

Optimization for Data Analysis

Optimization techniques are at the core of data science, including data analysis and machine learning. An understanding of basic optimization techniques and their fundamental properties provides important grounding for students, researchers, and practitioners in these areas. This text covers the fundamentals of optimization algorithms in a compact, self-contained way, focusing on the techniques most relevant to data science. An introductory chapter demonstrates that many standard problems in data science can be formulated as optimization problems. Next, many fundamental methods in optimization are described and analyzed, including gradient and accelerated gradient methods for unconstrained optimization of smooth (especially convex) functions; the stochastic gradient method, a workhorse algorithm in machine learning; the coordinate descent approach; several key algorithms for constrained optimization problems; algorithms for minimizing nonsmooth functions arising in data science; foundations of the analysis of nonsmooth functions and optimization duality; and the back-propagation approach, relevant to neural networks.

STEPHEN J. WRIGHT holds the George B. Dantzig Professorship, the Sheldon Lubar Chair, and the Amar and Balinder Sohi Professorship of Computer Sciences at the University of Wisconsin–Madison. He is a Discovery Fellow in the Wisconsin Institute for Discovery and works in computational optimization and its applications to data science and many other areas of science and engineering. Wright is also a fellow of the Society for Industrial and Applied Mathematics (SIAM) and recipient of the 2014 W. R. G. Baker Award from IEEE for most outstanding paper, the 2020 Khachiyan Prize by the INFORMS Optimization Society for lifetime achievements in optimization, and the 2020 NeurIPS Test of Time award. He is the author and coauthor of widely used textbooks and reference books in optimization, including *Primal Dual Interior-Point Methods* and *Numerical Optimization*.

BENJAMIN RECHT is Associate Professor in the Department of Electrical Engineering and Computer Sciences at the University of California, Berkeley. His research group studies how to make machine learning systems more robust to interactions with a dynamic and uncertain world by using mathematical tools from optimization, statistics, and dynamical systems. Recht is the recipient of a Presidential Early Career Award for Scientists and Engineers, an Alfred P. Sloan Research Fellowship, the 2012 SIAM/MOS Lagrange Prize in Continuous Optimization, the 2014 Jamon Prize, the 2015 William O. Baker Award for Initiatives in Research, and the 2017 and 2020 NeurIPS Test of Time awards.

Cambridge University Press
978-1-316-51898-4 — Optimization for Data Analysis
Stephen J. Wright , Benjamin Recht
Frontmatter
[More Information](#)

Optimization for Data Analysis

STEPHEN J. WRIGHT
University of Wisconsin–Madison

BENJAMIN RECHT
University of California, Berkeley



Cambridge University Press
978-1-316-51898-4 — Optimization for Data Analysis
Stephen J. Wright, Benjamin Recht
Frontmatter
[More Information](#)

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom
One Liberty Plaza, 20th Floor, New York, NY 10006, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre,
New Delhi – 110025, India
103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org

Information on this title: www.cambridge.org/9781316518984

DOI: 10.1017/9781009004282

© Stephen J. Wright and Benjamin Recht 2022

This publication is in copyright. Subject to statutory exception and to the provisions of relevant collective licensing agreements, no reproduction of any part may take place without the written permission of Cambridge University Press.

First published 2022

Printed in the United Kingdom by TJ Books Ltd, Padstow Cornwall

A catalogue record for this publication is available from the British Library.

Library of Congress Cataloging-in-Publication Data

Names: Wright, Stephen J., 1960– author. | Recht, Benjamin, author.

Title: Optimization for data analysis / Stephen J. Wright and Benjamin Recht.

Description: New York : Cambridge University Press, [2021] | Includes bibliographical references and index.

Identifiers: LCCN 2021028671 (print) | LCCN 2021028672 (ebook) |

ISBN 9781316518984 (hardback) | ISBN 9781009004282 (epub)

Subjects: LCSH: Big data. | Mathematical optimization. | Quantitative research. | Artificial intelligence. | BISAC: MATHEMATICS / General | MATHEMATICS / General

Classification: LCC QA76.9.B45 W75 2021 (print) | LCC QA76.9.B45 (ebook) | DDC 005.7–dc23

LC record available at <https://lcn.loc.gov/2021028671>

LC ebook record available at <https://lcn.loc.gov/2021028672>

ISBN 978-1-316-51898-4 Hardback

Cambridge University Press has no responsibility for the persistence or accuracy of URLs for external or third-party internet websites referred to in this publication and does not guarantee that any content on such websites is, or will remain, accurate or appropriate.

Cover image courtesy of © Isaac Sparks

Contents

<i>Preface</i>	<i>page ix</i>
1 Introduction	1
1.1 Data Analysis and Optimization	1
1.2 Least Squares	4
1.3 Matrix Factorization Problems	5
1.4 Support Vector Machines	6
1.5 Logistic Regression	9
1.6 Deep Learning	11
1.7 Emphasis	13
2 Foundations of Smooth Optimization	15
2.1 A Taxonomy of Solutions to Optimization Problems	15
2.2 Taylor's Theorem	16
2.3 Characterizing Minima of Smooth Functions	18
2.4 Convex Sets and Functions	20
2.5 Strongly Convex Functions	22
3 Descent Methods	26
3.1 Descent Directions	27
3.2 Steepest-Descent Method	28
3.2.1 General Case	28
3.2.2 Convex Case	29
3.2.3 Strongly Convex Case	30
3.2.4 Comparison between Rates	32
3.3 Descent Methods: Convergence	33
3.4 Line-Search Methods: Choosing the Direction	36
3.5 Line-Search Methods: Choosing the Steplength	38

3.6	Convergence to Approximate Second-Order Necessary Points	42
3.7	Mirror Descent	44
3.8	The KL and PL Properties	51
4	Gradient Methods Using Momentum	55
4.1	Motivation from Differential Equations	56
4.2	Nesterov's Method: Convex Quadratics	58
4.3	Convergence for Strongly Convex Functions	62
4.4	Convergence for Weakly Convex Functions	66
4.5	Conjugate Gradient Methods	68
4.6	Lower Bounds on Convergence Rates	70
5	Stochastic Gradient	75
5.1	Examples and Motivation	76
5.1.1	Noisy Gradients	76
5.1.2	Incremental Gradient Method	77
5.1.3	Classification and the Perceptron	77
5.1.4	Empirical Risk Minimization	78
5.2	Randomness and Steplength: Insights	80
5.2.1	Example: Computing a Mean	80
5.2.2	The Randomized Kaczmarz Method	82
5.3	Key Assumptions for Convergence Analysis	85
5.3.1	Case 1: Bounded Gradients: $L_g = 0$	86
5.3.2	Case 2: Randomized Kaczmarz: $B = 0, L_g > 0$	86
5.3.3	Case 3: Additive Gaussian Noise	86
5.3.4	Case 4: Incremental Gradient	87
5.4	Convergence Analysis	87
5.4.1	Case 1: $L_g = 0$	89
5.4.2	Case 2: $B = 0$	90
5.4.3	Case 3: B and L_g Both Nonzero	92
5.5	Implementation Aspects	93
5.5.1	Epochs	93
5.5.2	Minibatching	94
5.5.3	Acceleration Using Momentum	94
6	Coordinate Descent	100
6.1	Coordinate Descent in Machine Learning	101
6.2	Coordinate Descent for Smooth Convex Functions	103
6.2.1	Lipschitz Constants	104
6.2.2	Randomized CD: Sampling with Replacement	105
6.2.3	Cyclic CD	110

6.2.4	Random Permutations CD: Sampling without Replacement	112
6.3	Block-Coordinate Descent	113
7	First-Order Methods for Constrained Optimization	118
7.1	Optimality Conditions	118
7.2	Euclidean Projection	120
7.3	The Projected Gradient Algorithm	122
7.3.1	General Case: A Short-Step Approach	123
7.3.2	General Case: Backtracking	124
7.3.3	Smooth Strongly Convex Case	125
7.3.4	Momentum Variants	126
7.3.5	Alternative Search Directions	126
7.4	The Conditional Gradient (Frank–Wolfe) Method	127
8	Nonsmooth Functions and Subgradients	132
8.1	Subgradients and Subdifferentials	134
8.2	The Subdifferential and Directional Derivatives	137
8.3	Calculus of Subdifferentials	141
8.4	Convex Sets and Convex Constrained Optimization	144
8.5	Optimality Conditions for Composite Nonsmooth Functions	146
8.6	Proximal Operators and the Moreau Envelope	148
9	Nonsmooth Optimization Methods	153
9.1	Subgradient Descent	155
9.2	The Subgradient Method	156
9.2.1	Steplengths	158
9.3	Proximal-Gradient Algorithms for Regularized Optimization	160
9.3.1	Convergence Rate for Convex f	162
9.4	Proximal Coordinate Descent for Structured Nonsmooth Functions	164
9.5	Proximal Point Method	167
10	Duality and Algorithms	170
10.1	Quadratic Penalty Function	170
10.2	Lagrangians and Duality	172
10.3	First-Order Optimality Conditions	174
10.4	Strong Duality	178
10.5	Dual Algorithms	179
10.5.1	Dual Subgradient	179
10.5.2	Augmented Lagrangian Method	180

10.5.3	Alternating Direction Method of Multipliers	181
10.6	Some Applications of Dual Algorithms	182
10.6.1	Consensus Optimization	182
10.6.2	Utility Maximization	184
10.6.3	Linear and Quadratic Programming	185
11	Differentiation and Adjoints	188
11.1	The Chain Rule for a Nested Composition of Vector Functions	188
11.2	The Method of Adjoints	190
11.3	Adjoints in Deep Learning	191
11.4	Automatic Differentiation	192
11.5	Derivations via the Lagrangian and Implicit Function Theorem	195
11.5.1	A Constrained Optimization Formulation of the Progressive Function	195
11.5.2	A General Perspective on Unconstrained and Constrained Formulations	197
11.5.3	Extension: Control	197
	Appendix	200
A.1	Definitions and Basic Concepts	200
A.2	Convergence Rates and Iteration Complexity	203
A.3	Algorithm 3.1 Is an Effective Line-Search Technique	204
A.4	Linear Programming Duality, Theorems of the Alternative	205
A.5	Limiting Feasible Directions	208
A.6	Separation Results	209
A.7	Bounds for Degenerate Quadratic Functions	213
	<i>Bibliography</i>	216
	<i>Index</i>	223

Preface

Optimization formulations and algorithms have long played a central role in data analysis and machine learning. Maximum likelihood concepts date to Gauss and Laplace in the late 1700s; problems of this type drove developments in unconstrained optimization in the latter half of the 20th century. Mangasarian's papers in the 1960s on pattern separation using linear programming made an explicit connection between machine learning and optimization in the early days of the former subject. During the 1990s, optimization techniques (especially quadratic programming and duality) were key to the development of support vector machines and kernel learning. The period 1997–2010 saw many synergies emerge between regularized / sparse optimization, variable selection, and compressed sensing. In the current era of deep learning, two optimization techniques—stochastic gradient and automatic differentiation (a.k.a. back-propagation)—are essential.

This book is an introduction to the basics of continuous optimization, with an emphasis on techniques that are relevant to data analysis and machine learning. We discuss basic algorithms, with analysis of their convergence and complexity properties, mostly (though not exclusively) for the case of convex problems. An introductory chapter provides an overview of the use of optimization in modern data analysis, and the final chapter on differentiation provides several perspectives on gradient calculation for functions that arise in deep learning and control. The chapters in between discuss gradient methods, including accelerated gradient and stochastic gradient; coordinate descent methods; gradient methods for problems with simple constraints; theory and algorithms for problems with convex nonsmooth terms; and duality-based methods for constrained optimization problems. The material is suitable for a one-quarter or one-semester class at advanced undergraduate or early graduate level. We and our colleagues have made extensive use of drafts of this material in the latter setting.

This book has been a work in progress since about 2010, when we began to revamp our optimization courses, trying to balance the viewpoints of practical optimization techniques against renewed interest in non-asymptotic analyses of optimization algorithms. At that time, the flavor of analysis of optimization algorithms was shifting to include a greater emphasis on worst-case complexity. But algorithms were being judged more by their worst-case bounds rather than by their performance on practical problems in applied sciences. This book occupies a middle ground between analysis and practice.

Beginning with our courses CS726 and CS730 at University of Wisconsin, we began writing notes, problems, and drafts. After Ben moved to UC Berkeley in 2013, these notes became the core of the class EECS227C. Our material drew heavily from the evolving theoretical understanding of optimization algorithms. For instance, in several parts of the text, we have made use of the excellent slides written and refined over many years by Lieven Vandenberghé for the UCLA course ECE236C. Our presentation of accelerated methods reflects a trend in viewing optimization algorithms as dynamical systems, and was heavily influenced by collaborative work with Laurent Lessard and Andrew Packard. In choosing what material to include, we tried to not be distracted by methods that are not widely used in practice but also to highlight how theory can guide algorithm selection and design by applied researchers.

We are indebted to many other colleagues whose input shaped the material in this book. Moritz Hardt initially inspired us to try to write down our views after we presented a review of optimization algorithms at the bootcamp for the Simons Institute Program on Big Data in Fall 2013. He has subsequently provided feedback on the presentation and organization of drafts of this book. Ashia Wilson was Ben's TA in EECS227C, and her input and notes helped us to clarify our pedagogical messages in several ways. More recently, Martin Wainwright taught EECS227C and provided helpful feedback, and Jelena Diakonikolas provided corrections for the early chapters after she taught CS726. André Wibisono provided perspectives on accelerated gradient methods, and Ching-pei Lee gave useful advice on coordinate descent. We are also indebted to the many students who took CS726 and CS730 at Wisconsin and EECS227C at Berkeley who found typos and beta-tested homework problems, and who continue to make this material a joy to teach. Finally, we would like to thank the Simons Institute for supporting us on multiple occasions, including Fall 2017 when we both participated in their program on Optimization.

Madison, Wisconsin, USA
Berkeley, California, USA