

Zhouchen Lin • Huan Li • Cong Fang

Accelerated Optimization for Machine Learning

First-Order Algorithms



Springer

Zhouchen Lin 
Key Lab. of Machine Perception
School of EECS
Peking University
Beijing, Beijing, China

Huan Li
College of Computer Science
and Technology
Nanjing University of Aeronautics
and Astronautics
Nanjing, Jiangsu, China

Cong Fang
School of Engineering and Applied Science
Princeton University
Princeton, NJ, USA

ISBN 978-981-15-2909-2 ISBN 978-981-15-2910-8 (eBook)
<https://doi.org/10.1007/978-981-15-2910-8>

© Springer Nature Singapore Pte Ltd. 2020

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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

Contents

1	Introduction	1
1.1	Examples of Optimization Problems in Machine Learning	1
1.2	First-Order Algorithm	3
1.3	Sketch of Representative Works on Accelerated Algorithms	3
1.4	About the Book	6
	References	6
2	Accelerated Algorithms for Unconstrained Convex Optimization	11
2.1	Accelerated Gradient Method for Smooth Optimization	11
2.2	Extension to Composite Optimization	17
2.2.1	Nesterov’s First Scheme	17
2.2.2	Nesterov’s Second Scheme	21
2.2.3	Nesterov’s Third Scheme	25
2.3	Inexact Proximal and Gradient Computing	27
2.3.1	Inexact Accelerated Gradient Descent	36
2.3.2	Inexact Accelerated Proximal Point Method	37
2.4	Restart	38
2.5	Smoothing for Nonsmooth Optimization	39
2.6	Higher Order Accelerated Method	44
2.7	Explanation: A Variational Perspective	49
2.7.1	Discretization	50
	References	54
3	Accelerated Algorithms for Constrained Convex Optimization	57
3.1	Some Facts for the Case of Linear Equality Constraint	57
3.2	Accelerated Penalty Method	61
3.2.1	Generally Convex Objectives	66
3.2.2	Strongly Convex Objectives	67
3.3	Accelerated Lagrange Multiplier Method	67
3.3.1	Recovering the Primal Solution	69
3.3.2	Accelerated Augmented Lagrange Multiplier Method	71

3.4	Alternating Direction Method of Multiplier and Its Non-ergodic Accelerated Variant	72
3.4.1	Generally Convex and Nonsmooth Case	78
3.4.2	Strongly Convex and Nonsmooth Case	79
3.4.3	Generally Convex and Smooth Case	81
3.4.4	Strongly Convex and Smooth Case	83
3.4.5	Non-ergodic Convergence Rate	85
3.5	Primal-Dual Method.....	95
3.5.1	Case 1: $\mu_g = \mu_h = 0$	97
3.5.2	Case 2: $\mu_g > 0, \mu_h = 0$	99
3.5.3	Case 3: $\mu_g = 0, \mu_h > 0$	101
3.5.4	Case 4: $\mu_g > 0, \mu_h > 0$	101
3.6	Faster Frank–Wolfe Algorithm	103
	References	107
4	Accelerated Algorithms for Nonconvex Optimization	109
4.1	Proximal Gradient with Momentum	109
4.1.1	Convergence Theorem	110
4.1.2	Another Method: Monotone APG.....	117
4.2	AGD Achieves Critical Points Quickly	118
4.2.1	AGD as a Convexity Monitor.....	118
4.2.2	Negative Curvature Descent	121
4.2.3	Accelerating Nonconvex Optimization	122
4.3	AGD Escapes Saddle Points Quickly.....	125
4.3.1	Almost Convex Case	126
4.3.2	Very Nonconvex Case.....	127
4.3.3	AGD for Nonconvex Problems	129
	References	133
5	Accelerated Stochastic Algorithms.....	137
5.1	The Individually Convex Case	139
5.1.1	Accelerated Stochastic Coordinate Descent	139
5.1.2	Background for Variance Reduction Methods	147
5.1.3	Accelerated Stochastic Variance Reduction Method	152
5.1.4	Black-Box Acceleration	160
5.2	The Individually Nonconvex Case.....	161
5.3	The Nonconvex Case	168
5.3.1	SPIDER.....	168
5.3.2	Momentum Acceleration	175
5.4	Constrained Problem	177
5.5	The Infinite Case	202
	References	206
6	Accelerated Parallel Algorithms	209
6.1	Accelerated Asynchronous Algorithms	209
6.1.1	Asynchronous Accelerated Gradient Descent	210

6.1.2	Asynchronous Accelerated Stochastic Coordinate Descent	223
6.2	Accelerated Distributed Algorithms	236
6.2.1	Centralized Topology	236
6.2.2	Decentralized Topology	241
	References	254
7	Conclusions	257
	References	258
A	Mathematical Preliminaries	261
A.1	Notations	261
A.2	Algebra and Probability	263
A.3	Convex Analysis	264
A.4	Nonconvex Analysis	270
	References	272
	Index	273

Acronyms

AAAI	Association for the Advancement of Artificial Intelligence
AACD	Asynchronous Accelerated Coordinate Descent
AAGD	Asynchronous Accelerated Gradient Descent
AASCD	Asynchronous Accelerated Stochastic Coordinate Descent
AC-AGD	Almost Convex Accelerated Gradient Descent
Acc-ADMM	Accelerated Alternating Direction Method of Multiplier
Acc-SADMM	Accelerated Stochastic Alternating Direction Method of Multiplier
Acc-SDCA	Accelerated Stochastic Dual Coordinate Ascent
ADMM	Alternating Direction Method of Multiplier
AGD	Accelerated Gradient Descent
APG	Accelerated Proximal Gradient
ASCD	Accelerated Stochastic Coordinate Descent
ASGD	Asynchronous Stochastic Gradient Descent
ASVRG	Asynchronous Stochastic Variance Reduced Gradient
DSCAD	Distributed Stochastic Communication Accelerated Dual
ERM	Empirical Risk Minimization
EXTRA	EXact firST-order Algorithm
GIST	General Iterative Shrinkage and Thresholding
IC	Individually Convex
IFO	Incremental First-order Oracle
INC	Individually Nonconvex
iPiano	Inertial Proximal Algorithms for Nonconvex Optimization
IQC	Integral Quadratic Constraint
KKT	Karush–Kuhn–Tucker
KŁ	Kurdyka–Łojasiewicz
LASSO	Least Absolute Shrinkage and Selection Operator
LMI	Linear Matrix Inequality
MISO	Minimization by Incremental Surrogate Optimization
NC	Negative Curvature/Nonconvex
NCD	Negative Curvature Descent
PCA	Principal Component Analysis

PG	Proximal Gradient
SAG	Stochastic Average Gradient
SAGD	Stochastic Accelerated Gradient Descent
SCD	Stochastic Coordinate Descent
SDCA	Stochastic Dual Coordinate Ascent
SGD	Stochastic Gradient Descent
SPIDER	Stochastic Path-Integrated Differential Estimator
SVD	Singular Value Decomposition
SVM	Support Vector Machine
SVRG	Stochastic Variance Reduced Gradient
SVT	Singular Value Thresholding
VR	Variance Reduction