top$ are the rows of A and $b = (b_1, \ldots, b_n)^\top$. Note that the feasible region $P = \{x \in \mathbb{R}^d : Ax \geq b\}$ is a polyhedron, the intersection of half-spaces $\{x \in \mathbb{R}^d : a_i^\top x \geq b_i\}$ for $i = 1, \ldots, n$. Hence, the linear program is the problem of finding a point in a polyhedron that minimizes a linear function. Figure 5.3 describes a linear program with 2 variables.

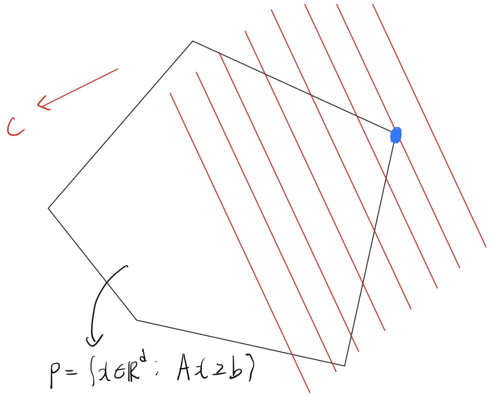

Figure 5.3: Geometric picture of a linear program

4.1 LP duality

The dual linear program of (P) is given by

$$\begin{aligned} \text{maximize} \quad & b^\top y \\ \text{subject to} \quad & A^\top y = c, \\ & y \geq 0. \end{aligned} \tag{D}$$

Weak LP duality states that $\mathrm{OPT}(P) \geq \mathrm{OPT}(D)$, where $\mathrm{OPT}(P)$ and $\mathrm{OPT}(D)$ denote the optimal values of (P) and (D), respectively. *Strong LP duality* states that if (P) is feasible and bounded, then (D) is feasible and bounded and $\mathrm{OPT}(P) = \mathrm{OPT}(D)$.

4.2 Example: facility location

Recall that the facility location problem can be modeled as follows.

$$\begin{aligned} \text{minimize} \quad & \max_{i=1,\ldots,n} \|x - x^i\|_1 \\ \text{subject to} \quad & x \in \mathbb{R}^d \end{aligned}$$

where $x^1, \ldots, x^n \in \mathbb{R}^d$ are the locations of households and we use the ℓ_1 norm for the norm $\|\cdot\|$. In fact, the problem is equivalent to a linear program. Why? We first introduce an auxiliary variable to replace the objective. Then the problem is equivalent to

$$\begin{aligned} \text{minimize} \quad & t \\ \text{subject to} \quad & t \geq \max_{i=1,\ldots,n} \|x - x^i\|_1. \end{aligned}$$

Basically, minimizing t forces minimizing the original objective $\max_{i=1,\ldots,n} \|x - x^i\|_1$. Moreover, $t \geq \max_{i=1,\ldots,n} \|x - x^i\|_1$ is equvialent to imposing $t \geq \|x - x^i\|_1$ for $i = 1, \ldots, n$. The next step is to replace $t \geq \|x - x^i\|_1$ by a set of linear inequalities. Note that

$$\|x - x^i\|_1 = \sum_{j=1}^d |x_j - x_j^i|$$

where $x = (x_1, \ldots, x_d)^\top$ and $x^i = (x_1^i, \ldots, x_d^i)^\top$. By introducing an auxiliary variable s_{ij} for each absolute value term $|x_j - x_j^i|$, we replace $t \geq \sum_{j=1}^d |x_j - x_j^i|$ by $t \geq \sum_{j=1}^d s_{ij}$ and $s_{ij} \geq |x_j - x_j^i|$. Moreover, $s_{ij} \geq |x_j - x_j^i|$ is equivalent to $s_{ij} \geq x_j - x_j^i \geq -s_{ij}$. Therefore, we obtain

$$\begin{aligned} \text{minimize} \quad & t \\ \text{subject to} \quad & t \geq \sum_{j=1}^d s_{ij} && \text{for } i = 1, \ldots, n, \\ & s_{ij} \geq x_j - x_j^i \geq -s_{ij} && \text{for } i = 1, \ldots, n,\ j = 1, \ldots, d \end{aligned}$$

which is a linear program.