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.

https://doi.org/10.1017/9781009004282.001 Published online by Cambridge University Press

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 Vandenberghe 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