TUP-Springer Project

Springer Optimization and Its Applications 103

李力 著
Li Li

凸优化应用讲义

Selected Applications of Convex Optimization

清华大学出版社

Springer

Springer Optimization and Its Applications 103

李力 著
Li Li

凸优化应用讲义

Selected Applications of Convex Optimization

清华大学出版社
北京

Springer

内 容 简 介

凸优化理论和方法能够解决一大类常见的优化问题。本书介绍了凸优化在支撑向量机、参数估计、范数逼近、控制器设计等问题中的应用，以期读者掌握将实际问题转换（或近似转换）成凸优化问题的基本知识和基本方法，能够灵活使用凸优化理论和方法解决实际问题。

本书潜在的读者包括运筹优化方向、机器学习方向、统计方向、控制方向、信号处理方向的研究生和高年级本科生。读者需对凸优化理论和线性代数理论有一定的了解。

图书在版编目(CIP)数据

凸优化应用讲义：英文/李力著. --北京：清华大学出版社，2015
ISBN 978-7-302-39029-9

Ⅰ. ①凸…　Ⅱ. ①李…　Ⅲ. ①凸分析－英文　Ⅳ. ①O174.13

中国版本图书馆 CIP 数据核字(2015)第 017185 号

责任编辑：王一玲
封面设计：傅瑞学
责任印制：王静怡

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　**邮　　编**：100084
社 总 机：010-62770175　**邮　　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课件下载：http://www.tup.com.cn，010-62775954
印 刷 者：北京鑫丰华彩印有限公司
装 订 者：三河市溧源装订厂
经　　销：全国新华书店
开　　本：155mm×235mm　**印　张**：9.5　**字　数**：166 千字
版　　次：2015 年 6 月第 1 版　**印　次**：2015 年 6 月第 1 次印刷
印　　数：1～1500
定　　价：69.00 元

产品编号：057858-01

Springer Optimization and Its Applications

Volume 103

Aims and Scope

Optimization has been expanding in all directions at an astonishing rate during the last few decades. New algorithmic and theoretical techniques have been developed, the diffusion into other disciplines has proceeded at a rapid pace, and our knowledge of all aspects of the field has grown even more profound. At the same time, one of the most striking trends in optimization is the constantly increasing emphasis on the interdisciplinary nature of the field. Optimization has been a basic tool in all areas of applied mathematics, engineering, medicine, economics, and other sciences.

The series *Springer Optimization and Its Applications* publishes undergraduate and graduate textbooks, monographs and state-of-the-art expository work that focus on algorithms for solving optimization problems and also study applications involving such problems. Some of the topics covered include nonlinear optimization (convex and nonconvex), network flow problems, stochastic optimization, optimal control, discrete optimization, multi-objective programming, description of software packages, approximation techniques and heuristic approaches.

More information about this series at http://www.springer.com/series/7393

To My Beloved Parents!

Preface

This book focuses on the applications of convex optimization, a special class of mathematical optimization techniques. Related problems arise in a variety of applications and can be numerically solved in an efficient way. Likewise, diversified applications have contributed to convex optimization and urged the development of new optimization techniques. This intergrowth continues to produce new achievements. John Tukey said, "*The best thing about being a statistician is that you get to play in everyone's backyard.*" While, in our humble opinion, "*The best thing about being a convex optimization researcher is that you get to help build many ones' backyards.*"

As with any science worth its salt, convex optimization has a coherent formal theory and a rambunctious experimental wing. There are already a number of textbooks on the theory of convex optimization. So, we address the applications in this book.

To thoroughly survey applications of convex optimization in this book is certainly impossible due to the width and depth of covered topics. Here, we emphasize our discussions on formulating empirical problems into formal convex optimization problems in six representative categories.

In Chap. 2, we introduce the supporting vector machines that are supervised learning models for data classification. In Chap. 3, we study the parameter estimation problems particularly on maximum likelihood estimation and expectation maximization algorithms. In Chap. 4, we discuss the norm approximation and some widely adopted regularization tricks with a special emphasis on recently hottest sparsity features. In Chap. 5, we present the positive semidefinite programming problems, linear matrix inequalities, and their applications in control theory. In Chap. 6, we give some interesting examples of convex relaxation methods. In Chap. 7, we visit some frequently used geometric problems that can be handled as convex optimization problems.

We believe that these representative applications well demonstrate the interplay of convex optimization theory and applications.

The selection of topics is significantly influenced by the valuable textbook *Convex Optimization* written by Prof. Stephen Boyd and Prof. Lieven Vandenberghe

(a book that we require all my graduate students to carefully read for at least one time). *Convex Optimization* contains much more topics than this book. However, we make notably deeper discussions in each category that are mentioned above and arrange some issues in a more synthetic way.

Similar to the writing purpose of *Convex Optimization*, our major goal is to help the reader develop the skills needed to recognize and formulate convex optimization problems that might be encountered in practice, since experiences of many researchers had proven the great advantages to recognizing or formulating a problem as a convex optimization problem. All derivation processes are presented in details so that readers can teach themselves without any difficulties.

The draft of this book had been used for five years in a graduate course "Convex Optimization and Applications" in Tsinghua University. The required background of readers includes a solid knowledge of advanced calculus, linear algebra, basic probability theory, and basic knowledge of convex optimization. This book is not a text primarily about convex analysis or the mathematics of convex optimization. Indeed, we hope this book will be used as a supplement textbook for several types of courses, including operations research, computer science, statistics, data mining, and many fields of science and engineering. Our teaching experiences showed that the covered materials also serve as useful references for beginners who are major in machine learning and control theory field. Moreover, upon the requests of many readers, we plan to enrich the topics of this book in the coming years so that it could serve students in other fields.

Finally, we would like to thank Prof. Stephen Boyd at the Department of Electrical Engineering, Stanford University, for the helpful discussions. We would like to thank Dr. Xiaoling Huang, Dr. Zhengpeng Wu, Mr. Kaidi Yang, Mr. Yingde Chen, as well as Mr. Wei Guo for typing some parts of this book and Mr. Jiajie Zhang for developing most Matlab code snippets as well as all figures for this book. We would also like to thank many of our students for debugging the typos of this book. Without their kind help, we cannot finish the tedious writing tasks in such a short time.

Due to our limited knowledge, this book might contain mistakes and typos. Please be kind to e-mail us when you find any problems within this book. We are more than happy to revise this book upon your notice and extend its contents in the following teaching process.

Beijing, China Li Li

Contents

Chapter 1
Preliminary Knowledge

Abstract Convex analysis and convex optimization are the basis for our following discussions. However, we will not recapitulate all the related issues in this book. Indeed, we just briefly review the minimum required preliminary knowledge. Several useful conclusions of linear algebra are also mentioned. Finally, we introduce the CVX toolbox, based on which we will write the sample Matlab code snippets in this book.

1.1 Nomenclatures

In this book, matrices are denoted by capital letters $A, B, C, \ldots, X, Y, Z$. The transpose of a matrix A is denoted as A^T. The conjugate transpose of a matrix A is denoted as A^*. The inverse of a matrix A is denoted as A^{-1}. The determinant of a matrix A is denoted as $\det(A)$. The trace of a matrix A is denoted as $\mathrm{Tr}(A)$.

Multivariate vectors are denoted by bold minuscule letters $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \ldots, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$. All vectors are column vectors, except we have special declarations.

One-dimensional variables are denoted by minuscule letters $a, b, c, \ldots, x, y, z$.

The sets are usually denoted by capital letters $\Gamma, \ldots, \Omega$.

$\mathbf{0}$ is a column vector of all 0 with appropriate dimensions, and $\mathbf{1}$ is a column vector of all 1 with appropriate dimensions. I often denotes certain identity matrices with proper dimensions.

A single bar $|\cdot|$ is used to denote a vector norm, absolute value, or complex modulus, while a double bar $\|\cdot\|$ is reserved for denoting a matrix norm.

The 0-norm of a vector $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]^T \in \mathbb{R}^n$ is defined as the number of the nonzero entries of $\boldsymbol{x}$.

The 1-norm of a vector $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]^T \in \mathbb{R}^n$ is defined as

$$|\boldsymbol{x}|_1 = \sum_{i=1}^{n} |x_i| \tag{1.1}$$

L. Li, *Selected Applications of Convex Optimization*, Springer Optimization and Its Applications 103, DOI 10.1007/978-3-662-46356-7_1

The 2-norm (also called Euclidean norm) of a vector $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]^T \in \mathbb{R}^n$ is defined as

$$|\boldsymbol{x}|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} \tag{1.2}$$

Clearly, we have

$$|\boldsymbol{x}|_2^2 = \sum_{i=1}^{n} x_i^2 = \boldsymbol{x}^T \boldsymbol{x} \tag{1.3}$$

The general ℓ_p-norm of a vector $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]^T \in \mathbb{R}^n$ with $p \in [1, +\infty)$ is defined as

$$|\boldsymbol{x}|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} \tag{1.4}$$

The ∞-norm of a $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]^T \in \mathbb{R}^n$ is defined as

$$|\boldsymbol{x}|_\infty = \max \{|x_i|\} \tag{1.5}$$

1.2 Convex Sets and Convex Functions

Definition 1.1 A set Ω is a *convex set* if for all $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ in Ω and $\lambda \in [0, 1]$, we have $\lambda \boldsymbol{x}_1 + (1 - \lambda)\boldsymbol{x}_2 \in \Omega$.

Geometrically, that means that given any two points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ in Ω, all points on the line segment joining $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are also in Ω.

Definition 1.2 Suppose $\Omega \subset \mathbb{R}^n$ is a convex set and $f(\boldsymbol{x}) : \Omega \mapsto \mathbb{R}$ is a *convex function*, if its epigraph (the set of points on or above the graph of the function) is a convex set. Or equivalently, if for all $\boldsymbol{x}, \boldsymbol{y} \in \Omega$ and $\lambda \in [0, 1]$, we have

$$f(\lambda \boldsymbol{x} + (1 - \lambda)\boldsymbol{y}) \leq \lambda f(\boldsymbol{x}) + (1 - \lambda) f(\boldsymbol{y}) \tag{1.6}$$

A function f is a *strictly convex function* if strict inequality holds in (1.6), whenever $\boldsymbol{x} \neq \boldsymbol{y}$ and $0 < \lambda < 1$.

Definition 1.3 Suppose $\Omega \subset \mathbb{R}^n$ is a convex set and $g(\boldsymbol{x}) : \Omega \mapsto \mathbb{R}$ is a *concave function*, if $f(\boldsymbol{x}) = -g(\boldsymbol{x}) : \Omega \mapsto \mathbb{R}$ is a convex function. Or equivalently, if for all $\boldsymbol{x}, \boldsymbol{y} \in \Omega$ and $\lambda \in [0, 1]$, we have

$$g(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \geq \lambda g(\boldsymbol{x}) + (1-\lambda) g(\boldsymbol{y}) \tag{1.7}$$

A function f is a *strictly concave function* if strict inequality holds in (1.7), whenever $\boldsymbol{x} \neq \boldsymbol{y}$ and $0 < \lambda < 1$.

The basic inequality (1.6)–(1.7) is sometimes called *Jensen's inequality* [5]. It can be easily extended to convex combinations of more than two points.

Theorem 1.1 (Jensen's Inequality) *Suppose $C \subset \mathbb{R}^n$ is a convex set, $f(\boldsymbol{x}) : C \mapsto \mathbb{R}$ is a convex function on its domain C, we have*

$$f(\lambda_1 \boldsymbol{x}_1 + \cdots + \lambda_k \boldsymbol{x}_k) \leq \lambda_1 f(\boldsymbol{x}) + \cdots + \lambda_k f(\boldsymbol{x}_k) \tag{1.8}$$

for all $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \in C$, $\lambda_i \geq 0$ and $\sum_{i=1}^m \lambda_i = 1$, $m \in \mathbb{N}$.

Proof Let us use mathematical induction to prove it. For the case $k = 2$, the statement is true by definition.

Assume the statement is true for the case $k \geq 2$, and then for the case $k+1$, we have

$$\begin{aligned}
& f(\lambda_1 \boldsymbol{x}_1 + \lambda_2 \boldsymbol{x}_2 + \cdots + \lambda_k \boldsymbol{x}_k + \lambda_{k+1} \boldsymbol{x}_{k+1}) \\
\leq\ & (1-\lambda_{k+1}) f\left(\frac{\lambda_1 \boldsymbol{x}_1 + \lambda_2 \boldsymbol{x}_2 + \cdots + \lambda_k \boldsymbol{x}_k}{1-\lambda_{k+1}}\right) + \lambda_{k+1} f(\boldsymbol{x}_{k+1}) \\
\leq\ & (1-\lambda_{k+1}) \left[\frac{\lambda_1}{1-\lambda_{k+1}} f(\boldsymbol{x}_1) + \frac{\lambda_2}{1-\lambda_{k+1}} f(\boldsymbol{x}_2) + \cdots + \frac{\lambda_k}{1-\lambda_{k+1}} f(\boldsymbol{x}_k)\right] \\
& + \lambda_{k+1} f(\boldsymbol{x}_{k+1}) \\
=\ & \lambda_1 f(\boldsymbol{x}_1) + \lambda_2 f(\boldsymbol{x}_2) + \cdots + \lambda_{k+1} f(\boldsymbol{x}_{k+1})
\end{aligned}$$

So, the statement holds true for the case $k+1$.

By mathematical induction, the statement holds for any $k \in \mathbb{N}\backslash\{1\}$. □

Corollary 1.1 *Suppose $C \subset \mathbb{R}^n$ is a convex set, $f(\boldsymbol{x}) : C \mapsto \mathbb{R}$ is a concave function on its domain C, and we have*

$$f(\lambda_1 \boldsymbol{x}_1 + \cdots + \lambda_k \boldsymbol{x}_k) \geq \lambda_1 f(\boldsymbol{x}) + \cdots + \lambda_k f(\boldsymbol{x}_k) \tag{1.9}$$

for all $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \in C$, $\lambda_i \geq 0$ and $\sum_{i=1}^m \lambda_i = 1$, $m \in \mathbb{N}$.

Based on Jensen's inequality, we can easily obtain the useful *Gibbs' inequality*.

Theorem 1.2 (Gibbs' Inequality) *Suppose that $\boldsymbol{p} = \{p_1, \ldots, p_n\}$ is a discrete probability distribution. Thus, we have $\sum_{i=1}^n p_i = 1$, $p_i > 0$ for $i = 1, \ldots, n$. For*

another discrete probability distribution $\boldsymbol{q} = \{q_1, \ldots, q_n\}$*, the following inequality holds*

$$-\sum_{i=1}^{n} p_i \ln p_i \leq -\sum_{i=1}^{n} p_i \ln q_i \tag{1.10}$$

with equality if and only if

$$p_i = q_i, \ \forall\, i \tag{1.11}$$

If $p(\boldsymbol{x})$ *is a probability distribution function for* $\boldsymbol{x}$ *on set* Ω *and* $q(\boldsymbol{x})$ *is another probability distribution function on* Ω*, the following inequality holds*

$$-\int_{\Omega} [p(\boldsymbol{x}) \ln p(\boldsymbol{x}) - p(\boldsymbol{x}) \ln q(\boldsymbol{x})]\, d\boldsymbol{x} \leq 0 \tag{1.12}$$

with equality if and only if

$$p(\boldsymbol{x}) = q(\boldsymbol{x}), \ \forall\, \boldsymbol{x} \tag{1.13}$$

Proof We only prove the discretized version and the proof for continuous version is similar.

Let Λ denote the set of all i for which p_i is nonzero. Since ln function is concave, we have

$$-\sum_{i\in\Lambda} p_i \ln \frac{q_i}{p_i} \geq -\ln\left(\sum_{i\in\Lambda} p_i \frac{q_i}{p_i}\right) = -\ln\left(\sum_{i\in\Lambda} q_i\right) \geq -\ln\left(\sum_{i} q_i\right) = -\ln 1 = 0 \tag{1.14}$$

So, we have

$$-\sum_{i\in\Lambda} p_i \ln q_i \geq -\sum_{i\in\Lambda} p_i \ln p_i \tag{1.15}$$

and therefore

$$-\sum_{i=1}^{n} p_i \ln q_i \geq -\sum_{i=1}^{n} p_i \ln p_i \tag{1.16}$$

since the right-hand side does not grow, and meanwhile the left-hand side may grow or remain unchanged.

Trivially, we can find the equality holds only if $p_i = q_i$, for $i = 1, \ldots, n$. □

Along with Gibbs' inequality, we can define a widely used measure of distribution differences as below [6, 7].

Definition 1.4 The nonnegative quantity $D\left(p(x) \,||\, q(x)\right)$ is called *Kullback-Leibler (K-L) divergence* of $q(x)$ from $p(x)$. For discrete probability distributions $p(x)$ and $q(x)$, the K-L divergence of $q(x)$ from $p(x)$ is defined to be

$$D\left(p(x) \,||\, q(x)\right) = \sum_i p_i \ln\left[\frac{p_i}{q_i}\right] \tag{1.17}$$

For continuous probability distributions $p(x)$ and $q(x)$, the K-L divergence of $q(x)$ from $p(x)$ is defined to be

$$D\left(p(x) \,||\, q(x)\right) = \int p(x) \ln\left[\frac{p(x)}{q(x)}\right] dx \tag{1.18}$$

We will discuss an important application of Kullback-Leibler divergence in Sect. 3.4.

More applications of Jensen's inequality can be found in [8].

1.3 Convex Optimization

1.3.1 Gradient Descent and Coordinate Descent

Let us consider a minimization problem without constraints

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) \tag{1.19}$$

where $\boldsymbol{x} \in \mathbb{R}^n$ is the optimization variable and the function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is the objective function.

Suppose we start from a point $\boldsymbol{x}_0 \in \mathbb{R}^n$. If the function $f(\boldsymbol{x})$ is defined and differentiable in a neighborhood of $\boldsymbol{x}_0$, then $f(\boldsymbol{x})$ decreases fastest if we move in the direction of the negative gradient of $\nabla f(\boldsymbol{x})$ at $\boldsymbol{x}_0$. When we move a small enough distance, in other words, for a small enough $\gamma_0 > 0$, we reach a new point $\boldsymbol{x}_1$

$$\boldsymbol{x}_1 = \boldsymbol{x}_0 - \gamma_0 \nabla f(\boldsymbol{x}_0) \tag{1.20}$$

and it follows that

$$f(\boldsymbol{x}_0) \geq f(\boldsymbol{x}_1) \tag{1.21}$$

Consider the sequence $\{\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ such that

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \gamma_k \nabla f(\boldsymbol{x}_k), \ \gamma_k > 0 \tag{1.22}$$

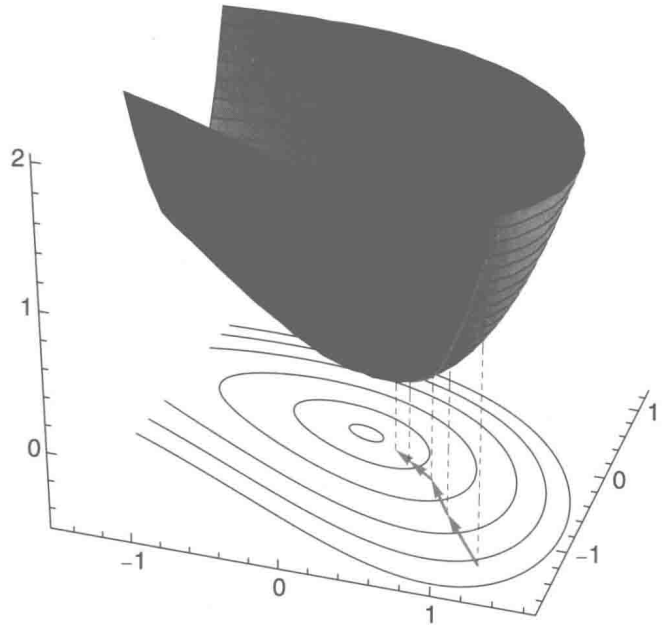

Fig. 1.1 An illustration of gradient descent algorithm

we have

$$f(\boldsymbol{x}_0) \geq f(\boldsymbol{x}_1) \geq \cdots f(\boldsymbol{x}_k) \geq f(\boldsymbol{x}_{k+1}) \geq \cdots \tag{1.23}$$

and hopefully this sequence converges to the desired local minimum [9]; see Fig. 1.1 for an illustration.

We call such search algorithm as *gradient descent* algorithm. When the function $f(\boldsymbol{x})$ is convex, all local minima are meanwhile the global minima, and gradient descent algorithm can converge to the global solution.

Another widely used search strategy is *coordinate descent* algorithm. It is based on the fact that the minimization of a multivariate function can be achieved by iteratively minimizing it along one direction at each time [10]. For the above problem (1.17), we can iterate through each direction, one at a time, minimizing the objective function with respect that coordinate direction as

$$\boldsymbol{x}_i^{k+1} = \arg\min_{y \in \mathbb{R}} f\left(x_1^{k+1}, \ldots, x_{i-1}^{k+1}, y, x_{i+1}^{k}, \ldots, x_n^{k}\right) \tag{1.24}$$

This will generate a sequence $\{\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k, \ldots\}$ such that $f(\boldsymbol{x}_0) \geq f(\boldsymbol{x}_1) \geq \cdots f(\boldsymbol{x}_k) \geq \cdots$. If this multivariate function is a convex function, this sequence will finally reach the global optimal solution; see Fig. 1.2 for an illustration.