2017

Inference And Learning: Computational Difficulty And Efficiency

Tengyuan Liang
University of Pennsylvania, tengyuan@wharton.upenn.edu

Inference And Learning: Computational Difficulty And Efficiency

Abstract

In this thesis, we mainly investigate two collections of problems: statistical network inference and model selection in regression. The common feature shared by these two types of problems is that they typically exhibit an interesting phenomenon in terms of computational difficulty and efficiency.

For statistical network inference, our goal is to infer the network structure based on a noisy observation of the network. Statistically, we model the network as generated from the structural information with the presence of noise, for example, planted submatrix model (for bipartite weighted graph), stochastic block model, and Watts-Strogatz model. As the relative amount of ``signal-to-noise'' varies, the problems exhibit different stages of computational difficulty. On the theoretical side, we investigate these stages through characterizing the transition thresholds on the ``signal-to-noise'' ratio, for the aforementioned models. On the methodological side, we provide new computationally efficient procedures to reconstruct the network structure for each model.

For model selection in regression, our goal is to learn a ``good'' model based on a certain model class from the observed data sequences (feature and response pairs), when the model can be misspecified. More concretely, we study two model selection problems: to learn from general classes of functions based on i.i.d. data with minimal assumptions, and to select from the sparse linear model class based on possibly adversarially chosen data in a sequential fashion. We develop new theoretical and algorithmic tools beyond empirical risk minimization to study these problems from a learning theory point of view.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Statistics

First Advisor
Tony T. Cai

Second Advisor
Alexander Rakhlin

Subject Categories
Computer Sciences | Statistics and Probability

INFERENCE AND LEARNING: COMPUTATIONAL DIFFICULTY AND EFFICIENCY

Tengyuan Liang

A DISSERTATION

in

Statistics

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2017

Supervisor of Dissertation	Co-Supervisor of Dissertation
T. Tony Cai Dorothy Silberberg Professor; Professor of Statistics	Alexander Rakhlin Associate Professor of Statistics, Computer and Information Science

Graduate Group Chairperson

Catherine Schrand
Celia Z. Moh Professor;
Professor of Accounting

Dissertation Committee

Mark G. Low, Walter C. Bladstrom Professor; Professor of Statistics

Elchanan Mossel, Professor of Mathematics, Massachusetts Institute of Technology

INFERENCE AND LEARNING: COMPUTATIONAL DIFFICULTY AND EFFICIENCY

Dedicated to my parents.

ACKNOWLEDGEMENT (optional)

I would like to thank Tony Cai, for being an excellent advisor. Tony offered kind support and insightful advice during my Ph.D. studies. Tony is very generous about his time spent with students. I am fortunate to come to Penn to work with him. More importantly, I learned from him how to be a productive researcher.

I would like to thank Alexander (Sasha) Rakhlin, for being an excellent advisor as well. It was always enjoyable to discuss problems with Sasha and to hear his unique and insightful feedback. I am indebted to him for showing me many interesting aspects about academia, and for introducing me to the larger research community.

I would like to thank Mark Low for his dedication to nurturing Ph.D. students. Mark cares deeply about the overall development of the young generation. I would also like to thank Mark for bringing me to nice recitals in Kimmel center and Curtis School.

In addition, I would like to thank Elchanan Mossel for stimulating discussions during his stay at Penn. It was always very interesting to hear his insightful and sharp questions during the seminar talk.

I am grateful to Larry Brown, Abba Krieger, Ed George, Mike Steele, Nancy Zhang and Dylan Small for their encouragement.

In the end, I would like to thank my fellow students and postdocs at Penn — Wenxin Zhou, Yupeng Chen, Yang Jiang, Yin Xia, Xiaodong Li, Jiaming Xu, Weijie Su, Daniel McCarthy, Matt Olsen, Colin Fogarty, Veronika Rockova, Hyunseung Kang, Sameer Deshpande, Justin Khim, Min Xu, Yuancheng Zhu, Yiran Chen, Shi Gu, Xiang Fang, Yang Liu and Xingtan Zhang. I value the friendship I carried on with Kyle Luh, Zhengying Liu, Shuting Lu and Gongchen Sun, whom I met during undergraduate years.

ABSTRACT

INFERENCE AND LEARNING: COMPUTATIONAL DIFFICULTY AND EFFICIENCY

Tengyuan Liang

T. Tony Cai

Alexander Rakhlin

In this thesis, we mainly investigate two collections of problems: statistical network inference and model selection in regression. The common feature shared by these two types of problems is that they typically exhibit an interesting phenomenon in terms of computational difficulty and efficiency. For statistical network inference, our goal is to infer the network structure based on a noisy observation of the network. Statistically, we model the network as generated from the structural information with the presence of noise, for example, planted submatrix model (for bipartite weighted graph), stochastic block model, and Watts-Strogatz model. As the relative amount of "signal-to-noise" varies, the problems exhibit different stages of computational difficulty. On the theoretical side, we investigate these stages through characterizing the transition thresholds on the "signal-to-noise" ratio, for the aforementioned models. On the methodological side, we provide new computationally efficient procedures to reconstruct the network structure for each model. For model selection in regression, our goal is to learn a "good" model based on a certain model class from the observed data sequences (feature and response pairs), when the model can be misspecified. More concretely, we study two model selection problems: to learn from general classes of functions based on i.i.d. data with minimal assumptions, and to select from the sparse linear model class based on possibly adversarially chosen data in a sequential fashion. We develop new theoretical and algorithmic tools beyond empirical risk minimization to study these problems from a learning theory point of view.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1 : Introduction

The unprecedented “data deluge” emerging in science and engineering in recent years poses many challenges to the development of theory and methodology. Specifically, part of the new features of these challenges can be summarized as follows:

- **Inference** How to make inference on a large amount of unknown variables or parameters with limited amount of partial, noisy, and indirect measurements (observations)?
- **Learning** How to predict as well as a class of complex models (non-linear and non-convex), using heterogeneous data that can be arbitrary or even adversarial, in various specific protocols (online, bandit feedback, partial information)?
- **Computation** Computation adds a new dimension to this modern challenge. As the scale of the data increases significantly, how to take computational and memory budgets into account when designing inference and learning algorithms?

The goal of this thesis is to investigate these features for two collections of problems: statistical network inference, and model selection in regression. We will start with two concrete data examples as a gentle introduction to motivate our studies, then we will move to the outline of the remaining chapters before stepping into detailed discussions.

Structural inference of C. elegans neuron network In this paragraph, we will use the C. elegans neuron network (Pavlovic et al., 2014) as an example to illustrate the importance of structural inference for complex networks, which is the focus of Chapter 2. C. elegans is one of the simplest organisms with a nervous system whose neuronal “wiring diagram” (neural connections) has been completed (with around 300 neurons and 2000 edges). In practice, it is important to understand the structural information of the complex systems/networks based on the connection patterns. From the statistical side, researchers often model the network as generated from some statistical model (for example, stochastic block models, and the Watts-Strogatz small world model), with the goal of inferring the structure from the data. Brute-force maximum likelihood based approach is usually computationally prohibitive with large scale networks. Therefore, it is important to understand when we can solve the inference problem in a computationally efficient manner. Here we illustrate some of the methods proposed in Chapter 2 on this C.elegans network.

We apply the algorithm proposed in Chapter 2.1 on discovering the community structure on this dataset. Figure 1 summarizes the community structure we discover.

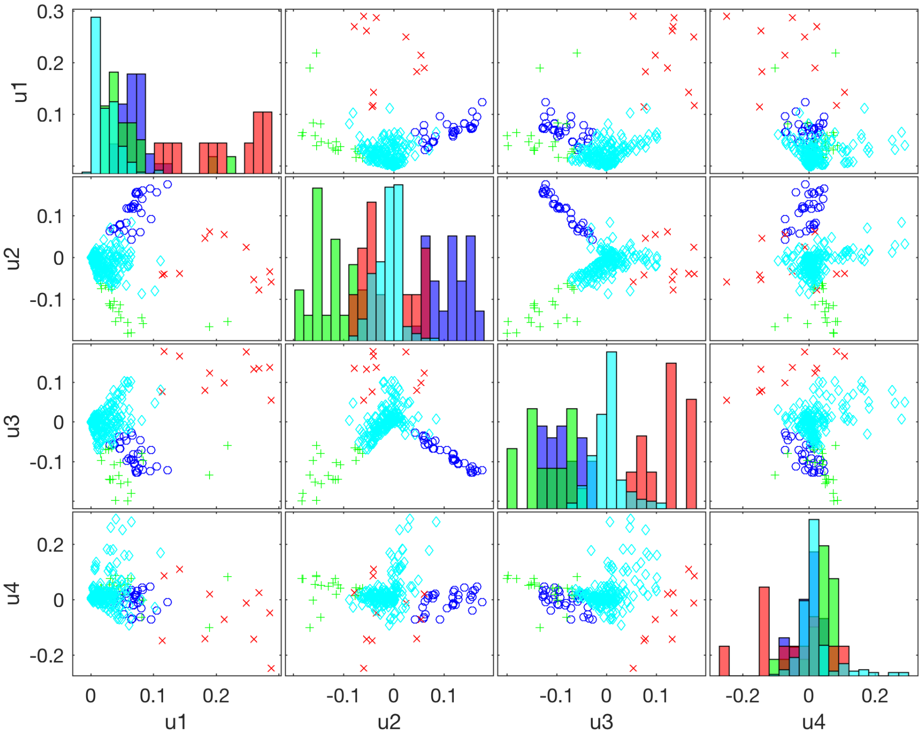

Figure 1: Plot of the community structure in the space spanned by the top four left singular vectors. Here each color represents a community and $u_i, 1 \leq i \leq 4$ denotes the singular vectors respectively. For instance, the top right subfigure denotes projecting the nodes onto the space spanned by (u_4, u_1).

We also apply the algorithm proposed in Chapter 2.3 on reconstructing the ring lattice structure on this dataset. Figure 2 describes the result. On the left, we solve for the ring embedding layout by the algorithm proposed in Chapter 2.3, and plot the connection among the nodes. On the right, we randomly permute the nodes as the circle layout (what one expects to see without the ring lattice structure). As a contrast, one can clearly see the edges are much better organized in the left compared to the right, suggesting the presence of Watts-Strogatz ring lattice structure.

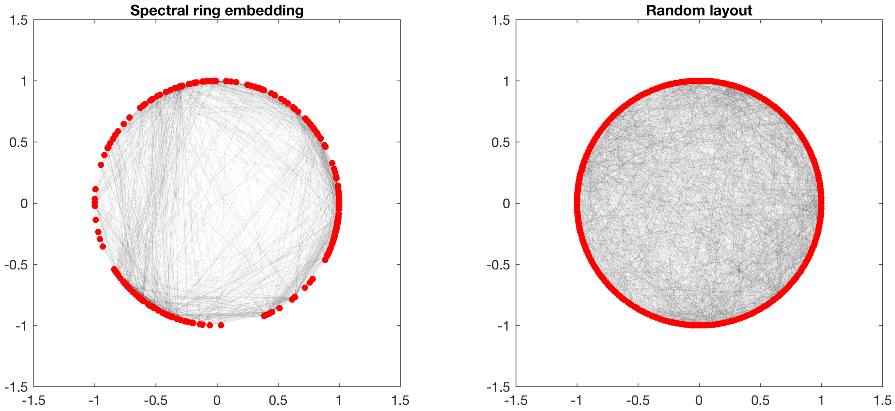

Figure 2: Plot of the ring lattice structure. On the left, the layout is returned by our spectral embedding algorithm; on the right, the layout is returned by a random permutation.

Model selection for fitting neuronal data In this paragraph, we will employ simulated neuronal data (Kaufman et al., 2005) to motivate the focus of Chapter 3 — model selection in regression. The task is to predict the firing rate of neurons based on the simulated data as good as the best one among a collection of models. Specifically, in Figure 3, we fitted three models, cubic smoothing spline, polynomial regression of degree 4, and Bayesian adaptive regression splines (BARS). Traditionally, model selection is typically addressed using cross-validation (CV) in a general way. However, it is questionable whether CV will provide the optimal behavior, let alone the computational burden when facing a large collection of models.

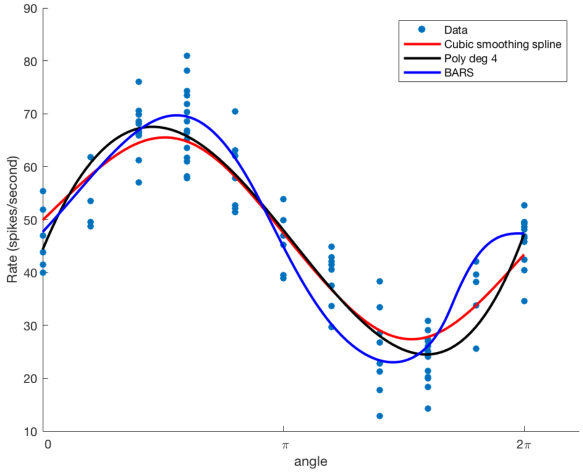

Figure 3: Simulated neuron data with three fitting models: cubic smoothing spline, polynomial regression of degree 4, and Bayesian adaptive regression splines (BARS).

We apply the two-step Star algorithm proposed in Chapter 3.1 on the neuronal dataset, and summarize the result in Figure 4. Here each data point corresponds to one experiment, whose x-axis denotes the mean square error (MSE) of the cross-validation, and y-axis denotes the MSE of the Star algorithm. As one can see, clearly the Star algorithm outperforms the CV significantly over many experiments.

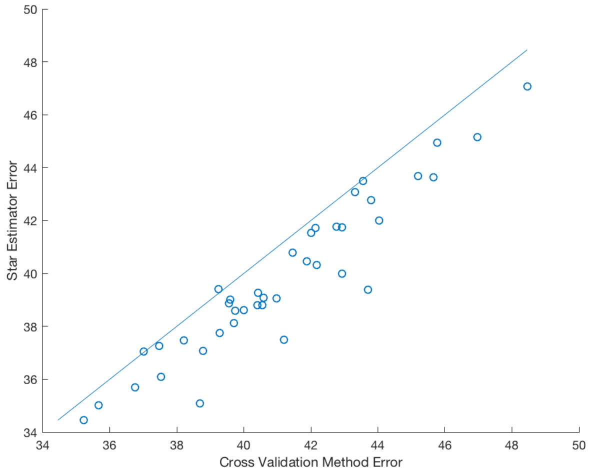

Figure 4: Plot of MSE of Star algorithm compared to cross validation.

1.1. Outline

1.1.1. Chapter 2

Historically, statisticians and information theorists mostly focus on analyzing the threshold that separates "signal" and "noise" over random instances (in the average-case sense), overlooking computation complexity. However, computer scientists are usually concerned with quantifying problems according to their computational difficulty in the worst-case sense. In this chapter, we will investigate the intersection of the above perspectives. For an average instance, we would like to: (a) identify the statistical threshold that describes the solvability of the problem information-theoretically, and (b) quantify the computational threshold that sheds light on the difficulty for polynomial-time algorithms (inside the statistically solvable phase).

In Cai et al. (2015a), we examined the above two thresholds on the problem of submatrix localization with background noise. We discovered that, quite surprisingly, there is always an intrinsic gap between computational and statistical thresholds under standard computational hardness assumption. There is a non-vanishing phase quantifying the price to pay for pursuing polynomial run-time. We established the computational optimality in two stages: (a) we provided a new average-case reduction to the hidden clique problem; (b) we proposed a simple near-linear time algorithm that achieves the computational threshold. Overall, this work illustrates that for certain statistical problems, there are more structured phases inside the statistically-solvable phase.

Motivated by the fact that real network datasets always contain side information (partial labels, nodes' features) in addition to the connections, our work Cai et al. (2016c) studied the computational difficulty of partially-labeled stochastic block models when a vanishing portion of true labels are revealed. One the one hand, we derived and analyzed a new local algorithm — linearized message passing — that achieves exponential decaying error for node inference down to the well-known Kesten-Stigum (K-S) threshold. One the other hand, we proved that this K-S threshold is indeed the barrier for all local algorithms (heuristically believed to be powerful polynomial time algorithms) through a minimax lower bound. Whether the gap between K-S and information-theoretic threshold (with growing number of blocks) is inevitable for polynomial time algorithms, remains an open problem.

In Cai et al. (2016a), we initiated the investigation of the corresponding thresholds for detection and structural reconstruction in Watts-Strogatz (W-S) small world networks. The W-S model with neighborhood size k and rewiring probability β can be viewed as a continuous interpolation between a deterministic ring lattice graph and the Erdős-Rényi random graph. We studied both the computational and statistical aspects of detecting the deterministic ring lattice structure (or local geographical links) in the presence of random connections (or long range links), and for its recovery. We partitioned parameter space (k, β) into several phases according to the difficulty of the problem, and proposed distinct methods that mathematically achieve the corresponding thresholds separating the phases. We implemented our spectral ring embedding algorithm on the Les Misérables co-appearance network.

1.1.2. Chapter 3

In this chapter we will study regression and model selection problem focusing on two aspects: (a) model misspecification; (b) model class can be non-convex. For the first point, classic decision theory is concerned with making decisions from i.i.d. data generated from a well-specified statistical model. However, one should be agnostic about these two assumptions: (a) the i.i.d. data may be generated from a mis-specified model; (b) the underlying stochastic process generating the data is non-i.i.d., even adversarially chosen by oblivious nature. Statistical learning theory and online learning provide handful tools to solve these two problems respectively. For the second point, when the model class is non-convex (such as sparse linear regression and finite aggregation), we would like to understand its consequences on the estimation/prediction procedure and accuracy, as well as the computation difficulty.

In Liang et al. (2015), we revisited the regression problem under square loss in the statistical learning setting (i.i.d., mis-specified model), for general classes of functions that can be unbounded and non-convex. We introduced a new notion of offset Rademacher complexity, and showed that the excess loss can be upper bounded by this new complexity through a novel geometric inequality. We achieved this goal through: (a) proposing a novel two-step estimator; (b) adopting the symmetrization and chaining tools in empirical processes theory to this offset complexity. We showed that localization for unbounded class is automatic through this offset analysis. This new framework recovers the sharp rates in parametric regression, finite aggregation, and non-parametric regression simultaneously.

Cai et al. (2016b) presents a unified geometric framework for the statistical analysis of a general ill-posed linear inverse model which includes as special cases noisy compressed sensing, sign vector recovery, trace regression, orthogonal matrix estimation, and noisy matrix completion. We propose computationally feasible convex programs for statistical inference including estimation, confidence intervals and hypothesis testing. A theoretical framework is developed to characterize the local estimation rate of convergence and to provide statistical inference guarantees. Our results are built based on the local conic geometry and duality. The difficulty of statistical inference is captured by the geometric characterization of the local tangent cone through the Gaussian width and Sudakov estimate.

Online sparse linear regression is an online problem where an algorithm repeatedly chooses a subset of coordinates to observe in an adversarially chosen feature vector, makes a real-valued prediction, receives the true label, and incurs the squared loss. The goal is to design an online learning algorithm with sublinear regret to the best sparse linear predictor in hindsight. Without any assumptions, this problem is known to be computationally intractable. In Kale et al. (2017), we make the assumption that data matrix satisfies restricted isometry property, and show that this assumption leads to computationally efficient algorithms with sublinear regret for two variants of the problem. In the first variant, the true label is generated according to a sparse linear model with additive Gaussian noise. In the second, the true label is chosen adversarially.

CHAPTER 2 : Statistical Network Inference and Computation

2.1. Submatrix Localization and Bi-Clustering

2.1.1. Introduction

The "signal + noise" model

$$X = M + Z, \tag{2.1}$$

where M is the signal of interest and Z is noise, is ubiquitous in statistics and is used in a wide range of applications. Such a "signal + noise" model has been well studied in statistics in a number of settings, including nonparametric regression where M is a function, and the Gaussian sequence model where M is a finite or an infinite dimensional vector. See, for example, Tsybakov (2009); Johnstone (2013) and the references therein. In nonparametric regression, the structural knowledge on M is typically characterized by smoothness, and in the sequence model the structural knowledge on M is often described by sparsity. Fundamental statistical properties such as the minimax estimation rates and the signal detection boundaries have been established under these structural assumptions.

For a range of contemporary applications in statistical learning and signal processing, M and Z in the "signal + noise" model (2.1) are high-dimensional matrices (Tufts and Shah, 1993; Drineas et al., 2006; Donoho and Gavish, 2014; Chandrasekaran et al., 2009; Candès et al., 2011). In this setting, many new interesting problems arise under a variety of structural assumptions on M and the distribution of Z. Examples include sparse principal component analysis (PCA) (Vu and Lei, 2012; Berthet and Rigollet, 2013b; Birnbaum et al., 2013; Cai et al., 2013, 2015b), low-rank matrix de-noising (Donoho and Gavish, 2014), matrix factorization and decomposition (Chandrasekaran et al., 2009; Candès et al., 2011; Agarwal et al., 2012), non-negative PCA (Zass and Shashua, 2006; Montanari and Richard, 2014), submatrix detection and localization (Butucea and Ingster, 2013; Butucea et al., 2013), synchronization and planted partition (Javanmard et al., 2015; Decelle et al., 2011), among many others. In the conventional statistical framework, the goal is developing optimal statistical procedures (for estimation, testing, etc), where optimality is understood with respect to the sample size and parameter space.

When the dimensionality of the data becomes large as in many contemporary applications, the computational concerns associated with the statistical procedures come to the forefront. After all, statistical methods are useful in practice only if they can be computed within a reasonable amount of time. A fundamental question is: Is there a price to pay for statistical performance if one only considers computable (polynomial-time) procedures? This question is particularly relevant for non-convex problems with combinatorial structures. These problems pose a significant computational challenge because naive methods based on exhaustive search are typically not computationally efficient. Trade-off between computational efficiency and statistical accuracy in high-dimensional inference has drawn increasing attention in the literature. In particular, Chandrasekaran et al. (2012) and Wainwright (2014) considered a general class of linear inverse problems, with different emphasis on geometry of convex relaxation and decomposition of statistical and computational errors. Chandrasekaran and Jordan (2013) studied an approach for trading off computational demands with statistical accuracy via relaxation hierarchies. Berthet and Rigollet (2013a); Ma and Wu (2013a); Zhang et al. (2014b) focused on computational difficulties for various statistical problems, such as detection and regression.

In the present thesis, we study the interplay between computational efficiency and statistical accuracy in submatrix localization based on a noisy observation of a large matrix. The problem considered in this thesis is formalized as follows.

Problem Formulation

Consider the matrix X of the form

$$X = M + Z, \quad \text{where} \quad M = \lambda \cdot 1_{R_m} 1_{C_n}^T \tag{2.2}$$

and $1_{R_m} \in \mathbb{R}^m$ denotes a binary vector with 1 on the index set R_m and zero otherwise. Here, the entries Z_{ij} of the noise matrix are i.i.d. zero-mean sub-Gaussian random variables with parameter σ (defined formally in Equation (2.5)). Given the parameters $m, n, k_m, k_n, \lambda/\sigma$, the set of all distributions described above – for all possible choices of R_m and C_n – forms the submatrix model $\mathcal{M}(m, n, k_m, k_n, \lambda/\sigma)$.

This model can be further extended to multiple submatrices where

$$M = \sum_{s=1}^{r} \lambda_s \cdot 1_{R_s} 1_{C_s}^T \tag{2.3}$$

where $|R_s| = k_s^{(m)}$ and $|C_s| = k_s^{(n)}$ denote the support set of the s-th submatrix. For simplicity, we first focus on the single submatrix and then extend the analysis to the model (2.3) in Section 2.1.2.

There are two fundamental questions associated with the submatrix model (2.2). One is the *detection* problem: given one observation of the X matrix, decide whether it is generated from a distribution in the submatrix model or from the pure noise model. Precisely, the detection problem considers testing of the hypotheses

$$H_0 : M = \mathbf{0} \quad \text{v.s.} \quad H_\alpha : M \in \mathcal{M}(m, n, k_m, k_n, \lambda/\sigma).$$

The other is the *localization* problem, where the goal is to exactly recover the signal index sets R_m and C_n (the support of the mean matrix M). It is clear that the localization problem is at least as hard (both computationally and statistically) as the detection problem. The focus of the current thesis is on the *localization* problem. As we will show in this thesis, the localization problem requires larger signal to noise ratio, as well as novel algorithm and analysis to exploit the submatrix structure.

Main Results

To state our main results, let us first define a hierarchy of algorithms in terms of their worst-case running time on instances of the submatrix localization problem:

$$\mathsf{LinAlg} \subset \mathsf{PolyAlg} \subset \mathsf{ExpoAlg} \subset \mathsf{AllAlg}.$$

The set LinAlg contains algorithms $\mathcal{A}$ that produce an answer (in our case, the localization subset $\hat{R}_m^{\mathcal{A}}, \hat{C}_n^{\mathcal{A}}$) in time linear in $m \times n$ (the minimal computation required to read the matrix). The classes $\mathsf{PolyAlg}$ and $\mathsf{ExpoAlg}$ of algorithms, respectively, terminate in polynomial and exponential time, while AllAlg has no restriction.

Combining Theorem 2.1.3 and 2.1.4 in Section 2.1.2 and Theorem 2.1.5 in Section 2.1.3, the statistical and computational boundaries for submatrix localization can be summarized as follows. The notations $\gtrsim, \lesssim, \asymp$ are formally defined in Section 2.1.1.

Theorem 2.1.1 (Computational and Statistical Boundaries)**.** *Consider the submatrix localization problem under the model* (2.2)*. The computational boundary* $\mathsf{SNR_c}$ *for the dense case when* $\min\{k_m, k_n\} \gtrsim \max\{m^{1/2}, n^{1/2}\}$ *is*

$$\mathsf{SNR_c} \asymp \sqrt{\frac{m \vee n}{k_m k_n}} + \sqrt{\frac{\log n}{k_m} \vee \frac{\log m}{k_n}},$$

in the sense that

$$\begin{aligned}
&\varlimsup_{m,n,k_m,k_n \to \infty} \inf_{\mathcal{A} \in \mathsf{LinAlg}} \sup_{M \in \mathcal{M}} \mathbb{P}\left(\hat{R}_m^{\mathcal{A}} \neq R_m \text{ or } \hat{C}_n^{\mathcal{A}} \neq C_n\right) = 0, \qquad \text{if } \frac{\lambda}{\sigma} \gtrsim \mathsf{SNR_c} \\
&\varliminf_{m,n,k_m,k_n \to \infty} \inf_{\mathcal{A} \in \mathsf{PolyAlg}} \sup_{M \in \mathcal{M}} \mathbb{P}\left(\hat{R}_m^{\mathcal{A}} \neq R_m \text{ or } \hat{C}_n^{\mathcal{A}} \neq C_n\right) > 0, \qquad \text{if } \frac{\lambda}{\sigma} \lesssim \mathsf{SNR_c}
\end{aligned} \tag{2.4}$$

where (2.4) *holds under the Hidden Clique hypothesis* $\mathsf{HC_l}$ *(see Section 2.1.2). For the sparse case when* $\max\{k_m, k_n\} \precsim \min\{m^{1/2}, n^{1/2}\}$*, the computational boundary is* $\mathsf{SNR_c} = \Theta^*(1)$*, more precisely*

$$1 \precsim \mathsf{SNR_c} \precsim \sqrt{\log \frac{m \vee n}{k_m k_n}}.$$

The statistical boundary $\mathsf{SNR_s}$ *is*

$$\mathsf{SNR_s} \asymp \sqrt{\frac{\log n}{k_m} \vee \frac{\log m}{k_n}},$$

in the sense that

$$\begin{aligned} &\overline{\lim_{m,n,k_m,k_n \to \infty}} \inf_{\mathcal{A} \in \mathsf{ExpoAlg}} \sup_{M \in \mathcal{M}} \mathbb{P}\left(\hat{R}_m^{\mathcal{A}} \neq R_m \ \textit{or}\ \hat{C}_n^{\mathcal{A}} \neq C_n\right) = 0, \qquad \textit{if}\ \frac{\lambda}{\sigma} \succsim \mathsf{SNR_s} \\ &\underline{\lim_{m,n,k_m,k_n \to \infty}} \inf_{\mathcal{A} \in \mathsf{AllAlg}} \sup_{M \in \mathcal{M}} \mathbb{P}\left(\hat{R}_m^{\mathcal{A}} \neq R_m \ \textit{or}\ \hat{C}_n^{\mathcal{A}} \neq C_n\right) > 0, \qquad \textit{if}\ \frac{\lambda}{\sigma} \precsim \mathsf{SNR_s} \end{aligned}$$

under the minimal assumption $\max\{k_m, k_n\} \precsim \min\{m, n\}$.

If we parametrize the submatrix model as $m = n, k_m \asymp k_n \asymp k = \Theta^*(n^\alpha), \lambda/\sigma = \Theta^*(n^{-\beta})$, for some $0 < \alpha, \beta < 1$, we can summarize the results of Theorem 2.1.1 in a phase diagram, as illustrated in Figure 5.

To explain the diagram, consider the following cases. First, the statistical boundary is

$$\sqrt{\frac{\log n}{k_m} \vee \frac{\log m}{k_n}},$$

which gives the line separating the red and the blue regions. For the dense regime $\alpha \geq 1/2$, the computational boundary given by Theorem 2.1.1 is

$$\sqrt{\frac{m \vee n}{k_m k_n}} + \sqrt{\frac{\log n}{k_m} \vee \frac{\log m}{k_n}},$$

which corresponds to the line separating the blue and the green regions. For the sparse

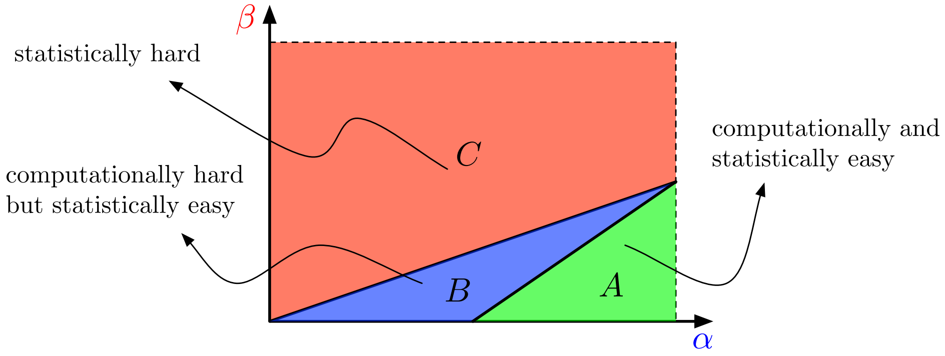

Figure 5: Phase diagram for submatrix localization. Red region (C): statistically impossible, where even without computational budget, the problem is hard. Blue region (B): statistically possible but computationally expensive (under the hidden clique hypothesis), where the problem is hard to all polynomial time algorithm but easy with exponential time algorithm. Green region (A): statistically possible and computationally easy, where a fast polynomial time algorithm will solve the problem.

regime $\alpha < 1/2$, the computational boundary is $\Theta(1) \precsim \mathsf{SNR_c} \precsim \Theta(\sqrt{\log \frac{m \vee n}{k_m k_n}})$, which is the horizontal line connecting $(\alpha = 0, \beta = 0)$ to $(\alpha = 1/2, \beta = 0)$.

As a key part of Theorem 2.1.1, we provide linear time spectral algorithm that will succeed in localizing the submatrix with high probability in the regime above the computational threshold. Furthermore, the method is data-driven and adaptive: it does not require the prior knowledge on the size of the submatrix. This should be contrasted with the method of Chen and Xu (2014) which requires the prior knowledge of k_m, k_n; furthermore, the running time of their SDP-based method is superlinear in nm. Under the hidden clique hypothesis, we prove that below the computational threshold there is no polynomial time algorithm that can succeed in localizing the submatrix. We remark that the computational lower bound for *localization* requires distinct new techniques compared to the lower bound for *detection*; the latter has been resolved in Ma and Wu (2013a).

Beyond localization of one single submatrix, we generalize both the computational and statistical story to a growing number of submatrices in Section 2.1.2. As mentioned earlier, the statistical boundary for one single submatrix localization has been investigated by Butucea et al. (2013) in the Gaussian case. Our result focuses on the computational intrinsic difficulty of localization for a growing number of submatrices, at the expense of not providing the exact constants for the thresholds.

The phase transition diagram in Figure 5 for localization should be contrasted with the corresponding result for detection, as shown in (Butucea and Ingster, 2013; Ma and Wu, 2013a). For a large enough submatrix size (as quantified by $\alpha > 2/3$), the computationally-intractable-but-statistically-possible region collapses for the detection problem, but not for localization. In plain words, detecting the presence of a large submatrix becomes both computationally and statistically easy beyond a certain size, while for localization there is always a gap between statistically possible and computationally feasible regions. This phenomenon also appears to be distinct to that of other problems like estimation of sparse principal components (Cai et al., 2013), where computational and statistical easiness coincide with each other over a large region of the parameter spaces.

Prior Work

There is a growing body of work in statistical literature on submatrix problems. Arias-Castro et al. (2011) studied the detection problem for a cluster inside a large matrix. Butucea and Ingster (2013); Butucea et al. (2013) formulated the submatrix detection and localization problems under Gaussian noise and determined sharp statistical transition boundaries. For the detection problem, Ma and Wu (2013a) provided a computational lower bound result under the assumption that hidden clique detection is computationally difficult.

Shabalin et al. (2009) provided a fast iterative maximization algorithm to heuristically solve the submatrix localization problem. Balakrishnan et al. (2011); Kolar et al. (2011) focused on statistical and computational trade-offs for the submatrix localization problem. Under the sparse regime $k_m \precsim m^{1/2}$ and $k_n \precsim n^{1/2}$, the entry-wise thresholding turns out to be the "near optimal" polynomial-time algorithm (which we will show to be a de-noised spectral algorithm that perform slightly better in Section 2.1.2). However, for the dense regime when $k_m \succsim m^{1/2}$ and $k_n \succsim n^{1/2}$, the algorithms provided in Kolar et al. (2011) are not optimal in the sense that there are other polynomial-time algorithm that can succeed in finding the submatrix with smaller SNR. Concurrently with our work, Chen and Xu (2014) provided a convex relaxation algorithm that improves the SNR boundary of Kolar et al. (2011) in the dense regime. On the computational downside, the implementation of the method requires a full SVD on each iteration, and therefore does not scale well with the dimensionality of the problem. Furthermore, there is no computational lower bound in the literature to guarantee the optimality of the SNR boundary achieved in Chen and Xu (2014). A problem similar to submatrix localization is that of clique finding in random graph. Deshpande and Montanari (2013) presented an iterative approximate message passing algorithm to solve the latter problem with sharp boundaries on SNR.

We would like to emphasize on the differences between the localization and the detection problems. In terms of the theoretical results, unlike detection, there is always a gap between statistically optimal and computationally feasible regions for localization. This non-vanishing computational-to-statistical-gap phenomenon also appears in the community detection literature with growing number of communities (Decelle et al., 2011). In terms of the methodology, for detection, combining the results in (Donoho and Jin, 2004; Ma and Wu, 2013a), there is no loss in treating M in model (2.2) as a vector and applying the higher criticism method (Donoho and Jin, 2004) to the vectorized matrix for the problem of submatrix detection, in the computationally efficient region. In fact, the procedure achieves sharper constants in the Gaussian setting. However, in contrast, we will show that for localization, it is crucial to utilize the matrix structure, even in the computationally efficient region.

Organization

The section is organized as follows. Section 2.1.2 establishes the computational boundary, with the computational lower bounds given in Section 2.1.2 and upper bound results in Sections 2.1.2-2.1.2. An extension to the case of multiple submatrices is presented in Section

2.1.2. The upper and lower bounds for statistical boundary for multiple submatrices are discussed in Section 2.1.3. A short discussion is given in Section 2.1.4. Technical proofs are deferred to Appendix.

Notation

Let $[m]$ denote the index set $\{1, 2, \ldots, m\}$. For a matrix $X \in \mathbb{R}^{m \times n}$, $X_{i\cdot} \in \mathbb{R}^n$ denotes its i-th row and $X_{\cdot j} \in \mathbb{R}^m$ denotes its j-th column. For any $I \subseteq [m], J \subseteq [n]$, X_{IJ} denotes the submatrix corresponding to the index set $I \times J$. For a vector $v \in \mathbb{R}^n$, $\|v\|_{\ell_p} = (\sum_{i \in [n]} |v_i|^p)^{1/p}$ and for a matrix $M \in \mathbb{R}^{m \times n}$, $\|M\|_{\ell_p} = \sup_{v \neq \mathbf{0}} \|Mv\|_{\ell_p} / \|v\|_{\ell_p}$. When $p = 2$, the latter is the usual spectral norm, abbreviated as $\|M\|_2$. The nuclear norm of a matrix M is convex surrogate for the rank, with the notation to be $\|M\|_*$. The Frobenius norm of a matrix M is defined as $\|M\|_F = \sqrt{\sum_{i,j} M_{ij}^2}$. The inner product associated with the Frobenius norm is defined as $\langle A, B \rangle = \mathsf{tr}(A^T B)$.

Denote the asymptotic notation $a(n) = \Theta(b(n))$ if there exist two universal constants c_l, c_u such that $c_l \leq \varliminf_{n \to \infty} a(n)/b(n) \leq \varlimsup_{n \to \infty} a(n)/b(n) \leq c_u$. Θ^* is asymptotic equivalence hiding logarithmic factors in the following sense: $a(n) = \Theta^*(b(n))$ iff there exists $c > 0$ such that $a(n) = \Theta(b(n) \log^c n)$. Additionally, we use the notation $a(n) \asymp b(n)$ as equivalent to $a(n) = \Theta(b(n))$, $a(n) \gtrsim b(n)$ iff $\lim_{n \to \infty} a(n)/b(n) = \infty$ and $a(n) \precsim b(n)$ iff $\lim_{n \to \infty} a(n)/b(n) = 0$.

We define the zero-mean sub-Gaussian random variable $\mathbf{z}$ with sub-Gaussian parameter σ in terms of its Laplacian

$$\mathbb{E} e^{\lambda \mathbf{z}} \leq \exp(\sigma^2 \lambda^2 / 2), \quad \text{for all} \quad \lambda > 0, \tag{2.5}$$

then we have

$$\mathbb{P}(|\mathbf{z}| > \sigma t) \leq 2 \cdot \exp(-t^2/2).$$

We call a random vector $Z \in \mathbb{R}^n$ isotropic with parameter σ if

$$\mathbb{E}(v^T Z)^2 = \sigma^2 \|v\|_{\ell_2}^2, \quad \text{for all} \quad v \in \mathbb{R}^n.$$

Clearly, Gaussian and Bernoulli measures, and more general product measures of zero-mean sub-Gaussian random variables satisfy this isotropic definition up to a constant scalar factor.

2.1.2. Computational Boundary

We characterize in this section the computational boundaries for the submatrix localization problem. Sections 2.1.2 and 2.1.2 consider respectively the computational lower bound and upper bound. The computational lower bound given in Theorem 2.1.2 is based on the hidden clique hypothesis.

Algorithmic Reduction and Computational Lower Bound

Theoretical Computer Science identifies a range of problems which are believed to be "hard," in the sense that in the worst-case the required computation grows exponentially with the size of the problem. Faced with a new computational problem, one might try to reduce any of the "hard" problems to the new problem, and therefore claim that the new problem is as hard as the rest in this family. Since statistical procedures typically deal with a random (rather than worst-case) input, it is natural to seek token problems that are believed to be computationally difficult on average with respect to some distribution on instances. The hidden clique problem is one such example (for recent results on this problem, see Feldman et al. (2013); Deshpande and Montanari (2013)). While there exists a quasi-polynomial algorithm, no polynomial-time method (for the appropriate regime, described below) is known. Following several other works on reductions for statistical problems, we work under the hypothesis that no polynomial-time method exists.

Let us make the discussion more precise. Consider the hidden clique model $\mathcal{G}(N, \kappa)$ where N is the total number of nodes and κ is the number of clique nodes. In the hidden clique model, a random graph instance is generated in the following way. Choose κ clique nodes uniformly at random from all the possible choices, and connect all the edges within the clique. For all the other edges, connect with probability $1/2$.

Hidden Clique Hypothesis for Localization ($\mathsf{HC_l}$) Consider the random instance of hidden clique model $\mathcal{G}(N,\kappa)$. For any sequence $\kappa(N)$ such that $\kappa(N) \leq N^{\beta}$ for some $0 < \beta < 1/2$, there is no randomized polynomial time algorithm that can find the planted clique with probability tending to 1 as $N \to \infty$. Mathematically, define the randomized polynomial time algorithm class $\mathsf{PolyAlg}$ as the class of algorithms $\mathcal{A}$ that satisfies

$$\overline{\lim_{N,\kappa(N)\to\infty}} \sup_{\mathcal{A}\in\mathsf{PolyAlg}} \mathbb{E}_{\mathsf{Clique}}\mathbb{P}_{\mathcal{G}(N,\kappa)|\mathsf{Clique}}\,(\text{runtime of } \mathcal{A} \text{ not polynomial in } N) = 0.$$

Then

$$\underline{\lim_{N,\kappa(N)\to\infty}} \inf_{\mathcal{A}\in\mathsf{PolyAlg}} \mathbb{E}_{\mathsf{Clique}}\mathbb{P}_{\mathcal{G}(N,\kappa)|\mathsf{Clique}}\,(\text{clique set returned by } \mathcal{A} \text{ not correct}) > 0,$$

where $\mathbb{P}_{\mathcal{G}(N,\kappa)|\mathsf{Clique}}$ is the (possibly more detailed due to randomness of algorithm) σ-field conditioned on the clique location and $\mathbb{E}_{\mathsf{Clique}}$ is with respect to uniform distribution over all possible clique locations.

Hidden Clique Hypothesis for Detection ($\mathsf{HC_d}$) Consider the hidden clique model $\mathcal{G}(N,\kappa)$. For any sequence of $\kappa(N)$ such that $\kappa(N) \leq N^{\beta}$ for some $0 < \beta < 1/2$, there is no randomized polynomial time algorithm that can distinguish between

$$\mathsf{H}_0 : \ \mathcal{P}_{\mathsf{ER}} \quad \text{v.s.} \quad \mathsf{H}_\alpha : \ \mathcal{P}_{\mathsf{HC}}$$

with probability going to 1 as $N \to \infty$. Here $\mathcal{P}_{\mathsf{ER}}$ is the Erdős-Rényi model, while $\mathcal{P}_{\mathsf{HC}}$ is the hidden clique model with uniform distribution on all the possible locations of the clique.

More precisely,

$$\varliminf_{N,\kappa(N)\to\infty} \inf_{\mathcal{A}\in\mathsf{PolyAlg}} \mathbb{E}_{\mathsf{Clique}}\mathbb{P}_{\mathcal{G}(N,\kappa)|\mathsf{Clique}}\,(\text{detection decision returned by } \mathcal{A} \text{ wrong}) > 0,$$

where $\mathbb{P}_{\mathcal{G}(N,\kappa)|\mathsf{Clique}}$ and $\mathbb{E}_{\mathsf{Clique}}$ are the same as defined in HC_l.

The hidden clique hypothesis has been used recently by several authors to claim computational intractability of certain statistical problems. In particular, Berthet and Rigollet (2013a); Ma and Wu (2013a) assumed the hypothesis HC_d and Wang et al. (2014) used HC_l. Localization is harder than detection, in the sense that if an algorithm $\mathcal{A}$ solves the localization problem with high probability, it also correctly solves the detection problem. Assuming that no polynomial time algorithm can solve the detection problem implies impossibility results in localization as well. In plain language, HC_l is a milder hypothesis than HC_d.

We will provide computational lower bound result for localization in Theorem 2.1.2. In Appendix, we contrast the difference of lower bound constructions between localization and detection. The detection computational lower bound was proved in Ma and Wu (2013a). For the localization computational lower bound, to the best of our knowledge, there is no proof in the literature. Theorem 2.1.2 ensures the upper bound in Lemma 2.1.1 being sharp.

Theorem 2.1.2 (Computational Lower Bound for Localization)**.** *Consider the submatrix model* (2.2) *with parameter tuple* $(m = n, k_m \asymp k_n \asymp n^\alpha, \lambda/\sigma = n^{-\beta})$*, where* $\frac{1}{2} < \alpha < 1$, $\beta > 0$. *Under the computational assumption* HC_l*, if*

$$\frac{\lambda}{\sigma} \precsim \sqrt{\frac{m+n}{k_m k_n}} \quad \Rightarrow \quad \beta > \alpha - \frac{1}{2},$$

it is not possible to localize the true support of the submatrix with probability going to 1 *within polynomial time.*

Our algorithmic reduction for localization relies on a *bootstrapping* idea based on the matrix structure and a cleaning-up procedure. These two key ideas offer new insights in addition to the usual computational lower bound arguments. Bootstrapping introduces an additional randomness on top of the randomness in the hidden clique. Careful examination of these two σ-fields allows us to write the resulting object into mixture of submatrix models. For submatrix localization we need to transform back the submatrix support to the original hidden clique support exactly, with high probability. In plain language, even though we lose track of the exact location of the support when reducing the hidden clique to submatrix model, we can still recover the exact location of the hidden clique with high probability. For technical details of the proof, please refer to the Appendix.

Adaptive Spectral Algorithm and Upper Bound

In this section, we introduce linear time algorithm that solves the submatrix localization problem above the computational boundary $\mathsf{SNR}_{\mathsf{c}}$. Our proposed localization Algorithms 1 and 2 are motivated by the spectral algorithm in random graphs (McSherry, 2001; Ng et al., 2002).

```
Algorithm 1 Vanilla Spectral Projection Algorithm for Dense Regime
Input: $X \in \mathbb{R}^{m \times n}$ the data matrix.
Output: A subset of the row indexes $\hat{R}_m$ and a subset of column indexes $\hat{C}_n$ as the localization sets of the submatrix.
1. Compute top left and top right singular vectors $U_{\cdot 1}$ and $V_{\cdot 1}$, respectively (these correspond to the SVD $X = U \Sigma V^T$)
2. To compute $\hat{C}_n$, calculate the inner products $U_{\cdot 1}^T X_{\cdot j} \in \mathbb{R}, 1 \leq j \leq n$. These values form two data-driven clusters, and a cut at the largest gap between consecutive values returns the subsets $\hat{C}_n$ and $[n] \backslash \hat{C}_n$. Similarly, for the $\hat{R}_m$, calculate $X_{i \cdot} V_{\cdot 1} \in \mathbb{R}, 1 \leq i \leq m$ and obtain two separated clusters.
```

The proposed algorithm has several advantages over the localization algorithms that appeared in literature. First, it is a linear time algorithm (that is, $\Theta(mn)$ time complexity). The top singular vectors can be evaluated using fast iterative power methods, which is efficient both in terms of space and time. Secondly, this algorithm does not require the prior knowledge of k_m and k_n and automatically adapts to the true submatrix size.

Lemma 2.1.1 below justifies the effectiveness of the spectral algorithm.

Lemma 2.1.1 (Guarantee for Spectral Algorithm)**.** *Consider the submatrix model* (2.2)*, Algorithm 1 and assume* $\min\{k_m, k_n\} \gtrsim \max\{m^{1/2}, n^{1/2}\}$*. There exist a universal* $C > 0$ *such that when*

$$\frac{\lambda}{\sigma} \geq C \cdot \left(\sqrt{\frac{m \vee n}{k_m k_n}} + \sqrt{\frac{\log n}{k_m} \vee \frac{\log m}{k_n}} \right),$$

the spectral method succeeds in the sense that $\hat{R}_m = R_m, \hat{C}_n = C_n$ *with probability at least* $1 - m^{-c} - n^{-c} - 2\exp\left(-c(m+n)\right)$*.*

Remark 2.1.1. The theory and algorithm remain the same if the signal matrix M is more general in the following sense: M has rank one, its left and right singular vectors are sparse, and the nonzero entries of the singular vectors are of the same order. Mathematically, $M = \lambda\sqrt{k_m k_n} \cdot uv^T$, where u, v are unit singular vectors with k_m, k_n non-zero entries, and $|u|_{\max}/|u|_{\min} \leq c$ and $|v|_{\max}/|v|_{\min} \leq c$ for some constant $c \geq 1$. Here for a vector w, $|w|_{\max}$ and $|w|_{\min}$ denote respectively the largest and smallest magnitudes among the nonzero coordinates. When $c = 1$, the algorithm is fully data-driven and does not require the knowledge of $\lambda, \sigma, k_m, k_n$. When c is large but finite, one may require in addition the knowledge of k_m and k_n to perform the final cut to obtain $\hat{C}_n$ and $\hat{R}_m$.

Dense Regime

We are now ready to state the SNR boundary for polynomial-time algorithms (under an appropriate computational assumption), thus excluding the exhaustive search procedure. The results hold under the dense regime when $k \gtrsim n^{1/2}$.

Theorem 2.1.3 (Computational Boundary for Dense Regime)**.** *Consider the submatrix model* (2.2) *and assume* $\min\{k_m, k_n\} \gtrsim \max\{m^{1/2}, n^{1/2}\}$*. There exists a critical rate*

$$\mathsf{SNR_c} \asymp \sqrt{\frac{m \vee n}{k_m k_n}} + \sqrt{\frac{\log n}{k_m} \vee \frac{\log m}{k_n}}$$

for the signal to noise ratio $\mathsf{SNR_c}$ *such that for* $\lambda/\sigma \gtrsim \mathsf{SNR_c}$*, both the adaptive linear time Algorithm 1 and convex relaxation (runs in polynomial time) will succeed in submatrix localization, i.e.,* $\hat{R}_m = R_m, \hat{C}_n = C_n$*, with high probability. For* $\lambda/\sigma \lesssim \mathsf{SNR_c}$*, there is no polynomial time algorithm that will work under the hidden clique hypothesis* $\mathsf{HC_l}$*.*

The proof of the above theorem is based on the theoretical justification of the spectral Algorithm 1 and the new computational lower bound result for localization in Theorem 2.1.2. We remark that the analyses can be extended to multiple, even growing number of submatrices case. We postpone a proof of this fact to Section 2.1.2 for simplicity and focus on the case of a single submatrix.

Sparse Regime

Under the sparse regime when $k \lesssim n^{1/2}$, a naive plug-in of Lemma 2.1.1 requires the $\mathsf{SNR_c}$ to be larger than $\Theta(n^{1/2}/k) \gtrsim \sqrt{\log n}$, which implies the vanilla spectral Algorithm 1 is outperformed by simple entrywise thresholding. However, a modified version with entrywise soft-thresholding as a preprocessing de-noising step turns out to provide near optimal performance in the sparse regime. Before we introduce the formal algorithm, let us define the soft-thresholding function at level t to be

$$\eta_t(y) = \mathrm{sign}(y)(|y| - t)_+. \tag{2.6}$$

Soft-thresholding as a de-noising step achieving optimal bias-and-variance trade-off has been widely understood in the wavelet literature, for example, see Donoho and Johnstone (1998).

Now we are ready to state the following de-noised spectral Algorithm 2 to localize the submatrix under the sparse regime when $k \lesssim n^{1/2}$.

Algorithm 2 De-noised Spectral Algorithm for Sparse Regime

Input: $X \in \mathbb{R}^{m\times n}$ the data matrix, a thresholding level $t = \Theta(\sigma\sqrt{\log \frac{m\vee n}{k_m k_n}})$.

Output: A subset of the row indexes $\hat{R}_m$ and a subset of column indexes $\hat{C}_n$ as the localization sets of the submatrix.

1. Soft-threshold each entry of the matrix X at level t, denote the resulting matrix as $\eta_t(X)$

2. Compute top left and top right singular vectors $U_{\cdot 1}$ and $V_{\cdot 1}$ of matrix $\eta_t(X)$, respectively (these correspond to the SVD $\eta_t(X) = U\Sigma V^T$)

3. To compute $\hat{C}_n$, calculate the inner products $U_{\cdot 1}^T \cdot \eta_t(X_{\cdot j}), 1 \le j \le n$. These values form two clusters. Similarly, for the $\hat{R}_m$, calculate $\eta_t(X_{i\cdot}) \cdot V_{\cdot 1}, 1 \le i \le m$ and obtain two separated clusters. A simple thresholding procedure returns the subsets $\hat{C}_n$ and $\hat{R}_m$.

Lemma 2.1.2 below provides the theoretical guarantee for the above algorithm when $k \precsim n^{1/2}$.

Lemma 2.1.2 (Guarantee for De-noised Spectral Algorithm)**.** *Consider the submatrix model* (2.2)*, soft-thresholded spectral Algorithm 2 with thresholded level σt, and assume* $\min\{k_m, k_n\} \precsim \max\{m^{1/2}, n^{1/2}\}$. *There exist a universal $C > 0$ such that when*

$$\frac{\lambda}{\sigma} \ge C \cdot \left(\left[\sqrt{\frac{m \vee n}{k_m k_n}} + \sqrt{\frac{\log n}{k_m} \vee \frac{\log m}{k_n}} \right] \cdot e^{-t^2/2} + t \right),$$

the spectral method succeeds in the sense that $\hat{R}_m = R_m, \hat{C}_n = C_n$ with probability at least $1 - m^{-c} - n^{-c} - 2\exp\left(-c(m+n)\right)$. *Further if we choose $\Theta(\sigma\sqrt{\log \frac{m\vee n}{k_m k_n}})$ as the optimal thresholding level, we have de-noised spectral algorithm works when*

$$\frac{\lambda}{\sigma} \succsim \sqrt{\log \frac{m \vee n}{k_m k_n}}.$$

Combining the hidden clique hypothesis $\mathsf{HC_l}$ together with Lemma 2.1.2, the following theorem holds under the sparse regime when $k \precsim n^{1/2}$.

Theorem 2.1.4 (Computational Boundary for Sparse Regime)**.** *Consider the submatrix model* (2.2) *and assume* $\max\{k_m, k_n\} \precsim \min\{m^{1/2}, n^{1/2}\}$*. There exists a critical rate for the signal to noise ratio* $\mathsf{SNR_c}$ *between*

$$1 \precsim \mathsf{SNR_c} \precsim \sqrt{\log \frac{m \vee n}{k_m k_n}}$$

such that for $\lambda/\sigma \succsim \sqrt{\log \frac{m\vee n}{k_m k_n}}$*, the linear time Algorithm 2 will succeed in submatrix localization, i.e.,* $\hat{R}_m = R_m, \hat{C}_n = C_n$*, with high probability. For* $\lambda/\sigma \precsim 1$*, there is no polynomial time algorithm that will work under the hidden clique hypothesis* $\mathsf{HC_l}$*.*

Remark 2.1.2. The upper bound achieved by the de-noised spectral Algorithm 2 is optimal in the two boundary cases: $k = 1$ and $k \asymp n^{1/2}$. When $k = 1$, both the information theoretic and computational boundary meet at $\sqrt{\log n}$. When $k \asymp n^{1/2}$, the computational lower bound and upper bound match in Theorem 2.1.4, thus suggesting the near optimality of Algorithm 2 within the polynomial time algorithm class. The potential logarithmic gap is due to the crudeness of the hidden clique hypothesis. Precisely, for $k = 2$, hidden clique is not only hard for $G(n, p)$ with $p = 1/2$, but also hard for $G(n, p)$ with $p = 1/\log n$. Similarly for $k = n^\alpha, \alpha < 1/2$, hidden clique is not only hard for $G(n, p)$ with $p = 1/2$, but also for some $0 < p < 1/2$.

Extension to Growing Number of Submatrices

The computational boundaries established in the previous sections for a single submatrix can be extended to non-overlapping multiple submatrices model (2.3). The non-overlapping assumption corresponds to that for any $1 \leq s \neq t \leq r$, $R_s \cap R_t = \emptyset$ and $C_s \cap C_t = \emptyset$. The Algorithm 3 below is an extension of the spectral projection Algorithm 1 to address the multiple submatrices localization problem.

Algorithm 3 Spectral Algorithm for Multiple Submatrices

Input: $X \in \mathbb{R}^{m\times n}$ the data matrix. A pre-specified number of submatrices r.

Output: A subset of the row indexes $\{\hat{R}^s_m, 1 \le s \le r\}$ and a subset of column indexes $\{\hat{C}^s_n, 1 \le s \le r\}$ as the localization of the submatrices.

1. Calculate top r left and right singular vectors in the SVD $X = U\Sigma V^T$. Denote these vectors as $U_r \in \mathbb{R}^{m\times r}$ and $V_r \in \mathbb{R}^{n\times r}$, respectively

2. For the $\hat{C}^s_n, 1 \le s \le r$, calculate the projection $U_r(U_r^T U_r)^{-1} U_r^T X_{\cdot j}, 1 \le j \le n$, run k-means clustering algorithm (with $k = r+1$) for these n vectors in $\mathbb{R}^m$. For the $\hat{R}^s_m, 1 \le s \le r$, calculate $V_r(V_r^T V_r)^{-1} V_r^T X_{i\cdot}^T, 1 \le i \le m$, run k-means clustering algorithm (with $k = r+1$) for these m vectors in $\mathbb{R}^n$ (while the effective dimension is $\mathbb{R}^r$).

We emphasize that the following Proposition 2.1.3 holds even when the number of submatrices r grows with m, n.

Lemma 2.1.3 (Spectral Algorithm for Non-overlapping Submatrices Case)**.** *Consider the non-overlapping multiple submatrices model* (2.3) *and Algorithm 3. Assume*

$$k_s^{(m)} \asymp k_m, k_s^{(n)} \asymp k_n, \lambda_s \asymp \lambda$$

for all $1 \le s \le r$ *and* $\min\{k_m, k_n\} \gtrsim \max\{m^{1/2}, n^{1/2}\}$*. There exist a universal* $C > 0$ *such that when*

$$\frac{\lambda}{\sigma} \ge C \cdot \left(\sqrt{\frac{r}{k_m \wedge k_n}} + \sqrt{\frac{\log n}{k_m}} \vee \sqrt{\frac{\log m}{k_n}} + \sqrt{\frac{m \vee n}{k_m k_n}} \right), \tag{2.7}$$

the spectral method succeeds in the sense that $\hat{R}_m^{(s)} = R_m^{(s)}, \hat{C}_n^{(s)} = C_n^{(s)}, 1 \le s \le r$ *with probability at least* $1 - m^{-c} - n^{-c} - 2\exp(-c(m+n))$*.*

Remark 2.1.3. Under the non-overlapping assumption, $rk_m \precsim m$, $rk_n \precsim n$ hold in most cases. Thus the first term in Equation (2.7) is dominated by the latter two terms. Thus a growing number r does not affect the bound in Equation (2.7) as long as the non-overlapping assumption holds.

2.1.3. Statistical Boundary

In this section we study the statistical boundary. As mentioned in the introduction, in the Gaussian noise setting, the statistical boundary for a single submatrix localization has been established in Butucea et al. (2013). In this section, we generalize to localization of a growing number of submatrices, as well as sub-Gaussian noise, at the expense of having non-exact constants for the threshold.

Information Theoretic Bound

We begin with the information theoretic lower bound for the localization accuracy.

Lemma 2.1.4 (Information Theoretic Lower Bound)**.** *Consider the submatrix model* (2.2) *with Gaussian noise* $Z_{ij} \sim \mathcal{N}(0, \sigma^2)$*. For any fixed* $0 < \alpha < 1$*, there exist a universal constant* C_α *such that if*

$$\frac{\lambda}{\sigma} \leq C_\alpha \cdot \sqrt{\frac{\log(m/k_m)}{k_n} + \frac{\log(n/k_n)}{k_m}}, \tag{2.8}$$

any algorithm $\mathcal{A}$ *will fail to localize the submatrix in the following minimax sense:*

$$\inf_{\mathcal{A} \in \mathsf{AllAlg}} \sup_{M \in \mathcal{M}} \quad \mathbb{P}\left(\hat{R}_m^{\mathcal{A}} \neq R_m \quad or \quad \hat{C}_n^{\mathcal{A}} \neq C_n\right) > 1 - \alpha - \frac{\log 2}{k_m \log(m/k_m) + k_n \log(n/k_n)}.$$

Combinatorial Search for Growing Number of Submatrices

Combinatorial search over all submatrices of size $k_m \times k_n$ finds the location with the strongest aggregate signal and is statistically optimal (Butucea et al., 2013; Butucea and Ingster, 2013). Unfortunately, it requires computational complexity $\Theta\left(\binom{m}{k_m} + \binom{n}{k_n}\right)$, which is exponential in k_m, k_n. The search Algorithm 4 was introduced and analyzed under the Gaussian setting for a single submatrix in Butucea and Ingster (2013), which can be used iteratively to solve multiple submatrices localization.

Algorithm 4 Combinatorial Search Algorithm

Input: $X \in \mathbb{R}^{m \times n}$ the data matrix.

Output: A subset of the row indexes $\hat{R}_m$ and a subset of column indexes $\hat{C}_n$ as the localization of the submatrix.

For all index subsets $I \times J$ with $|I| = k_m$ and $|J| = k_n$, calculate the sum of the entries in the submatrix X_{IJ}. Report the index subset $\hat{R}_m \times \hat{C}_n$ with the largest sum.

For the case of multiple submatrices, the submatrices can be extracted with the largest sum in a greedy fashion.

Lemma 2.1.5 below provides a theoretical guarantee for Algorithm 4 to achieve the information theoretic lower bound.

Lemma 2.1.5 (Guarantee for Search Algorithm). *Consider the non-overlapping multiple submatrices model* (2.3) *and iterative application of Algorithm 4 in a greedy fashion for r times. Assume*

$$k_s^{(m)} \asymp k_m, k_s^{(n)} \asymp k_n, \lambda_s \asymp \lambda$$

for all $1 \le s \le r$ and $\max\{k_m, k_n\} \precsim \min\{m, n\}$. There exists a universal constant $C > 0$ such that if

$$\frac{\lambda}{\sigma} \ge C \cdot \sqrt{\frac{\log(em/k_m)}{k_n} + \frac{\log(en/k_n)}{k_m}},$$

then Algorithm 4 will succeed in returning the correct location of the submatrix with probability at least $1 - \frac{2k_m k_n}{mn}$.

To complete Theorem 2.1.1, we include the following Theorem 2.1.5 capturing the statistical boundary. It is proved by exhibiting the information-theoretic lower bound Lemma 2.1.4 and analyzing Algorithm 4.

Theorem 2.1.5 (Statistical Boundary). *Consider the submatrix model* (2.2). *There exists a critical rate*

$$\mathsf{SNR}_s \asymp \sqrt{\frac{\log n}{k_m} \vee \frac{\log m}{k_n}}$$

for the signal to noise ratio, such that for any problem with $\lambda/\sigma \gtrsim \mathsf{SNR_s}$, the statistical search Algorithm 4 will succeed in submatrix localization, i.e., $\hat{R}_m = R_m, \hat{C}_n = C_n$, with high probability. On the other hand, if $\lambda/\sigma \precsim \mathsf{SNR_s}$, no algorithm will work (in the minimax sense) with probability tending to 1.

2.1.4. Discussion

Submatrix Localization v.s. Detection As pointed out in Section 2.1.1, for any $k = n^\alpha, 0 < \alpha < 1$, there is an intrinsic SNR gap between computational and statistical boundaries for submatrix localization. Unlike the submatrix detection problem where for the regime $2/3 < \alpha < 1$, there is no gap between what is computationally possible and what is statistical possible, the inevitable gap in submatrix localization is due to the combinatorial structure of the problem. This phenomenon is also seen in some network related problems, for instance, stochastic block models with a growing number of communities Decelle et al. (2011). Compared to the submatrix detection problem, the algorithm to solve the localization problem is more complicated and the techniques required for the analysis are much more involved.

Detection for Growing Number of Submatrices The current thesis solves localization of a growing number of submatrices. In comparison, for detection, the only known results are for the case of a single submatrix as considered in Butucea and Ingster (2013) for the statistical boundary and in Ma and Wu (2013a) for the computational boundary. The detection problem in the setting of a growing number of submatrices is of significant interest. In particular, it is interesting to understand the computational and statistical trade-offs in such a setting. This will need further investigation.

Estimation of the Noise Level σ Although Algorithms 1 and 3 do not require the noise level σ as an input, Algorithm 2 does require the knowledge of σ. The noise level σ can be estimated robustly. In the Gaussian case, a simple robust estimator of σ is the following median absolute deviation (MAD) estimator due to the fact that M is sparse $k^2/m^2 \ll 0.25$:

$$\begin{aligned}\hat{\sigma} &= \text{median}_{ij}|X_{ij} - \text{median}_{ij}(X_{ij})|/\Phi^{-1}(0.75) \\ &\approx 1.4826 \times \text{median}_{ij}|X_{ij} - \text{median}_{ij}(X_{ij})|.\end{aligned}$$

2.2. Semi-supervised Community Detection

2.2.1. Introduction

The stochastic block model (SBM) is a well-studied model that addresses the clustering phenomenon in large networks. Various phase transition phenomena and limitations for efficient algorithms have been established for this "vanilla" SBM (Coja-Oghlan, 2010; Decelle et al., 2011; Massoulié, 2014; Mossel et al., 2012, 2013a; Krzakala et al., 2013; Abbe et al., 2014; Hajek et al., 2014; Abbe and Sandon, 2015a; Deshpande et al., 2015). However, in real network datasets, additional side information is often available. This additional information may come, for instance, in the form of a small portion of revealed labels (or, community memberships), and this thesis is concerned with methods for incorporating this additional information to improve recovery of the latent community structure. Many global algorithms studied in the literature are based on spectral analysis (with belief propagation as a further refinement) or semi-definite programming. For these methods, it appears to be difficult to incorporate such additional side information, although some success has been reported (Cucuringu et al., 2012; Zhang et al., 2014a). Incorporating the additional information within local algorithms, however, is quite natural. In this thesis, we focus on local algorithms and study their fundamental limitations. Our model is a **partially labeled stochastic block model** (p-SBM), where δ portion of community labels are randomly revealed.

We address the following questions:

Phase Boundary Are there different phases of behavior in terms of the recovery guarantee, and what is the phase boundary for partially labeled SBM? How does the amount of additional information δ affect the phase boundary?

Inference Guarantee What is the optimal guarantee on the recovery results for p-SBM and how does it interpolate between weak and strong consistency known in the literature? Is there an efficient and near-optimal parallelizable algorithm?

Limitation for Local v.s. Global Algorithms While optimal local algorithms (belief propagation) are computationally efficient, some global algorithms may be computationally prohibitive. Is there a fundamental difference in the limits for local and global algorithms? An answer to this question gives insights on the computational and statistical trade-offs.

Problem Formulation

We define p-SBM with parameter bundle (n, k, p, q, δ) as follows. Let n denote the number of nodes, k the number of communities, p and q – the intra and inter connectivity probability, respectively. The proportion of revealed labels is denoted by δ. Specifically, one observes a partially labeled graph $G(V, E)$ with $|V| = n$, generated as follows. There is a latent disjoint partition $V = \bigcup_{l=1}^{k} V_l$ into k equal-sized groups,[1] with $|V_l| = n/k$. The partition information introduces the latent labeling $\ell(v) = l$ iff $v \in V_l$. For any two nodes $v_i, v_j, 1 \leq i, j \leq n$, there is an edge between them with probability p if v_i and v_j are in the same partition, and with probability q if not. Independently for each node $v \in V$, its true labeling is revealed with probability δ. Denote the set of labeled nodes V^{l}, its revealed labels $\ell(V^{\mathrm{l}})$, and unlabeled nodes by V^{u} (where $V = V^{\mathrm{l}} \cup V^{\mathrm{u}}$).

Equivalently, denote by $G \in \mathbb{R}^{n \times n}$ the adjacency matrix, and let $L \in \mathbb{R}^{n \times n}$ be the structural block matrix

$$L_{ij} = 1_{\ell(v_i) = \ell(v_j)},$$

where $L_{ij} = 1$ iff node i, j share the same labeling, $L_{ij} = 0$ otherwise. Then we have independently for $1 \leq i < j \leq n$

$$\begin{aligned} B_{ij} &\sim \mathsf{Bernoulli}(p) \quad \text{if } L_{ij} = 1, \\ B_{ij} &\sim \mathsf{Bernoulli}(q) \quad \text{if } L_{ij} = 0. \end{aligned}$$

Given the graph $G(V, E)$ and the partially revealed labels $\ell(V^{\mathrm{l}})$, we want to recover the re-

[1] The result can be generalized to the balanced case, $|V_l| \asymp n/k$, see Section 2.2.2.

maining labels efficiently and accurately. We are interested in the case when $\delta(n), p(n), q(n)$ decrease with n, and $k(n)$ can either grow with n or stay fixed.

Prior Work

In the existing literature on SBM without side information, there are two major criteria – weak and strong consistency. Weak consistency asks for recovery better than random guessing in a sparse random graph regime ($p \asymp q \asymp 1/n$), and strong consistency requires exact recovery for each node above the connectedness theshold ($p \asymp q \asymp \log n/n$). Interesting phase transition phenomena in weak consistency for SBM have been discovered in (Decelle et al., 2011) via insightful cavity method from statistical physics. Sharp phase transitions for weak consistency have been thoroughly investigated in (Coja-Oghlan, 2010; Mossel et al., 2012, 2013a,b; Massoulié, 2014). In particular, spectral algorithms on the non-backtracking matrix have been studied in (Massoulié, 2014) and the non-backtracking walk in (Mossel et al., 2013b). Spectral algorithms as initialization and belief propagation as further refinement to achieve optimal recovery was established in (Mossel et al., 2013a). The work of Mossel et al. (2012) draws a connection between SBM thresholds and broadcasting tree reconstruction thresholds through the observation that sparse random graphs are locally tree-like. Recent work of Abbe and Sandon (2015b) establishes the positive detectability result down to the Kesten-Stigum bound for all k via a detailed analysis of a modified version of belief propagation. For strong consistency, (Abbe et al., 2014; Hajek et al., 2014, 2015) established the phase transition using information theoretic tools and semi-definite programming (SDP) techniques. In the statistical literature, Zhang and Zhou (2015); Gao et al. (2015) investigated the mis-classification rate of the standard SBM.

Kanade et al. (2014) is one of the few papers that theoretically studied the partially labeled SBM. The authors investigated the stochastic block model where the labels for a vanishing fraction ($\delta \to 0$) of the nodes are revealed. The results focus on the asymptotic case when δ is sufficiently small and block number k is sufficiently large, with no specified growth rate dependence. Kanade et al. (2014) show that pushing below the Kesten-Stigum bound is possible in this setting, connecting to a similar phenomenon in k-label broadcasting processes (Mossel, 2001). In contrast to these works, the focus of our study is as follows. Given a certain parameter bundle p-SBM(n, k, p, q, δ), we investigate the recovery thresholds as the fraction of labeled nodes changes, and determine the fraction of nodes that local algorithms can recover.

The focus of this thesis is on local algorithms. These methods, naturally suited for distributed computation (Linial, 1992), provide efficient (sub-linear time) solutions to computationally hard combinatorial optimization problems on graphs. For some of these problems, they are good approximations to global algorithms. We refer to (Kleinberg, 2000) on the shortest path problem for small-world random graphs, (Gamarnik and Sudan, 2014) for the maximum independent set problem for sparse random graphs, (Parnas and Ron, 2007) on the minimum vertex cover problem, as well as (Nguyen and Onak, 2008).

Finally, let us briefly review the literature on broadcasting processes on trees, from which we borrow technical tools to study p-SBM. Consider a Markov chain on an infinite tree rooted at ρ with branching number b. Given the label of the root $\ell(\rho)$, each vertex chooses its label by applying the Markov rule M to its parent's label, recursively and independently. The process is called broadcasting process on trees. One is interested in reconstructing the root label $\ell(\rho)$ given all the n-th level leaf labels. Sharp reconstruction thresholds for the broadcasting process on general trees for the symmetric Ising model setting (each node's label is $\{+, -\}$) have been studied in (Evans et al., 2000). Mossel et al. (2003) studied a general Markov channel on trees that subsumes k-state Potts model and symmetric Ising model as special cases; the authors established non-census-solvability below the Kesten-Stigum bound. Janson and Mossel (2004) extended the sharp threshold to robust reconstruction cases, where the vertex' labels are contaminated with noise. In general, transition thresholds proved in the above literature correspond to the Kesten-Stigum bound $b|\lambda_2(M)|^2 = 1$ (Kesten and Stigum, 1966b,a). We remark that for a general Markov channel

M, $b|\lambda_2(M)|^2 < 1$ does not always imply non-solvability — even though it indeed implies non-census-solvability (Mossel et al., 2003) — which is equivalent to the extremality of free-boundary Gibbs measure. The non-solvability of the tree reconstruction problem below the Kesten-Stigum bound for a general Markov transition matrix $M \in \mathbb{R}^{k \times k}$ still remains open, especially for large k.

Our Contributions

This section summarizes the results. In terms of methodology, we propose a new efficient linearized message-passing Algorithm 5 that solves the label recovery problem of p-SBM in near-linear runtime. The algorithm shares the same transition boundary as the optimal local algorithm (belief propagation) and takes on a simple form of a weighted majority vote (with the weights depending on graph distance). This voting strategy is easy to implement (see Section 2.2.5). On the theoretical front, our focus is on establishing recovery guarantees according to the size of the Signal-to-Noise Ratio (SNR), defined as

$$\mathsf{SNR}(n, k, p, q, \delta) := (1-\delta)\frac{n(p-q)^2}{k^2(q + \frac{p-q}{k})}. \tag{2.9}$$

Phase Boundary For $k = 2$, the phase boundary for recovery guarantee is

$$\mathsf{SNR} = 1.$$

Above the threshold, the problem can be solved efficiently. Below the threshold, the problem is intrinsically hard. For growing k, on the one hand, a linearized message-passing algorithm succeeds when

$$\mathsf{SNR} > 1,$$

matching the well-established Kesten-Stigum bound for all k. On the other hand, no local algorithms work significantly better than random guessing if

$$\mathsf{SNR} < \frac{1}{4}.$$

Inference Guarantee Above the SNR phase boundary, Algorithm 5, a fast linearized message-passing algorithm $\hat{A}$ (with near-linear run-time $\mathcal{O}^*(n)$) provides near optimal recovery. For $k = 2$, under the regime $\mathsf{SNR} > 1$, the proportion of mis-classified labels is at most

$$\sup_{l \in \{+,-\}} \mathbb{P}_l(\hat{A} \neq l) \leq \exp\left(-\frac{\mathsf{SNR} - 1}{2 + o(1)}\right) \wedge \frac{1}{2}.$$

Thus when $\mathsf{SNR} \in (1, 2\log n)$, the recovery guarantee smoothly interpolates between weak and strong consistency. On the other hand, below the boundary $\mathsf{SNR} < 1$, all local algorithms suffer the minimax classification error at least

$$\inf_{\Phi} \sup_{l \in \{+,-\}} \mathbb{P}_l(\Phi \neq l) \geq \frac{1}{2} - \mathcal{O}\left(\sqrt{\frac{\delta}{1-\delta} \cdot \frac{\mathsf{SNR}}{1 - \mathsf{SNR}}}\right).$$

For growing k, above the phase boundary $\mathsf{SNR} > 1$, the proportion of mis-classified labels is at most

$$\sup_{l \in [k]} \mathbb{P}_l(\hat{A} \neq l) \leq (k-1) \cdot \exp\left(-\frac{\mathsf{SNR} - 1}{2 + o(1)}\right) \wedge \frac{k-1}{k}$$

via the approximate message-passing algorithm. However, below the boundary $\mathsf{SNR} < 1/4$, the minimax classification error is lower bounded by

$$\inf_{\Phi} \sup_{l \in [k]} \mathbb{P}_l(\Phi \neq l) \geq \frac{1}{2} - \mathcal{O}\left(\frac{\delta}{1-\delta} \cdot \frac{\mathsf{SNR}}{1 - 4 \cdot \mathsf{SNR}} \vee \frac{1}{k}\right).$$

Limitations of Local v.s. Global Algorithms It is known that the statistical boundary (limitation for global and possibly exponential time algorithms) for growing number of communities is $\mathsf{SNR} \asymp \mathcal{O}(\frac{\log k}{k})$ (Abbe and Sandon (2015b), weak consistency) and $\mathsf{SNR} \asymp \mathcal{O}(\frac{\log n}{k})$ (Chen and Xu (2014), strong consistency). We show in this thesis that the limitation for local algorithms (those that use neighborhood information up to depth $\log n$) is

$$\frac{1}{4} \leq \mathsf{SNR} \leq 1.$$

In conclusion, as k grows, *there is a factor k gap between the boundaries for global and local algorithms.* Local algorithms can be evaluated in near line time; however, the global algorithm achieving the statistical boundary requires exponential time.

To put our results in the right context, let us make comparisons with the known literature. Most of the literature studies the standard SBM with no side labeling information. Here, many algorithms that achieve the sharp phase boundary are either global algorithms, or a combination of global and local algorithms, see (Mossel et al., 2013b; Massoulié, 2014; Hajek et al., 2014; Abbe et al., 2014). However, from the theoretical perspective, it is not clear how to distinguish the limitation for global v.s. local algorithms through the above studies. In addition, from the model and algorithmic perspective, many global algorithms such as spectral (Coja-Oghlan, 2010; Massoulié, 2014) and semi-definite programming (Abbe et al., 2014; Hajek et al., 2014) are not readily applicable in a principled way when there is partially revealed labels.

We try to resolve the above concerns. First, we establish a detailed statistical inference guarantee for label recovery. Allowing for a vanishing δ amount of randomly revealed labels, we show that a fast local algorithm enjoys a good recovery guarantee that interpolates between weak and strong recovery precisely, down to the well-known Kesten-Stigum bound, for general k. The error bound $\exp(-(\mathsf{SNR}-1)/2)$ proved in this thesis improves upon the best known result of $(\mathsf{SNR}-1)^{-1}$ in the weak recovery literature. We also prove that the limitation for local algorithms matches the Kesten-Stigum bound, which is sub-optimal compared to the limitation for global algorithms, when k grows. We also remark that the boundary we establish matches the best known result for the standard SBM when we plug in $\delta = 0$.

We study the message-passing algorithms for multi-label broadcasting tree when a fraction of nodes' labels have been revealed. Unlike the usual asymptotic results for belief propagation and approximate message-passing, we prove *non-asymptotic concentration of measure phenomenon* for messages on multi-label broadcasting trees. As the tree structure encodes detailed dependence among random variables, proving the concentration phenomenon requires new ideas. We further provide a lower bound on belief propagation for multi-label broadcasting trees.

Organization

The rest of the section is organized as follows. Section 2.2.2 reviews the preliminary background and theoretical tools – broadcasting trees – that will be employed to solve the p-SBM problem. To better illustrate the main idea behind the theoretical analysis, we split the main result into two sections: Section 2.2.3 resolves the recovery transition boundary for $k = 2$, where the analysis is simple and best illustrates the main idea. In Section 2.2.4, we focus on the growing $k = k(n)$ case, where a modified algorithm and a more detailed analysis are provided. In the growing k case, we establish a distinct gap in phase boundaries between the global algorithms and local algorithms.

2.2.2. Preliminaries

Broadcasting Trees

First, we introduce the notation for the tree broadcasting process. Let $T_{\leq t}(\rho)$ denote the tree up to depth t with root ρ. The collection of revealed labels for a broadcasting tree $T_{\leq t}(\rho)$ is denoted as $\ell_{T_{\leq t}(\rho)}$ (this is a collection of random variables). The labels for the binary broadcasting tree are $[2] := \{+, -\}$ and for k-broadcasting tree $[k] := \{1, 2, \dots, k\}$. For a node v, the set of labeled children is denoted by $\mathcal{C}^{\mathrm{l}}(v)$ and unlabeled ones by $\mathcal{C}^{\mathrm{u}}(v)$. We also denote the depth-t children of v to be $\mathcal{C}_t(v)$. For a broadcasting tree T, denote by d its broadcasting number, whose rigorous definition is given in (Evans et al., 2000; Lyons and Peres, 2005). For a broadcasting tree with bias parameter θ, the labels are broadcasted in the following way: conditionally on the label of v,

$$\ell(u) = \begin{cases} \ell(v) & \text{w.p. } \theta + \frac{1-\theta}{k} \\ l \in [k] \backslash \ell(v) & \text{w.p. } \frac{1-\theta}{k} \end{cases}$$

for any $u \in \mathcal{C}(v)$. In words, the child copies the color of its parent with probability $\theta + \frac{1-\theta}{k}$, or changes to any of the remaining $k-1$ colors with equal probability $\frac{1-\theta}{k}$. For the node v, $N_{\mathcal{C}^{\mathrm{l}}(v)}(+)$ denotes the number of revealed positive nodes among its children. Similarly, we define $N_{\mathcal{C}^{\mathrm{l}}(v)}(l)$ for $l \in [k]$ in multi-label trees.

Local Tree-like Graphs & Local Algorithms

When viewed locally, stochastic block models share many properties with broadcasting trees. In fact, via the coupling lemma (see Lemma A.2.1) from (Mossel et al., 2012), one can show the graph generated from the stochastic block model is locally a tree-like graph. For the rest of the thesis, we abbreviate the following maximum coupling depth $\bar{t}_{n,k,p,q}$ as $\bar{t}$ (see Lemma A.2.1 for details).

Definition 2.2.1 ($\bar{t}$-Local Algorithm Class for p-SBM)**.** *The $\bar{t}$-local algorithm class is the collection of decentralized algorithms that run in parallel on nodes of the graph. To recover a node v's label in p-SBM, an algorithm may only utilize information (revealed labels, connectivity) of the local tree $T_{\leq \bar{t}}(v)$ rooted at v with depth at most $\bar{t}$.*

In view of the coupling result, for the stochastic block model p-SBM$(n, k = 2, p, q, \delta)$, as long as we focus on $\bar{t}$-local algorithms, we can instead study the binary-label broadcasting process Tree$_{k=2}(\theta, d, \delta)$ with broadcasting number $d = \frac{n}{2}(p+q)$ and bias parameter $\theta = \frac{p-q}{p+q}$. Similarly, for the multi-label model p-SBM(n, k, p, q, δ), we will study the k-label broadcasting process Tree$_k(\theta, d, \delta)$ with broadcasting number $d = n(q + \frac{p-q}{k})$ and bias parameter $\theta = \frac{p-q}{k(q+\frac{p-q}{k})}$. [2] For each layer of the broadcasting tree, δ portion of nodes'

[2] In the balanced SBM case, for each node, the local tree changes slightly with different branching number

labels are revealed. Our goal is to understand the condition under which message-passing algorithms on multi-label broadcasting trees succeed in recovering the root label.

Hyperbolic Functions and Other Notation

In order to introduce the belief propagation and message-passing algorithms, let us recall several hyperbolic functions that will be used frequently. As we show, linearization of the hyperbolic function induces a new approximate message-passing algorithm. Recall that

$$\tanh x = \frac{e^x - e^{-x}}{e^x + e^{-x}}, \quad \operatorname{arctanh} x = \frac{1}{2} \log \left(\frac{1+x}{1-x} \right),$$

and define

$$f_\theta(x) := 2 \operatorname{arctanh} \left(\theta \tanh \frac{x}{2} \right) = \log \frac{1 + \theta \cdot \frac{e^x - 1}{e^x + 1}}{1 - \theta \cdot \frac{e^x - 1}{e^x + 1}}. \tag{2.10}$$

The function $f_\theta : \mathbb{R} \to \mathbb{R}$ is a contraction with

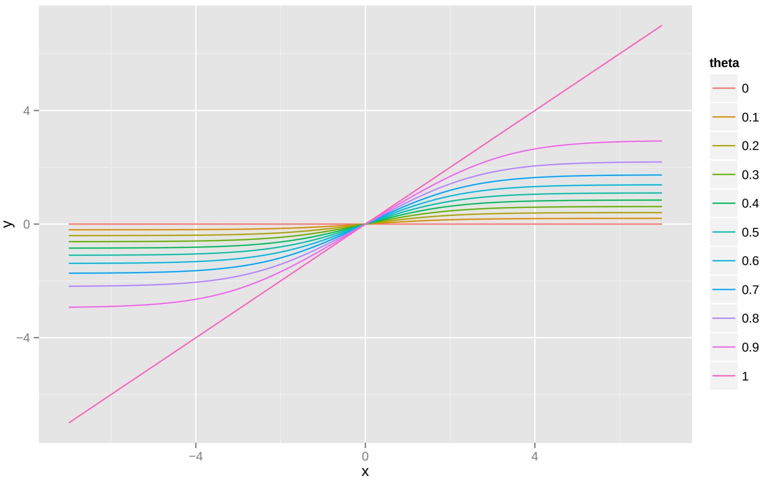

Figure 6: Function f_θ for $\theta \in [0, 1]$.

$$|f(x) - f(y)| \leq \theta |x - y|$$

and bias parameter.

since

$$\frac{df_\theta(x)}{dx} = \frac{2\theta}{(1-\theta^2)\cosh(x) + (1+\theta^2)} \leq \theta.$$

An illustration of f_θ is provided in Figure 6. The recursion rule for message passing can be written succinctly using the function f_θ, as we show in Section 2.2.3.

Let us collect a few remaining definitions. The moment generating function (MGF) for a random variable X is denoted by $\Psi_X(\lambda) = \mathbb{E}e^{\lambda X}$, for $\lambda > 0$, and the cumulant generating function is defined as $K_X(\lambda) = \log \Psi_X(\lambda)$. For asymptotic order of magnitude, we use $a(n) = \mathcal{O}(b(n))$ to mean $\forall n, a(n) \leq Cb(n)$ for some universal constant C, and use $\mathcal{O}^*(\cdot)$ to omit the poly-logarithmic dependence. As for notation $\precsim, \succsim$: $a(n) \precsim b(n)$ if and only if $\overline{\lim}_{n\to\infty} \frac{a(n)}{b(n)} \leq c$, with some constant $c > 0$, and vice versa. The square bracket $[\cdot]$ is used to represent the index set $[k] := [1, 2, \ldots, k]$; in particular when $k = 2$, $[2] := \{+, -\}$ for convenience.

2.2.3. Number of Communities $k = 2$: Message Passing with Partial Information

p-SBM Transition Thresholds

We propose a novel linearized message-passing algorithm to solve the p-SBM in near-linear time. The method employs Algorithm 7 and 8 as sub-routines, can run in parallel, and is easy to implement.

Algorithm 5 Message Passing for p-SBM

```
Data: A network graph $G(V,E)$ with partial label information, where $V = V^{\mathrm{l}} \cup V^{\mathrm{u}}$ is
        composed of labeled set and unlabeled set. Denote $\epsilon = o(1)$ small, and $\bar{t} \precsim \frac{\log n}{\log(n(p+q))}$.
Result: The labeling for each node $v \in V^{\mathrm{u}}$.
for each node $v \in V^{\mathrm{u}}$ in the unlabeled set, do
    open the tree neighborhood $T_{\leq \bar{t}}(v)$ induced by the graph $G(V,E)$   for each node $u \in$
    $\mathcal{C}_{(1-\epsilon)\bar{t}}(v)$, i.e., depth $(1-\epsilon)\bar{t}$ child of $v$, do
        focus on the subtree $T_{\leq \epsilon\bar{t}}(u)$,   initialize the message for $u$ via the labeled node $\in V^{\mathrm{l}}$
        in layer $\epsilon\bar{t}$ of the subtree [3]
    end
    run message-passing Algorithm 7 (Algorithm 8 for general $k$) on the tree $T_{\leq(1-\epsilon)\bar{t}}(v)$
    with initial message on layer $(1-\epsilon)\bar{t}$   output $\ell(v)$.
end
```

Now we are ready to present the main result.

Theorem 2.2.1 (Transition Thresholds for p-SBM: $k=2$)**.** *Consider the partially labeled stochastic block model $G(V,E)$ and its revealed labels $\ell(V^l)$ under the conditions (1) $np \asymp nq \precsim n^{o(1)}$ and (2) $\delta \succsim n^{-o(1)}$. For any node $\rho \in V^u$ and its locally tree-like neighborhood $T_{\leq \bar{t}}(\rho)$, define the maximum mis-classification error of a local estimator $\Phi : \ell_{T_{\leq \bar{t}}(\rho)} \to \{+,-\}$ as*

$$\mathsf{Err}(\Phi) := \max_{l \in \{+,-\}} \mathbb{P}\left(\Phi(\ell_{T_{\leq t}(\rho)}) \neq \ell(\rho) | \ell(\rho) = l\right).$$

The transition boundary for p-SBM depends on the value

$$\mathsf{SNR} = (1-\delta)\frac{n(p-q)^2}{2(p+q)}.$$

($k = 2$ in Eq. (2.9)). On the one hand, if

$$\mathsf{SNR} > 1, \tag{2.11}$$

the $\bar{t}$- local message-passing Algorithm 5 — denoted as $\hat{A}(\ell_{T_{\leq \bar{t}}(\rho)})$ — recovers the true labels of the nodes with mis-classification rate at most

$$\mathsf{Err}(\hat{A}) \leq \exp\left(-\frac{\mathsf{SNR} - 1}{2C + o_{\bar{t}}(1)}\right) \wedge \frac{1}{2}, \tag{2.12}$$

where $C > 0$ is a constant and $C \equiv 1$ if the local tree is regular. On the other hand, when

$$\mathsf{SNR} < 1, \tag{2.13}$$

for any $\bar{t}$-local estimator $\Phi : \ell_{T_{\leq \bar{t}}(\rho)} \to \{+, -\}$, the minimax mis-classification error is lower bounded as

$$\inf_{\Phi} \ \mathsf{Err}(\Phi) \geq \frac{1}{2} - C \cdot \sqrt{\frac{\delta}{1-\delta} \cdot \frac{\mathsf{SNR}}{1 - \mathsf{SNR}} \cdot \frac{(p+q)^2}{pq}} = \frac{1}{2} - C' \cdot \sqrt{\frac{\delta}{1-\delta} \cdot \frac{\mathsf{SNR}}{1 - \mathsf{SNR}}}.$$

The above lower bound in the regime $\delta = o(1)$ implies that no local algorithm using information up to depth $\bar{t}$ can do significantly better than $1/2 + \mathcal{O}(\sqrt{\delta})$, close to random guessing.

Let us compare the main result for p-SBM with the well-known result for the standard SBM with no partial label information. The boundary in Equations (2.11) and (2.13) is the phase transition boundary for the standard SBM when we plug in $\delta = 0$. This also matches the well-known Kesten-Stigum bound. For the standard SBM in $k = 2$ case, the Kesten-Stigum bound is proved to be sharp (even for global algorithms), see (Mossel et al., 2013b; Massoulié, 2014).

The interesting case is when there is a vanishing amount of revealed label information, i.e.,

$o(1) = \delta \gtrsim n^{-o(1)}$. In this case, the upper bound part of Theorem 2.2.1 states that this vanishing amount of initial information is enough to propagate the labeling information to all the nodes, above the same detection transition threshold as the vanilla SBM. However, the theoretical guarantee for the label propagation pushes beyond weak consistency (detection), explicitly interpolating between weak and strong consistency. The result provides a detailed understanding of the strength of the SNR threshold and its effect on percentage recovery guarantee, i.e., the inference guarantee. More concretely, for the regime $p = a/n, q = b/n$, the boundary

$$\mathsf{SNR} = (1-\delta)\frac{n(p-q)^2}{2(p+q)} > 1$$

which is equivalent to the setting $\frac{(a-b)^2}{2(a+b)} > \frac{1}{1-\delta}$. When $\delta = 0$, this matches the boundary for weak consistency in (Mossel et al., 2013b; Massoulié, 2014). In addition, $\mathsf{SNR} > 1 + 2\log n$ implies $\mathsf{Err}(\hat{A}) < 1/n \to 0$, which means strong consistency (recovery) in the regular tree case ($C \equiv 1$). This condition on SNR is satisfied, for instance, by taking $p = a\log n/n, q = b\log n/n$ and computing the relationship between a, b, and δ to ensure

$$\mathsf{SNR} = (1-\delta)\frac{n(p-q)^2}{2(p+q)} > 1 + 2\log n.$$

This relationship is precisely

$$\frac{\sqrt{a}-\sqrt{b}}{\sqrt{2}} > \sqrt{\frac{1+\frac{1}{2\log n}}{1-\delta}} \cdot \frac{\sqrt{a}+\sqrt{b}}{\sqrt{2(a+b)}} \gtrsim \sqrt{\frac{1}{1-\delta}}.$$

The above agrees with the scaling for strong recovery in (Abbe et al., 2014; Hajek et al., 2014).

The following sections are dedicated to proving the theorem. The upper bound is established in Corollary 2.2.1 through a linearized belief propagation that serves as a subroutine for Algorithm 5. The lower bound is established by employing the classic Le Cam's theory, as shown in Theorem 2.2.3.

Belief Propagation & Message Passing

In this section we introduce the belief propagation (BP) Algorithm 6 and motivate the new message-passing Algorithm 7 that, while being easier to analyze, mimics the behavior of BP. Algorithm 7 serves as the key building block for Algorithm 5.

Recall the definition of the partially revealed binary broadcasting tree $\mathrm{Tree}_{k=2}(\theta, d, \delta)$ with broadcasting number d. The root ρ is labeled $\ell(\cdot)$ with either $\{+, -\}$ equally likely, and the label is not revealed. The labels are broadcasted along the tree with a bias parameter $0 < \theta < 1$: for a child $v \in \mathcal{C}(u)$ of u, $\ell(v) = \ell(u)$ with probability $\frac{1+\theta}{2}$ and $\ell(v) = -\ell(u)$ with probability $\frac{1-\theta}{2}$. The tree is partially labeled in the sense that a fraction $0 < \delta < 1$ of labels are revealed for each layer and $\ell_{T_{\le t}(\rho)}$ stands for the revealed label information of tree rooted at ρ with depth $\le t$.

Let us formally introduce the BP algorithm, which is the Bayes optimal algorithm on trees. We define

$$M_i(\ell_{T_{\le i}(v)}) := \log \frac{\mathbb{P}\left(\ell(v) = +|\ell_{T_{\le i}(v)}\right)}{\mathbb{P}\left(\ell(v) = -|\ell_{T_{\le i}(v)}\right)}$$

as the belief of node v's label, and we abbreviate it as M_i when the context is clear. The belief depends on the revealed information $\ell_{T_{\le i}(v)}$. The following Algorithm 6 calculates the log ratio $M_t(\ell_{T_{\le t}(\rho)})$ based on the revealed labels up to depth t, recursively, as shown in Figure 7. The Algorithm is derived through Bayes' rule and simple algebra.

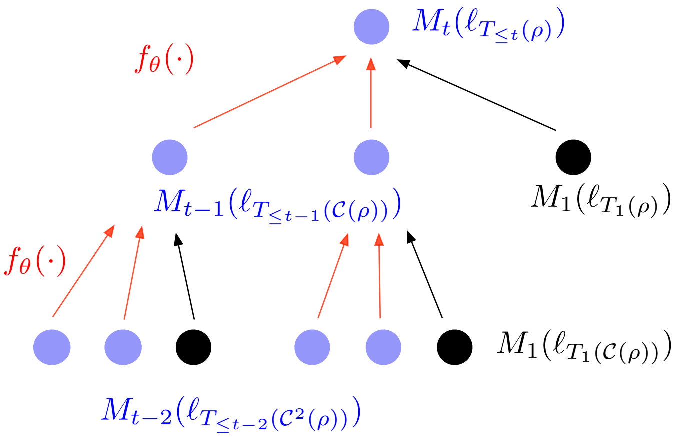
$M_t(\ell_{T_{\leq t}(\rho)})$
$f_\theta(\cdot)$
$M_{t-1}(\ell_{T_{\leq t-1}(\mathcal{C}(\rho))})$
$M_1(\ell_{T_1(\rho)})$
$f_\theta(\cdot)$
$M_1(\ell_{T_1(\mathcal{C}(\rho))})$
$M_{t-2}(\ell_{T_{\leq t-2}(\mathcal{C}^2(\rho))})$

Figure 7: Illustration of recursion in Eq. (2.14) for messages on a d-regular tree. Here $d = 3$ with two unlabeled children ($(1-\delta)d = 2$, denoted by blue) and one labeled child ($\delta d = 1$, denoted by black), and the depth is 2. $\mathcal{C}^t(\rho)$ denotes depth t children of the root ρ. The red arrows correspond to messages received from the labeled children and black arrow are from the unlabeled children.

Algorithm 6 Belief Propagation (BP) on Partially Labeled Binary Broadcasting Tree

```
Data: A partially labeled tree $T_{\leq t}(\rho)$ with depth $t$, with labels $\ell_{T_{\leq t}(\rho)}$, the root label $\ell(\rho)$
        is unknown.
Result: The logit of the posterior probability $M_t(\ell_{T_{\leq t}(\rho)}) = \log \frac{\mathbb{P}\left(\ell(\rho)=+|\ell_{T_{\leq t}(\rho)}\right)}{\mathbb{P}\left(\ell(\rho)=-|\ell_{T_{\leq t}(\rho)}\right)}$.
Initialization: $i = 1$, and $M_0(\ell_{T_{\leq 0}(v)}) = 0$, $M_1(\ell_{T_{\leq 1}(v)}) = \left(N_{\mathcal{C}^{\mathrm{l}}(v)}(+) - N_{\mathcal{C}^{\mathrm{l}}(v)}(-)\right) \log \frac{1+\theta}{1-\theta}$,
$\forall v \in T_{\leq t}(\rho)$
while $i \leq t$ do
    focus on $(t-i)$-th layer for $v \in \mathcal{C}_{t-i}(\rho)$ and $v$ unlabeled do
        update messages for the subtree:
            $$M_i(\ell_{T_{\leq i}(v)}) = M_1(\ell_{T_1(v)}) + \sum_{u \in \mathcal{C}^{\mathrm{u}}(v)} f_\theta\left(M_{i-1}(\ell_{T_{\leq i-1}(u)})\right) \quad (2.14)$$
        move one layer up: $i = i + 1$
    end
end
output $M_t(\ell_{T_{\leq t}(\rho)})$.
```

Here $f_\theta(\cdot)$ is the function defined in equation (2.10). The computational complexity of this algorithm is

$$\mathcal{O}\left(\frac{(\delta d+1)[(1-\delta)d]^t - d}{(1-\delta)d-1}\right).$$

While the method is Bayes optimal, the density of the messages M_i is difficult to analyze, due to the dependence on revealed labels and the non-linearity of f_θ. However, the following linearized version, Algorithm 7, shares many theoretical similarities with Algorithm 6, and is easier to analyze. Both Algorithms 6, 7 require the prior knowledge of θ.

Algorithm 7 Approximate Message Passing (AMP) on Partially Labeled Binary Broadcasting Tree

```
Data: A partially labeled tree $T_{\le t}(\rho)$ with depth $t$, with labels $\ell_{T_{\le t}(\rho)}$, the root label $\ell(\rho)$
      is unknown.
Result: Label $\ell(\rho) = \mathrm{sign}(M_t(\ell_{T_{\le t}(\rho)}))$.
Initialization: $i = 1$, and $M_0(\ell_{T_{\le 0}(v)}) = 0$, $M_1(\ell_{T_{\le 1}(v)}) = \Big(N_{\mathcal{C}^{\mathrm{l}}(v)}(+) - N_{\mathcal{C}^{\mathrm{l}}(v)}(-)\Big) \log\frac{1+\theta}{1-\theta}$,
$\forall v \in T_{\le t}(\rho)$
while $i \le t$ do
    focus on $(t-i)$-th layer  for $v \in \mathcal{C}_{t-i}(\rho)$ and $v$ unlabeled do
        update messages for the subtree:
            $$M_i(\ell_{T_{\le i}(v)}) = M_1(\ell_{T_1(v)}) + \theta \cdot \sum_{u \in \mathcal{C}^{\mathrm{u}}(v)} M_{i-1}(\ell_{T_{\le i-1}(u)})$$
        move one layer up: $i = i + 1$
    end
end
output $\ell(\rho) = \mathrm{sign}(M_t(\ell_{T_{\le t}(\rho)}))$.
```

Algorithm 7 can also be viewed as a weight-adjusted majority vote algorithm. We will prove in the next two sections that BP and AMP achieve the same transition boundary in the following sense. Above a certain threshold, the AMP algorithm succeeds, which implies that the optimal BP algorithm will also work. Below the same threshold, even the optimal BP algorithm will fail, and so does the AMP algorithm.

Concentration Phenomenon on Messages

We now prove Theorem 2.2.2, which shows the concentration of measure phenomenon for messages defined on the broadcasting tree. We focus on a simpler case of regular local trees, and the result will be generalized to Galton-Watson trees with a matching branching number.

We state the result under a stronger condition $\delta d \geq 1$. In the case when $\delta d = o(1)$, a separate trick, described in Remark 2.2.1 below, of aggregating the δ information in a subtree will work.

Theorem 2.2.2 (Concentration of Messages for AMP)**.** *Consider the Approximate Message Passing (AMP) Algorithm 7 on the binary-label broadcasting tree $\text{Tree}_{k=2}(\theta, d, \delta)$. Assume $\delta d \geq$. Define parameters $\{\mu_t, \sigma_t^2\}_{t\geq 0}$ as*

$$\mu_t = \mu_1 + \alpha \cdot \mu_{t-1}, \tag{2.15}$$

$$\sigma_t^2 = \sigma_1^2 + \alpha \cdot \sigma_{t-1}^2 + \alpha \cdot \mu_{t-1}^2, \tag{2.16}$$

with the initialization

$$\mu_1 = \theta \delta d \cdot \log \frac{1+\theta}{1-\theta}, \quad \sigma_1^2 = \delta d \cdot \log^2 \frac{1+\theta}{1-\theta},$$
$$\alpha := (1-\delta)\theta^2 d.$$

The explicit formulas for μ_t and σ_t^2 are

$$\mu_t = \frac{\alpha^t - 1}{\alpha - 1} \cdot \mu_1, \tag{2.17}$$

$$\sigma_t^2 = \frac{\alpha^t - 1}{\alpha - 1} \cdot \sigma_1^2 + \frac{\frac{\alpha^{2t} - \alpha^{t+1} + \alpha^t - \alpha}{\alpha - 1} - 2(t-1)\alpha^t}{(\alpha-1)^2} \cdot \mu_1^2. \tag{2.18}$$

For a certain depth t, conditionally on $\ell(\rho) = +$, the messages in Algorithm 7 concentrate as

$$\mu_t - x \cdot \sigma_t \leq M_t(\ell_{T_{\leq t}(\rho)}) \leq \mu_t + x \cdot \sigma_t,$$

and conditionally on $\ell(\rho) = -$,

$$-\mu_t - x \cdot \sigma_t \leq M_t(\ell_{T_{\leq t}(\rho)}) \leq -\mu_t + x \cdot \sigma_t,$$

both with probability at least $1 - 2\exp(x^2/2)$.

Using Theorem 2.2.2, we establish the following positive result for approximate message-passing.

Corollary 2.2.1 (Recovery Proportions for AMP, $\alpha > 1$)**.** *Assume*

$$\alpha := (1-\delta)\theta^2 d > 1,$$

and for any t define

$$\epsilon(t) = \frac{(\alpha-1)^2}{\theta^2 \delta d} \frac{1}{\alpha^t - 1} + \mathcal{O}(\alpha^{-t}), \quad \textit{with} \quad \lim_{t \to \infty} \epsilon(t) = 0.$$

Algorithm 7 recovers the label of the root node with probability at least

$$1 - \exp\left(-\frac{\alpha - 1}{2(1+\epsilon(t))}\right),$$

and its computational complexity is

$$\mathcal{O}\left(\frac{(\delta d + 1)[(1-\delta)d]^t - d}{(1-\delta)d - 1}\right).$$

Remark 2.2.1. *For the sparse case $\delta d = o(1)$, we employ the following technique. Take $t_0 > 0$ to be the smallest integer such that $\delta[(1-\delta)d]^{t_0} > 1$. For each leaf node v, open a* *depth t_0 subtree rooted at v, with the number of labeled nodes* $\mathsf{Poisson}(\delta[(1-\delta)d]^{t_0})$. *Then we have the following parameter updating rule*

$$\mu_t = \alpha \cdot \mu_{t-1}, \quad \sigma_t^2 = \alpha \cdot \sigma_{t-1}^2 + \alpha \cdot \mu_{t-1}^2,$$

with initialization

$$\begin{aligned} &\mu_1 = \theta^{t_0} \cdot \log \frac{1+\theta}{1-\theta}, \quad \sigma_1^2 = \log^2 \frac{1+\theta}{1-\theta}, \\ &\alpha := (1-\delta)\theta^2 d. \end{aligned}$$

The explicit formulas for μ_t and σ_t^2 based on the above updating rules are

$$\mu_t = \alpha^{t-1} \cdot \mu_1, \quad \sigma_t^2 = \alpha^{t-1} \cdot \sigma_1^2 + \frac{\alpha^{t-1}(\alpha^{t-1}-1)}{\alpha - 1} \cdot \mu_1^2.$$

Corollary 2.2.1 will change as follows: the value $\epsilon(t)$ is now

$$\epsilon(t) = \frac{1}{\theta^{2t_0}} \frac{1}{\alpha^{t-1}}, \quad \textit{with} \quad \lim_{t\to\infty} \epsilon(t) = 0.$$

This slightly modified algorithm recovers the label of the root node with probability at least $1 - \exp\left(-\frac{\alpha-1}{2(1+\epsilon(t))}\right)$.

Lower Bound for Local Algorithms: Le Cam's Method

In this section we show that the SNR threshold in Theorem 2.2.1 and Corollary 2.2.1 is sharp for all local algorithms. The limitation for local algorithms is proved along the lines of Le Cam's method. If we can show a small upper bound on total variation distance between two tree measures $\mu_{\ell_{\leq t}(+)}, \mu_{\ell_{\leq t}(-)}$, then no algorithm utilizing the information on the tree can distinguish these two measures well. Theorem 2.2.3 formalizes this idea.

Theorem 2.2.3 (Limits of Local Algorithms)**.** *Consider the following two measures of revealed labels defined on trees: $\mu^+_{\ell_{T_{\leq t}(\rho)}}, \mu^-_{\ell_{T_{\leq t}(\rho)}}$. Assume that $\delta d > 1$, $(1-\delta)\theta^2 d < 1$, and*

$2\delta d \log\left(1+\frac{4\theta^2}{1-\theta^2}\right) < [1-(1-\delta)\theta^2 d]^2$. *Then for any* $t > 0$, *the following bound on total variation holds*

$$d_{\rm TV}^2\left(\mu^+_{\ell_{T_{\leq t}(\rho)}}, \mu^-_{\ell_{T_{\leq t}(\rho)}}\right) \leq \frac{2\delta d \log\left(1+\frac{4\theta^2}{1-\theta^2}\right)}{1-(1-\delta)\theta^2 d}.$$

The above bound implies

$$\inf_{\Phi} \sup_{l(\rho)\in\{+,-\}} \mathbb{P}\left(\Phi(\ell_{T_{\leq t}(\rho)}) \neq \ell(\rho)\right) \geq \frac{1}{2} - C \cdot \left\{\frac{\delta d \log\left(1+\frac{4\theta^2}{1-\theta^2}\right)}{1-(1-\delta)\theta^2 d}\right\}^{1/2},$$

where $\Phi : \ell_{T \leq t}(\rho) \to \{+,-\}$ *is any estimator mapping the revealed labels in the local tree to a decision, and* $C > 0$ *is some universal constant.*

We defer the proof of the Theorem 2.2.3 to Appendix. Theorem 2.2.3 assures the optimality of Algorithm 7.

2.2.4. Growing Number of Communities

In this section, we extend the algorithmic and theoretical results to p-SBM with general k. There is a distinct difference between the case of large k and $k = 2$: there is a factor gap between the boundary achievable by local and global algorithms.

The main Algorithm that solves p-SBM for general k is still Algorithm 5, but this time it takes Algorithm 8 as a subroutine. We will first state Theorem 2.2.4, which summarizes the main result.

p-SBM Transition Thresholds

The transition boundary for partially labeled stochastic block model depends on the critical value SNR defined in Equation (2.9).

Theorem 2.2.4 (Transition Thresholds for p-SBM: general k). *Assume (1)* $np \asymp nq \precsim n^{o(1)}$, *(2)* $\delta \succsim n^{-o(1)}$, *(3)* $k \precsim n^{o(1)}$, *and consider the partially labeled stochastic block* *model $G(V, E)$ and the revealed labels $\ell(V^l)$. For any node $\rho \in V^u$ and its locally tree-like neighborhood $T_{\le \bar{t}}(\rho)$, define the maximum mis-classification error for a local estimator $\Phi : \ell_{T_{\le \bar{t}}(\rho)} \to [k]$ as*

$$\mathsf{Err}(\Phi) := \max_{l \in [k]} \mathbb{P}\left(\Phi(\ell_{T_{\le t}(\rho)}) \neq \ell(\rho) | \ell(\rho) = l\right).$$

On the one hand, if

$$\mathsf{SNR} > 1, \tag{2.19}$$

the $\bar{t}$- local message-passing Algorithm 5, denoted by $\hat{A}(\ell_{T_{\le \bar{t}}(\rho)})$, recovers the true labels of the nodes, with mis-classification rate at most

$$\mathsf{Err}(\hat{A}) \le (k-1)\exp\left(-\frac{\mathsf{SNR}-1}{2C + o_{\bar{t}}(1)}\right) \wedge \frac{k-1}{k}, \tag{2.20}$$

where $C \equiv 1$ if the local tree is regular. On the other hand, if

$$\mathsf{SNR} < \frac{1}{4}, \tag{2.21}$$

for any $\bar{t}$-local estimator $\Phi : \ell_{T_{\le \bar{t}}(\rho)} \to [k]$, the minimax mis-classification error is lower bounded as

$$\begin{aligned} \inf_{\Phi} \ \mathsf{Err}(\Phi) &\ge \frac{1}{2}\left(1 - C \cdot \frac{\delta}{1-\delta} \cdot \frac{\mathsf{SNR}}{1 - 4 \cdot \mathsf{SNR}} \cdot \frac{(p+q)(q + (p-q)/k)}{pq} - \frac{1}{k}\right) \\ &> \frac{1}{2} - C' \frac{\delta}{1-\delta} \cdot \frac{\mathsf{SNR}}{1 - 4 \cdot \mathsf{SNR}} \vee \frac{1}{k}, \end{aligned}$$

where $C = C' \equiv 1$ if the local tree is regular.

When $\delta = o(1)$ and $k > 2$, the above lower bound says that no local algorithm (that uses information up to depth $\bar{t}$) can consistently estimate the labels with vanishing error.

As we did for $k = 2$, let us compare the main result for p-SBM with the well-known result for the standard SBM with no partial label information. The boundary in Equation (2.19) matches the detection bound in (Abbe and Sandon, 2015b) for standard SBM when we plug in $\delta = 0$, which also matches the well-known Kesten-Stigum (K-S) bound. In contrast to the case $k = 2$, it is known that the K-S bound is not sharp when k is large, i.e., there exists an algorithm which can succeed below the K-S bound. A natural question is whether K-S bound is sharp within a certain local algorithm class. As we show in Equation (2.21), below a quarter of the K-S bound, the distributions (indexed by the root label) on the revealed labels for the local tree are bounded in the total variation distance sense, implying that no local algorithm can significantly push below the K-S bound. In summary, $1/4 \leq \mathsf{SNR} \leq 1$ is the limitation for local algorithms. Remarkably, it is known in the literature (Chen and Xu, 2014; Abbe and Sandon, 2015b) that information-theoretically the limitation for global algorithms is $\mathsf{SNR} = \mathcal{O}^*(1/k)$. This suggests a possible computational and statistical gap as k grows.

Belief Propagation & Message Passing

In this section, we investigate the message-passing Algorithm 8 for p-SBM with k blocks, corresponding to multi-label broadcasting trees. Denote $X_t^{(i)}(\ell_{T_{\leq t}(v)}) = \mathbb{P}\left(\ell(v) = i | \ell_{T_{\leq t}(v)}\right)$. For $u \in \mathcal{C}(v)$,

$$\mathbb{P}\left(\ell(u) = \ell(v) | \ell(v)\right) = \theta + \frac{1-\theta}{k}$$

$$\mathbb{P}\left(\ell(u) = l \in [k] \backslash \ell(v) | \ell(v)\right) = \frac{1-\theta}{k}.$$

For any $j \neq i \in [k]$ and general t, the following Lemma describes the recursion arising from the Bayes theorem.

Lemma 2.2.1. *It holds that*

$$\log \frac{X_t^{(i)}(\ell_{T_{\leq t}(v)})}{X_t^{(j)}(\ell_{T_{\leq t}(v)})} = \log \frac{X_1^{(i)}(\ell_{T_1(v)})}{X_1^{(j)}(\ell_{T_1(v)})} + \sum_{u \in \mathcal{C}^{\mathrm{u}}(v)} \log \frac{1 + \frac{k\theta}{1-\theta} X_{t-1}^{(i)}(\ell_{T_{\leq t-1}(u)})}{1 + \frac{k\theta}{1-\theta} X_{t-1}^{(j)}(\ell_{T_{\leq t-1}(u)})}.$$

The above belief propagation formula for $X_t^{(i)}(\ell_{T_{\leq t}(v)})$ is exact. However, it turns out analyzing the density of $X_t^{(i)}(\ell_{T_{\leq t}(v)})$ is hard. Inspired by the "linearization" trick for $k = 2$, we analyze the following linearized message-passing algorithm.

```
Algorithm 8 Approximate Message Passing on Partially Labeled $k$-Broadcasting Tree
Data: A partially labeled tree $T_{\leq t}(\rho)$ with depth $t$ and labels $\ell_{T_{\leq t}(\rho)}$, fixed $j \in [k]$.
Result: The messages $M_t^{(i \to j)}(\ell_{T_{\leq t}(v)})$, for any $i \in [k]/j$.
initialization: s = 1, and $M_0(\ell_{T_{\leq 0}(v)}) = 0, M_1^{(i \to j)}(\ell_{T_{\leq 1}(v)}) = \Big(N_{\mathcal{C}^{\rm l}(v)}(i) - N_{\mathcal{C}^{\rm l}(v)}(j)\Big) \log\Big(1 + \frac{k\theta}{1-\theta}\Big), \forall v, i \neq j$
while $s \leq t$ do
    focus on $(t-s)$-th layer  for $v \in \mathcal{C}_{t-s}(\rho)$ and $v$ unlabeled do
        update messages for the subtree: $M_s^{(i \to j)}(\ell_{T_{\leq s}(v)}) = M_1^{(i \to j)}(\ell_{T_1(v)}) + \theta \cdot \sum_{u \in \mathcal{C}^{\rm u}(v)} M_{s-1}^{(i \to j)}(\ell_{T_{\leq s-1}(u)})$  move one layer up: $s = s + 1$
    end
end
If $\max_{i \in [k]/j} M_t^{(i \to j)}(\ell_{T_{\leq t}(\rho)}) > 0$, output $\ell(\rho) = \arg\max_{i \in [k]/j} M_t^{(i \to j)}(\ell_{T_{\leq t}(\rho)})$; Else output $\ell(\rho) = j$.
```

For p-SBM with k blocks, Algorithm 5, which uses the above Algorithm 8 as a sub-routine, will succeed in recovering the labels in the regime above the threshold (2.19). The theoretical justification is given in the following sections.

Concentration Phenomenon on Messages

As in the case $k = 2$, here we provide the concentration result on the distribution of approximate messages recursively calculated based on the tree.

Theorem 2.2.5 (Concentration of Messages for k-AMP, $(1-\delta)\theta^2 d > 1$)**.** *Consider the Approximate Message Passing (AMP) Algorithm 8 on the k-label broadcasting tree $\mathrm{Tree}_k(\theta, d, \delta)$.*

Assume $\delta d \geq 1$. With the initial values

$$\mu_1 = \theta\delta d \cdot \log\left(1 + \frac{k\theta}{1-\theta}\right), \quad \sigma_1^2 = \delta d \cdot \log^2\left(1 + \frac{k\theta}{1-\theta}\right)$$

and the factor parameter

$$\alpha := (1-\delta)\theta^2 d,$$

the recursion of the parameters μ_t, σ_t^2 follows as in Eq. (2.15).

For a certain depth t, conditionally on $\ell(v) = l$, the moment generating function for $M_t^{(i\to j)}(\ell_{T_{\leq t}(v)})$ is upper bounded as

$$\Psi_{M_t^{(i\to j)}}(\lambda) \leq \begin{cases} e^{\lambda\mu_t} e^{\frac{\lambda^2\sigma_t^2}{2}}, & i = l \\ e^{\frac{\lambda^2\sigma_t^2}{2}}, & i, j \neq l \\ e^{-\lambda\mu_t} e^{\frac{\lambda^2\sigma_t^2}{2}}, & j = l \end{cases}$$

The message-passing Algorithm 8 succeeds in recovering the label with probability at least $1 - (k-1)\exp\left(-\frac{\alpha}{2(1+o(1))}\right)$ when $(1-\delta)\theta^2 d > 1$.

Again, from Theorem 2.2.5 we can easily get the following recovery proportion guarantee. For the message-passing Algorithm 7, assume

$$\alpha := (1-\delta)\theta^2 d > 1,$$

and define for any t

$$\epsilon(t) = \frac{(\alpha-1)^2}{\theta^2\delta d}\frac{1}{\alpha^t - 1} + \mathcal{O}(\alpha^{-t}), \quad \text{with} \quad \lim_{t\to\infty}\epsilon(t) = 0.$$

Then Algorithm 8 recovers the label of the root node with probability at least

$$1-(k-1)\exp\left(-\frac{(1-\delta)\theta^2 d-1}{2(1+\epsilon(t))}\right)$$

with time complexity

$$\mathcal{O}\left((k-1)\frac{(\delta d+1)[(1-\delta)d]^t-d}{(1-\delta)d-1}\right).$$

Multiple Testing Lower Bound on Local Algorithm Class

We conclude the theoretical study with a lower bound for local algorithms for k-label broadcasting trees. We bound the distributions of leaf labels, indexed by different root colors and show that in total variation distance, the distributions are indistinguishable (below the threshold in equation (2.21)) from each other as δ vanishes.

Theorem 2.2.6 (Limitation for Local Algorithms)**.** *Consider the following measures of revealed labels defined on trees indexed by the root's label: $\mu^{(i)}_{\ell_{T_{\leq t}(\rho)}}, i \in [k]$. Assume $\delta d > 1$, $(1-\delta)\theta^2 d < 1/4$ and*

$$2\delta d \log\left(1+\theta^2\left(\frac{1}{\theta+\frac{1-\theta}{k}}+\frac{1}{\frac{1-\theta}{k}}\right)\right) < [1-4(1-\delta)\theta^2 d]^2.$$

Then for any $t > 0$, the following bound on the χ^2 distance holds:

$$\begin{aligned}\max_{i,j\in[k]} \log\left(1+d_{\chi^2}\left(\mu^{(i)}_{\ell_{T_{\leq t}(\rho)}}, \mu^{(j)}_{\ell_{T_{\leq t}(\rho)}}\right)\right) &\leq \frac{2\delta d\log\left(1+\theta^2\left(\frac{1}{\theta+\frac{1-\theta}{k}}+\frac{1}{\frac{1-\theta}{k}}\right)\right)}{1-4(1-\delta)\theta^2 d} \\ &\leq k\cdot\frac{2\delta\theta^2 d}{1-4(1-\delta)\theta^2 d}\left(\frac{1}{1-\theta}+\frac{1}{k\theta+1-\theta}\right).\end{aligned}$$

Furthermore, it holds that

$$\inf_{\Phi}\sup_{l(\rho)\in[k]} \mathbb{P}\left(\Phi(\ell_{T_{\leq t}(\rho)}) \neq \ell(\rho)\right) \geq \frac{1}{2}\left(1-\frac{2\delta\theta^2 d}{1-4(1-\delta)\theta^2 d}\left(\frac{1}{1-\theta}+\frac{1}{k\theta+1-\theta}\right)-\frac{1}{k}\right),$$

where $\Phi : \ell_{T\leq t}(\rho) \to [k]$ is any local estimator mapping from the revealed labels to a *decision.*

The proof is based on a multiple testing argument in Le Cam's minimax lower bound theory. We would like to remark that condition $4 \cdot (1-\delta)\theta^2 d < 1$ can be relaxed to

$$(1-\delta)\theta^2 d \cdot \left(1 + 3(1-\theta)(1-\frac{2}{k})\right) < 1.$$

2.2.5. Numerical Studies

In this section we apply our approximate message-passing Algorithm 5 to the political blog dataset (Adamic and Glance, 2005), with a total of 1222 nodes. In the literature, the state-of-the-art result for a global algorithm appears in (Jin et al., 2015), where the mis-classification rate is $58/1222 = 4.75\%$. Here we run our message-passing Algorithm 5 with three different settings $\delta = 0.1, 0.05, 0.025$, replicating each experiment 50 times (we sample the revealed nodes independently in 50 experiments for each δ specification). As a benchmark, we compare our results to the spectral algorithm on the $(1-\delta)n$ sub-network. For our message-passing algorithm, we look at the local tree with depth 1 to 5. The results are summarized as boxplots in Figure 8. The left figure illustrates the comparison of AMP with depth 1 to 5 and the spectral algorithm, with red, green, blue boxes corresponding to $\delta = 0.025, 0.05, 0.1$, respectively. The right figure zooms in on the left plot with only AMP depth 2 to 4 and spectral, to better emphasize the difference. Remark that if we only look at depth 1, some of the nodes may have no revealed neighbors. In this setting, we classify this node as wrong (this explains why depth-1 error can be larger than 1/2). We present in this paragraph some of the statistics of the experiments, extracted from the above Figure 8. In the case $\delta = 0.1$, from depth 2-4, the AMP algorithm produces the mis-classification error rate (we took the median over the experiments for robustness) of $6.31\%, 5.22\%, 5.01\%$, while the spectral algorithm produces the error rate 6.68%. When $\delta = 0.05$, i.e. about 60 node labels revealed, the error rates are $7.71\%, 5.44\%, 5.08\%$ for the AMP algorithm with depth 2 to 4, contrasted to the spectral algorithm error 6.66%. In a

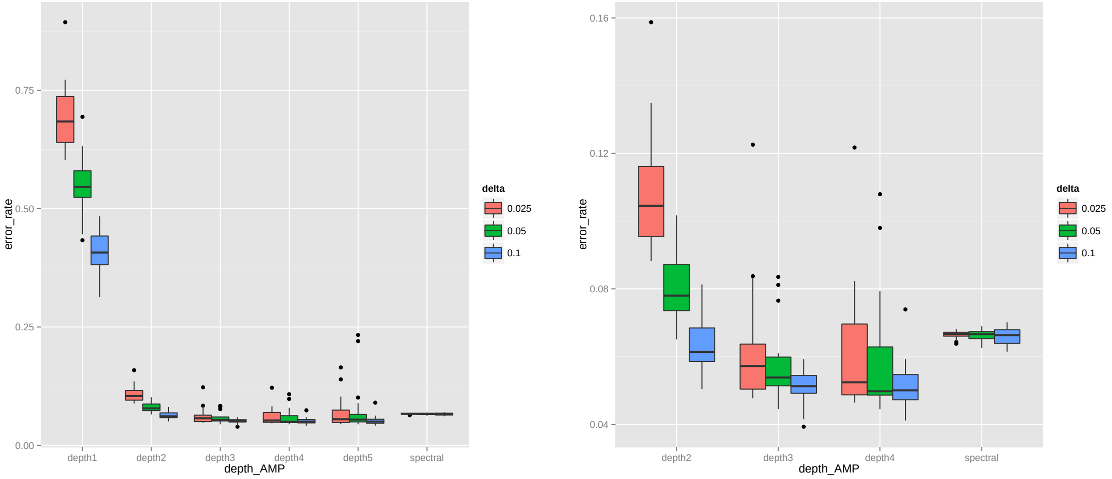

Figure 8: AMP algorithm on Political Blog Dataset.

more extreme case $\delta = 0.025$ when there are only ~ 30 node labels revealed, AMP depth 2-4 has error $10.20\%, 5.71\%, 5.66\%$, while spectral is 6.63%. In general, the AMP algorithm with depth 3-4 uniformly beats the vanilla spectral algorithm. Note our AMP algorithm is a distributed decentralized algorithm that can be run in parallel. We acknowledge that the error $\sim 5\%$ (when δ is very small) is still slightly worse than the state-of-the-art degree-corrected SCORE algorithm in (Jin et al., 2015), which is 4.75%.

2.3. Watts-Strogatz Small World Network

2.3.1. Introduction

The “small-world” phenomenon aims to describe real-world complex networks that exhibit both high clustering and short average length properties. While most of the pairs of nodes are not friends, any node can be reached from another in a small number of hops. The Watts-Strogatz (WS) model, introduced in (Watts and Strogatz, 1998; Newman and Watts, 1999), is a popular generative model for networks that exhibit the small-world phenomenon. The WS model interpolates between the two extremes—the regular lattice graph for high clustering on the one hand, and the random graph exhibiting the short chain property on the other. Considerable effort has been spent on studying the asymptotic statistical behavior (degree distribution, average path length, clustering coefficient, etc.) and the empirical performance of the WS model (Watts, 1999; Amaral et al., 2000; Barrat and Weigt, 2000; Latora and Marchiori, 2001; Van Der Hofstad, 2009). Successful applications of the WS model have been found in a range of disciplines, such as psychology (Milgram, 1967), epidemiology (Moore and Newman, 2000), medicine and health (Stam et al., 2007), to name a few. In one of the first algorithmic studies of small-world networks, Kleinberg (2000) investigated the theoretical difficulty of finding the shortest path between any two nodes when one is restricted to use local algorithms, and further extended the small-world notion to long range percolation on graphs (Benjamini and Berger, 2000; Coppersmith et al., 2002).

In the present thesis, we study detection and reconstruction of small-world networks. Our focus is on both statistical and computational aspects of these problems. Given a network, the first challenge is to detect whether it enjoys the small-world property (i.e., high clustering and short average path), or whether the observation may simply be explained by the Erdős-Rényi random graph (the null hypothesis). The second question is concerned with the reconstruction of the neighborhood structure if the network does exhibit the small-world phenomenon. In the language of social network analysis, the detection problem corresponds to detecting the existence of strong ties (close friend connections) in the presence of weak ties (random connections). The more difficult reconstruction problem corresponds to distinguishing between strong and weak ties. Statistical and computational difficulties of both detection and reconstruction are due to the latent high-dimensional permutation matrix which blurs the natural ordering of the ring structure on the nodes.

Let us parametrize the WS model in the following way: the number of nodes is denoted by n, the neighborhood size by k, and the rewiring probability by β. Provided the adjacency matrix $A \in \mathbb{R}^{n \times n}$, we are interested in identifying the tuples (n, k, β) when detection and reconstruction of the small-world random graph is possible. Specifically, we focus on the following two questions.

Detection Given the adjacency matrix A up to a permutation, when (in terms of n, k, β) and how (in terms of procedures) can one statistically distinguish whether it is a small-world graph ($\beta < 1$), or a random graph with matching degree ($\beta = 1$). What can be said if we restrict our attention to computationally efficient procedures?

Reconstruction Once the presence of the neighborhood structure is confirmed, when (in terms of n, k, β) and how (in terms of procedures) can one estimate the deterministic neighborhood structure? If one only aims to estimate the structure asymptotically consistently, are there computationally efficient procedures, and what are their limitations?

We address the above questions by presenting a phase diagram in Figure 9. The phase diagram divides the parameter space into four disjoint regions according to the difficulty of the problem. We propose distinct methods for the regions where solutions are possible.

Why the Small-World Model?

Finding and analyzing the appropriate statistical models for real-world complex networks is one of the main themes in network science. Many real empirical networks—for example, internet architecture, social networks, and biochemical pathways—exhibit two features simultaneously: high clustering among individual nodes and short distance between any two nodes. Consider the local tree rooted at a person. The high clustering property suggests prevalent existence of triadic closure, which significantly reduces the number of reachable people within a certain depth (in contrast to the regular tree case where this number grows exponentially with the depth), contradicting the short average length property. In a path-breaking paper, Watts and Strogatz (1998) provided a mathematical model that resolves the above seemingly contradictory notions. The solution is surprisingly simple — interpolating between structural ring lattice graph and a random graph. The ring lattice provides the strong ties (i.e., homophily, connection to people who are similar to us) and triadic closure, while the random graph generates the weak ties (connection to people who are otherwise far-away), preserving the local-regular-branching-tree-like structure that induces short paths between pairs.

We remark that one can find different notions of "small-worldness" in the existing literature. For instance, "small-world" refers to "short chain" in (Milgram, 1967; Kleinberg, 2000), while it refers to both "high clustering" and "short chain" in (Watts and Strogatz, 1998). We adopt the latter definition in the current study.

Rewiring Model

Let us now define the WS model. Consider a ring lattice with n nodes, where each node is connected with its k nearest neighbors ($k/2$ on the left and $k/2$ on the right, k even for convenience). The rewiring process consists of two steps. First, we erase each currently connected edge with probability β, independently. Next, we reconnect each edge pair with probability $\beta\frac{k}{n-1}$, allowing multiplicity.[4] The observed symmetric adjacency matrix $A \in \{0,1\}^{n\times n}$ has the following structure under some unobserved permutation matrix $P_\pi \in$

[4] The original rewiring process in Watts and Strogatz (1998) does not allow multiplicity; however, for the simplicity of technical analysis, we focus on reconnection allowing multiplicity. These two rewiring processes are asymptotically equivalent.

$\{0,1\}^{n\times n}$. For $1 \leq i < j \leq n$, the probability that

$$[P_\pi A P_\pi^T]_{ij} = 1$$

is given by

(i) $1-\beta(1-\beta\frac{k}{n-1})$, if $0 < |i-j| \leq \frac{k}{2} \mod n-1-\frac{k}{2}$

(ii) $\beta\frac{k}{n-1}$ otherwise,

and the entries are independent of each other. Equivalently, we have for $1 \leq i < j \leq n$

$$A_{ij} = \kappa\left([P_\pi B P_\pi^T]_{ij}\right), \tag{2.22}$$

where $\kappa(\cdot)$ is the entry-wise i.i.d. Markov channel,

$$\begin{aligned} \kappa(0) &\sim \mathsf{Bernoulli}\left(\beta\frac{k}{n-1}\right), \\ \kappa(1) &\sim \mathsf{Bernoulli}\left(1-\beta(1-\beta\frac{k}{n-1})\right), \end{aligned}$$

and $B \in \{0,1\}^{n\times n}$ indicates the support of the structural ring lattice

$$B_{ij} = \begin{cases} 1, & \text{if } 0 < |i-j| \leq \frac{k}{2} \mod n-1-\frac{k}{2} \\ 0, & \text{otherwise} \end{cases}. \tag{2.23}$$

We denote by $\mathsf{WS}(n,k,\beta)$ the distribution of the random graph generated from the rewiring model, and denote by $\mathsf{ER}(n,\frac{k}{n-1})$ the Erdős-Rényi random graph distribution (with matching average degree k). Remark that if $\beta = 1$, the small-world graph $\mathsf{WS}(n,k,\beta)$ reduces to $\mathsf{ER}(n,\frac{k}{n-1})$, with no neighborhood structure. In contrast, if $\beta = 0$, the small-world graph $\mathsf{WS}(n,k,\beta)$ corresponds to the deterministic ring lattice, without random connections. We focus on the dependence of the gap $1-\beta = o(1)$ on n and k, such that distinguishing between $\mathsf{WS}(n,k,\beta)$ and $\mathsf{ER}(n,\frac{k}{n-1})$ or reconstructing the ring lattice structure is statistically and computationally possible.

Summary of Results

The main theoretical and algorithmic results are summarized in this section. We first introduce several regions in terms of (n, k, β), according to the difficulty of the problem instance, and then we present the results using the phase diagram in Figure 9. Except for the 'impossible region', we will introduce algorithms with distinct computational properties. The 'impossible region' is defined through a lower bound, while the other regions are classified according to upper bounds on performance of respective procedures.

Impossible region: $1 - \beta \prec \sqrt{\frac{\log n}{n}} \vee \frac{\log n}{k}$. In this region, no multiple testing procedure (regardless of computational budget) can succeed in distinguishing, with vanishing error, among the class of models that includes all of $\mathsf{WS}(n, k, \beta)$ and $\mathsf{ER}(n, \frac{k}{n-1})$.

Hard region: $\sqrt{\frac{\log n}{n}} \vee \frac{\log n}{k} \preceq 1 - \beta \prec \sqrt{\frac{1}{k}} \vee \frac{\sqrt{\log n}}{k}$. It is possible to distringuish between $\mathsf{WS}(n, k, \beta)$ and $\mathsf{ER}(n, \frac{k}{n-1})$ statistically with vanishing error; however the evaluation of the test statistic (2.26) requires exponential time complexity, to the best of our knowledge.

Easy region: $\sqrt{\frac{1}{k}} \vee \frac{\sqrt{\log n}}{k} \preceq 1 - \beta \preceq \sqrt{\sqrt{\frac{\log n}{n}} \vee \frac{\log n}{k}}$. There exists an efficient spectral test that can distinguish between the small-world random graph $\mathsf{WS}(n, k, \beta)$ and the Erdős-Rényi graph $\mathsf{ER}(n, \frac{k}{n-1})$ in time nearly linear in the matrix size.

Reconstructable region: $\sqrt{\sqrt{\frac{\log n}{n}} \vee \frac{\log n}{k}} \prec 1 - \beta \preceq 1$. In this region, not only is it possible to detect the existence of the lattice structure in a small-world graph, but it is also possible to consistently reconstruct the neighborhood structure via a novel computationally efficient correlation thresholding procedure.

The following phase diagram provides an intuitive illustration of the above theoretical results. If we parametrize $k \asymp n^x, 0 < x < 1$ and $1 - \beta \asymp n^{-y}, 0 < y < 1$, each point $(x, y) \in [0, 1]^2$ corresponds to a particular problem instance with parameter bun-

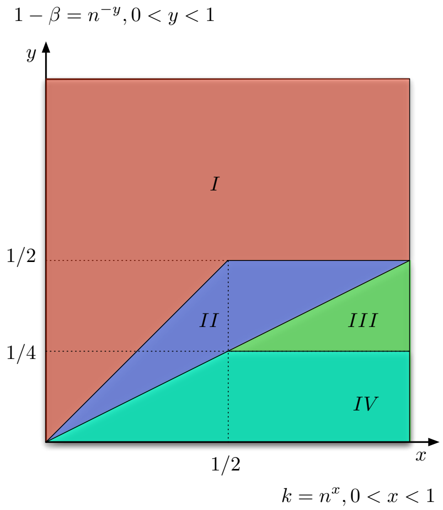

Figure 9: Phase diagram for small-world network: impossible region (red region I), hard region (blue region II), easy region (green region III), and reconstructable region (cyan region IV).

dle $(n, k = n^{x}, \beta = 1 - n^{-y})$. According to the location of (x, y), the difficulty of the problem changes; for instance, the larger the x and the smaller the y is, the easier the problem becomes. The various regions are: impossible region (red region I), hard region (blue region II), easy region (green region III), reconstructable region (cyan region IV).

Notation

$A, B, Z \in \mathbb{R}^{n\times n}$ denote symmetric matrices: A is the adjacency matrix, B is the structural signal matrix as in Equation (2.23), and $Z = A - \mathbb{E}A$ is the noise matrix. We denote the matrix of all ones by J. Notations $\preceq, \succeq, \prec, \succ$ denote the asymptotic order: $a(n) \preceq b(n)$ if and only if $\limsup_{n\to\infty} \frac{a(n)}{b(n)} \leq c$, with some constant $c > 0$, $a(n) \prec b(n)$ if and only if $\limsup_{n\to\infty} \frac{a(n)}{b(n)} = 0$. $C, C' > 0$ are universal constants that may change from line to line. For a symmetric matrix A, $\lambda_i(A)$, $1 \leq i \leq n$, denote the eigenvalues in a decreasing order. The inner-product $\langle A, B\rangle = \mathrm{tr}(A^T B)$ denotes both the Euclidian inner-product and matrix inner-product. For any integer n, $[n] := \{0, 1, \ldots, n-1\}$ denotes the index set. Denote the permutation in symmetric group $\pi \in S_n$ and its associated matrix form as P_π.

For a graph $G(V,E)$ generated from the Watts-Strogatz model $\mathsf{WS}(n,k,\beta)$ with associated permutation π, for each node $v_i \in V, 1 \leq i \leq |V|$, we denote

$$\mathcal{N}(v_i) := \left\{ v_j : 0 < |\pi^{-1}(i) - \pi^{-1}(j)| \leq \frac{k}{2} \mod n-1-\frac{k}{2} \right\},$$

the ring neighborhood of v_i before permutation π is applied.

Organization

The following sections are dedicated to the theoretical justification of the various regions in Section 2.3.1. Specifically, Section 2.3.2 establishes the boundary for the impossible region I, where the detection problem is information-theoretically impossible. We contrast the hard region II with the regions III and IV in Section 2.3.3; here, the difference arises in statistical and computational aspects of detecting the strong tie structure inside the random graph. Section 2.3.4 studies a correlation thresholding algorithm that reconstructs the neighborhood structure consistently when the parameters lie within the reconstructable region IV. We also study a spectral ordering algorithm which succeeds in reconstruction in a part of region III. Whether the remaining part of region III admits a recovery procedure is an open problem. Additional further directions are listed in Section 2.3.5.

2.3.2. The Impossible Region: Lower Bounds

We start with an information-theoretic result that describes the difficulty of distinguishing among a class of models. Theorem 2.3.1 below characterizes the impossible region, as in Section 2.3.1, in the language of minimax multiple testing error. The proof is postponed to Appendix.

Theorem 2.3.1 (Impossible Region)**.** *Consider the following statistical models: $\mathcal{P}_0$ denotes the distribution of the Erdős-Rényi random graph $\mathsf{ER}(n, \frac{k}{n-1})$, and $\mathcal{P}_\pi, \pi \in S_{n-1}$ denote distributions of the Watts-Strogatz small-world graph $\mathsf{WS}(n,k,\beta)$ as in Equation* (2.22) *with different permutations π. Consider any selector $\phi : \{0,1\}^{n\times n} \to S_{n-1} \cup \{0\}$ that maps the adjacency matrix $A \in \{0,1\}^{n\times n}$ to a decision in $S_{n-1} \cup \{0\}$. Then for any fixed* $0 < \alpha < 1/8$*,* *the following lower bound on multiple testing error holds:*

$$\varliminf_{n\to\infty} \min_{\phi} \max \left\{ \mathcal{P}_0(\phi \neq 0), \ \frac{1}{(n-1)!} \sum_{\pi \in S_{n-1}} \mathcal{P}_\pi(\phi \neq \pi) \right\} \geq 1 - 2\alpha,$$

when the parameters satisfy

$$1 - \beta \leq C_\alpha \cdot \sqrt{\frac{\log n}{n}} \quad or \quad 1 - \beta \leq C'_\alpha \cdot \frac{\log n}{k} \cdot \frac{1}{\log \frac{n \log n}{k^2}},$$

with constants C_α, C'_α that only depend on α. In other words, if

$$1 - \beta \prec \sqrt{\frac{\log n}{n}} \vee \frac{\log n}{k},$$

no multiple testing procedure can succeed in distinguishing, with vanishing error, the class of models containing all of $\mathsf{WS}(n,k,\beta)$ and $\mathsf{ER}(n, \frac{k}{n-1})$.

The missing latent random variable, the permutation matrix P_π, is the object we are interested in recovering. A permutation matrix P_π induces a certain distribution on the adjacency matrix A. Thus the parameter space of interest, including models $\mathsf{WS}(n,k,\beta)$ and $\mathsf{ER}(n, \frac{k}{n-1})$, is of cardinality $(n-1)! + 1$. Based on the observed adjacency matrix, distinguishing among the models $\mathsf{WS}(n,k,\beta)$ and $\mathsf{ER}(n, \frac{k}{n-1})$ is equivalent to a multiple testing problem. The impossible region characterizes the information-theoretic difficulty of this reconstruction problem by establishing the condition that ensures non-vanishing minimax testing error as $n, k(n) \to \infty$.

The "high dimensional" nature of this problem is mainly driven by the unknown permutation matrix, and this latent structure introduces difficulty both statistically and computationally. Statistically, via Le Cam's method, one can build a distance metric on permutation matrices using the distance between the corresponding measures (measures on adjacency matrices induced by the permutation structure). In order to characterize the intrinsic difficulty of estimating the permutation structure, one needs to understand the richness of the set of permutation matrices within certain distance to one particular element, a combinatorial task. The combinatorial nature of the problem makes the "naive" approach computationally intensive.

2.3.3. Hard vs. Easy Regions: Detection Statistics

This section studies the hard and easy regions in Section 2.3.1. First, we propose a near optimal test, the **maximum likelihood test**, that detects the ring structure above the information boundary derived in Theorem 2.3.1. However, the evaluation of the maximum likelihood test requires $\mathcal{O}(n^n)$ time complexity. The maximum likelihood test succeeds outside of region I, and, in particular, succeeds (statistically) in the hard region II. We then propose another efficient test, the **spectral test**, that detects the ring structure in time $\mathcal{O}^*(n^2)$ via the power method. The spectral test is motivated from the circulant structure of the signal matrix B (as in Eq. 2.23). The method succeeds in regions III and IV.

Theorem 2.3.2 combines the results of Lemma 2.3.1 and Lemma 2.3.2 below.

Theorem 2.3.2 (Detection: Easy and Hard Boundaries)**.** *Consider the following statistical models: $\mathcal{P}_0$ denotes the distribution of the Erdős-Rényi random graph $\mathsf{ER}(n, \frac{k}{n-1})$, and $\mathcal{P}_\pi, \pi \in S_{n-1}$ denote distributions of the Watts-Strogatz small-world graph $\mathsf{WS}(n,k,\beta)$. Consider any selector $\phi : \{0,1\}^{n\times n} \to \{0,1\}$ that maps an adjacency matrix to a binary decision.*

We say that minimax detection for the small-world random model is possible when

$$\lim_{n\to\infty} \min_{\phi} \max \left\{ \mathcal{P}_0(\phi \neq 0), \frac{1}{(n-1)!} \sum_{\pi \in S_{n-1}} \mathcal{P}_\pi(\phi \neq 1) \right\} = 0. \tag{2.24}$$

If the parameter (n,k,β) satisfies

$$\textit{hard boundary}: \quad 1-\beta \succeq \sqrt{\frac{\log n}{n}} \vee \frac{\log n}{k},$$

minimax detection is possible, and an exponential time **maximum likelihood test** (2.26) *ensures* (2.24)*. If, in addition, the parameter* (n, k, β) *satisfies*

$$\textit{easy boundary}: \quad 1-\beta \succeq \sqrt{\frac{1}{k}} \vee \frac{\sqrt{\log n}}{k},$$

then a near-linear time **spectral test** (2.28) *ensures* (2.24).

Proof of Theorem 2.3.2 consists of two parts, which will be addressed in the following two sections, respectively.

Maximum Likelihood Test

Consider the test statistic T_1 as the objective value of the following optimization

$$T_1(A) := \max_{P_\pi} \ \langle P_\pi B P_\pi^T, A \rangle, \tag{2.25}$$

where $P_\pi \in \{0,1\}^{n \times n}$ is taken over all permutation matrices and A is the observed adjacency matrix. The **maximum likelihood test** $\phi_1 : A \to \{0,1\}$ based on T_1 by

$$\begin{aligned} &\phi_1(A) \\ &= \begin{cases} 1 & \text{if } T_1(A) \geq \frac{k}{n-1} nk + 2\sqrt{\frac{k}{n-1} nk \cdot \log n!} + \frac{2}{3} \cdot \log n! \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{2.26}$$

The threshold is chosen as the rate $k^2 + \mathcal{O}\left(\sqrt{k^2 n \log \frac{n}{e}} \vee n \log \frac{n}{e}\right)$: if the objective value is of a greater order, then we believe the graph is generated from the small-world rewiring process with strong ties; otherwise we cannot reject the null, the random graph model with only weak ties.

Lemma 2.3.1 (Guarantee for Maximum Likelihood Test)**.** *The maximum likelihood test* ϕ_1 *in Equation* (2.26) *succeeds in detecting the small-world random structure when*

$$1-\beta \succeq \sqrt{\frac{\log n}{n}} \vee \frac{\log n}{k},$$

in the sense that

$$\lim_{n,k(n)\to\infty} \max \left\{ P_0(\phi_1 \neq 0),\ \frac{1}{(n-1)!}\sum_{i=1}^{(n-1)!} P_i(\phi_1 \neq 1) \right\} = 0.$$

Remark 2.3.1. *Lemma 2.3.1 can be viewed as the condition on the signal and noise separation. By solving the combinatorial optimization problem, the test statistic aggregates the signal that separates from the noise the most. An interesting open problem is: if we solve a relaxed version of the combinatorial optimization problem* (2.25) *in polynomial time, how much stronger the condition on* $1-\beta$ *needs to be to ensure power.*

Spectral Test

For the spectral test, we calculate the second largest eigenvalue of the adjacency matrix A as the test statistic

$$T_2(A) := \lambda_2(A). \tag{2.27}$$

The **spectral test** $\phi_2 : A \to \{0,1\}$ is

$$\phi_2(A) = \begin{cases} 1 & \text{if } T_2(A) \succeq \sqrt{k} \vee \sqrt{\log n} \\ 0 & \text{otherwise.} \end{cases} \tag{2.28}$$

Namely, if $\lambda_2(A)$ passes the threshold, we classify the graph as a small-world graph. Evaluation of (2.28) requires near-linear time $\mathcal{O}^*(n^2)$ in the size of the matrix.

Lemma 2.3.2 (Guarantee for Spectral Test)**.** *The second eigenvalue test* ϕ_2 *in Equation* (2.28) *satisfies*

$$\lim_{n,k(n)\to\infty} \max \left\{ P_0(\phi_2 \neq 0),\ \frac{1}{(n-1)!}\sum_{i=1}^{(n-1)!} P_i(\phi_2 \neq 1) \right\} = 0$$

whenever

$$1-\beta \succeq \sqrt{\frac{1}{k}} \vee \frac{\sqrt{\log n}}{k}.$$

The main idea behind Lemma 2.3.2 is as follows. Let us look at the expectation of the adjacency matrix,

$$\mathbb{E}A = (1-\beta)(1-\beta\frac{k}{n-1}) \cdot P_\pi^T B P_\pi + \beta\frac{k}{n-1} \cdot (J-I),$$

where J is the matrix of all ones. The main structure matrix $P_\pi^T B P_\pi$ is a permuted version of the *circulant matrix* (see e.g. (Gray, 2006)). The spectrum of the circulant matrix B is highly structured, and is of distinct nature in comparison to the noise matrix $A - \mathbb{E}A$.

2.3.4. Reconstructable Region: Fast Structural Reconstruction

In this section, we discuss reconstruction of the ring structure in the Watts-Strogatz model. We show that in the reconstructable region (region IV in Figure 9), a **correlation thresholding procedure** succeeds in reconstructing the ring neighborhood structure. As a by-product, once the neighborhood structure is known, one can distinguish between weak ties (random edges) and strong ties (neighborhood edges) for each node. A natural question is whether there is another algorithm that can work in a region (beyond region IV) where correlation thresholding fails. We show that in a certain regime with large k, a **spectral ordering procedure** outperforms the correlation thresholding procedure and succeeds in parts of regions III and IV (as depicted in Figure 10 below).

Correlation Thresholding

Consider the following **correlation thresholding procedure** for neighborhood reconstruction.

Algorithm 9 Correlation Thresholding for Neighborhood Reconstruction

```
Data: An adjacency matrix $A \in \mathbb{R}^{n \times n}$ for the graph $G(V, E)$.
Result: For each node $v_i, 1 \leq i \leq n$, an estimated set for neighborhood $\hat{\mathcal{N}}(v_i)$.
1. For each $v_i$, calculate correlations $\langle A_i, A_j \rangle$, $j \neq i$
2. Sort $\{\langle A_i, A_j \rangle, j \in [n] \backslash \{i\}\}$ in a decreasing order, select the largest $k$ ones to form the estimated set $\hat{\mathcal{N}}(v_i)$
Output: $\hat{\mathcal{N}}(v_i)$, for all $i \in [n]$
```

The following lemma proves consistency of the above Algorithm 9. Note the computational complexity is $\mathcal{O}(n \cdot \min\{\log n, k\})$ for each node using quick-sort, with a total runtime $\mathcal{O}^*(n^2)$.

Lemma 2.3.3 (Consistency of Correlation Thresholding)**.** *Consider the Watts-Strogatz random graph* $\mathsf{WS}(n, k, \beta)$. *Under the reconstructable regime IV (in Figure 9), that is,*

$$1 - \beta \succ \sqrt{\frac{\log n}{k}} \vee \left(\frac{\log n}{n}\right)^{1/4}, \tag{2.29}$$

correlation thresholding provides a consistent estimate of the neighborhood set $\mathcal{N}(v_i)$ *w.h.p in the sense that*

$$\lim_{n, k(n) \to \infty} \max_{i \in [n]} \frac{|\hat{\mathcal{N}}(v_i) \triangle \mathcal{N}(v_i)|}{|\mathcal{N}(v_i)|} = 0,$$

where $\triangle$ *denotes the symmetric set difference.*

The condition under which consistency of correlation thresholding is ensured corresponds to the reconstructable region in Figure 9. One may ask if there is another algorithm that can provide a consistent estimate of the neighborhood set beyond region IV. The answer is yes, and we will show in the following section that under the regime when k is large (for instance, $k \succeq n^{\frac{15}{16}}$), indeed it is possible to slightly improve on Algorithm 9.

Spectral Ordering

Consider the following **spectral ordering procedure**, which approximately reconstructs the ring lattice structure when k is large, i.e., $k \succ n^{\frac{7}{8}}$.

Algorithm 10 Spectral Reconstruction of Ring Structure

```
Data: An adjacency matrix $A \in \mathbb{R}^{n \times n}$ for the graph $G(V, E)$.
Result: A ring embedding of the nodes $V$.
1. Calculate top 3 eigenvectors in the SVD $A = U\Sigma U^T$. Denote second and third eigenvectors as $u \in \mathbb{R}^n$ and $v \in \mathbb{R}^n$, respectively
2. For each node $i$ and vector $A_{\cdot i} \in \mathbb{R}^n$, calculate the associated angle $\theta_i$ for the vector $(u^T A_{\cdot i}, v^T A_{\cdot i})$
Output: the sorted sequence $\{\theta_i\}_{i=1}^n$ and the corresponding ring embedding of the nodes.
        For each node $v_i$, $\hat{\mathcal{N}}(v_i)$ are the closest $k$ nodes in the ring embedding.
```

The following Lemma 2.3.4 shows that when k is large, Algorithm 10 also provides consistent reconstruction of the ring lattice. Its computational complexity is $\mathcal{O}^*(n^2)$.

Lemma 2.3.4 (Guarantee for Spectral Ordering)**.** *Consider the Watts-Strogatz graph* $\mathsf{WS}(n, k, \beta)$*. Assume k is large enough in the following sense:*

$$1 > \overline{\lim_{n,k(n)\to\infty}} \frac{\log k}{\log n} \geq \underline{\lim_{n,k(n)\to\infty}} \frac{\log k}{\log n} > \frac{7}{8}.$$

Under the regime

$$1 - \beta \succ \frac{n^{3.5}}{k^4}, \tag{2.30}$$

the spectral ordering provides consistent estimate of the neighborhood set $\mathcal{N}(v_i)$ w.h.p. in the sense that

$$\lim_{n,k(n)\to\infty} \max_{i\in[n]} \frac{|\hat{\mathcal{N}}(v_i) \triangle \mathcal{N}(v_i)|}{|\mathcal{N}(v_i)|} = 0,$$

where $\triangle$ denotes the symmetric set difference.

In Lemma 2.3.4, we can only prove consistency of spectral ordering under the technical condition that k is large. We do not believe this is due to an artifact of the proof. Even though the structural matrix (the signal) has large eigenvalues, the eigen-gap is not large enough. The spectral ordering succeeds when the spectral gap stands out over the noise

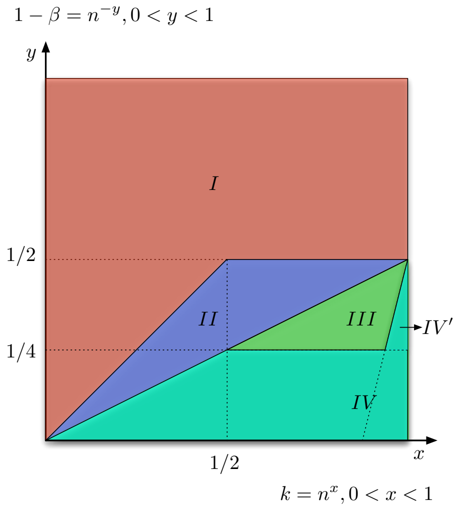

Figure 10: Phase diagram for small-world networks: impossible region (red region I), hard region (blue region II), easy region (green region III), and reconstructable region (cyan region IV and IV'). Compared to Figure 9, the spectral ordering procedure extends the reconstructable region (IV) when $k \succ n^{\frac{15}{16}}$ (IV').

level, which implies that k needs to be large enough.

Let us compare the region described in Equation (2.30) with the reconstructable region in Equation (2.29). We observe that spectral ordering pushes slightly beyond the reconstructable region when $k \succ n^{\frac{15}{16}}$, as shown in Figure 10.

Numerical Study

To see how the ring embedding Algorithm 10 performs on real dataset, we implemented it in Python, on the co-appearance network of characters in the novel Les Misérable[5] (Knuth, 1993). Figure 11 summarizes the visualization (zoom in for better resolution). Each node represents one character, and the color and size illustrate its degree, with darker color and larger size meaning higher degree. The lines connecting nodes on the ring represent co-appearance relationship in the chapters of the book, with the line width summarizing the co-appearance intensity. As one can see in the embedding, the obvious triangle is among

[5] The data is downloaded from Prof. Mark Newman's website `http://www-personal.umich.edu/~mejn/netdata/`.

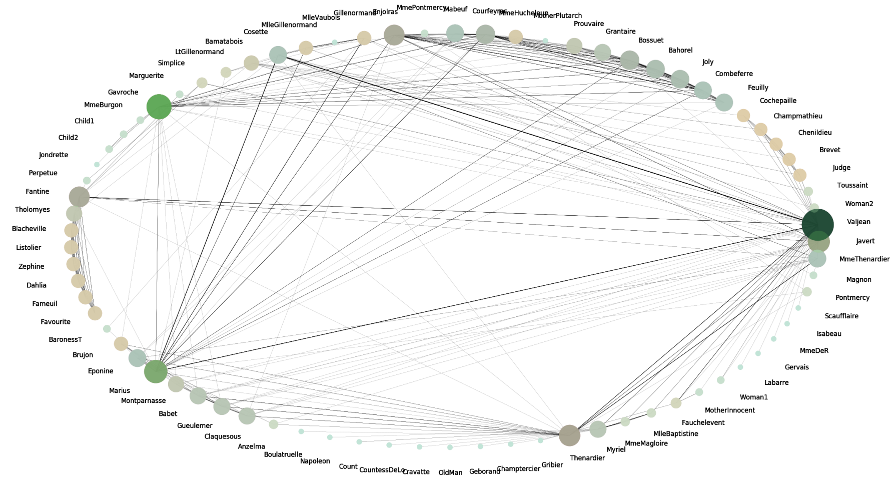

Figure 11: Ring embedding of Les Misérable co-appearance network.

the three main characters – Valjean, Marius and Cosette. In the ring embedding, Valjean and Javert are next to each other, so does Marius and Eponine, as they have very strong ties (enemies and friends) in the plot. The algorithm embeds the main characters – Valjean, Marius, Fantine, Thenadler, etc – in a rather spread out fashion on the ring, with each main character communicating with many other minor characters as in the novel. The structure assures the "short chain" property – any two characters can reach each other through these few main characters as middle points. One can also see many triadic closures in the ring neighborhood around main character, supporting the local "high clustering" feature.

2.3.5. Discussion

Comparison to stochastic block models Recently, stochastic block models (SBM) have attracted considerable amount of attention from researchers in various fields (Decelle et al., 2011; Massoulié, 2014; Mossel et al., 2013b). Community detection in stochastic block models focuses on recovering the hidden community structure obscured by noise in

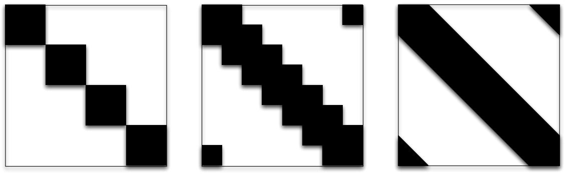

Figure 12: The structural matrices for stochastic block model (left), mixed membership SBM (middle), and small-world model (right). The black location denotes the support of the structural matrix.

the adjacency matrix and further concealed by the latent permutation on the nodes.

Detectability or weak recovery of the hidden community is one of the central question in studying SBM in the constant degree regime. Drawing insights from statistical physics, Decelle et al. (2011) conjectured a sharp transition threshold (also known as the Kesten-Stigum bound) for detection in the symmetric two-community case, above which recovering the community better than random guessing is possible, and below which – impossible. Massoulié (2014); Mossel et al. (2013b) proved the conjecture independently, one using spectral analysis on the non-backtracking matrix (Hashimoto, 1989), the other through analyzing non-backtracking walks. Later, for partial recovery and strong recovery (reconstruction) of multiple communities beyond the symmetric case, Abbe and Sandon (2015a) characterized the recovery threshold in terms of the Chernoff-Hellinger divergence.

The hidden community structure for classic SBM is illustrated in Figure 12 (left) as a block diagonal matrix. An interesting but theoretically more challenging extension to the classic SBM is the mixed-membership SBM, where each node may simultaneously belong to several communities. Consider an easy case of the model, where the mixed-membership occurs only within neighborhood communities, as shown in the middle image of Figure 12. The small-world network we are investigating in this thesis can be seen as an extreme case (shown on the right-most figure) of this easy mixed-membership SBM, where each node falls in effectively k local clusters. In the small-world network model, identifying the structural links and random links becomes challenging since there are many local clusters, in constrast to the relatively small number of communities in SBM. The multitude of local clusters makes it difficult to analyze the effect of the hidden permutation on the structural matrix. We view the current thesis as an initial attempt at tackling this problem.

Relations to graph matching Small world reconstruction, community membership reconstruction, planted clique localization etc., can be cast as solving for the latent permutation matrix P_π with different structural matrix B, in $\arg\max_{P_\pi \in \Pi} \langle P_\pi A P_\pi^T, B \rangle$ as suggested in Eq. (2.25). This is also called graph matching (Lyzinski et al., 2014). As one aims to match the observed adjacency matrix A to the structural matrix B via latent permutation matrix P_π. As written in the above form, the reconstruction task is reduced to a quadratic assignment problem (QAP), which is known to be NP-hard (Burkard et al., 1998; Cela, 2013). Due to the NP-hard nature of QAP, various relaxations on the permutation matrix constraints have been proposed: for instance, orthogonal matrices, doubly stochastic matrices, and SDP relaxations (Chandrasekaran et al., 2012). Quantifying the loss due to a relaxation for each model is a challenging task.

Reconstructable region We addressed the reconstruction problem via two distinct procedures: correlation thresholding and spectral ordering. Whether there exist other computationally efficient algorithms that can significantly improve upon the current reconstructable region is an open problem. Designing new algorithms requires a deeper insight into the structure of the small-world model, and will probably shed light on better algorithms for mixed membership models.

Robustness The test statistics and reconstruction procedure investigated in the current thesis are tailored to the W-S model assumptions. For example, the maximum likelihood test enjoys good properties when the model is well-specified. However, we acknowledge that real complex networks hardly satisfy the idealized assumptions. Therefore, when the model is mis-specified, designing robust procedures and tests with theoretical guarantees is an interesting further direction.

CHAPTER 3 : Regression, Learning and Model Selection

3.1. Learning under Square Loss with Offset Rademacher Complexity

3.1.1. Introduction

Determining the finite-sample behavior of risk in the problem of regression is arguably one of the most basic problems of Learning Theory and Statistics. This behavior can be studied in substantial generality with the tools of empirical process theory. When functions in a given convex class are uniformly bounded, one may verify the so-called "Bernstein condition." The condition—which relates the variance of the increments of the empirical process to their expectation—implies a certain localization phenomenon around the optimum and forms the basis of the analysis via *local Rademacher complexities*. The technique has been developed in (Koltchinskii and Panchenko, 2000; Koltchinskii, 2011b; Bousquet et al., 2002; Bartlett et al., 2005; Bousquet, 2002), among others, based on Talagrand's celebrated concentration inequality for the supremum of an empirical process.

In a recent pathbreaking paper, Mendelson (2014a) showed that a large part of this heavy machinery is not necessary for obtaining tight upper bounds on excess loss, even—and especially—if functions are unbounded. Mendelson observed that only one-sided control of the tail is required in the deviation inequality, and, thankfully, it is the tail that can be controlled under very mild assumptions.

In a parallel line of work, the search within the online learning setting for an analogue of "localization" has led to a notion of an "offset" Rademacher process (Rakhlin and Sridharan, 2014), yielding—in a rather clean manner—optimal rates for minimax regret in online supervised learning. It was also shown that the supremum of the offset process is a lower bound on the minimax value, thus establishing its intrinsic nature. The present thesis blends the ideas of Mendelson (2014a) and Rakhlin and Sridharan (2014). We introduce the notion of an offset Rademacher process for i.i.d. data and show that the supremum of this process upper bounds (both in expectation and in high probability) the excess risk of an empirical risk minimizer (for convex classes) and a two-step Star estimator of Audibert (2007) (for arbitrary classes). The statement holds under a weak assumption even if functions are not uniformly bounded.

The offset Rademacher complexity provides an intuitive alternative to the machinery of local Rademacher averages. Let us recall that the Rademacher process indexed by a function class $\mathcal{G} \subseteq \mathbb{R}^{\mathcal{X}}$ is defined as a stochastic process $g \mapsto \frac{1}{n}\sum_{t=1}^{n} \epsilon_t g(x_t)$ where $x_1, \ldots, x_n \in \mathcal{X}$ are held fixed and $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. Rademacher random variables. We define the offset Rademacher process as a stochastic process

$$g \mapsto \frac{1}{n}\sum_{t=1}^{n} \epsilon_t g(x_t) - c g(x_t)^2$$

for some $c \geq 0$. The process itself captures the notion of localization: when g is large in magnitude, the negative quadratic term acts as a compensator and "extinguishes" the fluctuations of the term involving Rademacher variables. The supremum of the process will be termed *offset Rademacher complexity*, and one may expect that this complexity is of a smaller order than the classical Rademacher averages (which, without localization, cannot be better than the rate of $n^{-1/2}$).

The self-modulating property of the offset complexity can be illustrated on the canonical example of a linear class $\mathcal{G} = \{x \mapsto w^{\mathsf{T}} x : w \in \mathbb{R}^p\}$, in which case the offset Rademacher complexity becomes

$$\frac{1}{n} \sup_{w \in \mathbb{R}^p} \left\{ w^{\mathsf{T}} \left(\sum_{t=1}^{n} \epsilon_t x_t \right) - c\|w\|_{\Sigma}^2 \right\} = \frac{1}{4cn} \left\| \sum_{t=1}^{n} \epsilon_t x_t \right\|_{\Sigma^{-1}}^2$$

where $\Sigma = \sum_{t=1}^{n} x_t x_t^{\mathsf{T}}$. Under mild conditions, the above expression is of the order $\mathcal{O}(p/n)$ in expectation and in high probability — a familiar rate achieved by the ordinary least squares, at least in the case of a well-specified model. We refer to Section 3.1.6 for the precise statement for both well-specified and misspecified case.

Our contributions can be summarized as follows. First, we show that offset Rademacher complexity is an upper bound on excess loss of the proposed estimator, both in expectation and in deviation. We then extend the chaining technique to quantify the behavior of the supremum of the offset process in terms of covering numbers. By doing so, we recover the rates of aggregation established in (Rakhlin et al., 2015) and, unlike the latter paper, the present method does not require boundedness (of the noise and functions). We provide a lower bound on minimax excess loss in terms of offset Rademacher complexity, indicating its intrinsic nature for the problems of regression. While our in-expectation results for bounded functions do not require any assumptions, the high probability statements rest on a lower isometry assumption that holds, for instance, for subgaussian classes. We show that offset Rademacher complexity can be further upper bounded by the fixed-point complexities defined by Mendelson Mendelson (2014a). We conclude with the analysis of ordinary least squares.

3.1.2. Problem Description and the Estimator

Let $\mathcal{F}$ be a class of functions on a probability space $(\mathcal{X}, P_X)$. The response is given by an unknown random variable Y, distributed jointly with X according to $P = P_X \times P_{Y|X}$. We observe a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ distributed i.i.d. according to P and aim to construct an estimator $\widehat{f}$ with small excess loss $\mathcal{E}(\widehat{f})$, where

$$\mathcal{E}(g) \triangleq \mathbb{E}(g - Y)^2 - \inf_{f \in \mathcal{F}} \mathbb{E}(f - Y)^2 \tag{3.1}$$

and $\mathbb{E}(f - Y)^2 = \mathbb{E}(f(X) - Y)^2$ is the expectation with respect to (X, Y). Let $\widehat{\mathbb{E}}$ denote the empirical expectation operator and define the following two-step procedure:

$$\widehat{g} = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \, \widehat{\mathbb{E}}(f(X) - Y)^2, \qquad \widehat{f} = \underset{f \in \operatorname{star}(\mathcal{F}, \widehat{g})}{\operatorname{argmin}} \, \widehat{\mathbb{E}}(f(X) - Y)^2 \tag{3.2}$$

where $\operatorname{star}(\mathcal{F}, g) = \{\lambda g + (1 - \lambda) f : f \in \mathcal{F}, \lambda \in [0, 1]\}$ is the star hull of $\mathcal{F}$ around g. (we abbreviate $\operatorname{star}(\mathcal{F}, 0)$ as $\operatorname{star}(\mathcal{F})$.) This two-step estimator was introduced (to the best of our knowledge) by Audibert (2007) for a finite class $\mathcal{F}$. We will refer to the procedure as the Star estimator. Audibert showed that this method is deviation-optimal for finite aggregation — the first such result, followed by other estimators with similar properties (Lecué and Mendelson, 2009; Dai et al., 2012) for the finite case. We present analysis that quantifies the behavior of this method for arbitrary classes of functions. The method has several nice features. First, it provides an alternative to the 3-stage discretization method of Rakhlin et al. (2015), does not require the prior knowledge of the entropy of the class, and goes beyond the bounded case. Second, it enjoys an upper bound of offset Rademacher complexity via relatively routine arguments under rather weak assumptions. Third, it naturally reduces to empirical risk minimization for convex classes (indeed, this happens whenever $\mathrm{star}(\mathcal{F}, \widehat{g}) = \mathcal{F}$).

Let f^* denote the minimizer

$$f^* = \operatorname*{argmin}_{f \in \mathcal{F}} \mathbb{E}(f(X) - Y)^2,$$

and let ξ denote the "noise"

$$\xi = Y - f^*.$$

We say that the model is misspecified if the regression function $\mathbb{E}[Y|X = x] \notin \mathcal{F}$, which means ξ is not zero-mean. Otherwise, we say that the model is well-specified.

3.1.3. A Geometric Inequality

We start by proving a geometric inequality for the Star estimator. This deterministic inequality holds conditionally on $X_1, \ldots, X_n$, and therefore reduces to a problem in $\mathbb{R}^n$.

Lemma 3.1.1 (Geometric Inequality)**.** *The two-step estimator $\widehat{f}$ in* (3.2) *satisfies*

$$\widehat{\mathbb{E}}(h - Y)^2 - \widehat{\mathbb{E}}(\widehat{f} - Y)^2 \geq c \cdot \widehat{\mathbb{E}}(\widehat{f} - h)^2 \tag{3.3}$$

for any $h \in \mathcal{F}$ and $c = 1/18$. If $\mathcal{F}$ is convex, (3.3) *holds with $c = 1$. Moreover, if $\mathcal{F}$ is a* *linear subspace,* (3.3) *holds with equality and* $c = 1$ *by the Pythagorean theorem.*

Remark 3.1.1. *In the absence of convexity of* $\mathcal{F}$*, the two-step estimator* $\widehat{f}$ *mimics the key Pythagorean identity, though with a constant* $1/18$*. We have not focused on optimizing* c *but rather on presenting a clean geometric argument.*

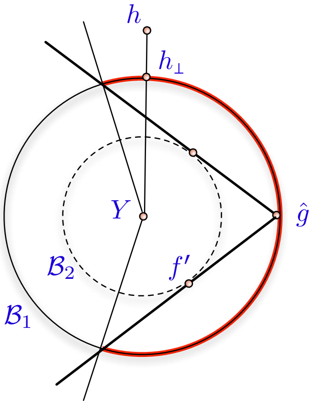

Figure 13: Proof of the geometry inequality. The solid and dotted balls are $\mathcal{B}(Y, \|\widehat{g} - Y\|_n)$ and $\mathcal{B}(Y, \|\widehat{f} - Y\|_n)$, respectively.

Proof of Lemma 3.1.1. Define the empirical ℓ_2 distance to be, for any f, g, $\|f\|_n := [\widehat{\mathbb{E}} f^2]^{1/2}$ and empirical product to be $\langle f, g\rangle_n := \widehat{\mathbb{E}}[fg]$. We will slightly abuse the notation by identifying every function with its finite-dimensional projection on $(X_1, \ldots, X_n)$.

Denote the ball (and sphere) centered at Y and with radius $\|\widehat{g} - Y\|_n$ to be $\mathcal{B}_1 := \mathcal{B}(Y, \|\widehat{g} - Y\|_n)$ (and $\mathcal{S}_1$, correspondingly). In a similar manner, define $\mathcal{B}_2 := \mathcal{B}(Y, \|\widehat{f} - Y\|_n)$ and $\mathcal{S}_2$. By the definition of the Star algorithm, we have $\mathcal{B}_2 \subseteq \mathcal{B}_1$. The statement holds with $c = 1$ if $\widehat{f} = \widehat{g}$, and so we may assume $\mathcal{B}_2 \subset \mathcal{B}_1$. Denote by $\mathcal{C}$ the conic hull of $\mathcal{B}_2$ with origin at $\widehat{g}$. Define the spherical cap outside the cone $\mathcal{C}$ to be $\mathcal{S} = \mathcal{S}_1 \setminus \mathcal{C}$ (drawn in red in Figure 13).

First, by the optimality of $\widehat{g}$, for any $h \in \mathcal{F}$, we have $\|h - Y\|_n^2 \geq \|\widehat{g} - Y\|_n^2$, i.e. any $h \in \mathcal{F}$ is not in the interior of $\mathcal{B}_1$. Furthermore, h is not in the interior of the cone $\mathcal{C}$, as otherwise there would be a point inside $\mathcal{B}_2$ strictly better than $\widehat{f}$. Thus $h \in (\mathrm{int}\mathcal{C})^c \cap (\mathrm{int}\mathcal{B}_1)^c$.

Second, $\widehat{f} \in \mathcal{B}_2$ and it is a contact point of $\mathcal{C}$ and $\mathcal{S}_2$. Indeed, $\widehat{f}$ is necessarily on a line segment between $\hat{g}$ and a point outside $\mathcal{B}_1$ that does not pass through the interior of $\mathcal{B}_2$ by optimality of $\widehat{f}$. Let K be the set of all contact points – potential locations of $\widehat{f}$.

Now we fix $h \in \mathcal{F}$ and consider the two dimensional plane $\mathcal{L}$ that passes through three points $(\hat{g}, Y, h)$, depicted in Figure 13. Observe that the left-hand-side of the desired inequality (3.3) is constant as $\widehat{f}$ ranges over K. To prove the inequality it therefore suffices to choose a value $f' \in K$ that maximizes the right-hand-side. The maximization of $\|h - f'\|^2$ over $f' \in K$ is achieved by $f' \in K \cap \mathcal{L}$. This can be argued simply by symmetry: the two-dimensional plane $\mathcal{L}$ intersects $\mathsf{span}(K)$ in a line and the distance between h and K is maximized at the extreme point of this intersection. Hence, to prove the desired inequality, we can restrict our attention to the plane $\mathcal{L}$ and f' instead of $\widehat{f}$.

For any $h \in \mathcal{F}$, define the projection of h onto the shell $\mathcal{L} \cap \mathcal{S}$ to be $h_\perp \in \mathcal{S}$. We first prove (3.3) for $h_\perp$ and then extend the statement to h. By the geometry of the cone,

$$\|f' - \widehat{g}\|_n \geq \frac{1}{2}\|\widehat{g} - h_\perp\|_n.$$

By triangle inequality,

$$\|f' - \widehat{g}\|_n \geq \frac{1}{2}\|\widehat{g} - h_\perp\|_n \geq \frac{1}{2}\left(\|f' - h_\perp\|_n - \|f' - \widehat{g}\|_n\right).$$

Rearranging,

$$\|f' - \widehat{g}\|_n^2 \geq \frac{1}{9}\|f' - h_\perp\|_n^2.$$

By the Pythagorean theorem,

$$\|h_\perp - Y\|_n^2 - \|f' - Y\|_n^2 = \|\widehat{g} - Y\|_n^2 - \|f' - Y\|_n^2 = \|f' - \widehat{g}\|_n^2 \geq \frac{1}{9}\|f' - h_\perp\|_n^2,$$

thus proving the claim for $h_\perp$ for constant $c = 1/9$.

We can now extend the claim to h. Indeed, due to the fact that $h \in (\mathrm{int}\mathcal{C})^c \cap (\mathrm{int}\mathcal{B}_1)^c$ and the geometry of the projection $h \to h_\perp$, we have $\langle h_\perp - Y, h_\perp - h\rangle_n \le 0$. Thus

$$\begin{aligned}
\|h - Y\|_n^2 - \|f' - Y\|_n^2 &= \|h_\perp - h\|_n^2 + \|h_\perp - Y\|_n^2 - 2\langle h_\perp - Y, h_\perp - h\rangle_n - \|f' - Y\|_n^2 \\
&\ge \|h_\perp - h\|_n^2 + (\|h_\perp - Y\|_n^2 - \|f' - Y\|_n^2) \\
&\ge \|h_\perp - h\|_n^2 + \frac{1}{9}\|f' - h_\perp\|_n^2 \ge \frac{1}{18}(\|h_\perp - h\|_n + \|f' - h_\perp\|_n)^2 \\
&\ge \frac{1}{18}\|f' - h\|_n^2.
\end{aligned}$$

This proves the claim for h with constant 1/18.

□

An upper bound on excess loss follows immediately from Lemma 3.1.1.

Corollary 3.1.1. *Conditioned on the data $\{X_n, Y_n\}$, we have a deterministic upper bound for the Star algorithm:*

$$\mathcal{E}(\widehat{f}) \le (\widehat{\mathbb{E}} - \mathbb{E})[2(f^* - Y)(f^* - \widehat{f})] + \mathbb{E}(f^* - \widehat{f})^2 - (1 + c) \cdot \widehat{\mathbb{E}}(f^* - \widehat{f})^2, \tag{3.4}$$

with the value of constant c given in Lemma 3.1.1.

Proof.

$$\begin{aligned}
\mathcal{E}(\widehat{f}) &= \mathbb{E}(\widehat{f}(X) - Y)^2 - \inf_{f\in\mathcal{F}} \mathbb{E}(f(X) - Y)^2 \\
&\le \mathbb{E}(\widehat{f} - Y)^2 - \mathbb{E}(f^* - Y)^2 + \left[\widehat{\mathbb{E}}(f^* - Y)^2 - \widehat{\mathbb{E}}(\widehat{f} - Y)^2 - c \cdot \widehat{\mathbb{E}}(\widehat{f} - f^*)^2\right] \\
&= (\widehat{\mathbb{E}} - \mathbb{E})[2(f^* - Y)(f^* - \widehat{f})] + \mathbb{E}(f^* - \widehat{f})^2 - (1 + c) \cdot \widehat{\mathbb{E}}(f^* - \widehat{f})^2.
\end{aligned}$$

□

An attentive reader will notice that the multiplier on the negative empirical quadratic term in (3.4) is slightly larger than the one on the expected quadratic term. This is the starting point of the analysis that follows.

3.1.4. Symmetrization

We will now show that the discrepancy in the multiplier constant in (3.4) leads to offset Rademacher complexity through rather elementary symmetrization inequalities. We perform this analysis both in expectation (for the case of bounded functions) and in high probability (for the general unbounded case). While the former result follows from the latter, the in-expectation statement for bounded functions requires no assumptions, in contrast to control of the tails.

Theorem 3.1.1. *Define the set $\mathcal{H} := \mathcal{F} - f^* + star(\mathcal{F} - \mathcal{F})$. The following expectation bound on excess loss of the Star estimator holds:*

$$\mathbb{E}\mathcal{E}(\widehat{f}) \leq (2M + K(2+c)/2) \cdot \mathbb{E} \sup_{h \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} 2\epsilon_i h(X_i) - c' h(X_i)^2 \right\}$$

where $\epsilon_1, \ldots, \epsilon_n$ are independent Rademacher random variables, $c' = \min\{\frac{c}{4M}, \frac{c}{4K(2+c)}\}$, $K = \sup_f |f|_\infty$, and $M = \sup_f |Y - f|_\infty$ almost surely.

The proof of the theorem involves an introduction of independent Rademacher random variables and two contraction-style arguments to remove the multipliers $(Y_i - f^*(X_i))$. These algebraic manipulations are postponed to the appendix.

The term in the curly brackets will be called an offset Rademacher process, and the expected supremum — an offset Rademacher complexity. While Theorem 3.1.1 only applies to bounded functions and bounded noise, the upper bound already captures the localization phenomenon, even for non-convex function classes (and thus goes well beyond the classical local Rademacher analysis).

As argued in (Mendelson, 2014a), it is the contraction step that requires boundedness of the functions when analyzing square loss. Mendelson uses a small ball assumption (a weak condition on the distribution, stated below) to split the analysis into the study of the multiplier and quadratic terms. This assumption allows one to compare the expected square of any function to its empirical version, to within a multiplicative constant that depends on the small ball property. In contrast, we need a somewhat stronger assumption that will allow us to take this constant to be at least $1 - c/4$. We phrase this condition—the lower isometry bound—as follows.

Definition 3.1.1 (Lower Isometry Bound)**.** *We say that a function class $\mathcal{F}$ satisfies the lower isometry bound with some parameters $0 < \eta < 1$ and $0 < \delta < 1$ if*

$$\mathbb{P}\left(\inf_{f \in \mathcal{F} \setminus \{0\}} \frac{1}{n} \sum_{i=1}^{n} \frac{f^2(X_i)}{\mathbb{E} f^2} \geq 1 - \eta \right) \geq 1 - \delta \tag{3.5}$$

for all $n \geq n_0(\mathcal{F}, \delta, \eta)$, where $n_0(\mathcal{F}, \delta, \eta)$ depends on the complexity of the class.

In general this is a mild assumption that requires good tail behavior of functions in $\mathcal{F}$, yet it is stronger than the small ball property. Mendelson Mendelson (2015) shows that this condition holds for heavy-tailed classes assuming the small ball condition plus a norm-comparison property $\|f\|_{\ell_q} \leq L\|f\|_{\ell_2}, \forall f \in \mathcal{F}$. We also remark that Assumption 3.1.1 holds for sub-gaussian classes $\mathcal{F}$ using concentration tools, as already shown in Lecué and Mendelson (2013). For completeness, let us also state the small ball property:

Definition 3.1.2 (Small Ball Property Mendelson (2014a,b))**.** *The class of functions $\mathcal{F}$ satisfies the small-ball condition if there exist constants $\kappa > 0$ and $0 < \epsilon < 1$ for every $f \in \mathcal{F}$,*

$$\mathbb{P}\big(|f(X)| \geq \kappa (\mathbb{E} f^2)^{1/2}\big) \geq \epsilon.$$

Armed with the lower isometry bound, we now prove that the tail behavior of the deterministic upper bound in (3.4) can be controlled via the tail behavior of offset Rademacher complexity.

Theorem 3.1.2. *Define the set $\mathcal{H} := \mathcal{F} - f^* + star(\mathcal{F} - \mathcal{F})$. Assume the lower isometry bound in Definition 3.1.1 holds with $\eta = c/4$ and some $\delta < 1$, where c is the constant in*

(3.3). Let $\xi_i = Y_i - f^*(X_i)$. Define

$$A := \sup_{h\in\mathcal{H}} \frac{\mathbb{E}h^4}{(\mathbb{E}h^2)^2} \quad \textit{and} \quad B := \sup_{X,Y} \mathbb{E}\xi^4.$$

Then there exist two absolute constants $c', \tilde{c} > 0$ (only depends on c), such that

$$\mathbb{P}\left(\mathcal{E}(\widehat{f}) > 4u\right) \leq 4\delta + 4\mathbb{P}\left(\sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^n \epsilon_i\xi_i h(X_i) - \tilde{c}\cdot h(X_i)^2 > u\right)$$

for any

$$u > \frac{32\sqrt{AB}}{c'} \cdot \frac{1}{n},$$

as long as $n > \frac{16(1-c')^2 A}{c'^2} \vee n_0(\mathcal{H}, \delta, c/4)$.

Theorem 3.1.2 states that excess loss is stochastically dominated by offset Rademacher complexity. We remark that the requirement in A, B holds under the mild moment conditions.

Remark 3.1.2. *In certain cases, Definition 3.1.1 can be shown to hold for $f \in \mathcal{F} \setminus r^*\mathcal{B}$ (rather than all $f \in \mathcal{F}$), for some critical radius r^*, as soon as $n \geq n_0(\mathcal{F}, \delta, \eta, r^*)$ (see Mendelson (2015)). In this case, the bound on the offset complexity is only affected additively by $(r^*)^2$.*

We postpone the proof of the Theorem to the appendix. In a nutshell, it extends the classical probabilistic symmetrization technique (Giné and Zinn, 1984; Mendelson, 2003) to the non-zero-mean offset process under the investigation.

3.1.5. Offset Rademacher Process: Chaining and Critical Radius

Let us summarize the development so far. We have shown that excess loss of the Star estimator is upper bounded by the (data-dependent) offset Rademacher complexity, both in expectation and in high probability, under the appropriate assumptions. We claim that the necessary properties of the estimator are now captured by the offset complexity, and we are now squarely in the realm of empirical process theory. In particular, we may want to quantify rates of convergence under complexity assumptions on $\mathcal{F}$, such as covering numbers. In contrast to local Rademacher analyses where one would need to estimate the data-dependent fixed point of the critical radius in some way, the task is much easier for the offset complexity. To this end, we study the offset process with the tools of empirical process theory.

Chaining Bounds

The first lemma describes the behavior of offset Rademacher process for a finite class.

Lemma 3.1.2. *Let $V \subset \mathbb{R}^n$ be a finite set of vectors of cardinality N. Then for any $C > 0$,*

$$\mathbb{E}_\epsilon \max_{v \in V} \left[\frac{1}{n} \sum_{i=1}^n \epsilon_i v_i - C v_i^2 \right] \leq \frac{1}{2C} \frac{\log N}{n}.$$

Furthermore, for any $\delta > 0$,

$$\mathbb{P}\left(\max_{v \in V} \left[\frac{1}{n} \sum_{i=1}^n \epsilon_i v_i - C v_i^2 \right] \geq \frac{1}{2C} \frac{\log N + \log 1/\delta}{n} \right) \leq \delta.$$

When the noise ξ is unbounded,

$$\begin{aligned} &\mathbb{E}_\epsilon \max_{v \in V} \left[\frac{1}{n} \sum_{i=1}^n \epsilon_i \xi_i v_i - C v_i^2 \right] \leq M \cdot \frac{\log N}{n}, \\ &\mathbb{P}_\epsilon \left(\max_{v \in V} \left[\frac{1}{n} \sum_{i=1}^n \epsilon_i \xi_i v_i - C v_i^2 \right] \geq M \cdot \frac{\log N + \log 1/\delta}{n} \right) \leq \delta, \end{aligned}$$

where

$$M := \sup_{v \in V \setminus \{0\}} \frac{\sum_{i=1}^n v_i^2 \xi_i^2}{2C \sum_{i=1}^n v_i^2}. \tag{3.6}$$

Armed with the lemma for a finite collection, we upper bound the offset Rademacher complexity of a general class through the chaining technique. We perform the analysis in expectation and in probability. Recall that a δ-cover of a subset S in a metric space (T, d) is a collection of elements such that the union of the δ-balls with centers at the elements contains S. A covering number at scale δ is the size of the minimal δ-cover.

One of the main objectives of symmetrization is to arrive at a stochastic process that can be studied conditionally on data, so that all the relevant complexities can be made sample-based (or, empirical). Since the functions only enter offset Rademacher complexity through their values on the sample $X_1, \ldots, X_n$, we are left with a finite-dimensional object. Throughout the thesis, we work with the empirical ℓ_2 distance

$$d_n(f,g) = \left(\frac{1}{n}\sum_{i=1}^{n}(f(X_i) - g(X_i))^2\right)^{1/2}.$$

The covering number of $\mathcal{G}$ at scale δ with respect to d_n will be denoted by $\mathcal{N}_2(\mathcal{G}, \delta)$.

Lemma 3.1.3. *Let $\mathcal{G}$ be a class of functions from $\mathcal{Z}$ to $\mathbb{R}$. Then for any $z_1, \ldots, z_n \in \mathcal{Z}$*

$$\mathbb{E}_\epsilon \sup_{g \in \mathcal{G}} \left[\frac{1}{n}\sum_{t=1}^{n} \epsilon_i g(z_i) - Cg(z_i)^2\right] \leq \inf_{\gamma \geq 0, \alpha \in [0,\gamma]} \left\{ \frac{(2/C)\log \mathcal{N}_2(\mathcal{G},\gamma)}{n} \right.$$
$$\left. + 4\alpha + \frac{12}{\sqrt{n}} \int_\alpha^\gamma \sqrt{\log \mathcal{N}_2(\mathcal{G},\delta)} d\delta \right\}$$

where $\mathcal{N}_2(\mathcal{G},\gamma)$ is an ℓ_2-cover of $\mathcal{G}$ on $(z_1, \ldots, z_n)$ at scale γ (assumed to contain $\mathbf{0}$).

Instead of assuming that $\mathbf{0}$ is contained in the cover, we may simply increase the size of the cover by 1, which can be absorbed by a small change of a constant.

Let us discuss the upper bound of Lemma 3.1.3. First, we may take $\alpha = 0$, unless the integral diverges (which happens for very large classes with entropy growth of $\log \mathcal{N}_2(\mathcal{G}, \delta) \sim \delta^{-p}$, $p \geq 2$). Next, observe that first term is precisely the rate of aggregation with a finite collection of size $\mathcal{N}_2(\mathcal{G}, \gamma)$. Hence, the upper bound is an optimal balance of the following procedure: cover the set at scale γ and pay the rate of aggregation for this finite collection, plus pay the rate of convergence of ERM within a γ-ball. The optimal balance is given by some γ (and can be easily computed under assumptions on covering number behavior — see (Rakhlin and Sridharan, 2014)). The optimal γ quantifies the localization radius that arises from the curvature of the loss function. One may also view the optimal balance as the well-known equation

$$\frac{\log \mathcal{N}(\mathcal{G},\gamma)}{n} \asymp \gamma^2,$$

studied in statistics (Yang and Barron, 1999) for well-specified models. The present thesis, as well as (Rakhlin et al., 2015), extend the analysis of this balance to the misspecified case and non-convex classes of functions.

Now we provide a high probability analogue of Lemma 3.1.3.

Lemma 3.1.4. *Let $\mathcal{G}$ be a class of functions from $\mathcal{Z}$ to $\mathbb{R}$. Then for any $z_1,\ldots,z_n \in \mathcal{Z}$ and any $u>0$,*

$$\mathbb{P}_\epsilon\left(\sup_{g\in\mathcal{G}}\left[\frac{1}{n}\sum_{t=1}^{n}\epsilon_i g(z_i) - Cg(z_i)^2\right] > u\cdot \inf_{\alpha\in[0,\gamma]}\left\{4\alpha+\frac{12}{\sqrt{n}}\int_\alpha^\gamma \sqrt{\log\mathcal{N}_2(\mathcal{G},\delta)}d\delta\right\} + \frac{2}{C}\frac{\log\mathcal{N}_2(\mathcal{G},\gamma)+u}{n}\right)$$
$$\le \frac{2}{1-e^{-2}}\exp(-cu^2)+\exp(-u)$$

where $\mathcal{N}_2(\mathcal{G},\gamma)$ is an ℓ_2-cover of $\mathcal{G}$ on $(z_1,\ldots,z_n)$ at scale γ (assumed to contain $\mathbf{0}$) and $C,c>0$ are universal constants.

The above lemmas study the behavior of offset Rademacher complexity for abstract classes $\mathcal{G}$. Observe that the upper bounds in previous sections are in terms of the class $\mathcal{F}-f^*+\text{star}(\mathcal{F}-\mathcal{F})$. This class, however, is not more complex that the original class $\mathcal{F}$ (with the exception of a finite class $\mathcal{F}$). More precisely, the covering numbers of $\mathcal{F}+\mathcal{F}' := \{f+g : f\in\mathcal{F}, g\in\mathcal{F}'\}$ and $\mathcal{F}-\mathcal{F}' := \{f-g : f\in\mathcal{F}, g\in\mathcal{F}'\}$ are bounded as

$$\log\mathcal{N}_2(\mathcal{F}+\mathcal{F}',2\epsilon),\ \ \log\mathcal{N}_2(\mathcal{F}-\mathcal{F}',2\epsilon)\le \log\mathcal{N}_2(\mathcal{F},\epsilon)+\log\mathcal{N}_2(\mathcal{F}',\epsilon)$$

for any $\mathcal{F},\mathcal{F}'$. The following lemma shows that the complexity of the star hull $\text{star}(\mathcal{F})$ is also not significantly larger than that of $\mathcal{F}$.

Lemma 3.1.5 (Mendelson (2002), Lemma 4.5)**.** *For any scale $\epsilon>0$, the covering number of $\mathcal{F}\subset\mathcal{B}_2$ and that of star($\mathcal{F}$) are bounded in the sense*

$$\log\mathcal{N}_2(\mathcal{F},2\epsilon)\le\log\mathcal{N}_2(star(\mathcal{F}),2\epsilon)\le\log\frac{2}{\epsilon}+\log\mathcal{N}_2(\mathcal{F},\epsilon).$$

Critical Radius

Now let us study the critical radius of offset Rademacher processes. Let $\xi = f^* - Y$ and define

$$\alpha_n(\mathcal{H}, \kappa, \delta) \triangleq \inf \left\{ r > 0 : \mathbb{P}\left(\sup_{h \in \mathcal{H} \cap r\mathcal{B}} \left\{ \frac{1}{n} \sum_{i=1}^n 2\epsilon_i \xi_i h(X_i) - c' \frac{1}{n} \sum_{i=1}^n h^2(X_i) \right\} \leq \kappa r^2 \right) \geq 1 - \delta \right\}. \tag{3.7}$$

Theorem 3.1.3. *Assume $\mathcal{H}$ is star-shaped around 0 and the lower isometry bound holds for δ, ϵ. Define the critical radius*

$$r = \alpha_n(\mathcal{H}, c'(1-\epsilon), \delta).$$

Then we have with probability at least $1 - 2\delta$,

$$\sup_{h \in \mathcal{H}} \left\{ \frac{2}{n} \sum_{i=1}^n \epsilon_i \xi_i h(X_i) - c' \frac{1}{n} \sum_{i=1}^n h^2(X_i) \right\} = \sup_{h \in \mathcal{H} \cap r\mathcal{B}} \left\{ \frac{2}{n} \sum_{i=1}^n \epsilon_i \xi_i h(X_i) - c' \frac{1}{n} \sum_{i=1}^n h^2(X_i) \right\},$$

which further implies

$$\sup_{h \in \mathcal{H}} \left\{ \frac{2}{n} \sum_{i=1}^n \epsilon_i \xi_i h(X_i) - c' \frac{1}{n} \sum_{i=1}^n h^2(X_i) \right\} \leq r^2.$$

The first statement of Theorem 3.1.3 shows the self-modulating behavior of the offset process: there is a critical radius, beyond which the fluctuations of the offset process are controlled by those within the radius. To understand the second statement, we observe that the complexity α_n is upper bounded by the corresponding complexity in (Mendelson, 2014a), which is defined without the quadratic term subtracted off. Hence, offset Rademacher complexity is no larger (under our Assumption 3.1.1) than the upper bounds obtained by Mendelson (2014a) in terms of the critical radius.

In this section, we briefly describe several applications. The first is concerned with parametric regression.

Lemma 3.1.6. *Consider the parametric regression $Y_i = X_i^T\beta^* + \xi_i, 1 \le i \le n$, where ξ_i need not be centered. The offset Rademacher complexity is bounded as*

$$\mathbb{E}_\epsilon \sup_{\beta\in\mathbb{R}^p} \left\{ \frac{1}{n}\sum_{i=1}^n 2\epsilon_i\xi_i X_i^T\beta - C\beta^T X_i X_i^T\beta \right\} = \frac{\mathsf{tr}\left(G^{-1}H\right)}{Cn}$$

and

$$\mathbb{P}_\epsilon\left(\sup_{\beta\in\mathbb{R}^p} \left\{ \frac{1}{n}\sum_{i=1}^n 2\epsilon_i\xi_i X_i^T\beta - C\beta^T X_i X_i^T\beta \right\} \ge \frac{\mathsf{tr}\left(G^{-1}H\right)}{Cn} + \frac{\sqrt{\mathsf{tr}\left([G^{-1}H]^2\right)}}{n}(4\sqrt{2\log\frac{1}{\delta}} + 64\log\frac{1}{\delta}) \right) \le \delta$$

where $G := \sum_{i=1}^n X_i X_i^T$ is the Gram matrix and $H = \sum_{i=1}^n \xi_i^2 X_i X_i^T$. In the well-specified case (that is, ξ_i are zero-mean), assuming that conditional variance is σ^2, then conditionally on the design matrix, $\mathbb{E}G^{-1}H = \sigma^2 I_p$ and excess loss is upper bounded by order $\frac{\sigma^2 p}{n}$.

Proof. The offset Rademacher can be interpreted as the Fenchel-Legendre transform, where

$$\sup_{\beta\in\mathbb{R}^p} \left\{ \sum_{i=1}^n 2\epsilon_i\xi_i X_i^T\beta - C\beta^T X_i X_i^T\beta \right\} = \frac{\sum_{i,j=1}^n \epsilon_i\epsilon_j\xi_i\xi_j X_i^T G^{-1} X_j}{Cn}. \tag{3.8}$$

Thus we have in expectation

$$\mathbb{E}_\epsilon \frac{1}{n} \sup_{\beta\in\mathbb{R}^p} \left\{ \sum_{i=1}^n 2\epsilon_i\xi_i X_i^T\beta - C\beta^T X_i X_i^T\beta \right\} = \frac{\sum_{i=1}^n \xi_i^2 X_i^T G^{-1} X_i}{Cn} = \frac{\mathsf{tr}[G^{-1}(\sum_{i=1}^n \xi_i^2 X_i X_i^T)]}{Cn}. \tag{3.9}$$

For high probability bound, note the expression in Equation (3.8) is Rademacher chaos of order two. Define symmetric matrix $M \in \mathbb{R}^{n\times n}$ with entries

$$M_{ij} = \xi_i\xi_j X_i^T G^{-1} X_j$$

and define

$$Z = \sum_{i,j=1}^{n} \epsilon_i \epsilon_j \xi_i \xi_j X_i^T G^{-1} X_j = \sum_{i,j=1}^{n} \epsilon_i \epsilon_j M_{ij}.$$

Then

$$\mathbb{E} Z = \mathsf{tr}[G^{-1}(\sum_{i=1}^{n} \xi_i^2 X_i X_i^T)],$$

and

$$\mathbb{E} \sum_{i=1}^{n} (\sum_{j=1}^{n} \epsilon_j M_{ij})^2 = \|M\|_F^2 = \mathsf{tr}[G^{-1}(\sum_{i=1}^{n} \xi_i^2 X_i X_i^T) G^{-1}(\sum_{i=1}^{n} \xi_i^2 X_i X_i^T)].$$

Furthermore,

$$\|M\| \leq \|M\|_F = \sqrt{\mathsf{tr}[G^{-1}(\sum_{i=1}^{n} \xi_i^2 X_i X_i^T) G^{-1}(\sum_{i=1}^{n} \xi_i^2 X_i X_i^T)]}$$

We apply the concentration result in (Boucheron et al., 2013, Exercise 6.9),

$$\mathbb{P}\left(Z - \mathbb{E} Z \geq 4\sqrt{2} \|M\|_F \sqrt{t} + 64 \|M\| t \right) \leq e^{-t}. \tag{3.10}$$

□

For the finite dictionary aggregation problem, the following lemma shows control of offset Rademacher complexity.

Lemma 3.1.7. *Assume $\mathcal{F} \in \mathcal{B}_2$ is a finite class of cardinality N. Define $\mathcal{H} = \mathcal{F} - f^* + \text{star}(\mathcal{F} - \mathcal{F})$ which contains the Star estimator $\widehat{f} - f^*$ defined in Equation (3.2). The offset Rademacher complexity for $\mathcal{H}$ is bounded as*

$$\mathbb{E}_\epsilon \sup_{h \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} 2\epsilon_i \xi_i h(X_i) - C h(X_i)^2 \right\} \leq \tilde{C} \cdot \frac{\log(N \vee n)}{n}$$

and

$$\mathbb{P}_\epsilon \left(\sup_{h \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} 2\epsilon_i \xi_i h(X_i) - C h(X_i)^2 \right\} \leq \tilde{C} \cdot \frac{\log(N \vee n) + \log \frac{1}{\delta}}{n} \right) \leq \delta.$$

where $\tilde{C}$ is a constant depends on $K := 2(\sqrt{\sum_{i=1}^n \xi_i^2/n} + 2C)$ and

$$M := \sup_{h\in\mathcal{H}\setminus\{0\}} \frac{\sum_{i=1}^n h(X_i)^2\xi_i^2}{2C\sum_{i=1}^n h(X_i)^2}.$$

We observe that the bound of Lemma 3.1.7 is worse than the optimal bound of (Audibert, 2007) by an additive $\frac{\log n}{n}$ term. This is due to the fact that the analysis for finite case passes through the offset Rademacher complexity of the star hull, and for this case the star hull is more rich than the finite class. For this case, a direct analysis of the Star estimator is provided in (Audibert, 2007).

While the offset complexity of the star hull is crude for the finite case, the offset Rademacher complexity *does* capture the correct rates for regression with larger classes, initially derived in (Rakhlin et al., 2015). We briefly mention the result. The proof is identical to the one in (Rakhlin and Sridharan, 2014), with the only difference that offset Rademacher is defined in that paper as a sequential complexity in the context of online learning.

Corollary 3.1.2. *Consider the problem of nonparametric regression, as quantified by the growth*

$$\log\mathcal{N}_2(\mathcal{F},\epsilon) \le \epsilon^{-p}.$$

In the regime $p \in (0,2)$, the upper bound of Lemma 3.1.4 scales as $n^{-\frac{2}{2+p}}$. In the regime $p \ge 2$, the bound scales as $n^{-1/p}$, with an extra logarithmic factor at $p = 2$.

For the parametric case of $p = 0$, one may also readily estimate the offset complexity. Results for VC classes, sparse combinations of dictionary elements, and other parametric cases follow easily by plugging in the estimate for the covering number or directly upper bounding the offset complexity (see Rakhlin et al. (2015); Rakhlin and Sridharan (2014)).

3.1.7. Lower bound on Minimax Regret via Offset Rademacher Complexity

We conclude with a lower bound on minimax regret in terms of offset Rademacher complexity.

Theorem 3.1.4 (Minimax Lower Bound on Regret)**.** *Define the offset Rademacher complexity over $\mathcal{X}^{\otimes n}$ as*

$$\mathfrak{R}^{\circ}(n, \mathcal{F}) = \sup_{\{x_i\}_{i=1}^n \in \mathcal{X}^{\otimes n}} \mathbb{E}_\epsilon \sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^n 2\epsilon_i f(x_i) - f(x_i)^2 \right\}$$

then the following minimax lower bound on regret holds:

$$\inf_{\hat{g} \in \mathcal{G}} \sup_P \left\{ \mathbb{E}(\hat{g} - Y)^2 - \inf_{f \in \mathcal{F}} \mathbb{E}(f - Y)^2 \right\} \geq \mathfrak{R}^{\circ}((1+c)n, \mathcal{F}) - \frac{c}{1+c} \mathfrak{R}^{\circ}(cn, \mathcal{G}),$$

for any $c > 0$.

For the purposes of matching the performance of the Star procedure, we can take $\mathcal{G} = \mathcal{F} + \mathrm{star}(\mathcal{F} - \mathcal{F})$.

3.2. Geometric Inference for General High Dimensional Linear Models

3.2.1. Introduction

Driven by a wide range of applications, high-dimensional linear inverse problems such as noisy compressed sensing, sign vector recovery, trace regression, orthogonal matrix estimation, and noisy matrix completion have drawn significant recent interest in several fields, including statistics, applied mathematics, computer science, and electrical engineering. These problems are often studied in a case-by-case fashion, with the main focus on estimation. Although similarities in the technical analyses have been suggested heuristically, a general unified theory for statistical inference including estimation, confidence intervals and hypothesis testing is still yet to be developed.

In this thesis, we consider a general linear inverse model

$$Y = \mathcal{X}(M) + Z \tag{3.11}$$

where $M \in \mathbb{R}^p$ is the vectorized version of the parameter of interest, $\mathcal{X} : \mathbb{R}^p \to \mathbb{R}^n$ is a linear operator (matrix in $\mathbb{R}^{n \times p}$), and $Z \in \mathbb{R}^n$ is a noise vector. We observe $(\mathcal{X}, Y)$ and wish to recover the unknown parameter M. A particular focus is on the high-dimensional setting where the ambient dimension p of the parameter M is much larger than the sample size n, i.e., the dimension of Y. In such a setting, the parameter of interest M is commonly assumed to have, with respect to a given atom set $\mathcal{A}$, a certain low complexity structure which captures the true dimension of the statistical estimation problem. A number of high-dimensional inference problems actively studied in the recent literature can be seen as special cases of this general linear inverse model.

High Dimension Linear Regression/Noisy Compressed Sensing. In high-dimensional linear regression, one observes (X, Y) with

$$Y = XM + Z, \tag{3.12}$$

where $Y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$ with $p \gg n$, $M \in \mathbb{R}^p$ is a sparse signal, and $Z \in \mathbb{R}^n$ is a noise vector. The goal is to recover the unknown sparse signal of interest $M \in \mathbb{R}^p$ based on the observation (X, Y) through an efficient algorithm. Many estimation methods including ℓ_1-regularized procedures such as the Lasso and Dantzig Selector have been developed and analyzed. See, for example, Tibshirani (1996); Candès and Tao (2007); Bickel et al. (2009); Bühlmann and van de Geer (2011) and the references therein. Confidence intervals and hypothesis testing for high-dimensional linear regression have also been actively studied in the last few years. A common approach is to first construct a de-biased Lasso or de-biased scaled-Lasso estimator and then make inference based on the asymptotic normality of low-dimensional functionals of the de-biased estimator. See, for example, Bühlmann (2013); Zhang and Zhang (2014); van de Geer et al. (2014); Javanmard and Montanari (2014).

Trace Regression. Accurate recovery of a low-rank matrix based on a small number of linear measurements has a wide range of applications and has drawn much recent attention in several fields. See, for example, Recht et al. (2010); Koltchinskii (2011a); Rohde et al. (2011); Koltchinskii et al. (2011); Candes and Plan (2011). In trace regression, one observes (X_i, Y_i), $i = 1, ..., n$ with

$$Y_i = \mathsf{Tr}(X_i^T M) + Z_i, \tag{3.13}$$

where $Y_i \in \mathbb{R}$, $X_i \in \mathbb{R}^{p_1\times p_2}$ are measurement matrices, and Z_i are noise. The goal is to recover the unknown matrix $M \in \mathbb{R}^{p_1\times p_2}$ which is assumed to be of low rank. Here the dimension of the parameter M is $p \equiv p_1 p_2 \gg n$. A number of constrained and penalized nuclear minimization methods have been introduced and studied in both the noiseless and noisy settings. See the aforementioned references for further details.

Sign Vector Recovery. The setting of sign vector recovery is similar to the one for the high-dimensional regression except the signal of interest is a sign vector. More specifically, one observes (X, Y) with

$$Y = XM + Z \tag{3.14}$$

where $Y \in \mathbb{R}^n, X \in \mathbb{R}^{n \times p}$, $M \in \{+1, -1\}^p$ is a sign vector, and $Z \in \mathbb{R}^n$ is a noise vector. The goal is to recover the unknown sign signal M. Exhaustive search over the parameter set is computationally prohibitive. The noiseless case of (3.14), known as the generalized multi-knapsack problem (Khuri et al., 1994; Mangasarian and Recht, 2011), can be solved through an integer program which is known to be computationally difficult even for checking the uniqueness of the solution, see (Prokopyev et al., 2005; Valiant and Vazirani, 1986).

Orthogonal Matrix Recovery. In some applications the matrix of interest in trace regression is known to be an orthogonal/rotation matrix (Ten Berge, 1977; Gower and Dijksterhuis, 2004). More specifically, in orthogonal matrix recovery, we observe (X_i, Y_i), $i = 1, \ldots, n$ as in the trace regression model (3.13) where $X_i \in \mathbb{R}^{m \times m}$ are measurement matrices and $M \in \mathbb{R}^{m \times m}$ is an orthogonal matrix. The goal is to recover the unknown M using an efficient algorithm. Computational difficulties come in because of the non-convex constraint.

Other high-dimensional inference problems that are closely connected to the structured linear inverse model (3.11) include Matrix Completion Candes and Plan (2010); Chatterjee (2012); Cai and Zhou (2013), sparse and low rank decomposition in robust principal component analysis (Candès et al., 2011), and sparse noise and sparse parameter in demixing problem (Amelunxen et al., 2013), to name a few. We will discuss the connections in Section 3.2.3.

There are several fundamental questions for this general class of high-dimensional linear inverse problems:

Statistical Questions: How well can the parameter M be estimated? What is the intrinsic difficulty of the estimation problem? How to provide inference guarantees for M, i.e., confidence intervals and hypothesis testing, in general?

Computational Questions: Are there computationally efficient (polynomial time complexity) algorithms that are also sharp in terms of statistical estimation and inference?

High-Dimensional Linear Inverse Problems

Linear inverse problems have been well studied in the classical setting where the parameter of interest lies in a convex set. See, for example, Tikhonov and Arsenin (1977), O'Sullivan (1986), and Johnstone and Silverman (1990). In particular, for estimation of a linear functional over a convex parameter space, Donoho (1994) developed an elegant geometric characterization of the minimax theory in terms of the modulus of continuity. However, the theory relies critically on the convexity assumption of the parameter space. As shown in Cai and Low (2004a,b), the behavior of the functional estimation and confidence interval problems is significantly different even when the parameter space is the union of two convex sets. For the high-dimensional linear inverse problems considered in the present thesis, the parameter space is highly non-convex and the theory and techniques developed in the classical setting are not readily applicable.

For high-dimensional linear inverse problems such as those mentioned earlier, the parameter space has low-complexity and exhaustive search often leads to the optimal solution in terms of statistical accuracy. However, it is computationally prohibitive and requires the prior knowledge of the true low complexity. In recent years, relaxing the problem to a convex program such as ℓ_1 or nuclear norm minimization and then solving it with optimization techniques has proven to be a powerful approach in individual cases.

Unified approaches to signal recovery recently appeared both in the applied mathematics literature (Chandrasekaran et al., 2012; Amelunxen et al., 2013; Oymak et al., 2013) and in the statistics literature (Negahban et al., 2012). Oymak et al. (2013) studied the generalized LASSO problem through conic geometry with a simple bound in terms of the ℓ_2 norm of the noise vector (which may not vanish to 0 as sample size n increases). (Chandrasekaran et al., 2012) introduced the notion of atomic norm to define a low complexity structure and showed that Gaussian width captures the minimum sample size required to ensure recovery. Amelunxen et al. (2013) studied the phase transition for the convex algorithms for a wide range of problems. These suggest that the geometry of the local tangent cone determines the minimum number of samples to ensure successful recovery in the noiseless or deterministic noise settings. Negahban et al. (2012) studied the regularized M-estimation with a decomposable norm penalty in the additive Gaussian noise setting.

Another line of research is focused on a detailed analysis of the Empirical Risk Minimization (ERM) (Lecué and Mendelson, 2013). The analysis is based on the empirical processes theory, with a proper localized rather than global analysis. In addition to convexity, the ERM requires the prior knowledge on the size of the bounded parameter set of interest. This knowledge is not needed for the algorithm we propose in the present thesis.

Compared to estimation, there is a paucity of methods and theoretical results for confidence intervals and hypothesis testing for these linear inverse models. Specifically for high-dimensional linear regression, Bühlmann (2013) studied a bias correction method based on ridge estimation, while Zhang and Zhang (2014) proposed bias correction via score vector using scaled Lasso as the initial estimator. van de Geer et al. (2014); Javanmard and Montanari (2014) focused on de-sparsifying Lasso by constructing a near inverse of the Gram matrix; the first paper uses nodewise Lasso, while the other uses ℓ_∞ constrained quadratic programing, with similar theoretical guarantees. To the best of our knowledge, a unified treatment of inference procedures for general high-dimensional linear inverse models is yet to be developed.

Geometric Characterization of Linear Inverse Problems

We take a geometric perspective in studying the model (3.11). The parameter M inherits certain low complexity structure with respect to a given atom set in a high-dimensional space, thus introducing computationally difficult non-convex constraints. However, proper convex relaxation based on the atom structure provides a computationally feasible solution. For point estimation, we are interested in how the local convex geometry around the true parameter affects the estimation procedure and the intrinsic estimation difficulty. For inference, we develop general procedures induced by the convex geometry, addressing inferential questions such as confidence intervals and hypothesis testing. We are also interested in the sample size condition induced by the local convex geometry for valid inference guarantees. This local geometry plays a key role in our analysis.

Complexity measures such as Gaussian width and Rademacher complexity are well studied in the empirical processes theory (Ledoux and Talagrand, 1991; Talagrand, 1996a), and are known to capture the difficulty of the estimation problem. Covering/Packing entropy and volume ratio (Yang and Barron, 1999; Vershynin, 2011; Ma and Wu, 2013b) are also widely used in geometric functional analysis to measure the complexity. In this thesis, we will show how these geometric quantities affect the computationally efficient estimation/inference procedure, as well as the intrinsic difficulties.

Our Contributions

The main result can be summarized as follows:

> **Unified convex algorithms.** We propose a general computationally feasible convex program that provides near optimal rate of convergence simultaneously for a collection of high-dimensional linear inverse problems. We also study a general efficient convex program that leads to statistical inference for linear contrasts of M, such as confidence intervals and hypothesis testing. The point estimation and statistical inference are

adaptive in the sense that the difficulty (rate of convergence, conditions on sample size, etc.) automatically adapts to the low complexity structure of the true parameter.

Local geometric theory. A unified theoretical framework is provided for analyzing high-dimensional linear inverse problems based on the local conic geometry and duality. Local geometric complexities govern the difficulty of statistical inference for the linear inverse problems.

Specifically, on the local tangent cone $T_{\mathcal{A}}(M)$ (defined in (3.18)), geometric quantities such as the Gaussian width $w(B_2^p \cap T_{\mathcal{A}}(M))$ and Sudakov minoration estimate $e(B_2^p \cap T_{\mathcal{A}}(M))$ (both defined in Section 3.2.2; B_2^p denotes unit Euclidean ball in $\mathbb{R}^p$) capture the rate of convergence. In terms of the upper bound, with overwhelming probability, if $n \gtrsim w^2(B_2^p \cap T_{\mathcal{A}}(M))$, the estimation error under ℓ_2 norm for our algorithm is

$$\sigma \frac{\gamma_{\mathcal{A}}(M) w(\mathcal{X}\mathcal{A})}{\sqrt{n}},$$

where $\gamma_{\mathcal{A}}(M)$ is the local asphericity ratio defined in (3.25). A minimax lower bound for estimation over the local tangent cone $T_{\mathcal{A}}(M)$ is

$$\sigma \frac{e(B_2^p \cap T_{\mathcal{A}}(M))}{\sqrt{n}}.$$

For statistical inference, we establish valid asymptotic normality for any linear functional $\langle v, M \rangle$ (with $\|v\|_{\ell_1}$ bounded) of the parameter M under the condition

$$\lim_{n,p(n)\to\infty} \frac{\gamma_{\mathcal{A}}^2(M) w^2(\mathcal{X}\mathcal{A})}{\sqrt{n}} = 0,$$

which can be compared to the condition for point estimation consistency

$$\lim_{n,p(n)\to\infty} \frac{\gamma_{\mathcal{A}}(M) w(\mathcal{X}\mathcal{A})}{\sqrt{n}} = 0.$$

There is a critical difference on the sufficient conditions between valid inference and estimation consistency — more stringent condition on sample size n is required for inference beyond estimation. Intuitively, statistical inference is purely geometrized by Gaussian width and Sudakov minoration estimate.

Organization

The rest of the section is structured as follows. In Section 3.2.2, after notation, definitions, and basic convex geometry are reviewed, we formally present convex programs for recovering the parameter M, and for providing inference guarantees for M. The properties of the proposed procedures are then studied in Section 3.2.3 under the Gaussian setting, where a geometric theory is developed, along with the minimax lower bound, as well as the confidence intervals and hypothesis testing. Applications to particular high-dimensional estimation problems are caculated in Section 3.2.3. Section 3.2.4 extends the geometric theory beyond the Gaussian case. Further discussions appear in Section 3.2.5, and the proofs of the main results are given in Appendix and Supplement Cai et al. (2014).

3.2.2. Preliminaries and Algorithms

Let us first review notation and definitions that will be used in the rest of the section. We use $\|\cdot\|_{\ell_q}$ to denote the ℓ_q norm of a vector or induced norm of a matrix, and use B_2^p to denote the unit Euclidean ball in $\mathbb{R}^p$. For a matrix M, denote by $\|M\|_F$, $\|M\|_*$, and $\|M\|$ the Frobenius norm, nuclear norm, and spectral norm of M respectively. When there is no confusion, we also denote $\|M\|_F = \|M\|_{\ell_2}$ for a matrix M. For a vector $V \in \mathbb{R}^p$, denote its transpose by V^*. The inner product on vectors is defined as usual $\langle V_1, V_2 \rangle = V_1^* V_2$. For matrices $\langle M_1, M_2 \rangle = \mathsf{Tr}(M_1^* M_2) = \mathsf{Vec}(M_1)^* \mathsf{Vec}(M_2)$, where $\mathsf{Vec}(M) \in \mathbb{R}^{pq}$ denotes the vectorized version of matrix $M \in \mathbb{R}^{p \times q}$. $\mathcal{X} : \mathbb{R}^p \to \mathbb{R}^n$ denotes a linear operator from $\mathbb{R}^p$ to $\mathbb{R}^n$. Following the notation above, $M^* \in \mathbb{R}^{q \times p}$ is the adjoint (transpose) matrix of M and $\mathcal{X}^* : \mathbb{R}^n \to \mathbb{R}^p$ is the adjoint operator of $\mathcal{X}$ such that $\langle \mathcal{X}(V_1), V_2 \rangle = \langle V_1, \mathcal{X}^*(V_2) \rangle$.

For a convex compact set K in a metric space with the metric d, the ϵ-entropy for a convex compact set K with respect to the metric d is denoted in the following way: ϵ-

packing entropy $\log \mathcal{M}(K, \epsilon, d)$ is the logarithm of the cardinality of the largest ϵ-packing set. Similarly, ϵ-covering entropy $\log \mathcal{N}(K, \epsilon, d)$ is the log-cardinality of the smallest ϵ-covering set with respect to metric d. A well known result is $\mathcal{M}(K, 2\epsilon, d) \leq \mathcal{N}(K, \epsilon, d) \leq \mathcal{M}(K, \epsilon, d)$. When the metric d is the usual Euclidean distance, we will omit d in $\mathcal{M}(K, \epsilon, d)$ and $\mathcal{N}(K, \epsilon, d)$ and simply write $\mathcal{M}(K, \epsilon)$ and $\mathcal{N}(K, \epsilon)$.

For two sequences of positive numbers $\{a_n\}$ and $\{b_n\}$, we denote $a_n \gtrsim b_n$ and $a_n \lesssim b_n$ if there exist constants c_0, C_0 such that $\frac{a_n}{b_n} \geq c_0$ and $\frac{a_n}{b_n} \leq C_0$ respectively, for all n. We write $a_n \asymp b_n$ if $a_n \gtrsim b_n$ and $a_n \lesssim b_n$. Throughout the thesis, c, C denote constants that may vary from place to place.

Basic Convex Geometry

The notion of low complexity is based on a collection of basic atoms. We denote the collection of these basic atoms as an atom set $\mathcal{A}$, either countable or uncountable. A parameter M is of complexity k in terms of the atoms in $\mathcal{A}$ if M can be expressed as a linear combination of at most k atoms in $\mathcal{A}$, i.e., there exists a decomposition

$$M = \sum_{a \in \mathcal{A}} c_a(M) \cdot a, \text{ where } \sum_{a \in \mathcal{A}} 1_{\{c_a(M) \neq 0\}} \leq k.$$

In convex geometry (Pisier, 1999), the Minkowski functional (gauge) of a symmetric convex body K is defined as

$$\|x\|_K = \inf\{t > 0 : x \in tK\}.$$

Let $\mathcal{A}$ be a collection of atoms that is a compact subset of $\mathbb{R}^p$. Without loss of generality, assume $\mathcal{A}$ is contained inside ℓ_∞ ball. We assume that the elements of $\mathcal{A}$ are extreme points of the convex hull $\mathsf{conv}(\mathcal{A})$ (in the sense that for any $x \in \mathbb{R}^p$, $\sup\{\langle x, a\rangle : a \in \mathcal{A}\} = \sup\{\langle x, a\rangle : a \in \mathsf{conv}(\mathcal{A})\}$). The atomic norm $\|x\|_{\mathcal{A}}$ for any $x \in \mathbb{R}^p$ is defined as the gauge of $\mathsf{conv}(\mathcal{A})$:

$$\|x\|_{\mathcal{A}} = \inf\{t > 0 : x \in t\ \mathsf{conv}(\mathcal{A})\}.$$

As noted in Chandrasekaran et al. (2012), the atomic norm can also be written as

$$\|x\|_{\mathcal{A}} = \inf \left\{ \sum_{a \in \mathcal{A}} c_a : x = \sum_{a \in \mathcal{A}} c_a \cdot a,\ c_a \geq 0 \right\}. \tag{3.15}$$

The dual norm of this atomic norm is defined in the following way (since the atoms in $\mathcal{A}$ are the extreme points of $\mathsf{conv}(\mathcal{A})$),

$$\|x\|_{\mathcal{A}}^* = \sup\{\langle x, a\rangle : a \in \mathcal{A}\} = \sup\{\langle x, a\rangle : \|a\|_{\mathcal{A}} \leq 1\}. \tag{3.16}$$

We have the following ("Cauchy-Schwarz") symmetric relation for the norm and its dual

$$\langle x, y\rangle \leq \|x\|_{\mathcal{A}}^* \|y\|_{\mathcal{A}}. \tag{3.17}$$

It is clear that the unit ball with respect to the atomic norm $\|\cdot\|_{\mathcal{A}}$ is the convex hull of the set of atoms $\mathcal{A}$. The **tangent cone** at x with respect to the scaled unit ball $\|x\|_{\mathcal{A}}\ \mathsf{conv}(\mathcal{A})$ is defined to be

$$T_{\mathcal{A}}(x) = \mathsf{cone}\left\{h : \|x + h\|_{\mathcal{A}} \leq \|x\|_{\mathcal{A}}\right\}. \tag{3.18}$$

Also known as a recession cone, $T_{\mathcal{A}}(x)$ is the collection of directions where the atomic norm becomes smaller. The "size" of the tangent cone at the true parameter M will affect the difficulty of the recovery problem. We focus on the cone intersected with the unit ball $B_2^p \cap T_{\mathcal{A}}(M)$ in analyzing the complexity of the cone. See Figure 14 for an intuitive illustration.

It is helpful to look at the atom set, atomic norm and tangent cone geometry in a few

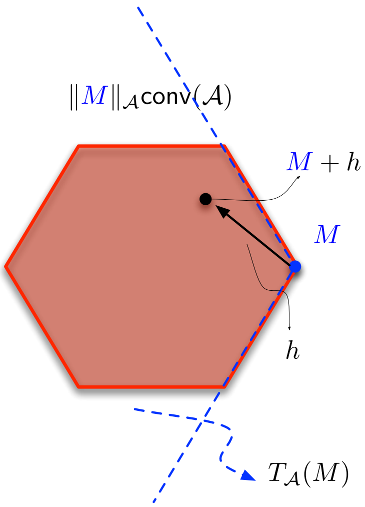

Figure 14: Tangent cone: general illustration in 2D. The red shaped area is the scaled convex hull of atom set. The blue dashed line forms the tangent cone at M. Black arrow denotes the possible directions inside the cone.

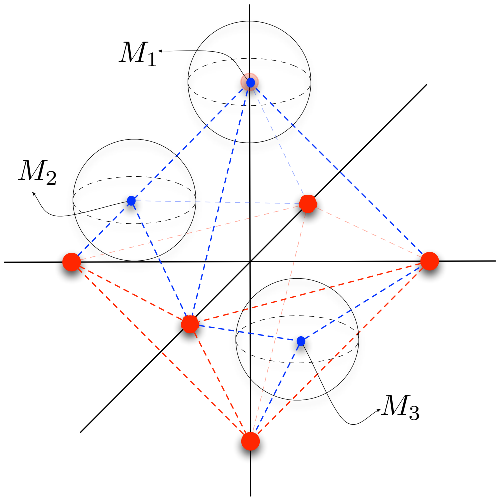

Figure 15: Tangent cone illustration in 3D for sparse regression. For three possible locations $M_i, 1 \leq i \leq 3$, the tangent cone are different, with cones becoming more complex as i increases.

examples to better illustrate the general model and notion of low complexity.

Example 3.2.1. For sparse signal recovery in high-dimensional linear regression, the atom set consists of the unit basis vectors $\{\pm e_i\}$, the atomic norm is the vector ℓ_1 norm, and its dual norm is the vector ℓ_∞ norm. The convex hull $\mathsf{conv}(\mathcal{A})$ is called the cross-polytope. Figure 15 illustrates this tangent cone for 3D ℓ_1 norm ball for 3 different cases $T_{\mathcal{A}}(M_i), 1 \leq i \leq 3$. The "angle" or "complexity" of the local tangent cone determines the difficulty of recovery. Previous work showed that the algebraic characterization (sparsity) of the parameter space drives the global rate, and we are arguing that the geometric characterization through the local tangent cone provides an intuitive and refined local approach.

Example 3.2.2. In trace regression and matrix completion, the goal is to recover low rank matrices. In such settings, the atom set consists of the rank one matrices (matrix manifold) $\mathcal{A} = \{uv^* : \|u\|_{\ell_2} = 1, \; \|v\|_{\ell_2} = 1\}$ and the atomic norm is the nuclear norm and the dual norm is the spectral norm. The convex hull $\mathsf{conv}(\mathcal{A})$ is called the nuclear norm ball of matrices. The position of the true parameter on the scaled nuclear norm ball determines the geometry of the local tangent cone, thus affecting the estimation difficulty.

Example 3.2.3. In integer programming, one would like to recover the sign vectors whose entries take on values ±1. The atom set is all sign vectors (cardinality 2^p) and the convex hull $\mathsf{conv}(\mathcal{A})$ is the hypercube. Tangent cones for each parameter have the same structure in this case.

Example 3.2.4. In orthogonal matrix recovery, the matrix of interest is constrained to be orthogonal. In this case, the atom set is all orthogonal matrices and the convex hull $\mathsf{conv}(\mathcal{A})$ is the spectral norm ball. Similar to sign vector recovery, the local tangent cones for each orthogonal matrix share similar geometric property.

Gaussian Width, Sudakov Estimate, and Other Geometric Quantities

We first introduce two complexity measures, the Gaussian width and Sudakov estimate.

Definition 3.2.1 (Gaussian Width)**.** *For a compact set $K \in \mathbb{R}^p$, the Gaussian width is defined as*

$$w(K) := \mathbb{E}_g \left[\sup_{v \in K} \langle g, v \rangle \right]. \tag{3.19}$$

where $g \sim N(0, I_p)$ is the standard multivariate Gaussian vector.

Gaussian width quantifies the probability that a randomly oriented subspace misses a convex subset. It was used in Gordon's analysis (Gordon, 1988), and was shown recently to play a crucial rule in linear inverse problems in various noiseless or deterministic noise settings, see, for example, Chandrasekaran et al. (2012); Amelunxen et al. (2013). Explicit upper bounds on the Gaussian width for different convex sets have been given in Chandrasekaran et al. (2012); Amelunxen et al. (2013). For example, if $M \in \mathbb{R}^p$ is a s-sparse vector, $w(B_2^p \cap T_{\mathcal{A}}(M)) \lesssim \sqrt{s \log p/s}$. When $M \in \mathbb{R}^{p \times q}$ is a rank-r matrix, $w(B_2^p \cap T_{\mathcal{A}}(M)) \lesssim \sqrt{r(p+q-r)}$. For sign vector in $\mathbb{R}^p$, $w(B_2^p \cap T_{\mathcal{A}}(M)) \lesssim \sqrt{p}$, while for orthogonal matrix in $\mathbb{R}^{m \times m}$, $w(B_2^p \cap T_{\mathcal{A}}(M)) \lesssim \sqrt{m(m-1)}$. See Section 3.4 propositions 3.10-3.14 in Chandrasekaran et al. (2012) for detailed calculations. The Gaussian width as a complexity measure of the local tangent cone will be used in the upper bound analysis in Sections 3.2.3 and 3.2.4.

Definition 3.2.2 (Sudakov Minoration Estimate)**.** *The Sudakov estimate of a compact set $K \in \mathbb{R}^p$ is defined as*

$$e(K) := \sup_{\epsilon} \ \epsilon\sqrt{\log \mathcal{N}(K, \epsilon)}. \tag{3.20}$$

where $\mathcal{N}(K, \epsilon)$ denotes the ϵ-covering number of set K with respect to the Euclidean norm.

Sudakov estimate has been used in the literature as a measure of complexity for a general functional class that nearly matches (from below) the expected supremum of a gaussian process. By balancing the cardinality of the covering set at scale ϵ and the covering radius ϵ, the estimate maximizes

$$\epsilon\sqrt{\log \mathcal{N}(B_2^p \cap T_{\mathcal{A}}(M), \epsilon)},$$

thus determining the complexity of the cone $T_{\mathcal{A}}(M)$. Sudakov estimate as a complexity measure of the local tangent cone is useful for the minimax lower bound analysis.

The following well-known result (Dudley, 1967; Ledoux and Talagrand, 1991) establishes a relation between the Gaussian width $w(\cdot)$ and Sudakov estimate $e(\cdot)$:

Lemma 3.2.1 (Sudakov Minoration and Dudley Entropy Integral)**.** *For any compact subset $K \subseteq \mathbb{R}^p$, there exist a universal constant $c > 0$ such that*

$$c \cdot e(K) \leq w(K) \leq 24 \int_0^\infty \sqrt{\log \mathcal{N}(K, \epsilon)} d\epsilon. \tag{3.21}$$

In the literature, another complexity measure—volume ratio—has also been used to characterize the minimax lower bounds (Ma and Wu, 2013b). Volume ratio has been studied in Pisier (1999) and Vershynin (2011). For a convex set $K \in \mathbb{R}^p$, volume ratio used in the present thesis is defined as follows.

Definition 3.2.3 (Volume Ratio). *The volume ratio is defined as*

$$v(K) := \sqrt{p}\left(\frac{\mathsf{vol}(K)}{\mathsf{vol}(B_2^p)}\right)^{\frac{1}{p}}. \tag{3.22}$$

The recovery difficulty of the linear inverse problem also depends on other geometric quantities defined on the local tangent cone $T_{\mathcal{A}}(M)$: the local isometry constants $\phi_{\mathcal{A}}(M, \mathcal{X})$ and $\psi_{\mathcal{A}}(M, \mathcal{X})$ and the local asphericity ratio $\gamma_{\mathcal{A}}(M)$. The **local isometry constants** are defined for the local tangent cone at the true parameter M as

$$\phi_{\mathcal{A}}(M, \mathcal{X}) := \inf\left\{\frac{\|\mathcal{X}(h)\|_{\ell_2}}{\|h\|_{\ell_2}} : h \in T_{\mathcal{A}}(M), h \neq 0\right\} \tag{3.23}$$

$$\psi_{\mathcal{A}}(M, \mathcal{X}) := \sup\left\{\frac{\|\mathcal{X}(h)\|_{\ell_2}}{\|h\|_{\ell_2}} : h \in T_{\mathcal{A}}(M), h \neq 0\right\}. \tag{3.24}$$

The local isometry constants measure how well the linear operator preserves the ℓ_2 norm within the local tangent cone. Intuitively, the larger the ψ or the smaller the ϕ is, the harder the recovery is. We will see later that the local isometry constants are determined by the Gaussian width under the Gaussian ensemble design.

The **local asphericity ratio** is defined as

$$\gamma_{\mathcal{A}}(M) := \sup\left\{\frac{\|h\|_{\mathcal{A}}}{\|h\|_{\ell_2}} : h \in T_{\mathcal{A}}(M), h \neq 0\right\} \tag{3.25}$$

and measures how extreme the atomic norm is relative to the ℓ_2 norm within the local tangent cone.

Point Estimation via Convex Relaxation

We now return to the linear inverse model (3.11) in the high-dimensional setting. Suppose we observe $(\mathcal{X},\ Y)$ as in (3.11) where the parameter of interest M is assumed to have low complexity with respect to a given atom set $\mathcal{A}$. The low complexity of M introduces a non-convex constraint, which leads to serious computational difficulties if solved directly. Convex relaxation is an effective and natural approach in such a setting. In most interesting cases, the atom set is not too rich in the sense that $\mathsf{conv}(\mathcal{A}) \subset B_2^p$. For such cases, we propose a generic convex constrained minimization procedure induced by the atomic norm and the corresponding dual norm to estimate M:

$$\hat{M} = \arg\min_{M} \{\|M\|_{\mathcal{A}} : \ \|\mathcal{X}^*(Y - \mathcal{X}(M))\|_{\mathcal{A}}^* \leq \lambda\} \tag{3.26}$$

where λ is a localization radius (tuning parameter) that depends on the sample size, noise level, and geometry of the atom set $\mathcal{A}$. An explicit formula for λ is given in (3.31) in the case of Gaussian noise. The atomic norm minimization (3.26) is a convex relaxation of the low complexity structure, and λ specifies the localization scale based on the noise. This generic convex program utilizes the duality and recovers the low complexity structure adaptively. The Dantzig selector for high-dimensional sparse regression (Candès and Tao, 2007) and the constrained nuclear norm minimization Candes and Plan (2011) for trace regression are particular examples of (3.26). The properties of the estimator $\hat{M}$ will be investigated in Sections 3.2.3 and 3.2.4.

In cases where the atomic norm ball is rich, i.e. $\mathsf{conv}(\mathcal{A}) \not\subset B_2^p$, a slightly stronger program

$$\hat{M} = \arg\min_{M} \{\|M\|_{\mathcal{A}} : \ \|\mathcal{X}^*(Y - \mathcal{X}(M))\|_{\mathcal{A}}^* \leq \lambda, \ \|\mathcal{X}^*(Y - \mathcal{X}(M))\|_{\ell_2} \leq \mu\} \tag{3.27}$$

with λ, μ as tuning parameters will yield optimal guarantees. The analysis of (3.27) is essentially the same as (3.26). For conciseness, we will present the main result for the interesting case (3.26). We remark that the atomic dual norm constraint is crucial for attaining optimal behavior unless $\mathsf{conv}(\mathcal{A}) \supset B_2^p$. For instance, the convex program in Chandrasekaran et al. (2012) with only the ℓ_2 constraint will lead to a suboptimal estimator.

Statistical Inference via Feasibility of Convex Program

In the high-dimensional setting, p-values as well as confidence intervals are important inferential questions beyond point estimation. In this section we will show how to perform statistical inference for the linear inverse model (3.11). Let $M \in \mathbb{R}^p$ be the vectorized parameter of interest, and $\{e_i, 1 \leq i \leq p\}$ are the corresponding basis vectors. Consider the following convex feasibility problem for matrix $\Omega \in \mathbb{R}^{p \times p}$, where each row $\Omega_{i\cdot}$ satisfies

$$\|\mathcal{X}^* \mathcal{X} \Omega_{i\cdot}^* - e_i\|_{\mathcal{A}}^* \leq \eta, \quad \forall 1 \leq i \leq p. \tag{3.28}$$

Here η is some tuning parameter that depends on the sample size and geometry of the atom set $\mathcal{A}$. One can also solve a stronger version of the above convex program for $\eta \in \mathbb{R}, \Omega \in \mathbb{R}^{p \times p}$ simultaneously:

$$(\Omega, \eta_n) = \underset{\Omega, \eta}{\arg\min} \left\{ \eta : \ \|\mathcal{X}^* \mathcal{X} \Omega_{i\cdot}^* - e_i\|_{\mathcal{A}}^* \leq \eta, \quad \forall 1 \leq i \leq p \right\}. \tag{3.29}$$

Built upon the constrained minimization estimator $\hat{M}$ in (3.26) and feasible matrix Ω in (3.29), the de-biased estimator for inference on parameter M is defined as

$$\tilde{M} := \hat{M} + \Omega \mathcal{X}^* (Y - \mathcal{X}(\hat{M})). \tag{3.30}$$

We will establish the asymptotic normality for linear contrast $\langle v, M \rangle$, where $v \in \mathbb{R}^p, \|v\|_{\ell_1} \leq \rho$, ρ does not grow with $n, p(n)$, and construct confidence intervals and hypothesis tests based on the asymptotic normality result. In the case of high-dimensional linear regression, de-biased estimators has been investigated in Bühlmann (2013); Zhang and Zhang (2014); van de Geer et al. (2014); Javanmard and Montanari (2014). The convex feasibility program we proposed here can be viewed as a unified treatment for general linear inverse models. We will show that under some conditions on the sample size and the local tangent cone, asymptotic confidence intervals and hypothesis tests are valid for linear contrast $\langle v, M \rangle$

which include as a special case the individual coordinates of M.

3.2.3. Local Geometric Theory: Gaussian Setting

We establish in this section a general theory of geometric inference in the Gaussian setting where the noise vector Z is Gaussian and the linear operator $\mathcal{X}$ is the Gaussian ensemble design (Definition 3.2.4). In analyzing Model 3.11, without loss of generality, we can scale $\mathcal{X}, Z$ simultaneously such that column ℓ_2 norm does not grow with n. In the stochastic noise setting, the noise $Z_i, 1 \leq i \leq n$ is scaled correspondingly to noise level $\sigma/\sqrt{n}$.

Definition 3.2.4 (Gaussian Ensemble Design)**.** *Let $\mathcal{X} \in \mathbb{R}^{n \times p}$ be the matrix form of the linear operator $\mathcal{X} : \mathbb{R}^p \to \mathbb{R}^n$. $\mathcal{X}$ is Gaussian ensemble if each element is an i.i.d Gaussian random variable with mean 0 and variance $\frac{1}{n}$.*

Our analysis is quite different from the case by case global analysis of the Dantzig selector, Lasso and nuclear norm minimization. We show a stronger result which adapts to the local tangent cone geometry. All the analyses in our theory are non-asymptotic, and the constants are explicit. Another advantage is that the local analysis yields robustness for a given parameter (with near but not exact low complexity), as the convergence rate is captured by the geometry of the associated local tangent cone at a given M. Later in Section 3.2.4 we will show how to extend the theory to a more general setting.

Local Geometric Upper Bound

For the upper bound analysis, we need to choose a suitable localization radius λ (in the convex program (3.26)) to guarantee that the true parameter M is in the feasible set with high probability. In the case of Gaussian noise the tuning parameter is chosen as

$$\lambda_{\mathcal{A}}(\mathcal{X}, \sigma, n) = \frac{\sigma}{\sqrt{n}} \left\{ w(\mathcal{X}\mathcal{A}) + \delta \cdot \sup_{v \in \mathcal{A}} \|\mathcal{X} v\|_{\ell_2} \right\} \asymp \frac{\sigma}{\sqrt{n}} w(\mathcal{X}\mathcal{A}) \qquad (3.31)$$

where $\mathcal{X}T$ is the image of the set T under the linear operator $\mathcal{X}$, and $\delta > 0$ can be chosen arbitrarily according to the probability of success we would like to attain (δ is commonly chosen at order $\sqrt{\log p}$). $\lambda_{\mathcal{A}}(\mathcal{X}, \sigma, n)$ is a global parameter that depends on the linear operator $\mathcal{X}$ and the atom set $\mathcal{A}$, but, importantly, not on the complexity of M. The following theorem geometrizes the local rate of convergence in the Gaussian case.

Theorem 3.2.1 (Gaussian Ensemble: Convergence Rate)**.** *Suppose we observe* $(\mathcal{X},\ Y)$ *as in* (3.11) *with the Gaussian ensemble design and* $Z \sim N(0, \frac{\sigma^2}{n} I_n)$*. Let* $\hat{M}$ *be the solution of* (3.26) *with* λ *chosen as in* (3.31)*. Let* $0 < c < 1$ *be a constant. For any* $\delta > 0$*, if*

$$n \geq \frac{4[w(B_2^p \cap T_{\mathcal{A}}(M)) + \delta]^2}{c^2} \vee \frac{1}{c},$$

then with probability at least $1 - 3\exp(-\delta^2/2)$*,*

$$\begin{aligned} &\|\hat{M} - M\|_{\mathcal{A}} \leq \gamma_{\mathcal{A}}(M) \cdot \|\hat{M} - M\|_{\ell_2}, \quad \textit{and further we have} \\ &\|\hat{M} - M\|_{\ell_2} \leq \frac{1}{1-c}\|\mathcal{X}(\hat{M} - M)\|_{\ell_2} \leq \frac{2\sigma}{(1-c)^2} \cdot \frac{\gamma_{\mathcal{A}}(M) w(\mathcal{X}\mathcal{A})}{\sqrt{n}}. \end{aligned}$$

Theorem 3.2.1 gives bounds for the estimation error under both the ℓ_2 norm loss and the atomic norm loss, as well as for the in sample prediction error. The upper bounds are determined by the geometric quantities $w(\mathcal{X}\mathcal{A}), \gamma_{\mathcal{A}}(M)$ and $w(B_2^p \cap T_{\mathcal{A}}(M))$. Take, for example, the estimation error under the ℓ_2 loss. Given any $\epsilon > 0$, the smallest sample size n to ensure the recovery error $\|\hat{M} - M\|_{\ell_2} \leq \epsilon$ with probability at least $1 - 3\exp(-\delta^2/2)$ is

$$n \geq \max\left\{\frac{4\sigma^2}{(1-c)^4} \cdot \frac{\gamma_{\mathcal{A}}^2(M) w^2(\mathcal{X}\mathcal{A})}{\epsilon^2},\ \frac{4w^2(B_2^p \cap T_{\mathcal{A}}(M))}{c^2}\right\}.$$

That is, the minimum sample size for guaranteed statistical accuracy is driven by two geometric terms $w(\mathcal{X}\mathcal{A})\gamma_{\mathcal{A}}(M)$ and $w(B_2^p \cap T_{\mathcal{A}}(M))$. We will see in Section 3.2.3 that these two rates match in a range of specific high-dimensional estimation problems.

The proof of Theorem 3.2.1 (and Theorem 3.2.4 in Section 3.2.4) relies on the following two key lemmas.

Lemma 3.2.2 (Choice of Tuning Parameter)**.** *Consider the linear inverse model* (3.11) *with* $Z \sim N(0, \frac{\sigma^2}{n} I_n)$*. For any* $\delta > 0$*, with probability at least* $1 - \exp(-\delta^2/2)$ *on the* σ*-field of* Z *(conditional on* $\mathcal{X}$*),*

$$\|\mathcal{X}^*(Z)\|_{\mathcal{A}}^* \leq \frac{\sigma}{\sqrt{n}} \left\{ w(\mathcal{X}\mathcal{A}) + \delta \cdot \sup_{v \in \mathcal{A}} \|\mathcal{X}v\|_{\ell_2} \right\}. \tag{3.32}$$

This lemma is proved in Appendix. The particular value of $\lambda_{\mathcal{A}}(\mathcal{X}, \sigma, n)$ for a range of examples will be calculated in Section 3.2.3.

The next lemma addresses the local behavior of the linear operator $\mathcal{X}$ around the true parameter M under the Gaussian ensemble design. We call a linear operator *locally near-isometric* if the local isometry constants are uniformly bounded. The following lemma tells us that in the most widely used Gaussian ensemble case, the local isometry constants are guaranteed to be bounded, given the sample size n is at least of order $[w(B_2^p \cap T_{\mathcal{A}}(M))]^2$. Hence, the difficulty of the problem is captured by the Gaussian width.

Lemma 3.2.3 (Local Isometry Bound for Gaussian Ensemble)**.** *Assume the linear operator* $\mathcal{X}$ *is the Gaussian ensemble design. Let* $0 < c < 1$ *be a constant. For any* $\delta > 0$*, if*

$$n \geq \frac{4[w(B_2^p \cap T_{\mathcal{A}}(M)) + \delta]^2}{c^2} \vee \frac{1}{c},$$

then with probability at least $1 - 2\exp(-\delta^2/2)$*, the local isometry constants are around 1 with*

$$\phi_{\mathcal{A}}(M, \mathcal{X}) \geq 1 - c \quad \textit{and} \quad \psi_{\mathcal{A}}(M, \mathcal{X}) \leq 1 + c.$$

Local Geometric Inference: Confidence Intervals and Hypothesis Testing

For statistical inference on the general linear inverse model, we would like to choose the smallest η in (3.28) to ensure that, under the Gaussian ensemble design, the feasibility set for (3.28) is non-empty with high probability. The following theorem establishes geometric inference for Model (3.11).

Theorem 3.2.2 (Geometric Inference)**.** *Suppose we observe* $(\mathcal{X},\ Y)$ *as in* (3.11) *with the Gaussian ensemble design and* $Z \sim N(0, \frac{\sigma^2}{n} I_n)$*. Let* $\hat{M} \in \mathbb{R}^p, \Omega \in \mathbb{R}^{p \times p}$ *be the solution of* (3.26) *and* (3.28) *, and let* $\tilde{M} \in \mathbb{R}^p$ *be the de-biased estimator as in* (3.30)*. Assume* $p \geq n \gtrsim w^2(B_2^p \cap T_{\mathcal{A}}(M))$*. If the tuning parameters* λ, η *are chosen with*

$$\lambda \asymp \frac{\sigma}{\sqrt{n}} w(\mathcal{X}\mathcal{A}), \quad \eta \asymp \frac{1}{\sqrt{n}} w(\mathcal{X}\mathcal{A}),$$

convex programs (3.26) *and* (3.28) *have non-empty feasibility set for* Ω *with high probability. The following decomposition*

$$\tilde{M} - M = \Delta + \frac{\sigma}{\sqrt{n}} \Omega \mathcal{X}^* W \tag{3.33}$$

holds, where $W \sim N(0, I_n)$ *is the standard Gaussian vector with*

$$\Omega \mathcal{X}^* W \sim N(0, \Omega \mathcal{X}^* \mathcal{X} \Omega^*)$$

and $\Delta \in \mathbb{R}^p$ *satisfies* $\|\Delta\|_{\ell_\infty} \precsim \gamma_{\mathcal{A}}^2(M) \cdot \lambda\eta \asymp \sigma \frac{\gamma_{\mathcal{A}}^2(M) w^2(\mathcal{X}\mathcal{A})}{n}$*. Suppose* $(n, p(n))$ *as a sequence satisfies*

$$\limsup_{n, p(n) \to \infty} \frac{\gamma_{\mathcal{A}}^2(M) w^2(\mathcal{X}\mathcal{A})}{\sqrt{n}} = 0,$$

then for any $v \in \mathbb{R}^p, \|v\|_{\ell_1} \leq \rho$ *with* ρ *finite, we have the asymptotic normality for the functional* $\langle v, \tilde{M} \rangle$*,*

$$\frac{\sqrt{n}}{\sigma} \Big(\langle v, \tilde{M} \rangle - \langle v, M \rangle \Big) = \sqrt{v^* [\Omega \mathcal{X}^* \mathcal{X} \Omega^*] v} \cdot Z_0 + o_p(1) \tag{3.34}$$

where $Z_0 \sim N(0,1)$ *and* $\lim_{n, p(n) \to \infty} o_p(1) = 0$ *means convergence in probability.*

It follows from Theorem 3.2.2 that a valid asymptotic $(1-\alpha)$-level confidence intervals for

$M_i, 1 \le i \le p$ (when v is taken as e_i in Theorem 3.2.2) is

$$\left[\tilde{M}_i + \Phi^{-1}\left(\frac{\alpha}{2}\right)\sigma\sqrt{\frac{[\Omega\mathcal{X}^*\mathcal{X}\Omega^*]_{ii}}{n}}, \quad \tilde{M}_i + \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sigma\sqrt{\frac{[\Omega\mathcal{X}^*\mathcal{X}\Omega^*]_{ii}}{n}} \right]. \tag{3.35}$$

If we are interested in a linear contrast $\langle v, M\rangle = v_0$, $\|v\|_{\ell_1} \le \rho$ with ρ fixed, consider the hypothesis testing problem

$$H_0 : \sum_{i=1}^{p} v_i M_i = v_0 \quad \text{v.s.} \quad H_\alpha : \sum_{i=1}^{p} v_i M_i \neq v_0.$$

The test statistic is $\frac{\sqrt{n}\left(\langle v,\tilde{M}\rangle - v_0\right)}{\sigma(v^*[\Omega\mathcal{X}^*\mathcal{X}\Omega^*]v)^{1/2}}$ and under the null, it follows an asymptotic standard normal distribution as $n \to \infty$. Similarly, the p-value is of the form $2-2\Phi^{-1}\left(\left|\frac{\sqrt{n}\left(\langle v,\tilde{M}\rangle - v_0\right)}{\sigma(v^*[\Omega\mathcal{X}^*\mathcal{X}\Omega^*]v)^{1/2}}\right|\right)$ as $n \to \infty$.

Note the asymptotic normality holds for any finite linear contrast, and the asymptotic variance nearly achieves the Fisher information lower bound, as Ω is an estimate of the inverse of $\mathcal{X}^*\mathcal{X}$. For fixed dimension inference, Fisher information lower bound is asymptotically optimal.

Remark 3.2.1. Note that the condition required for asymptotic normality and valid confidence intervals,

$$\lim_{n,p(n)\to\infty} \frac{\gamma_{\mathcal{A}}^2(M)w^2(\mathcal{X}\mathcal{A})}{\sqrt{n}} = 0,$$

is stronger than the one for estimation consistency of the parameter M under the ℓ_2 norm,

$$\lim_{n,p(n)\to\infty} \frac{\gamma_{\mathcal{A}}(M)w(\mathcal{X}\mathcal{A})}{\sqrt{n}} = 0.$$

For inference, we do require stronger condition in order to learn the order of the bias of the estimate. In the case when $n > p$ and the Gaussian ensemble design, $\mathcal{X}^*\mathcal{X}$ is non-singular with high probability. With the choice of $\Omega = (\mathcal{X}^*\mathcal{X})^{-1}$ and $\eta = 0$, for any $i \in [p]$, the following holds non-asymptotically,

$$\sqrt{n}(\tilde{M}_i - M_i) \sim N(0, \sigma^2[(\mathcal{X}^*\mathcal{X})^{-1}]_{ii}).$$

Extension: Correlated Design

The results in Section 3.2.3 and 3.2.3 can be extended beyond Gaussian ensemble (where $\mathbb{E}[\mathcal{X}^*\mathcal{X}] = I$) to Gaussian design with known covariance matrix Σ (where $\mathbb{E}[\mathcal{X}^*\mathcal{X}] = \Sigma$). Consider the following slightly modified point estimation and inference procedure (with tuning parameter λ, η)

$$\begin{aligned} \text{Point Estimation via } \hat{M} \quad & \hat{M} = \arg\min_M \{\|M\|_{\mathcal{A}} : \ \|\mathcal{X}^*(Y - \mathcal{X}(M))\|_{\mathcal{A}}^* \le \lambda\} \\ \text{Inference via } \tilde{M} \quad & \Omega : \|\mathcal{X}^*\mathcal{X}\Omega_{i\cdot}^* - \Sigma^{\frac{1}{2}} e_i\|_{\mathcal{A}}^* \le \eta, \quad \forall 1 \le i \le p \qquad (3.36) \\ & \tilde{M} := \hat{M} + \Sigma^{-\frac{1}{2}}\Omega\mathcal{X}^*(Y - \mathcal{X}(\hat{M})) \end{aligned}$$

where $\Omega \in \mathbb{R}^{p\times p}$ is an solution to the convex feasibility problem (3.36). Then the following Corollary holds.

Corollary 3.2.1. *Suppose we observe* $(\mathcal{X},\ Y)$ *as in* (3.11)*, where the Gaussian design* $\mathcal{X}$ *has covariance* Σ *and* $Z \sim N(0, \frac{\sigma^2}{n} I_n)$*. Consider the convex programs for estimation* $\hat{M}$ *and inference* $\tilde{M}$ *with the tuning parameters chosen as*

$$\lambda \asymp \frac{\sigma}{\sqrt{n}} w(\mathcal{X}\mathcal{A}), \quad \eta \asymp \frac{1}{\sqrt{n}} w(\mathcal{X}\mathcal{A}).$$

Under the condition $n \gtrsim w(B_2^p \cap \Sigma^{\frac{1}{2}} \circ T_{\mathcal{A}}(M))$*,* $\hat{M}$ *satisfies*

$$\|\hat{M} - M\|_{\ell_2} \precsim \sigma \frac{\gamma_{\mathcal{A}}(M) w(\mathcal{X}\mathcal{A})}{\sqrt{n}}, \quad \|\hat{M} - M\|_{\mathcal{A}} \precsim \sigma \frac{\gamma_{\mathcal{A}}^2(M) w(\mathcal{X}\mathcal{A})}{\sqrt{n}}.$$

Suppose $(n, p(n))$ as a sequence satisfies

$$\limsup_{n,p(n)\to\infty} \frac{\gamma_{\mathcal{A}}^2(M) w^2(\mathcal{X}\mathcal{A})}{\sqrt{n}} = 0,$$

then for any $v \in \mathbb{R}^p, \|v\|_{\ell_1} \leq \rho$ with ρ finite, we have the asymptotic normality for the functional $\langle \Sigma^{\frac{1}{2}} v, \tilde{M} \rangle$,

$$\frac{\sqrt{n}}{\sigma} \left(\langle \Sigma^{\frac{1}{2}} v, \tilde{M} \rangle - \langle \Sigma^{\frac{1}{2}} v, M \rangle \right) = \sqrt{v^* [\Omega \mathcal{X}^* \mathcal{X} \Omega^*] v} \cdot Z_0 + o_p(1)$$

where $Z_0 \sim N(0,1)$ and $\lim\limits_{n,p(n)\to\infty} o_p(1) = 0$ means convergence in probability.

Minimax Lower Bound for Local Tangent Cone

As seen in Section 3.2.3 and 3.2.3, the local tangent cone plays an important role in the upper bound analysis. In this section, we are interested in restricting the parameter space to the local tangent cone and seeing how the geometry of the cone affects the minimax lower bound.

Theorem 3.2.3 (Lower bound Based on Local Tangent Cone)**.** *Suppose we observe $(\mathcal{X}, Y)$ as in (3.11) with the Gaussian ensemble design and $Z \sim N(0, \frac{\sigma^2}{n} I_n)$. Let M be the true parameter of interest. Let $0 < c < 1$ be a constant. For any $\delta > 0$, if $n \geq \frac{4[w(B_2^p \cap T_{\mathcal{A}}(M)) + \delta]^2}{c^2} \vee \frac{1}{c}$. Then with probability at least $1 - 2\exp(-\delta^2/2)$,*

$$\inf_{\hat{M}} \sup_{M' \in T_{\mathcal{A}}(M)} \mathbb{E}_{\cdot|\mathcal{X}} \|\hat{M} - M'\|_{\ell_2}^2 \geq \frac{c_0 \sigma^2}{(1+c)^2} \cdot \left(\frac{e(B_2^p \cap T_{\mathcal{A}}(M))}{\sqrt{n}} \right)^2$$

for some universal constant $c_0 > 0$. Here $\mathbb{E}_{\cdot|\mathcal{X}}$ stands for the conditional expectation given the design matrix $\mathcal{X}$, and the probability statement is with respect to the distribution of $\mathcal{X}$ under the Gaussian ensemble design.

Recall Theorem 3.2.1, the local upper bound is basically determined by $\gamma_{\mathcal{A}}^2(M) w^2(\mathcal{X}\mathcal{A})$, which in many examples in Section 3.2.3 is of the rate $w^2(B_2^p \cap T_{\mathcal{A}}(M))$. The general relationship between these two quantities is given in Lemma 3.2.4 below, which is proved in Supplement Cai et al. (2014).

Lemma 3.2.4. *For any atom set $\mathcal{A}$, we have the following relation*

$$\gamma_{\mathcal{A}}(M)w(\mathcal{A}) \geq w(B_2^p \cap T_{\mathcal{A}}(M))$$

where $w(\cdot)$ is the Gaussian width and $\gamma_{\mathcal{A}}(M)$ is defined in (3.25).

From Theorem 3.2.3, the minimax lower bound for estimation over the local tangent cone is determined by the Sudakov estimate $e^2(B_2^p \cap T_{\mathcal{A}}(M))$. It follows directly from Lemma 3.2.1 that there exists a universal constant $c > 0$ such that $c \cdot e(B_2^p \cap T_{\mathcal{A}}(M)) \leq w(B_2^p \cap T_{\mathcal{A}}(M)) \leq 24 \int_0^\infty \sqrt{\log \mathcal{N}(B_2^p \cap T_{\mathcal{A}}(M), \epsilon)} d\epsilon$. Thus under the Gaussian setting, both in terms of the upper bound and lower bound, geometric complexity measures govern the difficulty of the estimation problem, through closely related quantities: Gaussian width and Sudakov estimate.

Application of the Geometric Approach

In this section we apply the general theory under the Gaussian setting to some of the actively studied high-dimensional problems mentioned in Section 3.2.1 to illustrate the wide applicability of the theory. The detailed proofs are deferred to Supplement Cai et al. (2014).

High-Dimensional Linear Regression We begin by considering the high-dimensional linear regression model (3.12) under the assumption that the true parameter $M \in \mathbb{R}^p$ is sparse, say $\|M\|_{l_0} = s$. Our general theory applying to the ℓ_1 minimization recovers the optimality results as in Dantzig selector and Lasso. In this case, it can be shown that $\gamma_{\mathcal{A}}(M)w(\mathcal{A})$ and $w(B_2^p \cap T_{\mathcal{A}}(M))$ are of the same rate $\sqrt{s \log p}$. See Supplement Cai et al. (2014) for the detailed calculations. The asphericity ratio $\gamma_{\mathcal{A}}(M) \leq 2\sqrt{s}$ reflects the sparsity of M through the local tangent cone and the Gaussian width $w(\mathcal{X}\mathcal{A}) \asymp \sqrt{\log p}$. The following corollary follows from the geometric analysis of the high-dimensional regression model.

Corollary 3.2.2. *Consider the linear regression model* (3.12). *Assume that* $\mathcal{X} \in \mathbb{R}^{n \times p}$ *is the Gaussian ensemble design and the parameter of interest* $M \in \mathbb{R}^p$ *is of sparsity* s. *Let* $\hat{M}$ *be the solution to the constrained* ℓ_1 *minimization* (3.26) *with* $\lambda = C_1 \sigma \sqrt{\frac{\log p}{n}}$. *If* $n \geq C_2 s \log p$, *then*

$$\|\hat{M} - M\|_{\ell_2} \precsim \sigma \sqrt{\frac{s \log p}{n}}, \ \|\hat{M} - M\|_{\ell_1} \precsim \sigma s \sqrt{\frac{\log p}{n}}, \ \|\mathcal{X}(\hat{M} - M)\|_{\ell_2} \precsim \sigma \sqrt{\frac{s \log p}{n}}.$$

with high probability, where $C_1, C_2 > 0$ *are some universal constants.*

For ℓ_2 norm consistency of the estimation for M, we require $\lim\limits_{n,p(n)\to\infty} \frac{s \log p}{n} = 0$. However, for valid inferential guarantee, the de-biased Dantzig selector type estimator $\tilde{M}$ satisfies asymptotic normality under the condition $\lim\limits_{n,p(n)\to\infty} \frac{s \log p}{\sqrt{n}} = 0$ through Theorem 3.2.2. Under this condition, the confidence interval given in (3.35) has asymptotic coverage probability of $(1-\alpha)$ and its expected length is at the parametric rate $\frac{1}{\sqrt{n}}$. Furthermore, the confidence intervals do not depend on the specific value of s. Results in Section 3.2.3 and 3.2.3 recover the best known result on confidence intervals as in Zhang and Zhang (2014); van de Geer et al. (2014); Javanmard and Montanari (2014). Our result is a generic procedure that compensates for the bias introduced by the point estimation convex program. All these procedures are driven by local geometry.

Low Rank Matrix Recovery We now consider the recovery of low-rank matrices under the trace regression model (3.13). The geometric theory leads to the optimal recovery results for nuclear norm minimization and penalized trace regression in the existing literature.

Assume the true parameter $M \in \mathbb{R}^{p \times q}$ has rank r. Let us examine the behavior of $\phi_{\mathcal{A}}(M, \mathcal{X})$, $\gamma_{\mathcal{A}}(M)$, and $\lambda_{\mathcal{A}}(\mathcal{X}, \sigma, n)$. Detailed calculations given in Supplement Cai et al. (2014) show that in this case $\gamma_{\mathcal{A}}(M) w(\mathcal{A})$ and $w(B_2^p \cap T_{\mathcal{A}}(M))$ are of the same order $\sqrt{r(p+q)}$. The asphericity ratio $\gamma_{\mathcal{A}}(M) \leq 2\sqrt{2r}$ characterizes the low rank structure and the Gaussian width $w(\mathcal{X}\mathcal{A}) \asymp \sqrt{p+q}$. We have the following corollary for low rank matrix recovery.

Corollary 3.2.3. *Consider the trace regression model* (3.13). *Assume that* $\mathcal{X} \in \mathbb{R}^{n \times pq}$ *is the Gaussian ensemble design and the true parameter* $M \in \mathbb{R}^{p \times q}$ *is of rank* r. *Let* $\hat{M}$ *be the solution to the constrained nuclear norm minimization* (3.26) *with* $\lambda = C_1 \sigma \sqrt{\frac{p+q}{n}}$. *If* $n \geq C_2 r(p+q)$, *then for some universal constants* $C_1, C_2 > 0$, *with high probability,*

$$\|\hat{M} - M\|_F \lesssim \sigma \sqrt{\frac{r(p+q)}{n}}, \ \|\hat{M} - M\|_* \lesssim \sigma r \sqrt{\frac{p+q}{n}}, \ \|\mathcal{X}(\hat{M} - M)\|_{\ell_2} \lesssim \sigma \sqrt{\frac{r(p+q)}{n}}.$$

For consistency under the Frobenius norm, the condition is $\lim_{n,p(n),q(n)\to\infty} \frac{\sqrt{r(p+q)}}{\sqrt{n}} = 0$. For statistical inference, Theorem 3.2.2 requires $\lim_{n,p(n),q(n)\to\infty} \frac{r(p+q)}{\sqrt{n}} = 0$, which is essentially $n \gtrsim pq$ (sample size is larger than the dimension) for $r = 1$. This phenomenon happens when the Gaussian width complexity of the rank-1 matrices is large, i.e., the atom set is too rich. We remark that in practice, convex program (3.29) can still be used for constructing confidence intervals and performing hypothesis testing. However, it is harder to provide sharp upper bound theoretically for the approximation error η in (3.29), for any given r, p, q.

Sign Vector Recovery We turn to the sign vector recovery model (3.14) where the parameter of interest $M \in \{+1, -1\}^p$ is a sign vector. The convex hull of the atom set is then the ℓ_∞ norm ball. Applying the general theory to the constrained ℓ_∞ norm minimization (3.27) leads to the optimal rates of convergence for the sign vector recovery. The calculations given in Supplement Cai et al. (2014) show that the asphericity ratio $\gamma_{\mathcal{A}}(M) \leq 1$ and the Gaussian width $w(\mathcal{X}B_2^p) \asymp \sqrt{p}$. Geometric theory when applied to sign vector recovery shows the following Corollary.

Corollary 3.2.4. *Consider the model* (3.14) *where the true parameter* $M \in \{+1, -1\}^p$ *is a sign vector. Assume that* $\mathcal{X} \in \mathbb{R}^{n \times p}$ *is the Gaussian ensemble design. Let* $\hat{M}$ *be the solution to the convex program* (3.27) *with* $\lambda = C_1 \sigma \frac{p}{\sqrt{n}}$ *and* $\mu = C_1 \sigma \sqrt{\frac{p}{n}}$. *If* $n \geq C_2 p$, *then for some* *universal constant $C > 0$, with high probability,*

$$\|\hat{M} - M\|_{\ell_2}, \|\hat{M} - M\|_{\ell_\infty}, \|\mathcal{X}(\hat{M} - M)\|_{\ell_2} \le C \cdot \sigma\sqrt{\frac{p}{n}}.$$

Orthogonal Matrix Recovery We now treat orthogonal matrix recovery using the spectral norm minimization. Please see Example 4 in Section 3.2.2 for details. Consider the same model as in trace regression, but the parameter of interest $M \in \mathbb{R}^{m\times m}$ is an orthogonal matrix. One can show that $w(B_2^p \cap T_{\mathcal{A}}(M))$ is of order $\sqrt{m^2}$ and $\gamma_{\mathcal{A}}(M) \le 1$. Applying the geometric analysis to the constrained spectral norm minimization (3.27) yields the following.

Corollary 3.2.5. *Consider the orthogonal matrix recovery model* (3.13)*. Assume that $\mathcal{X} \in \mathbb{R}^{n\times m^2}$ is the Gaussian ensemble matrix and the true parameter $M \in \mathbb{R}^{m\times m}$ is an orthogonal matrix. Let $\hat{M}$ be the solution to the program* (3.27) *with $\lambda = C_1\sigma\sqrt{\frac{m^3}{n}}$ and $\mu = C_1\sigma\sqrt{\frac{m^2}{n}}$. If $n \ge C_2 m^2$, then, with high probability,*

$$\|\hat{M} - M\|_F, \|\hat{M} - M\|, \|\mathcal{X}(\hat{M} - M)\|_{\ell_2} \le C \cdot \sigma\sqrt{\frac{m^2}{n}},$$

where $C > 0$ is some universal constant.

Other examples Other examples that can be formalized under the framework of the linear inverse model include permutation matrix recovery (Jagabathula and Shah, 2011), sparse plus low rank matrix recovery (Candès et al., 2011) and matrix completion (Candès and Recht, 2009). The convex relaxation of permutation matrix is double stochastic matrix; the atomic norm corresponding to sparse plus low rank atom set is the infimal convolution of the ℓ_1 norm and nuclear norm; for matrix completion, the design matrix can be viewed as a diagonal matrix with diagonal elements being independent Bernoulli random variables. See Section 3.2.5 for a discussion on further examples.

3.2.4. Local Geometric Theory: General Setting

We have developed in the last section a local geometric theory for the linear inverse model in the Gaussian setting. The Gaussian assumption on the design and noise enables us to carry out concrete and more specific calculations as seen in the examples given in Section 3.2.3, but the distributional assumption is not essential. In this section we extend this theory to the general setting.

General Local Upper Bound

We shall consider a fixed design matrix $\mathcal{X}$ (in the case of random design, results we will establish are conditional on the design) and condition on the event that the noise is controlled $\|\mathcal{X}^*(Z)\|_{\mathcal{A}}^* \leq \lambda_n$. We have seen in Lemma 3.2.2 of Section 3.2.3 how to choose λ_n to make this happen with overwhelming probability under Gaussian noise.

Theorem 3.2.4 (Geometrizing Local Convergence)**.** *Suppose we observe* $(\mathcal{X},\ Y)$ *as in* (3.11)*. Condition on the event that the noise vector Z satisfies, for some given choice of localization radius λ_n, $\|\mathcal{X}^*(Z)\|_{\mathcal{A}}^* \leq \lambda_n$. Let $\hat{M}$ be the solution to the convex program* (3.26) *with λ_n being the tuning parameter. Then the geometric quantities defined on the local tangent cone capture the local convergence rate for $\hat{M}$:*

$$\|\hat{M} - M\|_{\mathcal{A}} \leq \gamma_{\mathcal{A}}(M)\|\hat{M} - M\|_{\ell_2}, \quad \textit{and further}$$
$$\|\hat{M} - M\|_{\ell_2} \leq \frac{1}{\phi_{\mathcal{A}}(M,\mathcal{X})}\|\mathcal{X}(\hat{M} - M)\|_{\ell_2} \leq \frac{2\gamma_{\mathcal{A}}(M)\lambda_n}{\phi_{\mathcal{A}}^2(M,\mathcal{X})}$$

with the local asphericity ratio $\gamma_{\mathcal{A}}(M)$ defined in (3.25) *and the local lower isometry constant $\phi_{\mathcal{A}}(M,\mathcal{X})$ defined in* (3.23)*.*

Theorem 3.2.4 does not require distributional assumptions on the noise, nor does it impose conditions on the design matrix. Theorem 3.2.1 can be viewed as a special case where the local isometry constant $\phi_{\mathcal{A}}(M,\mathcal{X})$ and the local radius λ_n are calculated explicitly under the Gaussian assumption. Theorem 3.2.4 is proved in Appendix in a general form, which analyzes convex programs (3.26) and (3.27) simultaneously.

General Geometric Inference

Geometric inference can also be extended to other fixed designs when Z is Gaussian. We can modify the convex feasibility program (3.28) into the following stronger form

$$(\Omega, \eta_n) = \arg\min_{\Omega,\eta} \left\{\eta : \quad \|\mathcal{X}^* \mathcal{X} \Omega_{i\cdot}^* - e_i\|_{\mathcal{A}}^* \leq \eta, \quad \forall 1 \leq i \leq p\right\}. \tag{3.37}$$

Then the following theorem holds (proof is analogous to Theorem 3.2.2).

Theorem 3.2.5 (Geometric Inference)**.** *Suppose we observe* $(\mathcal{X},\ Y)$ *as in* (3.11) *with* $Z \sim N(0, \frac{\sigma^2}{n} I_n)$*. Let* $\hat{M}$ *be the solution to the convex program* (3.26)*. Denote* Ω *and* η_n *as the optimal solution to the convex program* (3.37)*, and* $\tilde{M}$ *as the de-biased estimator. The following decomposition*

$$\tilde{M} - M = \Delta + \frac{\sigma}{\sqrt{n}} \Omega \mathcal{X}^* W \tag{3.38}$$

holds, where $W \sim N(0, I_n)$ *is the standard Gaussian vector and*

$$\Omega \mathcal{X}^* W \sim N(0, \Omega \mathcal{X}^* \mathcal{X} \Omega^*).$$

Here the bias part $\Delta \in \mathbb{R}^p$ *satisfies, with high probability,*

$$\|\Delta\|_{\ell_\infty} \leq \frac{2 \cdot \gamma_{\mathcal{A}}^2(M)}{\phi_{\mathcal{A}}(M, \mathcal{X})} \cdot \lambda_n \eta_n,$$

provided we choose λ_n *as in Lemma 3.2.2.*

General Local Minimax Lower Bound

The lower bound given in the Gaussian case can also be extended to the general setting where the class of noise distributions contains the Gaussian distributions. We aim to geometrize the intrinsic difficulty of the estimation problem in a unified manner. We first present a general result for a convex cone T in the parameter space, which illustrates how the Sudakov estimate, volume ratio and the design matrix affect the minimax lower bound.

Theorem 3.2.6. *Let $T \in \mathbb{R}^p$ be a compact convex cone. The minimax lower bound for the linear inverse model* (3.11)*, if restricted to the cone T, is*

$$\inf_{\hat{M}} \sup_{M \in T} \mathbb{E}_{\cdot|\mathcal{X}} \|\hat{M} - M\|_{\ell_2}^2 \geq \frac{c_0 \sigma^2}{\psi^2} \cdot \left(\frac{e(B_2^p \cap T)}{\sqrt{n}} \vee \frac{v(B_2^p \cap T)}{\sqrt{n}} \right)^2$$

where $\hat{M}$ is any measurable estimator, $\psi = \sup_{v \in B_2^p \cap T} \|\mathcal{X}(v)\|_{\ell_2}$ and c_0 is a universal constant. Here $\mathbb{E}_{\cdot|\mathcal{X}}$ is conditioned on the design matrix. $e(\cdot)$ and $v(\cdot)$ denote the Sudakov estimate (3.20) *and volume ratio* (3.22)*. Then*

$$\inf_{\hat{M}} \sup_{M' \in T_{\mathcal{A}}(M)} \mathbb{E}_{\cdot|\mathcal{X}} \|\hat{M} - M'\|_{\ell_2}^2 \geq \frac{c_0 \sigma^2}{\psi_{\mathcal{A}}^2(M, \mathcal{X})} \cdot \left(\frac{e(B_2^p \cap T_{\mathcal{A}}(M))}{\sqrt{n}} \vee \frac{v(B_2^p \cap T_{\mathcal{A}}(M))}{\sqrt{n}} \right)^2.$$

Theorem 3.2.6 gives minimax lower bounds in terms of the Sudakov estimate and volume ratio. In the Gaussian setting, Lemma 3.2.3 shows that the local upper isometry constant satisfies $\psi_{\mathcal{A}}(M, \mathcal{X}) \leq 1 + c$ with probability at least $1 - 2\exp(-\delta^2/2)$, as long as

$$n \geq \frac{4[w(B_2^p \cap T_{\mathcal{A}}(M)) + \delta]^2}{c^2} \vee \frac{1}{c}.$$

We remark that $\psi_{\mathcal{A}}(M, \mathcal{X})$ can be bounded under more general design matrix $\mathcal{X}$. However, under the Gaussian design (even correlated design), the minimum sample size n to ensure that $\psi_{\mathcal{A}}(M, \mathcal{X})$ is upper bounded, is directly determined by Gaussian width of the tangent cone.

The geometric complexity of the lower bound provided by Theorem 3.2.6 matches $w(B_2^p \cap T_{\mathcal{A}}(M))$ if Sudakov minoration of Lemma 3.2.1 can be reversed on the tangent cone, in the sense that $w(B_2^p \cap T_{\mathcal{A}}(M)) \leq C \cdot e(B_2^p \cap T_{\mathcal{A}}(M))$. Further, recalling Urysohn's inequality we have $v(B_2^p \cap T_{\mathcal{A}}(M)) \leq w(B_2^p \cap T_{\mathcal{A}}(M))$. Hence, if the reverse Urysohn's inequality

$w(B_2^p \cap T_{\mathcal{A}}(M)) \leq C \cdot v(B_2^p \cap T_{\mathcal{A}}(M))$ holds for the local tangent cone, the obtained rate is, again, of the order $w(B_2^p \cap T_{\mathcal{A}}(M))$.

3.2.5. Discussion

This section presents a unified geometric characterization of the local estimation rates of convergence as well as statistical inference for high-dimensional linear inverse problems. Exploring other interesting applications that can be subsumed under the general framework is an interesting future research direction.

For statistical inference, both independent Gaussian design and correlated Gaussian design with known covariance Σ are considered. The case of unknown Σ is an interesting problem for future work.

The lower bound constructed in the current thesis can be contrasted with the lower bounds in Ye and Zhang (2010); Candes and Davenport (2013). Both the above two papers consider specifically the minimax lower bound for high-dimensional linear regression. We focus on a more generic perspective – lower bounds in Theorem 3.2.6 hold in general for arbitrary star-shaped body T, which includes $\ell_p, 0 \leq p \leq \infty$, balls and cones as special cases.

3.3. Adaptive Feature Selection: Efficient Online Sparse Regression

3.3.1. Introduction

In modern real-world sequential prediction problems, samples are typically high dimensional, and construction of the features may itself be a computationally intensive task. Therefore in sequential prediction, due to the computation and resource constraints, it is preferable to design algorithms that compute only a limited number of features for each new data example. One example of this situation, from (Cesa-Bianchi et al., 2011), is medical diagnosis of a disease, in which each feature is the result of a medical test on the patient. Since it is undesirable to subject a patient to a battery of medical tests, we would like to adaptively design diagnostic procedures that rely on only a few, highly informative tests.

Online sparse linear regression (OSLR) is a sequential prediction problem in which an algorithm is allowed to see only a small subset of coordinates of each feature vector. The problem is parameterized by 3 positive integers: d, the dimension of the feature vectors, k, the sparsity of the linear regressors we compare the algorithm's performance to, and k_0, a budget on the number of features that can be queried in each round by the algorithm. Generally we have $k \ll d$ and $k_0 \geq k$ but not significantly larger (our algorithms need[1] $k_0 = \tilde{O}(k)$).

In the OSLR problem, the algorithm makes predictions over a sequence of T rounds. In each round t, nature chooses a feature vector $x_t \in \mathbb{R}^d$, the algorithm chooses a subset of $\{1, 2, \ldots, d\}$ of size at most k' and observes the corresponding coordinates of the feature vector. It then makes a prediction $\widehat{y}_t \in \mathbb{R}$ based on the observed features, observes the true label y_t, and suffers loss $(y_t - \widehat{y}_t)^2$. The goal of the learner is to make the cumulative loss comparable to that of the best k-sparse linear predictor w in hindsight. The performance of the online learner is measured by the *regret*, which is defined as the difference between

[1] In this thesis, we use the $\tilde{O}(\cdot)$ notation to suppress factors that are polylogarithmic in the natural parameters of the problem.

the two losses:

$$\text{Regret}_T = \sum_{t=1}^{T} (y_t - \widehat{y}_t)^2 - \min_{w\colon \|w\|_0 \le k} \sum_{t=1}^{T} (y_t - \langle x_t, w \rangle)^2 \ .$$

The goal is to construct algorithms that enjoy regret that is sub-linear in T, the total number of rounds. A sub-linear regret implies that in the asymptotic sense, the average per-round loss of the algorithm approaches the average per-round loss of the best k-sparse linear predictor.

Sparse regression is in general a computationally hard problem. In particular, given k, $x_1, x_2, \ldots, x_T$ and $y_1, y_2, \ldots, y_T$ as inputs, the offline problem of finding a k-sparse w that minimizes the error $\sum_{t=1}^{T}(y_t - \langle x_t, w \rangle)^2$ does not admit a polynomial time algorithm under standard complexity assumptions Foster et al. (2015). This hardness persists even under the assumption that there exists a k-sparse w^* such that $y_t = \langle x_t, w^* \rangle$ for all t. Furthermore, the computational hardness is present even when the solution is required to be only $\widetilde{\mathcal{O}}(k)$-sparse solution and has to minimize the error only approximately; see Foster et al. (2015) for details. The hardness result was extended to online sparse regression by Foster et al. (2016). They showed that for all $\delta > 0$ there exists no polynomial-time algorithm with regret $\mathcal{O}(T^{1-\delta})$ unless $NP \subseteq BPP$.

Foster et al. (2016) posed the open question of what additional assumptions can be made on the data to make the problem tractable. In this thesis, we answer this open question by providing efficient algorithms with sublinear regret under the assumption that the matrix of feature vectors satisfies the *restricted isometry property* (RIP) (Candes and Tao, 2005). It has been shown that if RIP holds and there exists a sparse linear predictor w^* such that $y_t = \langle x_t, w^* \rangle + \eta_t$ where η_t is independent noise, the offline sparse linear regression problem admits computationally efficient algorithms, e.g., Candès and Tao (2007). RIP and related Restricted Eigenvalue Condition (Bickel et al., 2009) have been widely used as a standard assumption for theoretical analysis in the compressive sensing and sparse regression literature, in the offline case. In the online setting, it is natural to ask whether sparse regression avoids the computational difficulty under a proper form of RIP condition. In this thesis, we answer this question in a positive way, both in the realizable setting and in the agnostic setting. As a by-product, we resolve the adaptive feature selection problem as the efficient algorithms we propose in this thesis adaptively choose a different "sparse" subset of features to query at each round. This is closely related to attribute-efficient learning (see discussion in Section 3.3.1) and online model selection.

Summary of Results

We design polynomial-time algorithms for online sparse linear regression for two models for the sequence $(x_1, y_1), (x_2, y_2), \ldots, (x_T, y_T)$. The first model is called the *realizable* and the second is called *agnostic*. In both models, we assume that, after proper normalization, for all large enough t, the matrix X_t formed from the first t feature vectors $x_1, x_2, \ldots, x_t$ satisfies the restricted isometry property. The two models differ in the assumptions on y_t. The realizable model assumes that $y_t = \langle x_t, w^* \rangle + \eta_t$ where w^* is k-sparse and η_t is an independent noise. In the agnostic model, y_t can be arbitrary. The models and corresponding algorithms are presented in Sections 3.3.2 and 3.3.3 respectively. Interestingly enough, the algorithms and their corresponding analyses are completely different in the realizable and agnostic case.

Our algorithms allow for somewhat more flexibility than the problem definition: they are designed to work with a budget k_0 on the number of features that can be queried that may be larger than the sparsity parameter k of the comparator. The regret bounds we derive improve with increasing values of k_0. In the case when $k_0 \approx k$, the dependence on d in the regret bounds is polynomial, as can be expected in limited feedback settings (this is analogous to polynomial dependence on d in *bandit* settings). In the extreme case when $k_0 = d$, i.e. we have access to *all* the features, the dependence on the dimension d in the regret bounds we prove is only *logarithmic*. The interpretation is that if we have full access to the features, but the goal is to compete with just k sparse linear regressors, then the number of data points that need to be seen to achieve good predictive accuracy has only logarithmic dependence on d. This is analogous to the (offline) compressed sensing setting where the sample complexity bounds, under RIP, only depend logarithmically on d.

A major building block in the solution for the realizable setting (Section 3.3.2) consists of identifying the best k-sparse linear predictor for the past data at any round in the prediction problem. This is done by solving a sparse regression problem on the observed data. The solution of this problem cannot be obtained by a simple application of say, the Dantzig selector (Candès and Tao, 2007) since we do not observe the data matrix X, but rather a subsample of its entries. Our algorithm is a variant of the Dantzig selector that incorporates random sampling into the optimization, and computes a near-optimal solution by solving a linear program. The resulting algorithm has a regret bound of $\widetilde{\mathcal{O}}(\log T)$.

The algorithm for the agnostic setting relies on the theory of submodular optimization. The analysis in (Boutsidis et al., 2015) shows that the RIP assumption implies that the set function defined as the minimum loss achievable by a linear regressor restricted to the set in question satisfies a property called *weak supermodularity.* Weak supermodularity is a relaxation of standard supermodularity that is still strong enough to show performance bounds for the standard greedy feature selection algorithm for solving the sparse regression problem. We then employ a technique developed by Streeter and Golovin (2008) to construct an online learning algorithm that mimics the greedy feature selection algorithm. The resulting algorithm has a regret bound of $\widetilde{\mathcal{O}}(T^{2/3})$.

Related work

A related setting is attribute-efficient learning (Cesa-Bianchi et al., 2011; Hazan and Koren, 2012; Kukliansky and Shamir, 2015). This is a batch learning problem in which the examples are generated i.i.d., and the goal is to simply output a linear regressor using only a limited number of features per example with bounded excess risk compared to the optimal linear regressor, when given *full access* to the features at test time. Since the goal is not prediction but simply computing the optimal linear regressor, efficient algorithms exist and have been developed by the aforementioned papers.

Without any assumptions, only inefficient algorithms for the online sparse linear regression problem are known Zolghadr et al. (2013); Foster et al. (2016). Kale (2014) posed the open question of whether it is possible to design an efficient algorithm for the problem with a sublinear regret bound. This question was answered in the negative by Foster et al. (2016), who showed that efficiency can only be obtained under additional assumptions on the data. This thesis shows that the RIP assumption yields tractability in the online setting just as it does in the batch setting.

In the realizable setting, the linear program at the heart of the algorithm is motivated from Dantzig selection Candès and Tao (2007) and error-in-variable regression Rosenbaum and Tsybakov (2010); Belloni et al. (2016). The problem of finding the best sparse linear predictor when only a sample of the entries in the data matrix is available is also discussed by Belloni et al. (2016) (see also the references therein). In fact, these papers solve a more general problem where we observe a matrix Z rather than X that is an unbiased estimator of X. While we can use their results in a black-box manner, they are tailored for the setting where the variance of each Z_{ij} is constant and it is difficult to obtain the exact dependence on this variance in their bounds. In our setting, this variance can be linear in the dimension of the feature vectors, and hence we wish to control the dependence on the variance in the bounds. Thus, we use an algorithm that is similar to the one in Belloni et al. (2016), and provide an analysis for it (in the supplementary material). As an added bonus, our algorithm results in solving a linear program rather than a conic or general convex program, hence admits a solution that is more computationally efficient.

In the agnostic setting, the computationally efficient algorithm we propose is motivated from (online) supermodular optimization (Natarajan, 1995; Boutsidis et al., 2015; Streeter and Golovin, 2008). The algorithm is computationally efficient and enjoys sublinear regret under an RIP-like condition, as we will show in Section 3.3.3. This result can be contrasted with the known computationally prohibitive algorithms for online sparse linear regression

(Zolghadr et al., 2013; Foster et al., 2016), and the hardness result without RIP (Foster et al., 2015, 2016).

Notation and Preliminaries

For $d \in \mathbb{N}$, we denote by $[d]$ the set $\{1, 2, \ldots, d\}$. For a vector in $x \in \mathbb{R}^d$, denote by $x(i)$ its i-th coordinate. For a subset $S \subseteq [d]$, we use the notation $\mathbb{R}^S$ to indicate the vector space spanned by the coordinate axes indexed by S (i.e. the set of all vectors w supported on the set S). For a vector $x \in \mathbb{R}^d$, denote by $x(S) \in \mathbb{R}^d$ the projection of x on $\mathbb{R}^S$. That is, the coordinates of $x(S)$ are

$$x(S)(i) = \begin{cases} x(i) & \text{if } i \in S, \\ 0 & \text{if } i \notin S, \end{cases} \qquad \text{for } i = 1, 2, \ldots, d.$$

Let $\langle u, v \rangle = \sum_i u(i) \cdot v(i)$ be the inner product of vectors u and v.

For $p \in [0, \infty]$, the ℓ_p-norm of a vector $x \in \mathbb{R}^d$ is denoted by $\|x\|_p$. For $p \in (0, \infty)$, $\|x\|_p = (\sum_i |x_i|^p)^{1/p}$, $\|x\|_\infty = \max_i |x_i|$, and $\|x\|_0$ is the number of non-zero coordinates of x.

The following definition will play a key role:

Definition 3.3.1 (Restricted Isometry Property Candès and Tao (2007))**.** *Let $\epsilon \in (0, 1)$ and $k \geq 0$. We say that a matrix $X \in \mathbb{R}^{n \times d}$ satisfies* restricted isometry property *(RIP) with parameters (ϵ, k) if for any $w \in \mathbb{R}^d$ with $\|w\|_0 \leq k$ we have*

$$(1 - \epsilon) \|w\|_2 \leq \frac{1}{\sqrt{n}} \|Xw\|_2 \leq (1 + \epsilon) \|w\|_2 .$$

One can show that RIP holds with overwhelming probability if $n = \Omega(\epsilon^{-2} k \log(ed/k))$ and each row of the matrix is sampled independently from an isotropic sub-Gaussian distribution. In the realizable setting, the sub-Gaussian assumption can be relaxed to incorporate heavy tail distribution via the "small ball" analysis introduced in Mendelson (2014), since we only require one-sided lower isometry property.

Proper Online Sparse Linear Regression

We introduce a variant of online sparse regression (OSLR), which we call *proper online sparse linear regression (POSLR)*. The adjective "proper" is to indicate that the algorithm is required to output a weight vector in each round and its prediction is computed by taking an inner product with the feature vector.

We assume that there is an underlying sequence $(x_1, y_1), (x_2, y_2), \ldots, (x_T, y_T)$ of *labeled examples* in $\mathbb{R}^d \times \mathbb{R}$. In each round $t = 1, 2, \ldots, T$, the algorithm behaves according to the following protocol:

1. Choose a vector $w_t \in \mathbb{R}^d$ such that $\|w_t\|_0 \leq k$.
2. Choose $S_t \subseteq [d]$ of size at most k_0.
3. Observe $x_t(S_t)$ and y_t, and incur loss $(y_t - \langle x_t, w_t \rangle)^2$.

Essentially, the algorithm makes the prediction $\widehat{y}_t := \langle x_t, w_t \rangle$ in round t. The regret after T rounds of an algorithm with respect to $w \in \mathbb{R}^d$ is

$$\text{Regret}_T(w) = \sum_{t=1}^{T} (y_t - \langle x_t, w_t \rangle)^2 - \sum_{t=1}^{T} (y_t - \langle x_t, w \rangle)^2 \ .$$

The regret after T rounds of an algorithm with respect to the best k-sparse linear regressor is defined as

$$\text{Regret}_T = \max_{w\colon \|w\|_0 \leq k} \text{Regret}_T(w) \, .$$

Note that any algorithm for POSLR gives rise to an algorithm for OSLR. Namely, if an algorithm for POSLR chooses w_t and S_t, the corresponding algorithm for OSLR queries the coordinates $S_t \cup \{i \ : w_t(i) \neq 0\}$. The algorithm for OSLR queries at most $k_0 + k$ coordinates and has the same regret as the algorithm for POSLR.

Additionally, POSLR allows parameters settings which do not have corresponding counterparts in OSLR. Namely, we can consider the sparse "full information" setting where $k_0 = d$ and $k \ll d$.

We denote by X_t the $t \times d$ matrix of first t unlabeled samples i.e. rows of X_t are $x_1^T, x_2^T, \ldots, x_t^T$. Similarly, we denote by $Y_t \in \mathbb{R}^t$ the vector of first t labels $y_1, y_2, \ldots, y_t$. We use the shorthand notation X, Y for X_T and Y_T respectively.

In order to get computationally efficient algorithms, we assume that that for all $t \geq t_0$, the matrix X_t satisfies the restricted isometry condition. The parameter t_0 and RIP parameters k, ϵ will be specified later.

3.3.2. Realizable Model

In this section we design an algorithm for POSLR for the realizable model. In this setting we assume that there is a vector $w^* \in \mathbb{R}^d$ such that $\|w^*\|_0 \leq k$ and the sequence of labels $y_1, y_2, \ldots, y_T$ is generated according to the linear model

$$y_t = \langle x_t, w^* \rangle + \eta_t \ , \tag{3.39}$$

where $\eta_1, \eta_2, \ldots, \eta_T$ are independent random variables from $N(0, \sigma^2)$. We assume that the standard deviation σ, or an upper bound of it, is given to the algorithm as input. We assume that $\|w^*\|_1 \leq 1$ and $\|x_t\|_\infty \leq 1$ for all t.

For convenience, we use η to denote the vector $(\eta_1, \eta_2, \ldots, \eta_T)$ of noise variables.

Algorithm

The algorithm maintains an unbiased estimate $\widehat{X}_t$ of the matrix X_t. The rows of $\widehat{X}_t$ are vectors $\widehat{x}_1^T, \widehat{x}_2^T, \ldots, \widehat{x}_t^T$ which are unbiased estimates of $x_1^T, x_2^T, \ldots, x_t^T$. To construct the estimates, in each round t, the set $S_t \subseteq [d]$ is chosen uniformly at random from the collection of all subsets of $[d]$ of size k_0. The estimate is

$$\widehat{x}_t = \frac{d}{k_0} \cdot x_t(S_t). \tag{3.40}$$

To compute the predictions of the algorithm, we consider the linear program

$$\begin{aligned} \text{minimize } \|w\|_1 \text{ s.t. } & \left\| \frac{1}{t} \widehat{X}_t^T \left(Y_t - \widehat{X}_t w \right) + \frac{1}{t} \widehat{D}_t w \right\|_\infty \\ & \leq C \sqrt{\frac{d \log(td/\delta)}{tk_0}} \left(\sigma + \frac{d}{k_0} \right). \end{aligned} \tag{3.41}$$

Here, $C > 0$ is a universal constant, and $\delta \in (0,1)$ is the allowed failure probability. $\widehat{D}_t$, defined in equation (3.43), is a diagonal matrix that offsets the bias on the diag$(\widehat{X}_t^T \widehat{X}_t)$.

The linear program (3.41) is called the Dantzig selector. We denote its optimal solution by $\widehat{w}_{t+1}$. (We define $\widehat{w}_1 = 0$.)

Based on $\widehat{w}_t$, we construct $\widetilde{w}_t \in \mathbb{R}^d$. Let $|\widehat{w}_t(i_1)| \geq |\widehat{w}_t(i_2)| \geq \cdots \geq |\widehat{w}_t(i_d)|$ be the coordinates sorted according to the their absolute value, breaking ties according to their index. Let $\widetilde{S}_t = \{i_1, i_2, \ldots, i_k\}$ be the top k coordinates. We define $\widetilde{w}_t$ as

$$\widetilde{w}_t = \widehat{w}_t(\widetilde{S}_t). \tag{3.42}$$

The actual prediction w_t is either zero if $t \leq t_0$ or $\widetilde{w}_s$ for some $s \leq t$ and it gets updated whenever t is a power of 2.

The algorithm queries at most $k + k_0$ features each round, and the linear program can be solved in polynomial time using simplex method or interior point method. The algorithm solves the linear program only $\lceil \log_2 T \rceil$ times by using the same vector in the rounds $2^s, \ldots, 2^{s+1} - 1$. This lazy update improves both the computational aspects of the algorithm and the regret bound.

Algorithm 11 Dantzig Selector for POSLR

```
Require: $T$, $\sigma$, $t_0$, $k$, $k_0$
for $t = 1, 2, \ldots, T$ do
    if $t \le t_0$ then
        Predict $w_t = 0$
    else if $t$ is a power of 2 then
        Let $\widehat{w}_t$ be the solution of linear program (3.41)
        Compute $\widetilde{w}_t$ according to (3.42)
        Predict $w_t = \widetilde{w}_t$
    else
        Predict $w_t = w_{t-1}$
    end if
    Let $S_t \subseteq [d]$ be a random subset of size $k_0$
    Observe $x_t(S_t)$ and $y_t$
    Construct estimate $\widehat{x}_t$ according to (3.40)
    Append $\widehat{x}_t^T$ to $\widehat{X}_{t-1}$ to form $\widehat{X}_t \in \mathbb{R}^{t \times d}$
end for
```

Main Result

The main result in this section provides a logarithmic regret bound under the following assumptions [2]

- The feature vectors have the property that for any $t \ge t_0$, the matrix X_t satisfies the RIP condition with $(\frac{1}{5}, 3k)$, with $t_0 = \mathcal{O}(k \log(d) \log(T))$.
- The underlying POSLR online prediction problem has a sparsity budget of k and observation budget k_0.
- The model is realizable as defined in equation (3.39) with i.i.d unbiased Gaussian noise with standard deviation $\sigma = \mathcal{O}(1)$.

Theorem 3.3.1. *For any $\delta > 0$, with probability at least $1 - \delta$, Algorithm 11 satisfies*

$$\text{Regret}_T = \mathcal{O}\left(k^2 \log(d/\delta)(d/k_0)^3 \log(T)\right).$$

[2] A more precise statement with the exact dependence on the problem parameters can be found in the supplementary material.

The theorem asserts that an $\mathcal{O}(\log T)$ regret bound is efficiently achievable in the realizable setting. Furthermore when $k_0 = \Omega(d)$ the regret scales as $\log(d)$ meaning that we do not necessarily require $T \geq d$ to obtain a meaningful result. We note that the complete expression for arbitrary t_0, σ is given in the supplementary material.

The algorithm can be easily understood via the error-in-variable equation

$$\begin{aligned} y_t &= \langle x_t, w^* \rangle + \eta_t \ , \\ \widehat{x}_t &= x_t + \xi_t . \end{aligned}$$

with $\mathbf{E}[\xi_t] = \mathbf{E}[\widehat{x}_t - x_t] = 0$, where the expectation is taken over random sampling introduced by the algorithm when performing feature exploration. The learner observes y_t as well as the "noisy" feature vector $\widehat{x}_t$, and aims to recovery w^*.

As mentioned above, we (implicitly) need an unbiased estimator of $X_t^T X_t$. By taking $\widehat{X}_t^T \widehat{X}_t$ it is easy to verify that the off-diagonal entries are indeed unbiased however this is not the case for the diagonal. To this end we define $D_t \in \mathbb{R}^{d \times d}$ as the diagonal matrix compensating for the sampling bias on the diagonal elements of $\widehat{X}_t^T \widehat{X}_t$

$$D_t = \left(\frac{d}{k_0} - 1 \right) \cdot \mathrm{diag} \left(X_t^T X_t \right)$$

and the estimated bias from the observed data is

$$\widehat{D}_t = \left(1 - \frac{k_0}{d} \right) \cdot \mathrm{diag} \left(\widehat{X}_t^T \widehat{X}_t \right) . \tag{3.43}$$

Therefore, program (11) can be viewed as Dantzig selector with plug-in unbiased estimates for $X_t^T Y_t$ and $X_t^T X_t$ using limited observed features.

Sketch of Proof

The main building block in proving Theorem 3.3.1 is stated in Lemma 3.3.1. It proves that the sequence of solutions $\widehat{w}_t$ converges to the optimal response w^* based on which the signal y_t is created. More accurately, ignoring all second order terms, it shows that $\|\widehat{w}_t - w^*\|_1 \leq \mathcal{O}(1/\sqrt{t})$. In Lemma 3.3.2 we show that the same applies for the sparse approximation w_t of $\widehat{w}_t$. Now, since $\|x_t\|_\infty \leq 1$ we get that the difference between our response $\langle x_t, w_t \rangle$ and the (almost) optimal response $\langle x_t, w^* \rangle$ is bounded by $1/\sqrt{t}$. Given this, a careful calculation of the difference of losses leads to a regret bound w.r.t. w^*. Specifically, an elementary analysis of the loss expression leads to the equality

$$\text{Regret}_T(w^*) = \sum_{t=1}^{T} 2\eta_t \langle x_t, w^* - w_t \rangle + (\langle x_t, w^* - w_t \rangle)^2$$

A bound on both summands can clearly be expressed in terms of $| \langle x_t, w^* - w_t \rangle | = \mathcal{O}(1/\sqrt{t})$. The right summand requires a martingale concentration bound and the left is trivial. For both we obtain a bound of $\mathcal{O}(\log(T))$.

We are now left with two technicalities. The first is that w^* is not necessarily the empirically optimal response. To this end we provide, in the supplementary material, a constant (independent of T) bound on the regret of w^* compared to the empirical optimum. The second technicality is the fact that we do not solve for $\widehat{w}_t$ in every round, but in exponential gaps. This translates to an added factor of 2 to the bound $\|w_t - w^*\|_1$ that affects only the constants in the $\mathcal{O}(\cdot)$ terms.

Lemma 3.3.1 (Estimation Rates)**.** *Assume that the matrix $X_t \in \mathbb{R}^{t\times d}$ satisfies the RIP condition with $(\epsilon, 3k)$ for some $\epsilon < 1/5$. Let $\widehat{w}_{n+1} \in \mathbb{R}^d$ be the optimal solution of pro-*

gram (3.41). *With probability at least* $1-\delta$*,*

$$\|\widehat{w}_{t+1}-w^*\|_2 \le C \cdot \sqrt{\frac{d}{k_0} \cdot \frac{k \log(d/\delta)}{t}} \left(\sigma + \frac{d}{k_0}\right) \ ,$$
$$\|\widehat{w}_{t+1}-w^*\|_1 \le C \cdot \sqrt{\frac{d}{k_0} \frac{k^2 \log(d/\delta)}{t}} \left(\sigma + \frac{d}{k_0}\right) .$$

Here $C>0$ *is some universal constant and* σ *is the standard deviation of the noise.*

Note the $\widehat{w}_t$ may not be sparse; it can have many non-zero coordinates that are small in absolute value. However, we take the top k coordinates of $\widehat{w}_t$ in absolute value. Thanks to the Lemma 3.3.2 below, we lose only a constant factor $\sqrt{3}$.

Lemma 3.3.2. *Let* $\widehat{w} \in \mathbb{R}^d$ *be an arbitrary vector and let* $w^* \in \mathbb{R}^d$ *be a* k*-sparse vector. Let* $\widetilde{S} \subseteq [d]$ *be the top* k *coordinates of* $\widehat{w}$ *in absolute value. Then,*

$$\left\|\widehat{w}(\widetilde{S}) - w^*\right\|_2 \le \sqrt{3}\, \|\widehat{w} - w^*\|_2 .$$

3.3.3. Agnostic Setting

In this section we focus on the agnostic setting, where we don't impose any distributional assumption on the sequence. In this setting, there is no "true" sparse model, but the learner — with limited access to features — is competing with the best k-sparse model defined using full information $\{(x_t, y_t)\}_{t=1}^T$.

As before, we do assume that x_t and y_t are bounded. Without loss of generality, $\|x_t\|_\infty \le 1$, and $|y_t| \le 1$ for all t. Once again, without any regularity condition on the design matrix, Foster et al. (2016) have shown that achieving a sub-linear regret $\mathcal{O}(T^{1-\delta})$ is in general computationally hard, for any constant $\delta > 0$ unless NP $\subseteq$ BPP.

We give an efficient algorithm that achieves sub-linear regret under the assumption that the design matrix of any (sufficiently long) block of consecutive data points has bounded *restricted condition number*, which we define below:

Definition 3.3.2 (Restricted Condition Number)**.** *Let $k \in \mathbb{N}$ be a sparsity parameter. The* restricted condition number for sparsity k *of a matrix $X \in \mathbb{R}^{n \times d}$ is defined as*

$$\sup_{\substack{v,w:\ \|v\|=\|w\|=1,\\ \|v\|_0,\|w\|_0 \leq k}} \frac{\|Xv\|}{\|Xw\|}.$$

It is easy to see that if a matrix X satisfies RIP with parameters (ϵ, k), then its restricted condition number for sparsity k is at most $\frac{1+\epsilon}{1-\epsilon}$. Thus, having bounded restricted condition number is a weaker requirement than RIP.

We now define the *Block Bounded Restricted Condition Number Property* (BBRCNP):

Definition 3.3.3 (Block Bounded Restricted Condition Number Property)**.** *Let $\kappa > 0$ and $k \in \mathbb{N}$. A sequence of feature vectors $x_1, x_2, \ldots, x_T$ satisfies* BBRCNP *with parameters (κ, K) if there is a constant t_0 such that for any sequence of consecutive time steps $\mathcal{T}$ with $|\mathcal{T}| \geq t_0$, the restricted condition number for sparsity k of X, the design matrix of the feature vectors x_t for $t \in \mathcal{T}$, is at most κ.*

Note that in the random design setting where x_t, for $t \in [T]$, are isotropic sub-Gaussian vectors, $t_0 = O(\log T + k \log d)$ suffices to satisfy BBRCNP with high probability, where the $O(\cdot)$ notation hides a constant depending on κ.

We assume in this section that the sequence of feature vectors satisfies BBRCNP with parameters (κ, K) for some $K = \mathcal{O}(k \log(T))$ to be defined in the course of the analysis.

Algorithm

The algorithm in the agnostic setting is of distinct nature from that in the stochastic setting. Our algorithm is motivated from literature on maximization of sub-modular set function (Natarajan, 1995; Streeter and Golovin, 2008; Boutsidis et al., 2015). Though the problem being NP-hard, greedy algorithm on sub-modular maximization provides provable good approximation ratio. Specifically, Streeter and Golovin (2008) considered online optimization of super/sub-modular set functions using expert algorithm as sub-routine. Natarajan (1995); Boutsidis et al. (2015) casted the sparse linear regression as maximization of weakly supermodular function. We will introduce an algorithm that blends various ideas from referred literature, to attack the online sparse regression with limited features.

First, let's introduce the notion of a weakly supermodular function.

Definition 3.3.4. *For parameters $k \in \mathbb{N}$ and $\alpha \geq 1$, a set function $g : [d] \to \mathbb{R}$ is (k, α)-weakly supermodular if for any two sets $S \subseteq T \subseteq [d]$ with $|T| \leq k$, the following two inequalities hold:*

1. **(monotonicity)** *$g(T) \leq g(S)$, and*

2. **(approximately decreasing marginal gain)**

$$g(S) - g(T) \leq \alpha \sum_{i \in T \setminus S} [g(S) - g(S \cup \{i\})].$$

The definition is slightly stronger than that in Boutsidis et al. (2015). We will show that sparse linear regression can be viewed as weakly supermodular minimization in Definition 3.3.4 once the design matrix has bounded restricted condition number.

Now we outline the algorithm (see Algorithm 12). We divide the rounds $1, 2, \ldots, T$ into mini-batches of size B each (so there are T/B such batches). The b-th batch thus consists of the examples (x_t, y_t) for $t \in \mathcal{T}_b := \{(b-1)B + 1, (b-1)B + 1, \ldots, bB\}$. Within the b-th batch, our algorithm queries the same subset of features of size at most k_0.

The algorithm consists of few key steps. First, one can show that under BBRCNP, as long as B is large enough, the loss within batch b defines a weakly supermodular set function

$$g_t(S) = \frac{1}{B} \inf_{w \in \mathbb{R}^S} \sum_{t \in \mathcal{T}_b} (y_t - \langle x_t, w \rangle)^2.$$

Therefore, we can formulate the original online sparse regression problem into online weakly supermodular minimization problem. For the latter problem, we develop an online greedy algorithm along the lines of (Streeter and Golovin, 2008). We employ $k_1 = \mathcal{O}^*(k)$ budgeted experts algorithms (Amin et al., 2015), denoted BEXP, with budget parameter[3] $\frac{k_0}{k_1}$. The precise characteristics of BEXP are given in Theorem 3.3.2 (adapted from Theorem 2 in (Amin et al., 2015)).

Theorem 3.3.2. *For the problem of prediction from expert advice, let there be d experts, and let $k \in [d]$ be a budget parameter. In each prediction round t, the BEXP algorithm chooses an expert j_t and a set of experts U_t containing j_t of size at most k, obtains as feedback the losses of all the experts in U_t, suffers the loss of expert j_t, and guarantees an expected regret bound of $2\sqrt{\frac{d\log(d)}{k}T}$ over T prediction rounds.*

At the beginning of each mini-batch b, the BEXP algorithms are run. Each BEXP algorithm outputs a set of coordinates of size $\frac{k_0}{k_1}$ as well as a special coordinate in that set. The union of all of these sets is then used as the set of features to query throughout the subsequent mini-batch. Within the mini-batch, the algorithm runs the standard Vovk-Azoury-Warmuth algorithm for linear prediction with square loss *restricted* to set of special coordinates output by all the BEXP algorithms.

At the end of the mini-batch, every BEXP algorithm is provided carefully constructed losses for each coordinate that was output as feedback. These losses ensure that the set of special coordinates chosen by the BEXP algorithms mimic the greedy algorithm for weakly supermodular minimization.

Main Result

In this section, we will show that Algorithm 12 achieves sublinear regret under BBRCNP.

Theorem 3.3.3. *Suppose the sequence of feature vectors satisfies BBRCNP with parameters $(\kappa, k_1 + k)$ for $k_1 = \frac{1}{3}\kappa^2 k \log(T)$, and assume that T is large enough so that $t_0 \leq (\frac{k_0 T}{\kappa^2 dk})^{1/3}$. Then if Algorithm 12 is run with parameters $B = (\frac{k_0 T}{\kappa^2 dk})^{1/3}$ and k_1 as specified above, its*

[3] We assume, for convenience, that k_0 is divisible by k_1.

Algorithm 12 Online Greedy Algorithm for POSLR

```
Require: Mini-batch size $B$, sparsity parameters $k_0$ and $k_1$
Set up $k_1$ budgeted prediction algorithms $\mathsf{BEXP}^{(i)}$ for $i \in [k_1]$, each using the coordinates in $[d]$ as "experts" with a per-round budget of $\frac{k_0}{k_1}$.
for $b = 1, 2, \ldots, T/B$ do
    For each $i \in [k_1]$, obtain a coordinate $j_b^{(i)}$ and subset of coordinates $U_b^{(i)}$ from $\mathsf{BEXP}^{(i)}$ such that $j_b^{(i)} \in U_b^{(i)}$.
    Define $V_b^{(0)} = \emptyset$ and for each $i \in [k_1]$ define $V_b^{(i)} = \{j_b^{(i')} \mid i' \leq i\}$.
    Set up the Vovk-Azoury-Warmuth (VAW) algorithm for predicting using the features in $V_b^{(k_1)}$.
    for $t \in \mathcal{T}_b$ do
        Set $S_t = \bigcup_{i \in [k_1]} U_b^{(i)}$, obtain $x_t(S_t)$, and pass $x_t(V_b^{(k_1)})$ to VAW.
        Set $w_t$ to be the weight vector output by VAW.
        Obtain the true label $y_t$ and pass it to VAW.
    end for
    Define the function
        $$g_b(S) = \frac{1}{B} \inf_{w \in \mathbb{R}^S} \sum_{t \in \mathcal{T}_b} (y_t - \langle x_t, w \rangle)^2. \quad (3.44)$$
    For each $j \in U_b^{(i)}$, compute $g_b(V_b^{(i-1)} \cup \{j\})$ and pass it $\mathsf{BEXP}^{(i)}$ as the loss for expert $j$.
end for
```

expected regret is at most $\tilde{O}((\frac{\kappa^8 d k^4}{k_0})^{1/3} T^{2/3})$.

Proof. The proof relies on a number of lemmas whose proofs can be found in the supplementary material. We begin with the connection between sparse linear regression, weakly supermodular function and RIP, formally stated in Lemma 3.3.3. This lemma is a direct consequence of Lemma 5 in (Boutsidis et al., 2015).

Lemma 3.3.3. *Consider a sequence of examples* $(x_t, y_t) \in \mathbb{R}^d \times \mathbb{R}$ *for* $t = 1, 2, \ldots, B$, *and let* X *be the design matrix for the sequence. Consider the set function associated with least squares optimization:*

$$g(S) = \inf_{w \in \mathbb{R}^S} \frac{1}{B} \sum_{t=1}^{B} (y_t - \langle x_t, w \rangle)^2.$$

Suppose the restricted condition number of X *for sparsity* k *is bounded by* κ. *Then* $g(S)$ *is*

(k, κ^2)-weakly supermodular.

Even though minimization of weakly supermodular functions is NP-hard, the greedy algorithm provides a good approximation, as shown in the next lemma.

Lemma 3.3.4. *Consider a (k, α)-weakly supermodular set function $g(\cdot)$. Let $j^* := \arg\min_j g(\{j\})$. Then, for any subset V of size at most k, we have*

$$g(\{j^*\}) - g(V) \leq \left(1 - \tfrac{1}{\alpha|V|}\right)[g(\emptyset) - g(V)].$$

The BEXP algorithms essentially implement the greedy algorithm in an online fashion. Using the properties of the BEXP algorithm, we have the following regret guarantee:

Lemma 3.3.5. *Suppose the sequence of feature vectors satisfies BBRCNP with parameters $(\epsilon, k_1 + k)$. Then for any set V of coordinates of size at most k, we have*

$$\begin{aligned}&\mathbf{E}\left[\sum_{b=1}^{T/B} g_b(V_b^{(k_1)}) - g_b(V)\right]\\ &\leq \sum_{b=1}^{T/B}\left(1 - \tfrac{1}{\kappa^2|V|}\right)^{k_1}[g_b(\emptyset) - g_b(V)] + 2\kappa^2 k\sqrt{\tfrac{dk_1\log(d)T}{k_0 B}}.\end{aligned}$$

Finally, within every mini-batch, the VAW algorithm guarantees the following regret bound, an immediate consequence of Theorem 11.8 in Cesa-Bianchi and Lugosi (2006):

Lemma 3.3.6. *Within every batch b, the VAW algorithm generates weight vectors w_t for $t \in \mathcal{T}_b$ such that*

$$\sum_{t\in\mathcal{T}_b}(y_t - \langle x_t, w_t\rangle)^2 - Bg_b(V_b^{(k_1)}) \leq O(k_1\log(B)).$$

We can now prove Theorem 3.3.3. Combining the bounds of lemma 3.3.5 and 3.3.6, we conclude that for any subset of coordinates V of size at most k, we have

$$\begin{aligned}
&\mathbf{E}\left[\sum_{t=1}^{T}(y_t - \langle x_t, w_t\rangle)^2\right] && (3.45)\\
&\leq \sum_{b=1}^{T/B} Bg_b(V) + B(1 - \tfrac{1}{\kappa^2|V|})^{k_1}[g_b(\emptyset) - g_b(V)] && (3.46)\\
&+ O\left(\kappa^2 k\sqrt{\tfrac{dk_1 \log(d)BT}{k_0}} + \frac{T}{B}k_1 \log(B)\right). && (3.47)
\end{aligned}$$

Finally, note that

$$\sum_{b=1}^{T/B} Bg_b(V) \leq \inf_{w\in\mathbb{R}^V} \sum_{t=1}^{T}(y_t - \langle x_t, w\rangle)^2,$$

and

$$\sum_{b=1}^{T/B} B(1 - \tfrac{1}{\kappa^2|V|})^{k_1}[g_b(\emptyset) - g_b(V)] \leq T \cdot \exp(-\tfrac{k_1}{\kappa^2 k}),$$

because $g_b(\emptyset) \leq 1$. Using these bounds in (3.47), and plugging in the specified values of B and k_1, we get the stated regret bound. □

3.3.4. Conclusions and Future Work

In this section, we gave computationally efficient algorithms for the online sparse linear regression problem under the assumption that the design matrices of the feature vectors satisfy RIP-type properties. Since the problem is hard without any assumptions, our work is the first one to show that assumptions that are similar to the ones used to sparse recovery in the batch setting yield tractability in the online setting as well.

Several open questions remain in this line of work and will be the basis for future work. Is it possible to improve the regret bound in the agnostic setting? Can we give matching lower bounds on the regret in various settings? Is it possible to relax the RIP assumption on the design matrices and still have efficient algorithms?

APPENDIX

A.1. Appendix for Section 2.1

We prove in this section the main results given in the paper. We first collect and prove a few important technical lemmas that will be used in the proofs of the main results.

Prerequisite Lemmas

We start with the following version of the Wedin's Theorem.

Lemma A.1.1 (Davis-Kahan-Wedin's Type Perturbation Bound)**.** *It holds that*

$$\sqrt{\|\sin\Phi\|_F^2 + \|\sin\Theta\|_F^2} \leq \frac{\sqrt{2}\|E\|_F}{\delta}$$

and also the following holds for 2*-norm (or any unitary invariant norm)*

$$\max\{\|\sin\Phi\|_2, \|\sin\Theta\|_2\} \leq \frac{\|E\|_2}{\delta}.$$

We will then introduce some concentration inequalities. Lemmas A.1.2 and A.1.3 are concentration of measure results from random matrix theory.

Lemma A.1.2 (Vershynin (2010), Theorem 39)**.** *Let $Z \in \mathbb{R}^{m\times n}$ be a matrix whose rows $Z_{i\cdot}$ are independent sub-Gaussian isotropic random vectors in $\mathbb{R}^n$ with parameter σ. Then for every $t \geq 0$, with probability at least $1 - 2\exp(-ct^2)$ one has*

$$\|Z\|_2 \leq \sigma(\sqrt{m} + C\sqrt{n} + t)$$

where $C, c > 0$ are some universal constants.

Lemma A.1.3 (Hsu et al. (2012), Projection Lemma)**.** *Assume $Z \in \mathbb{R}^n$ is an isotropic sub-Gaussian vector with i.i.d. entries and parameter σ. $\mathcal{P}$ is a projection operator to a* *subspace of dimension r, then we have the following concentration inequality*

$$\mathbb{P}(\|\mathcal{P}Z\|_{\ell_2}^2 \geq \sigma^2(r + 2\sqrt{rt} + 2t)) \leq \exp(-ct),$$

where $c > 0$ is a universal constant.

The proof of this lemma is a simple application of Theorem 2.1 inHsu et al. (2012) for the case that $\mathcal{P}$ is a rank r positive semidefinite projection matrix.

The following two are standard Chernoff-type bounds for bounded random variables.

Lemma A.1.4 (Hoeffding (1963), Hoeffding's Inequality)**.** *Let $X_i, 1 \leq i \leq n$ be independent random variables. Assume $a_i \leq X_i \leq b_i, 1 \leq i \leq n$. Then for $S_n = \sum_{i=1}^n X_i$*

$$\mathbb{P}\left(|S_n - \mathbb{E}S_n| > u\right) \leq 2\exp\left(-\frac{2u^2}{\sum_{i=1}^n (b_i - a_i)^2}\right). \tag{A.1}$$

Lemma A.1.5 (Bennett (1962), Bernstein's Inequality)**.** *Let $X_i, 1 \leq i \leq n$ be independent zero-mean random variables. Suppose $|X_i| \leq M, 1 \leq i \leq n$. Then*

$$\mathbb{P}\left(\sum_{i=1}^n X_i > u\right) \leq \exp\left(-\frac{u^2/2}{\sum_{i=1}^n \mathbb{E}X_i^2 + Mu/3}\right). \tag{A.2}$$

We will end this section stating the Fano's information inequality, which plays a key role in many information theoretic lower bounds.

Lemma A.1.6 (Tsybakov (2009) Corollary 2.6)**.** *Let $\mathcal{P}_0, \mathcal{P}_1, \ldots, \mathcal{P}_M$ be probability measures on the same probability space $(\Theta, \mathcal{F})$, $M \geq 2$. If for some $0 < \alpha < 1$*

$$\frac{1}{M+1}\sum_{i=0}^M KL(\mathcal{P}_i||\bar{\mathcal{P}}) \leq \alpha \cdot \log M \tag{A.3}$$

where

$$\bar{\mathcal{P}} = \frac{1}{M+1}\sum_{i=0}^M \mathcal{P}_i.$$

Then

$$p_{e,M} \geq \bar{p}_{e,M} \geq \frac{\log(M+1) - \log 2}{\log M} - \alpha \tag{A.4}$$

where $p_{e,M}$ is the minimax error for the multiple testing problem.

Main Proofs

Proof of Lemma 2.1.1. Recall the matrix form of the submatrix model, with the SVD decomposition of the mean signal matrix M

$$X = \lambda\sqrt{k_m k_n} U V^T + Z.$$

The largest singular value of $\lambda U V^T$ is $\lambda\sqrt{k_m k_n}$, and all the other singular values are 0s. Davis-Kahan-Wedin's perturbation bound tells us how close the singular space of X is to the singular space of M. Let us apply the derived Lemma A.1.1 to $X = \lambda\sqrt{k_m k_n} U V^T + Z$. Denote the top left and right singular vector of X as $\tilde{U}$ and $\tilde{V}$. One can see that $\mathbb{E}\|Z\|_2 \asymp \sigma(\sqrt{m} + \sqrt{n})$ under very mild finite fourth moment conditions through a result in (Latała, 2005). Lemma A.1.2 provides a more explicit probabilisitic bound for the concentration of the largest singular value of i.i.d sub-Gaussian random matrix. Because the rows $Z_{i\cdot}$ are sampled from product measure of mean zero sub-Gaussians, they naturally satisfy the isotropic condition. Hence, with probability at least $1 - 2\exp(-c(m+n))$, via Lemma A.1.2, we reach

$$\|Z\|_2 \leq C \cdot \sigma(\sqrt{m} + \sqrt{n}). \tag{A.5}$$

Using Weyl's interlacing inequality, we have

$$|\sigma_i(X) - \sigma_i(M)| \leq \|Z\|_2$$

and thus

$$\sigma_1(X) \geq \lambda\sqrt{k_m k_n} - \|Z\|_2$$

$$\sigma_2(X) \leq \|Z\|_2.$$

Applying Lemma A.1.1, we have

$$\max\Big\{|\sin\angle(U,\tilde{U})|, |\sin\angle(V,\tilde{V})|\Big\} \leq \frac{C\sigma(\sqrt{m}+\sqrt{n})}{\lambda\sqrt{k_m k_n} - C\sigma(\sqrt{m}+\sqrt{n})} \asymp \frac{\sigma(\sqrt{m}+\sqrt{n})}{\lambda\sqrt{k_m k_n}}.$$

In addition

$$\|U - \tilde{U}\|_{\ell_2} = \sqrt{2 - 2\cos\angle(U,\tilde{U})} = 2|\sin\frac{1}{2}\angle(U,\tilde{U})|,$$

which means

$$\max\Big\{\|U - \tilde{U}\|_{\ell_2}, \|V - \tilde{V}\|_{\ell_2}\Big\} \leq C \cdot \frac{\sigma(\sqrt{m}+\sqrt{n})}{\lambda\sqrt{k_m k_n}}.$$

And according to the definition of the canonical angles, we have

$$\max\Big\{\|UU^T - \tilde{U}\tilde{U}^T\|_2, \|VV^T - \tilde{V}\tilde{V}^T\|_2\Big\} \leq C \cdot \frac{\sigma(\sqrt{m}+\sqrt{n})}{\lambda\sqrt{k_m k_n}}.$$

Now let us assume we have two observations of X. We use the first observation $\tilde{X}$ to solve for the singular vectors $\tilde{U}, \tilde{V}$, we use the second observation X to project to the singular vectors $\tilde{U}, \tilde{V}$. We can use Tsybakov's sample cloning argument (Tsybakov (2014), Lemma 2.1) to create two independent observations of X when noise is Gaussian as follows. Create a pure Gaussian matrix Z' and define $X_1 = X + Z' = M + (Z + Z')$ and $X_2 = X - Z' = M + (Z - Z')$, making X_1, X_2 independent with the variance being doubled. This step is not essential because we can perform random subsampling as in Vu (2014); having two observations instead of one does not change the picture statistically or computationally. Recall $X = M + Z = \lambda\sqrt{k_m k_n} UV^T + Z$.

Define the projection operator to be $\mathcal{P}$, we start the analysis by decomposing

$$\|\mathcal{P}_{\tilde{U}} X_{\cdot j} - M_{\cdot j}\|_{\ell_2} \leq \|\mathcal{P}_{\tilde{U}}(X_{\cdot j} - M_{\cdot j})\|_{\ell_2} + \|(\mathcal{P}_{\tilde{U}} - I)M_{\cdot j}\|_{\ell_2} \tag{A.6}$$

for $1 \leq j \leq n$.

For the first term of (A.6), note that $X_{\cdot j} - M_{\cdot j} = Z_{\cdot j} \in \mathbb{R}^m$ is an i.i.d. isotropic sub-Gaussian vector, and thus we have through Lemma A.1.3, for $t = (1 + 1/c)\log n$, $Z_{\cdot j} \in \mathbb{R}^m, 1 \leq j \leq n$ and $r = 1$

$$\mathbb{P}\left(\|\mathcal{P}_{\tilde{U}}(X_{\cdot j} - M_{\cdot j})\|_{\ell_2} \geq \sigma\sqrt{r}\sqrt{1 + 2\sqrt{1+1/c} \cdot \sqrt{\frac{\log n}{r}} + 2(1+1/c) \cdot \frac{\log n}{r}}\right) \leq n^{-c-1}. \tag{A.7}$$

We invoke the union bound for all $1 \leq j \leq n$ to obtain

$$\max_{1 \leq j \leq n} \|\mathcal{P}_{\tilde{U}}(X_{\cdot j} - M_{\cdot j})\|_{\ell_2} \leq \sigma\sqrt{r} + \sqrt{2(1+1/c)} \cdot \sigma\sqrt{\log n} \tag{A.8}$$

$$\leq \sigma + C \cdot \sigma\sqrt{\log n} \tag{A.9}$$

with probability at least $1 - n^{-c}$.

For the second term $M_{\cdot j} = \tilde{X}_{\cdot j} - \tilde{Z}_{\cdot j}$ of (A.6), there are two ways of upper bounding it. The first approach is to split

$$\|(\mathcal{P}_{\tilde{U}} - I)M\|_2 \leq \|(\mathcal{P}_{\tilde{U}} - I)\tilde{X}\|_2 + \|(\mathcal{P}_{\tilde{U}} - I)\tilde{Z}\|_2 \leq 2\|\tilde{Z}\|_2. \tag{A.10}$$

The first term of (A.10) is $\sigma_2(\tilde{X}) \leq \sigma_2(M) + \|\tilde{Z}\|_2$ through Weyl's interlacing inequality, while the second term is bounded by $\|\tilde{Z}\|_2$. We also know that $\|\tilde{Z}\|_2 \leq C_3 \cdot \sigma(\sqrt{m} + \sqrt{n})$. Recall the definition of the induced ℓ_2 norm of a matrix $(\mathcal{P}_{\tilde{U}} - I)M$:

$$\|(\mathcal{P}_{\tilde{U}} - I)M\|_2 \geq \frac{\|(\mathcal{P}_{\tilde{U}} - I)MV\|_{\ell_2}}{\|V\|_{\ell_2}} = \|(\mathcal{P}_{\tilde{U}} - I)\lambda\sqrt{k_m k_n}U\|_{\ell_2} \geq \sqrt{k_n}\|(\mathcal{P}_{\tilde{U}} - I)M_{\cdot j}\|_{\ell_2}.$$

In the second approach, the second term of (A.6) can be handled through perturbation Sin Theta Theorem A.1.1:

$$\|(\mathcal{P}_{\tilde{U}} - I)M_{\cdot j}\|_{\ell_2} = \|(\mathcal{P}_{\tilde{U}} - \mathcal{P}_U)M_{\cdot j}\|_{\ell_2} \leq \|\tilde{U}\tilde{U}^T - UU^T\|_2 \cdot \|M_{\cdot j}\|_{\ell_2} \leq C\frac{\sigma\sqrt{m+n}}{\lambda\sqrt{k_m k_n}}\lambda\sqrt{k_m}.$$

This second approach will be used in the multiple submatrices analysis.

Combining all the above, we have with probability at least $1 - n^{-c} - m^{-c}$, for all $1 \leq j \leq n$

$$\|\mathcal{P}_{\tilde{U}} X_{\cdot j} - M_{\cdot j}\|_{\ell_2} \leq C \cdot \left(\sigma\sqrt{\log n} + \sigma\sqrt{\frac{m \vee n}{k_n}}\right). \tag{A.11}$$

Similarly we have for all $1 \leq i \leq m$,

$$\|\mathcal{P}_{\tilde{V}} X_{i\cdot}^T - M_{i\cdot}^T\|_{\ell_2} \leq C \cdot \left(\sigma\sqrt{\log m} + \sigma\sqrt{\frac{m \vee n}{k_m}}\right). \tag{A.12}$$

Clearly we know that for $i \in R_m$ and $i' \in [m]\backslash R_m$

$$\|M_{i\cdot}^T - M_{i'\cdot}^T\|_{\ell_2} = \lambda\sqrt{k_n}$$

and for $j \in C_n$ and $j' \in [n]\backslash C_n$

$$\|M_{\cdot j} - M_{\cdot j'}\|_{\ell_2} = \lambda\sqrt{k_m}$$

Thus if

$$\lambda\sqrt{k_m} \geq 6C \cdot \left(\sigma\sqrt{\log n} + \sigma\sqrt{\frac{m \vee n}{k_n}}\right) \tag{A.13}$$

$$\lambda\sqrt{k_n} \geq 6C \cdot \left(\sigma\sqrt{\log m} + \sigma\sqrt{\frac{m \vee n}{k_m}}\right) \tag{A.14}$$

hold, we have

$$2 \max_{i,i' \in R_m} \|\mathcal{P}_{\tilde{V}} X_{i\cdot}^T - \mathcal{P}_{\tilde{V}} X_{i'\cdot}^T\| \leq \min_{i \in R_m, i' \in [m] \backslash R_m} \|\mathcal{P}_{\tilde{V}} X_{i\cdot}^T - \mathcal{P}_{\tilde{V}} X_{i'\cdot}^T\|$$

Therefore we have got $d_i = X_{i\cdot} \tilde{V} \in \mathbb{R}$ (a one dimensional line along direction $\tilde{V}$) such that on this line, data forms two data-driven clusters in the sense that

$$2 \max_{i,i' \in R_m} |d_i - d_{i'}| \leq \min_{i \in R_m, i' \in [m] \backslash R_m} |d_i - d_{i'}|.$$

In this case, the largest adjacent gap in $d_i, i \in [m]$ (data-driven) suggests the cut-off (without requiring the knowledge of λ, σ, k_m). And the simple cut-off clustering recovers the nodes exactly.

In summary, if

$$\lambda \geq C \cdot \sigma \left(\sqrt{\frac{\log n}{k_m}} + \sqrt{\frac{\log m}{k_n}} + \sqrt{\frac{m+n}{k_m k_n}} \right),$$

the spectral algorithm succeeds with probability at least

$$1 - m^{-c} - n^{-c} - 2\exp\left(-c(m+n)\right).$$

□

Proof of Theorem 2.1.2. Computational lower bound for localization (support recovery) is of different nature than the computational lower bound for detection (two point testing). The idea is to design a randomized polynomial time algorithmic reduction to relate an instance of *hidden clique* problem to our submatrix localization problem. The proof proceeds in the following way: we will construct a randomized polynomial time transformation $\mathcal{T}$ to map a random instance of $\mathcal{G}(N, \kappa)$ to a random instance of our submatrix $\mathcal{M}(m = n, k_m \asymp k_n \asymp k, \lambda/\sigma)$ (abbreviated as $\mathcal{M}(n, k, \lambda/\sigma)$). Then we will provide a quantitative computational lower bound by showing that if there is a polynomial time algorithm that pushes below the hypothesized computational boundary for localization in the submatrix model, there will be a polynomial time algorithm that solves hidden clique localization with high probability (a contradiction to $\mathsf{HC_l}$).

Denote the randomized polynomial time transformation as

$$\mathcal{T} : \mathcal{G}(N, \kappa(N)) \to M(n, k = n^{\alpha}, \lambda/\sigma = n^{-\beta}).$$

There are several stages for the construction of the algorithmic reduction. First we define a graph $\mathcal{G}^e(N, \kappa(N))$ that is stochastically equivalent to the hidden clique graph $\mathcal{G}(N, \kappa(N))$, but is easier for theoretical analysis. $\mathcal{G}^e$ has the property: each node independently has the probability $\kappa(N)/N$ to be a clique node, and with the remaining probability a non-clique node. Using Bernstein's inequality and the inequality (A.20) proved below, with probability at least $1 - 2N^{-1}$ the number of clique nodes κ^e in $\mathcal{G}^e$

$$\kappa \left(1 - \sqrt{\frac{4 \log N}{\kappa}}\right) \leq \kappa^e \leq \kappa \left(1 + \sqrt{\frac{4 \log N}{\kappa}}\right) \Rightarrow \kappa^e \asymp \kappa \tag{A.15}$$

as long as $\kappa \gtrsim \log N$.

Consider a hidden clique graph $\mathcal{G}^e(2N, 2\kappa(N))$ with $N = n$ and $\kappa(N) = \kappa$. Denote the set of clique nodes for $\mathcal{G}^e(2N, 2\kappa(N))$ to be $C_{N,\kappa}$. Represent the hidden clique graph using the symmetric adjacency matrix $G \in \{-1, 1\}^{2N \times 2N}$, where $G_{ij} = 1$ if $i, j \in C_{N,\kappa}$, otherwise with equal probability to be either -1 or 1. As remarked before, with probability at least $1 - 2N^{-1}$, we have planted $2\kappa(1 \pm o(1))$ clique nodes in graph $\mathcal{G}^e$ with $2N$ nodes. Take out the upper-right submatrix of G, denote as G_{UR} where U is the index set $1 \leq i \leq N$ and R is the index set $N + 1 \leq j \leq 2N$. Now G_{UR} has independent entries.

The construction of $\mathcal{T}$ employs the *Bootstrapping* idea. Generate l^2 (with $l \asymp n^{\beta}, 0 < \beta < 1/2$) matrices through bootstrap subsampling as follows. Generate $l - 1$ independent index vectors $\psi^{(s)} \in \mathbb{R}^n, 1 \leq s < l$, where each element $\psi^{(s)}(i), 1 \leq i \leq n$ is a random draw with replacement from the row indices $[n]$. Denote vector $\psi^{(0)}(i) = i, 1 \leq i \leq n$ as the original index set. Similarly, we can define independently the column index vectors $\phi^{(t)}, 1 \leq t < l$. We remark that these bootstrap samples can be generated in polynomial time $\Omega(l^2 n^2)$. The transformation is a weighted average of l^2 matrices of size $n \times n$ generated based on the original adjacency matrix G_{UR}.

$$\mathcal{T}: \quad M_{ij} = \frac{1}{l} \sum_{0 \leq s,t < l} (G_{UR})_{\psi^{(s)}(i)\phi^{(t)}(j)}, \quad 1 \leq i, j \leq n. \tag{A.16}$$

Recall that $C_{N,\kappa}$ stands for the clique set of the hidden clique graph. We define the row candidate set $R_l := \{i \in [n] : \exists\ 0 \leq s < l, \psi^{(s)}(i) \in C_{N,\kappa}\}$ and column candidate set $C_l := \{j \in [n] : \exists\ 0 \leq t < l, \phi^{(t)}(j) \in C_{N,\kappa}\}$. Observe that $R_l \times C_l$ are the indices where the matrix M contains signal.

There are two cases for M_{ij}, given the candidate set $R_l \times C_l$. If $i \in R_l$ and $j \in C_l$, namely when (i, j) is a clique edge in at least one of the l^2 matrices, then $\mathbb{E}[M_{ij}|\mathcal{G}^e] \geq l^{-1}$ where the expectation is taken over the bootstrap σ-field conditioned on the candidate set $R_l \times C_l$ and the original σ-field of $\mathcal{G}^e$. Otherwise $\mathbb{E}[M_{ij}|\mathcal{G}^e] = l(\frac{|E|}{N^2-\kappa^2} - \frac{1}{2})$ for $(i, j) \notin R_l \times C_l$, where $|E|$ is a Binomial$(N^2 - \kappa^2, 1/2)$. With high probability, $\mathbb{E}[M_{ij}|\mathcal{G}^e] \asymp \frac{l}{\sqrt{N^2-\kappa^2}} \asymp \frac{l}{n} = o(\frac{1}{l})$. Thus the mean separation between the signal position and non-signal position is $\frac{1}{\ell} - \frac{l}{n} \asymp \frac{1}{\ell}$. Note in the submatrix model, it does not matter if the noise has mean zero or not (since we can subtract the mean)– only the signal separation matters.

Now let us discuss the independence issue in M through our Bootstrapping construction. Clearly due to sampling with replacement and bootstrapping, condition on $\mathcal{G}^e$, we have independence among samples for the same location (i, j)

$$(G_{UR})_{\psi^{(s)}(i)\phi^{(t)}(j)} \perp (G_{UR})_{\psi^{(s')}(i)\phi^{(t')}(j)}.$$

For the independence among entries in one Bootstrapped matrix, clearly

$$(G_{UR})_{\psi^{(s)}(i)\phi^{(t)}(j)} \perp (G_{UR})_{\psi^{(s)}(i')\phi^{(t)}(j')}.$$

The only case where there might be a weak dependence is between

$$(G_{UR})_{\psi^{(s)}(i)\phi^{(t)}(j)}, (G_{UR})_{\psi^{(s)}(i)\phi^{(t)}(j')}$$

and $(G_{UR})_{\psi^{(s)}(i)\phi^{(t)}(j)}, (G_{UR})_{\psi^{(s)}(i)\phi^{(t')}(j)}$. The way to eliminate the weak dependence is through Vu (2008)'s result on universality of random discrete graphs. Vu (2008) showed that random regular graph $\mathcal{G}(n, n/2)$ shares many similarities with Erdős-Rényi random graph $\mathcal{G}(n, 1/2)$: for instance, top and second eigenvalues ($n/2$ and $\sqrt{n}$ respectively), limiting spectral distribution, sandwich conjecture, determinant, etc. Let us consider the case where the upper-right of the adjacency matrix G consists of random bi-regular graph with a planted clique. We assume that the hidden clique hypothesis for $k \precsim \sqrt{n}$ is still valid for the following random graph: for a $n \times n$ adjacency matrix G, first find a clique/principal submatrix of size k uniformly randomly and connect density, for the remaining part of the matrix, sample a random regular graph of $\mathcal{G}(n-k, \frac{n-k}{2})$ and a random bi-regular graph of size $k \times (n-k)$ with left regular degree $n/2 - k$ and right regular degree $k/2$ (here degree test will not work in this graph and spectral barrier still suggests $k \precsim \sqrt{n}$ is hard due to universality result of random discrete graphs). In the bootstrapping step, conditionally on the row $\psi^{(s)}(i)$ being not a clique, $(G_{UR})_{\psi^{(s)}(i)\phi^{(t)}(j)} \perp (G_{UR})_{\psi^{(s)}(i)\phi^{(t)}(j')} | \psi^{(s)}(i)$, and each one is a Rademacher random variable (regardless of the choice of $\psi^{(s)}(i)$), which implies $(G_{UR})_{\psi^{(s)}(i)\phi^{(t)}(j)} \perp (G_{UR})_{\psi^{(s)}(i)\phi^{(t)}(j')}$ holds unconditionally. Thus in the bootstrapping procedure, we have independence among entries within the matrix unconditionally.

Let us move to verify the sub-Gaussianity of M matrix. Note that for the index i, j that is not a clique for any of the matrices, M_{ij} is sub-Gaussian, due to Hoeffding's inequality

$$\mathbb{P}\left(|M_{ij} - \mathbb{E}M_{ij}| \geq u\right) \leq 2\exp(-u^2/2). \tag{A.17}$$

For the index i, j being a clique in at least one of the matrices, we claim the number of matrices has (i, j) being clique is $O^*(1)$. Due to Bernstein's inequality, we have $\max_i |\{0 \leq s < l : \psi^{(s)}(i) \in C_{N,\kappa}\}| \leq \frac{\kappa l}{n} + \frac{8}{3}\log n$ with probability at least $1 - n^{-1}$. This further implies there are at least $l^2 - (\frac{\kappa l}{n} + \frac{8}{3}\log n)^2$ many independent Rademacher random variables in each i, j position, thus

$$\mathbb{P}\left(|M_{ij} - \mathbb{E}M_{ij}| \geq u\right) \leq 2\exp\left(-(1 - C\cdot(\kappa n^{-1} + l^{-1}\log n)^2)u^2/2\right). \tag{A.18}$$

Up to now we have proved that when i, j is a signal node for M, then $O^*(1)l^{-1} \geq \mathbb{E}M_{ij} \geq l^{-1}$. Thus the sub-Gaussian parameter is $\sigma = 1 - o(1)$ because $\kappa n^{-1}, l^{-1}\log n$ are both $o(1)$. The constructed $M(n, k, \lambda/\sigma)$ matrix satisfies the submatrix model with $\lambda/\sigma \asymp l^{-1}$ and sub-Gaussian parameter $\sigma = 1 - o(1)$.

Let us estimate the corresponding k in the submatrix model. We need to bound the order of the cardinality of R_l, denoted as $|R_l|$. The total number of positions with signal (at least one clique node inside) is

$$\mathbb{E}|R_l| = \mathbb{E}|\{1 \leq i \leq n : i \in R_l\}| = n\left[1 - (1 - \kappa/n)^l\right].$$

Thus we have the two sided bound

$$\kappa l\left(1 - \frac{\kappa l}{2n}\right) \leq \mathbb{E}|R_l| \leq \kappa l$$

which is of the order $k := \kappa l$. Let us provide a high probability bound on $|R_l|$. By Bernstein's inequality

$$\mathbb{P}\left(||R_l| - \mathbb{E}|R_l|| > u\right) \le 2\exp\left(-\frac{u^2/2}{\kappa l + u/3}\right). \tag{A.19}$$

Thus if we take $u = \sqrt{4\kappa l \log n}$, as long as $\log n = o(\kappa l)$,

$$\mathbb{P}\left(||R_l| - \mathbb{E}|R_l|| > \sqrt{4\kappa l \log n}\right) \le 2n^{-1}. \tag{A.20}$$

So with probability at least $1 - 2n^{-1}$, the number of positions that contain signal nodes is bounded as

$$\kappa l\left(1 - \frac{\kappa l}{n}\right)\left(1 - \sqrt{\frac{4\log n}{\kappa l}}\right) < |R_l| < \kappa l\left(1 + \sqrt{\frac{4\log n}{\kappa l}}\right) \Rightarrow |R_l| \asymp \kappa l. \tag{A.21}$$

Equation (A.21) implies that with high probability

$$\kappa l(1 - o(1)) \le |R_l| \le \kappa l(1 + o(1)),$$
$$\kappa l(1 - o(1)) \le |C_l| \le \kappa l(1 + o(1)).$$

The above means, in the submatrix parametrization, $k_m \asymp k_n \asymp \kappa l \asymp n^{\alpha}$, $\lambda/\sigma \asymp l^{-1} \asymp n^{-\beta}$, which implies $\kappa \asymp n^{\alpha-\beta}$.

Suppose there exists a polynomial time algorithm $\mathcal{A}_M$ that pushes below the computational boundary. In other words,

$$n^{-\beta} \asymp \frac{\lambda}{\sigma} \precsim \sqrt{\frac{m+n}{k_m k_n}} \asymp n^{(1-2\alpha)/2} \Rightarrow \beta > \alpha - \frac{1}{2}$$

with the last inequality having a slack $\epsilon > 0$. More precisely, $\mathcal{A}_M$ returns two estimated index sets $\hat{R}_n$ and $\hat{C}_n$ corresponding to the location of the submatrix (and correct with probability going to 1) under the regime $\beta = \alpha - 1/2 + \epsilon$. Suppose under some conditions, this algorithm $\mathcal{A}_M$ can be modified to a randomized polynomial time algorithm $\mathcal{A}_{\mathcal{G}}$ that correctly identifies the hidden clique nodes with high probability. It means in the corresponding hidden clique graph $\mathcal{G}(2N, 2\kappa)$, $\mathcal{A}_{\mathcal{G}}$ also pushes below the computational boundary of hidden clique by the amount ϵ:

$$\kappa(N) = 2\kappa \asymp (2n)^{\alpha-\beta} \asymp n^{1/2-\epsilon} \precsim n^{1/2} \asymp N^{\frac{1}{2}}.$$

In summary, the quantitative computational lower bound implies that if the computational boundary for submatrix localization is pushed below by an amount ϵ in the power, the *hidden clique* boundary is correspondingly improved by ϵ.

Now let us show that any algorithm $\mathcal{A}_M$ that localizes the submatrix introduces a randomized algorithm that finds the hidden clique nodes with probability tending to 1. The algorithm relies on the following simple lemma.

Lemma A.1.7. *For the hidden clique model $\mathcal{G}(N, \kappa)$, suppose an algorithm provides a candidate set S of size k that contains the true clique subset. If*

$$\kappa \geq C\sqrt{k \log N}$$

then by looking at the adjacency matrix restricted to S we can recover the clique subset exactly with high probability.

The proof of Lemma A.1.7 is immediate. If i is a clique node, then $\min_i \sum_{j \in C} G_{ij} \geq \kappa - C/2 \cdot \sqrt{k \log N}$. If i is not a clique node, then $\max_i \sum_{j \in C} G_{ij} \leq C/2 \cdot \sqrt{k \log N}$. The proof is completed.

Algorithm $\mathcal{A}_M$ provides candidate sets R_l, C_l of size k, inside which κ are correct clique nodes, and thus exact recovery can be completed through Lemma A.1.7 since $\kappa \succsim (k \log N)^{1/2}$ (since $\kappa \asymp n^{1/2-\epsilon} \succsim k^{1/2} \asymp n^{\alpha/2}$ when ϵ is small). The algorithm $\mathcal{A}_M$ induces another randomized polynomial time algorithm $\mathcal{A}_{\mathcal{G}}$ that solves the hidden clique problem $\mathcal{G}(2N, 2\kappa)$ with $\kappa \precsim N^{1/2}$. The algorithm $\mathcal{A}_{\mathcal{G}}$ returns the support $\hat{C}_{N,\kappa}$ that coincides with the true support $C_{N,\kappa}$ with probability going to 1 (a contradiction to the hidden clique hypothesis $\mathsf{HC_l}$). We conclude that, under the hypothesis, there is no polynomial time algorithm $\mathcal{A}_M$ that can push below the computational boundary $\lambda \precsim \sqrt{\frac{m+n}{k_m k_n}}$.

□

Proof of Theorem 2.1.3 is a direct result of Lemma 2.1.1 and Theorem 2.1.2. Proof of Theorem 2.1.4 is obvious based on Lemma 2.1.2 and the hidden clique hypothesis $\mathsf{HC_l}$. Proof of Theorem 2.1.5 combines the result of Lemma 2.1.5 and Lemma 2.1.4.

A.2. Appendix for Section 2.2

We will start with two useful Lemmas. Lemma A.2.1 couples the local behavior of a stochastic block model with that of a Galton-Watson branching process. Lemma A.2.2 is the well-known Hoeffding's inequality.

Lemma A.2.1 (Proposition 4.2 in (Mossel et al., 2012))**.** *Take* $t = \bar{t}_{n,k,p,q} \precsim \frac{\log n}{\log[kn(q+\frac{p-q}{k})]}$. *There exists a coupling between* (G,σ) *and* (T,ℓ) *such that* $(G_{\le t}, \sigma_{G_{\le t}}) = (T_{\le t}, \ell_{T_{\le t}})$ *asymptotically almost surely. Here* (T,ℓ) *corresponds to the broadcast process on a Galton-Watson tree process* T *with offspring distribution* $\mathsf{Poisson}\left(n(q+\frac{p-q}{k})\right)$, *and* (G,σ) *corresponds to the SBM and its labels.*

Lemma A.2.2 (Hoeffding's Inequality)**.** *Let* X *be any real-valued random variable with expected value* $\mathbb{E}X = 0$ *and such that* $a \le X \le b$ *almost surely. Then, for all* $\lambda > 0$,

$$\mathbb{E}\left[e^{\lambda X}\right] \le \exp\left(\frac{\lambda^2(b-a)^2}{8}\right).$$

Now we are ready to prove the main theoretical results. First, we focus on the $k = 2$ case and prove the broadcasting tree version of Theorem 2.2.2 and Theorem 2.2.3, under the assumption the tree is regular. Later, based on these two theorems, Theorem 2.2.1 for p-SBM ($k = 2$) is proved. Similarly for general k case, we will first prove Theorem 2.2.5 and Theorem 2.2.6.

Proof of Theorem 2.2.2. We focus on a regular tree where each node has $(1-\delta)d$ unlabeled children and δd labeled children. For $t = 1$, the results follow from Hoeffding's inequality directly because

$$M_1(\ell_{T_1(\rho)}) = \Big(N_{\mathcal{C}^{\mathrm{l}}(\rho)}(+) - N_{\mathcal{C}^{\mathrm{l}}(\rho)}(-)\Big) \log \frac{1+\theta}{1-\theta}.$$

Let us use induction to prove the remaining claim. Assume for tree with depth $t-1$ rooted from u, for any $\lambda > 0$

$$\mathbb{E}\left[e^{\lambda M_{t-1}(\ell_{T_{\le t-1}(u)})} | \ell(u) = +\right] \le e^{\lambda \mu_{t-1}} \cdot e^{\frac{\lambda^2}{2}\sigma_{t-1}^2},$$
$$\mathbb{E}\left[e^{\lambda M_{t-1}(\ell_{T_{\le t-1}(u)})} | \ell(u) = -\right] \le e^{-\lambda \mu_{t-1}} \cdot e^{\frac{\lambda^2}{2}\sigma_{t-1}^2}.$$

These will further imply, conditionally on $\ell(u) = +$,

$$M_{t-1}(\ell_{T_{\le t-1}(u)}) \in \mu_{t-1} \pm x \cdot \sigma_{t-1};$$

and conditionally on $\ell(v) = -$,

$$M_{t-1}(\ell_{T_{\le t-1}(u)}) \in -\mu_{t-1} \pm x \cdot \sigma_{t-1};$$

both with probability at least $1 - 2\exp(x^2/2)$. Now, recall the recursion for AMP:

$$M_t(\ell_{T_{\le t}(v)}) = M_1(\ell_{T_1(v)}) + \theta \cdot \sum_{u \in \mathcal{C}^{\mathrm{u}}(v)} M_{t-1}(\ell_{T_{\le t-1}(u)}).$$

For the moment generating function we have

$$\begin{aligned}
&\mathbb{E}\left[e^{\lambda M_t(\ell_{T_{\leq t}(v)})}|\ell(v)=+\right] \\
&\leq e^{\lambda\left(\theta\delta d\cdot\log\frac{1+\theta}{1-\theta}\right)} e^{\frac{\lambda^2}{2}\left(\sqrt{\delta d}\cdot\log\frac{1+\theta}{1-\theta}\right)^2} \cdot \prod_{u\in\mathcal{C}^{\mathrm{u}}(v)} \mathbb{E}\left[e^{\lambda\theta M_{t-1}(\ell_{T_{\leq t-1}(u)})}|\ell(v)=+\right] \\
&= e^{\lambda\mu_1} e^{\frac{\lambda^2}{2}\sigma_1^2} \cdot \prod_{u\in\mathcal{C}^{\mathrm{u}}(v)} \mathbb{E}\left[e^{\lambda\theta M_{t-1}(\ell_{T_{\leq t-1}(u)})}|\ell(v)=+\right].
\end{aligned}$$

The last term in the previous equation can be written as

$$\begin{aligned}
&\mathbb{E}\left[e^{\lambda\theta M_{t-1}(\ell_{T_{\leq t-1}(u)})}|\ell(v)=+\right] \\
&\leq e^{\frac{\lambda^2\theta^2}{2}\sigma_{t-1}^2} \cdot \left\{\frac{1+\theta}{2}e^{\lambda\theta\mu_{t-1}} + \frac{1-\theta}{2}e^{-\lambda\theta\mu_{t-1}}\right\} \qquad \text{(A.22)} \\
&\leq e^{\frac{\lambda^2\theta^2}{2}\sigma_{t-1}^2} \cdot e^{\lambda\theta\left(\frac{1+\theta}{2}\mu_{t-1}-\frac{1-\theta}{2}\mu_{t-1}\right)} \cdot e^{\frac{\lambda^2\theta^2}{2}\mu_{t-1}^2} \qquad \text{(A.23)} \\
&= e^{\frac{\lambda^2\theta^2}{2}\sigma_{t-1}^2} \cdot e^{\lambda\theta^2\mu_{t-1}} \cdot e^{\frac{\lambda^2\theta^2}{2}\mu_{t-1}^2}
\end{aligned}$$

where equation (A.22) to (A.23) relies on Hoeffding's lemma: for a random variable $Y = \theta\mu_{t-1}$ with probability $\frac{1+\theta}{2}$ and $Y = -\theta\mu_{t-1}$ with probability $\frac{1-\theta}{2}$,

$$\Psi_Y(\lambda) = \mathbb{E}e^{\lambda Y} \leq e^{\lambda\mathbb{E}Y} e^{\frac{\lambda^2}{2}\theta^2\mu_{t-1}^2} = e^{\lambda\left(\frac{1+\theta}{2}\theta\mu_{t-1}-\frac{1-\theta}{2}\theta\mu_{t-1}\right)} e^{\frac{\lambda^2}{2}\theta^2\mu_{t-1}^2}.$$

Thus

$$\begin{aligned}
&\mathbb{E}\left[e^{\lambda M_t(\ell_{T_{\leq t}(v)})}|\ell(v)=+\right] \\
&\leq e^{\frac{\lambda^2}{2}\sigma_1^2} \cdot e^{\lambda\mu_1} \cdot \left\{e^{\frac{\lambda^2\theta^2}{2}\sigma_{t-1}^2} \cdot e^{\lambda\theta^2\mu_{t-1}} \cdot e^{\frac{\lambda^2\theta^2}{2}\mu_{t-1}^2}\right\}^{(1-\delta)d} \\
&= e^{\lambda(\mu_1+\alpha\mu_{t-1})} \cdot e^{\frac{\lambda^2}{2}(\sigma_1^2+\alpha\sigma_{t-1}^2+\alpha\mu_{t-1}^2)} \\
&= e^{\lambda\mu_t} \cdot e^{\frac{\lambda^2}{2}\sigma_t^2}.
\end{aligned}$$

When $\ell(v) = -$, we have

$$
\begin{aligned}
&\mathbb{E}\left[e^{\lambda M_t(\ell_{T_{\leq t}(v)})} | \ell(v) = -\right] \\
&\leq e^{\frac{\lambda^2}{2}\sigma_1^2} \cdot e^{-\lambda \mu_1} \cdot \prod_{u \in \mathcal{C}^{\mathrm{u}}(v)} \mathbb{E}\left[e^{\lambda \theta M_{t-1}(\ell_{T_{\leq t-1}(u)})} | \ell(v) = -\right] \\
&\leq e^{\frac{\lambda^2}{2}\sigma_1^2} \cdot e^{-\lambda \mu_1} \cdot \left\{ e^{\frac{\lambda^2 \theta^2}{2}\sigma_{t-1}^2} \cdot \left\{ \frac{1+\theta}{2} e^{-\lambda \theta \mu_{t-1}} + \frac{1-\theta}{2} e^{\lambda \theta \mu_{t-1}} \right\} \right\}^{(1-\delta)d} \\
&\leq e^{\frac{\lambda^2}{2}\sigma_1^2} \cdot e^{-\lambda \mu_1} \cdot \left\{ e^{\frac{\lambda^2 \theta^2}{2}\sigma_{t-1}^2} \cdot e^{-\lambda \theta^2 \mu_{t-1}} \cdot e^{\frac{\lambda^2 \theta^2}{2}\mu_{t-1}^2} \right\}^{(1-\delta)d} \\
&= e^{-\lambda \mu_t} \cdot e^{\frac{\lambda^2}{2}\sigma_t^2}.
\end{aligned}
$$

This completes the proof. □

Proof of Theorem 2.2.3. Define the measure $\mu^+_{\ell_{T_{\leq t}(\rho)}}$ on the revealed labels, for a depth t tree rooted from ρ with label $\ell(\rho) = +$ (and similarly define $\mu^-_{\ell_{T_{\leq t}(\rho)}}$). We have the following recursion formula

$$
\mu^+_{\ell_{T_{\leq t}(\rho)}} = \left(\frac{1+\theta}{2}\right)^{N_{\mathcal{C}^l(\rho)}} \left(\frac{1-\theta}{2}\right)^{\delta d - N_{\mathcal{C}^l(\rho)}} \prod_{v \in \mathcal{C}^u(\rho)} \left[\frac{1+\theta}{2} \cdot \mu^+_{\ell_{\leq t-1}(v)} + \frac{1-\theta}{2} \cdot \mu^-_{\ell_{\leq t-1}(v)}\right].
$$

Recall that the χ^2 distance between two absolute continuous measures $\mu(x), \nu(x)$ is

$$
d_{\chi^2}(\mu, \nu) = \int \frac{\mu^2}{\nu} dx - 1,
$$

and we have the total variation distance between these two measures is upper bounded by the χ^2 distance

$$
d_{TV}\left(\mu^+_{\ell_{T_{\leq t}(\rho)}}, \mu^-_{\ell_{T_{\leq t}(\rho)}}\right) \leq \sqrt{d_{\chi^2}\left(\mu^+_{\ell_{T_{\leq t}(\rho)}}, \mu^-_{\ell_{T_{\leq t}(\rho)}}\right)}.
$$

Let us upper bound the symmetric version of χ^2 distance defined as

$$d_{\chi^2}^t := \max\left\{ d_{\chi^2}\left(\mu^+_{\ell_{T_{\leq t}(\rho)}}, \mu^-_{\ell_{T_{\leq t}(\rho)}}\right),\ d_{\chi^2}\left(\mu^-_{\ell_{T_{\leq t}(\rho)}}, \mu^+_{\ell_{T_{\leq t}(\rho)}}\right)\right\}.$$

Note that

$$\begin{aligned}
&d_{\chi^2}\left(\mu^+_{\ell_{T_{\leq t}(\rho)}}, \mu^-_{\ell_{T_{\leq t}(\rho)}}\right) \\
&= \left(1 + \frac{4\theta^2}{1-\theta^2}\right)^{\delta d} \cdot \left[1 + d_{\chi^2}\left(\frac{1+\theta}{2} \cdot \mu^+_{\ell_{\leq t-1}(v)} + \frac{1-\theta}{2} \cdot \mu^-_{\ell_{\leq t-1}(v)}, \frac{1+\theta}{2} \cdot \mu^-_{\ell_{\leq t-1}(v)} + \frac{1-\theta}{2} \cdot \mu^+_{\ell_{\leq t-1}(v)}\right)\right]^{(1-\delta)d} - 1,
\end{aligned}$$

and for the RHS, we have the expression

$$\begin{aligned}
&d_{\chi^2}\left(\frac{1+\theta}{2} \cdot \mu^+_{\ell_{\leq t-1}(v)} + \frac{1-\theta}{2} \cdot \mu^-_{\ell_{\leq t-1}(v)}, \frac{1+\theta}{2} \cdot \mu^-_{\ell_{\leq t-1}(v)} + \frac{1-\theta}{2} \cdot \mu^+_{\ell_{\leq t-1}(v)}\right) \\
&= \theta^2 \int \frac{(\mu^+_{\ell_{\leq t-1}(v)} - \mu^-_{\ell_{\leq t-1}(v)})^2}{\frac{1+\theta}{2} \cdot \mu^-_{\ell_{\leq t-1}(v)} + \frac{1-\theta}{2} \cdot \mu^+_{\ell_{\leq t-1}(v)}} dx.
\end{aligned}$$

Recalling Jensen's inequality, RHS of the above equation is further upper bounded by

$$\begin{aligned}
\text{RHS} &\leq \theta^2 \int (\mu^+_{\ell_{\leq t-1}(v)} - \mu^-_{\ell_{\leq t-1}(v)})^2 \cdot \left[\frac{1+\theta}{2} \cdot \frac{1}{\mu^-_{\ell_{\leq t-1}(v)}} + \frac{1-\theta}{2} \cdot \frac{1}{\mu^+_{\ell_{\leq t-1}(v)}}\right] dx \\
&= \theta^2 \left[\frac{1+\theta}{2} d_{\chi^2}\left(\mu^+_{\ell_{T_{\leq t}(\rho)}}, \mu^-_{\ell_{T_{\leq t}(\rho)}}\right) + \frac{1-\theta}{2} d_{\chi^2}\left(\mu^-_{\ell_{\leq t-1}(v)}, \mu^+_{\ell_{\leq t-1}(v)}\right)\right].
\end{aligned}$$

Thus

$$\begin{aligned}
&\max\left\{ d_{\chi^2}\left(\frac{1+\theta}{2} \cdot \mu^+_{\ell_{\leq t-1}(v)} + \frac{1-\theta}{2} \cdot \mu^-_{\ell_{\leq t-1}(v)}, \frac{1+\theta}{2} \cdot \mu^-_{\ell_{\leq t-1}(v)} + \frac{1-\theta}{2} \cdot \mu^+_{\ell_{\leq t-1}(v)}\right),\right. \\
&\qquad \left. d_{\chi^2}\left(\frac{1+\theta}{2} \cdot \mu^-_{\ell_{\leq t-1}(v)} + \frac{1-\theta}{2} \cdot \mu^+_{\ell_{\leq t-1}(v)}, \frac{1+\theta}{2} \cdot \mu^+_{\ell_{\leq t-1}(v)} + \frac{1-\theta}{2} \cdot \mu^-_{\ell_{\leq t-1}(v)}\right)\right\} \\
\leq &\theta^2 \max\left\{ d_{\chi^2}\left(\mu^+_{\ell_{\leq t-1}(v)}, \mu^-_{\ell_{\leq t-1}(v)}\right), d_{\chi^2}\left(\mu^-_{\ell_{\leq t-1}(v)}, \mu^+_{\ell_{\leq t-1}(v)}\right)\right\} = \theta^2 d_{\chi^2}^{t-1}.
\end{aligned}$$

Therefore, we have

$$\log\left(1 + d_{\chi^2}^t\right) \leq \delta d \cdot \log\left(1 + \frac{4\theta^2}{1-\theta^2}\right) + (1-\delta)d \cdot \log\left(1 + \theta^2 \cdot d_{\chi^2}^{t-1}\right).$$

If $(1-\delta)\theta^2 d < 1$, denote the fixed point of the above equation as c^* (the existence is manifested by the following bound (A.26)), i.e.,

$$\log(1+c^*) = \delta d \cdot \log\left(1 + \frac{4\theta^2}{1-\theta^2}\right) + (1-\delta)d \cdot \log(1+\theta^2 \cdot c^*).$$

Due to the fact that $x - \frac{1}{2}x^2 \le \log(1+x) \le x$, we have the following upper bound

$$c^* - \frac{1}{2}(c^*)^2 \le \log(1+c^*) = \delta d \cdot \log\left(1+\frac{4\theta^2}{1-\theta^2}\right) + (1-\delta)d \cdot \log(1+\theta^2 c^*) \tag{A.24}$$

$$\le \delta d \cdot \log\left(1+\frac{4\theta^2}{1-\theta^2}\right) + (1-\delta)\theta^2 d \cdot c^* \tag{A.25}$$

$$\text{and thus} \quad c^* \le \frac{\delta d \cdot \log\left(1+\frac{4\theta^2}{1-\theta^2}\right)}{1-(1-\delta)\theta^2 d} \cdot \frac{2}{1+\sqrt{1-2\frac{\delta d \cdot \log\left(1+\frac{4\theta^2}{1-\theta^2}\right)}{(1-(1-\delta)\theta^2 d)^2}}}. \tag{A.26}$$

If we have

$$d_{\chi^2}^{t-1} \le c^*$$

it is easy to see that

$$\begin{aligned}\log\left(1+d_{\chi^2}^t\right) &\le \delta d \cdot \log\left(1+\frac{4\theta^2}{1-\theta^2}\right) + (1-\delta)d \cdot \log\left(1+\theta^2 \cdot d_{\chi^2}^{t-1}\right)\\ &\le \delta d \cdot \log\left(1+\frac{4\theta^2}{1-\theta^2}\right) + (1-\delta)d \cdot \log(1+\theta^2 \cdot c^*) = \log(1+c^*),\end{aligned}$$

which implies $d_{\chi^2}^t \le c^*$. Therefore we only need to verify $d_{\chi^2}^1 \le c^*$, which is trivial. Thus we have the bound,

$$\limsup_{t\to\infty}\ d_{\chi^2}^t \le c^* \le 2\frac{\delta d \cdot \log\left(1+\frac{4\theta^2}{1-\theta^2}\right)}{1-(1-\delta)\theta^2 d},$$

provided $\frac{2\delta d\cdot\log\left(1+\frac{4\theta^2}{1-\theta^2}\right)}{(1-(1-\delta)\theta^2 d)^2} < 1$. So far we have proved

$$\limsup_{t\to\infty}\ d_{TV}^t \le \limsup_{t\to\infty}\ \left(d_{\chi^2}^t\right)^{1/2} \le \left\{\frac{2\delta d\cdot\log\left(1+\frac{4\theta^2}{1-\theta^2}\right)}{1-(1-\delta)\theta^2 d}\right\}^{1/2}.$$

Through Le Cam's Lemma, the error rate, for all local algorithms, is at least

$$\inf_{\Phi}\ \sup_{l\in\{+,-\}}\ \mathbb{P}_l(\Phi\neq l) \ge \frac{1-\left\{\frac{2\delta d\cdot\log\left(1+\frac{4\theta^2}{1-\theta^2}\right)}{1-(1-\delta)\theta^2 d}\right\}^{1/2}}{2}.$$

□

Proof of Theorem 2.2.5. For $\alpha > 1$:

Use induction analysis. For $t = 1$, the result follows from Hoeffding's lemma. Assume results hold for $t-1$, then if above the fraction label is l,

$$\begin{aligned}
&\mathbb{E}\left[e^{\lambda M_t(\ell_{T_{\le t}(v)})}|\ell(v)=l\right]\\
&\le e^{\frac{\lambda^2}{2}\sigma_1^2}\cdot e^{\lambda\mu_1}\cdot\prod_{u\in\mathcal{C}^{\mathrm{u}}(v)}\mathbb{E}\left[e^{\lambda\theta M_{t-1}(\ell_{T_{\le t-1}(u)})}|\ell(v)=l\right]\\
&= e^{\frac{\lambda^2}{2}\sigma_1^2}\cdot e^{\lambda\mu_1}\cdot\prod_{u\in\mathcal{C}^{\mathrm{u}}(v)}\left[\left(\theta+\frac{1-\theta}{k}\right)\cdot e^{\lambda\theta\mu_{t-1}}e^{\frac{\lambda^2\theta^2\sigma_{t-1}^2}{2}}+\frac{1-\theta}{k}\cdot e^{-\lambda\theta\mu_{t-1}}e^{\frac{\lambda^2\theta^2\sigma_{t-1}^2}{2}}\right.\\
&\qquad\qquad\qquad\qquad\qquad\qquad\left.+\frac{(k-2)(1-\theta)}{k}\cdot e^{\frac{\lambda^2\theta^2\sigma_{t-1}^2}{2}}\right]\\
&\le e^{\lambda(\mu_1+\alpha\mu_{t-1})}\cdot e^{\frac{\lambda^2}{2}\cdot\left(\sigma_1^2+(1-\delta)d\theta^2\sigma_{t-1}^2\right)}\cdot e^{\frac{\lambda^2}{2}\cdot(1-\delta)d\theta^2\mu_{t-1}^2}\\
&\le e^{\lambda(\mu_1+\alpha\mu_{t-1})}\cdot e^{\frac{\lambda^2}{2}\cdot\left(\sigma_1^2+(1-\delta)d\theta^2\sigma_{t-1}^2+(1-\delta)d\theta^2\mu_{t-1}^2\right)}
\end{aligned}$$

where the last step uses Hoeffding's Lemma A.2.2. When none of the labels is l, we have the following bound

$$
\begin{aligned}
&\mathbb{E}\left[e^{\lambda M_t(\ell_{T_{\le t}(v)})}|\ell(v)=l\right]\\
&\le e^{\frac{\lambda^2}{2}\sigma_1^2}\cdot\prod_{u\in\mathcal{C}^{\mathrm{u}}(v)}\left[\frac{1-\theta}{k}\cdot e^{\lambda\theta\mu_{t-1}}e^{\frac{\lambda^2\theta^2\sigma_{t-1}^2}{2}}+\frac{1-\theta}{k}\cdot e^{-\lambda\theta\mu_{t-1}}e^{\frac{\lambda^2\theta^2\sigma_{t-1}^2}{2}}\right.\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\left.+\left(\theta+\frac{(k-2)(1-\theta)}{k}\right)\cdot e^{\frac{\lambda^2\theta^2\sigma_{t-1}^2}{2}}\right]\\
&\le e^{\frac{\lambda^2}{2}\cdot\left(\sigma_1^2+(1-\delta)d\theta^2\sigma_{t-1}^2\right)}\cdot e^{\frac{\lambda^2}{2}\cdot(1-\delta)d\theta^2\mu_{t-1}^2}.
\end{aligned}
$$

Proof is completed. □

Proof of Theorem 2.2.6. Borrowing the idea from Proof A.2, we can study the following testing problem:

$$
\begin{aligned}
&d_{\chi^2}\left(\mu^{(i)}_{\ell_{T_{\le t}(\rho)}},\mu^{(j)}_{\ell_{T_{\le t}(\rho)}}\right)\\
&=\left(1+\theta^2\left(\frac{1}{\theta+\frac{1-\theta}{k}}+\frac{1}{\frac{1-\theta}{k}}\right)\right)^{\delta d}\left[1+d_{\chi^2}\left(\theta\mu^{(i)}_{\ell_{\le t-1}(v)}+(1-\theta)\bar{\mu}_{\ell_{\le t-1}(v)},\theta\mu^{(j)}_{\ell_{\le t-1}(v)}+(1-\theta)\bar{\mu}_{\ell_{\le t-1}(v)}\right)\right]^{(1-\delta)d}-1
\end{aligned}
$$

We know

$$
\begin{aligned}
&d_{\chi^2}\left(\theta\mu^{(i)}_{\ell_{\le t-1}(v)}+(1-\theta)\bar{\mu}_{\ell_{\le t-1}(v)},\theta\mu^{(j)}_{\ell_{\le t-1}(v)}+(1-\theta)\bar{\mu}_{\ell_{\le t-1}(v)}\right)\\
&=\int\frac{\theta^2(\mu^{(i)}_{\ell_{\le t-1}(v)}-\mu^{(j)}_{\ell_{\le t-1}(v)})^2}{\theta\mu^{(j)}_{\ell_{\le t-1}(v)}+(1-\theta)\bar{\mu}_{\ell_{\le t-1}(v)}}dx\\
&\le\theta^2\left[(\theta+\frac{1-\theta}{k})d_{\chi^2}\left(\mu^{(i)}_{\ell_{\le t-1}(v)},\mu^{(j)}_{\ell_{\le t-1}(v)}\right)+\frac{1-\theta}{k}d_{\chi^2}\left(\mu^{(j)}_{\ell_{\le t-1}(v)},\mu^{(i)}_{\ell_{\le t-1}(v)}\right)\right.\\
&\left.+\frac{1-\theta}{k}\sum_{l\in[k]\setminus\{i,j\}}2\left(d_{\chi^2}\left(\mu^{(i)}_{\ell_{\le t-1}(v)},\mu^{(l)}_{\ell_{\le t-1}(v)}\right)+d_{\chi^2}\left(\mu^{(j)}_{\ell_{\le t-1}(v)},\mu^{(l)}_{\ell_{\le t-1}(v)}\right)\right)\right]\\
&\le\theta^2(1+\frac{3(1-\theta)(k-2)}{k})\cdot d^{t-1}_{\chi^2}
\end{aligned}
$$

Thus define

$$
d^t_{\chi^2}:=\max_{i,j\in[k],i\neq j}d_{\chi^2}\left(\mu^{(i)}_{\ell_{T_{\le t}(\rho)}},\mu^{(j)}_{\ell_{T_{\le t}(\rho)}}\right)
$$

Then

$$\log(1+d_{\chi^2}^t) \leq \delta d \cdot \log\left(1+\theta^2\left(\frac{1}{\theta+\frac{1-\theta}{k}}+\frac{1}{\frac{1-\theta}{k}}\right)\right)+(1-\delta)d\cdot\log\left(1+\theta^2(1+\frac{3(1-\theta)(k-2)}{k})\cdot d_{\chi^2}^{t-1}\right)$$

Thus if

$$(1-\delta)\theta^2 d(1+\frac{3(1-\theta)(k-2)}{k})<1,$$

denote c^* as the fixed point of the equation

$$\log(1+c^*)=\delta d\cdot\log\left(1+\theta^2\left(\frac{1}{\theta+\frac{1-\theta}{k}}+\frac{1}{\frac{1-\theta}{k}}\right)\right)+(1-\delta)d\cdot\log\left(1+\theta^2(1+\frac{3(1-\theta)(k-2)}{k})\cdot c^*\right).$$

We have the following upper bounds for c^* via the fact that $x-\frac{1}{2}x^2<\log(1+x)<x$

$$c^*-\frac{1}{2}(c^*)^2 \leq \delta d\cdot\log\left(1+\theta^2\left(\frac{1}{\theta+\frac{1-\theta}{k}}+\frac{1}{\frac{1-\theta}{k}}\right)\right)+(1-\delta)\theta^2 d(1+\frac{3(1-\theta)(k-2)}{k})\cdot c^*.$$

The above equation implies $c^*<\frac{2\delta d\cdot\log\left(1+\theta^2\left(\frac{1}{\theta+\frac{1-\theta}{k}}+\frac{1}{\frac{1-\theta}{k}}\right)\right)}{1-(1-\delta)\theta^2 d(1+\frac{3(1-\theta)(k-2)}{k})}$, and

$$\log(1+d_{\chi^2}^t)\leq\frac{2\delta d\cdot\log\left(1+\theta^2\left(\frac{1}{\theta+\frac{1-\theta}{k}}+\frac{1}{\frac{1-\theta}{k}}\right)\right)}{1-(1-\delta)\theta^2 d(1+\frac{3(1-\theta)(k-2)}{k})}.$$

Invoke the following Lemma from Tsybakov (2009)'s Proposition 2.4.

Lemma A.2.3 (Tsybakov (2009), Proposition 2.4). *Let $P_0, P_1, \ldots, P_{k-1}$ be probability measures on $(\mathcal{X}, \mathcal{A})$ satisfying*

$$\frac{1}{k-1}\sum_{i=1}^{k-1} d_{\chi^2}(P_j, P_0) \leq (k-1)\cdot\alpha_*$$

then we have for any selector $\psi : \mathcal{X} \to [k]$

$$\max_{i \in [k]} P_i(\psi \neq i) \geq \frac{1}{2}[1 - \alpha_* - \frac{1}{k-1}]$$

Since we have $\frac{1}{k-1} \sum_{i \in [k] \setminus j} d_{\chi^2} \left(\mu^{(i)}_{\ell_{T_{\leq t}(\rho)}}, \mu^{(j)}_{\ell_{T_{\leq t}(\rho)}} \right) \leq \alpha \cdot (k-1)$, we apply Lemma A.2.3 and obtain

$$\inf_{\Phi} \sup_{l \in [k]} \mathbb{P}\left(\Phi \neq l\right) \geq \frac{1}{2}\left(1 - \alpha - \frac{1}{k-1}\right),$$

where $\alpha = \frac{\delta}{1-\delta} \cdot \frac{\mathsf{SNR}}{1 - 4 \cdot \mathsf{SNR}} \cdot \frac{2(p+q)(q+p/(k-1))}{pq}$.

□

A.3. Appendix for Section 2.3

Proof of Theorem 2.3.1. Denote the circulant matrix by B (it is B_π for any $\pi \in S_{n-1}$). The log-likelihood for WS model on symmetric matrix X (with diagonal elements being 0) is

$$\begin{aligned} \log \mathcal{L}_{n,k,\beta}(X|B) &= \log \frac{1 - \beta(1 - \beta\frac{k}{n-1})}{\beta(1 - \beta\frac{k}{n-1})} \cdot \langle X, B \rangle \\ &+ \log \frac{\beta\frac{k}{n-1}}{1 - \beta\frac{k}{n-1}} \cdot \langle X, J - I - B \rangle \\ &+ nk \log(\beta(1 - \beta\frac{k}{n-1})) + n(n-1-k)\log(1 - \beta\frac{k}{n-1}) \end{aligned}$$

For the Erdős-Rényi model, the log likelihood is

$$\log \mathcal{L}_{n,k}(X) = \log \frac{\frac{k}{n-1}}{1 - \frac{k}{n-1}} \cdot \langle X, J - I \rangle + n(n-1)\log(1 - \frac{k}{n-1}).$$

The Kullback-Leibler divergence between these two models is

$$
\begin{aligned}
&\mathrm{KL}(P_B||P_0) = \mathbb{E}_{X\sim P_B} \log \frac{P_B(X)}{P_0(X)} \\
&= \mathbb{E}_{X\sim P_B} \left\{ -\left(\log \frac{\frac{k}{n-1}}{1-\frac{k}{n-1}} - \log \frac{\beta\frac{k}{n-1}}{1-\beta\frac{k}{n-1}} \right) \cdot \langle X, J-I \rangle \right. \\
&\quad -n(n-1)\log(1-\frac{k}{n-1}) \\
&\quad + \left(\log \frac{1-\beta(1-\beta\frac{k}{n-1})}{\beta(1-\beta\frac{k}{n-1})} - \log \frac{\beta\frac{k}{n-1}}{1-\beta\frac{k}{n-1}} \right) \cdot \langle X, B \rangle \\
&\quad \left. +nk\log(\beta(1-\beta\frac{k}{n-1})) + n(n-1-k)\log(1-\beta\frac{k}{n-1}) \right\}
\end{aligned}
$$

which is equal to

$$
\begin{aligned}
& -\left(\log \frac{\frac{k}{n-1}}{1-\frac{k}{n-1}} - \log \frac{\beta\frac{k}{n-1}}{1-\beta\frac{k}{n-1}} \right) \cdot \\
&\qquad\qquad \left\langle (1-\beta)(1-\beta\frac{k}{n-1})B + \beta\frac{k}{n-1}(J-I), J-I \right\rangle \\
& + \left(\log \frac{1-\beta(1-\beta\frac{k}{n-1})}{\beta(1-\beta\frac{k}{n-1})} - \log \frac{\beta\frac{k}{n-1}}{1-\beta\frac{k}{n-1}} \right) \cdot \\
&\qquad\qquad \left\langle (1-\beta)(1-\beta\frac{k}{n-1})B + \beta\frac{k}{n-1}(J-I), B \right\rangle \\
& - n(n-1)\log(1-\frac{k}{n-1}) + nk\log(\beta(1-\beta\frac{k}{n-1})) \\
& + n(n-1-k)\log(1-\beta\frac{k}{n-1}) \\
= {} & n(n-1)\log\frac{1-\beta\frac{k}{n-1}}{1-\frac{k}{n-1}} - nk\log\frac{1}{\beta} \qquad (A.27) \\
& - \left[\log\frac{1}{\beta} + \log\frac{1-\beta\frac{k}{n-1}}{1-\frac{k}{n-1}} \right] nk \left[1-(1-\beta)\beta\frac{k}{n-1} \right] \\
& + \left[\log\frac{1}{\beta} + \log\frac{1-\beta(1-\beta\frac{k}{n-1})}{\beta\frac{k}{n-1}} \right] nk \left[1-\beta(1-\beta\frac{k}{n-1}) \right] \\
= {} & -\log\frac{1}{\beta}\cdot nk \left[1+\beta-\beta\frac{k}{n-1} \right] \\
& + \log\frac{1-\beta\frac{k}{n-1}}{1-\frac{k}{n-1}} \cdot n \left[(n-1-k) + (1-\beta)\beta\frac{k^2}{n-1} \right] \\
& + \log\frac{1-\beta(1-\beta\frac{k}{n-1})}{\beta\frac{k}{n-1}} \cdot nk \left[1-\beta(1-\beta\frac{k}{n-1}) \right]. \qquad (A.28)
\end{aligned}
$$

Via the inequality $\log(1+x) < x$ for all $x > -1$, we can further simplify the above expression as

$$\begin{aligned}
&\mathrm{KL}(P_B||P_0)\\
&\leq nk(1-\beta)\left[-\beta+\beta\frac{k}{n-1}+(1-\beta)\beta\frac{k^2}{n(n-1-k)}\right]\\
&\quad+\frac{(1-\beta)(1-\beta\frac{k}{n-1})}{\beta\frac{k}{n-1}}nk\left[(1-\beta)+\beta^2\frac{k}{n-1}\right]\\
&\leq nk(1-\beta)\left[(1-\beta)\beta\frac{k}{n-1}+(1-\beta)\beta\frac{k^2}{n(n-1-k)}\right]\\
&\quad+\frac{(1-\beta)^2(1-\beta\frac{k}{n-1})}{\beta}n(n-1)\leq C\cdot n^2(1-\beta)^2, \qquad (A.29)
\end{aligned}$$

where $0 < C < \frac{1}{2}\frac{k^2}{n(n-1)}+\frac{1}{\beta}$ is some universal constant (note we are interested in the case when β is close to 1).

When $k \preceq n^{1/2}$, the above bound can be further strengthened, in the following sense (recall equation (A.28)):

$$\begin{aligned}
&\mathrm{KL}(P_B||P_0)\\
&\leq nk(1-\beta)\left[-\beta+\beta\frac{k}{n-1}+(1-\beta)\beta\frac{k^2}{n(n-1-k)}\right]\\
&+\log\frac{1-\beta(1-\beta\frac{k}{n-1})}{\beta\frac{k}{n-1}}\cdot nk\left[1-\beta(1-\beta\frac{k}{n-1})\right]\\
&\leq\left\{\log\frac{1-\beta(1-\beta\frac{k}{n-1})}{\beta\frac{k}{n-1}}\cdot\frac{1-\beta(1-\beta\frac{k}{n-1})}{\beta\frac{k}{n-1}}\right\}\cdot k^2\beta\frac{n}{n-1}.
\end{aligned}$$

Denote $t := \frac{1-\beta(1-\beta\frac{k}{n-1})}{\beta\frac{k}{n-1}} = \frac{1-\beta}{\beta}\frac{n-1}{k}+\beta$. Thus we have

$$\mathrm{KL}(P_B||P_0)\leq t\log t\cdot k^2\beta\frac{n}{n-1}. \qquad (A.30)$$

Suppose for some constant $\alpha_* > 0$, and $\alpha = \alpha_* \cdot \frac{1}{\beta}(1-\frac{1}{n})^2$, we have the following

$$t \leq \alpha \frac{n \log \frac{n}{e}}{k^2} \cdot \frac{1}{\log \alpha \frac{n \log \frac{n}{e}}{k^2}} \tag{A.31}$$

$$\text{and} \quad t \log t \leq \alpha \frac{n \log \frac{n}{e}}{k^2} \cdot \left(1 - \frac{\log \log \alpha \frac{n \log \frac{n}{e}}{k^2}}{\log \alpha \frac{n \log n}{k^2}}\right) < \alpha \frac{n \log \frac{n}{e}}{k^2}. \tag{A.32}$$

Plugging in the expression for t into (A.31), if

$$\begin{aligned} \frac{1-\beta}{\beta} &\leq \alpha(1 + \frac{1}{n-1}) \cdot \frac{\log \frac{n}{e}}{k} \cdot \frac{1}{\log \alpha \frac{n \log \frac{n}{e}}{k^2}} - \frac{k}{n-1} \\ &\asymp \frac{\log n}{k} \frac{1}{\log \frac{n \log \frac{n}{e}}{k^2}} \end{aligned} \tag{A.33}$$

we have

$$t \leq \alpha \frac{n \log \frac{n}{e}}{k^2} \cdot \frac{1}{\log \alpha \frac{n \log \frac{n}{e}}{k^2}} \quad \Rightarrow \quad t \log t < \alpha \frac{n \log \frac{n}{e}}{k^2}$$

which further implies, via Equation (A.29),

$$\begin{aligned} \frac{1}{(n-1)!} \sum_{\pi \in S_{n-1}} \mathrm{KL}(P_{B_\pi} || P_0) &\leq t \log t \cdot k^2 \beta \frac{n}{n-1} \\ &\leq \alpha_* \cdot \log(n-1)!. \end{aligned}$$

Recalling the bound on KL-divergence, if

$$1 - \beta \leq \sqrt{\frac{\alpha_*}{C} \cdot \frac{(n-1) \log \frac{n}{e}}{n^2}} \asymp \sqrt{\frac{\log n}{n}} \tag{A.34}$$

we have

$$\frac{1}{(n-1)!} \sum_{\pi \in S_{n-1}} \mathrm{KL}(P_{B_\pi} || P_0) \leq n^2 (1-\beta)^2 \leq \alpha_* \cdot \log(n-1)!.$$

We invoke the following Lemma on minimax error through Kullbak-Leibler divergence.

Lemma A.3.1 (Tsybakov (2009), Proposition 2.3)**.** *Let P_0, P_1,..., P_M be probability mea-*

sures on $(\mathcal{X}, \mathcal{A})$ *satisfying*

$$\frac{1}{M}\sum_{j=1}^{M} \mathrm{KL}(P_j||P_0) \leq \alpha \cdot \log M \tag{A.35}$$

with $0 < \alpha < \frac{1}{8}$*. Then for any* $\psi : \mathcal{X} \to [M+1]$

$$\max \left\{ P_0(\psi \neq 0), \frac{1}{M}\sum_{j=1}^{M} P_j(\psi \neq j) \right\} \\ \geq \frac{\sqrt{M}}{\sqrt{M}+1}\left(1 - 2\alpha - \sqrt{\frac{2\alpha}{\log M}}\right).$$

Hence, if either one of the conditions in Equations (A.33) and (A.34) holds, we have

$$\frac{1}{(n-1)!}\sum_{\pi \in S_{n-1}} \mathrm{KL}(P_{B_\pi}||P_0) \leq \alpha_* \cdot \log(n-1)!. \tag{A.36}$$

Putting everything together, Equation (A.36) holds whenever if

$$1 - \beta \prec \sqrt{\frac{\log n}{n}} \vee \frac{\log n}{k}.$$

Applying Lemma A.3.1, we complete the proof:

$$\underline{\lim}_{n\to\infty} \ \min_{\phi} \ \max \left\{ P_0(\phi \neq 0), \ \frac{1}{(n-1)!}\sum_{i=1}^{(n-1)!} P_i(\phi \neq i) \right\} \\ \geq \underline{\lim}_{n\to\infty} \frac{\sqrt{(n-1)!}}{1+\sqrt{(n-1)!}}\left(1 - 2\alpha - \sqrt{\frac{2\alpha}{\log(n-1)!}}\right) = 1 - 2\alpha.$$

□

Proof of Lemma 2.3.1. Let us state the well-known Bernstein's inequality (Boucheron et al. (2013), Theorem 2.10), which will be used in the proof of this lemma.

Lemma A.3.2 (Bernstein's inequality)**.** *Let* $X_1, \ldots, X_n$ *be independent bounded real-valued* *random variables. Assume that there exist positive numbers v and c such that*

$$\begin{aligned}&\sum_{i=1}^{n}\mathbb{E}[X_i^2] \le v,\\ &X_i \le 3c, \forall 1 \le i \le n \quad a.s.\end{aligned}$$

then we have, for all $t > 0$,

$$\mathbb{P}\left(\sum_{i=1}^{n}(X_i - \mathbb{E}X_i) \ge \sqrt{2vt} + ct\right) \le e^{-t}. \tag{A.37}$$

First, let us consider the case when the adjacency matrix A is generated from the Erdős-Rényi random graph $\mathsf{ER}(n, \frac{k}{n-1})$. For any P_π with $\pi \in S_{n-1}$, we know $\langle P_\pi B P_\pi^T, A\rangle$ has the same distribution as $\langle B, A\rangle$. Thus, in view of Lemma A.3.2,

$$\begin{aligned}\langle P_\pi B P_\pi^T, A\rangle \overset{\text{in law}}{=\!=} \langle B, A\rangle &= 2\sum_{i>j} A_{ij}B_{ij}\\ &= 2\sum_{i>j}\mathbb{E}[A_{ij}]B_{ij} + 2\sum_{i>j}(A_{ij} - \mathbb{E}[A_{ij}])B_{ij}\\ &\le \frac{k}{n-1}nk + 2\sqrt{\frac{k}{n-1}nkt} + \frac{2}{3}t\end{aligned}$$

with probability at least $1 - \exp(-t)$. Indeed, there are $nk/2$ non-zero $B_{i,j}, i > j$, and it is clear that $A_{ij} \sim \mathsf{Bernoulli}(\frac{k}{n-1})$ and $2\sum_{i>j}\mathbb{E}[A_{ij}]B_{ij} = nk\frac{k}{n-1}$, implying the choice of $c = \frac{1}{3}$ and

$$v = \sum_{i<j}\mathbb{E}[(A_{ij}B_{ij})^2] = \sum_{i<j}\mathbb{E}[A_{ij}^2]B_{ij} = \frac{nk}{2}\frac{k}{n-1}$$

in Lemma A.3.2. Via the union bound, taking $t = \log n!$, we have

$$\begin{aligned}&\max_{P_\pi}\ \langle P_\pi B P_\pi^T, A\rangle\\ &\le \frac{k}{n-1}nk + 2\sqrt{\frac{k}{n-1}nk \cdot \log n!} + \frac{2}{3}\cdot \log n!\end{aligned}$$

with probability at least $1-(n-1)!\exp(-\log n!) = 1-\frac{1}{n}$.

Alternatively, suppose A is from the small-world rewiring model $\mathsf{WS}(n,k,\beta)$, with permutation being identity. With probability at least $1-\exp(-\log n) = 1-\frac{1}{n}$,

$$\begin{aligned}\max_{P_\pi}\ \langle P_\pi B P_\pi^T, A\rangle &\geq \langle B, A\rangle \\ &= \langle B, \mathbb{E}[A]\rangle + \langle B, A-\mathbb{E}[A]\rangle \\ &\geq (1-\beta+\beta^2\frac{k}{n-1})nk - \sqrt{nk\cdot\log n}\end{aligned}$$

where the last step follows via Hoeffding's inequality: it is clear that for (i,j) with $B_{ij}\neq 0$,

$$\mathbb{E}[A_{ij}] = 1-\beta+\beta^2\frac{k}{n-1},$$

and $0\leq A_{ij}\leq 1$ almost surely.

Therefore if there exist a threshold $T>0$ such that

$$\begin{aligned}(1-\beta+\beta^2\frac{k}{n-1})nk - \sqrt{nk\cdot\log n} &> T \\ \text{and } T > \frac{k}{n-1}nk + 2\sqrt{\frac{k}{n-1}nk\cdot\log n!} &+ \frac{2}{3}\cdot\log n!\end{aligned} \tag{A.38}$$

we have that

$$\begin{aligned}&\lim_{n,k(n)\to\infty}\ \max\left\{P_0(\phi_1\neq 0),\ \frac{1}{(n-1)!}\sum_{i=1}^{(n-1)!}P_i(\phi_1\neq 1)\right\}\\ &\leq \lim_{n,k(n)\to\infty}\ \frac{1}{n} = 0.\end{aligned}$$

The detailed calculation of Equation (A.38) yields that the test succeeds with high probability whenever

$$1-\beta \succeq \sqrt{\frac{\log n}{n}} \vee \frac{\log n}{k}.$$

□

Proof of Lemma 2.3.2. Under the model $\mathsf{WS}(n, k, \beta)$ with permutation P_π,

$$A = (1-\beta)(1-\beta\frac{k}{n-1}) \cdot P_\pi^T B P_\pi + \beta\frac{k}{n-1} \cdot (J-I) + Z$$

where $J = 11^T \in \mathbb{R}^{n\times n}$, B is the ring structured signal matrix defined in Equation (2.23), and Z is a zero-mean noise random matrix.

We first study the random fluctuation part, $Z = A - \mathbb{E}A$. Let us bound the expectation $\mathbb{E}\|A - \mathbb{E}A\|$ as the first step, for any adjacency matrix $A \in \mathbb{R}^{n\times n}$ using the symmetrization trick. Denote $A' \sim A$ as the independent copy of A sharing the same distribution. Take $E, G \in \mathbb{R}^{n\times n}$ as random symmetric Rademacher and Gaussian matrices with entries E_{ij}, G_{ij} being, respectively, independent Rademacher and Gaussian. Denoting matrix Hadamard product as $A \circ B$, we have

$$\begin{aligned}
&\mathbb{E}\|A - \mathbb{E}A\| = \mathbb{E} \sup_{\|v\|_{\ell_2}=1} \langle (A - \mathbb{E}A)v, v\rangle \\
&= \mathbb{E} \sup_{\|v\|_{\ell_2}=1} \langle (A - \mathbb{E}_{A'}A')v, v\rangle \le \mathbb{E}_A\mathbb{E}_{A'} \sup_{\|v\|_{\ell_2}=1} \langle (A - A')v, v\rangle \\
&= \mathbb{E}_E\mathbb{E}_A\mathbb{E}_{A'} \sup_{\|v\|_{\ell_2}=1} \langle [E \circ (A - A')]v, v\rangle \\
&\le \mathbb{E}_A\mathbb{E}_E \sup_{\|v\|_{\ell_2}=1} \langle [E \circ A]v, v\rangle + \mathbb{E}_{A'}\mathbb{E}_E \sup_{\|v\|_{\ell_2}=1} \langle [-E \circ A']v, v\rangle \\
&= 2\mathbb{E}_A\mathbb{E}_E \sup_{\|v\|_{\ell_2}=1} \langle [E \circ A]v, v\rangle \\
&\le \frac{2}{\sqrt{2/\pi}} \cdot \mathbb{E}_A\mathbb{E}_E \sup_{\|v\|_{\ell_2}=1} \langle [\mathbb{E}_G[|G|] \circ E \circ A]v, v\rangle \\
&\le \sqrt{\frac{\pi}{2}} \cdot \mathbb{E}_A\mathbb{E}_E\mathbb{E}_G \sup_{\|v\|_{\ell_2}=1} \langle [|G| \circ E \circ A]v, v\rangle \\
&= \sqrt{\frac{\pi}{2}} \cdot \mathbb{E}_A\mathbb{E}_G \sup_{\|v\|_{\ell_2}=1} \langle [G \circ A]v, v\rangle \\
&= \sqrt{\frac{\pi}{2}} \cdot \mathbb{E}_A \left(\mathbb{E}_G\|G \circ A\|\right).
\end{aligned}$$

Recall the following Lemma from (Bandeira and van Handel, 2014).

Lemma A.3.3 (Bandeira and van Handel (2014), Theorem 1.1)**.** *Let X be the $n \times n$ symmetric random matrix with $X = G \circ A$, where $G_{ij}, i < j$ are i.i.d. $N(0,1)$ and A_{ij} are given scalars. Then*

$$\mathbb{E}_G\|X\| \precsim \max_i \sqrt{\sum_j A_{ij}^2} + \max_{ij} |A_{ij}| \cdot \sqrt{\log n}.$$

Thus via Jensen's inequality and the above Lemma, we upper bound

$$\begin{aligned}
&\mathbb{E}\|A - \mathbb{E}A\| \leq \sqrt{\frac{\pi}{2}} \cdot \mathbb{E}_A \left(\mathbb{E}_G \|G \circ A\|\right) \\
&\precsim \mathbb{E}_A \left[\max_i \sqrt{\sum_j A_{ij}^2} + \max_{ij} |A_{ij}| \cdot \sqrt{\log n} \right] \\
&\leq \sqrt{\mathbb{E}_A \max_i \sum_j A_{ij}^2} + \sqrt{\log n} \\
&\leq \sqrt{k + C_1 2\sqrt{k \log n} + C_2 \log n} + \sqrt{\log n} \asymp \sqrt{k} \vee \sqrt{\log n},
\end{aligned}$$

where the last step uses Bernstein inequality Lemma A.3.2. Moving from expectation $\mathbb{E}\|A - \mathbb{E}A\|$ to concentration on $\|A - \mathbb{E}A\|$ is through Talagrand's concentration inequality (see, Talagrand (1996b) and Tao (2012) Theorem 2.1.13), since $\|\cdot\|$ is $1-$Lipschitz convex function in our case (and the entries are bounded), thus with probability at least $1 - \frac{1}{n}$,

$$\|A - \mathbb{E}A\| \leq \mathbb{E}\|A - \mathbb{E}A\| + C \cdot \sqrt{\log n} \asymp \sqrt{k} \vee \sqrt{\log n}.$$

Now let us study the structural signal part. Matrix B is of the form circulant matrix, the associated polynomial is

$$f(x) = (x + x^{n-k/2}) \cdot \frac{x^{k/2} - 1}{x - 1}.$$

The eigenvectors can be analytically calculated: collect for all $j = 0, 1, ..., n/2$

$$(\cos 0, \cos \frac{2\pi j}{n}, \cos \frac{2\pi 2j}{n}, \dots, \cos \frac{2\pi nj}{n})$$

and

$$(\sin 0, \cos \frac{2\pi j}{n}, \sin \frac{2\pi 2j}{n}, \dots, \sin \frac{2\pi nj}{n})$$

and the corresponding eigenvalues are

$$\lambda_j = f(w_j) = 2 \sum_{i=1}^{k/2} \cos \left(i \frac{2\pi j}{n} \right).$$

Let us first assume $\frac{k}{n} \leq \frac{1}{2}$, thus the second largest eigenvalue

$$\lambda_2 = 2 \sum_{i=1}^{k/2} \cos \left(i \frac{2\pi}{n} \right) = \frac{2 \sin \frac{k\pi}{2n}}{\sin \frac{\pi}{n}} \cos \frac{(k+2)\pi}{2n} \asymp k.$$

Now if there exist a threshold $T > 0$ such that w.h.p., the second eigenvalue of the adjacency matrix generated from WS model A_{WS} separates from that of the adjacency matrix generated from ER model A_{ER} in the following sense

$$\lambda_2(A_{\mathsf{WS}}) > T > \lambda_2(A_{\mathsf{ER}}),$$

we have

$$\lim_{n,k(n)\to\infty} \max \left\{ P_0(\phi_2 \neq 0), \ \frac{1}{(n-1)!} \sum_{i=1}^{(n-1)!} P_i(\phi_2 \neq 1) \right\} = 0.$$

Using Weyl's interlacing inequality, we have

$$\begin{aligned} \lambda_2(A_{\mathsf{WS}}) &\geq \lambda_2(\mathbb{E}[A_{\mathsf{WS}}]) - \|Z\| \\ &\geq (1-\beta)(1-\beta\frac{k}{n-1})\lambda_2 - \sqrt{k} \vee \sqrt{\log n}, \end{aligned}$$

while

$$\lambda_2(A_{\mathsf{ER}}) \le \sqrt{k} \vee \sqrt{\log n}.$$

Therefore, we have the condition for which the second eigenvalue test succeeds:

$$(1-\beta)(1-\beta\frac{k}{n-1})\lambda_2 > \sqrt{k} \vee \sqrt{\log n}$$

which means

$$(1-\beta)(1-\beta\frac{k}{n-1}) > \frac{\sqrt{k} \vee \sqrt{\log n}}{\frac{2\sin\frac{k\pi}{2n}}{\sin\frac{\pi}{n}}\cos\frac{(k+2)\pi}{2n}} \asymp \sqrt{\frac{1}{k}} \vee \frac{\sqrt{\log n}}{k}.$$

□

Proof of Lemma 2.3.3. Take any two rows $A_{i\cdot}, A_{j\cdot}$ of the adjacency matrix. Define the distance $x = |\pi^{-1}(i) - \pi^{-1}(j)|_{\text{ring}}$. Equivalently, the Hamming distance of the corresponding signal vectors satisfies $\mathrm{H}(B_{i\cdot}, B_{j\cdot}) = 2x$. Therefore the union of signal nodes for i,j-th row is of cardinality $|S_i \cup S_j| = k + x$, common signal nodes are of cardinality $|S_i \cap S_j| = k - x$, unique signal nodes are of cardinality $|S_i \triangle S_j| = 2x$, and $|S_i^c \cap S_j^c| = n - k - x - 2$. Each signal coordinate is 1 with probability $p = 1 - \beta(1 - \beta\frac{k}{n-1})$, while non-signal coordinate is 1 with probability $q = \beta\frac{k}{n-1}$, and we have

$$\langle A_{i\cdot}, A_{j\cdot}\rangle = \sum_{l \in S_i \cap S_j} A_{il}A_{jl} + \sum_{l \in S_i \triangle S_j} A_{il}A_{jl} + \sum_{l \in S_i^c \cap S_j^c} A_{il}A_{jl}.$$

Observe as long as $l \neq i, j$, A_{il} and A_{jl} are independent, and $\{A_{il}A_{jl}, l \in [n]\backslash\{i,j\}\}$ are independent of each other.

Let us bound each term via Bernstein's inequality Lemma A.3.2,

$$\sum_{l \in S_i \cap S_j} A_{il} A_{jl} \in p^2 |S_i \cap S_j| \pm \left(\sqrt{2p^2 |S_i \cap S_j| t} + \frac{1}{3} t \right)$$
$$\sum_{l \in S_i \triangle S_j} A_{il} A_{jl} \in pq |S_i \triangle S_j| \pm \left(\sqrt{2pq |S_i \triangle S_j| t} + \frac{1}{3} t \right)$$
$$\sum_{l \in S_i^c \cap S_j^c} A_{il} A_{jl} \in q^2 |S_i^c \cap S_j^c| \pm \left(\sqrt{2q^2 |S_i^c \cap S_j^c| t} + \frac{1}{3} t \right)$$

with probability at least $1 - 6\exp(-t)$. We take $t = (2+\epsilon)\log n$ for any $\epsilon > 0$, such that with probability at least $1 - Cn^{-\epsilon}$, the above bound holds for all pairs (i, j).

Thus for all $|\pi^{-1}(i) - \pi^{-1}(j)|_{\text{ring}} > k$ pairs,

$$\begin{aligned} \langle A_{i\cdot}, A_{j\cdot} \rangle &\le 2kpq + (n-2k-2)q^2 \\ &\quad + \left(\sqrt{4kpqt} + \sqrt{2(n-2k-2)q^2 t} + t \right), \end{aligned}$$

for $|\pi^{-1}(i) - \pi^{-1}(j)|_{\text{ring}} \le x$ pairs

$$\begin{aligned} &\langle A_{i\cdot}, A_{j\cdot} \rangle \ge (k-x)p^2 + 2xpq + (n-k-x-2)q^2 \\ &- \left(\sqrt{2(k-x)p^2 t} + \sqrt{4xpqt} + \sqrt{2(n-k-x-2)q^2 t} + t \right). \end{aligned}$$

Thus, with $t = (2+\epsilon)\log n$, $p = 1 - \beta(1 - \beta\frac{k}{n-1})$ and $q = \beta\frac{k}{n-1}$, if $x < x_0$ with

$$\frac{x_0}{k} := 1 - C_1 \sqrt{\frac{\log n}{k}} \frac{1}{1-\beta} - C_2 \sqrt{\frac{\log n}{n}} \frac{1}{(1-\beta)^2},$$

we have

$$
\begin{aligned}
(k-x)(p-q)^2 &\geq 2t + (2\sqrt{2}+1)\left(\sqrt{kp^2}+\sqrt{nq^2}\right)\sqrt{2t} \\
&\geq 2t + \Big(\sqrt{2kpq} + \sqrt{(n-2k-2)q^2} + \sqrt{(k-x)p^2} \\
&\qquad + \sqrt{2xpq} + \sqrt{(n-k-x-2)q^2}\Big)\sqrt{2t},
\end{aligned}
$$

which further implies,

$$
\min_{j:|\pi^{-1}(i)-\pi^{-1}(j)|_{\text{ring}}\leq x_0} \langle A_{i\cdot}, A_{j\cdot}\rangle \geq \max_{j\notin\mathcal{N}(v_i)} \langle A_{i\cdot}, A_{j\cdot}\rangle, \forall i
$$

$$
\begin{aligned}
\max_{i\in[n]} \frac{|\hat{\mathcal{N}}(v_i)\triangle\mathcal{N}(v_i)|}{|\mathcal{N}(v_i)|} &\leq \frac{k-x_0}{k} \\
&= C_1\sqrt{\frac{\log n}{k}}\frac{1}{1-\beta} + C_2\sqrt{\frac{\log n}{n}}\frac{1}{(1-\beta)^2}.
\end{aligned}
$$

Therefore we can reconstruct the neighborhood consistently, under the condition

$$
1-\beta \succ \sqrt{\frac{\log n}{k}} \vee \left(\frac{\log n}{n}\right)^{1/4}.
$$

□

Proof of Lemma 2.3.4. Since eigen structure is not affected by permutation, we will work under the case when the true permutation is identity. We work under a mild technical assumption that we have two independent observation of the adjacency matrix, one used for calculating the eigen-vector, the other used for projection to reduce dependency. Note this technical assumption only affect the signal $(1-\beta)$ to noise (k/n) ratio by a constant factor. Recall that $A = M + Z$, where $M = (1-\beta)(1-\beta\frac{k}{n-1})\cdot B + \beta\frac{k}{n-1}\cdot(J-I)$ is the signal matrix. Denote the eigenvectors of M to be $U \in \mathbb{R}^{n\times n}$, and eigenvectors of A to be $\hat{U} \in \mathbb{R}^{n\times n}$. Denote the projection matrix corresponding the subspace of the second and third eigenvector $U_{\cdot 2}, U_{\cdot 3}$ to be H. Similarly $\hat{H}$ denotes the projection matrix to the 2-dim space spanned by $\hat{U}_{\cdot 2}, \hat{U}_{\cdot 3}$.

Classic Davis-Kahan perturbation bound informs us that two dimensional subspace $\hat{H}$ and H are close in spectral norm

$$\|\hat{H} - H\| \leq \frac{\|Z\|}{\Delta\lambda - \|Z\|},$$

where the spectral gap $\Delta\lambda$ of M is

$$\begin{aligned}
\Delta\lambda &:= (1-\beta)(1-\beta\frac{k}{n-1}) \cdot (\lambda_2 - \lambda_3) \\
&= (1-\beta)(1-\beta\frac{k}{n-1}) \cdot \left[2\sum_{i=1}^{k/2} \cos\left(i\frac{2\pi}{n}\right) - 2\sum_{i=1}^{k/2} \cos\left(i\frac{2\pi \cdot 2}{n}\right) \right] \\
&= (1-\beta)(1-\beta\frac{k}{n-1}) \left[\frac{2\sin\frac{k\pi}{2n}}{\sin\frac{\pi}{n}} \cos\frac{(k+2)\pi}{2n} - \frac{2\sin\frac{k\pi}{n}}{\sin\frac{2\pi}{n}} \cos\frac{(k+2)\pi}{n} \right] \\
&\asymp (1-\beta)(1-\beta\frac{k}{n-1})\frac{k^3}{n^2}.
\end{aligned}$$

From the proof of Lemma 2.3.2, we know with high probability

$$\|Z\| \preceq \sqrt{k} \vee \sqrt{\log n}.$$

Note we have for the true signal matrix M and true projection H

$$\begin{aligned}
HM_{\cdot i} &= \langle U_{\cdot 2}, M_{\cdot i} \rangle \cdot U_{\cdot 2} + \langle U_{\cdot 3}, M_{\cdot i} \rangle \cdot U_{\cdot 3}, \\
&= \frac{(1-\beta)\lambda_2}{\sqrt{n}} \cos\frac{(i-1)2\pi}{n} \cdot U_{\cdot 2} + \frac{(1-\beta)\lambda_2}{\sqrt{n}} \sin\frac{(i-1)2\pi}{n} \cdot U_{\cdot 3};
\end{aligned} \tag{A.39}$$

however, one only observes the noisy version $\hat{H}A_{\cdot i} \in \mathbb{R}^n$ (of the signal $HM_{\cdot i} \in \mathbb{R}^n$), which satisfies the equality

$$\hat{H}A_{\cdot i} = HM_{\cdot i} + (\hat{H} - H)M_{\cdot i} + \hat{H}Z_{\cdot i}.$$

Hence we have, uniformly for all i,

$$\begin{aligned}\|\hat{H}A_{\cdot i} - HM_{\cdot i}\| &\le \|(\hat{H} - H)M_{\cdot i}\| + \|\hat{H}Z_{\cdot i}\| \\ &\le \|\hat{H} - H\|\|M_{\cdot i}\| + \|\hat{H}Z_{\cdot i}\| \\ &\le \frac{\sqrt{k} \vee \sqrt{\log n}}{\Delta\lambda - \sqrt{k} \vee \sqrt{\log n}} \cdot \sqrt{k}(1-\beta) + C\sqrt{\log n}\end{aligned}$$

with probability $1-n^{-c}$, with some constants $c, C > 0$. Here the last line follows from Davis-Kahan bound on $\|\hat{H} - H\|$ and Azuma-Hoeffding's inequality for $\langle \hat{U}_{\cdot 2}, Z_{\cdot i}\rangle$ and $\langle \hat{U}_{\cdot 3}, Z_{\cdot i}\rangle$ condition on $\hat{H}$. Denote this stochastic error as

$$\begin{aligned}\delta &:= \frac{\sqrt{k} \vee \sqrt{\log n}}{\Delta\lambda - \sqrt{k} \vee \sqrt{\log n}} \cdot \sqrt{k}(1-\beta) + C\sqrt{\log n}, \\ &\asymp \frac{k}{\frac{k^3}{n^2}} = \frac{n^2}{k^2}.\end{aligned}$$

The second line follows under the condition $1-\beta \gtrsim \frac{n^2}{k^{2.5}}$, which is ensured under Eq. (A.40).

For any i, j with $|j-i|_{\text{ring}} = m$, Eq. (A.39) together with simple geometry implies

$$\begin{aligned}&\|HM_{\cdot i} - HM_{\cdot j}\| \\ &= \frac{(1-\beta)\lambda_2}{\sqrt{n}} \cdot \left[\left(\cos\frac{(i-1)2\pi}{n} - \cos\frac{(j-1)2\pi}{n}\right)^2\right. \\ &\quad \left. + \left(\sin\frac{(i-1)2\pi}{n} - \sin\frac{(j-1)2\pi}{n}\right)^2\right]^{1/2} \\ &= \frac{(1-\beta)\lambda_2}{\sqrt{n}} \cdot 2\sin\frac{m\pi}{n}.\end{aligned}$$

Therefore, fix any i, for $j \notin \mathcal{N}(v_i)$ not in i's neighborhood, using triangle inequality we have

$$\begin{aligned}
&\min_{j \notin \mathcal{N}(v_i)} \|\hat{H}A_{\cdot i} - \hat{H}A_{\cdot j}\| \geq \min_{j \notin \mathcal{N}(v_i)} \|HM_{\cdot i} - HM_{\cdot j}\| - 2\delta \\
&\geq \frac{(1-\beta)\lambda_2}{\sqrt{n}} \cdot 2\sin\frac{k\pi}{n} - 2\delta \\
&= \left(\frac{(1-\beta)\lambda_2}{\sqrt{n}} \cdot 2\sin\frac{k\pi}{n} - 4\delta\right) + 2\delta \\
&\geq \max_{|j-i|_{\text{ring}} < m} \|\hat{H}A_{\cdot i} - \hat{H}A_{\cdot j}\|
\end{aligned}$$

with

$$m = \frac{n}{\pi}\arcsin\left(\sin\frac{k\pi}{n} - 2\delta\frac{\sqrt{n}}{\lambda_2}\frac{1}{1-\beta}\right).$$

Therefore the following bound on symmetric set difference holds

$$\begin{aligned}
\max_{i\in[n]} \frac{|\hat{\mathcal{N}}(v_i) \triangle \mathcal{N}(v_i)|}{|\mathcal{N}(v_i)|} &\leq 1 - \frac{\arcsin\left(\sin\frac{k\pi}{n} - 2\delta\frac{\sqrt{n}}{\lambda_2}\frac{1}{1-\beta}\right)}{\frac{k\pi}{n}} \\
&\leq C' \cdot \frac{\frac{n^2}{k^2}\frac{\sqrt{n}}{k}\frac{1}{1-\beta}}{\frac{k}{n}} \asymp \frac{n^{3.5}}{k^4}\frac{1}{1-\beta}.
\end{aligned}$$

In summary under the condition

$$1 - \beta \succ \frac{n^{3.5}}{k^4}, \tag{A.40}$$

one can recover the neighborhood consistently w.h.p. in the sense

$$\lim_{n,k(n)\to\infty} \max_{i\in[n]} \frac{|\hat{\mathcal{N}}(v_i) \triangle \mathcal{N}(v_i)|}{|\mathcal{N}(v_i)|} = 0.$$

□

Proof of Theorem 3.1.1. Since $\widehat{f}$ is in the star hull around $\widehat{g}$, $\widehat{f}$ must lie in the set $\mathcal{H} := \mathcal{F} + \mathrm{star}(\mathcal{F} - \mathcal{F})$. Hence, in view of (3.4), excess loss $\mathcal{E}(\widehat{f})$ is upper bounded by

$$
\begin{aligned}
&\sup_{f\in\mathcal{H}} \left\{ (\widehat{\mathbb{E}} - \mathbb{E})[2(f^* - Y)(f^* - f)] + \mathbb{E}(f^* - f)^2 - (1+c)\cdot \widehat{\mathbb{E}}(f^* - f)^2 \right\} && \text{(A.41)}\\
&\le \sup_{f\in\mathcal{H}} \left\{ (\widehat{\mathbb{E}} - \mathbb{E})[2(f^* - Y)(f^* - f)] + (1+c/4)\mathbb{E}(f^* - f)^2 - (1+3c/4)\cdot \widehat{\mathbb{E}}(f^* - f)^2 \right.\\
&\qquad\qquad\qquad\qquad \left. -(c/4)\left(\widehat{\mathbb{E}}(f^* - f)^2 + \mathbb{E}(f^* - f)^2\right) \right\}\\
&\le \sup_{f\in\mathcal{H}} \left\{ (\widehat{\mathbb{E}} - \mathbb{E})[2(f^* - Y)(f^* - f)] - (c/4)\left(\widehat{\mathbb{E}}(f^* - f)^2 + \mathbb{E}(f^* - f)^2\right) \right\} && \text{(A.42)}\\
&+ \sup_{f\in\mathcal{H}} \left\{ (1+c/4)\mathbb{E}(f^* - f)^2 - (1+3c/4)\cdot \widehat{\mathbb{E}}(f^* - f)^2 \right\} && \text{(A.43)}
\end{aligned}
$$

We invoke the supporting Lemma A.4.1 (stated and proved below) for the term (A.43):

$$
\begin{aligned}
&\mathbb{E}\sup_{f\in\mathcal{H}} \left\{ (1+c/4)\mathbb{E}(f^* - f)^2 - (1+3c/4)\cdot \widehat{\mathbb{E}}(f^* - f)^2 \right\} && \text{(A.44)}\\
&\le \frac{K(2+c)}{2}\cdot \mathbb{E}\sup_{f\in\mathcal{H}} \frac{1}{n}\left\{ \sum_{i=1}^n 2\epsilon_i(f(X_i) - f^*(X_i)) - \frac{c}{4K(2+c)}\cdot \sum_{i=1}^n (f(X_i) - f^*(X_i))^2 \right\}. && \text{(A.45)}
\end{aligned}
$$

Let $\widehat{\mathbb{E}}'$ stand for empirical expectation with respect to an independent copy $(X_1', \dots, X_n')$. For the term (A.42), Jensen's inequality yields

$$
\begin{aligned}
&\mathbb{E}\sup_{f\in\mathcal{H}} \left\{ (\widehat{\mathbb{E}} - \mathbb{E})[2(f^* - Y)(f^* - f)] - (c/4)\left(\widehat{\mathbb{E}}(f^* - f)^2 + \mathbb{E}(f^* - f)^2\right) \right\}\\
&\le \mathbb{E}\sup_{f\in\mathcal{H}} \left\{ (\widehat{\mathbb{E}} - \widehat{\mathbb{E}}')[2(f^* - Y)(f^* - f)] - (c/4)\left(\widehat{\mathbb{E}}(f^* - f)^2 + \widehat{\mathbb{E}}'(f^* - f)^2\right) \right\}.
\end{aligned}
$$

When introducing i.i.d. Rademacher random variables, we observe that the quadratic term remains unchanged by renaming X_i and X_i', and thus the preceding expression is upper bounded by

$$2\mathbb{E}\sup_{f\in\mathcal{H}}\left\{\frac{1}{n}\sum_{i=1}^{n}2\epsilon_i(f^*(X_i)-Y_i)(f^*(X_i)-f(X_i))-(c/4)(f^*(X_i)-f(X_i))^2\right\}.$$

Using a contraction technique as in the proof of Lemma A.4.1, we obtain an upper bound of

$$2M\cdot\mathbb{E}\sup_{f\in\mathcal{H}}\frac{1}{n}\left\{\sum_{i=1}^{n}2\epsilon_i(f^*(X_i)-f(X_i))-\frac{c}{4M}\cdot\sum_{i=1}^{n}(f^*(X_i)-f(X_i))^2\right\} \tag{A.46}$$

Combining the bounds yields the statement of the theorem. □

Lemma A.4.1. *For any class $\mathcal{F}$ of uniformly bounded functions with $K=\sup_{f\in\mathcal{F}}|f|_\infty$, for any $f^*\in\mathcal{F}$, and for any $c>0$, it holds that*

$$\begin{aligned}&\mathbb{E}\sup_{f\in\mathcal{F}}\left\{\mathbb{E}(f-f^*)^2-(1+2c)\widehat{\mathbb{E}}(f-f^*)^2\right\}\\&\le c\cdot\mathbb{E}\sup_{f\in\mathcal{F}}\frac{1}{n}\left\{\frac{4K(1+c)}{c}\sum_{i=1}^{n}\epsilon_i(f(X_i)-f^*(X_i))-\sum_{i=1}^{n}(f(X_i)-f^*(X_i))^2\right\}.\end{aligned}$$

Proof of Lemma A.4.1. We write

$$\begin{aligned}&\mathbb{E}\sup_{f\in\mathcal{F}}\left\{\mathbb{E}(f-f^*)^2-(1+2c)\widehat{\mathbb{E}}(f-f^*)^2\right\}\\&=\mathbb{E}\sup_{f\in\mathcal{F}}\left\{(1+c)\mathbb{E}(f-f^*)^2-(1+c)\widehat{\mathbb{E}}(f-f^*)^2-c\mathbb{E}(f-f^*)^2-c\widehat{\mathbb{E}}(f-f^*)^2\right\}\end{aligned}$$

which, by Jensen's inequality, is upper bounded by

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left\{(1+c)(\widehat{\mathbb{E}}'(f-f^*)^2-\widehat{\mathbb{E}}(f-f^*)^2)-c\widehat{\mathbb{E}}'(f-f^*)^2-c\widehat{\mathbb{E}}(f-f^*)^2\right\}$$

We recall that $\widehat{\mathbb{E}}'$ is an empirical mean operator with respect to an independent copy $(X_1',\dots,X_n')$. Writing out the empirical expectations in the above expression, the above is equal to

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left\{\frac{1+c}{n}\sum_{i=1}^{n}\epsilon_i\Big((f(X_i')-f^*(X_i'))^2-(f(X_i)-f^*(X_i))^2\Big)-c\widehat{\mathbb{E}}'(f-f^*)^2-c\widehat{\mathbb{E}}(f-f^*)^2\right\}$$
$$\leq 2\cdot\mathbb{E}\sup_{f\in\mathcal{F}}\left\{\frac{1+c}{n}\sum_{i=1}^{n}\epsilon_i(f(X_i)-f^*(X_i))^2-c\widehat{\mathbb{E}}(f-f^*)^2\right\}$$

with the last expectation taken over ϵ_i and data X_i, $1\leq i\leq n$.

We proceed with a contraction-style proof. Condition on $X_1,\ldots,X_n$ and $\epsilon_2,\ldots,\epsilon_n$, and write out the expectation with respect to ϵ_1:

$$\begin{aligned}&\frac{1}{2}\sup_{f\in\mathcal{F}}\left\{\frac{1+c}{n}\sum_{i=2}^{n}\epsilon_i(f(X_i)-f^*(X_i))^2-c\widehat{\mathbb{E}}(f-f^*)^2+\frac{1+c}{n}(f(X_1)-f^*(X_1))^2\right\}\\&+\frac{1}{2}\sup_{g\in\mathcal{F}}\left\{\frac{1+c}{n}\sum_{i=2}^{n}\epsilon_i(g(X_i)-f^*(X_i))^2-c\widehat{\mathbb{E}}(g-f^*)^2-\frac{1+c}{n}(g(X_1)-f^*(X_1))^2\right\}\\&\leq\frac{1}{2}\sup_{f,g\in\mathcal{F}}\left\{\frac{1+c}{n}\sum_{i=2}^{n}\epsilon_t(f(X_i)-f^*(X_i))^2-c\widehat{\mathbb{E}}(f-f^*)^2+\frac{1+c}{n}\sum_{i=2}^{n}\epsilon_t(g(X_i)-f^*(X_i))^2\right.\\&\qquad\left.-c\widehat{\mathbb{E}}(g-f^*)^2+\frac{4K(1+c)}{n}|f(X_1)-g(X_1)|\right\}\end{aligned}$$

The absolute value can be dropped since the expression is symmetric in f,g. We obtain an upper bound of

$$\begin{aligned}&\frac{1}{2}\sup_{f,g\in\mathcal{F}}\left\{\frac{1+c}{n}\sum_{i=2}^{n}\epsilon_t(f(X_i)-f^*(X_i))^2-c\widehat{\mathbb{E}}(f-f^*)^2+\frac{1+c}{n}\sum_{i=2}^{n}\epsilon_t(g(X_i)-f^*(X_i))^2\right.\\&\qquad\left.-c\widehat{\mathbb{E}}(g-f^*)^2+\frac{4K(1+c)}{n}(f(X_1)-g(X_1))\right\}\\&=\mathbb{E}_{\epsilon_1}\sup_{f\in\mathcal{F}}\left\{\frac{1+c}{n}\sum_{i=2}^{n}\epsilon_i(f(X_i)-f^*(X_i))^2-c\widehat{\mathbb{E}}(f-f^*)^2+\frac{4K(1+c)}{n}\epsilon_1 f(X_1)\right\}\end{aligned}$$

Proceeding in this fashion for ϵ_2 until ϵ_n, we conclude

$$\begin{aligned}
&\mathbb{E} \sup_{f \in \mathcal{F}} \left\{ \mathbb{E}(f - f^*)^2 - (1 + 2c)\widehat{\mathbb{E}}(f - f^*)^2 \right\} \\
&\leq \mathbb{E} \sup_{f \in \mathcal{F}} \left\{ \frac{4K(1+c)}{n} \sum_{i=1}^{n} \epsilon_t(f(X_i) - f^*(X_i)) - \frac{c}{n} \sum_{i=1}^{n} (f(X_i) - f^*(X_i))^2 \right\}
\end{aligned}$$

where we added f^* back in for free since random signs are zero-mean. □

Proof of Theorem 3.1.2. We start with the deterministic upper bound (A.41) on excess loss (see the proof of Theorem 3.1.1):

$$\sup_{h \in \mathcal{H}} \left\{ (\widehat{\mathbb{E}} - \mathbb{E})[2\xi h] + \mathbb{E}h^2 - (1 + c) \cdot \widehat{\mathbb{E}}h^2 \right\} \tag{A.47}$$

where $h = f - f^* \in \mathcal{H}$. Define

$$\begin{aligned}
U_{X_i,Y_i}(h) &= 2\xi_i h(X_i) - \mathbb{E}[2\xi h] + \mathbb{E}h^2 - (1 + c) \cdot h(X_i)^2, \\
V_{X_i,Y_i}(h) &= 2\xi_i h(X_i) - \mathbb{E}[2\xi h] - \mathbb{E}h^2 + (1 - c') \cdot h(X_i)^2.
\end{aligned}$$

where c' will be specified later. We now prove a version of probabilistic symmetrization lemma Giné and Zinn (1984); Mendelson (2003) for

$$\mathbb{P}\left(\sup_{h \in \mathcal{H}} \sum_{i=1}^{n} U_{X_i,Y_i}(h) > x \right). \tag{A.48}$$

Note that unlike the usual applications of the technique in the literature, we perform symmetrization with the quadratic terms. Define

$$\mathcal{B} = \left\{ \sup_{h \in \mathcal{H}} \sum_{i=1}^{n} U_{X_i,Y_i}(h) > x \right\}, \quad \beta = \inf_{h \in \mathcal{H}} \mathbb{P}\left(\sum_{i=1}^{n} V_{X_i,Y_i}(h) < \frac{x}{2} \right). \tag{A.49}$$

Clearly for $\{X_i, Y_i\}_{i=1}^n \in \mathcal{B}$, there exists a $h \in \mathcal{H}$ satisfies condition in $\mathcal{B}$. If in addition h

satisfies

$$\sum_{i=1}^{n} V_{X_i',Y_i'}(h) < \frac{x}{2}$$

then

$$\sum_{i=1}^{n} U_{X_i,Y_i}(h) - V_{X_i',Y_i'}(h) > \frac{x}{2}$$

and therefore

$$\sup_{h\in\mathcal{H}} \sum_{i=1}^{n} U_{X_i,Y_i}(h) - V_{X_i',Y_i'}(h) > \frac{x}{2}.$$

The latter can be written as

$$\sup_{h\in\mathcal{H}} \left\{ \sum_{i=1}^{n} 2\xi_i h(X_i) - 2\xi_i' h(X_i') + 2\mathbb{E}h^2 - (1+c)\cdot h(X_i)^2 - (1-c')\cdot h(X_i')^2 \right\} > \frac{x}{2}.$$

Then for this particular h,

$$\begin{aligned}
\beta = \inf_{g\in\mathcal{H}} \mathbb{P}\left(\sum_{i=1}^{n} V_{X_i',Y_i'}(g) < \frac{x}{2}\right) &\le \mathbb{P}\left(\sum_{i=1}^{n} V_{X_i',Y_i'}(h) < \frac{x}{2}\right) \\
\le \mathbb{P}\left(\sum_{i=1}^{n} U_{X_i,Y_i}(h) - V_{X_i',Y_i'}(h) > \frac{x}{2}\right) &\le \mathbb{P}\left(\sup_{h\in\mathcal{H}} \sum_{i=1}^{n} U_{X_i,Y_i}(h) - V_{X_i',Y_i'}(h) > \frac{x}{2}\right).
\end{aligned}$$

Note that the right-hand-side does not depend on h. We integrate over $\{X_i, Y_i\}_{i=1}^{n} \in \mathcal{B}$ to obtain

$$\begin{aligned}
&\beta \cdot \mathbb{P}\left(\sup_{h\in\mathcal{H}} \sum_{i=1}^{n} U_{X_i,Y_i}(h) > x\right) \\
&\le \mathbb{P}\left(\sup_{h\in\mathcal{H}} n \cdot \left\{2(\widehat{\mathbb{E}} - \widehat{\mathbb{E}}')[\xi h] + 2\mathbb{E}h^2 - (1+c)\cdot \widehat{\mathbb{E}}h^2 - (1-c')\cdot \widehat{\mathbb{E}}'h^2\right\} > \frac{x}{2}\right) \qquad \text{(A.50)}
\end{aligned}$$

Next, we apply Assumption 3.1.1 with $\epsilon = c/4 = 1/72$ to terms in (A.50) to construct an offset Rademacher process. Note

$$\frac{2}{1-\epsilon} < 2(1+2\epsilon) = 2 + c.$$

We can now choose $\tilde{c}, c' > 0$ in that satisfy

$$\frac{2}{1-\epsilon} \leq 2 + c - c' - 2\tilde{c} \quad \iff \quad 1 - (1 - c' - \tilde{c})(1-\epsilon) \leq (1 + c - \tilde{c})(1-\epsilon) - 1. \tag{A.51}$$

Choose b now such that

$$1 - (1 - c' - \tilde{c})(1-\epsilon) \leq b \leq (1 + c - \tilde{c})(1-\epsilon) - 1. \tag{A.52}$$

Then we have on the set $\mathcal{H}$, applying lower isometry bound and Eq. (A.52), with probability at least $1 - 2\delta$,

$$\begin{aligned} \widehat{\mathbb{E}}(f - f^*)^2 \geq (1-\epsilon) \cdot \mathbb{E}(f - f^*)^2 \quad &\implies \quad (1+b)\mathbb{E}h^2 - (1+c) \cdot \widehat{\mathbb{E}}h^2 \leq -\tilde{c} \cdot \widehat{\mathbb{E}}h^2, \\ \widehat{\mathbb{E}}'(f - f^*)^2 \geq (1-\epsilon) \cdot \mathbb{E}(f - f^*)^2 \quad &\implies \quad (1-b)\mathbb{E}h^2 - (1-c') \cdot \widehat{\mathbb{E}}'h^2 \leq -\tilde{c} \cdot \widehat{\mathbb{E}}'h^2. \end{aligned}$$

Thus we can continue bounding the expression in (A.50) as

$$\begin{aligned} &\sup_{h \in \mathcal{H}} n \cdot \left\{ 2(\widehat{\mathbb{E}} - \widehat{\mathbb{E}}')[\xi h] + 2\mathbb{E}h^2 - (1+c) \cdot \widehat{\mathbb{E}}h^2 - (1-c') \cdot \widehat{\mathbb{E}}'h^2 \right\} \\ &= \sup_{h \in \mathcal{H}} n \cdot \left\{ 2(\widehat{\mathbb{E}} - \widehat{\mathbb{E}}')[\xi h] + (1+b)\mathbb{E}h^2 - (1+c) \cdot \widehat{\mathbb{E}}h^2 + (1-b)\mathbb{E}h^2 - (1-c') \cdot \widehat{\mathbb{E}}'h^2 \right\} \\ &\leq \sup_{h \in \mathcal{H}} n \cdot \left\{ 2(\widehat{\mathbb{E}} - \widehat{\mathbb{E}}')[\xi h] - \tilde{c} \cdot \widehat{\mathbb{E}}h^2 - \tilde{c} \cdot \widehat{\mathbb{E}}'h^2 \right\} \end{aligned}$$

For the probability of deviation, we obtain

$$
\begin{aligned}
&\beta \cdot \mathbb{P}\left(\sup_{h\in\mathcal{H}} \sum_{i=1}^{n} U_{X_i,Y_i}(h) > x\right) \\
&\le \mathbb{P}\left(\sup_{h\in\mathcal{H}} n \cdot \left\{2(\widehat{\mathbb{E}} - \widehat{\mathbb{E}}')[\xi h] - \tilde{c} \cdot \mathbb{E}h^2 - \tilde{c} \cdot \widehat{\mathbb{E}}'h^2\right\} > \frac{x}{2}\right) + 2\delta \\
&= \mathbb{P}\left(\sup_{h\in\mathcal{H}} n \cdot \left\{2(\widehat{\mathbb{E}} - \widehat{\mathbb{E}}')[\epsilon\xi h] - \tilde{c} \cdot \widehat{\mathbb{E}}h^2 - \tilde{c} \cdot \mathbb{E}'h^2\right\} > \frac{x}{2}\right) + 2\delta \\
&\le 2\mathbb{P}\left(\sup_{h\in\mathcal{H}} \left\{\sum_{i=1}^{n} 2\epsilon_i\xi_i h(X_i) - \tilde{c} \cdot \sum_{i=1}^{n} h(X_i)^2\right\} > \frac{x}{4}\right) + 2\delta.
\end{aligned}
$$

To estimate β, write

$$
\begin{aligned}
\beta &= \inf_{h\in\mathcal{H}} \mathbb{P}\left(\sum_{i=1}^{n} V_{X_i,Y_i}(h) < \frac{x}{2}\right) && \text{(A.53)} \\
&= 1 - \sup_{h\in\mathcal{H}} \mathbb{P}\left(\sum_{i=1}^{n} 2\xi_i h(X_i) - \mathbb{E}[2\xi h] - \mathbb{E}h^2 + (1-c') \cdot h(X_i)^2 \ge \frac{x}{2}\right). && \text{(A.54)}
\end{aligned}
$$

Let's bound the last term in above equation, for any $h \in \mathcal{H}$

$$
\begin{aligned}
&\mathbb{P}\left((\widehat{\mathbb{E}} - \mathbb{E})[2\xi h] + (1-c')\widehat{\mathbb{E}}h^2 - \mathbb{E}h^2 > \frac{x}{2n}\right) && \text{(A.55)} \\
\le &\mathbb{P}\left((\widehat{\mathbb{E}} - \mathbb{E})[2\xi h] > \frac{x}{2n} + \frac{c'}{2}\mathbb{E}h^2\right) + \mathbb{P}\left((\widehat{\mathbb{E}} - \mathbb{E})[h^2] > \frac{c'}{2(1-c')}\mathbb{E}h^2\right). && \text{(A.56)} \\
& && \text{(A.57)}
\end{aligned}
$$

Define

$$A := \sup_{h\in\mathcal{H}} \frac{\mathbb{E}h^4}{(\mathbb{E}h^2)^2} \quad \text{and} \quad B := \sup_{X,Y} \mathbb{E}\xi^4.$$

Then for the second term in Eq (A.56), using Chebyshev's inequality

$$\mathbb{P}\left((\widehat{\mathbb{E}} - \mathbb{E})[h^2] > \frac{c'}{2(1-c')}\mathbb{E}h^2\right) \le \frac{4(1-c')^2 A}{c'^2 n} \le 1/4$$

if

$$n \ge \frac{16(1-c')^2 A}{c'^2}.$$

For the first term in Eq (A.56), note

$$\mathsf{Var}[2\xi h] \leq 4\mathbb{E}[\xi^2 h^2] \leq 4\sqrt{AB} \cdot \mathbb{E}h^2$$

and thus through Chebyshev inequality

$$\begin{aligned}\mathbb{P}\left((\widehat{\mathbb{E}} - \mathbb{E})[2\xi h] > \frac{x}{2n} + \frac{c'}{2}\mathbb{E}h^2\right) &\leq \frac{4\sqrt{AB} \cdot \mathbb{E}h^2}{n\left(\frac{x}{2n} + \frac{c'}{2}\mathbb{E}h^2\right)^2} \\ &\leq \frac{4\sqrt{AB} \cdot \mathbb{E}h^2}{n \cdot 4\frac{x}{2n} \cdot \frac{c'}{2}\mathbb{E}h^2} \leq \frac{1}{4}\end{aligned}$$

if

$$x \geq \frac{16\sqrt{AB}}{c'}.$$

Assemble above bounds, for any $h \in \mathcal{H}$

$$\sup_{h\in\mathcal{H}} \mathbb{P}\left(\sum_{i=1}^n 2\xi_i h(X_i) - \mathbb{E}[2\xi h] - \mathbb{E}h^2 + (1-c') \cdot h(X_i)^2 \geq \frac{x}{2}\right) \leq \frac{1}{2}$$

which further implies $\beta \geq 1/2$ for any $x > \frac{16\sqrt{AB}}{c'}$ and whenever

$$n > \frac{16(1-c')^2 A}{c'^2}.$$

Under the above regime,

$$\frac{1}{2}\mathbb{P}\left(\sup_{h\in\mathcal{H}} \sum_{i=1}^n U_{X_i,Y_i}(h) > x\right) \leq 2\mathbb{P}\left(\sup_{h\in\mathcal{H}}\left\{\sum_{i=1}^n \epsilon_i\xi_i h(X_i) - \tilde{c} \cdot \sum_{i=1}^n h(X_i)^2\right\} > \frac{x}{4}\right) + 2\delta$$

and so

$$\begin{aligned}&\mathbb{P}\left(\sup_{h\in\mathcal{H}} \sum_{i=1}^n U_{X_i,Y_i}(h) > 4t\right) \\ &\leq 4\mathbb{P}\left(\sup_{h\in\mathcal{H}}\left\{\sum_{i=1}^n \epsilon_i\xi_i h(X_i) - \tilde{c} \cdot \sum_{i=1}^n h(X_i)^2\right\} > t\right) + 4\delta.\end{aligned}$$

We conclude by writing

$$\begin{aligned}
&\mathbb{P}\left(\sup_{h\in\mathcal{H}}(\widehat{\mathbb{E}}-\mathbb{E})[2\xi h]+\mathbb{E}h^2-(1+c)\cdot\widehat{\mathbb{E}}h^2>4t\right)\\
&\leq 4\mathbb{P}\left(\sup_{h\in\mathcal{H}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\xi_i h(X_i)-\tilde{c}\cdot\sum_{i=1}^{n}h(X_i)^2>t\right)+4\delta.
\end{aligned}$$

□

Proof of Lemma 3.1.2. Using a standard argument,

$$\mathbb{E}_\epsilon \max_{v\in V}\left[\sum_{i=1}^{n}\epsilon_i v_i - Cv_i^2\right]\leq\frac{1}{\lambda}\log\sum_{v\in V}\mathbb{E}_\epsilon\exp\left\{\sum_{i=1}^{n}\lambda\epsilon_i v_i-\lambda C v_i^2\right\}.$$

For any $v\in V$,

$$\mathbb{E}_\epsilon\exp\left\{\sum_{i=1}^{n}\lambda\epsilon_i v_i-\lambda C v_i^2\right\}\leq\exp\left\{\sum_{i=1}^{n}\lambda^2 v_i^2/2-\lambda C v_i^2\right\}\leq 1$$

by setting $\lambda=2C$. The first claim follows. For the second claim,

$$\begin{aligned}
&\Pr\max_{v\in V}\left[\sum_{i=1}^{n}\epsilon_i v_i - Cv_i^2\right]\geq\frac{1}{2C}\log(N/\delta)\\
\leq&\mathbb{E}\exp\left\{\lambda\max_{v\in V}\left[\sum_{i=1}^{n}\epsilon_i v_i - Cv_i^2\right]-\lambda\frac{1}{2C}\log(N/\delta)\right\}\\
\leq&\sum_{v\in V}\mathbb{E}\exp\left\{\lambda\left[\sum_{i=1}^{n}\epsilon_i v_i - Cv_i^2\right]-\lambda\frac{1}{2C}\log(N/\delta)\right\}\\
\leq&\sum_{v\in V}\exp\left\{-\log(N/\delta)\right\}=\delta.
\end{aligned}$$

Now let's move to the case where ξ, the noise is unbounded.

$$\mathbb{E}_\epsilon \frac{1}{n} \max_{v \in V} \left\{ \sum_{i=1}^n \epsilon_i \xi_i v_i - C v_i^2 \right\} \leq \frac{1}{n\lambda} \log \mathbb{E}_\epsilon \sum_{v \in V} \exp \left(\lambda \sum_{i=1}^n \epsilon_i \xi_i v_i - \lambda C v_i^2 \right)$$

$$\leq \frac{1}{n\lambda} \log \sum_{v \in V} \exp \left(\sum_{i=1}^n \frac{\lambda^2}{2} \xi_i^2 v_i^2 - \lambda C v_i^2 \right) \leq \max_{v \in V \setminus \{0\}} \frac{\sum_{i=1}^n v_i^2 \xi_i^2}{2C \sum_{i=1}^n v_i^2} \cdot \frac{\log N}{n}$$

if we take $\lambda = \min_{v \in V \setminus \{0\}} \frac{2C \sum_{i=1}^n v_i^2}{\sum_{i=1}^n v_i^2 \xi_i^2}$. The high probability statement follows also use this particular choice of λ.

□

Proof of Lemma 3.1.3. The proof proceeds as in (Rakhlin and Sridharan, 2014). Fix $\gamma \in [0, 1]$. By definition of a cover, there exists a set $V \subset \mathbb{R}^n$ vectors of size $N = \mathcal{N}_2(\mathcal{G}, \gamma)$ with the following property: for any $g \in \mathcal{G}$, there exists a $v = v[g] \in V$ such that

$$\frac{1}{n} \sum_{i=1}^n (g(z_i) - v_i)^2 \leq \gamma^2.$$

Then we may write,

$$\mathbb{E}_\epsilon \sup_{g \in \mathcal{G}} \left[\sum_{t=1}^n \epsilon_i g(z_i) - C g(z_i)^2 \right] \tag{A.58}$$

$$\leq \mathbb{E}_\epsilon \sup_{g \in \mathcal{G}} \left[\frac{1}{n} \sum_{t=1}^n \epsilon_i (g(z_i) - v[g]_i) \right] + \mathbb{E}_\epsilon \sup_{g \in \mathcal{G}} \left[\sum_{t=1}^n (C/4) v[g]_i^2 - C g(z_i)^2 \right] \tag{A.59}$$

$$+ \mathbb{E}_\epsilon \sup_{g \in \mathcal{G}} \left[\sum_{t=1}^n \epsilon_i v[g]_i - (C/4) v[g]_i^2 \right] \tag{A.60}$$

We now argue that the second term is nonpositive. More precisely, we claim that for any $g \in \mathcal{G}$,

$$\frac{1}{4} \sum_{t=1}^n v[g]_i^2 \leq \sum_{t=1}^n g(z_i)^2 \tag{A.61}$$

for some element $v[g] \in V \cup \{\mathbf{0}\}$. First, consider the case $\sum_{t=1}^n g(z_i)^2 \leq \gamma^2$. Then $v[g] = \mathbf{0}$

is an element γ-close to values of g on the sample, and (A.61) is trivially satisfied. Next, consider the case $\sum_{t=1}^{n} g(z_i)^2 > \gamma^2$ and write $u = (g(z_1), \ldots, g(z_n))$. The triangle inequality for the Euclidean norm yields

$$\|v[g]\| \leq \|v[g] - u\| + \|u\| \leq \gamma + \|u\| \leq 2\|u\|,$$

establishing non-positivity of the second term in (A.58). The third term in (A.58) is upper bounded with the help of Lemma 3.1.2 as

$$\mathbb{E}_\epsilon \max_{g\in\mathcal{G}} \left[\sum_{t=1}^{n} \epsilon_i v[g]_i - (C/4)v[g]_i^2\right] \leq \frac{2}{C}\log \mathcal{N}_2(\mathcal{G},\gamma)$$

Finally, the first term in (A.58) is upper bounded using the standard chaining technique, keeping in mind that the ℓ_2-diameter of the indexing set is at most γ. □

Proof of Lemma 3.1.4. The proof is similar to the proof of Lemma 3.1.3. We proceed with the following decomposition:

$$\sup_{g\in\mathcal{G}} \left[\frac{1}{n}\sum_{t=1}^{n} \epsilon_i g(z_i) - Cg(z_i)^2\right] \leq \sup_{g\in\mathcal{G}} \left[\frac{1}{n}\sum_{t=1}^{n} \epsilon_i (g(z_i) - v[g]_i)\right] + \sup_{g\in\mathcal{G}} \left[\frac{1}{n}\sum_{t=1}^{n} \epsilon_i v[g]_i - \frac{C}{4}v[g]_i^2\right].$$

For the first term, we can employ the traditional high probability chaining bound. For some $c > 0$, the following holds,

$$\mathbb{P}_\epsilon\left(\sup_{g\in\mathcal{G}} \left[\frac{1}{n}\sum_{t=1}^{n} \epsilon_i (g(z_i) - v[g]_i)\right] > u \cdot \inf_{\alpha\in[0,\gamma]}\left\{4\alpha + \frac{12}{\sqrt{n}}\int_\alpha^\gamma \sqrt{\log \mathcal{N}_2(\mathcal{G},\delta)}d\delta\right\}\right)$$
$$\leq \frac{2}{1-e^{-2}}\exp(-cu^2).$$

For the second term,

$$\mathbb{P}_\epsilon\left(\sup_{g\in\mathcal{G}} \left[\frac{1}{n}\sum_{t=1}^{n} \epsilon_i v[g]_i - (C/4)v[g]_i^2\right] > \frac{2}{C}\frac{\log \mathcal{N}_2(\mathcal{G},\gamma) + u}{n}\right) \leq \exp(-u).$$

Combining the above two bounds, we have

$$\begin{aligned}
&\mathbb{P}_\epsilon\left(\sup_{g\in\mathcal{G}}\left[\frac{1}{n}\sum_{t=1}^n \epsilon_i g(z_i) - Cg(z_i)^2\right] > u\cdot \inf_{\alpha\in[0,\gamma]}\left\{4\alpha + \frac{12}{\sqrt{n}}\int_\alpha^\gamma \sqrt{\log\mathcal{N}_2(\mathcal{G},\delta)}d\delta\right\}\right.\\
&\qquad\qquad\qquad\qquad\qquad\qquad\left.+\frac{2}{C}\frac{\log\mathcal{N}_2(\mathcal{G},\gamma)+u}{n}\right)\\
&\leq \mathbb{P}_\epsilon\left(\sup_{g\in\mathcal{G}}\left[\frac{1}{n}\sum_{t=1}^n \epsilon_i (g(z_i) - v[g]_i)\right] > u\cdot \inf_{\alpha\in[0,\gamma]}\left\{4\alpha + \frac{12}{\sqrt{n}}\int_\alpha^\gamma \sqrt{\log\mathcal{N}_2(\mathcal{G},\delta)}d\delta\right\}\right)\\
&\quad + \mathbb{P}_\epsilon\left(\sup_{g\in\mathcal{G}}\left[\frac{1}{n}\sum_{t=1}^n \epsilon_i v[g]_i - (C/4)v[g]_i^2\right] > \frac{2}{C}\frac{\log\mathcal{N}_2(\mathcal{G},\gamma)+u}{n}\right)\\
&\leq \frac{2}{1-e^{-2}}\exp(-cu^2) + \exp(-u).
\end{aligned}$$

□

Proof of Theorem 3.1.3. Denote by $\mathcal{B}$ the unit ball with respect to ℓ_2 distance, $\mathcal{B} = \{h : (\mathbb{E}h^2)^{1/2} \leq 1\}$, and let $\mathcal{S}$ denote the unit sphere. Choosing any $h \in \mathcal{H}\backslash r\mathcal{B}$, we have $\|h\|_{\ell_2} > r \triangleq \alpha_n(\mathcal{H}, \kappa', \delta)$ with k' to be chosen later. Under the assumption that $\mathcal{H}$ is star-shaped, we know $h_r := r/\|h\|_{\ell_2}\cdot h \in \mathcal{H}$, thus

$$\begin{aligned}
&\frac{2}{n}\sum_{i=1}^n \epsilon_i\xi_i h(X_i) - c'\frac{1}{n}\sum_{i=1}^n h^2(X_i)\\
=&\frac{\|h\|_{\ell_2}}{r}\frac{2}{n}\sum_{i=1}^n \epsilon_i\xi_i h_r(X_i) - \left(\frac{\|h\|_{\ell_2}}{r}\right)^2 c'\frac{1}{n}\sum_{i=1}^n h_r^2(X_i)\\
=&\frac{\|h\|_{\ell_2}}{r}\left\{\frac{2}{n}\sum_{i=1}^n \epsilon_i\xi_i h_r(X_i) - c'\frac{1}{n}\sum_{i=1}^n h_r^2(X_i)\right\} - \frac{\|h\|_{\ell_2}}{r}\left(\frac{\|h\|_{\ell_2}}{r}-1\right)c'\frac{1}{n}\sum_{i=1}^n h_r^2(X_i).
\end{aligned}$$

Comparing the supremum of the offset Rademacher process outside the ball $r\mathcal{B}$ with the one inside the ball $r\mathcal{B}$, we have

$$\begin{aligned}
&\sup_{h\in\mathcal{H}\setminus r\mathcal{B}}\left\{\frac{2}{n}\sum_{i=1}^{n}\epsilon_i\xi_i h(X_i)-c'\frac{1}{n}\sum_{i=1}^{n}h^2(X_i)\right\}-\sup_{h\in\mathcal{H}\cap r\mathcal{B}}\left\{\frac{2}{n}\sum_{i=1}^{n}\epsilon_i\xi_i h(X_i)-c'\frac{1}{n}\sum_{i=1}^{n}h^2(X_i)\right\}\\
&\le\sup_{h\in\mathcal{H}\setminus r\mathcal{B}}\left\{\left(\frac{\|h\|_{\ell_2}}{r}-1\right)\sup_{h_r\in\mathcal{H}\cap r\mathcal{B}}\left\{\frac{2}{n}\sum_{i=1}^{n}\epsilon_i\xi_i h_r(X_i)-c'\frac{1}{n}\sum_{i=1}^{n}h_r^2(X_i)\right\}\right.\\
&\qquad\qquad\left.-\frac{\|h\|_{\ell_2}}{r}\left(\frac{\|h\|_{\ell_2}}{r}-1\right)\inf_{h_r\in\mathcal{H}\cap r\mathcal{S}}\left\{c'\frac{1}{n}\sum_{i=1}^{n}h_r^2(X_i)\right\}\right\}\\
&\le\sup_{h\in\mathcal{H}\setminus r\mathcal{B}}\left\{\left(\frac{\|h\|_{\ell_2}}{r}-1\right)\left\{\sup_{h_r\in\mathcal{H}\cap r\mathcal{B}}\left\{\frac{2}{n}\sum_{i=1}^{n}\epsilon_i\xi_i h_r(X_i)-c'\frac{1}{n}\sum_{i=1}^{n}h_r^2(X_i)\right\}\right.\right.\\
&\qquad\qquad\qquad\left.\left.-\inf_{h_r\in\mathcal{H}_r\cap r\mathcal{S}}\left\{c'\frac{1}{n}\sum_{i=1}^{n}h_r^2(X_i)\right\}\right\}\right\}.
\end{aligned}\tag{A.62}$$

If

$$\kappa' r^2\le c'(1-\epsilon)r^2,$$

we can apply the lower isometry bound 3.1.1 and conclude

$$\sup_{h\in\mathcal{H}\cap r\mathcal{B}}\left\{\frac{2}{n}\sum_{i=1}^{n}\epsilon_i\xi_i h(X_i)-c'\frac{1}{n}\sum_{i=1}^{n}h^2(X_i)\right\}\le k'r^2\le c'(1-\epsilon)r^2\le\inf_{h_r\in\mathcal{H}\cap r\mathcal{S}}\left\{c'\frac{1}{n}\sum_{i=1}^{n}h_r^2(X_i)\right\}$$

with probability at least $1-2\delta$.

Under this event, the difference of terms in (A.62) is smaller than 0, and we conclude

$$\begin{aligned}
&\sup_{h\in\mathcal{H}\setminus r\mathcal{B}}\left\{\frac{2}{n}\sum_{i=1}^{n}\epsilon_i\xi_i h(X_i)-c'\frac{1}{n}\sum_{i=1}^{n}h^2(X_i)\right\}-\sup_{h\in\mathcal{H}\cap r\mathcal{B}}\left\{\frac{2}{n}\sum_{i=1}^{n}\epsilon_i\xi_i h(X_i)-c'\frac{1}{n}\sum_{i=1}^{n}h^2(X_i)\right\}\\
&\le\sup_{h\in\mathcal{H}\setminus r\mathcal{B}}\left\{\left(\frac{\|h\|_{\ell_2}}{r}-1\right)\left\{\sup_{h_r\in\mathcal{H}\cap r\mathcal{B}}\left\{\frac{2}{n}\sum_{i=1}^{n}\epsilon_i\xi_i h_r(X_i)-c'\frac{1}{n}\sum_{i=1}^{n}h_r^2(X_i)\right\}\right.\right.\\
&\qquad\qquad\qquad\left.\left.-\inf_{h_r\in\mathcal{H}_r\cap r\mathcal{S}}\left\{c'\frac{1}{n}\sum_{i=1}^{n}h_r^2(X_i)\right\}\right\}\right\}\\
&\le\sup_{h\in\mathcal{H}\setminus r\mathcal{B}}\left\{\left(\frac{\|h\|_{\ell_2}}{r}-1\right)\left(\kappa' r^2-c'(1-\epsilon)r^2\right)\right\}\le 0
\end{aligned}$$

Thus the excess loss is upper bounded by the offset Rademacher process, and the latter is further bounded by the process restricted within the critical radius:

$$\sup_{h\in\mathcal{H}}\left\{\frac{2}{n}\sum_{i=1}^{n}\epsilon_i\xi_i h(X_i)-c'\frac{1}{n}\sum_{i=1}^{n}h^2(X_i)\right\}\le\sup_{h\in\mathcal{H}\cap r\mathcal{B}}\left\{\frac{2}{n}\sum_{i=1}^{n}\epsilon_i\xi_i h(X_i)-c'\frac{1}{n}\sum_{i=1}^{n}h^2(X_i)\right\}$$
$$\le\alpha_n^2(\mathcal{H},c'(1-\epsilon),\delta)$$

with probability at least $1-2\delta$.

□

Proof of Theorem 3.1.4. Denote $\mathcal{F}\subset\mathcal{G}=\mathcal{F}+\mathrm{star}(\mathcal{F}-\mathcal{F})$. The minimax excess loss can be written as

$$\begin{aligned}&\inf_{\hat{g}\in\mathcal{G}}\sup_{P}\left\{\mathbb{E}(\hat{g}-Y)^2-\inf_{f\in\mathcal{F}}\mathbb{E}(f-Y)^2\right\}\\&=\inf_{\hat{g}\in\mathcal{G}}\sup_{P}\left\{\left\{-\mathbb{E}2Y\hat{g}+\mathbb{E}\hat{g}^2\right\}+\sup_{f\in\mathcal{F}}\left\{\mathbb{E}2Yf-\mathbb{E}f^2\right\}\right\}.\end{aligned}$$

Now let's construct a particular distribution P in the following way: take any $x_1,x_2,...,x_{(1+c)n}\in\mathcal{X}$ and let P_X be the uniform distribution on these $(1+c)n$ points. For any $\epsilon=(\epsilon_1,\ldots,\epsilon_{(1+c)n})\in\{\pm1\}^{(1+c)n}$, denote the distribution P_ϵ of (X,Y) indexed by ϵ to be: X is sampled from P_X, and $Y_{|X=x_i}=\epsilon_i,\ \forall 1\le i\le(1+c)n$. Note here $\hat{g}:(X,Y)^{\otimes n}\to\mathcal{F}+\mathrm{star}(\mathcal{F}-\mathcal{F})$. Now we proceed with this particular distribution

$$\begin{aligned}&\inf_{\hat{g}\in\mathcal{G}}\sup_{P}\left\{\left\{-\mathbb{E}2Y\hat{g}+\mathbb{E}\hat{g}^2\right\}+\sup_{f\in\mathcal{F}}\left\{\mathbb{E}2Yf-\mathbb{E}f^2\right\}\right\}\\&\ge\inf_{\hat{g}\in\mathcal{G}}\sup_{\{x_i\}_{i=1}^{(1+c)n}\in\mathcal{X}^{\otimes(1+c)n}}\mathbb{E}_\epsilon\left\{\left\{-\mathbb{E}2Y\hat{g}+\mathbb{E}\hat{g}^2\right\}+\sup_{f\in\mathcal{F}}\left\{\mathbb{E}2Yf-\mathbb{E}f^2\right\}\right\}\\&\ge\sup_{\{x_i\}_{i=1}^{(1+c)n}\in\mathcal{X}^{\otimes(1+c)n}}\mathbb{E}_\epsilon\left\{\sup_{f\in\mathcal{F}}\frac{1}{(1+c)n}\left\{\sum_{i=1}^{(1+c)n}2\epsilon_i f(x_i)-f(x_i)^2\right\}\right\}\\&\qquad-\sup_{\hat{g}\in\mathcal{G}}\sup_{\{x_i\}_{i=1}^{(1+c)n}\in\mathcal{X}^{\otimes(1+c)n}}\mathbb{E}_\epsilon\left\{2\mathbb{E}Y\hat{g}-\mathbb{E}\hat{g}^2\right\}.\end{aligned}$$

Note that the first term is exactly $\mathfrak{R}^{o}((1+c)n,\mathcal{F})$. Let us upper bound the second term.

Denote the indices of a uniform n samples from $(1+c)n$ samples $\{x_i\}_{i=1}^{(1+c)n}$ with replacement as $i_1, i_2, \ldots, i_n$, and I be the set of unique indices $|I| \leq n$. Observe that $\hat{g}$ is a function of (x_I, Y_I) only, independent of $\epsilon_j, j \notin I$.

$$\begin{aligned}
&\sup_{\hat{g}\in\mathcal{G}} \sup_{\{x_i\}_{i=1}^{(1+c)n}\in\mathcal{X}^{\otimes(1+c)n}} \mathbb{E}_\epsilon \left\{2\mathbb{E}Y\hat{g} - \mathbb{E}\hat{g}^2\right\} \\
&\leq \sup_{\hat{g}\in\mathcal{G}} \sup_{\{x_i\}_{i=1}^{(1+c)n}\in\mathcal{X}^{\otimes(1+c)n}} \mathbb{E}_\epsilon \mathbb{E}_{i_1,\ldots,i_n} \left\{\frac{1}{(1+c)n}\sum_{i=1}^{(1+c)n}\left\{2\epsilon_i\hat{g}(x_i) - \hat{g}(x_i)^2\right\}\right\} \\
&= \sup_{\hat{g}\in\mathcal{G}} \sup_{\{x_i\}_{i=1}^{(1+c)n}\in\mathcal{X}^{\otimes(1+c)n}} \mathbb{E}_{i_1,\ldots,i_n} \mathbb{E}_\epsilon \left\{\frac{1}{(1+c)n}\sum_{i=1}^{(1+c)n}\left\{2\epsilon_i\hat{g}(x_i) - \hat{g}(x_i)^2\right\}\right\} \qquad \text{(A.63)}
\end{aligned}$$

Conditionally on $i_1, i_2, ..., i_n$,

$$\frac{1}{(1+c)n}\sum_{i\notin I}\left\{2\epsilon_i\hat{g}(x_i) - \hat{g}(x_i)^2\right\} = 0 - \frac{1}{(1+c)n}\sum_{i\notin I}\hat{g}(x_i)^2 < 0.$$

Expression in (A.63) is upper bounded by

$$\begin{aligned}
&\sup_{\hat{g}\in\mathcal{G}} \sup_{\{x_i\}_{i=1}^{(1+c)n}\in\mathcal{X}^{\otimes(1+c)n}} \mathbb{E}_{i_1,\ldots,i_n} \mathbb{E}_\epsilon \left\{\frac{1}{(1+c)n}\sum_{i\in I}\left\{2\epsilon_i\hat{g}(x_i) - \hat{g}(x_i)^2\right\}\right\} \\
&\leq \sup_{\hat{g}\in\mathcal{G}} \mathbb{E}_{i_1,\ldots,i_n} \sup_{\{x_i\}_{i=1}^{|I|}\in\mathcal{X}^{\otimes|I|}} \mathbb{E}_\epsilon \left\{\frac{1}{(1+c)n}\sum_{i\in I}\left\{2\epsilon_i\hat{g}(x_i) - \hat{g}(x_i)^2\right\}\right\} \\
&\leq \sup_{\hat{g}\in\mathcal{G}} \sup_{\{x_i\}_{i=1}^{n}\in\mathcal{X}^{\otimes n}} \mathbb{E}_\epsilon \left\{\frac{1}{(1+c)n}\sum_{i=1}^{cn}\left\{2\epsilon_i\hat{g}(x_i) - \hat{g}(x_i)^2\right\}\right\} \\
&\leq \sup_{\{x_i\}_{i=1}^{n}\in\mathcal{X}^{\otimes n}} \mathbb{E}_\epsilon \sup_{g\in\mathcal{G}} \left\{\frac{1}{(1+c)n}\sum_{i=1}^{cn}\left\{2\epsilon_i g(x_i) - g(x_i)^2\right\}\right\} \\
&= \frac{c}{1+c}\mathfrak{R}^{\mathrm{o}}(cn, \mathcal{G}).
\end{aligned}$$

Thus the claim holds.

□

Proof of Lemma 3.1.7. From Lemma 3.1.5, we know for $\mathcal{H} = \mathcal{F} - f^* + \mathrm{star}(\mathcal{F} - \mathcal{F})$,

$$\log \mathcal{N}_2(\mathcal{H}, 8\epsilon) \leq \log \mathcal{N}_2(\mathcal{F} - f^*, 4\epsilon) + \log \mathcal{N}_2(\mathrm{star}(\mathcal{F} - \mathcal{F}), 4\epsilon) \leq \log \frac{2}{\epsilon} + 3 \log \mathcal{N}_2(\mathcal{F}, \epsilon).$$

Consider the δ-covering net of $\mathcal{H}$, where for any $h \in \mathcal{H}$, $v[h]$ is the closest point on the net.

$$\begin{aligned}
&\frac{1}{n} \sup_{h \in \mathcal{H}} \left\{ \sum_{i=1}^n 2\epsilon_i \xi_i h(X_i) - C h(X_i)^2 \right\} \\
&\leq \frac{1}{n} \sup_{h \in \mathcal{H}} \left\{ \sum_{i=1}^n 2\epsilon_i \xi_i (h(X_i) - v[h]) - C(h(X_i)^2 - v[h]^2) \right\} + \frac{1}{n} \sup_{v \in \mathcal{N}_2(\mathcal{H}, \delta)} \left\{ \sum_{i=1}^n 2\epsilon_i \xi_i v - C v^2 \right\} \\
&\leq 2(\sqrt{\sum_{i=1}^n \xi_i^2 / n} + 2C) \cdot \delta + \frac{1}{n} \sup_{v \in \mathcal{N}_2(\mathcal{H}, \delta)} \left\{ \sum_{i=1}^n 2\epsilon_i \xi_i v - C v^2 \right\}.
\end{aligned}$$

The second term is the offset Rademacher for a finite set of cardinality at most $\log(16/\delta) + 3 \log N$, thus applying Lemma 3.1.2,

$$\begin{aligned}
\mathbb{E}_\epsilon \frac{1}{n} \sup_{h \in \mathcal{H}} \left\{ \sum_{i=1}^n 2\epsilon_i \xi_i h(X_i) - C h(X_i)^2 \right\} &\leq \inf_{\delta > 0} \left\{ K \cdot \delta + M \cdot \frac{3 \log N + \log(16/\delta)}{n} \right\} \\
&\leq \tilde{C} \cdot \frac{\log(N \vee n)}{n}
\end{aligned}$$

where $K := 2(\sqrt{\sum_{i=1}^n \xi_i^2 / n} + 2C)$ and M is defined in Equation (3.6). We also have the high probability bound via Lemma 3.1.2:

$$\mathbb{P}_\epsilon \left(\frac{1}{n} \sup_{h \in \mathcal{H}} \left\{ \sum_{i=1}^n 2\epsilon_i \xi_i h(X_i) - C h(X_i)^2 \right\} \leq \tilde{C} \cdot \frac{\log(N \vee n) + u}{n} \right) \leq e^{-u}.$$

□

A.5. Appendix for Section 3.2

The proofs of the main results are divided into several parts. For the upper bound of point estimation, we will first prove Theorem 3.2.4 and then two lemmas, Lemma 3.2.3 and Lemma 3.2.2 (these two Lemmas are included in Supplement Cai et al. (2014). Theorem

3.2.1 is then easy to prove. As for the statistical inference, Theorem 3.2.2 is proved based on Theorem 3.2.1. For the lower bound of point estimation, Theorem 3.2.3 is a direct result combining Lemma 3.2.3 and Theorem 3.2.6, which is proved in Supplement Cai et al. (2014). Proofs of Corollaries are deferred to Supplement Cai et al. (2014).

Proof of Theorem 3.2.4. We will prove a stronger version of the Theorem, analyzing both (3.26) and (3.27). The proof is clean and in a general fashion, following directly from the assumptions of the theorem and the definitions:

$$\begin{aligned} \|\mathcal{X}^*(Y - \mathcal{X}M)\|_{\mathcal{A}}^* \leq \lambda, \|\mathcal{X}^*(Y - \mathcal{X}M)\|_{\ell_2} \leq \mu &\quad \text{Assumption of the Theorem} \\ \|\mathcal{X}^*(Y - \mathcal{X}\hat{M})\|_{\mathcal{A}}^* \leq \lambda, \|\mathcal{X}^*(Y - \mathcal{X}\hat{M})\|_{\ell_2} \leq \mu &\quad \text{Constraint in program} \\ \|\hat{M}\|_{\mathcal{A}} \leq \|M\|_{\mathcal{A}} &\quad \text{Definition of minimizer} \end{aligned}$$

Thus we have

$$\|\mathcal{X}^*\mathcal{X}(\hat{M} - M)\|_{\mathcal{A}}^* \leq 2\lambda, \|\mathcal{X}^*\mathcal{X}(\hat{M} - M)\|_{\ell_2} \leq 2\mu \ \text{ and } \ \hat{M} - M \in T_{\mathcal{A}}(M). \tag{A.64}$$

The first equation is due to triangle inequality and second one due to Tangent cone definition. Define $H = \hat{M} - M \in T_{\mathcal{A}}(M)$. According to the "Cauchy-Schwarz" (3.17) relation between atomic norm and its dual,

$$\|\mathcal{X}(H)\|_{\ell_2}^2 = \langle \mathcal{X}(H), \mathcal{X}(H) \rangle = \langle \mathcal{X}^*\mathcal{X}(H), H \rangle \leq \|\mathcal{X}^*\mathcal{X}(H)\|_{\mathcal{A}}^* \|H\|_{\mathcal{A}}.$$

Using the earlier result $\|\mathcal{X}^*\mathcal{X}(H)\|_{\mathcal{A}}^* \leq 2\lambda$, as well as the following two equations for any $H \in T_{\mathcal{A}}(M)$

$$\begin{aligned} \phi_{\mathcal{A}}(M, \mathcal{X})\|H\|_{\ell_2} \leq \|\mathcal{X}(H)\|_{\ell_2} &\qquad \text{local isometry constant} \\ \|H\|_{\mathcal{A}} \leq \gamma_{\mathcal{A}}(M)\|H\|_{\ell_2} &\qquad \text{local asphericity ratio} \end{aligned}$$

we get the following self-bounding relationship

$$\begin{aligned}
&\phi_{\mathcal{A}}^2(M,\mathcal{X})\|H\|_{\ell_2}^2 \leq \|\mathcal{X}(H)\|_{\ell_2}^2 \leq 2\lambda\|H\|_{\mathcal{A}} \leq 2\lambda\gamma_{\mathcal{A}}(M)\|H\|_{\ell_2}, \\
&\phi_{\mathcal{A}}^2(M,\mathcal{X})\|H\|_{\ell_2}^2 \leq \|\mathcal{X}(H)\|_{\ell_2}^2 \leq 2\mu\|H\|_{\ell_2}.
\end{aligned}$$

Thus $\|H\|_{\ell_2} \leq \frac{2}{\phi_{\mathcal{A}}^2(M,\mathcal{X})}\min\{\gamma_{\mathcal{A}}(M)\lambda, \mu\}$. The proof is then completed by simple algebra. Note here under the Gaussian setting, we can plug in $\lambda \asymp w(\mathcal{X}\mathcal{A})/\sqrt{n}$ and $\mu \asymp w(\mathcal{X}B_2^p)/\sqrt{n}$ using Lemma 3.2.2. □

Proof of Theorem 3.2.1. Theorem 3.2.1 is a special case of Theorem 3.2.4 under Gaussian setting, combining with Lemma 3.2.3 and Lemma 3.2.2. All we need to show is a good control of λ_n and $\phi_{\mathcal{A}}(M,\mathcal{X})$ with probability at least $1-3\exp(-\delta^2/2)$ under Gaussian ensemble and Gaussian noise. We bound λ_n with probability at least $1-\exp(-\delta^2/2)$ via Lemma 3.2.2. For $\phi_{\mathcal{A}}(M,\mathcal{X})$, we can lower bound by $1-c$ with probability at least $1-2\exp(-\delta^2/2)$. Let's define good event to be when

$$\lambda_n \leq \frac{\sigma}{\sqrt{n}}\left\{\mathbb{E}_g\left[\sup_{v\in\mathcal{A}}\langle g, \mathcal{X}v\rangle\right] + \delta\cdot\sup_{v\in\mathcal{A}}\|\mathcal{X}v\|_{\ell_2}\right\}$$

and $1-c \leq \phi_{\mathcal{A}}(M,\mathcal{X}) \leq \psi_{\mathcal{A}}(M,\mathcal{X}) \leq 1+c$ both hold. It is easy to see this good event holds with probability $1-3\exp(-\delta^2/2)$. Thus all we need to prove is $\max_{z\in\mathcal{A}}\|\mathcal{X}z\| \leq 1+c$ under the good event.

According to Lemma 3.2.3, $\max_{z\in\mathcal{A}}\|\mathcal{X}z\|/\|z\| \leq 1+c$ is satisfied under the condition $n \geq \frac{1}{c^2}[w(B_2^p\cap\mathcal{A})+\delta]^2$. As we know for any M, the unit atomic norm ball $\mathsf{conv}(\mathcal{A})$ is contained in $2B_2^p$ and $T_{\mathcal{A}}(M)$, which means $B_2^p\cap\mathcal{A}\subset 2B_2^p\cap T_{\mathcal{A}}(M)$, thus $w(B_2^p\cap\mathcal{A}) \leq 2w(B_2^p\cap T_{\mathcal{A}}(M))$ (monotonic property of Gaussian width). So we have for any M, if $n \geq \frac{4}{c^2}[w(B_2^p\cap T_{\mathcal{A}}(M))+\delta]^2 \vee \frac{1}{c}$. we have the following two bounds with probability at least

$1 - 2\exp(-\delta^2/2)$

$$\max_{z\in\mathcal{A}} \|\mathcal{X}z\| \leq 1 + c$$

$$1 - c \leq \phi_{\mathcal{A}}(M, \mathcal{X}) \leq \psi_{\mathcal{A}}(M, \mathcal{X}) \leq 1 + c. \tag{A.65}$$

Now plugging (A.65) into the expression of Lemma 3.2.2, together with Lemma 3.2.3, Theorem 3.2.4 reduces to Theorem 3.2.1. □

Proof of Theorem 3.2.2. We first prove that, with high probability, the convex program (3.28) is indeed feasible with $\Omega = I_n$. Equivalently we establish that, with high probability, for any $1 \leq i \leq p$, $\|\mathcal{X}^*\mathcal{X}e_i - e_i\|_{\mathcal{A}}^* \leq \eta$ for some proper choice of η. Here $\mathcal{X} \in \mathbb{R}^{n\times p}$, and the entries $\mathcal{X}_{ij} \overset{iid}{\sim} N(0, 1/n)$. Denote $g = \sqrt{n}\mathcal{X}_{\cdot i}$ as a scaling version of the i-th column of $\mathcal{X}$, $g \sim N(0, I_n)$ and $g' \sim N(0, I_n)$ being an independent copy. Below $O_p(\cdot)$ denotes the asymptotic order in probability. We have, for all $1 \leq i \leq p$ uniformly,

$$\begin{aligned}
\|\mathcal{X}^*\mathcal{X}e_i - e_i\|_{\mathcal{A}}^* &= \sup_{v\in\mathcal{A}} \langle \mathcal{X}^*\mathcal{X}e_i - e_i, v\rangle = \sup_{v\in\mathcal{A}} \langle \mathcal{X}^* g - e_i, v\rangle/\sqrt{n} \\
&\leq \sup_{v\in\mathcal{A}} \langle \mathcal{X}_{(-i)}^* g, v\rangle/\sqrt{n} + \sup_{v\in\mathcal{A}} \left(\frac{1}{n}\sum_{j=1}^n g_j^2 - 1\right) v_i \\
&\overset{w.h.p}{\precsim} \frac{w(\mathcal{X}_{(-i)}\mathcal{A})}{\sqrt{n}} + O_p(\sqrt{\log p/n}) \quad \text{invoking Lemma 3.2.2} \\
&\leq \frac{w(\mathcal{X}\mathcal{A})}{\sqrt{n}} + \frac{\mathbb{E}_{g'} \sup_{v\in\mathcal{A}} \sum_{k=1}^n g'_k \mathcal{X}_{ki}(-v_i)}{\sqrt{n}} + O_p(\sqrt{\log p/n}) \\
&\leq \frac{w(\mathcal{X}\mathcal{A})}{\sqrt{n}} + \frac{\sqrt{\mathbb{E}_{g'}(\sum_{k=1}^n g'_k \mathcal{X}_{ki})^2 \cdot \sup_{v\in\mathcal{A}} v_i^2}}{\sqrt{n}} + O_p(\sqrt{\log p/n}) \\
&\leq \frac{w(\mathcal{X}\mathcal{A})}{\sqrt{n}} + \sqrt{\frac{1 + O_p(\sqrt{\log p/n})}{n}} + O_p(\sqrt{\log p/n})
\end{aligned} \tag{A.66}$$

where $\mathcal{X}_{(-i)}$ is the linear operator setting i-th column to be all zeros. We applied Lemma 3.2.2 in establishing the above bounds.

For the de-biased estimate $\tilde{M}$, we have $\tilde{M} = \hat{M} + \Omega\mathcal{X}^*(Y - \mathcal{X}(\hat{M}))$ and $\tilde{M} - M = (\Omega\mathcal{X}^*\mathcal{X} -$

$I_p)(M-\hat{M})+\Omega\mathcal{X}^*Z := \Delta + \frac{\sigma}{\sqrt{n}}\Omega\mathcal{X}^*W$. Then for any $1 \le i \le p$, from the Cauchy-Schwartz relationship (3.17),

$$|\Delta_i| = |\langle \mathcal{X}^*\mathcal{X}\Omega_{i\cdot}^* - e_i, M-\hat{M}\rangle| \le \|\mathcal{X}^*\mathcal{X}\Omega_{i\cdot}^* - e_i\|_{\mathcal{A}}^* \|M-\hat{M}\|_{\mathcal{A}} \le \sigma\frac{\gamma_{\mathcal{A}}^2(M)w^2(\mathcal{X}\mathcal{A})}{n}. \tag{A.67}$$

The last line invokes the consistency result in Theorem 3.2.1, $\|\hat{M}-M\|_{\mathcal{A}} \precsim \sigma\frac{\gamma_{\mathcal{A}}^2(M)w(\mathcal{X}\mathcal{A})}{\sqrt{n}}$. Thus we have $\|\Delta\|_{\ell_\infty} \precsim \sigma\frac{\gamma_{\mathcal{A}}^2(M)w^2(\mathcal{X}\mathcal{A})}{n}$. For any linear contrast $\|v\|_{\ell_1} \le \rho$, we have $\frac{\sqrt{n}}{\sigma}v^*(\tilde{M}-M) = v^*\Omega\mathcal{X}^*W + \frac{\sqrt{n}}{\sigma}v^*\Delta$,

$$\limsup_{n,p(n)\to\infty} \frac{\sqrt{n}}{\sigma}v^*\Delta \le \limsup_{n,p(n)\to\infty} \frac{\sqrt{n}}{\sigma}\|v\|_{\ell_1}\|\Delta\|_{\ell_\infty} \le \rho\cdot \limsup_{n,p(n)\to\infty} \frac{\gamma_{\mathcal{A}}^2(M)w^2(\mathcal{X}\mathcal{A})}{\sqrt{n}} = 0,$$

and $v^*\Omega\mathcal{X}^*W \sim N(0, v^*[\Omega\mathcal{X}^*\mathcal{X}\Omega^*]v)$. □

Proof of Theorem 3.2.3. Theorem 3.2.3 is a special case of Theorem 3.2.6, combining with Lemma 3.2.3 (both in Supplement Cai et al. (2014). Plug in the general convex cone T by local tangent cone $T_{\mathcal{A}}(M)$, then upper bound $\psi_{\mathcal{A}}(M,\mathcal{X}) \le 1+c$ with high probability via Lemma 3.2.3. □

A.6. Appendix for Section 3.3

A.6.1. Proofs for Realizable Setting

Proof of Lemma 3.3.1. Let $\Delta := \widehat{w} - w^*$ be the difference between the true answer and solution to the optimization problem. Let S to be the support of w^* and let $S^c = [d]\backslash S$ be the complements of S. Consider the permutation $i_1, \ldots, i_{d-k}$ of S^c for which $|\Delta(i_j)| \ge |\Delta(i_{j+1})|$ for all j. That is, the permutation dictated by the magnitude of the entries of Δ outside of S. We split S^c into subsets of size k according to this permutation: Define S_j, for $j \ge 1$ as $\{i_{(j-1)k+1}, \ldots, i_{jk}\}$. For convenience we also denote by S_{01} the set $S \cup S_1$.

Now, consider the matrix $X_{S_{01}} \in \mathbb{R}^{t\times|S_{01}|}$ whose columns are those of X with indices S_{01}.

The Restricted Isometry Property of X dictates that for any vector $c \in \mathbb{R}^{S_{01}}$,

$$(1-\epsilon)\left\|c\right\|_2 \leq \frac{1}{\sqrt{n}}\left\|X_{S_{01}}c\right\|_2 \leq (1+\epsilon)\left\|c\right\|_2 .$$

Let $V \subseteq \mathbb{R}^t$ be the subspace of dimension $|S_{01}|$ that is the image of the linear operator $X_{S_{01}}$, and let $P_V \in \mathbb{R}^{t \times t}$ be the projection matrix onto that subspace. We have, for any vector $z \in \mathbb{R}^t$ that

$$(1-\epsilon)\left\|P_V z\right\| \leq \frac{1}{\sqrt{n}}\left\|X_{S_{01}}^T z\right\| \leq (1+\epsilon)\left\|P_V z\right\|$$

We apply this to $z = X\Delta$ and conclude that

$$\left\|P_V X\Delta\right\| \leq \frac{1}{\sqrt{t}(1-\epsilon)}\left\|X_{S_{01}}^T X\Delta\right\| \tag{A.68}$$

We continue to lower bound the quantity of $\|P_V X\Delta\|$. We decompose $P_V X\Delta$ as

$$P_V X\Delta = P_V X\Delta(S_{01}) + \sum_{j\geq 2} P_V X\Delta(S_j) \tag{A.69}$$

Now, according to the definition of V we that there exist vectors $\{c_j\}_{j\geq 2}$ in $\mathbb{R}^{|S_{01}|}$ for which

$$P_V X\Delta(S_j) = X_{S_{01}}c_j$$

We now invoke Lemma 1.1 from Candes and Tao (2005) stating that for any S', S'' with $|S'| + |S''| \leq 3k$ it holds that

$$\forall c, c' \quad \frac{1}{n}\left\langle X_{S'}c, X_{S''}c'\right\rangle \leq (2\epsilon - \epsilon^2)\left\|c\right\|_2 \left\|c'\right\|_2$$

We apply this for S_{01}, S_j, $j \geq 2$ and conclude that

$$\left\|P_V X\Delta(S_j)\right\|_2^2 = \langle P_V X\Delta(S_j), X\Delta(S_j)\rangle \leq 2\epsilon t \left\|c_j\right\|_2 \cdot \|\Delta(S_j)\| \leq \frac{2\epsilon\sqrt{t}}{1-\epsilon}\left\|P_V X\Delta(S_j)\right\|_2 \cdot \left\|\Delta(S_j)\right\|_2 .$$

Dividing through by $\|P_V X\Delta(S_j)\|_2$, we get

$$\|P_V X\Delta(S_j)\| \leq \frac{2\epsilon\sqrt{t}}{1-\epsilon}\|\Delta(S_j)\|. \tag{A.70}$$

Let us now bound the sum $\|\Delta(S_j)\|$. By the definition of S_j we know that any element $i \in S_j$ has the property $\Delta(i) \leq (1/k)\|\Delta(S_{j-1})\|_1$. Hence

$$\sum_{j\geq 2}\|\Delta(S_j)\| \leq (1/\sqrt{k})\sum_{j\geq 1}\|\Delta(S_j)\|_1 = (1/\sqrt{k})\|\Delta(S^c)\|_1$$

We now combine this inequality with Equations (A.68), (A.69) and (A.70)

$$\begin{aligned}
\frac{1}{t}\left\|X_{S_{01}}^T X\Delta\right\| &\geq \frac{1-\epsilon}{\sqrt{t}}\|P_V X\Delta\| \\
&\geq \frac{1-\epsilon}{\sqrt{t}}\|P_V X\Delta(S_{01})\| - \frac{1-\epsilon}{\sqrt{n}}\sum_{j\geq 2}\|P_V X\Delta(S_j)\| \\
&\geq \frac{1-\epsilon}{\sqrt{t}}\|X\Delta(S_{01})\| - 2\epsilon\sum_{j\geq 2}\|\Delta(S_j)\| \\
&\geq \frac{1-\epsilon}{\sqrt{t}}\|X\Delta(S_{01})\| - \frac{2\epsilon}{\sqrt{k}}\|\Delta(S^c)\|_1
\end{aligned}$$

The third inequality holds since $X\Delta(S_{01}) \in V$ hence $P_V X\Delta(S_{01)} = X\Delta(S_{01})$. We continue to bound the expression by claiming that $\|\Delta(S)\|_1 \geq \|\Delta(S^c)\|_1$. This holds since in S^c, $\widehat{w}_{S^c} = \Delta(S^c)$ hence

$$\|w^*\|_1 = \|\widehat{w} - \Delta(S^c) - \Delta(S)\|_1 \leq \|\widehat{w}\|_1 + (\|\Delta(S)\|_1 - \|\Delta(S^c)\|_1)$$

Now, the optimality of $\widehat{w}$ implies $\|\widehat{w}\|_1 \leq \|w^*\|_1$, hence indeed $\|\Delta(S)\|_1 \geq \|\Delta(S^c)\|_1$.

$$\|\Delta(S^c)\|_1 \leq \|\Delta(S)\|_1 \leq \sqrt{k}\|\Delta(S)\|_2 \leq \|\Delta(S_{01})\|_2 \leq \frac{\sqrt{k}}{(1-\epsilon)\sqrt{t}}\|X\Delta(S_{01})\|$$

We continue the chain of inequalities

$$\begin{aligned}\frac{1}{t}\left\|X_{S_{01}}^T X\Delta\right\| &\geq \frac{1-\epsilon}{\sqrt{n}}\left\|X\Delta(S_{01})\right\| - \frac{2\epsilon}{\sqrt{k}}\left\|\Delta(S^c)\right\|_1 \\ &\geq \left\|X\Delta(S_{01})\right\|\left(\frac{1-\epsilon}{\sqrt{n}} - \frac{2\epsilon}{\sqrt{k}}\cdot\frac{\sqrt{k}}{(1-\epsilon)\sqrt{n}}\right) \\ &= \frac{(1-\epsilon)^2-2\epsilon}{(1-\epsilon)\sqrt{t}}\left\|X\Delta(S_{01})\right\|\end{aligned}$$

Rearranging we conclude that

$$\begin{aligned}\left\|\Delta(S_{01})\right\| &\leq \frac{1}{(1-\epsilon)\sqrt{t}}\left\|X\Delta(S_{01})\right\| && \text{(RIP of } X\text{)} \\ &\leq \frac{1}{((1-\epsilon)^2-2\epsilon)t}\left\|X_{S_{01}}^T X\Delta\right\| \\ &\leq \frac{\sqrt{2k}}{(1-4\epsilon)t}\left\|X^T X\Delta\right\|_\infty && \text{(since for any } z\in\mathbb{R}^{2k},\ \|z\|_2\leq\sqrt{2k}\,\|z\|_\infty\text{)} \\ &\leq C\sqrt{\frac{dk\log(d/\delta)}{tk_0}}\left(\sigma+\frac{d}{k_0}\left\|w^*\right\|_1\right) && \text{(Lemma A.6.1 and } \epsilon<1/5\text{)}\end{aligned}$$

for some constant C. We continue our bound on $\|\Delta\|$ by showing that $\|\Delta(S_{01}^c)\| \leq \|\Delta(S_{01})\|$

$$\left\|\Delta(S_{01}^c)\right\|_2^2 \overset{(i)}{\leq} \left\|\Delta(S^c)\right\|_1^2\cdot\sum_{j\geq k+1}\frac{1}{j^2} \leq \frac{1}{k}\left\|\Delta(S^c)\right\|_1^2 \leq \frac{1}{k}\left\|\Delta(S)\right\|_1^2 \leq \left\|\Delta(S)\right\|_2^2.$$

Inequality (i) holds due to the following: Let α_i be the absolute value of the i'th largest (in absolute value) element of $\Delta(S^c)$. It obviously holds that $\alpha_i \leq \left\|\Delta(S^c)\right\|_1/i$. Now, according to the definition of S_{01} we have that $\left\|\Delta(S_{01}^c)\right\|_2^2 = \sum_{j\geq k+1}\alpha_i^2$ and the inequality follows. Hence,

$$\left\|\Delta(S_{01}^c)\right\|_2 \leq \left\|\Delta(S)\right\|_2 \leq \left\|\Delta(S_{01})\right\|_2.$$

We conclude that

$$\|\Delta\|_2 \leq \sqrt{2}\left\|\Delta(S_{01})\right\|_2 \leq C\sqrt{\frac{dk\log(d/\delta)}{tk_0}}\left(\sigma+\frac{d}{k_0}\left\|w^*\right\|_1\right)$$

for some universal constant $C > 0$. Since $\|\Delta(S)\|_1 \geq \|\Delta(S^c)\|_1$ and $|S| \leq k$ we get that

$$\|\Delta\|_1 \leq 2\|\Delta(S)\|_1 \leq 2\sqrt{k}\,\|\Delta(S)\|_2 \leq 2\sqrt{k}\,\|\Delta\|_2$$

and the claim follows. □

Proof of Lemma 3.3.2. Let S be the support of w^*. We can decompose the square of the left hand side as

$$\left\|\widehat{w}(\widetilde{S}) - w^*\right\|_2^2 = \sum_{i \in S \cap \widetilde{S}} (\widehat{w}(i) - w^*(i))^2 + \sum_{i \in \widetilde{S} \setminus S} (\widehat{w}(i))^2 + \sum_{i \in S \setminus \widetilde{S}} (w^*(i))^2.$$

We upper bound the last sum on the right hand side as

$$\begin{aligned}
\sum_{i \in S \setminus \widetilde{S}} (w^*(i))^2 &= \sum_{i \in S \setminus \widetilde{S}} \left[(\widehat{w}(i) - w^*(i)) + (\widehat{w}(i))\right]^2 \\
&\leq 2 \sum_{i \in S \setminus \widetilde{S}} (\widehat{w}(i) - w^*(i))^2 + (\widehat{w}(i))^2 \\
&\leq 2 \sum_{i \in S \setminus \widetilde{S}} (\widehat{w}(i) - w^*(i))^2 + 2 \sum_{i \in \widetilde{S} \setminus S} (\widehat{w}(i))^2 \,,
\end{aligned}$$

where first inequality follows from the elementary inequality $(a+b)^2 \leq 2a^2 + 2b^2$ and the second inequality is due to the fact that $\widetilde{S}$ contains top k entries of $\widehat{w}$ in absolute value and

$|S \setminus \widetilde{S}| = |\widetilde{S} \setminus S|$. Hence,

$$\begin{aligned}
\left\| \widehat{w}(\widetilde{S}) - w^* \right\|_2^2 &= \sum_{i \in S \cap \widetilde{S}} (\widehat{w}(i) - w^*(i))^2 + \sum_{i \in \widetilde{S} \setminus S} (\widehat{w}(i))^2 + \sum_{i \in S \setminus \widetilde{S}} (w^*(i))^2 \\
&\leq \sum_{i \in S \cap \widetilde{S}} (\widehat{w}(i) - w^*(i))^2 + 2 \sum_{i \in S \setminus \widetilde{S}} (\widehat{w}(i) - w^*(i))^2 + 3 \sum_{i \in \widetilde{S} \setminus S} (\widehat{w}(i))^2 \\
&\leq 2 \sum_{i \in S \cap \widetilde{S}} (\widehat{w}(i) - w^*(i))^2 + 2 \sum_{i \in S \setminus \widetilde{S}} (\widehat{w}(i) - w^*(i))^2 + 3 \sum_{i \in \widetilde{S} \setminus S} (\widehat{w}(i))^2 \\
&= 2 \sum_{i \in S} (\widehat{w}(i) - w^*(i))^2 + 3 \sum_{i \in \widetilde{S} \setminus S} (\widehat{w}(i))^2 \\
&\leq 3 \sum_{i=1}^{d} (\widehat{w}(i) - w^*(i))^2 \\
&= 3 \left\| \widehat{w} - w^* \right\|_2^2 .
\end{aligned}$$

Taking square root finishes the proof. □

Lemma A.6.1. *There exists a universal constant $C > 0$ such that, with probability at least $1 - \delta$, the convex program* (3.41) *is feasible and its optimal solution $\widehat{w}$ satisfies*

$$\left\| \frac{1}{t} X_t^T X_t (\widehat{w} - w^*) \right\|_\infty \leq C \sqrt{\frac{d \log(d/\delta)}{t k_0}} \left(\sigma + \frac{d}{k_0} \left\| w^* \right\|_1 \right) .$$

We note that the above lemma is beyond simple triangle inequality on the feasibility constraints, as the left hand side depends on actual design matrix X_t which we do not observe, instead of $\widehat{X}_t$.

Proof. To simplify notation, we drop subscript t. Namely, let $X = X_t$, $\widehat{X} = X_t$ and $\widehat{D} = \widehat{D}_t$, and also let $\eta = (\eta_1, \eta_2, \ldots, \eta_t)$ be the vector of noise variables.

First, we show that w^* satisfies the constraint of (3.41) with probability at least $1 - \delta$. We upper bound

$$\begin{aligned}\left\|\frac{1}{t}\widehat{X}^T(Y-\widehat{X}w^*)+\frac{1}{t}\widehat{D}w^*\right\|_\infty &= \left\|\left[\frac{1}{t}\widehat{X}^T(X-\widehat{X})+\frac{1}{t}\widehat{D}\right]w^*+\frac{1}{t}\widehat{X}^T\eta\right\|_\infty \\ &\le \left\|\left[\frac{1}{t}\widehat{X}^T(X-\widehat{X})+\frac{1}{t}\widehat{D}\right]w^*\right\|_\infty+\frac{1}{t}\left\|\widehat{X}^T\eta\right\|_\infty\end{aligned}$$

We first bound the left summand. By Lemma A.6.2, we have

$$\begin{aligned}\left\|\left[\frac{1}{t}\widehat{X}^T(X-\widehat{X})+\frac{1}{t}\widehat{D}\right]w^*\right\|_\infty &\le \|w^*\|_1\cdot\left\|\frac{1}{t}\widehat{X}^T(X-\widehat{X})+\frac{1}{t}\widehat{D}\right\|_\infty \\ &\le \|w^*\|_1\left(\left\|\frac{1}{t}X^T(\widehat{X}-X)\right\|_\infty+\left\|\frac{1}{t}(\widehat{X}-X)^T(\widehat{X}-X)-\frac{1}{t}\widehat{D}\right\|_\infty\right) \\ &\le \|w^*\|_1\,C\cdot\sqrt{\frac{d^3\log(d/\delta)}{t{k_0}^3}}.\end{aligned}$$

For the right summand, since η is vector of i.i.d Gaussians with variance σ^2, with probability at least $1-\delta$,

$$\frac{1}{t}\left\|\widehat{X}^T\eta\right\|_\infty \le C\frac{\sigma}{t}\sqrt{\log(d/\delta)}\cdot\max_{i\in[d]}\left\|\widehat{X}_{(i)}\right\|_2$$

where $\widehat{X}_{(1)},\widehat{X}_{(2)},\ldots,\widehat{X}_{(d)}$ are the columns of $\widehat{X}$. Since the absolute value of the entries of $\widehat{X}$ is at most d/k_0, we have $\left\|\widehat{X}_{(i)}\right\|_2\le\sqrt{td/k_0}$ and thus

$$\frac{1}{t}\left\|\widehat{X}^T\eta\right\|_\infty \le C\sigma\sqrt{\frac{d\log(d/\delta)}{tk_0}}.$$

Combining the inequalities so far provides

$$\left\|\frac{1}{t}\widehat{X}^T(Y-\widehat{X}w^*)+\frac{1}{t}\widehat{D}w^*\right\|_\infty \le C\sqrt{\frac{d\log(d/\delta)}{tk_0}}\left(\sigma+\frac{d}{k_0}\|w^*\|_1\right)$$

and hence conclude the constraint of the optimization problem (3.41) is satisfied (at least) by w^* and thus the optimization problem is feasible.

Now consider the vector $\Delta := \widehat{w} - w^*$, we have

$$\begin{aligned}
\left\|\frac{1}{t}X^T X\Delta\right\|_\infty &\le \left\|\frac{1}{t}(\widehat{X}^T\widehat{X} - \widehat{D})\Delta\right\|_\infty + \left\|\frac{1}{t}(\widehat{X}^T\widehat{X} - \widehat{D} - X^T X)\Delta\right\|_\infty \\
&\le \left\|\frac{1}{t}(\widehat{X}^T\widehat{X} - \widehat{D})\Delta\right\|_\infty + \left\|\frac{1}{t}(\widehat{X} - X)^T X\Delta\right\|_\infty \\
&\quad + \left\|\frac{1}{t}X^T(\widehat{X} - X)\Delta\right\|_\infty + \left\|\left(\frac{1}{t}(\widehat{X} - X)^T(\widehat{X} - X) - \frac{1}{t}\widehat{D}\right)\Delta\right\|_\infty .
\end{aligned}$$

According to Lemma A.6.2 we have

$$\begin{aligned}
&\left\|\frac{1}{t}X^T(\widehat{X} - X)\Delta\right\|_\infty \le \left\|\frac{1}{t}X^T(\widehat{X} - X)\right\|_\infty \|\Delta\|_1 \\
&\le C\sqrt{\frac{d\log(d/\delta)}{tk_0}}(\|w^*\|_1 + \|\widehat{w}\|_1) \le 2C\sqrt{\frac{d\log(d/\delta)}{tk_0}} \cdot \|w^*\|_1
\end{aligned}$$

where the last inequality is by the optimality of $\widehat{w}$. The same argument provides an identical bound for $\left\|\frac{1}{t}(\widehat{X} - X)^T X\Delta\right\|_\infty$. The last summand can also be bounded by using Lemma A.6.2 and the optimality of $\widehat{w}$.

$$\left\|\left(\frac{1}{t}(\widehat{X} - X)^T(\widehat{X} - X) - \frac{1}{t}\widehat{D}\right)\Delta\right\|_\infty \le 2C\sqrt{\frac{d^3\log(d/\delta)}{t{k_0}^3}} \cdot \|w^*\|_1$$

Finally, according to the feasibility of $\widehat{w}$ and w^* we may bound the first summand

$$\left\|\left(\frac{1}{t}\widehat{X}^T\widehat{X} - \frac{1}{t}\widehat{D}\right)\Delta\right\|_\infty \le 2C\sqrt{\frac{d\log(d/\delta)}{tk_0}}\left(\sigma + \frac{d}{k_0}\|w^*\|_1\right),$$

and reach the final bound. □

Lemma A.6.2. *For any $t \ge t_0$, with probability at least $1-\delta$, the following two inequalities hold*

$$\begin{aligned}
\left\|\frac{1}{t}(\widehat{X}_t - X_t)^T(\widehat{X}_t - X_t) - \frac{1}{t}\widehat{D}_t\right\|_\infty &\le C\sqrt{\frac{d^3\log(d/\delta)}{t{k_0}^3}}\,, \\
\left\|\frac{1}{t}X_t^T(\widehat{X}_t - X_t)\right\|_\infty &\le C\sqrt{\frac{d\log(d/\delta)}{tk_0}}\,,
\end{aligned}$$

where $\|\cdot\|_\infty$ *denotes the maximum of the absolute values of the entries of a matrix.*

Proof. Throughout we use that $|x_s(i)| \leq 1$ for all $i \in [d]$ and all $s \in [t]$, and (2) $(\widehat{x}_s(i) - x_s(i))^2 - \frac{1}{t}D_{ii}$ is unbiased with absolute value of at most $(d/k_0)^2$ and variance of at most $(d/k_0)^3$. For the first term, let's bound

$$\left[\frac{1}{t}(\widehat{X} - X)^T(\widehat{X} - X) - \frac{1}{t}\widehat{D}\right]_{ij} = \frac{1}{t}\sum_{s=1}^{t}(\widehat{x}_s(i) - x_s(i))(\widehat{x}_s(j) - x_s(j)) - \frac{1}{t}\widehat{D}_{ij}$$

For $i = j$, we have

$$\mathbf{E}\left[\left((\widehat{x}_s(i) - x_s(i))^2 - \frac{1}{t}D_{ii}\right)^2\right] \leq \mathbf{E}\left[(\widehat{x}_s(i) - x_s(i))^4\right] \leq (d/k_0)^3$$

$$(\widehat{x}_s(i) - x_s(i))^2 - \frac{1}{t}D_{ii} \leq (d/k_0)^2, \quad \mathbf{E}\left[(\widehat{x}_s(i) - x_s(i))^2 - \frac{1}{t}D_{ii}\right] = 0$$

Hence, by Bernstein's inequality, for any $v > 0$,

$$\Pr\left[\left|\frac{1}{t}\sum_{s=1}^{t}(\widehat{x}_s(i) - x_s(i))^2 - \frac{1}{t}D_{ii}\right| > v\right] \leq 2\exp\left(-\frac{v^2 t}{(d/k_0)^3 + (d/k_0)^2 v/3}\right).$$

It follows that for any $\delta > 0$, with probability at least $1 - \delta$ it holds for all $i \in [d]$ that,

$$\left|\frac{1}{t}\sum_{s=1}^{t}(\widehat{x}_s(i) - x_s(i))^2 - \frac{1}{t}D_{ii}\right| \leq \mathcal{O}\left(\frac{\log(d/\delta)d^2}{t{k_0}^2} + \sqrt{\frac{\log(d/\delta)d^3}{t{k_0}^3}}\right).$$

Similarly we have $\frac{1}{t}(\widehat{D}_{ii} - D_{ii}) \leq \mathcal{O}\left(\frac{\log(d/\delta)d^2}{t{k_0}^2} + \sqrt{\frac{\log(d/\delta)d^3}{t{k_0}^3}}\right)$.

For $i \neq j$ we use an analogous argument, only now the variance term in Bernstein's inequality is $(d/k_0)^2$ rather than $(d/k_0)^3$, hence only reach a tighter bound.

For the second term, we again bound via Bernstein's inequality as

$$\left[\frac{1}{t}X^T(\widehat{X} - X)\right]_{ij} = \frac{1}{t}\sum_{s=1}^{t}x_s(i)(\widehat{x}_s(j) - x_s(j)) \leq \mathcal{O}\left(\sqrt{\frac{d\log(d/\delta)}{tk_0}} + \frac{d\log(d/\delta)}{tk_0}\right)$$

The claim now follows by noticing that for large enough t, the dominating terms are those that scale as $1/\sqrt{t}$. □

Proof of Theorem 3.3.1. By Lemma 3.3.1,

$$\|w_{t+1} - w^*\|_2 \leq \mathcal{O}\left(\sqrt{\frac{d}{k_0}\frac{k\log(d/\delta)}{t}}(\sigma + \frac{d}{k_0}\|w^*\|_1)\right) .$$

We have

$$\begin{aligned}
\text{Regret}_T(w^*) - \text{Regret}_{t_0}(w^*) &= \sum_{t=t_0+1}^{T} (y_t - \langle x_t, w_t\rangle)^2 - (y_t - \langle x_t, w^*\rangle)^2 \\
&= \sum_{t=t_0+1}^{T} (\langle x_t, w^* - w_t\rangle + \eta_t)^2 - \eta_t^2 \\
&= \sum_{t=t_0+1}^{T} (\langle x_t, w^* - w_t\rangle + 2\eta_t)\, \langle x_t, w^* - w_t\rangle \\
&= \sum_{t=t_0+1}^{T} 2\eta_t \langle x_t, w^* - w_t\rangle + (\langle x_t, w^* - w_t\rangle)^2 \ ,
\end{aligned}$$

where we used that $y_t = \langle x_t, w_t\rangle + \eta_t$. To bound the regret we require the upper bound, that occurs with probability of at least $1 - \delta$, $\forall t \geq t_0$,

$$|\langle x_t, w^* - w_t\rangle| \overset{(i)}{\leq} \|x_t\|_\infty \sqrt{\|w_t - w^*\|_0} \cdot \|w_t - w^*\|_2 \overset{(ii)}{\leq} \mathcal{O}\left(k \cdot \sqrt{\frac{d}{k_0}\frac{\log(\log(T)d/\delta)}{t}}\left(\sigma + \frac{d}{k_0}\right)\right) .$$

Inequality (i) holds since $\langle a, b\rangle \leq \|a(S)\|_2 \cdot \|b\|_2$ with S being the support of b and $\|a(S)\|_2 \leq \|a\|_\infty \sqrt{|S|}$. Inequality (ii) follows from Lemma 3.3.1 and Lemma 3.3.2, and a union bound over the $\lceil \log(T) \rceil$ many times the vector w_t is updated. Now, for the left summand of the regret bound we have by Martingale concentration inequality that w.p. $1-\delta$

$$\begin{aligned}\sum_{t=t_0+1}^{T} 2\eta_t \langle x_t, w_t - w^* \rangle &\leq \mathcal{O}\left(\sigma \sqrt{\log(1/\delta) \sum_{t=t_0+1}^{T} \langle x_t, w_t - w^* \rangle^2}\right) \\ &= \mathcal{O}\left(\sigma \sqrt{\log(1/\delta)\log(T)k^2 \cdot \frac{d\log(d\log(T)/\delta)}{k_0}\left(\sigma + \frac{d}{k_0}\right)^2}\right).\end{aligned}$$

The right summand is bounded as

$$\sum_{t=t_0+1}^{T} \langle x_t, w^* - w_t \rangle^2 = \mathcal{O}\left(k^2 \cdot \frac{d\log(d\log(T)/\delta)}{k_0}\left(\sigma + \frac{d}{k_0}\right)^2 \cdot \log(T)\right).$$

Clearly, the right summand dominates the left one.

It remains to bound the regret in first t_0 rounds. Since $w_t = 0$ for $t \leq t_0$, we have

$$\mathrm{Regret}_{t_0}(w^*) = \sum_{t=1}^{t_0} 2\eta_t \langle x_t, w^* \rangle + (\langle x_t, w^* \rangle)^2 \leq \mathcal{O}\left(\sigma\sqrt{t_0 \log(1/\delta)} + t_0\right).$$

Here, we used that $|\langle x_t, w^* \rangle| \leq 1$ since $\|x_t\|_\infty \leq 1$ and $\|w^*\|_1 \leq 1$. We also used that $\eta_t \langle x_t, w^* \rangle \sim N(0, \sigma^2 \langle x_t, w^* \rangle^2)$ and $\eta_1 \langle x_1, w^* \rangle, \eta_2 \langle x_2, w^* \rangle, \ldots, \eta_{t_0} \langle x_{t_0}, w^* \rangle$ are independent. Thus their sum is a Gaussian with variance at most $\sigma^2 t_0$.

Collecting all the terms along with Lemma A.6.3, bounding the difference $\mathrm{Regret}_T - \mathrm{Regret}_T(w^*)$, gives

$$\mathrm{Regret}_T \leq \left(t_0 + \sqrt{t_0 \log(1/\delta)} + k^2 \cdot \frac{d\log(d\log(T)/\delta)}{k_0}\left(\sigma + \frac{d}{k_0}\right)^2 \cdot \log(T)\right) \tag{A.71}$$

□

Lemma A.6.3. *In the realizable case, w.p. at least $1-\delta$ we have for any sequence of w_t that* $\mathrm{Regret}_T - \mathrm{Regret}_T(w^*) = O(\sigma^2 k \log(d/\delta))$.

Proof. It is an easy exercise to show that $\mathrm{Regret}_T - \mathrm{Regret}_T(w^*)$ is equal to the regret on an algorithm that always plays w^*. We thus continue to bound the regret of w^*.

Let $\Delta \in \mathbb{R}^d$ be the difference between w^* and $\tilde{w}$, the empirical optimal solution for the sparse regression problem. The loss associated with w^* is clearly $\|\eta\|^2$, where η is the noise term $y = Xw^* + \eta$. The loss associated with $\tilde{w}$ is

$$\|X(w^* + \Delta) - Xw^* - \eta\|^2 = \|\eta - X\Delta\|^2 = \|\eta - X_{\tilde{S}}\Delta\|^2$$

where $\tilde{S}$ is the support of Δ, having a cardinality of at most $2k$. The closed form solution for the least-squares problem dictates that

$$\|\eta - X_{\tilde{S}}\Delta\|^2 \geq \|\eta - X_{\tilde{S}}X_{\tilde{S}}^{\dagger}\eta\|^2 = \|\eta\|^2 - \|X_{\tilde{S}}X_{\tilde{S}}^{\dagger}\eta\|^2 \ .$$

Here, $A^\dagger$ is the pseudo inverse of a matrix A and X_S is the matrix obtained from the columns of X whose indices are in S. It follows that the regret of w^* is bounded by

$$\|X_{\tilde{S}}X_{\tilde{S}}^{\dagger}\eta\|^2$$

for some subset $\tilde{S}$ of size at most $2k$. To bound this quantity we use a high probability bound for $\|X_S X_S^{\dagger}\eta\|^2$ for a fixed set S, and take a union bound over all possible sets of cardinality $2k$. For a fixed set S we have that $\|X_S X_S^{\dagger}\eta\|^2/\sigma^2$ is distributed according to the χ^2_{2k} distribution. The tail bounds of this distribution suggest that

$$\Pr\left[\|X_S X_S^{\dagger}\eta\|^2 > 2k\sigma^2 + 2\sigma^2\sqrt{2kx} + 2\sigma^2 x\right] \leq \exp(-x)$$

meaning that with probability at least $1 - \delta/d^{2k}$ we have

$$\|X_S X_S^{\dagger}\eta\|^2 < 2k\sigma^2 + 2\sigma^2\sqrt{2k \cdot 2k \cdot \log(d/\delta)} + 2\sigma^2 \cdot 2k \cdot \log(d/\delta) = O(\sigma^2 k \log(d/\delta))$$

Taking a union bound over all possible subsets of size $\leq 2k$ we get that w.p. at least $1 - \delta$ the regret of w^* is at most $O(\sigma^2 k \log(d/\delta))$. □

A.6.2. Proofs for Agnostic Setting

We begin with an auxiliary lemma for Lemma 3.3.3, informally proving that for any matrix $\bar{X}$ with BBRCNP (Definition 3.3.3) and vector y, the set function

$$g(S) = \inf_{w \in \mathbb{R}^S} \|y - \bar{X}w\|^2$$

is weakly supermodular. Its proof can be found in (Boutsidis et al., 2015), yet for completeness we provide it here as well.

Lemma A.6.4. *[Lemma 5 in (Boutsidis et al., 2015)] Let $\bar{X}$ be a matrix whose columns have 2-norm at most 1 and y be a vector with $\|y\|_\infty \leq 1$ of dimension matching the number of rows in X. the set function*

$$g(S) = \inf_{w \in \mathbb{R}^S} \|y - Xw\|^2$$

is α-weakly supermodular for sparsity k for $\alpha = \max_{S:|S|\leq k} 1/\sigma_{\min}(X_S)^2$, where X_S is the submatrix of X obtained by choosing the columns indexed by S, and $\sigma_{\min}(A)$ is the smallest singular value of A.

Proof. Firstly, the well known closed form solution for the least-squares problem informs us that

$$\begin{aligned} g(S) &= \inf_{w \in \mathbb{R}^S} \|y - Xw\|^2, \\ &= y^T[I - (X_S^T)^\dagger X_S^T]y. \end{aligned}$$

We use the notation $A^\dagger$ for the pseudoinverse of a matrix A. That is, if the singular value decomposition of A is $A = \sum_i \sigma_i u_i v_i^T$ with $\sigma_i > 0$ then $A^\dagger = \sum_i \sigma_i^{-1} v_i u_i^T$.

Let us first estimate $g(S) - g(T)$, for sets $S \subset T$. For brevity, define H_S as the projection matrix $X_S X_S^\dagger$ projecting onto the column space of X_S. Denote by $Z_{T \setminus S}$ the matrix whose columns are those of $X_{T\setminus S}$ projected away from the span of X_S, and normalized. Namely, writing x_i as the i'th column of X, $\zeta_i = \|(I - H_S)x_i\|$, $z_i = (I - H_S)x_i/\zeta_i$, and $Z_{T\setminus S}$'s columns are $\{z_i\}_{i\in T\setminus S}$. Notice that the columns of $Z_{T\setminus S}$ and X_S are orthogonal, hence according to the Pythagorean theorem it holds that

$$g(S) = \|y\|^2 - \|H_S y\|^2, \quad g(T) = \|y\|^2 - \|H_S y\|^2 - \|Z_{T\setminus S} Z_{T\setminus S}^{\dagger} y\|^2$$

meaning that $g(S) - g(T) = \|Z_{T\setminus S} Z_{T\setminus S}^{\dagger} y\|^2$. In particular, this implies that for any $j \notin S$ it holds that $g(S) - g(S \cup \{j\}) = (z_j^T y)^2$, since z_j is a unit vector. Let us now decompose $g(S) - g(T)$.

$$g(S) - g(T) = \|Z_{T\setminus S} Z_{T\setminus S}^{\dagger} y\|^2 = \|(Z_{T\setminus S}^T)^{\dagger} Z_{T\setminus S}^T y\|^2 \le \|(Z_{T\setminus S}^T)^{\dagger}\|^2 \cdot \|Z_{T\setminus S}^T y\|^2$$

The norm used in the last inequality is the matrix operator norm. We now bound both factors of the product on the RHS separately. For the first factor, we claim that $\|(Z_{T\setminus S}^T)^{\dagger}\| = \|Z_{T\setminus S}^{\dagger}\| \le \|X_T^{\dagger}\|$. To see this, consider a vector $w \in \mathbb{R}^{|T\setminus S|}$, for convenience denote its entries by $\{w(i)\}_{i\in T\setminus S}$, and write $z_i = (x_i - \sum_{j\in S} \alpha_{ij} x_j)/\zeta_i$. We have

$$Z_{T\setminus S} w = \sum_{i\in T\setminus S} z_i w(i) = \sum_{i\in T\setminus S} x_i w(i)/\zeta_i - \sum_{j\in S} x_j \sum_{i\in T\setminus S} w(i)\alpha_{ij}/\zeta_i = X_T w'$$

for the vector $w' \in \mathbb{R}^{|T|}$ defined as $w'(i) = w(i)/\zeta_i$ for $i \in T\setminus S$ and $w'(j) = -\sum_{i\in T\setminus S} w(i)\alpha_{ij}/\zeta_i$ for $j \in S$. Since $\zeta_i \le \|x_i\| \le 1$ we must have $\|w'\| \ge \|w\|$. Consider now the unit vector w for which $\|Z_{T\setminus S} w\| = \|Z_{T\setminus S}^{\dagger}\|^{-1}$, that is, the unit norm singular vector corresponding to the smallest non-zero singular value of $Z_{T\setminus S}$. For this w, and its corresponding vector w', we have

$$\|Z_{T\setminus S}^{\dagger}\|^{-1} = \|Z_{T\setminus S} w\| = \|X_T w'\| \ge \sigma_{\min}(X_T)\|w'\| \ge \sigma_{\min}(X_T)\|w\| = \sigma_{\min}(X_T).$$

It follows that

$$\|(Z_{T\setminus S}^T)^\dagger\|^2 = \|Z_{T\setminus S}^\dagger\|^2 \leq 1/\sigma_{\min}(X_T)^2$$

We continue to bound the right factor of product.

$$\|Z_{T\setminus S}^T y\|^2 = \sum_{i\in T\setminus S} (z_i^T y)^2 = \sum_{i\in T\setminus S} g(S) - g(S\cup\{i\}).$$

By combining the inequalities we obtained the required result:

$$g(S) - g(T) \leq \left(1/\sigma_{\min}(X_T)^2\right) \sum_{i\in T\setminus S} g(S) - g(S\cup\{i\}).$$

□

Proof of Lemma 3.3.3. We would like to apply Lemma A.6.4 on the design matrix X. The only catch is that the columns of X may not be bounded by 1 in norm. To remedy this, let j be the index of the column with the maximum norm and consider the matrix $\bar{X} = \frac{1}{\|X_j\|}X$ instead (here, X_j is the j-th column of X; note that $X_j = Xe_j$ for the j-th standard basis vector e_j). Now, for any subset S of coordinates,

$$\inf_{w\in\mathbb{R}^S} \|y - \bar{X}w\|^2 = \inf_{w\in\mathbb{R}^S} \|y - Xw\|^2.$$

Thus, we conclude that the set function of interest, $g(S) = \inf_{w\in\mathbb{R}^S} \|y - Xw\|^2$, is α-weakly supermodular for sparsity k for $\alpha = \max_{S:|S|\leq k} \|\bar{X}_S^\dagger\|_2^2$. For any subset of coordinates S of size at most k, let w be a unit norm right singular vector of $\bar{X}_S$ corresponding to the smallest singular value, so that $\|\bar{X}_S^\dagger\|_2 = \frac{1}{\|\bar{X}_S w\|}$. But $\frac{1}{\|\bar{X}_S w\|} = \frac{\|Xe_j\|}{\|Xw'\|}$, where w' is the vector w extended to all coordinates by padding with zeros.

Since the restricted condition number of X for sparsity k is bounded by κ we conclude that $\frac{\|Xe_j\|}{\|Xw'\|} \leq \kappa$. Since this bound holds for any subset S of size at most k, we conclude that $\alpha \leq \kappa^2$. □

Proof of Lemma 3.3.4. By the α-weak supermodularity of g, we have

$$g(\emptyset) - g(V) \leq \alpha \cdot \sum_{j \in V} [g(\emptyset) - g(\{j\})]$$
$$\leq \alpha |V| \cdot [(g(\emptyset) - g(V)) - (g(\{j^*\}) - g(V))].$$

Rearranging, we get the claimed bounds. □

The following lemma gives a useful property of weakly supermodular functions.

Lemma A.6.5. *Let $g(\cdot)$ be a (k, α)-weakly supermodular set function and U be a subset with $|U| < k$. Then $g'(S) := g(U \cup S)$ is $(k - |U|, \alpha)$-weakly supermodular.*

Proof. For any two subsets $S \subseteq T$ with $|T| \leq k - |U|$, we have

$$g'(S) - g'(T) = g(U \cup S) - g(U \cup T) \leq \alpha \sum_{j \in (T \cup U) \setminus (S \cup U)} [g(U \cup S) - g(U \cup S \cup \{j\})]$$
$$\leq \alpha \sum_{j \in T \setminus S} [g(U \cup S) - g(U \cup S \cup \{j\})] = \alpha \sum_{j \in T \setminus S} [g'(S) - g'(S \cup \{j\})].$$

□

Proof of Lemma 3.3.5. For $i \in \{0, 1, \ldots, k_1\}$, define the set function $g_b^{(i)}$ as $g_b^{(i)}(S) = g_b(S \cup V_b^{(i)})$.

First, we analyze the performance of the BEXP algorithms. Fix any $i \in [k_1]$ and consider $\mathsf{BEXP}^{(i)}$. Conceptually, for any $j \in [d]$, the loss of expert j at the end of mini-batch b is $g_b(V_b^{(i-1)} \cup j)$ (note that this loss is only evaluated for $j \in U_b^{(i)}$ in the algorithm). To bound the regret, we need to bound the magnitude of the losses. Note that for any subset S, we have $0 \leq g_b(S) \leq \frac{1}{B} \sum_{t \in \mathcal{T}_b} y_t^2 \leq 1$. Thus, the regret guarantee of BEXP (Theorem 3.3.2) implies that for any $i \in [k_1]$ and any $j \in [d]$, we have

$$\mathbf{E}\left[\sum_{b=1}^{T/B} g_b(V_b^{(i-1)} \cup \{j_b^{(i)}\})\right] \leq \sum_{b=1}^{T/B} g_b(V_b^{(i-1)} \cup \{j\}) + 2\sqrt{\tfrac{dk_1 \log(d) T}{k_0 B}}.$$

The expectation above is conditioned on the randomness in $V_b^{(i-1)}$, for $b \in [T/B]$. Rewriting the above inequality using the $g^{(i-1)}$ and $g^{(i)}$ functions, and using the fact that $V_b^{(i-1)} \cup \{j_b^{(i)}\} = V_b^{(i)}$, we get

$$\mathbf{E}\left[\sum_{b=1}^{T/B} g_b^{(i)}(\emptyset)\right] \leq \sum_{b=1}^{T/B} g_b^{(i-1)}(\{j\}) + 2\sqrt{\tfrac{dk_1 \log(d) T}{k_0 B}}. \tag{A.72}$$

Next, since we assumed that the sequence of feature vectors satisfies BBRCNP with parameters $(\epsilon, k_1 + k)$, Lemma 3.3.3 implies that the set function g_b defined in (3.44) is $(k_1 + k, \kappa^2)$-weakly supermodular for $\kappa = \frac{1+\epsilon}{1-\epsilon}$. By Lemma A.6.5, the set function $g_b^{(i)}$ is (k, κ^2)-weakly supermodular (since $|V_b^{(i)}| \leq k_1$).

It is easy to check that the sum of weakly supermodular functions is also weakly supermodular (with the same parameters), and hence $\sum_{b=1}^{T/B} g_b^{(i-1)}$ is also (k, κ^2)-weakly supermodular. Hence, by Lemma 3.3.4, if $j^* = \arg\min_j \sum_{b=1}^{T/B} g_b^{(i-1)}(\{j\})$, we have, for any subset V of size at most k,

$$\sum_{b=1}^{T/B} g_b^{(i-1)}(\{j^*\}) - g_b^{(i-1)}(V) \leq (1 - \tfrac{1}{\kappa^2 |V|})[\sum_{b=1}^{T/B} g_b^{(i-1)}(\emptyset) - g_b^{(i-1)}(V)].$$

Since $g_b(V) \geq g_b(V \cup V_b^{(i-1)}) = g_b^{(i-1)}(V)$, the above inequality implies that

$$\sum_{b=1}^{T/B} g_b^{(i-1)}(\{j^*\}) - g_b(V) \leq (1 - \tfrac{1}{\kappa^2 |V|})[\sum_{b=1}^{T/B} g_b^{(i-1)}(\emptyset) - g_b(V)].$$

Combining this bound with (A.72) for $j = j^*$, we get

$$\mathbf{E}\left[\sum_{b=1}^{T/B} g_b^{(i)}(\emptyset) - g_b(V)\right] \leq (1 - \tfrac{1}{\kappa^2|V|})[\sum_{b=1}^{T/B} g_b^{(i-1)}(\emptyset) - g_b(V)] + 2\sqrt{\tfrac{dk_1 \log(d)T}{k_0 B}}.$$

Applying this bound recursively for $i \in [k_1]$ and simplifying, we get

$$\mathbf{E}\left[\sum_{b=1}^{T/B} g_b^{(k_1)}(\emptyset) - g_b(V)\right] \leq (1 - \tfrac{1}{\kappa^2|V|})^{k_1}[\sum_{b=1}^{T/B} g_b^{(0)}(\emptyset) - g_b(V)] + 2\kappa^2|V|\sqrt{\tfrac{dk_1 \log(d)T}{k_0 B}}.$$

Using the definitions of $g_b^{(k_1)}$ and $g_b^{(0)}$, and the fact that $|V| \leq k$, we get the claimed bound. □

BIBLIOGRAPHY

E. Abbe and C. Sandon. Community detection in general stochastic block models: fundamental limits and efficient recovery algorithms. *arXiv preprint arXiv:1503.00609*, 2015a.

E. Abbe and C. Sandon. Detection in the stochastic block model with multiple clusters: proof of the achievability conjectures, acyclic bp, and the information-computation gap. *arXiv preprint arXiv:1512.09080*, 2015b.

E. Abbe, A. S. Bandeira, and G. Hall. Exact recovery in the stochastic block model. *arXiv preprint arXiv:1405.3267*, 2014.

L. A. Adamic and N. Glance. The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, pages 36–43. ACM, 2005.

A. Agarwal, S. Negahban, M. J. Wainwright, et al. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *The Annals of Statistics*, 40(2):1171–1197, 2012.

L. A. N. Amaral, A. Scala, M. Barthelemy, and H. E. Stanley. Classes of small-world networks. *Proceedings of the national academy of sciences*, 97(21):11149–11152, 2000.

D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: A geometric theory of phase transitions in convex optimization. *arXiv preprint arXiv:1303.6672*, 2013.

K. Amin, S. Kale, G. Tesauro, and D. S. Turaga. Budgeted prediction with expert advice. In *AAAI*, pages 2490–2496, 2015.

E. Arias-Castro, E. J. Candès, A. Durand, et al. Detection of an anomalous cluster in a network. *The Annals of Statistics*, 39(1):278–304, 2011.

J. Audibert. Progressive mixture rules are deviation suboptimal. *Advances in Neural Information Processing Systems*, 20(2), 2007.

S. Balakrishnan, M. Kolar, A. Rinaldo, A. Singh, and L. Wasserman. Statistical and computational tradeoffs in biclustering. In *NIPS 2011 Workshop on Computational Trade-offs in Statistical Learning*, 2011.

A. S. Bandeira and R. van Handel. Sharp nonasymptotic bounds on the norm of random matrices with independent entries. *arXiv preprint arXiv:1408.6185*, 2014.

A. Barrat and M. Weigt. On the properties of small-world network models. *The European Physical Journal B-Condensed Matter and Complex Systems*, 13(3):547–560, 2000.

P. Bartlett, O. Bousquet, and S. Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

A. Belloni, M. Rosenbaum, and A. B. Tsybakov. Linear and conic programming estimators in high dimensional errors-in-variables models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016. ISSN 1467-9868.

I. Benjamini and N. Berger. The diameter of long-range percolation clusters on finite cycles. *arXiv preprint math/0012070*, 2000.

G. Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962.

Q. Berthet and P. Rigollet. Computational lower bounds for sparse pca. *arXiv preprint arXiv:1304.0828*, 2013a.

Q. Berthet and P. Rigollet. Optimal detection of sparse principal components in high dimension. *The Annals of Statistics*, 41(4):1780–1815, 2013b.

P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

A. Birnbaum, I. M. Johnstone, B. Nadler, and D. Paul. Minimax bounds for sparse pca with noisy high-dimensional data. *Annals of statistics*, 41(3):1055, 2013.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. OUP Oxford, 2013.

O. Bousquet. *Concentration inequalities and empirical processes theory applied to the analysis of learning algorithms*. PhD thesis, Ecole Polytechnique, 2002.

O. Bousquet, V. Koltchinskii, and D. Panchenko. Some local measures of complexity of convex hulls and generalization bounds. In Springer, editor, *Computational Learning Theory*, pages 164–171, Sydney, Australia, 2002.

C. Boutsidis, E. Liberty, and M. Sviridenko. Greedy minimization of weakly supermodular set functions. *arXiv preprint arXiv:1502.06528*, 2015.

P. Bühlmann. Statistical significance in high-dimensional linear models. *Bernoulli*, 19(4): 1212–1242, 2013.

P. Bühlmann and S. van de Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer, 2011.

R. E. Burkard, E. Çela, P. M. Pardalos, and L. S. Pitsoulis. *The quadratic assignment problem*. Springer, 1998.

C. Butucea and Y. I. Ingster. Detection of a sparse submatrix of a high-dimensional noisy matrix. *Bernoulli*, 19(5B):2652–2688, 2013.

C. Butucea, Y. I. Ingster, and I. Suslina. Sharp variable selection of a sparse submatrix in a high-dimensional noisy matrix. *arXiv preprint arXiv:1303.5647*, 2013.

T. T. Cai and M. G. Low. An adaptation theory for nonparametric confidence intervals. *The Annals of Statistics*, 32:1805–1840, 2004a.

T. T. Cai and M. G. Low. Minimax estimation of linear functionals over nonconvex parameter spaces. *The Annals of Statistic*, 32(2):552–576, 2004b.

T. T. Cai and W. Zhou. Matrix completion via max-norm constrained optimization. *arXiv preprint arXiv:1303.0341*, 2013.

T. T. Cai, Z. Ma, and Y. Wu. Sparse pca: Optimal rates and adaptive estimation. *The Annals of Statistics*, 41(6):3074–3110, 2013.

T. T. Cai, T. Liang, and A. Rakhlin. Supplement to "geometric inference for general high-dimensional linear inverse problems". Technical report, 2014.

T. T. Cai, T. Liang, and A. Rakhlin. Computational and statistical boundaries for submatrix localization in a large noisy matrix. *The Annals of Statistics, to appear*, 2015a.

T. T. Cai, Z. Ma, and Y. Wu. Optimal estimation and rank detection for sparse spiked covariance matrices. *Probability Theory and Related Fields*, page to appear, 2015b.

T. T. Cai, T. Liang, and A. Rakhlin. On detection and structural reconstruction of small-world random networks. *IEEE Transactions on Network Science and Engineering, to appear*, 2016a.

T. T. Cai, T. Liang, and A. Rakhlin. Geometric inference for general high-dimensional linear inverse problems. *The Annals of Statistics*, 44(4):1536–1563, 2016b.

T. T. Cai, T. Liang, and A. Rakhlin. Inference via message passing on partially labeled stochastic block models. *arXiv preprint arXiv:1603.06923*, 2016c.

E. J. Candes and M. A. Davenport. How well can we estimate a sparse vector? *Applied and Computational Harmonic Analysis*, 34(2):317–323, 2013.

E. J. Candes and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6): 925–936, 2010.

E. J. Candes and Y. Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *Information Theory, IEEE Transactions on*, 57(4):2342–2359, 2011.

E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

E. J. Candes and T. Tao. Decoding by linear programming. *IEEE transactions on information theory*, 51(12):4203–4215, 2005.

E. J. Candès and T. Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35:2313–2351, 2007.

E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

E. Cela. *The quadratic assignment problem: theory and algorithms*, volume 1. Springer Science & Business Media, 2013.

N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games.* Cambridge University Press, 2006.

N. Cesa-Bianchi, S. Shalev-Shwartz, and O. Shamir. Efficient learning with partially observed attributes. *Journal of Machine Learning Research*, 12(Oct):2857–2878, 2011.

V. Chandrasekaran and M. I. Jordan. Computational and statistical tradeoffs via convex relaxation. *Proceedings of the National Academy of Sciences*, 110(13):E1181–E1190, 2013.

V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky. Sparse and low-rank matrix decompositions. In *Communication, Control, and Computing, 2009. Allerton 2009. 47th Annual Allerton Conference on*, pages 962–967. IEEE, 2009.

V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

S. Chatterjee. Matrix estimation by universal singular value thresholding. *arXiv preprint arXiv:1212.1247*, 2012.

Y. Chen and J. Xu. Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *arXiv preprint arXiv:1402.1267*, 2014.

A. Coja-Oghlan. Graph partitioning via adaptive spectral techniques. *Combinatorics, Probability and Computing*, 19(02):227–284, 2010.

D. Coppersmith, D. Gamarnik, and M. Sviridenko. The diameter of a long range percolation graph. In *Proceedings of the thirteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 329–337. Society for Industrial and Applied Mathematics, 2002.

M. Cucuringu, A. Singer, and D. Cowburn. Eigenvector synchronization, graph rigidity and the molecule problem. *Information and Inference*, 1(1):21–67, 2012.

D. Dai, P. Rigollet, and T. Zhang. Deviation optimal learning using greedy Q-aggregation. *Annals of Statistics*, 2012.

A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6):066106, 2011.

Y. Deshpande and A. Montanari. Finding hidden cliques of size\ sqrt {N/e} in nearly linear time. *arXiv preprint arXiv:1304.7047*, 2013.

Y. Deshpande, E. Abbe, and A. Montanari. Asymptotic mutual information for the two-groups stochastic block model. *arXiv preprint arXiv:1507.08685*, 2015.

D. Donoho and M. Gavish. Minimax risk of matrix denoising by singular value thresholding. *The Annals of Statistics*, 42(6):2413–2440, 2014.

D. Donoho and J. Jin. Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics*, pages 962–994, 2004.

D. L. Donoho. Statistical estimation and optimal recovery. *The Annals of Statistics*, pages 238–270, 1994.

D. L. Donoho and I. M. Johnstone. Minimax estimation via wavelet shrinkage. *The Annals of Statistics*, 26(3):879–921, 1998.

P. Drineas, R. Kannan, and M. W. Mahoney. Fast monte carlo algorithms for matrices ii: Computing a low-rank approximation to a matrix. *SIAM Journal on Computing*, 36(1): 158–183, 2006.

R. M. Dudley. The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967.

W. Evans, C. Kenyon, Y. Peres, and L. J. Schulman. Broadcasting on trees and the ising model. *Annals of Applied Probability*, pages 410–433, 2000.

V. Feldman, E. Grigorescu, L. Reyzin, S. Vempala, and Y. Xiao. Statistical algorithms and a lower bound for detecting planted cliques. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 655–664. ACM, 2013.

D. Foster, H. Karloff, and J. Thaler. Variable selection is hard. In *COLT*, volume 40, pages 696–709. JMLR.org, 2015.

D. Foster, S. Kale, and H. Karloff. Online sparse linear regression. In *COLT*, volume 49. JMLR.org, 2016.

D. Gamarnik and M. Sudan. Limits of local algorithms over sparse random graphs. In *Proceedings of the 5th conference on Innovations in theoretical computer science*, pages 369–376. ACM, 2014.

C. Gao, Z. Ma, A. Y. Zhang, and H. H. Zhou. Achieving optimal misclassification proportion in stochastic block model. *arXiv preprint arXiv:1505.03772*, 2015.

E. Giné and J. Zinn. Some limit theorems for empirical processes. *Annals of Probability*, 12(4):929–989, 1984.

Y. Gordon. *On Milman's inequality and random subspaces which escape through a mesh in* $\mathbb{R}^n$. Springer, 1988.

J. C. Gower and G. B. Dijksterhuis. *Procrustes problems*, volume 3. Oxford University Press Oxford, 2004.

R. M. Gray. *Toeplitz and circulant matrices: A review.* now publishers inc, 2006.

B. Hajek, Y. Wu, and J. Xu. Achieving exact cluster recovery threshold via semidefinite programming. *arXiv preprint arXiv:1412.6156*, 2014.

B. Hajek, Y. Wu, and J. Xu. Achieving exact cluster recovery threshold via semidefinite programming: Extensions. *arXiv preprint arXiv:1502.07738*, 2015.

K.-i. Hashimoto. Zeta functions of finite graphs and representations of p-adic groups. *Automorphic forms and geometry of arithmetic varieties.*, pages 211–280, 1989.

E. Hazan and T. Koren. Linear regression with limited observation. In *ICML*, 2012.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

D. Hsu, S. M. Kakade, and T. Zhang. A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab*, 17(6), 2012.

S. Jagabathula and D. Shah. Inferring rankings using constrained sensing. *Information Theory, IEEE Transactions on*, 57(11):7288–7306, 2011.

S. Janson and E. Mossel. Robust reconstruction on trees is determined by the second eigenvalue. *Annals of probability*, pages 2630–2649, 2004.

A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

A. Javanmard, A. Montanari, and F. Ricci-Tersenghi. Phase transitions in semidefinite relaxations. *arXiv preprint arXiv:1511.08769*, 2015.

J. Jin et al. Fast community detection by score. *The Annals of Statistics*, 43(1):57–89, 2015.

I. M. Johnstone. Gaussian estimation: Sequence and wavelet models. *Manuscript*, 2013.

I. M. Johnstone and B. W. Silverman. Speed of estimation in positron emission tomography and related inverse problems. *The Annals of Statistics*, pages 251–280, 1990.

S. Kale. Open problem: Efficient online sparse regression. In *COLT*, pages 1299–1301, 2014.

S. Kale, Z. Karnin, T. Liang, and D. Pal. Adaptive feature selection: Computationally efficient online sparse linear regression under rip. Technical report, 2017.

V. Kanade, E. Mossel, and T. Schramm. Global and local information in clustering labeled block models. *arXiv preprint arXiv:1404.6325*, 2014.

C. G. Kaufman, V. Ventura, and R. E. Kass. Spline-based non-parametric regression for periodic functions and its application to directional tuning of neurons. *Statistics in medicine*, 24(14):2255–2265, 2005.

H. Kesten and B. P. Stigum. Additional limit theorems for indecomposable multidimensional galton-watson processes. *The Annals of Mathematical Statistics*, pages 1463–1481, 1966a.

H. Kesten and B. P. Stigum. A limit theorem for multidimensional galton-watson processes. *The Annals of Mathematical Statistics*, 37(5):1211–1223, 1966b.

S. Khuri, T. Bäck, and J. Heitkötter. The zero/one multiple knapsack problem and genetic algorithms. In *Proceedings of the 1994 ACM symposium on Applied computing*, pages 188–193. ACM, 1994.

J. Kleinberg. The small-world phenomenon: An algorithmic perspective. In *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pages 163–170. ACM, 2000.

D. E. Knuth. *The Stanford GraphBase: a platform for combinatorial computing*, volume 37. Addison-Wesley Reading, 1993.

M. Kolar, S. Balakrishnan, A. Rinaldo, and A. Singh. Minimax localization of structural information in large noisy matrices. In *Advances in Neural Information Processing Systems*, pages 909–917, 2011.

V. Koltchinskii. Von neumann entropy penalization and low-rank matrix estimation. *The Annals of Statistics*, 39(6):2936–2973, 2011a.

V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008*, volume 2033. Springer, 2011b.

V. Koltchinskii and D. Panchenko. Rademacher processes and bounding the risk of function learning. *High Dimensional Probability*, II:443–459, 2000.

V. Koltchinskii, K. Lounici, and A. B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *The Annals of Statistics*, 39(5):2302–2329, 2011.

F. Krzakala, C. Moore, E. Mossel, J. Neeman, A. Sly, L. Zdeborová, and P. Zhang. Spectral redemption in clustering sparse networks. *Proceedings of the National Academy of Sciences*, 110(52):20935–20940, 2013.

D. Kukliansky and O. Shamir. Attribute efficient linear regression with distribution-dependent sampling. In *ICML*, pages 153–161, 2015.

R. Latała. Some estimates of norms of random matrices. *Proceedings of the American Mathematical Society*, 133(5):1273–1282, 2005.

V. Latora and M. Marchiori. Efficient behavior of small-world networks. *Physical review letters*, 87(19):198701, 2001.

G. Lecué and S. Mendelson. Aggregation via empirical risk minimization. *Probability Theory and Related Fields*, 145(3):591–613, 2009.

G. Lecué and S. Mendelson. Learning subgaussian classes: Upper and minimax bounds. *arXiv preprint arXiv:1305.4825*, 2013.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer, 1991.

T. Liang, A. Rakhlin, and K. Sridharan. Learning with square loss: Localization through offset rademacher complexity. In *Proceedings of The 28th Conference on Learning Theory*, pages 1260–1285, 2015.

N. Linial. Locality in distributed graph algorithms. *SIAM Journal on Computing*, 21(1): 193–201, 1992.

R. Lyons and Y. Peres. Probability on trees and networks, 2005.

V. Lyzinski, D. E. Fishkind, and C. E. Priebe. Seeded graph matching for correlated erdös-rényi graphs. *Journal of Machine Learning Research*, 15(1):3513–3540, 2014.

Z. Ma and Y. Wu. Computational barriers in minimax submatrix detection. *arXiv preprint arXiv:1309.5914*, 2013a.

Z. Ma and Y. Wu. Volume ratio, sparsity, and minimaxity under unitarily invariant norms. *arXiv preprint arXiv:1306.3609*, 2013b.

O. L. Mangasarian and B. Recht. Probability of unique integer solution to a system of linear equations. *European Journal of Operational Research*, 214(1):27–30, 2011.

L. Massoulié. Community detection thresholds and the weak ramanujan property. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pages 694–703. ACM, 2014.

F. McSherry. Spectral partitioning of random graphs. In *Foundations of Computer Science, 2001. Proceedings. 42nd IEEE Symposium on*, pages 529–537. IEEE, 2001.

S. Mendelson. Improving the sample complexity using global data. *Information Theory, IEEE Transactions on*, 48(7):1977–1991, 2002.

S. Mendelson. A few notes on statistical learning theory. In S. Mendelson and A. J. Smola, editors, *Advanced Lectures in Machine Learning, LNCS 2600, Machine Learning Summer School 2002, Canberra, Australia, February 11-22*, pages 1–40. Springer, 2003.

S. Mendelson. Learning without Concentration. In *Conference on Learning Theory*, 2014a.

S. Mendelson. Learning without Concentration for General Loss Functions. *ArXiv e-prints*, Oct. 2014b.

S. Mendelson. Learning without concentration. *arXiv preprint arXiv:1401.0304*, 2014.

S. Mendelson. On aggregation for heavy-tailed classes, 2015. Preprint.

S. Milgram. The small world problem. *Psychology today*, 2(1):60–67, 1967.

A. Montanari and E. Richard. Non-negative principal component analysis: Message passing algorithms and sharp asymptotics. *arXiv preprint arXiv:1406.4775*, 2014.

C. Moore and M. E. Newman. Epidemics and percolation in small-world networks. *Physical Review E*, 61(5):5678, 2000.

E. Mossel. Reconstruction on trees: beating the second eigenvalue. *Annals of Applied Probability*, pages 285–300, 2001.

E. Mossel, Y. Peres, et al. Information flow on trees. *The Annals of Applied Probability*, 13 (3):817–844, 2003.

E. Mossel, J. Neeman, and A. Sly. Stochastic block models and reconstruction. *arXiv preprint arXiv:1202.1499*, 2012.

E. Mossel, J. Neeman, and A. Sly. Belief propagation, robust reconstruction, and optimal recovery of block models. *arXiv preprint arXiv:1309.1380*, 2013a.

E. Mossel, J. Neeman, and A. Sly. A proof of the block model threshold conjecture. *arXiv preprint arXiv:1311.4115*, 2013b.

B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2):227–234, 1995.

S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 2012.

M. E. Newman and D. J. Watts. Scaling and percolation in the small-world network model. *Physical Review E*, 60(6):7332, 1999.

A. Y. Ng, M. I. Jordan, Y. Weiss, et al. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856, 2002.

H. N. Nguyen and K. Onak. Constant-time approximation algorithms via local improvements. In *Foundations of Computer Science, 2008. FOCS'08. IEEE 49th Annual IEEE Symposium on*, pages 327–336. IEEE, 2008.

F. O'Sullivan. A statistical perspective on ill-posed inverse problems. *Statistical Science*, 1 (4):502–518, 1986.

S. Oymak, C. Thrampoulidis, and B. Hassibi. Simple bounds for noisy linear inverse problems with exact side information. *arXiv preprint arXiv:1312.0641*, 2013.

M. Parnas and D. Ron. Approximating the minimum vertex cover in sublinear time and a connection to distributed algorithms. *Theoretical Computer Science*, 381(1):183–196, 2007.

D. M. Pavlovic, P. E. Vértes, E. T. Bullmore, W. R. Schafer, and T. E. Nichols. Stochastic blockmodeling of the modules and core of the caenorhabditis elegans connectome. *PloS one*, 9(7):e97584, 2014.

G. Pisier. *The volume of convex bodies and Banach space geometry*, volume 94. Cambridge University Press, 1999.

O. A. Prokopyev, H.-X. Huang, and P. M. Pardalos. On complexity of unconstrained hyperbolic 0–1 programming problems. *Operations Research Letters*, 33(3):312–318, 2005.

A. Rakhlin and K. Sridharan. Online non-parametric regression. In *Conference on Learning Theory*, 2014.

A. Rakhlin, K. Sridharan, and A. Tsybakov. Empirical entropy, minimax regret and minimax risk. *Bernoulli*, 2015. Forthcoming.

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.

A. Rohde, A. B. Tsybakov, et al. Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930, 2011.

M. Rosenbaum and A. B. Tsybakov. Sparse recovery under matrix uncertainty. *The Annals of Statistics*, 38(5):2620–2651, 2010.

A. A. Shabalin, V. J. Weigman, C. M. Perou, and A. B. Nobel. Finding large average submatrices in high dimensional data. *The Annals of Applied Statistics*, pages 985–1012, 2009.

C. Stam, B. Jones, G. Nolte, M. Breakspear, and P. Scheltens. Small-world networks and functional connectivity in alzheimer's disease. *Cerebral cortex*, 17(1):92–99, 2007.

M. J. Streeter and D. Golovin. An online algorithm for maximizing submodular functions. In *NIPS*, pages 1577–1584, 2008.

M. Talagrand. Majorizing measures: the generic chaining. *The Annals of Probability*, 24 (3):1049–1103, 1996a.

M. Talagrand. A new look at independence. *The Annals of probability*, pages 1–34, 1996b.

T. Tao. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.

J. M. Ten Berge. Orthogonal procrustes rotation for two or more matrices. *Psychometrika*, 42(2):267–276, 1977.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

A. Tikhonov and V. Y. Arsenin. *Methods for solving ill-posed problems.* John Wiley and Sons, Inc, 1977.

A. B. Tsybakov. *Introduction to nonparametric estimation*, volume 11. Springer Series in Statistics, 2009.

A. B. Tsybakov. Aggregation and minimax optimality in high-dimensional estimation. 2014.

D. W. Tufts and A. A. Shah. Estimation of a signal waveform from noisy data using low-rank approximation to a data matrix. *Signal Processing, IEEE Transactions on*, 41(4): 1716–1721, 1993.

L. G. Valiant and V. V. Vazirani. Np is as easy as detecting unique solutions. *Theoretical Computer Science*, 47:85–93, 1986.

S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42 (3):1166–1202, 2014.

R. Van Der Hofstad. Random graphs and complex networks. *Available on http://www. win. tue. nl/rhofstad/NotesRGCN. pdf*, page 11, 2009.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

R. Vershynin. Lectures in geometric functional analysis. *Unpublished manuscript. Available at http://www-personal. umich. edu/˜ romanv/papers/GFA-book/GFA-book. pdf*, 2011.

V. Vu. Random discrete matrices. In *Horizons of combinatorics*, pages 257–280. Springer, 2008.

V. Vu. A simple svd algorithm for finding hidden partitions. *arXiv preprint arXiv:1404.3918*, 2014.

V. Q. Vu and J. Lei. Minimax rates of estimation for sparse pca in high dimensions. *arXiv preprint arXiv:1202.0786*, 2012.

M. J. Wainwright. Structured regularizers for high-dimensional problems: Statistical and computational issues. *Annual Review of Statistics and Its Application*, 1:233–253, 2014.

T. Wang, Q. Berthet, and R. J. Samworth. Statistical and computational trade-offs in estimation of sparse principal components. *The Annals of Statistics, to appear*, 2014.

D. J. Watts. *Small worlds: the dynamics of networks between order and randomness.* Princeton university press, 1999.

D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393 (6684):440–442, 1998.

Y. Yang and A. Barron. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics*, pages 1564–1599, 1999.

F. Ye and C.-H. Zhang. Rate minimaxity of the lasso and dantzig selector for the lq loss in lr balls. *The Journal of Machine Learning Research*, 11:3519–3540, 2010.

R. Zass and A. Shashua. Nonnegative sparse pca. In *Advances in Neural Information Processing Systems*, pages 1561–1568, 2006.

A. Y. Zhang and H. H. Zhou. Minimax rates of community detection in stochastic block models. *arXiv preprint arXiv:1507.05313*, 2015.

C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

P. Zhang, C. Moore, and L. Zdeborová. Phase transitions in semisupervised clustering of sparse networks. *Physical Review E*, 90(5):052802, 2014a.

Y. Zhang, M. J. Wainwright, and M. I. Jordan. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. *arXiv preprint arXiv:1402.1918*, 2014b.

N. Zolghadr, G. Bartók, R. Greiner, A. György, and C. Szepesvári. Online learning with costly features and labels. In *NIPS*, pages 1241–1249, 2013.