Preface

Algorithms have never been more important. As the recipes of computer programs, algorithms rule our lives. Although they can be forces for both good and evil, this is not a book about ethics. It is about thoughtful algorithm design. Poorly designed algorithms offer vague solutions to vague problems. Well-designed algorithms have a clear objective in mind. A good algorithm makes efficient use of computer time and storage and terminates in a finite number of steps with a guaranteed solution or an adequate approximate solution. Many problems can be posed as optimization of some objective. To the man or woman on the street, optimality involves selecting the best or being the best. Physicists, mathematicians, statisticians, and indeed scientists and engineers of all stripes know that optimality considerations underpin all of physical reality. The famous quote of Leonhard Euler summarizes the consensus: "...nothing at all takes place in the universe in which some rule of maximum or minimum does not appear."

The MM principle is a device for creating optimization algorithms satisfying the ascent or descent property. In minimizing an objective function, an MM algorithm operates on a simpler surrogate function that majorizes the objective. Majorization is understood here to mean a combination of tangency and domination. Minimizing the surrogate drives the objective downhill. The celebrated EM (expectation-maximization) principle [154] of computational statistics is a special case of the MM principle that depends on missing data, either concretely or abstractly. Despite the lesser generality of the EM principle, the literature on EM algorithms far outpaces the literature on MM algorithms. As of the beginning of 2016, the EM paper of Dempster, Laird, and Rubin [66] is the second most cited paper in statistics.

Tracing the origin of an idea is never easy. It can take decades to distill the essence and discard the inessentials. Specific MM and EM algorithms appeared years before the MM principle was well understood [100, 152, 190, 207, 214]. The entire category of projected gradient and proximal gradient algorithms [91, 143] can be motivated from the MM perspective, but the early emphasis on operators and fixed points obscured this distinction. The MM principle was clearly stated in the numerical analysis text of Ortega and Rheinboldt [168] in 1970. Unfortunately, they did not pursue practical applications. The year 1977 saw the publication of two seminal papers. Dempster et al. [66] formally named the EM algorithm and applied it to a wide variety of statistical problems. De Leeuw [53] understood the MM principle divorced from missing data, but he focused entirely on multidimensional scaling. Much of the EM machinery was anticipated by Baum [8], Beale and Little [10], Orchard and Woodbury [166], and Sundberg [196]. The EM principle had an immediate and substantial impact on computational statistics. The more general MM principle was much slower to take hold. The papers [60, 55, 105, 121] by the Dutch school of statisticians solidified its position in psychometrics and econometrics. In the early literature the MM principle went by the less descriptive name "iterative majoriza-

tion." Despite the painfully gradual recognition of the MM principle, it has now safely emerged from obscurity [114, 135, 212].

Nonetheless, there is a real need for a sustained exposition of the MM principle. Creation of EM algorithms calls on skills in calculating conditional expectations. Design of MM algorithms more often requires dexterity with inequalities and acquaintance with the convex calculus. This distinction partially explains the rapid advance of the EM algorithm among statisticians. Owing to the many fine expositions of the EM algorithm and my desire for brevity, the EM algorithm is downplayed here. The relevant chapters from my previous books [130, 131, 132] will reassure the reader of my lack of prejudice in this matter. Despite my obvious omissions, many traditional EM algorithms are derived in the text from the MM perspective. This is not simply a matter of one-upmanship. Alternative derivations clarify known results and suggest new applications.

My hope in writing this book is to stimulate the creation of new MM algorithms. In high-dimensional problems, standard methods of optimization such as scoring or Newton's method are simply impractical. The matrices that specify curvature are too big to store and invert. Likewise, the penalties and constraints that enforce sparsity and parsimony are inconsistent with smooth optimization. These vexing limitations have forced a re-evaluation of optimization theory and practice that will play out over the next few decades. The scientific and technical worlds need new tools for solving non-traditional problems. The MM principle is an obvious candidate. Readers should also keep in mind that the MM principle plays well with other methods of optimization. For instance, the ECM algorithm [156] combines the EM principle and block ascent.

This book is part textbook and part research monograph. Graduate students and even ambitious undergraduates can learn from it, not just MM theory but also general optimization theory as well. The exercises at the end of each chapter are arranged thematically rather than by level of difficulty. Many exercises are supplied with hints. I am occasionally surprised by students' novel solutions that ignore the hints. The prerequisite for reading the book is a good background in elementary analysis, linear algebra, and statistics. Despite my best intentions, the book is not entirely self-contained. Puzzled readers are advised to visit Wikipedia as needed. The appendices cover interesting extensions and background some readers may lack.

Chapter 1 highlights some easy applications of the MM principle. All readers should start here; many will stop here with no permanent harm. Chapters 2 and 3 review material on convexity, inequalities, nonsmooth analysis, and convex calculus. These topics are coming to the fore, and no serious theoretician can avoid their mastery. Experts will regret my many omissions, and novices will most likely feel overwhelmed with the pace. Unfortunately, there is no royal road to mathematics and no substitute for working through the details of proofs and exercises. Chapter 4 collects many of the available techniques for majorization and minorization, including the missing data paradigm. Chapter 5 features proximal maps and projection operators and their applications to the construction of MM algorithms. Proximal operators and Fenchel conjugates encode the solutions to entire families of optimization problems. Chapter 6 focuses on applications to regression and multivariate analysis. Statisticians and data scientists will want to spend time in this unruly garden. Chapter 7, the final chapter, provides an introduction to convergence theory and algorithm acceleration. The emphasis here is on breadth and not depth.

For the record, let me mention some notational conventions and computing choices. All vectors and matrices appear in boldface. The * superscript indicates a vector or matrix transpose. The Euclidean norm of a vector $\boldsymbol{x}$ is denoted by $||\boldsymbol{x}||$ and the spectral and Frobenius norms of a matrix $\boldsymbol{M}$ by $||\boldsymbol{M}||$ and $||\boldsymbol{M}||_F$, respectively. All positive semidefinite matrices are symmetric by definition. For a smooth real-valued function $f(\boldsymbol{x})$, we write

its gradient (column vector of partial derivatives) as $\nabla f(\boldsymbol{x})$, its first differential (row vector of derivatives) as $df(\boldsymbol{x}) = \nabla f(\boldsymbol{x})^*$, and its second differential (Hessian matrix) as $d^2 f(\boldsymbol{x})$. Many of the algorithms featured have been programmed in Julia, a new language that may well supplant MATLAB® and R over the next decade. These programs can be downloaded from the SIAM website at www.siam.org/books/ot147.

Many colleagues had a hand in making this a better book. My former postdoctoral fellows Eric Chi, Tongtong Wu, and Hua Zhou suggested topics, helped with figures, and carefully proofread the text. UCLA Biomathematics students Kevin Keys, Julian Landaw, Alfonso Landeros, Joe Larson, Mauricio Loya, Bhaven Mistry, Lindsay Riley, Trevor Shaddox, Tim Stutz, Elif Tekin, Yuxi Tian, Rory Wasiolek, and Song Xu suffered through earlier drafts and corrected a host of mistakes. Let me single out Kevin, Tim, and Rory for their special editorial efforts. My debt to my former UCLA colleague Jan de Leeuw is enormous. He more than anyone else deserves the credit for launching the MM algorithm. My exposition of Hadamard semidifferentials and the KKT rule was heavily influenced by Michel Delfour of the Université de Montréal. Our correspondence was a delight. Finally, let me thank the Stanford University Statistics Department for hosting my 2014–2015 sabbatical, during which I wrote the majority of the text.

I dedicate this book to my former mentors Gian-Carlo Rota, Carol Newton, and Robert Elston. Gian-Carlo was a wonderful doctoral advisor, urbane, philosophical, generous, a splendid writer, and an inventive mathematician. Carol guided my conversion from pure to applied mathematics. She had a vision of computational biology long before it became popular. I admired her for her kindness, wise counsel, and persistence. Robert, the only surviving member of the trio, guided my growth in genetics and applied statistics. He was there for me at a crucial juncture of my career. I look with amazement on all the lives Robert touched with such deftness and grace, and I wonder whether I will ever come remotely close in my own mentoring. I hope you, the reader, may benefit from the kind of angelic guidance I received. It makes a difference.