

## DOCTORAL THESIS

### Study on efficient sparse and low-rank optimization and its applications

Lou, Jian

*Date of Award:*  
2018

[Link to publication](#)

#### **General rights**

Copyright and intellectual property rights for the publications made accessible in HKBU Scholars are retained by the authors and/or other copyright owners. In addition to the restrictions prescribed by the Copyright Ordinance of Hong Kong, all users and readers must also observe the following terms of use:

- Users may download and print one copy of any publication from HKBU Scholars for the purpose of private study or research
- Users cannot further distribute the material or use it for any profit-making activity or commercial gain
- To share publications in HKBU Scholars with others, users are welcome to freely distribute the permanent URL assigned to the publication

# Abstract

Sparse and low-rank models have been becoming fundamental machine learning tools and have wide applications in areas including computer vision, data mining, bioinformatics and so on. It is of vital importance, yet of great difficulty, to develop efficient optimization algorithms for solving these models, especially under practical design considerations of computational, communicational and privacy restrictions for ever-growing larger scale problems. This thesis proposes a set of new algorithms to improve the efficiency of the sparse and low-rank models optimization.

First, facing a large number of data samples during training of empirical risk minimization (ERM) with structured sparse regularization, the gradient computation part of the optimization can be computationally expensive and becomes the bottleneck. Therefore, I propose two gradient efficient optimization algorithms to reduce the total or per-iteration computational cost of the gradient evaluation step, which are new variants of the widely used generalized conditional gradient (GCG) method and incremental proximal gradient (PG) method, correspondingly. In detail, I propose a novel algorithm under GCG framework that requires optimal count of gradient evaluations as proximal gradient. I also propose a refined variant for a type of gauge regularized problem, where approximation techniques are allowed to further accelerate linear subproblem computation. Moreover, under the incremental proximal gradient framework, I propose to approximate the composite penalty by its proximal average under incremental gradient framework, so that a trade-off is made between precision and efficiency. Theoretical analysis and empirical studies show the efficiency of the proposed methods.

Furthermore, the large data dimension (e.g. the large frame size of high-resolution image and video data) can lead to high per-iteration computational complexity, thus results into poor-scalability of the optimization algorithm from practical perspective. In particular, in spectral k-support norm regularized robust low-rank matrix and tensor optimization, traditional proximal map based alternating direction method of multipliers (ADMM) requires to evaluate a super-linear complexity subproblem in each iteration. I propose a set of per-iteration computational efficient alternatives to reduce the cost to linear and nearly linear with respect to the input data dimension for matrix and tensor case, correspondingly. The proposed algorithms consider the dual objective of the original problem that can take advantage of the more computational efficient linear oracle of the spectral k-support norm to be evaluated. Further, by studying the sub-gradient of the loss of the dual objective, a line-search strategy is adopted in the algorithm to enable it to adapt to the Hölder smoothness. The overall convergence rate is also provided. Experiments on various computer vision and image processing applications demonstrate the superior prediction performance and computation efficiency of the proposed algorithm.

In addition, since machine learning datasets often contain sensitive individual information, privacy-preserving becomes more and more important during sparse optimization. I provide two differentially private optimization algorithms under two common large-scale machine learning computing contexts, i.e., distributed and streaming optimization, correspondingly. For the distributed setting, I develop a new algorithm with 1) guaranteed strict differential privacy requirement, 2) nearly optimal utility and 3) reduced uplink communication complexity, for a nearly unexplored context with features partitioned among different parties under privacy restriction. For the streaming setting, I propose to improve the utility of the private algorithm by trading the privacy of distant input instances, under the differential privacy restriction. I show that the proposed method can either solve the private approximation function by a projected gradient update for projection-friendly constraints, or by a conditional gradient step for linear oracle-friendly constraint, both of which improve the regret

bound to match the nonprivate optimal counterpart.

# Table of Contents

<b>Declaration</b>	i
<b>Abstract</b>	ii
<b>Acknowledgments</b>	v
<b>Table of Contents</b>	vi
<b>List of Tables</b>	xiv
<b>List of Figures</b>	xv
<b>Chapter 1 Introduction</b>	1
1.1 Background . . . . .	1
1.2 Motivations . . . . .	4
1.2.1 Large Scale Structured Sparse Empirical Risk Minimization . . . . .	4
1.2.2 Robust Low Rank Matrix/Tensor Optimization . . . . .	4
1.2.3 Differentially Private Optimization . . . . .	5
1.3 Main Contributions and Organization of this Thesis . . . . .	6
<b>Chapter 2 Related Work</b>	10
2.1 Conditional Gradient Algorithms . . . . .	10
2.1.1 Basic Conditional Gradient Algorithm . . . . .	10
2.1.2 Generalized Conditional Gradient . . . . .	11
2.1.3 Approximate Linear Operator Evaluation . . . . .	13

2.2	Efficient Proximal Gradient Algorithms	15
2.2.1	Proximal Average	15
2.2.2	Incremental Gradient Descent Methods	16
2.3	Robust Low Rank Matrix Learning	17
2.3.1	Notation for Matrix	17
2.3.2	Robust Low Rank Subspace Learning	18
2.3.3	Spectral k-Support Norm	18
2.4	Robust Low Rank Tensor Learning	20
2.4.1	Notation for Tensor	20
2.4.2	Tensor Singular Value Decomposition (t-SVD) Algebraic Framework	20
2.4.3	Tensor Tubal Rank Definition and Tensor Nuclear Norm	23
2.5	Differentially Private Learning	24
2.5.1	Differentially Private Optimization for Feature-wise Distributed Dataset	25
2.5.2	Differentially Private Streaming Convex Optimization	26
<b>Chapter 3 Efficient Generalized Conditional Gradient with Gradient Sliding for Composite Optimization</b>		
3.1	Introduction	28
3.2	Preliminary	30
3.2.1	Conditional Gradient Sliding Algorithm (CGS)	30
3.3	The Proposed Algorithm	33
3.3.1	General GCG-GS	33
3.3.2	Refined GCG-GS for Gauge Regularized Problem	37
3.4	Experiment	41
3.5	Summary	43
<b>Chapter 4 Proximal Average Approximated Incremental Gradient Method for Composite Penalty Regularized Empirical Risk</b>		

<b>Minimization</b>	<b>45</b>
4.1 Introduction	45
4.2 Preliminaries	49
4.3 Accelerated Proximal Average Approximated Incremental Gradient	
for ERM with Convex Composite Penalty	51
4.3.1 Overlapping Group Lasso and Graph-guided Fused Lasso	51
4.3.2 Incremental Gradient Proximal Average for Convex Composite	
Penalty Regularized ERM	52
4.3.3 Analysis of IncrePA-cvx	54
4.3.4 Discussion	55
4.4 Incremental Proximal Average for Nonconvex Composite Penalty	
Regularized ERM	56
4.4.1 Two Examples of Nonconvex Composite Penalties in Struc-	
tured Sparse Estimation	57
4.4.2 Related Work	58
4.4.3 Nonconvex Extension of Incremental Gradient with PA	58
4.4.4 Analysis of IncrePA-ncvx	60
4.5 Experiments	62
4.5.1 Experiment 1: Solving general convex loss function with convex	
composite penalty	63
4.5.2 Experiment 2: Solving strongly convex loss function with	
convex composite penalty	66
4.5.3 Experiment 3: Solving nonconvex composite penalty of capped	
$\ell_1$ overlapping group lasso	68
4.5.4 Experiment 4: Solving nonconvex composite penalty of capped	
$\ell_1$ graph-guided lasso	70
4.6 Summary	71
<b>Chapter 5 Scalable Spectral k-Support Norm Regularization for Ro-</b>	
<b>bust low-rank Subspace Learning</b>	<b>73</b>

5.1	Introduction	73
5.2	Preliminary: Scalable Algorithm with Spectral $k$ -Support Norm	76
5.3	The Proposed Method	77
5.3.1	Formulation I: Usage with Squared Spectral $k$ -Support Norm	78
5.3.2	Formulation II: Usage with Spectral	
	$k$ -Support Norm	82
5.3.3	Algorithm	84
5.4	Experiment	89
5.4.1	Synthetic Data	90
5.4.2	Background Modeling on Surveillance	
	Videos	93
5.4.3	Face Reconstruction	94
5.5	Summary	95
<b>Chapter 6 Robust Low-Rank Tensor Minimization via a New Tensor</b>		
	<b>Spectral <math>k</math>-Support Norm</b>	<b>96</b>
6.1	Introduction	96
6.2	Additional Notation	102
6.3	A New Convex Relaxation for Robust Tensor Recovery: Tensor Spec-	
	tral $k$ -Support Norm	102
6.3.1	The New Tensor Spectral $k$ -Support Norm	102
6.3.2	Robust low-rank Tensor Recovery with TSP- $k$ Norm	105
6.4	An ADMM-based Optimization Algorithm for Medium-size Data	106
6.4.1	Proximal Operator for TSP- $k$ Norm-based Regularizer	106
6.4.2	Preconditioned ADMM-based Optimization Algorithm	109
6.4.3	Computational Complexity	109
6.5	A Greedy Dual Optimization Algorithm for Larger-Size Data	110
6.5.1	Equivalent Dual Reformulation	111
6.5.2	Polar Operator for TSP- $k$ Norm	112
6.5.3	Universal Primal-Dual Optimization Based Algorithm	113



6.5.4	Complexity and Convergence Analysis	115
6.6	Experiment	116
6.6.1	Experiments on medium size datasets	116
6.6.2	Experiments on Larger Size Datasets	122
6.7	Summary	126
<b>Chapter 7 Uplink Communication Efficient Differentially Private Sparse</b>		
<b>Optimization with Feature-wise Distributed Data</b>		<b>127</b>
7.1	Introduction	127
7.2	Additional Notation and Preliminary	130
7.2.1	Sampling Distributions and Known Convergence Results for	
Block Coordinate Frank-Wolfe Algorithms		130
7.3	Block-Coordinate Frank-Wolfe under Arbitrary Sampling	133
7.3.1	Algorithm Description	133
7.3.2	Expected Curvature	135
7.3.3	Convergence Analysis for BCFW with Arbitrary Sampling	138
7.4	Uplink Communication Efficient Differentially Private BCFW with	
Distributed Features		139
7.4.1	Algorithm Description	140
7.4.2	Analysis	142
7.5	Summary	144
<b>Chapter 8 Differentially Private Streaming Convex Optimization with</b>		
<b>Decayed Privacy</b>		<b>146</b>
8.1	Introduction	146
8.2	Preliminary	149
8.2.1	Problem Setup	149
8.2.2	Differentially Private Follow The Approximate Leader	149
8.3	Proposed Method	151

8.3.1 Window Tree Mechanism for Private Gradient Summation	
with Decayed Privacy	152
8.3.2 Window Differentially Private COCO with Projection	153
8.3.3 Window Differentially Private COCO with Linear Oracle	156
8.4 Summary	159
<b>Chapter 9 Conclusions</b>	<b>161</b>
9.1 Conclusions	161
9.2 Future Work	164
<b>Bibliography</b>	<b>166</b>
<b>Appendix</b>	<b>183</b>
<b>Chapter A Proofs of Materials in Chapter 3</b>	<b>184</b>
A.1 Proof of Theorem 3.3.1 and Corollary 3.3.2	184
A.1.1 Optimal FO Evaluation	184
A.1.2 Evaluation bound on LO	186
A.2 Proof of Theorem 3.3.3	188
A.2.1 Outer Loop Analysis	188
A.2.2 Inner Loop Analysis	188
<b>Chapter B Proofs of Materials in Chapter 4</b>	<b>192</b>
B.1 Proof of Theorem 4.3.2 and Theorem 4.3.3	192
B.1.1 Proof of Theorem 4.3.2: General Convex Loss Function Case	193
B.1.2 Proof of Theorem 4.3.3: Strongly Convex Loss Function Case	195
<b>Chapter C Proofs of Materials in Chapter 6</b>	<b>199</b>
C.1 Proof of Materials in Section 6.3	199
C.1.1 Proof of Proposition 6.3.1	199
C.1.2 Proof of Proposition 6.3.2	200
C.1.3 Proof of Proposition 6.3.4	202

<b>C.2 Proof Materials in Section 6.4</b>	202
C.2.1 Proof of Proposition 6.4.1	202
C.2.2 Detailed derivation of preconditioned ADMM	202
C.2.3 Computational Complexity Analysis for Algorithm 14	203
<b>C.3 Proof of Materials in Section 6.5</b>	205
C.3.1 Proof of Proposition 6.5.1	205
C.3.2 Proof of Proposition 6.5.2	206
C.3.3 Proof of Proposition 6.5.3	207
C.3.4 Proof of Corollary 6.5.1	207
C.3.5 Line-search Subroutine	208
C.3.6 Computational Complexity Analysis for Algorithm 16	210
<b>Chapter D Proofs of Materials in Chapter 7</b>	<b>212</b>
D.1 Preliminary	212
D.2 Proofs of Materials in Section 7.3	214
D.2.1 Proof of Proposition 7.3.2	214
D.2.2 Proof of Proposition 7.3.3	215
D.2.3 Proof of Proposition 7.3.4	215
D.2.4 Proof of Proposition 7.3.5	216
D.2.5 Proof of Theorem 7.3.6	217
D.3 Proofs of Materials in Section 7.4	220
D.3.1 Lemma D.3.1 and Proof: Private feature sharing	220
D.3.2 Lemma D.3.2 and Proof: Private index computing	221
D.3.3 Proof of Theorem 7.4.1	221
D.3.4 Proof of Theorem 7.4.2	221
<b>Chapter E Proofs of Materials in Chapter 8</b>	<b>225</b>
E.1 Additional Materials to Section 8.3.1	225
E.1.1 Window Tree Mechanism with Gamma Noise Perturbation	225
E.1.2 Window Tree Mechanism with Gaussian Noise Perturbation	226

<b>E.2 Additional Materials to Section 8.3.2</b> . . . . .	227
<b>E.2.1 Proof of Theorem 8.3.7 and 8.3.8</b> . . . . .	228
<b>E.3 Additional Materials to Section 8.3.3</b> . . . . .	229
<b>E.3.1 Main Proof of Theorem 8.3.10 and 8.3.11</b> . . . . .	229
<b>Chapter F Publication List</b>	<b>236</b>
<b>F.1 Published Papers</b> . . . . .	236
<b>F.2 Submitted Papers</b> . . . . .	237
<b>CURRICULUM VITAE</b>	<b>238</b>