

FIRST-ORDER METHODS IN OPTIMIZATION



MOS-SIAM Series on Optimization

This series is published jointly by the Mathematical Optimization Society and the Society for Industrial and Applied Mathematics. It includes research monographs, books on applications, textbooks at all levels, and tutorials. Besides being of high scientific quality, books in the series must advance the understanding and practice of optimization. They must also be written clearly and at an appropriate level for the intended audience.

Editor-in-Chief

Katya Scheinberg
Lehigh University

Editorial Board

Santanu S. Dey, *Georgia Institute of Technology*
Maryam Fazel, *University of Washington*
Andrea Lodi, *University of Bologna*
Arkadi Nemirovski, *Georgia Institute of Technology*
Stefan Ulbrich, *Technische Universität Darmstadt*
Luis Nunes Vicente, *University of Coimbra*
David Williamson, *Cornell University*
Stephen J. Wright, *University of Wisconsin*

Series Volumes

Beck, Amir, *First-Order Methods in Optimization*
Terlaky, Tamás, Anjos, Miguel F., and Ahmed, Shabbir, editors, *Advances and Trends in Optimization with Engineering Applications*
Todd, Michael J., *Minimum-Volume Ellipsoids: Theory and Algorithms*
Bienstock, Daniel, *Electrical Transmission System Cascades and Vulnerability: An Operations Research Viewpoint*
Koch, Thorsten, Hiller, Benjamin, Pfetsch, Marc E., and Schewe, Lars, editors, *Evaluating Gas Network Capacities*
Corberán, Ángel, and Laporte, Gilbert, *Arc Routing: Problems, Methods, and Applications*
Toth, Paolo, and Vigo, Daniele, *Vehicle Routing: Problems, Methods, and Applications, Second Edition*
Beck, Amir, *Introduction to Nonlinear Optimization: Theory, Algorithms, and Applications with MATLAB*
Attouch, Hedy, Buttazzo, Giuseppe, and Michaille, Gérard, *Variational Analysis in Sobolev and BV Spaces: Applications to PDEs and Optimization, Second Edition*
Shapiro, Alexander, Dentcheva, Darinka, and Ruszczyński, Andrzej, *Lectures on Stochastic Programming: Modeling and Theory, Second Edition*
Locatelli, Marco and Schoen, Fabio, *Global Optimization: Theory, Algorithms, and Applications*
De Loera, Jesús A., Hemmecke, Raymond, and Köppe, Matthias, *Algebraic and Geometric Ideas in the Theory of Discrete Optimization*
Blekherman, Grigoriy, Parrilo, Pablo A., and Thomas, Rekha R., editors, *Semidefinite Optimization and Convex Algebraic Geometry*
Delfour, M. C., *Introduction to Optimization and Semidifferential Calculus*
Ulbrich, Michael, *Semismooth Newton Methods for Variational Inequalities and Constrained Optimization Problems in Function Spaces*
Biegler, Lorenz T., *Nonlinear Programming: Concepts, Algorithms, and Applications to Chemical Processes*
Shapiro, Alexander, Dentcheva, Darinka, and Ruszczyński, Andrzej, *Lectures on Stochastic Programming: Modeling and Theory*
Conn, Andrew R., Scheinberg, Katya, and Vicente, Luis N., *Introduction to Derivative-Free Optimization*
Ferris, Michael C., Mangasarian, Olvi L., and Wright, Stephen J., *Linear Programming with MATLAB*
Attouch, Hedy, Buttazzo, Giuseppe, and Michaille, Gérard, *Variational Analysis in Sobolev and BV Spaces: Applications to PDEs and Optimization*
Wallace, Stein W. and Ziemba, William T., editors, *Applications of Stochastic Programming*
Grötschel, Martin, editor, *The Sharpest Cut: The Impact of Manfred Padberg and His Work*
Renegar, James, *A Mathematical View of Interior-Point Methods in Convex Optimization*
Ben-Tal, Aharon and Nemirovski, Arkadi, *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications*
Conn, Andrew R., Gould, Nicholas I. M., and Toint, Philippe L., *Trust-Region Methods*

FIRST-ORDER METHODS IN OPTIMIZATION

Amir Beck

Tel-Aviv University
Tel-Aviv
Israel



Society for Industrial and Applied Mathematics
Philadelphia



Mathematical
Optimization Society

Mathematical Optimization Society
Philadelphia

Copyright © 2017 by the Society for Industrial and Applied Mathematics and the Mathematical Optimization Society

10 9 8 7 6 5 4 3 2 1

All rights reserved. Printed in the United States of America. No part of this book may be reproduced, stored, or transmitted in any manner without the written permission of the publisher. For information, write to the Society for Industrial and Applied Mathematics, 3600 Market Street, 6th Floor, Philadelphia, PA 19104-2688 USA.

Trademarked names may be used in this book without the inclusion of a trademark symbol. These names are used in an editorial context only; no infringement of trademark is intended.

<i>Publisher</i>	Kivmars Bowling
<i>Acquisitions Editor</i>	Paula Callaghan
<i>Developmental Editor</i>	Gina Rinelli Harris
<i>Managing Editor</i>	Kelly Thomas
<i>Production Editor</i>	Louis R. Primus
<i>Copy Editor</i>	Bruce Owens
<i>Production Manager</i>	Donna Witzleben
<i>Production Coordinator</i>	Cally Shrader
<i>Compositor</i>	Cheryl Hufnagle
<i>Graphic Designer</i>	Lois Sellers

Library of Congress Cataloging-in-Publication Data

Please visit www.siam.org/books/mo25 to view the CIP data.



For
My wife, Nili
My daughters, Noy and Vered
My parents, Nili and Itzhak



Contents

Preface	xi
1 Vector Spaces	1
1.1 Definition	1
1.2 Dimension	2
1.3 Norms	2
1.4 Inner Products	2
1.5 Affine Sets and Convex Sets	3
1.6 Euclidean Spaces	3
1.7 The Space \mathbb{R}^n	4
1.8 The Space $\mathbb{R}^{m \times n}$	6
1.9 Cartesian Product of Vector Spaces	7
1.10 Linear Transformations	8
1.11 The Dual Space	9
1.12 The Bidual Space	10
1.13 Adjoint Transformations	11
1.14 Norms of Linear Transformations	12
2 Extended Real-Valued Functions	13
2.1 Extended Real-Valued Functions and Closedness	13
2.2 Closedness versus Continuity	18
2.3 Convex Functions	21
2.4 Support Functions	26
3 Subgradients	35
3.1 Definitions and First Examples	35
3.2 Properties of the Subdifferential Set	39
3.3 Directional Derivatives	44
3.4 Computing Subgradients	53
3.5 The Value Function	67
3.6 Lipschitz Continuity and Boundedness of Subgradients	71
3.7 Optimality Conditions	72
3.8 Summary of Weak and Strong Subgradient Calculus Results	84
4 Conjugate Functions	87
4.1 Definition and Basic Properties	87
4.2 The Biconjugate	89

4.3	Conjugate Calculus Rules	91
4.4	Examples	93
4.5	Infimal Convolution and Conjugacy	102
4.6	Subdifferentials of Conjugate Functions	104
5	Smoothness and Strong Convexity	107
5.1	L -Smooth Functions	107
5.2	Strong Convexity	117
5.3	Smoothness and Strong Convexity Correspondence	123
6	The Proximal Operator	129
6.1	Definition, Existence, and Uniqueness	129
6.2	First Set of Examples of Proximal Mappings	131
6.3	Prox Calculus Rules	135
6.4	Prox of Indicators—Orthogonal Projections	146
6.5	The Second Prox Theorem	157
6.6	Moreau Decomposition	160
6.7	The Moreau Envelope	163
6.8	Miscellaneous Prox Computations	170
6.9	Summary of Prox Computations	177
7	Spectral Functions	179
7.1	Symmetric Functions	179
7.2	Symmetric Spectral Functions over \mathbb{S}^n	182
7.3	Symmetric Spectral Functions over $\mathbb{R}^{m \times n}$	188
8	Primal and Dual Projected Subgradient Methods	195
8.1	From Gradient Descent to Subgradient Descent	195
8.2	The Projected Subgradient Method	201
8.3	The Stochastic Projected Subgradient Method	221
8.4	The Incremental Projected Subgradient Method	229
8.5	The Dual Projected Subgradient Method	232
9	Mirror Descent	247
9.1	From Projected Subgradient to Mirror Descent	247
9.2	Convergence Analysis	252
9.3	Mirror Descent for the Composite Model	260
10	The Proximal Gradient Method	269
10.1	The Composite Model	269
10.2	The Proximal Gradient Method	270
10.3	Analysis of the Proximal Gradient Method—The Nonconvex Case	272
10.4	Analysis of the Proximal Gradient Method—The Convex Case	281
10.5	The Proximal Point Method	288
10.6	Convergence of the Proximal Gradient Method—The Strongly Convex Case	288
10.7	The Fast Proximal Gradient Method—FISTA	290
10.8	Smoothing	304

10.9	Non-Euclidean Proximal Gradient Methods	315
11	The Block Proximal Gradient Method	331
11.1	Decomposition Methods	331
11.2	Model and Assumptions	332
11.3	The Toolbox	333
11.4	The Cyclic Block Proximal Gradient Method	338
11.5	The Randomized Block Proximal Gradient Method	347
12	Dual-Based Proximal Gradient Methods	353
12.1	The Primal and Dual Models	353
12.2	The Dual Proximal Gradient Method	355
12.3	Fast Dual Proximal Gradient	358
12.4	Examples I	360
12.5	The Dual Block Proximal Gradient Method	369
12.6	Examples II	374
13	The Generalized Conditional Gradient Method	379
13.1	The Frank–Wolfe/Conditional Gradient Method	379
13.2	The Generalized Conditional Gradient Method	380
13.3	The Strongly Convex Case	391
13.4	The Randomized Generalized Block Conditional Gradient Method	400
14	Alternating Minimization	405
14.1	The Method	405
14.2	Coordinate-wise Minima	407
14.3	The Composite Model	411
14.4	Convergence in the Convex Case	413
14.5	Rate of Convergence in the Convex Case	415
15	ADMM	423
15.1	The Augmented Lagrangian Method	423
15.2	Alternating Direction Method of Multipliers (ADMM)	425
15.3	Convergence Analysis of AD-PMM	427
15.4	Minimizing $f_1(\mathbf{x}) + f_2(\mathbf{Ax})$	432
A	Strong Duality and Optimality Conditions	439
B	Tables	443
C	Symbols and Notation	451
D	Bibliographic Notes	457
	Bibliography	463
	Index	473

Preface

This book, as the title suggests, is about first-order methods, namely, methods that exploit information on values and gradients/subgradients (but not Hessians) of the functions comprising the model under consideration. First-order methods go back to 1847 with the work of Cauchy on the steepest descent method. With the increase in the amount of applications that can be modeled as large- or even huge-scale optimization problems, there has been a revived interest in using simple methods that require low iteration cost as well as low memory storage.

The primary goal of the book is to provide in a self-contained manner a comprehensive study of the main first-order methods that are frequently used in solving large-scale problems. This is done by gathering and reorganizing in a unified manner many results that are currently scattered throughout the literature. Special emphasis is placed on rates of convergence and complexity analysis. Although the name of the book is “first-order methods in optimization,” two disclaimers are in order. First, we will actually also consider methods that exploit additional operations at each iteration such as prox evaluations, linear oracles, exact minimization w.r.t. blocks of variables, and more, so perhaps a more suitable name would have been “simple methods in optimization.” Second, in order to be truly self-contained, the first part of the book (Chapters 1–7) is actually purely theoretical and contains essential topics that are crucial for the developments in the algorithmic part (Chapters 8–15).

The book is intended for students and researchers with a background in advanced calculus and linear algebra, as well as prior knowledge in the fundamentals of optimization (some convex analysis, optimality conditions, and duality). A MATLAB toolbox implementing many of the algorithms described in the book was developed by the author and Nili Guttman-Beck and can be found at www.siam.org/books/mo25.

The outline of the book is as follows. Chapter 1 reviews important facts about vector spaces. Although the material is quite fundamental, it is advisable not to skip this chapter since many of the conventions regarding the underlying spaces used in the book are explained. Chapter 2 focuses on extended real-valued functions with a special emphasis on properties such as convexity, closedness, and continuity. Chapter 3 covers the topic of subgradients starting from basic definitions, continuing with directional derivatives, differentiability, and subdifferentiability and ending with calculus rules. Optimality conditions are derived for convex problems (Fermat’s optimality condition), but also for the nonconvex composite model, which will be discussed extensively throughout the book. Conjugate functions are the subject of Chapter 4, which covers several issues, such as Fenchel’s

inequality, the biconjugate, calculus rules, conjugate subgradient theorem, relations with the infimal convolution, and Fenchel's duality theorem. Chapter 5 covers two different but closely related subjects: smoothness and strong convexity—several characterizations of each of these concepts are given, and their relation via the conjugate correspondence theorem is established. The proximal operator is discussed in Chapter 6, which includes a large amount of prox computations as well as calculus rules. The basic properties of the proximal mapping (first and second prox theorems and Moreau decomposition) are proved, and the Moreau envelope concludes the theoretical part of the chapter. The first part of the book ends with Chapter 7, which contains a study of symmetric spectral functions. The second, algorithmic part of the book starts with Chapter 8 with primal and dual projected subgradient methods. Several stepsize rules are discussed, and complexity results for both the convex and the strongly convex cases are established. The chapter also includes discussions on the stochastic as well as the incremental projected subgradient methods. The non-Euclidean version of the projected subgradient method, a.k.a. the mirror descent method, is discussed in Chapter 9. Chapter 10 is concerned with the proximal gradient method as well as its many variants and extensions. The chapter also studies several theoretical results concerning the so-called gradient mapping, which plays an important part in the convergence analysis of proximal gradient-based methods. The extension of the proximal gradient method to the *block* proximal gradient method is discussed in Chapter 11, while Chapter 12 considers the dual proximal gradient method and contains a result on a primal-dual relation that allows one to transfer rate of convergence results from the dual problem to the primal problem. The generalized conditional gradient method is the topic of Chapter 13, which contains the basic rate of convergence results of the method, as well as its block version, and discusses the effect of strong convexity assumptions on the model. The alternating minimization method is the subject of Chapter 14, where its convergence (as well as divergence) in many settings is established and illustrated. The book concludes with a discussion on the ADMM method in Chapter 15.

My deepest thanks to Marc Teboulle, whose fundamental works in first-order methods form the basis of many of the results in the book. Marc introduced me to the world of optimization, and he is a constant source and inspiration and admiration. I would like to thank Luba Tetruashvili for reading the book and for her helpful remarks. It has been a pleasure to work with the extremely devoted and efficient SIAM staff. Finally, I would like to acknowledge the support of the Israel Science Foundation for supporting me while writing this book.