

Signal Processing and Networking for Big Data Applications

ZHU HAN

University of Houston

MINGYI HONG

Iowa State University

DAN WANG

The Hong Kong Polytechnic University



CAMBRIDGE
UNIVERSITY PRESS

Contents

Part I Overview of Big Data Applications

page 1

1	Introduction	3
1.1	Background	3
1.2	Market and Readership	4
1.3	Brief Description of the Chapters	5
1.4	Acknowledgments	10

2	Data Parallelism: The Supporting Architecture	11
2.1	Data Parallelism	11
2.2	Hadoop Distributed File System (HDFS)	13
2.3	MapReduce and Yarn	13
2.3.1	MapReduce	13
2.3.2	Yarn	15
2.4	Spark	16
2.5	GraphLab and Spark Streaming	17
2.5.1	GraphLab	17
2.5.2	Spark Streaming	18
2.6	Big Data in the Cloud	18
2.7	Advances in the Data Parallelism Architecture	20
2.7.1	Fast Job Completion Time	20
2.7.2	Data-Computation Contention	23
2.7.3	Optimizing Resource Utilization for Big Data Jobs	24
2.8	Supports for Signal Processing Algorithms	26
2.9	Summary	28

Part II Methodology and Mathematical Background	31
--	-----------

3	First-Order Methods	33
3.1	A Brief Introduction to Optimization	33
3.1.1	Definition and Notations	33
3.1.2	Unconstrained Optimization	35
3.1.3	Constrained Optimization	36

3.2	Gradient Descent Method	37
3.3	Block Coordinate Descent Method	41
3.4	Block Successive Upper-Bound Minimization Method	46
3.5	The Augmented Lagrangian Method	48
3.6	Alternating Direction Method of Multipliers	49
3.6.1	The Algorithm	49
3.6.2	Convergence Analysis	50
3.6.3	Solving Multiblock Problems and Beyond	53
3.7	Conclusion	56
4	Sparse Optimization	58
4.1	Sparse Optimization Models	59
4.2	Classic Solvers	61
4.3	The BSUM Method for Sparse Optimization	62
4.4	Shrinkage Operation	64
4.4.1	Generalizations of Shrinkage	65
4.5	Prox-Linear Algorithms	67
4.5.1	Forward-Backward Operator Splitting	68
4.5.2	Examples	69
4.5.3	Convergence Rates	71
4.6	Dual Algorithms	71
4.6.1	Dual Formulations	71
4.7	Bregman Method	72
4.7.1	Bregman Iterations and Denoising	74
4.7.2	Linearized Bregman and Augmented Models	76
4.7.3	Handling Complex Data and Variables	77
4.8	ADMM for Sparse Optimization	78
4.9	Homotopy Algorithms and Parametric Quadratic Programming	82
4.10	Continuation, Varying Stepsizes, and Line Search	84
4.11	Greedy Algorithms	85
4.11.1	Greedy Pursuit Algorithms	85
4.11.2	Iterative Support Detection	86
4.11.3	Hard Thresholding	87
4.12	Algorithms for Low-Rank Matrices	88
4.13	How to Choose an Algorithm	89
5	Sublinear Algorithms	92
5.1	Introduction	92
5.1.1	Organization	93
5.2	Foundations	94
5.2.1	Approximation and Randomization	94
5.2.2	Inequalities and Bounds	95
5.2.3	Classification of Sublinear Algorithms	96
5.3	Examples	97

5.3.1	Estimating the User Percentage: The Very First Example	97
5.3.2	Finding Distinct Elements	98
5.3.3	Computing the Number of Connected Components	101
5.3.4	The Two-Cat Problem	105
5.4	Summary	106
6	Tensor for Big Data	108
6.1	Tensor Basics	108
6.1.1	Tensor Definitions	108
6.1.2	Tensor Operations	110
6.2	Tensor Decomposition	111
6.2.1	Tucker3 and PARAFAC	111
6.2.2	Tensor Compression	113
6.2.3	Tensor Completion	116
6.3	Tensor Voting	117
6.3.1	Encode Normal Space with Tensor	118
6.3.2	Inferring Structure	120
6.3.3	Token Refinement	122
6.3.4	Token Decomposition	122
6.3.5	Inference Algorithm	123
6.4	Other Tensor Applications	123
7	Deep Learning and Applications	126
7.1	Introduction	127
7.1.1	History	127
7.1.2	Milestone Literature	127
7.1.3	Applications	128
7.2	Deep Learning Basics	130
7.2.1	Restricted Boltzmann Machine	131
7.2.2	Stochastic Gradient Descent and Backpropagation	133
7.2.3	Deep Learning Networks	134
7.3	Example 1: Apache Spark for MBD	137
7.3.1	Introduction	137
7.3.2	MBD Concepts and Features	138
7.3.3	Deep Learning in MBD Analytics	140
7.3.4	Spark-Based Deep Learning Framework for MBD	142
7.3.5	Prototyping a Context-Aware Activity Recognition Systems	145
7.3.6	Future Work	147
7.4	Example 2: User Moving Pattern Extraction	150
7.4.1	Introduction	150
7.4.2	Data Collection	151
7.4.3	Deep Learning Extraction	152
7.4.4	Experiment Results	157
7.5	Example 3: Combination with Nonparametric Bayesian Learning	159

7.5.1	Introduction	159
7.5.2	Generative Model	160
7.5.3	Inference Algorithm	162
7.5.4	Simulations and Discussions for Wireless Applications	165
7.6	Summary	166
Part III	Big Data Applications	169
8	Compressive Sensing-Based Big Data Analysis	171
8.1	Background	171
8.1.1	Steps of Compressive Sensing	172
8.1.2	Robust Compressive Sensing and Signal Recovery	174
8.2	Traditional Sensing vs. Compressive Sensing	176
8.3	Sparse Representation	177
8.3.1	Extensions of Sparse Models	179
8.4	CS Encoding and Decoding	180
8.4.1	The Null-Space Property (NSP)	182
8.4.2	The Restricted Isometry Principle (RIP)	183
8.4.3	The Spherical Section Property	185
8.4.4	“RIPless” Analysis	186
8.4.5	Non- ℓ_1 Decoding Methods	186
8.4.6	Examples	187
8.5	Compressive Sensing-Based Big Data Applications	188
8.5.1	CS-Based Analog-to-Digital Converter	188
8.5.2	Communications	190
8.5.3	Hyperspectral Imaging	194
8.5.4	Data Streaming	195
9	Distributed Large-Scale Optimization	196
9.1	Background	196
9.1.1	Problem Formulation	196
9.1.2	Applications	198
9.2	Distributed Gradient/Subgradient Methods	199
9.2.1	Algorithm Description	199
9.2.2	Convergence Analysis and Variants	200
9.2.3	Convergence Rate Analysis	205
9.3	ADMM-Based Methods	208
9.3.1	Problem Formulation	208
9.3.2	Distributed Implementation	210
9.3.3	Convergence and Variations	212
9.4	Other Types of First-Order Method	216
9.5	Special Case: Global Consensus Problem	220
9.6	Comparing Different Algorithms	222
9.6.1	Problem Types	222

9.6.2	Graph Types	222
9.6.3	Convergence Rates	224
10	Optimization of Finite Sums	225
10.1	The Incremental Gradient Algorithms	226
10.2	The Stochastic Gradient-Based Algorithms	228
10.2.1	The SAG-Based Method	229
10.2.2	The SVRG Algorithm	231
10.3	The Stochastic Algorithms in the Dual	233
10.4	Summary	235
11	Big Data Optimization for Communication Networks	236
11.1	Mobile Data Offloading in Software-Defined Networks	237
11.2	Scalable Service Management in Mobile Cloud Computing	241
11.2.1	Literature Review	242
11.2.2	System Model and Problem Formulation	243
11.2.3	Mobile Cloud Service Management	244
11.2.4	Maximizing Total Revenue	245
11.2.5	Numerical Results	246
11.3	Scalable Workload Management in Data Centers	249
11.3.1	Problem Formulation	250
11.4	Resource Allocation for Wireless Network Virtualization	254
11.4.1	Wireless Network Virtualization	255
11.4.2	Routing Model for Virtual Network	256
11.4.3	Joint Resource and Routing Optimization for Wireless Virtual Networks	257
11.5	Summary	258
12	Big Data Optimization for Smart Grid Systems	259
12.1	Introduction	259
12.2	Backgrounds	260
12.2.1	False Data Injection Attacks against State Estimation	260
12.2.2	Security-Constrained Optimal Power Flow	263
12.3	Sparse Optimization for False Data Injection Detection	265
12.3.1	Sparse Optimization Problem Formulation	265
12.3.2	Nuclear Norm Minimization	265
12.3.3	Low-Rank Matrix Factorization	267
12.3.4	Numerical Results	269
12.4	Distributed Parallel Approach for Security-Constrained Optimal Power Flow	275
12.4.1	An Introduction to ADMM	276
12.4.2	Distributed and Parallel Approach for SCOPF	277
12.4.3	Numerical Results	279
12.5	Concluding Remarks	282

13	Processing Large Data Sets in MapReduce	283
13.1	Introduction	283
13.1.1	The Data Skew Problem of MapReduce Jobs	283
13.1.2	Chapter Outline	284
13.2	Server Load Balancing: Analysis and Problem Formulation	285
13.2.1	Background and Motivation	285
13.2.2	Problem Formulation	287
13.2.3	Input Models	288
13.3	A 2-Competitive Fully Online Algorithm	288
13.4	A Sampling-Based Semi-Online Algorithm	289
13.4.1	Sample Size	290
13.4.2	Heavy Keys	291
13.4.3	A Sample-Based Algorithm	291
13.5	Performance Evaluation	293
13.5.1	Simulation Setup	293
13.5.2	Results on Synthetic Data	293
13.5.3	Results on Real Data	297
13.6	Summary	300
14	Massive Data Collection Using Wireless Sensor Networks	301
14.1	Introduction	301
14.1.1	Background and Related Work	302
14.1.2	Chapter Outline	303
14.2	System Architecture	303
14.2.1	Preliminaries	303
14.2.2	Network Construction	304
14.2.3	Specifying the Structure of the Layers	305
14.2.4	Data Collection and Aggregation	305
14.3	Evaluation of the Accuracy and the Number of Sensors Queried	306
14.3.1	MAX and MIN Queries	307
14.3.2	QUANTILE Queries	307
14.3.3	AVERAGE and SUM Queries	308
14.3.4	Effect of the Promotion Probability p	313
14.4	Energy Consumption	314
14.4.1	Overall Lifetime of the System	314
14.5	Evaluation Results	315
14.5.1	System Settings	315
14.5.2	Layers vs. Accuracy	315
14.6	Practical Variations of the Architecture	320
14.7	Summary	321
	<i>Bibliography</i>	322
	<i>Index</i>	361