

**OPTIMIZATION AND LEARNING UNDER  
UNCERTAINTY - A UNIFIED ROBUSTNESS  
PERSPECTIVE**

**YANG WENZHUO**

A THESIS SUBMITTED FOR THE DEGREE OF DOCTOR OF PHILOSOPHY


DEPARTMENT OF MECHANICAL ENGINEERING  
NATIONAL UNIVERSITY OF SINGAPORE

2016

## Declaration

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

A handwritten signature in black ink, reading "Yang Wenzhuo", written over a horizontal line.

Yang Wenzhuo

August 2016

## Acknowledgments

First and foremost, I would like to express my sincere gratitude to my advisor Prof. Xu Huan for the continuous support of my Ph.D study and related research, for his patience, motivation, broad range of knowledge and insightful technical guidance. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D study.

Besides my advisor, I would like to thank my co-supervisor Prof. Melvyn Sim, for his encouragement and guidance, but also for his knowledge of robust optimization which incited me to widen my research from various perspectives. My sincere thanks also goes to Prof. Shie Mannor of Technion – Israel Institute of Technology, who provided me an opportunity to join their team as a visiting student, and who helped me appreciate the beauty of online learning. I am fortunate to have had the chance to collaborate with him, an experience that helped produce a significant portion of this thesis.

I thank my fellow labmates Dr. Feng Jiashi, Dr. Lim Shiau Hong, Wang Yuxiang, Yu Pengqian, Qu Chao, Zhu Changbo, Fu Jie, Zhou Qiang, Dr. Zhang Ning, Dr. Le Thi Khanh Hien, Dr. Nguyen Viet Cuong, Zhao Renbo, Liao Lizi, etc., for the stimulating discussions, the weekly group meetings, and all the fun we have had in the last four years.

Last but not the least, I would like to thank my family for supporting me spiritually throughout writing this thesis and my life in general.

## Abstract

Robust decision making is ubiquitous in real-world applications in machine learning, operations research and finance, etc., due to the uncertainty and noise in practical data coming from measurement errors or malicious attacking. Robust optimization and distributionally robust optimization are two popular techniques in robust decision making. Robust optimization treats uncertain parameters by defining uncertainty sets consisting of their possible realizations and solving solutions with the worst-case realizations, while distributionally robust optimization takes the advantages of prior distributional knowledge about uncertain parameters by constructing ambiguity sets that are assumed to include the true distributions of uncertain parameters and computes solutions by minimizing the worst-case expected cost over the distributions in the ambiguity sets.

In this thesis, we first investigate the computational aspects of distributionally robust chance constrained optimization with non-linear uncertainties, and apply these robust decision making techniques in machine learning both theoretically and algorithmically, i.e., we provide a new robustness interpretation of a broad range of Lasso-like algorithms and regularized SVMs. Second, we study optimization with unknown parameters, which generalizes both stochastic linear optimization and linear bandits. We tackle it from a dynamic perspective – the decision maker can make a tentative decision, collect feedbacks about the decision and fine tune the decision – and develop two algorithms based on the epsilon-decreasing strategy and the upper confidence bound strategy, respectively. Finally, we study principal component analysis with noisy or incomplete data, and propose three robust principal component analysis algorithms that are able to handle outlying observation with solid theoretical guarantees.

# Contents

<b>List of Publications</b>	viii
<b>List of Figures</b>	ix
<b>List of Tables</b>	xi
<b>List of Algorithms</b>	xii
<b>1 Introduction</b>	<b>1</b>
1.1 Optimization with Uncertain Parameters	3
1.2 Optimization with Unknown Parameters	9
1.3 Principal Component Analysis with Noisy Observation	11
1.4 Structure of the Thesis	13
<b>2 Distributionally Robust Chance Constraints for Non-Linear Uncertainties</b>	<b>17</b>
2.1 Introduction	18
2.2 Formulation and Motivating Examples	22
2.3 The Chance Constraint Case	27
2.3.1 Equivalence to Robust Optimization	28
2.3.2 Tractability of Individual DRCC	31
2.4 Probabilistic Envelope Constraint	33
2.5 Chance Constraints: Beyond Mean and Variance	38
2.5.1 Uncertain Mean and Covariance	38
2.5.2 Known Mean and Support	40
2.5.3 Conservative Approximation	41
2.6 Joint Chance Constraint	44
2.6.1 The Bonferroni Approximation	45
2.6.2 The Worst-case CVaR Approximation	45
2.7 Simulation	48

---

2.8	Proofs of the Main Results	51
2.8.1	Proof of Corollary 2.1	51
2.8.2	Proofs for Section 2.3.2	53
2.8.3	Proofs of Results in Section 2.4	54
2.8.4	Proofs for Section 2.5	57
2.8.5	Proofs of Results in Section 2.6	60
2.9	Chapter Summary	62
<b>3</b>	<b>A Unified Robust Regression Model for Lasso-like Algorithms</b>	<b>63</b>
3.1	Introduction	63
3.2	Unified Robust Framework	65
3.2.1	Preliminary	65
3.2.2	Main Results	66
3.3	General Uncertainty Sets	75
3.4	Sparsity	80
3.5	Consistency	83
3.6	Chapter Summary	91
<b>4</b>	<b>A Distributionally Robust Optimization Interpretation For Regularized SVMs</b>	<b>92</b>
4.1	Introduction	92
4.2	Preliminaries of DRO	94
4.3	DRO Interpretation for Regularized SVMs	96
4.4	Robustness to Corruption of Features	97
4.5	Robust Classification via Chance Constraints	99
4.6	Experiments	102
4.7	Proofs of Technical Results	103
4.7.1	Proof of Theorem 4.1	105
4.7.2	Proof of Theorem 4.2	107
4.7.3	Proof of Corollary 4.2	111
4.7.4	Proof of Theorem 4.3	111

---

4.8	Chapter Summary	113
<b>5</b>	<b>The Coherent Loss Function for Classification</b>	<b>114</b>
5.1	Introduction	114
5.2	Coherent Classification Loss Function	117
5.2.1	Salient Properties and Representation Theorem	117
5.2.2	Minimal Coherent Classification Loss Function	119
5.2.3	Optimization With the Coherent Loss Function	121
5.3	Equivalent Formulation and Applications	123
5.4	Statistical Interpretation	128
5.5	Simulation	131
5.6	Proofs of Technical Results	132
5.6.1	Proof of Theorem 5.1	132
5.6.2	Proof of Theorem 5.2	142
5.6.3	Proof of Theorem 5.3	144
5.6.4	Proof of Theorem 5.5	146
5.7	Chapter Summary	148
<b>6</b>	<b>Online Linear Optimization with Unobserved Constraints</b>	<b>149</b>
6.1	Introduction	149
6.2	Related Work	151
6.3	Problem Setting	152
6.3.1	Learning Model	152
6.3.2	Assumptions	154
6.3.3	Performance Metric	155
6.4	Two Algorithms: LPUC-ED and LPUC-UCB	155
6.5	Regret Bound and Constraint Violation	159
6.6	Extension to General Cases	162
6.7	Experiments	166
6.8	Proofs of Technical Results	168
6.8.1	Proof of Proposition 6.1	173

---

6.8.2	Proof of Theorem 6.1	175
6.8.3	Proof of Theorem 6.2	182
6.8.4	Proof of Theorem 6.3	184
6.8.5	Proofs in Section 6.6	185
6.9	Chapter Summary	192
<b>7</b>	<b>A Unified Framework for Outlier-Robust PCA-like Algorithms</b>	<b>193</b>
7.1	Introduction	193
7.2	Unified Framework for Outlier-Robust PCA	196
7.2.1	Problem Setup	196
7.2.2	General Formulation of PCA-like Algorithms	197
7.2.3	Outlier-Robust PCA-like Algorithms	198
7.3	Theoretical Guarantees	200
7.3.1	Upper Bound of Subspace Distance	201
7.3.2	Lower Bound of Expressed Variance	204
7.3.3	Complexity	206
7.4	Experimental Results	206
7.5	Proofs of Section 7.3.1	211
7.6	Proofs in Section 7.3.2	218
7.7	Additional Lemmas	227
7.8	Chapter Summary	235
<b>8</b>	<b>Non-convex Outlier-Robust PCA</b>	<b>236</b>
8.1	Introduction	236
8.2	Problem Setting	238
8.3	Outlier Rejection and Outlier Reduction	240
8.4	Performance Guarantees	244
8.5	Experiments	247
8.5.1	Synthetic Data	248
8.5.2	Real-world Data	250
8.6	Proofs of Technical Results	255



---

8.7	Chapter Summary	273
<b>9</b>	<b>Online PCA with Imperfect Data</b>	<b>274</b>
9.1	Introduction	274
9.2	Problem Setting	276
9.3	Framework for Online Robust PCA	278
9.3.1	Missing Entries	280
9.3.2	Corrupted Entries	281
9.3.3	Limited Observation	283
9.4	Unknown Parameters	285
9.5	Experiments	287
9.5.1	Synthetic Data	287
9.5.2	Real-world Data	291
9.6	Proofs of Technical Results	293
9.7	Chapter Summary	308
<b>10</b>	<b>Conclusion</b>	<b>309</b>
	<b>References</b>	<b>311</b>

## List of Publications

- (1) **W. Yang** and H. Xu. Distributionally robust chance constraints for non-linear uncertainties. *Mathematical Programming*, 155(1-2):231-265, 2016, Springer Berlin Heidelberg.
- (2) H. Zhang, X. Shang, **W. Yang**, H. Xu, H. Luan and T. S. Chua. Online Collaborative Learning for Open-Vocabulary Visual Classifiers. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- (3) **W. Yang**, H. Xu. A Divide and Conquer Framework for Distributed Graph Clustering. *The International Conference on Machine Learning (ICML)*, 2015.
- (4) **W. Yang**, H. Xu. Streaming sparse principal component analysis. *The International Conference on Machine Learning (ICML)*, 2015.
- (5) **W. Yang**, H. Xu. A Unified Framework for Outlier-Robust PCA-like Algorithms. *The International Conference on Machine Learning (ICML)*, 2015.
- (6) **W. Yang**, M. Sim, H. Xu. The Coherent Loss Function for Classification. *The International Conference on Machine Learning (ICML)*, 2014.
- (7) **W. Yang**, H. Xu. A Unified Robust Regression Model for Lasso-like Algorithms. *The International Conference on Machine Learning (ICML)*, 2013.
- (8) **W. Yang**, G. Zhang, H. Bao, J. Kim, H. Y. Lee. Consistent depth maps recovery from a trinocular video sequence. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

## List of Figures

2.1	The transportation problem: the resulting allocations for different guarantees $\gamma = 0.1 - 0.8$ .	49
2.2	The transportation problem: the resulting allocations for decay rates $b = 0.1, 1.0$ and $10.0$ .	50
3.1	Preferred solutions of the two group Lasso.	70
4.1	The classification errors of SVM, MCF QUA, MCF LOG, FDROP and DROFD.	102
5.1	Illustration of the effect of outliers to the cumulative loss vs the coherent loss.	125
5.2	Performance comparison of cumulative loss approach vs coherent loss approach where bound $T$ is fixed and the fraction $k/n$ varies from $0.0$ to $1.0$ .	132
5.3	Performance comparison of cumulative loss approach vs coherent loss approach where $k/n$ is fixed and $T$ varies.	133
6.1	The empirical performance of LPUC-ED and LPUC-UCB.	167
6.2	A maximum flow problem with unknown edge capacities.	168
7.1	The empirical performance of OR-PCA, OR-SPCA, ROB-SPCA and FPS when the singular values of $\mathbf{A}$ are drawn uniformly.	208
7.2	The empirical performance of OR-PCA, OR-SPCA, ROB-SPCA and FPS when the singular values of $\mathbf{A}$ are drawn from the chi-square density.	209
7.3	The effect of the number of iterations on the expressed variance and sparsity.	209
7.4	the leading ten PCs extracted by OR-PCA, FPS and OR-SPCA.	210
7.5	The sample face images and the five leading PCs computed by non-negative SPCA and non-negative OR-SPCA.	210

---

8.1	The phase transition properties of Outlier Rejection, Outlier Reduction and Outlier Pursuit in the noiseless case.	248
8.2	The running time of Outlier Rejection and Outlier Reduction, and the relationship between rank $r^*$ and the largest tolerable fraction of outliers for Outlier Reduction.	249
8.3	The comparison between Outlier Rejection, Outlier Reduction and Outlier Pursuit in the noisy case.	250
8.4	The leading five principal components extracted by standard PCA and Outlier Rejection from the MNIST dataset	251
8.5	The leading five principal components extracted by Outlier Reduction from the MNIST dataset.	251
8.6	The leading PC for datasets “Hall”, “Escalator” and “Lobby”.	252
8.7	The dataset with 500 digit images of “1” to “5”.	253
8.8	The inlier samples detected by Outlier Rejection with $\eta = 0.012$ .	254
8.9	The outlier samples identified by Outlier Rejection with $\eta = 0.012$ .	254
8.10	The inlier samples detected by Outlier Rejection with $\eta = 0.01$ .	255
8.11	The outlier samples identified by Outlier Rejection with $\eta = 0.01$ .	256
9.1	Comparison between RO-PCA, MSG, Streaming PCA and Batch.	288
9.2	Comparison between RO-PCA, MSG, online robust PCA [FXMY14] and HR-PCA [XCM13] in the presence of corrupted samples.	290
9.3	Comparison between the three sampling schemes and the baseline that all the attributes are observed.	291
9.4	The empirical performance of RO-PCA and MSG when the digit images are corrupted.	292
9.5	The leading eight PCs extracted by RO-PCA and MSG when the dataset is a mixture of digit and face images.	292

## List of Tables

2.1	The running time for solving the transportation problem with different numbers of suppliers.	50
7.1	The words associated with the leading two sparse principal components extracted by large-scale SPCA and large-scale OR-SPCA.	211
9.1	The average representation errors for different algorithms on a real dataset of 1000 digit images with missing entries.	292
9.2	The average representation errors for different sampling schemes under the limited observation setting.	293

# List of Algorithms

6.1	Naive sampling approach	156
6.2	Linear programming with unobserved constraints via the epsilon-decreasing strategy (LPUC-ED)	158
6.3	Linear programming with unobserved constraints via UCB (LPUC-UCB)	159
6.4	Accelerated LPUC	160
7.1	Outlier-robust PCA-like algorithm	199
8.1	Outlier Rejection	242
8.2	Outlier Reduction	242
8.3	Row-space recovery	243
9.1	A framework for robust online PCA	279
9.2	$\mathbf{Z}_t$ with entry-wise corruption	281
9.3	$\mathbf{Z}_t$ with sample-wise corruption	282
9.4	$\mathbf{Z}_t$ under limited attribute observation	284
9.5	Estimate “missing” probability $\delta$	286
9.6	Estimate the second moment of $\mathbf{x}_t(i)$	287

# CHAPTER 1

## Introduction

In the last few years, we have witnessed the rise of the big data era – numerous artificial intelligent systems have been created to help people make decisions by extracting implicit meaningful information from a large amount of data collected from internet applications, consumer behavior analytics, financial systems and computational biology. Prominent examples include recommendation systems, stock market predictions and disease diagnosis. A nonnegligible fact is that practical data inevitably contains uncertainty, noise and even outliers due to spurious readings, measurement errors, malicious attacking or mislabeling. Substantial research indicates that ignoring the uncertainty and noise can significantly degrade the performance of artificial intelligent systems and lead to unreliable predictions or bad decisions. This thesis focuses on handling uncertain and noisy data in machine learning and decision making problems, which mainly includes two parts: 1) robust optimization for tackling uncertainties in machine learning problems, especially in regression and classification, and 2) robust dimensionality reduction algorithms typically used for data preprocessing in regression or classification to improve prediction accuracy.

Robust optimization has become a popular and widely applied technique for handling uncertainties in optimization problems. The key ingredient of this approach is to define the uncertainty sets consisting of possible realizations of the uncertain parameters in optimization and solve it with the worst-case realizations of the parameters. Previous work showed that robust optimization problems can be computationally tractable and yield more reliable results than the corresponding non-robust ones if the uncertainty sets are selected properly. When the uncertainty sets are chosen poorly, robust optimization can lead to overly conservative solutions or even be intractable. To ease the problem of “overly-conservative”,

distributionally robust optimization is a proper choice for decision-making by taking the advantages of prior distributional knowledge about parameters. This model assumes that the probability distribution of a certain parameter belongs to an ambiguity set containing all the distributions that are compatible with the prior information extracted by the decision maker and finds solutions by minimizing the worst-case expected cost over the distributions in the ambiguity set.

Regression and classification are two fundamental techniques in machine learning, whose goals are estimating prediction rules from the observed samples and the corresponding outputs by minimizing an empirical loss function such that with high probability the prediction for a new sample is close to its true output as much as possible. The observed samples usually contain noisy or missing attributes, which make regression and classification become decision-making problems with uncertain parameters. Therefore, robust optimization and distributionally robust optimization can be applied to mitigate the impact of the noisy samples.

In this thesis, we first investigate the computational aspects of distributionally robust chance constrained optimization problems that have successfully modeled robust support vector machine (SVM) and portfolio optimization under uncertainty. Previous research mainly focused on the case where the constraints are linear in both of the decision variables and the uncertain parameters. We instead consider the case where the constraints can be non-linear in the decision variable, and in particular to the uncertain parameters.

Second, by applying robust optimization and distributionally robust optimization to regression and classification problems, we provide a robustness interpretation of widely applied Lasso-like algorithms, e.g., group Lasso and fused Lasso, and establish the connection between regularized SVMs and the distributionally robust optimization framework. For classification, we also develop an axiomatic framework by proposing a set of salient properties on loss functions and then propose the coherent loss function, revealing a new interpretation for robust SVMs.

Third, we consider the case where the prior information about the uncertainty sets in robust optimization is not available to the decision maker. Specifically, we study optimization



problems with unknown parameters and develop several algorithms for solving them in an online learning fashion. Unlike robust optimization where the unknown parameters lie in some known uncertainty sets, we assume that no prior knowledge about the parameters is available but the feasibility of each constraint can be evaluated at any given measurement point. We show that this problem is a generalized version of stochastic linear optimization and linear bandit problems, and derive the finite time bounds on the regret and the constraint violation for the proposed algorithms.

Finally, besides regression and classification, we study another widely applied technique in machine learning and data analysis – principal component analysis (PCA). It is well known that PCA is notoriously fragile to outlying observations – its performance can dramatically degrade in the presence of even few corrupted samples due to the quadratic error criterion used. In order to handle outliers, we propose: 1) a unified framework for making a wide range of PCA-like algorithms – including the standard PCA, sparse PCA and non-negative sparse PCA, etc. – robust when facing a constant fraction of arbitrarily corrupted outliers; 2) two novel computationally efficient non-convex outlier-robust PCA algorithms with capability of exactly recovering the low-dimensional subspace spanned by the uncorrupted samples; 3) a unified paradigm of online robust PCA via online mirror descent by designing “robust gradients” in the dual space for mirror descent.

## 1.1 Optimization with Uncertain Parameters

Numerous real-world problems can be modeled as the following mathematical optimization problem:

$$\begin{aligned} & \text{Minimize: }_{\mathbf{x}} && f(\mathbf{x}) \\ & \text{Subject to: } && g(\mathbf{x}, \boldsymbol{\xi}) \leq \mathbf{0}, \end{aligned} \tag{1.1}$$

where  $\mathbf{x} \in \mathbb{R}^n$  is the decision variable,  $\boldsymbol{\xi} \in \mathbb{R}^k$  is a vector of the parameters of this problem and  $g(\cdot, \cdot) \in \mathbb{R}^m$  includes all the constraint functions. In real-world applications, problem parameters usually contain uncertainties due to measurement errors or noisy observation. In the seminal papers [BTN98, BTN99], Ben-Tal and Nemirovski pointed out that one cannot

ignore the situations where an acceptable solution should be feasible for all realizations of the parameters and even a small violation of the constraints can lead to meaningless solutions. More specifically, suppose that some prior knowledge about the uncertain parameter  $\boldsymbol{\xi}$  is available to the decision maker, i.e.,  $\boldsymbol{\xi}$  belongs to a known uncertainty set  $\mathcal{U} \subseteq \mathbb{R}^k$ , then the key technique in robust optimization is considering the robust counterpart which is given by the following semi-infinite constraint:

$$g(\mathbf{x}, \boldsymbol{\xi}) \leq \mathbf{0}, \forall \boldsymbol{\xi} \in \mathcal{U} \iff \sup_{\boldsymbol{\xi} \in \mathcal{U}} g(\mathbf{x}, \boldsymbol{\xi}) \leq \mathbf{0} \quad (1.2)$$

The advantages of constraint (1.2) are: 1) it robustifies the constraint in (1.1), in the sense that decision variable  $\mathbf{x}$  is feasible for all the realizations of  $\boldsymbol{\xi} \in \mathcal{U}$ ; 2) it is computationally tractable for a wide range of real-world applications when the uncertainty set is chosen properly; and 3) the uncertainty set is constructed more easily than the distribution of the uncertain parameters from a practical perspective.

This robust counterpart scheme for mathematical programming has been extensively studied in recent decades. In the papers [BTN98, BTN99, BTN00, BTNR02], Ben-Tal and Nemirovski showed that the robust formulations for linear programming (LP), quadratic programming (QP), second order cone programming (SOCP) with ellipsoidal or polyhedral uncertainty sets are tractable, i.e., they can be solved in polynomial time. Bertsimas and Sim [BS03, BS04] proposed a new robust approach for linear programming that is able to adjust the level of conservatism of the robust solutions in terms of probabilistic bounds of constraint violations. Although the robust counterpart is polynomial time solvable, it is more computationally expensive than the nominal problem, e.g., robust LP becomes SOCP and robust SOCP becomes semidefinite programming (SDP). To reduce its computational cost, Bertsimas and Sim [BS06] developed a relaxed robust counterpart for general conic optimization that preserves the computational complexity of the corresponding nominal problem and guarantees the feasibility of the robust solution with a certain probability. Besides LP, QP and SOCP, El Ghaoui et.al [EL97, EOL98] studied robust SDP problems and robust least square problems.

Although robust optimization has beauty for theoretical analysis and simplicity for practi-

cal use, it may lead to overly conservative solutions when the uncertainty set is designed poorly. Its another weakness is that it is difficult to cope with the prior knowledge about the distributions of problem parameters. In contrast with robust optimization, stochastic programming is a framework for handling uncertainties by taking prior distributional knowledge into account, e.g., [BL97, Pre95, KW94]. Suppose that parameter  $\xi$  has distribution  $\mathbb{P}$ , then stochastic programming involves the following constraint

$$\mathbb{E}_{\xi \sim \mathbb{P}}[g(\mathbf{x}, \xi)] \leq \mathbf{0}, \quad (1.3)$$

which guarantees that the expectation constraint should hold. Unfortunately, although the constraint function is convex when  $g(\mathbf{x}, \xi)$  is convex w.r.t.  $\mathbf{x}$ , it is quite computationally challenging to solve it. One possible approach for handling (1.3) is using Monte Carlo approximations [SdM00] which are often computationally costly. Another challenging problem is that in practice distribution  $\mathbb{P}$  is difficult to estimate given only a limited information about parameter  $\xi$ .

In order to address these issues, a robust formulation for stochastic programming called distributionally robust optimization was proposed. Scarf [Sca58] was among the first researchers to investigate distributionally robust optimization. Similar to robust optimization that defines an uncertainty set over problem parameter  $\xi$ , distributionally robust optimization defines a set of probability distributions  $\mathcal{P}$  called ambiguity set that is assumed to include the true distribution  $\mathbb{P}$  of parameter  $\xi$ , and then solve the problem with the worst-case realization in  $\mathcal{P}$ . The following is the distributionally robust counterpart:

$$\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}_{\xi \sim \mathbb{P}}[g(\mathbf{x}, \xi)] \leq \mathbf{0}. \quad (1.4)$$

In practice, the ambiguity set  $\mathcal{P}$  can be constructed from existing domain knowledge or statistical analysis, e.g., estimating the mean and covariance from observed data and taking all the distributions with the estimated mean and covariance as  $\mathcal{P}$ . This model has been extensively studied in recent years, e.g., [CE06, ZKR11, WKS13, DY10, GS10]. Delage *et al.* [DY10] proposed a new ambiguity set taking into account the knowledge of the distribution support and of a confidence region for its mean and its second moment matrix

and applied this model to solve data-driven problems where the knowledge of  $\xi$  can be derived from the historical data. Wiesemann *et al.* [WKS13] developed a unified framework for modeling distributionally robust optimization problems by standardizing ambiguity sets that contain all distributions with prescribed conic representable confidence sets and with mean values residing on an affine manifold. Goh *et al.* [GS10] proposed a simple LDR model to tractably approximate linear distributionally robust optimization problems.

Another paradigm for handling stochastic problem parameters is the celebrated chance constraint approach which has the following formulation:

$$\mathbb{P}_{\xi}[g(\mathbf{x}, \xi) \leq \mathbf{0}] \geq p, \quad (1.5)$$

for some value  $p \in (0, 1)$ . (1.5) ensures that the constraint holds with probability at least  $p$ . Chance constraints were first proposed by Charnes and Cooper [CC59], and since then there has been considerable work, e.g., Miller and Wagner [MW65], Prékopa [Pré70], Delage and Mannor [DM10], and many others. Similar to stochastic programming, it is usually difficult to accurately estimate the distribution of  $\xi$  in practical applications and optimization problems involving chance constraints are notoriously hard to solve, even when  $g(\cdot, \cdot)$  is bilinear and the distribution  $\xi$  is uniform [NS06]. The only known tractable case of this formulation is when  $g(\cdot, \cdot)$  is bilinear and  $\xi$  follows a radial distribution [CG06, AG03]).

To overcome these problems, distributionally robust chance constrained approach has been proposed, e.g., [CG06, EI06, DY10, ZKR11]. In this approach, similar to distributionally robust optimization, the distribution of the uncertain parameter is assumed to belong to a given set  $\mathcal{P}$ . Constraint (1.5) is then replaced with the following constraint

$$\inf_{\mu \in \mathcal{P}} \mathbb{P}_{\xi \sim \mu}[g(\mathbf{x}, \xi) \leq \mathbf{0}] \geq p, \quad (1.6)$$

which requires that for all possible probability distributions of the stochastic uncertainty, the chance constraint must hold. This approach also brings in computational advantages, e.g., Cheung *et al.* [CSW12] developed safe tractable approximations of chance constrained affinely perturbed linear matrix inequalities. Calafiore and El Ghaoui [CG06] showed that

when  $f(\cdot, \cdot)$  is bilinear and  $\mathfrak{P}$  is characterized by the mean and the variance, (1.6) can be converted into a tractable second order cone constraint. Most previous results on the tractability of (1.6) are restricted to the case that  $g(\cdot, \cdot)$  is bilinear. One exception is that Zymler *et al.* [ZKR11] showed that (1.6) is tractable when  $g(\mathbf{x}, \boldsymbol{\xi})$  is linear in the decision variable  $\mathbf{x}$  and quadratic or piecewise linear in  $\boldsymbol{\xi}$ . To the best of our knowledge, the general non-linear case is largely untouched. In Chapter 2, we show that (1.6) is tractable when  $g(\cdot, \cdot)$  concave-quasiconvex and establish a connection between (1.6) and a robust optimization formulation using a deterministic uncertainty model.

In recent decades, many researchers have applied the robust optimization framework in machine learning problems such as classification and regression, e.g., [BGJ<sup>+</sup>04, LEBJ03, SBS06, BTBBN11, Bha04, BPS04, TG07, GR06, XCM09, XCM10, LCG12, SAEK15]. For regression with noisy training samples, the standard learning algorithms can be robustified by directly applying robust optimization or distributionally robust optimization. Xu *et al.* [XCM10] showed that the standard Lasso – the  $l_1$  regularized linear regression – is equivalent to a robust linear regression formulation and such robustness interpretation implies the sparsity and the consistency of the standard Lasso. Shafieezadeh-Abadeh *et al.* [SAEK15] proposed a distributionally robust approach to logistic regression by using the Wasserstein distance to construct a ball in the space of probability distributions centered at the uniform distribution on the training samples, and showed that the proposed formulation is tractable and takes the popular regularized logistic regression problems as its special cases.

Inspired by the success of the standard Lasso, many regularization schemes were proposed to select solutions with more general sparse-like structures. For example, domain knowledge may indicate that the solution is group sparse, i.e., features can be grouped, and the features belonging to one group is likely to be either all non-active or all active. A prominent algorithm proposed to enforce this sparse-like structure is the group Lasso formulation [YL06], where the regularization term is the sum of the  $\ell_2$ -norms of the different groups of features. Other examples of Lasso-like algorithms include the fused Lasso [TSR<sup>+</sup>05] that encourages sparsity of the coefficients and also sparsity of their differences, the sparse group Lasso [FHT10] that encourages solutions that are sparse at both the group and individual feature levels. Although the standard Lasso has been extensively studied from the robustness

perspective [XCM10], the connection between robust optimization and other Lasso-like algorithms such as fused Lasso and group Lasso is still unclear. In Chapter 3, we develop a unified robust linear regression model and show that it is equivalent to a general regularization framework to encourage sparse-like structure that contains group Lasso and fused Lasso as specific examples, which provides a robustness interpretation of these widely applied Lasso-like algorithms, and allows us to construct novel generalizations of Lasso-like algorithms by considering different uncertainty sets.

For classification, the existing robust formulations take one of the two approaches. The first approach treats the problem from the robust optimization perspective similar to robust regression. Xu *et al.* [XCM09] established a strong connection between robust optimization and regularized SVMs and provided a robustness interpretation for the success of regularized SVMs. Globerson *et al.* [GR06] applied the robust optimization formulation to construct classifiers that are robust to deletion of features in test data. The second approach is based on the chance constraints or distributionally robust chance constraints. Lanckriet *et al.* [LEBJ03] considered a binary classification problem where the mean and the covariance of the samples are assumed to be known and then developed a robust classification approach by minimizing the worst-case misclassification probability of future samples via imposing distributionally robust chance constraints on linear decision rules. Shivaswamy *et al.* [SBS06] proposed a robust formulation for SVM with chance constraints and showed that the proposed formulation can be converted into a second order cone program.

A natural question hence emerges: does there exist a universal formulation that can unify all these approaches and inspire new algorithms? In Chapter 4, we show that robust classification via the distributionally robust optimization formulation gives a positive answer, and provide a new distributionally robust optimization interpretation for regularized SVMs which allows us to design new algorithms that are robust to feature corruption. In Chapter 5, we revisit the empirical loss minimization paradigm for classification and propose a new loss function called the coherent loss function defined by a set of salient properties on functions for classification. We show that the proposed approach yields a strictly tighter approximation to the empirical classification error than any convex cumulative loss approach, and provide a new interpretation for robust SVMs from the “coherent loss” perspective.

## 1.2 Optimization with Unknown Parameters

Recall that in robust optimization, we consider the following optimization problem

$$\begin{aligned} & \text{Minimize:}_{\mathbf{x}} && f(\mathbf{x}) \\ & \text{Subject to:} && g(\mathbf{x}, \boldsymbol{\xi}) \leq \mathbf{0}, \end{aligned}$$

where  $\boldsymbol{\xi} \in \mathbb{R}^k$  is a vector of the problem parameters which is assumed to belong to a known uncertainty set  $\mathcal{U}$ . But in some practical applications, we do not always have the exact knowledge about parameter  $\boldsymbol{\xi}$  or its uncertainty set  $\mathcal{U}$ . In other words, the constraints imposed on decision variable  $\mathbf{x}$  can be unknown in real-world problems. For example, the network flow problems, which are usually used to model traffic in a road system, packet flow through network and circulation with demands, can be formulated as linear optimization problems. The decision makers who are trying to find the maximum flow or the minimum cost flow do not always exactly know the capacities or costs of all the edges in the network, e.g., the decision makers do not know the traffic condition in all the roads before they determine the flow of vehicles through a transport network until those vehicles run on their roads and give the traffic report. This kind of examples can be easily duplicated in many applications in machine learning, operations research and finance.

Consider a simpler optimization problem where the constraints are linear w.r.t.  $\mathbf{x}$ , namely,

$$\begin{aligned} & \text{Minimize:}_{\mathbf{x}} && f(\mathbf{x}) \\ & \text{Subject to:} && \mathbf{A}^\top \mathbf{x} \leq \mathbf{b}, \mathbf{x} \in \mathcal{S}, \end{aligned} \tag{1.7}$$

where  $f(\cdot)$  is the cost function,  $\mathcal{S}$  is a closed convex set, and  $\mathbf{A}, \mathbf{b}$  are the parameters of the linear constraint. We assume that  $f(\cdot)$  and  $\mathcal{S}$  are known but  $\mathbf{A}, \mathbf{b}$  are unknown. To the best of our knowledge, this problem has not been explored yet. The most related problems are stochastic linear optimization and contextual linear bandit problems, which attempt to solve the following optimization problem in the online setting:

$$\begin{aligned} & \text{Minimize:}_{\mathbf{x}} && \mathbf{c}^\top \mathbf{x} \\ & \text{Subject to:} && \mathbf{x} \in \mathcal{S}, \end{aligned} \tag{1.8}$$

where the cost vector  $\mathbf{c}$  is assumed to be unknown. Clearly, by introducing a new decision variable  $\beta$  and converting Problem (1.8) into its epigraph form

$$\begin{aligned} & \text{Minimize: } \mathbf{x}, \beta && \beta \\ & \text{Subject to: } && \mathbf{c}^\top \mathbf{x} \leq \beta, \mathbf{x} \in \mathcal{S}, \end{aligned}$$

we know that Problem (1.8) is indeed a special case of Problem (1.7), which implies this problem has a close relationship with the literature of online learning.

The classical multi-armed bandit problem is one of the basic problems in online learning, where in each of  $T$  rounds a learner selects one of  $K$  arms (forming a discrete set  $\mathcal{S}$ ) and subsequently receives a reward independently drawn from an unknown distribution associated with the selected arm. The goal of the learner is to choose a sequence of arms to maximize the cumulated rewards over the  $T$  rounds. This problem has been extensively studied in decades, e.g., [Lai87, Agr95, ACBF02, CBL06, PCA07, BSSM08, MS11]. An extension of the classical multi-armed bandit problem is the contextual multi-armed bandit problem in which each arm associates with a  $d$ -dimensional feature vector called “context” and the reward corresponding to each arm depends on the feature vectors. The set of the feature vectors associated with the arms forms set  $\mathcal{S}$ . The learner’s aim is to explore the relationship between the feature vectors and rewards so that he can predict which arm could provide best reward by examining the feature vectors. The contextual bandits setting with linear payoff functions was first studied by [AL99, ACBF02] and further analyzed by [CLRS11, FCGS10, AYPS11]. In this setting, we assume that there exists an unknown  $\mathbf{c}$  such that the expected reward for an arm given feature vector  $\mathbf{x}$  is  $\mathbf{c}^\top \mathbf{x}$ . When  $\mathcal{S}$  is very large or even infinite, this problem is also called “stochastic linear optimization” [DKH07, DHK08, RT08, Sha13, Sha15a]. One of the most important example is the online linear programming problem as shown in (1.8). Different from the standard linear programming problem with known cost vector  $\mathbf{c}$ , the learner only observes noisy feedback about  $\mathbf{c}$  corresponding to the selected solution in each round.

In Chapter 6, we study Problem (1.7) with unknown constraints in the online fashion where the learner has to select a solution in each round and then receives the corresponding



feedback providing the information about the feasibility of the selected solution. To solve this problem, we develop two algorithms based on the epsilon-decreasing strategies and the upper confidence bound strategy, and provide the theoretical performance of the proposed algorithms. Based on these results, we show that the robust linear programming problems with unknown uncertainty sets can be solved in a data-driven manner as long as the feedback information about the feasibility of each robust constraint for any given input is available to the decision maker.

### 1.3 Principal Component Analysis with Noisy Observation

Besides regression and classification, dimensionality reduction is another fundamental technique in machine learning, mapping data from the original space onto the reduced space. It is well-known that regression and classification can be done more accurately in the reduced space than the original space. Principal component analysis (PCA) [Pea01] is arguably the most widely applied dimensionality reduction method, playing a significant role in a broad range of areas including machine learning, statistics, finance and many others. The standard PCA performs the spectral decomposition of the sample covariance matrix, selects the eigenvectors corresponding to the largest eigenvalues, and then constructs a low dimensional subspace based on the selected eigenvectors. It is well known that standard PCA, depending on different applications, may suffer from three weaknesses [MR14, XCM13, JL09]: 1) PCA is notoriously fragile to outliers – indeed, its performance can significantly degrade in the presence of even few corrupted samples, due to the quadratic error criterion used; 2) PCA cannot utilize additional information of the principal components: e.g., in certain applications, it is known that the principal components should lie in the positive orthant; 3) its output may lack interpretability since it does not encourage sparse solutions.

In recent years, numerous robust PCA algorithms have been proposed to address the first issue [DGK81, XY95, YW99, ITB03, Das03, XCM13, FXY12]. Among them, Xu *et al.* [XCM13] successfully tackles the case where a constant fraction of samples are corrupted in the high dimensional regime. Their proposed method is tractable, easily kernelizable, and is able to robustly estimate the principal components even in the face of a constant fraction

of outliers and very low signal-to-noise ratio. To address the second weakness, Montanari *et al.* [MR14] recently proposed a new algorithm called non-negative PCA which handles the case that the principal components are known to lie in the positive orthant. But similar to the standard PCA, this algorithm is sensitive to outliers. Indeed, the estimated principal components can be far from the true ones in the face of even few outliers. To address the third weakness, previous works focus on a class of methods called sparse PCA that adapt the standard PCA so that only a few of attributes of the resulting principle components are non-zero, e.g., [VCLR13, ZHT06, SH08, JYN08, BJNP13, VL13, dEJL07, TDT10]. For example, Vu *et al.* [VCLR13] proposed a convex relaxation formulation of sparse PCA based on a semi-definite program with a Fantope constraint and established theoretical guarantees in the outlier-free regime. Yet, one severe drawback of most sparse PCA algorithms is that they are sensitive to the existence of even few outliers. This is clearly undesirable, as in real-world applications, the existence of outliers is ubiquitous. Recently, several robust sparse PCA have been proposed [CFF13, WC12, HRS14] to handle outliers, but all of them are only evaluated by experiments and have no theoretical performance guarantees.

In Chapter 7, we theoretically address these issues of PCA simultaneously. Specifically, we propose a general framework for a wide range of PCA-like algorithms to make them provably robust to a constant fraction of arbitrary outliers. Our framework has the capability of converting a non-robust PCA-like algorithm such as non-negative PCA [MR14], sparse PCA [VCLR13, PDK13] or non-negative sparse PCA [APD14], into its outlier-robust variant.

Recently, borrowing ideas from compressive sensing, a prominent new approach for addressing the first issue has been extensively studied, which decomposes the noisy sample matrix  $\mathbf{X}$  into a low-rank matrix  $\mathbf{L}^*$  and a sparse matrix  $\mathbf{S}^*$  via nuclear norm minimization, e.g., [CR09, CLMW11, RFP10, CSPW11, XCS12]. Among them, Xu *et al.* [XCS12] proposed a nuclear norm based algorithm called Outlier Pursuit to handle corrupted samples, where they assumed that  $\mathbf{S}^*$  is column-wise sparse instead of entry-wise sparse. The goal of Outlier Pursuit is to exactly recover the column space of the low-rank matrix  $\mathbf{L}^*$  and identify the nonzero columns of  $\mathbf{S}^*$ . They proved that exact recovery can be achieved under mild conditions depending on the incoherence of the row space of  $\mathbf{L}^*$  and the fraction of outliers.

While nuclear norm based algorithms have elegant theoretical results, they can be difficult to apply to large-scale applications due to high computational cost.

In order to reduce the computational cost, in Chapter 8, we develop two novel non-convex algorithms for outlier-robust PCA called Outlier Rejection and Outlier Reduction, which involve alternating between estimating the low-rank column space of  $\mathbf{L}^*$  and identifying the outliers indicated by  $\mathbf{S}^*$ . In comparison with Outlier Pursuit, the proposed algorithms have much lower computational load, yet enjoy similar performance guarantees for the exact recovery of the true column space.

Besides developing non-convex variants of PCA algorithms, another approach for designing computational-efficient PCA algorithms is based on the online setting, or online PCA, where one receives a sample sequentially and this sample vanishes after it is collected unless it is stored in the memory, e.g., [WK08, MCJ13, ACLS12, ACS13, YX15a, Bra02, ACS13, Sha15b]. These algorithms typically take one of the two approaches: 1) block-wise stochastic power methods, e.g., the memory efficient PCA developed by Mitliagkas *et al.* [MCJ13] performs a power iteration update on the estimated PCs once a block of new samples are received; and 2) stochastic convex optimization, e.g., stochastic PCA proposed by Arora *et al.* [ACS13] performs a matrix stochastic gradient descent when a new sample arrives. The weakness of these algorithms is that they cannot handle outliers or missing entries existing in the received samples.

In Chapter 9, we consider a unified paradigm on online PCA via online mirror descent – a general framework for developing and analyzing first-order online learning algorithms, e.g., [SST11, SS12, OCC15]. By designing proper robust gradients used in mirror descent, we propose new online PCA algorithms that are robust to various types of data defect such as missing entries, corrupted attributes or outliers, and establish finite-sample performance guarantees, which is a distinctive feature of the proposed paradigm.

## 1.4 Structure of the Thesis

This thesis is organized as follows:

**Chapter 2. Distributionally Robust Chance Constraints for Non-Linear Uncertainties.** The computational aspects of distributionally robust chance constrained optimization are investigated in this chapter, where the uncertainty is characterized by its mean and variance, and the constraint function is non-linear – concave in the decision variables and quasi-convex in the uncertain parameters, in contrast to bilinear constraint functions considered in previous work. Furthermore, an equivalence relationship between distributionally robust chance constrained optimization and robust optimization is established, which links two broadly applied paradigms in decision making under uncertainty and extends previous results of the same spirit in the linear case to more general cases. Finally, a generalization of distributionally robust chance constraints called probabilistic envelope constraints is studied in the non-linear case.

**Chapter 3. A Unified Robust Regression Model for Lasso-like Algorithms.** In this chapter, a unified paradigm between robustness and regularization schemes for various sparse-like structures containing group Lasso and fused Lasso is established via a unified robust linear regression model. This model provides a robustness interpretation of these widely applied Lasso-like algorithms, and forms a new way to construct novel generalizations of Lasso-like algorithms by considering different uncertainty sets. Based on this robustness interpretation, the sparsity and statistical consistency properties of Lasso-like algorithms are explored from a new robustness perspective.

**Chapter 4. A Distributionally Robust Optimization Interpretation For Regularized SVMs.** Similar to Chapter 3, a unified framework based on distributionally robust optimization is proposed in this chapter for designing robust classification methods. This framework establishes a close relationship with previous robust classification approaches tackling data uncertainty using robust optimization, and provides a distributionally robust optimization interpretation for regularized SVMs and robust SVMs.

**Chapter 5. The Coherent Loss Function for Classification.** The goal of classification is to find a prediction rule leading to a small misclassification error, which can be achieved by minimizing the cumulative loss – the sum of convex surrogates of the 0-1 loss of each sample. In this chapter, instead of using the cumulative loss, a new loss function called

coherence loss function is proposed by developing an axiomatic framework based on the salient properties on loss functions for classification. This approach yields a strictly tighter approximation to the empirical classification error than any convex cumulative loss approach while preserving the convexity of the underlying optimization problem, and provides a new perspective on understanding the robust formulation of SVM proposed by Shivaswamy *et al.* [SBS06].

**Chapter 6. Online Linear Optimization with Unobserved Constraints.** In some practical applications, the exact knowledge about the problem parameters or their corresponding uncertainty sets is not always available to the decision maker. To address this issue, this chapter considers to solve optimization problems with unknown constraints. More specifically, we investigate online linear optimization with unknown constraints, where in each round the decision maker chooses a solution from the known decision set and subsequently receives some feedback information about the feasibility of her choice w.r.t. the additional unknown constraints. This model takes stochastic linear optimization problems and contextual linear bandit problems as its special cases. To solve it numerically, two algorithms are proposed, namely, LPUC-ED based on the epsilon-decreasing strategy and LPUC-UCB based on the upper confidence bound strategy. Finally, the finite time bounds on the regret and the constraint violation of the proposed algorithms are provided.

**Chapter 7. A Unified Framework for Outlier-Robust PCA-like Algorithms.** From this chapter on, we will focus on robust dimensionality reduction methods. One well-known weakness of the standard principal component analysis is that its performance dramatically degrades in the presence of even few corrupted samples. To address this issue, this chapter proposes a unified framework for robustifying a wide range of PCA-like algorithms when facing a constant fraction of arbitrarily corrupted outliers. This framework is inspired by HR-PCA [XCM13], but overcomes the drawbacks of HR-PCA and has the capability of converting a non-robust PCA-like algorithm such as non-negative PCA [MR14], sparse PCA [VCLR13, PDK13] or non-negative sparse PCA [APD14], into its outlier-robust variant. Furthermore, it is shown that the proposed framework has solid theoretical performance guarantees, i.e., its estimation error is upper bounded by a term depending on the intrinsic parameters of the data model, the underlying PCA-like algorithm and the fraction of

outliers.

**Chapter 8. Non-convex Outlier-Robust PCA.** Recently, a prominent new approach for robust PCA has been extensively studied, which tries to decompose the noisy sample matrix into a low-rank matrix and a sparse matrix via nuclear norm minimization, e.g., Outlier Pursuit proposed by Xu *et al.* [XCS12]. The problem of nuclear norm based algorithms is that it is difficult to apply them to large-scale applications due to high computational cost. In this chapter, we develop two computationally efficient non-convex outlier-robust PCA algorithms. These two algorithms can be viewed as non-convex counterparts of Outlier Pursuit, which alternatively estimate the low-dimensional subspace spanned by the principal components and mitigate the effect of outlier samples. It is shown that they own similar theoretical performance guarantees with much lower computational complexity compared to Outlier Pursuit.

**Chapter 9. Online PCA with Imperfect Data.** Online PCA is commonly applied in large-scale applications, where the samples are assumed to be collected sequentially. Various online PCA algorithms have been recently developed, but most of which are fragile to even few outliers. This chapter considers a unified paradigm on online PCA via online mirror descent, and then provides a systematic way to develop new robust online PCA algorithms by designing proper robust gradients in the dual space for mirror descent. Theoretical analysis shows that the proposed algorithms from this framework have finite sample performance guarantees.

**Chapter 10. Conclusion.** This chapter summarizes the thesis and discusses the future work.

# CHAPTER 2

## Distributionally Robust Chance Constraints for Non-Linear Uncertainties

This chapter investigates the computational aspects of distributionally robust chance constrained optimization problems. In contrast to previous research that mainly focused on the linear case (with a few exceptions discussed in detail below), we consider the case where the constraints can be non-linear to the decision variable, and in particular to the uncertain parameters. This formulation is of great interest as it can model non-linear uncertainties that are ubiquitous in applications. Our main result shows that distributionally robust chance constrained optimization is tractable, provided that the uncertainty is characterized by its mean and variance, and the constraint function is concave in the decision variables, and quasi-convex in the uncertain parameters. En route, we establish an equivalence relationship between distributionally robust chance constraint and the robust optimization framework that models uncertainty in a deterministic manner. This links two broadly applied paradigms in decision making under uncertainty and extends previous results of the same spirit in the linear case to more general cases. We then consider probabilistic envelope constraints, a generalization of distributionally robust chance constraints first proposed in Xu *et al.* [XCM12] for the linear case. We extend this framework to the non-linear case, and derive sufficient conditions that guarantee its tractability. Finally, we investigate tractable approximations of joint probabilistic envelope constraints, and provide the conditions when these approximation formulations are tractable.

## 2.1 Introduction

Many optimization and decision making problems, when facing stochastic parameter uncertainty, can be tackled via the celebrated *chance constraint* paradigm. Here, a deterministic constraint is relaxed, and instead is required to hold with a certain probability (w.r.t. the uncertain parameter). That is, given a constraint  $f(\mathbf{x}, \boldsymbol{\delta}) \geq \alpha$  where  $\mathbf{x}$  denotes the decision variable,  $\alpha \in \mathbb{R}$  denotes the target value, and  $\boldsymbol{\delta}$ , the uncertain parameter, follows a distribution  $\mu$ , one solves:

$$\mathbb{P}_{\boldsymbol{\delta} \sim \mu}[f(\mathbf{x}, \boldsymbol{\delta}) \geq \alpha] \geq p, \quad (2.1)$$

for some value  $p \in (0, 1)$ . Chance constraints were first proposed by Charnes and Cooper [CC59], and since then there has been considerable work, e.g., Miller and Wagner [MW65], Prékopa [Pré70], Delage and Mannor [DM10], and many others; we refer the reader to the textbook by Prékopa [Pre95] and references therein for a thorough review.

While the chance constraint formulation is conceptually intuitive, it has two disadvantages that limit its practical applications. First, it is usually difficult to obtain enough samples to accurately estimate the distribution  $\mu$ . Second, optimization problems involving chance constraints are notoriously hard to solve, even when  $f(\cdot, \cdot)$  is bilinear (i.e., linear in either argument) and  $\mu$  is a uniform distribution (Nemirovski and Shapiro [NS06]). Indeed, the only known tractable case of the chance constraint formulation is when  $f(\cdot, \cdot)$  is bilinear and  $\mu$  follows a radial distribution (Calafiore and El Ghaoui [CG06]; Alizadeh and Goldfarb [AG03]).

A natural extension of the chance constraint paradigm that overcomes the above mentioned problems is the *distributionally robust chance constrained* (DRCC) approach (e.g., Calafiore and El Ghaoui [CG06], Erdogan and Iyengar [EI06], Delage and Ye [DY10], Zymler *et al.* [ZKR11]). In this paradigm, the distribution of the uncertain parameter is not precisely known, but instead, it is assumed to belong to a given set  $\mathfrak{P}$ . Constraint (2.1) is then replaced with the following constraint

$$\inf_{\mu \in \mathfrak{P}} \mathbb{P}_{\boldsymbol{\delta} \sim \mu}[f(\mathbf{x}, \boldsymbol{\delta}) \geq \alpha] \geq p. \quad (2.2)$$



In words, (2.2) requires that for all possible probability distributions of the stochastic uncertainty, the chance constraint must hold. Typically,  $\mathfrak{P}$  is characterized by the mean, the covariance, and sometimes the support of the distribution as well, all of which can be readily estimated from finite samples. The DRCC approach also brings in computational advantages, e.g., Cheung *et al.* [CSW12] developed safe tractable approximations of chance constrained affinely perturbed linear matrix inequalities. A celebrated result by Calafiore and El Ghaoui [CG06] shows that when  $f(\cdot, \cdot)$  is bilinear and  $\mathfrak{P}$  is characterized by the mean and the variance, DRCC (2.2) can be converted into a *tractable* second order cone constraint.

Yet, most previous results on the tractability of DRCC are restricted to the case that  $f(\cdot, \cdot)$  is bilinear, whereas not much has been discussed when  $f(\cdot, \cdot)$  is non-linear. One exception that we are aware of is Zymler *et al.* [ZKR11], where they showed that DRCC is tractable when  $f(\mathbf{x}, \boldsymbol{\delta})$  is linear in the decision variable  $\mathbf{x}$  and *quadratic* or *piecewise linear* in the uncertainty  $\boldsymbol{\delta}$ . However, their method is built upon the S-lemma, and hence it is not clear how to extend the method to more general cases. Another one is Cheng *et al.* [CDL13] where they studied the knapsack problem with distributionally robust chance constraints when  $f(\mathbf{x}, \boldsymbol{\delta})$  is *piecewise linear* in the uncertainty  $\boldsymbol{\delta}$  and provided its equivalent formulation when the first and second moment and the support information of  $\boldsymbol{\delta}$  are known. To the best of our knowledge, the *general* non-linear case is largely untouched.

This chapter is devoted to analyzing the tractability of DRCC (and its variants) under general – i.e., *non-linear* –  $f(\cdot, \cdot)$ . This problem is of interest, because in many applications the uncertainty is inherently non-linear, and cannot be modeled using a bilinear  $f(\cdot, \cdot)$ , e.g., [BS73, KT11b, ZKR12]; see Section 2.2 for a more detailed discussion. In particular, we consider the following constraint

$$\inf_{\boldsymbol{\delta} \sim (0, \boldsymbol{\Sigma})} \mathbb{P}[f(\mathbf{x}, \boldsymbol{\delta}) \geq \alpha] \geq p, \quad (2.3)$$

where  $f(\mathbf{x}, \boldsymbol{\delta})$  is *concave* in  $\mathbf{x}$ , and *quasi-convex* in  $\boldsymbol{\delta}$ . Here, following the notations from Xu *et al.* [XCM12], we use  $(0, \boldsymbol{\Sigma})$  to denote all distributions with mean zero and variance  $\boldsymbol{\Sigma}$ , and let  $\boldsymbol{\delta} \sim (0, \boldsymbol{\Sigma})$  stand for  $\boldsymbol{\delta}$  follows some unknown distribution  $\mu$  that belongs to

$(0, \Sigma)$ . Notice that DRCC is a special case of *distributionally robust optimization* (e.g., [Sca58, Dup87, Pop07, DY10]) by setting the utility function to the indicator function. However, because the indicator function is neither convex nor concave in either argument, previous results on the tractability of DRO do not apply in our setup.

Our first contribution, presented in Section 2.3, establishes that Constraint (2.3), when  $f(\cdot, \cdot)$  is concave-quasiconvex, is tractable. En route, we derive an equivalence relationship between (2.3) and a robust optimization formulation using a deterministic uncertainty model (e.g., Ben-Tal *et al.* [BTN98, BTN99, BTdHV12] Bertsimas and Sim [BS04]). This result thus links the two arguably most widely used approaches in optimization under uncertainty, and extends previous results of the same spirit for the linear case (e.g., Delage and Mannor [DM10], Shivaswamy *et al.* [SBS06]).

Our second result, presented in Section 2.4, establishes the tractability of the probabilistic envelope model in the non-linear case. The probabilistic envelope model is proposed in Xu *et al.* [XCM12], based on the following observation: the chance constraint (2.1) only guarantees that the given constraint will be satisfied with probability  $p$  or violated with the remaining  $(1 - p)$  probability, but no control is provided on the degree of violation. To overcome this, Xu *et al.* [XCM12] proposed the probabilistic envelope constraint framework – essentially a set of infinite number of chance constraints at *all* levels of potential violation. That is, replace the single DRCC in (2.3) with the following

$$\inf_{\delta \sim (0, \Sigma)} \mathbb{P}[f(\mathbf{x}, \delta) \geq \alpha - s] \geq \mathfrak{B}(s), \quad \forall s \geq 0, \quad (2.4)$$

where  $\mathfrak{B}(s)$  is a given non-decreasing and right-continuous function of  $s$ . However, only the bilinear case has been investigated. In this chapter, we extend the probabilistic envelope constraint to non-linear uncertainties. We prove that the optimization problem involving the probabilistic envelope constraint (2.4) is tractable when  $f(\cdot, \cdot)$  is concave-quasiconvex and  $\mathfrak{B}(s)$  satisfies some weak conditions. Similarly as for the (single) DRCC case, we establish a linkage between probabilistic envelope constraints and the comprehensive robust optimization framework using a deterministic uncertainty model (Ben-tal *et al.* [BTBN06, BTBB10]).

It is worthwhile to note that the probabilistic envelope constraint is closely related to *stochastic dominance constraints* in the literature of stochastic programming (Dentcheva and Ruszczyński [DR03, DR04a, DR04b]); see Chapter 4 of the book by Shapiro *et al.* [SDR09] for more details. A stochastic dominance constraint refers to a constraint of the form  $X \succeq_{(k)} Y$  where  $X$  and  $Y$  are random variables and  $\succeq_{(k)}$  stands for  $k$ -th order stochastic dominance. Thus, a probabilistic envelop constraint is indeed a first-order stochastic dominance constraint with the right hand side is a random variable whose cumulative distribution function is  $\mathfrak{B}(s)$ . However, most of the literature in optimization with stochastic dominance constraints does not address this specific case and instead focuses on the second (or higher) order constraints case, a case that preserves convexity and is more amenable to analysis. As we restrict our attention to this specific case, we choose to use the name “probabilistic envelop constraint”.

Finally, we extend our results in two ways, namely, more flexible uncertainty modeling and joint constraints. In Section 2.5, we provide tractability results for the case where the mean and variance themselves are unknown, and the case that the mean and the support of the distribution of the uncertain parameters are known. For more general uncertainty models where exact results appear difficult, we provide a conservative approximation scheme based on CVaR approximation of the chance constraints. In Section 2.6, we extend the probabilistic envelope constraint formulation to its joint chance constraint counterpart. This typically leads to a computationally challenging problem, and we adopt the CVaR approximation approach proposed by Zymler *et al.* [ZKR11], and show that the joint probabilistic envelope constraint can be approximated tractably under some technical conditions.

**Notation.** We use lower-case boldface letters to denote column vectors, upper-case boldface letters to denote matrices, and the transpose (superscript  $\top$ ) of the column vectors to denote row vectors. The all-ones vector is denoted by  $\mathbf{1}$ . The space of symmetric matrices of dimension  $n$  is denoted by  $\mathbb{S}^n$ . For any two matrices  $\mathbf{X}, \mathbf{Y} \in \mathbb{S}^n$ ,  $\langle \mathbf{X}, \mathbf{Y} \rangle = \text{tr}(\mathbf{X}\mathbf{Y})$  denotes the trace scalar product, and the relation  $\mathbf{X} \succeq \mathbf{Y}$  ( $\mathbf{X} \succ \mathbf{Y}$ ) implies that  $\mathbf{X} - \mathbf{Y}$  is positive semi-definite (positive definite). Random variables are always represented by  $\delta$ . Finally, we call an optimization problem *tractable* if it can be solved in polynomial time and call a set *tractable* if it is convex and a polynomial-time separation oracle can be constructed.

## 2.2 Formulation and Motivating Examples

We first propose the distributionally robust chance constraint, the probabilistic envelope constraint and the joint probabilistic envelope constraint discussed in this chapter. For clarity, we repeat some of the definitions given in the introduction. Given a random variable  $\delta$  and a function  $f(\mathbf{x}, \delta)$ , a chance constraint places a lower-bound on the probability that the constraint reaches a certain target, which is defined as

$$\textbf{Distributionally Robust Chance Constraint: } \inf_{\delta \sim (0, \Sigma)} \mathbb{P}[f(\mathbf{x}, \delta) \geq \alpha] \geq p. \quad (2.5)$$

As discussed above, the distributionally robust chance constraint provides protection against noise by bounding the probability of failing to achieve a pre-defined target  $\alpha$ . It says nothing about what happens when, with probability at most  $(1 - p)$ , the target is not met. In particular, there is no control over the magnitude of violation of the constraint. To overcome this shortcoming, the probabilistic envelope constraint is proposed, which can enforce all levels of probabilistic guarantees. Given a non-decreasing function  $\mathfrak{B}(s)$ , the probabilistic envelope constraint can be written as

$$\textbf{Probabilistic Envelope Constraint: } \inf_{\delta \sim (0, \Sigma)} \mathbb{P}[f(\mathbf{x}, \delta) \geq \alpha - s] \geq \mathfrak{B}(s); \forall s \geq 0. \quad (2.6)$$

For example, if we want the probability of *large* constraint violation to decrease exponentially, then we can set  $\mathfrak{B}(s) = 1 - \gamma \exp(-\beta s)$ . Besides the individual probabilistic envelope constraint discussed above, we propose the following joint probabilistic envelope constraint (JPEC):

$$\textbf{JPEC: } \inf_{\delta \sim (0, \Sigma)} \mathbb{P}[f_i(\mathbf{x}, \delta) \geq \alpha_i - s, \forall i = 1, \dots, m] \geq \mathfrak{B}(s); \forall s \geq 0. \quad (2.7)$$

Computationally, the joint envelope constraint is more complicated. A common method to simplify it is to decompose it into  $m$  individual envelope constraints by applying Bonferroni's inequality. However, since Bonferroni's inequality is not tight, this approximation method is usually overly conservative. In this chapter, we use the worst-case CVaR method proposed

by Zymler *et al.* [ZKR11] to give a tractable and tighter approximation for this joint envelope constraint.

Although the three types of constraints (2.5), (2.6), and (2.7) above can be general, they may not be tractable due to the non-convex feasible sets. To ensure tractability, we focus on the “concave-quasiconvex” case, i.e., the function  $f$  is concave w.r.t. the decision variable, and quasi-convex w.r.t. the uncertain parameters, see the following for a precise description:

**Assumption 2.1.** *Let  $\mathbb{X}$  and  $\mathbb{Y}$  be two convex sets, and let  $f$  be a function mapping from  $\mathbb{X} \times \mathbb{Y}$  to  $\mathbb{R}$ ,*

1. *For each  $\mathbf{x} \in \mathbb{X}$ , the function  $f(\mathbf{x}, \cdot)$  is quasi-convex and continuous on  $\mathbb{Y}$ . For each  $\mathbf{y} \in \mathbb{Y}$ , the function  $f(\cdot, \mathbf{y})$  is concave on  $\mathbb{X}$ .*
2. *The uncertainty  $\delta$  is modeled as a random variable whose mean and variance are known but its distribution is unknown. Without loss of generality, we assume the mean is zero.*

Notice that Assumption 2.1 generalizes the case where  $f(\cdot, \cdot)$  is bilinear – a setup that previous literature mainly focused on – to the non-linear case. In particular, the uncertainty can be non-linear. Bi-linearity and non-linearity of uncertainty arises naturally in a broad range of applications, as we demonstrate by the following examples.

### Example 1: Classification Under Uncertainty

The goal of classification is to predict the unknown label  $y$  of an observed sample  $\mathbf{x}$ . The relationship between label  $y$  and sample  $\mathbf{x}$  can be learned from a finite set of samples  $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ . For a binary classification problem where  $y_i$  is chosen from  $\{-1, 1\}$ , we try to construct a hyperplane  $(\mathbf{w}, b)$  to separate the two classes, e.g., for linearly separable data set,  $y_i = 1$  if  $\mathbf{w}^\top \mathbf{x}_i + b \geq 0$  or  $-1$  otherwise. Therefore, after hyperplane  $(\mathbf{w}, b)$  is constructed, the decision rule can be  $h(\mathbf{x}) = \text{sign}(\mathbf{w}^\top \mathbf{x} + b)$ . Support Vector Machine (SVM) [CV95] is one of the most famous and widely applied classification algorithm. The formulation of the

$l_1$  regularized SVM is as follows:

$$\begin{aligned} \text{Minimize: } \mathbf{w}, b, \xi \quad & \sum_{i=1}^n \xi_i \\ \text{Subject to: } \quad & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \forall i = 1, \dots, n, \\ & \xi_i \geq 0, \quad \forall i = 1, \dots, n, \\ & \|\mathbf{w}\| \leq B. \end{aligned}$$

In this formulation, we assume that the observed samples  $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  are certain. When the samples are uncertain, e.g., we suppose that the true sample  $\tilde{\mathbf{x}}_i$  follows a certain distribution with mean  $\mathbf{x}_i$  and covariance  $\Sigma_i$  for each  $i = 1, \dots, n$ , then we can apply the following robust formulation for SVM [SBS06]:

$$\begin{aligned} \text{Minimize: } \mathbf{w}, b, \xi \quad & \sum_{i=1}^n \xi_i \\ \text{Subject to: } \quad & \inf_{\tilde{\mathbf{x}}_i \sim (\mathbf{x}_i, \Sigma_i)} \mathbb{P} \left[ y_i(\mathbf{w}^\top \tilde{\mathbf{x}}_i + b) \geq 1 - \xi_i \right] \geq 1 - \kappa, \quad \forall i = 1, \dots, n, \\ & \xi_i \geq 0, \quad \forall i = 1, \dots, n, \\ & \|\mathbf{w}\| \leq B, \end{aligned}$$

which means that even for the worst-case distribution the samples should be correctly classified with probability at least  $1 - \kappa$ . Clearly, the distributionally robust chance constraints above satisfy Assumption 2.1.

### Example 2: Portfolio Optimization

Consider a stylized portfolio optimization problem, where an amount is to be allocated to  $n$  stocks and held for a time period  $T$ . Denote the price of the  $i$ th stock after time  $T$  by  $S_i$ , and our goal is to maximize the Value at Risk (VaR) of the total return of the portfolio, which leads to the following formulation, for a fixed  $\gamma \in (0, 1)$

$$\begin{aligned} \text{Maximize: } \mathbf{x} \geq 0, z \quad & z \\ \text{Subject to: } \quad & \mathbb{P} \left[ \sum_i S_i x_i \geq z \right] \geq 1 - \gamma; \quad \mathbf{1}^\top \mathbf{x} = 1, \end{aligned}$$

where  $x_i$  is the allocation for the  $i$ th stock. It is well believed that the true drivers of the uncertainty in stock price is not the stock return  $S_i$  itself, but instead the compounded rates of return, i.e.,  $S_i = \exp(\delta_i)$  where  $\delta_i$  is the random variable to model and analyze. For example, the celebrated log-normal model, pioneered by Black and Scholes [BS73], models  $S_i$  as  $S_i = \exp((\mu_i - \sigma_i^2/2)T + \sqrt{T}\xi_i)$  where the vector  $\xi$  is Normally distributed with mean 0 and covariance matrix  $\mathbf{Q}$ . This can be rewritten as  $S_i = \exp(\delta_i)$  where  $\delta \sim \mathcal{N}((\mu_i - \sigma_i^2/2)T, T\mathbf{Q})$ .

One common criticism of the log-normal model is that it assumes  $\xi$  to be Gaussian, whereas empirical evidence suggests that  $\xi$  (and hence  $\delta$ ) is fat-tailed (e.g., Jansen and deVries [Jd91], Cont [Con01], Kawas and Thiele [KT11a]). Since the Gaussian assumption ignores the fat tails, it essentially leads the managers to take more risk than she is willing to accept. On the other hand, it remains controversial about what is the most appropriate fat-tail distribution to use in modeling returns [Fam65, Kon84, Jd91, Con01], and “this controversy has proven hard to resolve” as Jensen and de Vries stated [Jd91]. In light of this, one possible approach is to not commit to any distribution, but instead only require that the first two moments match. This leads to the following problem:

$$\begin{aligned} & \text{Maximize: } \mathbf{x} \geq 0, z & z \\ & \text{Subject to: } \inf_{\delta \sim ((\mu_i - \sigma_i^2/2)T, T\mathbf{Q})} \mathbb{P}\left[\sum_i \exp(\delta_i)x_i \geq z\right] \geq 1 - \gamma; \mathbf{1}^\top \mathbf{x} = 1, \end{aligned} \tag{2.8}$$

Observe that this formulation satisfies Assumption 2.1, i.e., the constraint is linear to the decision variable and non-linearly convex to the uncertain parameters, and the decision variables are non-negative.

In portfolio optimization, options are another cause of non-linearity of the uncertainty (Kawas and Thiele [KT11b], Zymmler *et al.* [ZKR12]). Suppose for each stock, the investor is allowed to purchase an European call option at the price of  $c_i$  per unit, which gives her the right to buy a unit of stock  $i$  at time  $T$  with the strike price  $p_i$ . Thus, denote the stock return as  $S_i$ , the return of this option is  $\max(S_i - p_i, 0)$ , since the investor will execute the

option if and only if  $S_i > p_i$ . The portfolio optimization problem is thus formulated as

$$\begin{aligned} \text{Maximize: } & \mathbf{x} \geq 0, \mathbf{y} \geq 0, z && z \\ \text{Subject to: } & \mathbb{P} \left[ \sum_i (S_i x_i + \max(S_i - p_i, 0) y_i) \geq z \right] \geq 1 - \gamma; \\ & \sum_i (x_i + c_i y_i) = 1, \end{aligned}$$

where  $\mathbf{y}$  is the investment of the European call options. Notice that the constraints are non-linear, yet convex to  $S_i$ . Indeed, following the previous argument, we may further model  $S_i = \exp(\delta_i)$ , and require that the first two moments of  $\boldsymbol{\delta}$  are known. This makes the probabilistic constraint again satisfy Assumption 2.1.

### Example 3: Transportation Problem

Solving multi-stage optimization problems may also result in non-linearity of uncertainty and decision variables. We illustrate this using a transportation decision problem. Given a directed graph  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , and let  $\mathcal{S} \subset \mathcal{V}$  be the set of source nodes, and  $\mathcal{D} \subset \mathcal{V}$  be the set of destination nodes, with  $\mathcal{S} \cap \mathcal{D} = \emptyset$ . One can think of each node in  $\mathcal{S}$  as a supplier, and each node in  $\mathcal{D}$  as a consumer.

The decision to make contains two stages: in the first stage, the decision maker needs to decide the required flow of each source node and each destination node, i.e.,  $s(i)$  for  $i \in \mathcal{S}$  and  $d(j)$  for  $j \in \mathcal{D}$ . One can think of this as deciding how much amount of good to order from each supplier, and how much to sell to each client. Certain linear constraints on the required flow are imposed: for example, the total supply equals to the total demand, and they must be larger than a minimum demand  $L$ , i.e.,  $\sum_{i \in \mathcal{S}} s(i) = \sum_{j \in \mathcal{D}} d(j) \geq L$ .

In the second stage, after all the ordered goods are produced by the suppliers, the decision maker needs to decide how to transport these goods, i.e., the flow on the network from sources to destinations, by solving a minimum cost flow problem given  $s_i$  and  $d_j$ . This can be formulated as a linear program, where the decision variable  $f(u, v)$  is the flow from node



$u$  to node  $v$ :

$$\begin{aligned}
\text{Minimize:} & \quad \sum_{(u \rightarrow v) \in \mathcal{E}} \delta(u, v) f(u, v) \\
\text{Subject to:} & \quad \sum_{(u \rightarrow v) \in \mathcal{E}} f(u, v) - \sum_{(v \rightarrow u) \in \mathcal{E}} f(v, u) = 0 \quad \forall u \notin \mathcal{S} \cup \mathcal{D}; \\
& \quad \sum_{(u \rightarrow v) \in \mathcal{E}} f(u, v) - \sum_{(v \rightarrow u) \in \mathcal{E}} f(v, u) = s(u) \quad \forall u \in \mathcal{S}; \\
& \quad \sum_{(u \rightarrow v) \in \mathcal{E}} f(u, v) - \sum_{(v \rightarrow u) \in \mathcal{E}} f(v, u) = -d(u) \quad \forall u \in \mathcal{D}; \\
& \quad f(u, v) \geq 0 \quad \forall (u \rightarrow v) \in \mathcal{E}; \\
& \quad f(u, v) = 0 \quad \forall (u \rightarrow v) \notin \mathcal{E}.
\end{aligned}$$

Denote the optimal value by  $h(\mathbf{s}, \mathbf{d}, \boldsymbol{\delta})$ . Suppose  $\boldsymbol{\delta}$  represents uncertain parameters whose values are only revealed at stage two, then to ensure that the total transportation cost is low with high probability, the first stage decision can be formulated using DRCC:

$$\begin{aligned}
\text{Maximize:}_{\mathbf{s} \geq 0, \mathbf{d} \geq 0, z} & \quad z \\
\text{Subject to:} & \quad \inf_{\boldsymbol{\delta} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})} \mathbb{P}[-h(\mathbf{s}, \mathbf{d}, \boldsymbol{\delta}) \geq z] \geq 1 - \gamma; \\
& \quad \sum_{i \in \mathcal{S}} s(i) = \sum_{j \in \mathcal{D}} d(j) \geq L.
\end{aligned}$$

It is easy to verify that  $-h(\mathbf{s}, \mathbf{d}, \boldsymbol{\delta})$  is non-linearly concave w.r.t. the decision variables  $(\mathbf{s}, \mathbf{d})$  and non-linearly convex w.r.t.  $\boldsymbol{\delta}$ . Thus, the above transportation problem satisfies Assumption 2.1.

## 2.3 The Chance Constraint Case

This section is devoted to the (individual) distributionally robust chance constraint case (2.5). Our main theorem shows that when function  $f(\mathbf{x}, \boldsymbol{\delta})$  satisfies Assumption 2.1, then a DRCC is equivalent to a robust optimization constraint. This bridges the two main approaches in optimization under uncertainty, namely, stochastic programming, and robust optimization. We then investigate the tractability of DRCC, providing sufficient conditions for the indi-

vidual DRCC (2.5) to be tractable.

### 2.3.1 Equivalence to Robust Optimization

In this subsection we show that DRCC is equivalent to robust optimization by analyzing the feasible set given by the constraint (2.5), which we denote by

$$S \triangleq \{\mathbf{x} \mid \inf_{\boldsymbol{\delta} \sim (0, \boldsymbol{\Sigma})} \mathbb{P}[f(\mathbf{x}, \boldsymbol{\delta}) \geq \alpha] \geq p\} = \{\mathbf{x} \mid \sup_{\boldsymbol{\delta} \sim (0, \boldsymbol{\Sigma})} \mathbb{P}[f(\mathbf{x}, \boldsymbol{\delta}) < \alpha] \leq 1 - p\}.$$

Our main tool to analyze  $S$  is the following result from Marshall and Olkin [MO60].

**Lemma 2.1.** *Let  $\boldsymbol{\delta} = (\delta_1, \dots, \delta_k)$  be a random vector with  $\mathbb{E}[\boldsymbol{\delta}] = 0$ ,  $\mathbb{E}[\boldsymbol{\delta}\boldsymbol{\delta}^\top] = \boldsymbol{\Sigma}$ , and  $T \subseteq \mathbb{R}^k$  be a closed convex set. Then we have*

$$\mathbb{P}[\boldsymbol{\delta} \in T] \leq \frac{1}{1 + \theta^2},$$

where  $\theta = \inf_{\mathbf{y} \in T} \sqrt{\mathbf{y}^\top \boldsymbol{\Sigma}^{-1} \mathbf{y}}$ , and the equality can always be attained.

Notice that one technical difficulty that we face to apply Lemma 2.1 is that the set  $\{\boldsymbol{\delta} \mid f(\mathbf{x}, \boldsymbol{\delta}) < \alpha\}$  may not be closed. Hence we extend Lemma 2.1 to the case where  $T$  is not necessarily closed:

**Lemma 2.2.** *Let  $T \subseteq \mathbb{R}^k$  be a convex set. Denote  $\theta = \inf_{\mathbf{y} \in T} \sqrt{\mathbf{y}^\top \boldsymbol{\Sigma}^{-1} \mathbf{y}}$ . Then we have*

$$\sup_{\boldsymbol{\delta} \sim (0, \boldsymbol{\Sigma})} \mathbb{P}[\boldsymbol{\delta} \in T] = \frac{1}{1 + \theta^2}.$$

*Proof.* When  $T$  is empty, we have  $\sup_{\boldsymbol{\delta} \sim (0, \boldsymbol{\Sigma})} \mathbb{P}[\boldsymbol{\delta} \in T] = 0$ . On the other hand,  $\theta = \inf_{\mathbf{y} \in T} \sqrt{\mathbf{y}^\top \boldsymbol{\Sigma}^{-1} \mathbf{y}} = +\infty$  which implies  $1/(1 + \theta^2) = 0$ . Hence the lemma holds.

When  $T$  is non-empty,  $T$  has a non-empty relative interior. Let  $\mathbf{x}_0$  be a point in the relative interior of  $T$ . Let  $\bar{T}$  be the closure of  $T$ , and for  $0 \leq \lambda < 1$  define  $\underline{T}(\lambda)$  by

$$\underline{T}(\lambda) = \{\lambda(\mathbf{x} - \mathbf{x}_0) + \mathbf{x}_0 \mid \mathbf{x} \in \bar{T}\}.$$

Thus, we have  $\underline{T}(\lambda)$  is closed, convex, and  $\underline{T}(\lambda) \subseteq T$ . Define

$$\bar{\theta} = \inf_{\mathbf{y} \in \bar{T}} \sqrt{\mathbf{y}^\top \boldsymbol{\Sigma}^{-1} \mathbf{y}}, \quad \theta = \inf_{\mathbf{y} \in T} \sqrt{\mathbf{y}^\top \boldsymbol{\Sigma}^{-1} \mathbf{y}}, \quad \underline{\theta}(\lambda) = \inf_{\mathbf{y} \in \underline{T}(\lambda)} \sqrt{\mathbf{y}^\top \boldsymbol{\Sigma}^{-1} \mathbf{y}},$$

and hence  $\underline{\theta}(\lambda) \geq \theta \geq \bar{\theta}$ . On the other hand, for any  $\mathbf{x} \in \bar{T}$ , one can construct a sequence  $\mathbf{x}_i \rightarrow \mathbf{x}$  such that  $\mathbf{x}_i \in \underline{\theta}(\lambda_i)$  for some  $\{\lambda_i\}_{i=1}^\infty$ , by the definition of  $\underline{T}(\lambda)$ . Thus, since  $\mathbf{y}^\top \boldsymbol{\Sigma}^{-1} \mathbf{y}$  is a continuous function of  $\mathbf{y}$ , we have  $\inf_{\lambda \in [0,1]} \underline{\theta}(\lambda) \leq \bar{\theta}$ , which implies  $\inf_{\lambda \in [0,1]} \underline{\theta}(\lambda) = \bar{\theta}$ . By Lemma 2.1, the following inequalities hold for  $0 \leq \lambda < 1$  since  $\underline{T}(\lambda)$  and  $\bar{T}$  are both closed convex sets:

$$\frac{1}{1 + \underline{\theta}(\lambda)^2} = \sup_{\boldsymbol{\delta} \sim (0, \boldsymbol{\Sigma})} \mathbb{P}[\boldsymbol{\delta} \in \underline{T}(\lambda)] \leq \sup_{\boldsymbol{\delta} \sim (0, \boldsymbol{\Sigma})} \mathbb{P}[\boldsymbol{\delta} \in T] \leq \sup_{\boldsymbol{\delta} \sim (0, \boldsymbol{\Sigma})} \mathbb{P}[\boldsymbol{\delta} \in \bar{T}] = \frac{1}{1 + \bar{\theta}^2}.$$

Since  $\sup_{\lambda \in [0,1]} \frac{1}{1 + \underline{\theta}(\lambda)^2} = \frac{1}{1 + \bar{\theta}^2}$ , we have

$$\sup_{\boldsymbol{\delta} \sim (0, \boldsymbol{\Sigma})} \mathbb{P}[\boldsymbol{\delta} \in T] = \sup_{\boldsymbol{\delta} \sim (0, \boldsymbol{\Sigma})} \mathbb{P}[\boldsymbol{\delta} \in \bar{T}]$$

which establishes the lemma.  $\square$

Now we are ready to present the main result of this subsection.

**Theorem 2.1.** *Suppose  $f(\mathbf{x}, \cdot)$  is quasi-convex for every  $\mathbf{x} \in \mathbb{X}$  and  $f(\cdot, \mathbf{y})$  is concave for every  $\mathbf{y} \in \mathbb{Y}$ , and let  $p \in (0, 1)$  and set  $r = p/(1 - p)$ , then the feasible set  $S$  of the DRCC (2.5) is convex and admits*

$$S = \{\mathbf{x} | \forall \mathbf{y} \text{ such that } \mathbf{y}^\top \boldsymbol{\Sigma}^{-1} \mathbf{y} < r \Rightarrow f(\mathbf{x}, \mathbf{y}) \geq \alpha\}.$$

If  $f(\mathbf{x}, \cdot)$  is further assumed to be continuous for every  $\mathbf{x} \in \mathbb{X}$ , then the distributionally robust chance constraint  $\inf_{\boldsymbol{\delta} \sim (0, \boldsymbol{\Sigma})} \mathbb{P}[f(\mathbf{x}, \boldsymbol{\delta}) \geq \alpha] \geq p$  is equivalent to

$$f(\mathbf{x}, \mathbf{y}) \geq \alpha, \quad \forall \mathbf{y} \in \boldsymbol{\Omega} \triangleq \{\mathbf{y} | \mathbf{y}^\top \boldsymbol{\Sigma}^{-1} \mathbf{y} \leq r\}.$$

*Proof.* Since  $f(\mathbf{x}, \cdot)$  is quasi-convex for each  $\mathbf{x} \in \mathbb{X}$ , the set  $T_{\mathbf{x}} \triangleq \{\mathbf{y} | f(\mathbf{x}, \mathbf{y}) < \alpha\}$  is convex

for fixed  $\mathbf{x}$ . Then from Lemma 2.2, the feasible set of the constraint (2.5) satisfies

$$\begin{aligned}
S &\triangleq \{\mathbf{x} \mid \inf_{\boldsymbol{\delta} \sim (0, \boldsymbol{\Sigma})} \mathbb{P}[f(\mathbf{x}, \boldsymbol{\delta}) \geq \alpha] \geq p\} = \{\mathbf{x} \mid \sup_{\boldsymbol{\delta} \sim (0, \boldsymbol{\Sigma})} \mathbb{P}[f(\mathbf{x}, \boldsymbol{\delta}) < \alpha] \leq 1 - p\} \\
&= \{\mathbf{x} \mid \sup_{\boldsymbol{\delta} \sim (0, \boldsymbol{\Sigma})} \mathbb{P}[\boldsymbol{\delta} \in T_{\mathbf{x}}] \leq 1 - p\} \stackrel{(a)}{=} \{\mathbf{x} \mid \inf_{\mathbf{y} \in T_{\mathbf{x}}} \mathbf{y}^\top \boldsymbol{\Sigma}^{-1} \mathbf{y} \geq r\} \\
&= \{\mathbf{x} \mid \inf_{f(\mathbf{x}, \mathbf{y}) < \alpha} \mathbf{y}^\top \boldsymbol{\Sigma}^{-1} \mathbf{y} \geq r\} = \{\mathbf{x} \mid \forall \mathbf{y} \text{ such that } f(\mathbf{x}, \mathbf{y}) < \alpha \Rightarrow \mathbf{y}^\top \boldsymbol{\Sigma}^{-1} \mathbf{y} \geq r\} \\
&= \{\mathbf{x} \mid \forall \mathbf{y} \text{ such that } \mathbf{y}^\top \boldsymbol{\Sigma}^{-1} \mathbf{y} < r \Rightarrow f(\mathbf{x}, \mathbf{y}) \geq \alpha\},
\end{aligned}$$

where (a) holds by Lemma 2.2. Since  $f(\cdot, \mathbf{y})$  is concave for every  $\mathbf{y}$ , we know that  $S$  is convex, as the property is preserved under arbitrary intersection. Hence we proved the first part:  $S = \{\mathbf{x} \mid \forall \mathbf{y} \text{ such that } \mathbf{y}^\top \boldsymbol{\Sigma}^{-1} \mathbf{y} < r \Rightarrow f(\mathbf{x}, \mathbf{y}) \geq \alpha\}$ .

To show the second part, further notice that  $p \in (0, 1)$  implies  $r > 0$ . Thus we have

$$S = \{\mathbf{x} \mid f(\mathbf{x}, \mathbf{y}) \geq \alpha, \forall \mathbf{y} \text{ such that } \mathbf{y}^\top \boldsymbol{\Sigma}^{-1} \mathbf{y} \leq r\},$$

where the equality holds because for each  $\mathbf{x} \in \mathbb{X}$ ,  $f(\mathbf{x}, \mathbf{y})$  and  $\mathbf{y}^\top \boldsymbol{\Sigma}^{-1} \mathbf{y}$  are both continuous in  $\mathbf{y}$  so that we can replace “<” by “≤” without effect on  $S$ .  $\square$

Thus the probabilistic uncertainty model is linked to the deterministic set based uncertainty model of robust optimization (e.g., Ben-Tal and Nemirovski [BTN98, BTN99], Bertsimas and Sim [BS04]). This result is in the spirit of past work that has linked chance constraints to robust optimization in the linear case (e.g., Delage and Mannor [DM10], Shivaswamy *et al.* [SBS06]).

Interestingly, based on the above theorem, we can establish an equivalence relationship between the distributionally robust chance constraint and the Worst Case Conditional Value at Risk (WCCVaR) in the convex case, which recovers a result first shown in [ZKR11] using a different proof.

**Corollary 2.1.** *Suppose  $f(\mathbf{x}, \cdot)$  is convex and continuous for every  $\mathbf{x} \in \mathbb{X}$ , then for  $p \in (0, 1)$ ,*

$$\inf_{\boldsymbol{\delta} \sim (0, \boldsymbol{\Sigma})} \mathbb{P}[f(\mathbf{x}, \boldsymbol{\delta}) \geq \alpha] \geq p \Leftrightarrow \sup_{\boldsymbol{\delta} \sim (0, \boldsymbol{\Sigma})} CVaR_{1-p}(-f(\mathbf{x}, \boldsymbol{\delta})) \leq -\alpha.$$

In the most general case, i.e.,  $f(\mathbf{x}, \boldsymbol{\delta})$  is quasi-convex, the equivalence shown in Corollary 2.1 does not hold. Consider a constraint with a random variable  $\delta$ :

$$\inf_{\delta \sim (0, \sigma)} \mathbb{P}[f(\mathbf{x}, \delta) \geq \alpha] \geq 0.5.$$

We now construct a function  $f(\mathbf{x}, \delta)$  that is quasi-convex but not convex w.r.t.  $\delta$ . In particular, we construct  $f(\mathbf{x}, \delta)$  that is decreasing (hence quasi-convex) and *concave* w.r.t.  $\delta$ , such that the DRCC above holds but the constraint on the worse-case CVaR does not hold. For simplicity, denote  $-f(\mathbf{x}, \cdot)$  by  $L(\cdot)$  and let  $\alpha = -\sigma$ . Define  $L(\cdot)$  as follows:

$$L(x) = \begin{cases} \sigma, & x \leq \sqrt{\sigma}; \\ x^2, & x > \sqrt{\sigma}. \end{cases}$$

It can be easily shown that the constraint  $\inf_{\delta \sim (0, \sigma)} \mathbb{P}[L(\delta) \leq \sigma] \geq 0.5$  holds. Consider a uniform distribution over the interval  $[-\sqrt{3\sigma}, \sqrt{3\sigma}]$  which has mean 0 and variance  $\sigma$ . By simple computation, we can see that  $\text{CVaR}_{0.5}(L(\delta)) > \sigma$  w.r.t. this uniform distribution when  $\sigma = 1$ .

### 2.3.2 Tractability of Individual DRCC

In this subsection we investigate the tractability of DRCC. We first provide sufficient conditions for optimization problems involving chance constraint (2.5) with function  $f(\mathbf{x}, \boldsymbol{\delta})$  being tractable. We then show that for the special case where  $f(\mathbf{x}, \boldsymbol{\delta}) = g(\boldsymbol{\delta})^\top \mathbf{x}$  and  $g(\boldsymbol{\delta})$  is linear or convex quadratic, we can convert (2.5) to an equivalent semi-definite constraint.

**Theorem 2.2.** *If function  $f(\mathbf{x}, \boldsymbol{\delta})$  satisfies Assumption 2.1, set  $\Psi \subseteq \mathbb{X}$  is tractable and  $p \in (0, 1)$ , then the following optimization problem*

$$\begin{aligned} & \text{Minimize: }_{\mathbf{x} \in \Psi} \quad \mathbf{c}^\top \mathbf{x} \\ & \text{Subject to:} \quad \inf_{\boldsymbol{\delta} \sim (0, \boldsymbol{\Sigma})} \mathbb{P}[f(\mathbf{x}, \boldsymbol{\delta}) \geq \alpha] \geq p \end{aligned} \tag{2.9}$$

*can be solved in polynomial time, if (i) for any fixed  $\boldsymbol{\delta}$  the super-gradient of  $f(\cdot, \boldsymbol{\delta})$  can*

be evaluated in polynomial time; and (ii) for any fixed  $\mathbf{x} \in \Psi$  the following optimization problems on  $\mathbf{y}$  can be solved in polynomial time,

$$\begin{aligned} & \text{Minimize:}_{\mathbf{y}} \quad f(\mathbf{x}, \mathbf{y}) \\ & \text{Subject to:} \quad \mathbf{y}^T \Sigma^{-1} \mathbf{y} \leq \frac{p}{1-p}. \end{aligned} \tag{2.10}$$

*Proof.* By Theorem 2.1, the feasible set  $S$  of the constraint (2.5) is given by

$$S = \{\mathbf{x} \mid f(\mathbf{x}, \mathbf{y}) \geq \alpha, \forall \mathbf{y} \text{ such that } \mathbf{y}^T \Sigma^{-1} \mathbf{y} \leq \frac{p}{1-p}\}.$$

To establish the theorem, it suffices to construct a polynomial-time *separation oracle* for  $S$  (Grötschel et al. [GLS88]). A “separation oracle” is a routine such that for  $\mathbf{x}^*$ , it can be verified in polynomial time that (a) whether  $\mathbf{x}^* \in S$  or not; and (b) if  $\mathbf{x}^* \notin S$ , a hyperplane that separates  $\mathbf{x}^*$  with  $S$ .

We now construct such a separation oracle. To verify the feasibility of  $\mathbf{x}^*$ , notice that  $\mathbf{x}^* \in S$  if and only if the optimal value of the optimization problem (2.10) is greater than or equal to  $\alpha$ , which can be verified by solving Problem (2.10) directly. By assumption, this can be done in polynomial time.

If  $\mathbf{x}^* \notin S$ , then by solving Problem (2.10), we can find in polynomial time  $\mathbf{y}^*$  such that  $f(\mathbf{x}^*, \mathbf{y}^*) < \alpha$ . Because  $f(\mathbf{x}, \mathbf{y})$  is concave in  $\mathbf{x}$  for each  $\mathbf{y} \in \mathbb{Y}$ , for any  $\mathbf{x} \in S$ , the following holds

$$f(\mathbf{x}^*, \mathbf{y}^*) + \nabla_x f(\mathbf{x}^*, \mathbf{y}^*)^\top (\mathbf{x} - \mathbf{x}^*) \geq f(\mathbf{x}, \mathbf{y}^*) \geq \alpha.$$

Thus, the hyperplane separating  $\mathbf{x}^*$  from the feasible set  $S$  is the following

$$f(\mathbf{x}^*, \mathbf{y}^*) + \nabla_x f(\mathbf{x}^*, \mathbf{y}^*)^\top (\mathbf{x} - \mathbf{x}^*) \geq \alpha,$$

which can be generated in polynomial time since the super-gradient of  $\mathbf{x}$  can be obtained in polynomial time.  $\square$

We now consider the special case that  $f(\mathbf{x}, \boldsymbol{\delta}) = g(\boldsymbol{\delta})^\top \mathbf{x}$  and each component  $g_i(\boldsymbol{\delta})$  of  $g(\boldsymbol{\delta})$  is either quadratic convex or linear.

**Corollary 2.2.** *If  $f(\mathbf{x}, \boldsymbol{\delta}) = g(\boldsymbol{\delta})^\top \mathbf{x}$  and satisfies Assumption 2.1 and each component of  $g(\boldsymbol{\delta})$  is a convex quadratic or linear function, i.e., it has the form  $g_i(\boldsymbol{\delta}) = \boldsymbol{\delta}^\top \mathbf{G}_i \boldsymbol{\delta} + \mathbf{p}_i^\top \boldsymbol{\delta} + q_i$ , where  $\mathbf{p}_i \in \mathbb{R}^n$ ,  $q_i \in \mathbb{R}$  and  $\mathbf{G}_i \in \mathbb{S}^n$  is a symmetric semi-definite matrix ( $G_i$  is zero if  $g_i(\boldsymbol{\delta})$  is linear), then the following optimization problem*

$$\begin{aligned} & \text{Minimize: } \mathbf{x} \in \Psi && c(\mathbf{x}) \\ & \text{Subject to: } && \inf_{\boldsymbol{\delta} \sim (0, \boldsymbol{\Sigma})} \mathbb{P}[g(\boldsymbol{\delta})^\top \mathbf{x} \geq \alpha] \geq p \end{aligned} \quad (2.11)$$

where  $p \in (0, 1)$ , is equivalent to

$$\begin{aligned} & \text{Minimize: } \mathbf{x} \in \Psi, \beta \geq 0 && c(\mathbf{x}) \\ & \text{Subject to: } && \begin{pmatrix} \beta \boldsymbol{\Sigma}^{-1} + G(\mathbf{x}) & \frac{1}{2} P(\mathbf{x}) \\ \frac{1}{2} P(\mathbf{x})^\top & Q(\mathbf{x}) - \frac{\beta p}{1-p} \end{pmatrix} \succeq 0, \end{aligned} \quad (2.12)$$

where  $G(\mathbf{x}) \triangleq \sum_{i=1}^n x_i \mathbf{G}_i$ ,  $P(\mathbf{x}) \triangleq \sum_{i=1}^n x_i \mathbf{p}_i$ , and  $Q(\mathbf{x}) \triangleq \sum_{i=1}^n x_i q_i - \alpha$ .

Notice that  $G(\mathbf{x})$ ,  $P(\mathbf{x})$  and  $Q(\mathbf{x})$  are all linear functions of  $\mathbf{x}$ , and hence the semi-definite constraint in Problem (2.12) is a linear matrix inequality. Compare to the result by Calafiore and El Ghaoui [CG06] which only considers the case where  $f(\cdot, \cdot)$  is bilinear, the result above holds when  $f(\mathbf{x}, \cdot)$  is convex quadratic. Zymler *et al.* [ZKR11] showed that DRCC is tractable when  $f(\mathbf{x}, \boldsymbol{\delta})$  is linear in  $\mathbf{x}$  and quadratic in  $\boldsymbol{\delta}$ . However, their method is built upon S-lemma, and hence it is not clear how to extend the method to more general cases. Our formulation needs stronger conditions –  $f(\mathbf{x}, \cdot)$  is *convex* quadratic – than [ZKR11], but the equivalent formulation is simpler than [ZKR11].

## 2.4 Probabilistic Envelope Constraint

Recall that the probabilistic envelope constraint refers to the following:

$$\inf_{\boldsymbol{\delta} \sim (0, \boldsymbol{\Sigma})} \mathbb{P}[f(\mathbf{x}, \boldsymbol{\delta}) \geq \alpha - s] \geq \mathfrak{B}(s); \quad \forall s \geq 0. \quad (2.13)$$

Here,  $s$  represents allowed magnitude of constraint violation, and  $\mathfrak{B}(s)$  is the probabilistic guarantee associated with a constraint violation no more than  $s$ . Hence,  $\mathfrak{B}(s) \in (0, 1)$  for all  $s \geq 0$ , and is assumed to be non-decreasing without loss of generality.

When  $f(\mathbf{x}, \boldsymbol{\delta})$  is bilinear, the envelope constraint (2.13) is shown to be equivalent to a comprehensive robust constraint, and proved to be tractable under mild technical conditions in Xu *et al.* [XCM12]. We consider in this section the tractability of (2.13) where  $f(\mathbf{x}, \boldsymbol{\delta})$  satisfies Assumption 2.1. For convenience of exposition, we rewrite (2.13) to an equivalent formulation as shown in the following lemma.

**Lemma 2.3.** *If  $\mathfrak{B}(s) : \mathbb{R}^+ \mapsto (0, 1)$  is a non-decreasing function that is continuous from the right, then the probabilistic envelope constraint (2.13) is equivalent to*

$$\inf_{\boldsymbol{\delta} \sim (0, \boldsymbol{\Sigma})} \mathbb{P}[f(\mathbf{x}, \boldsymbol{\delta}) \geq \alpha - t(r)] \geq \frac{r}{1+r}; \quad \forall r \geq 0. \quad (2.14)$$

Here  $t(r) \triangleq \mathfrak{B}^{-1}\left(\frac{r}{1+r}\right)$  and  $\mathfrak{B}^{-1}(x)$  is defined as

$$\mathfrak{B}^{-1}(x) \triangleq \begin{cases} \inf\{y \geq 0 \mid \mathfrak{B}(y) \geq x\} & \text{if } \exists y \text{ such that } \mathfrak{B}(y) \geq x; \\ +\infty & \text{otherwise.} \end{cases}$$

Furthermore,  $t(\cdot)$  is non-decreasing,  $t(0) = 0$ ,  $\lim_{r \uparrow +\infty} t(r) = +\infty$ , and  $t(\cdot)$  is continuous at the neighborhood of 0.

Hence in the sequel, we analyze the probabilistic envelope constraint (2.14) instead of (2.13). The following theorem shows that a probabilistic envelope constraint is equivalent to a comprehensive robust constraint proposed in Ben-Tal *et al.* [BTBN06], Ben-Tal *et al.* [BTEN09] and Ben-Tal *et al.* [BTBB10]. This thus extends previous results for affine cases in Xu *et al.* [XCM12] to general  $f(\cdot, \cdot)$  satisfying Assumption 2.1.

**Theorem 2.3.** *Suppose  $t : \mathbb{R}^+ \mapsto [0, +\infty)$  is non-decreasing,  $t(0) = 0$ ,  $\lim_{r \uparrow +\infty} t(r) = +\infty$  and continuous at the neighborhood of 0. Then if function  $f(\mathbf{x}, \boldsymbol{\delta})$  satisfies Assumption 2.1, the probabilistic envelope constraint*

$$\inf_{\boldsymbol{\delta} \sim (0, \boldsymbol{\Sigma})} \mathbb{P}[f(\mathbf{x}, \boldsymbol{\delta}) \geq \alpha - t(r)] \geq \frac{r}{1+r}; \quad \forall r \geq 0 \quad (2.15)$$



is equivalent to the comprehensive robust constraint

$$f(\mathbf{x}, \mathbf{y}) \geq \alpha - t(\|\mathbf{y}\|_{\Sigma^{-1}}^2), \quad \forall \mathbf{y} \in \mathbb{R}^n. \quad (2.16)$$

*Proof.* Define the feasible set of (2.15) as  $S$ . For any fixed  $r \geq 0$ , we have

$$\begin{aligned} S(r) &= \left\{ \mathbf{x} \mid \inf_{\boldsymbol{\delta} \sim (0, \boldsymbol{\Sigma})} \mathbb{P}[f(\mathbf{x}, \boldsymbol{\delta}) \geq \alpha - t(r)] \geq \frac{r}{1+r} \right\} \\ &= \{ \mathbf{x} \mid \forall \mathbf{y} \text{ such that } \mathbf{y}^\top \boldsymbol{\Sigma}^{-1} \mathbf{y} < r \Rightarrow f(\mathbf{x}, \mathbf{y}) + t(r) \geq \alpha \}. \end{aligned}$$

by Lemma 2.3 and Theorem 2.1. Thus, we have

$$\begin{aligned} S &= \left\{ \mathbf{x} \mid \inf_{\boldsymbol{\delta} \sim (0, \boldsymbol{\Sigma})} \mathbb{P}[f(\mathbf{x}, \boldsymbol{\delta}) \geq \alpha - t(r)] \geq \frac{r}{1+r}; \quad \forall r \geq 0 \right\} \\ &= \{ \mathbf{x} \mid \forall \mathbf{y} \text{ such that } \mathbf{y}^\top \boldsymbol{\Sigma}^{-1} \mathbf{y} < r \Rightarrow f(\mathbf{x}, \mathbf{y}) + t(r) \geq \alpha; \quad \forall r \geq 0 \}. \end{aligned}$$

Notice that without loss of generality, we can neglect the case  $r = 0$  in the right hand side, as  $\{ \mathbf{y} \mid \mathbf{y}^\top \boldsymbol{\Sigma}^{-1} \mathbf{y} < 0 \} = \emptyset$ . Thus we have

$$S = \{ \mathbf{x} \mid \forall \mathbf{y} \text{ such that } \mathbf{y}^\top \boldsymbol{\Sigma}^{-1} \mathbf{y} \leq r \Rightarrow f(\mathbf{x}, \mathbf{y}) + t(r) \geq \alpha; \quad \forall r \geq 0 \},$$

where in the last equality we use the fact that  $\forall \mathbf{x} \in \mathbb{X}$ ,  $f(\mathbf{x}, \mathbf{y})$  and  $\mathbf{y}^\top \boldsymbol{\Sigma}^{-1} \mathbf{y}$  are both continuous in  $\mathbf{y}$ , we can replace “<” by “ $\leq$ ” without effect on  $S$  as long as  $\{ \mathbf{y}^\top \boldsymbol{\Sigma}^{-1} \mathbf{y} < r \}$  is non-empty. By continuity of  $t(r)$  at  $r = 0$ , we further have

$$S = \{ \mathbf{x} \mid \forall (\mathbf{y}, r) \text{ such that } \mathbf{y}^\top \boldsymbol{\Sigma}^{-1} \mathbf{y} \leq r \Rightarrow f(\mathbf{x}, \mathbf{y}) + t(r) \geq \alpha \}.$$

The second equality holds because there exists no  $\mathbf{y}$  such that  $\mathbf{y}^\top \boldsymbol{\Sigma}^{-1} \mathbf{y} \leq r$  when  $r < 0$  so that the constraint  $r \geq 0$  can be removed. Hence the probabilistic envelope constraint is equivalent to

$$f(\mathbf{x}, \mathbf{y}) + t(r) \geq \alpha, \quad \forall (\mathbf{y}, r) \text{ such that } \|\mathbf{y}\|_{\Sigma^{-1}}^2 \leq r. \quad (2.17)$$

Notice that (2.17) is equivalent to constraint (2.16) by monotonicity of  $t(\cdot)$ .  $\square$

It is known that comprehensive robust optimization generalizes robust optimization (e.g.,

Ben-Tal *et al.* [BTBN06], Ben-Tal *et al.* [BTEN09] and Ben-Tal *et al.* [BTBB10]). Indeed, if  $t(\cdot)$  is taken to be an indicator function, i.e.,  $t(r) = 0$  for  $r \in [0, c]$  and  $+\infty$  for  $r > c$ , the formulation (2.16) recovers the standard robust optimization formulation with the ellipsoidal uncertainty set  $\Omega = \{\mathbf{y} | \mathbf{y}^\top \Sigma^{-1} \mathbf{y} \leq c\}$ . On the other hand, while robust optimization guarantees that the constraint is not violated for any realization of the uncertain parameters in the set  $\Omega$ , it makes no guarantees for realizations outside that set. In contrast, the comprehensive robust optimization formulation allows us to choose different functions  $t(\cdot)$ , in order to provide different levels of protection for different parameter realizations, as opposed to the “all-or-nothing” view of standard robust optimization.

We now investigate the tractability of probabilistic envelope chance constraints. We first consider the general case where  $f(\mathbf{x}, \boldsymbol{\delta})$  is an arbitrary “concave-quasiconvex” function. The following theorem is essentially an envelope constraint counterpart of Theorem 2.2.

**Theorem 2.4.** *If  $t(\cdot)$  satisfies the conditions in Theorem 2.3,  $f(\mathbf{x}, \boldsymbol{\delta})$  satisfies Assumption 2.1 and set  $\Psi \subseteq \mathbb{X}$  is tractable, then the optimization problem with a linear objective function and the probabilistic envelope constraint (2.13):*

$$\begin{aligned} & \text{Minimize}_{\mathbf{x} \in \Psi} \quad \mathbf{c}^\top \mathbf{x} \\ & \text{Subject to:} \quad \inf_{\boldsymbol{\delta} \sim (0, \Sigma)} \mathbb{P}[f(\mathbf{x}, \boldsymbol{\delta}) \geq \alpha - t(r)] \geq \frac{r}{1+r}; \quad \forall r \geq 0 \end{aligned} \tag{2.18}$$

can be solved in polynomial time if (1) one can provide the super-gradient of  $f(\mathbf{x}, \boldsymbol{\delta})$  at  $\mathbf{x}$  for fixed  $\boldsymbol{\delta}$  in polynomial time, and (2) for any fixed  $\mathbf{x}$  the following optimization problems can be solved in polynomial time:

$$\begin{aligned} & \text{Minimize}_{\mathbf{y}, r} \quad f(\mathbf{x}, \mathbf{y}) + t(r) \\ & \text{Subject to:} \quad \mathbf{y}^\top \Sigma^{-1} \mathbf{y} \leq r. \end{aligned} \tag{2.19}$$

*Proof.* By Theorem 2.3, the feasible set  $S$  can be rewritten as

$$S = \{\mathbf{x} | \forall (\mathbf{y}, r) \text{ such that } \mathbf{y}^\top \Sigma^{-1} \mathbf{y} \leq r \Rightarrow f(\mathbf{x}, \mathbf{y}) + t(r) \geq \alpha\}.$$

Similar to the proof of Theorem 2.2, we construct a separation oracle to prove tractability.

In order to verify the feasibility of a given  $\mathbf{x}^*$ , notice that  $\mathbf{x}^* \in S$  if and only if the optimal objective value of the optimization problem (2.19) is greater than or equal to  $\alpha$ , which can be verified by directly solving Problem (2.19). By assumption, this can be done in polynomial time.

If  $\mathbf{x}^* \notin S$ , then by solving Problem (2.19), we can find in polynomial time  $(\mathbf{y}^*, r^*)$  such that  $f(\mathbf{x}^*, \mathbf{y}^*) + t(r^*) < \alpha$ . Because  $f(\mathbf{x}, \mathbf{y})$  is concave in  $\mathbf{x}$  for each  $\mathbf{y} \in \mathbb{Y}$ , for any  $\mathbf{x} \in S$ , we have

$$f(\mathbf{x}^*, \mathbf{y}^*) + \nabla f(\mathbf{x}^*, \mathbf{y}^*)^\top (\mathbf{x} - \mathbf{x}^*) + t(r^*) \geq f(\mathbf{x}, \mathbf{y}^*) + t(r^*) \geq \alpha.$$

Hence the hyperplane separating  $\mathbf{x}^*$  from the feasible set  $S$  is the following:

$$f(\mathbf{x}^*, \mathbf{y}^*) + \nabla f(\mathbf{x}^*, \mathbf{y}^*)^\top (\mathbf{x} - \mathbf{x}^*) + t(r^*) \geq \alpha, \quad (2.20)$$

which can be generated in polynomial time since the super-gradient of  $\mathbf{x}$  can be obtained in polynomial time. This completes the proof.  $\square$

Our next result states that when  $f(\mathbf{x}, \boldsymbol{\delta}) = g(\boldsymbol{\delta})^\top \mathbf{x}$  and  $g_i(\boldsymbol{\delta})$  is quadratic, (2.14) can be converted to a semi-definite constraint.

**Corollary 2.3.** *Suppose  $t(\cdot)$  satisfies the conditions in Theorem 2.3 and is convex,  $f(\mathbf{x}, \boldsymbol{\delta}) = g(\boldsymbol{\delta})^\top \mathbf{x}$  satisfies Assumption 2.1 and  $\Psi \subseteq \mathbb{X}$  is tractable, then if each component  $g_i(\boldsymbol{\delta})$  of  $g(\boldsymbol{\delta})$  is linear or convex quadratic as in Corollary 2.2, the optimization problem (2.18) is equivalent to*

$$\begin{aligned} & \text{Minimize: }_{\mathbf{x} \in \Psi, \beta \geq 0} \mathbf{c}^\top \mathbf{x} \\ & \text{Subject to: } \begin{pmatrix} \beta \boldsymbol{\Sigma}^{-1} + G(\mathbf{x}) & \frac{1}{2}P(\mathbf{x}) \\ \frac{1}{2}P(\mathbf{x})^\top & Q(\mathbf{x}) - t^*(\beta) \end{pmatrix} \succeq 0 \end{aligned} \quad (2.21)$$

where  $t^*(\beta)$  is the conjugate function of  $t(r)$ , i.e.,  $t^*(\beta) \triangleq \sup_{r \geq 0} (\beta r - t(r))$ ; and  $P(\cdot)$ ,  $G(\cdot)$ ,  $Q(\cdot)$  are defined as in Corollary 2.2. Furthermore, the optimization problem (2.18) with a linear objective function and the probabilistic envelope constraint can be solved in polynomial time if for any  $\beta \geq 0$  the following optimization problem on  $r$  can be solved in

*polynomial time:*

$$\text{Minimize}_{r \geq 0} t(r) - \beta r. \quad (2.22)$$

In particular, when  $t(r)$  is a convex function, the optimization problems (2.19) and (2.22) are both convex and can be solved efficiently.

## 2.5 Chance Constraints: Beyond Mean and Variance

Thus far we have studied the setup that models unknown parameters as following an ambiguous distribution with known mean and covariance. In this section we extend our results to some other models of uncertain parameters – this includes the case where the mean and the covariance themselves are unknown and can only be estimated from data; and the case where other information of the uncertain parameter (e.g., the support) may be available. Specifically, we first show that the chance constraint (2.5) and the probabilistic envelope constraint (2.6) with *uncertain* mean and covariance are still tractable. Then we deal with the case where the mean and support of the uncertain parameter are known. Finally, we apply *distributionally robust optimization* to make a conservative approximation for constraints (2.5) and (2.6) when additional information on the uncertain parameter is available.

### 2.5.1 Uncertain Mean and Covariance

We first study the uncertain mean and covariance case. This model of ambiguity was first proposed and studied in [DY10] for *distributionally robust optimization*, and was also investigated for linear chance constraints in [XCM12]. We formulate the robust counterparts of the distributionally robust chance constraint (2.5) and the probabilistic envelope constraint (2.6) where the mean and covariance themselves are uncertain, and then show that optimization problems with these constraints are tractable under mild conditions. Based on Theorem 2.1 and Theorem 2.3, we can easily obtain the following corollaries. Corollary 2.4 and Corollary 2.5 show that the DRCC and the probabilistic envelope constraint with unknown mean and covariance is equivalent to a set of (infinitely many) *deterministic* constraints. Note that the uncertainty sets  $\mathcal{U}$  and  $\mathcal{S}$  can be arbitrary. Corollary 2.6 shows the

tractability of probabilistic envelope constraints.

**Corollary 2.4.** *If function  $f(\mathbf{x}, \boldsymbol{\delta})$  satisfies Assumption 2.1, then for  $p \in (0, 1)$  the chance constraint*

$$\inf_{\boldsymbol{\delta} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma}), \boldsymbol{\mu} \in \mathcal{U}, \boldsymbol{\Sigma} \in \mathcal{S}} \mathbb{P}[f(\mathbf{x}, \boldsymbol{\delta}) \geq \alpha] \geq p, \quad (2.23)$$

is equivalent to the constraint  $f(\mathbf{x}, \mathbf{y} + \boldsymbol{\mu}) \geq \alpha$ ,  $\forall \mathbf{y} \in \mathbb{R}^n, \boldsymbol{\mu} \in \mathcal{U}$  and  $\boldsymbol{\Sigma} \in \mathcal{S}$  such that  $\begin{pmatrix} \boldsymbol{\Sigma} & \mathbf{y} \\ \mathbf{y}^\top & \frac{p}{1-p} \end{pmatrix} \succeq 0$ , where  $\mathcal{U}$  and  $\mathcal{S}$  are the uncertainty sets of mean  $\boldsymbol{\mu}$  and covariance  $\boldsymbol{\Sigma}$ , respectively.

**Corollary 2.5.** *Suppose  $t : \mathbb{R}^+ \mapsto [0, +\infty)$  is non-decreasing,  $t(0) = 0$ ,  $\lim_{r \uparrow +\infty} t(r) = +\infty$  and is continuous at the neighborhood of zero. Then if function  $f(\mathbf{x}, \boldsymbol{\delta})$  satisfies Assumption 2.1, the probabilistic envelope constraint*

$$\inf_{\boldsymbol{\delta} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma}), \boldsymbol{\mu} \in \mathcal{U}, \boldsymbol{\Sigma} \in \mathcal{S}} \mathbb{P}[f(\mathbf{x}, \boldsymbol{\delta}) \geq \alpha - t(r)] \geq \frac{r}{1+r}; \quad \forall r \geq 0, \quad (2.24)$$

is equivalent to the constraint

$$\inf_{\boldsymbol{\mu} \in \mathcal{U}} f(\mathbf{x}, \mathbf{y} + \boldsymbol{\mu}) \geq \alpha - t\left(\inf_{\boldsymbol{\Sigma} \in \mathcal{S}} \|\mathbf{y}\|_{\boldsymbol{\Sigma}^{-1}}^2\right), \quad \forall \mathbf{y} \in \mathbb{R}^n, \quad (2.25)$$

where  $\mathcal{U}$  and  $\mathcal{S}$  are the uncertainty sets of mean  $\boldsymbol{\mu}$  and covariance  $\boldsymbol{\Sigma}$ , respectively.

**Corollary 2.6.** *Under the conditions of Corollary 2.5, an optimization problem with a linear objective function and the probabilistic envelope constraint (2.24) can be solved in polynomial time if one can provide the super-gradient of  $f(\mathbf{x}, \boldsymbol{\delta})$  at  $\mathbf{x}$  for fixed  $\boldsymbol{\delta}$  in polynomial time, and for any fixed  $\mathbf{x}$  the following optimization problem can be solved in polynomial time:*

$$\begin{aligned} & \text{Minimize: } f(\mathbf{x}, \mathbf{y} + \boldsymbol{\mu}) + t(r) \\ & \text{Subject to: } \begin{pmatrix} \boldsymbol{\Sigma} & \mathbf{y} \\ \mathbf{y}^\top & r \end{pmatrix} \succeq 0 \\ & \boldsymbol{\Sigma} \in \mathcal{S}, \boldsymbol{\mu} \in \mathcal{U}. \end{aligned} \quad (2.26)$$

From Corollary 2.6 we see that if  $t(\cdot)$  is convex, and  $\mathcal{U} \subseteq \mathbb{R}^n$  and  $\mathcal{S} \in \mathbb{S}_+^{n \times n}$  are both convex sets, then the optimization problem (2.26) is a SDP problem which can be solved efficiently.

The tractability result of the chance constraint (2.23) is a special case of Corollary 2.6, namely,  $t(r) = 0$  and  $r = \frac{p}{1-p}$ .

### 2.5.2 Known Mean and Support

We now investigate the case where the mean and the support of the uncertain parameter  $\delta$  are known. We show that the corresponding robust chance constraint can be reformulated as a set of infinitely many deterministic constraints, and is tractable under mild technical conditions. Unfortunately, it seems that these results can not be easily extended to the probabilistic envelope constraint case, which is hence left for future research.

**Theorem 2.5.** *Suppose the mean  $\mu$  and support  $\mathcal{S}$  of the uncertain parameter  $\delta$  are known and  $\mathcal{S}$  is a closed convex set. If  $f(\mathbf{x}, \cdot)$  is continuous and quasi-convex for every  $\mathbf{x} \in \mathbb{X}$ , then for  $p \in (0, 1]$ , the chance constraint*

$$\inf_{\delta \sim (\mu, \mathcal{S})} \mathbb{P}[f(\mathbf{x}, \delta) \geq \alpha] \geq p, \quad (2.27)$$

is equivalent to

$$f(\mathbf{x}, \delta_1) \geq \alpha, \quad \forall \delta_1, \delta_2 \text{ such that } (1-p)\delta_1 + p\delta_2 - \mu = 0, \delta_1 \in \mathcal{S}, \delta_2 \in \mathcal{S}. \quad (2.28)$$

**Theorem 2.6.** *If  $f(\mathbf{x}, \cdot)$  is quasi-convex and continuous for every  $\mathbf{x} \in \mathbb{X}$  and  $f(\cdot, \mathbf{y})$  is concave for every  $\mathbf{y} \in \mathbb{Y}$ , the mean  $\mu$  and support  $\mathcal{S}$  of the uncertain parameter  $\delta$  are known, and  $\mathcal{S}$  is a closed convex set, then for  $0 < p \leq 1$ , the optimization problem with a linear objective function and a chance constraint (2.27):*

$$\begin{aligned} \text{Minimize: } & \mathbf{c}^\top \mathbf{x} \\ \text{Subject to: } & \inf_{\delta \sim (\mu, \mathcal{S})} \mathbb{P}[f(\mathbf{x}, \delta) \geq \alpha] \geq p \end{aligned} \quad (2.29)$$

can be solved in polynomial time if (1) one can provide the super-gradient of  $f(\mathbf{x}, \mathbf{y})$  at  $\mathbf{x}$  for fixed  $\mathbf{y}$  in polynomial time, and (2) for any fixed  $\mathbf{x}$  the following optimization problems

can be solved in polynomial time:

$$\begin{aligned}
 & \text{Minimize}_{\delta_1, \delta_2} && f(\mathbf{x}, \delta_1) \\
 & \text{Subject to:} && (1 - p)\delta_1 + p\delta_2 - \boldsymbol{\mu} = 0 \\
 & && \delta_1, \delta_2 \in \mathcal{S}.
 \end{aligned} \tag{2.30}$$

*Proof.* From Theorem 2.5 we know that the chance constraint is satisfied if and only if the optimal value of (2.30) is greater than or equal to  $\alpha$ . Thus, the theorem can be proved following a similar argument as the proof of Corollary 2.6.  $\square$

### 2.5.3 Conservative Approximation

For general sets of ambiguous distributions, optimization problems involving chance constraints are notoriously hard to solve. Recall that CVaR provides a conservative approximation of chance constraints (Nemirovski *et al.* [NS06]), which allows us to apply DRO to approximately solve such problems. For completeness, we give the following lemma which is an extension of Nemirovski *et al.* [NS06]:

**Lemma 2.4.** *Suppose that  $\mathcal{D}$  is the ambiguity set of distributions of the uncertain parameter  $\boldsymbol{\delta}$ , then the chance constraint*

$$\sup_{\mathbb{P} \in \mathcal{D}} \mathbb{P}[f(\mathbf{x}, \boldsymbol{\delta}) \geq 0] \leq p, \tag{2.31}$$

can be conservatively approximated by

$$-tp + \gamma \leq 0, \quad \sup_{\mathbb{P} \in \mathcal{D}} \mathbb{E}_{\mathbb{P}}[[f(\mathbf{x}, \boldsymbol{\delta}) + t]_+] \leq \gamma, \tag{2.32}$$

where  $0 \leq p \leq 1$ ,  $t \in \mathbb{R}$  and  $\gamma \in \mathbb{R}$  are decision variables, and  $[x]_+ = \max\{x, 0\}$ . Here, by “conservative approximation” we mean that any solution that satisfies (2.32) also satisfies (2.31).

Wiesemann *et al.* [WKS13] proposed a unified framework for modeling and solving distribu-

tionally robust optimization problems by introducing standardized ambiguity sets

$$\mathcal{D} = \left\{ \mathbb{P} \in \mathcal{P}_0(\mathbb{R}^m, \mathbb{R}^n) : \begin{array}{l} \mathbb{E}_{\mathbb{P}}[\mathbf{A}\boldsymbol{\delta} + \mathbf{B}\boldsymbol{\mu}] = \mathbf{b} \\ \mathbb{P}[(\boldsymbol{\delta}, \boldsymbol{\mu}) \in \mathcal{C}_i] \in [\underline{p}_i, \overline{p}_i], \forall i \in \mathcal{I} \end{array} \right\}, \quad (2.33)$$

where  $\mathbb{P}$  represents a joint probability distribution of the random vector  $\boldsymbol{\delta} \in \mathbb{R}^m$  appearing in the constraint function  $f(\mathbf{x}, \boldsymbol{\delta})$  and some auxiliary random vector  $\boldsymbol{\mu} \in \mathbb{R}^n$ , with  $\mathbf{A} \in \mathbb{R}^{k \times m}$ ,  $\mathbf{B} \in \mathbb{R}^{k \times n}$ ,  $\mathbf{b} \in \mathbb{R}^k$ ,  $\mathcal{I} = \{1, \dots, I\}$ ,  $\underline{p}_i, \overline{p}_i \in [0, 1]$  and  $\mathcal{C}_i$  are the confidence sets.

Applying Theorem 1 and 5 in [WKS13], the constraint  $\sup_{\mathbb{P} \in \mathcal{D}} \mathbb{E}_{\mathbb{P}}[[f(\mathbf{x}, \boldsymbol{\delta}) + t]_+] \leq \gamma$  can be reformulated as a semi-infinite constraint system. For succinctness, we only present the conservative approximation of the chance constraints when  $\underline{p}_i = \overline{p}_i = 1$  and  $|\mathcal{I}| = 1$  to illustrate our approach.

**Theorem 2.7.** *If the ambiguity set  $\mathcal{P}$  can be converted into*

$$\mathcal{D} = \left\{ \mathbb{P} \in \mathcal{P}_0(\mathbb{R}^m, \mathbb{R}^n) : \mathbb{E}_{\mathbb{P}}[\mathbf{A}\boldsymbol{\delta} + \mathbf{B}\boldsymbol{\mu}] = \mathbf{b}, \mathbb{P}[(\boldsymbol{\delta}, \boldsymbol{\mu}) \in \mathcal{C}] = 1 \right\}$$

by the lifting theorem (Theorem 5 in [WKS13]) where  $g(\cdot)$  is a convex function, then the chance constraint  $\inf_{\mathbb{P} \in \mathcal{P}} \mathbb{P}[f(\mathbf{x}, \boldsymbol{\delta}) \geq \alpha] \geq p$  with  $p \in (0, 1)$  can be conservatively approximated by

$$(\mathbf{A}\boldsymbol{\delta} + \mathbf{B}\boldsymbol{\mu})^\top \boldsymbol{\beta} \geq \max \left[ -\lambda, \alpha - f(\mathbf{x}, \boldsymbol{\delta}) + \frac{\mathbf{b}^\top \boldsymbol{\beta}}{1-p} + \frac{p\lambda}{1-p} \right], \forall (\boldsymbol{\delta}, \boldsymbol{\mu}) \in \mathcal{C} \quad (2.34)$$

where  $\boldsymbol{\beta}, \lambda, \mathbf{x}$  are decision variables. Furthermore, the optimization problem with a linear objective function and the constraint (2.34) can be solved in polynomial time if (1) one can provide the super-gradient of  $f(\mathbf{x}, \mathbf{y})$  at  $\mathbf{x}$  for fixed  $\mathbf{y}$  in polynomial time, and (2) for any fixed  $(\mathbf{x}, \boldsymbol{\beta}, \lambda)$  the following optimization problems

$$\begin{aligned} & \text{Minimize}_{\boldsymbol{\delta}, \boldsymbol{\mu}} \quad (\mathbf{A}\boldsymbol{\delta} + \mathbf{B}\boldsymbol{\mu})^\top \boldsymbol{\beta} + \lambda \\ & \text{Subject to:} \quad (\boldsymbol{\delta}, \boldsymbol{\mu}) \in \mathcal{C}, \end{aligned} \quad (2.35)$$



and

$$\begin{aligned} \text{Minimize:}_{\delta, \mu} \quad & (\mathbf{A}\delta + \mathbf{B}\mu - \frac{\mathbf{b}}{1-p})^\top \boldsymbol{\beta} - \frac{p\lambda}{1-p} + f(\mathbf{x}, \delta) \\ \text{Subject to:} \quad & (\delta, \mu) \in \mathcal{C}, \end{aligned} \tag{2.36}$$

can be solved in polynomial time.

*Proof.* From Theorem 1 in [WKS13] and Lemma 2.4, the conservative approximation formulation can be easily obtained. The proof of the tractability result is similar to that of Corollary 2.6, and hence omitted.  $\square$

We now extend this result to the probabilistic envelope constraint case.

**Theorem 2.8.** *Suppose  $t : \mathbb{R}^+ \mapsto [0, +\infty)$  is convex, non-decreasing and continuous at the neighborhood of zero, and  $t(0) = 0$ ,  $\lim_{r \uparrow +\infty} t(r) = +\infty$ . If the ambiguity set  $\mathcal{D}$  satisfies the condition in Theorem 2.7, the probabilistic envelope constraint  $\inf_{\mathbb{P} \in \mathcal{D}} \mathbb{P}[f(\mathbf{x}, \boldsymbol{\delta}) \geq \alpha - t(r)] \geq r/(1+r)$ ,  $\forall r \geq 0$  can be conservatively approximated by*

$$\begin{aligned} (\mathbf{A}\delta + \mathbf{B}\mu)^\top \boldsymbol{\beta} \geq \\ \max \left[ -\lambda, \alpha - f(\mathbf{x}, \boldsymbol{\delta}) - t(r) + (1+r)\mathbf{b}^\top \boldsymbol{\beta} + r\lambda \right], \quad \forall (\delta, \mu) \in \mathcal{C}, r \geq 0. \end{aligned} \tag{2.37}$$

Furthermore, the optimization problem with a linear objective function and this probabilistic envelope constraint can be solved in polynomial time if one can provide the super-gradient of  $f(\mathbf{x}, \mathbf{y})$  at  $\mathbf{x}$  for fixed  $\mathbf{y}$  in polynomial time, and for any fixed  $(\mathbf{x}, \boldsymbol{\beta}, \lambda)$  the following optimization problems:

$$\begin{aligned} \text{Minimize:}_{\delta, \mu, r} \quad & (\mathbf{A}\delta + \mathbf{B}\mu)^\top \boldsymbol{\beta} + \lambda \\ \text{Subject to:} \quad & (\delta, \mu) \in \mathcal{C}, r \geq 0, \end{aligned} \tag{2.38}$$

and

$$\begin{aligned} \text{Minimize:}_{\delta, \mu, r} \quad & [\mathbf{A}\delta + \mathbf{B}\mu - (1+r)\mathbf{b}]^\top \boldsymbol{\beta} + f(\mathbf{x}, \boldsymbol{\delta}) + t(r) - r\lambda \\ \text{Subject to:} \quad & (\delta, \mu) \in \mathcal{C}, r \geq 0. \end{aligned} \tag{2.39}$$

can be solved in polynomial time.

*Proof.* From Theorem 2.7, the probabilistic envelope constraint can be conservatively ap-

proximated by

$$0 \leq \min_{r \geq 0} \max_{\beta, \lambda} \min_{(\delta, \mu) \in \mathcal{C}} (\mathbf{A}\delta + \mathbf{B}\mu)^\top \beta - \max \left[ -\lambda, \alpha - f(\mathbf{x}, \delta) - t(r) + (1+r)\mathbf{b}^\top \beta + r\lambda \right], \quad (2.40)$$

Furthermore, by switching “min” and “max”, this can be conservatively approximated by (2.37). Then following a similar proof as that of Corollary 2.6, we obtain the tractability result to complete the proof.  $\square$

## 2.6 Joint Chance Constraint

In this section we investigate the case of joint probabilistic envelope constraint (2.7) which can be reformulated as (from Lemma 2.3)

$$\inf_{\delta \sim (0, \Sigma)} \mathbb{P}[f_i(\mathbf{x}, \delta) \geq \alpha_i - t(r), \forall i = 1, \dots, m] \geq \frac{r}{1+r}; \forall r \geq 0, \quad (2.41)$$

where  $t(r) = \mathfrak{B}^{-1}(r/(1+r))$ . The optimization problem with the constraint (2.41) is usually intractable (e.g., Nemirovski and Shapiro [NS06]; Zymler *et al.* [ZKR11]), even when  $f(\mathbf{x}, \delta)$  is a bi-linear function, and approximation schemes are often used to tackle them. The most straightforward method to approximate the constraints (2.41) is to decompose them into several individual probabilistic envelope constraints using Bonferroni’s inequality (see below for details). A notable advantage of the Bonferroni approximation is that it is easy to implement and requires no assumptions on the function  $f_i(\mathbf{x}, \delta)$ .

However, the Bonferroni approximation can be overly conservative. Zymler *et al.* [ZKR11] proposed a tighter approximation method called *worst-case CVaR approximation* that outperforms other methods including the Bonferroni approximation (e.g. Nemirovski and Shapiro [NS06] and Chen *et al.* [CSST10]). In the rest of the section, we extend both the Bonferroni approximation and worst-case CVaR methods to JPEC. We also investigate the tractability of the two approximation schemes for  $f_i(\mathbf{x}, \delta)$  satisfying Assumption 2.1.

### 2.6.1 The Bonferroni Approximation

The Bonferroni approximation for the joint probabilistic envelope constraint (2.41) can be easily derived from Bonferroni's inequality. From Theorem 2.2 and Theorem 2.4, we know that the optimization problem with a set of probabilistic envelope constraints generated by the Bonferroni approximation method is tractable, under mild technical conditions. More specifically we have the following theorem:

**Theorem 2.9.** *Let  $t : \mathbb{R}^+ \mapsto [0, +\infty)$  be a non-decreasing function such that  $t(0) = 0$  and  $\lim_{r \uparrow +\infty} t(r) = +\infty$ , and  $\boldsymbol{\epsilon}$  be a constant vector such that  $\sum_{i=1}^m \epsilon_i = 1$  and  $\boldsymbol{\epsilon} \geq 0$ . The Bonferroni approximation of the joint probabilistic envelope constraint (2.41) which has the form*

$$\inf_{\boldsymbol{\delta} \sim (0, \boldsymbol{\Sigma})} \mathbb{P}[f_i(\mathbf{x}, \boldsymbol{\delta}) \geq \alpha_i - t(r)] \geq 1 - \frac{\epsilon_i}{(1+r)}; \quad \forall r \geq 0, \quad \forall i = 1, \dots, m. \quad (2.42)$$

is tractable if for each  $i$ , (1) one can provide the super-gradient of  $f_i(\mathbf{x}, \boldsymbol{\delta})$  at  $\mathbf{x}$  for fixed  $\boldsymbol{\delta}$  in polynomial time, and (2) for any fixed  $\mathbf{x}$  the following optimization problem can be solved in polynomial time:

$$\begin{aligned} & \text{Minimize}_{\mathbf{y}, r} && f_i(\mathbf{x}, \mathbf{y}) + t(r) \\ & \text{Subject to:} && \mathbf{y}^\top \boldsymbol{\Sigma}^{-1} \mathbf{y} \leq \frac{r+1}{\epsilon_i} - 1. \end{aligned} \quad (2.43)$$

*Proof.* Let  $r' = (1+r)/\epsilon_i - 1$ , then we have  $r'/(1+r') = 1 - \epsilon_i/(1+r)$ . Let  $t'(r') \triangleq t(r)$ , then we apply Theorem 2.4 to complete the proof.  $\square$

### 2.6.2 The Worst-case CVaR Approximation

Zymler *et al.* [ZKR11] developed a new approximation scheme for robust joint chance constraints termed *Worst-case CVaR* approximation. In this subsection we extend the worst-case CVaR approximation to JPEC (2.41). In contrast to the rest of the chapter, we focus on the linear-quadratic uncertainty case, namely,  $f(\mathbf{x}, \boldsymbol{\delta})$  is linear in  $\mathbf{x}$  for any fixed  $\boldsymbol{\delta}$  and quadratic (possibly non-convex) in  $\boldsymbol{\delta}$  for each  $\mathbf{x} \in \mathbb{X}$ . Then (2.41) can be rewritten respec-

tively as:

$$\inf_{\boldsymbol{\delta} \sim (0, \boldsymbol{\Sigma})} \mathbb{P}[\boldsymbol{\delta}^\top Q_i(\mathbf{x})\boldsymbol{\delta} + y_i(\mathbf{x})^\top \boldsymbol{\delta} + y_i^0(\mathbf{x}) + t(r) \leq 0, \forall i = 1, \dots, m] \geq \frac{r}{1+r}; \forall r \geq 0; \quad (2.44)$$

where  $Q_i(\mathbf{x})$ ,  $y_i^0(\mathbf{x})$  and  $y_i(\mathbf{x})$  are all linear functions for  $i = 1, \dots, m$ . Zymler *et al.* [ZKR11] provided the Worst-case CVaR approximation for the following robust joint chance constraint

$$\inf_{\boldsymbol{\delta} \sim (0, \boldsymbol{\Sigma})} \mathbb{P}[\boldsymbol{\delta}^\top Q_i(\mathbf{x})\boldsymbol{\delta} + y_i(\mathbf{x})^\top \boldsymbol{\delta} + y_i^0(\mathbf{x}) \leq 0, \forall i = 1, \dots, m] \geq p. \quad (2.45)$$

**Theorem 2.10.** [ZKR11] Let  $\mathcal{A} \triangleq \{\boldsymbol{\alpha} \in \mathbb{R}^m | \alpha_i > 0\}$ . For any fixed  $\mathbf{x}$  and  $\boldsymbol{\alpha} \in \mathcal{A}$ , the feasible set of the worst-case CVaR approximation for the constraint (2.45) is

$$Z^{JCC}(\boldsymbol{\alpha}) = \left\{ \mathbf{x} \in \mathbb{R}^n : \begin{array}{l} \exists (\beta, \mathbf{M}) \in \mathbb{R} \times \mathbb{S}^{k+1}, \\ \beta + \frac{1}{1-p} \langle \boldsymbol{\Omega}, \mathbf{M} \rangle \leq 0, \mathbf{M} \succeq 0 \\ \mathbf{M} - \begin{pmatrix} \alpha_i Q_i(\mathbf{x}) & \frac{1}{2} \alpha_i y_i(\mathbf{x}) \\ \frac{1}{2} \alpha_i y_i(\mathbf{x})^\top & \alpha_i y_i^0(\mathbf{x}) - \beta \end{pmatrix} \succeq 0 \\ \forall i = 1, \dots, m \end{array} \right\}, \quad (2.46)$$

where  $\boldsymbol{\Omega} = \text{diag}(\boldsymbol{\Sigma}, 1)$ .

Indeed, Zymler *et al.* [ZKR11] showed that the approximation quality of the worst-case CVaR is controlled by the parameter  $\boldsymbol{\alpha}$  and that the approximation becomes exact if  $\boldsymbol{\alpha}$  is chosen optimally. Notice that  $Z^{JCC}(\boldsymbol{\alpha})$  contains semi-definite constraints, and hence provides a tractable approximation to robust joint chance constraint. We now extend this methodology to the joint probabilistic envelope constraints (2.44). From Theorem 2.10, the

feasible set of the constraint (2.44) can be approximated as

$$Z^P(\boldsymbol{\alpha}) = \left\{ \mathbf{x} \in \mathbb{R}^n : \begin{array}{l} \text{For any } r \geq 0 \text{ we have} \\ \exists(\beta, \mathbf{M}) \in \mathbb{R} \times \mathbb{S}^{k+1}, \\ \beta + (r+1)\langle \boldsymbol{\Omega}, \mathbf{M} \rangle \leq 0, \mathbf{M} \succeq 0 \\ \mathbf{M} - \begin{pmatrix} \alpha_i Q_i(\mathbf{x}) & \frac{1}{2}\alpha_i y_i(\mathbf{x}) \\ \frac{1}{2}\alpha_i y_i(\mathbf{x})^\top & \alpha_i(y_i^0(\mathbf{x}) - t(r)) - \beta \end{pmatrix} \succeq 0 \\ \forall i = 1, \dots, m \end{array} \right\}. \quad (2.47)$$

Notice that in contrast to (2.46), (2.47) is defined through *uncountably many* sets of constraints, and hence we need the following theorem to establish the tractability of the set  $Z^P$ .

**Theorem 2.11.** *Fix  $\boldsymbol{\alpha} \in \mathcal{A}$ . The optimization problem with a linear objective function and the feasible set  $Z^P(\boldsymbol{\alpha})$  in (2.47) can be solved in polynomial time if for any fixed  $\mathbf{x}$ , the following optimization problem can be solved in polynomial time:*

$$\begin{aligned} \min_{\mathbf{Y}_i \succeq 0, r \geq 0} & -\operatorname{tr}\left(\sum_{i=1}^m \alpha_i \mathbf{Y}_i \mathbf{B}_i\right) + t(r) \sum_{i=1}^m \alpha_i \operatorname{tr}(\mathbf{Y}_i \mathbf{E}) \\ \text{s.t.} & \sum_{i=0}^m \mathbf{Y}_i = (r+1)\boldsymbol{\Omega}, \operatorname{tr}(\mathbf{E} \sum_{i=1}^m \mathbf{Y}_i) = 1, \end{aligned} \quad (2.48)$$

where  $\mathbf{B}_i = \begin{pmatrix} Q_i(\mathbf{x}) & \frac{1}{2}y_i(\mathbf{x}) \\ \frac{1}{2}y_i(\mathbf{x})^\top & y_i^0(\mathbf{x}) \end{pmatrix}$  and  $\mathbf{E} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$ .

Interestingly, Theorem 2.11 provides a tractability result for *individual* probabilistic envelope constraint.

**Corollary 2.7.** *If each component  $f_i(\cdot)$  of  $f(\cdot)$  is quadratic (and possibly non-convex), the optimization problem with a linear objective function and the probabilistic envelope constraint (2.6) can be solved in polynomial time if for any fixed  $\mathbf{x}$ , the following optimization problem*

can be solved in polynomial time:

$$\begin{aligned} \min_{\mathbf{Y} \succeq 0, r \geq 0} \quad & -\operatorname{tr}(\mathbf{Y}\mathbf{B}) + t(r) \\ \text{s.t.} \quad & \mathbf{Y} = (r+1)\mathbf{\Omega}, \operatorname{tr}(\mathbf{E}\mathbf{Y}) = 1, \end{aligned} \tag{2.49}$$

where  $\mathbf{B} = \begin{pmatrix} Q(\mathbf{x}) & \frac{1}{2}y(\mathbf{x}) \\ \frac{1}{2}y(\mathbf{x})^\top & y^0(\mathbf{x}) \end{pmatrix}$  and  $\mathbf{E} = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$ .

*Proof.* When  $m = 1$ ,  $\alpha$  can be chosen as  $\alpha = 1$  without effect on the optimal solution of (2.48). Then (2.48) can be simplified as (2.49).  $\square$

Notice that Corollary 2.7 does not require that  $f_i(\cdot)$  is a convex quadratic function, and hence, subject to the price of a more complex formulation, is more general than Corollary 2.3 that investigates the probabilistic envelope constraint under convex quadratic uncertainty.

## 2.7 Simulation

In this section we illustrate two proposed approaches – chance constraint (2.5) and probabilistic envelope constraint (2.6) using the synthetic transportation problem discussed in Section 2.2.

We consider the transportation problem where the graph  $\mathcal{G}$  is a bi-parti graph between sources and destinations, i.e.,  $\mathcal{V} = \mathcal{S} \cup \mathcal{D}$  and  $\mathcal{E} = \{(s \rightarrow d) | s \in \mathcal{S}, d \in \mathcal{D}\}$ . Let  $m = |\mathcal{S}|$  and  $n = |\mathcal{D}|$ , then the unit cost  $\boldsymbol{\delta}$  is an  $m \times n$  matrix, and the transportation problem can be rewritten as

$$\begin{aligned} \text{Maximize: } & \mathbf{s} \geq 0, \mathbf{d} \geq 0 \quad z \\ \text{Subject to:} & \inf_{\boldsymbol{\delta} \sim \mathcal{P}} \mathbb{P}[-h(\mathbf{s}, \mathbf{d}, \boldsymbol{\delta}) \geq z] \geq 1 - \gamma; \\ & \mathbf{1}_m^\top \mathbf{s} = \mathbf{1}_n^\top \mathbf{d} \geq L, \end{aligned}$$

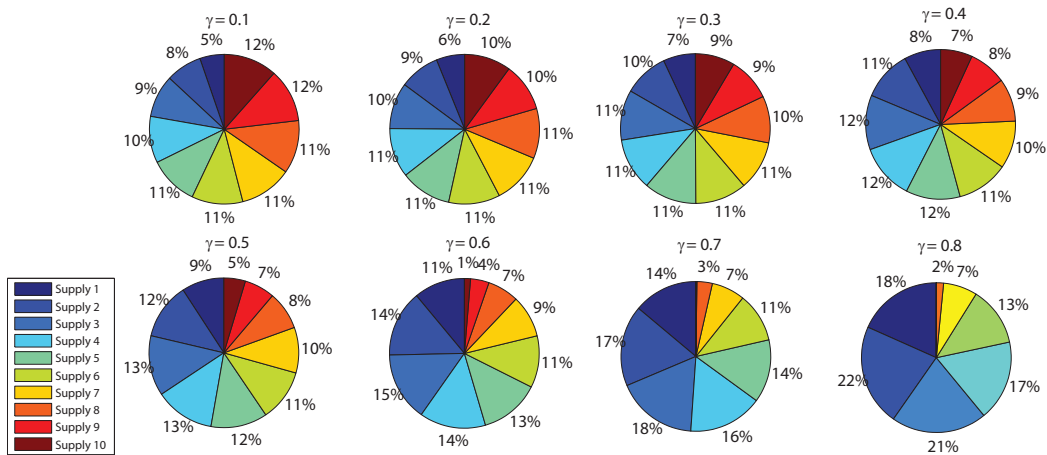
where  $\boldsymbol{\delta} \in \mathbb{R}^{m \times n}$ ,  $\mathbf{1}_m$  and  $\mathbf{1}_n$  are the all-one vectors with dimension  $m$  and  $n$  respectively.

The function  $h(\mathbf{s}, \mathbf{d}, \boldsymbol{\delta})$  is defined by

$$\begin{aligned} h(\mathbf{s}, \mathbf{d}, \boldsymbol{\delta}) = & \text{Minimize}_{\mathbf{F} \in \mathbb{R}^{m \times n}} \quad \text{tr}\langle \boldsymbol{\delta}, \mathbf{F} \rangle \\ \text{Subject to:} \quad & \mathbf{F}^\top \mathbf{1}_m = \mathbf{d}, \mathbf{F} \mathbf{1}_n = \mathbf{s}, \mathbf{F} \geq 0. \end{aligned}$$

By Theorem 2.2, one can solve this transportation problem by MATLAB and CVX [GB11]. We consider the case where there are 10 suppliers and 3 consumers, and the least demand  $L = 80$ . The mean  $M_{ij}$  and the variance  $\Sigma_{ij}$  of the transportation cost  $\delta_{ij}$  are set to  $100 + 0.1\sqrt{3(i-1)+j}$  and  $5/\sqrt{3(i-1)+j}$ , respectively. Then the transportation costs related to suppliers and consumers with lower serial numbers have smaller means but larger variances, i.e., lower mean cost but more risky.

Our first goal is to minimize the total cost to some fixed confidence parameter  $\gamma$ . Figure 2.1 shows the resulting allocations for different  $\gamma$ . As expected, small  $\gamma$  leads to more



**Figure 2.1:** The transportation problem: the resulting allocations for different guarantees  $\gamma = 0.1 - 0.8$ .

conservative allocations which tend to select supplies with higher mean costs and smaller variances, while large  $\gamma$  leads to less conservative allocations which select suppliers with lower mean costs and larger variances.

In this example, the algorithm takes about 40 seconds on a desktop PC with Intel i7 3.4GHz CPU and 8G memory. The computational time for solving the transportation problems of different numbers of suppliers is reported in Table 2.1. For a large-scale problem, i.e. the

Number of suppliers	10	50	100	200	500	1000
Running time (min)	0.88	3.33	4.01	6.21	15.16	34.78

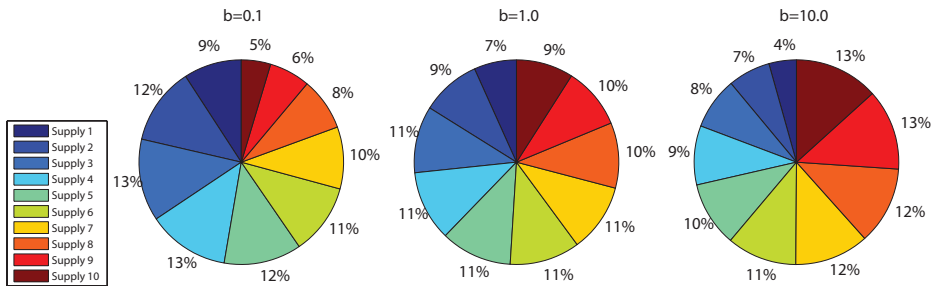
**Table 2.1:** The running time for solving the transportation problem with different numbers of suppliers.

number of suppliers is 1000, our algorithm finds the result in about 30 minutes. From the table, it appears that the computation time scales roughly linearly with respect to the number of suppliers. Note that one can use commercial solvers such as CPLEX instead of CVX to implement this algorithm, which is typically more computationally efficient.

Using the same notations, the transportation problem with probabilistic envelope constraints can be formulated as

$$\begin{aligned}
 & \text{Maximize}_{\mathbf{s}, \mathbf{d}} && z \\
 & \text{Subject to:} && \inf_{\boldsymbol{\delta} \sim \mathcal{P}} \mathbb{P}[-f(\mathbf{s}, \mathbf{d}, \boldsymbol{\delta}) \geq z - s] \geq \mathfrak{B}(s); \\
 & && \mathbf{1}_m^\top \mathbf{s} = \mathbf{1}_n^\top \mathbf{d} \geq D; \\
 & && \mathbf{s}, \mathbf{d} \geq 0.
 \end{aligned}$$

Our second goal is to minimize the total cost subject to a decaying probabilistic envelope  $\mathfrak{B}(s) = 1 - 1/(1 + b\sqrt{s + a/b^2})$  which implies  $t(r) = \max\{(r^2 - a)/b^2, 0\}$  by Lemma 2.3. We choose  $a = 1$  and  $b = 0.1, 1.0, 10.0$ , giving different rates of decay for the probability the constraint is violated at level  $s$  for each  $s$ . Based on Theorem 2.4, we can easily solve this problem. Figure 2.2 shows the resulting allocations. Clearly, larger  $b$  corresponds to a more



**Figure 2.2:** The transportation problem: the resulting allocations for decay rates  $b = 0.1, 1.0$  and  $10.0$ .



risk averse attitude towards large constraint violation so that the resulting allocation is more conservative and tends to choose suppliers with larger mean costs and smaller variances.

## 2.8 Proofs of the Main Results

### 2.8.1 Proof of Corollary 2.1

For clarity, we denote  $-f(\mathbf{x}, \boldsymbol{\delta})$  and  $-\alpha$  by  $L_{\mathbf{x}}(\boldsymbol{\delta})$  and  $\beta$ , respectively. Since  $f(\mathbf{x}, \boldsymbol{\delta})$  is convex w.r.t.  $\boldsymbol{\delta}$  for fixed  $\mathbf{x}$ ,  $L_{\mathbf{x}}(\boldsymbol{\delta})$  is a concave function. Then the equivalence to establish can be rewritten as

$$\sup_{\boldsymbol{\delta} \sim (0, \boldsymbol{\Sigma})} \mathbb{P}[L_{\mathbf{x}}(\boldsymbol{\delta}) > \beta] \leq 1 - p \Leftrightarrow \sup_{\boldsymbol{\delta} \sim (0, \boldsymbol{\Sigma})} \text{CVaR}_{1-p}(L_{\mathbf{x}}(\boldsymbol{\delta})) \leq \beta.$$

It is well known that  $\sup_{\boldsymbol{\delta} \sim (0, \boldsymbol{\Sigma})} \text{CVaR}_{1-p}(L_{\mathbf{x}}(\boldsymbol{\delta})) \leq \beta \Rightarrow \sup_{\boldsymbol{\delta} \sim (0, \boldsymbol{\Sigma})} \mathbb{P}[L_{\mathbf{x}}(\boldsymbol{\delta}) > \beta] \leq 1 - p$ . Besides,  $\sup_{\boldsymbol{\delta} \sim (0, \boldsymbol{\Sigma})} \mathbb{P}[L_{\mathbf{x}}(\boldsymbol{\delta}) > \beta] > 1 - p \Leftrightarrow \sup_{\boldsymbol{\delta} \sim (0, \boldsymbol{\Sigma})} \text{VaR}_{1-p}(L_{\mathbf{x}}(\boldsymbol{\delta})) > \beta$ , hence we only need to show that  $\sup_{\boldsymbol{\delta} \sim (0, \boldsymbol{\Sigma})} \text{CVaR}_{1-p}(L_{\mathbf{x}}(\boldsymbol{\delta})) > \beta \Rightarrow \sup_{\boldsymbol{\delta} \sim (0, \boldsymbol{\Sigma})} \text{VaR}_{1-p}(L_{\mathbf{x}}(\boldsymbol{\delta})) > \beta$ .

Since  $\sup_{\boldsymbol{\delta} \sim (0, \boldsymbol{\Sigma})} \text{CVaR}_{1-p}(L_{\mathbf{x}}(\boldsymbol{\delta})) > \beta$ , then there exists a probability distribution  $\mathbb{P}$  with zero mean, covariance  $\boldsymbol{\Sigma}$ , and  $\text{CVaR}_{1-p}(L_{\mathbf{x}}(\boldsymbol{\delta})) > \beta$  when  $\boldsymbol{\delta} \sim \mathbb{P}$ . Decompose  $\mathbb{P} = \mu_1 + \mu_2$  where the measure  $\mu_1$  constitutes a probability of  $p$  and the measure  $\mu_2$  constitutes a probability of  $1 - p$ , and that  $L_{\mathbf{x}}(y_1) \leq L_{\mathbf{x}}(y_2)$  for any  $y_1$  and  $y_2$  that belong to the support of  $\mu_1$  and  $\mu_2$  respectively. By the CVaR constraint, we have  $(\int_{\boldsymbol{\delta}} L_{\mathbf{x}}(\boldsymbol{\delta}) d\mu_2)/(1 - p) > \beta$ .

We now construct a new probability  $\bar{\mathbb{P}}$  as follows: let  $\mu'_2$  be a measure that put a probability mass of  $1 - p$  on  $\int_{\boldsymbol{\delta}} \boldsymbol{\delta} d\mu_2/(1 - p)$ , i.e., the conditional mean of  $\mu_2$ , and let  $\bar{\mathbb{P}} = \mu_1 + \mu'_2$ . Observe that  $\bar{\mathbb{P}}$  is a probability measure whose mean is the same as that of  $\mathbb{P}$ . Moreover, notice that  $\mu_2/(1 - p)$  is a probability measure, by concavity of  $L_{\mathbf{x}}(\cdot)$  we have that

$$L_{\mathbf{x}}\left(\int_{\boldsymbol{\delta}} \boldsymbol{\delta} d\mu_2/(1 - p)\right) \geq \left(\int_{\boldsymbol{\delta}} L_{\mathbf{x}}(\boldsymbol{\delta}) d\mu_2\right)/(1 - p) > \beta,$$

which implies that  $\text{VaR}_{1-p}(L_{\mathbf{x}}(\boldsymbol{\delta})) > \beta$  for  $\boldsymbol{\delta} \sim \bar{\mathbb{P}}$ .

We now show that this also implies that  $\sup_{\boldsymbol{\delta} \sim (0, \boldsymbol{\Sigma})} \text{VaR}_{1-p}(L_{\mathbf{x}}(\boldsymbol{\delta})) > \beta$ . Denote the covari-

ance w.r.t  $\bar{\mathbb{P}}$  by  $\bar{\Sigma}$  and recall that both  $\mathbb{P}$  and  $\bar{\mathbb{P}}$  are zero mean, then

$$\begin{aligned}
\Sigma - \bar{\Sigma} &= \int_{\delta} \delta \delta^{\top} d\mathbb{P} - \int_{\delta} \delta \delta^{\top} d\bar{\mathbb{P}} \\
&= \int_{\delta} \delta \delta^{\top} d\mu_1 + \int_{\delta} \delta \delta^{\top} d\mu_2 - \int_{\delta} \delta \delta^{\top} d\mu_1 - \int_{\delta} \delta \delta^{\top} d\mu_2' \\
&= \int_{\delta} \delta \delta^{\top} d\mu_2 - (1-p) \left[ \frac{\int_{\delta} \delta d\mu_2}{1-p} \right] \left[ \frac{\int_{\delta} \delta d\mu_2}{1-p} \right]^{\top} \\
&= \int_{\delta} \left\{ \delta - \left[ \frac{\int_{\delta} \delta d\mu_2}{1-p} \right] \right\} \left\{ \delta - \left[ \frac{\int_{\delta} \delta d\mu_2}{1-p} \right] \right\}^{\top} d\mu_2 \succeq 0,
\end{aligned}$$

where the third equality is due to the definition of  $\mu_2'$ . Note that from the construction of  $\bar{\mathbb{P}}$ , we have  $\sup_{\delta \sim (0, \bar{\Sigma})} \mathbb{P}[L_{\mathbf{x}}(\delta) > \beta] > 1-p$ . Denote the set  $\{\delta | L_{\mathbf{x}}(\delta) > \beta\}$  by  $T_{\mathbf{x}}$ . First, we consider the case where  $\bar{\Sigma}$  is full rank. From Lemma 2.2, we have  $\inf_{\mathbf{y} \in T_{\mathbf{x}}} \mathbf{y}^{\top} \bar{\Sigma}^{-1} \mathbf{y} < r \triangleq p/(1-p)$ . Since  $\bar{\Sigma} \preceq \Sigma$  and  $\bar{\Sigma}$  is full rank,  $\inf_{\mathbf{y} \in T_{\mathbf{x}}} \mathbf{y}^{\top} \Sigma^{-1} \mathbf{y} \leq \inf_{\mathbf{y} \in T_{\mathbf{x}}} \mathbf{y}^{\top} \bar{\Sigma}^{-1} \mathbf{y} < r$ , which implies that  $\sup_{\delta \sim (0, \Sigma)} \mathbb{P}[L_{\mathbf{x}}(\delta) > \beta] > 1-p$ , which establishes the theorem.

The case where  $\bar{\Sigma}$  is not full rank requires additional work, as Lemma 2.2 or Theorem 2.1 can not be applied directly. Consider the spectral decomposition  $\bar{\Sigma} = \mathbf{Q} \mathbf{\Lambda} \mathbf{Q}^{\top}$  and denote the pseudo inverse of  $\bar{\Sigma}$  by  $\bar{\Sigma}^+$ . Suppose that the top  $d$  diagonal entries of  $\mathbf{\Lambda}$  are non-zero. Let  $\mathbf{Q}_d$  be the submatrix of  $\mathbf{Q}$  by selecting the first  $d$  columns of  $\mathbf{Q}$  and  $\mathbf{\Lambda}_d$  be the top  $d \times d$  submatrix of  $\mathbf{\Lambda}$ . Denote the column space of  $\bar{\Sigma}$  by  $\mathcal{C}$ , and let  $\mathcal{Q} \triangleq \{\mathbf{z} | \mathbf{z} = \mathbf{Q}_d^{\top} \delta, \forall \delta \in T_{\mathbf{x}} \cap \mathcal{C}\}$ . Since there is no uncertainty in  $\mathcal{C}^{\perp}$  w.r.t  $\bar{\mathbb{P}}$ ,

$$\sup_{\mathbf{z} \sim (0, \mathbf{\Lambda}_d)} \mathbb{P}[\mathbf{z} \in \mathcal{Q}] = \sup_{\delta \sim (0, \bar{\Sigma})} \mathbb{P}[\delta \in T_{\mathbf{x}} \cap \mathcal{C}] = \sup_{\delta \sim (0, \bar{\Sigma})} \mathbb{P}[\delta \in T_{\mathbf{x}}] > 1-p.$$

From Lemma 2.2, we have  $\inf_{\mathbf{z} \in \mathcal{Q}} \mathbf{z}^{\top} \mathbf{\Lambda}_d^{-1} \mathbf{z} < r$ . In other words, there exists  $\mathbf{z} \in \mathcal{Q}$  such that  $\mathbf{z}^{\top} \mathbf{\Lambda}_d^{-1} \mathbf{z} < r$ , which implies that  $\mathbf{y}^{\top} \bar{\Sigma}^+ \mathbf{y} < r$  for  $\mathbf{y} \triangleq \mathbf{Q}_d \mathbf{z}$ . From the Schur complement, since  $\Sigma \succeq \bar{\Sigma} \succeq 0$ ,  $(\mathbf{I} - \bar{\Sigma} \bar{\Sigma}^+) \mathbf{y} = 0$  and  $r - \mathbf{y}^{\top} \bar{\Sigma}^+ \mathbf{y} = r - \mathbf{z}^{\top} \mathbf{\Lambda}_d^{-1} \mathbf{z} > 0$ , we have  $\begin{pmatrix} \Sigma & \mathbf{y} \\ \mathbf{y}^{\top} & r \end{pmatrix} \succeq$

$\begin{pmatrix} \bar{\Sigma} & \mathbf{y} \\ \mathbf{y}^{\top} & r \end{pmatrix} \succeq 0$ . Hence  $\inf_{\mathbf{y} \in T_{\mathbf{x}}} \mathbf{y}^{\top} \Sigma^{-1} \mathbf{y} < r$ , which implies that  $\sup_{\delta \sim (0, \Sigma)} \mathbb{P}[L_{\mathbf{x}}(\delta) > \beta] > 1-p$ .

## 2.8.2 Proofs for Section 2.3.2

**Proof of Corollary 2.2:**

By Theorem 2.1, the feasible set  $S = \{\mathbf{x} | \mathbf{x}^\top g(\mathbf{y}) \geq \alpha, \forall \mathbf{y}^\top \Sigma^{-1} \mathbf{y} \leq r\}$  where  $r = p/(1-p)$ . Hence, determining whether  $\mathbf{x} \in S$  is equivalent to determining whether the inner optimization problem  $\min_{\{\mathbf{y}^\top \Sigma^{-1} \mathbf{y} \leq r\}} \mathbf{x}^\top g(\mathbf{y}) - \alpha \geq 0$ . Rewrite the left hand side as an optimization problem on  $\mathbf{y}$ :

$$\begin{aligned} \text{Minimize: } & \mathbf{y}^\top G(\mathbf{x})\mathbf{y} + P(\mathbf{x})^\top \mathbf{y} + Q(\mathbf{x}) \\ \text{Subject to: } & \mathbf{y}^\top \Sigma^{-1} \mathbf{y} \leq r, \end{aligned} \quad (2.50)$$

by substituting  $g_i(\boldsymbol{\delta}) = \boldsymbol{\delta}^\top G_i \boldsymbol{\delta} + \mathbf{p}_i^\top \boldsymbol{\delta} + q_i$ . To prove Corollary 2.2, we need the following two results.

**Lemma 2.5.** *Fix  $\mathbf{x}$ . The optimal value of the optimization problem (2.50) equals that of the following SDP:*

$$\text{Maximize: } \beta \geq 0, t, \quad \text{subject to: } \begin{pmatrix} \beta \Sigma^{-1} + G(\mathbf{x}) & \frac{1}{2} P(\mathbf{x}) \\ \frac{1}{2} P(\mathbf{x})^\top & Q(\mathbf{x}) - t - \beta r \end{pmatrix} \succeq 0. \quad (2.51)$$

*Proof.* The dual problem of (2.50) is:  $\max_{\beta \geq 0} \min_{\mathbf{y}} \mathbf{y}^\top G(\mathbf{x})\mathbf{y} + P(\mathbf{x})^\top \mathbf{y} + Q(\mathbf{x}) + \beta \mathbf{y}^\top \Sigma^{-1} \mathbf{y} - \beta r$ . By taking minimum over  $\mathbf{y}$  and the Schur complement, this can be reformulated as the SDP (2.51). Notice that there exists  $\mathbf{y}$  such that  $\mathbf{y}^\top \Sigma^{-1} \mathbf{y} < r$  since  $r > 0$ , hence Slater's condition is satisfied for (2.50), and the strong duality holds.  $\square$

Thus,  $\mathbf{x} \in S$  if and only if the optimal value of problem (2.50) is greater than or equal to 0.

This means we can convert the constraint in  $S$  into a feasibility problem as follows:

**Lemma 2.6.** *Under the conditions of Corollary 2.2, and let  $r = p/(1-p)$ , we have the constraint*

$$\inf_{\boldsymbol{\delta} \sim (0, \Sigma)} \mathbb{P}[g(\boldsymbol{\delta})^\top \mathbf{x} \geq \alpha] \geq p, \quad (2.52)$$

is equivalent to the following problem

$$\text{Exist: } \beta \geq 0, \quad \text{s.t.:} \quad \begin{pmatrix} \beta \Sigma^{-1} + G(\mathbf{x}) & \frac{1}{2}P(\mathbf{x}) \\ \frac{1}{2}P(\mathbf{x})^\top & Q(\mathbf{x}) - \beta r \end{pmatrix} \succeq 0. \quad (2.53)$$

*Proof.* 1. Equation (2.52)  $\Rightarrow$  Equation (2.53): When Inequality (2.52) holds, the optimal value  $t$  of (2.50) must be greater than or equal to 0. So from Equation (2.51), we have

$$\begin{pmatrix} \beta \Sigma^{-1} + G(\mathbf{x}) & \frac{1}{2}P(\mathbf{x}) \\ \frac{1}{2}P(\mathbf{x})^\top & Q(\mathbf{x}) - \beta r \end{pmatrix} \succeq \begin{pmatrix} \beta \Sigma^{-1} + G(\mathbf{x}) & \frac{1}{2}P(\mathbf{x}) \\ \frac{1}{2}P(\mathbf{x})^\top & Q(\mathbf{x}) - t - \beta r \end{pmatrix} \succeq 0. \quad (2.54)$$

2. Equation (2.53)  $\Rightarrow$  Equation (2.52): Since the feasibility problem is solvable,  $t = 0$  must be a feasible solution of (2.51), which implies Inequality (2.52).

□

Lemma 2.6 immediately implies Corollary 2.2.

### 2.8.3 Proofs of Results in Section 2.4

#### Proof of Lemma 2.3:

We now show that the constraints (2.13) and (2.14) are equivalent.

1. (2.13)  $\Rightarrow$  (2.14): Since  $\lim_{y \rightarrow +\infty} \mathfrak{B}(y)$  may not converge to 1, we define  $\mathfrak{B}^{-1}(x) = +\infty$  when  $\{y \geq 0 \mid \mathfrak{B}(y) \geq x\} = \emptyset$ . Then if  $r/(1+r)$  is not in the range of  $\mathfrak{B}(s)$ , we have  $t(r) = +\infty$  so that the constraint (2.14) is always satisfied. Otherwise, suppose that  $y^* = t(r) = \inf\{y \geq 0 \mid \mathfrak{B}(y) \geq r/(1+r)\}$ , then we have

$$\inf_{\delta \sim (0, \Sigma)} \mathbb{P}[f(\mathbf{x}, \delta) \geq \alpha - t(r)] = \inf_{\delta \sim (0, \Sigma)} \mathbb{P}[f(\mathbf{x}, \delta) \geq \alpha - y^*] \geq \mathfrak{B}(y^*) \geq \frac{r}{1+r}.$$

2. (2.14)  $\Rightarrow$  (2.13): Since  $\mathfrak{B}(y^*) \in [0, 1)$  for any  $y^* \geq 0$ , there exists  $r^*$  such that  $\mathfrak{B}(y^*) = \frac{r^*}{1+r^*}$ . From the definition of  $t(r)$ , we have  $y^* \geq t(r^*) = \inf\{y \geq 0 \mid \mathfrak{B}(y) \geq r^*/(1+r^*)\}$ .

Hence the following inequality holds

$$\inf_{\boldsymbol{\delta} \sim (0, \boldsymbol{\Sigma})} \mathbb{P}[f(\mathbf{x}, \boldsymbol{\delta}) \geq \alpha - y^*] \geq \inf_{\boldsymbol{\delta} \sim (0, \boldsymbol{\Sigma})} \mathbb{P}[f(\mathbf{x}, \boldsymbol{\delta}) \geq \alpha - t(r^*)] \geq \frac{r^*}{1 + r^*} = \mathfrak{B}(y^*).$$

Furthermore,  $t(\cdot)$  is non-decreasing since both  $r/(1+r)$  and  $\mathfrak{B}(\cdot)$  are non-decreasing. By definition of  $t$ ,  $\mathfrak{B}(0) \geq 0$  leads to  $t(0) = 0$ ; and  $\mathfrak{B}(s) < 1$  for all  $s > 0$  leads to  $\mathfrak{B}^{-1}(1) = +\infty$  and hence  $\lim_{r \uparrow +\infty} t(r) = +\infty$ . Also,  $\mathfrak{B}(0) > 0$  implies for some  $\epsilon > 0$ ,  $\mathfrak{B}(0) \geq \epsilon$ , and hence  $t(\epsilon) = 0$ . Thus,  $t(\cdot)$  is continuous at a neighborhood of 0.

### Proof of Corollary 2.3:

The feasible set  $S = \{\mathbf{x} \mid \inf_{\boldsymbol{\delta} \sim (0, \boldsymbol{\Sigma})} \mathbb{P}[g(\boldsymbol{\delta})^\top \mathbf{x} \geq \alpha - t(r)] \geq \frac{r}{1+r}; \forall r \geq 0\}$  admits

$$\begin{aligned} S &\stackrel{(a)}{=} \{\mathbf{x} \mid \forall (\mathbf{y}, r) \text{ such that } \mathbf{y}^\top \boldsymbol{\Sigma}^{-1} \mathbf{y} \leq r \Rightarrow g(\mathbf{y})^\top \mathbf{x} \geq \alpha - t(r)\} \\ &= \{\mathbf{x} \mid \min_{\{\mathbf{y}, r \mid \mathbf{y}^\top \boldsymbol{\Sigma}^{-1} \mathbf{y} \leq r\}} g(\mathbf{y})^\top \mathbf{x} + t(r) - \alpha \geq 0\}, \end{aligned}$$

where (a) holds by Theorem 2.3. As each component  $g_i(\mathbf{y})$  of  $g(\mathbf{y})$  is linear or quadratic, i.e.,  $g_i(\mathbf{y}) = \mathbf{y}^\top \mathbf{G}_i \mathbf{y} + \mathbf{p}_i^\top \mathbf{y} + q_i$ , for fixed  $\mathbf{x}$  the inner optimization problem

$$\min_{\{\mathbf{y}, r \mid \mathbf{y}^\top \boldsymbol{\Sigma}^{-1} \mathbf{y} \leq r\}} g(\mathbf{y})^\top \mathbf{x} + t(r) - \alpha$$

can be rewritten as:

$$\begin{aligned} &\text{Minimize}_{:r \geq 0, \mathbf{y}} \quad \mathbf{y}^\top G(\mathbf{x}) \mathbf{y} + P(\mathbf{x})^\top \mathbf{y} + Q(\mathbf{x}) + t(r) \\ &\text{subject to:} \quad \mathbf{y}^\top \boldsymbol{\Sigma}^{-1} \mathbf{y} - r \leq 0, \end{aligned} \tag{2.55}$$

where  $G(\mathbf{x}) \triangleq \sum_{i=1}^n x_i \mathbf{G}_i$ ,  $P(\mathbf{x}) \triangleq \sum_{i=1}^n x_i \mathbf{p}_i$ , and  $Q(\mathbf{x}) \triangleq \sum_{i=1}^n x_i q_i - \alpha$ . Thus, in order to analyze  $S$ , we need to analyze the optimization problem (2.55). We have the following lemma:

**Lemma 2.7.** *For any fixed  $\mathbf{x}$ , the optimal value of problem (2.55) is equivalent to that of*

the following:

$$\text{Maximize: } \beta \geq 0, \eta, \quad \text{subject to: } \begin{pmatrix} \beta \Sigma^{-1} + G(\mathbf{x}) & \frac{1}{2}P(\mathbf{x}) \\ \frac{1}{2}P(\mathbf{x})^\top & Q(\mathbf{x}) - t^*(\beta) - \eta \end{pmatrix} \succeq 0, \quad (2.56)$$

where  $t^*(x)$  is the conjugate function of  $t(r)$  defined as  $t^*(x) = \sup_{r \geq 0} (xr - t(r))$ .

*Proof.* By the Schur complement and the strong duality of problem (2.55) (Slater's condition holds by picking  $r = 1$  and  $\mathbf{y} = 0$ ), one can easily obtain this lemma.  $\square$

From Lemma 2.7, the constraint  $\inf_{\delta \sim (0, \Sigma)} \mathbb{P}[g(\delta)^\top \mathbf{x} \geq \alpha - t(r)] \geq \frac{r}{1+r}$ ,  $\forall r \geq 0$ , is equivalent to a constraint that the optimal value of the optimization problem (2.56) is non-negative. Thus,  $\mathbf{x}$  belongs to the feasible set of the envelope constraint if and only if  $\alpha = 0$  is a feasible solution of (2.56), for the same  $\mathbf{x}$ . This means we can remove the  $-\alpha$  term from (2.56). That is, when each component  $g_i(\cdot)$  of  $g(\cdot)$  is linear or quadratic, the envelope constraint is equivalent to the following feasibility problem:

$$\text{exist: } \beta \geq 0, \quad \text{s.t.: } \begin{pmatrix} \beta \Sigma^{-1} + G(\mathbf{x}) & \frac{1}{2}P(\mathbf{x}) \\ \frac{1}{2}P(\mathbf{x})^\top & Q(\mathbf{x}) - t^*(\beta) \end{pmatrix} \succeq 0. \quad (2.57)$$

Hence the optimization problem (2.18) is equivalent to (2.21), which proves the first part of the Theorem.

To prove the second part of the Theorem, it suffices to show that Problem (2.21) can be solved in polynomial time. We show this by constructing a polynomial time separation oracle. For any  $(\beta, \mathbf{x})$ , if the optimization problem (2.22) can be solved in polynomial time, which implies  $t^*(\beta)$  can be computed in polynomial time, then it can be verified in polynomial time whether the constraint in (2.21) is satisfied or not, and hence the feasibility of  $(\beta, \mathbf{x})$  can be determined in polynomial time. Moreover, if  $(\beta_0, \mathbf{x}_0)$  is infeasible and let  $r_0$  be the optimal solution of the problem (2.22) (by assumption  $r_0$  can be found in polynomial

time), then we have

$$\begin{pmatrix} \beta_0 \Sigma^{-1} + G(\mathbf{x}_0) & \frac{1}{2}P(\mathbf{x}_0) \\ \frac{1}{2}P(\mathbf{x}_0)^\top & Q(\mathbf{x}_0) + t(r_0) - \beta_0 r_0 \end{pmatrix} \not\leq 0,$$

and we can find in polynomial time (e.g., by SVD) a vector  $(\mathbf{y}_0^\top, 1)$  such that

$$\begin{aligned} & (\mathbf{y}_0^\top, 1) \begin{pmatrix} \beta_0 \Sigma^{-1} + G(\mathbf{x}_0) & \frac{1}{2}P(\mathbf{x}_0) \\ \frac{1}{2}P(\mathbf{x}_0)^\top & Q(\mathbf{x}_0) + t(r_0) - \beta_0 r_0 \end{pmatrix} \begin{pmatrix} \mathbf{y}_0 \\ 1 \end{pmatrix} \\ & = (\mathbf{y}_0^\top \Sigma^{-1} \mathbf{y}_0 - r_0) \beta_0 + \mathbf{y}_0^\top G(\mathbf{x}_0) \mathbf{y}_0 + P(\mathbf{x}_0)^\top \mathbf{y}_0 + Q(\mathbf{x}_0) + t(r_0) < 0. \end{aligned}$$

Notice that for any feasible solution  $(\beta, \mathbf{x})$ , we must have

$$(\mathbf{y}_0^\top \Sigma^{-1} \mathbf{y}_0 - r_0) \beta + \mathbf{y}_0^\top G(\mathbf{x}) \mathbf{y}_0 + P(\mathbf{x})^\top \mathbf{y}_0 + Q(\mathbf{x}) + t(r_0) \geq 0.$$

Hence we have a separating hyperplane.

#### 2.8.4 Proofs for Section 2.5

##### Proof of Corollary 2.6:

As before, we construct a separation oracle to prove tractability. In order to verify the feasibility of a given  $\mathbf{x}^*$ , from Corollary 2.5 we know that  $\mathbf{x}^*$  is feasible if and only if the optimal value of the optimization problem (2.26) is greater than or equal to  $\alpha$ , which can be verified by directly solving Problem (2.26). By assumption, this can be done in polynomial time.

If  $\mathbf{x}^*$  is not feasible, then we can find in polynomial time  $(\mathbf{y}^*, r^*, \boldsymbol{\mu}^*, \Sigma^*)$  such that  $f(\mathbf{x}^*, \mathbf{y}^* + \boldsymbol{\mu}^*) + t(r^*) < \alpha$ . Because  $f(\mathbf{x}, \mathbf{y} + \boldsymbol{\mu})$  is concave in  $\mathbf{x}$  for fixed  $\mathbf{y}$  and  $\boldsymbol{\mu}$ , for any feasible  $\mathbf{x}$ , we have

$$f(\mathbf{x}^*, \mathbf{y}^* + \boldsymbol{\mu}^*) + \nabla f(\mathbf{x}^*, \mathbf{y}^* + \boldsymbol{\mu}^*)^\top (\mathbf{x} - \mathbf{x}^*) + t(r^*) \geq f(\mathbf{x}, \mathbf{y}^* + \boldsymbol{\mu}^*) + t(r^*) \geq \alpha.$$

Hence the hyperplane separating  $\mathbf{x}^*$  from the feasible set is the following:

$$f(\mathbf{x}^*, \mathbf{y}^* + \boldsymbol{\mu}^*) + \nabla f(\mathbf{x}^*, \mathbf{y}^* + \boldsymbol{\mu}^*)^\top (\mathbf{x} - \mathbf{x}^*) + t(r^*) \geq \alpha, \quad (2.58)$$

which can be generated in polynomial time since the super-gradient of  $\mathbf{x}$  can be obtained in polynomial time.

### Proof of Theorem 2.5:

If  $f(\mathbf{x}, \boldsymbol{\delta}) \geq \alpha$  for all  $\boldsymbol{\delta} \in \mathcal{S}$ , the constraints (2.27) and (2.28) are satisfied, so we only need to consider the case where there exists  $\boldsymbol{\delta} \in \mathcal{S}$  such that  $f(\mathbf{x}, \boldsymbol{\delta}) < \alpha$ . Note that (2.27) is equivalent to  $\sup_{\boldsymbol{\delta} \sim (\boldsymbol{\mu}, \mathcal{S})} \mathbb{P}[f(\mathbf{x}, \boldsymbol{\delta}) < \alpha] \leq 1 - p$ , then we can apply the following lemma:

**Lemma 2.8.** *If the conditions in Theorem 2.5 hold and  $\{\boldsymbol{\delta} : f(\mathbf{x}, \boldsymbol{\delta}) < \alpha\}$  is nonempty, then*

$$\sup_{\boldsymbol{\delta} \sim (\boldsymbol{\mu}, \mathcal{S})} \mathbb{P}[f(\mathbf{x}, \boldsymbol{\delta}) < \alpha] \leq 1 - p \quad (2.59)$$

is equivalent to

$$1 - p \geq \left\{ \begin{array}{l} \sup_{\theta, \boldsymbol{\delta}_1, \boldsymbol{\delta}_2} \theta \\ \text{such that } \theta \boldsymbol{\delta}_1 + (1 - \theta) \boldsymbol{\delta}_2 = \boldsymbol{\mu}, \\ \\ 0 \leq \theta \leq 1, \\ \\ f(\mathbf{x}, \boldsymbol{\delta}_1) < \alpha, \\ \\ \boldsymbol{\delta}_1, \boldsymbol{\delta}_2 \in \mathcal{S}. \end{array} \right. \quad (2.60)$$

*Proof.* Since  $\boldsymbol{\mu} \in \mathcal{S}$  and  $\{\boldsymbol{\delta} : f(\mathbf{x}, \boldsymbol{\delta}) < \alpha\}$  is not empty, the optimization problem in (2.60) is always feasible. To show the equivalence of (2.59) and (2.60), one needs to prove that the optimal objective value  $\theta^*$  of the optimization problem in (2.60) equals  $\zeta = \sup_{\boldsymbol{\delta} \sim (\boldsymbol{\mu}, \mathcal{S})} \mathbb{P}[f(\mathbf{x}, \boldsymbol{\delta}) < \alpha]$ .

The first step is to show  $\theta^* \leq \zeta$ : Since  $f(\mathbf{x}, \cdot)$  is continuous for fixed  $\mathbf{x} \in \mathbb{X}$  and  $\mathcal{S}$  is a closed convex set, for any  $\epsilon > 0$  there exists a feasible solution  $(\boldsymbol{\delta}'_1, \boldsymbol{\delta}'_2, \theta')$  such that  $|\theta' - \theta^*| < \epsilon$ . Construct a probability distribution  $\mathbb{P}'(\mathbf{x})$  such that  $\boldsymbol{\delta} = \boldsymbol{\delta}'_1$  with probability



$\theta'$  and  $\boldsymbol{\delta} = \boldsymbol{\delta}'_2$  with probability  $1 - \theta'$ , then we have  $\mathbb{P}' \in (\boldsymbol{\mu}, \mathcal{S})$ . By construction we have  $\theta' \leq \mathbb{P}'[f(\mathbf{x}, \boldsymbol{\delta}) < \alpha] \leq \zeta$  where the second inequality holds from  $\mathbb{P}' \in (\boldsymbol{\mu}, \mathcal{S})$ . Thus we have  $\theta^* \leq \zeta$  as  $\epsilon$  can be arbitrarily small.

The second step is to prove  $\theta^* \geq \zeta$ : Consider any probability distribution  $\bar{\mathbb{P}} \in (\boldsymbol{\mu}, \mathcal{S})$ , and define  $\bar{\theta} = \bar{\mathbb{P}}[f(\mathbf{x}, \boldsymbol{\delta}) < \alpha]$ ,  $\bar{\boldsymbol{\delta}}_1 = \mathbb{E}_{\bar{\mathbb{P}}}[\boldsymbol{\delta} | f(\mathbf{x}, \boldsymbol{\delta}) < \alpha]$  and  $\bar{\boldsymbol{\delta}}_2 = \mathbb{E}_{\bar{\mathbb{P}}}[\boldsymbol{\delta} | f(\mathbf{x}, \boldsymbol{\delta}) \geq \alpha]$ . We then have  $\bar{\boldsymbol{\delta}}_1 \bar{\theta} + \bar{\boldsymbol{\delta}}_2 (1 - \bar{\theta}) = \boldsymbol{\mu}$ ,  $f(\mathbf{x}, \bar{\boldsymbol{\delta}}_1) < \alpha$  and  $\bar{\boldsymbol{\delta}}_1, \bar{\boldsymbol{\delta}}_2 \in \mathcal{S}$ , or equivalently  $(\bar{\boldsymbol{\delta}}_1, \bar{\boldsymbol{\delta}}_2, \bar{\theta})$  is a feasible solution of the optimization problem in (2.60). Thus, we must have  $\bar{\mathbb{P}}[f(\mathbf{x}, \boldsymbol{\delta}) < \alpha] = \bar{\theta} \leq \theta^*$ , which implies that  $\theta^* \geq \zeta = \sup_{\boldsymbol{\delta} \sim (\boldsymbol{\mu}, \mathcal{S})} \mathbb{P}[f(\mathbf{x}, \boldsymbol{\delta}) < \alpha]$ . Therefore, (2.59) is equivalent to (2.60).  $\square$

From the equivalence shown in Lemma 2.8, we consider the following feasibility problem parameterized by  $\theta \in [0, 1]$ , denoted  $\mathcal{F}_\theta$ :

$$\begin{aligned} \text{exist:} & \quad \boldsymbol{\delta}_1, \boldsymbol{\delta}_2 \\ \text{such that:} & \quad \theta \boldsymbol{\delta}_1 + (1 - \theta) \boldsymbol{\delta}_2 = \boldsymbol{\mu}, \\ & \quad f(\mathbf{x}, \boldsymbol{\delta}_1) < \alpha, \\ & \quad \boldsymbol{\delta}_1, \boldsymbol{\delta}_2 \in \mathcal{S}. \end{aligned}$$

Then we have that for any  $0 \leq \theta_1 \leq \theta_2 \leq 1$ ,  $\mathcal{F}_{\theta_2}$  being feasible implies  $\mathcal{F}_{\theta_1}$  being feasible. To see this, let  $(\boldsymbol{\delta}_1^*, \boldsymbol{\delta}_2^*)$  be a feasible solution to  $\mathcal{F}_{\theta_2}$ . Hence we have  $\theta_2 \boldsymbol{\delta}_1^* + (1 - \theta_2) \boldsymbol{\delta}_2^* = \boldsymbol{\mu}$ . Let  $\boldsymbol{\delta}'_2$  be such that

$$\boldsymbol{\mu} - \boldsymbol{\delta}'_2 = (\boldsymbol{\mu} - \boldsymbol{\delta}_2^*) \times \frac{(1 - \theta_2) \theta_1}{(1 - \theta_1) \theta_2}.$$

Since  $\theta_2 \geq \theta_1$ , we have that  $\boldsymbol{\delta}'_2$  is on the line segment between  $\boldsymbol{\mu}$  and  $\boldsymbol{\delta}_2^*$ , and hence belongs to  $\mathcal{S}$  by its convexity. Furthermore, it is easy to check that  $(\boldsymbol{\delta}_1^*, \boldsymbol{\delta}'_2)$  is feasible to  $\mathcal{F}_{\theta_1}$ .

Thus, constraint (2.60) (and equivalently the chance constraint (2.27)) is equivalent to  $\mathcal{F}_{1-p+\epsilon}$  infeasible for all  $\epsilon > 0$ , i.e.,

$$\boldsymbol{\delta}_2 = -\frac{1-p+\epsilon}{p-\epsilon}(\boldsymbol{\delta}_1 - \boldsymbol{\mu}) + \boldsymbol{\mu} \notin \mathcal{S}, \quad \forall f(\mathbf{x}, \boldsymbol{\delta}_1) < \alpha \text{ and } \boldsymbol{\delta}_1 \in \mathcal{S}.$$

This further implies  $\mathcal{F}_{1-p}$  is infeasible, i.e.,

$$\boldsymbol{\delta}_2 = -\frac{1-p}{p}(\boldsymbol{\delta}_1 - \boldsymbol{\mu}) + \boldsymbol{\mu} \notin \mathcal{S}, \quad \forall f(\mathbf{x}, \boldsymbol{\delta}_1) < \alpha \text{ and } \boldsymbol{\delta}_1 \in \mathcal{S}. \quad (2.61)$$

To see this, we only need to show that  $\mathcal{F}_{1-p}$  being feasible implies that  $\mathcal{F}_{1-p+\varepsilon}$  is feasible for some  $\varepsilon > 0$ . Suppose that there exists  $\boldsymbol{\delta}_1^* \in \mathcal{S}$  such that  $\boldsymbol{\delta}_2^* = -\frac{1-p}{p}(\boldsymbol{\delta}_1^* - \boldsymbol{\mu}) + \boldsymbol{\mu} \in \mathcal{S}$  and  $f(\mathbf{x}, \boldsymbol{\delta}_1^*) < \alpha$ . By continuity of  $f(\mathbf{x}, \cdot)$ , we have that for a sufficiently small  $\eta > 0$ ,  $f(\mathbf{x}, \boldsymbol{\delta}'_1) < \alpha$  where  $\boldsymbol{\delta}'_1 \triangleq (1-\eta)\boldsymbol{\delta}_1^* + \eta\boldsymbol{\mu}$ . Note that  $\boldsymbol{\delta}'_1 \in \mathcal{S}$  and there exists  $\varepsilon > 0$  such that  $-\frac{1-p+\varepsilon}{p-\varepsilon}(\boldsymbol{\delta}'_1 - \boldsymbol{\mu}) + \boldsymbol{\mu} \in \mathcal{S}$ , which implies that  $\mathcal{F}_{1-p+\varepsilon}$  is feasible.

Finally, the constraint (2.61) can be rewritten as

$$0 < \min_{\boldsymbol{\delta}_1, \boldsymbol{\delta}_2} \|(1-p)\boldsymbol{\delta}_1 + p\boldsymbol{\delta}_2 - \boldsymbol{\mu}\|_2 \text{ s.t. } f(\mathbf{x}, \boldsymbol{\delta}_1) < \alpha, \boldsymbol{\delta}_1 \in \mathcal{S}, \boldsymbol{\delta}_2 \in \mathcal{S}, \quad (2.62)$$

which is equivalent to (2.28). Therefore, the theorem follows.

### 2.8.5 Proofs of Results in Section 2.6

#### Proof of Theorem 2.11:

The constraints in  $Z^P$  (2.47) requires that for any  $r \geq 0$ , we can find  $\beta$  and  $\mathbf{M}$  to satisfy  $\beta + (r+1)\langle \boldsymbol{\Omega}, \mathbf{M} \rangle \leq 0$  and the other  $(m+1)$  semi-definite constraints. This is equivalent to requiring that the following optimization problem has an optimal value less than or equal to 0 (notice that for any  $r \geq 0$ , finding  $\beta$  and  $\mathbf{M}$  to satisfy the  $(m+1)$  semi-definite constraints itself is trivial):

$$\begin{aligned} & \max_{r \geq 0} \min_{\mathbf{M} \succeq 0, \beta} \beta + (r+1)\langle \boldsymbol{\Omega}, \mathbf{M} \rangle \\ & \text{s.t. } \mathbf{M} - \begin{pmatrix} \alpha_i Q_i(\mathbf{x}) & \frac{1}{2}\alpha_i y_i(\mathbf{x}) \\ \frac{1}{2}\alpha_i y_i(\mathbf{x})^\top & \alpha_i (y_i^0(\mathbf{x}) - t(r)) - \beta \end{pmatrix} \succeq 0 \quad \forall i = 1, \dots, m. \end{aligned} \quad (2.63)$$

We analyze this requirement using duality. In order to find the dual problem of (2.63), it is more convenient for us to analyze the following problem:

$$\begin{aligned} \min_{r \geq 0} \max_{\mathbf{M} \succeq 0, \beta} & -\beta - (r+1)\langle \boldsymbol{\Omega}, \mathbf{M} \rangle \\ \text{s.t. } & \mathbf{M} - \begin{pmatrix} \alpha_i Q_i(\mathbf{x}) & \frac{1}{2} \alpha_i y_i(\mathbf{x}) \\ \frac{1}{2} \alpha_i y_i(\mathbf{x})^\top & \alpha_i (y_i^0(\mathbf{x}) - t(r)) - \beta \end{pmatrix} \succeq 0 \quad \forall i = 1, \dots, m. \end{aligned} \quad (2.64)$$

Consider the dual problem, the ‘‘max’’ part in (2.64) is equivalent to

$$L(r) = \min_{\lambda_i \geq 0} \max_{\beta, \mathbf{M}} -\beta - (r+1)\langle \boldsymbol{\Omega}, \mathbf{M} \rangle + \sum_{i=1}^m \lambda_i \lambda_{\min}(\mathbf{M} - \mathbf{S}_i + \beta \mathbf{E}) + \lambda_0 \lambda_{\min}(\mathbf{M}), \quad (2.65)$$

where the function  $\lambda_{\min}(\mathbf{X})$  denotes minimum eigenvalue of matrix  $\mathbf{X}$ , and  $\mathbf{S}_i \triangleq \alpha_i \mathbf{B}_i - \alpha_i t(r) \mathbf{E}$ . Further note that the function  $\lambda_{\min}(\mathbf{X})$  is equivalent to the following optimization problem:  $\min_{\mathbf{Y} \succeq 0, \text{tr}(\mathbf{Y})=1} \text{tr}(\mathbf{Y}\mathbf{X})$ . Thus (2.65) is equivalent to

$$L(r) = \min_{\lambda_i \geq 0} \max_{\beta, \mathbf{M}} \min_{\{\mathbf{Y}_i | \text{tr}(\mathbf{Y}_i)=1, \mathbf{Y}_i \succeq 0\}} -\beta - (r+1)\langle \boldsymbol{\Omega}, \mathbf{M} \rangle + \text{tr}(\lambda_0 \mathbf{Y}_0 \mathbf{M}) + \sum_{i=1}^m \text{tr}(\lambda_i \mathbf{Y}_i (\mathbf{M} - \mathbf{S}_i + \beta \mathbf{E})).$$

Notice that for any fixed  $\lambda$ , the objective function is continuous, convex w.r.t.  $(\mathbf{Y}_i)_{i=0}^m$  and concave w.r.t.  $(\beta, \mathbf{M})$ . Moreover, the feasible set of  $(\mathbf{Y}_i)_{i=0}^m$  is compact and does not depend on  $(\beta, \mathbf{M})$ . Hence Sion’s minimax theorem applies, and we have

$$L(r) = \min_{\lambda_i \geq 0} \min_{\{\mathbf{Y}_i | \text{tr}(\mathbf{Y}_i)=\lambda_i, \mathbf{Y}_i \succeq 0\}} \max_{\beta, \mathbf{M}} - \sum_{i=1}^m \text{tr}(\mathbf{Y}_i \mathbf{S}_i) + \langle \mathbf{M}, \sum_{i=0}^m \mathbf{Y}_i - (r+1) \boldsymbol{\Omega} \rangle + \beta (\langle \mathbf{E}, \sum_{i=1}^m \mathbf{Y}_i \rangle - 1).$$

Taking maximum over  $\beta$  and  $M$ , we have that  $L(r)$  is equivalent to the following optimization problem with variables  $\mathbf{Y}_i$  and  $\lambda_i$ :

$$\begin{aligned} L(r) = \min_{\lambda_i \geq 0} \min_{\{\mathbf{Y}_i | \text{tr}(\mathbf{Y}_i)=\lambda_i, \mathbf{Y}_i \succeq 0\}} & - \sum_{i=1}^m \text{tr}(\alpha_i \mathbf{Y}_i (\mathbf{B}_i - t(r) \mathbf{E})) \\ \text{s.t. } & \sum_{i=0}^m \mathbf{Y}_i = (r+1) \boldsymbol{\Omega}, \quad \text{tr}(\mathbf{E} \sum_{i=1}^m \mathbf{Y}_i) = 1. \end{aligned}$$

By taking minimum over  $\lambda_i$ ,  $\min_{r \geq 0} L(r)$  can be further reformulated as (2.48). Hence from the analysis above, we know that (2.64) is equivalent to (2.48). To complete the proof, we

construct a separation oracle of  $Z^P$  based on (2.48). Given  $\mathbf{x}$ , if the optimization problem (2.48) can be solved in polynomial time, then it can be verified whether  $\mathbf{x} \in Z^P$  or not in polynomial time since  $\mathbf{x}$  is feasible if and only if the optimal value of (2.48) is greater than or equal to 0. Furthermore, if  $\mathbf{x} \notin Z^P$ , let the optimal solution of (2.48) be  $(r^0, \{\mathbf{Y}_i^0\})$ , then we have  $-\sum_{i=1}^m \text{tr}(\alpha_i \mathbf{Y}_i^0 \mathbf{B}_i) + t(r^0) \sum_{i=1}^m \alpha_i \text{tr}(\mathbf{Y}_i^0 \mathbf{E}) < 0$  since  $\mathbf{x} \notin Z^P$ . On the other hand, for any  $\mathbf{x} \in Z^P$ , the following inequality must be satisfied

$$-\sum_{i=1}^m \text{tr}(\alpha_i \mathbf{Y}_i^0 \mathbf{B}_i) + t(r^0) \sum_{i=1}^m \alpha_i \text{tr}(\mathbf{Y}_i^0 \mathbf{E}) \geq 0,$$

which implies that a separating hyperplane can be generated in polynomial time.

## 2.9 Chapter Summary

The distributionally robust chance constraint formulation has been extensively studied. Yet, most previous work focused on the *linear* constraint function case. In this chapter, motivated by applications where uncertainty is inherently *non-linear*, we investigate the computational aspects of distributionally robust chance constrained optimization problems for the general function case. We show that the distributionally robust chance constrained optimization is tractable, provided that the uncertainty is characterized by its mean and variance, and the constraint function is concave-convex. This significantly expands the range of decision problems that can be modeled and solved efficiently via the DRCC framework. Along the way, we establish a relationship between the DRCC framework and robust optimization model, which links the stochastic model and the deterministic model of uncertainty. We then consider probabilistic envelope constraints, a generalization of distributionally robust chance constraint first proposed in Xu *et al.* [XCM12], and extend this framework to the non-linear case, and obtain conditions that guarantee its tractability. Finally, we discuss two extensions of our approach, provide approximation schemes for JPEC, and establish conditions to ensure these approximation formulations are tractable.

# CHAPTER 3

## A Unified Robust Regression Model for Lasso-like Algorithms

We develop a unified robust linear regression model and show that it is equivalent to a general regularization framework to encourage sparse-like structure that contains group Lasso and fused Lasso as specific examples. This provides a robustness interpretation of these widely applied Lasso-like algorithms, and allows us to construct novel generalizations of Lasso-like algorithms by considering different uncertainty sets. Using this robustness interpretation, we present new sparsity results, and establish the statistical consistency of the proposed regularized linear regression. This work extends a classical result from [XCM10] that relates standard Lasso with robust linear regression to learning problems with more general sparse-like structures, and provides new robustness-based tools to understand learning problems with sparse-like structures.

### 3.1 Introduction

In this chapter we establish a unified relationship between robustness and regularization schemes for various sparse-like structures, in the context of linear regression. Linear regression aims to find a vector  $\beta$  such that  $\mathbf{y} \approx \mathbf{X}\beta$ , for a given matrix  $\mathbf{X} \in \mathcal{R}^{n \times m}$  and vector  $\mathbf{y} \in \mathcal{R}^n$ . From a learning perspective, each row of  $\mathbf{X}$  represents a training sample, and the corresponding element of  $\mathbf{y}$  is the target value or response of this observed sample. Each column of  $\mathbf{X}$  corresponds to a feature, and the objective of linear regression is to obtain a set of weights so that the weighted sum of the feature values approximates the target value.

Regularized linear regression framework – where one finds the solution that minimizes a weighted combination of the residual norm and a certain regularization term, e.g., [TA77, Tib96] – is now a standard practice in machine learning and statistics for linear regression. Among different regularization schemes, the  $\ell_1$  regularized linear regression, also termed *Lasso* [Tib96, CDS99, EHJT04], is increasingly popular due to its tendency to select sparse solutions. Indeed, Lasso has been extremely successful in the high-dimensional regime, as it allows recovering the true solution  $\beta^*$  where the samples are significantly outnumbered by the dimensionality by exploiting sparse structure of  $\beta^*$ . Extensive effort has been made to explain the success of Lasso, e.g., [Tro06, Don06, Wai09, BRT09, Zha09], among which, one interesting result from [XCM10] showed that the success of Lasso is due to its robustness. In particular, they showed that Lasso is equivalent to a robust linear regression formulation, and such robustness interpretation implies the sparsity and the consistency of Lasso.

Inspired by the success of Lasso, numerous regularization schemes were proposed to select solutions with more general sparse-like structures. For example, domain knowledge may indicate that the solution is group sparse, i.e., features can be grouped, and the features belonging to one group is likely to be either all non-active (corresponding to the regressor having zero coefficients), or all active. One example of group sparsity appears is measuring gene expression, where experiments show that selecting a few genes that belong to the same functional groups can lead to increased interpretability of the predictive signature [RZD<sup>+</sup>07]. A prominent algorithm proposed to enforce this sparse-like structure is the group Lasso formulation [YL06], where the regularization term is the sum of the  $\ell_2$ -norms of the different groups of features, also called the  $\ell_1/\ell_2$ -norm. This formulation leads to a sparse selection of the *groups* of features. Other examples of Lasso-like algorithms include the fused Lasso [TSR<sup>+</sup>05] that encourages sparsity of the coefficients and also sparsity of their differences, the sparse group Lasso [FHT10] that encourages solutions that are sparse at both the group and individual feature levels, and many others.

This chapter attempts to explain the success of those Lasso-like algorithms in a unified way. Our approach is largely inspired by [XCM10] – we analyze these algorithms based on their robustness properties. In specific, our first result states that a wide range of regularized linear regression problems including the aforementioned ones, all have equivalent robust

regression reformulations. This provides a robustness re-interpretation of a class of regularized linear regression formulations for sparse-like structured solutions, and generalizes similar results of standard Lasso showed in [XCM10]. Moreover, our robustness interpretation leads to new formulation and new analysis. We derive new regularization variants of Lasso-like algorithms by considering different uncertainty sets of the robust linear regression formulation. We then present new sparsity results for the group Lasso, as well as proofs of consistency of Lasso-like algorithms, all based on the robustness interpretation. Since robustness is a geometric concept, our approach gives new analysis and new geometric intuition compared to previous methods.

**Notations.** We use lower-case boldface letters to denote column vectors and upper-case boldface letters to denote matrices. The operator vectorizing a matrix by stacking its columns is denoted by  $\text{vec}(\cdot)$ . For simplicity, we use  $\|\mathbf{X}\|_p$  to denote the  $\ell_p$ -norm of  $\text{vec}(\mathbf{X})$ , e.g.  $\|\mathbf{X}\|_2$  is the Frobenius norm  $\|\mathbf{X}\|_F$ , and  $\|\mathbf{X}\|_p^*$  to denote its dual norm. We denote the set  $\{1, \dots, m\}$  as  $[m]$  and call a subset  $g$  of  $[m]$  a *group*. The identity matrix is denoted by  $\mathbf{I}$ , the  $i$ th element of vector  $\mathbf{x}$  is denoted by  $x_i$ , and the  $i$ th column of matrix  $\mathbf{\Delta}$  is denoted by  $\mathbf{\Delta}_i$ . For vector  $\mathbf{x}$  and group  $g$ , we denote  $\mathbf{x}_g$  as the vector whose  $i$ th element is  $x_i$  if  $i \in g$  or 0 otherwise. Similarly, for matrix  $\mathbf{\Delta}$  and group  $g$ , we denote  $\mathbf{\Delta}_g$  as the matrix whose  $i$ th column is  $\mathbf{\Delta}_i$  if  $i \in g$  or  $\mathbf{0}$  otherwise.

## 3.2 Unified Robust Framework

This section presents the main result of this chapter – there exists a strong relationship between robust linear regression and several widely applied variants of Lasso.

### 3.2.1 Preliminary

We start by briefly review the result from [XCM10] that connects standard Lasso with robust regression. Robust linear regression considers the case that the observed data is corrupted by some (potentially malicious) disturbance. To protect against such disturbance, the following

min-max formulation is typically solved:

$$\min_{\beta \in \mathcal{R}^m} \{ \max_{\Delta \in U} \|\mathbf{y} - (\mathbf{X} + \Delta)\beta\|_p \}, \quad (3.1)$$

where  $U$  is the uncertainty set, or the set of admissible disturbances of the observed matrix  $\mathbf{X}$ . [XCM10] showed that the robust optimization above is equivalent to the  $\ell_1$ -norm regularized linear regression (standard Lasso) when the uncertainty set is defined by *feature wise* norm constraints:

**Theorem 3.1** ([XCM10]). *The robust regression problem (3.1) with the uncertainty set*

$$U = \{(\boldsymbol{\delta}_1, \dots, \boldsymbol{\delta}_m) \mid \|\boldsymbol{\delta}_i\|_2 \leq c_i, i = 1, \dots, m\},$$

for given  $c_i \geq 0$ , is equivalent to the following  $\ell_1$ -norm regularized regression problem:

$$\min_{\beta \in \mathcal{R}^m} \{ \|\mathbf{y} - \mathbf{X}\beta\|_2 + \sum_{i=1}^m c_i |\beta_i| \}.$$

It turns out Theorem 3.1 not only provides a new insight of Lasso from a robustness perspective, but is also a powerful tool to analyze the sparsity and consistency of Lasso, see [XCM10] for details.

### 3.2.2 Main Results

Given the success of the robust interpretation of Lasso, it is natural to ask whether different Lasso-like formulations such as the group Lasso or the fused Lasso can also be reformulated as robust linear regression problems by selecting appropriate uncertainty sets. We provide in this section an affirmative answer. To illustrate our general result, we first consider the overlapping group Lasso proposed in [YL06]. The following theorem shows that it is equivalent to a robust linear regression problem:

**Theorem 3.2.** *Let the uncertainty set be*

$$U = \{ \boldsymbol{\Delta}^{(1)} + \dots + \boldsymbol{\Delta}^{(t)} \mid \|\boldsymbol{\Delta}_{g_i}^{(i)}\|_2 \leq c_{g_i} \text{ and } \|\boldsymbol{\Delta}_{g_i^c}^{(i)}\|_2 = 0, \forall i \in [t] \}, \quad (3.2)$$



where matrix  $\Delta^{(i)} \in \mathcal{R}^{n \times m}$ ,  $\bigcup_{i=1}^t g_i = [m]$  and  $g_i^c = [m] \setminus g_i$ , then the robust regression (3.1) with  $U$  is equivalent to

$$\min_{\beta \in \mathcal{R}^m} \{ \|\mathbf{y} - \mathbf{X}\beta\|_2 + \sum_{i=1}^t c_{g_i} \|\beta_{g_i}\|_2 \}. \quad (3.3)$$

The regression formulations we consider slightly differ from the more widely used ones, as we minimize the norm of the error, rather than the squared norm. It is known that these two coincide up to a change of the regularization coefficient since the empirical error terms and the regularization terms we discuss are all convex.

Note that the groups defined in Theorem 3.2 are allowed to overlap. Theorem 3.2 shows that the group Lasso formulation is equivalent to the robust linear regression where the admissible disturbance is given by the norm constraints on each group  $g_i$ , as opposed to constraints on each feature in Theorem 3.1. Observe that by taking each feature as one group, Theorem 3.2 immediately implies Theorem 3.1.

We now present our main result that connects variants of Lasso-like algorithms with the robust linear regression framework. Consider the following uncertainty set:

$$U = \{ \Delta^{(1)} \mathbf{W}_1 + \dots + \Delta^{(t)} \mathbf{W}_t \mid \forall i \in [t], \forall g \in G_i, \|\Delta_g^{(i)}\|_p \leq c_g \}, \quad (3.4)$$

where matrix  $\mathbf{W}_i \in \mathcal{R}^{m \times m}$  is fixed,  $G_i$  is the set of the groups, and  $c_g$  provides the norm bound of group  $g$  of the disturbance. Notice that  $G_i$  may contain more than one groups, and two different groups  $g_1, g_2 \in G_i$  are allowed to overlap, i.e.,  $g_1 \cap g_2 \neq \emptyset$ . It is easy to see that such set contains the uncertainty set considered in Theorem 3.2 as a special case, i.e.  $G_i = \{g_i, g_i^c\}$  for  $i \in [t]$ . The next theorem shows that such uncertainty set provides a unified framework that “encodes” the ridge regression and many variants of Lasso-like algorithms.

**Theorem 3.3.** *The robust regression problem (3.1) with the uncertainty set (3.4) is equivalent to the convex regularized linear regression problem:*

$$\min_{\beta \in \mathcal{R}^m} \{ \|\mathbf{y} - \mathbf{X}\beta\|_p + \sum_{i=1}^t \max_{\forall g \in G_i, \|\alpha_g^{(i)}\|_p \leq c_g} \alpha^{(i)\top} \mathbf{W}_i \beta \}. \quad (3.5)$$

*Proof.* For any fixed  $\beta$ , we have

$$\begin{aligned} \max_{\Delta \in U} \|\mathbf{y} - (\mathbf{X} + \Delta)\beta\|_p &= \max_{\Delta \in U} \|\mathbf{y} - \mathbf{X}\beta - \sum_{i=1}^t \Delta^{(i)} \mathbf{W}_i \beta\|_p \\ &\leq \|\mathbf{y} - \mathbf{X}\beta\|_p + \max_{\Delta \in U} \sum_{i=1}^t \left\| \sum_{j=1}^m (\mathbf{W}_i \beta)_j \Delta_j^{(i)} \right\|_p \\ &\leq \|\mathbf{y} - \mathbf{X}\beta\|_p + \max_{\Delta \in U} \sum_{i=1}^t \sum_{j=1}^m |(\mathbf{W}_i \beta)_j| \|\Delta_j^{(i)}\|_p. \end{aligned}$$

For clarity, denote

$$\alpha^{(i)} \equiv [\text{sign}((\mathbf{W}_i \beta)_1) \cdot \|\Delta_1^{(i)}\|_p, \dots, \text{sign}((\mathbf{W}_i \beta)_m) \cdot \|\Delta_m^{(i)}\|_p]^\top.$$

From the definition of the uncertainty set  $U$ , we know that  $\|\Delta_g^{(i)}\|_p \leq c_g$  for any  $i \in [t]$  and  $g \in G_i$ . Thus,  $\|\alpha_g^{(i)}\|_p = \|\Delta_g^{(i)}\|_p \leq c_g$ , and we have

$$\begin{aligned} \|\mathbf{y} - \mathbf{X}\beta\|_p + \max_{\Delta \in U} \sum_{i=1}^t \sum_{j=1}^m |(\mathbf{W}_i \beta)_j| \|\Delta_j^{(i)}\|_p &= \|\mathbf{y} - \mathbf{X}\beta\|_p + \max_{\Delta \in U} \sum_{i=1}^t \alpha^{(i)\top} \mathbf{W}_i \beta \\ &\leq \|\mathbf{y} - \mathbf{X}\beta\|_p + \sum_{i=1}^t \max_{\forall g \in G_i, \|\alpha_g^{(i)}\|_p \leq c_g} \alpha^{(i)\top} \mathbf{W}_i \beta. \end{aligned}$$

On the other hand, let

$$\alpha_0^{(i)} = \arg \max_{\forall g \in G_i, \|\alpha_g^{(i)}\|_p \leq c_g} \alpha^{(i)\top} \mathbf{W}_i \beta$$

and

$$\mathbf{u} = \begin{cases} \frac{\mathbf{y} - \mathbf{X}\beta}{\|\mathbf{y} - \mathbf{X}\beta\|_p} & \text{if } \|\mathbf{y} - \mathbf{X}\beta\|_p \neq 0 \\ \text{any vector with unit } \ell_p \text{ norm} & \text{otherwise} \end{cases}$$

and then let

$$\Delta^{(i)} = -\mathbf{u} \cdot \alpha_0^{(i)\top}$$

From the definition above, we know that  $\|\Delta_g^{(i)}\|_p = \|\alpha_g^{(i)}\|_p \leq c_g$ . Thus, we have

$$\begin{aligned} \max_{\Delta \in U} \|\mathbf{y} - (\mathbf{X} + \Delta)\boldsymbol{\beta}\|_p &\geq \|\mathbf{y} - (\mathbf{X} + \sum_{i=1}^t \Delta^{(i)} \mathbf{W}_i)\boldsymbol{\beta}\|_p \\ &= \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \sum_{i=1}^t \alpha_0^{(i)\top} \mathbf{W}_i \boldsymbol{\beta}\|_p \\ &= \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p + \sum_{i=1}^t \max_{\forall g \in G_i, \|\alpha_g^{(i)}\|_p \leq c_g} \alpha^{(i)\top} \mathbf{W}_i \boldsymbol{\beta}, \end{aligned}$$

which establishes the theorem.  $\square$

Indeed, the regularized linear regression (3.5) is a generalization for Lasso. By setting  $t$ ,  $G_i$ ,  $\mathbf{W}_i$  and  $c_g$  to appropriate values, (3.5) can be reduced as *standard Lasso*, *group Lasso*, *fused Lasso*, *trend filtering*, among others.

**Corollary 3.1** (Ridge Regression). *Suppose that  $t = 1$ ,  $p = 2$ ,  $\mathbf{W}_1 = \mathbf{I}$ ,  $G_1 = \{[m]\}$  and  $c_g = c$ , then the robust regression problem (3.1) is equivalent to*

$$\min_{\boldsymbol{\beta} \in \mathcal{R}^m} \{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 + c\|\boldsymbol{\beta}\|_2\}. \quad (3.6)$$

Ridge regression has been well studied. It shrinks the regression coefficients  $\beta_1, \dots, \beta_m$  by penalizing their sizes (in terms of  $\ell_2$ -norm) to control the complexity of the regression model.

**Corollary 3.2** (Standard Lasso). *Suppose that  $t = 1$ ,  $\mathbf{W}_1 = \mathbf{I}$ ,  $G_1 = \{\{1\}, \dots, \{m\}\}$  and  $c_i = c_{\{i\}}$ , then the robust regression problem (3.1) is equivalent to*

$$\min_{\boldsymbol{\beta} \in \mathcal{R}^m} \{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p + \sum_{i=1}^m c_i |\beta_i|\}. \quad (3.7)$$

The main difference between the ridge regression and the standard Lasso is that the Lasso penalizes the  $\ell_1$ -norm of the coefficients. The Lasso's ability to recover sparse solutions has been extensively explored, and has found wide applications in statistics, signal processing, computer vision, bioinformatics, to name a few.

**Corollary 3.3** (Non-overlapping Group Lasso). *Suppose that  $t = 1$ ,  $\mathbf{W}_1 = \mathbf{I}$ ,  $G_1 =$*

$\{g_1, \dots, g_k\}$  and  $g_i \cap g_j = \emptyset$  for any  $i \neq j$ , then the robust regression problem (3.1) is equivalent to

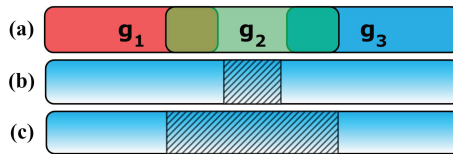
$$\min_{\beta \in \mathcal{R}^m} \{ \|\mathbf{y} - \mathbf{X}\beta\|_p + \sum_{i=1}^k c_{g_i} \|\beta_{g_i}\|_p^* \}. \quad (3.8)$$

The non-overlapping group Lasso is an extension of the standard Lasso, where non-overlapping group structure of features is known as the prior information. In particular, features are partitioned into known groups, and one seeks solutions that select few non-zero *groups*. Different from Lasso, group Lasso does not encourage sparsity inside each group.

**Corollary 3.4** (Overlapping Group Lasso [JOV09]). *Suppose that  $t = 1$ ,  $\mathbf{W}_1 = \mathbf{I}$ ,  $G_1 = \{g_1, \dots, g_k\}$ , and  $\bigcup_{i=1}^k g_i = [m]$ , then the robust regression problem (3.1) is equivalent to*

$$\min_{\beta \in \mathcal{R}^m} \{ \|\mathbf{y} - \mathbf{X}\beta\|_p + \min_{\sum \mathbf{v}_{g_i} = \beta, \text{supp}(\mathbf{v}_{g_i}) \subseteq g_i} \sum_{i=1}^k c_{g_i} \|\mathbf{v}_{g_i}\|_p^* \}. \quad (3.9)$$

Different from the overlapping group Lasso formulation (3.3) proposed in [YL06] that encourages solutions whose supports are in the *complement of a union of groups* (i.e, many groups are all zero), Formulation (3.9) tends to select solutions whose support is contained in a union of potentially overlapping groups. This is motivated by applications in bioinformatics, e.g., predicting the class of a tumor from gene expression measurements with microarrays, and simultaneously select a few genes to establish a predictive signature. Figure 3.1 illustrates the difference between two group Lasso formulations.



**Figure 3.1:** Preferred solutions of the two group Lasso. Hatched regions indicates non-zero coefficients and unhatched regions indicates zero coefficients. (a) Predefined groups of the coefficient  $\beta$ ; (b) One solution that [YL06] tends to select; (c) One solution that [JOV09] tends to select.

**Corollary 3.5** (Fused Lasso [TSR<sup>+</sup>05]). *Suppose that  $t = 2$ ,  $G_1 = G_2 = \{\{1\}, \dots, \{m\}\}$ ,*

and

$$\mathbf{W}_1 = \mathbf{I}, \quad \mathbf{W}_2 = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & -1 \\ 0 & \cdots & 0 & 0 & 0 \end{pmatrix},$$

then the robust regression problem (3.1) is equivalent to

$$\min_{\boldsymbol{\beta} \in \mathcal{R}^m} \{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p + \sum_{i=1}^m c_i |\boldsymbol{\beta}_i| + \sum_{i=1}^{m-1} c'_i |\boldsymbol{\beta}_i - \boldsymbol{\beta}_{i+1}| \}, \quad (3.10)$$

where  $c_i$  and  $c'_i$  are the “ $c_{\{i\}}$ ”s corresponding to the uncertainty sets of  $G_1$  and  $G_2$ , respectively.

The fused Lasso is motivated by protein mass spectroscopy and gene expression profiling. After estimating an order of data and putting correlated data near one another, solving it not only encourages sparsity in the coefficients  $\beta_1, \dots, \beta_m$  but also encourages sparsity in their differences, which implies that it tends to select a sparse solution in which nearby coefficients are similar to each other.

**Corollary 3.6** (Sparse Group Lasso [FHT10]). *Suppose that  $t = k + 1$ ,  $\bigcup_{i=1}^k g_i = [m]$  and  $g_i^c = [m] \setminus g_i$ . Let  $\mathbf{W}_i = \mathbf{I}$ ,  $G_i = \{g_i, g_i^c\}$ ,  $c_{g_i^c} = 0$  for  $i \in [k]$ , and let  $\mathbf{W}_{k+1} = \mathbf{I}$ ,  $G_{k+1} = \{\{1\}, \dots, \{m\}\}$ , then the robust regression problem (3.1) is equivalent to*

$$\min_{\boldsymbol{\beta} \in \mathcal{R}^m} \{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p + \sum_{i=1}^k c_{g_i} \|\boldsymbol{\beta}_{g_i}\|_p^* + \sum_{i=1}^m c_i |\boldsymbol{\beta}_i| \} \quad (3.11)$$

where  $c_i$  is equal to  $c_{\{i\}}$ .

The sparse group Lasso blends the standard Lasso with the group Lasso, and encourages solutions that are sparse at both the group and the individual feature levels. Notice that Equation (3.11) is equivalent to the elastic net [ZH05] when  $k = 1$  and  $p = 2$ .

**Corollary 3.7** (Generalized Lasso [TT11]). *Suppose that  $t = 1$ ,  $\mathbf{W}_1 = \mathbf{D}$ ,  $G_1 = \{\{1\}, \dots, \{m\}\}$ ,*

and  $c_{\{i\}} = \lambda$ , then the robust regression problem (3.1) is equivalent to

$$\min_{\beta \in \mathcal{R}^m} \{ \|\mathbf{y} - \mathbf{X}\beta\|_p + \lambda \|\mathbf{D}\beta\|_1 \}. \quad (3.12)$$

By making various choices of  $\mathbf{D}$ , the generalized Lasso can be reformulated as well-known problems in the literature: *trend filtering* [KKBG09], etc.

**Remark.** While the inner maximization of the robust linear regression problem (3.1) over the uncertainty set (3.4) is non-convex, Theorem 3.3 shows that it can be solved efficiently as it is equivalent to a convex optimization problem (3.5). In particular, by strong duality, the optimization problem (3.5) is equivalent to

$$\begin{aligned} \min_{\beta, \mathbf{v}_g^{(i)}} \quad & \|\mathbf{y} - \mathbf{X}\beta\|_p + \sum_{i=1}^t \sum_{g \in G_i} c_g \|\mathbf{v}_g^{(i)}\|_p^* \\ \text{s.t.} \quad & \sum_{g \in G_i} \|\mathbf{v}_g^{(i)}\|_p^* = \mathbf{W}_i \beta, \quad \forall i \in [t] \\ & \mathbf{A}_g^{(i)} \mathbf{v}_g^{(i)} = 0, \quad \forall i \in [t], \forall g \in G_i, \end{aligned}$$

where  $\mathbf{v}_g^{(i)} \in \mathcal{R}^m$  is a decision variable and  $\mathbf{A}_g^{(i)} \in \mathcal{R}^{(m-|g|) \times m}$  is a constant matrix defined as  $\mathbf{A}_g^{(i)} = (\mathbf{e}_{i_1}, \dots, \mathbf{e}_{i_k})^\top$  where  $k = m - |g|$ ,  $\{i_1, \dots, i_k\} = g^c$ , and  $\mathbf{e}_i$  is the  $i^{\text{th}}$  unit base vector. This is a linear constrained convex optimization problem which can be solved efficiently using off-the-shelf methods. In addition, for special case such as the non-overlapping group Lasso, more scalable codes are available, e.g., [MGB08, RF08].

### Proofs of the corollaries

To prove the corollaries shown above, we require the following lemma.

**Lemma 3.1.** *If any two different groups  $g_p$  and  $g_q$  in  $G_i$  in the uncertainty set  $U$  (3.4) are non-overlapping for  $i = 1, \dots, t$ , which means  $g_p \cap g_q = \emptyset$ , then the optimization problem*

(3.5) is equivalent to

$$\min_{\boldsymbol{\beta} \in \mathcal{R}^m} \{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p + \sum_{i=1}^t \sum_{g \in G_i} c_g \|(\mathbf{W}_i \boldsymbol{\beta})_g\|_p^* \} \quad (3.13)$$

*Proof.* Since any two different groups  $g_p$  and  $g_q$  in  $G_i$  are non-overlapping, we have

$$\sum_{i=1}^t \max_{\forall g \in G_i, \|\boldsymbol{\alpha}_g^{(i)}\|_p \leq c_g} \boldsymbol{\alpha}^{(i)\top} \mathbf{W}_i \boldsymbol{\beta} = \sum_{i=1}^t \sum_{g \in G_i} \max_{\|\boldsymbol{\alpha}_g^{(i)}\|_p \leq c_g} \boldsymbol{\alpha}_g^{(i)\top} (\mathbf{W}_i \boldsymbol{\beta})_g = \sum_{i=1}^t \sum_{g \in G_i} c_g \|(\mathbf{W}_i \boldsymbol{\beta})_g\|_p^* \quad (3.14)$$

Hence the lemma holds.  $\square$

By using Theorem 3.3 and Lemma 3.1, we have

1. *Proof of Corollary 3.1:*  $G_1 = \{[m]\}$  satisfies the condition of Lemma 3.1, so we have

$$\sum_{i=1}^t \sum_{g \in G_i} c_g \|(\mathbf{W}_i \boldsymbol{\beta})_g\|_p^* = \sum_{g \in G_1} c \| \boldsymbol{\beta}_g \|_2^* = c \|\boldsymbol{\beta}\|_2. \quad (3.15)$$

2. *Proof of Corollary 3.2:*  $G_1 = \{\{1\}, \dots, \{m\}\}$  satisfies the condition of Lemma 3.1, then

$$\sum_{i=1}^t \sum_{g \in G_i} c_g \|(\mathbf{W}_i \boldsymbol{\beta})_g\|_p^* = \sum_{g \in G_1} c_g \| \boldsymbol{\beta}_g \|_p^* = \sum_{i=1}^m c_i |\beta_i|. \quad (3.16)$$

3. *Proof of Corollary 3.3:*  $G_1 = \{g_1, \dots, g_k\}$  satisfies the condition of Lemma 3.1, so we have

$$\sum_{i=1}^t \sum_{g \in G_i} c_g \|(\mathbf{W}_i \boldsymbol{\beta})_g\|_p^* = \sum_{i=1}^k c_{g_i} \| \boldsymbol{\beta}_{g_i} \|_p^*. \quad (3.17)$$

4. *Proof of Theorem 3.2:*  $G_i = \{g_i, g_i^c\}$  satisfies the condition of Lemma 3.1 and  $c_{g_i^c} = 0$ , so that

$$\sum_{i=1}^t \sum_{g \in G_i} c_g \|(\mathbf{W}_i \boldsymbol{\beta})_g\|_p^* = \sum_{i=1}^k (c_{g_i} \| \boldsymbol{\beta}_{g_i} \|_p^* + c_{g_i^c} \| \boldsymbol{\beta}_{g_i^c} \|_p^*) = \sum_{i=1}^k c_{g_i} \| \boldsymbol{\beta}_{g_i} \|_p^*. \quad (3.18)$$

5. *Proof of Corollary 3.4:* The dual problem of the optimization problem

$$\min_{\sum \mathbf{v}_{g_i} = \boldsymbol{\beta}, \text{supp}(\mathbf{v}_{g_i}) \subseteq g_i} \sum_{i=1}^k c_{g_i} \|\mathbf{v}_{g_i}\|_p^*$$

can be formulated as

$$\begin{aligned} & \max_{\boldsymbol{\alpha}} \min_{\forall i, \text{supp}(\mathbf{v}_{g_i}) \subseteq g_i} \left\{ \sum_{i=1}^k c_{g_i} \|\mathbf{v}_{g_i}\|_p^* - \boldsymbol{\alpha}^\top \sum_{i=1}^k \mathbf{v}_{g_i} + \boldsymbol{\alpha}^\top \boldsymbol{\beta} \right\} \\ &= \max_{\boldsymbol{\alpha}} \left\{ \boldsymbol{\alpha}^\top \boldsymbol{\beta} + \min_{\forall i, \text{supp}(\mathbf{v}_{g_i}) \subseteq g_i} \left\{ \sum_{i=1}^k c_{g_i} \|\mathbf{v}_{g_i}\|_p^* - \boldsymbol{\alpha}^\top \mathbf{v}_{g_i} \right\} \right\} \\ &= \max_{\boldsymbol{\alpha}} \left\{ \boldsymbol{\alpha}^\top \boldsymbol{\beta} - \max_{\forall i, \text{supp}(\mathbf{v}_{g_i}) \subseteq g_i} \left\{ \sum_{i=1}^k \boldsymbol{\alpha}^\top \mathbf{v}_{g_i} - c_{g_i} \|\mathbf{v}_{g_i}\|_p^* \right\} \right\} \\ &= \max_{\forall i, \|\boldsymbol{\alpha}_{g_i}\| \leq c_{g_i}} \boldsymbol{\alpha}^\top \boldsymbol{\beta} \end{aligned} \quad (3.19)$$

Since the constraints in the primal problem satisfy Slater's condition, the strong duality holds. From the duality and the condition in Corollary 3.4, we have

$$\begin{aligned} & \min_{\boldsymbol{\beta} \in \mathcal{R}^m} \left\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p + \sum_{i=1}^t \max_{\forall g \in G_i, \|\boldsymbol{\alpha}_g^{(i)}\|_p \leq c_g} \boldsymbol{\alpha}^{(i)\top} \mathbf{W}_i \boldsymbol{\beta} \right\} \\ &= \min_{\boldsymbol{\beta} \in \mathcal{R}^m} \left\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p + \max_{\forall g \in G_1, \|\boldsymbol{\alpha}_g\|_p \leq c_g} \boldsymbol{\alpha}^\top \boldsymbol{\beta} \right\} \\ &= \min_{\boldsymbol{\beta} \in \mathcal{R}^m} \left\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p + \min_{\sum \mathbf{v}_{g_i} = \boldsymbol{\beta}, \text{supp}(\mathbf{v}_{g_i}) \subseteq g_i} \sum_{i=1}^k c_{g_i} \|\mathbf{v}_{g_i}\|_p^* \right\}. \end{aligned} \quad (3.20)$$

6. *Proof of Corollary 3.5:* From Theorem 3.2 and Lemma 3.1, we have

$$\begin{aligned} \sum_{i=1}^t \sum_{g \in G_i} c_g \|(\mathbf{W}_i \boldsymbol{\beta})_g\|_p^* &= \sum_{g \in G_1} c_g \|\boldsymbol{\beta}_g\|_p^* + \sum_{g \in G_2} c'_g \|(\mathbf{W}_2 \boldsymbol{\beta})_g\|_p^* \\ &= \sum_{i=1}^m c_i |\beta_i| + \sum_{i=1}^{m-1} c'_i |\beta_i - \beta_{i+1}|. \end{aligned} \quad (3.21)$$

7. *Proof of Corollary 3.6:* By using the proofs of Corollary 1 and Corollary 3, we can obtain Corollary 3.6.

8. *Proof of Corollary 3.7:*  $G_1 = \{\{1\}, \dots, \{m\}\}$  satisfies the condition of Lemma 3.1.



Since  $t = 1$ ,  $c_{\{i\}} = \lambda$  and  $\mathbf{W}_1 = \mathbf{D}$ , we have

$$\sum_{i=1}^t \sum_{g \in G_i} c_g \|(\mathbf{W}_i \boldsymbol{\beta})_g\|_p^* = \sum_{g \in G_1} \lambda \|(\mathbf{D} \boldsymbol{\beta})_g\|_p^* = \sum_{i=1}^m \lambda |(\mathbf{D} \boldsymbol{\beta})_i| = \lambda \|\mathbf{D} \boldsymbol{\beta}\|_1. \quad (3.22)$$

### 3.3 General Uncertainty Sets

As discussed above, we assume that the disturbance of each *group* is bounded individually, then the robust linear regression (3.1) can be reformulated as the regularized linear regression (3.5) which is a generalized formulation for Lasso-like algorithms. In this section, we provide a more generalized formulation of the uncertainty set.

Consider the following uncertainty set  $\hat{U}$ :

$$\hat{U} = \{\boldsymbol{\Delta}^{(1)} \mathbf{W}_1 + \cdots + \boldsymbol{\Delta}^{(t)} \mathbf{W}_t \mid \mathbf{c} \in Z; \forall i \in [t], \forall g \in G_i, \|\boldsymbol{\Delta}_g^{(i)}\|_p \leq c_g\}, \quad (3.23)$$

where  $G_i$  is the set of groups of disturbance  $\boldsymbol{\Delta}^{(i)}$ ,  $\mathbf{c}$  is the vector whose elements are the norm bounds  $c_g$  of all the groups contained in  $G_1, \dots, G_t$ , e.g.  $\mathbf{c} = (c_{g_1}, \dots, c_{g_n})$ , and  $Z$  is the feasible set of  $\mathbf{c}$ . If  $Z$  has only one element, then  $\hat{U}$  is equivalent to the uncertainty set  $U$  which is defined as (3.4) where  $c_g$  is fixed. Hence, the set  $\hat{U}$  is a very general formulation, and provides us with significant flexibility in designing uncertainty sets and equivalently new regression algorithms. In particular, we consider  $Z$  given by a set of convex constraints, i.e.,

$$Z = \{\mathbf{z} \in \mathcal{R}^k \mid f_i(\mathbf{z}) \leq 0, \forall i \in [q]; \mathbf{z} \geq \mathbf{0}\}, \quad (3.24)$$

where each  $f_i(\mathbf{z})$  is a convex function and  $k = \sum_{i=1}^t |G_i|$  ( $|G_i|$  is the cardinality of  $G_i$ ), and  $Z$  has non-empty relative interior.

Under these assumptions, we have the following theorem showing that the robust regression problem (3.1) with uncertainty set  $\hat{U}$  can be converted to a tractable convex optimization problem.

**Theorem 3.4.** *The robust regression problem with the uncertainty set (3.23)*

$$\min_{\beta \in \mathcal{R}^m} \{ \max_{\Delta \in \hat{U}} \|\mathbf{y} - (\mathbf{X} + \Delta)\beta\|_p \}$$

is equivalent to

$$\min_{\lambda \in \mathcal{R}_+^q, \kappa \in \mathcal{R}_+^k, \beta \in \mathcal{R}^m} \{ \|\mathbf{y} - \mathbf{X}\beta\|_p + v(\lambda, \kappa, \beta) \} \quad (3.25)$$

where

$$v(\lambda, \kappa, \beta) = \max_{\mathbf{c} \in \mathcal{R}^k} \left\{ \sum_{i=1}^t \max_{\forall g \in G_i, \|\alpha_g^{(i)}\|_p \leq c_g} \alpha^{(i)\top} \mathbf{W}_i \beta + \kappa^\top \mathbf{c} - \sum_{i=1}^q \lambda_i f_i(\mathbf{c}) \right\}.$$

Furthermore, the equivalent optimization problem (3.25) is convex and tractable.

*Proof.* From the definition of  $\hat{U}$ , we have

$$\begin{aligned} & \max_{\Delta \in \hat{U}} \|\mathbf{y} - (\mathbf{X} + \Delta)\beta\|_p \\ &= \max_{\mathbf{c} \in \mathcal{Z}} \max_{\forall i, \forall g \in G_i, \|\Delta_g^{(i)}\|_p \leq c_g} \|\mathbf{y} - (\mathbf{X} + \Delta)\beta\|_p \\ &= \|\mathbf{y} - \mathbf{X}\beta\|_p + \max_{\mathbf{c} \in \mathcal{Z}} \sum_{i=1}^t \max_{\forall g \in G_i, \|\alpha_g^{(i)}\|_p \leq c_g} \alpha^{(i)\top} \mathbf{W}_i \beta \\ &= \|\mathbf{y} - \mathbf{X}\beta\|_p + \max_{\mathbf{c} | \mathbf{c} \geq 0; f_i(\mathbf{c}) \leq 0} \sum_{i=1}^t \max_{\forall g \in G_i, \|\alpha_g^{(i)}\|_p \leq c_g} \alpha^{(i)\top} \mathbf{W}_i \beta \\ &= \|\mathbf{y} - \mathbf{X}\beta\|_p + \min_{\lambda \in \mathcal{R}_+^q, \kappa \in \mathcal{R}_+^k} \max_{\mathbf{c} \in \mathcal{R}^k} \left\{ \sum_{i=1}^t \max_{\forall g \in G_i, \|\alpha_g^{(i)}\|_p \leq c_g} \alpha^{(i)\top} \mathbf{W}_i \beta + \kappa^\top \mathbf{c} - \sum_{i=1}^q \lambda_i f_i(\mathbf{c}) \right\} \\ &= \|\mathbf{y} - \mathbf{X}\beta\|_p + \min_{\lambda \in \mathcal{R}_+^q, \kappa \in \mathcal{R}_+^k} v(\lambda, \kappa, \beta) \end{aligned} \quad (3.26)$$

Hence we establish the theorem by taking minimum over  $\beta$  on both sides. Now we show the optimization problem is convex and tractable. we first prove that  $v(\lambda, \kappa, \beta)$  is a convex

function of  $\boldsymbol{\lambda}, \boldsymbol{\kappa}, \boldsymbol{\beta}$ . Since

$$\begin{aligned}
v(\boldsymbol{\lambda}, \boldsymbol{\kappa}, \boldsymbol{\beta}) &= \max_{\substack{\mathbf{c} \in \mathcal{R}^k, \\ \forall i, g \in G_i, \|\boldsymbol{\alpha}_g^{(i)}\|_p \leq c_g}} \left\{ \sum_{i=1}^t \boldsymbol{\alpha}^{(i)\top} \mathbf{W}_i \boldsymbol{\beta} + \boldsymbol{\kappa}^\top \mathbf{c} - \sum_{i=1}^q \lambda_i f_i(\mathbf{c}) \right\} \\
&= \max_{\substack{\mathbf{c} \in \mathcal{R}^k, \\ \forall i, g \in G_i, \|\boldsymbol{\alpha}_g^{(i)}\|_p \leq c_g}} \mu(\boldsymbol{\lambda}, \boldsymbol{\kappa}, \boldsymbol{\beta}).
\end{aligned} \tag{3.27}$$

For fixed  $\mathbf{c}$  and  $\boldsymbol{\alpha}_g^{(i)}$ ,  $\mu(\boldsymbol{\lambda}, \boldsymbol{\kappa}, \boldsymbol{\beta})$  is a linear function of  $\boldsymbol{\lambda}, \boldsymbol{\kappa}, \boldsymbol{\beta}$ . Thus  $v(\boldsymbol{\lambda}, \boldsymbol{\kappa}, \boldsymbol{\beta})$  is convex, which implies the optimization problem is convex. By choosing parameter  $\gamma$ , the optimization problem can be reformulated as

$$\begin{aligned}
\min \quad & \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p \\
\text{s.t.} \quad & v(\boldsymbol{\lambda}, \boldsymbol{\kappa}, \boldsymbol{\beta}) \leq \gamma \\
& \boldsymbol{\lambda} \in \mathcal{R}_+^p, \boldsymbol{\kappa} \in \mathcal{R}_+^k, \boldsymbol{\beta} \in \mathcal{R}^m
\end{aligned}$$

To show the problem is tractable, it suffices to construct a polynomial-time *separation oracle* for the feasible set  $S$  (Grötschel et al. [GLS88]). A separation oracle is a routine such that for a solution  $(\boldsymbol{\lambda}_0, \boldsymbol{\kappa}_0, \boldsymbol{\beta}_0)$ , it can find, in polynomial time, that (a) whether  $(\boldsymbol{\lambda}_0, \boldsymbol{\kappa}_0, \boldsymbol{\beta}_0)$  belongs to  $S$  or not; and (b) if  $(\boldsymbol{\lambda}_0, \boldsymbol{\kappa}_0, \boldsymbol{\beta}_0) \notin S$ , a hyperplane that separates  $(\boldsymbol{\lambda}_0, \boldsymbol{\kappa}_0, \boldsymbol{\beta}_0)$  with  $S$ .

To verify the feasibility of  $(\boldsymbol{\lambda}_0, \boldsymbol{\kappa}_0, \boldsymbol{\beta}_0)$ , notice that  $(\boldsymbol{\lambda}_0, \boldsymbol{\kappa}_0, \boldsymbol{\beta}_0) \in S$  if and only if the optimal value of the optimization problem (3.27) is smaller than or equal to  $\gamma$ , which can be verified in polynomial time. If  $(\boldsymbol{\lambda}_0, \boldsymbol{\kappa}_0, \boldsymbol{\beta}_0) \notin S$ , then by solving (3.27), we can find in polynomial time  $\mathbf{c}_0, \boldsymbol{\alpha}_0^{(i)}$  such that

$$\sum_{i=1}^t \boldsymbol{\alpha}_0^{(i)\top} \mathbf{W}_i \boldsymbol{\beta}_0 + \boldsymbol{\kappa}_0^\top \mathbf{c}_0 - \sum_{i=1}^q \lambda_i f_i(\mathbf{c}_0) > \gamma.$$

which is the hyperplane separates  $(\boldsymbol{\lambda}_0, \boldsymbol{\kappa}_0, \boldsymbol{\beta}_0)$  with  $S$ . □

One interesting implication of Theorem 3.4 is that by choosing “proper” uncertainty sets, we can simplify (3.25) and obtain new regularized linear regression formulations. We provide

some examples to illustrate this in the rest of this section. The notations used follow those in Theorem 3.3.

**Corollary 3.8.** *Suppose that the uncertainty set  $\hat{U} = \{\Delta | \exists \mathbf{c} \in \mathcal{R}^m \text{ such that } \mathbf{c} \geq 0 \text{ and } \|\mathbf{c}_{g_i}\|_q^* \leq s_i, \forall i \in [k]; \|\Delta_j\| \leq c_j, \forall j \in [m]\}$ , then the equivalent linear regularized regression problem is*

$$\min_{\beta \in \mathcal{R}^m} \{ \|\mathbf{y} - \mathbf{X}\beta\|_p + \sum_{i=1}^k s_i \|\beta_{g_i}\|_q \},$$

where  $\|\cdot\|_q^*$  is the dual norm of  $\|\cdot\|_q$ ,  $\bigcup_{i=1}^k g_i = [m]$ , and  $g_i \cap g_j = \emptyset$  for  $i \neq j$ .

*Proof.* From Theorem 3.3 and Theorem 3.4, we have

$$\min_{\lambda \in \mathcal{R}_+, \kappa \in \mathcal{R}_+^m} v(\lambda, \kappa, \beta) = \min_{\lambda \in \mathcal{R}_+, \kappa \in \mathcal{R}_+^m} \max_{\mathbf{c} \in \mathcal{R}^m} \left\{ \sum_{i=1}^m (\kappa_i + |\beta_i|) c_i - \sum_{i=1}^k \lambda_i (\|\mathbf{c}_{g_i}\|_q^* + s_i) \right\}.$$

Define  $\mathbf{r}_{g_i}$  as the vector whose  $j^{\text{th}}$  elements is  $\kappa_j + |\beta_j|$  for all  $j \in g_i$ , then the equation above is equivalent to

$$\min_{\lambda \in \mathcal{R}_+, \kappa \in \mathcal{R}_+^m} \|\mathbf{r}_{g_i}\|_q \leq \lambda_i, \forall i \in [k] \quad \lambda^\top \mathbf{s} = \sum_{i=1}^k s_i \|\beta_{g_i}\|_q,$$

which establishes the corollary.  $\square$

This corollary interprets arbitrary norm-based regularizers for the non-overlapping group Lasso from a robust regression perspective. By choosing different norms that bound  $\mathbf{c}_{g_i}$  for  $i \in [k]$ , different regularization terms are obtained, which implies that the effect of the regularization term of Lasso is selecting a proper uncertainty set of the observed matrix.

For the overlapping group Lasso [YL06], the same result holds by adding more disturbances to the overlapping columns of the observed matrix.

**Corollary 3.9.** *Let  $g_1, \dots, g_t$  be  $t$  groups such that  $\bigcup_{i=1}^t g_i = [m]$ , and  $\bar{\Delta}_i$  be a  $n \times m$  matrix whose columns except the  $i$ th one are all zero. Suppose that  $\mathbf{c}_{g_i}$  is a  $|g_i|$  dimension vector whose elements give the norm bound of  $\bar{\Delta}_j$  for  $j \in g_i$ , e.g.  $\|\bar{\Delta}_j\|_2 \leq c_{g_i}^j$ , and  $\mathbf{c} = (\mathbf{c}_{g_1}, \dots, \mathbf{c}_{g_t})$ . We define the uncertainty set as  $\hat{U} = \{\sum_{i=1}^t \sum_{j \in g_i} \bar{\Delta}_j | \exists \mathbf{c} \text{ such that } \mathbf{c} \geq 0 \text{ and } \|\mathbf{c}_{g_i}\|_q^* \leq s_i, \forall i \in [t]; \|\bar{\Delta}_j\|_2 \leq c_{g_i}^j, \forall i \in [t], \forall j \in g_i\}$ , then the equivalent linear regular-*

ized regression problem is

$$\min_{\boldsymbol{\beta} \in \mathcal{R}^m} \{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p + \sum_{i=1}^t s_i \|\boldsymbol{\beta}_{g_i}\|_q \},$$

where  $\|\cdot\|_q^*$  is the dual norm of  $\|\cdot\|_q$ .

*Proof.* From Theorem 3.3 and Theorem 3.4, we have

$$\min_{\lambda \in \mathcal{R}_+, \boldsymbol{\kappa} \in \mathcal{R}_+^m} v(\lambda, \boldsymbol{\kappa}, \boldsymbol{\beta}) = \min_{\lambda \in \mathcal{R}_+, \boldsymbol{\kappa} \in \mathcal{R}_+^m} \max_{\mathbf{c} \in \mathcal{R}^m} \left\{ \sum_{j=1}^t \sum_{i \in g_j} (\kappa_i + |\beta_i|) c_i - \sum_{i=1}^t \lambda_i (\|\mathbf{c}_{g_i}\|_q^* + s_i) \right\}.$$

Define  $\mathbf{r}_{g_i}$  as the vector whose elements are  $\kappa_j + |\beta_j|$  for  $j \in g_i$ , then the equation above is equivalent to

$$\min_{\lambda \in \mathcal{R}_+, \boldsymbol{\kappa} \in \mathcal{R}_+^m} \left\{ \|\mathbf{r}_{g_i}\|_q \leq \lambda_i, \forall i \in [t] \right\} \boldsymbol{\lambda}^\top \mathbf{s} = \sum_{i=1}^t s_i \|\boldsymbol{\beta}_{g_i}\|_q,$$

which establishes the theorem.  $\square$

We now consider a polytope uncertainty set in which there exists an additional constraint bounding the total disturbance besides the norm bound for disturbance on each group.

**Corollary 3.10.** *Suppose that  $\hat{U} = \{\sum_{i=1}^t \boldsymbol{\Delta}^{(i)} \mid \exists 0 \leq \mathbf{c} \leq \mathbf{s} : \sum_{i=1}^t c_i/s_i \leq \theta; \|\boldsymbol{\Delta}_{g_i}^{(i)}\|_p \leq c_i, \|\boldsymbol{\Delta}_{g_i^c}^{(i)}\|_p = 0, \forall i \in [t]\}$ , then the equivalent linear regularized regression problem is as follows*

$$\begin{aligned} \min_{\boldsymbol{\beta}, \lambda} \quad & \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p + \sum_{i=1}^t [s_i \|\boldsymbol{\beta}_{g_i}\|_p^* - \lambda]_+ + \lambda\theta \\ \text{s.t.} \quad & \lambda \geq 0 \end{aligned} \tag{3.28}$$

where  $[x]_+ = \max\{x, 0\}$ .

*Proof.* From Theorem 3.2 and Theorem 3.4, we have

$$v(\boldsymbol{\lambda}, \boldsymbol{\kappa}, \boldsymbol{\beta}) = \max_{\mathbf{c}} \sum_{i=1}^t c_i \left( \|\boldsymbol{\beta}_{g_i}\|_p^* - \frac{\lambda}{s_i} \right) + (\boldsymbol{\kappa} - \bar{\boldsymbol{\lambda}})^\top \mathbf{c} + \bar{\boldsymbol{\lambda}}^\top \mathbf{s} + \lambda\theta.$$

Thus,  $v(\boldsymbol{\lambda}, \boldsymbol{\kappa}, \boldsymbol{\beta}) = \bar{\boldsymbol{\lambda}}^\top \mathbf{s} + \lambda\theta$  and  $\boldsymbol{\kappa}_i = \bar{\boldsymbol{\lambda}}_i + \lambda/s_i - \|\boldsymbol{\beta}_{g_i}\|_p^*$ ,  $\forall i \in [t]$ , which implies that the

robust regression is equivalent to

$$\begin{aligned} \min_{\boldsymbol{\beta}, \lambda} \quad & \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p + \bar{\boldsymbol{\lambda}}^\top \mathbf{s} + \lambda\theta \\ \text{s.t.} \quad & \|\boldsymbol{\beta}_{g_i}\|_p^* - \lambda/s_i \leq \bar{\boldsymbol{\lambda}}_i, \quad \forall i \in [t], \\ & \lambda \geq 0, \bar{\boldsymbol{\lambda}} \geq 0, \end{aligned}$$

which is also equivalent to (3.28).  $\square$

Notice that when  $\theta = t$ , the above formulation reduces to the overlapping group Lasso. On the other hand, when  $\theta = 0$ , it is equivalent to the linear least square problem. Hence, this formulation allows us to control the desired group sparsity level using only one parameter  $\theta$ .

### 3.4 Sparsity

The standard Lasso's ability to recover sparse solutions has been extensively studied [CDS99, FN03, CRT06, Tro04, Tro06], and the sparsity properties of the group Lasso have also been explored [HHM09, HZM09, Per11]. These results typically take one of two approaches – treating the problem from either a statistical or optimization perspective. In this section, we investigate the sparsity properties of the robust regression and equivalently non-overlapping/overlapping group Lasso from a robust optimization perspective, and provides a geometric interpretation for sparsity. We consider first the overlapping group Lasso.

**Theorem 3.5.** *For the overlapping group Lasso*

$$\min_{\boldsymbol{\beta} \in \mathcal{R}^m} \{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 + c \sum_{i=1}^t \|\boldsymbol{\beta}_{g_i}\|_2 \}$$

where  $\bigcup_{i=1}^t g_i = [m]$ , if there exists  $I \subset [t]$  such that for an orthonormal base  $\mathbf{V}$  of  $\text{span}(\{\mathbf{X}_j, j \in [m] \setminus \bigcup_{i \in I} g_i\} \cup \{\mathbf{y}\})$ , we have  $\|\mathbf{V}\mathbf{V}^\top \mathbf{X}_{g_i}\|_2 \leq c$  for  $i \in I$ , then any optimal solution  $\boldsymbol{\beta}^*$  satisfies that  $\boldsymbol{\beta}_{g_i}^* = \mathbf{0}$  for  $i \in I$ .

*Proof.* From Theorem 3.2, we know that the overlapping group Lasso is equivalent to

$$\min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta} \in U} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_2, \quad (3.29)$$

where the uncertainty set  $U$  is as follows

$$U = \left\{ \sum_{i=1}^t \boldsymbol{\Delta}^{(i)} \mid \forall i, \|\boldsymbol{\Delta}_{g_i}^{(i)}\|_2 \leq c \text{ and } \|\boldsymbol{\Delta}_{g_i^c}^{(i)}\|_2 = 0 \right\}.$$

Recall that it is allowed that  $g_i \cap g_j \neq \emptyset$  for  $i \neq j$ . We define group  $\hat{g}_i$  as

$$\hat{g}_i = \begin{cases} g_i & i \in I; \\ g_i - \bigcup_{j \in I} g_j & i \notin I, \end{cases}$$

and consider the following uncertainty set

$$\hat{U} = \left\{ \sum_{i=1}^t \boldsymbol{\Delta}^{(i)} \mid \forall i, \|\boldsymbol{\Delta}_{\hat{g}_i}^{(i)}\|_2 \leq c \text{ and } \|\boldsymbol{\Delta}_{\hat{g}_i^c}^{(i)}\|_2 = 0 \right\},$$

then we have  $\hat{U} \subseteq U$  since  $\hat{g}_i \subseteq g_i, \forall i \in [t]$ . Thus,

$$\min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta} \in \hat{U}} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_2 \leq \min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta} \in U} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_2. \quad (3.30)$$

Let  $\bar{\mathbf{X}}$  be a matrix whose  $i$ th column is

$$\bar{\mathbf{X}}_i = \begin{cases} \mathbf{X}_i & i \notin \bigcup_{j \in I} \hat{g}_j \\ \mathbf{X}_i - \mathbf{V}\mathbf{V}^\top \mathbf{X}_i & i \in \bigcup_{j \in I} \hat{g}_j, \end{cases} \quad (3.31)$$

then from the condition  $\|\mathbf{V}\mathbf{V}^\top \mathbf{X}_{\hat{g}_i}\|_2 \leq c$  for  $i \in I$ , we have  $\|(\mathbf{X} - \bar{\mathbf{X}})_{\hat{g}_i}\|_2 \leq c$  for  $i \in I$ .

Now let

$$\bar{U} = \left\{ \boldsymbol{\Delta}^{(1)} + \cdots + \boldsymbol{\Delta}^{(t)} \mid \|\boldsymbol{\Delta}_{\hat{g}_i}^{(i)}\|_2 \leq c \text{ and } \|\boldsymbol{\Delta}_{\hat{g}_i^c}^{(i)}\|_2 = 0 \text{ for } i \notin I; \|\boldsymbol{\Delta}^{(i)}\|_2 = 0 \text{ for } i \in I \right\},$$

and consider the following robust regression problem

$$\min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta} \in \bar{U}} \|\mathbf{y} - (\bar{\mathbf{X}} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_2,$$

which is equivalent to

$$\min_{\boldsymbol{\beta}} \{ \|\mathbf{y} - \bar{\mathbf{X}}\boldsymbol{\beta}\|_2 + c \sum_{i \notin I} \|\boldsymbol{\beta}_{\hat{g}_i}\|_2 \}. \quad (3.32)$$

We denote the optimal solution of (3.32) as  $\bar{\boldsymbol{\beta}}^*$ . From the definition of  $\bar{\mathbf{X}}$ , we know that each column of  $\bar{\mathbf{X}}_{\hat{g}_i}$  for  $i \in I$  is orthogonal to the span of  $\{\mathbf{X}_{\hat{g}_i}, i \notin I\} \cup \{\mathbf{y}\}$ . Hence by changing  $\bar{\boldsymbol{\beta}}_{\hat{g}_i}^*$  to 0 for all  $i \in I$ , the minimizing objective does not increase. This implies that the optimal solution  $\bar{\boldsymbol{\beta}}^*$  satisfies that  $\bar{\boldsymbol{\beta}}_{g_i}^* = 0$  for  $i \in I$ .

We now prove that  $\bar{\boldsymbol{\beta}}^*$  is also the optimal solution of the overlapping group Lasso. We first show that

$$\begin{aligned} \min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta} \in \bar{U}} \|\mathbf{y} - (\bar{\mathbf{X}} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_2 &\leq \min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta} \in \bar{U}} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_2 \\ &\leq \min_{\boldsymbol{\beta}} \max_{\boldsymbol{\Delta} \in \hat{U}} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_2. \end{aligned} \quad (3.33)$$

For any  $\mathbf{X}$ ,  $\bar{\mathbf{X}}$  (defined by (3.31)) and  $\bar{\boldsymbol{\Delta}} \in \bar{U}$  such that  $\|(\mathbf{X} - \bar{\mathbf{X}})_{\hat{g}_i}\|_p \leq c$  for  $i \in I$ , there exists  $\boldsymbol{\Delta} \in \hat{U}$  such that  $\bar{\mathbf{X}} + \bar{\boldsymbol{\Delta}} = \mathbf{X} + \boldsymbol{\Delta}$ , which implies  $\{\bar{\mathbf{X}} + \bar{\boldsymbol{\Delta}} | \bar{\boldsymbol{\Delta}} \in \bar{U}\} \subseteq \{\mathbf{X} + \boldsymbol{\Delta} | \boldsymbol{\Delta} \in \hat{U}\}$ . Thus, Inequality (3.33) holds. On the other hand, since  $\bar{\boldsymbol{\beta}}_{g_i}^* = 0$  for  $i \in I$ , we have

$$\begin{aligned} \max_{\boldsymbol{\Delta} \in \bar{U}} \|\mathbf{y} - (\bar{\mathbf{X}} + \boldsymbol{\Delta})\bar{\boldsymbol{\beta}}^*\|_2 &= \max_{\boldsymbol{\Delta} \in \bar{U}} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\bar{\boldsymbol{\beta}}^*\|_2 \\ &= \max_{\boldsymbol{\Delta} \in \hat{U}} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\bar{\boldsymbol{\beta}}^*\|_2, \end{aligned} \quad (3.34)$$

then for an arbitrary  $\boldsymbol{\beta}$ , the following inequality holds

$$\max_{\boldsymbol{\Delta} \in \bar{U}} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\bar{\boldsymbol{\beta}}^*\|_2 \leq \max_{\boldsymbol{\Delta} \in \hat{U}} \|\mathbf{y} - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_2,$$

which implies that  $\bar{\boldsymbol{\beta}}^*$  is the optimal solution of the overlapping group Lasso. Hence we establish the theorem.  $\square$

Theorem 3.5 gives a geometric interpretation of the sparsity properties of the overlapping group Lasso based on its robustness. Indeed, it shows that a set of *groups of features* all



receive zero weight if there exists an admissible perturbation of each group which makes their features orthogonal to the other ones. As a special case, if the groups are non-overlapping (i.e.,  $g_i \cap g_j = \emptyset$  for  $i \neq j$ ), we have the following theorem that shows the sparsity properties of the non-overlapping group Lasso.

**Corollary 3.11.** *If there exists  $I \subset [t]$  such that for an orthonormal base  $\mathbf{V}$  of  $\text{span}(\{\mathbf{X}_{g_j}, j \notin I\} \cup \{\mathbf{y}\})$ , we have  $\|\mathbf{V}\mathbf{V}^\top \mathbf{X}_{g_i}\|_2 \leq c$  for  $i \in I$ , then any optimal solution  $\beta^*$  of the non-overlapping group Lasso 3.8 satisfies that  $\beta_{g_i}^* = 0$  for  $i \in I$ .*

### 3.5 Consistency

In this section, we investigate the statistical properties of the regularized linear regression formulation (3.5), and show that it is asymptotically consistent by using the robust properties derived from its equivalence with the robust linear regression (3.1). The proofs of our results largely follow the same framework proposed in [XCM10]. The main idea of the proofs is as follows: We show that the robust optimization formulation (3.1) can be seen to be the maximum expected error with respect to a class of probability measures. This class includes a kernel density estimator, and using this, we can prove that the regularized linear regression is consistent. However, because the uncertainty set we consider is more complicated than the one investigated in [XCM10] (which corresponds to the standard Lasso), the construction of the class of probability measures is more involved.

Using the same notation, we define  $\bar{G}_i = \{g \in G_i | c_g \neq 0\}$  and assume that  $\bigcup_{i=1}^t \bar{G}_i = [m]$ , i.e., each feature is contained in at least one group to ensure that all features are regularized. We restrict our discussion to the case that  $\mathbf{W}_i = \mathbf{I}$  for  $i \in [t]$  and  $c_g$  for each group  $g$  equals either  $\sqrt{n}c_n$  ( $n$  is the number of the samples) or 0, and establish the statistical consistency of the regularized linear regression (3.5) from a distributional robustness argument. Let  $P$  be a probability measure with bounded support that generates i.i.d. samples  $(b_i, \mathbf{r}_i^\top)$ , and has a density  $f(\cdot)$ . Denote the set of the first  $n$  samples by  $S_n$  and define

$$\beta(c_n, S_n) = \arg \min_{\beta} \left\{ \sqrt{\frac{1}{n} \sum_{i=1}^n (b_i - \mathbf{r}_i^\top \beta)^2} + \sum_{i=1}^t \max_{\forall g \in \bar{G}_i, \|\alpha_g^{(i)}\|_2 \leq c_n} \alpha^{(i)\top} \beta \right\},$$

$$\boldsymbol{\beta}(P) = \arg \min_{\boldsymbol{\beta}} \left\{ \sqrt{\int_{b,\mathbf{r}} (b_i - \mathbf{r}_i^\top \boldsymbol{\beta})^2 dP(b, \mathbf{r})} \right\}.$$

Thus,  $\boldsymbol{\beta}(c_n, S_n)$  is the solution to the regularized linear regression (3.5) with the tradeoff parameter set to  $\sqrt{nc_n}$ , and  $\boldsymbol{\beta}(P)$  is the “true” optimal solution. We have the following consistency results.

**Theorem 3.6.** *Let  $\{c_n\}$  be such that  $c_n \downarrow 0$  and  $\lim_{n \rightarrow \infty} n(c_n)^{m+1} = \infty$ . Suppose there exists a constant  $H$  such that  $\|\boldsymbol{\beta}(c_n, S_n)\|_2 \leq H$  almost surely. Then*

$$\lim_{n \rightarrow \infty} \sqrt{\int_{b,\mathbf{r}} (b_i - \mathbf{r}_i^\top \boldsymbol{\beta}(c_n, S_n))^2 dP(b, \mathbf{r})} = \sqrt{\int_{b,\mathbf{r}} (b_i - \mathbf{r}_i^\top \boldsymbol{\beta}(P))^2 dP(b, \mathbf{r})},$$

*almost surely.*

Here is the sketch of the proof. We first show that the equivalent robust regression (3.1) over the training data is equal to the worst-case expected generalization error among a set of distributions. Then, we show that such set of distributions includes a kernel density estimator for the true (unknown) distribution of the samples. Finally, using the fact that the kernel density estimator converges to the true density function almost surely when  $c_n \downarrow 0$  and  $\lim_{n \rightarrow \infty} n(c_n)^{m+1} = \infty$ , we can prove the consistency.

In the first step of the above proof, the set of distributions is the union of classes of distributions corresponding to disturbance in hyper-rectangle Borel sets  $Z_1, \dots, Z_n$  centered at  $(b_i, \mathbf{r}_i^\top)$  with lengths depending on  $c_n$  and the constraints on the uncertainty set  $\boldsymbol{\Delta}$ . Since in [XCM10], only the constraint that the norm of each column of  $\boldsymbol{\Delta}$  is bounded is considered, such Borel sets can be easily constructed for the standard Lasso. In contrast, in this chapter, we consider the case where  $\boldsymbol{\Delta} = \sum_{i=1}^t \boldsymbol{\Delta}^{(i)}$  and the constraints are imposed on feature groups  $\boldsymbol{\Delta}_g^{(i)}$  for  $g \in G_i$ . Since two groups  $g_i$  and  $g_j$  may have overlapping elements, this case is much more general than [XCM10] and the construction of the Borel sets is more difficult. Yet, we can still show that such Borel sets can be constructed, and the kernel density estimator is included in the set of distributions formed by the constructed Borel sets.

Indeed, the assumption that  $\|\boldsymbol{\beta}(c_n, S_n)\|_2 \leq H$  in Theorem 3.6 can be removed, and the consistency result still holds. Notice that Theorem 3.6 implies that *standard Lasso, group*

*Lasso* and *sparse group Lasso* are all asymptotically consistent. Follow the same road map but with more involved analysis, one can show that that *fused Lasso* is also asymptotically consistent.

### Proof of consistency

Recall that the uncertainty set considered in this chapter is

$$U = \{\mathbf{\Delta}^{(1)}\mathbf{W}_1 + \cdots + \mathbf{\Delta}^{(t)}\mathbf{W}_t \mid \forall i, \forall g \in G_i, \|\mathbf{\Delta}_g^{(i)}\|_2 \leq c_g\} \quad (3.35)$$

where  $G_i$  is the set of the groups of  $\mathbf{\Delta}^{(i)}$  and  $c_g$  gives the bound of  $\mathbf{\Delta}_g^{(i)}$  for group  $g$ . We denote  $\bar{G}_i$  and  $\bar{G}_i^c$  as the set  $\{g \in G_i \mid c_g \neq 0\}$  and  $G_i - \bar{G}_i$ , respectively. In this theorem, we restrict our discussion to the case that  $\mathbf{W}_i = \mathbf{I}$  for  $i = 1, \dots, t$  and the bound  $c_g$  of  $\mathbf{\Delta}_g^{(i)}$  for each group  $g$  equals  $\sqrt{nc_n}$  or 0, so the uncertainty set can be rewritten as

$$U = \{\mathbf{\Delta}^{(1)} + \cdots + \mathbf{\Delta}^{(t)} \mid \forall i, \forall g \in \bar{G}_i, \|\mathbf{\Delta}_g^{(i)}\|_2 \leq \sqrt{nc_n}\} \quad (3.36)$$

Note that the constraint  $\|\mathbf{\Delta}\|_2 \leq \sqrt{nc_n}$  can be reformulated as the union of several element-wise constraints. Denote  $\mathcal{D} = \{\mathbf{D} \mid \sum_i \sum_j D_{ij}^2 = nc_n^2, D_{ij} \geq 0\}$  (we call an element  $\mathbf{D} \in \mathcal{D}$  *decomposition*), then we have

$$\{\mathbf{\Delta} \mid \|\mathbf{\Delta}\|_2 \leq \sqrt{nc_n}\} = \bigcup_{\mathbf{D} \in \mathcal{D}} \{\mathbf{\Delta} \mid \forall i, j, |\Delta_{ij}| \leq D_{ij}\}.$$

Similarly, the uncertainty set  $\{\mathbf{\Delta} \mid \|\mathbf{\Delta}_g\|_2 \leq \sqrt{nc_n}\}$  is equivalent to

$$\bigcup_{\mathbf{D} \in \mathcal{D}_g} \{\mathbf{\Delta} \mid \forall i, \forall j \in g, |\Delta_{ij}| \leq D_{ij}\},$$

where  $\mathcal{D}_g = \{\mathbf{D} \mid \sum_i \sum_{j \in g} D_{ij}^2 = nc_n^2, D_{ij} \geq 0\}$ . After the constraints of the uncertainty sets are decomposed into element-wise constraints, the set  $\{\mathbf{X} + \mathbf{\Delta}^{(1)} + \cdots + \mathbf{\Delta}^{(t)}\}$  can also be represented by an element-wise way. The notation is a little complicated so we first consider three simple cases:

- One uncertainty set  $\mathbf{\Delta}$  such that  $\|\mathbf{\Delta}\|_2 \leq c$ : for fixed  $\mathbf{D} \in \mathcal{D}$ , we have  $\{X_{ij} + \Delta_{ij}\} = [X_{ij} - D_{ij}, X_{ij} + D_{ij}]$ .
- Two uncertainty sets  $\mathbf{\Delta}^{(1)}$  and  $\mathbf{\Delta}^{(2)}$  such that  $\|\mathbf{\Delta}^{(1)}\|_2 \leq c$  and  $\|\mathbf{\Delta}^{(2)}\|_2 \leq c$ : for fixed  $\mathbf{D}^{(1)} \in \mathcal{D}$  and  $\mathbf{D}^{(2)} \in \mathcal{D}$ , we have  $\{X_{ij} + \Delta_{ij}^{(1)} + \Delta_{ij}^{(2)}\} = [X_{ij} - D_{ij}^{(1)} - D_{ij}^{(2)}, X_{ij} + D_{ij}^{(1)} + D_{ij}^{(2)}]$ .
- One uncertainty set  $\mathbf{\Delta}$  and two overlapping groups  $p$  and  $q$  such that  $\|\mathbf{\Delta}_p\|_2 \leq c$  and  $\|\mathbf{\Delta}_q\|_2 \leq c$ : for fixed  $\mathbf{P} \in \mathcal{D}_p$  and  $\mathbf{Q} \in \mathcal{D}_q$ , we have

$$\{X_{ij} + \Delta_{ij}\} = \begin{cases} [X_{ij} - P_{ij}, X_{ij} + P_{ij}] & j \in p, j \notin q \\ [X_{ij} - Q_{ij}, X_{ij} + Q_{ij}] & j \notin p, j \in q \\ [X_{ij} - \min\{P_{ij}, Q_{ij}\}, X_{ij} + \min\{P_{ij}, Q_{ij}\}] & j \in p, j \in q \end{cases}$$

Thus, if the decomposition  $\mathbf{D} \in \mathcal{D}_g$  for each  $\mathbf{\Delta}_g^{(i)}$  is fixed, we have  $\{X_{ij} + \Delta_{ij}^{(1)} + \dots + \Delta_{ij}^{(t)}\} = [X_{ij} - \gamma_{ij}, X_{ij} + \gamma_{ij}]$  where  $\gamma_{ij}$  is determined by the decomposition  $\mathbf{D}$ s. Since the number of the elements of  $\mathbf{\Delta}_g^{(i)}$  is less than or equal to  $mn$  ( $m$  is the feature dimension and  $n$  is the number of samples), there exists a decomposition  $\mathbf{D}$  for each  $\mathbf{\Delta}_g^{(i)}$  such that  $[X_{ij} - \frac{cn}{\sqrt{m}}, X_{ij} + \frac{cn}{\sqrt{m}}] \subseteq [X_{ij} - \gamma_{ij}, X_{ij} + \gamma_{ij}]$ . We now prove the theorem.

**Proposition 3.1.** [*XCM10*] Given a function  $h : \mathcal{R}^{m+1} \mapsto R$  and Borel sets  $Z_1, \dots, Z_n \subseteq \mathcal{R}^{m+1}$ , let

$$\mathcal{P}_n = \{\mu \in \mathcal{P} \mid \forall S \subseteq \{1, \dots, n\} : \mu(\bigcup_{i \in S} Z_i) \geq |S|/n\}.$$

The following holds

$$\frac{1}{n} \sum_{i=1}^n \sup_{(b_i, \mathbf{r}_i) \in Z_i} h(b_i, \mathbf{r}_i) = \sup_{\mu \in \mathcal{P}_n} \int_{\mathcal{R}^{m+1}} h(b_i, \mathbf{r}_i) d\mu(b_i, \mathbf{r}_i).$$

**Step 1:** Using the notation above, we first give the following corollary:

**Corollary 3.12.** *Given  $\mathbf{y} \in \mathcal{R}^n$ ,  $\mathbf{X} \in \mathcal{R}^{n \times m}$ , the following equation holds for any  $\boldsymbol{\beta} \in \mathcal{R}^m$ ,*

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 + \sqrt{\frac{n}{m}}c_n + \sum_{i=1}^t \max_{\forall g \in \bar{G}_i, \|\boldsymbol{\alpha}_g^{(i)}\|_2 \leq \sqrt{nc_n}} \boldsymbol{\alpha}^{(i)\top} \boldsymbol{\beta} = \sup_{\mu \in \hat{\mathcal{P}}(n)} \sqrt{n \int_{\mathcal{R}^{m+1}} (b' - \mathbf{r}'^\top \boldsymbol{\beta})^2 d\mu(b', \mathbf{r}')}$$
(3.37)

Here,

$$\hat{\mathcal{P}}(n) = \bigcup_{\mathcal{S} = \{\mathbf{D}_g^{(i)}\} | \mathbf{D}_g^{(i)} \in \mathcal{D}_g, \forall i, g \in \bar{G}_i} \mathcal{P}_n(\mathbf{X}, \mathcal{S}, \mathbf{y}, c_n)$$

$$\begin{aligned} \mathcal{P}_n(\mathbf{X}, \mathcal{S}, \mathbf{y}, c_n) &= \{\mu \in \mathcal{P} \mid Z_i = [y_i - \frac{c_n}{\sqrt{m}}, y_i + \frac{c_n}{\sqrt{m}}] \times \prod_{j=1}^m [X_{ij} - \gamma_{ij}, X_{ij} + \gamma_{ij}]; \\ &\quad \forall S \subseteq \{1, \dots, n\} : \mu(\bigcup_{i \in S} Z_i) \geq |S|/n\}, \end{aligned}$$

where  $\gamma_{ij}$  depends on the “decomposition” set  $\mathcal{S}$ .

*Proof.* The right hand side of Equation (3.37) is equal to

$$\sup_{\mathcal{S} = \{\mathbf{D}_g^{(i)}\} | \forall i, g \in \bar{G}_i, \mathbf{D}_g^{(i)} \in \mathcal{D}_g} \left\{ \sup_{\mu \in \mathcal{P}_n(\mathbf{X}, \mathcal{S}, \mathbf{y}, c_n)} \sqrt{n \int_{\mathcal{R}^{m+1}} (b' - \mathbf{r}'^\top \boldsymbol{\beta})^2 d\mu(b', \mathbf{r}')} \right\}.$$

From Theorem 3.2, we know that the left hand side is equal to

$$\begin{aligned} &\sup_{\forall i, g \in G_i, \|\boldsymbol{\delta}_y\|_2 \leq \sqrt{\frac{n}{m}}c_n, \|\boldsymbol{\Delta}_g^{(i)}\|_2 \leq \sqrt{nc_n}} \|\mathbf{y} + \boldsymbol{\delta}_y - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_2 \\ &= \sup_{\forall i, g \in G_i, \mathbf{D}_g^{(i)} \in \mathcal{D}_g} \left\{ \sup_{\|\boldsymbol{\delta}_y\|_2^2 \leq \frac{n}{m}c_n^2, |\boldsymbol{\Delta}_g^{(i)}| \leq \mathbf{D}_g^{(i)}} \|\mathbf{y} + \boldsymbol{\delta}_y - (\mathbf{X} + \boldsymbol{\Delta})\boldsymbol{\beta}\|_2 \right\} \\ &= \sup_{\forall i, g \in G_i, \mathbf{D}_g^{(i)} \in \mathcal{D}_g} \sqrt{\sum_{i=1}^n \sup_{(b_i, \mathbf{r}_i) \in [y_i - c_n/\sqrt{m}, y_i + c_n/\sqrt{m}] \times \prod_{j=1}^m [X_{ij} - \gamma_{ij}, X_{ij} + \gamma_{ij}]} (b_i - \mathbf{r}_i^\top \boldsymbol{\beta})}. \end{aligned}$$

Furthermore, applying Proposition 3.1 yields

$$\begin{aligned} &\sqrt{\sum_{i=1}^n \sup_{(b_i, \mathbf{r}_i) \in [y_i - c_n/\sqrt{m}, y_i + c_n/\sqrt{m}] \times \prod_{j=1}^m [X_{ij} - \gamma_{ij}, X_{ij} + \gamma_{ij}]} (b_i - \mathbf{r}_i^\top \boldsymbol{\beta})} \\ &= \sqrt{\sup_{\mu \in \mathcal{P}(\mathbf{X}, \mathcal{S}, \mathbf{y}, c_n)} n \int_{\mathcal{R}^{m+1}} (b' - \mathbf{r}'^\top \boldsymbol{\beta})^2 d\mu(b', \mathbf{r}')} \\ &= \sup_{\mu \in \mathcal{P}(\mathbf{X}, \mathcal{S}, \mathbf{y}, c_n)} \sqrt{n \int_{\mathcal{R}^{m+1}} (b' - \mathbf{r}'^\top \boldsymbol{\beta})^2 d\mu(b', \mathbf{r}')} \end{aligned}$$

which proves the corollary.  $\square$

**Step 2:** As [XCM10], we consider the following kernel estimator given samples  $(b_i, \mathbf{r}_i)_{i=1}^n$ ,

$$h_n(b, \mathbf{r}) = (nc^{m+1})^{-1} \sum_{i=1}^n K\left(\frac{b - b_i, \mathbf{r} - \mathbf{r}_i}{c}\right) \quad (3.38)$$

where  $K(\mathbf{x}) = I_{[-1,1]^{m+1}}(\mathbf{x})/2^{m+1}$ , and  $c = \frac{c_n}{\sqrt{m}}$ .

Observe that the estimated distribution above belongs to the set of distributions

$$\begin{aligned} \mathcal{P}_n(\mathbf{X}, \mathcal{S}, \mathbf{y}, c_n) &= \{\mu \in \mathcal{P} \mid Z_i = [y_i - \frac{c_n}{\sqrt{m}}, y_i + \frac{c_n}{\sqrt{m}}] \times \prod_{j=1}^m [X_{ij} - \gamma_{ij}, X_{ij} + \gamma_{ij}]; \\ &\quad \forall S \subseteq \{1, \dots, n\} : \mu(\bigcup_{i \in S} Z_i) \geq |S|/n\} \end{aligned}$$

and hence belongs to  $\hat{\mathcal{P}}(n) = \bigcup_{S=\{\mathbf{D}_g^{(i)}\} | \mathbf{D}_g^{(i)} \in \mathcal{D}_g, \forall i, g \in \bar{G}_i} \mathcal{P}_n(\mathbf{X}, \mathcal{S}, \mathbf{y}, c_n)$ .

**Step 3:** Combining the last two steps, and using the fact that  $\int_{b, \mathbf{r}} |h_n(b, \mathbf{r}) - h(b, \mathbf{r})| d(b, \mathbf{r})$  goes to zero almost surely when  $c \downarrow 0$  and  $nc^{m+1} \uparrow \infty$  or equivalently  $c_n \downarrow 0$  and  $nc_n^{m+1} \uparrow \infty$ .

Now we prove consistency of robust regression.

*Proof.* Let  $f(\cdot)$  be the true probability density function of the samples, and  $\hat{\mu}_n$  be the estimated distribution using Equation (3.38) given  $S_n$  and  $c_n$ , and denote its density function as  $f_n(\cdot)$ . The condition that  $\|\beta(c_n, S_n)\|_2 \leq H$  almost surely and  $P$  has a bounded support implies that there exists a universal constant  $C$  such that

$$\max_{b, \mathbf{r}} (b - \mathbf{r}^\top \beta(c_n, S_n))^2 \leq C$$

almost surely.

By Corollary 3.12 and  $\hat{\mu}_n \in \hat{P}(n)$ , we have

$$\begin{aligned}
& \sqrt{\int_{b,\mathbf{r}} (b - \mathbf{r}^\top \boldsymbol{\beta}(c_n, S_n))^2 d\hat{\mu}_n(b, \mathbf{r})} \\
& \leq \sup_{\mu \in \hat{P}(n)} \sqrt{\int_{b,\mathbf{r}} (b - \mathbf{r}^\top \boldsymbol{\beta}(c_n, S_n))^2 d\mu_n(b, \mathbf{r})} \\
& = \frac{\sqrt{n}}{n} \sqrt{\sum_{i=1}^n (b_i - \mathbf{r}_i^\top \boldsymbol{\beta}(c_n, S_n))^2 + \sum_{i=1}^t \max_{\forall g \in \bar{G}_i, \|\boldsymbol{\alpha}_g^{(i)}\|_2 \leq c_n} \boldsymbol{\alpha}^{(i)\top} \boldsymbol{\beta} + \frac{1}{\sqrt{m}} c_n} \\
& \leq \frac{\sqrt{n}}{n} \sqrt{\sum_{i=1}^n (b_i - \mathbf{r}_i^\top \boldsymbol{\beta}(P))^2 + \sum_{i=1}^t \max_{\forall g \in \bar{G}_i, \|\boldsymbol{\alpha}_g^{(i)}\|_2 \leq c_n} \boldsymbol{\alpha}^{(i)\top} \boldsymbol{\beta} + \frac{1}{\sqrt{m}} c_n}
\end{aligned}$$

Notice that,  $\sum_{i=1}^t \max_{\forall g \in \bar{G}_i, \|\boldsymbol{\alpha}_g^{(i)}\|_2 \leq c_n} \boldsymbol{\alpha}^{(i)\top} \boldsymbol{\beta} + \frac{1}{\sqrt{m}} c_n$  converges to 0 as  $c_n \downarrow 0$  almost surely, so the right-hand side converges to  $\sqrt{\int_{b,\mathbf{r}} (b - \mathbf{r}^\top \boldsymbol{\beta}(P))^2 dP(b, \mathbf{r})}$  as  $n \uparrow \infty$  and  $c_n \downarrow 0$  almost surely. Furthermore, we have

$$\begin{aligned}
& \int_{b,\mathbf{r}} (b - \mathbf{r}^\top \boldsymbol{\beta}(c_n, S_n))^2 dP(b, \mathbf{r}) \\
& \leq \int_{b,\mathbf{r}} (b - \mathbf{r}^\top \boldsymbol{\beta}(c_n, S_n))^2 d\hat{\mu}_n(b, \mathbf{r}) + \max_{b,\mathbf{r}} (b - \mathbf{r}^\top \boldsymbol{\beta}(c_n, S_n))^2 \cdot \int_{b,\mathbf{r}} |f_n(b, \mathbf{r}) - f(b, \mathbf{r})| d(b, \mathbf{r}) \\
& \leq \int_{b,\mathbf{r}} (b - \mathbf{r}^\top \boldsymbol{\beta}(c_n, S_n))^2 d\hat{\mu}_n(b, \mathbf{r}) + C \int_{b,\mathbf{r}} |f_n(b, \mathbf{r}) - f(b, \mathbf{r})| d(b, \mathbf{r}),
\end{aligned}$$

where the last inequality follows from the definition of  $C$ . Notice that  $\int_{b,\mathbf{r}} |f_n(b, \mathbf{r}) - f(b, \mathbf{r})| d(b, \mathbf{r})$  goes to zero almost surely when  $c_n \downarrow 0$  and  $nc_n^{m+1} \uparrow \infty$ . Hence the theorem follows.  $\square$

As mentioned above, the assumption that  $\|\boldsymbol{\beta}(c_n, S_n)\|_2 \leq H$  in Theorem 7 can be removed, then we have

**Theorem 3.7.** *Let  $\{c_n\}$  converge to zero sufficiently slowly. Then*

$$\lim_{n \rightarrow \infty} \sqrt{\int_{b,\mathbf{r}} (b_i - \mathbf{r}_i^\top \boldsymbol{\beta}(c_n, S_n))^2 dP(b, \mathbf{r})} = \sqrt{\int_{b,\mathbf{r}} (b_i - \mathbf{r}_i^\top \boldsymbol{\beta}(P))^2 dP(b, \mathbf{r})}$$

*almost surely.*

To prove this theorem, we establish the following lemma first.

**Lemma 3.2.** *Partition the support of  $P$  as  $V_1, \dots, V_T$  such that the  $l_\infty$  radius of each set*

is less than  $\frac{c_n}{\sqrt{m}}$ . If a distribution  $\mu$  satisfies

$$\mu(V_t) = \#\{(b_i, \mathbf{r}_i^\top) \in V_t\}/n; \quad t = 1, \dots, T, \quad (3.39)$$

then  $\mu \in \hat{P}(n)$ .

*Proof.* Let  $Z_i = [y_i - \frac{c_n}{\sqrt{m}}, y_i + \frac{c_n}{\sqrt{m}}] \times \prod_{j=1}^m [X_{ij} - \frac{c_n}{\sqrt{m}}, X_{ij} + \frac{c_n}{\sqrt{m}}]$ , recall that  $X_{ij}$  is the  $j$ th element of  $\mathbf{r}_i$ . Notice that the  $l_\infty$  radius of  $V_t$  is less than  $\frac{c_n}{\sqrt{m}}$ , we have

$$(b_i, \mathbf{r}_i^\top) \in V_t \Rightarrow V_t \subseteq Z_i.$$

Therefore, for any  $S \subseteq \{1, \dots, n\}$ , the following holds

$$\begin{aligned} & \mu\left(\bigcup_{i \in S} Z_i\right) \geq \mu\left(\bigcup V_t \mid \exists i \in S : (b_i, \mathbf{r}_i^\top) \in V_t\right) \\ &= \sum_{t \mid \exists i \in S : (b_i, \mathbf{r}_i^\top) \in V_t} \mu(V_t) = \sum_{t \mid \exists i \in S : (b_i, \mathbf{r}_i^\top) \in V_t} \#\{(b_i, \mathbf{r}_i^\top) \in V_t\}/n \geq |S|/n. \end{aligned}$$

Hence  $\mu \in P_n(\mathbf{X}, \mathcal{S}, \mathbf{y}, c_n)$  which implies  $\mu \in \hat{P}(n)$ .  $\square$

Partition the support of  $P$  into  $T$  subsets such that the  $l_\infty$  radius of each set is less than  $\frac{c_n}{\sqrt{m}}$ . Denote  $\tilde{\mathcal{P}}(n)$  as the set of probability measures satisfying Equation (3.39). Hence  $\tilde{\mathcal{P}}(n) \subseteq \hat{\mathcal{P}}(n)$  by Lemma 1. Further notice that there exists a universal constant  $K$  such that  $\|\boldsymbol{\beta}(c_n, S_n)\|_2 \leq K/c_n$  due to the fact that the square loss of the solution  $\boldsymbol{\beta} = 0$  is bounded by a constant only depends on the support of  $P$ . Thus, there exists a constant  $C$  such that  $\max_{b, \mathbf{r}} (b - \mathbf{r}^\top \boldsymbol{\beta}(c_n, S_n))^2 \leq C/c_n^2$ . Follow a similar argument as the proof of Theorem 6, we have

$$\begin{aligned} & \sup_{\mu \in \tilde{\mathcal{P}}(n)} \sqrt{\int_{b, \mathbf{r}} (b - \mathbf{r}^\top \boldsymbol{\beta}(c_n, S_n))^2 d\mu_n(b, \mathbf{r})} \\ & \leq \frac{\sqrt{n}}{n} \sqrt{\sum_{i=1}^n (b_i - \mathbf{r}_i^\top \boldsymbol{\beta}(P))^2} + \sum_{i=1}^t \max_{\forall g \in \tilde{G}_i, \|\boldsymbol{\alpha}_g^{(i)}\|_2 \leq c_n} \boldsymbol{\alpha}^{(i)\top} \boldsymbol{\beta} + \frac{1}{\sqrt{m}} c_n \end{aligned} \quad (3.40)$$



and

$$\begin{aligned}
& \int_{b, \mathbf{r}} (b - \mathbf{r}^\top \boldsymbol{\beta}(c_n, S_n))^2 dP(b, \mathbf{r}) \\
& \leq \inf_{\mu_n \in \tilde{\mathcal{P}}(n)} \left\{ \int_{b, \mathbf{r}} (b - \mathbf{r}^\top \boldsymbol{\beta}(c_n, S_n))^2 d\mu_n(b, \mathbf{r}) + \max_{b, \mathbf{r}} (b - \mathbf{r}^\top \boldsymbol{\beta}(c_n, S_n))^2 \int_{b, \mathbf{r}} |f_{\mu_n}(b, \mathbf{r}) - f(b, \mathbf{r})| d(b, \mathbf{r}) \right\} \\
& \leq \sup_{\mu_n \in \tilde{\mathcal{P}}(n)} \int_{b, \mathbf{r}} (b - \mathbf{r}^\top \boldsymbol{\beta}(c_n, S_n))^2 d\mu_n(b, \mathbf{r}) + 2C/c_n^2 \inf_{\mu_n \in \tilde{\mathcal{P}}(n)} \int_{b, \mathbf{r}} |f_{\mu_n}(b, \mathbf{r}) - f(b, \mathbf{r})| d(b, \mathbf{r}),
\end{aligned}$$

here  $f_\mu$  stands for the density function of a measure  $\mu$ . Notice that  $\tilde{\mathcal{P}}(n)$  is the set of distributions satisfying Equation (3.39), hence  $\inf_{\mu_n \in \tilde{\mathcal{P}}(n)} \int_{b, \mathbf{r}} |f_{\mu_n}(b, \mathbf{r}) - f(b, \mathbf{r})| d(b, \mathbf{r})$  is upper-bounded by  $\sum_{t=1}^T |P(V_t) - \#\{(b_i, \mathbf{r}_i^\top) \in V_t\}|/n$ , which goes to zero as  $n$  increases for any fixed  $c_n$ . Therefore,

$$2C/c_n^2 \inf_{\mu_n \in \tilde{\mathcal{P}}(n)} \int_{b, \mathbf{r}} |f_{\mu_n}(b, \mathbf{r}) - f(b, \mathbf{r})| d(b, \mathbf{r}) \rightarrow 0,$$

if  $c_n \downarrow 0$  sufficiently slow. Combining this with Inequality (3.40) proves the theorem.

## 3.6 Chapter Summary

In this chapter, we investigated a unified approach to explain the success of algorithms that encourage various sparse-like structures based on the concept of *robustness*. In particular, we considered robust linear regression where the perturbations are constrained with respect to *each group of features*, and show that this formulation is equivalent to a regularized linear regression framework that contains several widely used Lasso-like algorithms such as fused Lasso. This hence provides a robustness based interpretation of such algorithms. Moreover, we established sparsity property and statistical consistency of group Lasso from this robustness perspective. The main thrust of this work is to extend a classical result that relates standard Lasso with robust linear regression [XCM10] to learning problems with more general sparse-like structures. Achieving this makes it possible to understand these problems by analyzing the respective uncertainty sets, and will eventually enable us to design new algorithms to specific learning tasks that has superior performance than existing approaches.

# CHAPTER 4

## A Distributionally Robust Optimization Interpretation For Regularized SVMs

Distributionally robust optimization (DRO) is an effective framework for decision-making under uncertainty. It is topical in operation research but has not attracted much attention in the machine learning community yet. In this chapter, we present a unified framework using DRO for designing robust classification methods that generalize well to test data by handling uncertainties in training data. Indeed, we show that previous robust classification approaches tackling data uncertainty using robust optimization fits in the DRO framework. Based on this framework, we provide a DRO interpretation for regularized SVMs and propose new novel robust classification algorithms which are robust to feature corruption in test data.

### 4.1 Introduction

This chapter considers classification problems under uncertainties. The presence of uncertainty and noise in real world classification tasks is inevitable due to sampling errors, measurement errors, etc. Take face recognition as an example: face images submitted to a face recognition system may contain salt and pepper noise or illumination changes due to hardware failures or external disturbances [Bov05]. Similarly, in the optical character recognition problem, characters of the same label may have small variations due to translations, rotations or slanted versions [LJB<sup>+</sup>95]. For standard classification algorithms such as the Support Vector Machine (SVM) [CV95], the generalization performance

is guaranteed when training and testing samples (noisy or not) are drawn i.i.d., but not when training samples contain large outliers or testing samples are corrupted. Therefore, many studies have proposed classification algorithms aiming to handle uncertainties, and showed that better performance can be achieved than the classifiers that ignore the uncertainty [vdMCTW13, SBS06, BTBBN11, BGJ+04].

These works typically take one of two approaches. The first approach treats the problem from the robust optimization perspective. [XCM09] established a strong connection between robust optimization and regularized SVMs and provided a robustness interpretation for the success of regularized SVMs. [GR06] applied the robust optimization formulation to construct classifiers that are robust to deletion of features in test data. The second approach tackles data uncertainty using chance constraints. [SBS06] considered two cases: 1) the first and second moments of the uncertainty are known, and 2) the uncertainty follows a Gaussian distribution. They showed that SVMs with chance constraints under these two cases can be converted into a second order cone program. Beyond these two cases, [BTBBN11] considered some more sophisticated and hence less conservative chance constraint formulations and provided a convex second order cone program relaxation.

Recently, a new approach for handling uncertainty in test data is proposed in [vdMCTW13]. To develop predictors that generalize well to test data, the authors trained a classifier using infinitely many training data obtained from corrupting the existing finite training examples with a fixed noise distribution. Based on this idea, the robust predictors can be solved by minimizing the expected value of the loss function under the corruption distribution. However, there are two drawbacks of this approach: 1) the precise distribution of the noise is usually unknown; 2) one of the widely used cost functions for classification – hinge loss – is not considered.

A natural question hence emerges: does there exist a “universal” formulation that can unify all these approaches, overcome their shortcomings, and inspire new algorithms? In this chapter, we show that the *distributionally robust optimization (DRO)* formulation provides a positive answer. There are three advantages of applying the DRO formulation to handle uncertainties: 1) The robust optimization formulation [XCM09] and the “expected value”

formulation [vdMCTW13] turn to be two special cases of the DRO formulation by specializing one of the DRO’s “key points” – *ambiguity set* of distributions of uncertainties; 2) the hinge loss can be applied in the DRO formulation, which leads to a DRO interpretation of SVMs; 3) it encourages new robust classification algorithms by selecting different loss functions and ambiguity sets. Our robust classifiers are derived from this DRO perspective, i.e., via directly requesting robustness in the formulation, rather than designing new regularization terms to handle noises as in [HZH11], [MT13] and many others.

In particular, under the unifying umbrella of distributionally robust optimization, we make the following contributions. We first show that the DRO formulation with the hinge loss and the *mean absolute deviation* (MAD) ambiguity set is equivalent to the regularized SVM. We then propose a novel learning algorithm which is robust to corruption of features. Finally, we study robust classification via distributionally robust chance constraints.

**Notation.** We use lower-case boldface letters to denote column vectors and upper-case boldface letters to denote matrices and use  $\circ$  to denote the element-wise multiplication operation. For simplicity, we denote the set  $\{1, \dots, N\}$  by  $[N]$  and use  $\|\cdot\|_p$  to denote the  $\ell_p$ -norm, and  $\|\cdot\|_p^*$  to denote its dual norm. The identity matrix is denoted by  $\mathbf{I}$  and the indicator function is denoted by  $\mathbf{1}[\cdot]$ . Besides,  $[x]_+ \triangleq \max\{x, 0\}$ .  $\tilde{\mathbf{x}}_i \sim \mathbb{P}$  means random variable  $\tilde{\mathbf{x}}_i$  follows distribution  $\mathbb{P}$ .

## 4.2 Preliminaries of DRO

Distributionally robust optimization is a framework for decision-making under uncertainty where the uncertain data is governed by an *unknown* probability distribution. DRO has been extensively studied in the operation research community for many years [DY10, WKS13, Sca58, GS10]. Mathematically, DRO considers the minimization problem

$$\min_{\mathbf{x}} \sup_{\mathbb{P} \in \mathcal{D}} \mathbb{E}_{\tilde{\mathbf{z}} \sim \mathbb{P}} [v(\mathbf{x}, \tilde{\mathbf{z}})], \quad (4.1)$$

where  $\mathbf{x}$  is the decision vector,  $v(\cdot)$  is the cost function and  $\tilde{\mathbf{z}}$  is the random vector with an unknown distribution  $\mathbb{P}$  which is known to belong to a set of probability distributions

$\mathcal{D}$ , termed ‘‘ambiguity set’’. This formulation means that one wants to minimize the worst case expected cost over the choice of a distribution in the ambiguity set. This formulation overcomes two shortcomings of the standard stochastic programming (SP) approach for handling uncertainties: 1) SP can often be computationally difficult, and 2) the estimation of the true distribution is usually imprecise due to limited information about the stochastic variables.

The general setup we consider is as follows: we are given  $N$  training samples  $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$  where  $\mathbf{x}_i \in \mathbb{R}^n$  and  $y_i \in \{-1, 1\}$ , and the loss function is  $L(\mathbf{x}, y; \boldsymbol{\theta})$  where  $\boldsymbol{\theta}$  is the classifier to learn. The training samples  $\mathbf{x}$  are noisy, hence we consider the following distributionally robust optimization formulation for the classification problem:

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^N \sup_{\mathbb{P} \in \mathcal{D}(\mathbf{x}_i)} \mathbb{E}_{\tilde{\mathbf{x}}_i \sim \mathbb{P}} L(\tilde{\mathbf{x}}_i, y_i; \boldsymbol{\theta}) \quad (4.2)$$

where  $\mathcal{D}(\mathbf{x}_i)$  is the ambiguity set of probability distributions of the  $i$ th training sample and  $\tilde{\mathbf{x}}_i$  is a random variable following distribution  $\mathbb{P} \in \mathcal{D}(\mathbf{x}_i)$ . Formulation (4.2) hence minimizes the worst case expected loss over the choice of possible distributions of the training samples. By specifying different ambiguity sets  $\mathcal{D}$  and different loss functions, we can obtain different classification algorithms. Two widely used ambiguity sets in literature are the following:

$$\mathcal{D} = \left\{ \mathbb{P} : \mathbb{P}[\tilde{\mathbf{z}} \in \mathcal{S}] = 1, (\mathbb{E}[\tilde{\mathbf{z}}] - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}_0^{-1} (\mathbb{E}[\tilde{\mathbf{z}}] - \boldsymbol{\mu}_0) \leq \gamma_1, \mathbb{E}[(\tilde{\mathbf{z}} - \boldsymbol{\mu}_0)^\top (\tilde{\mathbf{z}} - \boldsymbol{\mu}_0)] \leq \gamma_2 \boldsymbol{\Sigma}_0 \right\},$$

$\mathcal{D} = \{\mathbb{P} : \mathbb{E}_{\mathbb{P}}[\|\tilde{\mathbf{z}} - \mathbf{m}\|] \leq \mathbf{f} \text{ for } \mathbf{m}, \mathbf{f} \in \mathbb{R}^n\}$ . We focus on the MAD ambiguity set in this chapter due to its computational efficiency. As we will see below, the constraint

$$\sup_{\mathbb{P} \in \mathcal{D}(\mathbf{x}_i)} \mathbb{E}_{\tilde{\mathbf{x}}_i \sim \mathbb{P}} L(\tilde{\mathbf{x}}_i, y_i; \boldsymbol{\theta}) \leq \xi_i$$

can be reformulated as linear constraints when  $L(\mathbf{x}, y; \boldsymbol{\theta})$  is the hinge loss and  $\mathcal{D}$  is MAD. In contrast, if  $\mathcal{D}$  is the first type of ambiguity set, this constraint can be reformulated as semi-definite constraints. While SDP formulation can be solved in polynomial time, it is not suitable for machine learning problems due to poor scalability. Recall that the distribution of the noise  $\tilde{\mathbf{z}}_i$  belongs to  $\mathcal{D}$ . We consider two types of noise: ‘‘additive noise’’ where the true

sample  $\tilde{\mathbf{x}}_i = \mathbf{x}_i + \tilde{\mathbf{z}}_i$ , and “multiplicative noise” where  $\tilde{\mathbf{x}}_i = \mathbf{x}_i \circ \tilde{\mathbf{z}}_i$ . As we study classification problems, we mainly consider the hinge loss.

Before concluding this section, we briefly explain the intuitive reason why DRO works for machine learning problems. Recall that in supervised learning, the ultimate goal is to minimize expected error w.r.t. the generative distribution, whereas we only have the empirical distribution of the training samples, which is indeed an approximation of the generative distribution. Thus, solving DRO accounts for the difference between the distribution we want to solve, and the distribution we have access to, and hence intuitively controls overfitting. And thus it is not surprising that, as we show below, the standard regularization schemes to control overfitting often have equivalent DRO re-formulations.

### 4.3 DRO Interpretation for Regularized SVMs

In this section, we consider a DRO formulation with hinge loss function and “*additive noise*”, and show that this formulation is equivalent to the standard regularized SVM. We remark that [XCM09] has shown that a robust optimization formulation is equivalent to the regularized SVM *when the training samples are non-separable*. While this result coincides with ours at a high level, our result holds regardless of whether the samples are separable or not.

**Theorem 4.1.** *The distributionally robust formulation for SVM*

$$\min_{\mathbf{w}, b} \frac{1}{N} \sum_{i=1}^N \sup_{\mathbb{P} \in \mathcal{D}} \mathbb{E}_{\tilde{\mathbf{z}} \sim \mathbb{P}} [1 - y_i (\mathbf{w}^\top (\mathbf{x}_i + \tilde{\mathbf{z}}) + b)]_+ \quad (4.3)$$

*is equivalent to the following regularized SVM*

$$\min_{\mathbf{w}, b} \frac{1}{N} \sum_{i=1}^N [1 - y_i (\mathbf{w}^\top (\mathbf{x}_i + \mathbf{m}) + b)]_+ + \mathbf{f}^\top |\mathbf{w}|, \quad (4.4)$$

*when the ambiguity set  $\mathcal{D}$  is the mean absolute deviation  $\mathcal{D} = \{\mathbb{P} : \mathbb{E}_{\mathbb{P}}[|\tilde{\mathbf{z}} - \mathbf{m}|] \leq \mathbf{f}\}$ . Fur-*

thermore, if  $\mathcal{D} = \{\mathbb{P} : \mathbb{E}_{\mathbb{P}}[|\tilde{\mathbf{z}} - \mathbf{m}|] \leq \mathbf{f} \text{ and } \mathbf{f} \in \mathcal{F}\}$ , then (4.3) is equivalent to

$$\min_{\mathbf{w}, b} \frac{1}{N} \sum_{i=1}^N [1 - y_i(\mathbf{w}^\top (\mathbf{x}_i + \mathbf{m}) + b)]_+ + \sup_{\mathbf{f} \in \mathcal{F}} \mathbf{f}^\top |\mathbf{w}|, \quad (4.5)$$

where  $|\mathbf{w}|$  stands for taking the absolute value of each element in  $\mathbf{w}$ .

*Proof.* See Section 4.7. □

This theorem provides a distributional robustness based interpretation for regularized SVMs: suppose that the true sample  $\mathbf{x}_i + \tilde{\mathbf{z}}$  follows an unknown distribution belonging to a known ambiguity set, then regularization of SVMs is indeed a direct result of minimizing the worst-case expected hinge loss error. Moreover, a large value of  $\mathbf{f}_i$  means the  $i$ th feature is heavily corrupted and hence not reliable, which implies that the corresponding weight  $\mathbf{w}_i$  should be set to a small value. This is consistent with the regularized formulation (4.4), where  $\mathbf{w}_i$  tends to be small if  $\mathbf{f}_i$  is very large. Various different regularized classification algorithms, including the standard SVM, can be obtained by selecting a different ambiguity set  $\mathcal{D}$ .

**Corollary 4.1.** *The distributionally robust formulation (4.3) is equivalent to*

$$\min_{\mathbf{w}, b} \frac{1}{N} \sum_{i=1}^N [1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b)]_+ + \alpha \|\mathbf{w}\|_p + \beta \|\mathbf{w}\|_q,$$

when  $\mathcal{D} = \{\mathbb{P} \in \mathcal{P}_0 : \mathbb{E}_{\mathbb{P}}[|\tilde{\mathbf{z}}|] \leq \mathbf{f}, \mathbf{f} \in \mathcal{F}\}$ , where  $\mathcal{F} = \{\mathbf{f} : \mathbf{f} = \mathbf{f}_1 + \mathbf{f}_2, \|\mathbf{f}_1\|_p^* \leq \alpha, \|\mathbf{f}_2\|_q^* \leq \beta\}$ .

*Proof.* From Theorem 4.1, the regularization term in (4.5) becomes

$$\sup_{\mathbf{f} : \mathbf{f} = \mathbf{f}_1 + \mathbf{f}_2, \|\mathbf{f}_1\|_p^* \leq \alpha, \|\mathbf{f}_2\|_q^* \leq \beta} \mathbf{f}^\top |\mathbf{w}| = \sup_{\mathbf{f}_1, \mathbf{f}_2 : \|\mathbf{f}_1\|_p^* \leq \alpha, \|\mathbf{f}_2\|_q^* \leq \beta} \mathbf{f}_1^\top |\mathbf{w}| + \mathbf{f}_2^\top |\mathbf{w}| = \alpha \|\mathbf{w}\|_p + \beta \|\mathbf{w}\|_q.$$

Hence we obtain this corollary. □

## 4.4 Robustness to Corruption of Features

The previous section considers the “additive noise”. We now turn to discuss the “multiplicative noise”, and use it to develop learning algorithms robust to corruption of features. Our

approach is inspired by [GR06], where the authors proposed a robust-optimization based approach to develop classifiers that are robust to *deletion* of features in test data. In contrast, we generalize this approach to incorporate distribution information using the DRO framework. There are two advantages to incorporate distribution information: First, the features in test data may not be completely deleted. Instead, they may be magnified or shrunk due to measurement errors, for example due to changes of illumination in computer vision applications. Second, even if the features are completely deleted, we may still obtain certain distribution information. For example, in the “pepper noise” model considered in [GR06], each feature is deleted with probability  $K/n$ . Thus, it is clear that the expected value of this multiplicative noise is  $K/n$ , a piece of distribution information that one can make use of. More precisely, we consider the following DRO formulation to develop classifiers that are robust to multiplicative disturbances of features in test data:

$$\min_{\mathbf{w}, b} \sum_{i=1}^N \sup_{\mathbb{P} \in \mathcal{D}} \mathbb{E}_{\tilde{\mathbf{z}} \sim \mathbb{P}} [1 - y_i(\mathbf{w}^\top (\mathbf{x}_i \circ \tilde{\mathbf{z}}) + b)]_+ + \frac{1}{2C} \|\mathbf{w}\|_2^2. \quad (4.6)$$

Here  $\mathcal{D}$  is ambiguity set of probability distributions of disturbance. Observe that (4.6) reduces to the robust optimization approach [GR06] when the ambiguity set  $\mathcal{D}$  is  $\{\mathbb{P} : \mathbb{P}(\mathcal{Z}) = 1\}$  where  $\mathcal{Z} \triangleq \{\mathbf{z} | \mathbf{z} \in \{0, 1\}^n, \mathbf{1}^\top \mathbf{z} = n - K\}$  and  $K$  is the maximum number of deleted features. We now consider a more general ambiguity set  $\mathcal{D}$  where the support and expectation of a certain linear transformation of the multiplicative noise are bounded.

**Theorem 4.2.** *If the ambiguity set  $\mathcal{D} = \{\mathbb{P} : \mathbb{E}_{\mathbb{P}}[\mathbf{A}\tilde{\mathbf{z}}] \geq \mathbf{v}$  and  $\mathbb{P}[\mathbf{0} \leq \tilde{\mathbf{z}} \leq \mathbf{t}] = 1\}$  where  $\mathbf{A} \in \mathbb{R}^{p \times n}$ ,  $\mathbf{v} \in \mathbb{R}^p$  and  $\mathbf{t} \geq \mathbf{0}$ , then the distributionally robust optimization problem (4.6) is equivalent to*

$$\begin{aligned} \min_{\mathbf{w}, b, \boldsymbol{\lambda}_i, \boldsymbol{\varphi}_i, \boldsymbol{\beta}_i} \quad & \sum_{i=1}^N \max\{1 - y_i((\mathbf{w} \circ \mathbf{t})^\top \mathbf{x}_i + b) + \mathbf{t}^\top \boldsymbol{\lambda}_i, \mathbf{t}^\top \boldsymbol{\varphi}_i\} + \sum_{i=1}^N (\mathbf{A}\mathbf{t} - \mathbf{v})^\top \boldsymbol{\beta}_i + \frac{1}{2C} \|\mathbf{w}\|_2^2 \\ \text{s.t.} \quad & \boldsymbol{\lambda}_i + \mathbf{A}^\top \boldsymbol{\beta}_i \geq y_i(\mathbf{w} \circ \mathbf{x}_i), \boldsymbol{\varphi}_i + \mathbf{A}^\top \boldsymbol{\beta}_i \geq \mathbf{0}, \forall i \in [N] \\ & \boldsymbol{\lambda}_i \geq \mathbf{0}, \boldsymbol{\varphi}_i \geq \mathbf{0}, \boldsymbol{\beta}_i \geq \mathbf{0}, \forall i \in [N] \end{aligned}$$

*Proof.* See Section 4.7. □

Thus, we obtain a more general scheme to handle multiplicative noise. Based on the appli-



cations, we can select proper  $\mathbf{A}$ ,  $\mathbf{v}$  and  $\mathbf{t}$  to develop classifiers that generalize well to test data. For instance, the next corollary considers the ambiguity set where the expectation of the multiplicative noise for each feature is bounded below by  $1 - K/n$ . Observe that this set contains the ‘‘pepper’’ noise model [GR06].

**Corollary 4.2.** *If the ambiguity set  $\mathcal{D} = \{\mathbb{P} : \mathbb{E}_{\mathbb{P}}[\tilde{\mathbf{z}}] \geq (1 - \frac{K}{n})\mathbf{1} \text{ and } \mathbb{P}[\mathbf{0} \leq \tilde{\mathbf{z}} \leq \mathbf{1}] = 1\}$ , then the distributionally robust optimization problem (4.6) is equivalent to*

$$\begin{aligned} \min_{\mathbf{w}, b, \boldsymbol{\lambda}_i, \boldsymbol{\beta}_i} \quad & \sum_{i=1}^N \max\{1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b) + \|\boldsymbol{\lambda}_i\|_1, 0\} + \sum_{i=1}^N \frac{K}{n} \|\boldsymbol{\beta}_i\|_1 + \frac{1}{2C} \|\mathbf{w}\|_2^2 \\ \text{s.t.} \quad & \boldsymbol{\lambda}_i + \boldsymbol{\beta}_i \geq y_i(\mathbf{w} \circ \mathbf{x}_i), \boldsymbol{\lambda}_i \geq \mathbf{0}, \boldsymbol{\beta}_i \geq \mathbf{0}, \quad \forall i \in [N]. \end{aligned} \quad (4.7)$$

Furthermore, its dual problem is a quadratic programming problem

$$\begin{aligned} \min_{\boldsymbol{\alpha}, \mathbf{r}_i} \quad & \frac{C}{2} \left\| \sum_{i=1}^N y_i \mathbf{x}_i \circ (\alpha_i \mathbf{1} - \mathbf{r}_i) \right\|_2^2 - \sum_{i=1}^N \alpha_i \\ \text{s.t.} \quad & 0 \leq \boldsymbol{\alpha} \leq \mathbf{1}, \sum_{i=1}^N y_i \alpha_i = 0, \\ & \mathbf{r}_i \leq \frac{K}{n} \mathbf{1}, 0 \leq \mathbf{r}_i \leq \alpha_i \mathbf{1}, \quad \forall i \in [N]. \end{aligned} \quad (4.8)$$

Suppose that  $(\boldsymbol{\alpha}^*, \mathbf{r}_i^*)$  is the optimal solution to the dual problem, then the primal optimal solution of  $\mathbf{w}$  is  $\mathbf{w}^* = C \sum_{i=1}^N y_i \mathbf{x}_i \circ (\alpha_i^* \mathbf{1} - \mathbf{r}_i^*)$ .

*Proof.* See Section 4.7. □

## 4.5 Robust Classification via Chance Constraints

Besides robust optimization and distributionally robust optimization, chance constraint is another classical approach for handling uncertainty, which requires that a stochastic constraint is satisfied with a certain probability. A straight-forward formulation of chance

constrained SVM is as follows:

$$\begin{aligned}
\min_{\mathbf{w}, b, \xi} \quad & \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{i=1}^N \xi_i \\
\text{s.t.} \quad & \mathbb{P}[y_i(\mathbf{w}^\top \tilde{\mathbf{x}}_i + b) \leq 1 - \xi_i] \leq \kappa_i, \forall i \in [N] \\
& \xi_i \geq 0, \forall i \in [N],
\end{aligned} \tag{4.9}$$

where  $\tilde{\mathbf{x}}_i$  is a random variable whose distribution is known. Yet, chance constrained optimization problems are notoriously difficult to solve, also the distribution of the uncertainty is often not exactly known. A natural way to extend chance constraints and avoid its intractability is to replace the chance constraint with distributionally robust chance constraint:

$$\sup_{\mathbb{P} \in \mathcal{D}(\mathbf{x}_i)} \mathbb{P}[y_i(\mathbf{w}^\top \tilde{\mathbf{x}}_i + b) \leq 1 - \xi_i] \leq \kappa_i, \forall i \in [N]. \tag{4.10}$$

In words, we require that the worst-case probability (under a class of distributions) of the constraint being satisfied is above a given threshold. [SBS06] studied this formulation where the mean and variance of  $\tilde{\mathbf{x}}_i$  are known and show that it can be reformulated as a SOCP. However, their result can not extend to the case where other useful information such as the support of  $\tilde{\mathbf{x}}_i$  is available. In this section, we propose a DRO-based conservative approximation for the constraint (4.10).

**Lemma 4.1.** *Suppose that  $f : \mathbb{R}^m \times \mathbb{R}^n \rightarrow \mathbb{R}$  and  $\mathcal{D}$  is the ambiguity set of distributions, then the chance constraint*

$$\sup_{\mathbb{P} \in \mathcal{D}} \mathbb{P}[f(\mathbf{x}, \tilde{\mathbf{z}}) \geq 0] \leq \alpha \tag{4.11}$$

can be conservatively approximated by

$$\begin{cases} -t\alpha + \gamma \leq 0 \\ \sup_{\mathbb{P} \in \mathcal{D}} \mathbb{E}_{\mathbb{P}}[[f(\mathbf{x}, \tilde{\mathbf{z}}) + t]_+] \leq \gamma, \end{cases} \tag{4.12}$$

where  $t \in \mathbb{R}$ ,  $\gamma \in \mathbb{R}$  and  $\mathbf{x} \in \mathbb{R}^n$  are decision variables. Here, conservative approximation means that any solution satisfies (4.12) also satisfies (4.11).

*Proof.* For any  $t > 0$ , we have

$$\begin{aligned} \sup_{\mathbb{P} \in \mathcal{D}} \mathbb{P}[f(\mathbf{x}, \tilde{\mathbf{z}}) \geq 0] &= \sup_{\mathbb{P} \in \mathcal{D}} \mathbb{P}[tf(\mathbf{x}, \tilde{\mathbf{z}}) \geq 0] \\ &= \sup_{\mathbb{P} \in \mathcal{D}} \mathbb{E}_{\mathbb{P}}[\mathbf{1}(tf(\mathbf{x}, \tilde{\mathbf{z}}) \geq 0)] \leq \sup_{\mathbb{P} \in \mathcal{D}} \mathbb{E}_{\mathbb{P}}[[1 + tf(\mathbf{x}, \tilde{\mathbf{z}})]_+]. \end{aligned}$$

Thus,  $\inf_{t>0} \{\sup_{\mathbb{P} \in \mathcal{D}} \mathbb{E}_{\mathbb{P}}[[1 + tf(\mathbf{x}, \tilde{\mathbf{z}})]_+] - \alpha\} \leq 0$  implies Inequality (4.11). By changing  $t$  to  $1/t$ , we have the following inequality

$$\inf_{t>0} \{-t\alpha + \sup_{\mathbb{P} \in \mathcal{D}} \mathbb{E}_{\mathbb{P}}[[t + f(\mathbf{x}, \tilde{\mathbf{z}})]_+]\} \leq 0.$$

Note that since  $0 \leq \alpha \leq 1$ ,  $-t\alpha + \sup_{\mathbb{P} \in \mathcal{D}} \mathbb{E}_{\mathbb{P}}[[t + f(\mathbf{x}, \tilde{\mathbf{z}})]_+]$  is always greater than or equal to 0 if  $t \leq 0$ , so that the inequality above can be rewritten as

$$\inf_{t \in \mathbb{R}} \{-t\alpha + \sup_{\mathbb{P} \in \mathcal{D}} \mathbb{E}_{\mathbb{P}}[[t + f(\mathbf{x}, \tilde{\mathbf{z}})]_+]\} \leq 0.$$

Therefore, we obtain the constraints (4.12).  $\square$

Using Lemma 4.1, we have the following theorem which generalizes the results in [SBS06]. Interestingly, this also provides a distributionally robust optimization interpretation of a certain kind of regularized SVMs.

**Theorem 4.3.** *If  $0 < \kappa_i \leq 1$  and the ambiguity set  $\mathcal{D}(\mathbf{x}_i) = \{\mathbb{P} : \mathbb{E}_{\mathbb{P}}[|\tilde{\mathbf{x}}_i - \mathbf{x}_i|] \leq \mathbf{f} \text{ and } \mathbf{f} \in \mathcal{F}\}$ , then the classification problem (4.9) with constraint (4.10) can be conservatively approximated by*

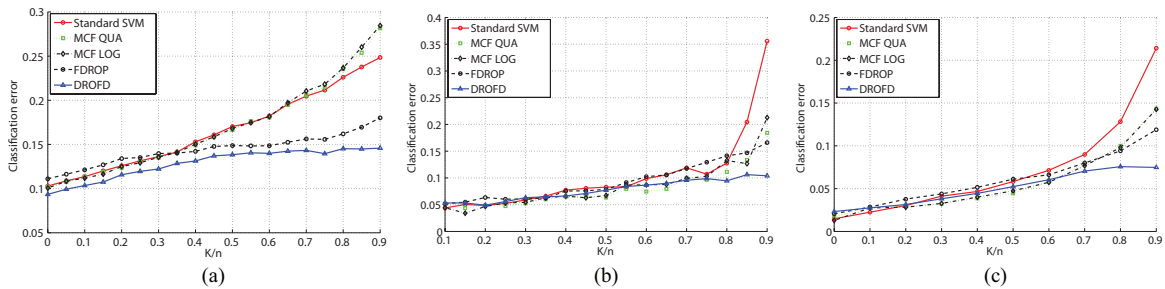
$$\begin{aligned} \min \quad & \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{i=1}^N \xi_i \\ \text{s.t.} \quad & \xi_i \geq 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b) + \frac{1}{\kappa_i} \sup_{\mathbf{f} \in \mathcal{F}} \mathbf{f}^\top |\mathbf{w}|, \quad \forall i \in [N] \\ & \xi_i \geq 0, \quad \forall i \in [N]. \end{aligned}$$

*Proof.* See Section 4.7.  $\square$

## 4.6 Experiments

We conducted several experiments using both synthetic and real data sets to evaluate our classification algorithms. For real datasets (UCI [AN07] and MNIST [LJB<sup>+</sup>95]), the feature values are normalized to  $[-1, 1]$ .

The experiments were conducted on synthetic data, “breast cancer” data from UCI and “MNIST”, to evaluate our classification algorithm proposed in Section 4.4. The synthetic data is generated as follows: The training, validation and test data are uniformly drawn from  $[-10, 10]^n$  and the labels are assigned according to a logistic regression rule  $p(y = 1|\mathbf{x}) \propto \exp(r\mathbf{w}^\top \mathbf{x}) / (1 + \exp(r\mathbf{w}^\top \mathbf{x}))$  where  $n = 20$ ,  $r = 1$  and  $\mathbf{w}$  is a sparse vector whose element is uniformly drawn from  $[-1, 1]$  with probability 25% or 0 otherwise. For synthetic data, we draw 200 training samples, 400 validation samples, and 2000 test samples. For UCI datasets, we use 50, 20 and 30 percent of data as training data, validation data and test data, respectively. For MNIST dataset, we randomly choose 200 and 600 samples from the training samples as the training data and validation data. Besides, all the test data are corrupted by some multiplicative noise (e.g., illumination changes) whose mean is  $1 - \alpha$  and support is  $[0, 1]$ , in particular, the noise is uniformly drawn from  $[0, 1 - \alpha]$  with probability  $\alpha$ , or uniformly drawn from  $[1 - \alpha, 1]$  otherwise. Similar as [WKS13], the validation data is also corrupted by the same noise and used to determine the best parameters. The parameters of the algorithms are selected by cross-validation (parameter  $K$  in FDROP and DROFD is chosen from 1 to  $n$ ). We repeated the experiments 20 times and computed the average classification errors. Figure 4.1(a) shows the simulation results



**Figure 4.1:** The classification errors of SVM, MCF QUA, MCF LOG, FDROP and DROFD: the x-axis is the corruption level  $\alpha = K/n$ . (a) Simulation results. (b) “Breast cancer”. (c) “MNIST”

on synthetic data in which we compare our method (DROFD) with standard SVM, FDROP [GR06] and MCF algorithms with quadratic loss and logistic loss [WKS13]. Clearly, our method outperforms all these methods, which shows that that DROFD is more robust to disturbance of features. Figure 4.1(b)(c) show the results of real data from which shows that DROFD has smaller classification error especially when corruption level  $\alpha$  is large.

## 4.7 Proofs of Technical Results

Recall that distributionally robust optimization concerns the following formulation

$$\sup_{\mathbb{Q} \in \mathcal{P}} \mathbb{E}_{\mathbb{Q}}[v(\mathbf{x}, \tilde{\mathbf{z}})] \leq w, \quad (4.13)$$

where  $\mathbf{x}$  is the decision vector, and  $\tilde{\mathbf{z}}$  is the random vector with distribution  $\mathbb{Q}$  that belongs to an ambiguity set  $\mathcal{P}$ . Although this constraint is intractable in general, it may become computationally tractable under specific assumptions about the ambiguity set  $\mathcal{P}$  and the constraint function  $v$ . For example, [WKS13] consider the following ambiguity set

$$\mathcal{P} = \left\{ \mathbb{P} \in \mathcal{P}_0(\mathbb{R}^P, \mathbb{R}^Q) : \begin{array}{l} \mathbb{E}_{\mathbb{P}}[\mathbf{A}\tilde{\mathbf{z}} + \mathbf{B}\tilde{\mathbf{u}}] = \mathbf{b} \\ \mathbb{P}[(\tilde{\mathbf{z}}, \tilde{\mathbf{u}}) \in \mathcal{C}_i] \in [\underline{p}_i, \overline{p}_i], \forall i \in \mathcal{I} \end{array} \right\}, \quad (4.14)$$

where  $\mathbb{P}$  represents a joint probability distribution of random vector  $\tilde{\mathbf{z}} \in \mathbb{R}^P$  and some auxiliary random vector  $\tilde{\mathbf{u}} \in \mathbb{R}^Q$ , and  $\mathbf{A} \in \mathbb{R}^{K \times P}$ ,  $\mathbf{B} \in \mathbb{R}^{K \times Q}$ ,  $\mathbf{b} \in \mathbb{R}^K$ ,  $\mathcal{I} = \{1, \dots, I\}$  and  $\underline{p}_i, \overline{p}_i \in [0, 1]$ . The confidence sets  $\mathcal{C}_i$  are defined by

$$\mathcal{C}_i = \{(\mathbf{z}, \mathbf{u}) \in \mathbb{R}^P \times \mathbb{R}^Q : \mathbf{C}_i \mathbf{z} + \mathbf{D}_i \mathbf{u} \preceq_{K_i} \mathbf{c}_i\}, \quad (4.15)$$

where  $\mathbf{C} \in \mathbb{R}^{L_i \times P}$ ,  $\mathbf{D} \in \mathbb{R}^{L_i \times Q}$ ,  $\mathbf{c}_i \in \mathbb{R}^{L_i}$  and  $K_i$  are proper cones.

[WKS13] requires that the ambiguity set  $\mathcal{P}$  satisfies the following conditions:

- (C1) The confidence set  $\mathcal{C}_I$  is bounded and has probability one, that is,  $\underline{p}_I = \overline{p}_I = 1$ ;
- (C2) There is a probability distribution  $\mathbb{P} \in \mathcal{P}$  such that  $\mathbb{P}[(\tilde{\mathbf{z}}, \tilde{\mathbf{u}}) \in \mathcal{C}_i] \in (\underline{p}_i, \overline{p}_i)$  whenever

$$\underline{p}_i < \overline{p}_i;$$

(N) For all  $i, i' \in \mathcal{I}$  and  $i \neq i'$ , we have either  $\mathcal{C}_i \Subset \mathcal{C}_{i'}$ ,  $\mathcal{C}_{i'} \Subset \mathcal{C}_i$  or  $\mathcal{C}_i \cap \mathcal{C}_{i'} = \emptyset$ ;

where  $A \Subset B$  means set  $A$  is strictly included in set  $B$  or  $A$  is contained in the interior of  $B$ .

Under these assumptions, they showed the following theorems:

**Theorem 4.4.** (Theorem 5, [WKS13]) Let  $\mathbf{f} \in \mathbb{R}^P$  and  $g : \mathbb{R}^P \rightarrow \mathbb{R}^Q$  be a function with a conic representable  $\mathcal{K}$ -epigraph, and consider the ambiguity set

$$\mathcal{P}' = \left\{ \mathbb{P} \in \mathcal{P}_0(\mathbb{R}^P) : \begin{array}{l} \mathbb{E}_{\mathbb{P}}[g(\tilde{\mathbf{z}})] \preceq_{\mathcal{K}} \mathbf{f} \\ \mathbb{P}[\tilde{\mathbf{z}} \in \mathcal{C}_i] \in [\underline{p}_i, \overline{p}_i], \forall i \in \mathcal{I} \end{array} \right\}$$

as well as the lifted ambiguity set

$$\mathcal{P} = \left\{ \mathbb{P} \in \mathcal{P}_0(\mathbb{R}^P, \mathbb{R}^Q) : \begin{array}{l} \mathbb{E}_{\mathbb{P}}[\tilde{\mathbf{u}}] = \mathbf{f} \\ \mathbb{P}[g(\tilde{\mathbf{z}}) \preceq_{\mathcal{K}} \tilde{\mathbf{u}}] = 1 \\ \mathbb{P}[\tilde{\mathbf{z}} \in \mathcal{C}_i] \in [\underline{p}_i, \overline{p}_i], \forall i \in \mathcal{I} \end{array} \right\}$$

which involves the auxiliary random vector  $\tilde{\mathbf{u}} \in \mathbb{R}^Q$ . We then have that (i)  $\mathcal{P}' = \prod_{\tilde{\delta}} \mathcal{P}$  and (ii)  $\mathcal{P}$  can be reformulated as an instance of the standardized ambiguity set (4.14).

**Theorem 4.5.** Assume that the conditions (C1), (C2) and (N) hold and that the constraint function  $v(\mathbf{x}, \mathbf{z})$  is convex in  $\mathbf{z}$ . Then, the distributionally robust constraint (4.13) is satisfied for the ambiguity set (4.14) if and only if the semi-infinite constraint system

$$\begin{aligned} \mathbf{b}^\top \boldsymbol{\beta} + \sum_{i \in \mathcal{I}} [\overline{p}_i \boldsymbol{\kappa}_i - \underline{p}_i \boldsymbol{\lambda}_i] &\leq w, \\ [\mathbf{A}\mathbf{z} + \mathbf{B}\mathbf{u}]^\top \boldsymbol{\beta} + \sum_{i' \in \mathcal{A}(i)} [\boldsymbol{\kappa}_{i'} - \boldsymbol{\lambda}_{i'}] &\geq v(\mathbf{x}, \mathbf{z}), \forall (\mathbf{z}, \mathbf{u}) \in \mathcal{C}_i, i \in \mathcal{I} \end{aligned} \tag{4.16}$$

is satisfied by some  $\boldsymbol{\beta} \in \mathbb{R}^K$  and  $\boldsymbol{\kappa}, \boldsymbol{\lambda} \in \mathbb{R}_+^I$ , where  $\mathcal{A}(i) = \{i\} \cup \{i' \in \mathcal{I} : \mathcal{C}_i \Subset \mathcal{C}_{i'}\}$ .

## 4.7.1 Proof of Theorem 4.1

*Proof.* From the lifting theorem (Theorem 4.4), the mean absolute deviation ambiguity set  $\mathcal{D}$  can be rewritten as  $\mathcal{D} = \{\mathbb{P} \in \mathcal{P}_0(\mathbb{R}^n, \mathbb{R}^n) : \mathbb{E}_{\mathbb{P}}[\tilde{\mathbf{u}}] = \mathbf{f}, \mathbb{P}[(\tilde{\mathbf{z}}, \tilde{\mathbf{u}}) \in \mathcal{C}] = 1\}$  where  $\mathcal{C} = \{(\mathbf{z}, \mathbf{u}) : |\mathbf{z} - \mathbf{m}| \leq \mathbf{u}\}$ . Since the hinge loss function is convex in  $\tilde{\mathbf{z}}$  and the mean absolute deviation satisfies the conditions (C1), (C2) and (N), then from Theorem 4.5, the optimization problem (4.3) is equivalent to

$$\begin{aligned} \min \quad & \frac{1}{N} \sum_{i=1}^N \varepsilon_i \\ \text{s.t.} \quad & \mathbf{f}^\top \boldsymbol{\beta}_i + \alpha_i \leq \varepsilon_i, \quad \forall i \in [N] \\ & \boldsymbol{\beta}_i^\top \mathbf{u} + \alpha_i \geq \max\{1 - y_i(\mathbf{w}^\top(\mathbf{x}_i + \mathbf{z}) + b), 0\}, \quad \forall (\mathbf{z}, \mathbf{u}) \in \mathcal{C}, \quad \forall i \in [N]. \end{aligned}$$

For simplicity, let  $\mathbf{C} = (\mathbf{I}, -\mathbf{I})^\top$ ,  $\mathbf{D} = (-\mathbf{I}, -\mathbf{I})^\top$  and  $\mathbf{c} = (\mathbf{m}^\top, -\mathbf{m}^\top)^\top$  where  $\mathbf{I}$  is the  $n \times n$  identity matrix. Then for each  $i \in [N]$ , the constraints can be rewritten as

$$\begin{cases} \mathbf{f}^\top \boldsymbol{\beta}_i + \alpha_i \leq \varepsilon_i \\ \boldsymbol{\beta}_i^\top \mathbf{u} + \alpha_i \geq \max\{1 - y_i(\mathbf{w}^\top(\mathbf{x}_i + \mathbf{z}) + b), 0\}, \quad \forall \mathbf{Cz} + \mathbf{Du} \leq \mathbf{c}. \end{cases}$$

or equivalently

$$\begin{cases} \mathbf{f}^\top \boldsymbol{\beta}_i + \alpha_i \leq \varepsilon_i \\ \boldsymbol{\beta}_i^\top \mathbf{u} + \alpha_i \geq 0, \quad \forall \mathbf{Cz} + \mathbf{Du} \leq \mathbf{c} \\ \boldsymbol{\beta}_i^\top \mathbf{u} + \alpha_i \geq 1 - y_i(\mathbf{w}^\top(\mathbf{x}_i + \mathbf{z}) + b), \quad \forall \mathbf{Cz} + \mathbf{Du} \leq \mathbf{c}. \end{cases}$$

The last two constraints can be reformulated as (for clearness, we ignore the subscript  $i$ )

$$\begin{cases} -\alpha \leq \min_{\mathbf{z}, \mathbf{u}} \boldsymbol{\beta}^\top \mathbf{u} \text{ s.t. } \mathbf{Cz} + \mathbf{Du} \leq \mathbf{c} \\ 1 - y(\mathbf{w}^\top \mathbf{x} + b) - \alpha \leq \min_{\mathbf{z}, \mathbf{u}} \boldsymbol{\beta}^\top \mathbf{u} + y\mathbf{w}^\top \mathbf{z} \text{ s.t. } \mathbf{Cz} + \mathbf{Du} \leq \mathbf{c}. \end{cases}$$

From the duality, we have

$$\begin{cases} \alpha \geq \mathbf{c}^\top \boldsymbol{\lambda}, \mathbf{C}^\top \boldsymbol{\lambda} = 0, \boldsymbol{\beta} + \mathbf{D}^\top \boldsymbol{\lambda} = 0, \boldsymbol{\lambda} \geq 0 \\ 1 - y(\mathbf{w}^\top \mathbf{x} + b) - \alpha \leq -\mathbf{c}^\top \mathbf{v}, \boldsymbol{\beta} + \mathbf{D}^\top \mathbf{v} = 0, y\mathbf{w} + \mathbf{C}^\top \mathbf{v} = 0, \mathbf{v} \geq 0, \end{cases}$$

which is equivalent to

$$\begin{cases} \alpha \geq 0, \boldsymbol{\beta} \geq 0 \\ 1 - y(\mathbf{w}^\top \mathbf{x} + b) - \alpha \leq y\mathbf{w}^\top \mathbf{m}, \boldsymbol{\beta} + \mathbf{D}^\top \mathbf{v} = 0, y\mathbf{w} + \mathbf{C}^\top \mathbf{v} = 0, \mathbf{v} \geq 0, \end{cases}$$

Combine with the first constraint  $\mathbf{f}^\top \boldsymbol{\beta} + \alpha \leq \varepsilon$ , we have

$$\begin{cases} 1 - y(\mathbf{w}^\top (\mathbf{x} + \mathbf{m}) + b) \leq \varepsilon - \mathbf{f}^\top \boldsymbol{\beta} \\ \boldsymbol{\beta} + \mathbf{D}^\top \mathbf{v} = 0 \\ y\mathbf{w} + \mathbf{C}^\top \mathbf{v} = 0 \\ \varepsilon - \mathbf{f}^\top \boldsymbol{\beta} \geq 0 \\ \mathbf{v} \geq 0 \end{cases}$$

By eliminating  $\boldsymbol{\beta}$ , we have

$$\begin{cases} 1 - y(\mathbf{w}^\top (\mathbf{x} + \mathbf{m}) + b) \leq \varepsilon + \mathbf{f}^\top \mathbf{D}^\top \mathbf{v} \\ y\mathbf{w} + \mathbf{C}^\top \mathbf{v} = 0 \\ \varepsilon + \mathbf{f}^\top \mathbf{D}^\top \mathbf{v} \geq 0 \\ \mathbf{v} \geq 0 \end{cases}$$

Then let  $\mathbf{v} = (\mathbf{v}_1^\top, \mathbf{v}_2^\top)^\top$  and  $\boldsymbol{\nu} = \mathbf{v}_1 + \mathbf{v}_2$  where  $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{R}^P$ , the constraints above can be rewritten as

$$\begin{cases} 1 - y(\mathbf{w}^\top (\mathbf{x} + \mathbf{m}) + b) \leq \varepsilon - \mathbf{f}^\top (\mathbf{v}_1 + \mathbf{v}_2) \\ y\mathbf{w} = \mathbf{v}_2 - \mathbf{v}_1 \\ \varepsilon \geq \mathbf{f}^\top (\mathbf{v}_1 + \mathbf{v}_2) \\ \mathbf{v}_1, \mathbf{v}_2 \geq 0 \end{cases}$$



Since  $\boldsymbol{\nu} + y\mathbf{w} = 2\mathbf{v}_2 \geq 0$  and  $\boldsymbol{\nu} - y\mathbf{w} = 2\mathbf{v}_1 \geq 0$ , we have

$$\begin{cases} 1 - y(\mathbf{w}^\top(\mathbf{x} + \mathbf{m}) + b) \leq \varepsilon - \mathbf{f}^\top \boldsymbol{\nu} \\ -\boldsymbol{\nu} \leq \mathbf{w} \leq \boldsymbol{\nu} \\ \varepsilon \geq \mathbf{f}^\top \boldsymbol{\nu} \end{cases}$$

which is also equivalent to (since  $\mathbf{f} \geq 0$ )

$$\begin{cases} 1 - y(\mathbf{w}^\top(\mathbf{x} + \mathbf{m}) + b) \leq \varepsilon - \mathbf{f}^\top |\mathbf{w}| \\ \varepsilon - \mathbf{f}^\top |\mathbf{w}| \geq 0. \end{cases}$$

Therefore, (4.3) is equivalent to the regularized SVM (4.4).

If  $\mathcal{D} = \{\mathbb{P} \in \mathcal{P}_0(\mathbb{R}^n) : \mathbb{P} \in \mathcal{D}(\mathbf{f}) \text{ and } \mathbf{f} \in \mathcal{F}\}$  where  $\mathcal{D}(\mathbf{f}) = \{\mathbb{P} \in \mathcal{P}_0(\mathbb{R}^n) : \mathbb{E}_{\mathbb{P}}[|\tilde{\mathbf{z}} - \mathbf{m}|] \leq \mathbf{f}\}$ ,

(4.3) is equivalent to

$$\min_{\mathbf{w}, b} \frac{1}{N} \sum_{i=1}^N \sup_{\mathbf{f} \in \mathcal{F}} \sup_{\mathbb{P} \in \mathcal{D}(\mathbf{f})} \mathbb{E}_{\tilde{\mathbf{z}} \sim \mathbb{P}} \max\{1 - y_i(\mathbf{w}^\top(\mathbf{x}_i + \tilde{\mathbf{z}}) + b), 0\}. \quad (4.17)$$

From the proof above, we know that (4.17) is equivalent to

$$\min_{\mathbf{w}, b} \frac{1}{N} \sum_{i=1}^N \sup_{\mathbf{f} \in \mathcal{F}} \{\max\{1 - y_i(\mathbf{w}^\top(\mathbf{x}_i + \mathbf{m}) + b), 0\} + \mathbf{f}^\top |\mathbf{w}|\}. \quad (4.18)$$

Hence (4.3) is equivalent to (4.5).  $\square$

#### 4.7.2 Proof of Theorem 4.2

**Theorem 4.6.** *If the ambiguity set  $\mathcal{D} = \{\mathbb{P} : \mathbb{E}_{\mathbb{P}}[\mathbf{A}\tilde{\mathbf{z}}] \geq \mathbf{v} \text{ and } \mathbb{P}[\mathbf{0} \leq \tilde{\mathbf{z}} \leq \mathbf{t}] = 1\}$  where  $\mathbf{A} \in \mathbb{R}^{p \times n}$ ,  $\mathbf{v} \in \mathbb{R}^p$  and  $\mathbf{t} \geq 0$ , then the distributionally robust optimization problem (4.6) is*

equivalent to

$$\begin{aligned}
\min \quad & \sum_{i=1}^N \max\{1 - y_i((\mathbf{w} \circ \mathbf{t})^\top \mathbf{x}_i + b) + \mathbf{t}^\top \boldsymbol{\lambda}_i, \mathbf{t}^\top \boldsymbol{\varphi}_i\} + \sum_{i=1}^N (\mathbf{A}\mathbf{t} - \mathbf{v})^\top \boldsymbol{\beta}_i + \frac{1}{2C} \|\mathbf{w}\|_2^2 \\
\text{s.t.} \quad & \boldsymbol{\lambda}_i + \mathbf{A}^\top \boldsymbol{\beta}_i \geq y_i(\mathbf{w} \circ \mathbf{x}_i), \boldsymbol{\varphi}_i + \mathbf{A}^\top \boldsymbol{\beta}_i \geq 0, \forall i \in [N] \\
& \boldsymbol{\lambda}_i \geq 0, \boldsymbol{\varphi}_i \geq 0, \boldsymbol{\beta}_i \geq 0, \forall i \in [N]
\end{aligned} \tag{4.19}$$

Furthermore, its dual problem is a quadratic programming problem:

$$\begin{aligned}
\min \quad & \frac{C}{2} \left\| \sum_{i=1}^N y_i \mathbf{x}_i \circ (\alpha_i \mathbf{t} - \mathbf{r}_i) \right\|_2^2 - \sum_{i=1}^N \alpha_i \\
\text{s.t.} \quad & 0 \leq \boldsymbol{\alpha} \leq \mathbf{1}, \sum_{i=1}^N y_i \alpha_i = 0, \mathbf{A}(\mathbf{r}_i + \boldsymbol{\xi}_i) \leq \mathbf{A}\mathbf{t} - \mathbf{v}, \forall i \in [N] \\
& 0 \leq \mathbf{r}_i \leq \alpha_i \mathbf{t}, 0 \leq \boldsymbol{\xi}_i \leq (1 - \alpha_i) \mathbf{t}, \forall i \in [N]
\end{aligned} \tag{4.20}$$

Suppose that  $(\alpha_i, \mathbf{r}_i, \boldsymbol{\xi}_i)$  is the optimal solution, then we have  $\mathbf{w} = C \sum_{i=1}^N y_i \mathbf{x}_i \circ (\alpha_i \mathbf{t} - \mathbf{r}_i)$ .

*Proof.* Replace  $\tilde{\mathbf{z}}$  by  $\mathbf{t} - \tilde{\mathbf{z}}$ , then the optimization problem (4.6) can be reformulated as

$$\min_{\mathbf{w}, b} \sum_{i=1}^N \sup_{\mathbb{P} \in \mathcal{D}} \mathbb{E}_{\tilde{\mathbf{z}} \sim \mathbb{P}} \max\{1 - y_i(\mathbf{w}^\top (\mathbf{x}_i \circ (\mathbf{t} - \tilde{\mathbf{z}})) + b), 0\} + \frac{1}{2C} \|\mathbf{w}\|_2^2, \tag{4.21}$$

and the ambiguity set  $\mathcal{D}$  can be rewritten as

$$\mathcal{D} = \{\mathbb{P} \in \mathcal{P}_0(\mathbb{R}^n) : \mathbb{E}_{\mathbb{P}}[\mathbf{A}\tilde{\mathbf{z}}] \leq \mathbf{A}\mathbf{t} - \mathbf{v} \text{ and } \mathbb{P}[\mathbf{0} \leq \tilde{\mathbf{z}} \leq \mathbf{t}] = 1\}.$$

From the lifting theorem (Theorem 4.4), the ambiguity set  $\mathcal{D}$  can be rewritten as  $\mathcal{D} = \{\mathbb{P} \in \mathcal{P}_0(\mathbb{R}^n, \mathbb{R}^p) : \mathbb{E}_{\mathbb{P}}[\tilde{\mathbf{u}}] = \mathbf{A}\mathbf{t} - \mathbf{v} \text{ and } \mathbb{P}[\mathbf{A}\tilde{\mathbf{z}} \leq \tilde{\mathbf{u}}] = 1 \text{ and } \mathbb{P}[\mathbf{0} \leq \tilde{\mathbf{z}} \leq \mathbf{t}] = 1\}$ , which is equivalent to  $\{\mathbb{P} \in \mathcal{P}_0(\mathbb{R}^n, \mathbb{R}^p) : \mathbb{E}_{\mathbb{P}}[\tilde{\mathbf{u}}] = \mathbf{A}\mathbf{t} - \mathbf{v} \text{ and } \mathbb{P}[\mathbf{A}\tilde{\mathbf{z}} \leq \tilde{\mathbf{u}}, \mathbf{0} \leq \tilde{\mathbf{z}} \leq \mathbf{t}] = 1\}$ . Since the constraint function is convex in  $\tilde{\mathbf{z}}$  and  $\mathcal{D}$  satisfies the conditions (C1), (C2) and (N), then

from Theorem 4.5, (4.21) is equivalent to

$$\begin{aligned} \min \quad & \sum_{i=1}^N \varepsilon_i \\ \text{s.t.} \quad & (\mathbf{A}\mathbf{t} - \mathbf{v})^\top \boldsymbol{\beta}_i + \alpha_i \leq \varepsilon_i, \quad \forall i \in [N] \\ & \boldsymbol{\beta}_i^\top \mathbf{u} + \alpha_i \geq \max\{1 - y_i(\mathbf{w}^\top(\mathbf{x}_i \circ (\mathbf{t} - \mathbf{z})) + b), 0\}, \quad \forall \mathbf{A}\mathbf{z} \leq \mathbf{u}, \mathbf{0} \leq \mathbf{z} \leq \mathbf{t}, \quad \forall i \in [N]. \end{aligned}$$

For each  $i \in [N]$ , the constraints above can be rewritten as (we ignore the subscript  $i$  for clearness)

$$\begin{cases} (\mathbf{A}\mathbf{t} - \mathbf{v})^\top \boldsymbol{\beta} + \alpha \leq \varepsilon \\ \boldsymbol{\beta}^\top \mathbf{u} + \alpha \geq 0, \quad \forall \mathbf{A}\mathbf{z} \leq \mathbf{u}, \mathbf{0} \leq \mathbf{z} \leq \mathbf{t} \\ \boldsymbol{\beta}^\top \mathbf{u} + \alpha \geq 1 - y((\mathbf{w} \circ \mathbf{x})^\top(\mathbf{t} - \mathbf{z}) + b), \quad \forall \mathbf{A}\mathbf{z} \leq \mathbf{u}, \mathbf{0} \leq \mathbf{z} \leq \mathbf{t}. \end{cases}$$

The last two constraints can be reformulated as

$$\begin{cases} -\alpha \leq \min_{\mathbf{z}, \mathbf{u}} \boldsymbol{\beta}^\top \mathbf{u} \text{ s.t. } \mathbf{A}\mathbf{z} \leq \mathbf{u}, \mathbf{0} \leq \mathbf{z} \leq \mathbf{t} \\ 1 - y((\mathbf{w} \circ \mathbf{t})^\top \mathbf{x} + b) - \alpha \leq \min_{\mathbf{z}, \mathbf{u}} \boldsymbol{\beta}^\top \mathbf{u} - y((\mathbf{w} \circ \mathbf{x})^\top \mathbf{z}) \text{ s.t. } \mathbf{A}\mathbf{z} \leq \mathbf{u}, \mathbf{0} \leq \mathbf{z} \leq \mathbf{t}. \end{cases}$$

From the duality, we have

$$\begin{cases} \alpha - \mathbf{t}^\top \boldsymbol{\varphi} \geq 0, \boldsymbol{\varphi} + \mathbf{A}^\top \boldsymbol{\beta} \geq 0, \boldsymbol{\beta} \geq 0, \boldsymbol{\varphi} \geq 0 \\ \alpha \geq 1 - y((\mathbf{w} \circ \mathbf{t})^\top \mathbf{x} + b) + \mathbf{t}^\top \boldsymbol{\lambda}, \boldsymbol{\lambda} + \mathbf{A}^\top \boldsymbol{\beta} \geq y(\mathbf{w} \circ \mathbf{x}), \boldsymbol{\lambda} \geq \mathbf{0}, \boldsymbol{\beta} \geq 0. \end{cases}$$

Combine with the first constraint  $(\mathbf{A}\mathbf{t} - \mathbf{v})^\top \boldsymbol{\beta} + \alpha \leq \varepsilon$ , we have

$$\begin{cases} \varepsilon \geq \mathbf{t}^\top \boldsymbol{\varphi} + (\mathbf{A}\mathbf{t} - \mathbf{v})^\top \boldsymbol{\beta} \\ \varepsilon \geq 1 - y((\mathbf{w} \circ \mathbf{t})^\top \mathbf{x} + b) + \mathbf{t}^\top \boldsymbol{\lambda} + (\mathbf{A}\mathbf{t} - \mathbf{v})^\top \boldsymbol{\beta} \\ \boldsymbol{\lambda} + \mathbf{A}^\top \boldsymbol{\beta} \geq y(\mathbf{w} \circ \mathbf{x}) \\ \boldsymbol{\varphi} + \mathbf{A}^\top \boldsymbol{\beta} \geq 0, \\ \boldsymbol{\lambda} \geq 0, \boldsymbol{\varphi} \geq 0, \boldsymbol{\beta} \geq 0. \end{cases}$$

Hence (4.6) is equivalent to (4.19).

The Lagrangian  $L$  of the optimization problem (4.19) is as follows:

$$\begin{aligned}
L = & \frac{1}{2C} \mathbf{w}^\top \mathbf{w} + \sum_{i=1}^N \varepsilon_i + \sum_{i=1}^N (\mathbf{A} \mathbf{t} - \mathbf{v})^\top \boldsymbol{\beta}_i + \sum_{i=1}^N \alpha_i (1 - y_i ((\mathbf{w} \circ \mathbf{t})^\top \mathbf{x}_i + b) + \mathbf{t}^\top \boldsymbol{\lambda}_i - \varepsilon_i) + \\
& \sum_{i=1}^N \delta_i (\mathbf{t}^\top \boldsymbol{\varphi}_i - \varepsilon_i) + \sum_{i=1}^N \mathbf{r}_i^\top (y_i (\mathbf{w} \circ \mathbf{x}_i) - \boldsymbol{\lambda}_i - \mathbf{A}^\top \boldsymbol{\beta}_i) - \\
& \sum_{i=1}^N \boldsymbol{\xi}_i^\top (\boldsymbol{\varphi}_i + \mathbf{A}^\top \boldsymbol{\beta}_i) - \sum_{i=1}^N (\mathbf{v}_{1i}^\top \boldsymbol{\lambda}_i + \mathbf{v}_{2i}^\top \boldsymbol{\varphi}_i + \mathbf{v}_{3i}^\top \boldsymbol{\beta}_i),
\end{aligned}$$

where  $\alpha_i \geq 0, \delta_i \geq 0, \mathbf{r}_i \geq 0, \boldsymbol{\xi}_i \geq 0$  and  $\mathbf{v}_{1i}, \mathbf{v}_{2i}, \mathbf{v}_{3i} \geq 0$ . To obtain the minimum of  $L$ , ones can compute the derivatives w.r.t the primal variables

$$\begin{aligned}
\frac{\partial L}{\partial \varepsilon_i} &= \sum_{i=1}^N (1 - \alpha_i - \delta_i) = 0 \\
\frac{\partial L}{\partial \boldsymbol{\beta}_i} &= \sum_{i=1}^N (\mathbf{A} \mathbf{t} - \mathbf{v} - \mathbf{A} \mathbf{r}_i - \mathbf{A} \boldsymbol{\xi}_i - \mathbf{v}_{3i}) = 0 \\
\frac{\partial L}{\partial b} &= \sum_{i=1}^N y_i \alpha_i = 0 \\
\frac{\partial L}{\partial \boldsymbol{\lambda}_i} &= \sum_{i=1}^N (\alpha_i \mathbf{t} - \mathbf{r}_i - \mathbf{v}_{1i}) = 0 \\
\frac{\partial L}{\partial \boldsymbol{\varphi}_i} &= \sum_{i=1}^N (\delta_i \mathbf{t} - \boldsymbol{\xi}_i - \mathbf{v}_{2i}) = 0 \\
\frac{\partial L}{\partial \mathbf{w}} &= \frac{1}{C} \mathbf{w} - \sum_{i=1}^N \alpha_i y_i (\mathbf{t} \circ \mathbf{x}_i) + \sum_{i=1}^N y_i (\mathbf{r}_i \circ \mathbf{x}_i) = 0.
\end{aligned}$$

Since  $\mathbf{t} \geq 0, \delta_i \geq 0$  and  $\mathbf{v}_{1i}, \mathbf{v}_{2i}, \mathbf{v}_{3i} \geq 0$ , we have

$$\begin{aligned}
0 &\leq \boldsymbol{\alpha} \leq \mathbf{1}, \\
\sum_{i=1}^N y_i \alpha_i &= 0, \\
\mathbf{A}(\mathbf{r}_i + \boldsymbol{\xi}_i) &\leq \mathbf{A} \mathbf{t} - \mathbf{v}, \quad \forall i \in [N] \\
0 &\leq \mathbf{r}_i \leq \alpha_i \mathbf{t}, \quad \forall i \in [N], \\
0 &\leq \boldsymbol{\xi}_i \leq (1 - \alpha_i) \mathbf{t}, \quad \forall i \in [N],
\end{aligned}$$

and  $\mathbf{w} = C \sum_{i=1}^N y_i \mathbf{x}_i \circ (\alpha_i \mathbf{t} - \mathbf{r}_i)$ . □

## 4.7.3 Proof of Corollary 4.2

From Theorem 4.2, when  $\mathbf{t} = \mathbf{1}$ ,  $\mathbf{A} = \mathbf{I}$  and  $\mathbf{v} = (1 - \frac{K}{n})\mathbf{1}$ , the optimization problem (4.19) can be reformulated as

$$\begin{aligned} \min \quad & \sum_{i=1}^N \max\{1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b) + \|\boldsymbol{\lambda}_i\|_1, \|\boldsymbol{\varphi}_i\|_1\} + \sum_{i=1}^N \frac{K}{n} \|\boldsymbol{\beta}_i\|_1 + \frac{1}{2C} \|\mathbf{w}\|_2^2 \\ \text{s.t.} \quad & \boldsymbol{\lambda}_i + \boldsymbol{\beta}_i \geq y_i(\mathbf{w} \circ \mathbf{x}_i), \quad \forall i \in [N] \\ & \boldsymbol{\varphi}_i + \boldsymbol{\beta}_i \geq 0, \quad \forall i \in [N] \\ & \boldsymbol{\lambda}_i \geq 0, \boldsymbol{\varphi}_i \geq 0, \boldsymbol{\beta}_i \geq 0, \quad \forall i \in [N], \end{aligned}$$

which implies that  $\boldsymbol{\varphi}_i = 0$ . Thus, we obtain (4.7).

By substituting  $\mathbf{t} = \mathbf{1}$ ,  $\mathbf{A} = \mathbf{I}$  and  $\mathbf{v} = (1 - \frac{K}{n})\mathbf{1}$  into (4.20), we can obtain its dual problem

$$\begin{aligned} \min \quad & \frac{C}{2} \left\| \sum_{i=1}^N y_i \mathbf{x}_i \circ (\alpha_i \mathbf{1} - \mathbf{r}_i) \right\|_2^2 - \sum_{i=1}^N \alpha_i \\ \text{s.t.} \quad & 0 \leq \boldsymbol{\alpha} \leq \mathbf{1}, \sum_{i=1}^N y_i \alpha_i = 0, \\ & \mathbf{r}_i + \boldsymbol{\xi}_i \leq \frac{K}{n} \mathbf{1}, \quad \forall i \in [N] \\ & 0 \leq \mathbf{r}_i \leq \alpha_i \mathbf{1}, \quad \forall i \in [N], \\ & 0 \leq \boldsymbol{\xi}_i \leq (1 - \alpha_i) \mathbf{1}, \quad \forall i \in [N], \end{aligned}$$

which implies that  $\boldsymbol{\xi}_i = 0$ , so that we obtain (4.8).

## 4.7.4 Proof of Theorem 4.3

*Proof.* Let  $\mathcal{D}(\mathbf{f}) = \{\mathbb{P} \in \mathcal{P}_0(\mathbb{R}^n) : \mathbb{E}_{\mathbb{P}}[\|\tilde{\mathbf{z}}\|] \leq \mathbf{f}\}$ , we first consider the constraint

$$\sup_{\mathbb{P} \in \mathcal{D}(\mathbf{f})} \mathbb{E}_{\mathbb{P}}[1 - \xi - y(\mathbf{w}^\top (\mathbf{x} + \tilde{\mathbf{z}}) + b) + t]_+ \leq \varepsilon.$$

Follow the proof of Theorem 4.1, it can be shown that this constraint is equivalent to

$$\begin{cases} \mathbf{f}^\top \boldsymbol{\beta} + \alpha \leq \varepsilon \\ \boldsymbol{\beta}^\top \mathbf{u} + \alpha \geq 0, \forall -\mathbf{u} \leq \mathbf{z} \leq \mathbf{u} \\ \boldsymbol{\beta}^\top \mathbf{u} + \alpha \geq 1 - \xi - y(\mathbf{w}^\top (\mathbf{x} + \mathbf{z}) + b) + t, \forall -\mathbf{u} \leq \mathbf{z} \leq \mathbf{u}, \end{cases}$$

which can be reformulated as

$$\begin{cases} \mathbf{f}^\top \boldsymbol{\beta} + \alpha \leq \varepsilon \\ -\alpha \leq \min_{\mathbf{z}, \mathbf{u}} \boldsymbol{\beta}^\top \mathbf{u} \text{ s.t. } -\mathbf{u} \leq \mathbf{z} \leq \mathbf{u} \\ 1 - \xi - y(\mathbf{w}^\top \mathbf{x} + b) + t - \alpha \leq \min_{\mathbf{z}, \mathbf{u}} \boldsymbol{\beta}^\top \mathbf{u} + y\mathbf{w}^\top \mathbf{z} \text{ s.t. } -\mathbf{u} \leq \mathbf{z} \leq \mathbf{u}. \end{cases}$$

By duality and simple calculation, we have

$$\begin{cases} -\boldsymbol{\beta} \leq \mathbf{w} \leq \boldsymbol{\beta} \\ \alpha \geq 0 \\ \alpha \leq \varepsilon - \mathbf{f}^\top \boldsymbol{\beta} \\ \alpha \geq 1 - y(\mathbf{w}^\top \mathbf{x} + b) + t - \xi, \end{cases}$$

or equivalently

$$\begin{cases} \varepsilon \geq \mathbf{f}^\top |\mathbf{w}| \\ \xi \geq 1 - y(\mathbf{w}^\top \mathbf{x} + b) + t + \mathbf{f}^\top |\mathbf{w}| - \varepsilon. \end{cases}$$

Hence the constraints  $\sup_{\mathbb{P} \in \mathcal{D}(\mathbf{x})} \mathbb{E}_{\mathbb{P}}[[1 - \xi - y(\mathbf{w}^\top \tilde{\mathbf{x}} + b) + t]_+] \leq \gamma$  and  $-t\kappa + \gamma \leq 0$  are equivalent to

$$\begin{cases} \gamma \geq \sup_{\mathbf{f} \in \mathcal{F}} \mathbf{f}^\top |\mathbf{w}| \\ \xi \geq 1 - y(\mathbf{w}^\top \mathbf{x} + b) + \left(\frac{1}{\kappa} - 1\right)\gamma + \sup_{\mathbf{f} \in \mathcal{F}} \mathbf{f}^\top |\mathbf{w}| \end{cases}$$

Since  $0 < \kappa \leq 1$ , the constraints above are equivalent to  $\xi \geq 1 - y(\mathbf{w}^\top \mathbf{x} + b) + \frac{1}{\kappa} \sup_{\mathbf{f} \in \mathcal{F}} \mathbf{f}^\top |\mathbf{w}|$ , so that we obtain this theorem.  $\square$

## 4.8 Chapter Summary

In this chapter, we presented a new framework for robust classification based on distributionally robust optimization, and showed that distributionally robust optimization can be a powerful tool to design robust classification algorithms that appropriately handle uncertainties in training and testing data. In particular, we provided a distributionally robust optimization interpretation for the regularized SVM, i.e., the DRO formulation with the hinge loss and the mean absolute deviation ambiguity set is equivalent to the regularized SVM. We then proposed a new robust classification algorithm that is robust to feature corruption of test data and developed a new robust formulation for classification based on distributionally robust chance constraints.

# CHAPTER 5

## The Coherent Loss Function for Classification

A prediction rule in classification that aims to achieve the lowest probability of misclassification involves minimizing over a non-convex, 0-1 loss function, which is typically a computationally intractable optimization problem. To address the intractability, previous methods consider minimizing the *cumulative loss* – the sum of convex surrogates of the 0-1 loss of each sample. We revisit this paradigm and develop instead an *axiomatic* framework by proposing a set of salient properties on functions for classification and then propose the *coherent loss* approach, which is a tractable upper-bound of the empirical classification error over the *entire* sample set. We show that the proposed approach yields a strictly tighter approximation to the empirical classification error than any convex cumulative loss approach while preserving the convexity of the underlying optimization problem, and this approach for binary classification also has a robustness interpretation which builds a connection to robust SVMs.

### 5.1 Introduction

The goal of supervised learning is to predict an unobserved output value  $y$  from an observed input  $\mathbf{x}$ . This is achieved by learning a function relationship  $y \approx f(\mathbf{x})$  from a set of observed training examples  $\{(y_i, \mathbf{x}_i)\}_{i=1}^m$ . The quality of predictor  $f(\cdot)$  is often measured by some loss function  $\ell(f(\mathbf{x}), y)$ . A typical statistical setup in machine learning assumes that all training data and testing samples are i.i.d. samples drawn from an unknown distribution  $\mu$ , and the goal is to find a predictor  $f(\cdot)$  such that the expected loss  $\mathbb{E}_{(y, \mathbf{x}) \sim \mu} \ell(f(\mathbf{x}), y)$  is minimized.



Since  $\mu$  is unknown, the expected loss is often replaced by the empirical loss

$$L_{emp}(f) \triangleq \frac{1}{m} \sum_{i=1}^m \ell(f(\mathbf{x}_i), y_i). \quad (5.1)$$

Minimizing  $L_{emp}(f)$ , as well as numerous regularization based variants of it, is one of the fundamental cornerstones of statistical machine learning, e.g., [VL63, VC91, PRMN04].

This chapter focuses on binary classification problems, where  $y \in \{-1, +1\}$ . A point  $(y, \mathbf{x})$  is correctly predicted if  $\text{sign}(f(\mathbf{x})) = y$ , and its classification error is given by the 0-1 loss  $\ell(f(\mathbf{x}_i), y_i) = \mathbf{1}(y \neq \text{sign}(f(\mathbf{x}))) = \mathbf{1}(yf(\mathbf{x}) \leq 0)$ . Due to the non-convexity of the indicator function, minimizing the empirical classification error  $\sum_i \mathbf{1}(y_i f(\mathbf{x}_i) \leq 0)$  is known to be NP-hard even to approximate [ABSS97, BDEL03]. A number of methods have been proposed to mitigate this computational difficulty, all based on the idea that to minimize the ‘‘cumulative loss’’, which is the sum of individual losses given by,

$$L_\phi(f) \triangleq \frac{1}{m} \sum_{i=1}^m \phi(yf(\mathbf{x}_i))$$

where  $\phi(\cdot)$  is a convex upper bound of the classification error  $\mathbf{1}(yf(\mathbf{x}) \leq 0)$ . For example, AdaBoost [FS97, FHT00, SS99] employs the exponential loss function  $\exp(-yf(\mathbf{x}))$ , and Support Vector Machines (SVMs) [BGV92, CV95] employ a hinge-loss function  $\max\{1 - yf(\mathbf{x}), 0\}$ .

In this chapter we revisit this paradigm, and introduce a notion termed *coherent loss*, as opposed to cumulative loss used in the conventional approach. Briefly speaking, instead of using an upper bound of the *individual* classification error (the 0-1 loss), we propose to use a tractable upper bound of the *total* empirical classification error for the whole training set. That is, we look for  $\Phi : \mathfrak{R}^m \mapsto \mathfrak{R}$  such that

$$\Phi(c_1, \dots, c_m) \geq \frac{1}{m} \sum_{i=1}^m \mathbf{1}(c_i \leq 0), \quad \forall (c_1, \dots, c_m) \in \mathfrak{R}^m.$$

Intuitively, since coherent loss functions are more general than cumulative loss functions, one may expect to obtain a tighter and still tractable bound of the empirical classification error via coherent loss function. We formalize this intuition in this chapter. Specifically,

our contributions include the followings.

In Section 5.2, we consider a principled approach by formalizing the salient properties of functions, termed as *coherent classification loss* functions, that could be used to quantify the performance of a classification rule. These functions have dual-representations which enable us to identify the *minimal coherent classification loss* function, which, loosely speaking, is the coherent classification loss function that best approximates the 0-1 loss, which also achieves a tighter bound of the empirical classification error than any convex cumulative loss. We show that optimizing this function is equivalent to a convex optimization problem, and hence tractable.

In Section 5.3, we consider an equivalent form of the coherent loss function and then provide several applications of this loss function in classification problems. We remark that a tighter approximation of the 0-1 loss can potentially reduce the impact of outliers on the classification accuracy. Cumulative loss function may significantly deviate from the 0-1 loss when  $c \ll 0$ . Consequently, a misclassified outlier can incur a huge loss, and prevents an otherwise perfect prediction rule from being selected. This sensitivity can be mitigated by a tighter approximation.

Section 5.4 provides a statistical interpretation of minimizing the coherent loss function. Section 5.5 reports the experimental results which show that our classification method outperforms the standard SVM when additional constraints are imposed on the decision function.

**Notations:** We use boldface letters to represent column vectors, and capital letters for matrices. We reserve  $\mathbf{e}$  for special vectors:  $\mathbf{e}_i$  is the vector whose  $i$ -th entry is 1, and the rest are 0;  $\mathbf{e}_N$ , where  $N$  is an index set, is the vector that for all  $i \in N$ , the corresponding entry equals 1, and zero otherwise;  $\mathbf{e}^n \in \mathfrak{R}^n$  is the vector with all entries equal to 1. The  $i$ -th entry of a vector  $\mathbf{x}$  is denoted by  $x_i$ . We use  $[c]_+$  to denote  $\max\{0, c\}$  and  $\mathbf{1}[\cdot]$  to denote the indicator function, and let  $\mathcal{P}_n$  be the set of all  $n \times n$  permutation matrices and  $\mathbf{I}_n$  be the  $n \times n$  identity matrix.

## 5.2 Coherent Classification Loss Function

We now propose the notion of coherent classification loss functions based on an axiomatic approach. Along the way, we show the existence of a “tight” coherent classification loss function which can achieve better approximation of the empirical classification error than *any* convex cumulative loss. The definition of the coherent classification loss function is motivated from analyzing the salient properties of functions used to quantify the performance of a classification rule. A natural approach is to elicit these properties from the *classification error*. Specifically, given  $u_1, \dots, u_m$  where  $u_i$  the “decision value” of the  $i$ th sample, e.g.  $u_i = y_i f(\mathbf{x}_i)$ , we suppose that these  $m$  samples are divided into  $n$  groups  $\mathcal{G} = \{\mathcal{G}_1, \dots, \mathcal{G}_n\}$  satisfying that

$$\mathcal{G}_i \neq \emptyset, \mathcal{G}_i \subseteq \{1, \dots, m\}, \text{ and } \bigcup_{i=1}^n \mathcal{G}_i = \{1, \dots, m\}.$$

Denote by  $\mathcal{S}$  the set of all feasible groups  $\mathcal{G}$ . We define the classification error  $\varrho : \Re^m \times \mathcal{S} \mapsto [0, 1]$  as follows:

$$\varrho(u_1, \dots, u_m, \mathcal{G}) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}[u_j < 0, \exists j \in \mathcal{G}_i]. \quad (5.2)$$

The definition of  $\varrho(\cdot, \cdot)$  ensures that the classification error is nonzero as long as there exists a group so that one of its samples has a negative decision value. Clearly, when  $\mathcal{G}_i = \{i\}$ , (5.2) is reduced to the classical zero-one loss function:

$$\varrho(u_1, \dots, u_m) = \frac{1}{m} \sum_{i=1}^m \mathbf{1}[u_i < 0], \quad (5.3)$$

where the input  $\mathcal{G}$  is ignored for notational simplicity. We will next propose a set of properties and that functions endowed with these properties are known as coherent classification loss functions.

### 5.2.1 Salient Properties and Representation Theorem

We elicit the five salient properties from the classification error as follows. Consider  $\rho(\cdot, \cdot) : \Re^m \times \mathcal{S} \rightarrow [0, 1]$ .

**Property 1** (Complete classification).  $\rho(\mathbf{u}, \mathcal{G}) = 0$  if and only if  $\mathbf{u} \geq \mathbf{0}$ .

Complete classification essentially says if every sample is correctly classified, then it is optimal.

**Property 2** (Misclassification avoidance). *If  $\mathbf{u} < \mathbf{0}$ , then  $\rho(\mathbf{u}, \mathcal{G}) = 1$ .*

This property states that if all samples are misclassified, then it is the worst classification and hence  $\rho(\cdot, \cdot)$  achieves the maximal value.

**Property 3** (Monotonicity). *If  $\min_{j \in \mathcal{G}_i} u_j \geq \min_{j \in \mathcal{G}_i} w_j$  for all  $i = 1, \dots, n$ , then  $\rho(\mathbf{u}, \mathcal{G}) \leq \rho(\mathbf{w}, \mathcal{G})$ .*

Monotonicity requires that if a decision better classifies every group of samples, then it is more desirable. When  $G_i = i$ , this property simply means that  $\mathbf{u} \geq \mathbf{v}$  implies  $\rho(\mathbf{u}) \leq \rho(\mathbf{v})$ .

**Property 4** (Order invariance). *For any permutations  $\pi$  of  $\{1, \dots, m\}$  and  $\tau$  of  $\{1, \dots, n\}$ , let  $\tilde{\mathcal{G}}_{\tau(i)} = \{\pi(j) : j \in \mathcal{G}_i\}$  for  $i = 1, \dots, n$  and  $\tilde{\mathbf{u}} = (u_{\pi(1)}, \dots, u_{\pi(m)})^\top$ , then we have  $\rho(\mathbf{u}, \mathcal{G}) = \rho(\tilde{\mathbf{u}}, \tilde{\mathcal{G}})$ .*

Order invariance essentially states that the order of the samples does not matter. This is natural in the classification problem, since each sample is drawn i.i.d., and is treated equally.

**Property 5** (Scale invariance). *For all  $\alpha > 0$ ,  $\rho(\alpha\mathbf{u}, \mathcal{G}) = \rho(\mathbf{u}, \mathcal{G})$ .*

Scale invariance is a property that the classification error function satisfies. It essentially means that changing the scale does not affect the preference between classifiers. While it may be debatable whether scale invariance is as necessary as other properties, indeed as we show later in this section, this property can be relaxed.

**Definition 5.1** (Coherent Classification Loss). *A function  $\rho(\mathbf{u}, \mathcal{G}) : \mathbb{R}^m \times \mathcal{S} \rightarrow [0, 1]$  is a coherent classification loss function (CCLF) if it satisfies Property 1 to 5, and is quasi-convex and lower semi-continuous w.r.t.  $\mathbf{u}$ .*

Here, quasi-convexity and semi-continuity are introduced to for tractability. Our first result is a (dual) representation theorem of any CCLF. We need the following definition first.

**Definition 5.2** (Admissible Class). *A class of sets  $\mathbf{V}_k \subseteq \mathbb{R}^n$  parameterized by  $k \in [0, 1]$  is called admissible class, if they satisfy the following properties:*

1. *For any  $k \in [0, 1]$ ,  $\mathbf{V}_k$  is a closed, convex cone, and is order invariant. Here, being order invariant means that  $\mathbf{v} \in \mathbf{V}$  implies  $\mathbf{P}\mathbf{v} \in \mathbf{V}$  for any  $\mathbf{P} \in \mathcal{P}_n$ ;*

2.  $k \leq k'$  implies  $\mathbf{V}_k \subseteq \mathbf{V}_{k'}$ ;
3.  $\mathbf{V}_1 = \text{cl}(\lim_{k \uparrow 1} \mathbf{V}_k)$  and  $\mathbf{V}_0 = \lim_{k \downarrow 0} \mathbf{V}_k$ .
4.  $\mathbf{V}_1 = \mathfrak{R}_+^n$ ;
5. For any  $\lambda > 0$ , we have  $\lambda \mathbf{e}^n \in \mathbf{V}_0$ .

**Theorem 5.1** (Representation Theorem). *A function  $\rho(\cdot, \cdot)$  is a CCLF if and only if it can be written as*

$$\rho(\mathbf{u}, \mathcal{G}) = 1 - \sup\{k \in [0, 1] : \sup_{\mathbf{v} \in \mathbf{V}_k} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \leq 0, \tilde{\mathbf{u}} = (\min_{j \in \mathcal{G}_1} u_j, \dots, \min_{j \in \mathcal{G}_n} u_j)^\top\}, \quad (5.4)$$

for an admissible class  $\{\mathbf{V}_k\}$ . Here sup over an empty set is set as 0.

*Proof.* We sketch the proof and leave the details in the appendix. The “if” part is relatively easy, by checking that any function  $\rho(\mathbf{u}, \mathcal{G}) = 1 - \sup\{k \in [0, 1] : \sup_{\mathbf{v} \in \mathbf{V}_k} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \leq 0, \tilde{\mathbf{u}} = (\min_{j \in \mathcal{G}_1} u_j, \dots, \min_{j \in \mathcal{G}_n} u_j)^\top\}$  for some admissible class  $\{\mathbf{V}_k\}$  satisfies all properties required for a CCLF.

The “only if” part requires more work. We want to show that given a function  $\rho(\cdot, \cdot)$  which is a CCLF, it can be represented as (5.4) for some admissible class  $\{\mathbf{V}_k\}$ . The proof consists of three steps: We first show that  $\rho(\cdot, \cdot)$  can be represented as  $\rho(\mathbf{u}, \mathcal{G}) = 1 - \sup\{k \in [0, 1] : \sup_{\mathbf{v} \in \overline{\mathbf{V}}_k} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \leq 0, \tilde{\mathbf{u}} = (\min_{j \in \mathcal{G}_1} u_j, \dots, \min_{j \in \mathcal{G}_n} u_j)^\top\}$ , for some  $\{\overline{\mathbf{V}}_k\}$ , not necessarily admissible. This essentially follows from a result in [BS09]. We then show that we can replace  $\overline{\mathbf{V}}_k$  by a class of closed, convex, order-invariant, cones  $\mathbf{V}_k$ . Specifically, we can pick  $\mathbf{V}_k \triangleq \text{cl}(\text{cc}(\text{or}(\overline{\mathbf{V}}_k)))$ , where  $\text{or}(\cdot)$  (respectively  $\text{cc}(\cdot)$ ) is the minimal **order** invariant (respectively, **convex cone**) superset. Finally we show that  $\{\mathbf{V}_k\}$  is admissible, by checking that all properties in Definition 5.2 are satisfied, to complete the proof.  $\square$

### 5.2.2 Minimal Coherent Classification Loss Function

This section shows that among all CCLF functions that upper-bound the classification error, there exists a minimal (i.e., best) one.

**Theorem 5.2.** Define  $\bar{\rho}(\cdot, \cdot) : \mathfrak{R}^m \times \mathcal{S} \mapsto [0, 1]$  as follows

$$\bar{\rho}(\mathbf{u}, \mathcal{G}) = \frac{\max\{t : \sum_{i=1}^t \tilde{u}_{(i)} < 0\}}{n},$$

where  $\tilde{\mathbf{u}} = (\min_{j \in \mathcal{G}_1} u_j, \dots, \min_{j \in \mathcal{G}_n} u_j)^\top$ ,  $\{\tilde{u}_{(i)}\}$  is a permutation of  $\{\tilde{u}_i\}$  in a non-decreasing order, and max over an empty set is taken as zero. Then the following holds.

1.  $\bar{\rho}(\cdot, \cdot)$  is a CCLF, and is an upper-bound of the classification error, i.e.,  $\bar{\rho}(\mathbf{u}, \mathcal{G}) \geq \varrho(\mathbf{u}, \mathcal{G})$ ,  $\forall \mathbf{u} \in \mathfrak{R}^m$ .
2. Let  $\bar{\mathbf{V}}_k \subset \mathfrak{R}^n$  satisfy that if  $k = 0$ , then  $\bar{\mathbf{V}}_k = \text{conv}\{\lambda \mathbf{e}^n | \lambda > 0\}$ ; and if  $\frac{s}{n} < k \leq \frac{s+1}{n}$  for  $s = 0, \dots, n-1$ , then

$$\bar{\mathbf{V}}_k = \text{conv}\{\lambda \mathbf{e}_N | \forall \lambda > 0, \forall N : |N| = n - s\},$$

where  $N$  is an index set. Then  $\{\bar{\mathbf{V}}_k\}$  is an admissible class corresponding to  $\bar{\rho}(\cdot, \cdot)$ .

3.  $\bar{\rho}(\cdot, \cdot)$  is the tightest CCLF bound. That is, if  $\rho'(\cdot, \cdot)$  is a CCLF function and satisfies  $\rho'(\mathbf{u}, \mathcal{G}) \geq \varrho(\mathbf{u}, \mathcal{G})$  for all  $\mathbf{u} \in \mathfrak{R}^m$ , then  $\rho'(\mathbf{u}, \mathcal{G}) \geq \bar{\rho}(\mathbf{u}, \mathcal{G})$  for all  $\mathbf{u} \in \mathfrak{R}^m$ .

*Proof.* We provide a sketch of the proof and leave the details to the appendix. Claim 1 is relatively straightforward. It is also easy to see that  $\bar{\mathbf{V}}_k$  is an admissible set. So one only needs to show that  $\bar{\mathbf{V}}_k$  is the set corresponding to  $\bar{\rho}(\cdot, \cdot)$ , to establish Claim 2. To show Claim 3, we let  $\{\mathbf{V}'_k\}$  be an admissible class corresponding to  $\rho'(\cdot, \cdot)$ , and show that  $\lambda \mathbf{e}_N \in \mathbf{V}'_k$ , which further implies  $\bar{\mathbf{V}}_k \subseteq \mathbf{V}'_k$ . This establishes claim 3.  $\square$

We next show that *scale invariance* can be relaxed. Indeed, for any quasi-convex upper bound of classification error that satisfies other properties, the minimal CCLF is a tighter bound.

**Theorem 5.3.** Let  $\hat{\rho}(\mathbf{u}, \mathcal{G}) : \mathfrak{R}^m \times \mathcal{S} \mapsto [0, 1]$  be a quasi-convex function w.r.t.  $\mathbf{u}$  that satisfies complete classification, misclassification avoidance, monotonicity, order invariance, and that  $\hat{\rho}(\mathbf{u}, \mathcal{G}) \geq \varrho(\mathbf{u}, \mathcal{G})$ . Then there exists a CCLF  $\rho(\cdot, \cdot)$  such that

$$\varrho(\mathbf{u}, \mathcal{G}) \leq \bar{\rho}(\mathbf{u}, \mathcal{G}) \leq \rho(\mathbf{u}, \mathcal{G}) \leq \hat{\rho}(\mathbf{u}, \mathcal{G}), \quad \forall \mathbf{u} \in \mathfrak{R}^m \text{ and } \mathcal{G} \in \mathcal{S}.$$

*Proof.* We sketch the proof. The main idea is to construct such a function

$$\rho(\mathbf{u}, \mathcal{G}) \triangleq \lim_{\epsilon \downarrow 0} [\min_{\gamma > 0} \hat{\rho}((\mathbf{u} + \epsilon)/\gamma, \mathcal{G})],$$

and show that  $\rho(\cdot, \cdot)$  is a CCLF and  $\hat{\rho}(\mathbf{u}, \mathcal{G}) \geq \rho(\mathbf{u}, \mathcal{G}) \geq \varrho(\mathbf{u}, \mathcal{G})$ . Finally, since  $\bar{\rho}(\cdot, \cdot)$  is the minimal CCLF, this completes the proof.  $\square$

One important property of  $\bar{\rho}(\cdot, \cdot)$  is that it achieves better approximation of the empirical classification error than any *convex cumulative loss*.

**Theorem 5.4.** *If  $f(\cdot)$  is a convex function and an upper bound of the 0-1 loss function, then for any  $\mathbf{u} = (u_1, \dots, u_m)$  and  $\mathcal{G} \in \mathcal{S}$ , we have  $\varrho(\mathbf{u}, \mathcal{G}) \leq \bar{\rho}(\mathbf{u}, \mathcal{G}) \leq \frac{1}{n} \sum_{i=1}^n f(\tilde{u}_i)$  where  $\tilde{\mathbf{u}} = (\min_{j \in \mathcal{G}_1} u_j, \dots, \min_{j \in \mathcal{G}_n} u_j)^\top$ .*

*Proof.* Without loss of generality, assume  $(\tilde{u}_1, \dots, \tilde{u}_m)$  are in a non-decreasing order. Let  $p \triangleq \max\{i : \tilde{u}_i < 0\}$  and  $q \triangleq \max\{t : \sum_{i=1}^t \tilde{u}_i < 0\}$ , then  $\sum_{i=1}^q \tilde{u}_i = \sum_{i=1}^p \tilde{u}_i + \sum_{i=p+1}^q \tilde{u}_i < 0$ . Since  $f(\cdot)$  is convex and  $f(x) \geq \mathbf{1}[x \leq 0]$ , there exists  $k \leq 0$  such that  $f(x) \geq \max\{kx + 1, 0\}$  (this can be done for example by taking  $k$  as a subgradient of  $f(x)$  at  $x = 0$ ). If  $k = 0$ , then  $\frac{1}{n} \sum_{i=1}^n f(\tilde{u}_i) \geq 1 \geq \bar{\rho}(\mathbf{u}, \mathcal{G})$ , the theorem holds. Otherwise  $k < 0$ , we have

$$\begin{aligned} \sum_{i=1}^n f(\tilde{u}_i) &\geq \sum_{i=1}^p (k\tilde{u}_i + 1) + \sum_{i=p+1}^n f(\tilde{u}_i) = p + k \sum_{i=1}^p \tilde{u}_i + \sum_{i=p+1}^n f(\tilde{u}_i) \\ &> p - k \sum_{i=p+1}^q \tilde{u}_i + \sum_{i=p+1}^n f(\tilde{u}_i) \geq p + \sum_{i=p+1}^q (f(\tilde{u}_i) - k\tilde{u}_i). \end{aligned}$$

Note that  $\tilde{u}_i \geq 0$  for  $i = p+1, \dots, m$ , then if  $\tilde{u}_i \geq -\frac{1}{k}$ ,  $f(\tilde{u}_i) - k\tilde{u}_i \geq -k\tilde{u}_i \geq 1$ . Otherwise  $f(\tilde{u}_i) - k\tilde{u}_i \geq k\tilde{u}_i + 1 - k\tilde{u}_i = 1$ . Hence,  $p + \sum_{i=p+1}^q (f(\tilde{u}_i) - k\tilde{u}_i) \geq p + (q - p) = q$ . By the definition of  $\bar{\rho}(\mathbf{u}, \mathcal{G})$ , the theorem holds.  $\square$

### 5.2.3 Optimization With the Coherent Loss Function

We now discuss the computational issue of optimization of the minimal CCLF  $\bar{\rho}(\cdot, \cdot)$ . Indeed, we show that this can be converted to a tractable convex optimization problem. Specifically,

for fixed  $\mathcal{G}$ , we consider the following problem on variables  $(\mathbf{u}, \mathbf{w})$ :

$$\begin{aligned} \min \quad & \bar{\rho}(\mathbf{u}, \mathcal{G}) \\ \text{s.t.} \quad & f_j(\mathbf{u}, \mathbf{w}) \leq 0, \quad j = 1, \dots, k, \end{aligned} \tag{5.5}$$

where  $f_j(\cdot, \cdot)$  are convex functions. We have the following theorem.

**Theorem 5.5.** *Assume that all the feasible solutions  $(\mathbf{u}, \mathbf{w})$  to Problem 5.5 satisfy that  $\mathbf{u} > \mathbf{0}$  or  $\mathbf{u} \not\geq \mathbf{0}$ . Let  $(\mathbf{s}^*, \mathbf{t}^*, h^*)$  be an optimal solution to the following optimization problem:*

$$\begin{aligned} \min_{h, \mathbf{s}, \mathbf{t}} \quad & \frac{1}{n} \sum_{i=1}^n [1 - \min_{j \in \mathcal{G}_i} s_j]_+ \\ \text{s.t.} \quad & h f_j(\mathbf{s}/h, \mathbf{t}/h) \leq 0, \quad j = 1, \dots, k; \\ & h > 0. \end{aligned} \tag{5.6}$$

Then  $(\mathbf{s}^*/h^*, \mathbf{t}^*/h^*)$  is an optimal solution to Problem (5.5).

*Proof.* We provide a sketch of the proof. We first show that the level set of Problem (5.5), i.e.,  $\mathcal{U}_i \triangleq \{(\mathbf{u}, \mathbf{w}) : \bar{\rho}(\mathbf{u}) \leq 1 - i/n; f_j(\mathbf{u}, \mathbf{w}) \leq 0, \forall j\}$  for  $i = 1, \dots, n$ , equals the following

$$\{(\mathbf{u}, \mathbf{w}) \mid \exists d : \sum_{i=1}^n [d - \min_{j \in \mathcal{G}_i} u_j]_+ \leq (n - i + 1)d \text{ and } f_j(\mathbf{u}, \mathbf{w}) \leq 0, \forall j.\}$$

This can be proved by applying the Theorem 5.2, and then using duality of linear program.

This set can further be shown equivalent to the feasible set of

$$\begin{aligned} \sum_{i=1}^n [1 - \frac{\min_{j \in \mathcal{G}_i} u_j}{d}]_+ &\leq (m - i + 1) \\ f_j(\mathbf{u}, \mathbf{w}) &\leq 0, \quad j = 1, \dots, k. \end{aligned} \tag{5.7}$$

Thus, finding the optimal solution to Problem (5.5) is equivalent to solve the following problem

$$\begin{aligned} \min \quad & \sum_{i=1}^m [1 - \frac{\min_{j \in \mathcal{G}_i} u_j}{d}]_+ \\ \text{s.t.} \quad & f_j(\mathbf{u}, \mathbf{w}) \leq 0, \quad j = 1, \dots, n; \\ & d > 0. \end{aligned} \tag{5.8}$$



Then let  $h = 1/d$ ,  $\mathbf{s} = h\mathbf{u}$  and  $\mathbf{t} = h\mathbf{w}$ , the theorem is established.  $\square$

Notice that  $hf_j(\mathbf{s}/h, \mathbf{t}/h)$  is the perspective function of  $f_j(\cdot, \cdot)$ , and is hence jointly convex to  $(h, \mathbf{s}, \mathbf{t})$  [BV04]. Thus, Problem (5.6) is equivalent to a tractable convex optimization problem.

### 5.3 Equivalent Formulation and Applications

From Theorem 5.5, when there is no  $(\mathbf{u}, \mathbf{w})$  such that  $\mathbf{u} \geq \mathbf{0}$  and  $f_j(\mathbf{u}, \mathbf{w}) \leq 0$  for  $j = 1, \dots, k$ , Problem (5.5) is equivalent to minimizing the following optimization problem:

$$\begin{aligned} \min \quad & \Phi(\mathbf{u}, \mathcal{G}) \\ \text{s.t.} \quad & f_j(\mathbf{u}, \mathbf{w}) \leq 0, \quad j = 1, \dots, k, \end{aligned} \tag{5.9}$$

where  $\Phi(\mathbf{u}, \mathcal{G})$  is defined by

$$\Phi(\mathbf{u}, \mathcal{G}) \triangleq \min_{\gamma > 0} \frac{1}{n} \sum_{i=1}^n [1 - \min_{j \in \mathcal{G}_i} u_j / \gamma]_+. \tag{5.10}$$

From this formulation, we also show, from another perspective, that minimizing the coherent loss function is equivalent to minimizing a “tighter” upper bound of the 0-1 loss function, or in other words, the coherent loss function achieves better approximation of the empirical classification error than any convex cumulative loss.

**Theorem 5.6.** *Let  $\phi : \mathfrak{R} \mapsto \mathfrak{R}^+$  be a non-increasing, convex function that satisfies*

$$\phi(c) \geq \mathbf{1}(c \leq 0), \quad \forall c \in \mathfrak{R}.$$

For all  $\mathbf{u} \in \mathfrak{R}^m$ , let  $\tilde{\mathbf{u}} = (\min_{j \in \mathcal{G}_1} u_j, \dots, \min_{j \in \mathcal{G}_n} u_j)^\top$ , then we have

$$\frac{1}{n} \sum_{i=1}^n \mathbf{1}(\tilde{u}_i \leq 0) \leq \Phi(\mathbf{u}, \mathcal{G}) \leq \frac{1}{n} \sum_{i=1}^n \phi(\tilde{u}_i).$$

*Proof.* Recall that the hinge-loss  $\phi_1^*(c) \triangleq [1 - c]_+$  is the tightest convex bound of 0-1 loss which has a derivative (or sub-gradient)  $-1$  at  $c = 0$  (e.g., [SS02]). That is, if a convex

function  $\phi(\cdot)$  satisfies  $\phi(c) \geq \mathbf{1}(c \leq 0), \forall c$ , and also satisfies  $-1 \in \partial\phi(0)$ , then  $\phi_1^*(c) \leq \phi(c)$  for all  $c$ . Similarly,  $\phi_\gamma^*(c) \triangleq \max[1 - c/\gamma]_+$  is the tightest convex bound of 0-1 loss with a derivative  $-1/\gamma$  at  $x = 0$ . Since  $\phi(\cdot)$  is non-increasing, it can not have positive derivative at  $c = 0$ . Thus,  $\Phi(\cdot, \cdot)$  is a tighter bound than any non-increasing, convex cumulative loss functions.  $\square$

When  $\mathcal{G} = \{\{1\}, \{2\}, \dots, \{m\}\}$ , Problem (5.9) is equivalent to the following convex optimization problem

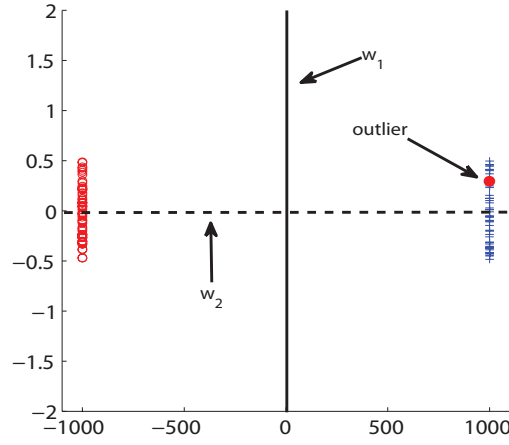
$$\begin{aligned} \min_{h, \mathbf{s}, \mathbf{t}} \quad & \frac{1}{m} \sum_{i=1}^m [1 - s_i]_+ \\ \text{s.t.} \quad & hf_j(\mathbf{s}/h, \mathbf{t}/h) \leq 0, \quad j = 1, \dots, n; \\ & h > 0, \end{aligned} \tag{5.11}$$

which can be solved efficiently. We now provide some applications of the proposed coherent loss function.

At first, we illustrate with an example, that the proposed coherent loss function can be more robust to outliers. Let  $\mathbf{u}_1, \mathbf{u}_2 \in \mathfrak{R}^{100}$  be the followings:  $\mathbf{u}_1 = (-1000, 1000, 1000, \dots, 1000)$ , and  $\mathbf{u}_2 = (+1, -1, +1, -1, \dots, +1, -1)$ . In this case,  $\mathbf{u}_2$  appears to be a less favorable classification since 50% of samples are misclassified. It is easy to check that  $\mathbf{u}_1$  incurs a much larger hinge-loss than  $\mathbf{u}_2$ , even though only one sample is misclassified. In contrast, the coherent loss of  $\mathbf{u}_1$  is no more than 0.02 (take  $\gamma = 1/1000$ ), and that of  $\mathbf{u}_2$  is at least 0.5 (since 50% samples are misclassified, and the coherent loss is an upper bound). Thus, the coherent loss is more robust in this example, partly because it better approximates the 0-1 loss, and hence is less affected by large outliers. See Figure 5.1.

### Example: linear SVM

We illustrate the proposed method with the linear classification problem, and in particular, the linear Support Vector Machines algorithm (SVMs) [BGV92, CV95, SS02]. Given  $m$  training samples  $(y_i, \mathbf{x}_i)_{i=1}^m$ , the goal is to find a hyperplane that correctly classify as many



**Figure 5.1:** Illustration of the effect of outliers to the cumulative loss vs the coherent loss. Here,  $\mathbf{w}_1$  has a margin  $\mathbf{u}_1$ , and  $\mathbf{w}_2$  has a margin  $\mathbf{u}_2$ . The cumulative loss approach will pick  $\mathbf{w}_2$ , where the proposed method will pick  $\mathbf{w}_1$ , which is a better classification.

training samples as possible with a large margin, which leads to the following formulation:

$$\begin{aligned} \min \quad & \frac{1}{m} \sum_{i=1}^m \mathbf{1}[y_i(\mathbf{w}^\top \mathbf{x}_i + b) \leq 0] \\ \text{s.t.} \quad & \|\mathbf{w}\|_2 \leq C \end{aligned} \quad (5.12)$$

for a given  $C > 0$ . Since the objective function is non-convex, Problem (5.12) is an intractable problem. Hence, SVM uses the hinge-loss function  $\phi_1^*(c) = [1 - c]_+$  as a convex surrogate.

Following the proposed coherent loss function approach, we minimize the 0-1 loss function with margin  $a \geq 0$ :  $\frac{1}{m} \sum_{i=1}^m \mathbf{1}[y_i(\mathbf{w}^\top \mathbf{x}_i + b) \leq a]$  and replace this objective function by the coherent loss function  $\bar{\rho}(\mathbf{u})$  where  $u_i = y_i(\mathbf{w}^\top \mathbf{x}_i + b) - a$  (Margin  $a$  makes the condition in Theorem 5.5 hold, and the approximation of this 0-1 loss function by using the hinge-loss function still leads to the standard SVM). Then we obtain the following formulation,

$$\begin{aligned} \min_{\mathbf{w}, b, \gamma > 0} \quad & \frac{1}{m} \sum_{i=1}^m [1 - (y_i(\mathbf{w}^\top \mathbf{x}_i + b) - a)/\gamma]_+ \\ \text{s.t.} \quad & \|\mathbf{w}\|_2 \leq C. \end{aligned} \quad (5.13)$$

As discussed above, we can change variables  $h = 1/\gamma$ ,  $\hat{\mathbf{w}} = \mathbf{w}/\gamma$  and  $\hat{b} = b/\gamma$ , and simplify Formulation (5.13) as the following:

$$\begin{aligned} \min_{\hat{\mathbf{w}}, \hat{b}, h > 0} \quad & \frac{1}{m} \sum_{i=1}^m [1 + ah - y_i(\hat{\mathbf{w}}^\top \mathbf{x}_i + \hat{b})]_+ \\ \text{s.t.} \quad & \|\hat{\mathbf{w}}\|_2 \leq hC. \end{aligned}$$

An interesting thing is that this formulation is also equivalent to the robust formulation of SVM [SBS06], which provides another interpretation for the coherent loss function.

**Proposition 5.1.** *Problem (5.13) is equivalent to the following optimization problem:*

$$\begin{aligned} \min_{\mathbf{w}, b, \xi} \quad & \sum_{i=1}^m \xi_i \\ \text{s.t.} \quad & \inf_{\tilde{\mathbf{x}}_i \sim (\mathbf{x}_i, \mathbf{I})} \mathbb{P}[y_i(\mathbf{w}^\top \tilde{\mathbf{x}}_i + b) \geq 1 - \xi_i] \geq 1 - \kappa, \quad i = 1, \dots, m, \end{aligned} \tag{5.14}$$

where  $\tilde{\mathbf{x}}_i \sim (\mathbf{x}_i, \mathbf{I})$  denotes a family of distributions which have a common mean  $\mathbf{x}_i$  and covariance  $\mathbf{I}$ , and  $\kappa = a^2/(a^2 + C^2)$ .

*Proof.* Theorem 1 in [SBS06] shows that Problem 5.14 is equivalent to

$$\begin{aligned} \min_{\mathbf{w}, b, \xi} \quad & \sum_{i=1}^m \xi_i \\ \text{s.t.} \quad & y_i(\mathbf{w}^\top \tilde{\mathbf{x}}_i + b) \geq 1 - \xi_i + \gamma \|\mathbf{w}\|_2, \quad i = 1, \dots, m, \\ & \xi_i \geq 0, \quad i = 1, \dots, m, \end{aligned}$$

where  $\gamma = \sqrt{\kappa/(1 - \kappa)}$ . When  $\kappa = a^2/(a^2 + C^2)$ , we have  $\gamma = a/C$ , which implies that the formulation above is equivalent to

$$\min_{\mathbf{w}, b, \xi} \quad \sum_{i=1}^m [1 + \frac{a}{C} \|\mathbf{w}\|_2 - y_i(\mathbf{w}^\top \tilde{\mathbf{x}}_i + b)]_+.$$

Therefore, by moving  $\frac{a}{C} \|\mathbf{w}\|_2$  into the constraint, we obtain this result.  $\square$

We next consider the case where one may like to impose additional constraints on  $\mathbf{w}$ . For instance, if the first feature is measured from a less reliable source, then an ideal classification

rule should discount the importance of the first feature, by imposing a constraint like  $|w_1| \leq 0.001$ . Thus, the linear classification problem becomes

$$\begin{aligned} \min_{\mathbf{w}, b} \quad & \frac{1}{m} \sum_{i=1}^m \mathbf{1}[y_i(\mathbf{w}^\top \mathbf{x}_i + b) \leq a] \\ \text{s.t.} \quad & \|\mathbf{w}\|_2 \leq C \\ & A\mathbf{w} \leq \mathbf{d}. \end{aligned}$$

Using the coherent loss to replace the objective function, and simplifying the resulting formulation, we obtain the following second order cone program

$$\begin{aligned} \min_{\hat{\mathbf{w}}, \hat{b}, h > 0} \quad & \frac{1}{m} \sum_{i=1}^m [1 + ah - y_i(\hat{\mathbf{w}}^\top \mathbf{x}_i + \hat{b})]_+ \\ \text{s.t.} \quad & \|\hat{\mathbf{w}}\|_2 - Ch \leq 0 \\ & A\hat{\mathbf{w}} \leq \mathbf{d}h. \end{aligned}$$

Finally, we remark that the coherent loss approach can be kernelized, since a representation theorem [SS02] still holds if the coherent loss function is used.

### Example: Multi-class SVM

The coherent loss function can also be applied in multi-class classification problems. The main idea of previous approaches [LS06, LLW04, CS02] of multi-class SVMs is solving one single regularization problem by imposing a penalty on the values of  $f_y(\mathbf{x}) - f_z(\mathbf{x})$  for sample  $(\mathbf{x}, y)$  where  $f_y(\cdot)$  and  $f_z(\cdot)$  are decision function for class  $y$  and  $z$ , respectively. Suppose that the training samples are drawn from  $k$  different classes and the decision function  $f_y(\mathbf{x}) = \mathbf{w}_y^\top \mathbf{x} + b_y$  for each  $y = 1, \dots, k$ . Consider the following 0-1 loss penalty formulation:

$$\begin{aligned} \min_{f_i} \quad & \frac{1}{m} \sum_{i=1}^m \mathbf{1} \left[ \min_{z \in [k], z \neq y_i} \{f_{y_i}(\mathbf{x}_i) - f_z(\mathbf{x}_i)\} \leq a \right] \\ \text{s.t.} \quad & G_i(\mathbf{w}_i) \leq C, \quad i = 1, \dots, k \\ & \sum_{i=1}^k f_i = 0, \end{aligned}$$

where  $\sum_{i=1}^k f_i = (\sum_{i=1}^k \mathbf{w}_i, \sum_{i=1}^k b_i)$ ,  $G_i(\cdot)$  is convex (e.g.  $G_i(\cdot) = \|\cdot\|_2$ ) and margin  $a \geq 0$ , then we can apply the coherent loss function approach to make an approximation:

$$\begin{aligned} \min_{f_i, \gamma > 0} \quad & \frac{1}{m} \sum_{i=1}^m \left[ 1 - \frac{\min_{z \in [k], z \neq y_i} \{f_{y_i}(\mathbf{x}_i) - f_z(\mathbf{x}_i)\} - a}{\gamma} \right]_+ \\ \text{s.t.} \quad & G_i(\mathbf{w}_i) \leq C, \quad i = 1, \dots, k \\ & \sum_{i=1}^k f_i = 0, \end{aligned}$$

which can be simplified as the following:

$$\begin{aligned} \min_{\hat{f}_i, h > 0} \quad & \frac{1}{m} \sum_{i=1}^m \left[ 1 + ah + \max_{z \in [k], z \neq y_i} \{\hat{f}_z(\mathbf{x}_i) - \hat{f}_{y_i}(\mathbf{x}_i)\} \right]_+ \\ \text{s.t.} \quad & hG_i(\hat{\mathbf{w}}_i/h) \leq hC, \quad i = 1, \dots, k \\ & \sum_{i=1}^k \hat{f}_i = 0, \end{aligned}$$

where  $\hat{f}_i(\mathbf{x}) = \hat{\mathbf{w}}_i^\top \mathbf{x} + \hat{b}_i$ . Clearly, this is a convex optimization problem and can be solved efficiently.

## 5.4 Statistical Interpretation

In this section, we provide a statistical interpretation of minimizing the coherent loss function. As standard in learning theory, we assume that the training samples are drawn i.i.d. from an unknown distribution  $\mathbb{P}$ , and the goal is to find a predictor  $f(\cdot)$  such that the classification error of  $f$  given below is as small as possible:

$$L(f(\cdot)) = \mathbb{E}_{(\tilde{\mathbf{x}}, \tilde{y}) \sim \mathbb{P}} [I(f(\tilde{\mathbf{x}}), \tilde{y})].$$

Here  $(\tilde{\mathbf{x}}, \tilde{y}) \sim \mathbb{P}$  means sample  $(\tilde{\mathbf{x}}, \tilde{y})$  follows the distribution  $\mathbb{P}$ , and  $I(f(\tilde{\mathbf{x}}), \tilde{y}) = \mathbf{1}[\tilde{y}f(\tilde{\mathbf{x}}) \leq 0]$ . Recall that when  $\mathcal{G} = \{\{1\}, \{2\}, \dots, \{m\}\}$ , minimizing the coherent loss function is

equivalent to minimizing the following function

$$\Phi(\mathbf{u}) = \min_{\gamma > 0} \frac{1}{m} \sum_{i=1}^m \phi_{\gamma}(u_i),$$

where  $\phi_{\gamma}(u) = [1 - u/\gamma]_+$ . Let  $\eta(\mathbf{x}) = \mathbb{P}[\tilde{y} = 1 | \tilde{\mathbf{x}} = \mathbf{x}]$  be the probability that sample  $\mathbf{x}$  belongs to the first class, then the optimal Bayes error  $L^* = L(2\eta(\cdot) - 1)$ . We now develop an upper bound of the difference between  $L(f(\cdot))$  and  $L^*$  by using similar techniques in [Zha04].

For fixed  $\gamma$ , denote the expected loss of  $f(\cdot)$  w.r.t  $\phi_{\gamma}(\cdot)$  by

$$Q_{\gamma}(f(\cdot)) = \mathbb{E}_{(\tilde{\mathbf{x}}, \tilde{y}) \sim \mathbb{P}}[\phi_{\gamma}(\tilde{y}f(\tilde{\mathbf{x}}))],$$

and define two quantities

$$Q_{\gamma}(\eta, f) = \eta\phi_{\gamma}(f) + (1 - \eta)\phi_{\gamma}(-f), \quad \Delta Q_{\gamma}(\eta, f) = Q_{\gamma}(\eta, f) - Q_{\gamma}(\eta, f_{\gamma}^*(\eta)),$$

where  $f_{\gamma}^*(\eta) = \arg \min_f Q_{\gamma}(\eta, f)$ . By simple calculation, we know that  $f_{\gamma}^*(\eta) = \text{sign}(2\eta - 1)\gamma$  when  $\phi_{\gamma}(u) = [1 - u/\gamma]_+$ . Then we have the following lemma.

**Lemma 5.1.** *For  $\gamma > 0$ , we have  $\Delta Q_{\gamma}(\eta, 0) = |2\eta - 1|$ .*

*Proof.* From the definition of  $Q_{\gamma}(\eta, f)$  and  $\Delta Q_{\gamma}(\eta, f)$ , we have

$$\begin{aligned} \Delta Q_{\gamma}(\eta, f) &= \eta(\phi_{\gamma}(f) - \phi_{\gamma}(f_{\gamma}^*(\eta))) + (1 - \eta)(\phi_{\gamma}(-f) - \phi_{\gamma}(-f_{\gamma}^*(\eta))) \\ &= \eta[1 - f/\gamma]_+ + (1 - \eta)[1 + f/\gamma]_+ - \eta[1 - \text{sign}(2\eta - 1)]_+ - (1 - \eta)[1 + \text{sign}(2\eta - 1)]_+ \\ &= \eta[1 - f/\gamma]_+ + (1 - \eta)[1 + f/\gamma]_+ - 1 + |2\eta - 1|. \end{aligned}$$

This implies that  $\Delta Q_{\gamma}(\eta, 0) = |2\eta - 1|$ . □

By applying Lemma 5.1, we can bound the classification error of  $f(\cdot)$  w.r.t  $\phi_{\gamma}(\cdot)$  in terms of  $\mathbb{E}_{\tilde{\mathbf{x}}} \Delta Q_{\gamma}(\eta(\tilde{\mathbf{x}}), f(\tilde{\mathbf{x}}))$ .

**Theorem 5.7.** For any  $\gamma > 0$  and any measurable function  $f(x)$ , we have

$$L(f(\cdot)) - L^* \leq \mathbb{E}_{\tilde{\mathbf{x}}} \Delta Q_\gamma(\eta(\tilde{\mathbf{x}}), f(\tilde{\mathbf{x}})) = \mathbb{E}_{\tilde{\mathbf{x}}} [Q_\gamma(\eta(\tilde{\mathbf{x}}), f(\tilde{\mathbf{x}})) + |2\eta(\tilde{\mathbf{x}}) - 1| - 1].$$

*Proof.* By definition of  $L(\cdot)$ , it is easy to verify that

$$\begin{aligned} L(f(\cdot)) - L(2\eta(\cdot) - 1) &= \mathbb{E}_{\eta(X) \geq 0.5, f(X) < 0} (2\eta(X) - 1) + \mathbb{E}_{\eta(X) < 0.5, f(X) \geq 0} (1 - 2\eta(X)) \\ &\leq \mathbb{E}_{(2\eta(X) - 1)f(X) \leq 0} |2\eta(X) - 1|. \end{aligned}$$

From Lemma 5.1, i.e.,  $\Delta Q_\gamma(\eta, 0) = |2\eta - 1|$ , we have

$$L(f(\cdot)) - L^* \leq \mathbb{E}_{(2\eta(\tilde{\mathbf{x}}) - 1)f(\tilde{\mathbf{x}}) \leq 0} \Delta Q_\gamma(\eta(\tilde{\mathbf{x}}), 0).$$

To complete the proof, since  $\Delta Q_\gamma(\eta, f) = Q_\gamma(\eta, f) - Q_\gamma(\eta, f_\gamma^*(\eta))$ , it suffices to show that  $Q_\gamma(\eta(\mathbf{x}), 0) \leq Q_\gamma(\eta(\mathbf{x}), f(\mathbf{x}))$  for all  $\mathbf{x}$  such that  $(2\eta(\mathbf{x}) - 1)f(\mathbf{x}) \leq 0$ . To see this, we consider three scenarios:

- $\eta > 0.5$ : We have  $f_\gamma^*(\eta) = \text{sign}(2\eta - 1)\gamma > 0$ . In addition,  $(2\eta - 1)f \leq 0$  implies  $f \leq 0$ . Since  $0 \in [f, f_\gamma^*(\eta)]$  and the convexity of  $Q_\gamma(\eta, f)$  w.r.t.  $f$ , we have  $Q_\gamma(\eta, 0) \leq \max\{Q_\gamma(\eta, f), Q_\gamma(\eta, f_\gamma^*(\eta))\} = Q_\gamma(\eta, f)$ .
- $\eta < 0.5$ : In this case we have  $f_\gamma^*(\eta) < 0$  and  $f \geq 0$ , which leads to  $0 \in [f_\gamma^*(\eta), f]$ , which implies  $Q_\gamma(\eta, 0) \leq \max\{Q_\gamma(\eta, f), Q_\gamma(\eta, f_\gamma^*(\eta))\} = Q_\gamma(\eta, f)$ .
- $\eta = 0.5$ : Note that  $f_\gamma^* = 0$ , which implies that  $Q_\gamma(\eta, 0) \leq Q_\gamma(\eta, f)$  for all  $f$ .

From the proof of Lemma 5.1, we have  $\Delta Q_\gamma(\eta, f) = Q_\gamma(\eta, f) + |2\eta - 1| - 1$ . Hence the theorem holds.  $\square$

**Corollary 5.1.** For any measurable function  $f(x)$ ,

$$L(f(\cdot)) - L^* \leq \min_{\gamma > 0} \mathbb{E}_{\tilde{\mathbf{x}}} [Q_\gamma(\eta(\tilde{\mathbf{x}}), f(\tilde{\mathbf{x}})) + |2\eta(\tilde{\mathbf{x}}) - 1| - 1]. \quad (5.15)$$

*Proof.* Since Theorem 5.7 holds for any  $\gamma > 0$ , we obtain this corollary.  $\square$

For samples  $\{\mathbf{x}_i, y_i\}_{i=1}^m$ , since  $\eta(\mathbf{x}_i) = y_i \in \{1, -1\}$ , the empirical estimation of the bound in



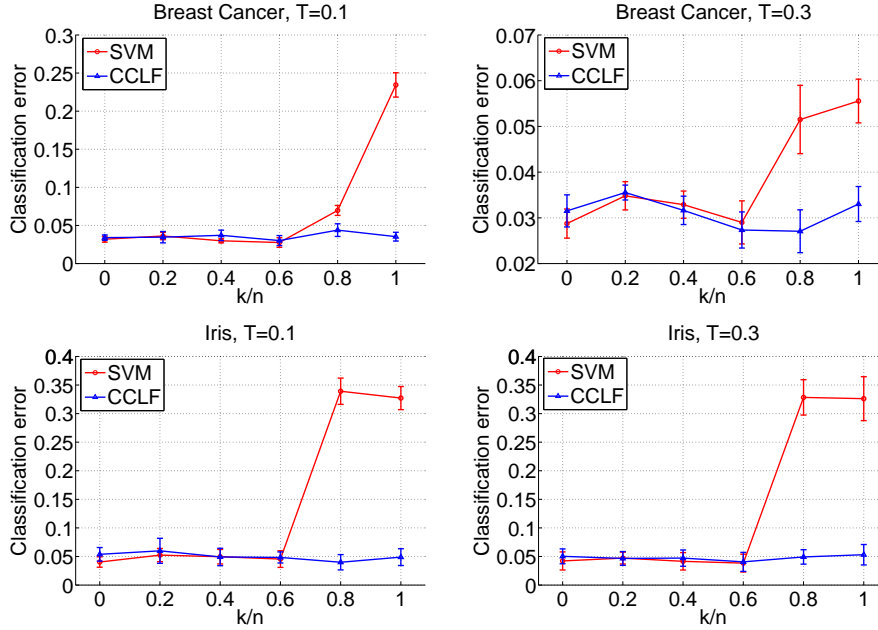
(5.15) is  $\Phi(\mathbf{u}) = \min_{\gamma>0} \frac{1}{m} \sum_{i=1}^m \phi_{\gamma}(u_i)$  where  $u_i = y_i f(\mathbf{x}_i)$ , which implies that minimizing the coherent loss function is equivalent to minimizing the empirical bound of the difference between  $L(f(\cdot))$  and  $L^*$ .

## 5.5 Simulation

We report some numerical simulation results in this section to illustrate the proposed approach. Besides the regularization constraints (e.g.  $\|\mathbf{w}\| \leq C$  for binary-class SVMs and  $\|\mathbf{w}_i\| \leq C, i = 1, \dots, k$  for multi-class SVMs), we consider the case where additional linear constraints are also imposed on the coefficient  $\mathbf{w}$ . For clarity, we choose a simple additional constraint  $\|\mathbf{A}\mathbf{w}\|_{\infty} \leq T$  to compare the performance of the cumulative loss formulation (SVM) and our coherent loss formulation (CCLF) for binary-class and multi-class classification, where  $\mathbf{A} = [\mathbf{I}_k, \mathbf{0}] \in \mathbb{R}^{k \times n}$ . In other words, the constraint ensures that the maximum of the first  $k$  elements of  $\mathbf{w}$  is bounded by  $T$ . We now compare their performance under two cases: 1)  $k$  is fixed,  $T$  varies; 2)  $T$  is fixed,  $k$  varies.

Three binary-class datasets “Breast cancer”, “Ionosphere” and “Diabetes”, and two multi-class datasets “Wine” and “Iris” from UCI [AN07] are used, where we randomly pick 50% as training samples, 20% as validation samples, and the rest as testing samples. For the cumulative loss formulation approach, parameter  $C$  is determined by cross-validation. For the coherent loss formulation approach, parameter  $C$  is fixed while parameter  $a$  is determined by cross-validation. For each  $T$ , we repeated the experiments 20 times and computed the average classification errors. To solve the resulting optimization problems, we use CVX [GB11, GB08], and Gurobi [GO13] as the solver.

Figure 5.3 shows the simulation results under fixed  $k$ . Clearly, when additional constraints are imposed, it appears that the coherent loss approach consistently outperforms the cumulative loss approach. When  $T$  is small, the cumulative loss approach performs much worse. When  $T$  becomes large, its performance can be close to the coherent loss approach. Figure 5.2 provides the results under fixed  $T$ , which shows that the coherent loss and cumulative loss approaches have similar performance when  $k/n$  is small but the coherent loss



**Figure 5.2:** Performance comparison of cumulative loss approach vs coherent loss approach where bound  $T$  is fixed and the fraction  $k/n$  varies from 0.0 to 1.0. Left and right columns report the classification errors for the two cases  $T = 0.1$  and  $T = 0.3$ .

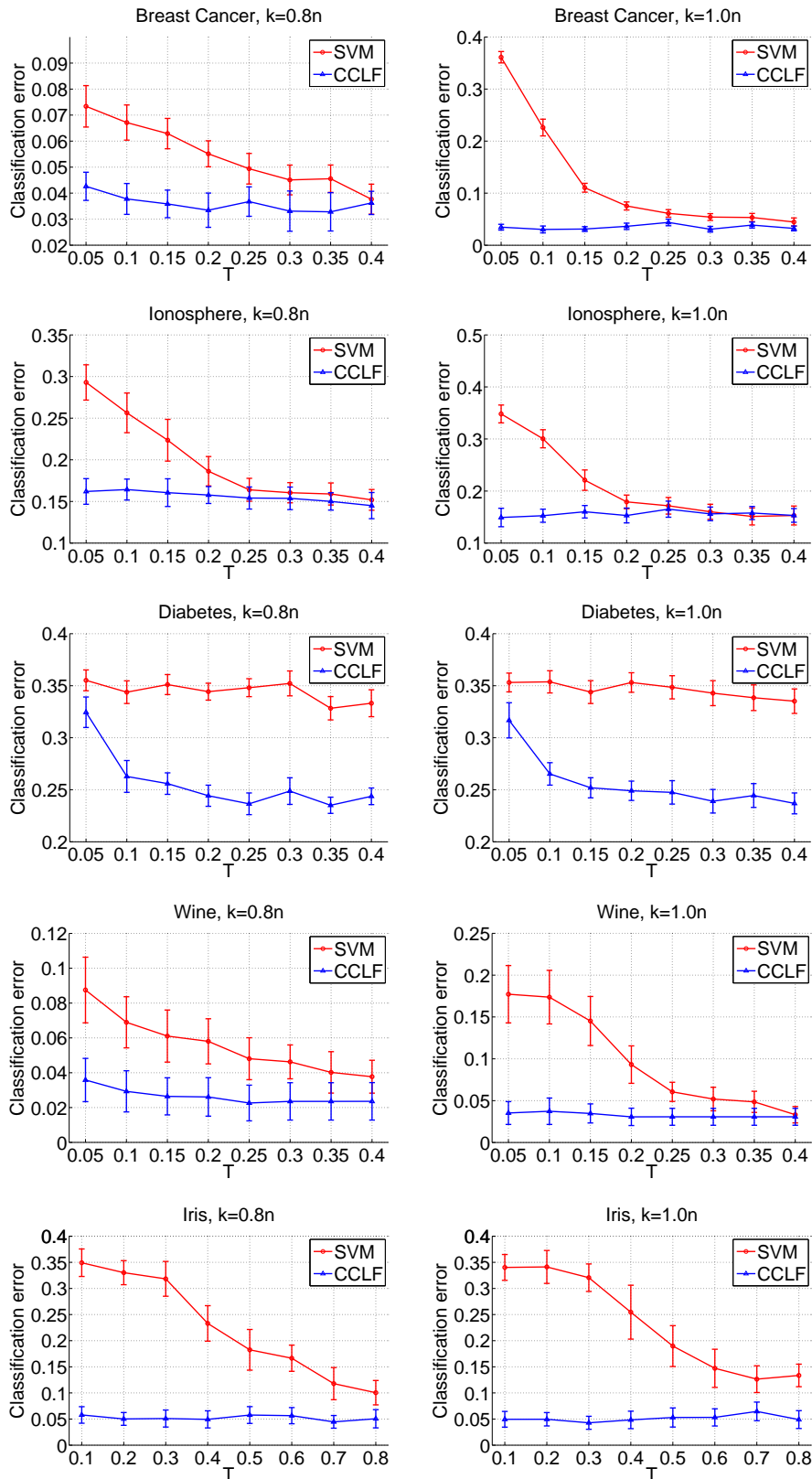
approach outperforms the cumulative loss approach when  $k/n$  is large. We believe that these phenomena are due to the fact that the coherent loss is a better approximation for the empirical classification error.

## 5.6 Proofs of Technical Results

### 5.6.1 Proof of Theorem 5.1

*Proof. Step 1 – the “if” part.* Given a function  $\rho(\mathbf{u}, \mathcal{G}) = 1 - \sup\{k \in [0, 1] : \sup_{\mathbf{v} \in \mathbf{V}_k} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \leq 0, \tilde{\mathbf{u}} = (\min_{j \in \mathcal{G}_1} u_j, \dots, \min_{j \in \mathcal{G}_n} u_j)^\top\}$  for some admissible class  $\{\mathbf{V}_k\}$ , we show that  $\rho(\cdot, \cdot)$  satisfies all properties required for a CCLF.

**Step 1.1 – Complete Classification:** If  $\mathbf{u} \geq 0$ , then by  $\mathbf{V}_1 = \mathbb{R}_+^n$  we have that  $\mathbf{v}^\top \tilde{\mathbf{u}} \geq 0$  for all  $\mathbf{v} \in \mathbf{V}_1$ , which implies that  $\sup_{\mathbf{v} \in \mathbf{V}_1} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \leq 0$ . Hence  $\rho(\mathbf{u}, \mathcal{G}) = 0$ . Conversely, if  $\mathbf{u} \not\geq 0$ , without loss of generality we assume that there exists  $j \in \mathcal{G}_1$  such that  $u_j < 0$ , then



**Figure 5.3:** Performance comparison of cumulative loss approach vs coherent loss approach. Left and right columns report the classification errors for the two cases  $k = 0.8n$  and  $k = n$  (recall that  $k$  and  $n$  are the numbers of the rows and columns of matrix  $\mathbf{A}$ , respectively). The four rows, from top to bottom, report results for *Breast Cancer*, *Ionosphere*, *Diabetes*, *Wine* and *Iris*, respectively.

we have

$$\sup_{\mathbf{v} \in \mathbf{V}_1} (-\mathbf{v}^\top \tilde{\mathbf{u}}) = \sup_{\mathbf{v} \in \mathfrak{R}_+^n} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \geq -\mathbf{e}_1^\top \tilde{\mathbf{u}} > 0.$$

This inequality, combined with  $\mathbf{V}_1 = \text{cl}(\lim_{k \uparrow 1} \mathbf{V}_k)$ , leads to that  $\exists \delta > 0$  such that

$$\sup_{\mathbf{v} \in \mathbf{V}_{1-\delta}} (-\mathbf{v}^\top \tilde{\mathbf{u}}) > 0,$$

which implies that  $\rho(\mathbf{u}, \mathcal{G}) > 0$ . This shows that  $\rho(\cdot, \cdot)$  satisfies *complete classification*.

**Step 1.2 – Misclassification avoidance:** Fix  $\mathbf{u}$  such that  $\mathbf{u} < \mathbf{0}$  which implies  $\tilde{\mathbf{u}} < \mathbf{0}$ .

Since  $\mathbf{e} \in \mathbf{V}_0$ , we have

$$\sup_{\mathbf{v} \in \mathbf{V}_0} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \geq -\mathbf{e}^\top \tilde{\mathbf{u}} > 0.$$

Hence  $\rho(\mathbf{u}, \mathcal{G}) = 1$ . Thus,  $\rho(\cdot, \cdot)$  satisfies *misclassification avoidance*.

**Step 1.3 – Monotonicity:** Note that  $\min_{j \in \mathcal{G}_i} u_j \geq \min_{j \in \mathcal{G}_i} w_j$  for all  $i = 1, \dots, n$  implies  $\tilde{\mathbf{u}} \geq \tilde{\mathbf{w}}$ . Then for any  $k \in [0, 1]$ , since  $\mathbf{V}_k \subseteq \mathbf{V}_1 = \mathfrak{R}_+^n$ , we have that  $-\mathbf{v}^\top \tilde{\mathbf{u}} \leq -\mathbf{v}^\top \tilde{\mathbf{w}}$  for any  $\mathbf{v} \in \mathbf{V}_k$ . Thus,

$$\sup_{\mathbf{v} \in \mathbf{V}_k} (-\mathbf{v}^\top \tilde{\mathbf{w}}) \leq 0 \implies \sup_{\mathbf{v} \in \mathbf{V}_k} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \leq 0.$$

Hence  $\rho(\mathbf{u}, \mathcal{G}) \leq \rho(\mathbf{w}, \mathcal{G})$ . Thus,  $\rho(\cdot, \cdot)$  satisfies *monotonicity*.

**Step 1.4 – Order & scale invariance:** Order invariance follows directly from the fact that  $\mathbf{V}_k$  is order invariant for all  $k$ . Scale invariant holds because for  $\alpha > 0$  and  $k \in [0, 1]$ ,

$$\sup_{\mathbf{v} \in \mathbf{V}_k} (-\mathbf{v}^\top \mathbf{u}) \leq 0 \iff \sup_{\mathbf{v} \in \mathbf{V}_k} (-\mathbf{v}^\top \alpha \mathbf{u}) \leq 0.$$

**Step 1.5 – Quasi-convexity:** To show quasi-convexity in  $\mathbf{u}$ , let  $c = \max(\rho(\mathbf{u}, \mathcal{G}), \rho(\mathbf{w}, \mathcal{G}))$  and without loss of generality assume  $c < 1$  since otherwise the claim trivially holds. Thus we have that for any  $\epsilon > 0$

$$\sup_{\mathbf{v} \in \mathbf{V}_{1-c-\epsilon}} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \leq 0 \text{ and } \sup_{\mathbf{v} \in \mathbf{V}_{1-c-\epsilon}} (-\mathbf{v}^\top \tilde{\mathbf{w}}) \leq 0,$$

which implies that for  $\alpha \in [0, 1]$

$$\sup_{\mathbf{v} \in \mathbf{V}_{1-c-\epsilon}} \{-\mathbf{v}^\top [\alpha \tilde{\mathbf{u}} + (1-\alpha) \tilde{\mathbf{w}}]\} \leq 0.$$

Recall that  $\tilde{\mathbf{u}} = (\min_{j \in \mathcal{G}_1} u_j, \dots, \min_{j \in \mathcal{G}_n} u_j)^\top$  and  $\tilde{\mathbf{w}} = (\min_{j \in \mathcal{G}_1} w_j, \dots, \min_{j \in \mathcal{G}_n} w_j)^\top$ . Let  $\mathbf{y} = \alpha \mathbf{u} + (1-\alpha) \mathbf{w}$  and  $\tilde{\mathbf{y}} = (\min_{j \in \mathcal{G}_1} y_j, \dots, \min_{j \in \mathcal{G}_n} y_j)^\top$ . Then we have  $\tilde{\mathbf{y}} \geq \alpha \tilde{\mathbf{u}} + (1-\alpha) \tilde{\mathbf{w}}$  which implies that

$$\sup_{\mathbf{v} \in \mathbf{V}_{1-c-\epsilon}} \{-\mathbf{v}^\top \tilde{\mathbf{y}}\} \leq \sup_{\mathbf{v} \in \mathbf{V}_{1-c-\epsilon}} \{-\mathbf{v}^\top [\alpha \tilde{\mathbf{u}} + (1-\alpha) \tilde{\mathbf{w}}]\} \leq 0.$$

Thus, we have  $\rho(\alpha \mathbf{u}_1 + (1-\alpha) \mathbf{u}_2, \mathcal{G}) \leq c$  since  $\epsilon$  can be arbitrarily close to 0. The quasi-convexity holds.

**Step 1.6 – Lower semi-continuity:** We show that  $\rho(\mathbf{u}^*, \mathcal{G}) \leq \liminf_i \rho(\mathbf{u}_i, \mathcal{G})$  for  $\mathbf{u}_i \xrightarrow{i} \mathbf{u}^*$ . Let  $c > \liminf_i \rho(\mathbf{u}_i, \mathcal{G})$ , then there exists an infinite sub-sequence  $\{\mathbf{u}_{i_j}\}$  such that  $\rho(\mathbf{u}_{i_j}, \mathcal{G}) < c$ . That is

$$-\mathbf{v}^\top \tilde{\mathbf{u}}_{i_j} \leq 0; \quad \forall \mathbf{v} \in \mathbf{V}_{1-c}, \forall j.$$

Note that  $\mathbf{u}_{i_j} \rightarrow \mathbf{u}^*$ , hence

$$-\mathbf{v}^\top \tilde{\mathbf{u}}^* \leq 0; \quad \forall \mathbf{v} \in \mathbf{V}_{1-c},$$

i.e.,  $\rho(\mathbf{u}^*, \mathcal{G}) \leq c$ . Since  $c$  can be arbitrarily close to  $\liminf_i \rho(\mathbf{u}_i, \mathcal{G})$ , the semi-continuity follows.

**Step 2 – the “only if” part.** Given a function  $\rho(\cdot, \cdot)$  which is a CCLF, we show that it can be represented as

$$\rho(\mathbf{u}, \mathcal{G}) = 1 - \sup\{k \in [0, 1] : \sup_{\mathbf{v} \in \mathbf{V}_k} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \leq 0, \tilde{\mathbf{u}} = (\min_{j \in \mathcal{G}_1} u_j, \dots, \min_{j \in \mathcal{G}_n} u_j)^\top\},$$

for some admissible class  $\{\mathbf{V}_k\}$ . This consists of three steps. We first show that  $\rho(\cdot, \cdot)$  can be represented as

$$\rho(\mathbf{u}, \mathcal{G}) = 1 - \sup\{k \in [0, 1] : \sup_{\mathbf{v} \in \tilde{\mathbf{V}}_k} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \leq 0, \tilde{\mathbf{u}} = (\min_{j \in \mathcal{G}_1} u_j, \dots, \min_{j \in \mathcal{G}_n} u_j)^\top\},$$

for some  $\{\bar{\mathbf{V}}_k\}$ . Here  $\{\bar{\mathbf{V}}_k\}$  is not necessarily admissible, but satisfies  $\bar{\mathbf{V}}_k \subseteq \bar{\mathbf{V}}_{k'}$  for all  $k \leq k'$ . We then show that we can replace  $\bar{\mathbf{V}}_k$  by a class of closed, convex, order-invariant, cones  $\mathbf{V}_k$ . Finally we show that  $\{\mathbf{V}_k\}$  is admissible to complete the proof.

**Step 2.1:** The representability of  $\rho(\cdot, \cdot)$  follows from the following lemma which is a variant of Theorem 2 in [BS09].

**Lemma 5.2.** *Given a CCLF  $\rho(\cdot, \cdot)$ , then there exists  $\{\bar{\mathbf{V}}_k\}$  that satisfies  $\bar{\mathbf{V}}_k \subseteq \bar{\mathbf{V}}_{k'}$  for all  $k \leq k'$ , such that*

$$\rho(\mathbf{u}, \mathcal{G}) = 1 - \sup\{k \in [0, 1] : \sup_{\mathbf{v} \in \bar{\mathbf{V}}_k} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \leq 0, \tilde{\mathbf{u}} = (\min_{j \in \mathcal{G}_1} u_j, \dots, \min_{j \in \mathcal{G}_n} u_j)^\top\}.$$

*Proof.* We recall the definition of the collective satisfying measure in [BS09].

**Definition 5.3.** *Let  $\mathcal{U}$  be the set of random variables on the probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ . A function  $\bar{\rho}(\cdot) : \mathcal{U} \rightarrow [0, 1]$  is a collective satisfying measure if the following holds for all  $U, U' \in \mathcal{U}$ .*

1. *If  $U \geq 0$ , then  $\bar{\rho}(U) = 1$ ;*
2. *If  $U < 0$ , then  $\bar{\rho}(U) = 0$ ;*
3. *If  $U \geq U'$  then  $\bar{\rho}(U) \geq \bar{\rho}(U')$ ;*
4.  *$\lim_{\alpha \geq 0} \bar{\rho}(U + \alpha) = \bar{\rho}(U)$ ;*
5. *If  $\lambda \in [0, 1]$ , then  $\bar{\rho}(\lambda U + (1 - \lambda)U') \geq \min(\bar{\rho}(U), \bar{\rho}(U'))$ ;*
6. *If  $\lambda > 0$ , then  $\bar{\rho}(\lambda U) = \bar{\rho}(U)$ .*

We now consider  $\mathcal{U}$  – a special set of random variables defined on the probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  with  $\Omega = \{1, \dots, m\}$ . Note that each random variable  $U : \Omega \mapsto \mathfrak{R}$  can be represented as a vector  $\mathbf{u} \in \mathfrak{R}^m$  where  $u_i = U(i)$ . Let  $\mathcal{G}$  be a partition of the set  $\{1, \dots, m\}$ , namely,  $\mathcal{G} = \{\mathcal{G}_1, \dots, \mathcal{G}_n\}$  satisfying that

$$\mathcal{G}_i \neq \emptyset, \mathcal{G}_i \subseteq \{1, \dots, m\}, \bigcup_{i=1}^n \mathcal{G}_i = \{1, \dots, m\}.$$

We define another random variable  $\tilde{U}$  on probability space  $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$  with  $\tilde{\Omega} = \{1, \dots, n\}$  by taking  $\tilde{U}(i) = \min_{j \in \mathcal{G}_i} u_j$ . The mapping from  $U$  to  $\tilde{U}$  is denoted by  $g$ , i.e.,  $\tilde{U} = g(U, \mathcal{G})$ . Let  $\hat{\rho}(\cdot)$  be a collective satisfying measure on  $\mathcal{U}$  that satisfies all the properties given by Definition 5.3 and another property that for all  $U, U' \in \mathcal{U}$ , if  $\tilde{U} \geq \tilde{U}'$ , then  $\hat{\rho}(U) \geq \hat{\rho}(U')$ .

**Theorem 5.8.** *The collective satisfying measure  $\hat{\rho}(\cdot)$  can be represented as*

$$\hat{\rho}(U) = \sup\{k \in [0, 1] : \sup_{\mathbb{Q} \in \mathcal{Q}_k} \mathbb{E}_{\mathbb{Q}}(-\tilde{U}) \leq 0, \tilde{U} = g(U, \mathcal{G})\},$$

for a class of sets of probability measures  $\mathcal{Q}_k$  satisfying  $\mathcal{Q}_k \subseteq \mathcal{Q}_{k'}$  for  $k \leq k'$ .

*Proof.* Let  $X$  be a random variable on probability space  $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$ . For  $k \in [0, 1]$ , we define

$$\mu_k(X) = \inf\{a : \hat{\rho}(\hat{X} + a) \geq k \text{ for some r.v. } \hat{X} \text{ such that } X = g(\hat{X}, \mathcal{G})\}. \quad (5.16)$$

Then we have

$$\hat{\rho}(U) = \sup\{k : \mu_k(\tilde{U}) \leq 0, k \in [0, 1]\}. \quad (5.17)$$

To verify this equality, note that

$$\begin{aligned} & \sup\{k : \mu_k(\tilde{U}) \leq 0, k \in [0, 1]\} \\ &= \sup\{k : \exists a \leq 0, \hat{U} \text{ such that } \hat{\rho}(\hat{U} + a) \geq k \text{ and } \tilde{U} = g(\hat{U}, \mathcal{G})\} \\ &= \sup\{\hat{\rho}(\hat{U} + a) : a \leq 0, \tilde{U} = g(\hat{U}, \mathcal{G})\} \\ &= \sup\{\hat{\rho}(\hat{U}) : \tilde{U} = g(\hat{U}, \mathcal{G})\} \end{aligned}$$

where the last equality holds due to the monotonicity of  $\hat{\rho}(\cdot)$ . Since  $\tilde{U} = g(\hat{U}, \mathcal{G})$  and  $\tilde{U} = g(U, \mathcal{G})$ , by the additional property of  $\hat{\rho}(\cdot)$  given above, we have  $\hat{\rho}(U) = \hat{\rho}(\hat{U})$ . Hence (5.17) holds. We next verify that  $\mu_k(\cdot)$  defined by (5.16) is a coherent risk measure. Recall the definition of coherent risk measure:

**Definition 5.4.** *Let  $\mathcal{U}$  be the set of random variables on the probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ . A function  $\mu(\cdot) : \mathcal{U} \rightarrow \mathfrak{R}$  is a coherent risk measure if the following holds for all  $X, Y \in \mathcal{U}$ .*

1. If  $X \geq Y$  then  $\mu(X) \leq \mu(Y)$ ;

2. If  $c \in \mathfrak{R}$ , then  $\mu(X + c) = \mu(X) - c$ ;
3. If  $\lambda \in [0, 1]$ , then  $\mu(\lambda X + (1 - \lambda)Y) \leq \lambda\mu(X) + (1 - \lambda)\mu(Y)$ ;
4. If  $\lambda > 0$ , then  $\mu(\lambda X) = \lambda\mu(X)$ .

We now verify that  $\mu_k$  satisfies these properties. For random variables  $X$  and  $Y$ , let  $\hat{X}$  and  $\hat{Y}$  be any random variables satisfying that  $X = g(\hat{X}, \mathcal{G})$  and  $Y = g(\hat{Y}, \mathcal{G})$ . If  $X \geq Y$ , then by the property of  $\hat{\rho}(\cdot)$ , we have  $\hat{\rho}(\hat{X}) \geq \hat{\rho}(\hat{Y})$ , which implies  $\mu_k(X) \leq \mu_k(Y)$ . Hence Property 1 holds. Property 2 can be easily seen from the definition of  $\mu_k(\cdot)$ . For Property 3, note that for all  $\epsilon > 0$ , we have

$$\hat{\rho}(\hat{X} + \mu_k(X) + \epsilon) \geq k, \quad \hat{\rho}(\hat{Y} + \mu_k(Y) + \epsilon) \geq k.$$

On the other hand, since  $\hat{\rho}$  is quasi-concave, we have

$$\hat{\rho}(\lambda(\hat{X} + \mu_k(X)) + (1 - \lambda)(\hat{Y} + \mu_k(Y) + \epsilon)) \geq \min\{\hat{\rho}(\hat{X} + \mu_k(X) + \epsilon), \hat{\rho}(\hat{Y} + \mu_k(Y) + \epsilon)\} \geq k. \quad (5.18)$$

Now consider special  $\hat{X}$  and  $\hat{Y}$  such that  $\hat{X}(i) = X(j)$  for  $i \in \mathcal{G}_j$  and  $\hat{Y}(i) = Y(j)$  for  $i \in \mathcal{G}_j$ . Clearly, these  $\hat{X}$  and  $\hat{Y}$  are the “smallest”, namely, for all  $\tilde{X}$  and  $\tilde{Y}$  such that  $X = g(\tilde{X}, \mathcal{G})$  and  $Y = g(\tilde{Y}, \mathcal{G})$ , we have  $\tilde{X} \geq \hat{X}$  and  $\tilde{Y} \geq \hat{Y}$ . This implies

$$\begin{aligned} \mu_k(\lambda X + (1 - \lambda)Y) &= \inf\{a : \hat{\rho}(\hat{Z} + a) \geq k \text{ for some r.v. } \hat{Z} \text{ such that } \lambda X + (1 - \lambda)Y = g(\hat{Z}, \mathcal{G})\} \\ &\leq \inf\{a : \hat{\rho}(\lambda\hat{X} + (1 - \lambda)\hat{Y} + a) \geq k\} \\ &\leq \lambda\mu_k(X) + (1 - \lambda)\mu_k(Y), \end{aligned}$$

where the last inequality follows from (5.18). For the last property, note that for  $\lambda > 0$ ,

$$\begin{aligned} \mu_k(\lambda X) &= \inf\{a : \hat{\rho}(\lambda\hat{X} + a) \geq k \text{ for some r.v. } \hat{X} \text{ such that } \lambda X = g(\lambda\hat{X}, \mathcal{G})\} \\ &= \inf\{a : \hat{\rho}(\lambda\hat{X} + \lambda a) \geq k \text{ for some r.v. } \hat{X} \text{ such that } \lambda X = g(\lambda\hat{X}, \mathcal{G})\} \\ &= \inf\{\lambda a : \hat{\rho}(\hat{X} + a) \geq k \text{ for some r.v. } \hat{X} \text{ such that } X = g(\hat{X}, \mathcal{G})\} \\ &= \lambda\mu_k(X). \end{aligned}$$



Hence  $\mu_k(\cdot)$  is a coherent risk measure. It is known that coherent risk measure  $\mu_k(\cdot)$  can be written in the form

$$\mu_k(X) = \sup_{\mathbb{Q} \in \mathcal{Q}_k} \mathbb{E}_{\mathbb{Q}}(-X)$$

for a family of generating measures  $\mathcal{Q}_k$ . By combining this formula with (5.17), Theorem 5.8 can be obtained.  $\square$

We now turn to the proof of Lemma 5.2. Given a CCLF  $\rho(\cdot, \cdot)$ , for fixed  $\mathcal{G}$ , we define  $\bar{\rho} : \mathcal{U} \mapsto \mathfrak{R}$  as following

$$\bar{\rho}(U) = 1 - \rho(\mathbf{u}, \mathcal{G}); \quad \text{where } u_i = U(i), \quad i = 1, \dots, m.$$

It is straightforward to check that  $\bar{\rho}(\cdot)$  has all the properties of the collective satisfying measure  $\hat{\rho}(\cdot)$ . Thus, Theorem 5.8 states there exists a class of sets of probability measure  $\mathcal{Q}_k$  such that

$$1 - \rho(\mathbf{u}) = \bar{\rho}(U) = \sup\{k \in [0, 1] : \sup_{\mathbb{Q} \in \mathcal{Q}_k} \mathbb{E}_{\mathbb{Q}}(-\tilde{U}) \leq 0, \tilde{U} = g(U, \mathcal{G})\}.$$

Note that any probability measure  $Q$  on  $\tilde{\Omega} = \{1, \dots, n\}$  can be represented by a vector  $\mathbf{v} \in \mathfrak{R}^n$  such that  $v_i = Q(i)$ . Thus  $\mathbb{E}_Q(-\tilde{U}) = -\mathbf{v}^\top \tilde{\mathbf{u}}$  where  $\mathbf{v}$  and  $\tilde{\mathbf{u}}$  are the vector form for  $Q$  and  $\tilde{U}$  respectively. Hence we have there exists  $\bar{\mathbf{V}}_k$  such that

$$\rho(\mathbf{u}) = 1 - \sup\{k \in [0, 1] : \sup_{\mathbf{v} \in \bar{\mathbf{V}}_k} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \leq 0, \tilde{\mathbf{u}} = (\min_{j \in \mathcal{G}_1} u_j, \dots, \min_{j \in \mathcal{G}_n} u_j)^\top\}.$$

Note that for  $k \leq k'$ ,  $\bar{\mathbf{V}}_k \subseteq \bar{\mathbf{V}}_{k'}$  since  $\mathcal{Q}_k \subseteq \mathcal{Q}_{k'}$ . This concludes the proof of Lemma 5.2.  $\square$

**Step 2.2:** We construct the admissible class  $\{\mathbf{V}_k\}$  as follows. Define  $\hat{\mathbf{V}}_k \triangleq \text{cl}(\text{cc}(\text{or}(\bar{\mathbf{V}}_k)))$ . Then we let  $\mathbf{V}_k \triangleq \hat{\mathbf{V}}_k$  for  $k \in (0, 1)$ , and  $\mathbf{V}_0 \triangleq \bigcap_{k \in (0, 1)} \hat{\mathbf{V}}_k$ , and  $\mathbf{V}_1 \triangleq \text{cl}(\bigcup_{k \in (0, 1)} \hat{\mathbf{V}}_k)$ . Here  $\text{or}(\cdot)$  (respectively  $\text{cc}(\cdot)$ ) is the minimal **order** invariant (respectively, **convex cone**) superset, defined as

$$\text{or}(S) = \{\mathbf{P}\mathbf{v} : \mathbf{P} \in \mathcal{P}_n, \mathbf{v} \in S\}, \quad \text{cc}(S) = \left\{ \sum_{i=1}^k \lambda_i \mathbf{v}_i \mid k \in \mathbb{N}, \mathbf{v}_i \in S, \lambda_i \geq 0 \right\},$$

where  $\mathcal{P}_n$  is the set of all  $n \times n$  permutation matrices. Let

$$\rho'(\mathbf{u}, \mathcal{G}) = 1 - \sup\{k \in [0, 1] : \sup_{\mathbf{v} \in \hat{\mathbf{V}}_k} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \leq 0, \tilde{\mathbf{u}} = (\min_{j \in \mathcal{G}_1} u_j, \dots, \min_{j \in \mathcal{G}_n} u_j)^\top\},$$

and observe that  $\bar{\mathbf{V}}_k \subseteq \hat{\mathbf{V}}_k$ , hence  $\rho(\mathbf{u}) \leq \rho'(\mathbf{u})$ . To show that  $\rho(\mathbf{u}) \geq \rho'(\mathbf{u})$ , it suffices to show that for any  $k, \epsilon$  and  $\tilde{\mathbf{u}}$ , the following holds,

$$\sup_{\mathbf{v} \in \bar{\mathbf{V}}_k} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \leq 0 \implies \sup_{\mathbf{v} \in \hat{\mathbf{V}}_{k-\epsilon}} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \leq 0.$$

Note that  $\sup_{\mathbf{v} \in \bar{\mathbf{V}}_k} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \leq 0$  implies  $\rho(\mathbf{u}) \leq 1 - k$ . Hence by order invariance of  $\rho(\cdot, \cdot)$ , we have

$$\sup_{\mathbf{v} \in \bar{\mathbf{V}}_{k-\epsilon}} \sup_{\mathbf{P} \in \mathcal{P}_n} (-\mathbf{v}^\top \mathbf{P} \tilde{\mathbf{u}}) \leq 0,$$

which is equivalent to

$$\sup_{\mathbf{v} \in \text{or}(\bar{\mathbf{V}}_{k-\epsilon})} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \leq 0.$$

By definition of  $\text{cc}(\cdot)$  and the continuity of  $-\mathbf{v}^\top \tilde{\mathbf{u}}$  w.r.t.  $\mathbf{v}$ , this leads to

$$\sup_{\mathbf{v} \in \text{cl}(\text{or}(\bar{\mathbf{V}}_{k-\epsilon}))} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \leq 0.$$

Therefore we have  $\rho(\mathbf{u}, \mathcal{G}) = \rho'(\mathbf{u}, \mathcal{G})$ . Finally note that  $\hat{\mathbf{V}}_k \subseteq \hat{\mathbf{V}}_{k'}$  for  $k \leq k'$ , which leads to the following

$$\begin{aligned} \sup_{\mathbf{v} \in \hat{\mathbf{V}}_0} (-\mathbf{v}^\top \tilde{\mathbf{u}}) &\leq \sup_{\mathbf{v} \in \bigcap_{k \in (0,1)} \hat{\mathbf{V}}_k} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \leq \sup_{\mathbf{v} \in \hat{\mathbf{V}}_\epsilon} (-\mathbf{v}^\top \tilde{\mathbf{u}}), \\ \sup_{\mathbf{v} \in \hat{\mathbf{V}}_{1-\epsilon}} (-\mathbf{v}^\top \tilde{\mathbf{u}}) &\leq \sup_{\mathbf{v} \in \bigcup_{k \in (0,1)} \hat{\mathbf{V}}_k} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \leq \sup_{\mathbf{v} \in \hat{\mathbf{V}}_1} (-\mathbf{v}^\top \tilde{\mathbf{u}}). \end{aligned}$$

By definitions of  $\mathbf{V}_0$  and  $\mathbf{V}_1$ , together with the fact (due to continuity)

$$\sup_{\mathbf{v} \in \text{cl}(\bigcup_{k \in (0,1)} \hat{\mathbf{V}}_k)} (-\mathbf{v}^\top \tilde{\mathbf{u}}) = \sup_{\mathbf{v} \in \bigcup_{k \in (0,1)} \hat{\mathbf{V}}_k} (-\mathbf{v}^\top \tilde{\mathbf{u}}),$$

we conclude that

$$\rho(\mathbf{u}, \mathcal{G}) = 1 - \sup\{k \in [0, 1] : \sup_{\mathbf{v} \in \mathbf{V}_k} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \leq 0, \tilde{\mathbf{u}} = (\min_{j \in \mathcal{G}_1} u_j, \dots, \min_{j \in \mathcal{G}_n} u_j)^\top\}.$$

**Step 2.3:** Now we check that  $\{\mathbf{V}_k\}$  is indeed admissible. Property 1-3 are straightforward from the definition of  $\mathbf{V}_k$ . To see that  $\mathbf{V}_0$  is closed, recall that the intersection of a class of closed sets is closed.

We next show Property 4:  $\mathbf{V}_1 = \mathfrak{R}_+^n$ . By definition of  $\mathbf{V}_1$ , we have

$$\lim_{k \rightarrow 1} \sup_{\mathbf{v} \in \mathbf{V}_k} (-\mathbf{v}^\top \tilde{\mathbf{u}}) = \sup_{\mathbf{v} \in \mathbf{V}_1} (-\mathbf{v}^\top \tilde{\mathbf{u}}).$$

Hence  $\rho(\mathbf{u}) = 0$  if and only if  $\sup_{\mathbf{v} \in \mathbf{V}_1} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \leq 0$ . Therefore, by the property of *complete classification* we have the following

$$\sup_{\mathbf{v} \in \mathbf{V}_1} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \leq 0 \iff \tilde{\mathbf{u}} \geq 0 \iff \mathbf{u} \geq 0. \quad (5.19)$$

Denote the dual cone of a cone  $\mathbf{C}$  by  $\mathbf{C}^*$  and recall that for any  $k$ ,  $\mathbf{V}_k$  is a closed convex cone, hence we have

$$(\mathbf{V}_1^*)^* = \mathbf{V}_1.$$

The definition of dual cone states that

$$\mathbf{V}_1^* = \{\mathbf{u} : \mathbf{u}^\top \mathbf{v} \geq 0, \forall \mathbf{v} \in \mathbf{V}_1\},$$

which combined with Equation (5.19) implies that  $\mathbf{V}_1^* = \mathfrak{R}_+^n$ . Since  $\mathfrak{R}_+^n$  is self-dual, we have  $\mathbf{V}_1 = \mathfrak{R}_+^n$ .

We now turn to Property 5. Fix  $k > 0$ . Consider  $\mathbf{u} = -\mathbf{e}^m$ , which means  $\tilde{\mathbf{u}} = -\mathbf{e}^n$ . By misclassification avoidance,  $\rho(\mathbf{u}, \mathcal{G}) = 1$ , which means there exists  $\mathbf{v} \in \mathbf{V}_k$  such that

$\mathbf{v}^\top \tilde{\mathbf{u}} < 0$ , i.e.,  $\sum_{i=1}^n v_i > 0$ . Define a permutation matrix  $\mathbf{P} \in \mathcal{P}_n$ :

$$\mathbf{P} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & 0 & \cdots & 1 \\ 1 & 0 & 0 & \cdots & 0 \end{pmatrix}$$

Thus, by order invariance of  $\mathbf{V}_k$ ,  $\mathbf{P}^t \mathbf{v} \in \mathbf{V}_k$  for  $t = 0, \dots, n-1$ . By convexity, this implies  $\frac{1}{n} \sum_{t=0}^{n-1} \mathbf{P}^t \mathbf{v} \in \mathbf{V}_k$ . Note that  $\frac{1}{n} \sum_{t=0}^{n-1} \mathbf{P}^t \mathbf{v} = [\frac{1}{n} \sum_{i=1}^n v_i] \mathbf{e}^n$ , thus

$$\frac{\sum_{i=1}^n v_i}{n} \mathbf{e}^n \in \mathbf{V}_k.$$

Since  $\sum_{i=1}^n v_i > 0$  and  $\mathbf{V}_k$  is a cone, we have  $\lambda \mathbf{e}^n \in \mathbf{V}_k$  for all  $\lambda \geq 0$  and  $k > 0$ . By definition of  $\mathbf{V}_0$ , this implies  $\lambda \mathbf{e}^n \in \mathbf{V}_0$ .  $\square$

### 5.6.2 Proof of Theorem 5.2

*Proof. Claim 1:* We check that all conditions of Definition 5.1 are satisfied by  $\bar{\rho}(\cdot)$ . The only condition needs a proof is the semi-continuity. Consider a sequence  $\mathbf{u}^j \rightarrow \mathbf{u}^0$ , and let  $t^0 = \max\{t : \sum_{i=1}^t \tilde{u}_{(i)}^0 < 0\}$ . Without loss of generality we let  $\tilde{u}_1^0 \leq \tilde{u}_2^0 \leq \dots \leq \tilde{u}_n^0$ . Thus we have that  $\sum_{i=1}^{t^0} \tilde{u}_i^0 < 0$ . This implies that  $\limsup_j \sum_{i=1}^{t^0} \tilde{u}_i^j < 0$ , which further leads to  $\liminf_j (\max\{t : \sum_{i=1}^t \tilde{u}_{(i)}^j < 0\}) \geq t^0$ . Hence  $\liminf_j \bar{\rho}(\mathbf{u}^j, \mathcal{G}) \geq \bar{\rho}(\mathbf{u}^0, \mathcal{G})$ , which established the semi-continuity. Thus, we conclude that  $\bar{\rho}(\cdot)$  is a CCLF. Further, observe that  $\max\{t : \sum_{i=1}^t \tilde{u}_{(i)} < 0\} \geq \sum_{i=1}^n \mathbf{1}(u_j < 0, \exists j \in \mathcal{G}_i)$ , which established the first claim.

**Claim 2:** It is straightforward to check that  $\bar{\mathbf{V}}_k$  satisfies all conditions of Definition 5.2, and hence is an admissible set. Thus, we proceed to show that  $\bar{\mathbf{V}}_k$  is an admissible set corresponding to  $\bar{\rho}(\cdot)$ , i.e., to show

$$\bar{\rho}(\mathbf{u}) = 1 - \sup\{k \in [0, 1] : \sup_{\mathbf{v} \in \bar{\mathbf{V}}_k} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \leq 0, \tilde{\mathbf{u}} = (\min_{j \in \mathcal{G}_1} u_j, \dots, \min_{j \in \mathcal{G}_n} u_j)^\top\}.$$

Fix a  $\mathbf{u} \in \mathfrak{R}^m$ . If  $\mathbf{u} \geq 0$ , then we have  $\bar{\rho}(\mathbf{u}) = 0$ , as well as  $\sup_{\mathbf{v} \in \bar{\mathbf{V}}_1} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \leq 0$ , and hence

the equivalence holds trivially. Thus we assume  $\mathbf{u} \not\geq 0$ , and let  $t^0 = \max\{t : \sum_{i=1}^t \tilde{u}_{(i)} < 0\}$ .

By definition we have

$$\bar{\mathbf{V}}_{1-t^0/n} = \text{conv} \{ \lambda \mathbf{e}_{N'} : \lambda > 0, |N'| = t^0 + 1 \}.$$

Note that by definition of  $t^0$

$$\min_{|N'|=t^0+1} \sum_{i \in N'} \tilde{u}_i \geq 0,$$

which implies that

$$\sup_{\mathbf{v} \in \{ \mathbf{e}_{N'} : |N'|=t^0+1 \}} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \leq 0.$$

This leads to

$$\sup_{\mathbf{v} \in \bar{\mathbf{V}}_{1-t^0/n}} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \leq 0. \quad (5.20)$$

On the other hand for arbitrarily small  $\epsilon > 0$ , by definition

$$\bar{\mathbf{V}}_{1-t^0/n+\epsilon} = \text{conv} \{ \lambda \mathbf{e}_N : \lambda > 0, |N| = t^0 \}.$$

Because  $\min_{N:|N|=t^0} \sum_{i \in N} \tilde{u}_i < 0$ , we have

$$\sup_{\mathbf{v} \in \bar{\mathbf{V}}_{1-t^0/n+\epsilon}} (-\mathbf{v}^\top \tilde{\mathbf{u}}) > 0.$$

Combining with Equation (5.20) we established the second claim.

**Claim 3:** Let  $\rho'(\cdot)$  be a CCLF satisfying that  $\rho'(\mathbf{u}, \mathcal{G}) \geq \varrho(\mathbf{u}, \mathcal{G})$  for all  $\mathbf{u} \in \mathfrak{R}^m$ , and let  $\{\mathbf{V}'_k\}$  be its corresponding admissible set. Thus, it suffices to show that  $\bar{\mathbf{V}}_k \subseteq \mathbf{V}'_k$  for all  $k$ . This holds trivially for  $k = 0$ , since  $\rho'(\mathbf{u}, \mathcal{G}) = 1$  for all  $\mathbf{u} < \mathbf{0}$  implies that  $\lambda \mathbf{e}^n \in \mathbf{V}'_0$ . When  $k > 0$ , let  $s/n < k \leq (s+1)/n$  for some integer  $s$ . Then, since  $\mathbf{V}'_k$  is an order-invariant convex cone, it suffices to show that  $\mathbf{e}_{[1:n-s]} \in \mathbf{V}'_k$  to establish the third claim. Consider  $\mathbf{u}^*$  such that  $\tilde{\mathbf{u}}^* = -\mathbf{e}_{[1:n-s]}$ . Then, by  $\rho'(\mathbf{u}^*, \mathcal{G}) \geq \sum_i \mathbf{1}(\tilde{u}_i^* < 0)/n = 1 - s/n > 1 - k$ , we have

$$\sup_{\mathbf{v} \in \mathbf{V}'_k} (-\mathbf{v}^\top \tilde{\mathbf{u}}^*) > 0 \implies \exists \mathbf{v}^* \in \mathbf{V}'_k : \sum_{i=1}^{n-s} v_i^* > 0.$$

Define a permutation matrix  $\mathbf{P}$ :

$$\mathbf{P} = \begin{pmatrix} \mathbf{P}_1 & 0_{(n-s) \times s} \\ 0_{(n-s) \times s} & 0_{s \times s} \end{pmatrix}$$

where  $\mathbf{P}_1$  is a  $(n-s) \times (n-s)$  matrix:

$$\mathbf{P}_1 = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & 0 & \cdots & 1 \\ 1 & 0 & 0 & \cdots & 0 \end{pmatrix}$$

Thus, by order invariance of  $\mathbf{V}'_k$ ,  $\mathbf{P}^t \mathbf{v}^* \in \mathbf{V}'_k$  for  $t = 0, \dots, n-s-1$ . By convexity, this implies  $\frac{1}{n-s} \sum_{t=0}^{n-s-1} \mathbf{P}^t \mathbf{v}^* \in \mathbf{V}'_k$ . Note that  $\frac{1}{n-s} \sum_{t=0}^{n-s-1} \mathbf{P}^t \mathbf{v}^* = \frac{1}{n-s} [\sum_{i \in [1:n-s]} v_i^*] \mathbf{e}_{[1:n-s]}$ , thus

$$\frac{\sum_{i=1}^{n-s} v_i^*}{n-s} \mathbf{e}_{[1:n-s]} \in \mathbf{V}'_k.$$

Since  $\frac{\sum_{i=1}^{n-s} v_i^*}{n-s}$  is positive, and  $\mathbf{V}'_k$  is a cone, we have  $\mathbf{e}_{[1:n-s]} \in \mathbf{V}'_k$ , which completes the proof.  $\square$

### 5.6.3 Proof of Theorem 5.3

*Proof.* We prove the theorem by constructing such a function  $\rho(\cdot, \cdot)$ . To do this, first consider  $\tilde{\rho} : \mathcal{R}^m \times \mathcal{S} \mapsto [0, 1]$  defined as

$$\tilde{\rho}(\mathbf{u}, \mathcal{G}) = \min_{\gamma > 0} \hat{\rho}(\mathbf{u}/\gamma, \mathcal{G}).$$

Then it is easy to check that  $\tilde{\rho}(\cdot)$  satisfies complete classification, misclassification avoidance, monotonicity, order invariance, and scale invariance. To see that  $\tilde{\rho}(\mathbf{u}, \mathcal{G}) \geq \varrho(\mathbf{u}, \mathcal{G})$ , note that if  $\tilde{\mathbf{u}}$ , i.e.,  $\tilde{\mathbf{u}} = (\min_{j \in \mathcal{G}_1} u_j, \dots, \min_{j \in \mathcal{G}_n} u_j)^\top$ , has  $t$  negative coefficients, then for any

$\gamma > 0$ ,  $\tilde{\mathbf{u}}/\gamma$  also has  $t$  negative coefficients, which means

$$\hat{\rho}(\mathbf{u}/\gamma, \mathcal{G}) \geq t/n.$$

Taking minimization over  $\gamma$ , we have  $\tilde{\rho}(\mathbf{u}, \mathcal{G}) \geq \varrho(\mathbf{u}, \mathcal{G})$  holds. Finally, we show quasi-convexity of  $\tilde{\rho}(\cdot)$ . Fix  $\mathbf{u}_1, \mathbf{u}_2$ , and  $\alpha \in [0, 1]$ , let  $\gamma_1, \gamma_2$  be  $\epsilon$ -optimal, i.e.,

$$\hat{\rho}(\mathbf{u}_i/\gamma_i, \mathcal{G}) \leq \tilde{\rho}(\mathbf{u}_i, \mathcal{G}) + \epsilon, \quad i = 1, 2.$$

Since  $\hat{\rho}(\mathbf{u}, \mathcal{G})$  is quasi-convex w.r.t.  $\mathbf{u}$ , we have

$$\begin{aligned} \hat{\rho}\left(\frac{\alpha\mathbf{u}_1 + (1-\alpha)\mathbf{u}_2}{\alpha\gamma_1 + (1-\alpha)\gamma_2}, \mathcal{G}\right) &= \hat{\rho}\left(\frac{\alpha\gamma_1}{\alpha\gamma_1 + (1-\alpha)\gamma_2} \cdot \frac{\mathbf{u}_1}{\gamma_1} + \frac{(1-\alpha)\gamma_2}{\alpha\gamma_1 + (1-\alpha)\gamma_2} \cdot \frac{\mathbf{u}_2}{\gamma_2}, \mathcal{G}\right) \\ &\leq \max\left\{\hat{\rho}\left(\frac{\mathbf{u}_1}{\gamma_1}, \mathcal{G}\right), \hat{\rho}\left(\frac{\mathbf{u}_2}{\gamma_2}, \mathcal{G}\right)\right\} \end{aligned}$$

which implies

$$\begin{aligned} \tilde{\rho}(\alpha\mathbf{u}_1 + (1-\alpha)\mathbf{u}_2, \mathcal{G}) &\leq \hat{\rho}\left(\frac{\alpha\mathbf{u}_1 + (1-\alpha)\mathbf{u}_2}{\alpha\gamma_1 + (1-\alpha)\gamma_2}, \mathcal{G}\right) \leq \max\left\{\hat{\rho}\left(\frac{\mathbf{u}_1}{\gamma_1}, \mathcal{G}\right), \hat{\rho}\left(\frac{\mathbf{u}_2}{\gamma_2}, \mathcal{G}\right)\right\} \\ &\leq \max\{\tilde{\rho}(\mathbf{u}_1, \mathcal{G}), \tilde{\rho}(\mathbf{u}_2, \mathcal{G})\} + \epsilon. \end{aligned}$$

Hence  $\tilde{\rho}(\cdot)(\mathbf{u}, \mathcal{G})$  is quasi-convex w.r.t.  $\mathbf{u}$ . Note that the only property that is not satisfied is the semi-continuity. To handle this, define  $\rho : \mathcal{R}^m \times \mathcal{S} \mapsto [0, 1]$  as

$$\rho(\mathbf{u}, \mathcal{G}) = \lim_{\epsilon \downarrow 0} \tilde{\rho}(\mathbf{u} + \epsilon \mathbf{e}^m, \mathcal{G})$$

Because of monotonicity of  $\tilde{\rho}(\cdot)$ ,  $\rho(\cdot, \cdot)$  is well-defined. In addition, it can be shown that  $\rho(\cdot, \cdot)$  is lower-semicontinuous. Complete classification, misclassification avoidance, monotonicity, order invariance, scale invariance, and quasi-convexity all follows easily from the fact that same property holds for  $\tilde{\rho}(\cdot)$ . Thus,  $\rho(\cdot, \cdot)$  is a CCLF w.r.t.  $m$ . Next, we show that

$$\hat{\rho}(\mathbf{u}, \mathcal{G}) \geq \rho(\mathbf{u}, \mathcal{G}) \geq \varrho(\mathbf{u}, \mathcal{G}).$$

The first inequality holds due to  $\hat{\rho}(\mathbf{u}, \mathcal{G}) \geq \tilde{\rho}(\mathbf{u}, \mathcal{G}) \geq \tilde{\rho}(\mathbf{u} + \epsilon \mathbf{e}^m, \mathcal{G})$ . The second inequality holds because for any  $\mathbf{u}$ , there exists  $\epsilon > 0$  small enough such that  $\varrho(\mathbf{u} + \epsilon \mathbf{e}^m, \mathcal{G}) = \varrho(\mathbf{u}, \mathcal{G})$ . Thus, taking limit over  $\tilde{\rho}(\mathbf{u} + \epsilon \mathbf{e}^m, \mathcal{G}) \geq \varrho(\mathbf{u} + \epsilon \mathbf{e}^m, \mathcal{G})$  establishes the second inequality. Recall that  $\bar{\rho}(\mathbf{u}, \mathcal{G})$  is the minimal CCLF, we establish the lemma by

$$\varrho(\mathbf{u}, \mathcal{G}) \leq \bar{\rho}(\mathbf{u}, \mathcal{G}) \leq \rho(\mathbf{u}, \mathcal{G}).$$

□

#### 5.6.4 Proof of Theorem 5.5

*Proof.* To prove Theorem 5.5, we start with establishing the following lemma. Observe that  $\bar{\rho}(\mathbf{u}, \mathcal{G})$  only takes value in  $\{0, \frac{1}{n}, \frac{2}{n}, \dots, 1\}$ .

**Lemma 5.3.** *The level set of Problem (5.5), i.e.,  $\mathcal{U}_i \triangleq \{(\mathbf{u}, \mathbf{w}) : \bar{\rho}(\mathbf{u}, \mathcal{G}) \leq 1 - \frac{i}{n}; f_j(\mathbf{u}, \mathbf{w}) \leq 0, \forall j\}$  for  $i = 1, \dots, n$ , equals the following*

$$\{(\mathbf{u}, \mathbf{w}) : \exists d \text{ such that } \sum_{i=1}^n [d - \min_{j \in \mathcal{G}_i} u_j]_+ \leq (n - i + 1)d; f_j(\mathbf{u}, \mathbf{w}) \leq 0, \forall j.\}$$

*Proof.* Let  $\tilde{\mathbf{u}} = (\min_{j \in \mathcal{G}_1} u_j, \dots, \min_{j \in \mathcal{G}_n} u_j)^\top$ . From Property 2 of Theorem 5.2, we have that  $\mathcal{U}_i$  equals to the feasible set of the following program

$$\sup_{\mathbf{v} \in \bar{\mathbf{V}}_{i/n}} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \leq 0; f_j(\mathbf{u}, \mathbf{w}) \leq 0, j = 1, \dots, k.$$

Recall that  $\bar{\mathbf{V}}_{i/n} = \text{conv} \{\lambda \mathbf{e}_N \mid \lambda > 0, |\lambda| = n - i + 1\}$  we have that  $\sup_{\mathbf{v} \in \bar{\mathbf{V}}_{i/n}} (-\mathbf{v}^\top \tilde{\mathbf{u}}) \leq 0$  is equivalent to

$$\inf_{\mathbf{v}: \mathbf{0} \leq \mathbf{v} \leq \mathbf{e}, \mathbf{e}^\top \mathbf{v} = n - i + 1} \mathbf{v}^\top \tilde{\mathbf{u}} \geq 0,$$

which left-hand-side by duality theorem is equivalent to the following optimization problem



on  $(\mathbf{c}, d)$

$$\begin{aligned} \text{Maximize:} & \quad \sum_{i=1}^n c_i + (n - i + 1)d \\ \text{Subject to:} & \quad c_i + d \leq \tilde{u}_i, \quad c_i \leq 0, \quad i = 1, \dots, n. \end{aligned}$$

Thus we have  $\mathbf{u} \in \mathcal{U}_i$  if and only if there exists  $\mathbf{c}$ ,  $d$ , and  $\mathbf{w}$  such that

$$\begin{aligned} \mathbf{e}^\top \mathbf{c} + (n - i + 1)d &\geq 0; \\ \mathbf{c} + d\mathbf{e} &\leq \tilde{\mathbf{u}}; \\ \mathbf{c} &\leq \mathbf{0}; \\ f_j(\mathbf{u}, \mathbf{w}) &\leq 0, \quad j = 1, \dots, k. \end{aligned}$$

Note that this can be further simplified, since optimal  $c_i = -[d - \tilde{u}_i]_+$ , as

$$\begin{aligned} \sum_{i=1}^n [d - \tilde{u}_i]_+ &\leq (n - i + 1)d \\ f_j(\mathbf{u}, \mathbf{w}) &\leq 0, \quad j = 1, \dots, k. \end{aligned} \tag{5.21}$$

This establishes the lemma. □

Now we turn to prove Theorem 5.5. When all feasible solutions  $\mathbf{u}, \mathbf{w}$ , i.e.,  $f_j(\mathbf{u}, \mathbf{w}) \leq 0$  for all  $j = 1, \dots, k$ , satisfy that  $\mathbf{u} > 0$  or  $\mathbf{u} \not\geq \mathbf{0}$ , we only need to consider the feasible solutions to (5.21) with  $d > 0$ . Hence the feasible set to Problem (5.21) is equivalent to that of

$$\begin{aligned} \sum_{i=1}^n [1 - \tilde{u}_i/d]_+ &\leq (n - i + 1) \\ f_j(\mathbf{u}, \mathbf{w}) &\leq 0, \quad j = 1, \dots, k. \end{aligned}$$

Thus, finding the optimal solution to Problem (5.5) is equivalent to solve the following

$$\begin{aligned} \text{Minimize:} & \quad \sum_{i=1}^n [1 - \tilde{u}_i/d]_+ \\ \text{Subject to:} & \quad f_j(\mathbf{u}, \mathbf{w}) \leq 0, \quad j = 1, \dots, k; \\ & \quad d > 0. \end{aligned} \tag{5.22}$$

By a change of variable where we let  $h = 1/d$ ,  $\mathbf{s} = h\mathbf{u}$ ,  $\mathbf{t} = h\mathbf{w}$ , this is equivalent to

$$\begin{aligned} \text{Minimize:} \quad & \sum_{i=1}^n [1 - \min_{j \in \mathcal{G}_i} s_j]_+ \\ \text{Subject to:} \quad & hf_j(\mathbf{s}/h, \mathbf{t}/h) \leq 0, \quad j = 1, \dots, k; \\ & h > 0. \end{aligned}$$

Hence Theorem 5.5 is established. □

## 5.7 Chapter Summary

In this chapter, we revisit the standard cumulative-loss approach in dealing with the non-convexity of the 0-1 loss function in classification, namely minimizing the sum of convex surrogates for each sample. We propose the notion of *coherent loss*, which is a tractable upper-bound of the total classification error for the entire sample set. This approach yields a strictly tighter approximation to the 0-1 loss than any cumulative loss, while preserving the tractability of the resulting optimization problem. The formulation obtained by applying the coherent loss to binary classification also has a robustness interpretation, which builds a strong connection between the coherent loss and robust SVMs. Finally, we remark that the coherent loss approach has favorable statistical properties and the simulation results show that it can outperform the standard SVM when additional constraints are imposed.

## Online Linear Optimization with Unobserved Constraints

We consider online linear programming with unobserved constraints (LPUC) – a generalization of stochastic linear optimization – where in each round a learner chooses a solution and subsequently receives some feedback about the *feasibility* of the selected solution w.r.t. the unknown constraints, e.g., indicating which constraint is violated or how much the solution deviates from the feasibility set. To tackle this problem, we develop two algorithms, namely, LPUC-ED based on the epsilon-decreasing strategy and LPUC-UCB based on the upper confidence bound strategy, and derive finite time bounds on the regret and the constraint violation. Numerical experiments show satisfactory empirical performance of the proposed algorithms and validate our theoretical results.

### 6.1 Introduction

Linear programming (LP), which optimizes a linear objective subject to linear equality and linear inequality constraints, is undoubtedly the most extensively studied and widely applied optimization formulation and has been applied in machine learning, operations research, finance, and beyond. A vanilla LP problem can be readily solved via the simplex method or interior point methods, when the objective function and the constraints are known to the decision maker [BT97]. In many cases, however, such exact knowledge may not be available. Optimization under uncertainty is a fast growing research field, and classical methods such as *stochastic programming* [BL97] and *robust optimization* [BTN98] take a static view –

some historical observations of the parameters are given, based on which the decision maker attempts to obtain a decision. In this paper, we tackle LPUC from a dynamic perspective: the decision maker can make a tentative decision, collect feedback information about the decision, and fine tune the decision, essentially solving the LP problem via *trial and error*.

We motivate our setup using the following example. Network flow problems, often used to model traffic in a road system, packet flow through network and circulation with demands, etc., can be formulated as LP problems. The decision maker who aims to find the maximum flow or the minimum-cost flow does not always know the capacities or costs of all the edges in the network exactly. To see this, imagine a decision maker who determines how to dispatch vehicles on a transportation network, it is not surprising that she does not know the precise traffic condition in each and every road when she makes the decision. Instead, such information is available only when these vehicles are on the roads (and can then provide accurate traffic reports). The goal of this paper is then to develop methods to leverage such post-decision information to obtain near optimal solutions in a learning fashion.

Specifically, we study linear programming problems with unknown constraints (LPUC). To gather the information about the unknown constraints, we consider an online setting where the decision maker or learner selects a solution in each round and then receives corresponding feedbacks providing information about the feasibility of the selected solution. As an example, consider that routers forward data packets through a data network and observe packet delays due to congestion (i.e, flows exceed the edge capacities). The goal is to find solutions close to the optimal solution of the unknown LP. Without loss of generality, we assume that the objective function is known because otherwise we can convert the original problem into its epigraph form [BV04]. This model generalizes both stochastic linear optimization [DHK08, RT08] and multi-armed bandit problems [FCGS10], allowing to tackle a broad class of problems. The main complicating factor in this model is that the selected solutions are not always feasible to the original problem due to the unobserved constraints. To the best of our knowledge, this problem has not been explored yet.

To tackle this problem, we develop two algorithms – LPUC-ED based on the epsilon-decreasing strategy [KP00] and LPUC-UCB based on the upper confidence bound strategy

[ACBF02, AMS09, FCGS10, AYPS11]. We measure their performance using *two metrics simultaneously*, namely, *regret* – the difference between the learner’s cumulated cost and the cost of the optimal strategy, and *constraint violation* – an indicator of level of constraint violation over the  $T$  rounds. We show that the regret and constraint violation of LPUC-ED are  $O(dT^{\frac{2}{3}} \log T)$  and  $O(dT^{\frac{2}{3}})$  respectively, whereas those of LPUC-UCB are both  $O(d\sqrt{T} \log T)$ . LPUC-UCB achieves a better regret than LPUC-ED and matches the lower bound of the linear bandit problem [DHK08] up to a logarithmic factor, but is computationally more demanding than LPUC-ED.

**Notations:** We use boldface lower-case letters to represent column vectors and capital letters for matrices, and use  $[c]_+$  to denote  $\max\{0, c\}$ . For matrix  $\mathbf{M}$ ,  $\|\mathbf{M}\|_2$  denotes its spectral norm. We use  $\mathbf{e}_1, \dots, \mathbf{e}_d$  to represent the standard basis of  $\mathbb{R}^d$  and define  $\mathbf{e}_{d+1} \triangleq \mathbf{0}$  for convenience, and use  $\mathcal{S}_{d-1}(B)$  to denote the unit sphere  $\{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 = B\}$ .

## 6.2 Related Work

LPUC can be viewed as a generalized version of stochastic linear optimization and linear bandits. In the classical multi-armed bandit problem, in each of  $T$  rounds, the learner selects one of  $K$  arms and subsequently receives a reward independently drawn from an unknown distribution associated with the selected arm. The goal of the learner is to choose a sequence of arms to maximize the cumulated rewards over the  $T$  rounds. This problem has been extensively studied for decades, e.g., [Lai87, ACBF02, CBL06, PCA07, BSSM08, MS11], and various efficient algorithms based on upper confidence bound (UCB) or Thompson sampling (TS) have been proposed, e.g., [Agr95, ACBF02, CL11, AG12].

An extension of the classical multi-armed bandit problem is the contextual multi-armed bandit problem in which each arm associates with a  $d$ -dimensional feature vector called “context” and the reward corresponding to each arm depends on the associated feature vector. The learner’s aim is to explore the relationship between the feature vectors and rewards so that she can predict which arm could provide best reward by examining the feature vectors. The contextual bandits setting with linear payoff functions was studied by

[AL99, ACBF02] and further analyzed by [CLRS11, FCGS10, AYPS11]. In this setting, the learner competes with the set of all linear predictors on the feature vectors, e.g., we assume that there exists an unknown parameter  $\mathbf{c}$  such that the expected reward for an arm given feature vector  $\mathbf{x}$  is  $\mathbf{c}^\top \mathbf{x}$ .

When  $\mathcal{S}$  – the set of the feature vectors associated with the arms – is very large or even infinite, this problem is also called “stochastic linear optimization” [DKH07, DHK08, RT08, Sha13, Sha15b]. One of the most important examples is online linear programming, in which the aim is to minimize the cost function  $\mathbf{c}^\top \mathbf{x}$  with the constraint  $\mathbf{x} \in \mathcal{S}$  where  $\mathcal{S}$  is specified by known linear inequality constraints. Different from the linear programming problem with known cost vector  $\mathbf{c}$ , the learner only observes noisy feedback about  $\mathbf{c}$  corresponding to the value of the objective for the selected solution. Compared with the previous work, this paper considers online linear programming with additional unknown constraints besides  $\mathbf{x} \in \mathcal{S}$  and tries to find its optimal solution instead of only minimizing the regret.

## 6.3 Problem Setting

Consider the following linear programming problem:

$$\min \mathbf{c}^\top \mathbf{x}, \text{ s.t. } \mathbf{A}^\top \mathbf{x} \leq \mathbf{b}, \mathbf{x} \in \mathcal{S}, \quad (6.1)$$

where  $\mathbf{c} \in \mathbb{R}^d$ ,  $\mathbf{A} \in \mathbb{R}^{d \times m}$ ,  $\mathbf{b} \in \mathbb{R}^m$ , and  $\mathcal{S}$  is a convex polytope. This paper concerns to find its optimal solution in the case that  $\mathbf{A}$  and  $\mathbf{b}$  are unknown but  $\mathbf{c}$  is known, and assumes that for any input  $\mathbf{x} \in \mathcal{S}$  the system provides us some feedback about  $\mathbf{A}^\top \mathbf{x} - \mathbf{b}$  indicating how much  $\mathbf{x}$  deviates from the feasibility set.

### 6.3.1 Learning Model

More generally, we consider to solve a sequence of linear programming problems  $\{\mathcal{P}_1, \dots, \mathcal{P}_T\}$ .

For each  $t$ ,  $\mathcal{P}_t$  has the following formulation:

$$\min \mathbf{c}^\top \mathbf{x}, \text{ s.t. } \mathbf{A}^\top \mathbf{x} \leq \mathbf{b}, \mathbf{x} \in \mathcal{S}_t, \quad (6.2)$$

where  $\mathcal{S}_t$  is a convex polytope and  $\mathbf{A}, \mathbf{b}, \mathbf{c}$  are shared for all  $t$ . Let  $\mathbf{x}^*(t)$  be the optimal solution of  $\mathcal{P}_t$ . Clearly, when  $\mathcal{S}_t = \mathcal{S}$ , Problem (6.2) is reduced to (6.1). We suppose that the constraint parameters  $\mathbf{A}, \mathbf{b}$  are unknown, and tackle this problem in the following online setting. In each round  $t$ , the learner receives linear program  $\mathcal{P}_t$  and chooses a solution  $\mathbf{x}(t)$  for  $\mathcal{P}_t$ . After  $\mathbf{x}(t)$  is submitted, she receives the corresponding feedback  $\mathbf{r}(t)$  whose  $i^{\text{th}}$  entry  $r_i(t) = f(\mathbf{a}_i^\top \mathbf{x}(t) - b_i) + \xi_i(t)$ , where  $\mathbf{a}_i$  is the  $i$ th column of  $\mathbf{A}$ ,  $\xi_i(t)$  is a random noise with mean  $\mathbf{0}$  and  $f(\cdot)$  is a non-decreasing function. Without loss of generality, we assume that  $f(0) = 0$ . The goal of the learner is to find the optimal solution of  $\mathcal{P}_t$  as  $t$  grows.

If the cost vector  $\mathbf{c}$  in Problem (6.2) is also unknown, one can convert Problem (6.2) into its epigraph form. This means that the problem studied in this paper is more general than linear stochastic bandits discussed in [DHK08, RT08, FCGS10] in which  $\mathbf{c}$  is unknown and no additional constraints such as  $\mathbf{A}^\top \mathbf{x} \leq \mathbf{b}$  are imposed. More specifically, recall that in linear stochastic bandit problems, the learner has to choose an action  $\mathbf{x}(t)$  given decision set  $\mathcal{S}_t$  in the  $t^{\text{th}}$  round and then receives feedback  $r(t) = \mathbf{c}^\top \mathbf{x}(t) + \xi(t)$  where  $\xi(t)$  is a random noise. This is equivalent to that she selects a solution  $\{\mathbf{x}(t), \beta(t)\}$  of the following linear programming problem

$$\min \beta, \text{ s.t. } \mathbf{c}^\top \mathbf{x} - \beta \leq 0, \mathbf{x} \in \mathcal{S}_t,$$

and subsequently observes feedback  $r(t) = \mathbf{c}^\top \mathbf{x}(t) - \beta(t) + \xi(t)$ . Therefore, the linear stochastic bandit problem is indeed a special case of our setting.

The feedback  $r_i(t) = f(\mathbf{a}_i^\top \mathbf{x}(t) - b_i) + \xi_i(t)$  has been considered in the generalized linear bandit model proposed by [FCGS10]. They showed that this model has a strong connection with the generalized linear models and allows to model various feedback structures. For example, the simplest choice of  $f(\cdot)$  is  $f(x) = x$ , leading to the linear bandit feedback. When only the signs of  $\mathbf{a}_i^\top \mathbf{x}(t) - b_i$  are revealed in each round, namely, the system tells us which constraints are violated for solution  $\mathbf{x}(t)$ , one suitable choice is  $f(x) = (\exp(x) - 1)/(\exp(x) + 1)$  which is an approximation of the sign function. With this kind of feedback, they generalized the linear multi-armed bandit to the non-linear case and developed a new algorithm called GLM-UCB.

### 6.3.2 Assumptions

This problem is hard to solve if there are no assumptions on decision sets  $\mathcal{S}_t$ , noise  $\xi_i(t)$  and function  $f(\cdot)$ . In order to achieve meaningful empirical and theoretical results, we make the following assumptions. Roughly speaking, we assume that 1)  $\mathcal{S}_t$  is bounded, 2)  $\xi_i(t)$  is supported on a bounded interval and has mean  $\mathbf{0}$ , 3)  $f(\cdot)$  is strictly increasing, and 4) the constraint in (6.2) is feasible and regular.

**Assumption 6.1.** *For any  $t = 1, \dots, T$ , Problem (6.2) is always feasible, i.e.,  $\mathcal{S}_t \cap \{\mathbf{x} : \mathbf{A}^\top \mathbf{x} \leq \mathbf{b}\} \neq \emptyset$ , and there exists constants  $L$  and  $B$  so that  $\|\mathbf{x}\|_2 \leq L$  for any  $\mathbf{x} \in \mathcal{S}_t$  and  $[-B, B]^d \subseteq \mathcal{S}_t$ .*

Note that when the intersection of sets  $\mathcal{S}_t$  for  $t = 1, \dots, T$  is nonempty, the assumption  $[-B, B]^d \subseteq \mathcal{S}_t$  can be always satisfied by shifting  $\mathcal{S}_t$ . Since the objective function and the constraints in (6.2) are linear, this assumption can be made without loss of generality.

**Assumption 6.2.** *The function  $f(\cdot)$  is continuously differentiable, Lipschitz continuous with constant  $l_\mu$ , and satisfies  $c_\mu = \inf_{\mathbf{x} \in \bigcup_{t=1}^T \mathcal{S}_t, (\mathbf{a}, b) \in \bigcup_{i=1}^m \mathcal{A}_i} \left. \frac{df(z)}{dz} \right|_{z=\mathbf{a}^\top \mathbf{x} - b} > 0$ .*

Here  $\mathcal{A}_i$  represents the admissible sets for  $\mathbf{a}_i$  and  $b_i$ , i.e.,  $(\mathbf{a}_i, b_i) \in \mathcal{A}_i$ . Assumption 6.2 implies that function  $f(\cdot)$  has an inverse, which means  $\mathbf{a}_i^\top \mathbf{x}(t) - b_i$  can be evaluated via measuring  $f(\mathbf{a}_i^\top \mathbf{x}(t) - b_i)$ .

**Assumption 6.3.** *Random variables  $\xi_i(t)$  for  $i = 1, \dots, m$  are i.i.d. with mean 0 and support  $[-R, R]$ .*

Actually, when  $\xi_i(t)$  is a  $R$ -sub-Gaussian random variable, our results can still hold.

**Assumption 6.4.** *The constraint  $\mathbf{A}^\top \mathbf{x} \leq \mathbf{b}$ ,  $\mathbf{x} \in \mathcal{S}_t$  is regular, i.e.,  $\mathbf{b}$  is an interior point of  $\{\mathbf{A}^\top \mathbf{x} + \mathbf{z} : \mathbf{x} \in \mathcal{S}_t, \mathbf{z} \in \mathbb{R}_+^m\}$  where  $\mathbb{R}_+^m$  denotes the non-negative orthant in  $\mathbb{R}^m$ .*

The ‘‘regular’’ assumption implies that this constraint is still feasible when  $\mathbf{A}$  or  $\mathbf{b}$  has some small perturbation.



### 6.3.3 Performance Metric

Recall that the desirable solutions should be approximately feasible and optimal at the same time. To measure “optimality”, we consider the following *absolute regret* (or *regret* for short):

$$\text{Regret}(T) = \sum_{t=1}^T |\mathbf{c}^\top \mathbf{x}(t) - \mathbf{c}^\top \mathbf{x}^*(t)|,$$

which is different from the traditional regret that sums  $\mathbf{c}^\top \mathbf{x}(t) - \mathbf{c}^\top \mathbf{x}^*(t)$  without taking their absolute values. The reason why we use the absolute regret is as follows. In the linear bandit problem where decision set  $\mathcal{S}_t$  is given and there are no additional unobserved constraints, it is guaranteed that  $\mathbf{c}^\top \mathbf{x}(t) \geq \mathbf{c}^\top \mathbf{x}^*(t)$  for all  $t$ . This is not the case in our setting however, due to the existence of the additional unknown constraint  $\mathbf{A}^\top \mathbf{x} \leq \mathbf{b}$ , that is,  $\mathbf{c}^\top \mathbf{x}(t)$  can be much less than  $\mathbf{c}^\top \mathbf{x}^*(t)$  because the precise information about the feasibility of  $\mathbf{x}(t)$  is not available to the learner so that  $\mathbf{x}(t)$  can be infeasible w.r.t. the unknown constraint, resulting in a lower cost than  $\mathbf{c}^\top \mathbf{x}^*(t)$  and making the traditional regret meaningless as the sum contains both positive and negative terms.

To measure “feasibility”, we define the following metric called *constraint violation* indicating the level that constraint  $\mathbf{A}^\top \mathbf{x} \leq \mathbf{b}$  is violated:

$$\text{Violation}(T) = \sum_{i=1}^m \sum_{t=1}^T [\mathbf{a}_i^\top \mathbf{x}(t) - b_i]_+.$$

Measuring constraint violation is necessary because playing an infeasible solution may have additional penalties in practical applications, e.g., congestion and delays. Therefore, our aim is to design policies with both the regret and constraint violation growing sub-linearly in  $T$ .

## 6.4 Two Algorithms: LPUC-ED and LPUC-UCB

We first provide a naive sampling approach to solve this problem. In the first  $T - 1$  rounds, the learner randomly draws  $\mathbf{x}(t)$  from decision sets  $\mathcal{S}_t$  regardless of whether constraint  $\mathbf{A}^\top \mathbf{x}(t) \leq \mathbf{b}$  are satisfied or not, and then uses these selected solutions  $\mathbf{x}(t)$  and the cor-

responding feedbacks to estimate  $\mathbf{A}$  and  $\mathbf{b}$ . Finally, in the  $T^{\text{th}}$  round, he solves (6.2) with the obtained estimation of  $\mathbf{A}$  and  $\mathbf{b}$  and takes its optimal solution as  $\mathbf{x}(T)$ . One example of this approach is shown in Algorithm 6.1 where we assume that  $f(x) = x$ . We will show in

---

**Algorithm 6.1:** Naive sampling approach
 

---

**Input** : Vector  $\mathbf{c} \in \mathbf{R}^d$ , bounded sets  $\mathcal{S}_t$ , parameter  $B$ .

- 1 **for**  $t = 1$  to  $T - 1$  **do**
- 2     Draw  $\mathbf{x}(t)$  from  $\mathcal{S}_{d-1}(B)$  uniformly at random;
- 3     Play  $\mathbf{x}(t)$  and receive  $\mathbf{r}(t)$ ;
- 4 **end**
- 5 Let  $\hat{\Sigma} = \frac{1}{T-1} \sum_{t=1}^{T-1} \mathbf{x}(t)\mathbf{x}(t)^\top$  and calculate  $\hat{\mathbf{A}} = \hat{\Sigma}^{-1} \sum_{t=1}^{T-1} \mathbf{x}(t)\mathbf{r}(t)^\top$  and  $\hat{\mathbf{b}} = -\frac{1}{T-1} \sum_{t=1}^{T-1} \mathbf{r}(t)$ ;
- 6 Play  $\mathbf{x}(T)$  – the optimal solution of Problem (6.2) with  $\mathbf{A} = \hat{\mathbf{A}}$  and  $\mathbf{b} = \hat{\mathbf{b}}$ .

---

the next section that although this algorithm can guarantee  $|\mathbf{c}^\top \mathbf{x}(T) - \mathbf{c}^\top \mathbf{x}^*(T)| = O(\frac{1}{\sqrt{T}})$ , both of its regret and constraint violation can be  $\Omega(T)$ . This happens because there is no tradeoff between “exploitation” and “exploration” – making good decisions and probing more information about the constraints, e.g., the first  $T - 1$  rounds make the regret grow linearly in  $T$ , where all  $\mathbf{x}(t)$  are randomly drawn from  $\mathcal{S}_{d-1}(B)$  to explore  $\mathbf{A}$  and  $\mathbf{b}$ .

One possible approach to make a good balance between exploitation and exploration is as follows. In each round, with some probability  $p$  the learner tries to explore more information about the constraints, while with probability  $1 - p$  he chooses the optimal solution of (6.2) with the current estimates of  $\mathbf{A}$  and  $\mathbf{b}$ . Based on this idea, we propose Algorithm 6.2 – linear programming with unobserved constraints via the epsilon-decreasing strategy (LPUC-ED) (“epsilon-decreasing” means the exploration probability  $p$  decreases as  $t$  grows).

In the  $t^{\text{th}}$  round, the first step of Algorithm 6.2 is to estimate  $\mathbf{A}$  and  $\mathbf{b}$  based on the information  $(\mathbf{x}(1), \dots, \mathbf{x}(t-1), \mathbf{r}(1), \dots, \mathbf{r}(t-1))$  obtained before round  $t$ . For convenience, we let  $\mathbf{y}(t) = (\mathbf{x}(t)^\top, -1)^\top$  and define several useful quantities:

$$\mathbf{M}_t \triangleq \sum_{k=1}^{t-1} \mathbf{y}(k)\mathbf{y}(k)^\top, \quad g_t(\mathbf{z}) \triangleq \sum_{k=1}^{t-1} f(\mathbf{z}^\top \mathbf{y}(k))\mathbf{y}(k), \quad (6.3)$$

and

$$\mathbf{g}_t^i \triangleq \sum_{k=1}^{t-1} r_i(k) \mathbf{y}(k), \quad \forall i = 1, \dots, m.$$

Suppose that the admissible set for  $(\mathbf{a}_i, b_i)$  is  $\mathcal{A}_i$  which is known. The new estimates of  $\mathbf{a}_i$  and  $b_i$  in round  $t$  can be computed by solving the following optimization problem:

$$\mathbf{a}_i(t), b_i(t) = \arg \min_{(\mathbf{a}, b) \in \mathcal{A}_i} \|g_t((\mathbf{a}^\top, b)^\top) - \mathbf{g}_t^i\|_{\mathbf{M}_t^{-1}}. \quad (6.4)$$

As discussed in [FCGS10], this problem can be easily solved via Newton's method.

The second step is to select proper  $\mathbf{x}(t)$  by solving Problem (6.2) with the current estimates of  $\mathbf{A}$  and  $\mathbf{b}$  as shown in (6.5), which involves two issues: 1) Problem (6.5) is not always feasible since  $\mathbf{a}_i(t)$  and  $b_i(t)$  are not identical to  $\mathbf{a}_i$  and  $b_i$ , and 2) the feedback corresponding to  $\mathbf{x}(t)$  conveys less new information about  $\mathbf{A}$  and  $\mathbf{b}$  as  $\mathbf{x}(t)$  becomes closer to  $\mathbf{x}^*(t)$ , which means exploration is required. To address these two issues, we sample  $\mathbf{x}(t)$  from the uniform distribution on  $\mathcal{S}_{d-1}(B)$  to explore more information about  $\mathbf{A}$  and  $\mathbf{b}$  when Problem (6.5) is infeasible or  $\tilde{\eta}(t)$  – a Bernoulli random variable with parameter  $p(t)$  – equals 1.

We prove in the next section that the regret and the constraint violation for Algorithm 6.2 are at most  $O(dT^{\frac{2}{3}} \log T)$  and  $O(dT^{\frac{2}{3}})$ , respectively. Then the question is: Can we obtain a regret bound better than  $O(dT^{\frac{2}{3}} \log T)$ ?

To achieve a better regret bound, we develop Algorithm 6.3 – Linear programming with unobserved constraints via UCB (LPUC-UCB) – that chooses  $\mathbf{x}(t)$  by solving a non-convex optimization problem as shown in (6.6) without explicitly exploring the information about the constraints via sampling  $\mathbf{x}(t)$  from  $\mathcal{S}_{d-1}(B)$ .

In general, it is difficult to find the global optimal solution of Problem (6.6) due to the non-convexity of its constraints. But in some special cases, e.g.,  $f(\cdot)$  is convex or  $\mathcal{S}_t$  is discrete, it can be solved efficiently. When  $f(\cdot)$  is convex, Problem (6.6) is a DC (difference of convex functions) programming problem that can be solved by many DC algorithms [HN99, AT05, MTA10] proposed in recent years. The DC algorithms have been successfully applied to a large number of non-convex optimization problems to which they quite often find global optimal solutions efficiently. When  $\mathcal{S}_t$  is a finite discrete set, one can solve (6.6)

---

**Algorithm 6.2:** Linear programming with unobserved constraints via the epsilon-decreasing strategy (LPUC-ED)

---

**Input** : Cost vector  $\mathbf{c} \in \mathbf{R}^d$ , decision sets  $\mathcal{S}_t$ .

- 1 Play  $\mathbf{e}_1, \dots, \mathbf{e}_{d+1}$  and receive  $\mathbf{r}(1), \dots, \mathbf{r}(d+1)$ ;
- 2 **for**  $t = d+2$  to  $T$  **do**
- 3     Calculate  $\mathbf{M}_t$  and  $\mathbf{g}_t^i$  for  $i \in [m]$ ;
- 4     Compute  $\mathbf{a}_i(t), b_i(t)$  for  $i \in [m]$  via solving (6.4);
- 5     Compute the optimal solution  $\hat{\mathbf{x}}(t)$  of the following linear program:
 
$$\begin{aligned} \min \quad & \mathbf{c}^\top \mathbf{x} \\ \text{s.t.} \quad & \mathbf{a}_i(t)^\top \mathbf{x} \leq b_i(t), \quad \forall i \in [m], \\ & \mathbf{x} \in \mathcal{S}_t. \end{aligned} \tag{6.5}$$
- 6     Set variable  $\eta(t)$  to
 
$$\eta(t) = \begin{cases} \tilde{\eta}(t), & \text{Problem (6.5) is feasible,} \\ 1, & \text{otherwise} \end{cases}$$

where  $\tilde{\eta}(t)$  is drawn from Bernoulli distribution with success probability  $p(t) \propto 1/t^{1/3}$ ;
- 7     Play  $\mathbf{x}(t) = [1 - \eta(t)]\hat{\mathbf{x}}(t) + \eta(t)\tilde{\mathbf{x}}(t)$  and receive  $\mathbf{r}(t)$ , where  $\tilde{\mathbf{x}}(t)$  follows the uniform distribution on  $\mathcal{S}_{d-1}(B)$ ;
- 8 **end**

---

by evaluating each element in  $\mathcal{S}_t$ , i.e., selecting the one that is feasible and has the smallest objective value. In the next section, we show that the regret bound and the constraint violation of Algorithm 6.3 are  $O(d\sqrt{T} \log T)$ .

Note that in each round, both of Algorithm 6.2 and Algorithm 6.3 require to solve Problem (6.4) to update the estimates of  $\mathbf{A}$  and  $\mathbf{b}$ , which could have high computational cost when

---

**Algorithm 6.3:** Linear programming with unobserved constraints via UCB (LPUC-UCB)

---

**Input** : Cost vector  $\mathbf{c} \in \mathbf{R}^d$ , decision sets  $\mathcal{S}_t$  and parameter  $\theta(t)$ .

1 Play  $\mathbf{e}_1, \dots, \mathbf{e}_{d+1}$  and receive  $\mathbf{r}(1), \dots, \mathbf{r}(d+1)$ ;

2 **for**  $t = d+2$  to  $T$  **do**

3     Calculate  $\mathbf{M}_t$  and  $\mathbf{g}_t^i$  for  $i \in [m]$ ;

4     Compute  $\mathbf{a}_i(t), b_i(t)$  for  $i \in [m]$  via solving (6.4);

5     Solve the optimization problem:

$$\begin{aligned}
 \min \quad & \mathbf{c}^\top \mathbf{x} \\
 \text{s.t.} \quad & f(\mathbf{a}_i(t)^\top \mathbf{x} - b_i(t)) \leq \theta(t) \|\mathbf{y}\|_{\mathbf{M}_t^{-1}}, i \in [m], \\
 & \mathbf{y} = (\mathbf{x}^\top, -1)^\top, \\
 & \mathbf{x} \in \mathcal{S}_t,
 \end{aligned} \tag{6.6}$$

and denote the optimal solution by  $\hat{\mathbf{x}}(t)$ ;

6     Play  $\mathbf{x}(t) = \hat{\mathbf{x}}(t)$  and receive  $\mathbf{r}(t)$ ;

7 **end**

---

$d$  and  $T$  are large. By following the idea in [AYPS11], we propose Algorithm 6.4 to accelerate these two algorithms. Instead of computing  $\mathbf{a}_i(t), b_i(t)$  in each round, Algorithm 6.4 recomputes them only when  $\det(\mathbf{M}_t)$  increases by a constant factor  $1 + \gamma$ . It can be shown that (6.4) only needs to be solved  $O(\log T)$  times and hence saves computation.

## 6.5 Regret Bound and Constraint Violation

We now provide the upper bounds for the regret and the constraint violation of Algorithms 6.2, 6.3 and 6.4. Before the main theorems are given, we first show that the naive sampling approach is not able to achieve sub-linear regret.

**Proposition 6.1.** *Under Assumptions 6.1-6.4, the followings hold: 1) for  $\delta > 0$ , when  $T$  is*

**Algorithm 6.4:** Accelerated LPUC

---

**Input** : Cost vector  $\mathbf{c} \in \mathbf{R}^d$ , decision sets  $\mathcal{S}_t$ , and parameter  $\gamma > 0$ .

- 1 Play  $\mathbf{e}_1, \dots, \mathbf{e}_{d+1}$  and receive  $\mathbf{r}(1), \dots, \mathbf{r}(d+1)$ ;
- 2 **for**  $t = d+2$  to  $T$  **do**
- 3     Compute  $\mathbf{M}_t$  and  $\mathbf{g}_t^i$  for  $i \in [m]$ ;
- 4     **if**  $\det(\mathbf{M}_t) > (1 + \gamma) \det(\mathbf{M}_{t-1})$  **then**
- 5         Update  $\mathbf{a}_i(t), b_i(t)$  for  $i \in [m]$  via solving (6.4);
- 6     **end**
- 7     Run steps (4), (5) and (6) in Algorithm 6.2;
- 8     Run steps (4) and (5) in Algorithm 6.3;
- 9 **end**

---

large enough,  $\mathbf{x}(T)$  generated by Algorithm 6.1 satisfies  $|\mathbf{c}^\top \mathbf{x}(T) - \mathbf{c}^\top \mathbf{x}^*(T)| = O(\sqrt{\frac{md}{T} \log \frac{md}{\delta}})$  with probability at least  $1 - 3\delta$ , and 2) there exists an instance of Problem (6.2) so that the regret and the constraint violation of Algorithm 6.1 are both  $\Omega(T)$ .

In the following parts, we assume that the decision sets  $\mathcal{S}_t$  are convex polytopes, under which Problem (6.2) is a standard linear programming problem. We will discuss the case where  $\mathcal{S}_t$  are bounded convex sets in the next section. Theorem 6.1 provides the upper bounds for the regret and the constraint violation of Algorithm 6.2.

**Theorem 6.1.** *Under Assumptions 6.1-6.4, there exist constants  $c, c_1, c_2, c_3$  so that for  $0 < \delta < 1$ , when*

$$\theta(t) \triangleq \frac{c_1 l_\mu R}{c_\mu} \sqrt{d \log \frac{2m(L^2 + 1)t}{\delta d}},$$

and

$$T \geq T_0 \triangleq c_2 \left( \frac{l_\mu R d}{c_\mu^2} \sqrt{m \log \frac{m^{3/2}}{c_\mu \delta}} \right)^3,$$

with probability at least  $1 - 2\delta - cT^{-9}$  the regret and the constraint violation of Algorithm

6.2 satisfy

$$\begin{aligned} \text{Regret}(T) &\leq 2T_0L\|\mathbf{c}\|_2 + \\ &\quad c_3 \left[ \frac{\theta(T)\sqrt{md}}{c_\mu} + L\|\mathbf{c}\|_2 \right] T^{2/3} \sqrt{\log T}, \\ \text{Violation}(T) &\leq (T_0 + c_3T^{2/3}) \sum_{i=1}^m (L\|\mathbf{a}_i\|_2 + |b_i|) + \\ &\quad \frac{10m\theta(T)}{c_\mu} \sqrt{dT \log T}. \end{aligned}$$

**Remark.** The proof can be found in the appendix which also shows that if all the feasible sets of Problem (6.2) for  $t = 1, \dots, T$  contain  $\mathcal{S}_{d-1}(B)$ , the term  $c_3T^{2/3}$  in the upper bound of  $\text{Violation}(T)$  can be removed, which leads to a  $O(d\sqrt{T} \log T)$  constraint violation.

The following theorem demonstrates the theoretical performance guarantee of Algorithm 6.3.

**Theorem 6.2.** *Under Assumptions 6.1-6.4, there exist constants  $c_1, c_2$  so that for  $0 < \delta < 1$ , when*

$$\theta(t) = \frac{c_1 l_\mu R}{c_\mu} \sqrt{d \log \frac{2m(L^2 + 1)t}{\delta d}}, \quad T > d + 1,$$

*with probability at least  $1 - 2\delta$  the regret and the constraint violation of Algorithm 6.3 satisfy*

$$\begin{aligned} \text{Regret}(T) &\leq 2(d+1)L\|\mathbf{c}\|_2 + \frac{c_2\theta(T)\|\mathbf{c}\|_2}{c_\mu} \sqrt{dT \log T}, \\ \text{Violation}(T) &\leq (d+1) \sum_{i=1}^m (L\|\mathbf{a}_i\|_2 + |b_i|) + \\ &\quad \frac{20m\theta(T)}{c_\mu} \sqrt{dT \log T}. \end{aligned}$$

**Remark.** Theorem 6.2 states that the upper bounds for the regret and the constraint violation of Algorithm 6.3 are at most  $O(d\sqrt{T} \log T)$ . We will show in the next section that this is also true when  $\mathcal{S}_t$  are bounded convex sets. [DHK08] have proved that the regret for the linear bandit problem with arbitrary compact decision sets has a  $\Omega(d\sqrt{T})$  lower bound. Since the problem discussed here is a general form of the linear bandit problem, the upper bounds achieved by Algorithm 6.3 are nearly optimal, i.e., they match the lower bound up to a logarithmic factor.

The next theorem shows the regret bounds achieved by Algorithm 6.4 are essentially the same as those for Algorithms 6.2 and 6.3.

**Theorem 6.3.** *Denote by  $R(T)$  and  $V(T)$  the upper bounds for the regret and the constraint violation of Algorithm 6.2 (or Algorithm 6.3), then the regret and the constraint violation of Algorithm 6.4 corresponding to Algorithm 6.2 (or Algorithm 6.3) are at most  $\sqrt{1+\gamma}R(T)$  and  $\sqrt{1+\gamma}V(T)$ , respectively.*

## 6.6 Extension to General Cases

The previous sections mainly focus on studying the linear programming problem with unknown constraints when decision sets  $\mathcal{S}_t$  are convex polytopes and the additional constraints are linear. We now consider several extensions:

**Case 1:** Decision sets  $\mathcal{S}_t$  are discrete, each of which contains  $K$  elements, namely,  $\mathcal{S}_t = \{\mathbf{x}_{t,1}, \dots, \mathbf{x}_{t,K}\}$  for  $t = 1, \dots, T$ . In this case, one can directly apply Algorithm 6.3 since the optimal solution of (6.5) can be efficiently solved by evaluating each element in  $\mathcal{S}_t$  and selecting the one that is feasible and has the smallest objective value.

**Case 2:** Decision sets  $\mathcal{S}_t$  are closed bounded convex sets beyond convex polytopes. Actually, our algorithms are still workable when  $\mathcal{S}_t$  are arbitrary compact sets. But Problems (6.5) and (6.6) become much harder to solve if  $\mathcal{S}_t$  are non-convex, so only convex decision sets are considered here.

**Case 3:** The linear constraint  $\mathbf{A}^\top \mathbf{x} \leq \mathbf{b}$  in Problem (6.2) is replaced by  $\mathbf{A}^\top \phi(\mathbf{x}) \leq \mathbf{b}$  where  $\mathbf{A}$  and  $\mathbf{b}$  are unknown while mapping  $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^{d_\phi}$  is known. Note that the constraints considered here is more general than those of linear programming. Suppose that one wants to minimize a linear cost function with the following robust linear constraints:

$$(\mathbf{a}_i + \boldsymbol{\delta})^\top \mathbf{x} \leq b_i, \quad \forall \boldsymbol{\delta} \in \mathcal{U}, \quad \forall i = 1, \dots, m,$$

where parameter  $\boldsymbol{\delta}$  is uncertain and  $\mathcal{U}$  is the corresponding uncertainty set. Under the setting described in Section 6.3.1, for solution  $\mathbf{x}(t)$  chosen in the  $t^{\text{th}}$  round, we assume that



the feedback corresponding to  $\mathbf{x}(t)$  is  $\mathbf{r}(t) = (r_1(t), \dots, r_m(t))$  where

$$r_i(t) = \max_{\boldsymbol{\delta} \in \mathcal{U}} f((\mathbf{a}_i + \boldsymbol{\delta})^\top \mathbf{x}(t) - b_i) + \xi_i(t),$$

i.e.,  $\mathbf{r}(t)$  reveals the information about the feasibility of  $\mathbf{x}(t)$  under the robust linear constraints. Different uncertainty sets lead to different functions  $\phi(\cdot)$ :

**Example 1:** When the uncertainty set  $\mathcal{U} = \{\boldsymbol{\delta} : \|\boldsymbol{\delta}\| \leq \lambda\}$ , the robust linear constraints become

$$\max_{\|\boldsymbol{\delta}\| \leq \lambda} \{(\mathbf{a}_i + \boldsymbol{\delta})^\top \mathbf{x}\} \leq b_i, \quad \forall i = 1, \dots, m.$$

For each  $i = 1, \dots, m$ , we have

$$\max_{\|\boldsymbol{\delta}\| \leq \lambda} \{(\mathbf{a}_i + \boldsymbol{\delta})^\top \mathbf{x}\} \leq b_i \Leftrightarrow \mathbf{a}_i^\top \mathbf{x} + \max_{\|\boldsymbol{\delta}\|_2 \leq \lambda} \boldsymbol{\delta}^\top \mathbf{x} \leq b_i \Leftrightarrow \mathbf{a}_i^\top \mathbf{x} + \lambda \|\mathbf{x}\|_* \leq b_i,$$

where  $\|\cdot\|_*$  is the dual norm of  $\|\cdot\|$ . Therefore, the robust linear constraints are equivalent to  $\mathbf{a}_i^\top \mathbf{x} + \lambda \|\mathbf{x}\|_* \leq b_i$  for  $i = 1, \dots, m$ . Suppose that parameters  $\mathbf{a}_i$ ,  $b_i$  and the “radius” of the uncertainty set  $\lambda$  are unknown, we can define  $\phi(\mathbf{x}) = (\mathbf{x}^\top, \|\mathbf{x}\|_*)^\top$ .

**Example 2:** When  $\mathcal{U} = \{\boldsymbol{\delta} : |\delta_i| \leq \lambda_i, \forall i = 1, \dots, d\}$ , the robust linear constraints are equivalent to

$$\max_{-\lambda \leq \boldsymbol{\delta} \leq \lambda} \{(\mathbf{a}_i + \boldsymbol{\delta})^\top \mathbf{x}\} \leq b_i, \quad \forall i = 1, \dots, m,$$

where  $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_d)^\top$ . Thus, for each  $i = 1, \dots, m$ , we have

$$\begin{aligned} & \mathbf{a}_i^\top \mathbf{x} + \max_{-\lambda \leq \boldsymbol{\delta} \leq \lambda} \boldsymbol{\delta}^\top \mathbf{x} \leq b_i \\ \Leftrightarrow & \mathbf{a}_i^\top \mathbf{x} + \min_{\boldsymbol{\mu} \geq 0, \boldsymbol{\nu} \geq 0} \max_{\boldsymbol{\delta}} \{\boldsymbol{\delta}^\top \mathbf{x} + \boldsymbol{\mu}^\top (\boldsymbol{\lambda} - \boldsymbol{\delta}) + \boldsymbol{\nu}^\top (\boldsymbol{\lambda} + \boldsymbol{\delta})\} \leq b_i \\ \Leftrightarrow & \mathbf{a}_i^\top \mathbf{x} + \boldsymbol{\lambda}^\top (\boldsymbol{\mu} + \boldsymbol{\nu}) \leq b_i, \quad \exists \boldsymbol{\mu} \geq 0, \boldsymbol{\nu} \geq 0 \text{ such that } \mathbf{x} + \boldsymbol{\nu} = \boldsymbol{\mu} \\ \Leftrightarrow & \mathbf{a}_i^\top \mathbf{x} + \boldsymbol{\lambda}^\top (\mathbf{x} + 2\boldsymbol{\nu}) \leq b_i, \quad \exists \boldsymbol{\nu} \geq [-\mathbf{x}]_+ \\ \Leftrightarrow & (\mathbf{a}_i + \boldsymbol{\lambda})^\top \mathbf{x} + 2\boldsymbol{\lambda}^\top [-\mathbf{x}]_+ \leq b_i, \end{aligned}$$

where  $[-\mathbf{x}]_+ = ([-x_1]_+, \dots, [-x_d]_+)^\top$ . If parameters  $\mathbf{a}_i$ ,  $b_i$  and  $\boldsymbol{\lambda}$  are unknown, we can take  $\phi(\mathbf{x}) = (\mathbf{x}^\top, [-\mathbf{x}]_+^\top)^\top$ .

Example 1 and Example 2 show that the robust linear programming problems with unknown uncertainty sets can be solved in a data-driven way as long as one can obtain the feedback information about the feasibility of each robust constraint for any given input.

We now consider the first case where decision sets  $\mathcal{S}_t = \{\mathbf{x}_{t,1}, \dots, \mathbf{x}_{t,K}\}$  for some finite number  $K$ . Note that the initialization step (Step 1) in Algorithm 6.2 or Algorithm 6.3 is no longer available since there is no guarantee that  $\mathbf{e}_1, \dots, \mathbf{e}_{d+1}$  belong to  $\mathcal{S}_t$ . Therefore, in the first  $d + 1$  rounds, we draw samples from  $\mathcal{S}_t$  uniformly at random instead of choosing  $\mathbf{e}_1, \dots, \mathbf{e}_{d+1}$ .

To achieve meaningful regret bounds, we assume that there exists a constant  $c_0 > 0$  so that  $\mathbf{M}_{d+1}$  computed according to (6.3) satisfies  $\lambda_{\min}(\mathbf{M}_{d+1}) \geq c_0$ . Conditioned on this  $\mathbf{M}_{d+1}$ , we provide the following theorem which provides the performance guarantee for Algorithm 6.3 in this case.

**Theorem 6.4.** *Suppose that Assumptions 6.2 and 6.3 hold and  $\mathcal{S}_t$  is a finite set, i.e.,  $\mathcal{S}_t = \{\mathbf{x}_{t,1}, \dots, \mathbf{x}_{t,K}\}$ , satisfying that Problem (6.2) with  $\mathcal{S}_t$  is feasible and  $\|\mathbf{x}\|_2 \leq L$  for any  $\mathbf{x} \in \mathcal{S}_t$ . Then there exist constants  $c$  so that for  $0 < \delta < 1$ , when*

$$\theta(t) = \frac{cl_\mu R}{c_\mu} \sqrt{d \log \frac{2m(L^2 + 1)t}{\delta}}, \quad T > d + 1,$$

with probability at least  $1 - 2\delta$  the regret and the constraint violation of Algorithm 6.3 satisfy

$$\begin{aligned} \text{Regret}(T) &= \sum_{t=1}^T \mathbf{c}^\top \mathbf{x}(t) - \mathbf{c}^\top \mathbf{x}^*(t) \leq 2(d+1)L\|\mathbf{c}\|_2, \\ \text{Violation}(T) &\leq (d+1) \sum_{i=1}^m (L\|\mathbf{a}_i\|_2 + |b_i|) + \\ &\quad \frac{20m\theta(T)}{c_\mu} \sqrt{dT \log T}. \end{aligned}$$

Note that the regret defined in Theorem 6.4 differs from the regret considered in the previous sections, because when  $\mathcal{S}_t$  is discrete, the gap between  $\mathbf{x}(t)$  and  $\mathbf{x}^*(t)$  can be quite large even when both of  $\mathbf{a}_i(t)$  and  $b_i(t)$  are close to  $\mathbf{a}_i$  and  $\mathbf{b}_i$  respectively. As discussed in Section 6.3.1, Algorithm 6.2 can also be applied to solve the online linear bandit problem, leading to a  $O(d\sqrt{T} \log T)$  regret bound, implied by Theorem 6.4.

For the second case where decision sets  $\mathcal{S}_t$  are bounded and convex, Theorem 6.5 states that Algorithms 6.2 and 6.3 can obtain the same upper bounds for the regret and the constraint violation as those in Theorems 6.1 and 6.2.

**Theorem 6.5.** *Under Assumptions 6.1-6.4, when the decision sets  $\mathcal{S}_t$  are convex, Theorem 6.1 and Theorem 6.3 still hold.*

We now consider the third case where the linear constraint  $\mathbf{A}^\top \mathbf{x} \leq \mathbf{b}$  in Problem (6.2) is replaced by  $\mathbf{A}^\top \phi(\mathbf{x}) \leq \mathbf{b}$ . For the sake of analysis, we make the following assumptions:

**Assumption 6.5.** *For  $t = 1, \dots, T$ , the feasible set of Problem (6.2) is nonempty, i.e.,  $\mathcal{S}_t \cap \{\mathbf{x} : \mathbf{A}^\top \phi(\mathbf{x}) \leq \mathbf{b}\} \neq \emptyset$ .*

**Assumption 6.6.** *For  $i = 1, \dots, m$ ,  $\mathbf{a}_i^\top \phi(\mathbf{x}) - b_i$  is a convex function in  $\mathbf{x}$  for any  $(\mathbf{a}_i, b_i) \in \mathcal{A}_i$ , and there exists vector  $\bar{\mathbf{x}} \in \mathbb{R}^d$  such that  $\mathbf{a}_i^\top \phi(\bar{\mathbf{x}}) < b_i$ .*

**Assumption 6.7.** *For  $t = 1, \dots, T$ , there exists a constant  $L$  such that  $\|\phi(\mathbf{x})\|_2 \leq L$  for any  $\mathbf{x} \in \mathcal{S}_t$ .*

Assumption 6.6 ensures that Problem (6.2) is a convex optimization problem, and the constraint  $\mathbf{a}_i^\top \phi(\mathbf{x}) - b_i$  has a strictly feasible point. Obviously, Example 1 and Example 2 discussed above satisfy this assumption.

For notational simplicity, we let  $\mathbf{y}(t) = (\phi(\mathbf{x}(t))^\top, -1)^\top$  for  $t = 1, \dots, T$  and  $\mathbf{y} = (\phi(\mathbf{x})^\top, -1)^\top$ . In order to solve this problem, we make a slight modification to Algorithm 6.3, namely, we change the constraints in (6.6) into

$$f(\mathbf{a}_i(t)^\top \phi(\mathbf{x}) - b_i(t)) \leq \theta(t) \|\mathbf{y}\|_{\mathbf{M}_t^{-1}}, \quad \forall i \in [m],$$

where  $\mathbf{M}_t$  can be calculated according to (6.3) using these  $\mathbf{y}(t)$ . Then Theorem 6.6 provides the performance guarantee for the modified version of Algorithm 6.3:

**Theorem 6.6.** *Under Assumptions 6.2-6.3 and 6.5-6.7, there exists constant  $c$  so that for  $0 < \delta < 1$ , when*

$$\theta(t) = \frac{cl_\mu R}{c_\mu} \sqrt{d_\phi \log \frac{2m(L^2 + 1)t}{\delta}},$$

*the upper bounds of the regret and constraint violation for the modified version of Algorithm 6.3 are the same as those in Theorem 6.2 with  $d = d_\phi$ .*

This problem can also be solved by applying Algorithm 6.2 if we can sample “good” points from decision sets  $\mathcal{S}_t$ . Suppose that for  $t = 1, \dots, T$ , there exist set  $\mathcal{T} \subseteq \mathcal{S}_t$  and distribution  $\mathcal{D}$  over  $\mathcal{T}$  such that for samples  $\mathbf{x}_1, \dots, \mathbf{x}_n$  drawn from distribution  $\mathcal{D}$  independently, there exists constant  $c$  so that the following holds with high probability:  $\lambda_{\min}(\frac{1}{n} \sum_{i=1}^n \mathbf{y}_i \mathbf{y}_i^\top) \geq c$  where  $\mathbf{y}_i = (\phi(\mathbf{x}_i)^\top, -1)^\top$ . Then we make the following changes to Algorithm 6.2: 1) the constraints in (6.5) are replaced with

$$\mathbf{a}_i(t)^\top \phi(\mathbf{x}) \leq b_i(t), \quad \forall i \in [m],$$

and 2)  $\tilde{\mathbf{x}}(t)$  are sampled according to  $\mathcal{D}$  instead of the uniform distribution on  $\mathcal{S}_{d-1}(B)$ .

**Theorem 6.7.** *Under Assumptions 6.2, 6.3 and 6.5-6.7, the upper bounds of the regret and constraint violation for the modified version of Algorithm 6.2 are the same as those in Theorem 6.1 with  $d = d_\phi$ .*

## 6.7 Experiments

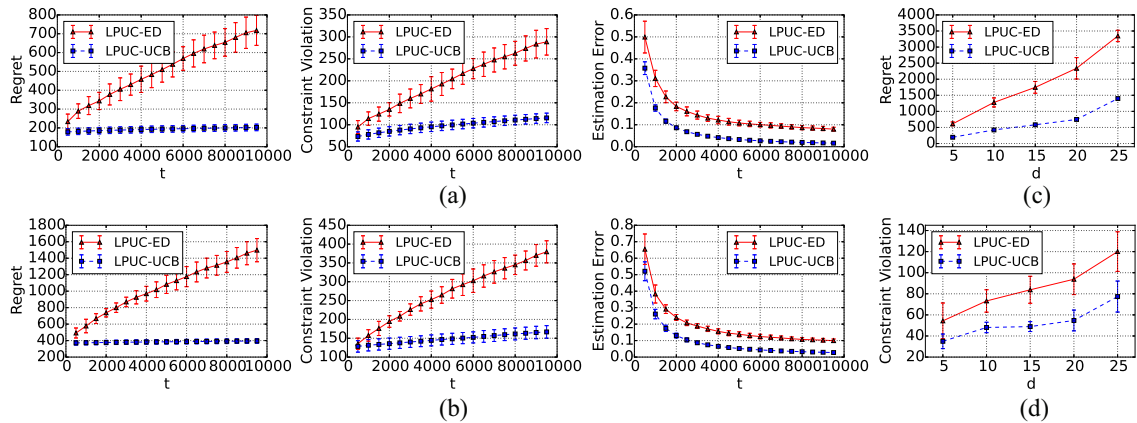
In this section, we investigate the empirical performance of our algorithms on synthetic data. The linear programming problems are randomly generated as follows: 1) cost vector  $\mathbf{c}$  is sampled from  $[-1, 1]^d$  uniformly at random, 2)  $\mathbf{b}$  is uniformly drawn from  $[0, 2]^m$ , 3) each column of  $\mathbf{A}$  is sampled from  $\mathcal{S}_{d-1}(1)$  uniformly at random, and 4)  $\boldsymbol{\xi}(t)$  is set to  $0.01\boldsymbol{\mu}$  where  $\boldsymbol{\mu}$  follows the standard Gaussian distribution  $\mathcal{N}(0, \mathbf{I})$ . We let  $\mathcal{A}_i$  – the admissible set for  $\mathbf{a}_i$  and  $b_i$  – be  $[-5, 5]^{d+1}$  and  $\mathcal{S}_t$  – the decision set in round  $t$  – be  $[-5, 5]^d$ . We repeat each test 10 times and report the average results.

In the experiments, Problem (6.4) is solved via the L-BFGS-B algorithm [BLN95]. For LPUC-ED,  $p(t)$  and  $B$  are set to  $0.1/t^3$  and 5, respectively. For LPUC-UCB, in the  $t^{\text{th}}$  round, the non-convex optimization problem (6.6) is solved by two steps: 1) we compute  $\tilde{\mathbf{x}}_t$  – the optimal solution of (6.6) with  $\theta(t) = 0$  (if  $\tilde{\mathbf{x}}_t$  does not exist,  $\tilde{\mathbf{x}}_t$  is set to  $\mathbf{x}_{t-1}$ ), and 2) by taking  $\tilde{\mathbf{x}}_t$  as the initial solution, we use the SciPy optimization package [JOP<sup>+</sup>] to solve (6.6) with  $\theta(t) = 0.01\sqrt{\log t}$ . In the initialization step, one can also uniformly draw 20 samples from  $\mathcal{S}_{d-1}(B)$  and then play these samples in the first 20 rounds, which leads to

better performance.

In the first experiment, the linear programming problem is generated with  $d = 10$  and  $m = 20$ . We compare the acceleration versions of LPUC-ED and LPUC-UCB with parameter  $\gamma = 0.01$ . The empirical performance is measured by three quantities: the regret, the constraint violation and the estimation error. The estimation error is the difference between the true optimal solution  $\mathbf{x}^*$  of (6.2) and the average of the solutions up to time  $T$ , namely,  $\|\bar{\mathbf{x}} - \frac{1}{T} \sum_{i=1}^T \mathbf{x}(T)\|_2$ .

For input  $\mathbf{x}(t)$ , we consider two different feedbacks: 1) linear feedback  $r_i(t) = \mathbf{a}_i^\top \mathbf{x}(t) - b_i + \xi_i(t)$ , and 2) sign feedback  $r_i(t) = -1$  if  $\mathbf{a}_i^\top \mathbf{x}(t) - b_i \leq 0$  or 1 otherwise. The sign feedback only indicates which constraint is violated, e.g., observing congestion or delays in the network. In the algorithms, we use  $f(x) = (\exp(x) - 1)/(\exp(x) + 1)$  to approximate the sign function. Figures 6.1(a) and 6.1(b) show the empirical performance of LPUC-ED and LPUC-UCB. Obviously, their regrets and constraint violations are sub-linear in  $T$ , and LPUC-UCB has a significantly better performance than LPUC-ED. This is consistent with our theoretical analysis, i.e., the regret and the constraint violation of LPUC-ED are  $O(T^{\frac{2}{3}})$  due to the sampling procedure for exploration, while those of LPUC-UCB are  $O(\sqrt{T})$  since it implicitly does the exploration by solving (6.6). From the estimation errors we observe that  $\bar{\mathbf{x}} = \frac{1}{T} \sum_{i=1}^T \mathbf{x}(T)$  converges to  $\mathbf{x}^*$  as  $T$  goes to infinity.

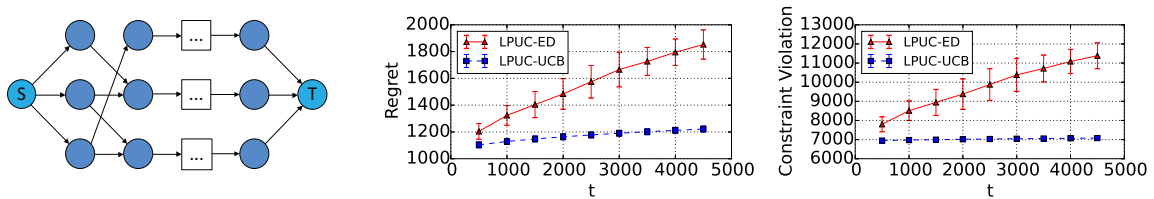


**Figure 6.1:** We compare the empirical performance of LPUC-ED and LPUC-UCB. (a) Linear feedback  $\mathbf{r}(t) = \mathbf{A}^\top \mathbf{x}(t) - \mathbf{b} + \boldsymbol{\xi}(t)$  for any  $\mathbf{x}(t)$ . (b) Sign feedback  $\mathbf{r}(t) = \text{sign}(\mathbf{A}^\top \mathbf{x}(t) - \mathbf{b})$ . (c)(d) The regret and the constraint violation against dimension  $d$  with linear feedbacks.

In the second experiment, we investigate the performance of LPUC-ED and LPUC-UCB for

different  $d$  and  $m$ . In particular,  $d = 5, 10, 15, 20, 25$  and  $m = d$ . The linear programming problems are generated with these  $d$  and  $m$ . For any input  $\mathbf{x}(t)$ , the feedback  $r_i(t)$  is  $\mathbf{a}_i^\top \mathbf{x}(t) - b_i + \xi_i(t)$ . Figures 6.1(c) and 6.1(d) show the regrets and constraint violations for LPUC-ED and LPUC-UCB when  $d$  varies from 5 to 25. We observe that both of the regret and constraint violation grow nearly linearly in  $d$ . Similar to the first experiment, LPUC-UCB clearly outperforms LPUC-ED.

The third experiment solves a maximum flow problem with unknown edge capacities. The structure of the network is shown in Figure 6.2, which contains two terminal nodes and 10 layers of internal nodes where each layer has 3 nodes. Each terminal node is fully connected to the nodes in the closest layer with edge capacity 10. The edges between the nodes in one layer and those in the next layer are formed with probability 0.2, whose capacities are uniformly drawn from integers 1 – 10. We assume that the capacities are unavailable to the decision maker. For any input flow  $f$ , suppose that  $f_i$  runs through edge  $i$  whose capacity is  $u_i$ , the decision maker receives a piecewise-linear feedback  $r_i = \max\{f_i - u_i, 0\}$ , namely, the delay is reported if the edge is jammed. In LPUC-ED and LPUC-UCB, we set  $f(x) = x \exp(5x)/(1 + \exp(5x))$  to approximate this piecewise-linear feedback. Figure 6.2 shows their empirical performance. Clearly, their regrets and the constraint violations are sublinear in  $T$ .



**Figure 6.2:** A maximum flow problem with unknown edge capacities. The graph contains two terminal nodes and 10 layers of internal nodes.

## 6.8 Proofs of Technical Results

Before the main proofs are shown, we first provide several useful lemmas.

**Lemma 6.1.** *Suppose  $\mathbf{z}_1, \dots, \mathbf{z}_n$  are independently drawn from the uniform distribution over the sphere  $\mathcal{S}_{d-1}(B)$ . Let  $\mathbf{y}_i \triangleq (\mathbf{z}_i^\top, -1)^\top$  and  $\mathbf{Y} \triangleq \sum_{i=1}^n \mathbf{y}_i \mathbf{y}_i^\top$ . Then the following*

inequality holds with probability at least  $1 - 2d \exp(-\frac{cn}{d^2})$ :

$$\min\left\{\frac{nB^2}{4d}, \frac{n}{2}\right\} \leq \lambda_{\min}(\mathbf{Y}), \lambda_{\max}(\mathbf{Y}) \leq \max\left\{\frac{3nB^2}{4d}, \frac{3n}{2}\right\},$$

where  $c$  is a universal constant.

*Proof.* Since  $\mathbf{z}_i$  is drawn from the uniform distribution over  $\mathcal{S}_{d-1}(B)$ , we have  $\mathbb{E}[\mathbf{z}_i] = 0$  and  $\mathbb{E}[\mathbf{z}_i \mathbf{z}_i^\top] = \frac{B^2}{d} \mathbf{I}$ . Define  $\mathbf{Z}_i \triangleq \mathbf{z}_i \mathbf{z}_i^\top$  and  $\bar{\mathbf{Z}} \triangleq \mathbb{E}[\mathbf{z}_i \mathbf{z}_i^\top]$ , then we have

$$\|\mathbf{Z}_i - \bar{\mathbf{Z}}\|_2 \leq B^2 + \frac{B^2}{d} \leq 2B^2$$

and

$$\begin{aligned} \|\mathbb{E}[(\sum_{i=1}^n \mathbf{Z}_i - n\bar{\mathbf{Z}})^2]\|_2 &= \|\mathbb{E}[(\sum_{i=1}^n \mathbf{Z}_i)^2] - n^2 \bar{\mathbf{Z}}^2\|_2 \\ &= n \|\mathbb{E}[\mathbf{Z}_i^2 + \mathbf{Z}_i \sum_{j \neq i} \mathbf{Z}_j] - n \bar{\mathbf{Z}}^2\|_2 \\ &= n \|\mathbb{E}[\|\mathbf{z}_i\|_2 \cdot \mathbf{Z}_i] - \bar{\mathbf{Z}}^2\|_2 \\ &= n \|B^2 \bar{\mathbf{Z}} - \bar{\mathbf{Z}}^2\|_2 \\ &\leq \frac{nB^4}{d} \left(1 - \frac{1}{d}\right) \leq \frac{nB^4}{d}. \end{aligned}$$

By the matrix Bernstein inequality, the following inequality holds for all  $\beta \geq 0$ :

$$\mathbb{P}\left[\left\|\sum_{i=1}^n \mathbf{Z}_i - \frac{nB^2}{d} \mathbf{I}\right\|_2 \geq \beta\right] \leq 2d \exp\left(-\frac{\beta^2/2}{nB^4/d + 2B^2\beta/3}\right).$$

By setting  $\beta = \frac{nB^2}{2d}$ , we have

$$\frac{nB^2}{2d} \leq \lambda_{\min}\left(\sum_{i=1}^n \mathbf{Z}_i\right), \lambda_{\max}\left(\sum_{i=1}^n \mathbf{Z}_i\right) \leq \frac{3nB^2}{2d} \quad (6.7)$$

holds with probability at least  $1 - 2d \exp(-\frac{n}{16d})$ . Similarly, note that

$$\|\mathbf{z}_i\|_2 = B, \quad \|\mathbb{E}[(\sum_{i=1}^n \mathbf{z}_i)^\top (\sum_{i=1}^n \mathbf{z}_i)]\|_2 = nB^2.$$

By the matrix Bernstein inequality, for all  $\beta \geq 0$ ,

$$\mathbb{P}[\|\sum_{i=1}^n \mathbf{z}_i\|_2 \geq \beta] \leq (d+1) \exp\left(-\frac{\beta^2/2}{nB^2 + B\beta/3}\right).$$

Therefore, when  $\beta = \min\{\frac{nB^2}{4d}, \frac{1}{2}n\}$ , we have

$$\|\sum_{i=1}^n \mathbf{z}_i\|_2 \leq \min\left\{\frac{nB^2}{4d}, \frac{1}{2}n\right\} \quad (6.8)$$

holds with probability at least

$$1 - (d+1) \max\left\{\exp\left(-\frac{n \min\{B^2, 1\}}{64d^2}\right), \exp\left(-\frac{n}{8B(B+1)}\right)\right\}.$$

Recall that

$$\mathbf{Y} = \sum_{i=1}^n \mathbf{y}_i \mathbf{y}_i^\top = \begin{pmatrix} \sum_{i=1}^n \mathbf{z}_i & -\sum_{i=1}^n \mathbf{z}_i \\ -\sum_{i=1}^n \mathbf{z}_i^\top & n \end{pmatrix}.$$

We define

$$\mathbf{Q}_1 \triangleq \begin{pmatrix} \sum_{i=1}^n \mathbf{z}_i & 0 \\ 0 & n \end{pmatrix}, \quad \mathbf{Q}_2 \triangleq \begin{pmatrix} 0 & -\sum_{i=1}^n \mathbf{z}_i \\ -\sum_{i=1}^n \mathbf{z}_i^\top & 0 \end{pmatrix},$$

then from the Weyl's inequality,

$$\lambda_{\min}(\mathbf{Q}_1) + \lambda_{\min}(\mathbf{Q}_2) \leq \lambda_{\min}(\mathbf{Y}) \leq \lambda_{\min}(\mathbf{Q}_1) + \lambda_{\max}(\mathbf{Q}_2).$$

From Inequalities (6.7) and (6.8), there exists constant  $c$  so that  $\min\{\frac{nB^2}{4d}, \frac{n}{2}\} \leq \lambda_{\min}(\mathbf{Y}) \leq \max\{\frac{3nB^2}{4d}, \frac{3n}{2}\}$  holds with probability at least  $1 - 2d \exp(-\frac{cn}{d^2})$ .

Similarly, we can prove that  $\min\{\frac{nB^2}{4d}, \frac{n}{2}\} \leq \lambda_{\max}(\mathbf{Y}) \leq \max\{\frac{3nB^2}{4d}, \frac{3n}{2}\}$ .  $\square$

**Lemma 6.2.** *Suppose  $\mathbf{z}_1, \dots, \mathbf{z}_n$  are independently drawn from the uniform distribution over the sphere  $\mathcal{S}_{d-1}(B)$  and  $\eta_1, \dots, \eta_n$  are independently drawn from Bernoulli distribution where the success probability of  $\eta_i$  is  $p_i = 1/i^{1/3}$  for  $i = 1, \dots, n$ . Define  $\mathbf{y}_i \triangleq (\mathbf{z}_i^\top, -1)^\top$  and  $\mathbf{Y} \triangleq \sum_{i=1}^n \eta_i \mathbf{y}_i \mathbf{y}_i^\top$ , then we have*

$$\lambda_{\min}(\mathbf{Y}) \geq c_1 \min\left\{\frac{B^2}{d}, 1\right\} n^{2/3}$$



holds with probability at least

$$1 - 2 \max\left\{d \exp\left(-\frac{c_2 n^{2/3}}{d^2}\right), \exp(-c_3 n^{1/3})\right\},$$

where  $c_1, c_2, c_3$  are universal constants.

*Proof.* Note that  $\mathbb{E}[\sum_{i=1}^n \eta_i] = \sum_{i=1}^n \frac{1}{i^{1/3}}$ , so there exist constants  $c_1$  and  $c_2$  such that  $c_1 n^{2/3} \leq \mathbb{E}[\sum_{i=1}^n \eta_i] \leq c_2 n^{2/3}$ . By Hoeffding's inequality,

$$\mathbb{P}\left[\left|\sum_{i=1}^n \eta_i - \mathbb{E}\left[\sum_{i=1}^n \eta_i\right]\right| \geq \beta\right] \leq 2 \exp\left(-\frac{2\beta^2}{n}\right).$$

By taking  $\beta = \frac{1}{2}c_1 n^{2/3}$ , we have

$$\frac{1}{2}c_1 n^{2/3} \leq \sum_{i=1}^n \eta_i \leq \left(\frac{1}{2}c_1 + c_2\right)n^{2/3}$$

holds with probability at least  $1 - 2 \exp(-c_3 n^{1/3})$ . Then this lemma can be obtained by following the proof of Lemma 6.1.  $\square$

**Lemma 6.3.** [Rob77] *The linear system*

$$\mathbf{A}^\top \mathbf{x} \leq \mathbf{b}, \mathbf{x} \in \mathcal{S}$$

is regular, i.e.,  $\mathbf{b}$  is an interior point of  $\{\mathbf{A}^\top \mathbf{x} + \mathbf{z} : \mathbf{x} \in \mathcal{S}, \mathbf{z} \in \mathbb{R}_+^m\}$ , if and only if there exists some constant  $\rho > 0$  such that for any  $\hat{\mathbf{A}}, \hat{\mathbf{b}}$  with  $\max\{\|\mathbf{A} - \hat{\mathbf{A}}\|_2, \|\mathbf{b} - \hat{\mathbf{b}}\|_2\} < \rho$ , the system

$$\hat{\mathbf{A}}^\top \mathbf{x} \leq \hat{\mathbf{b}}, \mathbf{x} \in \mathcal{S}$$

is solvable.

**Lemma 6.4.** [Ren94] *If the linear programming problem (6.1) and its dual problem are both feasible, then the following statement is true: there exists some constant  $\rho$  so that for any*

$\hat{\mathbf{A}}, \hat{\mathbf{b}}, \hat{\mathbf{c}}$  with  $\Delta \triangleq \max\{\|\mathbf{A} - \hat{\mathbf{A}}\|_2, \|\mathbf{b} - \hat{\mathbf{b}}\|_2, \|\mathbf{c} - \hat{\mathbf{c}}\|_2\} < \rho$ , we have

$$\begin{aligned} & |opt(\hat{\mathbf{A}}, \hat{\mathbf{b}}, \hat{\mathbf{c}}) - opt(\mathbf{A}, \mathbf{b}, \mathbf{c})| \\ & \leq \|\hat{\mathbf{A}} - \mathbf{A}\|_2 \cdot \frac{\|\mathbf{b}\|_2 + \|\hat{\mathbf{b}} - \mathbf{b}\|_2}{\rho - \Delta} \cdot \frac{\|\mathbf{c}\|_2 + \|\hat{\mathbf{c}} - \mathbf{c}\|_2}{\rho - \Delta} \cdot \frac{\max\{\|\mathbf{A}\|_2, \|\mathbf{b}\|_2, \|\mathbf{c}\|_2\}}{\rho} \\ & + \|\hat{\mathbf{b}} - \mathbf{b}\|_2 \cdot \frac{\|\mathbf{c}\|_2 + \|\hat{\mathbf{c}} - \mathbf{c}\|_2}{\rho - \Delta} \cdot \frac{\max\{\|\mathbf{A}\|_2, \|\mathbf{b}\|_2, \|\mathbf{c}\|_2\}}{\rho} \\ & + \|\hat{\mathbf{c}} - \mathbf{c}\|_2 \cdot \frac{\|\mathbf{b}\|_2 + \|\hat{\mathbf{b}} - \mathbf{b}\|_2}{\rho - \Delta} \cdot \frac{\max\{\|\mathbf{A}\|_2, \|\mathbf{b}\|_2, \|\mathbf{c}\|_2\}}{\rho}, \end{aligned}$$

where  $opt(\cdot)$  denotes the optimal value of a certain linear program.

**Lemma 6.5.** [Ren94] *If the linear programming problem (6.2) and its dual problem are both feasible, then there exists some constant  $\rho$  depending on  $\mathbf{A}, \mathbf{b}, \mathbf{c}$  so that for any feasible solution  $\hat{\mathbf{x}}$  of linear system  $\mathbf{A}^\top \mathbf{x} \leq \hat{\mathbf{b}}, \mathbf{x} \in \mathcal{S}_t$ , there exists a feasible solution  $\tilde{\mathbf{x}}$  of (6.2) satisfying*

$$\|\hat{\mathbf{x}} - \tilde{\mathbf{x}}\|_2 \leq \rho \|\hat{\mathbf{b}} - \mathbf{b}\|_2.$$

**Lemma 6.6.** [AYPS11] *Let  $\{\mathcal{F}_t\}_{t=1}^\infty$  be a filtration. Let  $\{\eta_t\}_{t=1}^\infty$  be a real-valued stochastic process such that  $\eta_t$  is  $\mathcal{F}_t$ -measurable and  $\eta_t$  is conditionally  $R$ -sub-Gaussian for some  $R \geq 0$ . Let  $\{\mathbf{x}_t\}_{t=1}^\infty$  be a  $\mathbb{R}^d$  valued stochastic process such that  $\mathbf{x}_t$  is  $\mathcal{F}_{t-1}$ -measurable. Assume that  $\mathbf{V}$  is a  $d \times d$  positive definite matrix. For any  $t \geq 0$ , define*

$$\bar{\mathbf{V}}_t = \mathbf{V} + \sum_{k=1}^t \mathbf{x}_k \mathbf{x}_k^\top, \quad \mathbf{y}_t = \sum_{k=1}^t \eta_k \mathbf{x}_k,$$

then for any  $\delta > 0$ , with probability at least  $1 - \delta$ , for all  $t \geq 0$ ,

$$\|\mathbf{y}_t\|_{\bar{\mathbf{V}}_t}^2 \leq 2R^2 \log \left( \frac{\det(\bar{\mathbf{V}}_t)^{1/2} \det(\mathbf{V})^{1/2}}{\delta} \right).$$

**Lemma 6.7.** *Suppose that  $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$  and  $\|\mathbf{x}_k\|_2 \leq L$  for  $k = 1, \dots, n$ . Let  $\mathbf{M}_t = \mathbf{M} + \sum_{k=1}^t \mathbf{x}_k \mathbf{x}_k^\top$  for some positive definite matrix  $\mathbf{M}$ , then*

$$\det(\mathbf{M}_t) \leq \left( \frac{\text{tr}(\mathbf{M}) + tL^2}{d} \right)^d.$$

*Proof.* Let  $\lambda_1, \dots, \lambda_d$  be the eigenvalues of  $\mathbf{M}_t$ . Note that  $\det(\mathbf{M}_t) = \prod_{k=1}^d \lambda_k$  and  $\text{tr}(\mathbf{M}_t) =$

$\sum_{k=1}^t \lambda_k$ , by the inequality of arithmetic and geometric means, we have

$$\det(\mathbf{M}_t) \leq \left( \frac{\text{tr}(\mathbf{M}_t)}{d} \right)^d = \left( \frac{\text{tr}(\mathbf{M}) + \sum_{k=1}^t \|\mathbf{x}_k\|_2^2}{d} \right)^d \leq \left( \frac{\text{tr}(\mathbf{M}) + tL^2}{d} \right)^d.$$

Hence the lemma holds.  $\square$

### 6.8.1 Proof of Proposition 6.1

*Proof.* We first show that  $|\mathbf{c}^\top \mathbf{x}(T) - \mathbf{c}^\top \mathbf{x}^*(T)| = O(\frac{1}{\sqrt{T}})$ . Since  $\mathbf{r}(t) = \mathbf{A}^\top \mathbf{x}(t) - \mathbf{b} + \boldsymbol{\xi}(t)$  and  $\mathbf{x}(t)$  are drawn from  $\mathcal{S}_{d-1}(B)$  uniformly at random for  $t = 1, \dots, T-1$ , we have

$$\mathbb{E}[\hat{\mathbf{b}}] = \mathbf{b}, \text{ and } \mathbb{E}[\hat{\boldsymbol{\Sigma}}] = \frac{B^2}{d} \mathbf{I}.$$

Note that  $|r_i(t) + b_i| \leq |\mathbf{a}_i^\top \mathbf{x}(t)| + |\xi_i(t)| \leq B\|\mathbf{a}_i\|_2 + R$ . Then by the Hoeffding's inequality, for any  $i = 1, \dots, m$ , the following inequality holds:

$$\mathbb{P}[|\hat{b}_i - b_i| \geq \beta] \leq 2 \exp\left(-\frac{(T-1)\beta^2}{(\max_i B\|\mathbf{a}_i\|_2 + R)^2}\right).$$

For any constant  $\delta > 0$ , let  $D = (\max_i B\|\mathbf{a}_i\|_2 + R)$  and  $\beta = D\sqrt{\frac{1}{T-1} \log \frac{m}{\delta}}$ , then

$$|\hat{b}_i - b_i| \leq D\sqrt{\frac{1}{T-1} \log \frac{m}{\delta}}$$

holds with probability at least  $1 - \frac{2\delta}{m}$ . By the union bound,

$$\|\hat{\mathbf{b}} - \mathbf{b}\|_2 \leq D\sqrt{\frac{m}{T-1} \log \frac{m}{\delta}}.$$

holds with probability at least  $1 - 2\delta$ .

On the other hand, from the definition of  $\hat{\mathbf{A}}$ , we know that

$$\hat{\mathbf{a}}_i - \mathbf{a}_i = \hat{\boldsymbol{\Sigma}}^{-1} \left[ \frac{1}{T-1} \sum_{t=1}^{T-1} (b_i + \xi_i(t)) \mathbf{x}(t) \right],$$

which implies

$$\|\hat{\mathbf{a}}_i - \mathbf{a}_i\|_2 \leq \lambda_{\min}(\hat{\boldsymbol{\Sigma}})^{-1} \left\| \frac{1}{T-1} \sum_{t=1}^{T-1} (b_i + \xi_i(t)) \mathbf{x}(t) \right\|_2.$$

From Inequality (6.7), we know that  $\frac{B^2}{2d} \leq \lambda_{\min}(\hat{\boldsymbol{\Sigma}}) \leq \frac{3B^2}{2d}$  holds with probability at least  $1 - 2d \exp(-\frac{T-1}{16d})$ . Define  $\mathbf{y}_i(t) = (b_i + \xi_i(t)) \mathbf{x}(t)$ , then by the matrix Bernstein inequality,

$$\begin{aligned} \mathbb{P}\left[\left\| \frac{1}{T-1} \sum_{t=1}^{T-1} \mathbf{y}_i(t) \right\|_2 \geq \beta\right] &\leq (d+1) \exp\left(-\frac{(T-1)\beta^2/2}{(b_i^2 + R^2)B^2/d + (|b_i| + R)\beta/3}\right) \\ &\leq (d+1) \exp\left(-\frac{(T-1)\beta^2/2}{2(|b_i| + R)^2 B^2/d + (|b_i| + R)\beta/3}\right). \end{aligned}$$

Let  $G = \max_i |b_i| + R$  and

$$\beta = 3BG \sqrt{\frac{1}{d(T-1)} \log \frac{(d+1)m}{\delta}},$$

then when  $T-1 \geq \frac{1}{4B^2} d \log \frac{(d+1)m}{\delta}$ , we have

$$\left\| \frac{1}{T-1} \sum_{t=1}^{T-1} (b_i + \xi_i(t)) \mathbf{x}(t) \right\|_2 \leq \beta$$

holds with probability at least  $1 - \frac{\delta}{m}$ . By the union bound, with probability at least  $1 - \delta$  the following inequality holds for all  $i = 1, \dots, m$ :

$$\|\hat{\mathbf{a}}_i - \mathbf{a}_i\|_2 \leq \frac{6G}{B} \sqrt{\frac{d}{T-1} \log \frac{(d+1)m}{\delta}},$$

which implies

$$\|\hat{\mathbf{A}} - \mathbf{A}\|_2 \leq \frac{6G}{B} \sqrt{\frac{md}{T-1} \log \frac{(d+1)m}{\delta}}.$$

Thus, by applying Lemma 6.3 and Lemma 6.4, when  $T$  is large enough, i.e., for constant  $\rho$ ,

$$T-1 \geq \frac{4 \max\{D, 6G/B\}}{\rho^2} \cdot md \log \frac{(d+1)m}{\delta},$$

we have

$$|\mathbf{c}^\top \mathbf{x}(T) - \mathbf{c}^\top \mathbf{x}^*(T)| = O\left(\sqrt{\frac{md}{T} \log \frac{md}{\delta}}\right) \approx O\left(\frac{1}{\sqrt{T}}\right).$$

We now prove that there exists an instance of Problem (6.2) so that the regret and the constraint violation of Algorithm 6.1 are  $\Omega(T)$ . Consider the following linear programming problem:

$$\begin{aligned} \min \quad & -x \\ \text{s.t.} \quad & 2x \leq 1, \quad x \in [-1, 1], \end{aligned}$$

which means that  $\mathbf{c} = -1$ ,  $\mathbf{A} = 2$ ,  $\mathbf{b} = 1$  and  $\mathcal{S}_t = [-1, 1]$ . Obviously, for  $t = 1, \dots, T$ , the optimal solutions  $\mathbf{x}^*(t)$  are always 0.5. When  $B = 1$ ,  $\mathbf{x}(t)$  is draw from  $\{-1, 1\}$  with equal probability, implying that  $|\mathbf{c}^\top \mathbf{x}(t) - \mathbf{c}^\top \mathbf{x}^*(t)|$  equals 0.5 or 1.5 with equal probability and  $[\mathbf{A}^\top \mathbf{x}(t) - \mathbf{b}]_+$  equals 0 or 1 with equal probability. Thus, it can be easily verified that

$$\begin{aligned} \text{Regret}(T) &\geq \sum_{t=1}^{T-1} |\mathbf{c}^\top \mathbf{x}(t) - \mathbf{c}^\top \mathbf{x}^*(t)| = \Omega(T), \\ \text{Violation}(T) &\geq \sum_{t=1}^{T-1} [\mathbf{A}^\top \mathbf{x}(t) - \mathbf{b}]_+ = \Omega(T). \end{aligned}$$

Hence we obtain this proposition. □

### 6.8.2 Proof of Theorem 6.1

*Proof.* In the following proofs, the constants may change from line to line. For convenience, let  $\bar{\mathbf{a}}_i = (\mathbf{a}_i^\top, b_i)^\top$ ,  $\bar{\mathbf{a}}_i(t) = (\mathbf{a}_i(t)^\top, b_i(t))^\top$  and let  $\bar{\mathbf{x}} = (\mathbf{x}^\top, -1)^\top$  for any  $\mathbf{x} \in \mathbb{R}^d$ .

When  $t \geq d + 2$ , from the definition of  $\mathbf{M}_t$ , we have

$$\mathbf{M}_t = \sum_{k=1}^{d+1} \mathbf{y}(k)\mathbf{y}(k)^\top + \sum_{k=d+2}^{t-1} \mathbf{y}(k)\mathbf{y}(k)^\top.$$

Let  $\mathbf{M} \triangleq \sum_{k=1}^{d+1} \mathbf{y}(k)\mathbf{y}(k)^\top$ . Since  $\mathbf{y}(k) = (B\mathbf{e}_k^\top, -1)^\top$  for  $k \leq d + 1$ , one can easily verify that  $\det(\mathbf{M}) = B^{2d}$ .

Since  $f(\cdot)$  is Lipschitz continuous (Assumption 6.2), we have

$$|f(\bar{\mathbf{a}}_i(t)^\top \bar{\mathbf{x}}) - f(\bar{\mathbf{a}}_i^\top \bar{\mathbf{x}})| \leq l_\mu |(\bar{\mathbf{a}}_i(t) - \bar{\mathbf{a}}_i)^\top \bar{\mathbf{x}}|.$$

Since  $f(\cdot)$  is continuously differentiable (Assumption 6.2),  $\nabla g_t(\cdot)$  is continuous. Therefore,

$$g_t(\bar{\mathbf{a}}_i(t)) - g_t(\bar{\mathbf{a}}_i) = \mathbf{G}_t(\bar{\mathbf{a}}_i(t) - \bar{\mathbf{a}}_i)$$

where  $\mathbf{G}_t = \int_0^1 \nabla g_t(s\bar{\mathbf{a}}_i(t) + (1-s)\bar{\mathbf{a}}_i) ds$ . Note that

$$\nabla g_t(\mathbf{z}) = \sum_{k=1}^{t-1} \nabla f(\mathbf{z}^\top \mathbf{y}(k)) \cdot \mathbf{y}(k) \mathbf{y}(k)^\top$$

and  $\nabla f(\mathbf{z}^\top \mathbf{y}(k)) \geq c_\mu$  (Assumption 6.2), we have

$$\mathbf{G}_t \succeq c_\mu \mathbf{M}_t \succ 0,$$

which implies

$$\begin{aligned} |f(\bar{\mathbf{a}}_i(t)^\top \bar{\mathbf{x}}) - f(\bar{\mathbf{a}}_i^\top \bar{\mathbf{x}})| &\leq l_\mu |(\bar{\mathbf{a}}_i(t) - \bar{\mathbf{a}}_i)^\top \bar{\mathbf{x}}| \\ &= l_\mu |\bar{\mathbf{x}}^\top \mathbf{G}_t^{-1} (g_t(\bar{\mathbf{a}}_i(t)) - g_t(\bar{\mathbf{a}}_i))| \\ &\leq \frac{l_\mu}{c_\mu} \|\bar{\mathbf{x}}\|_{\mathbf{M}_t^{-1}} \|g_t(\bar{\mathbf{a}}_i(t)) - g_t(\bar{\mathbf{a}}_i)\|_{\mathbf{M}_t^{-1}}. \end{aligned}$$

From Step 3 of Algorithm 6.2 which estimates  $\mathbf{A}$  and  $\mathbf{b}$  by solving Problem (6.4), we know that

$$\|g_t(\bar{\mathbf{a}}_i(t)) - \mathbf{g}_t^i\|_{\mathbf{M}_t^{-1}} \leq \|g_t(\bar{\mathbf{a}}_i) - \mathbf{g}_t^i\|_{\mathbf{M}_t^{-1}}.$$

Therefore,

$$\begin{aligned} |f(\bar{\mathbf{a}}_i(t)^\top \bar{\mathbf{x}}) - f(\bar{\mathbf{a}}_i^\top \bar{\mathbf{x}})| &\leq \frac{2l_\mu}{c_\mu} \|\bar{\mathbf{x}}\|_{\mathbf{M}_t^{-1}} \|\mathbf{g}_t^i - g_t(\bar{\mathbf{a}}_i)\|_{\mathbf{M}_t^{-1}} \\ &\leq \frac{2l_\mu}{c_\mu} \|\bar{\mathbf{x}}\|_{\mathbf{M}_t^{-1}} \left\| \sum_{k=1}^{t-1} \xi_i(k) \mathbf{y}(k) \right\|_{\mathbf{M}_t^{-1}}. \end{aligned} \tag{6.9}$$

We now bound  $\left\| \sum_{k=1}^{t-1} \xi_i(k) \mathbf{y}(k) \right\|_{\mathbf{M}_t^{-1}}$ . Note that

$$\begin{aligned} \left\| \sum_{k=1}^{t-1} \xi_i(k) \mathbf{y}(k) \right\|_{\mathbf{M}_t^{-1}} &\leq \left\| \sum_{k=1}^{d+1} \xi_i(k) \mathbf{y}(k) \right\|_{\mathbf{M}_t^{-1}} + \left\| \sum_{k=d+2}^{t-1} \xi_i(k) \mathbf{y}(k) \right\|_{\mathbf{M}_t^{-1}} \\ &\leq \left\| \sum_{k=1}^{d+1} \xi_i(k) \mathbf{y}(k) \right\|_{\mathbf{M}^{-1}} + \left\| \sum_{k=d+2}^{t-1} \xi_i(k) \mathbf{y}(k) \right\|_{\mathbf{M}_t^{-1}} \end{aligned}$$

where the last inequality holds because  $\mathbf{M}_t \succeq \mathbf{M} \succ 0$ . Recall that for  $k \leq d+1$ ,  $\mathbf{y}(k) = (B\mathbf{e}_k^\top, -1)^\top$  (notice that we define  $\mathbf{e}_{d+1} \triangleq \mathbf{0}$ ), we have

$$\sum_{k=1}^{d+1} \xi_i(k) \mathbf{y}(k) = (B\xi_i(1), B\xi_i(2), \dots, B\xi_i(d), -\sum_{k=1}^{d+1} \xi_i(k))^\top,$$

and

$$\mathbf{M} = \sum_{k=1}^{d+1} \mathbf{y}(k) \mathbf{y}(k)^\top = \begin{pmatrix} B^2 & 0 & \cdots & -B \\ 0 & B^2 & \cdots & -B \\ \vdots & \vdots & \ddots & \vdots \\ -B & -B & \cdots & d+1 \end{pmatrix}.$$

By simple calculation, one can obtain

$$\mathbf{M}^{-1} = \frac{1}{B^2} \begin{pmatrix} 2 & 1 & \cdots & B \\ 1 & 2 & \cdots & B \\ \vdots & \vdots & \ddots & \vdots \\ B & B & \cdots & 1 \end{pmatrix}.$$

Then we have

$$\begin{aligned} \left\| \sum_{k=1}^{d+1} \xi_i(k) \mathbf{y}(k) \right\|_{\mathbf{M}^{-1}}^2 &= \sum_{k=1}^d \xi_i(k)^2 + \left( \sum_{k=1}^d \xi_i(k) \right)^2 + \\ &2 \left( \sum_{k=1}^d \xi_i(k) \right) \left( \sum_{k=1}^{d+1} \xi_i(k) \right) + \frac{1}{B^2} \left( \sum_{k=1}^{d+1} \xi_i(k) \right)^2 \end{aligned}$$

Since  $\xi_i(1), \dots, \xi_i(d+1)$  are independently drawn from  $[-R, R]$  with mean 0, by the Hoeffding's inequality, for  $\delta > 0$

$$\left| \sum_{k=1}^d \xi_i(k) \right|, \left| \sum_{k=1}^{d+1} \xi_i(k) \right| \leq cR \sqrt{d \log \frac{m}{\delta}}$$

holds with probability at least  $1 - \frac{\delta}{m}$ , where  $c$  is a universal constant. Therefore,

$$\begin{aligned} \left\| \sum_{k=1}^{d+1} \xi_i(k) \mathbf{y}(k) \right\|_{\mathbf{M}^{-1}} &\leq \sqrt{dR^2 + c\left(3 + \frac{1}{B^2}\right)R^2 d \log \frac{m}{\delta}} \\ &\leq cR \sqrt{d \log \frac{m}{\delta}}. \end{aligned} \quad (6.10)$$

The next step is to bound  $\left\| \sum_{k=d+2}^{t-1} \xi_i(k) \mathbf{y}(k) \right\|_{\mathbf{M}_t^{-1}}$ . Lemma 6.6 states that for any  $\delta > 0$ , the following inequality holds with probability at least  $1 - \frac{\delta}{m}$ ,

$$\left\| \sum_{k=d+2}^{t-1} \xi_i(k) \mathbf{y}(k) \right\|_{\mathbf{M}_t^{-1}}^2 \leq 2R^2 \log \left( \frac{m \det(\mathbf{M}_t)^{\frac{1}{2}} \det(\mathbf{M})^{\frac{1}{2}}}{\delta} \right).$$

By Assumption 6.1, i.e.,  $\|\mathbf{x}\|_2 \leq L$  for any  $\mathbf{x} \in \mathcal{S}_t$ , and Lemma 6.7, we have  $\det(\mathbf{M}_t) \leq \left( \frac{\text{tr}(\mathbf{M}) + t(L^2 + 1)}{d+1} \right)^{d+1}$ . Therefore,

$$\left\| \sum_{k=d+2}^{t-1} \xi_i(k) \mathbf{y}(k) \right\|_{\mathbf{M}_t^{-1}}^2 \leq 2R^2 \left[ \log \frac{m}{\delta} + d \log B + d \log \left( B^2 + 1 + \frac{t(L^2 + 1)}{d+1} \right) \right]. \quad (6.11)$$

By combining Inequalities (6.10) and (6.11), we know that there exists constant  $c$  such that the following inequality holds with probability at least  $1 - \frac{2\delta}{m}$ :

$$\begin{aligned} \left\| \sum_{k=1}^{t-1} \xi_i(k) \mathbf{y}(k) \right\|_{\mathbf{M}_t^{-1}} &\leq cR \sqrt{d \left[ \log \frac{m}{\delta} + \log \left( B^2 + 1 + \frac{t(L^2 + 1)}{d} \right) \right]} \\ &\leq cR \sqrt{d \left[ \log \frac{m}{\delta} + \log(L^2 + 1) + \log \left( 1 + \frac{t}{d} \right) \right]} \\ &\leq cR \sqrt{d \log \frac{2m(L^2 + 1)t}{\delta d}}, \end{aligned} \quad (6.12)$$

where the last two inequalities hold since  $L \geq B$  and  $t > d$ .

Let  $\theta(t) = \frac{c\mu R}{c_\mu} \sqrt{d \log \frac{2m(L^2 + 1)t}{\delta d}}$ , then from Inequalities (6.9) and (6.12), we have

$$|f(\bar{\mathbf{a}}_i(t)^\top \bar{\mathbf{x}}) - f(\bar{\mathbf{a}}_i^\top \bar{\mathbf{x}})| \leq \theta(t) \|\bar{\mathbf{x}}\|_{\mathbf{M}_t^{-1}} \quad (6.13)$$

holds for  $i = 1, \dots, m$  with probability at least  $1 - 2\delta$ .



When  $\bar{\mathbf{x}} = (\mathbf{z}, -1)^\top$  with  $\mathbf{z} = \frac{\mathbf{a}_i(t) - \mathbf{a}_i}{\|\mathbf{a}_i(t) - \mathbf{a}_i\|_2}$ , Inequality (6.13) implies

$$|f(\mathbf{a}_i(t)^\top \mathbf{z} - b_i(t)) - f(\mathbf{a}_i^\top \mathbf{z} - b_i)| \leq \theta(t) \sqrt{\frac{2}{\lambda_{\min}(\mathbf{M}_t)}}.$$

Similarly, when  $\bar{\mathbf{x}} = (\mathbf{0}^\top, -1)^\top$ , we have

$$|f(-b_i(t)) - f(-b_i)| \leq \theta(t) \sqrt{\frac{1}{\lambda_{\min}(\mathbf{M}_t)}}.$$

Then by Assumption 6.2, we have

$$|f(\mathbf{a}_i(t)^\top \mathbf{z} - b_i(t)) - f(\mathbf{a}_i^\top \mathbf{z} - b_i)| \geq c_\mu \|\mathbf{a}_i(t) - \mathbf{a}_i\|_2 + b_i - b_i(t)$$

and

$$|f(-b_i(t)) - f(-b_i)| \geq c_\mu |b_i(t) - b_i|.$$

Therefore,

$$\max\{\|\mathbf{a}_i(t) - \mathbf{a}_i\|_2, |b_i(t) - b_i|\} \leq \frac{3\theta(t)}{c_\mu} \sqrt{\frac{1}{\lambda_{\min}(\mathbf{M}_t)}}. \quad (6.14)$$

In order to bound the right hand side of (6.14), we need to provide a lower bound of  $\lambda_{\min}(\mathbf{M}_t)$ . Recall that  $\mathbf{x}(t) = [1 - \eta(t)]\hat{\mathbf{x}}(t) + \eta(t)\tilde{\mathbf{x}}(t)$  for  $t \geq d + 2$ . Let  $\hat{\mathbf{y}}(t) = (\hat{\mathbf{x}}(t)^\top, -1)^\top$  and  $\tilde{\mathbf{y}}(t) = (\tilde{\mathbf{x}}(t)^\top, -1)^\top$ , then

$$\begin{aligned} \mathbf{M}_t &= \sum_{k=1}^{d+1} \mathbf{y}(k)\mathbf{y}(k)^\top + \sum_{k=d+2}^{t-1} \mathbf{y}(k)\mathbf{y}(k)^\top \\ &\geq \sum_{k=d+2}^{t-1} [(1 - \eta(t))\hat{\mathbf{y}}(t)\hat{\mathbf{y}}(t)^\top + \eta(t)\tilde{\mathbf{y}}(t)\tilde{\mathbf{y}}(t)^\top] \\ &\geq \sum_{k=d+2}^{t-1} \eta(t)\tilde{\mathbf{y}}(t)\tilde{\mathbf{y}}(t)^\top. \end{aligned}$$

Since  $\tilde{\eta}(t)$  is drawn from Bernoulli distribution with success probability  $p(t) \propto \frac{1}{t^{1/3}}$  and  $\tilde{\mathbf{x}}(t)$  follows the uniform distribution on  $\mathcal{S}_{d-1}(B)$ , by Lemma 6.2, when  $t > d + 1$ ,

$$\lambda_{\min}(\mathbf{M}_t) \geq c_1 \min\left\{\frac{B^2}{d}, 1\right\} t^{2/3} \quad (6.15)$$

holds with probability at least

$$1 - 2 \max \left\{ d \exp\left(-\frac{c_2 t^{2/3}}{d^2}\right), \exp(-c_3 t^{1/3}) \right\}$$

for constants  $c_1, c_2$  and  $c_3$ . Thus, there exists constant  $c_4$  such that when  $t \geq c_4 d^3 \log^3 t$ , Inequality (6.15) holds with probability at least  $1 - \frac{c_5}{t^{10}}$ . Then by the union bound, Inequality (6.15) holds for all  $t$  such that  $T \geq t \geq c_4 d^3 \log^3 t$  with probability at least  $1 - \frac{c_5}{T^{10}}$ .

By Inequalities (6.14) and (6.15), there exists constant  $c_6$  such that

$$\max\{\|\mathbf{a}_i(t) - \mathbf{a}_i\|_2, |b_i(t) - b_i|\} \leq \frac{c_6 \theta(t) \sqrt{d}}{c_\mu t^{1/3}},$$

which implies

$$\|\mathbf{A}(t) - \mathbf{A}\|_2 \leq \frac{c_6 \theta(t) \sqrt{md}}{c_\mu t^{1/3}}, \|\mathbf{b}(t) - \mathbf{b}\|_2 \leq \frac{c_6 \theta(t) \sqrt{md}}{c_\mu t^{1/3}}. \quad (6.16)$$

By Lemma 6.3 and Lemma 6.4, there exists constant  $\rho$  so that when

$$\max\{\|\mathbf{A}(t) - \mathbf{A}\|_2, \|\mathbf{b}(t) - \mathbf{b}\|_2\} \leq \frac{\rho}{2},$$

Linear program (6.5) is feasible and

$$|\mathbf{c}^\top \hat{\mathbf{x}}(t) - \mathbf{c}^\top \mathbf{x}^*(t)| \leq c_7 \max\{\|\mathbf{A}(t) - \mathbf{A}\|_2, \|\mathbf{b}(t) - \mathbf{b}\|_2\},$$

where  $c_7$  is a constant depending on  $\mathbf{A}, \mathbf{b}, \mathbf{c}$  and  $\rho$ . Thus, from the upper bounds of  $\|\mathbf{A}(t) - \mathbf{A}\|_2$  and  $\|\mathbf{b}(t) - \mathbf{b}\|_2$  as shown in (6.16), we know that when  $\frac{t}{\theta(t)^3} \geq \left(\frac{2c_6 \sqrt{md}}{c_\mu \rho}\right)^3$ ,

$$|\mathbf{c}^\top \hat{\mathbf{x}}(t) - \mathbf{c}^\top \mathbf{x}^*(t)| \leq \frac{c_7 \theta(t) \sqrt{md}}{c_\mu t^{1/3}}. \quad (6.17)$$

Let  $T_0$  be the minimum value of  $t$  such that  $\frac{t}{\theta(t)^3} \geq \left(\frac{2c_6 \sqrt{md}}{c_\mu \rho}\right)^3$  and  $\frac{t}{\log^3 t} \geq c_4 d^3$ , one can

easily verify that there exists constant  $c_8$  so that

$$T_0 \leq c_8 \left( \frac{l_\mu R d}{c_\mu^2} \sqrt{m \log \frac{m^{3/2}}{c_\mu \delta}} \right)^3.$$

Then by applying Inequality (6.17), we can develop an upper bound of the regret. Since  $U \triangleq 2L\|\mathbf{c}\|_2$  is an upper bound for  $|\mathbf{c}^\top \mathbf{x}(t) - \mathbf{c}^\top \mathbf{x}^*(t)|$ , we have

$$\begin{aligned} \text{Regret}(T) &= \sum_{t=1}^{T_0} |\mathbf{c}^\top \mathbf{x}(t) - \mathbf{c}^\top \mathbf{x}^*(t)| + \sum_{t=T_0+1}^T |\mathbf{c}^\top \mathbf{x}(t) - \mathbf{c}^\top \mathbf{x}^*(t)| \\ &\leq T_0 U + \sum_{t=T_0+1}^T \eta(t) |\mathbf{c}^\top \hat{\mathbf{x}}(t) - \mathbf{c}^\top \mathbf{x}^*(t)| + \sum_{t=T_0+1}^T [1 - \eta(t)] |\mathbf{c}^\top \hat{\mathbf{x}}(t) - \mathbf{c}^\top \mathbf{x}^*(t)| \\ &\leq [T_0 + \sum_{t=1}^T \eta(t)] U + \sum_{t=T_0+1}^T [1 - \eta(t)] |\mathbf{c}^\top \hat{\mathbf{x}}(t) - \mathbf{c}^\top \mathbf{x}^*(t)| \\ &\leq [T_0 + c_9 T^{2/3}] U + \sum_{t=1}^T \frac{c_7 \theta(t) \sqrt{m d}}{c_\mu t^{1/3}} \\ &\leq 2T_0 L \|\mathbf{c}\|_2 + \left[ \frac{c_7 \theta(T) \sqrt{m d}}{c_\mu} + c_9 L \|\mathbf{c}\|_2 \right] \sqrt{T^{4/3} \log T} \end{aligned}$$

holds with probability at least  $1 - 2\delta - cT^{-9}$  for some constant  $c$ .

We now show the upper bound of ‘‘constraint violation’’. For the  $i$ th constraint, we have

$$\begin{aligned} \text{Violation}_i(T) &= \sum_{t=1}^{T_0} [\mathbf{a}_i^\top \mathbf{x}(t) - b_i]_+ + \sum_{t=T_0+1}^T [\mathbf{a}_i^\top \mathbf{x}(t) - b_i]_+ \\ &\leq \sum_{t=1}^{T_0} [\mathbf{a}_i^\top \mathbf{x}(t) - b_i]_+ + \sum_{t=T_0+1}^T [\mathbf{a}_i(t)^\top \mathbf{x}(t) - b_i(t)]_+ \\ &\quad + \sum_{t=T_0+1}^T [\mathbf{a}_i^\top \mathbf{x}(t) - b_i - \mathbf{a}_i(t)^\top \mathbf{x}(t) + b_i(t)]_+ \\ &\leq T_0 (L \|\mathbf{a}_i\|_2 + |b_i|) + \sum_{t=T_0+1}^T \eta(t) [\mathbf{a}_i(t)^\top \hat{\mathbf{x}}(t) - b_i(t)]_+ \\ &\quad + \sum_{t=T_0+1}^T |\mathbf{a}_i^\top \mathbf{x}(t) - b_i - \mathbf{a}_i(t)^\top \mathbf{x}(t) + b_i(t)| \\ &\leq (T_0 + c_9 T^{2/3}) (L \|\mathbf{a}_i\|_2 + |b_i|) + \sum_{t=T_0+1}^T |\bar{\mathbf{a}}_i^\top \bar{\mathbf{x}}(t) - \bar{\mathbf{a}}_i(t)^\top \bar{\mathbf{x}}(t)| \end{aligned}$$

By Assumption 6.2 and Inequality (6.13),

$$|\bar{\mathbf{a}}_i^\top \bar{\mathbf{x}}(t) - \bar{\mathbf{a}}_i(t)^\top \bar{\mathbf{x}}(t)| \leq \frac{1}{c_\mu} |f(\bar{\mathbf{a}}_i(t)^\top \bar{\mathbf{x}}) - f(\bar{\mathbf{a}}_i^\top \bar{\mathbf{x}})| \leq \frac{\theta(t)}{c_\mu} \|\bar{\mathbf{x}}\|_{\mathbf{M}_t^{-1}} \quad (6.18)$$

As shown in [CLRS11],

$$\sum_{t=1}^T \|\bar{\mathbf{x}}(t)\|_{\mathbf{M}_t^{-1}} \leq 5\sqrt{(d+1)T \log T} \leq 10\sqrt{dT \log T},$$

implying that

$$\sum_{t=T_0+1}^T |\bar{\mathbf{a}}_i^\top \bar{\mathbf{x}}(t) - \bar{\mathbf{a}}_i(t)^\top \bar{\mathbf{x}}(t)| \leq \frac{10\theta(T)}{c_\mu} \sqrt{dT \log T}.$$

Therefore, we have

$$\text{Violation}_i(T) = (T_0 + c_9 T^{2/3})(L\|\mathbf{a}_i\|_2 + |b_i|) + \frac{10\theta(T)}{c_\mu} \sqrt{dT \log T}.$$

Hence we obtain this theorem.  $\square$

### 6.8.3 Proof of Theorem 6.2

*Proof.* From the proof of Theorem 6.1, we know that

$$|[f(\bar{\mathbf{a}}_i(t)^\top \bar{\mathbf{x}})]_+ - [f(\bar{\mathbf{a}}_i^\top \bar{\mathbf{x}})]_+| \leq |f(\bar{\mathbf{a}}_i(t)^\top \bar{\mathbf{x}}) - f(\bar{\mathbf{a}}_i^\top \bar{\mathbf{x}})| \leq \theta(t) \|\bar{\mathbf{x}}\|_{\mathbf{M}_t^{-1}} \quad (6.19)$$

holds for all  $i = 1, \dots, m$  with probability at least  $1 - \delta$ , where  $\theta(t) = \frac{c_\mu R}{c_\mu} \sqrt{d \log \frac{2m(L^2+1)t}{\delta d}}$ ,  $\bar{\mathbf{a}}_i = (\mathbf{a}_i^\top, b_i)^\top$ ,  $\bar{\mathbf{a}}_i(t) = (\mathbf{a}_i(t)^\top, b_i(t))^\top$  and  $\bar{\mathbf{x}} = (\mathbf{x}^\top, -1)^\top$ .

Recall that Step 4 of Algorithm 6.3 needs to solve Problem (6.6). By Inequality (6.19), for any feasible solution  $\mathbf{x}$  of Problem (6.2), i.e.,  $f(\bar{\mathbf{a}}_i^\top \bar{\mathbf{x}}) \leq 0$ , we have

$$f(\bar{\mathbf{a}}_i(t)^\top \bar{\mathbf{x}}) \leq f(\bar{\mathbf{a}}_i^\top \bar{\mathbf{x}}) + \theta(t) \|\bar{\mathbf{x}}\|_{\mathbf{M}_t^{-1}} \leq \theta(t) \|\bar{\mathbf{x}}\|_{\mathbf{M}_t^{-1}},$$

which implies that  $\mathbf{x}$  is also a feasible solution of Problem (6.6). Hence w.h.p. the regret

bound satisfies

$$\begin{aligned}
\text{Regret}(T) &= \sum_{t=1}^{d+1} |\mathbf{c}^\top \mathbf{x}(t) - \mathbf{c}^\top \mathbf{x}^*(t)| + \sum_{t=d+2}^T |\mathbf{c}^\top \mathbf{x}(t) - \mathbf{c}^\top \mathbf{x}^*(t)| \\
&\leq (d+1) \|\mathbf{c}\|_2 \|\mathbf{x}(t) - \mathbf{x}^*(t)\|_2 + \sum_{t=d+2}^T |\mathbf{c}^\top \mathbf{x}(t) - \mathbf{c}^\top \mathbf{x}^*(t)| \\
&\leq 2(d+1)L \|\mathbf{c}\|_2 + \sum_{t=d+2}^T |\mathbf{c}^\top \mathbf{x}(t) - \mathbf{c}^\top \mathbf{x}^*(t)|.
\end{aligned}$$

In order to bound  $|\mathbf{c}^\top \mathbf{x}(t) - \mathbf{c}^\top \mathbf{x}^*(t)|$ , we consider the following linear program:

$$\begin{aligned}
&\min \quad \mathbf{c}^\top \mathbf{x} \\
&\text{s.t.} \quad f(\mathbf{a}_i^\top \mathbf{x} - b_i) \leq 2\theta(t) \|\bar{\mathbf{x}}(t)\|_{\mathbf{M}_t^{-1}}, \forall i \in [m], \\
&\quad \mathbf{x} \in \mathcal{S}_t.
\end{aligned} \tag{6.20}$$

We denote the optimal solution of (6.20) by  $\hat{\mathbf{x}}$ . By Inequality (6.19), with probability at least  $1 - \delta$ ,  $\mathbf{x}(t)$  satisfies that

$$f(\bar{\mathbf{a}}_i^\top \bar{\mathbf{x}}(t)) \leq f(\bar{\mathbf{a}}_i(t)^\top \bar{\mathbf{x}}(t)) + \theta(t) \|\bar{\mathbf{x}}(t)\|_{\mathbf{M}_t^{-1}} \leq 2\theta(t) \|\bar{\mathbf{x}}(t)\|_{\mathbf{M}_t^{-1}}$$

for all  $i = 1, \dots, m$ . Therefore,  $\mathbf{x}(t)$  is a feasible solution of (6.20). Recall that any feasible solution of (6.2) is also a feasible solution of (6.6). Thus,

$$\mathbf{c}^\top \hat{\mathbf{x}} \leq \mathbf{c}^\top \mathbf{x}(t) \leq \mathbf{c}^\top \mathbf{x}^*(t).$$

By Lemma 6.5, there exists a feasible solution  $\tilde{\mathbf{x}}$  of (6.2) and a constant  $c_1$  so that

$$\|\hat{\mathbf{x}} - \tilde{\mathbf{x}}\|_2 \leq c_1 f^{-1}(2\theta(t) \|\bar{\mathbf{x}}(t)\|_{\mathbf{M}_t^{-1}}).$$

Thus, we have

$$\begin{aligned}
|\mathbf{c}^\top \mathbf{x}(t) - \mathbf{c}^\top \mathbf{x}^*(t)| &\leq |\mathbf{c}^\top \hat{\mathbf{x}} - \mathbf{c}^\top \mathbf{x}^*(t)| \\
&= \mathbf{c}^\top \mathbf{x}^*(t) - \mathbf{c}^\top \tilde{\mathbf{x}} + \mathbf{c}^\top \tilde{\mathbf{x}} - \mathbf{c}^\top \hat{\mathbf{x}} \\
&\leq \mathbf{c}^\top \tilde{\mathbf{x}} - \mathbf{c}^\top \hat{\mathbf{x}} \\
&\leq \frac{c_1 \theta(t) \|\mathbf{c}\|_2}{c_\mu} \|\tilde{\mathbf{x}}(t)\|_{\mathbf{M}_t^{-1}}
\end{aligned}$$

for some constant  $c_1$ . Recall that  $\sum_{t=1}^T \|\tilde{\mathbf{x}}(t)\|_{\mathbf{M}_t^{-1}} \leq 10\sqrt{dT \log T}$  as shown in [CLRS11], we have

$$\begin{aligned}
\text{Regret}(T) &\leq 2(d+1)L\|\mathbf{c}\|_2 + \sum_{t=d+2}^T \frac{c_1 \theta(t) \|\mathbf{c}\|_2}{c_\mu} \|\tilde{\mathbf{x}}\|_{\mathbf{M}_t^{-1}} \\
&\leq 2(d+1)L\|\mathbf{c}\|_2 + \frac{c_1 \theta(T) \|\mathbf{c}\|_2}{c_\mu} \sqrt{dT \log T}.
\end{aligned}$$

For the ‘‘constraint violation’’, by Assumption 6.2 and Inequality (6.13), we have

$$\begin{aligned}
\text{Violation}_i(T) &= \sum_{t=1}^{d+1} [\mathbf{a}_i^\top \mathbf{x}(t) - b_i]_+ + \sum_{t=d+2}^T [\mathbf{a}_i^\top \mathbf{x}(t) - b_i]_+ \\
&\leq \sum_{t=1}^{d+1} [\mathbf{a}_i^\top \mathbf{x}(t) - b_i]_+ + \sum_{t=d+2}^T [\mathbf{a}_i(t)^\top \mathbf{x}(t) - b_i(t)]_{++} \\
&\quad \sum_{t=d+2}^T [\mathbf{a}_i^\top \mathbf{x}(t) - b_i - \mathbf{a}_i(t)^\top \mathbf{x}(t) + b_i(t)]_+ \\
&\leq (d+1)(L\|\mathbf{a}_i\|_2 + |b_i|) + \sum_{t=d+2}^T \frac{2\theta(t)}{c_\mu} \|\tilde{\mathbf{x}}\|_{\mathbf{M}_t^{-1}}, \\
&\leq (d+1)(L\|\mathbf{a}_i\|_2 + |b_i|) + \frac{20\theta(T)}{c_\mu} \sqrt{dT \log T}
\end{aligned}$$

holds w.h.p., where  $c$  is a constant. Hence we obtain this theorem.  $\square$

#### 6.8.4 Proof of Theorem 6.3

**Lemma 6.8.** [AYPS11] *Let  $\mathbf{A}$ ,  $\mathbf{B}$  and  $\mathbf{C}$  be positive semi-definite matrices such that  $\mathbf{A} = \mathbf{B} + \mathbf{C}$ . Then we have that*

$$\sup_{\mathbf{x} \neq \mathbf{0}} \frac{\mathbf{x}^\top \mathbf{A} \mathbf{x}}{\mathbf{x}^\top \mathbf{B} \mathbf{x}} \leq \frac{\det(\mathbf{A})}{\det(\mathbf{B})}.$$

We now prove Theorem 6.3.

*Proof.* We use the same notation as that in the proof of Theorem 6.1 and let  $\tau(t) = \max\{\tau : \tau \leq t, \det(\mathbf{M}_\tau) > (1 + \gamma) \det(\mathbf{M}_{\tau-1})\}$ . Then we have that,

$$\begin{aligned} |f(\bar{\mathbf{a}}_i(t)^\top \bar{\mathbf{x}}) - f(\bar{\mathbf{a}}_i^\top \bar{\mathbf{x}})| &\leq l_\mu |(\bar{\mathbf{a}}_i(t) - \bar{\mathbf{a}}_i)^\top \bar{\mathbf{x}}| \\ &\leq l_\mu |(\bar{\mathbf{a}}_i(t) - \bar{\mathbf{a}}_i)^\top \mathbf{M}_t^{\frac{1}{2}} \mathbf{M}_t^{-\frac{1}{2}} \bar{\mathbf{x}}| \\ &\leq l_\mu \|(\bar{\mathbf{a}}_i(t) - \bar{\mathbf{a}}_i)^\top \mathbf{M}_t^{\frac{1}{2}}\|_2 \|\bar{\mathbf{x}}\|_{\mathbf{M}_t^{-1}}. \end{aligned}$$

By Lemma 6.8 and the definition of  $\tau(t)$ ,

$$\begin{aligned} (\bar{\mathbf{a}}_i(t) - \bar{\mathbf{a}}_i)^\top \mathbf{M}_t (\bar{\mathbf{a}}_i(t) - \bar{\mathbf{a}}_i) &\leq (\bar{\mathbf{a}}_i(t) - \bar{\mathbf{a}}_i)^\top \mathbf{M}_{\tau(t)} (\bar{\mathbf{a}}_i(t) - \bar{\mathbf{a}}_i) \cdot \frac{\det(\mathbf{M}_t)}{\det(\mathbf{M}_{\tau(t)})} \\ &= (\bar{\mathbf{a}}_i(\tau(t)) - \bar{\mathbf{a}}_i)^\top \mathbf{M}_{\tau(t)} (\bar{\mathbf{a}}_i(\tau(t)) - \bar{\mathbf{a}}_i) \cdot \frac{\det(\mathbf{M}_t)}{\det(\mathbf{M}_{\tau(t)})} \\ &\leq (1 + \gamma) (\bar{\mathbf{a}}_i(\tau(t)) - \bar{\mathbf{a}}_i)^\top \mathbf{M}_{\tau(t)} (\bar{\mathbf{a}}_i(\tau(t)) - \bar{\mathbf{a}}_i) \end{aligned}$$

which implies that

$$\|(\bar{\mathbf{a}}_i(t) - \bar{\mathbf{a}}_i)^\top \mathbf{M}_t^{\frac{1}{2}}\|_2 \leq \sqrt{1 + \gamma} \|(\bar{\mathbf{a}}_i(\tau(t)) - \bar{\mathbf{a}}_i)^\top \mathbf{M}_{\tau(t)}^{\frac{1}{2}}\|_2.$$

Then from the proof of Theorem 6.1, we know that

$$(\bar{\mathbf{a}}_i(t) - \bar{\mathbf{a}}_i)^\top \mathbf{M}_t (\bar{\mathbf{a}}_i(t) - \bar{\mathbf{a}}_i) \leq \sqrt{1 + \gamma} \theta(t) \|\bar{\mathbf{x}}\|_{\mathbf{M}_t^{-1}}.$$

Finally, we can obtain this theorem by following the proofs of Theorem 6.1 or Theorem 6.2. □

### 6.8.5 Proofs in Section 6.6

**Theorem 6.8.** *Consider the following convex optimization problem*

$$\begin{aligned} \min \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & g_i(\mathbf{x}) \leq 0, \quad \forall i = 1, \dots, m, \\ & \mathbf{x} \in \mathcal{X}, \end{aligned}$$

where  $f(\cdot)$ ,  $g_i(\cdot)$  are convex functions,  $\mathcal{X}$  is a nonempty convex set and  $\mathcal{X} \subseteq \cap_i \text{dom}(g_i) \cap \text{dom}(f)$ . If its optimal value is finite and the Slater's condition is satisfied, namely, there exists a vector  $\bar{\mathbf{x}} \in \mathcal{X}$  such that  $g_i(\bar{\mathbf{x}}) < 0$  for all  $i = 1, \dots, m$ , then there is no duality gap and the set of dual optimal solutions is nonempty and bounded.

*Proof.* The proof of this theorem is standard. For clearness and completeness, we provide its proof here. Consider the set  $\mathcal{V} = \{(\mathbf{u}, w) : g(\mathbf{x}) \leq \mathbf{u}, f(\mathbf{x}) \leq w, \mathbf{x} \in \mathcal{X}\}$  and denote by  $f^*$  the optimal value. Since  $f^*$  is optimal, the vector  $(\mathbf{0}, f^*)$  is not in the interior of  $\mathcal{V}$ . Thus, by the supporting hyperplane theorem, there exists a hyperplane passing through  $(\mathbf{0}, f^*)$  and supporting  $\mathcal{V}$ , namely, there exists  $(\boldsymbol{\mu}, \mu_0)$  with  $(\boldsymbol{\mu}, \mu_0) \neq \mathbf{0}$  such that

$$\boldsymbol{\mu}^\top \mathbf{u} + \mu_0 w \geq \mu_0 f^*, \quad \forall (\mathbf{u}, w) \in \mathcal{V}, \quad (6.21)$$

which implies that  $\boldsymbol{\mu} \geq \mathbf{0}$  and  $\mu_0 \geq 0$ . Suppose that  $\mu_0 = 0$ , then  $\boldsymbol{\mu} \neq \mathbf{0}$  and

$$\inf_{(\mathbf{u}, w) \in \mathcal{V}} \boldsymbol{\mu}^\top \mathbf{u} \geq 0. \quad (6.22)$$

On the other hand, since  $\boldsymbol{\mu} \geq \mathbf{0}$  and  $\boldsymbol{\mu} \neq \mathbf{0}$ , by the definition of  $\mathcal{V}$ , we have

$$\inf_{(\mathbf{u}, w) \in \mathcal{V}} \boldsymbol{\mu}^\top \mathbf{u} = \inf_{\mathbf{x} \in \mathcal{X}} \boldsymbol{\mu}^\top g(\mathbf{x}) \leq \boldsymbol{\mu}^\top g(\bar{\mathbf{x}}) < 0,$$

which contradicts with (6.22). Hence  $\mu_0 > 0$ . Let  $\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu}/\mu_0$ , Inequality (6.21) implies

$$\inf_{(\mathbf{u}, w) \in \mathcal{V}} \tilde{\boldsymbol{\mu}}^\top \mathbf{u} + w \geq f^*.$$

Therefore,  $h(\tilde{\boldsymbol{\mu}}) = \inf_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) + \tilde{\boldsymbol{\mu}}^\top g(\mathbf{x}) \geq f^*$  which implies that the dual optimal value  $h^* \geq f^*$ . On the other hand, by the weak duality  $h^* \leq f^*$ , we have  $h^* = f^*$  and  $\tilde{\boldsymbol{\mu}}$  is a dual optimal solution. For any dual optimal solution  $\tilde{\boldsymbol{\mu}}$ , we have

$$h^* = \inf_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) + \tilde{\boldsymbol{\mu}}^\top g(\mathbf{x}) \leq f(\bar{\mathbf{x}}) + \tilde{\boldsymbol{\mu}}^\top g(\bar{\mathbf{x}}) \leq f(\bar{\mathbf{x}}) + \max_i \tilde{\mu}_i g_i(\bar{\mathbf{x}}) \cdot \|\tilde{\boldsymbol{\mu}}\|_1.$$

Thus, we have  $\|\tilde{\boldsymbol{\mu}}\|_1 \leq \frac{f(\bar{\mathbf{x}}) - h^*}{\min_i \{-g_i(\bar{\mathbf{x}})\}} = \frac{f(\bar{\mathbf{x}}) - f^*}{\min_i \{-g_i(\bar{\mathbf{x}})\}}$ . □



**Lemma 6.9.** Denote by  $\mathcal{P}(\boldsymbol{\delta})$  the following optimization problem

$$\begin{aligned} \min \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & g_i(\mathbf{x}) \leq \delta_i, \quad \forall i = 1, \dots, m, \\ & \mathbf{x} \in \mathcal{X}, \end{aligned}$$

and denote by  $\text{opt}(\boldsymbol{\delta})$  its optimal value. Then if  $\mathcal{P}(\mathbf{0})$  satisfies the conditions of Theorem 6.8 and  $\boldsymbol{\delta} \geq 0$ , we have  $\text{opt}(\mathbf{0}) - \text{opt}(\boldsymbol{\delta}) \leq c \|\boldsymbol{\delta}\|_\infty$  for some constant  $c$  depending on  $\mathcal{P}(\mathbf{0})$ .

*Proof.* Since  $\mathcal{P}(\mathbf{0})$  satisfies the conditions of Theorem 6.8 and  $\boldsymbol{\delta} \geq 0$ ,  $\mathcal{P}(\boldsymbol{\delta})$  also satisfies these conditions. Let  $\tilde{\boldsymbol{\mu}}$  be the dual optimal solution of  $\mathcal{P}(\mathbf{0})$ . By the strong duality shown in Theorem 6.8,

$$\begin{aligned} \text{opt}(\boldsymbol{\delta}) &= \max_{\boldsymbol{\mu} \geq 0} \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) + \boldsymbol{\mu}^\top g(\mathbf{x}) - \boldsymbol{\mu}^\top \boldsymbol{\delta} \\ &\geq \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) + \tilde{\boldsymbol{\mu}}^\top g(\mathbf{x}) - \tilde{\boldsymbol{\mu}}^\top \boldsymbol{\delta} \\ &= \text{opt}(\mathbf{0}) - \tilde{\boldsymbol{\mu}}^\top \boldsymbol{\delta}. \end{aligned}$$

Therefore,  $\text{opt}(\mathbf{0}) - \text{opt}(\boldsymbol{\delta}) \leq \|\tilde{\boldsymbol{\mu}}\|_1 \|\boldsymbol{\delta}\|_\infty$ . Since  $\|\tilde{\boldsymbol{\mu}}\|_1$  is bounded, we obtain this lemma.  $\square$

**Lemma 6.10.** For  $\mathbf{A} \in \mathbb{R}^{n \times m}$ ,  $\mathbf{b} \in \mathbb{R}^m$ , function  $\phi(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^n$ , and nonempty set  $\mathcal{X}$ , if there exists  $\bar{\mathbf{x}} \in \mathcal{X}$  such that  $\mathbf{A}^\top \phi(\bar{\mathbf{x}}) < \mathbf{b}$ , then there exists a constant  $\rho$  so that

$$\hat{\mathbf{A}}^\top \phi(\mathbf{x}) \leq \hat{\mathbf{b}}, \quad \mathbf{x} \in \mathcal{X}$$

is always strictly feasible whenever  $\max\{\|\hat{\mathbf{a}}_i - \mathbf{a}_i\|_2, |\hat{b}_i - b_i|\} < \rho$  where  $\mathbf{a}_i$  and  $\hat{\mathbf{a}}_i$  are the  $i^{\text{th}}$  columns of  $\mathbf{A}$  and  $\hat{\mathbf{A}}$ , respectively.

*Proof.* By the assumption above,  $\mathbf{a}_i^\top \phi(\bar{\mathbf{x}}) < b_i$  for all  $i = 1, \dots, m$ . Note that for any  $\hat{\mathbf{a}}_i$  and  $\hat{b}_i$ ,  $\hat{\mathbf{a}}_i^\top \phi(\bar{\mathbf{x}}) < \hat{b}_i$  holds as long as  $(\hat{\mathbf{a}}_i - \mathbf{a}_i)^\top \phi(\bar{\mathbf{x}}) - (\hat{b}_i - b_i) < b_i - \mathbf{a}_i^\top \phi(\bar{\mathbf{x}})$ . Therefore, if

$$\max\{\|\hat{\mathbf{a}}_i - \mathbf{a}_i\|_2, |\hat{b}_i - b_i|\} < \frac{b_i - \mathbf{a}_i^\top \phi(\bar{\mathbf{x}})}{2} \min\left\{1, \frac{1}{\|\phi(\bar{\mathbf{x}})\|_2}\right\},$$

$\bar{\mathbf{x}}$  is a feasible solution of  $\hat{\mathbf{a}}_i^\top \phi(\bar{\mathbf{x}}) < \hat{b}_i$ . Then by taking  $\rho = \min_i \frac{b_i - \mathbf{a}_i^\top \phi(\bar{\mathbf{x}})}{2} \min\left\{1, \frac{1}{\|\phi(\bar{\mathbf{x}})\|_2}\right\}$ ,

we obtain this lemma. □

**Lemma 6.11.** *Denote by  $\mathcal{P}(\boldsymbol{\delta})$  the following optimization problem*

$$\begin{aligned} \min \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & g_i(\mathbf{x}) \leq \delta_i, \quad \forall i = 1, \dots, m, \\ & \mathbf{x} \in \mathcal{X}, \end{aligned}$$

and denote by  $\text{opt}(\boldsymbol{\delta})$  its optimal value. If  $\mathcal{P}(\mathbf{0})$  satisfies the conditions of Theorem 6.8, then there exist constant  $\rho$  and vector  $\bar{\mathbf{x}} \in \mathcal{X}$  which only depend on  $\mathcal{P}(\mathbf{0})$  so that for any  $\|\boldsymbol{\delta}\|_\infty \leq \rho$ , the following inequality holds

$$|\text{opt}(\boldsymbol{\delta}) - \text{opt}(\mathbf{0})| \leq \frac{2|f(\bar{\mathbf{x}}) - \text{opt}(\mathbf{0})|}{\min_i \{-g_i(\bar{\mathbf{x}})\}} \cdot \|\boldsymbol{\delta}\|_\infty.$$

*Proof.* Since  $\mathcal{P}(\mathbf{0})$  satisfies the Slater's condition, there exists  $\bar{\mathbf{x}} \in \mathcal{X}$  so that  $g_i(\bar{\mathbf{x}}) < 0$  for  $i = 1, \dots, m$ . Thus, for any  $\boldsymbol{\delta}$  such that  $\|\boldsymbol{\delta}\|_\infty \leq \rho \triangleq \frac{1}{2} \min_i |g_i(\bar{\mathbf{x}})|$ ,  $\mathcal{P}(\boldsymbol{\delta})$  is feasible and satisfies the Slater's condition.

Let  $\boldsymbol{\mu}_0$  and  $\boldsymbol{\mu}_\delta$  be the dual optimal solutions of  $\mathcal{P}(\mathbf{0})$  and  $\mathcal{P}(\boldsymbol{\delta})$ , respectively. Then from Theorem 6.8, we know that

$$\begin{aligned} \text{opt}(\boldsymbol{\delta}) &= \max_{\boldsymbol{\mu} \geq 0} \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) + \boldsymbol{\mu}^\top g(\mathbf{x}) - \boldsymbol{\mu}^\top \boldsymbol{\delta} \\ &\geq \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) + \boldsymbol{\mu}_0^\top g(\mathbf{x}) - \boldsymbol{\mu}_0^\top \boldsymbol{\delta} \\ &= \text{opt}(\mathbf{0}) - \boldsymbol{\mu}_0^\top \boldsymbol{\delta}, \end{aligned}$$

and

$$\begin{aligned} \text{opt}(\mathbf{0}) &= \max_{\boldsymbol{\mu} \geq 0} \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) + \boldsymbol{\mu}^\top g(\mathbf{x}) \\ &\geq \min_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}) + \boldsymbol{\mu}_\delta^\top g(\mathbf{x}) - \boldsymbol{\mu}_\delta^\top \boldsymbol{\delta} + \boldsymbol{\mu}_\delta^\top \boldsymbol{\delta} \\ &= \text{opt}(\boldsymbol{\delta}) + \boldsymbol{\mu}_\delta^\top \boldsymbol{\delta}. \end{aligned}$$

Theorem 6.8 also shows that

$$\boldsymbol{\mu}_0 \leq \frac{f(\bar{\mathbf{x}}) - \text{opt}(\mathbf{0})}{\min_i \{-g_i(\bar{\mathbf{x}})\}}, \quad \boldsymbol{\mu}_\delta \leq \frac{f(\bar{\mathbf{x}}) - \text{opt}(\boldsymbol{\delta})}{\min_i \{-g_i(\bar{\mathbf{x}}) + \delta_i\}}.$$

When  $\text{opt}(\boldsymbol{\delta}) \leq \text{opt}(\mathbf{0})$ ,

$$|\text{opt}(\boldsymbol{\delta}) - \text{opt}(\mathbf{0})| \leq \|\boldsymbol{\mu}_0\|_1 \|\boldsymbol{\delta}\|_\infty.$$

When  $\text{opt}(\boldsymbol{\delta}) > \text{opt}(\mathbf{0})$ ,

$$\begin{aligned} |\text{opt}(\boldsymbol{\delta}) - \text{opt}(\mathbf{0})| &\leq \max\{\|\boldsymbol{\mu}_0\|_1, \|\boldsymbol{\mu}_\delta\|_1\} \cdot \|\boldsymbol{\delta}\|_\infty \\ &\leq \|\boldsymbol{\mu}_0\|_1 \|\boldsymbol{\delta}\|_\infty \max\left\{1, \frac{\min_i |g_i(\bar{\mathbf{x}})|}{\min_i \{|g_i(\bar{\mathbf{x}})| + \delta_i\}}\right\}. \end{aligned}$$

Recall that  $\|\boldsymbol{\delta}\|_\infty \leq \frac{1}{2} \min_i |g_i(\bar{\mathbf{x}})|$ , then

$$|\text{opt}(\boldsymbol{\delta}) - \text{opt}(\mathbf{0})| \leq 2\|\boldsymbol{\mu}_0\|_1 \|\boldsymbol{\delta}\|_\infty.$$

Hence we obtain this lemma. □

**Lemma 6.12.** Denote by  $\mathcal{P}(\mathbf{A}, \mathbf{b})$  the following optimization problem

$$\begin{aligned} \min \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & \mathbf{A}^\top \phi(\mathbf{x}) \leq \mathbf{b}, \\ & \mathbf{x} \in \mathcal{X}, \end{aligned} \tag{6.23}$$

where  $f(\cdot)$  is a convex function,  $\mathcal{X}$  is a bounded convex set, and for  $i = 1, \dots, m$ ,  $\mathbf{a}_i$  – the  $i^{\text{th}}$  column of  $\mathbf{A}$  – satisfies that  $\mathbf{a}_i \in \mathcal{A}$  for some set  $\mathcal{A}$  such that  $\mathbf{a}^\top \phi(\mathbf{x})$  is a convex function in  $\mathbf{x}$  for any  $\mathbf{a} \in \mathcal{A}$ . Let  $\text{opt}(\mathbf{A}, \mathbf{b})$  be the optimal value of Problem (6.23). If there exists a constant  $L$  such that  $\|\phi(\mathbf{x})\|_2 \leq L$  for any  $\mathbf{x} \in \mathcal{X}$  and Problem (6.23) satisfies the Slater's condition, then there exist constants  $\rho$  and  $c$  so that for any  $\hat{\mathbf{A}}$  and  $\hat{\mathbf{b}}$  satisfying that  $\max\{\|\mathbf{a}_i - \hat{\mathbf{a}}_i\|_2, |\hat{b}_i - b_i|\} < \rho$  and  $\hat{\mathbf{a}}_i \in \mathcal{A}$  for  $i = 1, \dots, m$ , the following inequality holds

$$|\text{opt}(\hat{\mathbf{A}}, \hat{\mathbf{b}}) - \text{opt}(\mathbf{A}, \mathbf{b})| \leq c[\|\hat{\mathbf{b}} - \mathbf{b}\|_\infty + L \max_i \|\hat{\mathbf{a}}_i - \mathbf{a}_i\|_2].$$

*Proof.* By Lemma 6.10, there exists constant  $\rho_1$  such that when  $\max\{\|\mathbf{a}_i - \hat{\mathbf{a}}_i\|_2, |\hat{b}_i - b_i|\} < \rho_1$ ,  $\mathcal{P}(\hat{\mathbf{A}}, \hat{\mathbf{b}})$  is feasible and satisfies the Slater's condition. Let  $\hat{\mathbf{x}}$  be the optimal solution of  $\mathcal{P}(\hat{\mathbf{A}}, \hat{\mathbf{b}})$ . Since

$$\mathbf{A}^\top \phi(\hat{\mathbf{x}}) \leq \mathbf{b} + (\hat{\mathbf{b}} - \mathbf{b} - (\hat{\mathbf{A}} - \mathbf{A})^\top \phi(\hat{\mathbf{x}})),$$

by Lemma 6.11, we know that there exists constants  $\rho_2$  and  $c$  such that when  $\|\hat{\mathbf{b}} - \mathbf{b}\|_\infty + \|(\hat{\mathbf{A}} - \mathbf{A})^\top \phi(\hat{\mathbf{x}})\|_\infty < \rho_2$ , the following inequality holds

$$|\text{opt}(\hat{\mathbf{A}}, \hat{\mathbf{b}}) - \text{opt}(\mathbf{A}, \mathbf{b})| \leq c[\|\hat{\mathbf{b}} - \mathbf{b}\|_\infty + \|(\hat{\mathbf{A}} - \mathbf{A})^\top \phi(\hat{\mathbf{x}})\|_\infty].$$

By the assumption that there exists a constant  $L$  such that  $\|\phi(\mathbf{x})\|_2 \leq L$  for any  $\mathbf{x} \in \mathcal{X}$ , we have

$$\|(\hat{\mathbf{A}} - \mathbf{A})^\top \phi(\hat{\mathbf{x}})\|_\infty \leq \max_i \{\|\hat{\mathbf{a}}_i - \mathbf{a}_i\|_2 \|\phi(\hat{\mathbf{x}})\|_2\} \leq \max_i \{L \|\hat{\mathbf{a}}_i - \mathbf{a}_i\|_2\}.$$

Then by taking  $\rho = \min\{\rho_1, \frac{\rho_2}{L+1}\}$ , we obtain this lemma.  $\square$

### Proof of Theorem 6.4

We use the same notations as that in the proof of Theorem 6.1. Equation (6.9) shows that for any  $\bar{\mathbf{x}}$ ,

$$|f(\bar{\mathbf{a}}_i(t)^\top \bar{\mathbf{x}}) - f(\bar{\mathbf{a}}_i^\top \bar{\mathbf{x}})| \leq \frac{2l_\mu}{c_\mu} \|\bar{\mathbf{x}}\|_{\mathbf{M}_t^{-1}} \left\| \sum_{k=1}^{t-1} \xi_i(k) \mathbf{y}(k) \right\|_{\mathbf{M}_t^{-1}}.$$

Since  $\mathbf{M}_t \succeq \mathbf{M}_{d+1} \succ 0$ ,

$$\left\| \sum_{k=1}^{t-1} \xi_i(k) \mathbf{y}(k) \right\|_{\mathbf{M}_t^{-1}} \leq \left\| \sum_{k=1}^{d+1} \xi_i(k) \mathbf{y}(k) \right\|_{\mathbf{M}_{d+1}^{-1}} + \left\| \sum_{k=d+2}^{t-1} \xi_i(k) \mathbf{y}(k) \right\|_{\mathbf{M}_t^{-1}}.$$

Recall that  $\lambda_{\min}(\mathbf{M}_{d+1}) \geq c_0 > 0$  and  $\{\mathbf{y}(1), \dots, \mathbf{y}(d+1)\}$  are fixed conditioned on  $\{\mathbf{x}(1), \dots, \mathbf{x}(d+1)\}$ .

Since  $\|\mathbf{y}(k)\|_2 \leq \sqrt{L^2 + 1}$ ,  $|\xi_i(k)| \leq R$  and

$$\left\| \sum_{k=1}^{d+1} \xi_i(k) \mathbf{y}(k) \right\|_{\mathbf{M}_{d+1}^{-1}} = \left\| \sum_{k=1}^{d+1} \xi_i(k) \mathbf{M}_{d+1}^{-\frac{1}{2}} \mathbf{y}(k) \right\|_2,$$

by the matrix Bernstein inequality, one can easily verify that for  $0 < \delta < 1$  there exists a

constant  $c$  such that

$$\left\| \sum_{k=1}^{d+1} \xi_i(k) \mathbf{y}(k) \right\|_{\mathbf{M}_{d+1}^{-1}} \leq cR \sqrt{d \log \frac{md}{\delta}}$$

holds with probability at least  $1 - \frac{\delta}{m}$ .

By Lemma 6.6 and Lemma 6.7, we have that for any  $\delta > 0$ , with probability at least  $1 - \frac{\delta}{m}$ ,

$$\left\| \sum_{k=d+2}^{t-1} \xi_i(k) \mathbf{y}(k) \right\|_{\mathbf{M}_t^{-1}}^2 \leq 2R^2 \left[ \log \frac{m}{\delta} + d \log c_0 + d \log \left( L^2 + 1 + \frac{t(L^2 + 1)}{d+1} \right) \right].$$

Therefore,

$$\begin{aligned} \left\| \sum_{k=1}^{t-1} \xi_i(k) \mathbf{y}(k) \right\|_{\mathbf{M}_t^{-1}} &\leq cR \sqrt{d \log \frac{md}{\delta} + d \log \left( L^2 + 1 + \frac{t(L^2 + 1)}{d+1} \right)} \\ &\leq cR \sqrt{d \log \frac{md}{\delta} + d \log \frac{2t(L^2 + 1)}{d}} \\ &= cR \sqrt{d \log \frac{2mt(L^2 + 1)}{\delta}}. \end{aligned}$$

Then let  $\theta(t) = \frac{cl_\mu R}{c_\mu} \sqrt{d \log \frac{2mt(L^2 + 1)}{\delta}}$ , we have

$$|f(\bar{\mathbf{a}}_i(t)^\top \bar{\mathbf{x}}) - f(\bar{\mathbf{a}}_i^\top \bar{\mathbf{x}})| \leq \theta(t) \|\bar{\mathbf{x}}\|_{\mathbf{M}_t^{-1}}.$$

By following the proof of Theorem 6.2, we know that

$$\begin{aligned} \text{Regret}(T) &= \sum_{t=1}^{d+1} \mathbf{c}^\top \mathbf{x}(t) - \mathbf{c}^\top \mathbf{x}^*(t) + \sum_{t=d+2}^T \mathbf{c}^\top \mathbf{x}(t) - \mathbf{c}^\top \mathbf{x}^*(t) \\ &\leq (d+1) \|\mathbf{c}\|_2 \|\mathbf{x}(t) - \mathbf{x}^*(t)\|_2 \leq 2(d+1)L \|\mathbf{c}\|_2, \end{aligned}$$

and

$$\text{Violation}_i(T) \leq (d+1)(L \|\mathbf{a}_i\|_2 + |b_i|) + \frac{20\theta(T)}{c_\mu} \sqrt{dT \log T}$$

hold with probability at least  $1 - \delta$ .

**Proofs of Theorems 6.5, 6.6 and 6.7**

These theorems can be easily proved by applying Lemmas 6.9-6.12 and following the same procedures of the proofs of Theorem 6.1, Theorem 6.2 and Theorem 6.4.

**6.9 Chapter Summary**

We proposed two algorithms LPUC-ED and LPUC-UCB to solve the online linear programming problem with unobserved constraints which generalized the stochastic linear optimization problem studied by [DHK08]. Both of the algorithms have sublinear bounds on the regret and the constraint violation. The numerical experiments demonstrated their good empirical performance and validated our theoretical results. For future work, we will try to develop algorithms that achieve  $O(\sqrt{T} \log T)$  bounds and are computationally efficient.

# CHAPTER 7

## A Unified Framework for Outlier-Robust PCA-like Algorithms

We propose a unified framework for making a wide range of PCA-like algorithms – including the standard PCA, sparse PCA and non-negative sparse PCA, etc. – robust when facing a constant fraction of arbitrarily corrupted outliers. Our analysis establishes solid performance guarantees of the proposed framework: its estimation error is upper bounded by a term depending on the intrinsic parameters of the data model, the selected PCA-like algorithm and the fraction of outliers. Our experiments on synthetic and real-world datasets demonstrate that the outlier-robust PCA-like algorithms derived from our framework have outstanding performance.

### 7.1 Introduction

Principal component analysis (PCA) [Pea01], arguably the most widely applied dimension reduction method, plays a significant role in data analysis in a broad range of areas including machine learning, statistics, finance, biostatistics and many others. The standard PCA performs the spectral decomposition of the sample covariance matrix, selects the eigenvectors corresponding to the largest eigenvalues, and then constructs a low dimensional subspace based on the selected eigenvectors. It is well known that standard PCA, depending on different applications, may suffer from three weaknesses [MR14, XCM13, JL09]: 1) PCA is notoriously fragile to outliers – indeed, its performance can significantly degrade in the presence of even few corrupted samples, due to the quadratic error criterion used; 2) PCA

cannot utilize additional information of the principal components: e.g., in certain applications, it is known that the principal components should lie in the positive orthant; 3) its output may lack interpretability since it does not encourage sparse solutions.

Many efforts have been made to mitigate these weaknesses of PCA. In recent years, numerous robust PCA algorithms have been proposed to address the first issue [DGK81, XY95, YW99, ITB03, Das03, XCM13, FXY12]. Among them, [XCM13] successfully tackles the case where a *constant fraction* of samples are corrupted in the *high dimensional regime*. Their proposed method, termed HR-PCA (which stands for High-dimensional Robust PCA), is tractable, easily kernelizable, and is able to robustly estimate the principal components even in the face of a constant fraction of outliers and very low signal-to-noise ratio. To overcome the computational issue of HR-PCA, Feng *et al.* [FXY12] proposed a deterministic approach (DHR-PCA) that dramatically reduces the computational work. However, neither HR-PCA nor DHR-PCA deals with the last two weaknesses mentioned above.

To address the second weakness, [MR14] recently proposed a new algorithm called *non-negative PCA* which handles the case that the principal components are known to lie in the positive orthant, and showed that near-optimal non-negative principal components can be extracted in nearly linear time. But similar to the standard PCA, this algorithm is sensitive to outliers. Indeed, the estimated principal components can be far from the true ones in the face of even few outliers.

To address the third weakness, previous works focus on a class of methods called sparse PCA that adapt the standard PCA so that only a few of attributes of the resulting principle components are non-zero, e.g., [VCLR13, ZHT06, SH08, JYN08, BJNP13, VL13, dEJL07, TDT10]. Some of these methods are based on non-convex optimization formulations [JTU03, MWA05] while others use  $\ell_1$ -norm regularization [ZHT06]. Recently, Vu *et al.* [VCLR13] proposed FPS – a convex relaxation formulation of sparse principal subspace estimation based on a semi-definite program with a *Fantope* constraint and established theoretical guarantees in the outlier-free regime. Yet, one severe drawback of most sparse PCA algorithms is that the output can be sensitive to the existence of even few outliers. This is clearly undesirable, as in real-world applications, the existence of outliers is ubiquitous.



Recently, several robust sparse PCA have been proposed [CFF13, WC12, HRS14] to handle outliers, but all of them are only evaluated by experiments and have no theoretical performance guarantees.

This chapter is the first attempt to theoretically address these issues of PCA simultaneously. In specific, we propose a general framework for a wide range of PCA-like algorithms to make them provably *robust to a constant fraction of arbitrary* outliers. Our framework is inspired by HR-PCA [XCM13, FXY12], but overcomes the drawbacks of HR-PCA and has the capability of converting a non-robust PCA-like algorithm such as non-negative PCA [MR14], sparse PCA [VCLR13, PDK13] or non-negative sparse PCA [APD14], into its outlier-robust variant.

The analysis of our proposed framework is novel and different from that of HR-PCA. We analyze its performance using two performance metrics: the subspace distance and the expressed variance. We show that the subspace distance between its estimated principal components and the ground-truth under the spiked model can be upper bounded by a term depending on the parameters of the spike model, the selected PCA-like algorithm and the fraction of outliers. The analysis of subspace distance in the presence of outliers is new to the best of our knowledge. Moreover, while the analysis of expressed variance for HR-PCA exists in literature, our analysis of the expressed variance of this framework is more general, in that it shows that maximal robustness can be achieved for a wide range of PCA-like algorithms besides HR-PCA. Our numerical experiments results show that when outliers exist, the outlier-robust PCA-like algorithms developed from our framework outperform their non-robust counterparts considerably.

**Notation.** We use lower-case boldface letters to denote column vectors and upper-case boldface letters to denote matrices. In this chapter,  $\|\mathbf{M}\|_2$  is the spectral norm,  $\|\mathbf{M}\|_*$  is the nuclear norm,  $\|\mathbf{M}\|_1$  is the element-wise  $\ell_1$  norm,  $\|\mathbf{M}\|_\infty$  is the element-wise infinity norm, and  $\|\mathbf{M}\|_F$  is the Frobenius norm. We use  $\|\mathbf{M}\|_0$  to denote the number of non-zero entries in  $\mathbf{M}$ , and use subscript  $(\cdot)$  to represent order statistics of a random variable. For example, let  $v_1, \dots, v_n \in \mathbb{R}$ , then  $v_{(1)}, \dots, v_{(n)}$  is a permutation of  $v_1, \dots, v_n$  in a non-decreasing order. For matrix  $\mathbf{X}$ , the first  $k$  singular values of  $\mathbf{X}$  are denoted by  $\lambda_1(\mathbf{X}), \dots, \lambda_k(\mathbf{X})$ .

## 7.2 Unified Framework for Outlier-Robust PCA

In this section, we present our framework for outlier-robust PCA-like algorithms. We first describe the problem setup and necessary assumptions, and then show the details of the algorithm along with the key intuition underlying it.

### 7.2.1 Problem Setup

Suppose there are  $n$  samples  $\mathcal{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_n \in \mathbb{R}^p\}$  which consist of  $t$  authentic samples  $\mathbf{z}_1, \dots, \mathbf{z}_t \in \mathbb{R}^p$  and  $n - t$  outliers  $\mathbf{o}_1, \dots, \mathbf{o}_{n-t} \in \mathbb{R}^p$ . The outliers are *arbitrary*. We denote the fraction of outliers by  $\rho = (n - t)/n$  and assume that  $\rho < 0.5$ . The authentic samples  $\mathbf{z}_i$  are generated according to  $\mathbf{z}_i = \mathbf{A}\mathbf{x}_i + \mathbf{n}_i$  where  $\mathbf{x}_i \in \mathbb{R}^d$  are i.i.d. samples of a random vector  $\mathbf{x}$  with mean 0 and variance  $\mathbf{I}_d$  and  $\mathbf{n}_i$  are independent realizations of standard Gaussian  $\mathcal{N}(0, \mathbf{I}_p)$ . The matrix  $\mathbf{A} \in \mathbb{R}^{p \times d}$  and the distribution of  $\mathbf{x}$  (denoted by  $\nu$ ) are unknown. The covariance of  $\mathbf{z}$  is denoted by  $\mathbf{\Sigma}$ . Since  $\mathbf{z} = \mathbf{A}\mathbf{x} + \mathbf{n}$ ,  $\mathbf{\Sigma} = \mathbb{E}[\mathbf{z}\mathbf{z}^\top] = \mathbf{A}\mathbf{A}^\top + \mathbf{I}_p$ . We denote the one-dimensional marginal of  $\nu$  along direction  $\mathbf{v} \in \mathcal{S}_d$  by  $\bar{\nu}_{\mathbf{v}}$ , and assume that  $\bar{\nu}_{\mathbf{v}}(\{0\}) < 0.5$  for all  $\mathbf{v} \in \mathcal{S}_d$  and it is sub-Gaussian, i.e., there exists  $\theta > 0$  such that  $\bar{\nu}_{\mathbf{v}}((-\infty, x] \cup [x, +\infty)) \leq \exp(1 - x^2/\theta)$  for all  $x > 0$ . Clearly, both assumptions are satisfied if  $\nu$  is Gaussian.

We make the following two assumptions: 1)  $\mathbf{A}$  is full row rank and  $n > d$ . This essentially means the intrinsic dimension of the authentic samples (ignoring the noise) is indeed  $d$ . 2) The projection  $\mathbf{\Pi}(k) = \mathbf{U}(k)\mathbf{U}(k)^\top$  onto the subspace spanned by the eigenvectors  $\mathbf{U}(k)$  of  $\mathbf{\Sigma}$  corresponding to its  $k$  largest eigenvalues satisfies  $\|\mathbf{\Pi}(k)\|_0 \leq \beta^2$ , where  $\|\mathbf{\Pi}(k)\|_0$  is the number of nonzero entries of  $\mathbf{\Pi}(k)$ . Our goal is to approximately recover  $\mathbf{\Pi}(k)$  even though the samples contain a non-negligible fraction of arbitrary outliers. For convenience, we let  $\mathbf{\Pi} \triangleq \mathbf{\Pi}(k)$  in the following sections. In the followings, “with high probability” means with probability at least  $1 - c \max\{p^{-10}, n^{-10}\}$  for some constant  $c$ .

### 7.2.2 General Formulation of PCA-like Algorithms

Many kinds of PCA-like algorithms have been proposed in recent decades, e.g., sparse PCA [ZHT06, PDK13], non-negative PCA [MR14], etc., which play a significant role in machine learning, computer vision, statistics and data analysis. In this section, we consider a general formulation as shown below for a wide range of these algorithms:

$$\max_{\mathbf{X} \in \mathcal{C}} \langle \hat{\Sigma}, \mathbf{X} \rangle - \mu \|\mathbf{X}\|_1, \quad (7.1)$$

where  $\hat{\Sigma}$  is the empirical sample covariance matrix,  $\mathcal{C}$  includes the constraints imposed on  $\mathbf{X}$ , and  $\mu$  is the weight of the regularization term. Typically,  $\mu$  is less than a certain universal constant. To see that this formulation can model most PCA-like algorithms proposed in literature, let  $k$  be the number of the principal components one wants to extract and  $\mathcal{F}(k)$  be the set  $\{\mathbf{X} : 0 \preceq \mathbf{X} \preceq \mathbf{I}_p, \text{tr}(\mathbf{X}) = k\}$  which includes the matrices that lie in the convex hull of all feasible projection matrices. Thus, the following algorithms are all equivalent to Formulation (7.1) for appropriate  $k$ ,  $\mathcal{C}$  and  $\mu$ :

1. Standard PCA [Pea01]:  $k = d$ ,  $\mathcal{C} = \mathcal{F}(k)$  and  $\mu = 0$ ;
2. Non-negative PCA [MR14]:  $k = 1$ ,  $\mathcal{C} = \{\mathbf{u}\mathbf{u}^\top : \|\mathbf{u}\|_2 \leq 1, \mathbf{u} \geq \mathbf{0}\}$  and  $\mu = 0$ ;
3. Sparse PCA [PDK13]:  $k = 1$ ,  $\mathcal{C} = \{\mathbf{u}\mathbf{u}^\top : \|\mathbf{u}\|_0 \leq \gamma, \|\mathbf{u}\|_2 \leq 1\}$  and  $\mu = 0$ ;
4. Fantope projection and selection (FPS) [VCLR13]:  $k = d$ ,  $\mathcal{C} = \mathcal{F}(k)$  and  $\mu \asymp \sqrt{\frac{\log p}{n}}$ ;
5. Non-negative sparse PCA [APD14]:  $k = 1$ ,  $\mathcal{C} = \{\mathbf{u}\mathbf{u}^\top : \|\mathbf{u}\|_0 \leq \gamma, \|\mathbf{u}\|_2 \leq 1, \mathbf{u} \geq \mathbf{0}\}$  and  $\mu = 0$ ;
6. Large-scale sparse PCA [ZE11]:  $k = 1$ ,  $\mathcal{C} = \{\mathbf{X} : \mathbf{X} \succeq \mathbf{0}, \text{tr}(\mathbf{X}) = 1\}$ ,  $\mu > 0$ .

Since the feasible set  $\mathcal{C}$  in (7.1) may be non-convex, the global optimum of (7.1) may not be achievable. Therefore, there are two important issues: 1) whether a PCA-like algorithm can probably find an optimal or near-optimal solution of (7.1), and 2) whether its solution converges to the ground truth. We call the PCA-like algorithms that can find optimal or near-optimal solutions of (7.1) “workable” algorithms, formally defined as:

**Definition 7.1.** A PCA-like algorithm is  $(\eta, \gamma)$ -workable if there exist positive numbers

$\eta \leq 1$  and  $\gamma \leq p$  such that with high probability its output  $\hat{\mathbf{X}}$  satisfies  $\|\hat{\mathbf{X}}\|_0 \leq \gamma^2$  and

$$\langle \hat{\Sigma}, \hat{\mathbf{X}} \rangle - \mu \|\hat{\mathbf{X}}\|_1 \geq (1 - \eta) \left[ \langle \hat{\Sigma}, \mathbf{\Pi} \rangle - \mu \|\mathbf{\Pi}\|_1 \right].$$

Note that  $\eta$  indicates the accuracy of the solution  $\hat{\mathbf{X}}$ , e.g.,  $\eta = 0$  means  $\hat{\mathbf{X}}$  is optimal, while  $\eta = 0.5$  means the cost value corresponding to  $\hat{\mathbf{X}}$  is half of the optimum. Parameter  $\gamma$  bounds the sparsity of  $\hat{\mathbf{X}}$ . For the first five algorithms mentioned above, previous works have proved that all of these algorithms are workable. In particular,  $\eta = 0, \gamma = p$  for standard PCA and FPS,  $0 < \eta < 1, \gamma = p$  for non-negative PCA, and  $0 < \eta < 1, \gamma \ll p$  for sparse PCA and non-negative sparse PCA. For large-scale sparse PCA, no performance guarantees are known, but our experiments show that this algorithm can still be put into our framework to achieve robustness.

### 7.2.3 Outlier-Robust PCA-like Algorithms

Our framework is inspired by HR-PCA [XCM13]. Therefore, before presenting its details, we briefly explain the intuition behind HR-PCA. HR-PCA iteratively performs PCA to compute principal components (PCs) and then randomly removes one point with a probability proportional to its magnitude after projected on the found PCs. HR-PCA works for the following intuitive reasons. In each iteration, a PC is computed either due to true samples which implies it is a “good” direction; or due to large outliers in which case the random removal scheme will remove an outlier with high probability. Thus, for at least one iteration, the algorithm will find a good direction, say  $\mathbf{w}^t$ . Among all the directions found in the algorithm, the final output of HR-PCA is the one with the largest *Robust Variance Estimator* (RVE). RVE measures the projection variance of the  $(n - \hat{t})$ -smallest points: A large RVE means that many of the points have a large variance in this direction, while a small RVE indicates otherwise. This makes sure that the final output is close to  $\mathbf{w}^t$ , and hence a good direction. A variant of HR-PCA is called deterministic HR-PCA or DHR-PCA [FXY12]. Instead of removing one point, DHR-PCA decreases the weights of all samples according to their magnitudes after projected on the found PCs in each iteration to reduce computational cost.

**Algorithm 7.1:** Outlier-robust PCA-like algorithm

- 
- Input** : Contaminated sample-set  $\mathcal{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_n\}$  and parameters  $k, T, \hat{t}, \mu$ .
- Output:** The estimated principal components.
- 1 Initialize:  $s = 0, \text{Opt} = 0$ ;  $\hat{\mathbf{y}}_i = \mathbf{y}_i$  and  $\alpha_i = 1$  for  $i = 1, \dots, n$ ;
  - 2 **for**  $s = 1$  to  $T$  **do**
  - 3     Compute the weighted empirical covariance matrix  $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n \alpha_i \hat{\mathbf{y}}_i \hat{\mathbf{y}}_i^\top$ ;
  - 4     Solve the PCA-like problem 7.1 and denote the output by  $\hat{\mathbf{X}}$ ;
  - 5     If  $\bar{V}_{\hat{t}}(\hat{\mathbf{X}}) > \text{Opt}$ , let  $\text{Opt} = \bar{V}_{\hat{t}}(\hat{\mathbf{X}})$  and  $\mathbf{X}^* = \hat{\mathbf{X}}$ , where  

$$\bar{V}_{\hat{t}}(\hat{\mathbf{X}}) \triangleq \frac{1}{\hat{t}} \sum_{i=1}^{\hat{t}} \langle \mathbf{y} \mathbf{y}^\top, \hat{\mathbf{X}} \rangle_{(i)}$$
;
  - 6     Update  $\alpha_i = \left(1 - \frac{\langle \mathbf{y}_i \mathbf{y}_i^\top, \hat{\mathbf{X}} \rangle}{\max_{\{i | \alpha_i \neq 0\}} \langle \mathbf{y}_i \mathbf{y}_i^\top, \hat{\mathbf{X}} \rangle}\right) \alpha_i$ ;
  - 7 **end**
  - 8 Perform SVD on  $\mathbf{X}^*$  and denote the top  $k$  eigenvectors by  $\mathbf{w}_1^*, \dots, \mathbf{w}_k^*$ ;
  - 9 Return  $\mathbf{w}_1^*, \dots, \mathbf{w}_k^*$  and  $\mathbf{X}^*$ .
- 

HR-PCA and DHR-PCA only focus on making standard PCA robust to outliers but say nothing about whether it is possible to improve the robustness of non-negative PCA or sparse PCA. In this chapter, we propose a more general framework as shown in Algorithm 7.1 for developing outlier-robust PCA-like algorithms. In Algorithm 7.1, the weighted covariance matrix acts as a robust covariance estimator [Rou85, RD98, CH00], and  $\bar{V}_{\hat{t}}(\mathbf{X})$  is the *Robust Variance Estimator* which is defined as  $\bar{V}_{\hat{t}}(\mathbf{X}) \triangleq \frac{1}{\hat{t}} \sum_{i=1}^{\hat{t}} \langle \mathbf{y} \mathbf{y}^\top, \mathbf{X} \rangle_{(i)}$ , where  $\mathbf{y} \in \mathcal{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_n\}$ . Intuitively, the term  $\langle \mathbf{y} \mathbf{y}^\top, \mathbf{X} \rangle$  imitates the magnitude of  $\mathbf{y}$  after it is projected on the column subspace of  $\mathbf{X}$ , so this RVE measures the projection variance similar to the one in HR-PCA. As we show below, a PCA-like algorithm becomes outlier-robust if it is integrated into this general robustness framework. For example, DHR-PCA can be easily deduced from this framework by solving the standard PCA in Step 3.

### 7.3 Theoretical Guarantees

We now present the performance guarantees of Algorithm 7.1 with a  $(\eta, \gamma)$ -workable PCA-like algorithm. Typically, there are two ways to measure the performance of PCA-like algorithms [XCM13, VCLR13]. The first one, termed the *subspace distance* (S.D.), measures the distance between the subspace spanned by the estimated PCs and the subspace spanned by the true PCs. The second one, termed the *expressed variance* (E.V.), measures the portion of the signal  $\mathbf{A}\mathbf{x}$  being expressed by the estimated principle components. Formally, we have:

**Definition 7.2.** Let  $\mathbf{M}_1, \mathbf{M}_2$  be two symmetric matrices and  $\mathcal{M}_1, \mathcal{M}_2$  be their respective  $k$ -dimensional principal subspaces, then the subspace distance is  $S.D. \triangleq \sin \Theta(\mathcal{M}_1, \mathcal{M}_2)$ .

**Definition 7.3.** The expressed variance of  $\mathbf{w}_1, \dots, \mathbf{w}_k$  is defined as  $E.V. \triangleq \frac{\sum_{i=1}^k \mathbf{w}_i^\top \mathbf{A} \mathbf{A}^\top \mathbf{w}_i}{\sum_{i=1}^k \lambda_i(\mathbf{A} \mathbf{A}^\top)}$ .

Notice that a smaller S.D. or a larger E.V. indicates a more desirable solution. Also,  $S.D. \geq 0$  and  $E.V. \leq 1$  with equality achieved when the vectors  $\mathbf{w}_1, \dots, \mathbf{w}_k$  span the same space as the true PCs. Thus, to provide performance guarantees of the proposed algorithms, we lower bound the expressed variance as well as upper bound the subspace distance for the output. This is different from [XCM13] and [FXY12] which only analyzed the expressed variance (of HR-PCA and DHR-PCA respectively).

To analyze the performance of Algorithm 7.1, the following ‘‘tail weight’’ function the first appeared in [XCM13] is required.

**Definition 7.4.** ([XCM13]) For any  $\gamma \in [0, 1]$  and  $\mathbf{v} \in \mathcal{S}_d$ , let  $\delta_\gamma \triangleq \min\{\delta \geq 0 \mid \bar{\nu}_\mathbf{v}([-\delta, \delta]) \geq \gamma\}$  and  $\gamma_\nu^- = \bar{\nu}_\mathbf{v}((-\delta, \delta))$ . Then the ‘‘tail weight’’ functions  $\mathcal{V}_\mathbf{v}$  is defined as follows:

$$\mathcal{V}_\mathbf{v}(\gamma) \triangleq \lim_{\epsilon \downarrow 0} \int_{-\delta_\gamma + \epsilon}^{\delta_\gamma - \epsilon} x^2 \bar{\nu}_\mathbf{v}(dx) + (\gamma - \gamma_\nu^-) \delta_\gamma^2.$$

We define  $\mathcal{V}^+(\gamma) \triangleq \sup_{\mathbf{v} \in \mathcal{S}_d} \mathcal{V}_\mathbf{v}(\gamma)$  and  $\mathcal{V}^-(\gamma) \triangleq \inf_{\mathbf{v} \in \mathcal{S}_d} \mathcal{V}_\mathbf{v}(\gamma)$ . In the following subsections, we assume that the feasible set  $\mathcal{C}$  in (7.1) is a subset of  $\mathcal{F}(k)$  – the convex hull of all the feasible projection matrices. This is not a restrictive condition. Indeed all the algorithms listed in Section 7.2.2 except large-scale sparse PCA meet this condition.

### 7.3.1 Upper Bound of Subspace Distance

We first bound the subspace distance for Algorithm 7.1. The following lemma relates the subspace distance with the Frobenius norm of  $\mathbf{X}^* - \mathbf{\Pi}$  so that we only need to bound  $\|\mathbf{X}^* - \mathbf{\Pi}\|_F$ .

**Lemma 7.1.** [VCLR13] *If  $\mathcal{M}$  is the principal  $d$ -dimensional subspace of  $\Sigma$  and  $\mathcal{M}^*$  is the principal  $k$ -dimensional subspace of  $\mathbf{X}^*$ , then*

$$\sin \Theta(\mathcal{M}, \mathcal{M}^*) \leq \sqrt{2} \|\mathbf{X}^* - \mathbf{\Pi}\|_F.$$

In the following parts, we let  $\delta_k(\mathbf{A}\mathbf{A}^\top) \triangleq \lambda_k(\mathbf{A})^2 - \lambda_{k+1}(\mathbf{A})^2$  and let

$$f(B) = \min \left\{ 2B\|\mathbf{A}\|_2^2 + c_1\tau, \gamma B\|\mathbf{A}\|_2^2 + c_2\gamma(d\|\mathbf{A}\|_2 + 1) \right\},$$

where  $\tau = \max\{\frac{p}{n}, 1\}$  and  $c$  is a universal constant. Notice that  $f(B)$  is upper bounded by  $2B\|\mathbf{A}\|_2^2 + c_1$  when  $p = O(n)$  and by  $\gamma B\|\mathbf{A}\|_2^2 + c_2\gamma(d\|\mathbf{A}\|_2 + 1)$  when  $p = \Omega(n)$  and  $\gamma \ll p$  for some constants  $c_1$  and  $c_2$ . Therefore, in the high dimensional case where  $p \gg n$ , when sparse PCA algorithms are applied, i.e.,  $\gamma \ll p$ ,  $f(B)$  can still be small, compared with  $\frac{p}{n}$ .

We now provide our first main theorem which states that the output  $\hat{\mathbf{X}}$  in Step 3 will be close to the true projection matrix after a certain number of iterations.

**Theorem 7.1.** *Suppose that  $\rho < 0.5$  and  $\log p \leq n$ , then there exists a finite number  $s \leq n$  such that the output  $\mathbf{X}_s$  of the PCA-like algorithm in the  $s^{\text{th}}$  stage satisfies the following inequality with high probability,*

$$\|\mathbf{X}_s - \mathbf{\Pi}\|_F \leq R(\mu) + \sqrt{k} \min_{1 \leq \kappa > 2\rho} \sqrt{\frac{f(B_1) + \eta\beta B_0}{\delta_k(\mathbf{A}\mathbf{A}^\top)}}, \quad (7.2)$$

where

$$R(\mu) \triangleq \begin{cases} \frac{8(\gamma[\epsilon_0(\|\mathbf{A}\|_2^2 + 1) - \mu]_+ + \mu\beta)}{\delta_k(\mathbf{A}\mathbf{A}^\top)}, & \mu \neq 0 \\ \min \left\{ \frac{8\epsilon_0\gamma(\|\mathbf{A}\|_2^2 + 1)}{\delta_k(\mathbf{A}\mathbf{A}^\top)}, 2\sqrt{\frac{\epsilon_1 k(\|\mathbf{A}\|_2^2 + 1)}{\delta_k(\mathbf{A}\mathbf{A}^\top)}} \right\}, & \mu = 0, \end{cases}$$

$$\epsilon_0 = c_0 \sqrt{\frac{\log p}{n}}, \epsilon_1 = c_1 \sqrt{\frac{p}{n}}, B_0 = c_2(\|\mathbf{A}\|_2^2 + 1), B_1 = \kappa + 1 - \mathcal{V}^-(1 - \frac{\rho}{\kappa(1-\rho)}) + \epsilon_0 + c_3 \left( \frac{d \log^3 n}{n} \right)^{\frac{1}{4}},$$

and  $c_0, c_1, c_2, c_3$  are universal constants.

**Remark.** The upper bound of  $\|\mathbf{X}_s - \mathbf{\Pi}\|_F$  involves three terms: 1)  $R(\mu)$ :  $R(\mu)$  is related to the weight of the regularization term in (7.1). A positive  $\mu$  can encourage sparse solutions. From the formulation of  $R(\mu)$ , we know that setting  $\mu$  to  $\epsilon_0(\|\mathbf{A}\|_2^2 + 1)$  when  $\mu$  is non-zero leads to a tighter bound. 2)  $f(B_1)$ :  $B_1$  involves  $\rho$  – the fraction of outliers, and decreases when  $\rho$  decreases. Clearly,  $B_1 \rightarrow 0$  when  $\rho, \frac{d \log^3 n}{n}$  and  $\frac{\log p}{n}$  converge to zero. 3)  $\eta \beta B_0$ : This term contains  $\eta$ , i.e., the accuracy of the selected PCA-like algorithm. When the optimal solution of (7.1) can be achieved, this term becomes zero.

Theorem 7.1 tells us that a good solution  $\mathbf{X}_s$  can be generated for some iteration  $s$ . However, such  $s$  is not specified. Thus, one can not take  $\mathbf{X}_s$  as the output; instead, one can choose a solution that is close to  $\mathbf{X}_s$  as the output. In Algorithm 7.1, the solution with the maximal RVE is selected as the final output  $\mathbf{X}^*$ . Other methods can also be applied in practical applications based on specific information. The following theorem provides the estimation error of  $\mathbf{X}^*$ .

**Theorem 7.2.** *Suppose that  $\rho < 0.5$  and  $\log p \leq n$ , the following holds with high probability,*

$$\|\mathbf{X}^* - \mathbf{\Pi}\|_F \leq \sqrt{\frac{2[(dB_2 + kB_4)\lambda_1(\mathbf{A})^2 + kf(B_3)]}{\delta_k(\mathbf{A}\mathbf{A}^\top)}}, \quad (7.3)$$

where  $B_2$  is the right hand side of (7.2),

$$B_3 = 2 - \mathcal{V}^-\left(\frac{\hat{t}}{t}\right) - \mathcal{V}^-\left(\frac{\hat{t} - \rho n}{t}\right) + c_0 \left(\frac{d \log^3 n}{n}\right)^{\frac{1}{4}},$$

$B_4 = \min\{c_1 \sqrt{\frac{p}{n}}, c_2 \gamma \sqrt{\frac{\log p}{n}}\}$ , and  $c_0, c_1, c_2$  are universal constants.

**Remark.** This upper bound contains three terms: 1)  $B_2$  is the upper bound of  $\|\mathbf{X}_s - \mathbf{\Pi}\|_F$  as shown in Theorem 7.1. 2)  $B_3$  involves  $\rho$  and parameter  $\hat{t}$ , which becomes small when  $\rho$  decreases and  $\hat{t}$  approaches  $t$ . 3)  $B_4$  converges to zero as  $\frac{p}{n} \rightarrow 0$  or  $\gamma \sqrt{\frac{\log p}{n}} \rightarrow 0$ . To achieve consistency, one should ensure that  $\frac{p}{n} \rightarrow 0$  for the standard PCA where  $\gamma = p$ , and  $\frac{\gamma^2 \log p}{n} \rightarrow 0$  for sparse PCA where  $\gamma \ll p$ .

The following corollaries provide more interpretable bounds of the subspace distance for the



standard PCA, FPS and sparse PCA discussed in Section 7.2.2.

**Corollary 7.1.** *Suppose that  $\rho < 0.5$  and  $\log p \leq n$ , then when the PCA-like algorithm is the standard PCA [Pea01], the following holds with high probability,*

$$\|\mathbf{X}^* - \mathbf{\Pi}\|_F \leq \sqrt{\frac{4d(B_2 + B_3 + \epsilon)\lambda_1(\mathbf{A})^2 + cd\tau}{\lambda_d(\mathbf{A})^2}}, \quad (7.4)$$

where

$$B_2 = 2\sqrt{\frac{\epsilon d(\lambda_1(\mathbf{A})^2 + 1)}{\lambda_d(\mathbf{A})^2}} + \min_{1 \geq \kappa > 2\rho} \sqrt{\frac{2dB_1\lambda_1(\mathbf{A})^2 + cd\tau}{\lambda_d(\mathbf{A})^2}},$$

$B_1$  is defined in Theorem 7.1,  $B_3$  is defined in Theorem 7.2,  $\epsilon = c_1\sqrt{\frac{p}{n}}$ ,  $\tau = \max\{\frac{p}{n}, 1\}$  and  $c, c_0, c_1$  are universal constants.

The standard PCA imposes no constraint on the sparsity of its solution, so when the ambient dimension  $p$  grows faster than the sample number  $n$ , the bound in Corollary 7.1 will go to infinity. One way to encourage sparsity is to impose a “soft” constraint which upper bounds the  $l_1$ -norm of the solution, e.g., FPS.

**Corollary 7.2.** *Suppose that  $\rho < 0.5$  and  $\log p \leq n$ , then when the PCA-like algorithm is FPS [VCLR13], the following holds with high probability,*

$$\|\mathbf{X}^* - \mathbf{\Pi}\|_F \leq \sqrt{\frac{4d(B_2 + B_3 + \epsilon_1)\lambda_1(\mathbf{A})^2 + cd\tau}{\lambda_d(\mathbf{A})^2}}, \quad (7.5)$$

where

$$B_2 = \frac{\epsilon_0(\lambda_1(\mathbf{A})^2 + 1)\beta}{\lambda_d(\mathbf{A})^2} + \min_{1 \geq \kappa > 2\rho} \sqrt{\frac{2dB_1\lambda_1(\mathbf{A})^2 + cd\tau}{\lambda_d(\mathbf{A})^2}},$$

$B_1$  is defined in Theorem 7.1,  $B_3$  is defined in Theorem 7.2,  $\epsilon_0 = c_0\sqrt{\frac{\log p}{n}}$ ,  $\epsilon_1 = c_1\sqrt{\frac{p}{n}}$ ,  $\tau = \max\{\frac{p}{n}, 1\}$  and  $c, c_0, c_1$  are universal constants.

Notice that  $p$  cannot grow faster than  $\frac{n\lambda_d(\mathbf{A})^2}{d}$  due to the existence of outliers, but the first term in  $B_2$  in Corollary 7.2 involves  $\frac{\log p}{n}$  instead of  $\frac{p}{n}$ , which is much smaller than that in Corollary 7.1. Thus, the soft constraint is helpful if the true solution is indeed sparse. When the selected PCA-like algorithm has a “hard” constraint on the sparsity, e.g., the ones proposed by [PDK13] and [APD14],  $p$  can grow much faster than  $n$ .

**Corollary 7.3.** *Suppose that  $\rho < 0.5$  and  $\log p \leq n$ , then when  $\gamma \geq \beta$  and the PCA-like*

algorithm is sparse PCA [PDK13] or non-negative sparse PCA [APD14], the following holds with high probability,

$$\|\mathbf{X}^* - \mathbf{\Pi}\|_F \leq \sqrt{\frac{2[(dB_2 + \beta B_3 + \beta \epsilon_0)\lambda_1(\mathbf{A})^2 + c\beta(d\lambda_1(\mathbf{A}) + 1)]}{\delta_1(\mathbf{A}\mathbf{A}^\top)}}, \quad (7.6)$$

where

$$B_2 = \frac{\epsilon_0(\lambda_1(\mathbf{A})^2 + 1)\beta}{\delta_1(\mathbf{A}\mathbf{A}^\top)} + \sqrt{\gamma} \min_{1 \geq \kappa > 2\rho} \sqrt{\frac{B_1\lambda_1(\mathbf{A})^2 + c(d\lambda_1(\mathbf{A}) + 1) + \eta B_0}{\delta_1(\mathbf{A}\mathbf{A}^\top)}}.$$

$B_0, B_1$  are defined in Theorem 7.1,  $B_3$  is defined in Theorem 7.2,  $\epsilon_0 = c_0\sqrt{\frac{\log p}{n}}$ , and  $c, c_0$  are universal constants.

Recall that  $\mathbf{\Sigma} = \mathbf{A}\mathbf{A}^\top + I_p$ . The bound shown in Corollary 7.3 can be finite regardless of the magnitude of the existing outliers, e.g., when  $d, \beta\sqrt{\frac{\log p}{n}}, (B_1 + B_3 + \eta)\gamma, \frac{\gamma}{\lambda_1(A)}$  and  $\frac{\lambda_1(\mathbf{\Sigma})}{\lambda_1(\mathbf{\Sigma}) - \lambda_2(\mathbf{\Sigma})}$  are bounded from above.

### 7.3.2 Lower Bound of Expressed Variance

[XCM13] and [FXY12] provided lower bounds of E.V. when the standard PCA is selected in Algorithm 7.1. We now show that E.V. can be bounded from below when other PCA-like algorithm of form (7.1) (and workable) are used in Algorithm 7.1. Let  $H^* \triangleq \langle \mathbf{A}\mathbf{A}^\top, \mathbf{X}^* \rangle$  and  $\bar{H} \triangleq \langle \mathbf{A}\mathbf{A}^\top, \mathbf{\Pi} \rangle$ , then we have the following theorem.

**Theorem 7.3.** *Suppose that  $\rho < 0.5$ . For any  $\kappa$ , there exists a constant  $c$  such that the following inequalities hold w.h.p,*

$$\begin{aligned} E.V. \geq & \frac{(1 - \eta)\mathcal{V}^-\left(\frac{\hat{t}}{t} - \frac{\rho}{1 - \rho}\right)\mathcal{V}^-\left(1 - \frac{\rho}{\kappa(1 - \rho)}\right)}{(1 + \kappa)\mathcal{V}^+\left(\frac{\hat{t}}{t}\right)} \\ & - \frac{c}{\mathcal{V}^+(0.5)} \left[ \left(\frac{k \min\{\tau, \gamma\varsigma\}}{\bar{H}}\right)^{\frac{1}{2}} + \left(\frac{d \log^3 n}{n}\right)^{\frac{1}{4}} \right] \\ & - \frac{2(1 - \eta)\mu\beta\sqrt{k}}{\mathcal{V}^+\left(\frac{\hat{t}}{t}\right)\bar{H}} - \max\{1 - \lambda_k(\mathbf{X}^*), \lambda_{k+1}(\mathbf{X}^*)\}, \end{aligned} \quad (7.7)$$

where  $\tau = \max\{\frac{p}{n}, 1\}$  and  $\varsigma = \max\{\frac{\log p}{n}, 1\}$ .

As discussed in Section 7.2.2,  $\mathbf{X}^*$  has the form  $\mathbf{X}^* = \mathbf{u}\mathbf{u}^\top$  for the standard PCA, non-negative PCA [MR14], sparse PCA [PDK13] and non-negative sparse PCA [APD14], which implies that the last term in (7.7) vanishes for these four algorithms when  $k = 1$ . But for FPS [VCLR13], this term may not be zero. The following lemma shows that it can converge to zero under certain circumstances.

**Lemma 7.2.** *Suppose that  $\mathcal{S}$  is a sequence of matrices such that for any  $\mathbf{S}_n \in \mathcal{S}$ ,  $\mathbf{S}_n \in \mathbb{S}_+^{p \times p}$  and  $\lambda_d(\mathbf{S}_n) - \lambda_{d+1}(\mathbf{S}_n) \geq \delta > 0$ . Let*

$$\mathbf{X}_n \triangleq \arg \max_{\mathbf{X} \in \mathcal{F}(d)} \langle \mathbf{S}_n, \mathbf{X} \rangle - \mu_n \|\mathbf{X}\|_1,$$

then if  $\mu_n \rightarrow 0$  as  $n \rightarrow +\infty$  and  $pd^{3/2} = o(\frac{1}{\mu_n})$ , we have  $\lambda_d(\mathbf{X}_n) \rightarrow 1$  and  $\lambda_{d+1}(\mathbf{X}_n) \rightarrow 0$  as  $n \uparrow +\infty$ .

The following result shows the asymptotic bound of the expressed variance in which we assume that the last term in (7.7) converges to zero as  $n$  goes to infinity. This condition holds for all the algorithms mentioned above.

**Theorem 7.4.** *(Asymptotic Bound): Consider a sequence of  $\{\mathcal{Y}_i, d_i, n_i, p_i, \mu_i, \beta_i, \gamma_i\}$ , where the asymptotic scaling satisfies*

$$n_i \uparrow +\infty, \lim_{i \uparrow +\infty} \frac{\log p_i}{n_i} \leq +\infty, \lim_{i \uparrow +\infty} \frac{\min\{p_i/n_i, \gamma_i\}}{\sum_{j=1}^k \lambda_j(\mathbf{A}_i)^2} \downarrow 0,$$

$$\frac{n_i}{d_i \log^3 n_i} \uparrow +\infty, \frac{d_i}{\sum_{j=1}^k \lambda_j(\mathbf{A}_i)^2} \downarrow 0, \mu_i \beta_i \downarrow 0,$$

Let  $\rho^* = \limsup \rho_i \leq 0.5$  and suppose  $\hat{t} > 0.5n$ , then if  $\lambda_k(\mathbf{X}^*) \rightarrow 1$  and  $\lambda_{k+1}(\mathbf{X}^*) \rightarrow 0$  as  $n_i \uparrow +\infty$ , the following holds in probability when  $i \uparrow +\infty$ ,

$$\liminf_i E.V \geq (1 - \eta) \max_{\kappa} \frac{\mathcal{V}^- \left(1 - \frac{\rho^*}{(1-\rho^*)\kappa}\right) \mathcal{V}^- \left(\frac{\hat{t}}{\hat{t}} - \frac{\rho^*}{1-\rho^*}\right)}{(1 + \kappa) \mathcal{V}^+ \left(\frac{\hat{t}}{\hat{t}}\right)}.$$

Furthermore, if  $\bar{\mu}_{\mathbf{v}}(\{0\}) = 0$  for all  $\mathbf{v} \in \mathcal{S}_d$ , then the breakdown point is  $\rho^* = 0.5$ .

**Corollary 7.4.** *Under the settings of the above theorem, the following holds in probability*

for some constant  $C$  when  $i \uparrow +\infty$ ,

$$\liminf_i E.V \geq (1 - \eta) \left[ \frac{\mathcal{V}^-(\hat{t}) - C\sqrt{\theta\rho^* \log(1/2\rho^*)}}{\mathcal{V}^+(\hat{t})} \right].$$

### 7.3.3 Complexity

Recall that Algorithm 7.1 is an iterative algorithm that solves a PCA-like algorithm in each iteration. Theoretically, the number of iterations required to generate a good solution is bounded by  $n$ . But in practice, one can stop the algorithm at any time as long as the output of the robust variance estimator is good enough. We will show in the experiments that 5-10 iterations are sufficient to achieve a good solution. Since the time and space complexity of Algorithm 7.1 mainly depends on performing the selected PCA-like algorithm, this means the computational cost of Algorithm 7.1 is about 5-10 times higher than the non-robust PCA-like algorithm – robustness is not a free lunch, but you don’t pay much.

## 7.4 Experimental Results

In this section, we show that our framework indeed makes PCA-like algorithms more robust to outliers. We refer to the selected PCA-like algorithm in Step 3 in Algorithm 7.1 as  $\mathcal{A}$  and consider four algorithms induced from our framework: 1) OR-PCA:  $\mathcal{A}$  is the standard PCA. OR-PCA has been extensively studied in [XCM13]. 2) OR-SPCA:  $\mathcal{A}$  is FPS [VCLR13] to encourage sparse solutions. 3) Nonnegative OR-SPCA:  $\mathcal{A}$  is non-negative sparse PCA [APD14]. 4) Large-scale OR-SPCA:  $\mathcal{A}$  is the algorithm proposed by [ZE11] which is able to handle high dimensional data. Although this algorithm has no performance guarantees, it does work well in the experiments.

Firstly, we illustrate the performance of OR-PCA and OR-SPCA via numerical results on synthetic and real data. For synthetic data, we generate matrix  $\mathbf{A}$  via the following three steps: 1) randomly generate sparse orthogonal matrices  $\mathbf{U} \in \mathbb{R}^{p \times d}$  and  $\mathbf{V} \in \mathbb{R}^{d \times d}$  such that  $\|\mathbf{U}\|_{2,0} = \beta$  where  $\|\mathbf{U}\|_{2,0}$  is the number of non-zero rows in  $\mathbf{U}$ ; 2) generate a diagonal matrix  $\mathbf{S}$  whose diagonal entries are drawn from (a) the uniform distribution over  $[1, 2]$  or

(b) the chi-square density  $\frac{x^{-0.5}e^{-0.5x}}{\sqrt{2}\Gamma(0.5)}$  where  $x$  is chosen from 0.05 to  $0.05d$  using step-size 0.05; 3) finally, let  $\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^\top$ . The  $t$  authentic samples  $\mathbf{z}_i$  are generated by the function  $\mathbf{z}_i = \mathbf{A}\mathbf{x}_i + \mathbf{n}_i$  where  $\mathbf{x}_i \sim \mathcal{N}(0, I_d)$ ,  $\mathbf{n}_i \sim \mathcal{N}(0, \sigma^2 I_p)$ . A  $\rho$  fraction outliers  $\mathbf{o}_i$  are generated with a uniform distribution over  $[-c, c]^p$  where  $c$  is a constant.

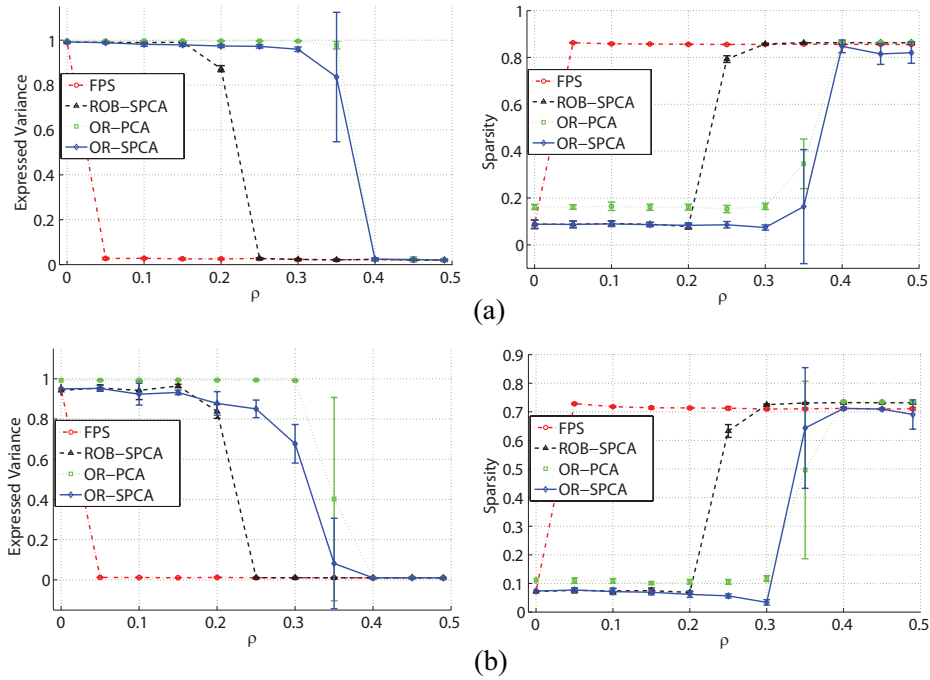
We make a comparison between OR-PCA, OR-SPCA, FPS and ROB-SPCA. ROB-SPCA is developed based on [HRS14], which uses ROBPCA [HRB05] to estimate the robust sample covariance and then applies FPS to compute the principal components. The performance is evaluated by the “expressed variance” and “sparsity”. The sparsity is defined by

$$\text{Sparsity} \triangleq |(i, j) : |X_{ij}| > 0.001|/p^2,$$

where  $\mathbf{X}$  is the projection matrix generated by each algorithm.

In the first experiment, we compare the performance of each algorithm when  $\rho$  varies while the other parameters are fixed. The parameters for generating test data are set as follows:  $d = 10$ ,  $\sigma = 0.05$ ,  $\beta = 0.3p$ . Parameter  $T$  and  $\hat{t}$  for OR-PCA and OR-SPCA are set to 10 and  $\rho n$ , respectively. Parameter  $\mu$  for FPS and OR-SPCA is  $0.2\sqrt{\frac{\log p}{n}}$ . For each parameter setup, we report the average results of 10 tests. Figures 7.1 and 7.2 show the performance of these four algorithms. Clearly, FPS easily breaks down, even when there exists only a small fraction of outliers. ROB-SPCA breaks down when  $\rho$  is larger than 0.25. Actually, most of robust PCA algorithms based on ROBPCA do not work well when the fraction of outliers exceeds 0.25 [XCM13]. One can also observe that OR-PCA and OR-SPCA are much more robust than the other two algorithms, and OR-SPCA can generate more sparse solutions than OR-PCA without significant decrease in the expressed variance, which implies that our framework has the capability of converting a non-robust SPCA algorithm, e.g., FPS, into a robust one.

In the second experiment, we investigate the number of the iterations required in Algorithm 7.1 to achieve good performance. We take OR-SPCA as an example. Figure 7.3 shows the effect of the number of iterations on the expressed variance and sparsity for OR-SPCA under three cases that  $p = 600$ ,  $p = 800$  and  $p = 1000$ , from which we observe that only 5 iterations are required for OR-SPCA to generate acceptable results in all three cases.

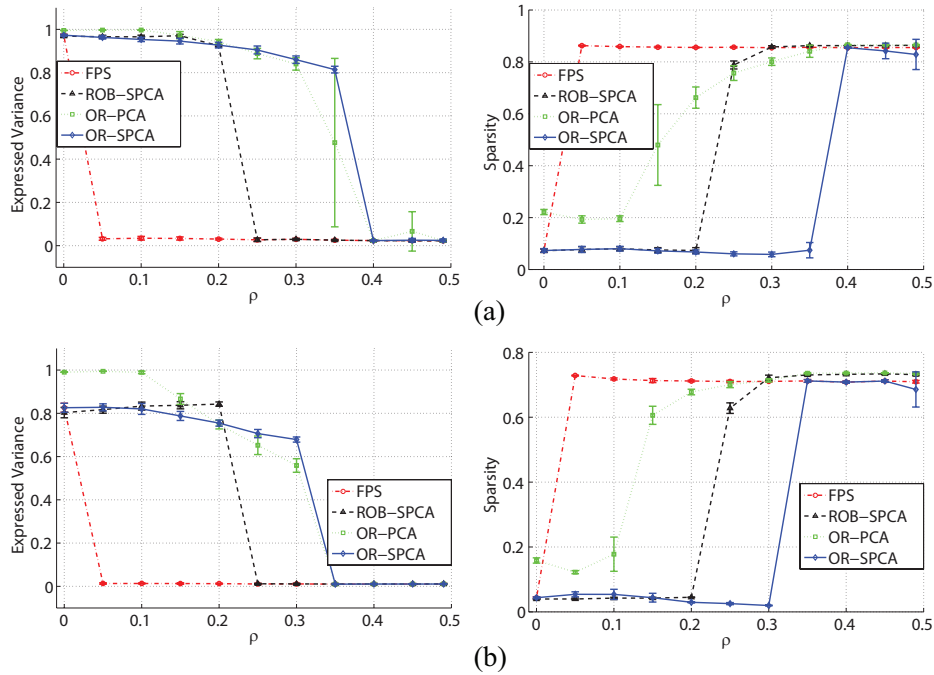


**Figure 7.1:** The performance of OR-PCA, OR-SPCA, ROB-SPCA and FPS under (a)  $p = 500, n = 300, c = 5$  and (b)  $p = 1000, n = 300, c = 5$ . The singular values of  $\mathbf{A}$  are uniformly drawn from  $[1, 2]$ .

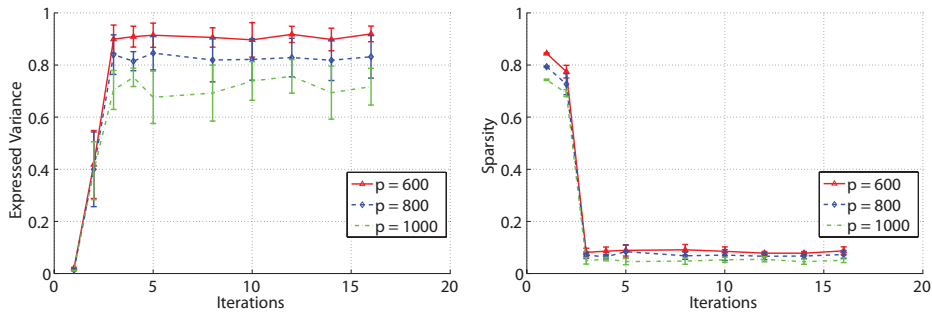
Empirically, we observe that 5-10 iterations are enough for Algorithm 7.1 to compute good results in practical applications. Hence in the following experiments on real data, parameter  $T$  is set to 10.

In the third experiment, we show the performance of OR-SPCA, OR-PCA and FPS on a real dataset of 600 samples in which 75% of samples are drawn from MNIST [LJB<sup>+</sup>95] and 25% of samples are drawn from the CBCL face image dataset [Sun96]. We take the digit images as the authentic samples and the face images as the outliers. Each image in this dataset is converted into a vector with dimension 784. Figure 7.4 shows the leading ten principal components extracted by FPS, OR-PCA and OR-SPCA. It can be observed that OR-SPCA can generate more interpretable results than OR-PCA, i.e., each PC corresponds to some strokes. Notice that the principal components extracted by OR-SPCA are more reliable than FPS. For example, the third principal component extracted by FPS clearly mixes digits with faces, which is obviously unreliable.

Secondly, we evaluate the performance of the non-negative OR-SPCA on the real world



**Figure 7.2:** The performance of OR-PCA, OR-SPCA, ROB-SPCA and FPS under (a)  $p = 500, n = 300, c = 5$  and (b)  $p = 1000, n = 300, c = 5$ . The singular values of  $\mathbf{A}$  are drawn from the chi-square density.



**Figure 7.3:** The effect of the number of iterations on the expressed variance and sparsity.  $n, \rho$  and  $c$  are fixed:  $n = 300, \rho = 0.1, c = 5$ .

dataset constructing by mixing 2429 images in the CBCL face image dataset with 125 digit images randomly drawn from the MNIST dataset. We take the face images as the authentic samples and the digit images as the outliers. Each image in this dataset is converted into a vector with dimension 361. We compare non-negative OR-SPCA with non-negative SPCA. Figure 7.5 shows the sample images and the five leading PCs computed by non-negative SPCA and non-negative OR-SPCA. Clearly, non-negative SPCA fails in the face of these



**Figure 7.4:** We plot the leading ten PCs extracted by OR-PCA, FPS and OR-SPCA. (a) shows a couple of sample images. (b), (c) and (d) show the results of OR-PCA, FPS and OR-SPCA, respectively.

“digit” outliers, while non-negative OR-SPCA can still extract good principal components that are close to the ones generated by applying non-negative SPCA on the clean data, i.e., 2429 face images only.



**Figure 7.5:** We plot (a) five samples in the dataset, (b) the five leading PCs extracted by non-negative SPCA on the clean data (2429 face images), and the five leading PCs extracted by (c) non-negative SPCA and (d) non-negative OR-SPCA on the dirty data (2429 face images plus 125 outliers).

Finally, we use the NYTimes news article dataset from the UCI Machine Learning Repository [FA10], which contains 300000 articles and a dictionary of 102660 unique words, to illustrate the performance of Algorithm 7.1 on large-scale data. 3000 random vectors whose entries are randomly drawn from the uniform distribution with support  $[0, 100]$  are added into the NYTimes dataset, which are taken as outliers. We choose large-scale SPCA (LS-SPCA) proposed by [ZE11] as  $\mathcal{A}$  and compare the corresponding large-scale OR-SPCA (LS-OR-SPCA) with it. Table 7.1 provides the leading two sparse PCs in which the first two columns shows the two leading PCs extracted by LS-SPCA on the dataset without outliers, and the other four columns presents the leading PCs extracted by LS-SPCA and LS-OR-SPCA on the dataset with outliers. The ground truth is obtained by performing large-scale sparse PCA on the clean data. Clearly, the results of LS-SPCA are meaning-



**Table 7.1:** The words associated with the leading two sparse principal components extracted by large-scale SPCA and large-scale OR-SPCA.

Ground-truth		LS-SPCA		LS-OR-SPCA	
1st PC	2st PC	1st PC	2st PC	1st PC	2st PC
million	point	site	fire	percent	team
percent	play	summer	scientist	company	player
business	team	contract	oil	million	season
company	season	system	prices	market	game
market	game	person	district	money	play

less when outliers exist, whereas LS-OR-SPCA can generate quite similar results to the ground-truth where the first PC is about business and the second PC is about sports.

## 7.5 Proofs of Section 7.3.1

**Lemma 7.3.** (Lemma 3.1, [VCLR13]) Let  $\Sigma$  be a symmetric matrix and  $\Pi$  be the projection onto the subspace spanned by the eigenvectors of  $\Sigma$  corresponding to its  $k$  largest eigenvalues  $\lambda_1 \geq \lambda_2 \geq \dots$ . If  $\delta = \lambda_k(\Sigma) - \lambda_{k+1}(\Sigma) > 0$ , then

$$\frac{\delta}{2} \|\Pi - \mathbf{X}\|_F^2 \leq \langle \Sigma, \Pi - \mathbf{X} \rangle$$

for all  $\mathbf{X}$  satisfying  $0 \preceq \mathbf{X} \preceq \mathbf{I}$  and  $\text{tr}(\mathbf{X}) = k$ .

**Lemma 7.4.** The event  $\mathcal{E}(s)$  is true for some  $1 \leq s \leq s_0$  where  $s_0 = \frac{\rho n(1+\kappa)}{\kappa}$ .

### Proof of Lemma 7.4

*Proof.* If  $\mathcal{E}(s)$  is false, then

$$\sum_{i \in \mathcal{Z}} \alpha_i(s) \langle \mathbf{y}_i \mathbf{y}_i^\top, \mathbf{X}_s \rangle < \frac{1}{\kappa} \sum_{i \in \mathcal{O}} \alpha_i(s) \langle \mathbf{y}_i \mathbf{y}_i^\top, \mathbf{X}_s \rangle.$$

Let  $\Delta\alpha_i \triangleq \frac{\langle \mathbf{y}_i \mathbf{y}_i^\top, \hat{\mathbf{X}} \rangle}{\max_{\{i|\alpha_i \neq 0\}} \langle \mathbf{y}_i \mathbf{y}_i^\top, \hat{\mathbf{X}} \rangle} \alpha_i$ , if  $\bigcup_{s=1}^{s_0} \mathcal{E}(s)$  is false, we have

$$\sum_{s=1}^{s_0} \sum_{i \in \mathcal{Z}} \Delta\alpha_i(s) < \frac{1}{\kappa} \sum_{s=1}^{s_0} \sum_{i \in \mathcal{O}} \Delta\alpha_i(s).$$

From the algorithm, at least one  $\alpha$  is eliminated in each iteration. Thus, we have

$$\sum_{s=1}^{s_0} \sum_{i=1}^n \Delta\alpha_i(s) \geq s_0,$$

which implies that

$$\sum_{s=1}^{s_0} \sum_{i \in \mathcal{Z}} \Delta\alpha_i(s) + \sum_{s=1}^{s_0} \sum_{i \in \mathcal{O}} \Delta\alpha_i(s) \geq s_0.$$

Hence

$$\frac{1}{\kappa} \sum_{s=1}^{s_0} \sum_{i \in \mathcal{O}} \Delta\alpha_i(s) + \sum_{s=1}^{s_0} \sum_{i \in \mathcal{O}} \Delta\alpha_i(s) \geq s_0.$$

Note that  $\rho n \geq \sum_{s=1}^{s_0} \sum_{i \in \mathcal{O}} \Delta\alpha_i(s)$ , then  $\rho n \geq \frac{\kappa s_0}{1+\kappa}$ , so  $s_0 \leq \frac{\rho n(1+\kappa)}{\kappa}$ .  $\square$

### Proof of Theorem 7.1

*Proof.* Note that for any  $1 \leq \bar{s} \leq s$ , the event  $\mathcal{E}(\bar{s})$  is false, which implies that

$$\sum_{i \in \mathcal{Z}} \alpha_i(\bar{s}) \langle \mathbf{y}_i \mathbf{y}_i^\top, \mathbf{X}_{\bar{s}} \rangle < \frac{1}{\kappa} \sum_{i \in \mathcal{O}} \alpha_i(\bar{s}) \langle \mathbf{y}_i \mathbf{y}_i^\top, \mathbf{X}_{\bar{s}} \rangle,$$

Thus, we have

$$\sum_{i \in \mathcal{Z}} \Delta\alpha_i(\bar{s}) < \frac{1}{\kappa} \sum_{i \in \mathcal{O}} \Delta\alpha_i(\bar{s}).$$

Since  $\alpha_i(s) = 1 - \sum_{k=1}^{s-1} \Delta\alpha_i(k)$ ,

$$\sum_{i \in \mathcal{Z}} \alpha_i(s) = t - \sum_{i \in \mathcal{Z}} \sum_{k=1}^{s-1} \Delta\alpha_i(k) > t - \frac{1}{\kappa} \sum_{k=1}^{s-1} \sum_{i \in \mathcal{O}} \Delta\alpha_i(k) \geq t - \frac{\rho n}{\kappa}.$$

Hence for any  $\mathbf{X} \in \mathcal{F}(k)$ , we have

$$\begin{aligned}
& \sum_{i \in \mathcal{Z}} \alpha_i(s) \langle \mathbf{y}_i \mathbf{y}_i^\top, \mathbf{X} \rangle - \sum_{i=1}^{t-\rho n/\kappa} \langle \mathbf{z} \mathbf{z}^\top, \mathbf{X} \rangle_{(i)} \\
&= \sum_{i=1}^{t-\rho n/\kappa} (\alpha_{j(i)}(s) - 1) \langle \mathbf{z}_{j(i)} \mathbf{z}_{j(i)}^\top, \mathbf{X} \rangle + \sum_{i=t-\rho n/\kappa+1}^t \alpha_{j(i)}(s) \langle \mathbf{z}_{j(i)} \mathbf{z}_{j(i)}^\top, \mathbf{X} \rangle \\
&\geq \sum_{i=1}^{t-\rho n/\kappa} (\alpha_{j(i)}(s) - 1) \langle \mathbf{z} \mathbf{z}^\top, \mathbf{X} \rangle_{(t-\rho n/\kappa)} + \sum_{i=t-\rho n/\kappa+1}^t \alpha_{j(i)}(s) \langle \mathbf{z} \mathbf{z}^\top, \mathbf{X} \rangle_{(t-\rho n/\kappa)} \\
&= \left( \sum_{i \in \mathcal{Z}} \alpha_i - \left(t - \frac{\rho n}{\kappa}\right) \right) \langle \mathbf{z} \mathbf{z}^\top, \mathbf{X} \rangle_{(t-\rho n/\kappa)} \geq 0.
\end{aligned} \tag{7.8}$$

Since  $\mathbf{X}_s$  is the optimal solution of the PCA-like algorithm and event  $\mathcal{E}(s)$  is true, we have

$$\begin{aligned}
& \frac{1}{n} \left\langle \sum_{i=1}^n \alpha_i(s) \mathbf{y}_i \mathbf{y}_i^\top, \mathbf{X}_s \right\rangle - \mu \|\mathbf{X}_s\|_1 \\
&\geq (1 - \eta) \left[ \frac{1}{n} \left\langle \sum_{i=1}^n \alpha_i(s) \mathbf{y}_i \mathbf{y}_i^\top, \mathbf{\Pi} \right\rangle - \mu \|\mathbf{\Pi}\|_1 \right] \\
&\geq (1 - \eta) \left[ \frac{1}{n} \left\langle \sum_{i \in \mathcal{Z}} \alpha_i(s) \mathbf{y}_i \mathbf{y}_i^\top, \mathbf{\Pi} \right\rangle - \mu \|\mathbf{\Pi}\|_1 \right] \\
&\geq \frac{1}{n} \left\langle \sum_{i \in \mathcal{Z}} \alpha_i(s) \mathbf{y}_i \mathbf{y}_i^\top, \mathbf{\Pi} \right\rangle - \mu \|\mathbf{\Pi}\|_1 - \eta \left( \mu + \left\| \frac{1}{n} \sum_{i \in \mathcal{Z}} \alpha_i(s) \mathbf{y}_i \mathbf{y}_i^\top \right\|_\infty \right) \|\mathbf{\Pi}\|_1
\end{aligned}$$

and

$$\begin{aligned}
& \frac{1}{n} \left\langle \sum_{i=1}^n \alpha_i(s) \mathbf{y}_i \mathbf{y}_i^\top, \mathbf{X}_s \right\rangle - \mu \|\mathbf{X}_s\|_1 \\
&= \frac{1}{n} \left\langle \sum_{i \in \mathcal{Z}} \alpha_i(s) \mathbf{y}_i \mathbf{y}_i^\top, \mathbf{X}_s \right\rangle + \frac{1}{n} \left\langle \sum_{i \in \mathcal{O}} \alpha_i(s) \mathbf{y}_i \mathbf{y}_i^\top, \mathbf{X}_s \right\rangle - \mu \|\mathbf{X}_s\|_1 \\
&\leq \frac{1 + \kappa}{n} \left\langle \sum_{i \in \mathcal{Z}} \alpha_i(s) \mathbf{y}_i \mathbf{y}_i^\top, \mathbf{X}_s \right\rangle - \mu \|\mathbf{X}_s\|_1.
\end{aligned}$$

Denote  $(\mu + \|\frac{1}{n} \sum_{i \in \mathcal{Z}} \alpha_i(s) \mathbf{y}_i \mathbf{y}_i^\top\|_\infty) \|\mathbf{\Pi}\|_1$  by  $B$ , we have

$$\frac{1}{n} \left\langle \sum_{i \in \mathcal{Z}} \alpha_i(s) \mathbf{y}_i \mathbf{y}_i^\top, \mathbf{X}_s - \mathbf{\Pi} \right\rangle - \mu \|\mathbf{X}_s\|_1 + \mu \|\mathbf{\Pi}\|_1 + \frac{\kappa}{n} \left\langle \sum_{i \in \mathcal{Z}} \alpha_i(s) \mathbf{y}_i \mathbf{y}_i^\top, \mathbf{X}_s \right\rangle + \eta B \geq 0.$$

Since Inequality 7.8 holds and  $0 \leq \alpha_i \leq 1$ ,

$$\frac{1}{n} \left\langle \sum_{i \in \mathcal{Z}} \mathbf{y}_i \mathbf{y}_i^\top, \mathbf{X}_s \right\rangle - \frac{1}{n} \sum_{i=1}^{t-\rho n/\kappa} \langle \mathbf{z} \mathbf{z}^\top, \mathbf{\Pi} \rangle_{(i)} - \mu \|\mathbf{X}_s\|_1 + \mu \|\mathbf{\Pi}\|_1 + \frac{\kappa}{n} \left\langle \sum_{i \in \mathcal{Z}} \mathbf{y}_i \mathbf{y}_i^\top, \mathbf{X}_s \right\rangle + \eta B \geq 0$$

or equivalently,

$$\frac{1}{n} \left\langle \sum_{i=1}^t \mathbf{z}_i \mathbf{z}_i^\top, \mathbf{X}_s - \mathbf{\Pi} \right\rangle - \mu \|\mathbf{X}_s\|_1 + \mu \|\mathbf{\Pi}\|_1 + \frac{\kappa}{n} \left\langle \sum_{i=1}^t \mathbf{z}_i \mathbf{z}_i^\top, \mathbf{X}_s \right\rangle + \frac{1}{n} \sum_{i=1}^{\rho n/\kappa} \langle \mathbf{z} \mathbf{z}^\top, \mathbf{\Pi} \rangle_{[i]} + \eta B \geq 0$$

Let  $\mathbf{\Delta} = \mathbf{X}_s - \mathbf{\Pi}$  and  $\mathbf{W} = \frac{1}{t} \sum_{i=1}^t \mathbf{z}_i \mathbf{z}_i^\top - \mathbf{\Sigma}$ , then

$$\langle \mathbf{W} + \mathbf{\Sigma}, \mathbf{\Delta} \rangle - \frac{n\mu}{t} \|\mathbf{\Pi} + \mathbf{\Delta}\|_1 + \frac{n\mu}{t} \|\mathbf{\Pi}\|_1 + \kappa \langle \mathbf{W} + \mathbf{\Sigma}, \mathbf{X}_s \rangle + \frac{1}{t} \sum_{i=1}^{\rho n/\kappa} \langle \mathbf{z} \mathbf{z}^\top, \mathbf{\Pi} \rangle_{[i]} + \frac{n\eta B}{t} \geq 0.$$

Since  $-\langle \mathbf{\Sigma}, \mathbf{\Delta} \rangle \geq \frac{\delta}{2} \|\mathbf{\Delta}\|_F^2$  where  $\delta = \lambda_k(\mathbf{\Sigma}) - \lambda_{k+1}(\mathbf{\Sigma})$  (Lemma 7.3), we have

$$\langle \mathbf{W}, \mathbf{\Delta} \rangle - \frac{n\mu}{t} \|\mathbf{\Pi} + \mathbf{\Delta}\|_1 + \frac{n\mu}{t} \|\mathbf{\Pi}\|_1 + \kappa \langle \mathbf{W} + \mathbf{\Sigma}, \mathbf{X}_s \rangle + \frac{1}{t} \sum_{i=1}^{\rho n/\kappa} \langle \mathbf{z} \mathbf{z}^\top, \mathbf{\Pi} \rangle_{[i]} + \frac{n\eta B}{t} \geq \frac{\delta}{2} \|\mathbf{\Delta}\|_F^2. \quad (7.9)$$

For simplicity, we let

$$T = \kappa \langle \mathbf{W} + \mathbf{\Sigma}, \mathbf{X}_s \rangle + \frac{1}{t} \sum_{i=1}^{\rho n/\kappa} \langle \mathbf{z} \mathbf{z}^\top, \mathbf{\Pi} \rangle_{[i]} + \frac{n\eta B}{t}.$$

We first consider the case that  $\mu \neq 0$ . Since  $\langle \mathbf{W}, \mathbf{\Delta} \rangle \leq \|\mathbf{W}\|_\infty \|\mathbf{\Delta}\|_1$  and  $n \geq t \geq 0.5n$ ,

$$[\|\mathbf{W}\|_\infty - \mu]_+ \|\mathbf{\Delta}\|_1 + \mu \|\mathbf{\Delta}\|_1 - \mu \|\mathbf{\Pi} + \mathbf{\Delta}\|_1 + \mu \|\mathbf{\Pi}\|_1 + T \geq \frac{\delta}{4} \|\mathbf{\Delta}\|_F^2.$$

Let  $N$  be the subset of indices of the nonzero entries of  $\mathbf{\Pi}$ , since  $\|\mathbf{\Pi}\|_0 \leq \beta^2$  and  $\|\mathbf{\Delta}_N\|_1 \leq \beta \|\mathbf{\Delta}_N\|_F \leq \beta \|\mathbf{\Delta}\|_F$ ,

$$\|\mathbf{\Delta}\|_1 - \|\mathbf{\Pi} + \mathbf{\Delta}\|_1 + \|\mathbf{\Pi}\|_1 = \|\mathbf{\Delta}_N\|_1 - \|\mathbf{\Pi}_N + \mathbf{\Delta}_N\|_1 + \|\mathbf{\Pi}_N\|_1 \leq 2\|\mathbf{\Delta}_N\|_1.$$

Also note that  $\mathbf{\Delta}$  has at most  $\gamma^2 + \beta^2$  non-zero entries, so  $\|\mathbf{\Delta}\|_1 \leq \sqrt{\gamma^2 + \beta^2} \|\mathbf{\Delta}\|_F \leq 2\gamma \|\mathbf{\Delta}\|_F$ . Thus,

$$2(\gamma[\|\mathbf{W}\|_\infty - \mu]_+ + \mu\beta) \|\mathbf{\Delta}\|_F + T \geq \frac{\delta}{4} \|\mathbf{\Delta}\|_F^2,$$

which implies that

$$\|\Delta\|_F \leq \frac{8(\gamma[\|\mathbf{W}\|_\infty - \mu]_+ + \mu\beta)}{\delta} + 2\sqrt{\frac{T}{\delta}} \leq \frac{8(\gamma[c_0\zeta\sqrt{\frac{\log p}{n}}\|\Sigma\|_2 - \mu]_+ + \mu\beta)}{\delta} + 2\sqrt{\frac{T}{\delta}},$$

where the last inequality follows from Lemma 7.17.

We now consider the case that  $\mu = 0$ , then (7.9) becomes  $\langle \mathbf{W}, \Delta \rangle + T \geq \frac{\delta}{2} \|\Delta\|_F^2$ . Since  $\langle \mathbf{W}, \Delta \rangle \leq \min\{\|\mathbf{W}\|_\infty \|\Delta\|_1, \|\mathbf{W}\|_2 \|\Delta\|_*\} \leq 2 \min\{\gamma \|\mathbf{W}\|_\infty \|\Delta\|_F, k \|\mathbf{W}\|_2\}$ ,  $\|\Delta\|_F$  should satisfy that

$$2 \min\{\gamma \|\mathbf{W}\|_\infty \|\Delta\|_F, k \|\mathbf{W}\|_2\} + T \geq \frac{\delta}{2} \|\Delta\|_F^2.$$

By simple calculation, we have

$$\|\Delta\|_F \leq \min\left\{\frac{8\gamma \|\mathbf{W}\|_\infty}{\delta}, 2\sqrt{\frac{k \|\mathbf{W}\|_2}{\delta}}\right\} + 2\sqrt{\frac{T}{\delta}} \leq \min\left\{\frac{8c_0\gamma\zeta\sqrt{\frac{\log p}{n}}\|\Sigma\|_2}{\delta}, 2\sqrt{\frac{c_1 k \zeta \sqrt{\frac{p}{n}}\|\Sigma\|_2}{\delta}}\right\} + 2\sqrt{\frac{T}{\delta}},$$

where the last inequality follows from Lemma 7.13 and Lemma 7.17. Hence we have  $\|\Delta\|_F \leq R(\mu) + 2\sqrt{\frac{T}{\delta}}$  where

$$R(\mu) = \begin{cases} \frac{8(\gamma[c_0\sqrt{\frac{\log p}{n}}(\|\mathbf{A}\|_2^2+1) - \mu]_+ + \mu\beta)}{\delta}, & \mu \neq 0 \\ \min\left\{\frac{8c_0\gamma\sqrt{\frac{\log p}{n}}(\|\mathbf{A}\|_2^2+1)}{\delta}, 2\sqrt{\frac{c_1 k \sqrt{\frac{p}{n}}(\|\mathbf{A}\|_2^2+1)}{\delta}}\right\}, & \mu = 0. \end{cases}$$

We ignore  $\zeta$  in  $R(\mu)$  because it's a constant.

We now bound  $T$ . Notice that  $\|\mathbf{\Pi}\|_* \leq k$  and  $\|\mathbf{\Pi}\|_1 \leq \beta \|\mathbf{\Pi}\|_F \leq \beta k$ , from Lemma 7.16, the following inequality holds with high probability,

$$\frac{1}{n} \sum_{i=1}^{n-\bar{n}} \langle \mathbf{z}\mathbf{z}^\top, \mathbf{\Pi} \rangle_{[i]} \leq k \min\{2(1 - \mathcal{V}^-(\bar{n}/n) + \epsilon(d)) \|\mathbf{A}\|_2^2 + c\tau, \beta[(1 - \mathcal{V}^-(\bar{n}/n) + \epsilon(d)) \|\mathbf{A}\|_2^2 + c\phi(1 + \zeta d \|\mathbf{A}\|_2)]\}.$$

Since  $t \geq 0.5n$ , there exist constants  $c_1, c_2$  such that  $\|\mathbf{W} + \Sigma\|_2 \leq 2(1 + c_1\theta\sqrt{\frac{d}{n}})\|\mathbf{A}\|_2^2 + c\tau$  and  $\|\mathbf{W}\|_\infty \leq c_0\zeta\sqrt{\frac{\log p}{n}}(\|\mathbf{A}\|_2^2 + 1)$  hold with high probability (Lemma 7.14 and Lemma

7.17). Hence

$$\begin{aligned} \langle \mathbf{W} + \boldsymbol{\Sigma}, \mathbf{X}_s \rangle &\leq \min\{\|\mathbf{W} + \boldsymbol{\Sigma}\|_2 \|\mathbf{X}_s\|_*, \|\mathbf{W} + \boldsymbol{\Sigma}\|_\infty \|\mathbf{X}_s\|_1\} \\ &\leq k \min \left\{ 2 \left( 1 + c_1 \theta \sqrt{\frac{d}{n}} \right) \|\mathbf{A}\|_2^2 + c\tau, \gamma \left( 1 + c_0 \zeta \sqrt{\frac{\log p}{n}} \right) (\|\mathbf{A}\|_2^2 + 1) \right\} \end{aligned}$$

where the last inequality follows from  $\|\mathbf{X}_s\|_* \leq k$  and  $\|\mathbf{X}_s\|_1 \leq \gamma \|\mathbf{X}_s\|_F \leq \gamma k$ . Also notice that  $0 \leq \alpha_i(s) \leq 1$  and  $\|\boldsymbol{\Pi}\|_1 \leq \beta \|\boldsymbol{\Pi}\|_F = \beta k$ , we have

$$\frac{n\eta B}{t} \leq \eta\beta k \left( 2\mu + \left\| \frac{1}{t} \sum_{i \in \mathcal{Z}} \mathbf{y}_i \mathbf{y}_i^\top \right\|_\infty \right) \leq \eta\beta k \left( 2\mu + \left( c_0 \zeta \sqrt{\frac{\log p}{n}} + 1 \right) \|\boldsymbol{\Sigma}\|_2 \right).$$

Let  $B'_0 \triangleq 2\mu + \left( c_0 \zeta \sqrt{\frac{\log p}{n}} + 1 \right) (\|\mathbf{A}\|_2^2 + 1)$ , since  $\mu$  is less than some universal constant and  $\log p \leq n$ , there exists constant  $c_2$  such that  $B_0 \triangleq c_2 (\|\mathbf{A}\|_2^2 + 1) \geq B'_0$ . Let  $\epsilon_0 \triangleq c_0 \zeta \sqrt{\frac{\log p}{n}}$ ,  $\epsilon_1 \triangleq \epsilon_0 + \epsilon(d) + c_1 \theta \sqrt{\frac{d}{n}}$ . Since  $d < n$ ,  $\epsilon_1 \leq \epsilon_0 + c_1 \left( \frac{d \log^3 n}{n} \right)^{\frac{1}{4}}$ . Since  $\zeta$  is a constant, we have  $\kappa \leq 1$ ,  $\beta \leq \gamma$  and  $\log p \leq n$ ,

$$T = k \min \left\{ 2B_1 \|\mathbf{A}\|_2^2 + c\tau, \gamma B_1 \|\mathbf{A}\|_2^2 + c\gamma (d \|\mathbf{A}\|_2 + 1) \right\} + \eta\beta k B_0,$$

where  $B_1 = \kappa + 1 - \mathcal{V}^{-1} \left( 1 - \frac{\rho}{\kappa(1-\rho)} \right) + \epsilon_1$ . By minimizing  $T$  over  $\kappa$ , we can obtain this theorem.  $\square$

### Proof of Theorem 7.2

*Proof.* Under the conditions of Theorem 7.2, the conditions of Theorem 7.1 are satisfied, let

$$\boldsymbol{\Delta} = \mathbf{X}_s - \boldsymbol{\Pi}, \mathbf{W} = \frac{1}{t} \sum_{i=1}^t \mathbf{z}_i \mathbf{z}_i^\top - \boldsymbol{\Sigma}, \text{ and } f(B) = \min \left\{ 2B \|\mathbf{A}\|_2^2 + c\tau, \gamma B \|\mathbf{A}\|_2^2 + c\gamma (d \|\mathbf{A}\|_2 + 1) \right\}$$

then w.h.p

$$\|\boldsymbol{\Delta}\|_F \leq R(\mu) + \sqrt{k} \min_{1 \geq \kappa > 2\rho} \sqrt{\frac{f(B_1) + \eta\beta B_0}{\delta}} \triangleq B_2.$$

From the Algorithm, we know that

$$\frac{1}{\hat{t}} \sum_{i=1}^{\hat{t}} \langle \mathbf{z}\mathbf{z}^\top, \mathbf{X}^* \rangle_{(i)} \geq \frac{1}{\hat{t}} \sum_{i=1}^{\hat{t}} \langle \mathbf{y}\mathbf{y}^\top, \mathbf{X}^* \rangle_{(i)} \geq \frac{1}{\hat{t}} \sum_{i=1}^{\hat{t}} \langle \mathbf{y}\mathbf{y}^\top, \mathbf{X}_s \rangle_{(i)} \geq \frac{1}{\hat{t}} \sum_{i=1}^{\hat{t}-\rho n} \langle \mathbf{z}\mathbf{z}^\top, \mathbf{X}_s \rangle_{(i)}.$$

Hence we have

$$\begin{aligned} & \frac{1}{t} \sum_{i=1}^t \langle \mathbf{z}_i \mathbf{z}_i^\top, \mathbf{X}^* \rangle - \frac{1}{t} \sum_{i=1}^{t-\hat{t}} \langle \mathbf{z} \mathbf{z}^\top, \mathbf{X}^* \rangle_{[i]} + \frac{1}{t} \sum_{i=1}^{t-\hat{t}+\rho n} \langle \mathbf{z} \mathbf{z}^\top, \mathbf{X}_s \rangle_{[i]} \geq \frac{1}{t} \sum_{i=1}^t \langle \mathbf{z}_i \mathbf{z}_i^\top, \mathbf{X}_s \rangle \\ \Rightarrow & \langle \mathbf{W} + \boldsymbol{\Sigma}, \mathbf{X}^* \rangle + \frac{1}{t} \sum_{i=1}^{t-\hat{t}} \langle \mathbf{z} \mathbf{z}^\top, \mathbf{X}^* \rangle_{[i]} + \frac{1}{t} \sum_{i=1}^{t-\hat{t}+\rho n} \langle \mathbf{z} \mathbf{z}^\top, \mathbf{X}_s \rangle_{[i]} \geq \langle \mathbf{W} + \boldsymbol{\Sigma}, \mathbf{X}_s \rangle. \end{aligned}$$

let  $T \triangleq \frac{1}{t} \sum_{i=1}^{t-\hat{t}} \langle \mathbf{z} \mathbf{z}^\top, \mathbf{X}^* \rangle_{[i]} + \frac{1}{t} \sum_{i=1}^{t-\hat{t}+\rho n} \langle \mathbf{z} \mathbf{z}^\top, \mathbf{X}_s \rangle_{[i]}$  and  $\boldsymbol{\Delta}^* \triangleq \mathbf{X}^* - \boldsymbol{\Pi}$ . Note that  $\boldsymbol{\Sigma} = \mathbf{A} \mathbf{A}^\top + \mathbf{I}_p$ , from Lemma 7.3, we have

$$\langle \mathbf{W}, \boldsymbol{\Delta}^* - \boldsymbol{\Delta} \rangle + \|\mathbf{A} \mathbf{A}^\top\|_F \|\boldsymbol{\Delta}\|_F + \|\boldsymbol{\Delta}\|_* + T \geq \frac{\delta}{2} \|\boldsymbol{\Delta}^*\|_F^2,$$

where  $\delta = \lambda_k(\boldsymbol{\Sigma}) - \lambda_{k+1}(\boldsymbol{\Sigma})$ . Since  $\|\boldsymbol{\Delta}^*\|_* \leq \|\mathbf{X}^*\|_* + \|\boldsymbol{\Pi}\|_* \leq 2k$ ,  $\|\mathbf{A} \mathbf{A}^\top\|_F = \sqrt{\sum_{i=1}^d \lambda_i(\mathbf{A} \mathbf{A}^\top)^2} \leq d \|\mathbf{A}\|_2^2$  and  $\|\boldsymbol{\Delta}\|_F \leq B_2$ , we have

$$\|\boldsymbol{\Delta}^*\|_F \leq \sqrt{\frac{2}{\delta} (\langle \mathbf{W}, \boldsymbol{\Delta}^* - \boldsymbol{\Delta} \rangle + dB_2 \|\mathbf{A}\|_2^2 + T + 2k)}.$$

We first bound the term  $\langle \mathbf{W}, \boldsymbol{\Delta}^* - \boldsymbol{\Delta} \rangle$ . Notice that

$$\langle \mathbf{W}, \boldsymbol{\Delta}^* - \boldsymbol{\Delta} \rangle \leq \min\{\|\mathbf{W}\|_2 (\|\boldsymbol{\Delta}^*\|_* + \|\boldsymbol{\Delta}\|_*), \|\mathbf{W}\|_\infty (\|\boldsymbol{\Delta}^*\|_1 + \|\boldsymbol{\Delta}\|_1)\}.$$

Since  $\|\boldsymbol{\Delta}^*\|_* \leq 2k$  and  $\|\boldsymbol{\Delta}^*\|_1 \leq \|\mathbf{X}^*\|_1 + \|\boldsymbol{\Pi}\|_1 \leq \gamma \|\mathbf{X}^*\|_F + \beta \|\boldsymbol{\Pi}\|_F \leq k(\gamma + \beta) \leq 2k\gamma$  ( $\boldsymbol{\Delta}$  has similar inequalities), we have

$$\langle \mathbf{W}, \boldsymbol{\Delta}^* - \boldsymbol{\Delta} \rangle \leq 4k \min\{\|\mathbf{W}\|_2, \gamma \|\mathbf{W}\|_\infty\}.$$

From Lemma 7.13, there exists constant  $c_2$  such that  $\|\mathbf{W}\|_2 \leq c_1 \zeta \sqrt{\frac{p}{n}} \|\boldsymbol{\Sigma}\|_2 = c_1 \zeta \sqrt{\frac{p}{n}} (\|\mathbf{A}\|_2^2 + 1)$  holds with high probability, where  $\zeta = \max\{\theta, 2\}$ . From Lemma 7.17,  $\|\mathbf{W}\|_\infty \leq c_2 \zeta \|\boldsymbol{\Sigma}\|_2 \sqrt{\frac{\log p}{n}} = c_2 \zeta \sqrt{\frac{\log p}{n}} (\|\mathbf{A}\|_2^2 + 1)$  holds for constant  $c_2$ . Let  $B_4 \triangleq 4\zeta \min\{c_1 \sqrt{\frac{p}{n}}, c_2 \gamma \sqrt{\frac{\log p}{n}}\}$ , then

$$\|\boldsymbol{\Delta}^*\|_F \leq \sqrt{\frac{2}{\delta} [(dB_2 + kB_4) \|\mathbf{A}\|_2^2 + T + 2k + kB_4]}.$$

For term  $T$ , we follow the same proof of Theorem 7.1. The following inequality holds w.h.p,

$$T \leq k \min \{2B_3\|\mathbf{A}\|_2^2 + c\tau, \gamma B_3\|\mathbf{A}\|_2^2 + c\gamma(d\|\mathbf{A}\|_2 + 1)\},$$

where  $B_3 = 2 - \mathcal{V}^-(\frac{\hat{t}}{t}) - \mathcal{V}^-(\frac{\hat{t}-\rho n}{t}) + \epsilon(d)$ . Hence we have

$$\|\Delta^*\|_F \leq \sqrt{\frac{2}{\delta} [(dB_2 + kB_4)\|\mathbf{A}\|_2^2 + k \min \{2B_3 + c\tau, \gamma B_3 + c\gamma(d\|\mathbf{A}\|_2 + 1)\}]},$$

which establishes this theorem.  $\square$

## 7.6 Proofs in Section 7.3.2

Let  $H^* \triangleq \langle \mathbf{A}\mathbf{A}^\top, \mathbf{X}^* \rangle$ ,  $H_s \triangleq \langle \mathbf{A}\mathbf{A}^\top, \mathbf{X}_s \rangle$  and  $\bar{H} \triangleq \langle \mathbf{A}\mathbf{A}^\top, \mathbf{\Pi} \rangle$ . In order to bound E.V, we first bound  $|H^* - \sum_{i=1}^k \mathbf{w}_i^{*\top} \mathbf{A}\mathbf{A}^\top \mathbf{w}_i^*|$ , and then bound  $H^*/\bar{H}$ . This involves the following steps:

1. Bound  $|H^* - \sum_{i=1}^k \mathbf{w}_i^{*\top} \mathbf{A}\mathbf{A}^\top \mathbf{w}_i^*|$ .
2. Bound the robust variance estimator of the the authentic samples by applying the concentration inequalities (Theorem 7.7, Theorem 7.8 and Theorem 7.9, i.e. bounding  $\frac{1}{t} \sum_{i=1}^{\hat{t}} |\mathbf{w}^\top \mathbf{z}|_{(i)}^2$ ).
3. Show that with high probability, the algorithm finds a “good” solution within a bounded number of steps and then show that the “good” solution in previous step is close to the optimal solution and the final solution of our algorithm is close to this “good” solution.

### Step 1

**Lemma 7.5.** *For any  $\mathbf{X} \in \mathbb{R}^{p \times p}$  such that  $0 \preceq \mathbf{X} \preceq \mathbf{I}_p$  and  $\text{tr}(\mathbf{X}) = k$ , let  $\mathbf{w}_1, \dots, \mathbf{w}_k$  be the top  $k$  eigenvectors of  $\mathbf{X}$ , then*

$$\left| \langle \mathbf{A}\mathbf{A}^\top, \mathbf{X} \rangle - \sum_{i=1}^k \mathbf{w}_i^\top \mathbf{A}\mathbf{A}^\top \mathbf{w}_i \right| \leq \max\{1 - \lambda_k(\mathbf{X}), \lambda_{k+1}(\mathbf{X})\} \cdot \text{tr}(\mathbf{A}\mathbf{A}^\top),$$



where  $\lambda_k$  is the  $k^{\text{th}}$  largest eigenvalue of  $\mathbf{X}$ .

From this lemma, we have  $\text{E.V}\{\mathbf{w}_1^*, \dots, \mathbf{w}_k^*\} \geq \frac{H^*}{H} - \max\{1 - \lambda_k(\mathbf{X}), \lambda_{k+1}(\mathbf{X})\}$ .

### Step 2

From Theorem 7.7, Theorem 7.8, Theorem 7.9 and Lemma 7.17, the following inequalities hold with high probability for constant  $c$ ,  $c_1$  and  $c_2$ :

$$\begin{aligned}
(I) \quad & \sup_{\mathbf{w} \in \mathcal{S}_p} \frac{1}{t} \sum_{i=1}^t (\mathbf{w}^\top \mathbf{n}_i)^2 \leq c\tau, \\
(II) \quad & \sup_{\mathbf{w} \in \mathcal{S}_d} \left| \frac{1}{t} \sum_{i=1}^t (\mathbf{w}^\top \mathbf{x}_i)^2 - 1 \right| \leq c_1 \theta \sqrt{\frac{d}{n}} \triangleq \epsilon_0, \\
(III) \quad & \sup_{\mathbf{w} \in \mathcal{S}_d} \left| \frac{1}{t} \sum_{i=1}^{\bar{t}} |\mathbf{w}^\top \mathbf{x}_{(i)}|^2 - \mathcal{V}\left(\frac{\bar{t}}{t}\right) \right| \leq \frac{c_2 t (1 + \epsilon_0) \sqrt{d \log n / n}}{t - \bar{t}} \wedge c_2 \theta^{\frac{1}{2}} d^{\frac{1}{4}} (\log n)^{\frac{3}{4}} n^{-\frac{1}{4}} \triangleq \epsilon_1 \left(\frac{\bar{t}}{t}\right), \\
(IV) \quad & \left\| \frac{1}{t} \sum_{i=1}^t \mathbf{n}_i \mathbf{n}_i^\top \right\|_\infty \leq c\varsigma,
\end{aligned}$$

where  $\tau = \max\{\frac{p}{n}, 1\}$  and  $\varsigma = \max\{\sqrt{\frac{\log p}{n}}, 1\}$ . When  $\bar{t} = t$ , we can indeed sharpen the result of (III) by applying (II), so let  $\epsilon_1(1) = \epsilon_0$ . We have the following theorem:

**Theorem 7.5.** *There exists a constant  $c$  such that the following inequalities hold w.h.p,*

$$\begin{aligned}
& \|\mathbf{X}^{1/2} \mathbf{A}\|_F^2 \left( \mathcal{V}^-\left(\frac{\bar{t}}{t}\right) - \epsilon_1\left(\frac{\bar{t}}{t}\right) \right) - 2\|\mathbf{X}^{1/2} \mathbf{A}\|_F \sqrt{(1 + \epsilon_0)ck \min\{\tau, \gamma_\varsigma\}} \\
& \leq \frac{1}{t} \sum_{i=1}^{\bar{t}} \langle \mathbf{z} \mathbf{z}^\top, \mathbf{X} \rangle_{(i)} \\
& \leq \|\mathbf{X}^{1/2} \mathbf{A}\|_F^2 \left( \mathcal{V}^+\left(\frac{\bar{t}}{t}\right) + \epsilon_1\left(\frac{\bar{t}}{t}\right) \right) + 2\|\mathbf{X}^{1/2} \mathbf{A}\|_F \sqrt{(1 + \epsilon_0)ck \min\{\tau, \gamma_\varsigma\}} + ck \min\{\tau, \gamma_\varsigma\},
\end{aligned}$$

for any  $\bar{t} \leq t$  and  $\mathbf{X} \in \mathcal{F}(k)$ .

### Step 3

Suppose that a “good” solution  $\mathbf{X}_s$  is found at stage  $s$  ( $0 \leq s \leq s_0$ ), namely event  $\mathcal{E}(s)$  is true. We can bound  $H^*/\bar{H}$  by leveraging the relationship between  $\mathbf{X}_s$  and  $\mathbf{\Pi}$  and the connection between  $\mathbf{X}^*$  and  $\mathbf{X}_s$ .

**Lemma 7.6.** *If  $\|\mathbf{\Pi}\|_0 \leq \beta^2$  and  $\mathcal{E}(s)$  is true for  $s \leq s_0$ , there exists a constant  $c$  such that the following inequalities hold w.h.p,*

$$(1 + \epsilon_0)H_s + 2\sqrt{(1 + \epsilon_0)ck \min\{\tau, \gamma_S\}H_s} + ck \min\{\tau, \gamma_S\} \\ \geq \frac{1 - \eta}{\kappa + 1} \left[ \left( \mathcal{V}^-\left(1 - \frac{\rho}{(1 - \rho)\kappa}\right) - \epsilon_1\left(1 - \frac{\rho}{(1 - \rho)\kappa}\right) \right) \overline{H} - 2\sqrt{(1 + \epsilon_0)ck \min\{\tau, \gamma_S\}\overline{H}} - \frac{1}{1 - \rho}\mu\beta\sqrt{k} \right].$$

**Lemma 7.7.** *Fix  $\hat{t} \leq t$ , there exists constant  $c$  such that the following inequalities hold w.h.p,*

$$\left( \mathcal{V}^+\left(\frac{\hat{t}}{t}\right) + \epsilon_1\left(\frac{\hat{t}}{t}\right) \right) H^* + 2\sqrt{(1 + \epsilon_0)ck \min\{\tau, \gamma_S\}H^*} + ck \min\{\tau, \gamma_S\} \\ \geq \left( \mathcal{V}^-\left(\frac{\hat{t} - \rho n}{t}\right) - \epsilon_1\left(\frac{\hat{t} - \rho n}{t}\right) \right) H_s - 2\sqrt{(1 + \epsilon_0)ck \min\{\tau, \gamma_S\}H_s}.$$

**Theorem 7.6.** *Suppose  $\|\mathbf{\Pi}\|_0 \leq \beta^2$  and  $\rho \leq 0.5$ . For any  $\kappa$ , there exists a constant  $c$  such that the following inequalities hold w.h.p,*

$$\frac{H^*}{\overline{H}} \geq \frac{\mathcal{V}^-\left(\frac{\hat{t}}{t} - \frac{\rho}{1 - \rho}\right) \mathcal{V}^-\left(1 - \frac{\rho}{\kappa(1 - \rho)}\right)}{(1 + \kappa)\mathcal{V}^+\left(\frac{\hat{t}}{t}\right)} - \frac{10}{\mathcal{V}^+(0.5)} \left( \frac{ck \min\{\tau, \gamma_S\}}{\overline{H}} \right)^{1/2} \\ - \frac{c\{\theta^{\frac{1}{2}}d^{\frac{1}{4}}(\log^{\frac{3}{4}}n)n^{-\frac{1}{4}} \vee \theta[(1 + \kappa)/\kappa]^{\frac{3}{2}}(\log^{\frac{3}{2}}n)n^{-\frac{1}{2}}\}}{\mathcal{V}^+(0.5)} - \frac{2\mu\beta\sqrt{k}}{\mathcal{V}^+\left(\frac{\hat{t}}{t}\right)\overline{H}}$$

### Proof of Lemma 7.5

*Proof.* Since  $0 \preceq \mathbf{X} \preceq \mathbf{I}_p$ , we have

$$\left| \langle \mathbf{A}\mathbf{A}^\top, \mathbf{X} \rangle - \sum_{i=1}^k \mathbf{w}_i^\top \mathbf{A}\mathbf{A}^\top \mathbf{w}_i \right| \leq \|\mathbf{A}\mathbf{A}^\top\|_* \cdot \|\mathbf{X} - \sum_{i=1}^k \mathbf{w}_i \mathbf{w}_i^\top\|_2 \\ = \text{tr}(\mathbf{A}\mathbf{A}^\top) \cdot \left\| \sum_{i=1}^k (\lambda_i - 1) \mathbf{w}_i \mathbf{w}_i^\top + \sum_{i=k+1}^p \lambda_i \mathbf{w}_i \mathbf{w}_i^\top \right\|_2 = \text{tr}(\mathbf{A}\mathbf{A}^\top) \cdot \max\{1 - \lambda_k(\mathbf{X}), \lambda_{k+1}(\mathbf{X})\}.$$

Hence we obtain this lemma.  $\square$

### Proof of Lemma 7.2

*Proof.* Let  $\mathbf{S} = \mathbf{S}_n$ ,  $\mu = \mu_n$  and  $\mathbf{\Delta} = \mathbf{B}_n - \mathbf{A}_n$ , then  $\langle \mathbf{S}, \mathbf{\Delta} \rangle \geq 0$  and  $\langle \mathbf{S}, \mathbf{\Delta} \rangle \leq \mu \|\mathbf{B}_n\|_1 - \mu \|\mathbf{A}_n\|_1 \leq \mu \|\mathbf{B}_n\|_1$ . Since  $\text{tr}(\mathbf{B}_n) = d$  and  $\mathbf{B}_n \succeq 0$ ,  $\|\mathbf{B}_n\|_1 \leq p \|\mathbf{B}_n\|_F = p\sqrt{d}$ . Then we have

$0 \leq \langle \mathbf{S}, \mathbf{\Delta} \rangle \leq \mu p \sqrt{d}$ . Since  $\mathbf{A}_n, \mathbf{B}_n \in \mathcal{F}_d$ ,

$$0 \leq \text{tr}(\mathbf{S}\mathbf{\Delta}) \leq \mu p \sqrt{d}, \quad 0 \preceq \mathbf{B}_n - \mathbf{\Delta} \preceq \mathbf{I}_p, \quad \text{tr}(\mathbf{\Delta}) = 0.$$

By SVD decomposition,  $\mathbf{S} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$  where  $\mathbf{Q}$  is an orthogonal matrix and  $\mathbf{\Lambda}$  is a diagonal matrix. Let  $\bar{\mathbf{\Delta}} = \mathbf{Q}^\top \mathbf{\Delta} \mathbf{Q}$ , then,

$$0 \leq \text{tr}(\mathbf{\Lambda}\bar{\mathbf{\Delta}}) \leq \mu p \sqrt{d}, \quad 0 \preceq \mathbf{\Sigma} - \bar{\mathbf{\Delta}} \preceq \mathbf{I}_p, \quad \text{tr}(\bar{\mathbf{\Delta}}) = 0,$$

where  $\mathbf{\Sigma} = \begin{bmatrix} I_d & 0 \\ 0 & 0 \end{bmatrix}$ . Thus,  $0 \leq \sum_{i=1}^p \lambda_i \bar{\Delta}_{ii} \leq \mu p \sqrt{d}$  and

$$\begin{aligned} 0 &\leq \bar{\Delta}_{ii} \leq 1 \text{ for } 1 \leq i \leq d, \\ -1 &\leq \bar{\Delta}_{ii} \leq 0 \text{ for } d+1 \leq i \leq p, \end{aligned}$$

which implies that  $\sum_{i=d+1}^p |\bar{\Delta}_{ii}| \leq \frac{\mu p \sqrt{d}}{\delta}$  (otherwise,  $\sum_{i=1}^p \lambda_i \bar{\Delta}_{ii} \geq \sum_{i=1}^p \lambda_d \bar{\Delta}_{ii} + |\lambda_d - \lambda_{d+1}| \sum_{i=d+1}^p |\bar{\Delta}_{ii}| > \mu p \sqrt{d}$ ). Since  $\text{tr}(\bar{\mathbf{\Delta}}) = 0$ , we also have  $\sum_{i=1}^d \bar{\Delta}_{ii} \leq \frac{\mu p \sqrt{d}}{\delta}$ . Let  $\bar{\mathbf{\Delta}} =$

$\begin{bmatrix} \mathbf{\Delta}_1 & -\mathbf{D} \\ -\mathbf{D}^\top & -\mathbf{\Delta}_2 \end{bmatrix}$ , then  $0 \preceq \begin{bmatrix} \mathbf{I}_d - \mathbf{\Delta}_1 & \mathbf{D} \\ \mathbf{D}^\top & \mathbf{\Delta}_2 \end{bmatrix} \preceq \mathbf{I}_p$ , which implies that  $\mathbf{\Delta}_1 \succeq 0$  and  $\mathbf{\Delta}_2 \succeq 0$ .  
Hence

$$\begin{aligned} \|\bar{\mathbf{\Delta}}\|_F^2 &= \|\mathbf{\Delta}_1\|_F^2 + \|\mathbf{\Delta}_2\|_F^2 + 2\|\mathbf{D}\|_F^2 \\ &\leq \text{tr}(\mathbf{\Delta}_1)^2 + \text{tr}(\mathbf{\Delta}_2)^2 + 2 \sum_{i=1}^d \sum_{j=d+1}^p D_{ij}^2 \\ &\leq 2 \left( \frac{\mu \beta \sqrt{d}}{\delta} \right)^2 + 2 \sum_{i=1}^d \sum_{j=d+1}^p |(1 - \bar{\Delta}_{ii}) \bar{\Delta}_{jj}| \\ &\leq 2 \left( \frac{\mu \beta \sqrt{d}}{\delta} \right)^2 + 2 \sum_{i=1}^d \sum_{j=d+1}^p |\bar{\Delta}_{jj}| \\ &\leq 2 \left( \frac{\mu \beta \sqrt{d}}{\delta} \right)^2 + 2 \frac{\mu p d^{3/2}}{\delta}. \end{aligned}$$

Thus,  $\|\mathbf{A}_n - \mathbf{B}_n\|_F^2 = \|\mathbf{Q}\bar{\mathbf{A}}\mathbf{Q}^\top\|_F^2 = \|\bar{\mathbf{A}}\|_F^2 \leq 2\left(\frac{\mu p\sqrt{d}}{\delta}\right)^2 + 2\frac{\mu p d^{3/2}}{\delta} \rightarrow 0$  as  $\mu \rightarrow 0$  when  $p d^{3/2} = o(\frac{1}{\mu})$ .  $\square$

### Proof of Theorem 7.5

*Proof.* For an arbitrary  $\mathbf{w} \in \mathcal{S}_p$ , let  $j(i)$  be permutations of  $\{1, \dots, n\}$  such that  $(\mathbf{w}^\top \mathbf{x}_{j(i)})^2$  is non-decreasing. Thus,

$$\begin{aligned} \frac{1}{t} \sum_{i=1}^{\bar{t}} \langle \mathbf{z}\mathbf{z}^\top, \mathbf{X} \rangle_{(i)} &\leq \frac{1}{t} \sum_{i=1}^{\bar{t}} \text{tr} \left( (\mathbf{A}\mathbf{x}_{j(i)} + \mathbf{n}_{j(i)})^\top \mathbf{X} (\mathbf{A}\mathbf{x}_{j(i)} + \mathbf{n}_{j(i)}) \right) \\ &= \frac{1}{t} \sum_{i=1}^{\bar{t}} \text{tr} \left( \mathbf{x}_{j(i)}^\top \mathbf{A}^\top \mathbf{X} \mathbf{A} \mathbf{x}_{j(i)} + 2\mathbf{n}_{j(i)}^\top \mathbf{X} \mathbf{A} \mathbf{x}_{j(i)} + \mathbf{n}_{j(i)}^\top \mathbf{X} \mathbf{n}_{j(i)} \right) \\ &\leq \frac{1}{t} \sum_{i=1}^{\bar{t}} \mathbf{x}_{j(i)}^\top \mathbf{A}^\top \mathbf{X} \mathbf{A} \mathbf{x}_{j(i)} + \frac{2}{t} \sum_{i=1}^{\bar{t}} \langle \mathbf{X}^{1/2} \mathbf{A} \mathbf{x}_{j(i)}, \mathbf{X}^{1/2} \mathbf{n}_{j(i)} \rangle + \frac{1}{t} \sum_{i=1}^{\bar{t}} \langle \mathbf{X}^{1/2} \mathbf{n}_i, \mathbf{X}^{1/2} \mathbf{n}_i \rangle \\ &\leq \|\mathbf{A}^\top \mathbf{X} \mathbf{A}\|_* \cdot \frac{1}{t} \sum_{i=1}^{\bar{t}} \mathbf{x}_{j(i)} \mathbf{x}_{j(i)}^\top + \frac{2}{t} \sum_{i=1}^{\bar{t}} \|\mathbf{X}^{1/2} \mathbf{A} \mathbf{x}_i\|_2 \cdot \|\mathbf{X}^{1/2} \mathbf{n}_i\|_2 + \frac{1}{t} \sum_{i=1}^{\bar{t}} \|\mathbf{X}^{1/2} \mathbf{n}_i\|_2^2 \end{aligned}$$

Since  $\|\mathbf{A}^\top \mathbf{X} \mathbf{A}\|_* = \|\mathbf{X}^{1/2} \mathbf{A}\|_F^2$  and the Cauchy-Schwarz inequality holds, we have

$$\begin{aligned} \frac{1}{t} \sum_{i=1}^{\bar{t}} \langle \mathbf{z}\mathbf{z}^\top, \mathbf{X} \rangle_{(i)} &\leq \|\mathbf{X}^{1/2} \mathbf{A}\|_F^2 \cdot \sup_{\mathbf{w} \in \mathcal{S}_d} \frac{1}{t} \sum_{i=1}^{\bar{t}} (\mathbf{w}^\top \mathbf{x})_{(i)}^2 + \\ &\quad 2\sqrt{\frac{1}{t} \sum_{i=1}^{\bar{t}} \|\mathbf{X}^{1/2} \mathbf{A} \mathbf{x}_i\|_2^2} \cdot \sqrt{\frac{1}{t} \sum_{i=1}^{\bar{t}} \|\mathbf{X}^{1/2} \mathbf{n}_i\|_2^2} + \frac{1}{t} \sum_{i=1}^{\bar{t}} \|\mathbf{X}^{1/2} \mathbf{n}_i\|_2^2 \\ &\leq \|\mathbf{X}^{1/2} \mathbf{A}\|_F^2 \cdot \sup_{\mathbf{w} \in \mathcal{S}_d} \frac{1}{t} \sum_{i=1}^{\bar{t}} (\mathbf{w}^\top \mathbf{x})_{(i)}^2 + \\ &\quad 2\|\mathbf{X}^{1/2} \mathbf{A}\|_F \sqrt{\sup_{\mathbf{w} \in \mathcal{S}_d} \frac{1}{t} \sum_{i=1}^{\bar{t}} (\mathbf{w}^\top \mathbf{x}_i)^2} \cdot \sqrt{\frac{1}{t} \sum_{i=1}^{\bar{t}} \|\mathbf{X}^{1/2} \mathbf{n}_i\|_2^2} + \frac{1}{t} \sum_{i=1}^{\bar{t}} \|\mathbf{X}^{1/2} \mathbf{n}_i\|_2^2 \end{aligned}$$

Note that

$$\begin{aligned} \frac{1}{t} \sum_{i=1}^{\bar{t}} \|\mathbf{X}^{1/2} \mathbf{n}_i\|_2^2 &= \langle \mathbf{X}, \frac{1}{t} \sum_{i=1}^{\bar{t}} \mathbf{n}_i \mathbf{n}_i^\top \rangle \leq \min\{\|\mathbf{X}\|_* \cdot \frac{1}{t} \sum_{i=1}^{\bar{t}} \|\mathbf{n}_i \mathbf{n}_i^\top\|_2, \|\mathbf{X}\|_1 \cdot \frac{1}{t} \sum_{i=1}^{\bar{t}} \|\mathbf{n}_i \mathbf{n}_i^\top\|_\infty\} \\ &\leq \min\{k \frac{1}{t} \sum_{i=1}^{\bar{t}} \|\mathbf{n}_i \mathbf{n}_i^\top\|_2, \gamma k \frac{1}{t} \sum_{i=1}^{\bar{t}} \|\mathbf{n}_i \mathbf{n}_i^\top\|_\infty\}. \end{aligned}$$

Then from (I)(II)(III)(IV), we have w.h.p

$$\frac{1}{t} \sum_{i=1}^{\bar{t}} \langle \mathbf{z}\mathbf{z}^\top, \mathbf{X} \rangle_{(i)} \leq \|\mathbf{X}^{1/2} \mathbf{A}\|_F^2 \left( \mathcal{V}^+ \left( \frac{\bar{t}}{t} \right) + \epsilon_1 \left( \frac{\bar{t}}{t} \right) \right) + 2\|\mathbf{X}^{1/2} \mathbf{A}\|_F \sqrt{(1 + \epsilon_0)ck \min\{\tau, \gamma_S\}} + ck \min\{\tau, \gamma_S\}.$$

We now compute the lower bound. For an arbitrary  $\mathbf{w} \in \mathcal{S}_p$ , let  $k(i)$  be permutations of  $\{1, \dots, n\}$  such that  $\langle \mathbf{z}_{k(i)} \mathbf{z}_{k(i)}^\top, \mathbf{X} \rangle$  is non-decreasing, then

$$\begin{aligned} \frac{1}{t} \sum_{i=1}^{\bar{t}} \langle \mathbf{z}\mathbf{z}^\top, \mathbf{X} \rangle_{(i)} &= \frac{1}{t} \sum_{i=1}^{\bar{t}} \mathbf{x}_{k(i)}^\top \mathbf{A}^\top \mathbf{X} \mathbf{A} \mathbf{x}_{k(i)} + \frac{2}{t} \sum_{i=1}^{\bar{t}} \mathbf{n}_{k(i)}^\top \mathbf{X} \mathbf{A} \mathbf{x}_{k(i)} + \frac{1}{t} \sum_{i=1}^{\bar{t}} \mathbf{n}_{k(i)}^\top \mathbf{n}_{k(i)} \\ &\geq \langle \mathbf{X}, \frac{1}{t} \mathbf{A} \sum_{i=1}^{\bar{t}} \mathbf{x}_{k(i)} \mathbf{x}_{k(i)}^\top \mathbf{A}^\top \rangle - \frac{2}{t} \sum_{i=1}^{\bar{t}} \|\mathbf{X}^{1/2} \mathbf{A} \mathbf{x}_i\|_2 \cdot \|\mathbf{X}^{1/2} \mathbf{n}_i\|_2 \end{aligned}$$

Perform SVD on  $\mathbf{X}$ , we have  $\mathbf{X} = \sum_{i=1}^p \alpha_i \mathbf{v}_i \mathbf{v}_i^\top$ , then

$$\begin{aligned} \langle \mathbf{X}, \frac{1}{t} \mathbf{A} \sum_{i=1}^{\bar{t}} \mathbf{x}_{k(i)} \mathbf{x}_{k(i)}^\top \mathbf{A}^\top \rangle &= \sum_{j=1}^p \frac{\alpha_j}{t} \sum_{i=1}^{\bar{t}} (\mathbf{v}_j^\top \mathbf{A} \mathbf{x}_{k(i)})^2 \\ &\geq \sum_{j=1}^p \frac{\alpha_j}{t} \sum_{i=1}^{\bar{t}} (\mathbf{v}_j^\top \mathbf{A} \mathbf{x})_{(i)}^2 \geq \|\mathbf{v}_j^\top \mathbf{A}\|_2^2 \cdot \sum_{j=1}^p \frac{\alpha_j}{t} \sum_{i=1}^{\bar{t}} \left( \frac{\mathbf{v}_j^\top \mathbf{A}}{\|\mathbf{v}_j^\top \mathbf{A}\|_2} \mathbf{x} \right)_{(i)}^2 \end{aligned}$$

Then from Lemma 7.8 and Lemma 7.9 (Note that we assume  $\mathbf{v}_j^\top \mathbf{A} \neq 0$  in the last inequality.

We ignore the case that  $\mathbf{v}_j^\top \mathbf{A} = 0$  since the bound holds trivially),

$$\begin{aligned} &\frac{1}{t} \sum_{i=1}^{\bar{t}} \langle \mathbf{z}\mathbf{z}^\top, \mathbf{X} \rangle_{(i)} \\ &\geq \sum_{j=1}^p \alpha_j \|\mathbf{v}_j^\top \mathbf{A}\|_2^2 \left( \mathcal{V}^- \left( \frac{\bar{t}}{t} \right) - \epsilon_1 \left( \frac{\bar{t}}{t} \right) \right) - 2\|\mathbf{X}^{1/2} \mathbf{A}\|_F \sqrt{(1 + \epsilon_0)ck \min\{\tau, \gamma_S\}} \\ &= \text{tr}(\mathbf{A}^\top \cdot \sum_{j=1}^p \alpha_j \mathbf{v}_j \mathbf{v}_j^\top \cdot \mathbf{A}) \left( \mathcal{V}^- \left( \frac{\bar{t}}{t} \right) - \epsilon_1 \left( \frac{\bar{t}}{t} \right) \right) - 2\|\mathbf{X}^{1/2} \mathbf{A}\|_F \sqrt{(1 + \epsilon_0)ck \min\{\tau, \gamma_S\}} \\ &= \|\mathbf{X}^{1/2} \mathbf{A}\|_F^2 \left( \mathcal{V}^- \left( \frac{\bar{t}}{t} \right) - \epsilon_1 \left( \frac{\bar{t}}{t} \right) \right) - 2\|\mathbf{X}^{1/2} \mathbf{A}\|_F \sqrt{(1 + \epsilon_0)ck \min\{\tau, \gamma_S\}} \end{aligned}$$

Hence the theorem holds.  $\square$

**Proof of Lemma 7.6**

*Proof.* Since  $\mathcal{E}(s)$  is true, we have  $\sum_{i \in \mathcal{Z}} \alpha_i(s) \langle \mathbf{y}_i \mathbf{y}_i^\top, \mathbf{X}_s \rangle \geq \frac{1}{\kappa} \sum_{i \in \mathcal{O}} \alpha_i(s) \langle \mathbf{y}_i \mathbf{y}_i^\top, \mathbf{X}_s \rangle$ , which implies that

$$(\kappa+1) \sum_{i \in \mathcal{Z}} \alpha_i(s) \langle \mathbf{y}_i \mathbf{y}_i^\top, \mathbf{X}_s \rangle \geq \sum_{i=1}^n \alpha_i(s) \langle \mathbf{y}_i \mathbf{y}_i^\top, \mathbf{X}_s \rangle \geq (1-\eta) \left( \sum_{i=1}^n \alpha_i(s) \langle \mathbf{y}_i \mathbf{y}_i^\top, \mathbf{\Pi} \rangle - n\mu \|\mathbf{\Pi}\|_1 \right),$$

where the last inequality holds because  $\mathbf{X}_s$  is the  $(1-\eta)$ -optimal solution of the PCA-like algorithm at stage  $s$ . Note that  $\|\mathbf{\Pi}\|_1 \leq \beta \|\mathbf{\Pi}\|_F = \beta \sqrt{\text{tr}(\mathbf{\Pi}^2)} = \beta \sqrt{k}$ , then

$$\begin{aligned} & \frac{1}{t} \sum_{i=1}^t \langle \mathbf{z}_i \mathbf{z}_i^\top, \mathbf{X}_s \rangle \geq \frac{1}{t} \sum_{i \in \mathcal{Z}} \alpha_i(s) \langle \mathbf{y}_i \mathbf{y}_i^\top, \mathbf{X}_s \rangle \\ & \geq \frac{1-\eta}{\kappa+1} \left( \frac{1}{t} \sum_{i=1}^n \alpha_i(s) \langle \mathbf{y}_i \mathbf{y}_i^\top, \mathbf{\Pi} \rangle - \frac{n}{t} \mu \beta \sqrt{k} \right) \\ & \geq \frac{1-\eta}{\kappa+1} \left( \frac{1}{t} \sum_{i=1}^{t-\rho n/\kappa} \langle \mathbf{z}_i \mathbf{z}_i^\top, \mathbf{\Pi} \rangle_{(i)} - \frac{n}{t} \mu \beta \sqrt{k} \right), \end{aligned}$$

where the last inequality follows from Equation (7.8). From Theorem 7.5, the following inequality holds w.h.p,

$$\begin{aligned} & (1+\epsilon_0)H_s + 2\sqrt{(1+\epsilon_0)ck \min\{\tau, \gamma_S\}H_s} + ck \min\{\tau, \gamma_S\} \\ & \geq \frac{1-\eta}{\kappa+1} \left[ \left( \mathcal{V}^-\left(\frac{t-\rho n/\kappa}{t}\right) - \epsilon_1 \left(\frac{t-\rho n/\kappa}{t}\right) \right) \bar{H} - 2\sqrt{(1+\epsilon_0)ck \min\{\tau, \gamma_S\} \bar{H}} - \frac{n}{n-\rho n} \mu \beta \sqrt{k} \right] \\ & = \frac{1-\eta}{\kappa+1} \left[ \left( \mathcal{V}^-\left(1 - \frac{\rho}{(1-\rho)\kappa}\right) - \epsilon_1 \left(1 - \frac{\rho}{(1-\rho)\kappa}\right) \right) \bar{H} - 2\sqrt{(1+\epsilon_0)ck \min\{\tau, \gamma_S\} \bar{H}} - \frac{1}{1-\rho} \mu \beta \sqrt{k} \right]. \end{aligned}$$

Hence we obtain this lemma.  $\square$

**Proof of Lemma 7.7**

*Proof.* Since  $|\mathcal{O}| = |\mathcal{Y} \setminus \mathcal{Z}| = \rho n$ , we have

$$\sum_{i=1}^{\hat{t}-\rho n} \langle \mathbf{z} \mathbf{z}^\top, \mathbf{X}^* \rangle_{(i)} \leq \sum_{i=1}^{\hat{t}} \langle \mathbf{y} \mathbf{y}^\top, \mathbf{X}^* \rangle_{(i)} \leq \sum_{i=1}^{\hat{t}} \langle \mathbf{z} \mathbf{z}^\top, \mathbf{X}^* \rangle_{(i)}.$$

Since  $\mathbf{X}^*$  is the final output of this algorithm,  $\bar{V}_{\hat{t}}(\mathbf{X}^*) \geq \bar{V}_{\hat{t}}(\mathbf{X}_s)$ . Thus,

$$\frac{1}{\hat{t}} \sum_{i=1}^{\hat{t}} \langle \mathbf{z}\mathbf{z}^\top, \mathbf{X}^* \rangle_{(i)} \geq \bar{V}_{\hat{t}}(\mathbf{X}^*) \geq \bar{V}_{\hat{t}}(\mathbf{X}_s) \geq \frac{1}{\hat{t}} \sum_{i=1}^{\hat{t}-\rho n} \langle \mathbf{z}\mathbf{z}^\top, \mathbf{X}_s \rangle_{(i)}.$$

Then from Theorem 7.5, the following inequality holds w.h.p,

$$\begin{aligned} & \left( \mathcal{V}^+\left(\frac{\hat{t}}{t}\right) + \epsilon_1\left(\frac{\hat{t}}{t}\right) \right) H^* + 2\sqrt{(1+\epsilon_0)ck \min\{\tau, \gamma_S\} H^*} + ck \min\{\tau, \gamma_S\} \\ & \geq \left( \mathcal{V}^-\left(\frac{\hat{t}-\rho n}{t}\right) - \epsilon_1\left(\frac{\hat{t}-\rho n}{t}\right) \right) H_s - 2\sqrt{(1+\epsilon_0)ck \min\{\tau, \gamma_S\} H_s}. \end{aligned}$$

Therefore, this lemma holds.  $\square$

### Proof of Theorem 7.6

*Proof.* Recall that with high probability  $\mathcal{E}(s)$  is true for  $s \leq s_0$  and notice that we can assume  $\epsilon_0 \leq 1$  for large enough  $n$ . From Lemma 7.6 and Lemma 7.7, since  $\bar{H} \geq H^*$  and  $\bar{H} \geq H_s$ , the following inequalities hold w.h.p,

$$\begin{aligned} & \frac{1-\eta}{\kappa+1} \left[ \left( \mathcal{V}^-\left(1 - \frac{\rho}{(1-\rho)\kappa}\right) - \epsilon_1\left(1 - \frac{\rho}{(1-\rho)\kappa}\right) \right) \bar{H} - 2\sqrt{(1+\epsilon_0)ck \min\{\tau, \gamma_S\} \bar{H}} - \frac{1}{1-\rho} \mu\beta\sqrt{k} \right] \\ & \leq (1+\epsilon_0)H_s + 2\sqrt{(1+\epsilon_0)ck \min\{\tau, \gamma_S\} \bar{H}} + ck \min\{\tau, \gamma_S\}, \end{aligned}$$

and

$$\left( \mathcal{V}^-\left(\frac{\hat{t}-\rho n}{t}\right) - \epsilon_1\left(\frac{\hat{t}-\rho n}{t}\right) \right) H_s \leq \left( \mathcal{V}^+\left(\frac{\hat{t}}{t}\right) + \epsilon_1\left(\frac{\hat{t}}{t}\right) \right) H^* + 4\sqrt{(1+\epsilon_0)ck \min\{\tau, \gamma_S\} \bar{H}} + ck \min\{\tau, \gamma_S\}.$$

By re-organization, we have

$$\begin{aligned} \frac{1+\kappa}{1-\eta} (1+\epsilon_0)H_s & \geq \left( \mathcal{V}^-\left(1 - \frac{\rho}{(1-\rho)\kappa}\right) - \epsilon_1\left(1 - \frac{\rho}{(1-\rho)\kappa}\right) \right) \bar{H} - \frac{2\kappa+4}{1-\eta} \sqrt{(1+\epsilon_0)ck \min\{\tau, \gamma_S\} \bar{H}} - \\ & \quad \frac{\mu\beta\sqrt{k}}{1-\rho} - \frac{1+\kappa}{1-\eta} ck \min\{\tau, \gamma_S\} \end{aligned}$$

$$\begin{aligned} \left( \mathcal{V}^-\left(\frac{\hat{t}-\rho n}{t}\right) - \epsilon_1\left(\frac{\hat{t}-\rho n}{t}\right) \right) H_s & \leq \left( \mathcal{V}^+\left(\frac{\hat{t}}{t}\right) + \epsilon_1\left(\frac{\hat{t}}{t}\right) \right) H^* + \\ & \quad 4\sqrt{(1+\epsilon_0)ck \min\{\tau, \gamma_S\} \bar{H}} + ck \min\{\tau, \gamma_S\}. \end{aligned}$$

Let  $\epsilon_1 = c_2 \theta^{\frac{1}{2}} d^{\frac{1}{4}} (\log n)^{\frac{3}{4}} n^{-\frac{1}{4}}$ . Since  $\epsilon_1(\frac{\hat{t}-\rho n}{t}) \leq \epsilon_1$  and  $\epsilon_1(\frac{t-s_0}{t}) \leq \epsilon_1$ , we have

$$\begin{aligned} \frac{H^*}{\bar{H}} &\geq \frac{(1-\eta) \left( \mathcal{V}^-\left(1 - \frac{\rho}{(1-\rho)\kappa}\right) - \epsilon_1 \right) \left( \mathcal{V}^-\left(\frac{\hat{t}}{t} - \frac{\rho}{1-\rho}\right) - \epsilon_1 \right)}{(1+\kappa)(1+\epsilon_0) \left( \mathcal{V}^+\left(\frac{\hat{t}}{t}\right) + \epsilon_1 \right)} \\ &\quad - \frac{(2\kappa+4) \left( \mathcal{V}^-\left(\frac{\hat{t}}{t} - \frac{\rho}{1-\rho}\right) - \epsilon_1 \right) \sqrt{(1+\epsilon_0)ck \min\{\tau, \gamma_S\}}}{(1+\kappa)(1+\epsilon_0) \left( \mathcal{V}^+\left(\frac{\hat{t}}{t}\right) + \epsilon_1 \right)} \bar{H}^{-1/2} \\ &\quad - \frac{4(1+\kappa)(1+\epsilon_0) \sqrt{(1+\epsilon_0)ck \min\{\tau, \gamma_S\}}}{(1+\kappa)(1+\epsilon_0) \left( \mathcal{V}^+\left(\frac{\hat{t}}{t}\right) + \epsilon_1 \right)} \bar{H}^{-1/2} - \frac{\left( \mathcal{V}^-\left(\frac{\hat{t}}{t} - \frac{\rho}{1-\rho}\right) - \epsilon_1 + 1 + \epsilon_0 \right) ck \min\{\tau, \gamma_S\}}{(1+\epsilon_0) \left( \mathcal{V}^+\left(\frac{\hat{t}}{t}\right) + \epsilon_1 \right)} \bar{H}^{-1} \\ &\quad - \frac{(1-\eta) \frac{\mu\beta\sqrt{k}}{1-\rho}}{(1+\kappa)(1+\epsilon_0) \left( \mathcal{V}^+\left(\frac{\hat{t}}{t}\right) + \epsilon_1 \right)} \bar{H}^{-1}. \end{aligned}$$

Note that the last term

$$\frac{(1-\eta) \frac{\mu\beta\sqrt{k}}{1-\rho}}{(1+\kappa)(1+\epsilon_0) \left( \mathcal{V}^+\left(\frac{\hat{t}}{t}\right) + \epsilon_1 \right)} \bar{H}^{-1} \leq \frac{(1-\eta) \frac{\mu\beta\sqrt{k}}{1-\rho}}{\mathcal{V}^+\left(\frac{\hat{t}}{t}\right)} \bar{H}^{-1} \leq \frac{2(1-\eta)\mu\beta\sqrt{k}}{\mathcal{V}^+\left(\frac{\hat{t}}{t}\right)} \bar{H}^{-1}.$$

Since  $\epsilon_0 = c_1 \theta \sqrt{\frac{d}{n}} = \epsilon_0$ ,  $\epsilon_1(\frac{\hat{t}}{t}) = \frac{c_2 t(1+\epsilon_0) \sqrt{d \log n/n}}{t-\hat{t}} \wedge c_2 \theta^{\frac{1}{2}} d^{\frac{1}{4}} (\log n)^{\frac{3}{4}} n^{-\frac{1}{4}}$ , and  $\mathcal{V}_v(\kappa) - \mathcal{V}_v(\kappa - \epsilon) \leq C\theta\epsilon \log \epsilon$  by Lemma 7.10, we can follow the proof of Theorem 2 in [XCM13] and obtain that the following inequality holds w.h.p,

$$\begin{aligned} \frac{H^*}{\bar{H}} &\geq \frac{(1-\eta) \mathcal{V}^-\left(\frac{\hat{t}}{t} - \frac{\rho}{1-\rho}\right) \mathcal{V}^-\left(1 - \frac{\rho}{\kappa(1-\rho)}\right)}{(1+\kappa) \mathcal{V}^+\left(\frac{\hat{t}}{t}\right)} - \frac{10}{\mathcal{V}^+(0.5)} \left( \frac{ck \min\{\tau, \gamma_S\}}{\bar{H}} \right)^{1/2} \\ &\quad - \frac{c\{\theta^{\frac{1}{2}} d^{\frac{1}{4}} (\log^{\frac{3}{4}} n) n^{-\frac{1}{4}} \vee \theta[(1+\kappa)/\kappa]^{\frac{3}{2}} (\log^{\frac{3}{2}} n) n^{-\frac{1}{2}}\}}{\mathcal{V}^+(0.5)} - \frac{2(1-\eta)\mu\beta\sqrt{k}}{\mathcal{V}^+\left(\frac{\hat{t}}{t}\right) \bar{H}}. \end{aligned}$$

Hence this theorem holds.  $\square$



**Proof of Corollary 7.4**

*Proof.* When  $\kappa > 1$ , the corollary holds trivially. Hence, fix  $\kappa \leq 1$ . From Theorem 7.4, we have

$$\begin{aligned}
& \liminf_k \mathbb{E} \cdot \mathbb{V} \{ \mathbf{w}_1^*, \dots, \mathbf{w}_k^* \} \\
& \geq (1 - \eta) \max_{\kappa} \left[ \frac{\mathcal{V}^- \left( 1 - \frac{\rho^*}{(1-\rho^*)\kappa} \right)}{1 + \kappa} \right] \times \left[ \frac{\mathcal{V}^- \left( \frac{\hat{t}}{t} - \frac{\rho^*}{1-\rho^*} \right)}{\mathcal{V}^+ \left( \frac{\hat{t}}{t} \right)} \right] \\
& \geq (1 - \eta) \max_{\kappa} \left[ \frac{1}{1 + \kappa} - \frac{C\theta \frac{\rho^*}{(1-\rho^*)\kappa} \log \frac{(1-\rho^*)\kappa}{\rho^*}}{1 + \kappa} \right] \times \left[ \frac{\mathcal{V}^- \left( \frac{\hat{t}}{t} \right)}{\mathcal{V}^+ \left( \frac{\hat{t}}{t} \right)} - \frac{C\theta \frac{\rho^*}{1-\rho^*} \log \frac{1-\rho^*}{\rho^*}}{\mathcal{V}^+ \left( \frac{\hat{t}}{t} \right)} \right] \\
& \geq (1 - \eta) \max_{\kappa} \left[ 1 - \kappa - \frac{C\theta \rho^*}{\kappa} \log \frac{1}{2\rho^*} \right] \times \left[ \frac{\mathcal{V}^- \left( \frac{\hat{t}}{t} \right)}{\mathcal{V}^+ \left( \frac{\hat{t}}{t} \right)} - \frac{C\theta \rho^* \log \frac{1}{2\rho^*}}{\mathcal{V}^+ \left( \frac{\hat{t}}{t} \right)} \right] \\
& \geq (1 - \eta) \max_{\kappa} \left[ 1 - \kappa - \frac{C\theta \rho^*}{\kappa} \log \frac{1}{2\rho^*} \right] \times \left[ 1 - \frac{C\theta \rho^* \log \frac{1}{2\rho^*}}{\mathcal{V}^- \left( \frac{\hat{t}}{t} \right)} \right] \times \frac{\mathcal{V}^- \left( \frac{\hat{t}}{t} \right)}{\mathcal{V}^+ \left( \frac{\hat{t}}{t} \right)} \\
& \geq (1 - \eta) \max_{\kappa} \left[ 1 - \kappa - \left( \frac{1}{\kappa} + \frac{1}{\mathcal{V}^- \left( \frac{\hat{t}}{t} \right)} \right) C\theta \rho^* \log \frac{1}{2\rho^*} \right] \times \frac{\mathcal{V}^- \left( \frac{\hat{t}}{t} \right)}{\mathcal{V}^+ \left( \frac{\hat{t}}{t} \right)} \\
& \geq (1 - \eta) \max_{\kappa} \left[ 1 - \kappa - \frac{C\theta \rho^* \log \frac{1}{2\rho^*}}{\kappa \mathcal{V}^- \left( \frac{\hat{t}}{t} \right)} \right] \times \frac{\mathcal{V}^- \left( \frac{\hat{t}}{t} \right)}{\mathcal{V}^+ \left( \frac{\hat{t}}{t} \right)}.
\end{aligned}$$

The second inequality is due to Lemma 7.10 and  $\mathcal{V}^-(1) = 1$ . The third inequality is due to  $\rho^* < 0.5$  and  $\kappa \leq 1$ . The sixth inequality holds because  $\kappa \leq 1$  and  $\mathcal{V}^- \left( \frac{\hat{t}}{t} \right) \leq 1$ . Taking  $\kappa = \sqrt{\theta \rho^* \log \frac{1}{2\rho^*}}$ , we can obtain this corollary.  $\square$

**7.7 Additional Lemmas****Concentration Results for Isotropic Random Vectors**

**Lemma 7.8.** (Lemma 2, [XCM13]) For any  $0 \leq a_1 < a_2 < a_3 \leq 1$  and  $\mathbf{v} \in \mathcal{S}_d$ , we have

$$\frac{\mathcal{V}_{\mathbf{v}}(a_2) - \mathcal{V}_{\mathbf{v}}(a_1)}{a_2 - a_1} \leq \frac{\mathcal{V}_{\mathbf{v}}(a_3) - \mathcal{V}_{\mathbf{v}}(a_2)}{a_3 - a_2}.$$

**Lemma 7.9.** (Lemma 3, [XCM13]) 1) For any  $a \in [0, 1]$  and  $\mathbf{v} \in \mathcal{S}_d$ , we have  $\mathcal{V}_{\mathbf{v}}(a) \leq a$ .  
 2) For any  $0 \leq a_1 < a_2 \leq 1$  and  $\mathbf{v} \in \mathcal{S}_d$ , we have

$$\mathcal{V}_{\mathbf{v}}(a_2) - \mathcal{V}_{\mathbf{v}}(a_1) \leq \frac{a_2 - a_1}{1 - a_1}.$$

**Lemma 7.10.** For any  $1 > \epsilon > 0$  and  $\kappa \in [\epsilon, 1]$  and  $\mathbf{v} \in \mathcal{S}_d$ , we have  $\mathcal{V}_{\mathbf{v}}(\kappa) - \mathcal{V}_{\mathbf{v}}(\kappa - \epsilon) \leq C\theta\epsilon \log(1/\epsilon)$ .

*Proof.* By monotonicity, it suffices to prove the result for  $\kappa = 1$ . Notice that for  $K \geq 2\theta$ ,

$$\begin{aligned} & \mathcal{V}_{\mathbf{v}}(1) - \mathcal{V}_{\mathbf{v}}(1 - \epsilon) \\ & \leq \epsilon K^2 + \mathbb{E}_{x \sim \bar{\mu}_{\mathbf{v}}}[x^2 \cdot \mathbf{1}(x > K)] \\ & = \epsilon K^2 + \int_{K^2}^{\infty} \mathbb{P}_{x \sim \bar{\mu}_{\mathbf{v}}}[x^2 > z] dz \\ & \leq \epsilon K^2 + \int_{K^2}^{\infty} \exp(1 - z/\theta) dz \\ & = \epsilon K^2 + e_0 \theta \exp(-K^2/\theta) \end{aligned}$$

Let  $K^2 = \theta \log(1/\epsilon)$ , then we have  $\mathcal{V}_{\mathbf{v}}(1) - \mathcal{V}_{\mathbf{v}}(1 - \epsilon) \leq C\theta\epsilon \log(1/\epsilon)$ .  $\square$

**Theorem 7.7.** (Theorem 7(I), [XCM13]) Suppose random vector  $\mathbf{n}_i \sim \mathcal{N}(0, I_p)$ . Let  $\tau \triangleq \max\{p/n, 1\}$ . There exist a universal constant  $c > 0$  such that with high probability,

$$\sup_{\mathbf{w} \in \mathcal{S}_p} \frac{1}{t} \sum_{i=1}^t (\mathbf{w}^\top \mathbf{n}_i)^2 \leq c\tau.$$

**Theorem 7.8.** There exists an absolute constant  $C > 0$ , such that with high probability,

$$\sup_{\mathbf{v} \in \mathcal{S}_d} \left| \frac{1}{n} \sum_{i=1}^n (\mathbf{v}^\top \mathbf{x}_i)^2 - 1 \right| \leq C\theta \sqrt{\frac{d}{n}}.$$

*Proof.* The proof depends on the following Lemma (Lemma 14 in [LW12]).

**Lemma 7.11.** If  $\mathbf{X} \in \mathbb{R}^{n \times d}$  is a zero-mean sub-Gaussian matrix with parameters  $(\Sigma, \sigma^2)$ , then for any fixed (unit) vector  $\mathbf{v} \in \mathbb{R}^d$  and any  $t > 0$ , we have

$$\mathbb{P}[|\|\mathbf{X}\mathbf{v}\|_2^2 - \mathbb{E}[\|\mathbf{X}\mathbf{v}\|_2^2]| > nt] \leq 2 \exp\left(-cn \min\left(\frac{t^2}{\sigma^4}, \frac{t}{\sigma^2}\right)\right),$$

for a universal constant  $c$ .

Consider matrix  $\mathbf{Z} \in \mathbb{R}^{n \times d}$  where the  $i^{\text{th}}$  row is  $\mathbf{x}_i^\top$ , then for any fixed (unit) vector  $\mathbf{v} \in \mathbb{R}^d$  and any  $t > 0$ , there exists a universal constant  $c$  such that

$$\mathbb{P}[|\|\mathbf{Z}\mathbf{v}\|_2^2 - \mathbb{E}[\|\mathbf{Z}\mathbf{v}\|_2^2]| > nt] \leq 2 \exp\left(-cn \min\left(\frac{t^2}{\theta^2}, \frac{t}{\theta}\right)\right).$$

Let  $\mathcal{A}$  be a  $1/3$  cover of  $\mathcal{S}_d$ , then for any  $\mathbf{v} \in \mathcal{S}_d$ , there is some  $\mathbf{u} \in \mathcal{A}$  such that  $\|\mathbf{u} - \mathbf{v}\|_2 \leq 1/3$ .

It is known that  $|\mathcal{A}| \leq 9^d$ . Define  $\psi(\mathbf{v}_1, \mathbf{v}_2) = |\mathbf{v}_1^\top \left( \frac{\mathbf{Z}^\top \mathbf{Z}}{n} - \frac{\mathbb{E}[\mathbf{Z}^\top \mathbf{Z}]}{n} \right) \mathbf{v}_2|$ , then we have

$$\sup_{\mathbf{v} \in \mathcal{S}_d} \psi(\mathbf{v}, \mathbf{v}) \leq \max_{\mathbf{u} \in \mathcal{A}} \psi(\mathbf{u}, \mathbf{u}) + 2 \sup_{\mathbf{v} \in \mathcal{S}_d} \psi(\mathbf{v} - \mathbf{u}, \mathbf{u}) + \sup_{\mathbf{v} \in \mathcal{S}_d} \psi(\mathbf{v} - \mathbf{u}, \mathbf{v} - \mathbf{u}).$$

Since  $\|\mathbf{u} - \mathbf{v}\|_2 \leq \frac{1}{3}$ , we have

$$\sup_{\mathbf{v} \in \mathcal{S}_d} \psi(\mathbf{v}, \mathbf{v}) \leq \max_{\mathbf{u} \in \mathcal{A}} \psi(\mathbf{u}, \mathbf{u}) + \left(\frac{2}{3} + \frac{1}{9}\right) \sup_{\mathbf{v} \in \mathcal{S}_d} \psi(\mathbf{v}, \mathbf{v}).$$

Hence  $\sup_{\mathbf{v} \in \mathcal{S}_d} \psi(\mathbf{v}, \mathbf{v}) \leq \frac{9}{2} \max_{\mathbf{u} \in \mathcal{A}} \psi(\mathbf{u}, \mathbf{u})$ . By the lemma above and the union bound,

$$\mathbb{P}[\sup_{\mathbf{v} \in \mathcal{S}_d} \psi(\mathbf{v}, \mathbf{v}) > t] \leq \mathbb{P}\left[\frac{9}{2} \max_{\mathbf{u} \in \mathcal{A}} \psi(\mathbf{u}, \mathbf{u}) > t\right] \leq 9^d \cdot 2 \exp\left(-cn \min\left(\frac{t^2}{\theta^2}, \frac{t}{\theta}\right)\right).$$

Thus, we have

$$\mathbb{P}\left[\sup_{\mathbf{v} \in \mathcal{S}_d} \left|\frac{1}{n} \sum_{i=1}^n (\mathbf{v}^\top \mathbf{x}_i)^2 - 1\right| > t\right] \leq 2 \exp\left(-c_1 n \min\left(\frac{t^2}{\theta^2}, \frac{t}{\theta}\right) + c_2 d\right).$$

Let the right hand side be  $d^{-10}$ , then  $t = C\theta\sqrt{\frac{d}{n}}$  for constant  $C$  and large enough  $n$ .  $\square$

**Lemma 7.12.** *With high probability, the following holds uniformly over  $\bar{n} \leq n$  and  $\mathbf{v} \in \mathcal{S}_d$ ,*

$$\left| \frac{1}{n} \sum_{i=1}^{\bar{n}} [\mathbf{v}^\top \mathbf{x}]_{(i)}^2 - \mathcal{V}_{\mathbf{v}}(\bar{n}/n) \right| \leq \frac{Cn\sqrt{d \log n/n}}{n - \bar{n}},$$

for a universal constant  $C$ .

*Proof.* The proof is similar to the proof of Theorem 11 [XCM13]. We just need to replace  $\mathcal{V}$  with  $\mathcal{V}_{\mathbf{v}}$ .  $\square$

**Theorem 7.9.** *With high probability, the following holds uniformly over  $\bar{n} \leq n$  and  $\mathbf{v} \in \mathcal{S}_d$ ,*

$$\left| \frac{1}{\bar{n}} \sum_{i=1}^{\bar{n}} [\mathbf{v}^\top \mathbf{x}]_{(i)}^2 - \mathcal{V}_{\mathbf{v}}(\bar{n}/n) \right| \leq C \max \left\{ \theta \sqrt{\frac{d}{n}}, \theta^{1/2} d^{1/4} (\log n)^{3/4} n^{-1/4} \right\},$$

for a universal constant  $C$ .

*Proof.* Follow the proof of Corollary 5 in [XCM13]. As shown above, Theorem 7.8 and Lemma 7.12 hold w.h.p. Under the condition of Theorem 7.8 and Lemma 7.12, we define  $n_0$

$$n_0 = (1 - \Theta(\theta^{-1/2} d^{1/4} n^{-1/4} \log^{-1/4} n))n.$$

If  $\bar{n} \leq n_0$ , then Lemma 7.12 leads to

$$\left| \frac{1}{\bar{n}} \sum_{i=1}^{\bar{n}} [\mathbf{v}^\top \mathbf{x}]_{(i)}^2 - \mathcal{V}_{\mathbf{v}}(\bar{n}/n) \right| \leq C \theta^{1/2} d^{1/4} (\log n)^{3/4} n^{-1/4}.$$

If  $\bar{n} > n_0$ , we have

$$\begin{aligned} & \frac{1}{\bar{n}} \sum_{i=1}^{\bar{n}} [\mathbf{v}^\top \mathbf{x}]_{(i)}^2 - \mathcal{V}_{\mathbf{v}}(\bar{n}/n) \\ & \leq \left| \frac{1}{\bar{n}} \sum_{i=1}^{\bar{n}} [\mathbf{v}^\top \mathbf{x}]_{(i)}^2 - 1 \right| + |1 - \mathcal{V}_{\mathbf{v}}(\bar{n}/n)| \\ & \leq C_1 \theta \sqrt{\frac{d}{n}} + C_2 \theta \frac{n - n_0}{n} \log \frac{n}{n - n_0} \\ & \leq C \max \left\{ \theta \sqrt{\frac{d}{n}}, \theta^{1/2} d^{1/4} (\log n)^{3/4} n^{-1/4} \right\}. \end{aligned}$$

On the other hand,

$$\begin{aligned} & \mathcal{V}_{\mathbf{v}}(\bar{n}/n) - \frac{1}{\bar{n}} \sum_{i=1}^{\bar{n}} [\mathbf{v}^\top \mathbf{x}]_{(i)}^2 \\ & \leq \mathcal{V}_{\mathbf{v}}(\bar{n}/n) - \frac{1}{\bar{n}} \sum_{i=1}^{n_0} [\mathbf{v}^\top \mathbf{x}]_{(i)}^2 \\ & \leq \left| \frac{1}{\bar{n}} \sum_{i=1}^{n_0} [\mathbf{v}^\top \mathbf{x}]_{(i)}^2 - \mathcal{V}_{\mathbf{v}}(n_0/n) \right| + |\mathcal{V}_{\mathbf{v}}(n_0/n) - \mathcal{V}_{\mathbf{v}}(\bar{n}/n)| \\ & \leq C_1 \frac{n \sqrt{d \log n/n}}{n - n_0} + C_2 \theta \frac{n - n_0}{n} \log \frac{n}{n - n_0} \\ & \leq C \theta^{1/2} d^{1/4} (\log n)^{3/4} n^{-1/4}. \end{aligned}$$

Hence this theorem holds.  $\square$

### Concentration Results for Non-isotropic Random Vectors

**Lemma 7.13.** *There exists a constant  $c > 0$  such that with high probability,*

$$\sup_{\mathbf{v} \in \mathcal{S}_p} \left| \frac{1}{t} \sum_{i=1}^t (\mathbf{v}^\top \mathbf{z}_i)^2 - \mathbf{v}^\top \boldsymbol{\Sigma} \mathbf{v} \right| \leq c\zeta \|\boldsymbol{\Sigma}\|_2 \sqrt{\frac{p}{n}}.$$

*Proof.* Recall that  $\mathbf{z} = [\mathbf{A}, \mathbf{I}_p] \mathbf{u}$  where  $\mathbf{u}$  is a sub-gaussian random variable with mean zero and variance  $\mathbf{I}_{p+d}$ . Denote  $(\mathbf{A}, \mathbf{I}_p)^\top$  by  $\bar{\mathbf{A}}$ , since  $\bar{\mathbf{A}}^\top \bar{\mathbf{A}} = \boldsymbol{\Sigma}$  and  $d \leq p$ , then with high probability

$$\begin{aligned} & \sup_{\mathbf{v} \in \mathcal{S}_p} \left| \frac{1}{t} \sum_{i=1}^n (\mathbf{v}^\top \mathbf{z}_i)^2 - \mathbf{v}^\top \boldsymbol{\Sigma} \mathbf{v} \right| = \sup_{\mathbf{v} \in \mathcal{S}_p} \left| \frac{1}{t} \sum_{i=1}^n ((\bar{\mathbf{A}} \mathbf{v})^\top \mathbf{u})^2 - \mathbf{v}^\top \bar{\mathbf{A}}^\top \bar{\mathbf{A}} \mathbf{v} \right| \\ & \leq \sup_{\mathbf{v} \in \mathcal{S}_p \cap \{\mathbf{v}: \bar{\mathbf{A}} \mathbf{v} \neq \mathbf{0}\}} \|\bar{\mathbf{A}}^\top \bar{\mathbf{A}}\|_2 \cdot \left| \frac{1}{n} \sum_{i=1}^n \frac{((\bar{\mathbf{A}} \mathbf{v})^\top \mathbf{u})^2}{\|\bar{\mathbf{A}} \mathbf{v}\|^2} - 1 \right| \leq \|\boldsymbol{\Sigma}\|_2 \cdot \sup_{\mathbf{q} \in \mathcal{S}_{p+d}} \left| \frac{1}{n} \sum_{i=1}^n (\mathbf{q}^\top \mathbf{u})^2 - 1 \right| \\ & \leq c\zeta \|\boldsymbol{\Sigma}\|_2 \sqrt{\frac{p}{n}}. \end{aligned}$$

$\square$

**Lemma 7.14.** *Let  $\tau \triangleq \max\{p/n, 1\}$ . There exists a constant  $C > 0$  such that with high probability,*

$$\sup_{\mathbf{v} \in \mathcal{S}_p} \frac{1}{t} \sum_{i=1}^t (\mathbf{v}^\top \mathbf{z}_i)^2 \leq 2\|\mathbf{A}\|_2^2 (1 + c_1 \theta \sqrt{\frac{d}{n}}) + c_2 \tau.$$

*Proof.* Consider the following inequalities

$$\begin{aligned}
& \sup_{\mathbf{v} \in \mathcal{S}_p} \frac{1}{t} \sum_{i=1}^t (\mathbf{v}^\top \mathbf{z}_i)^2 = \sup_{\mathbf{v} \in \mathcal{S}_p} \frac{1}{t} \sum_{i=1}^t (\mathbf{v}^\top (\mathbf{A} \mathbf{x}_i + \mathbf{n}_i))^2 \\
& \leq \sup_{\mathbf{v} \in \mathcal{S}_p} \frac{1}{t} \sum_{i=1}^t (\mathbf{v}^\top \mathbf{A} \mathbf{x}_i)^2 + \sup_{\mathbf{v} \in \mathcal{S}_p} \frac{1}{t} \sum_{i=1}^t (\mathbf{v}^\top \mathbf{n}_i)^2 + \sup_{\mathbf{v} \in \mathcal{S}_p} \frac{2}{t} \sum_{i=1}^t (\mathbf{v}^\top \mathbf{A} \mathbf{x}_i \mathbf{n}_i^\top \mathbf{v}) \\
& \leq 2 \left( \sup_{\mathbf{v} \in \mathcal{S}_p} \frac{1}{t} \sum_{i=1}^t (\mathbf{v}^\top \mathbf{A} \mathbf{x}_i)^2 + \sup_{\mathbf{v} \in \mathcal{S}_p} \frac{1}{t} \sum_{i=1}^t (\mathbf{v}^\top \mathbf{n}_i)^2 \right) \\
& = 2 \left( \sup_{\mathbf{v} \in \mathcal{S}_p \cap \{\mathbf{v}: \mathbf{A}^\top \mathbf{v} \neq 0\}} \|\mathbf{v}^\top \mathbf{A}\|_2^2 \cdot \frac{1}{t} \sum_{i=1}^t \left( \frac{\mathbf{v}^\top \mathbf{A} \mathbf{x}_i}{\|\mathbf{v}^\top \mathbf{A}\|_2} \right)^2 + \sup_{\mathbf{v} \in \mathcal{S}_p} \frac{1}{t} \sum_{i=1}^t (\mathbf{v}^\top \mathbf{n}_i)^2 \right) \\
& \leq 2 \|\mathbf{A}\|_2^2 (1 + c_1 \theta \sqrt{\frac{d}{n}}) + c_2 \tau
\end{aligned}$$

where the last inequality follows from Theorem 7.7 and Theorem 7.8.  $\square$

**Lemma 7.15.** *Let  $\tau \triangleq \max\{p/n, 1\}$ . There exists a universal constant  $c$  such that with high probability the following holds uniformly over  $\bar{n} \leq n$ ,*

$$\sup_{\mathbf{v} \in \mathcal{S}_p} \frac{1}{n} \sum_{i=1}^{n-\bar{n}} |\mathbf{v}^\top \mathbf{z}|_{[i]}^2 \leq 2 \|\mathbf{A}\|_2^2 \left( 1 - \mathcal{V}^-(\bar{n}/n) + c_1 \theta \sqrt{\frac{d}{n}} + c_2 \theta^{1/2} d^{1/4} (\log n)^{3/4} n^{-1/4} \right) + c\tau.$$

*Proof.* Consider the following inequalities

$$\begin{aligned}
& \sup_{\mathbf{v} \in \mathcal{S}_p} \frac{1}{n} \sum_{i=1}^{n-\bar{n}} |\mathbf{v}^\top \mathbf{z}|_{[i]}^2 = \sup_{\mathbf{v} \in \mathcal{S}_p} \frac{1}{n} \sum_{i=1}^{n-\bar{n}} |\mathbf{v}^\top (\mathbf{A} \mathbf{x} + \mathbf{n})|_{[i]}^2 \\
& \leq 2 \left( \sup_{\mathbf{v} \in \mathcal{S}_p} \frac{1}{n} \sum_{i=1}^{n-\bar{n}} |\mathbf{v}^\top \mathbf{A} \mathbf{x}|_{[i]}^2 + \sup_{\mathbf{v} \in \mathcal{S}_p} \frac{1}{n} \sum_{i=1}^{n-\bar{n}} |\mathbf{v}^\top \mathbf{n}|_{[i]}^2 \right).
\end{aligned}$$

Note that

$$\begin{aligned}
& \sup_{\mathbf{v} \in \mathcal{S}_p} \frac{1}{n} \sum_{i=1}^{n-\bar{n}} |\mathbf{v}^\top \mathbf{A} \mathbf{x}|_{[i]}^2 \leq \sup_{\mathbf{v} \in \mathcal{S}_p \cap \{\mathbf{v}: \mathbf{A}^\top \mathbf{v} \neq 0\}} \|\mathbf{A} \mathbf{A}^\top\|_2 \cdot \frac{1}{n} \sum_{i=1}^{n-\bar{n}} \left| \frac{\mathbf{v}^\top \mathbf{A} \mathbf{x}}{\|\mathbf{v}^\top \mathbf{A}\|_2} \right|_{[i]}^2 \\
& = \|\mathbf{A} \mathbf{A}^\top\|_2 \cdot \sup_{\mathbf{v} \in \mathcal{S}_d} \left| \frac{1}{n} \sum_{i=1}^n |\mathbf{v}^\top \mathbf{x}|^2 - \frac{1}{n} \sum_{i=1}^{\bar{n}} |\mathbf{v}^\top \mathbf{x}|_{(i)}^2 \right|
\end{aligned}$$

From Theorem 7.8 and Theorem 7.9, we know that

$$\begin{aligned} & \sup_{\mathbf{v} \in \mathcal{S}_d} \left| \frac{1}{n} \sum_{i=1}^n |\mathbf{v}^\top \mathbf{x}|^2 - \frac{1}{n} \sum_{i=1}^{\bar{n}} |\mathbf{v}^\top \mathbf{x}|_{(i)}^2 - (1 - \mathcal{V}_{\mathbf{v}}(\bar{n}/n)) \right| \\ & \leq \sup_{\mathbf{v} \in \mathcal{S}_d} \left| \frac{1}{n} \sum_{i=1}^n |\mathbf{v}^\top \mathbf{x}|^2 - 1 \right| + \sup_{\mathbf{v} \in \mathcal{S}_d} \left| \frac{1}{n} \sum_{i=1}^{\bar{n}} |\mathbf{v}^\top \mathbf{x}|_{(i)}^2 - \mathcal{V}_{\mathbf{v}}(\bar{n}/n) \right| \\ & \leq c_1 \theta \sqrt{\frac{d}{n}} + c_2 \max \left\{ \theta \sqrt{\frac{d}{n}}, \theta^{1/2} d^{1/4} (\log n)^{3/4} n^{-1/4} \right\}, \end{aligned}$$

which implies that

$$\sup_{\mathbf{v} \in \mathcal{S}_p} \frac{1}{n} \sum_{i=1}^{n-\bar{n}} |\mathbf{v}^\top \mathbf{A} \mathbf{x}|_{[i]}^2 \leq \|\mathbf{A}\|_2^2 \cdot \left( 1 - \mathcal{V}^-(\bar{n}/n) + c_1 \theta \sqrt{\frac{d}{n}} + c_2 \theta^{1/2} d^{1/4} (\log n)^{3/4} n^{-1/4} \right). \quad (7.10)$$

Similarly, for the term  $\sup_{\mathbf{v} \in \mathcal{S}_p} \frac{1}{n} \sum_{i=1}^{n-\bar{n}} |\mathbf{v}^\top \mathbf{n}|_{[i]}^2$ , since  $n \sim \mathcal{N}(0, \mathbf{I}_p)$ , from Theorem 7.7 we have

$$\sup_{\mathbf{v} \in \mathcal{S}_p} \frac{1}{n} \sum_{i=1}^{n-\bar{n}} |\mathbf{v}^\top \mathbf{n}|_{[i]}^2 \leq \sup_{\mathbf{v} \in \mathcal{S}_p} \frac{1}{n} \sum_{i=1}^n |\mathbf{v}^\top \mathbf{n}|^2 \leq c\tau.$$

Hence we obtain this theorem.  $\square$

**Lemma 7.16.** *With high probability the following holds uniformly over  $\bar{n} \leq n$  for every matrix  $\mathbf{X} \in \mathbb{R}^{p \times p}$ ,*

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n-\bar{n}} \langle \mathbf{z} \mathbf{z}^\top, \mathbf{X} \rangle_{[i]} & \leq \min \left\{ [2(1 - \mathcal{V}^-(\bar{n}/n) + \epsilon(d)) \|\mathbf{A}\|_2^2 + c\tau] \|\mathbf{X}\|_*, \right. \\ & \left. [(1 - \mathcal{V}^-(\bar{n}/n) + \epsilon(d)) \|\mathbf{A}\|_2^2 + c\phi(1 + \zeta d \|\mathbf{A}\|_2)] \|\mathbf{X}\|_1 \right\}, \end{aligned}$$

where  $\epsilon(d) = c_1 \theta \sqrt{\frac{d}{n}} + c_2 \theta^{1/2} d^{1/4} (\log n)^{3/4} n^{-1/4}$ ,  $\tau = \max\{1, \frac{p}{n}\}$ ,  $\phi = \max\{1, \sqrt{\frac{\log p}{n}}\}$  and  $c, c_1, c_2$  are universal constants.

*Proof.* Let  $\{k(i)\}$  be the indices of the largest  $n - \bar{n}$  values of  $\langle \mathbf{z} \mathbf{z}^\top, \mathbf{X} \rangle$ , then

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^{n-\bar{n}} \langle \mathbf{z} \mathbf{z}^\top, \mathbf{X} \rangle_{[i]} = \left\langle \frac{1}{n} \sum_{i=1}^{n-\bar{n}} \mathbf{z}_{k(i)} \mathbf{z}_{k(i)}^\top, \mathbf{X} \right\rangle \\ & \leq \min \left\{ \left\| \frac{1}{n} \sum_{i=1}^{n-\bar{n}} \mathbf{z}_{k(i)} \mathbf{z}_{k(i)}^\top \right\|_\infty \|\mathbf{X}\|_1, \left\| \frac{1}{n} \sum_{i=1}^{n-\bar{n}} \mathbf{z}_{k(i)} \mathbf{z}_{k(i)}^\top \right\|_2 \|\mathbf{X}\|_* \right\}. \end{aligned}$$

Notice that  $\|\frac{1}{n} \sum_{i=1}^{n-\bar{n}} \mathbf{z}_{k(i)} \mathbf{z}_{k(i)}^\top\|_2$  can be bounded by Lemma 7.15, so we only need to bound  $\|\frac{1}{n} \sum_{i=1}^{n-\bar{n}} \mathbf{z}_{k(i)} \mathbf{z}_{k(i)}^\top\|_\infty$ . We have

$$\begin{aligned} & \left\| \frac{1}{n} \sum_{i=1}^{n-\bar{n}} \mathbf{z}_{k(i)} \mathbf{z}_{k(i)}^\top \right\|_\infty \\ & \leq \left\| \frac{1}{n} \sum_{i=1}^{n-\bar{n}} \mathbf{A} \mathbf{x}_{k(i)} \mathbf{x}_{k(i)}^\top \mathbf{A}^\top \right\|_\infty + \left\| \frac{1}{n} \sum_{i=1}^{n-\bar{n}} \mathbf{n}_{k(i)} \mathbf{n}_{k(i)}^\top \right\|_\infty + \left\| \frac{2}{n} \sum_{i=1}^{n-\bar{n}} \mathbf{A} \mathbf{x}_{k(i)} \mathbf{n}_{k(i)}^\top \right\|_\infty \\ & \leq \left\| \frac{1}{n} \sum_{i=1}^{n-\bar{n}} \mathbf{A} \mathbf{x}_{k(i)} \mathbf{x}_{k(i)}^\top \mathbf{A}^\top \right\|_2 + \left\| \frac{1}{n} \sum_{i=1}^n |\mathbf{n}_i| |\mathbf{n}_i|^\top \right\|_\infty + 2d \|\mathbf{A}\|_\infty \left\| \frac{1}{n} \sum_{i=1}^n |\mathbf{x}_i| |\mathbf{n}_i|^\top \right\|_\infty. \end{aligned}$$

Let  $\mathbf{P} \triangleq \frac{1}{n} \sum_{i=1}^n |\mathbf{n}_i| |\mathbf{n}_i|^\top$  and  $\mathbf{Q} \triangleq \frac{1}{n} \sum_{i=1}^n |\mathbf{x}_i| |\mathbf{n}_i|^\top$ , then

$$\mathbf{P}_{ij} = \frac{1}{n} \sum_{k=1}^n |\mathbf{n}_{ki} \mathbf{n}_{kj}| \leq \frac{2}{n} \sum_{k=1}^n (\mathbf{n}_{ki}^2 + \mathbf{n}_{kj}^2), \quad \mathbf{Q}_{ij} = \frac{1}{n} \sum_{k=1}^n |\mathbf{x}_{ki} \mathbf{n}_{kj}| \leq \frac{2}{n} \sum_{k=1}^n (\mathbf{x}_{ki}^2 + \mathbf{n}_{kj}^2).$$

Since  $\mathbf{x}_{ki}$  is a zero-mean sub-Gaussian random variable and  $\mathbf{n}_{ki}$  is a standard Gaussian random variable, from Proposition 5.16 [Ver12]

$$\begin{aligned} \mathbb{P} \left[ \left| \frac{1}{n} \sum_{k=1}^n \mathbf{n}_{ki}^2 - 1 \right| > t \right] & \leq 2 \exp(-c_1 \min\{nt^2, nt\}), \text{ and} \\ \mathbb{P} \left[ \left| \frac{1}{n} \sum_{k=1}^n \mathbf{x}_{ki}^2 - 1 \right| > t \right] & \leq 2 \exp(-c_2 \min\{\frac{nt^2}{\theta^2}, \frac{nt}{\theta}\}) \end{aligned}$$

for universal constant  $c_1, c_2$ . There exists  $c$  (may change from line to line) so that when  $t = c\sqrt{\frac{\log p}{n}}$ , by the union bound we have

$$\mathbb{P} \left[ \|\mathbf{P}\|_\infty > 1 + c\sqrt{\frac{\log p}{n}} \right] \leq p^{-10}, \text{ and } \mathbb{P} \left[ \|\mathbf{Q}\|_\infty > \zeta(1 + c\sqrt{\frac{\log p}{n}}) \right] \leq p^{-10}$$

where  $\zeta = \max\{\theta, 2\}$ . Let  $\phi = \max\{1, \frac{\log p}{n}\}$ , then with high probability

$$\|\mathbf{P}\|_\infty \leq c\phi, \text{ and } \|\mathbf{Q}\|_\infty \leq c\zeta\phi.$$

Thus,

$$\left\| \frac{1}{n} \sum_{i=1}^{n-\bar{n}} \mathbf{z}_{k(i)} \mathbf{z}_{k(i)}^\top \right\|_\infty \leq \left\| \frac{1}{n} \sum_{i=1}^{n-\bar{n}} \mathbf{A} \mathbf{x}_{k(i)} \mathbf{x}_{k(i)}^\top \mathbf{A}^\top \right\|_2 + c\phi(1 + \zeta d \|\mathbf{A}\|_2).$$

The first term on the right hand side can be bound by Equation (7.10). Hence we obtain



this lemma. □

**Lemma 7.17.** (Corollary 3.3, [VCLR13]) *There exists a universal constant  $c$  such that with probability at least  $1 - p^{-10}$ ,*

$$\left\| \frac{1}{n} \sum_{i=1}^n \mathbf{z}_i \mathbf{z}_i^\top - \boldsymbol{\Sigma} \right\|_\infty \leq c\zeta \|\boldsymbol{\Sigma}\|_2 \sqrt{\frac{\log p}{n}}.$$

## 7.8 Chapter Summary

In this chapter, we proposed a unified framework for making PCA-like algorithms robust to outliers. We provided theoretical performance analysis of the proposed framework using both the subspace distance and the expressed variance metrics. To the best of our knowledge, this is the first attempt to make a wide range of PCA-like algorithms provably robust to any constant fraction of arbitrarily corrupted samples. As an immediate result, our framework leads to robust sparse PCA and robust non-negative sparse PCA with theoretic guarantees – the first of its kind to the best of our knowledge. The experiments show that the outlier-robust PCA-like algorithms derived from our framework outperforms their non-robust version and other alternatives including HR-PCA and ROB-SPCA.

## Non-convex Outlier-Robust PCA

We develop new efficient algorithms for outlier-robust PCA whose aim is to exactly recover the low-dimensional subspace spanned by the uncorrupted samples and correctly identify the corrupted samples. Our algorithms are non-convex counterparts of Outlier Pursuit proposed by [XCS12], which alternatively estimate the low-dimensional subspace and mitigate the effect of corruption. They have much lower computational complexity compared to Outlier Pursuit. In particular, for a  $p \times n$  input matrix, the total operations required to obtain an estimated subspace with target rank  $r$  and estimation error  $\epsilon$  is  $O(rnp \log(1/\epsilon))$ , which is close to computation cost for the standard PCA. We establish theoretical performance guarantees for the proposed algorithm on the exact recovery of the true subspace under some mild assumptions on the fraction of the corrupted samples that are similar to those required by Outlier Pursuit. The numerical experiments on synthetic and real-world data illustrate their good empirical performance.

### 8.1 Introduction

Principal component analysis (PCA) [Pea01] is arguably the most widely applied dimensionality reduction method, playing a significant role in a broad range of areas including machine learning, statistics, finance and many others. The standard PCA is simple to implement by performing the eigenvalue decomposition of the sample covariance matrix. It is well known that PCA is sensitive to the presence of outliers, i.e., its performance degrades significantly even with a few corrupted samples, due to the quadratic error criterion used.

The pursuit of robust PCA algorithms has consistently attracted attention in statistics and

in machine learning, e.g., [DGK81, XY95, YW99, ITB03, Das03, XCM13, FXY12, YX15b]. In Chapter 7, we proposed a general framework that is tractable, computationally efficient, and provably robustify a wide range of PCA-like algorithms including the standard PCA and sparse PCA, even in the face of a constant fraction of samples are corrupted in the high dimensional regime. But these algorithms cannot guarantee the exact recovery of the subspace spanned by the true principal components.

Recently, borrowing ideas from compressive sensing, a prominent new approach for robust PCA is to decompose the noisy sample matrix  $\mathbf{X}$  into a low-rank matrix  $\mathbf{L}^*$  and a sparse matrix  $\mathbf{S}^*$  via *nuclear norm minimization*, e.g., [CR09, CLMW11, RFP10, CSPW11]. The seminal papers of [RFP10] and [CLMW11] showed that the exact recovery of the low-rank and sparse matrices can be achieved under some mild conditions on the incoherence of  $\mathbf{L}^*$  and the sparsity of  $\mathbf{S}^*$ . These papers assume that the support set of  $\mathbf{S}^*$  is uniformly distributed among all the sets of a certain cardinality, which is not suitable for handling outliers where there exist some columns whose entries are all corrupted. To address this issue, [XCS12] proposed a nuclear norm based algorithm called *Outlier Pursuit* to handle corrupted samples, where they assumed that  $\mathbf{S}^*$  is column-wise sparse instead of entry-wise sparse. The goal of Outlier Pursuit is to exactly recover the column space of the low-rank matrix  $\mathbf{L}^*$  and identify the nonzero columns of  $\mathbf{S}^*$ . They proved that exact recovery can be achieved under mild conditions depending on the incoherence of the row space of  $\mathbf{L}^*$  and the fraction of outliers. While nuclear norm based algorithms have elegant theoretical results, they can be difficult to apply to large-scale applications due to high computational cost. In particular, for a  $p \times n$  matrix, state-of-art numerical algorithms solving these formulations, typically based on ALM [LCM10, CLMW11] and ADMM [BPC<sup>+</sup>10], requires  $O(\min\{p^2n, pn^2\})$  computation *per iteration*.

In the last couple of years, computationally efficient robust PCA algorithms based on alternating minimization techniques have drawn much attention, e.g., [NNS<sup>+</sup>14, JNS13, Har13, ZWL15]. All these algorithms are designed to recover the low-rank matrix  $\mathbf{L}^*$  from  $\mathbf{X}$  with entry-wise sparse noise or missing entries, instead of column-wise corruption, i.e., outliers. In this chapter, we develop two novel non-convex algorithms for *outlier-robust* PCA called *Outlier Rejection* and *Outlier Reduction*, which involve alternating between estimating the

low-rank column space of  $\mathbf{L}^*$  and identifying the outliers indicated by  $\mathbf{S}^*$ . In comparison with Outlier Pursuit [XCS12], the proposed algorithms have much lower computational load, yet enjoy similar performance guarantees for the exact recovery of the true column space. Indeed, for a  $p \times n$  matrix  $\mathbf{L}^*$ , the overall computational complexity of the proposed algorithms is  $O(rnp \log(1/\epsilon))$  where  $r$  is the target rank and  $\epsilon$  is the estimation error, which is almost as low as the computational complexity of the standard PCA. Moreover, if the fraction of the outliers  $\rho$  is  $O(\frac{1}{\mu r^{*2}})$  where  $r^*$  is the rank of  $\mathbf{L}^*$  and  $\mu$  is the column-incoherence parameter that will be discussed in the following section, our algorithms guarantee to recover the true column space with an arbitrary small error. This condition on  $\rho$  is slightly stronger than that of Outlier Pursuit –  $O(\frac{1}{\mu r^*})$  to be specific, yet our experiments demonstrate that they outperform Outlier Pursuit in practice.

**Notations:** We use boldface lower-case letters to represent column vectors and capital letters for matrices. For matrix  $\mathbf{X}$ , three matrix norms are used:  $\|\mathbf{X}\|_2$  is the spectral norm,  $\|\mathbf{X}\|_F$  is the Frobenius norm,  $\|\mathbf{X}\|_{\infty,2}$  is the largest  $l_2$  norm of the columns. Additionally,  $\|\mathbf{X}\|_{0,2}$  denotes the number of the nonzero columns of  $\mathbf{X}$ ,  $\mathbf{X}_i$  denotes the  $i^{\text{th}}$  column of  $\mathbf{X}$  and  $\sigma_r(\mathbf{X})$  denotes the  $r^{\text{th}}$  largest singular value of  $\mathbf{X}$ .

## 8.2 Problem Setting

The outlier-robust PCA problem we consider in this chapter is the same as studied in [XCS12]. More specifically, suppose that we receive  $n$  samples  $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}$  in  $p$ -dimensional space, where a fraction  $1 - \rho$  of these samples lie in a  $r^*$ -dimensional true subspace of  $\mathbb{R}^p$  – these  $(1 - \rho)n$  samples are taken as inliers, and the remaining  $\rho n$  samples are arbitrarily located – these  $\rho n$  samples are taken as outliers. Our goal is to recovery the true subspace spanned by the inlier samples.

Let  $\mathbf{X} \in \mathbb{R}^{p \times n}$  be the data matrix formed by these  $n$  samples, each of whose columns is one of the samples and let  $\mathcal{C}$  be the set of indices corresponding to the outlier samples. In the noiseless case,  $\mathbf{X}$  can be decomposed as

$$\mathbf{X} = \mathbf{L}^* + \mathbf{S}^*,$$

where matrices  $\mathbf{L}^*$  and  $\mathbf{S}^*$  correspond to the inlier and outlier samples, respectively. That is,  $\mathbf{L}_i^* = \mathbf{x}_i$  for  $i \notin \mathcal{C}$  or  $\mathbf{0}$  otherwise and  $\mathbf{S}_i^* = \mathbf{x}_i$  for  $i \in \mathcal{C}$  or  $\mathbf{0}$  otherwise. Thus,  $\text{rank}(\mathbf{L}^*) = r^*$  and  $\|\mathbf{S}^*\|_{0,2} = |\mathcal{C}| = \rho n$ . Similarly, in the noisy case,  $\mathbf{X}$  can be decomposed as

$$\mathbf{X} = \mathbf{L}^* + \mathbf{S}^* + \mathbf{N},$$

where  $\mathbf{N}$  is any additional noise applied to the samples. Consider the singular value decomposition of  $\mathbf{L}^*$

$$\mathbf{L}^* = \mathbf{U}^* \mathbf{\Sigma}^* \mathbf{V}^{*\top},$$

where  $\mathbf{U}^* \in \mathbb{R}^{p \times r^*}$ ,  $\mathbf{V}^* \in \mathbb{R}^{n \times r^*}$  and  $\mathbf{\Sigma}^* \in \mathbb{R}^{r^* \times r^*}$ , then  $\mathbf{U}^*$  forms an orthonormal basis for the true column subspace we wish to reveal. It is well known that recovering a low rank matrix from column sparse corruption may not be well defined when the matrix is column-sparse itself. As an extreme example, if  $\mathbf{X}$  has only one nonzero column, then  $\mathbf{X}$  is both low-rank and column-sparse and hence it is impossible to identify the true column space of  $\mathbf{L}^*$ . To avoid such “low-rank” and “column-sparse” ambiguity, the following incoherent assumption is made [XCS12]:

**Assumption 8.1.**  $\mathbf{L}^*$  is  $\mu$ -column-incoherent, i.e., if  $\mathbf{L}^*$  with SVD  $\mathbf{L}^* = \mathbf{U}^* \mathbf{\Sigma}^* \mathbf{V}^{*\top}$  has  $(1 - \rho)n$  nonzero columns, then

$$\max_i \|\mathbf{e}_i^\top \mathbf{V}^*\|_2^2 \leq \frac{\mu r^*}{(1 - \rho)n},$$

where  $\{\mathbf{e}_1, \dots, \mathbf{e}_n\}$  is the standard basis of  $\mathbb{R}^n$ .

Obviously, if  $\mathbf{V}^*$  is perfectly column-incoherent, i.e.,  $\mathbf{V}^*$  has rank 1 with nonzero entries equal to  $\frac{1}{\sqrt{(1-\rho)n}}$ , the incoherent parameter  $\mu$  is 1, and if each column of  $\mathbf{V}^*$  aligns with a coordinate axis, then  $\mu$  equals  $\frac{(1-\rho)n}{r^*}$ . [CR09] proved that if the samples are generated according to some low-dimensional isometric distribution, then  $\mu = O(\max\{1, \frac{\log n}{r^*}\})$  with high probability. Thus, a smaller  $\mu$  means that the column support of each column of  $\mathbf{V}^*$  spreads out.

This condition is weaker than the incoherent conditions for matrix completion, e.g., [CR09, CT10, Gro11, CBSW14], and robust PCA, e.g., [WPM<sup>+</sup>09, CLMW11, NNS<sup>+</sup>14], which also

require row-incoherence, while we only need column-incoherence since our goal is to recover the *column space* of  $\mathbf{L}^*$ , instead of  $\mathbf{L}^*$ .

### 8.3 Outlier Rejection and Outlier Reduction

A direct approach to recover the column space of  $\mathbf{L}^*$  is to solve the following optimization problem:

$$\min_{\mathbf{L}, \mathbf{S}} \text{rank}(\mathbf{L}) + \lambda \|\mathbf{S}\|_{0,2}, \text{ s.t. } \mathbf{X} = \mathbf{L} + \mathbf{S}, \quad (8.1)$$

and then perform SVD of the optimal solution. Note that (8.1) is intractable due to the non-convexity of  $\text{rank}(\cdot)$  and  $\|\cdot\|_{0,2}$ . To address this issue, [XCS12] proposed a convex relaxation for (8.1) called *Outlier Pursuit*:

$$\min_{\mathbf{L}, \mathbf{S}} \|\mathbf{L}\|_* + \lambda \|\mathbf{S}\|_{1,2}, \text{ s.t. } \mathbf{X} = \mathbf{L} + \mathbf{S}, \quad (8.2)$$

where  $\|\cdot\|_*$  is the nuclear norm and  $\|\cdot\|_{1,2}$  denotes the sum of the  $l_2$ -norm of the columns. Under Assumption 8.1 and some mild conditions on the fraction of outlier samples, Outlier Pursuit is guaranteed to exactly recover of the true column space. Although it owns beautiful theoretical results, its practical applications are still limited due to its high computational cost. In particular, (8.2) is usually solved via the augmented Lagrangian multipliers (ALM) method [LCM10, CLMW11] or the alternating direction method of multipliers (ADMM) method [BPC<sup>+</sup>10], either of which requires  $O(\min\{p^2n, pn^2\})$  computations in each iteration that are unaffordable when faced with large-scale data.

In this section, we present two new computationally efficient algorithms for the outlier-robust PCA problem. Instead of considering convex surrogates of (8.1), we formulate this problem as the following non-convex feasibility problem: find  $\mathbf{L}, \mathbf{S}$  such that 1)  $\mathbf{X} = \mathbf{L} + \mathbf{S}$ , 2)  $\mathbf{L}$  lies in the set of low-rank matrices, 3)  $\mathbf{S}$  lies in the set of column-sparse matrices, and 4)  $\|(\mathbf{I} - \mathbf{L}\mathbf{L}^\dagger)\mathbf{L}^*\|_{\infty,2} \leq \epsilon$ , where  $\mathbf{L}^\dagger$  is the pseudo-inverse of  $\mathbf{L}$  and  $\epsilon$  is a certain constant. The last constraint relates to the estimation accuracy, ensuring that the  $l_2$ -norm of each inlier sample after projected onto the subspace orthogonal to the column space of  $\mathbf{L}$  is less than  $\epsilon$ . In other words, it guarantees that the true column space is approximately included

in the column space of  $\mathbf{L}$ . Clearly, when  $\text{rank}(\mathbf{L}) = r^*$  and  $\epsilon = 0$ , the column spaces of  $\mathbf{L}$  and  $\mathbf{L}^*$  are the same. Since  $\mathbf{L}^*$  is unknown, our plan is to solve solution  $\mathbf{L}$  based on the first three constraints while taking the last constraint as a metric for theoretical performance.

The algorithms proposed in this chapter are shown in Algorithm 8.1 and Algorithm 8.2, namely, outlier-robust PCA via *Outlier Rejection* and *Outlier Reduction*. The intuitive idea of our algorithms is as follows. In the preprocessing step, each nonzero column of  $\mathbf{X}$  is normalized to reduce the bad effect of outliers with large magnitudes. Then given the target rank  $r \geq r^*$ ,  $\mathbf{L}$  and  $\mathbf{S}$  can be solved alternatively: 1) update  $\mathbf{L}$  with fixed  $\mathbf{S}$ , i.e., compute  $\mathbf{L} = \mathbf{X} - \mathbf{S}$ , and 2) update  $\mathbf{S}$  with fixed  $\mathbf{L}$ , i.e., project each column of  $\mathbf{X}$  onto the subspace that is orthogonal to the subspace spanned by the top  $r$  left singular vectors of  $\mathbf{L}$  to determine  $\mathbf{S}$ . Therefore, the key component of the algorithms is to construct proper  $\mathbf{S}$  to guarantee that the subspace spanned by the top  $r$  singular vectors of  $\mathbf{L}$  becomes close to the true column space after several iterations.

In Outlier Rejection,  $\mathbf{S}$  is constructed by identifying the outlier samples revealed so far from the training samples. More specifically, at the  $t^{\text{th}}$  iteration, let  $\mathbf{U} \in \mathbb{R}^{p \times r}$  consists of the top  $r$  left singular vectors of  $\mathbf{L}$ , then  $\mathbf{S}_i$  – the  $i^{\text{th}}$  column of  $\mathbf{S}$  – equals  $\mathbf{X}_i$  if  $\|(\mathbf{I} - \mathbf{U}\mathbf{U}^\top)\mathbf{X}_i\|_2 > \epsilon$  for a certain threshold  $\epsilon$  or  $\mathbf{0}$  otherwise. This essentially means the samples deviating from  $\mathbf{U}_t$  too much are considered as outliers and removed in the next iteration. Intuitively, the more outliers are revealed, the smaller  $\|(\mathbf{I} - \mathbf{U}\mathbf{U}^\top)\mathbf{L}^*\|_{\infty,2}$  becomes, and vice versa.

In Outlier Reduction, instead of removing outlier samples, it tries to reduce the residual of each sample after projected onto the subspace spanned by  $\mathbf{U}$  – the top  $r$  left singular vectors of  $\mathbf{L}$ . Specifically, at the  $t^{\text{th}}$  iteration, if the residual  $\mathbf{R}_i \triangleq (\mathbf{I} - \mathbf{U}\mathbf{U}^\top)\mathbf{X}_i$  for the  $i^{\text{th}}$  sample is relatively large, namely,  $\|\mathbf{R}_i\|_2 > \epsilon$  for a certain threshold  $\epsilon$ , we reduce this residual from  $\mathbf{X}_i$  by setting  $\mathbf{S}_i$  to  $\mathbf{R}_i$ , or equivalently, we select  $\mathbf{S} = CT_\epsilon(\mathbf{X} - \mathbf{U}\mathbf{U}^\top\mathbf{X})$  where  $CT_\epsilon(\cdot)$  is a *column-wise truncation operator* defined as follows: For matrix  $\mathbf{Y} \in \mathbb{R}^{p \times n}$ ,  $CT_\epsilon(\mathbf{Y})$  returns a matrix with size  $p \times n$  whose  $i^{\text{th}}$  column is  $\mathbf{Y}_i$  if  $\|\mathbf{Y}_i\|_2 > \epsilon$  or  $\mathbf{0}$  otherwise. Since the corruption in the contaminated samples that affect the accuracy of column space estimation is reduced in one iteration, the resulting subspace spanned by  $\mathbf{U}$  will become closer to the true column space in the next iteration.

**Algorithm 8.1:** Outlier Rejection

---

**Input** : Matrix  $\mathbf{X}$ , target rank  $r$  and parameters  $\tau, \eta$ .  
**Output**: The estimated principal components.

- 1 Normalize the columns of  $\mathbf{X}$ , i.e.,  $\mathbf{X}_i = \frac{\mathbf{X}_i}{\|\mathbf{X}_i\|_2}$  when  $\mathbf{X}_i \neq \mathbf{0}$ ;
- 2 Initialize  $\mathbf{L}_1 = \mathbf{X}$ ,  $\epsilon_1 = 1$ ,  $T = \frac{\log \epsilon}{\log \tau} + 1$ ;
- 3 **for**  $t = 1$  *to*  $T$  **do**
- 4     Compute  $\mathbf{U}_t$  – the top  $r$  left singular vectors of  $\mathbf{L}_t$ ;
- 5     Construct  $\mathcal{A}_t = \{i : \|(\mathbf{X} - \mathbf{U}_t \mathbf{U}_t^\top \mathbf{X})_i\|_2 > \epsilon_t\}$ ;
- 6     Update  $\mathbf{L}_{t+1}$  so that  $\mathbf{L}_{t+1,i} = \mathbf{X}_i$  if  $i \notin \mathcal{A}_t$  or  $\mathbf{0}$  otherwise.
- 7     Set  $\epsilon_{t+1} = \tau \epsilon_t + \eta$ ;
- 8 **end**
- 9 Return  $\mathbf{U}_T$ .

---

**Algorithm 8.2:** Outlier Reduction

---

**Input** : Matrix  $\mathbf{X}$ , target rank  $r$  and parameters  $\tau, \eta$ .  
**Output**: The estimated principal components.

- 1 Normalize the columns of  $\mathbf{X}$ , i.e.,  $\mathbf{X}_i = \frac{\mathbf{X}_i}{\|\mathbf{X}_i\|_2}$  when  $\mathbf{X}_i \neq \mathbf{0}$ ;
- 2 Initialize  $\mathbf{S}_1 = \mathbf{0}$ ,  $\epsilon_1 = 1$ ,  $T = \frac{\log \epsilon}{\log \tau} + 1$ ;
- 3 **for**  $t = 1$  *to*  $T$  **do**
- 4     Compute  $\mathbf{L}_t = \mathbf{X} - \mathbf{S}_t$ ;
- 5     Compute  $\mathbf{U}_t$  – the top  $r$  left singular vectors of  $\mathbf{L}_t$ ;
- 6     Update  $\mathbf{S}_{t+1} = CT_{\epsilon_t}(\mathbf{X} - \mathbf{U}_t \mathbf{U}_t^\top \mathbf{X})$ .
- 7     Set  $\epsilon_{t+1} = \tau \epsilon_t + \eta$ ;
- 8 **end**
- 9 Return  $\mathbf{U}_T, \epsilon_T$ .

---

In both of Algorithm 8.1 and Algorithm 8.2, threshold  $\epsilon_t$  is a key parameter for achieving exact recovery of the true column space with a fast convergence rate. Note that  $\epsilon_t$  is



updated via  $\epsilon_{t+1} = \tau\epsilon_t + \eta$  for certain constants  $\tau < 1$  and  $\eta > 0$ . We will discuss how to select  $\tau$  and  $\eta$  in the next section. Obviously, the major difference between the two algorithms is that Outlier Rejection tries to directly identify the set of the outliers while Outlier Reduction tends to reduce the corruption containing in the contaminated samples. Empirically, Outlier Reduction is more robust to corruption than Outlier Rejection. For example, our experiments show that it can recover the true column space even when each sample contains some corrupted feature values.

Based on Algorithm 8.2, one can also recover the row space of  $\mathbf{L}^*$  as shown in Algorithm 8.3. The basic idea is that the outlier samples can be identified via projecting each sample onto the column space estimated by Algorithm 8.2, which is similar to the technique applied in Algorithm 8.1. Clearly, if all the outliers are identified and removed, one can recover the true row space.

---

**Algorithm 8.3:** Row-space recovery
 

---

**Input** : Matrix  $\mathbf{X}$ , target rank  $r$ , accuracy  $\epsilon$  and parameters  $\tau, \eta$ .

**Output:** The estimated projection matrix on the row space of  $\mathbf{L}^*$ .

- 1 Compute  $\mathbf{U}_T, \epsilon_T$  by running Algorithm 8.2 with input  $\mathbf{X}$  and parameters  $r, \tau, \eta$  and  $T \geq \log \epsilon / \log \tau + 1$ ;
  - 2 Let  $\mathcal{A} = \{i : \|(\mathbf{X} - \mathbf{U}_T \mathbf{U}_T^\top \mathbf{X})_i\|_2 / \|\mathbf{X}_i\|_2 > \epsilon_T\}$  and construct  $\mathbf{L}$  so that  $\mathbf{L}_i = \mathbf{X}_i$  if  $i \notin \mathcal{A}$  or  $\mathbf{0}$  otherwise;
  - 3 Compute  $\mathbf{V}_r$  – the top  $r$  right singular vectors of  $\mathbf{L}$ ;
  - 4 Return  $\mathbf{V}_r \mathbf{V}_r^\top$ .
- 

Note that each iteration in Algorithm 8.1 and Algorithm 8.2 has a  $O(rnp)$  computational complexity because it only involves the calculation of the top  $r$  left singular vectors of a  $p \times n$  matrix. Therefore, the overall computational complexity of our algorithms is  $O(rnp \log \epsilon / \log \tau)$ . In comparison with the Outlier Pursuit algorithm [XCS12] which requires  $O(\min\{p^2 n, pn^2\})$  operations in each iteration due to calculating the singular value decomposition of a  $p \times n$  matrix when it is solved via the ALM or ADMM method, our algorithms have a much lower computational cost and hence can be applied in large-scale

applications.

## 8.4 Performance Guarantees

We now provide theoretical performance guarantees for the proposed algorithms. Recall that the columns of sample matrix  $\mathbf{X}$  are normalized in the first step of Algorithms 8.1 and 8.2. Since this normalization step has no effect on the true column space, we assume that each nonzero column of  $\mathbf{X}$  is a unit vector and such normalized  $\mathbf{X}$  satisfies Assumption 8.1. In the following parts, we use  $\mathcal{L}$  to denote the set of full rank  $p \times r^*$  submatrices of  $\mathbf{L}^*$ , i.e., for any  $\mathbf{L} \in \mathcal{L}$ , the columns of  $\mathbf{L}$  are linearly independent and drawn from the columns of  $\mathbf{L}^*$ .

Theorem 8.1 shows the performance guarantee of Algorithm 8.1 in the noiseless case, stating that  $\|(\mathbf{I} - \mathbf{U}_T \mathbf{U}_T^\top) \mathbf{L}^*\|_{\infty,2}$  can be arbitrarily small as long as the fraction of outliers is upper bounded by a certain value depending on  $r^*$ . This means the true column space is approximately included in the subspace spanned by  $\mathbf{U}_T$ . Especially, the exact recovery can be achieved when target rank  $r$  equals  $r^*$ .

**Theorem 8.1.** *Suppose that  $\mathbf{X} = \mathbf{L}^* + \mathbf{S}^*$  where  $\mathbf{L}^*$  is  $\mu$ -column-incoherent and has rank  $r^*$ , and  $\mathbf{S}^*$  is supported on at most  $\rho n$  columns. Then as long as the fraction of outliers  $\rho$  satisfies*

$$\frac{\rho}{1-\rho} < \frac{1}{4\mu r^* (1 + \sqrt{r^*} / \max_{\mathbf{L} \in \mathcal{L}} \sigma_{r^*}(\mathbf{L}))^2}, \quad (8.3)$$

there exists  $\tau \in [2\sqrt{\frac{\mu r^* \rho}{1-\rho}} (1 + \frac{\sqrt{r^*}}{\max_{\mathbf{L} \in \mathcal{L}} \sigma_{r^*}(\mathbf{L})}), 1)$  so that for any  $\epsilon > 0$ , when  $r \geq r^*$  and  $\eta = 0$ , the output  $\mathbf{U}_T$  of Algorithm 8.1 satisfies

$$\|(\mathbf{I} - \mathbf{U}_T \mathbf{U}_T^\top) \mathbf{L}^*\|_{\infty,2} \leq \epsilon.$$

Furthermore, if  $r = r^*$ , we have

$$\|\mathbf{U}^* \mathbf{U}^{*\top} - \mathbf{U}_T \mathbf{U}_T^\top\|_2 \leq \frac{\sqrt{r^*} \epsilon}{\max_{\mathbf{L} \in \mathcal{L}} \sigma_r(\mathbf{L})}.$$

**Remark 1.** The term  $\max_{\mathbf{L} \in \mathcal{L}} \sigma_{r^*}(\mathbf{L})$  is determined by the intrinsic property of  $\mathbf{L}^*$ , which

could be a constant, e.g., when  $\Sigma^* = \sqrt{(1-\rho)n}$  and  $\mathbf{V}^* \in \mathbb{R}^{n \times 1}$  whose nonzero entries are  $\frac{1}{\sqrt{(1-\rho)n}}$ , one can easily verify that  $\mathbf{V}^*$  satisfies Assumption 8.1 and  $\max_{\mathbf{L} \in \mathcal{L}} \sigma_{r^*}(\mathbf{L}) = 1$ .

**Remark 2.** [XCS12] proved that Outlier Pursuit achieves exact recovery when  $\frac{\rho}{1-\rho} \leq \frac{9}{121\mu r^*}$ . Note that when  $\max_{\mathbf{L} \in \mathcal{L}} \sigma_{r^*}(\mathbf{L})$  is a constant, the required upper bound for  $\frac{\rho}{1-\rho}$  shown in (8.3) is about  $O(r^*)$  times large as that of Outlier Pursuit, which is a mild cost in the low-rank case where  $r^*$  is small. This highlights a tradeoff between computational efficiency and sample complexity: while our algorithms run much faster than Outlier Pursuit, it needs a stronger condition on the fraction of outliers. Yet, empirically, the experiments appear to suggest that our algorithms outperform Outlier Pursuit.

Theorem 8.2 provides a performance guarantee for Algorithm 8.1 in the noisy case. The major difference between Theorem 8.1 and Theorem 8.2 is that now parameter  $\eta$  need to be set to some positive constant instead of zero due to existence of additional noise  $\mathbf{N}$ .

**Theorem 8.2.** *Suppose that  $\mathbf{X} = \mathbf{L}^* + \mathbf{S}^* + \mathbf{N}$  where  $\mathbf{L}^*$  is  $\mu$ -column-incoherent and has rank  $r^*$ ,  $\mathbf{S}^*$  is supported on at most  $pn$  columns and  $\mathbf{N}$  is the additional noise. Then as long as the fraction of outliers  $\rho$  satisfies*

$$\frac{\rho}{1-\rho} < \frac{1}{4\mu r^*(1 + \sqrt{r^*}/\max_{\mathbf{L} \in \mathcal{L}} \sigma_{r^*}(\mathbf{L}))^2},$$

there exists  $\tau \in [2\sqrt{\frac{\mu r^* \rho}{1-\rho}}(1 + \frac{\sqrt{r^*}}{\max_{\mathbf{L} \in \mathcal{L}} \sigma_{r^*}(\mathbf{L})}), 1]$  so that for any  $\epsilon > 0$ , when  $r \geq r^*$ , and  $\eta \geq \varphi \|\mathbf{N}\|_{\infty, 2}$ , where

$$\varphi = 2\sqrt{\frac{\mu r^*}{1-\rho}} + 2r^* \sqrt{\frac{\mu \rho}{(1-\rho)\max_{\mathbf{L} \in \mathcal{L}} \sigma_{r^*}(\mathbf{L})^2}} + 1,$$

the output  $\mathbf{U}_T$  of Algorithm 8.1 satisfies

$$\|(\mathbf{I} - \mathbf{U}_T \mathbf{U}_T^\top) \mathbf{L}^*\|_{\infty, 2} \leq \epsilon + \left(\frac{\varphi}{1-\tau} + 1\right) \|\mathbf{N}\|_{\infty, 2}.$$

Furthermore, if  $r = r^*$ , we have

$$\|\mathbf{U}^* \mathbf{U}^{*\top} - \mathbf{U}_T \mathbf{U}_T^\top\|_2 \leq \frac{\sqrt{r}(\epsilon + (\frac{\varphi}{1-\tau} + 2) \|\mathbf{N}\|_{\infty, 2})}{\max_{\mathbf{L} \in \mathcal{L}} \sigma_r(\mathbf{L})}.$$

**Remark 3.** Due to existence of noise  $\mathbf{N}$ ,  $\eta$  should be greater than  $\varphi\|\mathbf{N}\|_{\infty,2}$  to ensure that most of the rejected samples in each iteration are outliers. The error bound for  $\|(\mathbf{I} - \mathbf{U}_T\mathbf{U}_T^\top)\mathbf{L}^*\|_{\infty,2}$  involves the magnitude of the noise applied to each sample. If this magnitude is not too large, our algorithm obtains good results.

Theorem 8.3 and Theorem 8.4 provide performance guarantees for Algorithm 8.2 in the noiseless case and the noisy case, respectively. Interestingly, the guarantees for Algorithm 8.1 and Algorithm 8.2 are the same although their proofs differ (refer to the appendix).

**Theorem 8.3.** *Suppose that  $\mathbf{X} = \mathbf{L}^* + \mathbf{S}^*$  where  $\mathbf{L}^*$  is  $\mu$ -column-incoherent and has rank  $r^*$ , and  $\mathbf{S}^*$  is supported on at most  $pn$  columns. Under the same conditions as Theorem 8.1, the output  $\mathbf{U}_T$  of Algorithm 8.2 satisfies*

$$\|(\mathbf{I} - \mathbf{U}_T\mathbf{U}_T^\top)\mathbf{L}^*\|_{\infty,2} \leq \epsilon.$$

Furthermore, if  $r = r^*$ , we have

$$\|\mathbf{U}^*\mathbf{U}^{*\top} - \mathbf{U}_T\mathbf{U}_T^\top\|_2 \leq \frac{\sqrt{r}\epsilon}{\max_{\mathbf{L} \in \mathcal{L}} \sigma_r(\mathbf{L})}.$$

**Theorem 8.4.** *Suppose that  $\mathbf{X} = \mathbf{L}^* + \mathbf{S}^* + \mathbf{N}$  where  $\mathbf{L}^*$  is  $\mu$ -column-incoherent and has rank  $r^*$ ,  $\mathbf{S}^*$  is supported on at most  $pn$  columns and  $\mathbf{N}$  is the additional noise. Under the same conditions as Theorem 8.2, the output  $\mathbf{U}_T$  of Algorithm 8.2 satisfies*

$$\|(\mathbf{I} - \mathbf{U}_T\mathbf{U}_T^\top)\mathbf{L}^*\|_{\infty,2} \leq \epsilon + \left(\frac{\varphi}{1-\tau} + 1\right)\|\mathbf{N}\|_{\infty,2}.$$

Furthermore, if  $r = r^*$ , we have

$$\|\mathbf{U}^*\mathbf{U}^{*\top} - \mathbf{U}_T\mathbf{U}_T^\top\|_2 \leq \frac{\sqrt{r}(\epsilon + (\frac{\varphi}{1-\tau} + 2)\|\mathbf{N}\|_{\infty,2})}{\max_{\mathbf{L} \in \mathcal{L}} \sigma_r(\mathbf{L})}.$$

We remark that while Algorithms 8.1 and 8.2 have same performance guarantees, their empirical performance may differ. Algorithm 8.1 tends to identify the set of outliers while Algorithm 8.2 prefers reducing the corruption. Consequently, Algorithm may be more robust in practice, e.g., our experiments show that Algorithm 8.2 is able to extract the background

for a video sequence where many sample images contain corrupted regions (foreground objects), but Algorithm 8.1 is not capable of doing this.

The following theorem shows when Algorithm 8.3 can exactly recover the row space of  $\mathbf{L}^*$ . Note that we do not assume the columns of  $\mathbf{X}$  are normalized here.

**Theorem 8.5.** *Suppose that  $\mathbf{X} = \mathbf{L}^* + \mathbf{S}^*$  where  $\mathbf{L}^*$  is  $\mu$ -column-incoherent and has rank  $r^*$ , and  $\mathbf{S}^*$  is supported on at most  $\rho n$  columns. Let  $\mathcal{L}$  be the set of full rank  $p \times r^*$  submatrices of  $\mathbf{L}^*$  after the normalization. If the fraction of outliers  $\rho$  satisfies*

$$\frac{\rho}{1 - \rho} < \frac{1}{4\mu r^* (1 + 2\sqrt{r^*} / \max_{\mathbf{L} \in \mathcal{L}} \sigma_{r^*}(\mathbf{L}))^2},$$

and there exists constant  $\delta > 0$  so that

$$\frac{\|\mathbf{U}^* \mathbf{U}^{*\top} \mathbf{S}_i^*\|_2}{\|\mathbf{S}_i^*\|_2} \geq \delta, \quad \forall i \in \mathcal{C},$$

then when  $r = r^*$ ,  $\eta = 0$ ,  $1 > \tau \geq 2\sqrt{\frac{\mu r^* \rho}{(1 - \rho)}} (1 + \frac{2\sqrt{r^*}}{\max_{\mathbf{L} \in \mathcal{L}} \sigma_{r^*}(\mathbf{L})})$  and  $\epsilon < \frac{\delta}{1 + \sqrt{r^*} / \max_{\mathbf{L} \in \mathcal{L}} \sigma_{r^*}(\mathbf{L})}$ , the output of Algorithm 8.3 satisfies  $\mathbf{V}_r \mathbf{V}_r^\top = \mathbf{V}^* \mathbf{V}^{*\top}$ .

**Remark 4.** The main idea behind this theorem is that since the outliers do not lie in the true column space, they can be identified by measuring the residual  $\|(\mathbf{I} - \mathbf{U}_T \mathbf{U}_T^\top) \mathbf{X}_i\|_2$ . When  $\mathbf{U}_T$  and  $\mathbf{U}^*$  are close enough, we can guarantee that  $\|(\mathbf{I} - \mathbf{U}_T \mathbf{U}_T^\top) \mathbf{X}_i\|_2 / \|\mathbf{X}_i\|_2 \leq \epsilon$  for all inlier samples while  $\|(\mathbf{I} - \mathbf{U}_T \mathbf{U}_T^\top) \mathbf{X}_i\|_2 / \|\mathbf{X}_i\|_2 > \epsilon$  for all outlier samples. Therefore, the true row space can be recovered after all the outliers are removed.

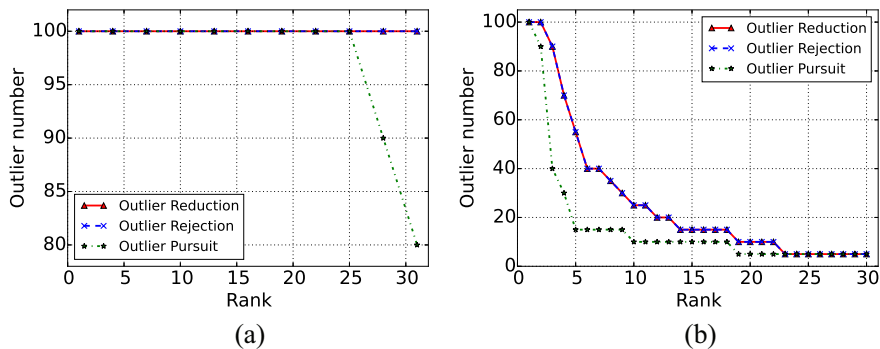
## 8.5 Experiments

We investigate the performance of our algorithms on a variety of simulated and real-world datasets. All the algorithms mentioned below are implemented in Python. The experiments are conducted on a desktop PC with an i7 3.4GHz CPU and 4G memory.

### 8.5.1 Synthetic Data

We first investigate the empirical performance of Outlier Rejection (Algorithm 8.1) and Outlier Reduction (Algorithm 8.2) on synthetic data. In order to compare these two algorithms with Outlier Pursuit [XCS12], we use the same scheme for generating test data as that stated in [XCS12]. For different  $r^*$  and number of outliers  $\rho n$ , the true low-rank matrix  $\mathbf{L}^*$  is generated according to  $\mathbf{L}^* = \mathbf{A}\mathbf{B}^\top$  where matrices  $\mathbf{A} \in \mathbb{R}^{p \times r^*}$  and  $\mathbf{B} \in \mathbb{R}^{(n-\rho n) \times r^*}$  whose entries are independently drawn from the standard Gaussian distribution  $\mathcal{N}(0, 1)$ . The outliers  $\{\mathbf{x}_i : i \in \mathcal{C}\}$  are generated either randomly, where each entry of  $\mathbf{x}_i$  follows  $\mathcal{N}(0, 1)$ , or adversarially, where each  $\mathbf{x}_i$  is an identical copy of a certain random Gaussian vector. In the following experiments, both of  $p$  and  $n$  are set to 400, and the target rank  $r$  and estimation error  $\epsilon$  for Outlier Rejection and Outlier Reduction are set to  $r^*$  and  $10^{-3}$ , respectively. We implement Outlier Pursuit based on the algorithm shown in Section VI in [XCS12].

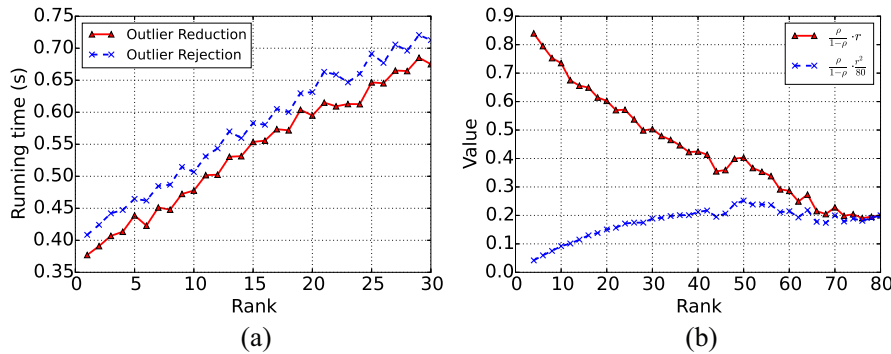
In the first experiment, we study the phase transition properties of these three algorithms in the noiseless case. An algorithm “succeeds” if the subspace spanned by the leading  $r^*$  left singular vectors of its output  $\mathbf{L}$  is included in the true column space of  $\mathbf{L}^*$ , i.e., the projection  $\mathbf{U}\mathbf{U}^\top$  onto the subspace spanned by the leading  $r^*$  left singular vectors of  $\mathbf{L}$  satisfies  $\|\mathbf{U}\mathbf{U}^\top - \mathbf{U}^*\mathbf{U}^{*\top}\|_2 \leq 10^{-3}$ . For different rank  $r^*$ , we report the maximum number of the outliers existing in the samples so that an algorithm can succeed. Parameters  $\tau$  and  $\eta$  in Algorithms 8.1 and 8.2 are set to 0.9 and 0, respectively.



**Figure 8.1:** The phase transition properties of Outlier Rejection, Outlier Reduction and Outlier Pursuit in the noiseless case. (a) Random outliers. (b) Identical outliers.

Figure 8.1 shows the empirical performance of these three algorithms when the outliers are generated (a) randomly and (b) adversarially. We observe that Outlier Rejection and Outlier Reduction have similar performance on this synthetic dataset, which matches our theoretical results that the two algorithms have same theoretical guarantees. When the outliers are randomly generated, Outlier Rejection and Outlier Reduction succeed even when rank  $r^* = 30$  and there are 100 outliers, while Outlier Pursuit can only tolerate 80 outliers when  $r^* = 30$ . When the outliers are adversarial, Outlier Rejection and Outlier Reduction consistently outperform Outlier Pursuit, e.g., they can succeed when rank  $r^* = 10$  with 25 outliers but Outlier Pursuit fails.

In the second experiment, we compare the running time of Outlier Rejection and Outlier Reduction as rank  $r^*$  varies and empirically verify the condition on the fraction of outliers  $\rho$ , i.e.,  $\frac{\rho}{1-\rho} \approx O(\frac{1}{r^{*2}})$ , as shown in Theorem 8.1 and Theorem 8.3. Each test is repeated 20 times and the average results are reported. Figure 8.2(a) plots the wall clock time

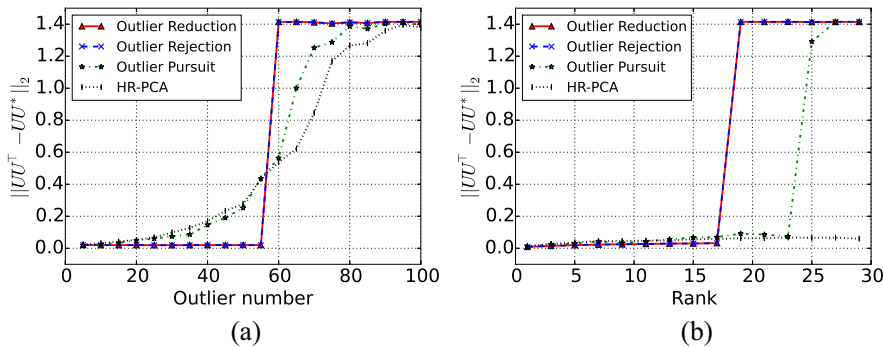


**Figure 8.2:** We plot (a) the running time of Outlier Rejection and Outlier Reduction as rank  $r^*$  increases and (b) the relationship between rank  $r^*$  and the largest tolerable fraction of outliers for Outlier Reduction.

of Outlier Rejection and Outlier Reduction against  $r^*$ . These two algorithms run much faster than Outlier Pursuit which takes around 30s to compute the solution. Clearly, their computational time increases as rank  $r^*$  grows, which is consistent with the fact that their computational cost grows linearly in rank  $r^*$ . We also observe that Outlier Reduction runs slightly slower than Outlier Rejection, as more operations are required by the truncation operator  $CT_\epsilon(\cdot)$  in Outlier Reduction. Figure 8.2(b) illustrates the relationship between rank  $r^*$  and the largest fraction of outliers  $\rho$  so that Outlier Reduction succeeds, in which

rank  $r^*$  varies from 4 to 80 and two values  $V_1 \triangleq \frac{\rho}{1-\rho} \cdot r^*$  and  $V_2 \triangleq \frac{\rho}{1-\rho} \cdot \frac{r^{*2}}{80}$  are plotted. From this figure we see that 1)  $V_1$  always decreases as  $r^*$  grows, implying that  $\frac{\rho}{1-\rho}$  should be  $o(\frac{1}{r^*})$ , and 2) when  $r^* \geq 30$ ,  $V_2$  becomes close to a constant, which empirically verifies that the upper bound for  $\frac{\rho}{1-\rho}$  shown in Theorem 8.1 and Theorem 8.2 is tight w.r.t.  $r^*$ .

The third experiment tests the performance of Outlier Rejection, Outlier Reduction, Outlier Pursuit and HR-PCA [FXY12] in the noisy case where the outliers are generated adversarially and each entry of noise  $\mathbf{N}$  is independently drawn from the Gaussian distribution  $\mathcal{N}(0, \sigma^2)$  with  $\sigma = 0.1$ . Parameters  $\tau$  and  $\eta$  for Outlier Rejection and Outlier Reduction are set to 0.9 and 0.01, respectively. Let  $\mathbf{U}$  be the leading  $r^*$  left singular vector of the output  $\mathbf{L}$  of a certain algorithm. Figure 8.3(a) and Figure 8.3(b) plot  $\|\mathbf{U}\mathbf{U}^\top - \mathbf{U}^*\mathbf{U}^{*\top}\|_2$  against the number of the outliers when rank  $r^* = 5$ , and against rank  $r^*$  when there exist 15 outliers, respectively. Obviously, when  $r^*$  is relatively small, e.g.,  $r^* \leq 15$ , our algorithms are able to



**Figure 8.3:** The comparison between Outlier Rejection, Outlier Reduction and Outlier Pursuit in the noisy case. (a) Rank  $r^* = 5$ . (b) Outlier number is 15.

generate more accurate solutions than Outlier Pursuit and HR-PCA. When  $r^*$  is large, e.g.,  $r^* \geq 20$ , and the number of outliers is small, Outlier Pursuit and HR-PCA obtain better results.

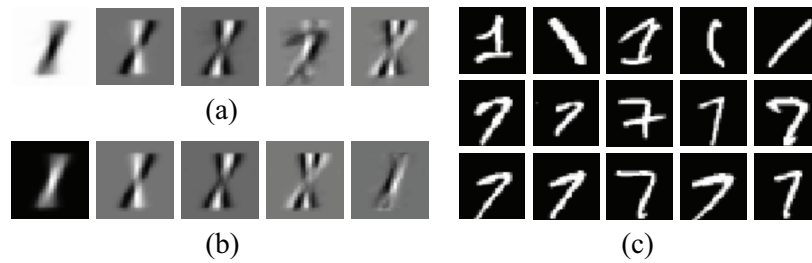
### 8.5.2 Real-world Data

We now investigate the performance of Outlier Rejection and Outlier Reduction on real-world datasets. The goal of these experiments is to show that Outlier Rejection can be used to identify outlier samples within the dataset and Outlier Reduction can be used to remove



anomalous parts in each samples, e.g., extracting the background in a video.

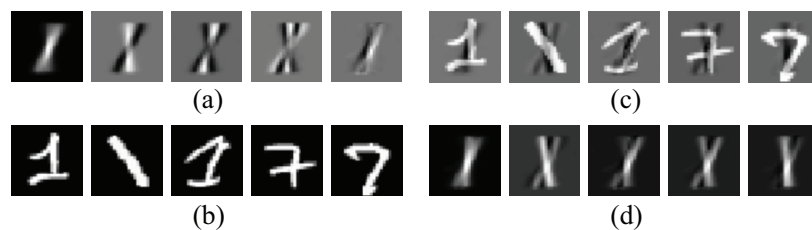
In the first experiment, we perform Outlier Rejection on a real-world dataset of 230 digit images drawn from MNIST [LJB+95] which contains 220 images of “1” and 10 images of “7”. We take “1”s as the inlier samples and “7”s as the outlier samples. Each of these images is converted into a 784-dimensional vector. The objective here is to identify all “7”s without knowing the labels of these images. For Outlier Rejection, target rank  $r$  is 5,  $\epsilon$  is set to 0.01,  $\tau$  is set to 0.98 and  $\eta$  lies in  $[0.01, 0.015]$ .  $\eta$  controls the number of the outliers one wants to detect. Typically, one can choose  $\eta = 0.012$ . Figure 8.4(a) and Figure 8.4(b) show



**Figure 8.4:** We plot the leading five principal components extracted by (a) standard PCA and (b) Outlier Rejection, and (c) the outliers identified by Outlier Rejection.

the leading five principal components extracted by standard PCA and Outlier Rejection, respectively. Note that the principal components extracted by Outlier Rejection is more reliable than standard PCA, e.g., the fourth principal component extracted by standard PCA clearly mixes “1”s with “7”s. Figure 8.4(c) shows the outliers identified by Outlier Rejection. We observe that all the “7”s and five “1”s are identified. These “1”s are identified as the outliers because they are written in a different way from the rest of “1”s.

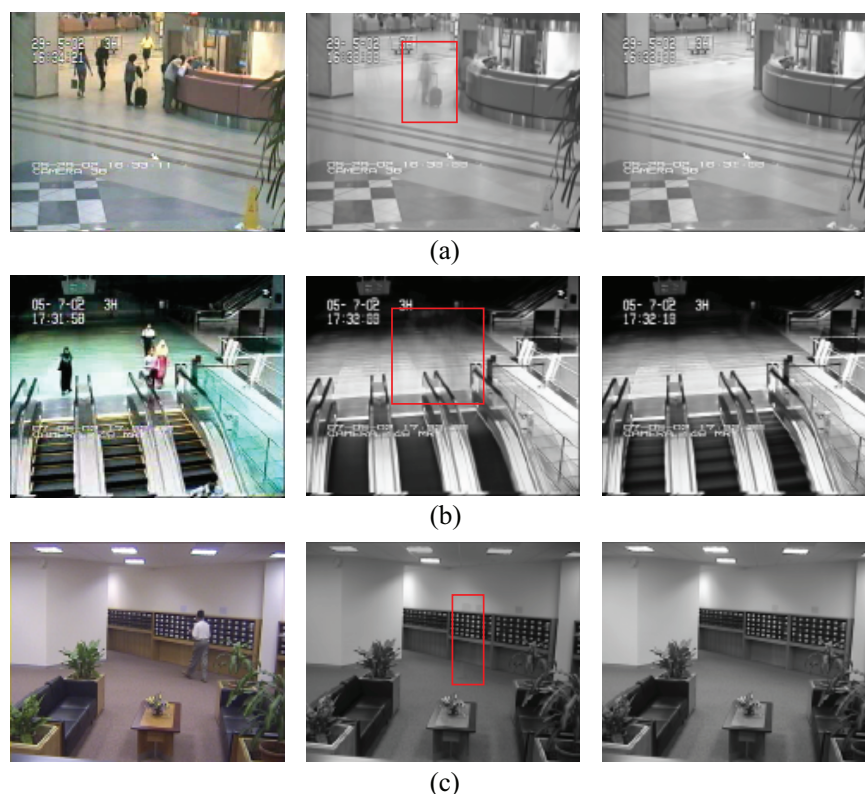
In the second experiment, we run Outlier Reduction with the same dataset and parameters as discussed above. Figure 8.5(a) plots the leading five principal components extracted by



**Figure 8.5:** (a) The leading five principal components extracted by Outlier Reduction. (b) Five “abnormal” samples. (c) Column-sparse component  $\mathbf{S}$ . (d) Low-rank matrix  $\mathbf{L}$ .

Outlier Reduction. Figure 8.5(b) gives five “abnormal” samples and Figures 8.5(c), 8.5(d) show the corresponding  $\mathbf{S}_i$  and  $\mathbf{L}_i$  computed by Step 3 and Step 5 in Algorithm 8.2. Similar to Outlier Rejection, the results obtained by Outlier Reduction are more reliable than those computed by standard PCA.

In the third experiment, we aim to extract the background in a video by computing the leading principal component of its frames via Outlier Reduction. We consider three benchmark datasets – “Hall”, “Escalator” and “Lobby” – that are used for the problem of foreground-background separation, in which the backgrounds are static and hence form low-rank components while the foregrounds are dynamic which can be taken as noise. This experiment shows that Outlier Reduction can still be applied and generate better results than standard PCA although these datasets do not follow the setting discussed in Section 8.2. Parameters  $\epsilon$ ,  $\tau$  and  $\eta$  for Outlier Reduction are set to 0.01, 0.98 and 0.001, respectively. Figure 8.6



**Figure 8.6:** We plot the results for (a) “Hall”, (b) “Escalator” and (c) “Lobby”. The left column shows the original frames. The middle column presents the leading PC extracted by standard PCA. The right column gives the leading PC extracted by Outlier Reduction.

shows the leading principal components extracted by standard PCA and Outlier Reduc-

tion. The red rectangles highlight artifacts produced by standard PCA, e.g., the shadows of people in the middle of the pictures. Obviously, Outlier Reduction obtains much better results.

In the final experiment, we test the effect of parameter  $\eta$  on the performance of Outlier Rejection. Recall that each iteration of Outlier Rejection takes the samples whose  $l_2$ -norms are greater than threshold  $\epsilon_t$  after projected onto the subspace orthogonal to the current estimated principal components as outliers. Since  $\epsilon_t$  is updated via  $\epsilon_{t+1} = \tau\epsilon_t + \eta$  where  $\tau < 1$ , we can guess that more samples will be identified as outliers as  $\eta$  becomes smaller. In

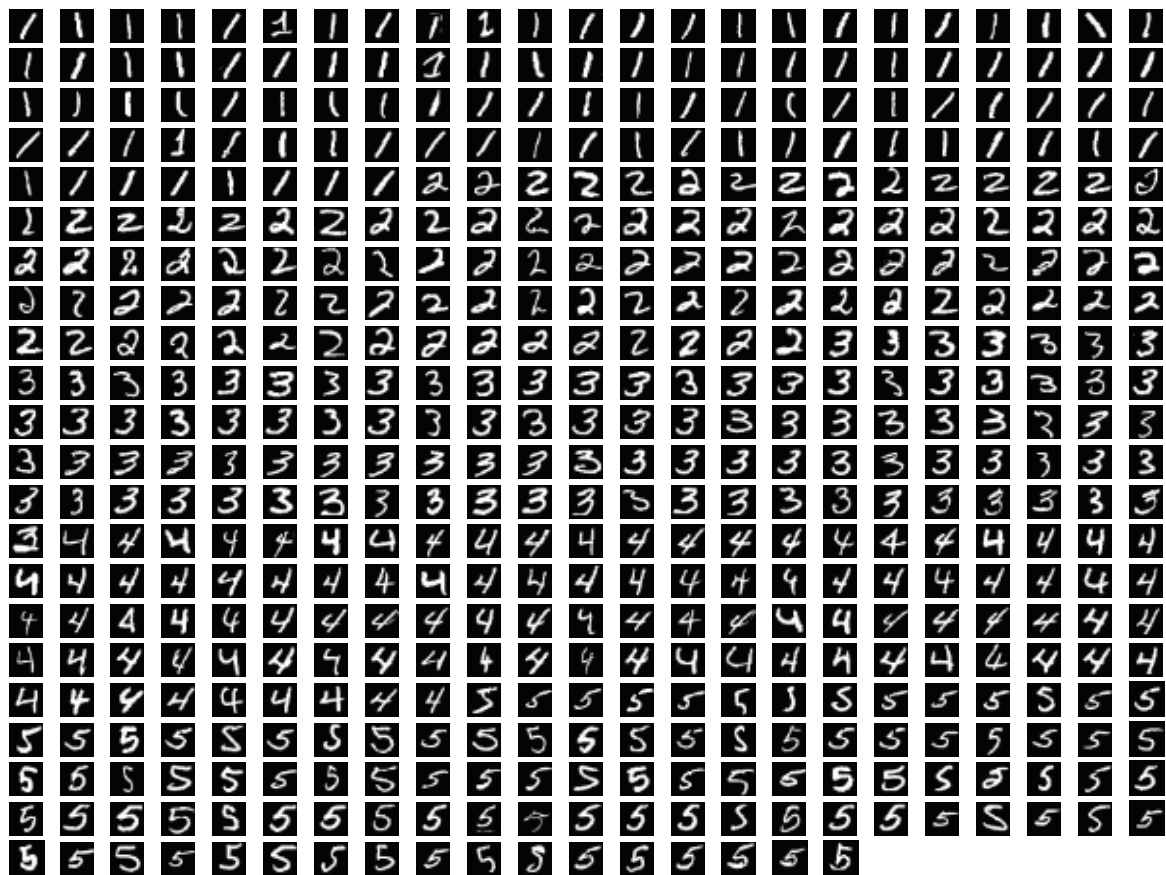


Figure 8.7: The dataset with 500 digit images of “1” to “5”.

this experiment, we construct a dataset containing 500 digit images of “1” to “5” where each digit has 100 images. Figure 8.7 shows the whole dataset from which we can observe that each digit can be written in many different ways, e.g., the first image of “1” and the sixth image of “1”. Our goal is to identify the digit images that are written quite differently from the others. We take these “abnormal” digit images as outliers and run Outlier Rejection

to detect them. We set the target rank  $r$  to 10 and parameters  $\epsilon$  and  $\tau$  to 0.1 and 0.98 respectively.

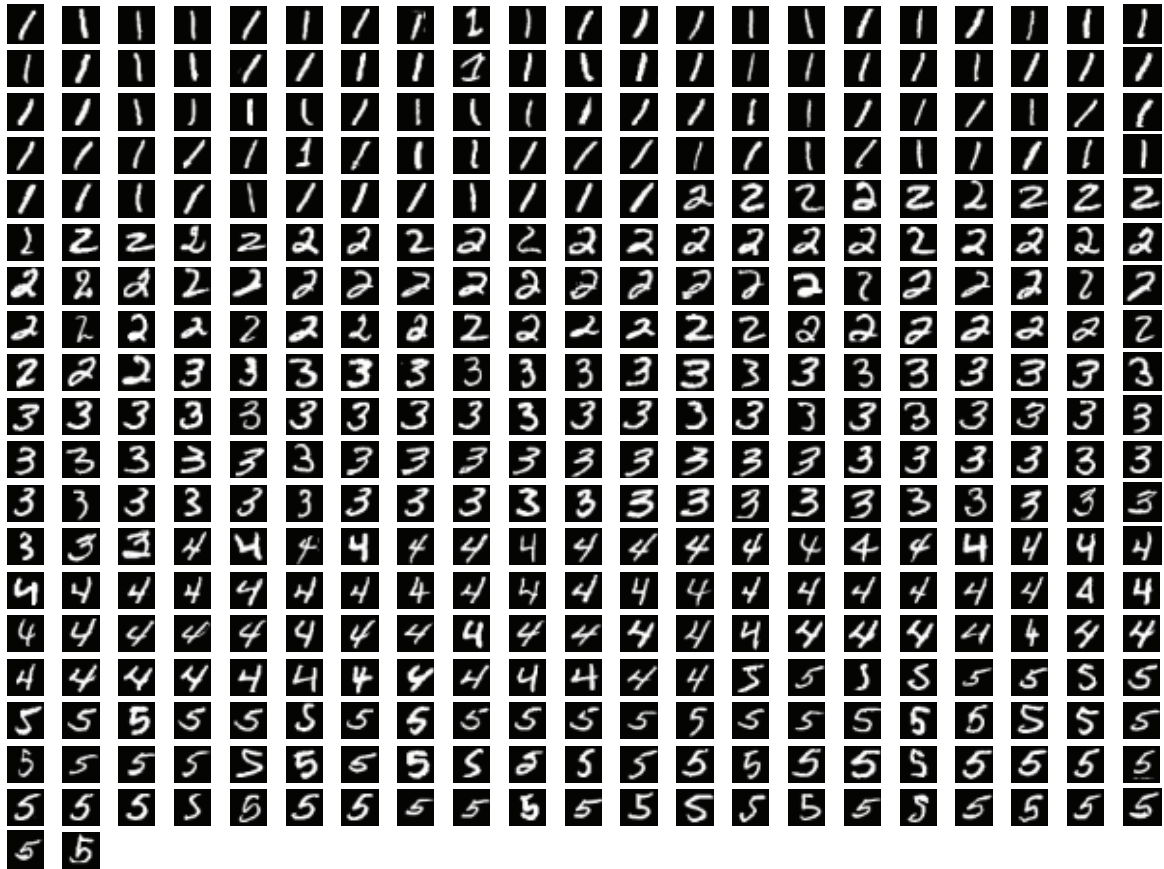


Figure 8.8: The inlier samples detected by Outlier Rejection with  $\eta = 0.012$ .

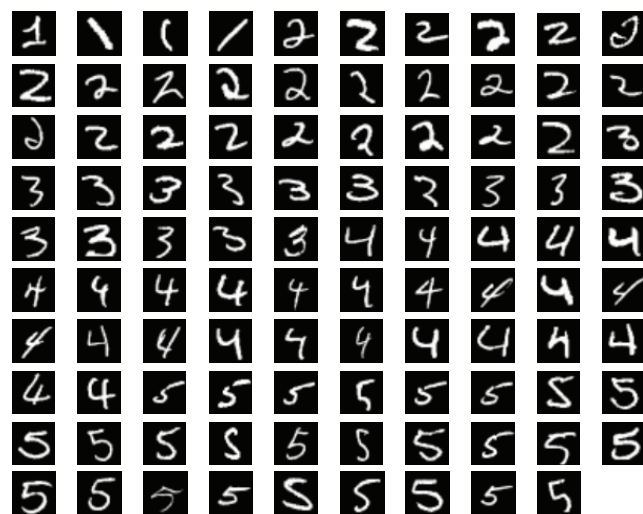


Figure 8.9: The outlier samples identified by Outlier Rejection with  $\eta = 0.012$ .

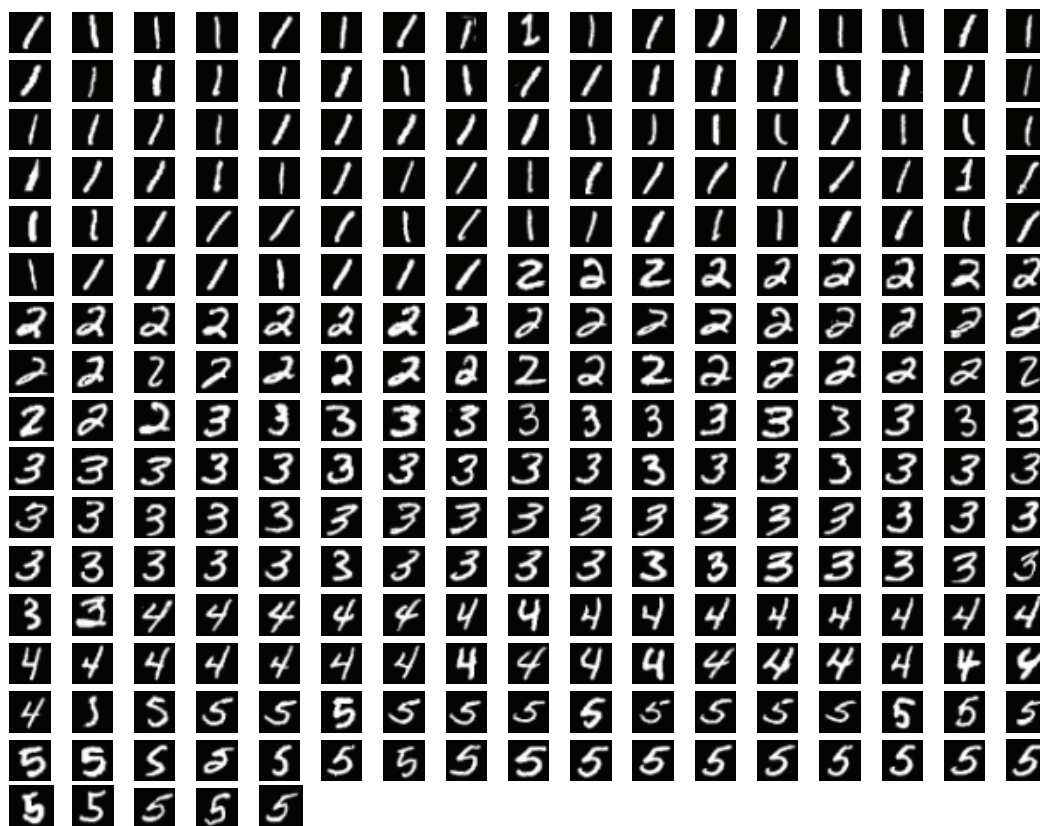


Figure 8.10: The inlier samples detected by Outlier Rejection with  $\eta = 0.01$ .

Figures 8.8 and 8.9 provide the inlier and outlier samples identified by Outlier Rejection when  $\eta = 0.012$  and Figures 8.10 and 8.11 show the results when  $\eta = 0.01$ . Obviously, when  $\eta$  increases, more “abnormal” hand-writing digit images are extracted, and the rest digit images look more “regular”, namely, they have similar shapes. Therefore, when the samples lie in a low dimensional subspace, Outlier Rejection can be used to identify the outlier samples.

## 8.6 Proofs of Technical Results

Before the main proofs are provided, we first give two useful lemmas.

**Lemma 8.1.** *Suppose that  $\mathbf{S} \in \mathbb{R}^{p \times n}$  satisfies  $\|\mathbf{S}\|_{0,2} \leq \rho n$ , then  $\|\mathbf{S}\|_2 \leq \sqrt{\rho n} \|\mathbf{S}\|_{\infty,2}$ .*

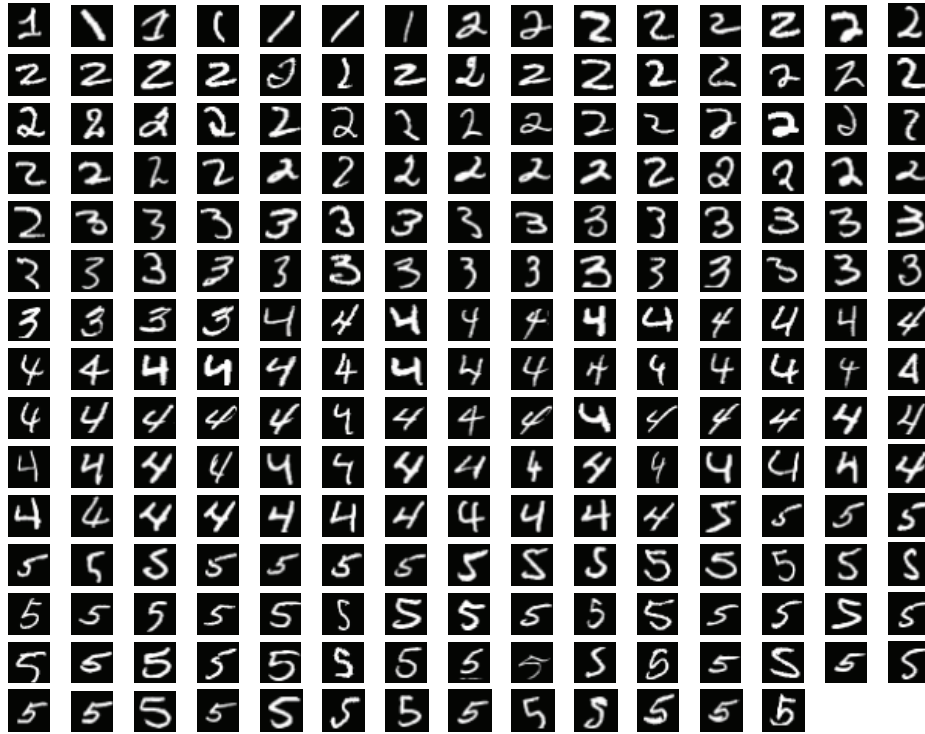


Figure 8.11: The outlier samples identified by Outlier Rejection with  $\eta = 0.01$ .

*Proof.* By the definition of the spectral norm,

$$\begin{aligned} \|\mathbf{S}\|_2^2 &= \max_{\|\mathbf{x}\|_2=1} \|\mathbf{S}\mathbf{x}\|_2^2 = \max_{\|\mathbf{x}\|_2=1} \left\| \sum_{i:\|\mathbf{S}_i\|_2 \neq 0} x_i \mathbf{S}_i \right\|_2^2 \\ &\leq \max_{\|\mathbf{x}\|_2=1} \left( \sum_{i:\|\mathbf{S}_i\|_2 \neq 0} x_i \|\mathbf{S}_i\|_2 \right)^2 \\ &\leq \max_{\|\mathbf{x}\|_2=1} \left( \sum_{i:\|\mathbf{S}_i\|_2 \neq 0} x_i \right)^2 \|\mathbf{S}\|_{\infty,2}^2 \end{aligned}$$

By the Cauchy-Schwarz inequality,

$$\max_{\|\mathbf{x}\|_2=1} \left( \sum_{i:\|\mathbf{S}_i\|_2 \neq 0} x_i \right)^2 \leq \rho n.$$

Hence this lemma holds. □

**Lemma 8.2.** Suppose that  $\mathbf{L} \in \mathbb{R}^{p \times n}$  has rank  $r$  and SVD  $\mathbf{L} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$  where  $\mathbf{U} \in \mathbb{R}^{p \times r}$  and  $\mathbf{V} \in \mathbb{R}^{n \times r}$ . For  $\bar{r} \geq r$ , if there exists an orthogonal matrix  $\bar{\mathbf{U}} \in \mathbb{R}^{p \times \bar{r}}$  satisfying  $\|(\mathbf{I} - \bar{\mathbf{U}}\bar{\mathbf{U}}^\top)\mathbf{L}\|_{\infty,2} \leq \epsilon$  for some constant  $\epsilon$ , then there exists matrix  $\hat{\mathbf{U}} \in \mathbb{R}^{p \times (\bar{r}-r)}$  so that

$\tilde{\mathbf{U}} = [\mathbf{U}, \hat{\mathbf{U}}]$  is orthogonal,  $\bar{\mathbf{U}}\bar{\mathbf{U}}^\top \hat{\mathbf{U}} = \hat{\mathbf{U}}$  and  $\|\bar{\mathbf{U}}\bar{\mathbf{U}}^\top - \tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top\|_2 \leq \frac{\sqrt{r}}{\max_{\hat{\mathbf{L}} \in \mathcal{L}} \sigma_r(\hat{\mathbf{L}})} \epsilon$ , where  $\mathcal{L}$  is the set of full rank  $p \times r$  submatrices of  $\mathbf{L}$ .

*Proof.* Since  $\|(\mathbf{I} - \bar{\mathbf{U}}\bar{\mathbf{U}}^\top)\mathbf{L}_i\|_2 \leq \epsilon$  for all  $i$ , by applying Lemma 8.1, we have that for any  $\hat{\mathbf{L}} \in \mathcal{L}$ ,

$$\|(\mathbf{I} - \bar{\mathbf{U}}\bar{\mathbf{U}}^\top)\hat{\mathbf{L}}\|_2 \leq \sqrt{r}\epsilon \Leftrightarrow \|(\mathbf{I} - \bar{\mathbf{U}}\bar{\mathbf{U}}^\top)\mathbf{U}\mathbf{U}^\top\hat{\mathbf{L}}\|_2 \leq \sqrt{r}\epsilon$$

implying that

$$\|(\mathbf{I} - \bar{\mathbf{U}}\bar{\mathbf{U}}^\top)\mathbf{U}\|_2 \leq \sqrt{r}\epsilon \|(\mathbf{U}^\top\hat{\mathbf{L}})^{-1}\|_2 = \sqrt{r}\epsilon \sigma_r(\hat{\mathbf{L}})^{-1}.$$

By the assumption that  $\text{rank}(\mathbf{U}) = r$  and  $\text{rank}(\bar{\mathbf{U}}) = \bar{r}$  where  $\bar{r} \geq r$ , we can construct  $\hat{\mathbf{U}} \in \mathbb{R}^{p \times (\bar{r}-r)}$  so that  $\tilde{\mathbf{U}} = [\mathbf{U}, \hat{\mathbf{U}}] \in \mathbb{R}^{p \times \bar{r}}$  is orthogonal and each column of  $\hat{\mathbf{U}}$  lies in the column space of  $\bar{\mathbf{U}}$ , i.e.,  $\bar{\mathbf{U}}\bar{\mathbf{U}}^\top \hat{\mathbf{U}} = \hat{\mathbf{U}}$ . Therefore, we have

$$\|(\mathbf{I} - \bar{\mathbf{U}}\bar{\mathbf{U}}^\top)\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top\|_2 = \|(\mathbf{I} - \bar{\mathbf{U}}\bar{\mathbf{U}}^\top)\tilde{\mathbf{U}}\|_2 = \|(\mathbf{I} - \bar{\mathbf{U}}\bar{\mathbf{U}}^\top)\mathbf{U}\|_2 \leq \sqrt{r}\epsilon \sigma_r(\hat{\mathbf{L}})^{-1}.$$

Let  $\bar{\mathbf{U}}_\perp$  be the orthogonal basis of the perpendicular subspace to the one spanned by the columns of  $\bar{\mathbf{U}}$ , then

$$\|\bar{\mathbf{U}}_\perp^\top \tilde{\mathbf{U}}\|_2 \leq \sqrt{r}\epsilon \sigma_r(\hat{\mathbf{L}})^{-1}.$$

By applying Theorem 2.6.1 in [GVL96],

$$\|\bar{\mathbf{U}}\bar{\mathbf{U}}^\top - \tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top\|_2 = \|\bar{\mathbf{U}}_\perp^\top \tilde{\mathbf{U}}\|_2 \leq \sqrt{r}\epsilon \sigma_r(\hat{\mathbf{L}})^{-1}.$$

Since this inequality holds for any  $\hat{\mathbf{L}} \in \mathcal{L}$ , we have

$$\|\bar{\mathbf{U}}\bar{\mathbf{U}}^\top - \tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top\|_2 \leq \frac{\sqrt{r}}{\max_{\hat{\mathbf{L}} \in \mathcal{L}} \sigma_r(\hat{\mathbf{L}})} \epsilon.$$

Hence we obtain this lemma. □

### Proof of Theorem 8.1

We prove this theorem using mathematical induction.

For  $t = 1$ , since  $\|\mathbf{X}_i\|_2 \leq 1$  for all  $i$  after the normalization step and  $\epsilon_1 = 1$ , the following inequality holds:

$$\|\mathbf{L}^* - \mathbf{U}_t \mathbf{U}_t^\top \mathbf{L}^*\|_{\infty,2} \leq \|\mathbf{L}^*\|_{\infty,2} \leq \epsilon_t.$$

For  $t > 1$ , suppose  $\|\mathbf{L}^* - \mathbf{U}_t \mathbf{U}_t^\top \mathbf{L}^*\|_{\infty,2} \leq \epsilon_t$ , then our goal is to show that after one iteration

$$\|\mathbf{L}^* - \mathbf{U}_{t+1} \mathbf{U}_{t+1}^\top \mathbf{L}^*\|_{\infty,2} \leq \epsilon_{t+1}$$

for some  $\epsilon_{t+1}$  satisfying  $\frac{\epsilon_{t+1}}{\epsilon_t} \leq \tau < 1$ .

In the noiseless case, we have  $\mathbf{X} = \mathbf{L}^* + \mathbf{S}^*$ . Let  $\mathcal{C}$  be the column support of  $\mathbf{S}^*$ , then

$$(\mathbf{X} - \mathbf{U}_t \mathbf{U}_t^\top \mathbf{X})_i = \begin{cases} (\mathbf{I} - \mathbf{U}_t \mathbf{U}_t^\top) \mathbf{S}_i^*, & i \in \mathcal{C}, \\ (\mathbf{I} - \mathbf{U}_t \mathbf{U}_t^\top) \mathbf{L}_i^*, & i \notin \mathcal{C}. \end{cases}$$

Recall that  $\mathcal{A}_t = \{i : \|(\mathbf{X} - \mathbf{U}_t \mathbf{U}_t^\top \mathbf{X})_i\|_2 > \epsilon_t\}$  shown in Step 4 of Algorithm 8.1. Since  $\|(\mathbf{I} - \mathbf{U}_t \mathbf{U}_t^\top) \mathbf{L}_i^*\|_2 \leq \epsilon_t$  holds for any  $i \notin \mathcal{C}$ , we have

$$\mathcal{A}_t \subseteq \mathcal{C}, \text{ and } \|(\mathbf{I} - \mathbf{U}_t \mathbf{U}_t^\top) \mathbf{S}_i^*\|_2 \leq \epsilon_t, \forall i \in \mathcal{A}_t^c \cap \mathcal{C}.$$

Recall that  $\text{rank}(\mathbf{U}_t) = r$ ,  $\text{rank}(\mathbf{U}^*) = r^*$  and  $r \geq r^*$ . By applying Lemma 8.2, we can construct  $\hat{\mathbf{U}} \in \mathbb{R}^{p \times (r-r^*)}$  so that  $\tilde{\mathbf{U}} = [\mathbf{U}^*, \hat{\mathbf{U}}]$  is orthogonal and each column of  $\hat{\mathbf{U}}$  lies in the column space of  $\mathbf{U}_t$ , i.e.,  $\mathbf{U}_t \mathbf{U}_t^\top \hat{\mathbf{U}} = \hat{\mathbf{U}}$ , and

$$\|\tilde{\mathbf{U}} \tilde{\mathbf{U}}^\top - \mathbf{U}_t \mathbf{U}_t^\top\|_2 \leq \frac{\sqrt{r^*}}{\max_{\mathbf{L} \in \mathcal{L}} \sigma_{r^*}(\mathbf{L})} \epsilon_t.$$

Therefore, for any  $i \in \mathcal{A}_t^c \cap \mathcal{C}$ ,

$$\begin{aligned} \|(\mathbf{I} - \tilde{\mathbf{U}} \tilde{\mathbf{U}}^\top) \mathbf{S}_i^*\|_2 &\leq \|(\mathbf{I} - \mathbf{U}_t \mathbf{U}_t^\top) \mathbf{S}_i^*\|_2 + \|(\tilde{\mathbf{U}} \tilde{\mathbf{U}}^\top - \mathbf{U}_t \mathbf{U}_t^\top) \mathbf{S}_i^*\|_2 \\ &\leq \epsilon_t + \|\tilde{\mathbf{U}} \tilde{\mathbf{U}}^\top - \mathbf{U}_t \mathbf{U}_t^\top\|_2 \\ &\leq \left(1 + \frac{\sqrt{r^*}}{\max_{\mathbf{L} \in \mathcal{L}} \sigma_{r^*}(\mathbf{L})}\right) \epsilon_t. \end{aligned} \tag{8.4}$$



By Step 5 of Algorithm 8.1,  $\mathbf{L}_{t+1}$  is constructed as follows:

$$\mathbf{L}_{t+1,i} = \begin{cases} \mathbf{0}, & i \in \mathcal{A}_t, \\ \mathbf{L}_i^*, & i \in \mathcal{A}_t^c \cap \mathcal{C}^c, \\ \mathbf{S}_i^*, & i \in \mathcal{A}_t^c \cap \mathcal{C}. \end{cases}$$

Let  $\hat{\mathbf{S}}_i^*$  be a  $p \times n$  matrix whose columns are  $\mathbf{0}$  except that the  $i$ th column equals  $\mathbf{S}_i^*$ , then we have

$$\begin{aligned} \mathbf{L}_{t+1} &= \mathbf{L}^* + \sum_{i \in \mathcal{A}_t^c \cap \mathcal{C}} \hat{\mathbf{S}}_i^* \\ &= \mathbf{L}^* + \tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top \sum_{i \in \mathcal{A}_t^c \cap \mathcal{C}} \hat{\mathbf{S}}_i^* + (\mathbf{I} - \tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top) \sum_{i \in \mathcal{A}_t^c \cap \mathcal{C}} \hat{\mathbf{S}}_i^* \\ &= \tilde{\mathbf{U}}\mathbf{R} + (\mathbf{I} - \tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top) \sum_{i \in \mathcal{A}_t^c \cap \mathcal{C}} \hat{\mathbf{S}}_i^*, \end{aligned}$$

where  $\mathbf{R} = \tilde{\mathbf{U}}^\top \mathbf{L}^* + \tilde{\mathbf{U}}^\top \sum_{i \in \mathcal{A}_t^c \cap \mathcal{C}} \hat{\mathbf{S}}_i^*$ .

For notational simplicity, we define

$$\mathbf{A} = \tilde{\mathbf{U}}\mathbf{R}, \quad \mathbf{B} = (\mathbf{I} - \tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top) \sum_{i \in \mathcal{A}_t^c \cap \mathcal{C}} \hat{\mathbf{S}}_i^*,$$

and let  $\mathcal{U}$  be the subspace spanned by  $\mathbf{U}_{t+1}$  and  $\mathcal{U}_\perp$  be the subspace orthogonal to  $\mathcal{U}$ , then

$$\begin{aligned} \|(\mathbf{I} - \mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top)\mathbf{A}\|_2 &= \max_{\|\mathbf{x}\|_2=1} \|\mathbf{x}^\top (\mathbf{I} - \mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top)\mathbf{A}\|_2 \\ &= \max_{\substack{\|\mathbf{y}\|_2=1, \mathbf{y} \in \mathcal{U}, \\ \|\mathbf{z}\|_2=1, \mathbf{z} \in \mathcal{U}_\perp, \\ \alpha^2 + \beta^2 = 1}} \|(\alpha\mathbf{y} + \beta\mathbf{z})^\top (\mathbf{I} - \mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top)\mathbf{A}\|_2 \\ &\leq \max_{\|\mathbf{z}\|_2=1, \mathbf{z} \in \mathcal{U}_\perp} \|\mathbf{z}^\top (\mathbf{I} - \mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top)\mathbf{A}\|_2 \\ &= \max_{\|\mathbf{z}\|_2=1, \mathbf{z} \in \mathcal{U}_\perp} \|\mathbf{z}^\top (\mathbf{L}_{t+1} - \mathbf{B})\|_2 \\ &\leq \max_{\|\mathbf{z}\|_2=1, \mathbf{z} \in \mathcal{U}_\perp} \|\mathbf{z}^\top \mathbf{L}_{t+1}\|_2 + \|\mathbf{B}\|_2 \\ &= \sigma_{r+1}(\mathbf{L}_{t+1}) + \|\mathbf{B}\|_2. \end{aligned}$$

Thus, by the Weyl's inequality, we have

$$\begin{aligned}
\|(\mathbf{I} - \mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top)\mathbf{A}\|_2 &\leq \sigma_{r+1}(\mathbf{A} + \mathbf{B}) + \|\mathbf{B}\|_2 \\
&\leq \sigma_{r+1}(\mathbf{A}) + 2\|\mathbf{B}\|_2 \\
&= 2\|\mathbf{B}\|_2,
\end{aligned} \tag{8.5}$$

where the equality holds because  $\text{rank}(\mathbf{A}) = r$ . Note that  $\mathbf{L}^*$  and  $\mathbf{S}^*$  have disjoint column supports, then

$$\begin{aligned}
\|(\mathbf{I} - \mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top)\mathbf{A}\|_2 &= \|(\mathbf{I} - \mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top)(\mathbf{L}^* + \tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top \sum_{i \in \mathcal{A}_i^c \cap \mathcal{C}} \hat{\mathbf{S}}_i^*)\|_2 \\
&= \max_{\|\mathbf{x}\|_2=1} \|(\mathbf{I} - \mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top)(\mathbf{L}^* + \tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top \sum_{i \in \mathcal{A}_i^c \cap \mathcal{C}} \hat{\mathbf{S}}_i^*)\mathbf{x}\|_2 \\
&\geq \max_{\|\mathbf{x}\|_2=1} \|(\mathbf{I} - \mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top)\mathbf{L}^*\mathbf{x}\|_2 \\
&= \|(\mathbf{I} - \mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top)\mathbf{L}^*\|_2,
\end{aligned} \tag{8.6}$$

By Lemma 8.1 and Inequalities (8.4), (8.5) and (8.6), we have

$$\begin{aligned}
\|(\mathbf{I} - \mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top)\mathbf{L}^*\|_2 &\leq 2\sqrt{\rho n} \max_{i \in \mathcal{A}_i^c \cap \mathcal{C}} \|(\mathbf{I} - \tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top)\mathbf{S}_i^*\|_2 \\
&\leq 2\sqrt{\rho n} \cdot \left(1 + \frac{\sqrt{r^*}}{\max_{\mathbf{L} \in \mathcal{L}} \sigma_{r^*}(\mathbf{L})}\right) \epsilon_t.
\end{aligned}$$

Thus, by the incoherent condition, i.e.,  $\max_i \|\mathbf{e}_i^\top \mathbf{V}^*\|_2^2 \leq \frac{\mu r^*}{(1-\rho)n}$ , we have

$$\begin{aligned}
\|(\mathbf{I} - \mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top)\mathbf{L}^*\|_{\infty,2} &= \max_i \|(\mathbf{I} - \mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top)\mathbf{U}^*\mathbf{\Sigma}^*\mathbf{V}^{*\top} \mathbf{e}_i\|_2 \\
&\leq \|(\mathbf{I} - \mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top)\mathbf{U}^*\mathbf{\Sigma}^*\|_2 \cdot \max_i \|\mathbf{V}^{*\top} \mathbf{e}_i\|_2 \\
&\leq \sqrt{\frac{\mu r^*}{(1-\rho)n}} \|(\mathbf{I} - \mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top)\mathbf{L}^*\|_2 \\
&\leq 2\sqrt{\frac{\mu r^* \rho}{1-\rho}} \left(1 + \frac{\sqrt{r^*}}{\max_{\mathbf{L} \in \mathcal{L}} \sigma_{r^*}(\mathbf{L})}\right) \epsilon_t.
\end{aligned} \tag{8.7}$$

Therefore, as long as  $2\sqrt{\frac{\mu r^* \rho}{(1-\rho)}} \left(1 + \frac{\sqrt{r^*}}{\max_{\mathbf{L} \in \mathcal{L}} \sigma_{r^*}(\mathbf{L})}\right) < 1$  or equivalently

$$\frac{\rho}{1-\rho} < \frac{1}{4\mu r^* (1 + \sqrt{r^*}/\max_{\mathbf{L} \in \mathcal{L}} \sigma_{r^*}(\mathbf{L}))^2},$$

there exists  $\tau < 1$  so that  $2\sqrt{\frac{\mu r^* \rho}{(1-\rho)}}(1 + \frac{\sqrt{r^*}}{\sigma_{r^*}(\mathbf{L})}) \leq \tau$  and

$$\|(\mathbf{I} - \mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top)\mathbf{L}^*\|_{\infty,2} \leq \tau\epsilon_t = \epsilon_{t+1},$$

implying that

$$\|(\mathbf{I} - \mathbf{U}_T\mathbf{U}_T^\top)\mathbf{L}^*\|_{\infty,2} \leq \tau^{T-1}\epsilon_1 = \tau^{T-1}.$$

By taking  $T \geq \log \epsilon / \log \tau + 1$ , we have

$$\|(\mathbf{I} - \mathbf{U}_T\mathbf{U}_T^\top)\mathbf{L}^*\|_{\infty,2} \leq \epsilon_T \leq \epsilon.$$

Finally, when  $r = r^*$ , by applying Lemma 8.2, we have

$$\|\mathbf{U}_T\mathbf{U}_T^\top - \mathbf{U}^*\mathbf{U}^{*\top}\|_2 \leq \frac{\sqrt{r}}{\max_{\mathbf{L} \in \mathcal{L}} \sigma_r(\mathbf{L})} \epsilon_T.$$

Hence we obtain this theorem.

### Proof of Theorem 8.2

The proof is similar to the proof of Theorem 8.1. We prove this theorem using mathematical induction.

For  $t = 1$ , since  $\epsilon_1 = 1$  and  $\|\mathbf{X}_i\|_2 \leq 1$  for all  $i$  after the normalization step, the following inequality holds:

$$\|(\mathbf{I} - \mathbf{U}_t\mathbf{U}_t^\top)(\mathbf{L}_i^* + \mathbf{N}_i)\|_2 \leq \|\mathbf{L}_i^* + \mathbf{N}_i\|_2 \leq \epsilon_t.$$

For  $t > 1$ , suppose  $\|(\mathbf{I} - \mathbf{U}_t\mathbf{U}_t^\top)(\mathbf{L}_i^* + \mathbf{N}_i)\|_2 \leq \epsilon_t$  for  $i \notin \mathcal{C}$ , our goal is to show that after one iteration

$$\|(\mathbf{I} - \mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top)(\mathbf{L}_i^* + \mathbf{N}_i)\|_2 \leq \epsilon_{t+1}$$

holds for some  $\epsilon_{t+1}$ .

Let  $\mathcal{C}$  be the column support of  $\mathbf{S}^*$ . In the noisy case, recall that  $\mathbf{X} = \mathbf{L}^* + \mathbf{S}^* + \mathbf{N}$ , then

$$(\mathbf{X} - \mathbf{U}_t \mathbf{U}_t^\top \mathbf{X})_i = \begin{cases} (\mathbf{I} - \mathbf{U}_t \mathbf{U}_t^\top)(\mathbf{S}_i^* + \mathbf{N}_i), & i \in \mathcal{C}, \\ (\mathbf{I} - \mathbf{U}_t \mathbf{U}_t^\top)(\mathbf{L}_i^* + \mathbf{N}_i), & i \notin \mathcal{C}. \end{cases}$$

Since  $\mathcal{A}_t = \{i : \|(\mathbf{X} - \mathbf{U}_t \mathbf{U}_t^\top \mathbf{X})_i\|_2 > \epsilon_t\}$  and  $\|(\mathbf{I} - \mathbf{U}_t \mathbf{U}_t^\top)(\mathbf{L}_i^* + \mathbf{N}_i)\|_2 \leq \epsilon_t$  for any  $i \notin \mathcal{C}$ , we have

$$\mathcal{A}_t \subseteq \mathcal{C}, \text{ and } \|(\mathbf{I} - \mathbf{U}_t \mathbf{U}_t^\top)(\mathbf{S}_i^* + \mathbf{N}_i)\|_2 \leq \epsilon_t, \forall i \in \mathcal{A}_t^c \cap \mathcal{C}. \quad (8.8)$$

Recall that  $\text{rank}(\mathbf{U}_t) = r$ ,  $\text{rank}(\mathbf{U}^*) = r^*$  and  $r \geq r^*$ . By Lemma 8.2, we can construct  $\hat{\mathbf{U}} \in \mathbb{R}^{p \times (r-r^*)}$  so that  $\tilde{\mathbf{U}} = [\mathbf{U}^*, \hat{\mathbf{U}}]$  is orthogonal and each column of  $\hat{\mathbf{U}}$  lies in the column space of  $\mathbf{U}_t$ , i.e.,  $\mathbf{U}_t \mathbf{U}_t^\top \hat{\mathbf{U}} = \hat{\mathbf{U}}$ , and

$$\|\tilde{\mathbf{U}} \tilde{\mathbf{U}}^\top - \mathbf{U}_t \mathbf{U}_t^\top\|_2 \leq \frac{\sqrt{r^*}}{\max_{\mathbf{L} \in \mathcal{L}} \sigma_{r^*}(\mathbf{L})} (\epsilon_t + \|\mathbf{N}\|_{\infty, 2}).$$

We define  $\alpha_t = \frac{\sqrt{r^*}}{\max_{\mathbf{L} \in \mathcal{L}} \sigma_{r^*}(\mathbf{L})} (\epsilon_t + \|\mathbf{N}\|_{\infty, 2})$ , then for any  $i \in \mathcal{A}_t^c \cap \mathcal{C}$ , we have

$$\begin{aligned} \|(\mathbf{I} - \tilde{\mathbf{U}} \tilde{\mathbf{U}}^\top)(\mathbf{S}_i^* + \mathbf{N}_i)\|_2 &\leq \epsilon_t + \|\tilde{\mathbf{U}} \tilde{\mathbf{U}}^\top - \mathbf{U}_t \mathbf{U}_t^\top\|_2 \\ &\leq \epsilon_t + \alpha_t. \end{aligned} \quad (8.9)$$

where the first inequality follows from Inequality (8.8) and  $\|\mathbf{S}_i^* + \mathbf{N}_i\| \leq 1$ . By Step 5 of Algorithm 8.1,  $\mathbf{L}_{t+1}$  satisfies

$$\mathbf{L}_{t+1, i} = \begin{cases} \mathbf{0}, & i \in \mathcal{A}_t, \\ \mathbf{L}_i^* + \mathbf{N}_i, & i \in \mathcal{A}_t^c \cap \mathcal{C}^c, \\ \mathbf{S}_i^* + \mathbf{N}_i, & i \in \mathcal{A}_t^c \cap \mathcal{C}. \end{cases}$$

Let  $\hat{\mathbf{S}}_i^*$  be a  $p \times n$  matrix whose columns are  $\mathbf{0}$  except that the  $i$ th column equals  $\mathbf{S}_i^*$  and  $\hat{\mathbf{N}}_i$

be a  $p \times n$  matrix whose columns are  $\mathbf{0}$  except that the  $i$ th column equals  $\mathbf{N}_i$ , then

$$\begin{aligned}
\mathbf{L}_{t+1} &= \mathbf{L}^* + \sum_{i \in \mathcal{A}_i^c \cap \mathcal{C}} \hat{\mathbf{S}}_i^* + \sum_{i \in \mathcal{A}_i^c} \hat{\mathbf{N}}_i \\
&= \mathbf{L}^* + \sum_{i \in \mathcal{C}^c} \hat{\mathbf{N}}_i + \tilde{\mathbf{U}} \tilde{\mathbf{U}}^\top \sum_{i \in \mathcal{A}_i^c \cap \mathcal{C}} (\hat{\mathbf{S}}_i^* + \hat{\mathbf{N}}_i) + \\
&\quad (\mathbf{I} - \tilde{\mathbf{U}} \tilde{\mathbf{U}}^\top) \sum_{i \in \mathcal{A}_i^c \cap \mathcal{C}} (\hat{\mathbf{S}}_i^* + \hat{\mathbf{N}}_i) \\
&= \tilde{\mathbf{U}} \mathbf{R} + \sum_{i \in \mathcal{C}^c} \hat{\mathbf{N}}_i + (\mathbf{I} - \tilde{\mathbf{U}} \tilde{\mathbf{U}}^\top) \sum_{i \in \mathcal{A}_i^c \cap \mathcal{C}} (\hat{\mathbf{S}}_i^* + \hat{\mathbf{N}}_i),
\end{aligned}$$

where  $\mathbf{R} = \tilde{\mathbf{U}}^\top \mathbf{L}^* + \tilde{\mathbf{U}}^\top \sum_{i \in \mathcal{A}_i^c \cap \mathcal{C}} (\hat{\mathbf{S}}_i^* + \hat{\mathbf{N}}_i)$ .

For notational simplicity, we define

$$\mathbf{A} = \tilde{\mathbf{U}} \mathbf{R}, \quad \mathbf{B} = (\mathbf{I} - \tilde{\mathbf{U}} \tilde{\mathbf{U}}^\top) \sum_{i \in \mathcal{A}_i^c \cap \mathcal{C}} (\hat{\mathbf{S}}_i^* + \hat{\mathbf{N}}_i),$$

and let  $\mathcal{U}$  be the subspace spanned by  $\mathbf{U}_{t+1}$  and  $\mathcal{U}_\perp$  be the subspace orthogonal to  $\mathcal{U}$ , then

$$\begin{aligned}
\|(\mathbf{I} - \mathbf{U}_{t+1} \mathbf{U}_{t+1}^\top) \mathbf{A}\|_2 &= \max_{\|\mathbf{x}\|_2=1} \|\mathbf{x}^\top (\mathbf{I} - \mathbf{U}_{t+1} \mathbf{U}_{t+1}^\top) \mathbf{A}\|_2 \\
&= \max_{\substack{\|\mathbf{y}\|_2=1, \mathbf{y} \in \mathcal{U}, \\ \|\mathbf{z}\|_2=1, \mathbf{z} \in \mathcal{U}_\perp, \\ \alpha^2 + \beta^2=1}} \|(\alpha \mathbf{y} + \beta \mathbf{z})^\top (\mathbf{I} - \mathbf{U}_{t+1} \mathbf{U}_{t+1}^\top) \mathbf{A}\|_2 \\
&\leq \max_{\|\mathbf{z}\|_2=1, \mathbf{z} \in \mathcal{U}_\perp} \|\mathbf{z}^\top (\mathbf{I} - \mathbf{U}_{t+1} \mathbf{U}_{t+1}^\top) \mathbf{A}\|_2 \\
&= \max_{\|\mathbf{z}\|_2=1, \mathbf{z} \in \mathcal{U}_\perp} \|\mathbf{z}^\top (\mathbf{L}_{t+1} - \mathbf{B} - \sum_{i \in \mathcal{C}^c} \hat{\mathbf{N}}_i)\|_2 \\
&\leq \max_{\|\mathbf{z}\|_2=1, \mathbf{z} \in \mathcal{U}_\perp} \|\mathbf{z}^\top \mathbf{L}_{t+1}\|_2 + \|\mathbf{B}\|_2 + \|\mathbf{N}\|_2 \\
&\leq 2(\|\mathbf{B}\|_2 + \|\mathbf{N}\|_2),
\end{aligned} \tag{8.10}$$

where the last inequality follows from the Weyl's inequality. Let  $\mathbf{P}_{t+1} = \mathbf{U}_{t+1} \mathbf{U}_{t+1}^\top$ , since

$\mathbf{L}^*$  is supported on  $\mathcal{C}^c$ , we have

$$\begin{aligned}
\|(\mathbf{I} - \mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top)\mathbf{A}\|_2 &= \|(\mathbf{I} - \mathbf{P}_{t+1})(\mathbf{L}^* + \tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top \sum_{i \in \mathcal{A}_i^c \cap \mathcal{C}} (\hat{\mathbf{S}}_i^* + \hat{\mathbf{N}}_i))\|_2 \\
&= \max_{\|\mathbf{x}\|_2=1} \|(\mathbf{I} - \mathbf{P}_{t+1})(\mathbf{L}^* + \tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top \sum_{i \in \mathcal{A}_i^c \cap \mathcal{C}} (\hat{\mathbf{S}}_i^* + \hat{\mathbf{N}}_i))\mathbf{x}\|_2 \\
&\geq \max_{\|\mathbf{x}\|_2=1} \|(\mathbf{I} - \mathbf{P}_{t+1})\mathbf{L}^*\mathbf{x}\|_2 \\
&= \|(\mathbf{I} - \mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top)\mathbf{L}^*\|_2.
\end{aligned} \tag{8.11}$$

By applying Lemma 8.1 and Inequalities (8.9), (8.10) and (8.11), we have

$$\begin{aligned}
\|(\mathbf{I} - \mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top)\mathbf{L}^*\|_2 &\leq 2\sqrt{\rho n} \max_{i \in \mathcal{A}_i^c \cap \mathcal{C}} \|(\mathbf{I} - \mathbf{U}^*\mathbf{U}^{*\top})(\mathbf{S}_i^* + \mathbf{N}_i)\|_2 + 2\|\mathbf{N}\|_2 \\
&\leq 2\sqrt{\rho n}(\epsilon_t + \alpha_t) + 2\|\mathbf{N}\|_2.
\end{aligned}$$

Thus, by the incoherent condition, i.e.,  $\max_i \|\mathbf{e}_i^\top \mathbf{V}^*\|_2^2 \leq \frac{\mu r^*}{(1-\rho)n}$ , for any  $i \notin \mathcal{C}$ ,

$$\begin{aligned}
&\|(\mathbf{I} - \mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top)(\mathbf{L}_i^* + \mathbf{N}_i)\|_2 \\
&\leq \|(\mathbf{I} - \mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top)\mathbf{U}^*\mathbf{\Sigma}^*\mathbf{V}^{*\top}\mathbf{e}_i\|_2 + \|\mathbf{N}_i\|_2 \\
&\leq \|(\mathbf{I} - \mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top)\mathbf{U}^*\mathbf{\Sigma}^*\|_2 \cdot \max_i \|\mathbf{V}^{*\top}\mathbf{e}_i\|_2 + \|\mathbf{N}_i\|_2 \\
&\leq \sqrt{\frac{\mu r^*}{(1-\rho)n}} \|(\mathbf{I} - \mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top)\mathbf{L}^*\|_2 + \|\mathbf{N}_i\|_2 \\
&\leq 2\sqrt{\frac{\mu r^* \rho}{1-\rho}}(\epsilon_t + \alpha_t) + 2\sqrt{\frac{\mu r^*}{(1-\rho)n}} \|\mathbf{N}\|_2 + \|\mathbf{N}\|_{\infty,2} \\
&\leq 2\sqrt{\frac{\mu r^* \rho}{1-\rho}}(\epsilon_t + \alpha_t) + (2\sqrt{\frac{\mu r^*}{1-\rho}} + 1)\|\mathbf{N}\|_{\infty,2}.
\end{aligned} \tag{8.12}$$

By substituting  $\alpha_t = \frac{\sqrt{r^*}}{\max_{\mathbf{L} \in \mathcal{L}} \sigma_{r^*}(\mathbf{L})}(\epsilon_t + \|\mathbf{N}\|_{\infty,2})$  into (8.12), for  $i \notin \mathcal{C}$  we have

$$\begin{aligned}
\|(\mathbf{I} - \mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top)(\mathbf{L}_i^* + \mathbf{N}_i)\|_2 &\leq 2\sqrt{\frac{\mu r^* \rho}{(1-\rho)}} \left(1 + \frac{\sqrt{r^*}}{\max_{\mathbf{L} \in \mathcal{L}} \sigma_{r^*}(\mathbf{L})}\right) \epsilon_t + \\
&\quad (2r^* \sqrt{\frac{\mu \rho}{(1-\rho) \max_{\mathbf{L} \in \mathcal{L}} \sigma_{r^*}(\mathbf{L})^2}} + 2\sqrt{\frac{\mu r^*}{1-\rho}} + 1) \|\mathbf{N}\|_{\infty,2}.
\end{aligned}$$

Thus, when

$$2\sqrt{\frac{\mu r^* \rho}{(1-\rho)}} \left(1 + \frac{\sqrt{r^*}}{\max_{\mathbf{L} \in \mathcal{L}} \sigma_{r^*}(\mathbf{L})}\right) \leq \tau < 1$$

and

$$\eta \geq \left(2\sqrt{\frac{\mu r^*}{1-\rho}} + 2r^* \sqrt{\frac{\mu \rho}{(1-\rho) \max_{\mathbf{L} \in \mathcal{L}} \sigma_{r^*}(\mathbf{L})^2}} + 1\right) \|\mathbf{N}\|_{\infty,2},$$

we have

$$\|(\mathbf{I} - \mathbf{U}_{t+1} \mathbf{U}_{t+1}^\top)(\mathbf{L}_i^* + \mathbf{N}_i)\|_2 \leq \tau \epsilon_t + \eta,$$

which implies that

$$\|(\mathbf{I} - \mathbf{U}_T \mathbf{U}_T^\top)(\mathbf{L}_i^* + \mathbf{N}_i)\|_2 \leq \tau^{T-1} + \frac{1 - \tau^{T-1}}{1 - \tau} \eta.$$

Therefore,

$$\|(\mathbf{I} - \mathbf{U}_T \mathbf{U}_T^\top) \mathbf{L}^*\|_{\infty,2} \leq \tau^{T-1} + \frac{\eta}{1 - \tau} + \|\mathbf{N}\|_{\infty,2}.$$

By taking  $T \geq \log \epsilon / \log \tau + 1$ , we have

$$\|(\mathbf{I} - \mathbf{U}_T \mathbf{U}_T^\top) \mathbf{L}^*\|_{\infty,2} \leq \epsilon + \frac{\eta}{1 - \tau} + \|\mathbf{N}\|_{\infty,2}.$$

Finally, when  $r = r^*$ , by applying Lemma 8.2 again, we obtain this theorem.

### Proof of Theorem 8.3

The proof of this theorem is almost the same as the proof of Theorem 8.1.

For  $t = 1$ , since  $\epsilon_1 = 1$  and  $\|\mathbf{X}_i\|_2 \leq 1$  for all  $i$  after the normalization step, the following inequality holds:

$$\|\mathbf{L}^* - \mathbf{U}_t \mathbf{U}_t^\top \mathbf{L}^*\|_{\infty,2} \leq \|\mathbf{L}^*\|_{\infty,2} \leq \epsilon_t.$$

For  $t > 1$ , suppose that  $\|\mathbf{L}^* - \mathbf{U}_t \mathbf{U}_t^\top \mathbf{L}^*\|_{\infty,2} \leq \epsilon_t$ , then our goal is to show that after one iteration

$$\|\mathbf{L}^* - \mathbf{U}_{t+1} \mathbf{U}_{t+1}^\top \mathbf{L}^*\|_{\infty,2} \leq \epsilon_{t+1}$$

holds for some  $\epsilon_{t+1}$  satisfying  $\frac{\epsilon_{t+1}}{\epsilon_t} \leq \tau < 1$ .

Recall that  $\mathbf{X} = \mathbf{L}^* + \mathbf{S}^*$ , then

$$(\mathbf{X} - \mathbf{U}_t \mathbf{U}_t^\top \mathbf{X})_i = \begin{cases} (\mathbf{I} - \mathbf{U}_t \mathbf{U}_t^\top) \mathbf{S}_i^*, & i \in \mathcal{C}, \\ (\mathbf{I} - \mathbf{U}_t \mathbf{U}_t^\top) \mathbf{L}_i^*, & i \notin \mathcal{C}. \end{cases}$$

We define  $\mathcal{A} = \{i : \|(\mathbf{X} - \mathbf{U}_t \mathbf{U}_t^\top \mathbf{X})_i\|_2 > \epsilon_t\}$ . Since  $\|(\mathbf{I} - \mathbf{U}_t \mathbf{U}_t^\top) \mathbf{L}_i^*\|_2 \leq \epsilon_t$  holds for any  $i \notin \mathcal{C}$ , we have

$$\mathcal{A} \subseteq \mathcal{C}, \text{ and } \|(\mathbf{I} - \mathbf{U}_t \mathbf{U}_t^\top) \mathbf{S}_i^*\|_2 \leq \epsilon_t, \forall i \in \mathcal{A}^c \cap \mathcal{C}.$$

Recall that  $\text{rank}(\mathbf{U}_t) = r$ ,  $\text{rank}(\mathbf{U}^*) = r^*$  and  $r \geq r^*$ . By applying Lemma 8.2, we can construct  $\hat{\mathbf{U}} \in \mathbb{R}^{p \times (r-r^*)}$  so that  $\tilde{\mathbf{U}} = [\mathbf{U}^*, \hat{\mathbf{U}}]$  is orthogonal and each column of  $\hat{\mathbf{U}}$  lies in the column space of  $\mathbf{U}_t$ , i.e.,  $\mathbf{U}_t \mathbf{U}_t^\top \hat{\mathbf{U}} = \hat{\mathbf{U}}$ , and

$$\|\tilde{\mathbf{U}} \tilde{\mathbf{U}}^\top - \mathbf{U}_t \mathbf{U}_t^\top\|_2 \leq \frac{\sqrt{r^*}}{\max_{\mathbf{L} \in \mathcal{L}} \sigma_{r^*}(\mathbf{L})} \epsilon_t. \quad (8.13)$$

Therefore, for  $i \in \mathcal{A}^c \cap \mathcal{C}$ ,

$$\begin{aligned} \|(\mathbf{I} - \tilde{\mathbf{U}} \tilde{\mathbf{U}}^\top) \mathbf{S}_i^*\|_2 &\leq \|(\mathbf{I} - \mathbf{U}_t \mathbf{U}_t^\top) \mathbf{S}_i^*\|_2 + \|\tilde{\mathbf{U}} \tilde{\mathbf{U}}^\top - \mathbf{U}_t \mathbf{U}_t^\top\|_2 \|\mathbf{S}_i^*\|_2 \\ &\leq \left(1 + \frac{\sqrt{r^*}}{\max_{\mathbf{L} \in \mathcal{L}} \sigma_{r^*}(\mathbf{L})}\right) \epsilon_t. \end{aligned} \quad (8.14)$$

Following from Step 3 and Step 5 of Algorithm 8.2,  $\mathbf{L}_{t+1}$  satisfies

$$\mathbf{L}_{t+1,i} = \begin{cases} \mathbf{U}_t \mathbf{U}_t^\top \mathbf{S}_i^*, & i \in \mathcal{A}, \\ \mathbf{L}_i^*, & i \in \mathcal{A}^c \cap \mathcal{C}^c, \\ \mathbf{S}_i^*, & i \in \mathcal{A}^c \cap \mathcal{C}. \end{cases}$$



Let  $\hat{\mathbf{S}}_i^*$  be a  $p \times n$  matrix whose columns are  $\mathbf{0}$  except that the  $i$ th column equals  $\mathbf{S}_i^*$ , then

$$\begin{aligned} \mathbf{L}_{t+1} &= \mathbf{L}^* + \sum_{i \in \mathcal{A}^c \cap \mathcal{C}} \hat{\mathbf{S}}_i^* + \mathbf{U}_t \mathbf{U}_t^\top \sum_{i \in \mathcal{A}} \hat{\mathbf{S}}_i^* \\ &= \mathbf{L}^* + \tilde{\mathbf{U}} \tilde{\mathbf{U}}^\top \left( \sum_{i \in \mathcal{A}^c \cap \mathcal{C}} \hat{\mathbf{S}}_i^* + \mathbf{U}_t \mathbf{U}_t^\top \sum_{i \in \mathcal{A}} \hat{\mathbf{S}}_i^* \right) + (\mathbf{I} - \tilde{\mathbf{U}} \tilde{\mathbf{U}}^\top) \left( \sum_{i \in \mathcal{A}^c \cap \mathcal{C}} \hat{\mathbf{S}}_i^* + \mathbf{U}_t \mathbf{U}_t^\top \sum_{i \in \mathcal{A}} \hat{\mathbf{S}}_i^* \right) \\ &= \tilde{\mathbf{U}} \mathbf{R} + \mathbf{B}, \end{aligned}$$

where

$$\mathbf{R} = \tilde{\mathbf{U}}^\top \mathbf{L}^* + \tilde{\mathbf{U}}^\top \left( \sum_{i \in \mathcal{A}^c \cap \mathcal{C}} \hat{\mathbf{S}}_i^* + \mathbf{U}_t \mathbf{U}_t^\top \sum_{i \in \mathcal{A}} \hat{\mathbf{S}}_i^* \right),$$

and

$$\mathbf{B} = (\mathbf{I} - \tilde{\mathbf{U}} \tilde{\mathbf{U}}^\top) \left( \sum_{i \in \mathcal{A}^c \cap \mathcal{C}} \hat{\mathbf{S}}_i^* + \mathbf{U}_t \mathbf{U}_t^\top \sum_{i \in \mathcal{A}} \hat{\mathbf{S}}_i^* \right).$$

Let  $\alpha_t = \frac{\sqrt{r^*}}{\max_{\mathbf{L} \in \mathcal{L}} \sigma_{r^*}(\mathbf{L})} \epsilon_t$ . By Inequalities (8.13) and (8.14), we have

$$\begin{aligned} \|\mathbf{B}\|_{\infty,2} &\leq \max \left\{ \max_{i \in \mathcal{A}^c \cap \mathcal{C}} \|(\mathbf{I} - \tilde{\mathbf{U}} \tilde{\mathbf{U}}^\top) \mathbf{S}_i^*\|_2, \max_{i \in \mathcal{A}} \|(\mathbf{I} - \tilde{\mathbf{U}} \tilde{\mathbf{U}}^\top) \mathbf{U}_t \mathbf{U}_t^\top \mathbf{S}_i^*\|_2 \right\} \\ &\leq \max \{ \epsilon_t + \alpha_t, \|(\mathbf{I} - \tilde{\mathbf{U}} \tilde{\mathbf{U}}^\top) \mathbf{U}_t \mathbf{U}_t^\top\|_2 \} \\ &= \max \{ \epsilon_t + \alpha_t, \|\tilde{\mathbf{U}} \tilde{\mathbf{U}}^\top - \mathbf{U}_t \mathbf{U}_t^\top\|_2 \} \\ &= \epsilon_t + \alpha_t. \end{aligned}$$

Then by applying Lemma 8.1 and Inequalities (8.5) and (8.6), we have

$$\|(\mathbf{I} - \mathbf{U}_{t+1} \mathbf{U}_{t+1}^\top) \mathbf{L}^*\|_2 \leq 2\sqrt{\rho n} \|\mathbf{B}\|_{\infty,2} \leq 2\sqrt{\rho n} (\epsilon_t + \alpha_t).$$

Thus, by the incoherent condition, i.e.,  $\max_i \|\mathbf{e}_i^\top \mathbf{V}^*\|_2^2 \leq \frac{\mu r^*}{(1-\rho)n}$ ,

$$\begin{aligned} \|(\mathbf{I} - \mathbf{U}_{t+1} \mathbf{U}_{t+1}^\top) \mathbf{L}^*\|_{\infty,2} &= \max_i \|(\mathbf{I} - \mathbf{U}_{t+1} \mathbf{U}_{t+1}^\top) \mathbf{U}^* \boldsymbol{\Sigma}^* \mathbf{V}^{*\top} \mathbf{e}_i\|_2 \\ &\leq \|(\mathbf{I} - \mathbf{U}_{t+1} \mathbf{U}_{t+1}^\top) \mathbf{U}^* \boldsymbol{\Sigma}^*\|_2 \cdot \max_i \|\mathbf{V}^{*\top} \mathbf{e}_i\|_2 \\ &\leq \sqrt{\frac{\mu r^*}{(1-\rho)n}} \|(\mathbf{I} - \mathbf{U}_{t+1} \mathbf{U}_{t+1}^\top) \mathbf{L}^*\|_2 \\ &\leq 2\sqrt{\frac{\mu r^* \rho}{1-\rho}} (\epsilon_t + \alpha_t). \end{aligned}$$

Substituting  $\alpha_t$  into this inequality, we have

$$\|(\mathbf{I} - \mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top)\mathbf{L}^*\|_{\infty,2} \leq 2\sqrt{\frac{\mu r^* \rho}{1-\rho}} \left(1 + \frac{\sqrt{r^*}}{\max_{\mathbf{L} \in \mathcal{L}} \sigma_{r^*}(\mathbf{L})}\right) \epsilon_t.$$

Therefore, as long as  $2\sqrt{\frac{\mu r^* \rho}{1-\rho}} \left(1 + \frac{\sqrt{r^*}}{\max_{\mathbf{L} \in \mathcal{L}} \sigma_{r^*}(\mathbf{L})}\right) < 1$  or equivalently

$$\frac{\rho}{1-\rho} < \frac{1}{4\mu r^* (1 + \sqrt{r^*}/\max_{\mathbf{L} \in \mathcal{L}} \sigma_{r^*}(\mathbf{L}))^2},$$

there exists  $\tau < 1$  so that  $2\sqrt{\frac{\mu r^* \rho}{1-\rho}} \left(1 + \frac{\sqrt{r^*}}{\max_{\mathbf{L} \in \mathcal{L}} \sigma_{r^*}(\mathbf{L})}\right) \leq \tau$  and

$$\|(\mathbf{I} - \mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top)\mathbf{L}^*\|_{\infty,2} \leq \tau \epsilon_t = \epsilon_{t+1},$$

which implies that

$$\|(\mathbf{I} - \mathbf{U}_T\mathbf{U}_T^\top)\mathbf{L}^*\|_{\infty,2} \leq \tau^{T-1} \epsilon_1 = \tau^{T-1}.$$

By taking  $T \geq \log \epsilon / \log \tau + 1$ , we have

$$\|(\mathbf{I} - \mathbf{U}_T\mathbf{U}_T^\top)\mathbf{L}^*\|_{\infty,2} \leq \epsilon_T \leq \epsilon.$$

Finally, by applying Lemma 8.2 again, we obtain this theorem.

### Proof of Theorem 8.4

The proof of this theorem is almost the same as the proof of Theorem 8.2.

For  $t = 1$ , since  $\epsilon_1 = 1$  and  $\|\mathbf{X}_i\|_2 \leq 1$  for all  $i$  after the normalization step, the following inequality holds:

$$\|(\mathbf{I} - \mathbf{U}_t\mathbf{U}_t^\top)(\mathbf{L}_i^* + \mathbf{N}_i)\|_2 \leq \|\mathbf{L}_i^* + \mathbf{N}_i\|_2 \leq \epsilon_t.$$

For  $t > 1$ , suppose that  $\|(\mathbf{I} - \mathbf{U}_t\mathbf{U}_t^\top)(\mathbf{L}_i^* + \mathbf{N}_i)\|_2 \leq \epsilon_t$  for  $i \notin \mathcal{C}$ , our goal is to prove that after one iteration

$$\|(\mathbf{I} - \mathbf{U}_{t+1}\mathbf{U}_{t+1}^\top)(\mathbf{L}_i^* + \mathbf{N}_i)\|_2 \leq \epsilon_{t+1}$$

holds for some  $\epsilon_{t+1}$ .

Recall that  $\mathbf{X} = \mathbf{L}^* + \mathbf{S}^* + \mathbf{N}$ , then

$$(\mathbf{X} - \mathbf{U}_t \mathbf{U}_t^\top \mathbf{X})_i = \begin{cases} (\mathbf{I} - \mathbf{U}_t \mathbf{U}_t^\top)(\mathbf{S}_i^* + \mathbf{N}_i), & i \in \mathcal{C}, \\ (\mathbf{I} - \mathbf{U}_t \mathbf{U}_t^\top)(\mathbf{L}_i^* + \mathbf{N}_i), & i \notin \mathcal{C}. \end{cases}$$

Let  $\mathcal{A} = \{i : \|(\mathbf{X} - \mathbf{U}_t \mathbf{U}_t^\top \mathbf{X})_i\|_2 > \epsilon_t\}$ . Since  $\|(\mathbf{I} - \mathbf{U}_t \mathbf{U}_t^\top)(\mathbf{L}_i^* + \mathbf{N}_i)\|_2 \leq \epsilon_t$  for any  $i \notin \mathcal{C}$ , we have

$$\mathcal{A} \subseteq \mathcal{C}, \text{ and } \|(\mathbf{I} - \mathbf{U}_t \mathbf{U}_t^\top)(\mathbf{S}_i^* + \mathbf{N}_i)\|_2 \leq \epsilon_t, \forall i \in \mathcal{A}^c \cap \mathcal{C}.$$

Recall that  $\text{rank}(\mathbf{U}_t) = r$ ,  $\text{rank}(\mathbf{U}^*) = r^*$  and  $r \geq r^*$ . By Lemma 8.2, we can construct  $\hat{\mathbf{U}} \in \mathbb{R}^{p \times (r-r^*)}$  so that  $\tilde{\mathbf{U}} = [\mathbf{U}^*, \hat{\mathbf{U}}]$  is orthogonal and each column of  $\hat{\mathbf{U}}$  lies in the column space of  $\mathbf{U}_t$ , i.e.,  $\mathbf{U}_t \mathbf{U}_t^\top \hat{\mathbf{U}} = \hat{\mathbf{U}}$ , and

$$\|\tilde{\mathbf{U}} \tilde{\mathbf{U}}^\top - \mathbf{U}_t \mathbf{U}_t^\top\|_2 \leq \frac{\sqrt{r^*}}{\max_{\mathbf{L} \in \mathcal{L}} \sigma_{r^*}(\mathbf{L})} (\epsilon_t + \|\mathbf{N}\|_{\infty, 2}). \quad (8.15)$$

We define  $\alpha_t = \frac{\sqrt{r^*}}{\max_{\mathbf{L} \in \mathcal{L}} \sigma_{r^*}(\mathbf{L})} (\epsilon_t + \|\mathbf{N}\|_{\infty, 2})$ , then for any  $i \in \mathcal{A}^c \cap \mathcal{C}$ , we have

$$\begin{aligned} \|(\mathbf{I} - \tilde{\mathbf{U}} \tilde{\mathbf{U}}^\top)(\mathbf{S}_i^* + \mathbf{N}_i)\|_2 &\leq \epsilon_t + \|\tilde{\mathbf{U}} \tilde{\mathbf{U}}^\top - \mathbf{U}_t \mathbf{U}_t^\top\|_2 \\ &\leq \epsilon_t + \alpha_t. \end{aligned} \quad (8.16)$$

where the first inequality follows from Inequality (8.8) and  $\|\mathbf{S}_i^* + \mathbf{N}_i\| \leq 1$ . From Step 3 and Step 5 of Algorithm 8.2,  $\mathbf{L}_{t+1}$  satisfies

$$\mathbf{L}_{t+1, i} = \begin{cases} \mathbf{U}_t \mathbf{U}_t^\top (\mathbf{S}_i^* + \mathbf{N}_i), & i \in \mathcal{A}, \\ \mathbf{L}_i^* + \mathbf{N}_i, & i \in \mathcal{A}^c \cap \mathcal{C}^c, \\ \mathbf{S}_i^* + \mathbf{N}_i, & i \in \mathcal{A}^c \cap \mathcal{C}. \end{cases}$$

Let  $\hat{\mathbf{S}}_i^*$  be a  $p \times n$  matrix whose columns are  $\mathbf{0}$  except that the  $i$ th column equals  $\mathbf{S}_i^*$  and

$\hat{\mathbf{N}}_i$  be a  $p \times n$  matrix whose columns are  $\mathbf{0}$  except that the  $i$ th column equals  $\mathbf{N}_i$ , then

$$\begin{aligned}
\mathbf{L}_{t+1} &= \mathbf{L}^* + \sum_{i \in \mathcal{A}^c \cap \mathcal{C}} \hat{\mathbf{S}}_i^* + \sum_{i \in \mathcal{A}^c} \hat{\mathbf{N}}_i + \mathbf{U}_t \mathbf{U}_t^\top \sum_{i \in \mathcal{A}} (\hat{\mathbf{S}}_i^* + \hat{\mathbf{N}}_i) \\
&= \mathbf{L}^* + \sum_{i \in \mathcal{C}^c} \hat{\mathbf{N}}_i + \tilde{\mathbf{U}} \tilde{\mathbf{U}}^\top \left[ \sum_{i \in \mathcal{A}^c \cap \mathcal{C}} (\hat{\mathbf{S}}_i^* + \hat{\mathbf{N}}_i) + \mathbf{U}_t \mathbf{U}_t^\top \sum_{i \in \mathcal{A}} (\hat{\mathbf{S}}_i^* + \hat{\mathbf{N}}_i) \right] + \\
&\quad (\mathbf{I} - \tilde{\mathbf{U}} \tilde{\mathbf{U}}^\top) \left[ \sum_{i \in \mathcal{A}^c \cap \mathcal{C}} (\hat{\mathbf{S}}_i^* + \hat{\mathbf{N}}_i) + \mathbf{U}_t \mathbf{U}_t^\top \sum_{i \in \mathcal{A}} (\hat{\mathbf{S}}_i^* + \hat{\mathbf{N}}_i) \right] \\
&= \tilde{\mathbf{U}} \mathbf{R} + \sum_{i \in \mathcal{C}^c} \hat{\mathbf{N}}_i + \mathbf{B},
\end{aligned}$$

where

$$\mathbf{R} = \tilde{\mathbf{U}}^\top \mathbf{L}^* + \tilde{\mathbf{U}}^\top \left[ \sum_{i \in \mathcal{A}^c \cap \mathcal{C}} (\hat{\mathbf{S}}_i^* + \hat{\mathbf{N}}_i) + \mathbf{U}_t \mathbf{U}_t^\top \sum_{i \in \mathcal{A}} (\hat{\mathbf{S}}_i^* + \hat{\mathbf{N}}_i) \right],$$

and

$$\mathbf{B} = (\mathbf{I} - \tilde{\mathbf{U}} \tilde{\mathbf{U}}^\top) \left[ \sum_{i \in \mathcal{A}^c \cap \mathcal{C}} (\hat{\mathbf{S}}_i^* + \hat{\mathbf{N}}_i) + \mathbf{U}_t \mathbf{U}_t^\top \sum_{i \in \mathcal{A}} (\hat{\mathbf{S}}_i^* + \hat{\mathbf{N}}_i) \right],$$

then from Inequalities (8.15) and (8.16),

$$\begin{aligned}
\|\mathbf{B}\|_{\infty,2} &\leq \max\{\epsilon_t + \alpha_t, \|(\mathbf{I} - \mathbf{U}^* \mathbf{U}^{*\top}) \mathbf{U}_t \mathbf{U}_t^\top\|_2\} \\
&= \max\{\epsilon_t + \alpha_t, \|\mathbf{U}^* \mathbf{U}^{*\top} - \mathbf{U}_t \mathbf{U}_t^\top\|_2\} \\
&\leq \epsilon_t + \alpha_t.
\end{aligned}$$

From Lemma 8.1 and Inequalities (8.10) and (8.11), we have

$$\begin{aligned}
\|(\mathbf{I} - \mathbf{U}_{t+1} \mathbf{U}_{t+1}^\top) \mathbf{L}^*\|_2 &\leq 2\sqrt{\rho n} \|\mathbf{B}\|_{\infty,2} + 2\|\mathbf{N}\|_2 \\
&\leq 2\sqrt{\rho n} (\epsilon_t + \alpha_t) + 2\|\mathbf{N}\|_2.
\end{aligned}$$

Thus, by the incoherent condition, i.e.,  $\max_i \|\mathbf{e}_i^\top \mathbf{V}^*\|_2^2 \leq \frac{\mu r^*}{(1-\rho)n}$ , for  $i \notin \mathcal{C}$ ,

$$\begin{aligned}
& \|(\mathbf{I} - \mathbf{U}_{t+1} \mathbf{U}_{t+1}^\top)(\mathbf{L}_i^* + \mathbf{N}_i)\|_2 \\
& \leq \|(\mathbf{I} - \mathbf{U}_{t+1} \mathbf{U}_{t+1}^\top) \mathbf{U}^* \boldsymbol{\Sigma}^* \mathbf{V}^{*\top} \mathbf{e}_i\|_2 + \|\mathbf{N}_i\|_2 \\
& \leq \|(\mathbf{I} - \mathbf{U}_{t+1} \mathbf{U}_{t+1}^\top) \mathbf{U}^* \boldsymbol{\Sigma}^*\|_2 \cdot \max_i \|\mathbf{V}^{*\top} \mathbf{e}_i\|_2 + \|\mathbf{N}_i\|_2 \\
& \leq \sqrt{\frac{\mu r^*}{(1-\rho)n}} \|(\mathbf{I} - \mathbf{U}_{t+1} \mathbf{U}_{t+1}^\top) \mathbf{L}^*\|_2 + \|\mathbf{N}_i\|_2 \\
& \leq 2\sqrt{\frac{\mu r^* \rho}{1-\rho}} (\epsilon_t + \alpha_t) + 2\sqrt{\frac{\mu r^*}{(1-\rho)n}} \|\mathbf{N}\|_2 + \|\mathbf{N}\|_{\infty,2} \\
& \leq 2\sqrt{\frac{\mu r^* \rho}{1-\rho}} (\epsilon_t + \alpha_t) + (2\sqrt{\frac{\mu r^*}{1-\rho}} + 1) \|\mathbf{N}\|_{\infty,2}.
\end{aligned} \tag{8.17}$$

Substituting  $\alpha_t = \frac{\sqrt{r^*}}{\max_{\mathbf{L} \in \mathcal{L}} \sigma_{r^*}(\mathbf{L})} (\epsilon_t + \|\mathbf{N}\|_{\infty,2})$  into this inequality, we have for  $i \notin \mathcal{C}$ ,

$$\begin{aligned}
\|(\mathbf{I} - \mathbf{U}_{t+1} \mathbf{U}_{t+1}^\top)(\mathbf{L}_i^* + \mathbf{N}_i)\|_2 & \leq 2\sqrt{\frac{\mu r^* \rho}{(1-\rho)}} \left(1 + \frac{\sqrt{r^*}}{\max_{\mathbf{L} \in \mathcal{L}} \sigma_{r^*}(\mathbf{L})}\right) \epsilon_t + \\
& \quad (2r^* \sqrt{\frac{\mu \rho}{(1-\rho) \max_{\mathbf{L} \in \mathcal{L}} \sigma_{r^*}(\mathbf{L})^2}} + 2\sqrt{\frac{\mu r^*}{1-\rho}} + 1) \|\mathbf{N}\|_{\infty,2}.
\end{aligned}$$

Therefore, when

$$2\sqrt{\frac{\mu r^* \rho}{(1-\rho)}} \left(1 + \frac{\sqrt{r^*}}{\max_{\mathbf{L} \in \mathcal{L}} \sigma_{r^*}(\mathbf{L})}\right) \leq \tau < 1$$

and

$$\eta \geq (2\sqrt{\frac{\mu r^*}{1-\rho}} + 2r^* \sqrt{\frac{\mu \rho}{(1-\rho) \max_{\mathbf{L} \in \mathcal{L}} \sigma_{r^*}(\mathbf{L})^2}} + 1) \|\mathbf{N}\|_{\infty,2},$$

we have

$$\|(\mathbf{I} - \mathbf{U}_{t+1} \mathbf{U}_{t+1}^\top)(\mathbf{L}_i^* + \mathbf{N}_i)\|_2 \leq \tau \epsilon_t + \eta,$$

which implies that

$$\|(\mathbf{I} - \mathbf{U}_T \mathbf{U}_T^\top)(\mathbf{L}_i^* + \mathbf{N}_i)\|_2 \leq \tau^{T-1} + \frac{1 - \tau^{T-1}}{1 - \tau} \eta.$$

Therefore,

$$\|(\mathbf{I} - \mathbf{U}_T \mathbf{U}_T^\top) \mathbf{L}^*\|_{\infty,2} \leq \tau^{T-1} + \frac{\eta}{1 - \tau} + \|\mathbf{N}\|_{\infty,2}.$$

By taking  $T \geq \log \epsilon / \log \tau + 1$ , we have

$$\|(\mathbf{I} - \mathbf{U}_T \mathbf{U}_T^\top) \mathbf{L}^*\|_{\infty, 2} \leq \epsilon + \frac{\eta}{1 - \tau} + \|\mathbf{N}\|_{\infty, 2}.$$

Finally, when  $r = r^*$ , by applying Lemma 8.2 again, we obtain this theorem.

### Proof of Theorem 8.5

Recall that  $r = r^*$  and  $\mathbf{U}_T, \epsilon_T$  are the outputs of Algorithm 8.2. Let  $\tilde{\mathbf{L}} \in \mathbb{R}^{p \times n}$  satisfy that  $\tilde{\mathbf{L}}_i = \mathbf{L}_i^* / \|\mathbf{L}_i^*\|_2$  and  $\mathcal{L}$  be the set of full rank  $p \times r$  submatrices of  $\tilde{\mathbf{L}}$ . From Theorem 8.3, we have that for any  $i \notin \mathcal{C}$ ,

$$\frac{\|(\mathbf{I} - \mathbf{U}_T \mathbf{U}_T^\top) \mathbf{L}_i^*\|_2}{\|\mathbf{L}_i^*\|_2} \leq \epsilon_T \leq \epsilon$$

and

$$\|\mathbf{U}^* \mathbf{U}^{*\top} - \mathbf{U}_T \mathbf{U}_T^\top\|_2 \leq \frac{\sqrt{r} \epsilon_T}{\max_{\mathbf{L} \in \mathcal{L}} \sigma_r(\mathbf{L})}.$$

Note that for  $i \in \mathcal{C}$ ,

$$\begin{aligned} \|(\mathbf{I} - \mathbf{U}_T \mathbf{U}_T^\top) \mathbf{S}_i^*\|_2 &\geq \|(\mathbf{I} - \mathbf{U}^* \mathbf{U}^{*\top}) \mathbf{S}_i^*\|_2 - \|(\mathbf{U}_T \mathbf{U}_T^\top - \mathbf{U}^* \mathbf{U}^{*\top}) \mathbf{S}_i^*\|_2 \\ &\geq \|(\mathbf{I} - \mathbf{U}^* \mathbf{U}^{*\top}) \mathbf{S}_i^*\|_2 - \frac{\sqrt{r} \epsilon_T \|\mathbf{S}_i^*\|_2}{\max_{\mathbf{L} \in \mathcal{L}} \sigma_r(\mathbf{L})} \end{aligned}$$

By the assumption  $\|(\mathbf{I} - \mathbf{U}^* \mathbf{U}^{*\top}) \mathbf{S}_i^*\|_2 / \|\mathbf{S}_i^*\|_2 \geq \delta$ , we have

$$\frac{\|(\mathbf{I} - \mathbf{U}_T \mathbf{U}_T^\top) \mathbf{S}_i^*\|_2}{\|\mathbf{S}_i^*\|_2} \geq \delta - \frac{\sqrt{r} \epsilon_T}{\max_{\mathbf{L} \in \mathcal{L}} \sigma_r(\mathbf{L})}$$

Therefore, as long as

$$\epsilon_T \leq \epsilon < \frac{\delta}{1 + \sqrt{r} / \max_{\mathbf{L} \in \mathcal{L}} \sigma_r(\mathbf{L})},$$

$\mathcal{A}$  is identical to  $\mathcal{C}$ , which implies that  $\mathbf{L} = \mathbf{L}^*$ . Hence we obtain this theorem.

## 8.7 Chapter Summary

In this chapter, we proposed two non-convex outlier-robust PCA algorithms – Outlier Rejection and Outlier Reduction and establish their performance guarantees on the exact recovery of the true column space. We showed that the exact recovery of the true column space can be achieved by our proposed algorithms if the fraction of outliers is  $O(\frac{1}{\mu^{r+2}})$  where  $\mu$  is the column-incoherence parameter. Our proposed algorithms have much lower computational cost than Outlier Pursuit and hence can be applied in large-scale applications, i.e., their overall computational complexity is  $O(rnp \log(1/\epsilon))$  where  $r$  is the target rank and  $\epsilon$  is the estimation error, which is much lower than  $O(\min\{np^2, n^2p\})$  – the computational complexity of Outlier Pursuit. For future work, we aim to develop possible variants of these algorithms so that the required condition for the exact recovery match the assumption for Outlier Pursuit to further bridge the gap between theory and practice.

# CHAPTER 9

## Online PCA with Imperfect Data

We propose a unified paradigm on online principal component analysis via online mirror descent with samples collected sequentially. By designing proper robust gradients in the dual space for mirror descent, we develop and analyze novel robust online PCA algorithms that are able to estimate the true principal components well even in the face of defect data such as samples with missing entries, arbitrary corruption, or limited attribute observation. We establish finite sample performance guarantees for the proposed algorithms, and conduct numerical experiments on synthetic and real-world data to demonstrate that they outperform existing methods in practice.

### 9.1 Introduction

Principal component analysis (PCA) [Pea01] is a fundamental method for dimensionality reduction, applied in a wide range of data analysis applications in machine learning, statistics and bioinformatics, to name a few. Standard PCA extracts the principal components (PCs) from a set of samples by computing the leading eigenvectors of the sample covariance matrix or the leading singular vectors of the sample matrix, which is computationally expensive and memory exhausting when faced with large-scale applications.

To address this issue, various computational-efficient PCA algorithms have been recently developed [WK08, MCJ13, ACLS12, ACS13, YX15a, Bra02, ACS13, Sha15b], most of which focus on an online setting where one receives a sample sequentially and this sample vanishes after it is collected unless it is stored in the memory. These algorithms typically take one of the two approaches: 1) block-wise stochastic power methods. For example, the memory



efficient PCA/sparse PCA algorithms developed by [MCJ13] and [YX15a] perform a power iteration update on the estimated PCs once a block of new samples are received; and 2) stochastic convex optimization. For example, the stochastic PCA algorithm proposed by [ACS13] performs a matrix stochastic gradient descent when a new sample arrives. The main benefit of the latter is that it converges faster than block-wise stochastic power methods [BDF13, Sha15b].

A second weakness of PCA, is that it is notoriously fragile to outliers or missing entries. Many efforts have been made to mitigate this weakness, by proposing robust variants of PCA [XY95, YW99, ITB03, Das03, XCM13, FXY12, FXMY14, YX15b]. Most of these algorithms require either explicitly computing the covariance matrix or storing all the samples, and hence cannot be implemented in an online manner. To our knowledge, the only existing work on online outlier-robust PCA is [FXMY14], which is based on probabilistically admitting each new sample depending on its variance along the current estimated PCs. This algorithm has several limitations: 1) It requires a reasonably good initial solution. 2) It only has asymptotic performance guarantees instead of finite-sample guarantees, and empirically the convergence is slow. 3) It cannot handle missing entries. On the other hand, [MCJ14] extended streaming PCA [MCJ13] to tackle samples with missing entries using an unbiased estimator of the covariance matrix. However, their algorithm cannot handle outliers.

Beyond missing entries or outliers, another interesting setup is that the decision maker can actively choose which entries to measure given certain budget. For example, in medical diagnosis and DNA sequencing, measuring all attributes may be infeasible due to time and cost limits, leading to the following question: can we efficiently estimate the true PCs when we can only choose a fraction of attributes to observe? In the context of linear regression – e.g., Ridge regression, Lasso and Support-vector regression – [CBSS10] and [HK12] first studied this *limited attribute observation* problem, and then [KS15] developed a distribution-dependent sampling scheme for sampling attributes to achieve better performance. For online PCA, however, to the best of our knowledge, this problem has not been explored yet.

In this chapter, we consider a unified paradigm on online PCA via *online mirror descent* – a general framework for developing and analyzing first-order online learning algorithms,

e.g., [SST11, SS12, OCC15]. By designing proper *robust gradients* used in mirror descent, we propose new online PCA algorithms that are robust to various types of data defect such as missing entries, corrupted attributes or outliers; we further develop efficient algorithms for online PCA in the limited attribute observation setting. We establish finite-sample performance guarantees, which is a distinctive feature of the proposed paradigm.

**Notations:** We use boldface lower-case letters to represent vectors and capital letters for matrices. For a matrix  $\mathbf{X}$ ,  $\|\mathbf{X}\|_2$ ,  $\|\mathbf{X}\|_1$  and  $\|\mathbf{X}\|_F$  denote its spectral norm, element-wise  $l_1$ -norm and Frobenius norm. For a vector  $\mathbf{x}$ ,  $\|\mathbf{x}\|_p$  denotes its  $l_p$ -norm. The inner product between two matrices  $\mathbf{X}, \mathbf{Y}$  is defined by  $\langle \mathbf{X}, \mathbf{Y} \rangle = \text{tr}(\mathbf{X}^\top \mathbf{Y})$ . For a symmetric matrix  $\mathbf{X}$ , we use  $\lambda_k(\mathbf{X})$  to denote its  $k$ th largest eigenvalue. We let  $\mathbb{S}^d$  be the set of  $d \times d$  symmetric matrices and use  $\mathbf{e}_1, \dots, \mathbf{e}_d$  to represent the standard basis of  $\mathbb{R}^d$ .

## 9.2 Problem Setting

We consider the streaming data model where one receives sample points  $\mathbf{x}_t \in \mathbb{R}^d$  drawn from an unknown distribution  $\mathcal{D}$  for  $t = 1, \dots, T$  and  $\mathbf{x}_t$  vanishes after it is collected unless it is stored in the memory. Our goal is to extract the leading  $k$  principal components of the received data. The standard PCA extracts principal components by solving the following optimization problem:

$$\begin{aligned} \max \quad & \mathbb{E}_{\mathbf{x} \sim \mathcal{D}}[\langle \mathbf{U}\mathbf{U}^\top, \mathbf{x}\mathbf{x}^\top \rangle] \\ \text{s.t.} \quad & \mathbf{U}^\top \mathbf{U} = \mathbf{I}_k, \mathbf{U} \in \mathbb{R}^{d \times k}, \end{aligned} \tag{9.1}$$

where each column of  $\mathbf{U}$  represents one principal component. A typical way to reformulate (9.1) into a convex optimization formulation is to relax the non-convex constraint  $\mathbf{U}^\top \mathbf{U} = \mathbf{I}_k$  as follows:

$$\begin{aligned} \max \quad & \mathbb{E}_{\mathbf{x} \sim \mathcal{D}}[\langle \mathbf{P}, \mathbf{x}\mathbf{x}^\top \rangle] \\ \text{s.t.} \quad & 0 \preceq \mathbf{P} \preceq \mathbf{I}_p, \text{tr}(\mathbf{P}) = k, \mathbf{P} \in \mathbb{S}^d, \end{aligned} \tag{9.2}$$

where the constraint in (9.2) is called the Fantope constraint. Denote the optimal solutions of Problems (9.1) and (9.2) by  $\mathbf{U}^*$  and  $\mathbf{P}^*$ , respectively. It can be proved that if  $\mathbb{E}_{\mathbf{x} \sim \mathcal{D}}[\mathbf{x}\mathbf{x}^\top]$

is available, then  $\mathbf{P}^* = \mathbf{U}^* \mathbf{U}^{*\top}$ , i.e.,  $\mathbf{P}^*$  is the projection matrix onto the subspace spanned by the leading  $k$  principal components [OW92].

In the online setting where a learner has to choose an estimate  $\mathbf{P}_t$  of  $\mathbf{P}^*$  after receiving a new sample, the theoretical performance of online PCA is usually measured by the regret – the difference between the learner’s total cost and the cost of the optimal strategy:

$$\text{regret}(T) = T \mathbb{E}_{\mathbf{x} \sim \mathcal{D}}[\langle \mathbf{P}^*, \mathbf{x} \mathbf{x}^\top \rangle] - \sum_{t=1}^T \mathbb{E}[\mathbb{E}_{\mathbf{x} \sim \mathcal{D}}[\langle \mathbf{P}_t, \mathbf{x} \mathbf{x}^\top \rangle]],$$

where  $\mathbf{P}_t$  is computed by a certain online PCA algorithm at time  $t$  and the outer expectation in the second term is taken with respect to the randomness in this algorithm. It has been shown that this regret has a lower bound  $\Omega(\sqrt{kT})$  and the lower bound is tight, i.e., there exists an online PCA algorithm so that  $\text{regret}(T) \leq O(\sqrt{kT})$  [NKW13].

Most of previous work assume that the complete information about sample  $\mathbf{x}_t$  is available, namely, all attributes of  $\mathbf{x}_t$  can be observed. However, in practical applications, some attributes of  $\mathbf{x}_t$  may be missing or corrupted because of sensor failure, or only a small subset of the attributes of  $\mathbf{x}_t$  is to be observed due to high measurement costs. In this chapter, we develop and analyze several online PCA algorithms for handling imperfect information.

For theoretical analysis, we make the following assumptions: 1) Samples  $\mathbf{x}_t$  are i.i.d. drawn from an unknown distribution  $\mathcal{D}$  and there exists constant  $B$  such that  $\|\mathbf{x}_t\|_2^2 \leq B$  for every  $\mathbf{x}_t$ . 2) The projection matrix  $\mathbf{P}^*$  onto the subspace spanned by the leading  $k$  eigenvectors of  $\Sigma \triangleq \mathbb{E}_{\mathbf{x} \sim \mathcal{D}}[\mathbf{x} \mathbf{x}^\top]$  satisfies  $\|\mathbf{P}^*\|_0 \leq \gamma^2$ , where  $\|\mathbf{P}^*\|_0$  is the number of nonzero entries of  $\mathbf{P}^*$  and  $\gamma \leq d$  indicates the sparsity of  $\mathbf{P}^*$ . 3) The gap between the  $k$ th and  $k+1$ th eigenvalues of  $\Sigma$ , i.e.,  $\Delta_k \triangleq \lambda_k(\Sigma) - \lambda_{k+1}(\Sigma)$ , is greater than 0. Our goal is to approximately recover  $\mathbf{P}^*$  with the received samples subject to missing or corrupted entries.

We use *subspace distance* to measure the performance of the proposed online PCA algorithms. Subspace distance measures the distance between the subspace spanned by the estimated principal components and the subspace spanned by the true ones. Below is the formal definition:

**Definition 9.1.** Let  $\mathbf{X}, \mathbf{Y}$  be two symmetric matrices and  $\mathcal{X}, \mathcal{Y}$  be their respective  $k$ -

dimensional principal subspaces, then the subspace distance is  $\sin \Theta(\mathcal{X}, \mathcal{Y})$  where  $\Theta(\mathcal{X}, \mathcal{Y})$  is the principal angle between  $\mathcal{X}$  and  $\mathcal{Y}$ .

For any symmetric matrix  $\mathbf{P} \in \mathbb{S}^{d \times d}$ , the following lemma relates the subspace distance between  $\mathbf{P}^*$  and  $\mathbf{P}$  with the Frobenius norm of  $\mathbf{P}^* - \mathbf{P}$ .

**Lemma 9.1.** [VCLR13] *If  $\mathcal{M}^*$  is the principal  $k$ -dimensional subspace of  $\Sigma$  and  $\mathcal{M}$  is the principal  $k$ -dimensional subspace of  $\mathbf{P}$ , then*

$$\sin \Theta(\mathcal{M}, \mathcal{M}^*) \leq \sqrt{2} \|\mathbf{P}^* - \mathbf{P}\|_F.$$

### 9.3 Framework for Online Robust PCA

Online mirror descent (OMD) is a popular framework for online convex optimization [SST11, SS12], which we use to solve the online PCA problem discussed in this chapter. Let  $\mathcal{F} \subseteq \{\mathbf{P} \in \mathbb{S}^d : 0 \preceq \mathbf{P} \preceq \mathbf{I}_d, \text{tr}(\mathbf{P}) = k\}$  be a closed convex set taken as the set of feasible projection matrices and  $f(\cdot)$  be a closed and strongly convex function with domain  $\mathcal{F}$ . Let  $f^*(\cdot)$  be the Fenchel conjugate of  $f(\cdot)$  which is defined by  $f^*(\mathbf{Y}) = \sup_{\mathbf{X} \in \mathcal{F}} \langle \mathbf{X}, \mathbf{Y} \rangle - f(\mathbf{X})$ , and let  $\nabla f^*(\mathbf{Y})$  be the subgradient of  $f^*(\cdot)$  at  $\mathbf{Y}$ . A well-known fact is  $\nabla f^*(\mathbf{Y}) = \arg \max_{\mathbf{X} \in \mathcal{F}} \langle \mathbf{X}, \mathbf{Y} \rangle - f(\mathbf{X})$  when  $f(\cdot)$  is strongly convex [OCC15].

Algorithm 9.1 presents the general framework for online PCA based on OMD. By selecting different  $f(\cdot)$  and  $\mathbf{Z}_t$ , various kinds of online PCA algorithms can be derived. For example, suppose that  $\mathbf{Z}_t = \mathbf{x}_t \mathbf{x}_t^\top$ , when  $f(\mathbf{X}) = \frac{1}{2} \|\mathbf{X}\|_F^2$ , we obtain MSG – a matrix stochastic gradient descent based online PCA algorithm [ACS13], and when  $f(\mathbf{X}) = \sum_{i=1}^d \lambda_i(\mathbf{X}) \log \lambda_i(\mathbf{X})$ , we derive a matrix exponentiated gradient descent based algorithm similar to those developed by [WK08]. The theoretical performance of this framework is given by Theorem 9.1 which holds for any  $\mathbf{Z}_t$ .

**Theorem 9.1.** *Suppose that  $f(\cdot)$  is  $\beta$ -strongly convex with respect to the norm  $\|\cdot\|$ . Let  $\|\cdot\|_*$  be the dual norm of  $\|\cdot\|$ . Then after  $T$  iterations with step size  $\eta$ , we have*

$$\sum_{t=1}^T \langle \mathbf{Z}_t, \mathbf{P}^* - \mathbf{P}_t \rangle \leq \frac{1}{\eta} [f(\mathbf{P}^*) + f^*(\mathbf{0})] + \frac{\eta}{2\beta} \sum_{t=1}^T \|\mathbf{Z}_t\|_*^2, \quad (9.3)$$

---

**Algorithm 9.1:** A framework for robust online PCA
 

---

**Input** : A strongly convex function  $f(\cdot)$ , a closed convex set  $\mathcal{F}$  and step size  $\eta > 0$ .

- 1 Initialize  $\Theta_1 = \mathbf{0}$ ;
- 2 **for**  $t = 1$  to  $T$  **do**
- 3     Compute  $\mathbf{P}_t = \nabla f^*(\Theta_t)$ ;
- 4     Receive  $\mathbf{Z}_t \in \mathbb{S}^d$ ;
- 5     Update  $\Theta_{t+1} = \Theta_t + \eta \mathbf{Z}_t$ ;
- 6 **end**
- 7 Return  $\bar{\mathbf{P}} = \frac{1}{T} \sum_{t=1}^T \mathbf{P}_t$ .

---

where  $\mathbf{P}^*$  is the optimal solution of (9.2).

Theorem 9.1 implies the following corollary showing that the regret bounds corresponding to various strong convex functions with respect to the  $p$ -Schatten norm have similar relationship with  $T$  and  $\mathbf{Z}_t$ .

**Corollary 9.1.** *Suppose that  $f(\mathbf{X})$  is  $\beta$ -strongly convex with respect to the  $p$ -Schatten norm  $\|\mathbf{X}\|_{S(p)} = \|\sigma(\mathbf{X})\|_p$  for  $p \geq 1$  where  $\sigma(\mathbf{X})$  is a vector containing the singular values of  $\mathbf{X}$  in descending order, and  $f^*(\mathbf{0}) = 0$ . If  $\mathbf{Z}_t$  can be decomposed into  $\mathbf{Z}_t = \frac{1}{2}(\hat{\mathbf{x}}_t \tilde{\mathbf{x}}_t^\top + \tilde{\mathbf{x}}_t \hat{\mathbf{x}}_t^\top)$  for some vectors  $\hat{\mathbf{x}}_t$  and  $\tilde{\mathbf{x}}_t$ , then after  $T$  iterations with step size  $\eta = \sqrt{\frac{2\beta f(\mathbf{P}^*)}{\sum_{t=1}^T \mathbb{E}[\|\hat{\mathbf{x}}_t\|_2^2 \|\tilde{\mathbf{x}}_t\|_2^2]}}$ , we have*

$$\mathbb{E}\left[\sum_{t=1}^T \langle \mathbf{Z}_t, \mathbf{P}^* - \mathbf{P}_t \rangle\right] \leq \sqrt{\frac{2f(\mathbf{P}^*) \sum_{t=1}^T \mathbb{E}[\|\hat{\mathbf{x}}_t\|_2^2 \|\tilde{\mathbf{x}}_t\|_2^2]}{\beta}},$$

where  $\mathbf{P}^*$  is the optimal solution of (9.2) and the expectation is taken with respect to the randomness in Algorithm 9.1.

Recall that  $f(\mathbf{X}) = \frac{1}{2}\|\mathbf{X}\|_F^2$  is  $1/2$ -strongly convex w.r.t. the Frobenius norm  $\|\mathbf{X}\|_F$ . Therefore, when  $\mathbf{Z}_t = \mathbf{x}_t \mathbf{x}_t^\top$  and  $\mathbb{E}_{\mathbf{x}_t \sim \mathcal{D}}[\|\mathbf{x}_t\|_2^4] \leq c$  for constant  $c$ , we have  $\text{regret}(T) = O(\sqrt{kT})$ . This bound holds when all attributes of samples  $\mathbf{x}_t$  are revealed and cannot be further improved. In this case, [ACS13] proposed an efficient implementation of this algorithm whose

memory- and computational-complexity are much lower than  $O(d^2)$ .

In this chapter, besides lower computational cost, we investigate another advantage of Algorithm 9.1, i.e., by selecting “robust”  $\mathbf{Z}_t$  we can develop online PCA algorithms that are able to tackle the case where the received samples have imperfect information. The main idea is to design proper  $\mathbf{Z}_t$  so that its expectation is closed to  $\Sigma$  when some attributes of the samples are corrupted or unobserved. With these  $\mathbf{Z}_t$ , the left hand side of (9.3) is approximately  $T\mathbb{E}[\langle \Sigma, \mathbf{P}^* - \bar{\mathbf{P}} \rangle]$  which can be used to bound  $\|\mathbf{P}^* - \bar{\mathbf{P}}\|_F$ .

### 9.3.1 Missing Entries

We first consider the case where the received samples contain many missing entries. Assume that the true sample  $\bar{\mathbf{x}}_t$  is drawn from distribution  $\mathcal{D}$  and the received sample  $\mathbf{x}_t$  is generated by erasing some entries of  $\bar{\mathbf{x}}_t$  with a certain probability, namely,  $\mathbf{x}_t(i) = \bar{\mathbf{x}}_t(i)$  with probability  $\delta$  or 0 otherwise for  $i = 1, \dots, d$ . Given  $\mathbf{x}_t$ , we construct  $\mathbf{Z}_t$  as follows:

$$\mathbf{Z}_t(q) = \frac{1}{q^2} \mathbf{x}_t \mathbf{x}_t^\top - \frac{1-q}{q^2} \text{diag}(\mathbf{x}_t \mathbf{x}_t^\top) \quad (9.4)$$

where parameter  $q \in [0, 1]$ . Note that  $\mathbf{Z}_t(\delta)$  is an unbiased estimator of  $\Sigma$ . Theorem 9.2 shows the performance guarantee of Algorithm 9.1 with  $\mathbf{Z}_t = \mathbf{Z}_t(q)$  under this setting.

**Theorem 9.2.** *Suppose that  $f(\mathbf{X}) = \frac{1}{2} \|\mathbf{X}\|_F^2$  and  $\mathbf{Z}_t$  is computed according to (9.4) with  $q \in [0, 1]$ , then after  $T$  iterations with step size  $\eta = \sqrt{\frac{k\beta q^2}{\max\{1, \delta/q^2\} \cdot 2\delta B^2 T}}$ , the output  $\bar{\mathbf{P}}$  of Algorithm 9.1 satisfies*

$$\mathbb{E}[\|\mathbf{P}^* - \bar{\mathbf{P}}\|_F^2] \leq \frac{2\sqrt{k}B}{\Delta_k} \left( \sqrt{\max\left\{1, \frac{\delta}{q^2}\right\} \frac{2\delta}{q^2 T}} + \frac{6|\delta - q|}{q^2 \delta^2} \right),$$

where the expectation is taken with respect to the randomness in the samples.

Obviously, Theorem 9.2 implies that this algorithm is guaranteed to converge as long as  $|\delta - q|$  decreases with  $T$ . In particular, when  $q = \delta$ , the expectation of the estimation error is  $O(\frac{1}{\delta} \sqrt{\frac{k}{T}})$ . In practice, since the exact value of probability  $\delta$  is usually unknown, we propose to let  $q$  be an empirical estimate of  $\delta$  as discussed in Section 9.4.

**Algorithm 9.2:**  $\mathbf{Z}_t$  with entry-wise corruption

**Input** : Block size  $m$  and parameter  $\rho$ .

- 1 Receive  $m$  samples  $\mathbf{x}_1, \dots, \mathbf{x}_m$  that form sample matrix  $\mathbf{X} \in \mathbb{R}^{d \times m}$  whose  $i$ th column is  $\mathbf{x}_i$ ;
- 2 Select the smallest  $m - \rho$  entries of  $|\mathbf{X}_{(i)}|$  for  $i = 1, \dots, d$ , where  $\mathbf{X}_{(i)}$  is the  $i$ th row of  $\mathbf{X}$ . Denote the selected indices of  $\mathbf{X}_{(i)}$  by  $\mathcal{I}_i$ ;
- 3 Construct  $\hat{\mathbf{X}}$  so that for the  $i$ th row  $\hat{\mathbf{X}}_{(i)}$  we have  $\hat{\mathbf{X}}_{(i)}(j) = \mathbf{X}_{(i)}(j)$  if  $j \in \mathcal{I}_i$  or 0 otherwise;
- 4 Return  $\hat{\mathbf{X}}\hat{\mathbf{X}}^\top/m$ .

**9.3.2 Corrupted Entries**

In the previous section, the observed entries of the received samples are not corrupted by noises or malicious attackers. In practical applications, one may face with contaminated samples due to sensor failures, malicious attacking or other reasons. In this section, we consider two corruption schemes – (a) “entry-wise” corruption: each entry of the received samples is corrupted with probability  $q$ , and (b) “sample-wise” corruption: the corrupted samples are uniformly distributed among the receive samples, i.e., with probability  $q$  the received sample at time  $t$  is corrupted. Note that the corruption can be *arbitrary*. We develop Algorithms 9.2 and 9.3 for constructing  $\mathbf{Z}_t$  to make Algorithm 9.1 robust to these two types of corruption. Algorithms 9.2 and 9.3 accept a block of samples and truncate the entries (*resp.* samples) with the largest  $\rho$  absolute-values (*resp.*  $\ell_2$ -norms) to zero. This truncation operation mitigates the impact of contamination.

Recall that the number of nonzero entries of  $\mathbf{P}^*$  is  $\gamma^2$  ( $\gamma^2$  measures the sparsity of  $\mathbf{P}^*$ ). Since  $\|\mathbf{P}^*\|_1 \leq \gamma\|\mathbf{P}^*\|_F \leq \gamma\sqrt{k}$ , one can add an additional constraint  $\|\mathbf{P}\|_1 \leq \gamma\sqrt{k}$  into  $\mathcal{F}$  to encourage Algorithm 9.1 to find sparse solutions. Then when  $\mathbf{Z}_t$  used in Algorithm 9.1 is constructed by Algorithm 9.2, we have the following theorem.

**Theorem 9.3.** *Suppose that the corrupted samples are generated according to the “entry-wise corruption” scheme and  $f(\mathbf{X}) = \frac{1}{2}\|\mathbf{X}\|_F^2$ . For any constant  $\delta > 0$ , if  $\mathbf{Z}_t$  is constructed*

**Algorithm 9.3:**  $\mathbf{Z}_t$  with sample-wise corruption

**Input** : Block size  $m$  and parameter  $\rho$ .

- 1 Receive  $m$  samples  $\mathbf{x}_1, \dots, \mathbf{x}_m$  and compute  $v_i = \|\mathbf{x}_i\|_2$  for  $i = 1, \dots, m$ ;
- 2 Select the smallest  $m - \rho$  elements of  $\{v_i\}$ . Let  $\mathcal{I}$  be the selected indices;
- 3 Construct  $\hat{\mathbf{X}}$  so that  $\hat{\mathbf{X}}_i = \mathbf{x}_i$  if  $i \in \mathcal{I}$  or  $\mathbf{0}$  otherwise;
- 4 Return  $\hat{\mathbf{X}}\hat{\mathbf{X}}^\top/m$ .

by Algorithm 9.2 with  $m \geq \frac{2(2+\delta)\log(dT)}{\delta^2q}$  and  $\rho \geq (1 + \delta)mq$ , then with probability at least  $1 - \frac{1}{dT}$  the output  $\bar{\mathbf{P}}$  of Algorithm 9.1 with these  $\mathbf{Z}_t$  satisfies

$$\mathbb{E}[\|\mathbf{P}^* - \bar{\mathbf{P}}\|_F^2] \leq \frac{2B}{\Delta_k} \left( \sqrt{\frac{2k}{T}} + \frac{12\rho\gamma\sqrt{k}}{m} \right) \quad (9.5)$$

where the expectation is taken w.r.t. the randomness in the samples.

Note that when  $\rho = (1 + \delta)mq$ , the right hand side of (9.5) becomes  $O(\sqrt{\frac{k}{T}}) + \frac{24(1+\delta)B\gamma\sqrt{k}q}{\Delta_k}$ . Therefore, for any  $\epsilon > 0$ , if  $q \leq \frac{\epsilon\Delta_k}{24(1+\delta)B\gamma\sqrt{k}\epsilon}$ , the estimation error is  $O(\sqrt{\frac{k}{T}}) + \epsilon$  regardless of the kind of corruption. The next theorem provides the performance guarantee for Algorithm 9.3 in the ‘‘sample-wise’’ corruption case.

**Theorem 9.4.** *Suppose that the corrupted samples are generated according to the ‘‘sample-wise corruption’’ scheme and  $f(\mathbf{X}) = \frac{1}{2}\|\mathbf{X}\|_F^2$ . For any constant  $\delta > 0$ , if  $\mathbf{Z}_t$  is constructed by Algorithm 9.3 with  $m \geq \frac{2(2+\delta)\log(T)}{\delta^2q}$  and  $\rho \geq (1 + \delta)mq$ , then with probability at least  $1 - \frac{1}{T}$  the output  $\bar{\mathbf{P}}$  of Algorithm 9.1 with these  $\mathbf{Z}_t$  satisfies*

$$\mathbb{E}[\|\mathbf{P}^* - \bar{\mathbf{P}}\|_F^2] \leq \frac{2B}{\Delta_k} \left( \sqrt{\frac{2k}{T}} + \frac{6\rho\sqrt{k}}{m} \right) \quad (9.6)$$

where the expectation is taken w.r.t. the randomness in the samples.

For any  $\epsilon > 0$ , when  $\rho = (1 + \delta)mq$  and  $q \leq \frac{\epsilon\Delta_k}{12B\sqrt{k}}$ , the estimation error is bounded by  $O(\sqrt{\frac{k}{T}}) + \epsilon$ . Note that the upper bound for  $q$  does not depend on  $d$  or  $\gamma$ , which means Algorithm 9.3 is robust to a constant fraction of outliers no matter how large dimension  $d$



is.

### 9.3.3 Limited Observation

In Section 9.3.1, the observed attributes of the received samples are determined by the data model that the learner cannot interfere with. In other cases, e.g., medical diagnosis, the learner is able to select which attributes to observe. This problem is an example of “learning with limited attribute observation” [BDD98]. More formally, at time  $t$ , sample  $\mathbf{x}_t$  is drawn from distribution  $\mathcal{D}$  and we are able to choose  $s$  attributes (or entries) of  $\mathbf{x}_t$  we wish to observe. The main questions are 1) which attributes do we choose to reveal? and 2) what convergence rate for Algorithm 9.1 can we achieve given these partially observed samples?

The key to solve this problem is to construct an unbiased estimator of  $\Sigma$  as  $\mathbf{Z}_t$ . One possible approach is shown in Algorithm 9.4 (without loss of generality, we assume that  $s$  is an even number), which samples  $s$  attributes of  $\mathbf{x}_t$  under a certain distribution, constructs two independent unbiased estimators of  $\mathbf{x}_t$ , e.g.,  $\hat{\mathbf{x}}_t$  and  $\tilde{\mathbf{x}}_t$ , and then returns  $\frac{1}{2}(\hat{\mathbf{x}}_t\tilde{\mathbf{x}}_t^\top + \tilde{\mathbf{x}}_t\hat{\mathbf{x}}_t^\top)$  which is an unbiased estimator of  $\mathbf{x}_t\mathbf{x}_t^\top$  conditioned on  $\mathbf{x}_t$ .

When applying Algorithm 9.4 to construct  $\mathbf{Z}_t$ , we need to choose a proper probability vector  $\mathbf{q}$  whose  $i$ th entry  $q_i$  indicates the probability that  $\mathbf{x}_t(i)$  is observed. Different  $\mathbf{q}$  can lead to quite different performance. For example, if the attributes of  $\mathbf{x}_t$  are sampled uniformly at random, i.e.,  $\mathbf{q} = (\frac{1}{d}, \dots, \frac{1}{d})$ , we have the following performance guarantee.

**Theorem 9.5.** *Suppose that  $f(\mathbf{X})$  is  $\beta$ -strongly convex with respect to the  $p$ -Schatten norm  $\|\sigma(\mathbf{X})\|_p$  for  $p \geq 1$  and  $f^*(\mathbf{0}) = 0$ . If  $\mathbf{Z}_t$  is constructed by Algorithm 9.4 with  $\mathbf{q} = (\frac{1}{d}, \dots, \frac{1}{d})$ , the output  $\bar{\mathbf{P}}$  of Algorithm 9.1 after  $T$  iterations with step size  $\eta = \frac{s}{3dB} \sqrt{\frac{2\beta f(\mathbf{P}^*)}{T}}$  satisfies*

$$\mathbb{E}[\|\mathbf{P}^* - \bar{\mathbf{P}}\|_F^2] \leq \frac{6dB}{\Delta_k s} \sqrt{\frac{2f(\mathbf{P}^*)}{\beta T}}.$$

Obviously, when  $f(\mathbf{X}) = \frac{1}{2}\|\mathbf{X}\|_F^2$ , the right hand side becomes  $O(\frac{d}{\Delta_k s} \sqrt{\frac{k}{T}})$ . Note that this bound involves the dimension  $d$ , which means the number of samples required to achieve an  $\epsilon$ -optimal solution could be undesirably large in the high dimensional regime, as  $T =$

**Algorithm 9.4:**  $\mathbf{Z}_t$  under limited attribute observation

---

**Input** : Number of observed attributes  $s$  and probability vector  $\mathbf{q} \in \mathbb{R}^d$ .

- 1 Initialize  $\hat{\mathbf{x}}_t = \mathbf{0}$  and  $\tilde{\mathbf{x}}_t = \mathbf{0}$ ;
- 2 **for**  $r = 1$  to  $s$  **do**
- 3     Choose  $i_{t,r} \in [d]$  with probability  $q_{i_{t,r}}$  and observe  $\mathbf{x}_t(i_{t,r})$  – the  $i_{t,r}$ th entry of  $\mathbf{x}_t$ ;
- 4     **if**  $r \leq s/2$  **then**
- 5         Let  $\hat{\mathbf{x}}_t = \hat{\mathbf{x}}_t + \frac{2}{sq_{i_{t,r}}} \mathbf{x}_t(i_{t,r}) \mathbf{e}_{i_{t,r}}$ ;
- 6     **else**
- 7         Let  $\tilde{\mathbf{x}}_t = \tilde{\mathbf{x}}_t + \frac{2}{sq_{i_{t,r}}} \mathbf{x}_t(i_{t,r}) \mathbf{e}_{i_{t,r}}$ ;
- 8     **end**
- 9 **end**
- 10 Return  $\frac{1}{2}(\hat{\mathbf{x}}_t \tilde{\mathbf{x}}_t^\top + \tilde{\mathbf{x}}_t \hat{\mathbf{x}}_t^\top)$ .

---

$$\Omega\left(\frac{kd^2}{\Delta_k^2 s^2 \epsilon^2}\right).$$

To mitigate dependence on  $d$ , we propose to sample the attributes in a distribution-dependent manner, i.e., the attributes with larger second moments  $\mathbb{E}_{\mathbf{x}_t \sim \mathcal{D}}[\mathbf{x}_t(i)^2]$  are sampled with relatively higher probabilities. The following theorem shows how such a strategy leads to better theoretical performance.

**Theorem 9.6.** *Suppose that  $f(\mathbf{X})$  is  $\beta$ -strongly convex with respect to the  $p$ -Schatten norm  $\|\sigma(\mathbf{X})\|_p$  for  $p \geq 1$  and  $f^*(\mathbf{0}) = 0$ . If  $\mathbf{Z}_t$  is constructed by Algorithm 9.4 with*

$$q_i = \frac{\mathbb{E}_{\mathbf{x}_t \sim \mathcal{D}}[\mathbf{x}_t(i)^2]^{\frac{1}{3}}}{\sum_{i=1}^d \mathbb{E}_{\mathbf{x}_t \sim \mathcal{D}}[\mathbf{x}_t(i)^2]^{\frac{1}{3}}}, \quad i = 1, \dots, d, \quad (9.7)$$

the output  $\bar{\mathbf{P}}$  of Algorithm 9.1 after  $T$  iterations with step size

$$\eta = \sqrt{\frac{\beta f(\mathbf{P}^*)}{\left[\frac{4B}{s^2} \left(\sum_{i=1}^d \mathbb{E}_{\mathbf{x}_t \sim \mathcal{D}}[\mathbf{x}_t(i)^2]^{\frac{1}{3}}\right)^3 + B^2\right] T}}$$

satisfies that

$$\begin{aligned} & \mathbb{E}[\|\mathbf{P}^* - \bar{\mathbf{P}}\|_F^2] \\ & \leq \frac{4}{\Delta_k} \sqrt{\frac{f(\mathbf{P}^*) \left[ \frac{4B}{s^2} \left( \sum_{i=1}^d \mathbb{E}_{\mathbf{x}_t \sim \mathcal{D}}[\mathbf{x}_t(i)^2]^{\frac{1}{3}} \right)^3 + B^2 \right]}{\beta T}}. \end{aligned}$$

By the power mean inequality, it can be easily verified that  $\left( \sum_{i=1}^d \mathbb{E}_{\mathbf{x}_t \sim \mathcal{D}}[\mathbf{x}_t(i)^2]^{\frac{1}{3}} \right)^3 \leq d^2 B$ , which means this bound is always as good as the one shown in Theorem 9.5. In the case where the second moments decay very fast or the samples are sparse, this bound can be much smaller than  $d^2 B$ . When the second moment of  $\mathbf{x}_t(i)$  is unknown, we can use an empirical estimate of  $\mathbb{E}_{\mathbf{x}_t \sim \mathcal{D}}[\mathbf{x}_t(i)^2]$  to compute  $\mathbf{q}$  as discussed in the next section.

## 9.4 Unknown Parameters

In previous sections, we discuss online PCA algorithms for handling missing entries and limited attribute observation, where some prior information about  $\delta$  – the probability that an attribute is observed – or  $\mathbb{E}[\mathbf{x}_t(i)^2]$  – the second moments of each attribute – are required, which may not be available in practical applications. To address this issue, we propose to estimate  $\delta$  or  $\mathbb{E}[\mathbf{x}_t(i)^2]$  from a set of training samples before running the specific online PCA algorithms. Given  $m_0$  training samples,  $\delta$  and  $\mathbb{E}[\mathbf{x}_t(i)^2]$  can be easily estimated by Algorithm 9.5 and Algorithm 9.6, respectively.

With the estimates of  $\delta$  and  $\mathbb{E}[\mathbf{x}_t(i)^2]$  computed by Algorithms 9.5 and 9.6, we have the following variants of Theorem 9.2 and Theorem 9.6.

**Theorem 9.7.** *Suppose that  $f(\mathbf{X}) = \frac{1}{2} \|\mathbf{X}\|_F^2$  and  $\mathbf{Z}_t$  is constructed by (9.4) with parameter  $q$  computed by Algorithm 9.5 with  $m_0 \geq \frac{12 \log d}{\delta d} T^{2\alpha}$  for constant  $\alpha \in [0, \frac{1}{2}]$ , then after  $T$  iterations with step size  $\eta = \sqrt{\frac{kq^2}{8B^2T}}$ , with probability at least  $1 - \frac{2}{d}$ , the output  $\bar{\mathbf{P}}$  of Algorithm 9.1 satisfies*

$$\mathbb{E}[\|\mathbf{P}^* - \bar{\mathbf{P}}\|_F^2] \leq \frac{8B\sqrt{k}}{\Delta_k \delta} \left( \sqrt{\frac{2}{T}} + \frac{9}{\delta^2} \cdot T^{-\alpha} \right),$$

where the expectation is taken with respect to the randomness in the samples.

Theorem 9.7 implies that the upper bound of  $\mathbb{E}[\|\mathbf{P}^* - \bar{\mathbf{P}}\|_F^2]$  decreases with  $T$  as long as  $m_0$

---

**Algorithm 9.5:** Estimate “missing” probability  $\delta$

---

**Input** : Block size  $m_0$ .

- 1 Initialize  $s = 0$ ;
- 2 **for**  $t = 1$  to  $m_0$  **do**
- 3     Receive sample  $\mathbf{x}_t$  and count the number of the observed entries in  $\mathbf{x}_t$  (denoted by  $n_t$ );
- 4     Set  $s = s + \frac{n_t}{d}$ ;
- 5 **end**
- 6 Return  $s/m_0$ .

---

– the number of samples for estimating  $\delta$  – is  $\Omega(\frac{\log d}{d}T^{2\alpha})$  where  $\frac{1}{2} \geq \alpha > 0$ . Therefore, a tradeoff can be made between  $m_0$  and this upper bound, namely, using more (less) samples to estimate  $\delta$  leads to a smaller (larger) bound.

**Theorem 9.8.** *Suppose that  $f(\mathbf{X})$  is  $\beta$ -strongly convex with respect to the  $p$ -Schatten norm  $\|\sigma(\mathbf{X})\|_p$  for  $p \geq 1$  and  $f^*(\mathbf{0}) = 0$ . Let  $\psi_1, \dots, \psi_d$  be the outputs of Algorithm 9.6 with  $m_0 \geq \frac{78d \log d}{s}$ . If  $\mathbf{Z}_t$  is constructed by Algorithm 9.4 with*

$$q_i = \psi_i^{\frac{1}{3}} / \sum_{i=1}^d \psi_i^{\frac{1}{3}}, \quad i = 1, \dots, d,$$

then after  $T$  iterations, the output  $\bar{\mathbf{P}}$  of Algorithm 9.1 with step size

$$\eta = \sqrt{\frac{\beta f(\mathbf{P}^*)}{\left[ \frac{8B}{s^2} (\sum_{i=1}^d \psi_i^{\frac{1}{3}})^3 + B^2 \right] T}}$$

satisfies

$$\mathbb{E}[\|\mathbf{P}^* - \bar{\mathbf{P}}\|_F^2] \leq \frac{4}{\Delta_k} \sqrt{\frac{f(\mathbf{P}^*) \left[ \frac{12B}{s^2} (\sum_{i=1}^d (\mathbb{E}[\mathbf{x}_t(i)^2] + \epsilon)^{\frac{1}{3}})^3 + B^2 \right]}{\beta T}},$$

where  $\epsilon = \frac{8Bd \log d}{3m_0 s}$ .

The upper bound of  $\mathbb{E}[\|\mathbf{P}^* - \bar{\mathbf{P}}\|_F^2]$  shown in Theorem 9.8 is slightly larger than the bound

---

**Algorithm 9.6:** Estimate the second moment of  $\mathbf{x}_t(i)$

---

**Input** : Block size  $m_0$  and number of observed entries  $s$ .

- 1 Initialize  $\mathbf{y} = \mathbf{0}$  and  $\mathbf{z} = \mathbf{0}$ ;
- 2 **for**  $t = 1$  to  $m_0$  **do**
- 3     **for**  $r = 1$  to  $s$  **do**
- 4         Choose  $i_{t,r} \in [d]$  uniformly at random and observe  $\mathbf{x}_t(i_{t,r})$ ;
- 5         Set  $\mathbf{y}(i_{t,r}) = \mathbf{y}(i_{t,r}) + \mathbf{x}_t(i_{t,r})^2$  and  $\mathbf{z}(i_{t,r}) = \mathbf{z}(i_{t,r}) + 1$ ;
- 6     **end**
- 7 **end**
- 8 Set  $\xi(i) = \mathbf{y}(i)/\mathbf{z}(i)$  for  $i \in [d]$ ;
- 9 Return  $\xi(i) + \frac{2Bd \log d}{m_0 s}$  for  $i \in [d]$ .

---

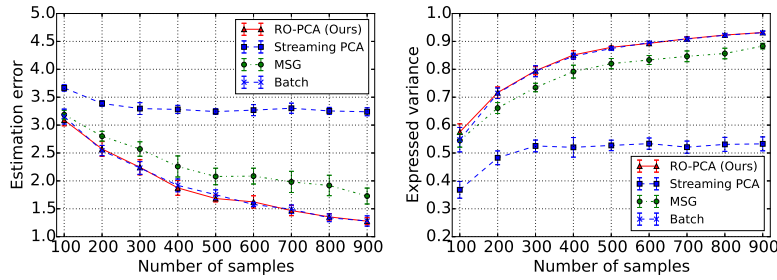
in Theorem 9.6. The gap between these two bounds exists due to the error of estimating the second moment  $\mathbb{E}[\mathbf{x}_t(i)^2]$ .

## 9.5 Experiments

We now investigate the empirical performance of the proposed algorithms on a variety of synthetic and real-world datasets. The experiments are conducted on a desktop PC with an i7 3.4GHz CPU and 8G memory.

### 9.5.1 Synthetic Data

The “inlier” samples with no missing or corrupted entries are generated according to the spike model, i.e., samples  $\mathbf{x}_t$  satisfy  $\mathbf{x}_t = \mathbf{A}\mathbf{z}_t + \boldsymbol{\epsilon}_t$  where matrix  $\mathbf{A} \in \mathbb{R}^{d \times k}$  is fixed,  $\mathbf{z}_t \in \mathbb{R}^k$  are independently sampled from the standard Gaussian distribution  $\mathcal{N}(\mathbf{0}, \mathbf{I}_k)$  and  $\boldsymbol{\epsilon}_t$  are the noises that are independent realizations of Gaussian distribution  $\mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_d)$  for some  $\sigma > 0$ . Matrix  $\mathbf{A}$  is generated by the following three steps: 1) Randomly generate sparse orthogonal matrices  $\mathbf{U} \in \mathbb{R}^{d \times k}$  and  $\mathbf{V} \in \mathbb{R}^{k \times k}$ ; 2) generate a diagonal matrix  $\mathbf{S}$  whose diagonal entries



**Figure 9.1:** Comparison between RO-PCA, MSG, Streaming PCA and Batch. The samples are generated under  $\delta = 0.2$ ,  $d = 100$ ,  $k = 10$  and  $\sigma = 0.05$ .

are drawn from the uniform distribution over  $[1, 2]$ ; 3) finally, set  $\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^\top$ . The empirical performance is measured by two quantities – *estimation error* and *expressed variance* [XCM13]. Let  $\mathbf{P}$  be the output of a certain algorithm. The estimation error is defined by  $\|\mathbf{P} - \mathbf{P}^*\|_F$  and the expressed variance is defined by  $\sum_{i=1}^k \mathbf{u}_i^\top \mathbf{A}\mathbf{A}^\top \mathbf{u}_i / \text{tr}(\mathbf{A}\mathbf{A}^\top)$  where  $\mathbf{u}_i$  are the eigenvectors of  $\mathbf{P}$ . In the following experiments, we choose function  $f(\mathbf{X}) = \frac{1}{2}\|\mathbf{X}\|_F^2$ . We repeat each test 10 times and report the average results.

In the first experiment, we test the performance of the algorithm discussed in Section 9.3.1 when the observed samples contain missing entries. We compare our algorithm (RO-PCA) with several existing algorithms: 1) Matrix stochastic gradient (MSG) [ACS13]. Actually, it can be easily verified that MSG is equivalent to Algorithm 9.1 with  $\mathbf{Z}_t = \mathbf{x}_t \mathbf{x}_t^\top$ . When the observed sample  $\mathbf{x}_t$  contains missing entries, we set these entries to zero and then perform the MSG update. 2) Streaming PCA with missing entries [MCJ14]. This algorithm is an extension of the memory efficient streaming PCA algorithm [MCJ13] based on the block-wise stochastic power method. Different from streaming PCA, it estimates the sample covariance matrix in each block by applying (9.4) that takes the “missing” probability into account. 3) The batch PCA. This is taken as the baseline for our experiments. It stores all the samples received from time  $t = 1$  to  $T$ , then use (9.4) to estimate the sample covariance matrix and finally performs the standard PCA to extract PCs. In the following experiment, we set  $q$  in (9.4) to  $\delta$  and the block size of streaming PCA to 100.

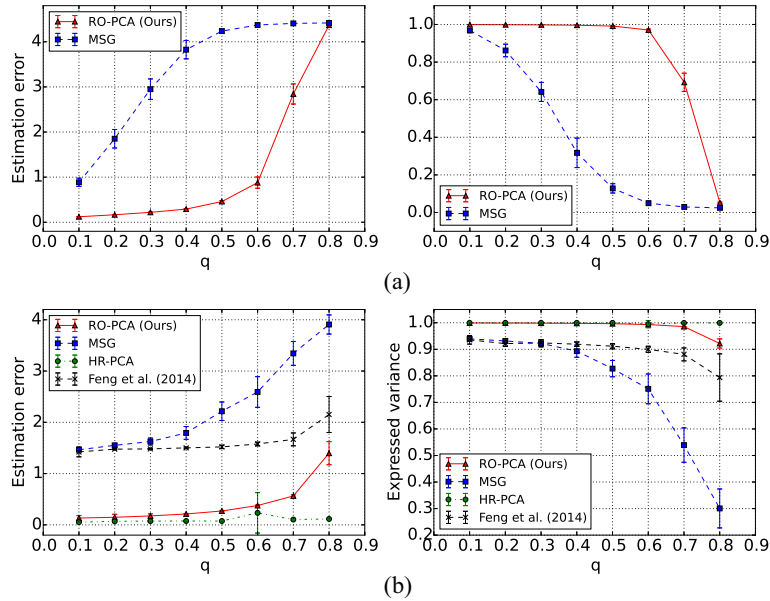
Figure 9.1 presents the results of these four algorithms when  $\delta = 0.2$ ,  $d = 100$ ,  $k = 10$  and  $T$  varies from 100 to 900. Because streaming PCA requires a good estimate of the sample

covariance matrix in each block, it needs a large number of samples ( $\sim 10000$  samples) to obtain acceptable results. When relatively fewer samples are received, its performance is much worse than the other algorithms. We observe our algorithm RO-PCA achieves comparable performance as the baseline algorithm, and clearly outperforms MSG.

The second experiment investigates the performance of Algorithms 9.2 and 9.3 when there exist corrupted samples. In the “entry-wise corruption” case, the corrupted samples are generated as follows. Sample  $\mathbf{x}_t$  is first generated according to the spike model discussed above and then each entry of  $\mathbf{x}_t$  is set to  $\xi$  with probability  $q$  where  $\xi$  is sampled from  $[-5, 5]$  uniformly at random. In the “sample-wise corruption” case, we assume that with probability  $q$  the received sample is replaced by a random vector drawn from  $[-5, 5]^d$  uniformly at random. We compare our algorithms (RO-PCA) with MSG, online robust PCA [FXMY14] and HR-PCA [XCM13]. HR-PCA is an offline outlier-robust PCA algorithm and is taken as the baseline.

Figure 9.2(a) shows the performance of RO-PCA and MSG when the received samples are entry-wise corrupted. Since HR-PCA and online robust PCA [FXMY14] can only handle sample-wise corruption, they are not considered in this case. Obviously, MSG easily breaks down, even when there exists only a small fraction of corruptions. RO-PCA is much more robust than MSG, which can generate meaningful results when  $q$  is less than 0.5. Figure 9.2(b) provides the comparison between RO-PCA, MSG, online robust PCA and HR-PCA in the sample-wise corruption case. Clearly, as  $q$  increases, the performance of MSG dramatically decreases, while RO-PCA is much more stable, which performs similarly to HR-PCA when  $q$  is less than 0.5 and consistently outperforms online robust PCA proposed by [FXMY14].

In the third experiment, we investigate the performance of Algorithm 9.4 when a limited number of attributes of received sample  $\mathbf{x}_t$  can be revealed. We assume that only 40% of the attributes of  $\mathbf{x}_t$  can be observed and  $\mathbf{x}_t$  is generated according to the spike model discussed above where  $d = 500$ ,  $k = 10$ ,  $\sigma = 0.1$  and  $\mathbf{A}$  is row-sparse, e.g., the number of nonzero rows of  $\mathbf{A}$  is  $0.3d$ . We compare three different attribute sampling schemes for Algorithm 9.4: 1) Uniform sampling. The attributes of each sample are sampled uniformly at random; 2) Distribution-dependent sampling with unknown second moments. The second moments

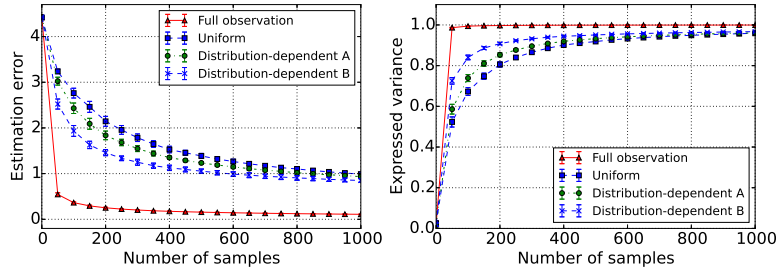


**Figure 9.2:** Comparison between RO-PCA, MSG, online robust PCA [FXMY14] and HR-PCA [XCM13] in the (a) entry-wise corruption case and (b) sample-wise corruption case. The  $x$ -axis is the probability that an attribute or sample is corrupted. The samples are generated under  $d = 500$ ,  $k = 10$  and  $\sigma = 0.05$ . Parameters  $T = 100$ ,  $m = 100$  and  $\rho = 1.2mq$ .

of the attributes are estimated via Algorithm 9.6. Then the probability vector  $\mathbf{q}$  used for attribute sampling are computed according to (9.7) with these estimated second moments; 3) Distribution-dependent sampling with known second moments. The probability vector  $\mathbf{q}$  is directly computed by utilizing these known second moments. For clarity, we denote these three schemes by “Uniform”, “Distribution-dependent A” and “Distribution-dependent B”, respectively.

Figure 9.3 shows the performance of Algorithm 9.4 under the three sampling schemes. We take Algorithm 9.1 with fully observed samples as the baseline algorithm. Clearly, the two distribution-dependent sampling schemes outperform the uniform sampling scheme, and the information about the second moments of the attributes are quite helpful, i.e., “Distribution-dependent B” has a much better performance than “Distribution-dependent A”. This is consistent with our theoretical results shown in Theorems 9.6 and 9.8.





**Figure 9.3:** Comparison between the three sampling schemes and the baseline that all the attributes are observed.

### 9.5.2 Real-world Data

We now show the performance of the proposed algorithms on the MNIST dataset [LJB<sup>+</sup>95]. Each digit image is represented by a 784-dimensional vector and the grayscale values of its pixels lie in  $[0, 1]$ . We first test our algorithms on 1000 digit images with missing or corrupted pixels. In the “missing entries” case, each pixel of the digit images is observed with probability  $q$ . In the “entry-wise corruption” case, for each pixel, it is corrupted by the white pixel, i.e., its grayscale value is set to 1, with probability  $q$ . Suppose that  $\mathbf{U}$  consists of the estimated principal components computed by a certain algorithm from  $T$  corrupted digit images  $\mathbf{x}_1, \dots, \mathbf{x}_T$  and let  $\hat{\mathbf{x}}_1, \dots, \hat{\mathbf{x}}_T$  be the corresponding non-corrupted digit images, then the performance of this algorithm is measured by the average representation error  $\frac{1}{T} \sum_{i=1}^T \|\hat{\mathbf{x}}_i - \mathbf{U}\mathbf{U}^\top \hat{\mathbf{x}}_i\|_2$ .

Table 9.1 shows the average representation errors of the outputs generated by RO-PCA, MSG and Batch when the digit images have missing pixels. Similar to the simulations, the outputs of all the algorithms become better as the probability that a pixel is revealed increases, and RO-PCA performs similar to Batch and outperforms MSG. Figure 9.4 illustrates the empirical performance of RO-PCA and MSG when the digit images contain corrupted pixels. Figure 9.4(a) and Figure 9.4(b) shows the leading five principal components extracted by RO-PCA and MSG, respectively, from which we observe that RO-PCA is more robust than MSG, i.e., the principal components extracted by MSG have more noisy entries than those extracted by RO-PCA.

**Table 9.1:** The average representation errors for different algorithms on a real dataset of 1000 digit images with missing entries.

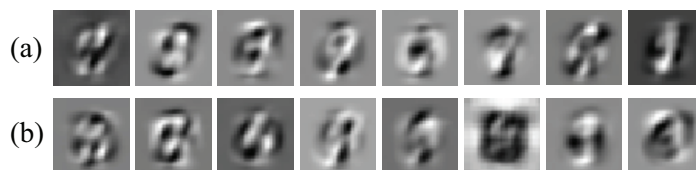
$q$	0.1	0.2	0.3	0.4	0.5	0.6
MSG	6.26	5.55	5.26	5.14	5.07	5.03
RO-PCA	6.11	5.45	5.25	5.13	5.06	5.04
Batch	6.09	5.44	5.25	5.12	5.06	5.03



**Figure 9.4:** The empirical performance of RO-PCA and MSG when the digit images are corrupted. We plot the leading five PCs extracted by (a) RO-PCA and (b) MSG.

The next experiment tests the performance of RO-PCA on a dataset considered in [YX15b] which mixes MNIST digit images with CBCL face images [Sun96]. We suppose that at time  $t$  sample  $\mathbf{x}_t$  is drawn from MNIST with probability 0.8 or from CBCL otherwise and take face images as outliers. Figure 9.5 shows the leading PCs extracted by RO-PCA and MSG. It can be observed that RO-PCA is more reliable than MSG, e.g., the sixth PC extracted by MSG mixes digits with faces, which is obviously unreliable.

Finally, we test the performance with  $\mathbf{Z}_t$  computed by Algorithm 9.4 when a limited number of pixels of the received digit images can be observed. Table 9.2 shows the average representation errors for the uniform sampling scheme and the distribution-dependent sampling scheme as the fraction of the observed pixels varies from 0.1 to 0.9, from which we observe that distribution-dependent sampling generates significantly better results than uniform



**Figure 9.5:** The leading eight PCs extracted by (a) RO-PCA and (b) MSG when the dataset is a mixture of digit and face images.

sampling.

**Table 9.2:** The average representation errors for different sampling schemes under the limited observation setting.

Fraction	0.1	0.3	0.5	0.7	0.9
Uniform	5.953	5.110	4.828	4.749	4.683
Dependent	5.662	4.853	4.716	4.661	4.617

## 9.6 Proofs of Technical Results

### Useful Lemmas

**Lemma 9.2.** (Lemma 3.1, [VCLR13]) Let  $\Sigma$  be a positive semidefinite matrix and  $\Pi$  be the projection onto the subspace spanned by the eigenvectors of  $\Sigma$  corresponding to its  $k$  largest eigenvalues  $\lambda_1 \geq \lambda_2 \geq \dots$ . If  $\delta = \lambda_k(\Sigma) - \lambda_{k+1}(\Sigma) > 0$ , then

$$\frac{\delta}{2} \|\Pi - \mathbf{X}\|_F^2 \leq \langle \Sigma, \Pi - \mathbf{X} \rangle$$

for all  $\mathbf{X}$  satisfying  $0 \preceq \mathbf{X} \preceq \mathbf{I}$  and  $\text{tr}(\mathbf{X}) = k$ .

**Lemma 9.3.** Suppose that  $\mathbf{Z}_1, \dots, \mathbf{Z}_T \in \mathbb{R}^{d \times d}$  are i.i.d. random variables with  $\mathbb{E}[\mathbf{Z}_t] = \Sigma$ , and  $\mathbf{P}_1, \dots, \mathbf{P}_T$  are generated by Algorithm 9.1 with  $\mathbf{Z}_1, \dots, \mathbf{Z}_T$ , satisfying that

$$\mathbb{E}\left[\sum_{t=1}^T \langle \mathbf{Z}_t, \mathbf{P}^* - \mathbf{P}_t \rangle\right] \leq G$$

for some constant  $G$ . Then

$$\mathbb{E}[\|\mathbf{P}^* - \bar{\mathbf{P}}\|_F^2] \leq \frac{2G}{(\lambda_k(\Sigma) - \lambda_{k+1}(\Sigma))T},$$

where  $\bar{\mathbf{P}} = \frac{1}{T} \sum_{t=1}^T \mathbf{P}_t$  and  $\lambda_k(\Sigma)$  is the  $k$ th largest eigenvalue of  $\Sigma$ .

*Proof.* Since  $\mathbf{Z}_t$  is independent with  $\mathbf{P}_t$  and  $\mathbb{E}[\mathbf{Z}_t] = \boldsymbol{\Sigma}$ , we have

$$\begin{aligned} \frac{G}{T} &\geq \mathbb{E}\left[\frac{1}{T} \sum_{t=1}^T \langle \mathbf{Z}_t, \mathbf{P}^* - \mathbf{P}_t \rangle\right] \\ &= \langle \mathbb{E}[\mathbf{Z}_t], \mathbf{P}^* \rangle - \langle \mathbb{E}[\mathbf{Z}_t], \mathbb{E}\left[\frac{1}{T} \sum_{t=1}^T \mathbf{P}_t\right] \rangle \\ &= \mathbb{E}[\langle \boldsymbol{\Sigma}, \mathbf{P}^* - \bar{\mathbf{P}} \rangle]. \end{aligned}$$

By Lemma 9.2,

$$\mathbb{E}[\langle \boldsymbol{\Sigma}, \mathbf{P}^* - \bar{\mathbf{P}} \rangle] \geq \mathbb{E}\left[\frac{1}{2}(\lambda_k(\boldsymbol{\Sigma}) - \lambda_{k+1}(\boldsymbol{\Sigma}))\|\mathbf{P}^* - \bar{\mathbf{P}}\|_F^2\right].$$

Hence we obtain this lemma. □

### Proof of Corollary 9.1

*Proof.* Since  $\mathbf{Z}_t$  can be decomposed into

$$\mathbf{Z}_t = \frac{1}{2}(\hat{\mathbf{x}}_t \tilde{\mathbf{x}}_t^\top + \tilde{\mathbf{x}}_t \hat{\mathbf{x}}_t^\top)$$

for some vectors  $\hat{\mathbf{x}}$  and  $\tilde{\mathbf{x}}$ , we have

$$\|\mathbf{Z}_t\|_{S(p)}^* \leq \|\hat{\mathbf{x}}_t \tilde{\mathbf{x}}_t^\top\|_{S(p)}^* = \|\sigma(\hat{\mathbf{x}}_t \tilde{\mathbf{x}}_t^\top)\|_p = \|\hat{\mathbf{x}}_t\|_2 \|\tilde{\mathbf{x}}_t\|_2.$$

By taking the expectation of the both sides in (9.3) and choosing

$$\eta = \sqrt{\frac{2\beta f(\mathbf{P}^*)}{\sum_{t=1}^T \mathbb{E}[\|\hat{\mathbf{x}}_t\|_2^2 \|\tilde{\mathbf{x}}_t\|_2^2]}}$$

we obtain this corollary. □

### Proof of Theorem 9.1

*Proof.* The proof follows from the standard procedure of deriving the regret bound for online mirror descent algorithms. Let  $\Delta_t = f^*(\boldsymbol{\Theta}_{t+1}) - f^*(\boldsymbol{\Theta}_t)$ . Since  $f(\cdot)$  is a closed and

$\beta$ -strongly convex function w.r.t.  $\|\cdot\|$ ,  $f^*(\cdot)$  is  $\frac{1}{\beta}$ -strongly smooth w.r.t.  $\|\cdot\|_*$ . Then

$$\begin{aligned} \sum_{t=1}^T \Delta_t &= \sum_{t=1}^T [f^*(\Theta_t + \eta \mathbf{Z}_t) - f^*(\Theta_t)] \\ &\leq \sum_{t=1}^T [\langle \nabla f^*(\Theta_t), \eta \mathbf{Z}_t \rangle + \frac{\eta^2}{2\beta} \|\mathbf{Z}_t\|_*^2] \\ &= \eta \sum_{t=1}^T [\langle \mathbf{P}_t, \mathbf{Z}_t \rangle + \frac{\eta}{2\beta} \|\mathbf{Z}_t\|_*^2]. \end{aligned}$$

On the other hand,

$$\begin{aligned} \sum_{t=1}^T \Delta_t &= f^*(\Theta_{T+1}) - f^*(\Theta_1) \\ &\geq \langle \mathbf{P}^*, \Theta_{T+1} \rangle - f(\mathbf{P}^*) - f^*(\mathbf{0}) \\ &= \eta \sum_{t=1}^T \langle \mathbf{P}^*, \mathbf{Z}_t \rangle - f(\mathbf{P}^*) - f^*(\mathbf{0}), \end{aligned}$$

where the inequality follows from the definition of the Fenchel conjugate. Hence we have

$$\sum_{t=1}^T \langle \mathbf{P}^*, \mathbf{Z}_t \rangle - \frac{f(\mathbf{P}^*) + f^*(\mathbf{0})}{\eta} \leq \sum_{t=1}^T [\langle \mathbf{P}_t, \mathbf{Z}_t \rangle + \frac{\eta}{2\beta} \|\mathbf{Z}_t\|_*^2],$$

which implies this theorem.  $\square$

**Proof of Theorem 9.2**

*Proof.* For simplicity, let  $\tilde{\mathbf{x}}_t = \frac{1}{q}\mathbf{x}_t$ , then

$$\begin{aligned}
\mathbb{E}[\|\mathbf{Z}_t(q)\|_F^2|\bar{\mathbf{x}}_t] &= \mathbb{E}[\text{tr}((\tilde{\mathbf{x}}_t\tilde{\mathbf{x}}_t^\top - (1-q)\text{diag}(\tilde{\mathbf{x}}_t\tilde{\mathbf{x}}_t^\top))^2)|\bar{\mathbf{x}}_t] \\
&= \mathbb{E}[\|\tilde{\mathbf{x}}_t\|_2^4 + (1-q)^2 \sum_{i=1}^d \tilde{\mathbf{x}}_t(i)^4 - 2(1-q) \sum_{i=1}^d \tilde{\mathbf{x}}_t(i)^4|\bar{\mathbf{x}}_t] \\
&= \mathbb{E}[\|\tilde{\mathbf{x}}_t\|_2^4 + (q^2 - 1) \sum_{i=1}^d \tilde{\mathbf{x}}_t(i)^4|\bar{\mathbf{x}}_t] \\
&= \mathbb{E}[q^2 \sum_{i=1}^d \tilde{\mathbf{x}}_t(i)^4 + \sum_{i \neq j} \tilde{\mathbf{x}}_t(i)^2 \tilde{\mathbf{x}}_t(j)^2|\bar{\mathbf{x}}_t] \\
&= \mathbb{E}[\sum_{i=1}^d \frac{\mathbf{x}_t(i)^4}{q^2} + \sum_{i \neq j} \frac{\mathbf{x}_t(i)^2 \mathbf{x}_t(j)^2}{q^4}|\bar{\mathbf{x}}_t] \\
&= \sum_{i=1}^d \frac{\bar{\mathbf{x}}_t(i)^4}{q^2} \cdot \delta + \sum_{i \neq j} \frac{\bar{\mathbf{x}}_t(i)^2 \bar{\mathbf{x}}_t(j)^2}{q^4} \cdot \delta^2 \\
&\leq \frac{\delta}{q^2} \sum_{i=1}^d \bar{\mathbf{x}}_t(i)^4 + \frac{\delta^2}{q^4} (B^2 - \sum_{i=1}^d \bar{\mathbf{x}}_t(i)^4)
\end{aligned}$$

where the last inequality holds since  $\|\bar{\mathbf{x}}_t\|_2^2 \leq B$ . Therefore, we have

$$\begin{aligned}
\mathbb{E}[\|\mathbf{Z}_t(q)\|_F^2|\bar{\mathbf{x}}_t] &\leq \frac{\delta^2}{q^4} B^2 + \frac{\delta}{q^2} \sum_{i=1}^d \bar{\mathbf{x}}_t(i)^4 \cdot (1 - \frac{\delta}{q^2}) \\
&\leq \frac{\delta^2}{q^4} B^2 + \max\{\frac{\delta}{q^2} (1 - \frac{\delta}{q^2}) B^2, 0\} \\
&\leq \max\left\{1, \frac{\delta}{q^2}\right\} \frac{\delta}{q^2} B^2.
\end{aligned} \tag{9.8}$$

Note that

$$\begin{aligned}
\mathbb{E}[\mathbf{Z}_t(q)|\bar{\mathbf{x}}_t] &= \mathbb{E}[\frac{1}{q^2} \mathbf{x}_t \mathbf{x}_t^\top - \frac{1-q}{q^2} \text{diag}(\mathbf{x}_t \mathbf{x}_t^\top)|\bar{\mathbf{x}}_t] \\
&= \frac{\delta^2}{q^2} \bar{\mathbf{x}}_t \bar{\mathbf{x}}_t^\top - \frac{\delta(\delta-q)}{q^2} \text{diag}(\bar{\mathbf{x}}_t \bar{\mathbf{x}}_t^\top),
\end{aligned}$$

which implies that  $\mathbf{Z}_t(\delta)$  is an unbiased estimator of  $\bar{\mathbf{x}}_t \bar{\mathbf{x}}_t^\top$ . Then consider the following inequalities

$$\begin{aligned} \sum_{t=1}^T \langle \mathbf{Z}_t(\delta), \mathbf{P}^* - \mathbf{P}_t \rangle &= \sum_{t=1}^T \langle \mathbf{Z}_t(q), \mathbf{P}^* - \mathbf{P}_t \rangle + \sum_{t=1}^T \langle \mathbf{Z}_t(\delta) - \mathbf{Z}_t(q), \mathbf{P}^* - \mathbf{P}_t \rangle \\ &\leq \sum_{t=1}^T \langle \mathbf{Z}_t(q), \mathbf{P}^* - \mathbf{P}_t \rangle + 2\sqrt{k} \sum_{t=1}^T \|\mathbf{Z}_t(\delta) - \mathbf{Z}_t(q)\|_F, \end{aligned}$$

where the last inequality holds because  $\|\mathbf{P}^*\|_F \leq \sqrt{k}$  and  $\|\mathbf{P}_t\|_F \leq \sqrt{k}$ . From the definition of  $\mathbf{Z}_t(q)$ ,

$$\begin{aligned} \|\mathbf{Z}_t(\delta) - \mathbf{Z}_t(q)\|_F &= \left\| \left( \frac{1}{q^2} - \frac{1}{\delta^2} \right) \mathbf{x}_t \mathbf{x}_t^\top - \left( \frac{1-q}{q^2} - \frac{1-\delta}{\delta^2} \right) \text{diag}(\mathbf{x}_t \mathbf{x}_t^\top) \right\|_F \\ &\leq \frac{|\delta - q|(2\delta + 2q - q\delta)}{q^2 \delta^2} \|\mathbf{x}_t \mathbf{x}_t^\top\|_F \\ &\leq \frac{3B|\delta - q|}{q^2 \delta^2}, \end{aligned}$$

where the last inequality holds since  $q, \delta \in [0, 1]$  and  $\|\mathbf{x}_t \mathbf{x}_t^\top\|_F \leq \|\bar{\mathbf{x}}_t \bar{\mathbf{x}}_t^\top\|_F \leq B$ . Then by Theorem 9.1 and Inequality (9.8), we have

$$\begin{aligned} \mathbb{E} \left[ \sum_{t=1}^T \langle \mathbf{Z}_t(\delta), \mathbf{P}^* - \mathbf{P}_t \rangle \right] &\leq \frac{k}{2\eta} + \frac{\delta B^2}{q^2} \max \left\{ 1, \frac{\delta}{q^2} \right\} \cdot \eta T + \frac{6\sqrt{k}|\delta - q|BT}{q^2 \delta^2} \\ &= \sqrt{\max \left\{ 1, \frac{\delta}{q^2} \right\} \cdot \frac{2\delta B^2}{q^2} \cdot kT} + \frac{6\sqrt{k}|\delta - q|BT}{q^2 \delta^2} \end{aligned}$$

with  $\eta = \sqrt{\frac{kq^2}{\max\{1, \delta/q^2\} \cdot 2\delta B^2 T}}$ . Finally, by Lemma 9.3, we have

$$\mathbb{E}[\|\mathbf{P}^* - \bar{\mathbf{P}}\|_F^2] \leq \frac{2}{\Delta_k} \left( \sqrt{\max \left\{ 1, \frac{\delta}{q^2} \right\} \cdot \frac{2\delta B^2}{q^2} \cdot \frac{k}{T}} + \frac{6\sqrt{k}|\delta - q|B}{q^2 \delta^2} \right)$$

where  $\Delta_k = \lambda_k(\boldsymbol{\Sigma}) - \lambda_{k+1}(\boldsymbol{\Sigma})$ . □

### Proof of Theorem 9.3

*Proof.* At time  $t$ , suppose that we receive a block of samples  $\mathbf{x}_1, \dots, \mathbf{x}_m$ . Since  $\mathbf{x}_1, \dots, \mathbf{x}_m$  may contain corrupted entries, we apply Algorithm 9.2 to construct  $\mathbf{Z}_t$  which satisfies  $\mathbf{Z}_t =$

$\hat{\mathbf{X}}\hat{\mathbf{X}}^\top/m$  where  $\hat{\mathbf{X}}$  is obtained by keeping the smallest  $m - \rho$  entries of each row of  $\mathbf{X}$  and setting the other entries to 0. Let  $\bar{\mathbf{x}}_1, \dots, \bar{\mathbf{x}}_m$  be the corresponding non-corrupted samples of  $\mathbf{x}_1, \dots, \mathbf{x}_m$  and  $\bar{\mathbf{X}}$  be their sample matrix. Define  $\bar{\mathbf{Z}}_t \triangleq \bar{\mathbf{X}}\bar{\mathbf{X}}^\top/m$ , then we have  $\mathbb{E}[\bar{\mathbf{Z}}_t] = \mathbb{E}_{\mathbf{x} \sim \mathcal{D}}[\mathbf{x}\mathbf{x}^\top]$ .

The first step of the proof is to bound  $\|\mathbf{Z}_t - \bar{\mathbf{Z}}_t\|_\infty = \max_{i,j} |\mathbf{Z}_t(i,j) - \bar{\mathbf{Z}}_t(i,j)|$ . For fixed  $i$  and  $j$ , we define three sets:

$$\mathcal{N} \triangleq \{k : \mathbf{X}_{(i)}(k) \text{ and } \mathbf{X}_{(j)}(k) \text{ are non-corrupted}\},$$

$$\mathcal{C} \triangleq \{k : \mathbf{X}_{(i)}(k) \text{ or } \mathbf{X}_{(j)}(k) \text{ is corrupted}\},$$

$$\mathcal{T} \triangleq \{k : \mathbf{X}_{(i)}(k) \text{ or } \mathbf{X}_{(j)}(k) \text{ is trimmed to zero}\}.$$

Assume that parameter  $\rho$  is greater than the number of corrupted entries in each row of  $\mathbf{X}$  (we will later show this assumption holds with high probability under some mild conditions on  $m$ ), then

$$\begin{aligned} m|\mathbf{Z}_t(i,j) - \bar{\mathbf{Z}}_t(i,j)| &= \sum_{k \in \mathcal{N} \cap \mathcal{T}^c} \hat{\mathbf{X}}_{(i)}(k)\hat{\mathbf{X}}_{(j)}(k) + \sum_{k \in \mathcal{C} \cap \mathcal{T}^c} \hat{\mathbf{X}}_{(i)}(k)\hat{\mathbf{X}}_{(j)}(k) - \\ &\quad \sum_{k \in \mathcal{N}} \bar{\mathbf{X}}_{(i)}(k)\bar{\mathbf{X}}_{(j)}(k) - \sum_{k \in \mathcal{C}} \bar{\mathbf{X}}_{(i)}(k)\bar{\mathbf{X}}_{(j)}(k) \\ &= \sum_{k \in \mathcal{C} \cap \mathcal{T}^c} \hat{\mathbf{X}}_{(i)}(k)\hat{\mathbf{X}}_{(j)}(k) - \sum_{k \in \mathcal{N} \cap \mathcal{T}} \bar{\mathbf{X}}_{(i)}(k)\bar{\mathbf{X}}_{(j)}(k) - \sum_{k \in \mathcal{C}} \bar{\mathbf{X}}_{(i)}(k)\bar{\mathbf{X}}_{(j)}(k) \\ &\leq 6\rho \max_{i \in [d]} \|\bar{\mathbf{X}}_{(i)}\|_\infty^2, \end{aligned}$$

where the last inequality holds because  $|\mathcal{C}| \leq 2\rho$  and  $|\mathcal{T}| \leq 2\rho$ . Recall that  $\|\bar{\mathbf{x}}_i\|_2^2 \leq B$  for  $i = 1, \dots, m$ , which implies  $\max_i \|\bar{\mathbf{X}}_{(i)}\|_\infty^2 \leq \max_{i \in [m]} \|\bar{\mathbf{x}}_i\|_2^2 \leq B$ . Therefore

$$\|\mathbf{Z}_t - \bar{\mathbf{Z}}_t\|_\infty \leq \frac{6\rho B}{m}.$$

The second step is applying Theorem 9.1 to show the performance guarantee of Algorithm 9.1 with these  $\mathbf{Z}_t$ . Recall that the domain set  $\mathcal{F} = \{\mathbf{P} : \mathbf{0} \leq \mathbf{P} \leq \mathbf{I}_d, \text{tr}(\mathbf{P}) = k, \|\mathbf{P}\|_1 \leq \gamma\sqrt{k}\}$



where  $\gamma$  is the number of nonzero entries of  $\mathbf{P}^*$ . Then

$$\begin{aligned} \sum_{t=1}^T \langle \bar{\mathbf{Z}}_t, \mathbf{P}^* - \mathbf{P}_t \rangle &= \sum_{t=1}^T \langle \mathbf{Z}_t, \mathbf{P}^* - \mathbf{P}_t \rangle + \sum_{t=1}^T \langle \bar{\mathbf{Z}}_t - \mathbf{Z}_t, \mathbf{P}^* - \mathbf{P}_t \rangle \\ &\leq \sum_{t=1}^T \langle \mathbf{Z}_t, \mathbf{P}^* - \mathbf{P}_t \rangle + \sum_{t=1}^T \|\bar{\mathbf{Z}}_t - \mathbf{Z}_t\|_\infty \|\mathbf{P}^* - \mathbf{P}_t\|_1 \\ &\leq \sum_{t=1}^T \langle \mathbf{Z}_t, \mathbf{P}^* - \mathbf{P}_t \rangle + \frac{12B\rho\gamma\sqrt{kT}}{m}. \end{aligned}$$

Thus, by Theorem 9.1, we have

$$\begin{aligned} \mathbb{E}\left[\sum_{t=1}^T \langle \bar{\mathbf{Z}}_t, \mathbf{P}^* - \mathbf{P}_t \rangle\right] &\leq \mathbb{E}\left[\sum_{t=1}^T \langle \mathbf{Z}_t, \mathbf{P}^* - \mathbf{P}_t \rangle\right] + \frac{12B\rho\gamma\sqrt{kT}}{m} \\ &\leq \frac{k}{2\eta} + \eta \sum_{t=1}^T \mathbb{E}[\|\mathbf{Z}_t\|_F^2] + \frac{12B\rho\gamma\sqrt{kT}}{m} \\ &\leq \frac{k}{2\eta} + \eta B^2 T + \frac{12B\rho\gamma\sqrt{kT}}{m} \end{aligned}$$

where the last inequality follows from

$$\begin{aligned} \|\mathbf{Z}_t\|_F^2 &\leq \frac{1}{m^2} \|\hat{\mathbf{X}}\hat{\mathbf{X}}^\top\|_F^2 \leq \frac{1}{m^2} \|\hat{\mathbf{X}}\|_F^4 \\ &= \frac{1}{m^2} \left[ \sum_{i=1}^d \sum_{j \in \{j: \hat{\mathbf{X}}_{(i)}(j) \text{ is not trimmed to } 0\}} \hat{\mathbf{X}}_{(i)}(j)^2 \right]^2 \\ &\leq \frac{1}{m^2} \left[ \sum_{i=1}^d \sum_{j \in \{j: \hat{\mathbf{X}}_{(i)}(j) \text{ is non-corrupted}\}} \hat{\mathbf{X}}_{(i)}(j)^2 \right]^2 \\ &\leq \frac{1}{m^2} \left[ \sum_{i=1}^d \sum_{j=1}^m \bar{\mathbf{X}}_{(i)}(j)^2 \right]^2 \\ &\leq \frac{1}{m^2} \left[ \sum_{j=1}^m \|\bar{\mathbf{X}}_j\|_2^2 \right]^2 \leq B^2. \end{aligned}$$

Therefore, when  $\eta = \sqrt{\frac{k}{2B^2T}}$ , we have

$$\mathbb{E}\left[\sum_{t=1}^T \langle \bar{\mathbf{Z}}_t, \mathbf{P}^* - \mathbf{P}_t \rangle\right] \leq B\sqrt{2kT} + \frac{12B\rho\gamma\sqrt{kT}}{m}.$$

Since  $\mathbb{E}[\bar{\mathbf{Z}}_t] = \boldsymbol{\Sigma}$ , by Lemma 9.3, we have

$$\mathbb{E}[\|\mathbf{P}^* - \bar{\mathbf{P}}\|_F^2] \leq \frac{2B}{\Delta_k} \left( \sqrt{\frac{2k}{T}} + \frac{12\rho\gamma\sqrt{k}}{m} \right)$$

where  $\Delta_k = \lambda_k(\boldsymbol{\Sigma}) - \lambda_{k+1}(\boldsymbol{\Sigma})$ .

The final step is to find the relationship between  $q$ ,  $m$  and  $\rho$ . For the case where each entry of the received samples is corrupted with probability  $q$ , let  $y_j$  be the indicator of the  $j$ th entry of  $\mathbf{X}_{(i)}$  being corrupted, i.e.,  $y_j = 1$  if  $\mathbf{X}_{(i)}(j)$  is corrupted or 0 otherwise. By the Chernoff bound, for all  $\delta > 0$ ,

$$\mathbb{P}\left[\sum_{j=1}^m y_j \geq (1 + \delta)mq\right] \leq \exp\left(-\frac{\delta^2}{2 + \delta}mq\right),$$

which implies that when  $m \geq \frac{2(2+\delta)\log(dT)}{\delta^2q}$ ,  $\sum_{i=1}^m y_j \leq (1 + \delta)mq$  holds with probability at least  $1 - \frac{1}{d^2T^2}$ . By the union bound, with probability at least  $1 - \frac{1}{dT}$ , the number of corrupted entries in each row of sample matrix  $\mathbf{X}$  is less than or equal to  $(1 + \delta)mq$  for all  $t = 1$  to  $T$ .

Therefore, as long as  $\rho \geq (1 + \delta)mq$ , we can guarantee that  $\rho$  is greater than the number of corrupted entries in each row of  $\mathbf{X}$  with high probability. Hence this theorem is obtained.  $\square$

### Proof of Theorem 9.4

*Proof.* Similar to the proof of Theorem 9.3, suppose that we receive a block of samples  $\mathbf{x}_1, \dots, \mathbf{x}_m$ , some of which are corrupted. So we apply Algorithm 9.3 to construct  $\mathbf{Z}_t$  which satisfies  $\mathbf{Z}_t = \hat{\mathbf{X}}\hat{\mathbf{X}}^\top/m$  where  $\hat{\mathbf{X}}$  is obtained by keeping the smallest  $m - \rho$  samples w.r.t.  $\|\cdot\|_2$  and setting the other columns to  $\mathbf{0}$ . Let  $\bar{\mathbf{x}}_1, \dots, \bar{\mathbf{x}}_m$  be the corresponding non-corrupted samples of  $\mathbf{x}_1, \dots, \mathbf{x}_m$  and  $\bar{\mathbf{X}}$  be their sample matrix. Define  $\bar{\mathbf{Z}}_t \triangleq \bar{\mathbf{X}}\bar{\mathbf{X}}^\top/m$ , then  $\mathbb{E}[\bar{\mathbf{Z}}_t] = \mathbb{E}_{\mathbf{x} \sim \mathcal{D}}[\mathbf{x}\mathbf{x}^\top]$ .

The first step of the proof is to bound  $\|\mathbf{Z}_t - \bar{\mathbf{Z}}_t\|_F$ . We first define the following three sets:

$$\begin{aligned}\mathcal{N} &\triangleq \{k : \mathbf{x}_k \text{ is non-corrupted}\}, \\ \mathcal{C} &\triangleq \{k : \mathbf{x}_k \text{ is corrupted}\}, \\ \mathcal{T} &\triangleq \{k : \mathbf{x}_k \text{ is trimmed to zero}\}.\end{aligned}$$

Assume that parameter  $\rho$  is greater than the number of corrupted columns of  $\mathbf{X}$  (we will later show this assumption holds with high probability), then

$$\begin{aligned}m\|\mathbf{Z}_t - \bar{\mathbf{Z}}_t\|_F &= \left\| \sum_{k \in \mathcal{N} \cap \mathcal{T}^c} \hat{\mathbf{x}}_k \hat{\mathbf{x}}_k^\top + \sum_{k \in \mathcal{C} \cap \mathcal{T}^c} \hat{\mathbf{x}}_k \hat{\mathbf{x}}_k^\top - \sum_{k \in \mathcal{N} \cup \mathcal{C}} \bar{\mathbf{x}}_k \bar{\mathbf{x}}_k^\top \right\|_F \\ &= \left\| \sum_{k \in \mathcal{C} \cap \mathcal{T}^c} \hat{\mathbf{x}}_k \hat{\mathbf{x}}_k^\top - \sum_{k \in \mathcal{N} \cap \mathcal{T}} \bar{\mathbf{x}}_k \bar{\mathbf{x}}_k^\top - \sum_{k \in \mathcal{C}} \bar{\mathbf{x}}_k \bar{\mathbf{x}}_k^\top \right\|_F \\ &\leq 3\rho \max_i \|\bar{\mathbf{x}}_i\|_2^2,\end{aligned}$$

where the last inequality holds because  $|\mathcal{C}| \leq \rho$  and  $|\mathcal{T}| \leq \rho$ . Recall that  $\|\bar{\mathbf{x}}_i\|_2^2 \leq B$  for  $i = 1, \dots, m$ , then

$$\|\mathbf{Z}_t - \bar{\mathbf{Z}}_t\|_F \leq \frac{3\rho B}{m}.$$

The second step is applying Theorem 9.1 to show the performance guarantee of Algorithm 9.1 with these  $\mathbf{Z}_t$ , namely,

$$\begin{aligned}\sum_{t=1}^T \langle \bar{\mathbf{Z}}_t, \mathbf{P}^* - \mathbf{P}_t \rangle &= \sum_{t=1}^T \langle \mathbf{Z}_t, \mathbf{P}^* - \mathbf{P}_t \rangle + \sum_{t=1}^T \langle \bar{\mathbf{Z}}_t - \mathbf{Z}_t, \mathbf{P}^* - \mathbf{P}_t \rangle \\ &\leq \sum_{t=1}^T \langle \mathbf{Z}_t, \mathbf{P}^* - \mathbf{P}_t \rangle + \sum_{t=1}^T \|\bar{\mathbf{Z}}_t - \mathbf{Z}_t\|_F \|\mathbf{P}^* - \mathbf{P}_t\|_F \\ &\leq \sum_{t=1}^T \langle \mathbf{Z}_t, \mathbf{P}^* - \mathbf{P}_t \rangle + \frac{6B\rho\sqrt{kT}}{m}.\end{aligned}$$

Thus, by Theorem 9.1, we have

$$\mathbb{E}\left[\sum_{t=1}^T \langle \bar{\mathbf{Z}}_t, \mathbf{P}^* - \mathbf{P}_t \rangle\right] \leq \frac{k}{2\eta} + \eta B^2 T + \frac{6B\rho\sqrt{kT}}{m}.$$

Therefore, when  $\eta = \sqrt{\frac{k}{2B^2T}}$ , we have

$$\mathbb{E}\left[\sum_{t=1}^T \langle \bar{\mathbf{Z}}_t, \mathbf{P}^* - \mathbf{P}_t \rangle\right] \leq B\sqrt{2kT} + \frac{6B\rho\sqrt{kT}}{m}.$$

Since  $\mathbb{E}[\bar{\mathbf{Z}}_t] = \boldsymbol{\Sigma}$ , by Lemma 9.3, we have

$$\mathbb{E}[\|\mathbf{P}^* - \bar{\mathbf{P}}\|_F^2] \leq \frac{2B}{\Delta_k} \left( \sqrt{\frac{2k}{T}} + \frac{6\rho\sqrt{k}}{m} \right)$$

where  $\Delta_k = \lambda_k(\boldsymbol{\Sigma}) - \lambda_{k+1}(\boldsymbol{\Sigma})$ .

The final step is to find the relationship between  $q$ ,  $m$  and  $\rho$ . Let  $y_i$  be the indicator of  $\mathbf{x}_i$  being corrupted, i.e.,  $y_i = 1$  if  $\mathbf{x}_i$  is corrupted or 0 otherwise. By the Chernoff bound, for all  $\delta > 0$ ,

$$\mathbb{P}\left[\sum_{i=1}^m y_i \geq (1 + \delta)mq\right] \leq \exp\left(-\frac{\delta^2}{2 + \delta}mq\right),$$

which implies that when  $m \geq \frac{2(2+\delta)\log(T)}{\delta^2q}$ ,  $\sum_{i=1}^m y_i \leq (1 + \delta)mq$  holds with probability at least  $1 - \frac{1}{T^2}$ . By the union bound, we know that with probability at least  $1 - \frac{1}{T}$  the number of corrupted samples in each block from time  $t = 1$  to  $T$  is less than or equal to  $(1 + \delta)mq$ .

Therefore, as long as  $\rho \geq (1 + \delta)mq$ , we can guarantee that  $\rho$  is greater than the number of corrupted entries in each row of  $\mathbf{X}$  with high probability. Hence we obtain this theorem.  $\square$

## Proofs of Theorem 9.5 and Theorem 9.6

Before the main proofs are given, we first provide several useful lemmas.

**Lemma 9.4.** *The vectors  $\hat{\mathbf{x}}_t$  and  $\tilde{\mathbf{x}}_t$  constructed by Algorithm 9.4 satisfy that*

$$\mathbb{E}[\|\hat{\mathbf{x}}_t\|_2^2 | \mathbf{x}_t] = \mathbb{E}[\|\tilde{\mathbf{x}}_t\|_2^2 | \mathbf{x}_t] = \frac{2}{s} \sum_{i=1}^d \frac{\mathbf{x}_t(i)^2}{q_i} + \frac{s-2}{s} \|\mathbf{x}_t\|_2^2.$$

*Proof.* Recall that  $\hat{\mathbf{x}}_t = \frac{2}{s} \sum_{r=1}^{s/2} \frac{1}{q_{i_t,r}} \mathbf{x}_t(i_{t,r}) \mathbf{e}_{i_t,r}$  where indices  $i_{t,r}$  are independently sampled

from [d] with probabilities  $q_{i_t,r}$ . For clarity, let  $l = s/2$ , then

$$\begin{aligned} l^2 \cdot \|\hat{\mathbf{x}}_t\|_2^2 &= \left( \sum_{r=1}^l \frac{1}{q_{i_t,r}} \mathbf{x}_t(i_{t,r}) \mathbf{e}_{i_{t,r}} \right)^\top \left( \sum_{r=1}^l \frac{1}{q_{i_t,r}} \mathbf{x}_t(i_{t,r}) \mathbf{e}_{i_{t,r}} \right) \\ &= \sum_{r=1}^l \frac{1}{q_{i_t,r}^2} \mathbf{x}_t(i_{t,r})^2 + \sum_{r \neq u} \frac{1}{q_{i_t,r} q_{i_t,u}} \mathbf{x}_t(i_{t,r}) \mathbf{x}_t(i_{t,u}) \mathbf{e}_{i_{t,r}}^\top \mathbf{e}_{i_{t,u}}. \end{aligned}$$

Note that  $i_{t,r}$  and  $i_{t,u}$  are independent when  $r \neq u$ . Thus we have

$$\begin{aligned} &\mathbb{E} \left[ \frac{1}{q_{i_{t,r}} q_{i_{t,u}}} \mathbf{x}_t(i_{t,r}) \mathbf{x}_t(i_{t,u}) \mathbf{e}_{i_{t,r}}^\top \mathbf{e}_{i_{t,u}} \mid \mathbf{x}_t \right] \\ &= \mathbb{E} \left[ \frac{1}{q_{i_{t,r}}} \mathbf{x}_t(i_{t,r}) \mathbf{e}_{i_{t,r}} \mid \mathbf{x}_t \right]^\top \mathbb{E} \left[ \frac{1}{q_{i_{t,u}}} \mathbf{x}_t(i_{t,u}) \mathbf{e}_{i_{t,u}} \mid \mathbf{x}_t \right] \\ &= \mathbf{x}_t \mathbf{x}_t^\top. \end{aligned}$$

Thus,  $\mathbb{E} [l^2 \|\hat{\mathbf{x}}_t\|_2^2 \mid \mathbf{x}_t] = l \cdot \sum_{i=1}^d \frac{\mathbf{x}_t(i)^2}{q_i} + l(l-1) \|\mathbf{x}_t\|_2^2$ , which implies that

$$\mathbb{E} [\|\hat{\mathbf{x}}_t\|_2^2 \mid \mathbf{x}_t] = \frac{1}{l} \sum_{i=1}^d \frac{\mathbf{x}_t(i)^2}{q_i} + \frac{l-1}{l} \|\mathbf{x}_t\|_2^2.$$

By substituting  $l$  by  $s/2$ , we obtain this lemma. □

We now prove Theorem 9.5 and Theorem 9.6.

*Proof.* By Corollary 9.1, we only need to find an upper bound of  $\mathbb{E}_{\mathbf{x}_t \sim \mathcal{D}} \mathbb{E} [\|\hat{\mathbf{x}}_t\|_2^2 \|\tilde{\mathbf{x}}_t\|_2^2 \mid \mathbf{x}_t]$ .

By Lemma 9.4 and  $\|\mathbf{x}_t\|_2^2 \leq B$ , we know that

$$\begin{aligned} &\mathbb{E}_{\mathbf{x}_t \sim \mathcal{D}} \mathbb{E} [\|\hat{\mathbf{x}}_t\|_2^2 \|\tilde{\mathbf{x}}_t\|_2^2 \mid \mathbf{x}_t] \\ &= \mathbb{E}_{\mathbf{x}_t \sim \mathcal{D}} \left[ \left( \frac{2}{s} \sum_{i=1}^d \frac{\mathbf{x}_t(i)^2}{q_i} + \frac{s-2}{s} \|\mathbf{x}_t\|_2^2 \right)^2 \right] \\ &\leq \mathbb{E}_{\mathbf{x}_t \sim \mathcal{D}} \left[ \left( \frac{2}{s} \sum_{i=1}^d \frac{\mathbf{x}_t(i)^2}{q_i} + \frac{s-2}{s} B \right)^2 \right]. \end{aligned} \tag{9.9}$$

Therefore, when  $q_i = \frac{1}{d}$  for all  $i = 1, \dots, d$ , we have

$$\mathbb{E}_{\mathbf{x}_t \sim \mathcal{D}} \mathbb{E} [\|\hat{\mathbf{x}}_t\|_2^2 \|\tilde{\mathbf{x}}_t\|_2^2 \mid \mathbf{x}_t] \leq \left( \frac{2d}{s} B + \frac{s-2}{s} B \right)^2 \leq \left( \frac{3dB}{s} \right)^2,$$

where the last inequality holds since  $d \geq s$ . Then from Corollary 9.1, we know that after  $T$  iterations with step size  $\eta = \frac{s}{3dB} \sqrt{\frac{2\beta f(\mathbf{P}^*)}{T}}$ , the outputs  $\mathbf{P}_t$  of Algorithm 9.1 satisfy that

$$\mathbb{E}\left[\sum_{t=1}^T \langle \mathbf{Z}_t, \mathbf{P}^* - \mathbf{P}_t \rangle\right] \leq \frac{3dB}{s} \sqrt{\frac{2f(\mathbf{P}^*)T}{\beta}}.$$

Then by applying Lemma 9.3, we obtain Theorem 9.5.

When  $\mathbf{q} \neq (\frac{1}{d}, \dots, \frac{1}{d})$ , Inequality (9.9) becomes

$$\begin{aligned} \mathbb{E}_{\mathbf{x}_t \sim \mathcal{D}} \mathbb{E}[\|\hat{\mathbf{x}}_t\|_2^2 \|\tilde{\mathbf{x}}_t\|_2^2 | \mathbf{x}_t] &\leq \mathbb{E}_{\mathbf{x}_t \sim \mathcal{D}} \left[ \left( \frac{2}{s} \sum_{i=1}^d \frac{\mathbf{x}_t(i)^2}{q_i} + B \right)^2 \right] \\ &\leq \frac{8}{s^2} \mathbb{E}_{\mathbf{x}_t \sim \mathcal{D}} \left[ \left( \sum_{i=1}^d \frac{\mathbf{x}_t(i)^2}{q_i} \right)^2 \right] + 2B^2 \\ &\leq \frac{8}{s^2} \mathbb{E}_{\mathbf{x}_t \sim \mathcal{D}} \left[ \|\mathbf{x}_t\|_2^2 \cdot \sum_{i=1}^d \frac{\mathbf{x}_t(i)^2}{q_i^2} \right] + 2B^2 \\ &\leq \frac{8B}{s^2} \sum_{i=1}^d \frac{\mathbb{E}_{\mathbf{x}_t \sim \mathcal{D}}[\mathbf{x}_t(i)^2]}{q_i^2} + 2B^2. \end{aligned} \tag{9.10}$$

Since  $\sum_{i=1}^d q_i = 1$ ,  $\sum_{i=1}^d \frac{\mathbb{E}_{\mathbf{x}_t \sim \mathcal{D}}[\mathbf{x}_t(i)^2]}{q_i^2}$  is minimized when

$$q_i = \frac{\mathbb{E}_{\mathbf{x}_t \sim \mathcal{D}}[\mathbf{x}_t(i)^2]^{\frac{1}{3}}}{\sum_{i=1}^d \mathbb{E}_{\mathbf{x}_t \sim \mathcal{D}}[\mathbf{x}_t(i)^2]^{\frac{1}{3}}}, \quad i = 1, \dots, d.$$

Then in this case, we have

$$\mathbb{E}_{\mathbf{x}_t \sim \mathcal{D}} \mathbb{E}[\|\hat{\mathbf{x}}_t\|_2^2 \|\tilde{\mathbf{x}}_t\|_2^2 | \mathbf{x}_t] \leq \frac{8B}{s^2} \left( \sum_{i=1}^d \mathbb{E}_{\mathbf{x}_t \sim \mathcal{D}}[\mathbf{x}_t(i)^2]^{\frac{1}{3}} \right)^3 + 2B^2.$$

By Corollary 9.1, we know that after  $T$  iterations with step size

$$\eta = \sqrt{\frac{2\beta f(\mathbf{P}^*)}{\left[ \frac{8B}{s^2} \left( \sum_{i=1}^d \mathbb{E}_{\mathbf{x}_t \sim \mathcal{D}}[\mathbf{x}_t(i)^2]^{\frac{1}{3}} \right)^3 + 2B^2 \right] T}},$$

the outputs  $\mathbf{P}_t$  of Algorithm 9.1 satisfy that

$$\mathbb{E}\left[\sum_{t=1}^T \langle \mathbf{Z}_t, \mathbf{P}^* - \mathbf{P}_t \rangle\right] \leq \sqrt{\frac{2f(\mathbf{P}^*) \left[ \frac{8B}{s^2} \left( \sum_{i=1}^d \mathbb{E}_{\mathbf{x}_t \sim \mathcal{D}} [\mathbf{x}_t(i)^2] \right)^{\frac{1}{3}} + 2B^2 \right] T}{\beta}}.$$

Therefore, by applying Lemma 9.3, we obtain Theorem 9.6.  $\square$

### Proof of Theorem 9.7

*Proof.* Recall that  $q$  is calculated by Algorithm 9.5. By the Chernoff bound, the following inequalities hold for all  $0 < \theta < 1$ ,

$$\begin{aligned} \mathbb{P}[qm_0d \geq (1 + \theta)m_0d\delta] &\leq \exp\left(-\frac{\theta^2}{2 + \theta}m_0d\delta\right), \\ \mathbb{P}[qm_0d \leq (1 - \theta)m_0d\delta] &\leq \exp\left(-\frac{\theta^2}{2 + \theta}m_0d\delta\right). \end{aligned}$$

Thus, for  $0 \leq \alpha \leq \frac{1}{2}$ ,  $\theta = \frac{1}{2}T^{-\alpha}$  and  $m_0 \geq \frac{12 \log d}{\delta d} T^{2\alpha}$ , we have

$$(1 - \theta)\delta \leq q \leq (1 + \theta)\delta \tag{9.11}$$

holds with probability at least  $1 - \frac{2}{d}$ . As shown in the proof of Theorem 9.2, for  $\eta > 0$ ,

$$\mathbb{E}\left[\sum_{t=1}^T \langle \mathbf{Z}_t(\delta), \mathbf{P}^* - \mathbf{P}_t \rangle\right] \leq \frac{k}{2\eta} + \frac{\delta B^2}{q^2} \max\left\{1, \frac{\delta}{q^2}\right\} \cdot \eta T + \frac{6\sqrt{k}|\delta - q|BT}{q^2\delta^2}.$$

When Inequality (9.11) holds, we have

$$\begin{aligned} \mathbb{E}\left[\sum_{t=1}^T \langle \mathbf{Z}_t(\delta), \mathbf{P}^* - \mathbf{P}_t \rangle\right] &\leq \frac{k}{2\eta} + \frac{\delta B^2}{q^2} \max\left\{1, \frac{\delta}{q^2}\right\} \cdot \eta T + \frac{6\sqrt{k}|\delta - q|BT}{q^2\delta^2} \\ &\leq \frac{k}{2\eta} + \frac{B^2}{q^2(1 - \theta)^2} \eta T + \frac{6\sqrt{k}\theta(1 + \theta)BT}{q^3} \\ &\leq \frac{k}{2\eta} + \frac{4B^2}{q^2} \cdot \eta T + \frac{9\sqrt{k}B}{2q^3} \cdot T^{1-\alpha} \\ &\leq \frac{2B}{q} \sqrt{2kT} + \frac{9\sqrt{k}B}{2q^3} \cdot T^{1-\alpha}, \end{aligned}$$

where the last equality holds when  $\eta = \sqrt{\frac{kq^2}{8B^2T}}$ . Finally, by  $q \geq (1 - \theta)\delta \geq \frac{1}{2}\delta$  and Lemma 9.3, we obtain this theorem.  $\square$

### Proof of Theorem 9.8

Before the main proof is given, we provide two useful concentration inequalities. The first one is based on the Chernoff bound and the second one is based on the Bernstein inequality.

**Lemma 9.5.** *Let  $n_i$  be the amount of times that the  $i$ th attribute is sampled by Algorithm 9.6. Then we have*

$$\frac{5m_0s}{6d} \leq n_i \leq \frac{7m_0s}{6d}, \quad \forall i = 1, \dots, d$$

holds with probability at least  $1 - 2d \exp(-\frac{m_0s}{78d})$ .

*Proof.* By the Chernoff bound, for any  $0 < \theta < 1$  and  $i \in [d]$ ,  $(1 - \theta)\frac{m_0s}{d} \leq n_i \leq (1 + \theta)\frac{m_0s}{d}$  holds with probability at least  $1 - 2 \exp(-\frac{\theta^2}{2+\theta} \cdot \frac{m_0s}{d})$ . Let  $\theta = \frac{1}{6}$ , then we can obtain this lemma by the union bound.  $\square$

**Lemma 9.6.** *Let  $x_1, \dots, x_n$  be i.i.d. random variables. Suppose that  $0 \leq x_i \leq B$  holds with probability one for  $i = 1, \dots, n$ , then with probability at least  $1 - \frac{2}{d}$ ,*

$$\frac{\mathbb{E}[x_i]}{2} - \frac{5B \log d}{3n} \leq \frac{1}{n} \sum_{i=1}^n x_i \leq \frac{3\mathbb{E}[x_i]}{2} + \frac{5B \log d}{3n}.$$

*Proof.* Let  $\bar{x}$  and  $\sigma^2$  be the mean and variance of  $x_1, \dots, x_n$ , respectively. Then by the Bernstein inequality, for all  $t \geq 0$ ,

$$\mathbb{P}\left[\left|\frac{1}{n} \sum_{i=1}^n x_i - \bar{x}\right| \geq t\right] \leq 2 \exp\left(-\frac{nt^2/2}{\sigma^2 + Bt/3}\right),$$

which implies that when  $t \geq \sqrt{\frac{2\sigma^2 \log d}{n}} + \frac{2B \log d}{3n}$ , we have  $|\frac{1}{n} \sum_{i=1}^n x_i - \bar{x}| \leq t$  holds with probability at least  $1 - \frac{2}{d}$ . Note that

$$\sigma^2 = \mathbb{E}[x_i^2] - \mathbb{E}[x_i]^2 \leq \mathbb{E}\left[\frac{x_i^2}{B^2}\right] \cdot B^2 \leq \bar{x} \cdot B.$$



The last inequality holds since  $0 \leq x_i \leq B$ . Thus,

$$\sqrt{\frac{2\sigma^2 \log d}{n}} \leq \sqrt{\frac{\bar{x} \cdot 2B \log d}{n}} \leq \frac{\bar{x}}{2} + \frac{B \log d}{n}.$$

Therefore, by setting  $t = \frac{\bar{x}}{2} + \frac{5B \log d}{3n}$ , we have

$$\frac{\bar{x}}{2} - \frac{5B \log d}{3n} \leq \frac{1}{n} \sum_{i=1}^n x_i \leq \frac{3\bar{x}}{2} + \frac{5B \log d}{3n}$$

holds with probability at least  $1 - \frac{2}{d}$ . □

We now show the main proof of this theorem.

*Proof.* By Lemma 9.5, when  $m_0 \geq \frac{78d \log d}{s}$ ,

$$\frac{5m_0 s}{6d} \leq \mathbf{z}(i) \leq \frac{7m_0 s}{6d}, \quad \forall i = 1, \dots, d \quad (9.12)$$

holds with probability at least  $1 - \frac{2}{d}$ . By combining this inequality with Lemma 9.6, we know that

$$\frac{\mathbb{E}[\mathbf{x}_t(i)^2]}{2} - \frac{2Bd \log d}{m_0 s} \leq \boldsymbol{\xi}(i) \leq \frac{3\mathbb{E}[\mathbf{x}_t(i)^2]}{2} + \frac{2Bd \log d}{m_0 s}. \quad (9.13)$$

We let

$$\epsilon = \frac{8Bd \log d}{3m_0 s} \quad \text{and} \quad q_i = \frac{(\boldsymbol{\xi}(i) + \frac{3}{4}\epsilon)^{\frac{1}{3}}}{\sum_{i=1}^d (\boldsymbol{\xi}(i) + \frac{3}{4}\epsilon)^{\frac{1}{3}}}. \quad (9.14)$$

From the proof of Theorem 9.6 we know that in order to derive the theoretical performance guarantee of Algorithm 9.1, we need to find an upper bound of  $\sum_{i=1}^d \frac{\mathbb{E}[\mathbf{x}_t(i)^2]}{q_i^2}$ . By (9.13) and (9.14),

$$\begin{aligned} \sum_{i=1}^d \frac{\mathbb{E}[\mathbf{x}_t(i)^2]}{q_i^2} &\leq 2 \sum_{i=1}^d \frac{\boldsymbol{\xi}(i) + \frac{3}{4}\epsilon}{q_i^2} \\ &= 2 \sum_{i=1}^d (\boldsymbol{\xi}(i) + \frac{3}{4}\epsilon) \cdot \frac{(\sum_{i=1}^d (\boldsymbol{\xi}(i) + \frac{3}{4}\epsilon)^{\frac{1}{3}})^2}{(\boldsymbol{\xi}(i) + \frac{3}{4}\epsilon)^{\frac{2}{3}}} \\ &\leq 2 \left( \sum_{i=1}^d (\boldsymbol{\xi}(i) + \frac{3}{4}\epsilon)^{\frac{1}{3}} \right)^3 \end{aligned}$$

By Theorem 9.1 and Inequality (9.10) in the proof of Corollary 9.1, we have

$$\mathbb{E}\left[\sum_{t=1}^T \langle \mathbf{Z}_t, \mathbf{P}^* - \mathbf{P}_t \rangle\right] \leq \frac{1}{\eta} f(\mathbf{P}^*) + \frac{\eta}{2\beta} \left( \frac{16B}{s^2} \left( \sum_{i=1}^d (\xi(i) + \frac{3}{4}\epsilon)^{\frac{1}{3}} \right)^3 + 2B^2 \right).$$

Hence when

$$\eta = \sqrt{\frac{\beta f(\mathbf{P}^*)}{\left[ \frac{8B}{s^2} \left( \sum_{i=1}^d (\xi(i) + \frac{3}{4}\epsilon)^{\frac{1}{3}} \right)^3 + B^2 \right] T}},$$

the regret bound becomes

$$\begin{aligned} \mathbb{E}\left[\sum_{t=1}^T \langle \mathbf{Z}_t, \mathbf{P}^* - \mathbf{P}_t \rangle\right] &\leq 2\sqrt{\frac{f(\mathbf{P}^*) \left[ \frac{8B}{s^2} \left( \sum_{i=1}^d (\xi(i) + \frac{3}{4}\epsilon)^{\frac{1}{3}} \right)^3 + B^2 \right] T}{\beta}} \\ &\leq 2\sqrt{\frac{f(\mathbf{P}^*) \left[ \frac{12B}{s^2} \left( \sum_{i=1}^d (\mathbb{E}[\mathbf{x}_t(i)^2] + \epsilon)^{\frac{1}{3}} \right)^3 + B^2 \right] T}{\beta}}, \end{aligned}$$

where the last inequality holds since  $\xi(i) + \frac{3}{4}\epsilon \leq \frac{3}{2}(\mathbb{E}[\mathbf{x}_t(i)^2] + \epsilon)$  by (9.13). Therefore, by Lemma 9.3, we obtain Theorem 9.8.  $\square$

## 9.7 Chapter Summary

In this chapter, we proposed new online PCA algorithms that are robust to partial observation or arbitrary corruption, developed a distribution-dependent sampling scheme for estimating PCs with limited attribute observation and established their finite sample performance guarantees. The experiments empirically validated their good performance.

# CHAPTER 10

## Conclusion

This thesis investigated the methodologies for decision making and machine learning problems, e.g., regression, classification and dimensionality reduction, with uncertain or noisy data. In this chapter, we summarize the main contributions of this thesis.

Many optimization and decision making problems with stochastic uncertain parameters can be tackled via the celebrated distributionally robust chance constraint paradigm, e.g., robust classification via minimizing the worst-case misclassification probability of future samples, and transportation problem with uncertain delivery costs. In Chapter 2, we addressed an open problem of distributionally robust chance constrained problems, namely, the tractability of distributionally robust chance constraints with non-linear constraint functions. We showed that distributionally robust chance constraints are computationally tractable when the uncertainty is characterized by its mean and covariance and the constraint function is concave in the decision variables and quasi-convex in the uncertain parameters. We then established a connection between distributionally robust chance constrained optimization and robust optimization, and extended probabilistic envelope constraints into the non-linear case.

In Chapters 3-5, we explored the relationship between robust/distributionally robust optimization and two widely applied machine learning techniques – 1) Lasso-like algorithms: We showed that a wide range of Lasso-like algorithms including group Lasso and fused Lasso fit in a unified robust linear regression model. This model allows us to develop new regularization variants of Lasso-like algorithms and theoretically analyze the sparsity and consistency properties of Lasso-like algorithms from a robustness perspective; and 2) Regu-

larized SVMs: We developed a unified framework using distributionally robust optimization for designing robust classification algorithms which provides a new robustness interpretation for regularized SVMs. A new perspective on understanding the robust formulation of SVMs is presented in Chapter 5 where we proposed the novel coherent loss approach for classification which yields a strictly tighter approximation to the empirical classification error than any convex cumulative loss approach.

Both of robust optimization and distributionally robust optimization require the prior knowledge about the uncertain parameters such as the uncertainty sets that the parameters belong to and the ambiguity sets that includes the true distributions of the parameters. In Chapter 6, we studied the optimization problems where such prior knowledge is unrevealed. To the best of our knowledge, this problem has not been well explored yet. We showed that this problem is a generalization of stochastic linear optimization and contextual linear bandit problems, and proposed two algorithms LPUC-ED and LPUC-UCB in the online setting, both of which own sub-linear bounds on the regret and the constraint violation.

Dimensionality reduction methods, e.g., principal component analysis, are widely applied for preprocessing data in machine learning problems such as regression and classification. In Chapters 7-9, we addressed the issue of standard PCA that it is fragile to the existence of outlying observations. We proposed a unified framework for making not only standard PCA but also PCA-like algorithms including sparse PCA and non-negative sparse PCA robust when facing a constant fraction of arbitrarily corrupted outliers. For large-scale applications, we developed two computationally efficient non-convex outlier-robust PCA algorithms – Outlier Rejection and Outlier Reduction – that guarantee the exact recovery of the low-dimensional subspace spanned by the uncorrupted samples, and a unified framework via online mirror descent for designing online robust PCA algorithms.

## References

- [ABSS97] S. Arora, L. Babai, J. Stern, and Z. Sweedyk. The hardness of approximate optima in lattices, codes, and systems of linear equations. *Journal of Computer and System Sciences*, 54:317–331, 1997.
- [ACBF02] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multi-armed bandit problem. *Machine Learning*, 47(2):235–256, 2002.
- [ACLS12] R. Arora, A. Cotter, K. Livescu, and N. Srebro. Stochastic optimization for PCA and PLS. In *Allerton Conference*, 2012.
- [ACS13] R. Arora, A. Cotter, and N. Srebro. Stochastic optimization of PCA with capped MSG. In *NIPS*, 2013.
- [AG03] F. Alizadeh and D. Goldfarb. Second-order cone programming. *Mathematical Programming*, 95(1):3–51, 2003.
- [AG12] S. Agrawal and N. Goyal. Analysis of Thompson sampling for the multi-armed bandit problem. In *COLT*, 2012.
- [Agr95] R. Agrawal. Sample mean based index policies with  $o(\log n)$  regret for the multi-armed bandit problem. *Advances in Applied Probability*, 27(4):1054–1078, 1995.
- [AL99] N. Abe and P. M. Long. Associative reinforcement learning using linear probabilistic concepts. In *ICML*, 1999.
- [AMS09] J. Y. Audibert, R. Munos, and C. Szepesvári. Exploration-exploitation trade-off using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.
- [AN07] A. Asuncion and D. J. Newman. UCI machine learning repository, 2007.

- [APD14] M. Asteris, D. S. Papailiopoulos, and A. G. Dimakis. Nonnegative sparse PCA with provable guarantees. In *ICML*, 2014.
- [AT05] L. Hoai An and P. D. Tao. The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems. *Annals of Operations Research*, 133:23–46, 2005.
- [AYPS11] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. In *NIPS*, 2011.
- [BDD98] S. Ben-David and E. Dichterman. Learning with restricted focus of attention. *Journal of Computer and System Sciences*, 56(3):277–298, 1998.
- [BDEL03] S. Ben-David, N. Eiron, and P. M. Long. On the difficulty of approximately maximizing agreements. *Journal of Computer and System Sciences*, 66:496–513, 2003.
- [BDF13] A. Balsubramani, S. Dasgupta, and Y. Freund. The fast convergence of incremental PCA. In *NIPS*, 2013.
- [BGJ<sup>+</sup>04] C. Bhattacharyya, L. R. Grate, M. I. Jordan, L. El Ghaoui, and I. S. Mian. Robust sparse hyperplane classifiers: Application to uncertain molecular profiling data. *Journal of Computational Biology*, 11(6):1073–1089, 2004.
- [BGV92] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, pages 144–152, New York, NY, 1992.
- [Bha04] C. Bhattacharyya. Robust classification of noisy data using second order cone programming approach. In *Proceedings International Conference on Intelligent Sensing and Information Processing*, pages 433–438, Chennai, India, 2004.
- [BJNP13] A. Birnbaum, I. M. Johnstone, B. Nadler, and D. Paul. Minimax bounds for sparse PCA with noisy high-dimensional data. *The Annals of Statistics*, 41(3):1055–1084, 2013.

- [BL97] J. R. Birge and F. Louveaux. *Introduction to Stochastic Programming*. Springer-Verlag, New York, 1997.
- [BLN95] R. H. Byrd, P. Lu, and J. Nocedal. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific and Statistical Computing*, 16(5):1190–1208, 1995.
- [Bov05] Alan C. Bovik. *Handbook of Image and Video Processing (Communications, Networking and Multimedia)*. Academic Press, Inc., Orlando, FL, USA, 2005.
- [BPC<sup>+</sup>10] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2010.
- [BPS04] C. Bhattacharyya, K. S. Pannagadatta, and A. J. Smola. A second order cone programming formulation for classifying missing data. In *NIPS*, 2004.
- [Bra02] M. Brand. Incremental singular value decomposition of uncertain data with missing values. In *ECCV*, 2002.
- [BRT09] P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37:1705–1732, 2009.
- [BS73] F. Black and M. Scholes. The pricing of options and corporate liabilities. *Journal of Political Economy*, 81:637–659, 1973.
- [BS03] D. Bertsimas and M. Sim. Robust discrete optimization and network flows. *Mathematical Programming, Series B*, 98:49–71, 2003.
- [BS04] D. Bertsimas and M. Sim. The price of robustness. *Operations Research*, 52(1):35–53, January 2004.
- [BS06] D. Bertsimas and M. Sim. Tractable approximations to robust conic optimization problems. *Mathematical Programming, Serial B*, 107(1):5–36, 2006.
- [BS09] D.B. Brown and M. Sim. Satisficing measures for analysis of risky positions. *Management Science*, 55(1):71–84, 2009.

- [BSSM08] S. Bubeck, G. Stoltz, Csaba Szepesvári, and R. Munos. Online optimization in x-armed bandits. In *NIPS*, 2008.
- [BT97] Dimitris Bertsimas and John Tsitsiklis. *Introduction to Linear Optimization*. Athena Scientific, 1st edition, 1997.
- [BTBB10] A. Ben-Tal, D. Bertsimas, and D. Brown. A soft robust model for optimization under ambiguity. *Operations Research*, 58:1220–1234, 2010.
- [BTBBN11] A. Ben-Tal, S. Bhadra, C. Bhattacharyya, and J. S. Nath. Chance constrained uncertain classification via robust optimization. *Math. Program.*, 127(1):145–173, 2011.
- [BTBN06] A. Ben-Tal, S. Boyd, and A. Nemirovski. Extending scope of robust optimization: comprehensive robust counterparts of uncertain problems. *Mathematical Programming, Series B*, 107:63–89, 2006.
- [BTdHV12] A. Ben-Tal, D. den Hertog, and J. P. Vial. Deriving robust counterparts of nonlinear uncertain inequalities. *Center Discussion Paper Series No. 2012-053*, 2012.
- [BTEN09] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust Optimization*. Princeton University Press, 2009.
- [BTN98] A. Ben-Tal and A. Nemirovski. Robust convex optimization. *Mathematics of Operations Research*, 23:769–805, 1998.
- [BTN99] A. Ben-Tal and A. Nemirovski. Robust solutions of uncertain linear programs. *Operations Research Letters*, 25(1):1–13, 1999.
- [BTN00] A. Ben-Tal and A. Nemirovski. Robust solutions of linear programming problems contaminated with uncertain data. *Mathematical Programming, Serial A*, 88:411–424, 2000.
- [BTNR02] A. Ben-Tal, A. Nemirovski, and C. Roos. Robust solutions of uncertain quadratic and conic-quadratic problems. *SIAM Journal on Optimization*, 13(2):535–560, 2002.



- [BV04] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.
- [CBL06] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.
- [CBSS10] N. Cesa-Bianchi, S. Shalev-Shwartz, and O. Shamir. Efficient learning with partially observed attributes. In *ICML*, 2010.
- [CBSW14] Y. Chen, S. Bhojanapalli, S. Sanghavi, and R. Ward. Coherent matrix completion. In *ICML*, 2014.
- [CC59] A. Charnes and W. W. Cooper. Chance constrained programming. *Management Science*, 6:73–79, 1959.
- [CDL13] J. Cheng, E. Delage, and A. Lisser. Distributionally robust stochastic knapsack problem. *working draft*, 2013.
- [CDS99] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1999.
- [CE06] G. Calafiore and L. El Ghaoui. On distributionally robust chance-constrained linear programs. *Journal of Optimization Theory and Applications*, 130:1–22, 2006.
- [CFF13] C. Croux, P. Filzmoser, and H. Fritz. Robust sparse principal component analysis. *Technometrics*, 55(2):202–214, 2013.
- [CG06] G. Calafiore and L. El Ghaoui. Distributionally robust chance constrained linear programs with applications. *Journal of Optimization Theory and Applications*, 130(1):1–22, 2006.
- [CH00] C. Croux and G. Haesbroeck. Principal component analysis based on robust estimators of the covariance or correlation matrix: Influence functions and efficiencies. *BIOMETRIKA*, 87:603–618, 2000.
- [CL11] O. Chapelle and L. Li. An empirical evaluation of Thompson sampling. In *NIPS*, 2011.

- [CLMW11] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM*, 58(3):717–772, 2011.
- [CLRS11] W. Chu, L. Li, L. Reyzin, and R. E. Schapire. Contextual bandits with linear payoff functions. *Journal of Machine Learning Research*, 15:208–214, 2011.
- [Con01] R. Cont. Empirical properties of asset returns: Stylized facts and statistical issues. *Quantitative Finance*, 1(2):223–236, 2001.
- [CR09] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9:717–772, 2009.
- [CRT06] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.
- [CS02] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *J. Mach. Learn. Res.*, 2:265–292, 2002.
- [CSPW11] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.
- [CSST10] W. Chen, M. Sim, J. Sun, and C. Teo. From CVaR to uncertainty set: implications in joint chance constrained optimization. *Operations Research*, 58(2):470–485, 2010.
- [CSW12] S. S. Cheung, A. M. So, and K. Wang. Linear matrix inequalities with stochastically dependent perturbations and applications to chance-constrained semidefinite optimization. *SIAM Journal on Optimization*, 22(4):1394–1430, 2012.
- [CT10] E. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

- [CV95] C. Cortes and V. N. Vapnik. Support vector networks. *Machine Learning*, 20:1–25, 1995.
- [Das03] S. Dasgupta. Subspace detection: A robust statistics formulation. *COLT*, 2003.
- [dEJL07] A. d’Aspremont, L. El Ghaoui, M. I. Jordan, and G. R. Lanckriet. A direct formulation for sparse PCA using semidefinite programming. *SIAM review*, 49(3):434–448, 2007.
- [DGK81] S. J. Devlin, R. Gnanadesikan, and J. R. Kettenring. Robust estimation of dispersion matrices and principal components. *Journal of the American Statistical Association*, 76(374):354–362, 1981.
- [DHK08] V. Dani, T. P. Hayes, and S. M. Kakade. Stochastic linear optimization under bandit feedback. In *COLT*, 2008.
- [DKH07] V. Dani, S. Kakade, and T. Hayes. The price of bandit information for online optimization. In *NIPS*, 2007.
- [DM10] E. Delage and S. Mannor. Percentile optimization for Markov decision processes with parameter uncertainty. *Operations Research*, 58(1):203–213, 2010.
- [Don06] D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.
- [DR03] D. Dentcheva and A. Ruszczyński. Optimization with stochastic dominance constraints. *SIAM Journal on Optimization*, 14(2):548–566, 2003.
- [DR04a] D. Dentcheva and A. Ruszczyński. Optimality and duality theory for stochastic optimization problems with nonlinear dominance constraints. *Mathematical Programming, Series A*, 99(2):329–350, 2004.
- [DR04b] D. Dentcheva and A. Ruszczyński. Semi-infinite probabilistic optimization: first order stochastic dominance constraints. *Optimization*, 53(5-6):433–451, 2004.

- [Dup87] J. Dupacová. The minimax approach to stochastic programming and an illustrative application. *Stochastics*, 20:73–88, 1987.
- [DY10] E. Delage and Y. Ye. Distributionally robust optimization under moment uncertainty with applications to data-driven problems. *Operations Research*, 58(1):595–612, 2010.
- [EHJT04] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.
- [EI06] E. Erdogvan and G. Iyengar. Ambiguous chance constrained problems and robust optimization. *Mathematical Programming, Ser. B*, 107:37–61, 2006.
- [EL97] L. El Ghaoui and H. Le Bret. Robust solutions to least-squares problems with uncertain data. *SIAM Journal on Matrix Analysis and Applications*, 18:1035–1064, 1997.
- [EOL98] L. El Ghaoui, F. Oustry, and H. Le Bret. Robust solutions to uncertain semidefinite programs. *SIAM Journal on Optimization*, 9(1):33–52, 1998.
- [FA10] A. Frank and A. Asuncion. UCI machine learning repository. 2010.
- [Fam65] E. Fama. The behavior of stock prices. *Journal of Business*, 38:34–105, 1965.
- [FCGS10] S. Filippi, O. Capp, A. Garivier, and C. Szepesvari. Parametric bandits: The generalized linear case. In *NIPS*, 2010.
- [FHT00] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: A statistical view of boosting. *Annals of Statistics*, 28:337–407, 2000.
- [FHT10] J. Friedman, T. Hastie, and R. Tibshirani. A note on the group Lasso and a sparse group Lasso. Technical report, Jan 2010.
- [FN03] A. Feuer and A. Nemirovski. On sparse representation in pairs of bases. *IEEE Transactions on Information Theory*, 49(6):1579–1581, 2003.

- [FS97] Y. Freund and R. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.
- [FXMY14] J. Feng, H. Xu, S. Mannor, and S. Yan. Online PCA for contaminated data. In *NIPS*, 2014.
- [FXY12] J. Feng, H. Xu, and S. Yan. Robust PCA in high-dimension: A deterministic approach. In *ICML*, 2012.
- [GB08] M. Grant and S. Boyd. Graph implementations for nonsmooth convex programs. In *Recent Advances in Learning and Control*, pages 95–110. Springer-Verlag Limited, 2008.
- [GB11] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 1.21. <http://cvxr.com/cvx>, 2011.
- [GLS88] M. Grötschel, L. Lovász, and A. Schrijver. *Geometric Algorithms and Combinatorial Optimization*, volume 2. Springer, 1988.
- [GO13] Inc. Gurobi Optimization. Gurobi optimizer reference manual, 2013.
- [GR06] Amir Globerson and Sam Roweis. Nightmare at test time: Robust learning by feature deletion. In *ICML*, 2006.
- [Gro11] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.
- [GS10] J. Goh and M. Sim. Distributionally robust optimization and its tractable approximations. *Operations Research*, 58:902–917, 2010.
- [GVL96] Gene H. Golub and Charles F. Van Loan. *Matrix Computations (3rd Ed.)*. Johns Hopkins University Press, 1996.
- [Har13] M. Hardt. On the provable convergence of alternating minimization for matrix completion. *CoRR abs/1312.0925*, 2013.

- [HHM09] J. Huang, X. Huang, and D. N. Metaxas. Learning with dynamic group sparsity. In *ICCV*, pages 64–71, 2009.
- [HK12] E. Hazan and T. Koren. Linear regression with limited observation. In *ICML*, 2012.
- [HN99] R. Horst and N.V.Thoai. DC programming: overview. *Journal of optimization theory and applications*, 103(1):1–43, 1999.
- [HRB05] M. Hubert, P. J. Rousseeuw, and K. Branden. ROBPCA: A new approach to robust principal component analysis. *Technometrics*, 47(1):64–79, 2005.
- [HRS14] M. Hubert, T. Reynkens, and E. Schmitt. Sparse PCA for high-dimensional data with outliers. *Technical Report*, 2014.
- [HZH11] R. He, W. Zheng, and B. Hu. Maximum correntropy criterion for robust face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1561–1576, 2011.
- [HZM09] J. Huang, T. Zhang, and D. Metaxas. Learning with structured sparsity. In *ICML*, pages 417–424, 2009.
- [Jd91] D. Jansen and C. deVries. On the frequency of large stock returns: Putting booms and busts into perspective. *Review of Economics and Statistics*, 73(1):18–24, 1991.
- [JL09] I. M. Johnstone and A. Y. Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486):682–693, 2009.
- [JNS13] P. Jain, P. Netrapalli, and S. Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the Forty-fifth Annual ACM Symposium on Theory of Computing, STOC*, pages 665–674, 2013.
- [JOP<sup>+</sup>] E. Jones, T. Oliphant, P. Peterson, et al. SciPy: Open source scientific tools for Python, 2001–.

- [JOV09] L. Jacob, G. Obozinski, and J. Vert. Group Lasso with overlap and graph Lasso. In *ICML*, 2009.
- [JTU03] I. T. Jolliffe, N. T. Trendafilov, , and M. Uddin. A modified principal component technique based on the Lasso. In *JCGS*, pages 531–547, 2003.
- [JYN08] M. Journee and R. Sepulchre Y. Nesterov, Peter Richtarik. Generalized power method for sparse principal component analysis. *Journal of Machine Learning Research*, pages 517–553, 2008.
- [KKBG09] S. Kim, K. Koh, S. Boyd, and D. Gorinevsky.  $\ell_1$  trend filtering. *SIAM Review*, 51(2):339–360, 2009.
- [Kon84] S. Kon. Models of stock returns - a comparison. *Journal of Finance*, 39:147–65, 1984.
- [KP00] V. Kuleshov and D. Precup. Algorithms for the multi-armed bandit problem. *Journal of Machine Learning Research*, 1:1–48, 2000.
- [KS15] D. Kukliansky and O. Shamir. Attribute efficient linear regression with distribution-dependent sampling. In *ICML*, 2015.
- [KT11a] B. Kawas and A. Thiele. A log-robust optimization approach to portfolio management. *OR Spectrum*, 33:207–233, 2011.
- [KT11b] B. Kawas and A. Thiele. Short sales in log-robust portfolio management. *European Journal of Operations Research*, 215:651–661, 2011.
- [KW94] P. Kall and S. W. Wallace. *Stochastic programming*. John Wiley & Sons, 1994.
- [Lai87] L. T. Lai. Adaptive treatment allocation and the multi-armed bandit problem. *Annals of Statistics*, 15(3):1091–1114, 1987.
- [LCG12] R. Livni, K. Crammer, and A. Globerson. A simple geometric interpretation of svm using stochastic adversaries. In *AISTATS*, 2012.

- [LCM10] Z. Lin, M. Chen, and Y. Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv:1009.5055*, 2010.
- [LEBJ03] G. R. Lanckriet, L. El Ghaoui, C. Bhattacharyya, and M. I. Jordan. A robust minimax approach to classification. *Journal of Machine Learning Research*, 3:555–582, 2003.
- [LJB<sup>+</sup>95] Y. LeCun, L. Jackel, L. Bottou, A. Brunot, C. Cortes, J. Denker, H. Drucker, I. Guyon, U. Müller, E. Säckinger, P. Simard, and V. Vapnik. Comparison of learning algorithms for handwritten digit recognition. In *International Conference on Artificial Neural Networks*, pages 53–60, 1995.
- [LLW04] Y. Lee, Y. Lin, and G. Wahba. Multicategory support vector machines, theory, and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99:67–81, 2004.
- [LS06] Y. Liu and X. Shen. Multicategory  $\varphi$ -learning. *Journal of the American Statistical Association*, 101(474):500–509, 2006.
- [ITB03] F. De la Torre and M. J. Black. A framework for robust subspace learning. *International Journal of Computer Vision*, 54(1/2/3):117–142, 2003.
- [LW12] P. Loh and M. J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *The Annals of Statistics*, 40(3):1637–1664, 2012.
- [MCJ13] I. Mitliagkas, C. Caramanis, and P. Jain. Memory limited, streaming PCA. In *NIPS*, 2013.
- [MCJ14] I. Mitliagkas, C. Caramanis, and P. Jain. Streaming PCA with many missing entries. In *KDD*, 2014.
- [MGB08] L. Meier, S. Van De Geer, and P. Bühlmann. The group Lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical*



- Methodology*), 70(1):53–71, 2008.
- [MO60] A. W. Marshall and I. Olkin. Multivariate Chebyshev inequalities. *Annals of Mathematical Statistics*, 31(4):1001–1014, 1960.
- [MR14] A. Montanari and E. Richard. Non-negative principal component analysis: Message passing algorithms and sharp asymptotics. In *arXiv:1406.4775*, 2014.
- [MS11] S. Mannor and O. Shamir. From bandits to experts: On the value of side-observations. In *NIPS*, 2011.
- [MT13] Q. Mao and I. W. Tsang. A feature selection method for multivariate performance measures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(9):2051–2063, 2013.
- [MTA10] T. Mamadou, P. D. Tao, and L. Hoai An. A DC programming approach for sparse eigenvalue problem. In *ICML*, 2010.
- [MW65] B. Miller and H. Wagner. Chance-constrained programming with joint constraints. *Operations Research*, 13:930–945, 1965.
- [MWA05] B. Moghaddam, Y. Weiss, and S. Avidan. Spectral bounds for sparse PCA: Exact and greedy algorithms. In *NIPS*, 2005.
- [NKW13] J. Nie, W. Kotlowski, and M. K. Warmuth. Online PCA with optimal regrets. *Lecture Notes in Computer Science*, 8139:98–112, 2013.
- [NNS<sup>+</sup>14] P. Netrapalli, U. N. Niranjan, S. Sanghavi, A. Anandkumar, and P. Jain. Non-convex robust PCA. In *NIPS*, 2014.
- [NS06] A. Nemirovski and A. Shapiro. Convex approximations of chance constrained programs. *SIAM Journal on Optimization*, 17(4):969–996, 2006.
- [OCC15] F. Orabona, K. Crammer, and N. Cesa-Bianchi. A generalized online mirror descent with applications to classification and regression. *Machine Learning*, 99(3):411–435, 2015.

- [OW92] M. Overton and R. Womersley. On the sum of the largest eigenvalues of a symmetric matrix. *SIAM Journal on Matrix Analysis and Applications*, 13(1):41–45, 1992.
- [PCA07] S. Pandey, D. Chakrabarti, and D. Agarwal. Multi-armed bandit problems with dependent arms. In *ICML*, 2007.
- [PDK13] D. S. Papailiopoulos, A. G. Dimakis, and S. Korokythakis. Sparse PCA through low-rank approximations. In *ICML*, 2013.
- [Pea01] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572, 1901.
- [Per11] D. Percival. Theoretical properties of the overlapping groups Lasso. *Electronic Journal of Statistics*, 2011.
- [Pop07] I. Popescu. Robust mean-covariance solutions for stochastic optimization. *Operations Research*, 55(1):98–112, 2007.
- [Pré70] A. Prékopa. On probabilistic constrained programming. In *Proceedings of the Princeton Symposium on Mathematical Programming*, pages 113–138, 1970.
- [Pre95] A. Prekopa. *Stochastic Programming*. Kluwer, 1995.
- [PRMN04] T. Poggio, R. Rifkin, S. Mukherjee, and P. Niyogi. General conditions for predictivity in learning theory. *Nature*, 428(6981):419–422, 2004.
- [RD98] P. J. Rousseeuw and K. V. Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41:212–223, 1998.
- [Ren94] J. Renegar. Some perturbation theory for linear programming. *Mathematical Programming*, 65:73–91, 1994.
- [RF08] V. Roth and B. Fischer. The group Lasso for generalized linear models: uniqueness of solutions and efficient algorithms. In *ICML*, pages 848–855, 2008.

- [RFP10] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.
- [Rob77] S. M. Robinson. A characterization of stability in linear programming. *Operations Research*, 25:435–447, 1977.
- [Rou85] P. J. Rousseeuw. Multivariate estimation with high breakdown point. *Mathematical statistics and applications*, 8, 1985.
- [RT08] P. Rusmevichientong and J. N. Tsitsiklis. Linearly parameterized bandits. *CoRR*, 2008.
- [RZD<sup>+</sup>07] F. Rapaport, A. Zinovyev, M. Dutreix, E. Barillot, and J. Vert. Classification of microarray data using gene networks. *BMC Bioinformatics*, 8, 2007.
- [SAEK15] S. Shafieezadeh-Abadeh, P. M. Esfahani, and D. Kuhn. Distributionally robust logistic regression. In *NIPS*, 2015.
- [SBS06] P. K. Shivaswamy, C. Bhattacharyya, and A. J. Smola. Second order cone programming approaches for handling missing and uncertain data. *Journal of Machine Learning Research*, 7:1283–1314, July 2006.
- [Sca58] H. Scarf. A min-max solution of an inventory problem. In *Studies in Mathematical Theory of Inventory and Production*, pages 201–209. Stanford University Press, 1958.
- [SdM00] A. Shapiro and T. Homem de Mello. On the rate of convergence of optimal solutions of monte carlo approximations of stochastic programs. *SIAM Journal on Optimization*, 11(1):70–86, 2000.
- [SDR09] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, Philadelphia, 2009.
- [SH08] H. Shen and J. Z. Huang. Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, 99:1015–1034, 2008.

- [Sha13] O. Shamir. On the complexity of bandit and derivative-free stochastic convex optimization. In *COLT*, 2013.
- [Sha15a] O. Shamir. On the complexity of bandit linear optimization. In *COLT*, 2015.
- [Sha15b] O. Shamir. A stochastic PCA and SVD algorithm with an exponential convergence rate. In *ICML*, 2015.
- [SS99] E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 37:297–336, 1999.
- [SS02] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, 2002.
- [SS12] S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2), 2012.
- [SST11] N. Srebro, K. Sridharan, and A. Tewari. On the universality of online mirror descent. In *NIPS*, 2011.
- [Sun96] K. Sung. Learning and example selection for object and pattern recognition. *PhD thesis, MIT*, 1996.
- [TA77] A. N. Tikhonov and V. Y. Arsenin. *Solutions of Ill-Posed Problems*. V. H. Winston & Sons, Washington, D.C.: John Wiley & Sons, New York,, 1977.
- [TDT10] M. Thiao, T. P. Dinh, and H. A. Thi. A DC programming approach for sparse eigenvalue problem. In *ICML*, 2010.
- [TG07] T. Trafalis and R. Gilbert. Robust support vector machines for classification and computational issues. *Optimization Methods and Software*, 22(1):187–198, February 2007.
- [Tib96] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.
- [Tro04] J. A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004.

- [Tro06] J. A. Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Transactions on Information Theory*, 51(3):1030–1051, 2006.
- [TSR<sup>+</sup>05] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight. Sparsity and smoothness via the fused Lasso. *Journal of the Royal Statistical Society Series B*, pages 91–108, 2005.
- [TT11] R. J. Tibshirani and J. Taylor. The solution path of the generalized Lasso. *The Annals of Statistics*, 39(3), 2011.
- [VC91] V. N. Vapnik and A. Chervonenkis. The necessary and sufficient conditions for consistency in the empirical risk minimization method. *Pattern Recognition and Image Analysis*, 1(3):260–284, 1991.
- [VCLR13] V. Q. Vu, J. Cho, J. Lei, and K. Robe. Fantope projection and selection: A near-optimal convex relaxation of sparse PCA. In *NIPS*, 2013.
- [vdMCTW13] Laurens van der Maaten, Minmin Chen, Stephen Tyree, and Kilian Q. Weinberger. Learning with marginalized corrupted features. In *ICML*, 2013.
- [Ver12] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing: Theory and Applications*, 2012.
- [VL63] V. N. Vapnik and A. Lerner. Pattern recognition using generalized portrait method. *Automation and Remote Control*, 24:744–780, 1963.
- [VL13] V. Q. Vu and Jing Lei. Minimax bounds for sparse PCA with noisy high-dimensional data. *The Annals of Statistics*, 41(6):2703–3110, 2013.
- [Wai09] M. Wainwright. Sharp thresholds for noisy and high-dimensional recovery of sparsity using  $\ell_1$ -constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, 55:2183–2202, 2009.
- [WC12] L. Wang and H. Cheng. Robust sparse PCA via weighted elastic net. In *Pattern Recognition*, pages 88–95. Springer, 2012.

- [WK08] M. K. Warmuth and D. Kuzmin. Randomized online pca algorithms with regret bounds that are logarithmic in the dimension. *Journal of Machine Learning Research*, 9:2287–2320, 2008.
- [WKS13] W. Wiesemann, D. Kuhn, and M. Sim. Distributionally robust convex optimization. 2013.
- [WPM<sup>+</sup>09] J. Wright, Y. Peng, Y. Ma, A. Ganesh, and S. Rao. Robust principal component analysis: Exact recovery of corrupted low-rank matrices by convex optimization. In *NIPS*, 2009.
- [XCM09] H. Xu, C. Caramanis, and S. Mannor. Robustness and regularization of support vector machines. *Journal of Machine Learning Research*, 10:1485–1510, 2009.
- [XCM10] H. Xu, C. Caramanis, and S. Mannor. Robust regression and Lasso. *IEEE Transactions on Information Theory*, 56(7):3561–3574, 2010.
- [XCM12] H. Xu, C. Caramanis, and S. Mannor. Optimization under probabilistic envelope constraints. *Operations Research*, 60:682–700, 2012.
- [XCM13] H. Xu, C. Caramanis, and S. Mannor. Outlier-robust PCA: the high-dimensional case. *IEEE Transactions on Information Theory*, 59(1):546–572, 2013.
- [XCS12] H. Xu, C. Caramanis, and S. Sanghavi. Robust PCA via outlier pursuit. *IEEE Transactions on Information Theory*, 58(5):3047–3064, 2012.
- [XY95] L. Xu and A. L. Yuille. Robust principal component analysis by self-organizing rules based on statistical physics approach. *IEEE Transactions on Neural Networks*, 6(1):131–143, 1995.
- [YL06] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68:49–67, 2006.

- [YW99] T. N. Yang and S. D. Wang. Robust algorithms for principal component analysis. *Pattern Recognition Letters*, 20(9):927–933, 1999.
- [YX15a] W. Yang and H. Xu. Streaming sparse principal component analysis. In *ICML*, 2015.
- [YX15b] W. Yang and H. Xu. A unified framework for outlier-robust PCA-like algorithms. In *ICML*, 2015.
- [ZE11] Y. Zhang and L. El Ghaoui. Large-scale sparse principal component analysis with application to text data. In *NIPS*, 2011.
- [ZH05] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.
- [Zha04] T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32:56–85, 2004.
- [Zha09] T. Zhang. Some sharp performance bounds for least squares regression with l1 regularization. *Annals of Statistics*, 37:2109–2144, 2009.
- [ZHT06] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. In *JCGS*, pages 265–286, 2006.
- [ZKR11] S. Zymler, D. Kuhn, and B. Rustem. Distributionally robust joint chance constraints with second-order moment information. *Mathematical Programming*, 2011.
- [ZKR12] S. Zymler, D. Kuhn, and B. Rustem. Worst-case value at risk of nonlinear portfolios. *Management Science*, 2012.
- [ZWL15] T. Zhao, Z. Wang, and H. Liu. A nonconvex optimization framework for low rank matrix estimation. In *NIPS*, 2015.