Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers
Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers

Stephen Boyd  
*Stanford University, USA*  
boyd@stanford.edu

Neal Parikh  
*Stanford University, USA*  
npparikh@cs.stanford.edu

Eric Chu  
*Stanford University, USA*  
echu508@stanford.edu

Borja Peleato  
*Stanford University, USA*  
peleato@stanford.edu

Jonathan Eckstein  
*Rutgers University, USA*  
jeckstei@rci.rutgers.edu

now

the essence of knowledge

Boston – Delft

Full text available at: http://dx.doi.org/10.1561/2200000016
Foundations and Trends® in
Machine Learning
Volume 3 Issue 1, 2010
Editorial Board

Editor-in-Chief:
Michael Jordan
Department of Electrical Engineering and Computer Science
Department of Statistics
University of California, Berkeley
Berkeley, CA 94720-1776

Editors

Peter Bartlett (UC Berkeley)                                       John Lafferty (Carnegie Mellon University)
Yoshua Bengio (Université de Montréal)                           Michael Littman (Rutgers University)
Avrim Blum (Carnegie Mellon University)                          Gabor Lugosi (Pompeu Fabra University)
Craig Boutilier (University of Toronto)                           David Madigan (Columbia University)
Stephen Boyd (Stanford University)                               Pascal Massart (Université de Paris-Sud)
Carla Brodley (Tufts University)                                  Andrew McCallum (University of
Inderjit Dhillon (University of Texas at Massachusetts Amherst)
Austin)                                                          Marina Meila (University of Washington)
Jerome Friedman (Stanford University)                             Andrew Moore (Carnegie Mellon
Kenji Fukumizu (Institute of Statistical University)              University)
Mathematics)                                                      John Platt (Microsoft Research)
Zoubin Ghahramani (Cambridge Luc de Raedt (Albert-Ludwigs Universitaet
University)                                                      Freiburg)
David Heckerman (Microsoft Research)                             Christian Robert (Université
Tom Heskes (Radboud University Nijmegen)                         Paris-Dauphine)
Geoffrey Hinton (University of Toronto)                           Sunita Sarawagi (IIT Bombay)
Aapo Hyvarinen (Helsinki Institute for Robert Schapire (Princeton University)
Information Technology)                                          Bernhard Schoelkopf (Max Planck Institute)
Leslie Pack Kaelbling (MIT)                                      Richard Sutton (University of Alberta)
Michael Kearns (University of Larry Wasserman (Carnegie Mellon
Pennsylvania)                                                    University)
Daphne Koller (Stanford University)                              Bin Yu (UC Berkeley)
Editorial Scope

Foundations and Trends® in Machine Learning will publish survey and tutorial articles in the following topics:

- Adaptive control and signal processing
- Applications and case studies
- Behavioral, cognitive and neural learning
- Bayesian learning
- Classification and prediction
- Clustering
- Data mining
- Dimensionality reduction
- Evaluation
- Game theoretic learning
- Graphical models
- Independent component analysis
- Inductive logic programming
- Kernel methods
- Markov chain Monte Carlo
- Model choice
- Nonparametric methods
- Online learning
- Optimization
- Reinforcement learning
- Relational learning
- Robustness
- Spectral methods
- Statistical learning theory
- Variational inference
- Visualization

Information for Librarians

Foundations and Trends® in Machine Learning, 2010, Volume 3, 4 issues. ISSN paper version 1935-8237. ISSN online version 1935-8245. Also available as a combined paper and online subscription.
Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers

Stephen Boyd\textsuperscript{1}, Neal Parikh\textsuperscript{2}, Eric Chu\textsuperscript{3} 
Borja Peleato\textsuperscript{4} and Jonathan Eckstein\textsuperscript{5}

\textsuperscript{1} Electrical Engineering Department, Stanford University, Stanford, CA 94305, USA, boyd@stanford.edu
\textsuperscript{2} Computer Science Department, Stanford University, Stanford, CA 94305, USA, npparikh@cs.stanford.edu
\textsuperscript{3} Electrical Engineering Department, Stanford University, Stanford, CA 94305, USA, echu508@stanford.edu
\textsuperscript{4} Electrical Engineering Department, Stanford University, Stanford, CA 94305, USA, peleato@stanford.edu
\textsuperscript{5} Management Science and Information Systems Department and RUTCOR, Rutgers University, Piscataway, NJ 08854, USA, jeckstei@rci.rutgers.edu

Abstract

Many problems of recent interest in statistics and machine learning can be posed in the framework of convex optimization. Due to the explosion in size and complexity of modern datasets, it is increasingly important to be able to solve problems with a very large number of features or training examples. As a result, both the decentralized collection or storage of these datasets as well as accompanying distributed solution methods are either necessary or at least highly desirable. In this
review, we argue that the alternating direction method of multipliers is well suited to distributed convex optimization, and in particular to large-scale problems arising in statistics, machine learning, and related areas. The method was developed in the 1970s, with roots in the 1950s, and is equivalent or closely related to many other algorithms, such as dual decomposition, the method of multipliers, Douglas–Rachford splitting, Spingarn’s method of partial inverses, Dykstra’s alternating projections, Bregman iterative algorithms for $\ell_1$ problems, proximal methods, and others. After briefly surveying the theory and history of the algorithm, we discuss applications to a wide variety of statistical and machine learning problems of recent interest, including the lasso, sparse logistic regression, basis pursuit, covariance selection, support vector machines, and many others. We also discuss general distributed optimization, extensions to the nonconvex setting, and efficient implementation, including some details on distributed MPI and Hadoop MapReduce implementations.
## Contents

1 Introduction 1

2 Precursors 5
   2.1 Dual Ascent 5
   2.2 Dual Decomposition 7
   2.3 Augmented Lagrangians and the Method of Multipliers 8

3 Alternating Direction Method of Multipliers 11
   3.1 Algorithm 11
   3.2 Convergence 13
   3.3 Optimality Conditions and Stopping Criterion 16
   3.4 Extensions and Variations 18
   3.5 Notes and References 21

4 General Patterns 23
   4.1 Proximity Operator 23
   4.2 Quadratic Objective Terms 24
   4.3 Smooth Objective Terms 28
   4.4 Decomposition 29

5 Constrained Convex Optimization 31
   5.1 Convex Feasibility 32
   5.2 Linear and Quadratic Programming 34
6  \( \ell_1 \)-Norm Problems  37
  6.1 Least Absolute Deviations  38
  6.2 Basis Pursuit  40
  6.3 General \( \ell_1 \) Regularized Loss Minimization  41
  6.4 Lasso  42
  6.5 Sparse Inverse Covariance Selection  44

7  Consensus and Sharing  47
  7.1 Global Variable Consensus Optimization  47
  7.2 General Form Consensus Optimization  52
  7.3 Sharing  55

8  Distributed Model Fitting  61
  8.1 Examples  62
  8.2 Splitting across Examples  64
  8.3 Splitting across Features  66

9  Nonconvex Problems  73
  9.1 Nonconvex Constraints  73
  9.2 Bi-convex Problems  76

10 Implementation  79
  10.1 Abstract Implementation  79
  10.2 MPI  81
  10.3 Graph Computing Frameworks  82
  10.4 MapReduce  83

11 Numerical Examples  89
  11.1 Small Dense Lasso  90
  11.2 Distributed \( \ell_1 \) Regularized Logistic Regression  94
  11.3 Group Lasso with Feature Splitting  97
  11.4 Distributed Large-Scale Lasso with MPI  99
  11.5 Regressor Selection  102
1

Introduction

In all applied fields, it is now commonplace to attack problems through data analysis, particularly through the use of statistical and machine learning algorithms on what are often large datasets. In industry, this trend has been referred to as ‘Big Data’, and it has had a significant impact in areas as varied as artificial intelligence, internet applications, computational biology, medicine, finance, marketing, journalism, network analysis, and logistics.

Though these problems arise in diverse application domains, they share some key characteristics. First, the datasets are often extremely large, consisting of hundreds of millions or billions of training examples; second, the data is often very high-dimensional, because it is now possible to measure and store very detailed information about each example; and third, because of the large scale of many applications, the data is often stored or even collected in a distributed manner. As a result, it has become of central importance to develop algorithms that are both rich enough to capture the complexity of modern data, and scalable enough to process huge datasets in a parallelized or fully decentralized fashion. Indeed, some researchers have suggested that even highly complex and structured problems may succumb most easily to relatively simple models trained on vast datasets.
Many such problems can be posed in the framework of convex optimization. Given the significant work on decomposition methods and decentralized algorithms in the optimization community, it is natural to look to parallel optimization algorithms as a mechanism for solving large-scale statistical tasks. This approach also has the benefit that one algorithm could be flexible enough to solve many problems.

This review discusses the alternating direction method of multipliers (ADMM), a simple but powerful algorithm that is well suited to distributed convex optimization, and in particular to problems arising in applied statistics and machine learning. It takes the form of a decomposition-coordination procedure, in which the solutions to small local subproblems are coordinated to find a solution to a large global problem. ADMM can be viewed as an attempt to blend the benefits of dual decomposition and augmented Lagrangian methods for constrained optimization, two earlier approaches that we review in §2. It turns out to be equivalent or closely related to many other algorithms as well, such as Douglas-Rachford splitting from numerical analysis, Spingarn’s method of partial inverses, Dykstra’s alternating projections method, Bregman iterative algorithms for $\ell_1$ problems in signal processing, proximal methods, and many others. The fact that it has been re-invented in different fields over the decades underscores the intuitive appeal of the approach.

It is worth emphasizing that the algorithm itself is not new, and that we do not present any new theoretical results. It was first introduced in the mid-1970s by Gabay, Mercier, Glowinski, and Marrocco, though similar ideas emerged as early as the mid-1950s. The algorithm was studied throughout the 1980s, and by the mid-1990s, almost all of the theoretical results mentioned here had been established. The fact that ADMM was developed so far in advance of the ready availability of large-scale distributed computing systems and massive optimization problems may account for why it is not as widely known today as we believe it should be.

The main contributions of this review can be summarized as follows:

(1) We provide a simple, cohesive discussion of the extensive literature in a way that emphasizes and unifies the aspects of primary importance in applications.
(2) We show, through a number of examples, that the algorithm is well suited for a wide variety of large-scale distributed modern problems. We derive methods for decomposing a wide class of statistical problems by training examples and by features, which is not easily accomplished in general.

(3) We place a greater emphasis on practical large-scale implementation than most previous references. In particular, we discuss the implementation of the algorithm in cloud computing environments using standard frameworks and provide easily readable implementations of many of our examples.

Throughout, the focus is on applications rather than theory, and a main goal is to provide the reader with a kind of ‘toolbox’ that can be applied in many situations to derive and implement a distributed algorithm of practical use. Though the focus here is on parallelism, the algorithm can also be used serially, and it is interesting to note that with no tuning, ADMM can be competitive with the best known methods for some problems.

While we have emphasized applications that can be concisely explained, the algorithm would also be a natural fit for more complicated problems in areas like graphical models. In addition, though our focus is on statistical learning problems, the algorithm is readily applicable in many other cases, such as in engineering design, multi-period portfolio optimization, time series analysis, network flow, or scheduling.

**Outline**

We begin in §2 with a brief review of dual decomposition and the method of multipliers, two important precursors to ADMM. This section is intended mainly for background and can be skimmed. In §3 we present ADMM, including a basic convergence theorem, some variations on the basic version that are useful in practice, and a survey of some of the key literature. A complete convergence proof is given in appendix A.

In §4 we describe some general patterns that arise in applications of the algorithm, such as cases when one of the steps in ADMM can

Full text available at: http://dx.doi.org/10.1561/2200000016
be carried out particularly efficiently. These general patterns will recur throughout our examples. In §5, we consider the use of ADMM for some generic convex optimization problems, such as constrained minimization and linear and quadratic programming. In §6, we discuss a wide variety of problems involving the $\ell_1$ norm. It turns out that ADMM yields methods for these problems that are related to many state-of-the-art algorithms. This section also clarifies why ADMM is particularly well suited to machine learning problems.

In §7, we present consensus and sharing problems, which provide general frameworks for distributed optimization. In §8, we consider distributed methods for generic model fitting problems, including regularized regression models like the lasso and classification models like support vector machines.

In §9, we consider the use of ADMM as a heuristic for solving some nonconvex problems. In §10, we discuss some practical implementation details, including how to implement the algorithm in frameworks suitable for cloud computing applications. Finally, in §11, we present the details of some numerical experiments.
References


References


117


References

References


References


References


References


124  References


References

126  References
