Alternating Direction Method of Multipliers for Machine Learning

Zhouchen Lin • Huan Li • Cong Fang

Alternating Direction Method of Multipliers for Machine Learning

Springer

Zhouchen Lin
Key Lab. of Machine Perception (MoE)
School of Artificial Intelligence
Peking University
Beijing, China

Huan Li
Institute of Robotics and Automatic
Information Systems
College of Artificial Intelligence
Nankai University
Tianjin, China

Cong Fang
Key Lab. of Machine Perception (MoE)
School of Artificial Intelligence
Peking University
Beijing, China

ISBN 978-981-16-9839-2 ISBN 978-981-16-9840-8 (eBook)
https://doi.org/10.1007/978-981-16-9840-8

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

To our families. Without your great support, this book would not exist and even our careers would be meaningless.

Foreword

Alternating direction method of multipliers (ADMM) is an important algorithm for solving constrained optimization problems. It particularly fits well for the machine learning community because the latter basically favors algorithms with low per-iteration cost and does not need high numerical precision. Due to its versatility and high usability, I would not hesitate to make it one of the top recommendations if one wants to develop a general-purpose optimization library or AI chip. So there has been renewed interest on ADMM since its successful application in solving low-rank models around 2010. Since then, ADMM has been extended significantly, going far beyond the traditional setting: deterministic, convex, two-blocks of variables, and centralized. Unfortunately, the vast literature on ADMM is scattered across various sources, making it difficult for non-experts to track the advances in this important optimization technique. This book resolves this issue in a timely manner. Its materials are quite comprehensive, covering ADMM for various situations: convex (and deterministic), nonconvex, stochastic, and distributed. It is self-contained, with detailed proofs, so that even a beginner can grasp the state-of-the-art quickly, not just the pseudo-codes but also the proof techniques. More importantly, this book has not simply compiled various papers together. It has actually completely rewritten the materials so that the notations are consistent and the deductions are smooth, removing the major obstacle of reading existing literatures. In addition, the book puts more emphasis on convergence *rates* rather than only convergence. This makes the theoretical analysis extremely informative to practitioners. I would say that this book is definitely a valuable reference for researchers and practitioners from multiple areas, including optimization, signal processing, and machine learning.

The authors, Zhouchen Lin, Huan Li, and Cong Fang, are experts in the intersection of optimization and machine learning. Besides contributing greatly to this field with technical papers, they have also endeavored a lot in sorting out valuable algorithms that are fit for engineers, making another kind of good contribution to the community. After their previous book, *Accelerated Optimization for Machine Learning: First-Order Algorithms*, which I like very much, I am happy to advocate their book once more.

Xi'an Jiaotong University, Xi'an, China Zongben Xu
October 2021

Foreword

With the advance of sensor, communication, and storage technologies, data acquisition has become more ubiquitous than any time in the past. This has enabled us to learn a considerable amount of valuable information from big diverse data sets for effective inference, estimation, tracking, and decision-making. Learning from data requires the proper modelling and analysis of big data sets, which are usually formulated as optimization problems. Consequently, large-scale optimization involving big data has attracted significant attention from various areas, including signal processing, machine learning, and operations research.

To minimize a cost function involving a large number of variables, the most popular approach is block coordinate descent/minimization (sometimes also called alternating optimization). However, if variables are coupled linearly, the block coordinate descent/minimization method no longer works. The alternating direction method of multipliers (ADMM) can be considered an extension of block coordinate descent/minimization method for linearly constrained optimization problems. Given the abundance of application problems that can be cast in the form of linearly constrained optimization problems (convex/nonconvex, smooth/nonsmooth), ADMM has been the method of choice for machine learning and signal processing problems involving big data. It is widely (sometimes wildly) applied in many different contexts, often times without sufficient theoretical underpinning on its convergence.

This book provides an excellent summary of the state of the art for the theoretical research on ADMM. It introduces the basic mathematical form of ADMM as well as its variations. The core material is on the convergence analysis of ADMM for different classes of linearly constrained optimization problems, including convex, nonconvex, deterministic, stochastic, and centralized/distributed. The mathematical treatment is concise, up to date, and rigorous. A nice bonus is the last chapter where the practical aspects of ADMM are discussed, which should be very valuable for practitioners or first-time users of ADMM.

The first author is a well-known researcher in optimization, particularly on optimization methods for machine learning applications. The text is written in a reader friendly manner, complete with appendices that introduce the mathematical tools and background for the convergence analysis covered in the book. It should be a valuable reference for both researchers and users on ADMM and will be a great read for graduate students in optimization, statistics, machine learning, and signal processing.

The Chinese University of Hong Kong, Shenzhen, China Zhi-Quan Luo
November 2021

Preface

Alternating direction method of multipliers (ADMM) is a magic algorithm to me. In my biased opinion, it is more or less a universal method for solving a wide range of constrained problems that ordinary practitioners in machine learning may encounter. Unlike gradient descent, which is roughly a universal method for unconstrained problems, ADMM appears to be more elegant yet less transparent. The secret lies in the Lagrange multiplier, which temporarily makes the constrained problem unconstrained, not only removing the difficulty in handling the constraints but also overcoming some inherent defects of the penalty method and the projected gradient descent, while non-experts are much easier to think of the latter two methods. The Lagrange multiplier also plays a central role in the proofs of convergence and convergence rate of ADMM. With possible exaggeration, I would say that one who cannot appreciate the beauty of ADMM cannot be a good researcher in optimization.

Since my first encounter with ADMM around 2009, I have seen that more and more machine learning people are using ADMM and extending its scope of applications. I also benefited a lot from and contributed a bit to the new developments. Yet, I also found that many engineers are not using ADMM correctly (the most notable example is to naively apply the ADMM for two blocks of variables to problems with more than two blocks). There has been an excellent tutorial on ADMM, *Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers*, written by Boyd et al. in 2011. Nonetheless, it is now 10 years old and does not cover the new developments, which I actually think are more important than the traditional ones for the wider applications of ADMM, because the new developments were done out of demands from real applications in signal processing and machine learning. So, I think that writing a new book on ADMM will be very useful for many people, including myself when teaching and advising students. My goal is to incorporate the most important aspects of the new developments in ADMM, rather than being confined to the traditional materials, which are typically for convex and two-block cases. Clearly, I am unable to review

Prof. Bingsheng He

all papers on ADMM. So, the strategy is to choose representative algorithms by their types (e.g., convex, nonconvex, stochastic, and distributed) instead. As a result, one should not be surprised that some variants of ADMM are not included (because they are not the most representative ones of their types but just discuss in more depth, or are too complex to use or analyze) while some variants of ADMM appear to be rather crude but they are still included (because that type of ADMM is less explored). Of course, personal flavor and limited knowledge also matter greatly. Finally, being self-contained is also very important. So, I also want to present proofs in detail.

I truly feel lucky as my former PhD students, Huan Li and Cong Fang, agreed to join this task even though they have graduated, and I have tortured them in the previous book, *Accelerated Optimization for Machine Learning: First-Order Algorithms*. I am also very lucky that more students contributed to the proofreading, including checking the proofs thoroughly (most of the proofs have undergone rewriting, rather than being directly copied from corresponding papers), which made the work less daunting. Nonetheless, the book is still far from being perfect. One of the major regrets is that using adaptive penalty is critical for speeding up convergence (see Sect. 7.1.2), but all the algorithms introduced in this book use a fixed penalty. Actually, most of the literatures do not consider adaptive penalty. Although it is quite possible that some of the algorithms introduced in this book may also work with adaptive penalties, we are unable to test which adaptive penalty strategy to use and then rewrite the proofs for adaptive penalties (actually drastic changes in the proofs may be necessary). The other regret is that we have to leave out learning-based ADMM, an emerging yet immature type of ADMM, as the theoretical guarantees are weak.

I expect that there will still be errors in the book despite the great efforts from my students and myself. So, if the readers detect any problem, please feel free to write an email to zclin2000@hotmail.com.

Finally, I would like to pay tribute to Prof. Bingsheng He. He has devoted most of his life to ADMM and contributed significantly to the research and education of ADMM. This book also introduces many of his works. I am glad to see that he has been well recognized in China, manifested by winning the "Operations Research" Research Award of the Operations Research Society of China in 2014. However, he is much less recognized internationally. I hope that my advocacy here could add to his credit.

Peking University, Beijing, China Zhouchen Lin
October 2021

Acknowledgments

We would like to thank all our collaborators and friends, especially Bingsheng He, Licheng Jiao, Junchi Li, Qing Ling, Guangcan Liu, Risheng Liu, Yuanyuan Liu, Canyi Lu, Zhi-Quan Luo, Yi Ma, Fanhua Shang, Zaiwen Wen, Xingyu Xie, Zongben Xu, Shuicheng Yan, Wotao Yin, Xiaoming Yuan, Xiaotong Yuan, Yaxiang Yuan, Tong Zhang, and Pan Zhou. We also thank Yiming Dong, Dr. Yuqing Hou, and Prof. Hongyang Zhang for proofreading, Yikang Li, Qiuhao Wang, Zhoutong Wu, Tong Yang, and Pengyun Yue for carefully checking all the proofs (in particular, Qiuhao Wang gave a much simpler proof for Lemma 3.16) and Celine Chang from Springer for helping the publication of the book. The authors are truly honored to have forewords written by Prof. Zongben Xu and Prof. Zhi-Quan Luo. This book is supported by National Natural Science Foundation of China under Grant Nos. 61625301 and 61731018.

About the Book

This book introduces the basic concepts of ADMM and its latest progress. Specifically, it introduces various ADMMs under different scenarios: convex and deterministic ADMM (Chap. 3), nonconvex and deterministic ADMM (Chap. 4), stochastic ADMM (Chap. 5), and distributed ADMM (Chap. 6). To make the book self-contained, it gives the detailed proofs of the convergence rates for most of the introduced algorithms.

This book serves as a useful reference to the recent advances in ADMM. It is appropriate for graduate students and researchers who are interested in optimization or practitioners who seek a powerful tool for optimization.

Contents

Acronyms

AAAI	Association for the Advancement of Artificial Intelligence
Acc-SADMM	Accelerated Stochastic Alternating Direction Method of Multipliers
ADMM	Alternating Direction Method of Multipliers
ALM	Augmented Lagrangian Method
DRS	Douglas-Rachford Splitting
KKT	Karush-Kuhn-Tucker
LADMM	Linearized Alternating Direction Method of Multipliers
MISO	Minimization by Incremental Surrogate Optimization
RPCA	Robust Principal Component Analysis
SADMM	Stochastic Alternating Direction Method of Multipliers
SAG	Stochastic Average Gradient
SDCA	Stochastic Dual Coordinate Ascent
SGD	Stochastic Gradient Descent
SPIDER	Stochastic Path-Integrated Differential EstimatoR
SVD	Singular Value Decomposition
SVRG	Stochastic Variance Reduced Gradient
VR	Variance Reduction