On the Statistical Limits of Convex Relaxations: A Case Study

Zhaoran Wang ZHAORAN@PRINCETON.EDU
Princeton University, NJ 08544 USA

Quanquan Gu QG5W@VIRGINIA.EDU
University of Virginia, VA 22904 USA

Han Liu HANLIU@PRINCETON.EDU
Princeton University, NJ 08544 USA

Abstract

Many high dimensional sparse learning problems are formulated as nonconvex optimization. A popular approach to solve these nonconvex optimization problems is through convex relaxations such as linear and semidefinite programming. In this paper, we study the statistical limits of convex relaxations. Particularly, we consider two problems: Mean estimation for sparse principal submatrix and edge probability estimation for stochastic block model. We exploit the sum-of-squares relaxation hierarchy to sharply characterize the limits of a broad class of convex relaxations. Our result shows statistical optimality needs to be compromised for achieving computational tractability using convex relaxations. Compared with existing results on computational lower bounds for statistical problems, which consider general polynomial-time algorithms and rely on computational hardness hypotheses on problems like planted clique detection, our theory focuses on a broad class of convex relaxations and does not rely on unproven hypotheses.

1. Introduction

A broad variety of high dimensional statistical problems are formulated as nonconvex optimization. For example, sparse estimation can be formulated as optimization under ℓ_0-norm constraints, where the ℓ_0-norm is a pseudo-norm defined as the number of nonzero elements in a vector. To solve these nonconvex optimization problems, a popular approach is to resort to convex relaxations. Particularly, for sparse estimation, significant progress has been made by using ℓ_1-norm as a convex relaxation for the nonconvex ℓ_0-norm (see, e.g., (Bühlmann & van de Geer, 2011; Chandrasekaran et al., 2012) and the references therein).

In this paper, we study the statistical limits of convex relaxations. In particular, we focus on the sum-of-squares (SoS) hierarchy of convex relaxations (Lasserre, 2001; Parrilo, 2000; 2003), which is made up of a sequence of increasingly tighter convex relaxations based on semidefinite programming. We study the SoS hierarchy because it attains tighter approximations than other hierarchies such as the hierarchies proposed by (Sherali & Adams, 1990) and (Lovász & Schrijver, 1991), as well as their extensions (see (Laurent, 2003) for a comparison). Hence, the estimators in the SoS hierarchy achieve superior statistical performance than the estimators within other weaker hierarchies, which suggests the statistical limits of the SoS hierarchy are also the limits of weaker hierarchies.

To demonstrate the statistical limits of convex relaxations, we focus on the examples of sparse principal submatrix estimation and stochastic block model estimation. In detail, for sparse principal submatrix estimation, we assume there is a $s^* \times s^*$ submatrix with elevated mean β^* on the diagonal of a $d \times d$ noisy symmetric matrix. For stochastic block model estimation, we assume there exists a dense subgraph with s^* nodes planted in an Erdős-Rényi graph with d nodes. We denote by β^* the edge probability of the subgraph. For both examples, our goal is to estimate β^* under a challenging regime where $s^* = o\big[(d/\sqrt{\log d})^{2/3}\big]$ and $\log d = o(s^*)$. We prove the following information-theoretic lower bound

$$\inf_{\widehat{\beta}} \sup_{\mathbb{P} \in \mathcal{P}(s^*, d)} \mathbb{E}_{\mathbb{P}} \big| \widehat{\beta} - \beta^* \big| \geq C \sqrt{1/s^* \cdot \log(d/s^*)}, \quad (1.1)$$

where $\widehat{\beta}$ denotes any estimator, $\mathcal{P}(s^*, d)$ is the distribution family to be specified later and C is an absolute constant. We prove that a computational intractable estimator $\widehat{\beta}^{\mathrm{scan}}$ (to be specified later) attains the lower bound in (1.1). In order to achieve computational tractability, we consider convex relaxations of $\widehat{\beta}^{\mathrm{scan}}$ that fall within the SoS and weaker hierarchies, which are denoted by $\mathcal{H}$. Let C' be a positive absolute constant. We prove that under certain

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

conditions,

$$\inf_{\widehat{\beta}\in\mathcal{H}} \sup_{\mathbb{P}\in\mathcal{P}(s^*,d)} \mathbb{E}_{\mathbb{P}}\big|\widehat{\beta}-\beta^*\big| \geq C'. \tag{1.2}$$

Together with (1.1), (1.2) illustrates the statistical limitations of a broad class of convex relaxations. Ignoring the logarithmic factor, (1.1) and (1.2) suggest there exists a gap of $\sqrt{s^*}$ between the limits for any estimator and the limits for estimators within the hierarchies of convex relaxations. Hence, this result shows statistical optimality must be sacrificed for gaining computational tractability with convex relaxations. For sparse principal submatrix estimation, we prove that a linear-time estimator within $\mathcal{H}$ attains the lower bound in (1.2) up to a logarithmic factor, and is therefore nearly optimal within a general family of convex relaxations.

Our work is closely related to a recent line of research on computational barriers for statistical problems (Arias-Castro & Verzelen, 2014; Berthet & Rigollet, 2013a;b; Cai et al., 2015; Chen & Xu, 2014; Gao et al., 2014; Hajek et al., 2014; Krauthgamer et al., 2013; Ma & Wu, 2013; Wang et al., 2014; Zhang et al., 2014). Under various computational hardness hypotheses on problems like planted clique detection, these works quantify the gap between the information-theoretic limits and the statistical accuracy achievable by polynomial-time algorithms. For this purpose, their proofs are based on polynomial-time reductions from hard computational problems to statistical problems. In contrast with these works, we focus on the statistical limits of a broad class of convex relaxations rather than all polynomial-time algorithms. Correspondingly, our theory does not hinge on unproven computational hardness hypotheses, and our proof is based on constructions rather than reductions. Also, based on another perspective, (Chandrasekaran & Jordan, 2013) study the tradeoffs between computational complexity and statistical performance for normal mean estimation via hierarchies of convex relaxations. Their results are based on hierarchies of convex constraints, which are obtained by successively weakening the cone representation of the original constraint set. In comparison, our results are based on hierarchies of convex relaxations of the optimization problem itself rather than the constraints, which are obtained by successively tightening a basic semidefinite relaxation using variable augmentation techniques. In addition, our work is connected to previous works on the SoS and other convex relaxation hierarchies (see, e.g., (Barak & Moitra, 2015; Barak & Steurer, 2014; Chlamtac & Tulsiani, 2012; Ma & Wigderson, 2015; Meka et al., 2015) and the references therein). In particular, the key construction of feasible solutions in our proof is based on the dual certificates designed for the maximum clique problem, which is proposed by (Meka et al., 2015).

The rest of this paper is organized as follows. In §2 we introduce the statistical models. In §3 we present the SoS hierarchy of convex relaxations and apply it to estimate the models in §2. In §4 we establish the main results and lay out the proofs in §5. In §6 we conclude the paper.

2. Statistical Model

In the sequel, we briefly introduce the statistical models considered in this paper. Then we present several common estimators for them.

2.1. Sparse Principal Submatrix Estimation

Let $\mathbf{X}\in\mathbb{R}^{d\times d}$ be a random matrix from distribution $\mathbb{P}$ and $\mathbb{E}(\mathbf{X})=\mathbf{\Theta}$. We assume there exists an index set $\mathcal{S}^*\subseteq[d]$ with $|\mathcal{S}^*|=s^*$ that satisfies $\Theta_{i,j}=\beta^*$ for $i\neq j$ and $(i,j)\in\mathcal{S}^*\times\mathcal{S}^*$, while $\Theta_{i,j}=0$ for $i\neq j$ and $(i,j)\notin\mathcal{S}^*\times\mathcal{S}^*$. Here $\beta^*\geq 0$ is the signal strength. For all $i<j$, we assume that $X_{i,j}$'s are independently sub-Gaussian with $\mathbb{E}(X_{i,j})=\Theta_{i,j}$ and $\|X_{i,j}-\Theta_{i,j}\|_{\psi_2}\leq 1$. In addition, we assume that $X_{i,i}=0$ and $X_{i,j}=X_{j,i}$. We aim to estimate the signal strength β^*. For simplicity, hereafter we assume s^* is known. We denote by $\mathcal{P}(s^*,d)$ the family of distribution $\mathbb{P}$'s satisfying the above constraints.

This estimation problem is closely related to the problems considered by (Butucea & Ingster, 2013; Butucea et al., 2013; Cai et al., 2015; Kolar et al., 2011; Ma & Wu, 2013; Shabalin et al., 2009; Sun & Nobel, 2013). These works consider the detection problem and the recovery of $\mathcal{S}^*$, while we consider the estimation of signal strength. Besides, we focus on symmetric $\mathbf{X}$ for simplicity.

We consider the following estimator for β^* proposed by (Butucea & Ingster, 2013),

$$\widehat{\beta}^{\mathrm{scan}}=\frac{1}{s^*(s^*-1)}\sup_{\substack{\mathcal{S}\subseteq[d]\\|\mathcal{S}|=s^*}}\sum_{(i,j)\in\mathcal{S}\times\mathcal{S}}X_{i,j}, \tag{2.1}$$

where $|\mathcal{S}|$ is the cardinality of set $\mathcal{S}$. The intuition behind $\widehat{\beta}^{\mathrm{scan}}$ is to exhaustively search all principal submatrices of cardinality s^* and calculate the average of all entires within each principal submatrix. In §4 we will prove that $\widehat{\beta}^{\mathrm{scan}}$ attains the information-theoretic lower bound for estimating β^* within $\mathcal{P}(s^*,d)$ under a challenging regime where $s^*=o\big[(d/\sqrt{\log d})^{2/3}\big]$. Nevertheless, it is computationally intractable to obtain $\widehat{\beta}^{\mathrm{scan}}$. In §3 we will introduce convex relaxations of $\widehat{\beta}^{\mathrm{scan}}$. We also consider the following computational tractable estimators

$$\widehat{\beta}^{\mathrm{avg}}=\frac{1}{s^*(s^*-1)}\sum_{i,j=1}^{d}X_{i,j},\ \widehat{\beta}^{\max}=\max_{i,j\in[d]}X_{i,j} \tag{2.2}$$

for further discussion in §4.

2.2. Stochastic Block Model

We consider the estimation of edge probability in a dense subgraph with s^* nodes planted within an Erdős-Rényi

graph with d nodes. If a pair of nodes are within the subgraph, they are independently connected with edge probability $\beta^* \in [0,1]$. Otherwise, they are independently connected with edge probability $\widetilde{\beta}^* \in [0, \beta^*]$. We denote $\mathcal{P}(s^*, d)$ to be the distribution family of graphs which satisfy the above constraints and by $\mathbf{A} \in \mathbb{R}^{d\times d}$ the adjacency matrix. We assume $A_{i,i} = 0$ for all $i \in [d]$ and s^* is known. Similar to principal submatrix estimation, we focus on the challenging regime with $s^* = o\big[(d/\log d)^{2/3}\big]$. Additionally, we assume $\log(d/s^*)/\big(s^*\widetilde{\beta}^*\big) = o(1)$ so that s^* is not too small.

This estimation problem is connected to the ones studied by (Arias-Castro & Verzelen, 2014; Bhaskara et al., 2010; Chen & Xu, 2014; Coja-Oghlan, 2010; Decelle et al., 2011; Fortunato, 2010; Hajek et al., 2014; Kučera, 1995; Massoulié, 2014; Meka et al., 2015; Mossel et al., 2012; 2013; Verzelen & Arias-Castro, 2013). However, we mainly focus on estimating the edge probability of the dense subgraph rather than detection or recovery of subgraphs. Also, we assume that the dense subgraph and its size are fixed rather than random as in some of the existing works. To estimate β^*, we use $\widehat{\beta}^{\rm scan}$ and $\widehat{\beta}^{\max}$ defined in (2.1) and (2.2) with $X_{i,j}$ replaced by $A_{i,j}$. Though stochastic model estimation is closely related to sparse principal submatrix estimation, in §4 we will illustrate that the respective upper and lower bounds have subtle differences because of the different deviations of Bernoulli random variables and general sub-Gaussian random variables, which possibly have unbounded support.

3. Convex Relaxation Hierarchy

In this section, we first introduce some specific notations which will greatly simplify our presentation. Then we introduce the SoS hierarchy for $\widehat{\beta}^{\rm scan}$ defined in (2.1).

Notation: We define a collection $\mathcal{C}$ to be an unordered array of elements, where each element can appear more than once. For instance, $\{1\}$, $\{1,2\}$ and $\{1,1\}$ are all collections. Let the summation between two collections be the combination of all elements in them, e.g., for $\mathcal{C}_1 = \{1,2\}$, $\mathcal{C}_2 = \{1,3\}$ we have $\mathcal{C}_1 + \mathcal{C}_2 = \{1,1,2,3\}$. Note that a collection is different from a set, because a set has distinct elements. Let the merge operation $M(\cdot)$ on a collection be the operation that eliminates the duplicate elements and outputs a set, e.g., for $\mathcal{C} = \{1,1,2,2,3\}$ we have $M(\mathcal{C}) = \{1,2,3\}$, which is a set. We use $|\mathcal{C}|$ and $|\mathcal{S}|$ to denote the cardinality of a collection and a set. Also, we denote by $\mathcal{C}_1 = \mathcal{C}_2$ if they contain the same elements. For integer $\ell \geq 0$, we define $d^{(\ell)} = \sum_{i=0}^{\ell} d^i$ for notational simplicity.

Note that $\widehat{\beta}^{\rm scan}$ in (2.1) can be reformulated as

$$\widehat{\beta}^{\rm scan} = \max_{\mathbf{v}\in\mathcal{V}_{s^*}} \mathbf{v}^\top \mathbf{X}\mathbf{v}/[s^*(s^*-1)], \tag{3.1}$$

$$\text{where } \mathcal{V}_{s^*} = \bigg\{\mathbf{v} : \mathbf{v} \in \{0,1\}^d,\ \sum_{i=1}^{d} v_i = s^*\bigg\}.$$

Because (3.1) involves maximizing a convex function subject to nonconvex constraints, it is computational intractable to solve. Note that $\mathbf{v}^\top\mathbf{X}\mathbf{v} = \mathrm{tr}\big(\mathbf{X}\mathbf{v}\mathbf{v}^\top\big)$ in (3.1). We can reparameterize $\mathbf{v}\mathbf{v}^\top$ to be a $d\times d$ positive semidefinite matrix with rank one. For notational simplicity, we define

$$\mathbf{Y} = \begin{bmatrix} 0 & \mathbf{0}_{1\times d} \\ \mathbf{0}_{d\times 1} & \mathbf{X} \end{bmatrix}, \quad v_0 = 1, \tag{3.2}$$

$$\mathbf{\Pi} = \big(1, \mathbf{v}^\top\big)^\top \big(1, \mathbf{v}^\top\big) = \begin{bmatrix} 1 & \Pi_{0,1} & \ldots & \Pi_{0,d} \\ \Pi_{1,0} & \Pi_{1,1} & \ldots & \Pi_{1,d} \\ \vdots & \vdots & \ddots & \vdots \\ \Pi_{d,0} & \Pi_{d,1} & \ldots & \Pi_{d,d} \end{bmatrix}.$$

Here $\mathbf{Y}, \mathbf{\Pi} \in \mathbb{R}^{(d+1)\times(d+1)}$ and $\mathbf{0}_{d_1\times d_2}$ denotes a $d_1\times d_2$ matrix whose entries are all zero. Meanwhile, note that $\mathcal{V}_{s^*}$ defined in (3.1) can be reformulated as

$$\mathcal{V}_{s^*} = \bigg\{\mathbf{v} : \sum_{i=1}^{d} v_i = s^*,\ v_i^2 - v_i = 0,\ \forall i \in [d]\bigg\}. \tag{3.3}$$

According to the reparametrization in (3.2), it holds that $\Pi_{i,j} = v_i v_j$ for all $i,j \in \{0,\ldots,d\}$. Hence, from (3.1) we obtain the following semidefinite program

$$\max_{\mathbf{\Pi}} \mathrm{tr}(\mathbf{Y}\mathbf{\Pi})/[s^*(s^*-1)], \tag{3.4}$$

$$\text{subject to } \sum_{i=1}^{d} \Pi_{i,0} = s^*,\ \Pi_{0,0} = 1,\ \mathbf{\Pi} \succeq \mathbf{0},$$

$$\Pi_{i,j} = \Pi_{j,i},\ \Pi_{i,i} = \Pi_{i,0} \text{ for all } i,j \in \{0,1,\ldots,d\},$$

in which $\sum_{i=1}^{d}\Pi_{i,0} = s^*$ corresponds to $\sum_{i=1}^{d} v_i = s^*$, $\Pi_{i,j} = \Pi_{j,i}$ corresponds to $v_i v_j = v_j v_i$, while $\Pi_{i,i} = \Pi_{i,0}$ corresponds to $v_i^2 - v_i = 0$. Note that if $\mathrm{rank}(\mathbf{\Pi}) = 1$, then from our reparametrization in (3.2), the maximum of (3.4) equals the maximum of (3.1). However, we drop this rank constraint since it is nonconvex, and hence (3.4) is a convex relaxation of (3.1).

The SoS hierarchy is obtained by increasingly tightening the basic semidefinite program in (3.4) using variable augmentation techniques. In particular, the reparametrization in (3.4) only involves the second order interaction between v_i and v_j. For integer $\ell \geq 1$, we consider a $d^{(\ell)} \times d^{(\ell)}$ matrix $\mathbf{\Pi}^{(\ell)}$, where $d^{(\ell)} = \sum_{i=0}^{\ell} d^i$ in our notations. For notational simplicity, we index the entries of $\mathbf{\Pi}^{(\ell)}$ using collections $\mathcal{C}_1$ and $\mathcal{C}_2$ with $|\mathcal{C}_1|, |\mathcal{C}_2| \leq \ell$, whose elements are indices $1,\ldots,d$. Our reparametrization takes the form

$$\Pi^{(\ell)}_{\mathcal{C}_1,\mathcal{C}_2} = \prod_{i\in\mathcal{C}_1} v_i \prod_{j\in\mathcal{C}_2} v_j = \prod_{i\in\mathcal{C}_1+\mathcal{C}_2} v_i. \tag{3.5}$$

In particular, for $\mathcal{C}=\emptyset$ we define $\prod_{i\in\mathcal{C}} v_i = 1$. The ℓ-th level SoS relaxation of (3.1) takes the form

$$\max_{\mathbf{\Pi}} \operatorname{tr}\big(\mathbf{Y}^{(\ell)}\mathbf{\Pi}^{(\ell)}\big)/[s^*(s^*-1)], \quad \text{subject to} \tag{3.6}$$

$$\sum_{i=1}^{d} \Pi^{(\ell)}_{\{i\}+\mathcal{C}_1,\mathcal{C}_2} = s^*\Pi^{(\ell)}_{\mathcal{C}_1,\mathcal{C}_2}, \text{ for all } |\mathcal{C}_1|\le \ell-1,\ |\mathcal{C}_2|\le\ell,$$

$$\Pi^{(\ell)}_{\{i,i\}+\mathcal{C}_1,\mathcal{C}_2} = \Pi^{(\ell)}_{\{i\}+\mathcal{C}_1,\mathcal{C}_2},$$
$$\text{for all } i\in[d],\ |\mathcal{C}_1|\le\ell-2,\ |\mathcal{C}_2|\le\ell,$$

$$\Pi^{(\ell)}_{\mathcal{C}_1,\mathcal{C}_2} = \Pi^{(\ell)}_{\mathcal{C}_1',\mathcal{C}_2'},$$
$$\text{for all } \mathcal{C}_1+\mathcal{C}_2=\mathcal{C}_1'+\mathcal{C}_2',\ |\mathcal{C}_1|,|\mathcal{C}_2|,|\mathcal{C}_1'|,|\mathcal{C}_2'|\le\ell,$$

$$\Pi^{(\ell)}_{\emptyset,\emptyset}=1,\ \mathbf{\Pi}^{(\ell)}\succeq\mathbf{0},$$

where $\mathbf{Y}^{(\ell)}\in\mathbb{R}^{d^{(\ell)}\times d^{(\ell)}}$ is defined as

$$\mathbf{Y}^{(\ell)} = \begin{bmatrix} 0 & \mathbf{0}_{1\times d} & \dots & \mathbf{0}_{1\times d^\ell} \\ \mathbf{0}_{d\times 1} & \mathbf{X} & \dots & \mathbf{0}_{d\times d^\ell} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}_{d^\ell\times 1} & \mathbf{0}_{d^\ell\times d} & \dots & \mathbf{0}_{d^\ell\times d^\ell} \end{bmatrix}.$$

Note that the first constraint in (3.6) is corresponding to the reparametrization in (3.5) and

$$\prod_{j\in\mathcal{C}} v_j\Big(\sum_{i=1}^{d} v_i\Big) = s^*\prod_{j\in\mathcal{C}} v_j, \quad \text{for all } |\mathcal{C}|\le 2\ell-1,$$

which is equivalent to $\sum_{i=1}^d v_i = s^*$ in (3.3). The second constraint corresponds to (3.5) and

$$\prod_{j\in\mathcal{C}} v_j\cdot v_i^2 = \prod_{j\in\mathcal{C}} v_j\cdot v_i, \quad \text{for all } |\mathcal{C}|\le 2\ell-2,$$

which is equivalent to $v_i^2 - v_i = 0$ in (3.3). Also, the third constraint corresponds to (3.5) and

$$\prod_{j\in\mathcal{C}_1+\mathcal{C}_2} v_j = \prod_{j\in\mathcal{C}_1'+\mathcal{C}_2'} v_j,$$
$$\text{for all } \mathcal{C}_1+\mathcal{C}_2=\mathcal{C}_1'+\mathcal{C}_2',\ |\mathcal{C}_1|,|\mathcal{C}_2|,|\mathcal{C}_1'|,|\mathcal{C}_2'|\le\ell.$$

The last constraint that $\Pi^{(\ell)}_{\emptyset,\emptyset}=1$ follows from (3.5) and our definition that $\prod_{i\in\mathcal{C}} v_i = 1$ for $\mathcal{C}=\emptyset$. For $\ell=1$, (3.6) reduces to the basic semidefinite relaxation in (3.4). We denote by $\widehat{\beta}^{(\ell)}_{\mathrm{SoS}}$ the maximum of (3.6). We have

$$\widehat{\beta}^{\mathrm{scan}} \le \cdots \le \widehat{\beta}^{(\ell)}_{\mathrm{SoS}} \le \cdots \le \widehat{\beta}^{(2)}_{\mathrm{SoS}} \le \widehat{\beta}^{(1)}_{\mathrm{SoS}},$$

since we have more constraints in (3.6) for a larger ℓ. Thus, for a larger ℓ (3.6) gives a tighter convex relaxation of (3.1). Meanwhile, note that the semidefinite program in (3.6) can be solved in $O\big(d^{O(\ell)}\big)$ operations. Hereafter we focus on the settings where ℓ does not increase with d.

(Laurent, 2003) proves that other existing convex relaxation hierarchies, such as Sherali-Adams and Lovász-Schrijver hierarchies as well as their extensions, are weaker than the SoS hierarchy in the sense that $\widehat{\beta}^{\mathrm{scan}} \le \widehat{\beta}^{(\ell)}_{\mathrm{SoS}} \le \widehat{\beta}^{(\ell)}_{\mathrm{other}}$, where $\widehat{\beta}^{(\ell)}_{\mathrm{other}}$ denotes the ℓ-th level of other weaker hierarchies.

Note that relaxing constraints and objectives in the convex relaxations also leads to looser approximations of $\widehat{\beta}^{\mathrm{scan}}$. Hence, we denote by $\mathcal{H}^{(\ell)}$ the class of estimator $\widehat{\beta}$'s that fall in the ℓ-th level of the SoS and weaker hierarchies, as well as their weakened versions obtained by relaxing constraints and objectives. By this definition, we have $\mathcal{H}^{(1)}\subseteq\mathcal{H}^{(2)}\cdots$. For example, for $\ell>1$ we can drop constraints in (3.6) to obtain (3.4), which corresponds to $\ell=1$. In particular, from (3.1) we have

$$\widehat{\beta}^{\mathrm{scan}} = \max_{\mathbf{v}\in\mathcal{V}_{s^*}} \frac{\mathbf{v}^\top\mathbf{X}\mathbf{v}}{s^*(s^*-1)} \le \max_{\mathbf{u},\mathbf{v}\in\mathcal{V}_{s^*}} \frac{\mathbf{u}^\top\mathbf{X}\mathbf{v}}{s^*(s^*-1)}$$
$$\le \max_{\mathbf{u},\mathbf{v}\in\bar{\mathcal{V}}_{s^*}} \frac{\mathbf{u}^\top\mathbf{X}\mathbf{v}}{s^*(s^*-1)} \le \max_{\mathbf{\Omega}\in\mathcal{W}_{s^*}} \frac{\operatorname{tr}(\mathbf{X}\mathbf{\Omega})}{s^*(s^*-1)}, \tag{3.7}$$

$$\text{where } \bar{\mathcal{V}}_{s^*} = \Big\{\mathbf{v} : \sum_{i=1}^{d} v_i = s^*,\ v_i\ge 0,\ \forall i\in[d]\Big\},$$

$$\mathcal{W}_{s^*} = \Big\{\mathbf{\Omega} : \sum_{i=1}^{d}\Omega_{i,j} = (s^*)^2,\ \Omega_{i,j}\ge 0,\ \forall i,j\in[d]\Big\}.$$

Here $\bar{\mathcal{V}}_{s^*}$ is a linear programming relaxation of $\mathcal{V}_{s^*}$. Note that the right-hand side of (3.7) is equal to $s^*/(s^*-1)\cdot\widehat{\beta}^{\max}$, where $\widehat{\beta}^{\max}$ is defined in (2.2). Therefore, $s^*/(s^*-1)\cdot\widehat{\beta}^{\max}$ can be viewed as a linear programming relaxation of $\widehat{\beta}^{\mathrm{scan}}$, which falls in $\mathcal{H}^{(1)}$ (see, e.g., §2 of (Chlamtac & Tulsiani, 2012) for details). It is also worth noting that the SoS hierarchy has several equivalent formulations. See, e.g., Theorem 2.7 of (Barak & Steurer, 2014) for a proof of such equivalence.

4. Main Result

As defined in §3, $\mathcal{H}^{(\ell)}$ denotes the ℓ-th level of the convex relaxation hierarchy for $\widehat{\beta}^{\mathrm{scan}}$ defined in (2.1). For stochastic block model, we replace $\mathbf{X}$ in (2.1) with the adjacency matrix $\mathbf{A}$ respectively.

4.1. Sparse Principal Submatrix Estimation

In the following, we present the main theoretical results for estimating the signal strength of sparse principal submatrix. In the sequel we establish the information-theoretic lower bound for estimating β^* within the distribution family $\mathcal{P}(s^*,d)$ defined in §2.1.

Theorem 1. For all estimators $\widehat{\beta}$ constructed using $\mathbf{X}\sim\mathbb{P}\in\mathcal{P}(s^*,d)$ and $s^*=o\big[(d/\sqrt{\log d})^{2/3}\big]$, there exists an absolute constant $C>0$ such that

$$\inf_{\widehat{\beta}}\sup_{\mathbb{P}\in\mathcal{P}(s^*,d)} \mathbb{E}_{\mathbb{P}}\big|\widehat{\beta}-\beta^*\big| \ge C\sqrt{1/s^*\cdot\log(d/s^*)}.$$

Proof. See §A for a detailed proof. □

In Theorem 1 we consider a challenging regime. More specifically, a straightforward calculation shows that

$\widehat{\beta}^{\mathrm{avg}}$ defined in (2.2) achieves the $d/(s^*)^2$ rate of convergence. For $s^* = o\big[(d/\sqrt{\log d})^{2/3}\big]$, we have $\sqrt{1/s^* \cdot \log(d/s^*)} = o\big[d/(s^*)^2\big]$. Thus, there exists a gap between the rate attained by $\widehat{\beta}^{\mathrm{avg}}$ and the information-theoretic lower bound. We will show that there is also such a gap for $\widehat{\beta}^{\max}$. The next proposition shows $\widehat{\beta}^{\mathrm{scan}}$ in (2.1) attains the information-theoretic lower bound in Theorem 1.

Proposition 2. For $\widehat{\beta}^{\mathrm{scan}}$ defined in (2.1) with $X_{i,j}$ being the (i,j)-th entry of $\mathbf{X} \sim \mathbb{P} \in \mathcal{P}(s^*, d)$, we have that

$$\big|\widehat{\beta}^{\mathrm{scan}} - \beta^*\big| \le C\sqrt{1/s^* \cdot \log(d/s^*)}$$

holds with probability at least $1 - 1/d$ for some absolute constant $C > 0$.

Proof. See §A for a detailed proof. □

Theorem 1 and Proposition 2 show $\widehat{\beta}^{\mathrm{scan}}$ is statistically optimal under the regime where $s^* = o\big[(d/\sqrt{\log d})^{2/3}\big]$. However, it is computationally intractable to obtain $\widehat{\beta}^{\mathrm{scan}}$. Thus, we consider the family of convex relaxations of $\widehat{\beta}^{\mathrm{scan}}$ within the ℓ-th level SoS and weaker hierarchies as well as their further relaxations, which is denoted by $\mathcal{H}^{(\ell)}$. In the sequel, we establish a minimax lower bound for the statistical performance of all estimators within $\mathcal{H}^{(\ell)}$. Recall that $\mathcal{P}(s^*, d)$ is the distribution family defined in §2.1.

Theorem 3. We assume that $s^* = o\big\{\big[d/(\log d)^2\big]^{1/2\ell}\big\}$. There is an absolute constant $C > 0$ such that

$$\inf_{\widehat{\beta} \in \mathcal{H}^{(\ell)}} \sup_{\mathbb{P} \in \mathcal{P}(s^*, d)} \mathbb{E}_{\mathbb{P}}\big|\widehat{\beta} - \beta^*\big| \ge C.$$

Proof. See §A for a detailed proof. □

Note that the regime considered in Theorem 3 is within the challenging regime considered in Theorem 1. Under this regime, Theorem 3 proves that any estimator within the convex relaxation hierarchy fails to attain a statistical rate that decreases when s^* is increasing. A comparison between Theorems 1 and 3 illustrates that there exists a gap of $\sqrt{s^*}$ (ignoring the $\log d$ factor) between the information-theoretic lower bound and the statistical rate achievable by a broad class of convex relaxations. In other words, to achieve computational tractability via convex relaxations, we have to compromise statistical optimality.

It is worth noting that this gap between computational tractability and statistical optimality is effective under the regime $s^* = o\big\{\big[d/(\log d)^2\big]^{1/2\ell}\big\}$, which shrinks as ℓ increases. However, ℓ cannot increase with d and s^*, because otherwise the computational complexity required to solve the convex relaxations increases exponentially, according to our discussion in §3. For ℓ being any constant, the regime in Theorem 3 is a nontrivial subset of the regime in Theorem 1. As will be shown in our proof, $s^* = o\big\{\big[d/(\log d)^2\big]^{1/2\ell}\big\}$ is a sufficient condition to establish the feasibility of the constructed solution. In fact, for $\ell = 2$, we can further relax this condition to $s^* = o\big(d^{1/3}/\log d\big)$ with the results of (Deshpande & Montanari, 2015). Under the regime in Theorem 3, the next proposition shows that $\widehat{\beta}^{\max}$ defined in (2.2) is nearly optimal under computational tractability constraints.

Proposition 4. For $\widehat{\beta}^{\max}$ in (2.2), where $X_{i,j}$ is the (i,j)-th entry of $\mathbf{X} \sim \mathbb{P} \in \mathcal{P}(s^*, d)$, we have

$$\big|\widehat{\beta}^{\max} - \beta^*\big| \le C\sqrt{\log d}$$

holds with probability at least $1 - 1/d$ for some absolute constant $C > 0$.

Proof. See §A for a detailed proof. □

According to (3.7) and the discussion in §3, we have $\widehat{\beta}^{\max} \in \mathcal{H}^{(1)} \subseteq \mathcal{H}^{(2)} \cdots$. Thus $\widehat{\beta}^{\max}$ attains the minimax lower bound with computational constraints in Theorem 3 for every ℓ up to a $\log d$ factor, which also suggests that the lower bound in Theorem 3 is tight. Meanwhile, note that the calculation of $\widehat{\beta}^{\max}$ in (2.2) requires $O(d^2)$ operations, which is linear in the size of input. In contrast, tighter approximations in the ℓ-th level SoS hierarchy require $O\big(d^{O(\ell)}\big)$ operations. In practice, such a computational complexity is in general higher than the complexity for calculating $\widehat{\beta}^{\max}$. Theorem 3 indicates that this extra computational cost can only result in limited possible improvements on the statistical rate of convergence, i.e., a $\log d$ factor.

It is worth noting the gap between the lower bounds in Theorems 1 and 3 vanishes when s^* is a constant that does not increase with d. In this case, $\widehat{\beta}^{\max}$ achieves the information-theoretic lower bound in Theorem 1. On the other hand, $\widehat{\beta}^{\mathrm{scan}}$ is computational tractable to obtain in this case.

4.2. Stochastic Block Model

In this section, we present the main theory for edge probability estimation in stochastic block model. Recall that $\mathcal{P}(s^*, d)$ is the distribution family defined in §2.2. The following lemma establishes the information-theoretic lower bound for estimating β^*. Recall $\widetilde{\beta}^*$ denotes the edge probability of the large Erdős-Rényi graph with d nodes.

Theorem 5. We assume that $s^* = o\big[(d/\sqrt{\log d})^{2/3}\big]$ and $\log(d/s^*)/\big(s^*\widetilde{\beta}^*\big) = o(1)$. There is an absolute constant $C > 0$ such that

$$\inf_{\widehat{\beta}} \sup_{\mathbb{P} \in \mathcal{P}(s^*, d)} \mathbb{E}_{\mathbb{P}}\big|\widehat{\beta} - \beta^*\big| \ge C\sqrt{1/s^* \cdot \log(d/s^*)}.$$

Proof. See §B for a detailed proof. □

Theorem 5 is similar to Theorem 1 but needs an extra condition that $\log(d/s^*)/\big(s^*\widetilde{\beta}^*\big) = o(1)$, which ensures s^* is not too small. Recall each entry of the adjacency matrix $\mathbf{A}$ is Bernoulli. (Arias-Castro & Verzelen, 2014) shows that a larger s^* guarantees the moderate deviation of the Bernoulli

distribution is in effect in the lower bound. Next, we prove $\widehat{\beta}^{\mathrm{scan}}$ achieves the information-theoretic lower bound in Theorem 5 and hence is optimal.

Proposition 6. For $\widehat{\beta}^{\mathrm{scan}}$ defined in (2.1), we have that with probability at least $1 - 1/d$,

$$\big|\widehat{\beta}^{\mathrm{scan}} - \beta^*\big| \le C\sqrt{1/s^* \cdot \log(d/s^*)}.$$

Proof. See §B for a detailed proof. □

The next theorem establishes the minimax lower bound on the statistical performance of convex relaxations within $\mathcal{H}^{(\ell)}$ defined in §3.

Theorem 7. For s^* and d sufficiently large and $s^* = o\big\{\big[d/(\log d)^2\big]^{1/2\ell}\big\}$, we have

$$\inf_{\widehat{\beta} \in \mathcal{H}^{(\ell)}} \sup_{\mathbb{P} \in \mathcal{P}(s^*, d)} \mathbb{E}_{\mathbb{P}}\big|\widehat{\beta} - \beta^*\big| \ge 1/4.$$

Proof. See §B for a detailed proof. □

Similar to Theorem 3, Theorem 7 shows the gap between statistical optimality and computational tractability. Note that $\beta^* \in [0, 1]$. Meanwhile, it is easy to show $\mathbb{P}\big(\widehat{\beta}^{\max} = 1\big) \ge 1 - \big(1 - \widetilde{\beta}^*\big)^{\frac{d^2-d}{2}}$. Therefore, $\widehat{\beta}^{\max}$ exactly attains such a minimax lower bound under computational constraints up to constants. From another point of view, for $s^* = o\big\{\big[d/(\log d)^2\big]^{1/2\ell}\big\}$, every estimators within $\mathcal{H}^{(\ell)}$ is at most as accurate as the trivial estimator $\widehat{\beta} = 1$.

Theorems 3 and 7 are similar. Note that for sparse principal submatrix estimation we consider sub-Gaussian entries, while in the adjacency matrix for stochastic block model each entry is Bernoulli. A direct way to establish Theorem 3 is to adapt the construction of $\mathbb{P}$ in the proof of Theorem 7, since Bernoulli is sub-Gaussian. However, as illustrated in §A the information-theoretic lower bound in Theorem 1 is established using the construction of $\mathbb{P}$ with unbounded support. Correspondingly, we use a construction of $\mathbb{P}$ with unbounded support to establish the lower bound with computational constraints in Theorem 3. By matching the constructions of $\mathbb{P} \in \mathcal{P}(s^*, d)$ in the proofs of Theorems 1 and 3, we can sharply characterize the existence of the $\sqrt{s^*}$ gap particularly for sub-Gaussian distributions with unbounded support.

5. Proof of Main Results

In the sequel, we present the proofs of the main results in §4. Due to space limit, we lay out the proof of Theorem 3 for sparse principal submatrix estimation, which is one of the major results, and defer the rest proofs to the appendix.

5.1. Proof of Theorem 3

Proof of Theorem 3. In this proof, we focus on specific distributions in $\mathcal{P}(s^*, d)$ with $\beta^* = 0$. We consider $X_{i,j}$'s ($i < j$) being sub-Gaussian random variables which satisfy the constraints in §2.1. In addition, we assume that $|X_{i,j}| \ge \nu$ almost surely and $\mathbb{P}(X_{i,j} > 0) = \mathbb{P}(X_{i,j} < 0) = 1/2$ for all $i < j$ and constant $\nu > 0$. Under such a distribution, we construct a matrix $\mathbf{\Pi}^{(\ell)} \in \mathbb{R}^{d^{(\ell)} \times d^{(\ell)}}$, which is a feasible solution to the ℓ-th level SoS program in (3.6) with high probability. We further prove that the objective value corresponding to $\mathbf{\Pi}^{(\ell)}$ is larger than ν, which indicates that the maximum of the corresponding SoS program is at least ν with high probability. In the rest of this proof, we denote $\mathbf{X} + \nu \cdot \mathbf{I}_d$ to be $\overline{\mathbf{X}}$.

Hereafter, we denote by $\overline{\mathbf{X}}_{\mathcal{S},\mathcal{S}'}$ the submatrix of $\overline{\mathbf{X}}$ whose row indices are in $\mathcal{S}$ and column indices are in $\mathcal{S}'$. For notational simplicity, we define the expansivity $\eta(\mathcal{S}, \overline{\mathbf{X}})$ of some set $\mathcal{S} \subseteq [d]$ to be the number of sets $\mathcal{S}' \subseteq [d]$ that satisfy $|\mathcal{S}'| = 2\ell$, $\mathcal{S} \subseteq \mathcal{S}'$ and $\mathrm{sign}\big(\overline{\mathbf{X}}_{\mathcal{S}',\mathcal{S}'}\big) = \mathbf{1}_{2\ell,2\ell}$. Here $\mathrm{sign}(\mathbf{X})$ is a matrix that satisfies $[\mathrm{sign}(\mathbf{X})]_{i,j} = 1$ if $X_{i,j} > 0$ and $[\mathrm{sign}(\mathbf{X})]_{i,j} = 0$ if $X_{i,j} \le 0$. Note that $\eta(\mathcal{S}, \overline{\mathbf{X}})$ is nonzero only if $\overline{X}_{i,j} > 0$ for all $i \in \mathcal{S}, j \in \mathcal{S}$. Hence, $\eta(\mathcal{S}, \overline{\mathbf{X}})$ gives the number of $\overline{\mathbf{X}}$'s submatrices that are extended from $\overline{\mathbf{X}}_{\mathcal{S},\mathcal{S}}$ and have size $2\ell \times 2\ell$ with all entries being positive. It is worth noting that by definition $\eta(\mathcal{S}, \overline{\mathbf{X}})$ is a random quantity, which depends on the random matrix $\overline{\mathbf{X}}$. Recall that each entry $\Pi^{(\ell)}_{\mathcal{C}_1,\mathcal{C}_2}$ of $\mathbf{\Pi}^{(\ell)}$ are indexed by collections $\mathcal{C}_1$ and $\mathcal{C}_2$, and $M(\mathcal{C}_1)$ and $M(\mathcal{C}_1)$ are the respective sets, which have distinct elements. Based on the construction of dual certificates of (Meka et al., 2015), we construct $\mathbf{\Pi}^{(\ell)}$ as

$$\Pi^{(\ell)}_{\mathcal{C}_1,\mathcal{C}_2} = \frac{\eta\big[M(\mathcal{C}_1) \cup M(\mathcal{C}_2), \overline{\mathbf{X}}\big]}{\eta\big(\emptyset, \overline{\mathbf{X}}\big)} \cdot \frac{s^*!/[s^* - |M(\mathcal{C}_1) \cup M(\mathcal{C}_2)|]!}{(2\ell)!/[2\ell - |M(\mathcal{C}_1) \cup M(\mathcal{C}_2)|]!}. \tag{5.1}$$

Now we verify $\mathbf{\Pi}^{(\ell)}$ defined in (5.1) satisfies all the constraints of the ℓ-th level SoS program in (3.6). First, we have $\Pi^{(\ell)}_{\emptyset,\emptyset} = 1$ from (5.1). Also, $\mathbf{\Pi}^{(\ell)}$ satisfies $\Pi^{(\ell)}_{\mathcal{C}_1,\mathcal{C}_2} = \Pi^{(\ell)}_{\mathcal{C}'_1,\mathcal{C}'_2}$ for $\mathcal{C}_1 + \mathcal{C}_2 = \mathcal{C}'_1 + \mathcal{C}'_2$, since

$$\begin{aligned} M(\mathcal{C}_1) \cup M(\mathcal{C}_2) &= M(\mathcal{C}_1 + \mathcal{C}_2) \\ &= M(\mathcal{C}'_1 + \mathcal{C}'_2) = M(\mathcal{C}'_1) \cup M(\mathcal{C}'_2) \end{aligned}$$

by the definition of the merge operation $M(\cdot)$. Meanwhile, it holds that $\Pi^{(\ell)}_{\mathcal{C}_1+\{i,i\},\mathcal{C}_2} = \Pi^{(\ell)}_{\mathcal{C}_1+\{i\},\mathcal{C}_2}$ for all $\mathcal{C}_1$ and $\mathcal{C}_2$ with $|\mathcal{C}_1| \le \ell - 2$ and $|\mathcal{C}_2| \le \ell$, since in (5.1) we have

$$M(\mathcal{C}_1 + \{i, i\}) \cup M(\mathcal{C}_2) = M(\mathcal{C}_1 + \{i\}) \cup M(\mathcal{C}_2).$$

Now we prove that $\sum_{i=1}^{d} \Pi^{(\ell)}_{\mathcal{C}_1+\{i\},\mathcal{C}_2} = s^* \Pi^{(\ell)}_{\mathcal{C}_1,\mathcal{C}_2}$ holds for all $|\mathcal{C}_1| \le \ell - 1$ and $|\mathcal{C}_2| \le \ell$. Let $\mathcal{C} = \mathcal{C}_1 + \mathcal{C}_2$, which

satisfies $|M(\mathcal{C})| \le |\mathcal{C}| \le 2\ell - 1$. By (5.1) we have

$$\sum_{i=1}^{d} \Pi^{(\ell)}_{\mathcal{C}_1+\{i\},\mathcal{C}_2} = \sum_{i=1}^{d} \frac{\eta\big[M(\mathcal{C}+\{i\}),\overline{\mathbf{X}}\big]}{\eta\big(\emptyset,\overline{\mathbf{X}}\big)} \cdot \frac{s^*!/[s^* - |M(\mathcal{C}+\{i\})|]!}{(2\ell)!/[2\ell - |M(\mathcal{C}+\{i\})|]!}, \tag{5.2}$$

where we use the fact that

$$M(\mathcal{C}_1+\{i\}) \cup M(\mathcal{C}_2) = M(\mathcal{C}_1+\mathcal{C}_2+\{i\}) = M(\mathcal{C}+\{i\}).$$

Also, note that $M(\mathcal{C}+\{i\}) = M(\mathcal{C})$ for $i \in M(\mathcal{C})$. In addition, it holds that $M(\mathcal{C}+\{i\}) = M(\mathcal{C}) \cup \{i\}$ and $|M(\mathcal{C}+\{i\})| = |M(\mathcal{C})| + 1$ for $i \notin M(\mathcal{C})$. From (5.2) we have

$$\sum_{i=1}^{d} \Pi^{(\ell)}_{\mathcal{C}_1+\{i\},\mathcal{C}_2} \tag{5.3}$$

$$= \overbrace{\sum_{i\in M(\mathcal{C})} \frac{\eta\big[M(\mathcal{C}),\overline{\mathbf{X}}\big]}{\eta\big(\emptyset,\overline{\mathbf{X}}\big)} \cdot \frac{s^*!/[s^* - |M(\mathcal{C})|]!}{(2\ell)!/[2\ell - |M(\mathcal{C})|]!}}^{\text{(i)}}$$

$$+ \underbrace{\sum_{i\notin M(\mathcal{C})} \frac{\eta\big[M(\mathcal{C})\cup\{i\},\overline{\mathbf{X}}\big]}{\eta\big(\emptyset,\overline{\mathbf{X}}\big)} \cdot \frac{s^*!/[s^* - |M(\mathcal{C})| - 1]!}{(2\ell)!/[2\ell - |M(\mathcal{C})| - 1]!}}_{\text{(ii)}}.$$

In the following, we characterize the relationship between $\eta\big[M(\mathcal{C}),\overline{\mathbf{X}}\big]$ and $\eta\big[M(\mathcal{C})\cup\{i\},\overline{\mathbf{X}}\big]$ with $i \notin M(\mathcal{C})$. We define $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_{\eta[M(\mathcal{C}),\overline{\mathbf{X}}]} \subseteq [d]$ to be the distinct sets that satisfy $|\mathcal{S}_j| = 2\ell - |M(\mathcal{C})|$, $M(\mathcal{C}) \cap \mathcal{S}_j = \emptyset$, as well as $\operatorname{sign}\big(\overline{\mathbf{X}}_{\mathcal{S}_j\cup M(\mathcal{C}),\mathcal{S}_j\cup M(\mathcal{C})}\big) = \mathbf{1}_{2\ell\times 2\ell}$ for every $j \in \big\{1,\ldots,\eta\big[M(\mathcal{C}),\overline{\mathbf{X}}\big]\big\}$. Setting $\mathcal{S}^\sharp = \cup_{j=1}^{\eta[M(\mathcal{C}),\overline{\mathbf{X}}]}\mathcal{S}_j$, we have that

$$\sum_{i\notin M(\mathcal{C})} \eta\big[M(\mathcal{C})\cup\{i\},\overline{\mathbf{X}}\big] = \sum_{i\in\mathcal{S}^\sharp} \eta\big[M(\mathcal{C})\cup\{i\},\overline{\mathbf{X}}\big]$$

$$= \sum_{i\in\mathcal{S}^\sharp} \sum_{j=1}^{\eta[M(\mathcal{C}),\overline{\mathbf{X}}]} \mathbb{1}(i\in\mathcal{S}_j) = \sum_{j=1}^{\eta[M(\mathcal{C}),\overline{\mathbf{X}}]} \sum_{i\in\mathcal{S}^\sharp} \mathbb{1}(i\in\mathcal{S}_j)$$

$$= \sum_{j=1}^{\eta[M(\mathcal{C}),\overline{\mathbf{X}}]} |\mathcal{S}_j| = \eta\big[M(\mathcal{C}),\overline{\mathbf{X}}\big]\cdot[2\ell - |M(\mathcal{C})|].$$

Here the first equality is from $\eta\big[M(\mathcal{C})\cup\{i\},\overline{\mathbf{X}}\big] = 0$ for $i \notin \mathcal{S}^\sharp$, since in this case

$$\operatorname{sign}\big(\overline{\mathbf{X}}_{M(\mathcal{C})\cup\{i\},M(\mathcal{C})\cup\{i\}}\big) \neq \mathbf{1}_{|M(\mathcal{C})\cup\{i\}|,|M(\mathcal{C})\cup\{i\}|}.$$

The second equality holds because to calculate $\eta\big[M(\mathcal{C})\cup\{i\},\overline{\mathbf{X}}\big]$, we only need to count the number of $\mathcal{S}_j$'s that include i. The last equality is from $|\mathcal{S}_j| = 2\ell - |M(\mathcal{C})|$.

Therefore, for term (ii) in (5.3) we have

$$\sum_{i\notin M(\mathcal{C})} \frac{\eta\big[M(\mathcal{C})\cup\{i\},\overline{\mathbf{X}}\big]}{\eta\big(\emptyset,\overline{\mathbf{X}}\big)} \cdot \frac{s^*!/[s^* - |M(\mathcal{C})| - 1]!}{(2\ell)!/[2\ell - |M(\mathcal{C})| - 1]!}$$

$$= \frac{\eta\big[M(\mathcal{C}),\overline{\mathbf{X}}\big]}{\eta\big(\emptyset,\overline{\mathbf{X}}\big)} \cdot (2\ell - |M(\mathcal{C})|) \cdot \frac{s^*!/[s^* - |M(\mathcal{C})| - 1]!}{(2\ell)!/[2\ell - |M(\mathcal{C})| - 1]!}$$

$$= \frac{\eta\big[M(\mathcal{C}),\overline{\mathbf{X}}\big]}{\eta\big(\emptyset,\overline{\mathbf{X}}\big)} \cdot \frac{s^*!/[s^* - |M(\mathcal{C})| - 1]!}{(2\ell)!/[2\ell - |M(\mathcal{C})|]!}. \tag{5.4}$$

Meanwhile, for term (i) in (5.3) we have

$$\sum_{i\in M(\mathcal{C})} \frac{\eta\big[M(\mathcal{C}),\overline{\mathbf{X}}\big]}{\eta\big(\emptyset,\overline{\mathbf{X}}\big)} \cdot \frac{s^*!/[s^* - |M(\mathcal{C})|]!}{(2\ell)!/[2\ell - |M(\mathcal{C})|]!} \tag{5.5}$$

$$= |M(\mathcal{C})| \cdot \frac{\eta\big[M(\mathcal{C}),\overline{\mathbf{X}}\big]}{\eta\big(\emptyset,\overline{\mathbf{X}}\big)} \cdot \frac{s^*!/[s^* - |M(\mathcal{C})|]!}{(2\ell)!/[2\ell - |M(\mathcal{C})|]!}$$

$$= (|M(\mathcal{C})| - s^*) \cdot \frac{\eta\big[M(\mathcal{C}),\overline{\mathbf{X}}\big]}{\eta\big(\emptyset,\overline{\mathbf{X}}\big)} \cdot \frac{s^*!/[s^* - |M(\mathcal{C})|]!}{(2\ell)!/[2\ell - |M(\mathcal{C})|]!}$$

$$+ s^* \cdot \frac{\eta\big[M(\mathcal{C}),\overline{\mathbf{X}}\big]}{\eta\big(\emptyset,\overline{\mathbf{X}}\big)} \cdot \frac{s^*!/[s^* - |M(\mathcal{C})|]!}{(2\ell)!/[2\ell - |M(\mathcal{C})|]!}$$

$$= -\frac{\eta\big[M(\mathcal{C}),\overline{\mathbf{X}}\big]}{\eta\big(\emptyset,\overline{\mathbf{X}}\big)} \cdot \frac{s^*!/[s^* - |M(\mathcal{C})| - 1]!}{(2\ell)!/[2\ell - |M(\mathcal{C})|]!}$$

$$+ s^* \cdot \frac{\eta\big[M(\mathcal{C}),\overline{\mathbf{X}}\big]}{\eta\big(\emptyset,\overline{\mathbf{X}}\big)} \cdot \frac{s^*!/[s^* - |M(\mathcal{C})|]!}{(2\ell)!/[2\ell - |M(\mathcal{C})|]!}.$$

Plugging (5.4) and (5.5) into (5.3), we obtain

$$\sum_{i=1}^{d} \Pi^{(\ell)}_{\mathcal{C}_1+\{i\},\mathcal{C}_2} = s^* \cdot \frac{\eta\big[M(\mathcal{C}),\overline{\mathbf{X}}\big]}{\eta\big(\emptyset,\overline{\mathbf{X}}\big)} \cdot \frac{s^*!/[s^* - |M(\mathcal{C})|]!}{(2\ell)!/[2\ell - |M(\mathcal{C})|]!}$$

$$= s^* \Pi^{(\ell)}_{\mathcal{C}_1,\mathcal{C}_2}.$$

Thus, we conclude that $\mathbf{\Pi}^{(\ell)}$ satisfies all the constraints of the ℓ-th level SoS program in (3.6) except $\mathbf{\Pi}^{(\ell)} \succeq \mathbf{0}$. We defer the verification of this constraint to the end of the proof. Next we calculate the value of objective function corresponding to $\mathbf{\Pi}^{(\ell)}$. Note that

$$\sum_{i,j=1}^{d} \overline{X}_{i,j} \cdot \Pi^{(\ell)}_{\{i\},\{j\}} = \sum_{i,j=1}^{d} \overline{X}_{i,j} \cdot \operatorname{sign}\big(\overline{X}_{i,j}\big) \cdot \Pi^{(\ell)}_{\{i\},\{j\}}$$

$$= \sum_{i,j=1}^{d} \overline{X}_{i,j} \cdot \mathbb{1}\big(\overline{X}_{i,j} > 0\big) \cdot \Pi^{(\ell)}_{\{i\},\{j\}},$$

where the first equality holds because by the definition of $\eta(\cdot,\cdot)$, it holds $\eta\big(\{i,j\},\overline{\mathbf{X}}\big) = 0$ for $\overline{X}_{i,j} \le 0$, which implies $\Pi^{(\ell)}_{\{i\},\{j\}} = 0$ correspondingly. Moreover, we have

$$\sum_{i,j=1}^{d} \Pi^{(\ell)}_{\{i\},\{j\}} = \sum_{j=1}^{d}\sum_{i=1}^{d} \Pi^{(\ell)}_{\{i\},\{j\}} = \sum_{j=1}^{d} s^* \Pi^{(\ell)}_{\emptyset,\{j\}}$$

$$= s^* \sum_{j=1}^{d} \Pi^{(\ell)}_{\{j\},\emptyset} = s^* \cdot s^* \Pi^{(\ell)}_{\emptyset,\emptyset} = (s^*)^2,$$

where the third and second last equalities are from the constraint $\sum_{i=1}^{d} \Pi^{(\ell)}_{\mathcal{C}_1+\{i\},\mathcal{C}_2} = s^* \Pi^{(\ell)}_{\mathcal{C}_1,\mathcal{C}_2}$, while the last is from $\Pi^{(\ell)}_{\emptyset,\emptyset} = 1$. Similarly, we have

$$\sum_{i=1}^{d} \Pi^{(\ell)}_{\{i\},\{i\}} = \sum_{i=1}^{d} \Pi^{(\ell)}_{\{i\},\emptyset} = s^* \Pi^{(\ell)}_{\emptyset,\emptyset} = s^*,$$

where the first equality follows from the constraints that $\Pi^{(\ell)}_{\mathcal{C}_1+\{i,i\},\mathcal{C}_2} = \Pi^{(\ell)}_{\mathcal{C}_1+\{i\},\mathcal{C}_2}$ and $\Pi^{(\ell)}_{\mathcal{C}_1,\mathcal{C}_2} = \Pi^{(\ell)}_{\mathcal{C}_1',\mathcal{C}_2'}$ for $\mathcal{C}_1 + \mathcal{C}_2 = \mathcal{C}_1' + \mathcal{C}_2'$, and the second is from $\sum_{i=1}^{d} \Pi^{(\ell)}_{\mathcal{C}_1+\{i\},\mathcal{C}_2} = s^* \Pi^{(\ell)}_{\mathcal{C}_1,\mathcal{C}_2}$. Recall that $|X_{i,j}| \geq \nu$ almost surely and the objective function is equivalent to

$$\begin{aligned}
&\frac{1}{s^*(s^*-1)} \sum_{i,j=1}^{d} X_{i,j} \Pi^{(\ell)}_{\{i\},\{j\}} \\
&= \frac{1}{s^*(s^*-1)} \sum_{i,j=1}^{d} \overline{X}_{i,j} \cdot \mathbb{1}\left(\overline{X}_{i,j} > 0\right) \cdot \Pi^{(\ell)}_{\{i\},\{j\}} \\
&\quad - \frac{\nu}{s^*(s^*-1)} \sum_{i=1}^{d} \Pi^{(\ell)}_{\{i\},\{i\}} \\
&\geq \frac{\nu}{s^*(s^*-1)} \sum_{i,j=1}^{d} \Pi^{(\ell)}_{\{i\},\{j\}} - \frac{\nu}{s^*(s^*-1)} \sum_{i=1}^{d} \Pi^{(\ell)}_{\{i\},\{i\}} \\
&= \frac{\nu[(s^*)^2 - s^*]}{s^*(s^*-1)} \geq \nu.
\end{aligned}$$

Hence, the objective value corresponding to $\mathbf{\Pi}^{(\ell)}$ is ν. Because $\widehat{\beta} \in \mathcal{H}^{(\ell)}$ is the maximum of the ℓ-th level SoS program or its relaxed versions, so far we obtain

$$\mathbb{P}\big(\widehat{\beta} \geq \nu \mid \mathbf{\Pi}^{(\ell)} \succeq 0\big) = 1. \tag{5.6}$$

In the sequel, we verify that $\mathbf{\Pi}^{(\ell)} \succeq 0$ holds with high probability. We invoke Theorem 2.5 of (Meka et al., 2015). They consider a matrix $\mathbf{M}^{(\ell)} \in \mathbb{R}^{\sum_{j=0}^{\ell} \binom{d}{j} \times \sum_{j=0}^{\ell} \binom{d}{j}}$ indexed by sets $\mathcal{S}_1, \mathcal{S}_2 \subseteq [d]$, which satisfies $M^{(\ell)}_{\mathcal{S}_1,\mathcal{S}_2} = \Pi^{(\ell)}_{\mathcal{C}_1,\mathcal{C}_2}$ for $\mathcal{S}_1 = M(\mathcal{C}_1)$ and $\mathcal{S}_2 = M(\mathcal{C}_2)$. Their result implies that under the distribution within $\mathcal{P}(s^*, d)$ specified at the beginning of our proof, $\mathbf{M}^{(\ell)} \succeq \mathbf{0}$ holds with probability at least $1/2$ for sufficiently large s^* and d, and $s^* = o\big\{\big[d/(\log d)^2\big]^{1/2\ell}\big\}$. Note $\mathbf{M}^{(\ell)}$ is a submatrix of $\mathbf{\Pi}^{(\ell)}$, i.e.,

$$\mathbf{M}^{(\ell)} = \mathbf{\Pi}^{(\ell)}_{\{\mathcal{C}:|\mathcal{C}|=|M(\mathcal{C})|\},\{\mathcal{C}:|\mathcal{C}|=|M(\mathcal{C})|\}}.$$

In other words, we can simultaneously permute the rows and columns of $\mathbf{\Pi}^{(\ell)}$, which are indexed by the collection $\mathcal{C}$'s that satisfy $|\mathcal{C}| = |M(\mathcal{C})|$, to the upper-left corner of $\mathbf{\Pi}^{(\ell)}$. Then $\mathbf{M}^{(\ell)}$ is identical to such a $\sum_{j=0}^{\ell} \binom{d}{j} \times \sum_{j=0}^{\ell} \binom{d}{j}$ upper-left submatrix of $\mathbf{\Pi}^{(\ell)}$. Meanwhile, note that by (5.1) we have

$$\begin{aligned}
&\mathbf{\Pi}^{(\ell)}_{\mathcal{C}_1,*} = \mathbf{\Pi}^{(\ell)}_{\mathcal{C}_2,*},\ \mathbf{\Pi}^{(\ell)}_{*,\mathcal{C}_1} = \mathbf{\Pi}^{(\ell)}_{*,\mathcal{C}_2}, \\
&\text{for all } |\mathcal{C}_1| = |M(\mathcal{C}_1)|,\ M(\mathcal{C}_1) = M(\mathcal{C}_2).
\end{aligned}$$

Here $\mathbf{\Pi}^{(\ell)}_{\mathcal{C},*}$ and $\mathbf{\Pi}^{(\ell)}_{*,\mathcal{C}}$ denote the row and column corresponding to collection $\mathcal{C}$. Thus, for any vector $\mathbf{u} \in \mathbb{R}^{d^{(\ell)}}$, we have

$$\begin{aligned}
&\mathbf{u}^\top \mathbf{\Pi}^{(\ell)} \mathbf{u} \qquad (5.7) \\
&= \mathbf{u}^\top \left[\sum_{\mathcal{C}_1:|\mathcal{C}_1|=|M(\mathcal{C}_1)|} \left(\sum_{\mathcal{C}_1':M(\mathcal{C}_1')=M(\mathcal{C}_1)} u_{\mathcal{C}_1'} \right) \mathbf{\Pi}^{(\ell)}_{*,\mathcal{C}_1} \right] \\
&= \sum_{\mathcal{C}_2} u_{\mathcal{C}_2} \left[\sum_{\mathcal{C}_1:|\mathcal{C}_1|=|M(\mathcal{C}_1)|} \left(\sum_{\mathcal{C}_1':M(\mathcal{C}_1')=M(\mathcal{C}_1)} u_{\mathcal{C}_1'} \right) \mathbf{\Pi}^{(\ell)}_{\mathcal{C}_2,\mathcal{C}_1} \right] \\
&= \sum_{\mathcal{C}_2:|\mathcal{C}_2|=|M(\mathcal{C}_2)|} \left(\sum_{\mathcal{C}_2':M(\mathcal{C}_2')=M(\mathcal{C}_2)} u_{\mathcal{C}_2'} \right) \\
&\qquad \left[\sum_{\mathcal{C}_1:|\mathcal{C}_1|=|M(\mathcal{C}_1)|} \left(\sum_{\mathcal{C}_1':M(\mathcal{C}_1')=M(\mathcal{C}_1)} u_{\mathcal{C}_1'} \right) \mathbf{\Pi}^{(\ell)}_{\mathcal{C}_2,\mathcal{C}_1} \right] \\
&= \bar{\mathbf{u}}^\top \mathbf{M}^{(\ell)} \bar{\mathbf{u}},
\end{aligned}$$

where $\bar{\mathbf{u}} \in \mathbb{R}^{\sum_{j=0}^{\ell} \binom{d}{j}}$ is indexed by sets and $\bar{u}_{\mathcal{S}} = \sum_{\mathcal{C}:M(\mathcal{C})=\mathcal{S}} u_{\mathcal{C}}$. Therefore, using (5.7) and the fact that $\mathbf{M}^{(\ell)} \succeq \mathbf{0}$ with probability at least $1/2$, we have $\mathbf{\Pi}^{(\ell)} \succeq \mathbf{0}$ holds with the same probability. Moreover, according to (5.6) and our setting that $\beta^* = 0$, by Markov's inequality we have

$$\begin{aligned}
&\mathbb{E}\big|\widehat{\beta} - \beta^*\big| \geq \nu \cdot \mathbb{P}\big(\widehat{\beta} \geq \nu\big) \qquad (5.8) \\
&\geq \nu \cdot \mathbb{P}\big(\widehat{\beta} \geq \nu \mid \mathbf{\Pi}^{(\ell)} \succeq \mathbf{0}\big) \cdot \mathbb{P}\big(\mathbf{\Pi}^{(\ell)} \succeq \mathbf{0}\big) \\
&\geq 1/2 \cdot \nu
\end{aligned}$$

for all $\widehat{\beta} \in \mathcal{H}^{(\ell)}$ and $s^* = o\big\{\big[d/(\log d)^2\big]^{1/2\ell}\big\}$. Recall ν is a positive constant and our construction of distributions are within $\mathcal{P}(s^*, d)$. Hence, we conclude the proof. □

6. Conclusions

In this paper, we investigate the statistical limits of convex relaxations for two statistical problems: mean estimation for sparse principal submatrix and edge probability estimation for stochastic block model. Different from existing works, which consider the statistical limits of general polynomial-time algorithms, we instead characterize the loss in statistical rates incurred by a broad family of convex relaxations. At the core of our main theoretical results is a construction-based proof, which does not hinge on any unproven hardness hypotheses. Our conclusion is that in order to attain computational tractability with convex relaxations, under particular regimes we have to compromise the statistical optimality.

References

Arias-Castro, Ery and Verzelen, Nicolas. Community detection in dense random networks. *The Annals of Statistics*, 42(3):940–969, 06 2014.

Barak, Boaz and Moitra, Ankur. Tensor prediction, rademacher complexity and random 3-xor. *arXiv preprint arXiv:1501.06521*, 2015.

Barak, Boaz and Steurer, David. Sum-of-squares proofs and the quest toward optimal algorithms. *arXiv preprint arXiv:1404.5236*, 2014.

Berthet, Quentin and Rigollet, Philippe. Computational lower bounds for sparse PCA. *arXiv preprint arXiv:1304.0828*, 2013a.

Berthet, Quentin and Rigollet, Philippe. Optimal detection of sparse principal components in high dimension. *The Annals of Statistics*, 41(4):1780–1815, 2013b.

Bhaskara, Aditya, Charikar, Moses, Chlamtac, Eden, Feige, Uriel, and Vijayaraghavan, Aravindan. Detecting high log-densities: An $o(n^{1/4})$ approximation for densest k-subgraph. In *ACM Symposium on Theory of Computing*, pp. 201–210, 2010.

Bühlmann, Peter and van de Geer, Sara. *Statistics for high-dimensional data: Methods, theory and applications*. Springer, 2011.

Butucea, Cristina and Ingster, Yuri I. Detection of a sparse submatrix of a high-dimensional noisy matrix. *Bernoulli*, 19(5B):2652–2688, 11 2013.

Butucea, Cristina, Ingster, Yuri I, and Suslina, Irina. Sharp variable selection of a sparse submatrix in a high-dimensional noisy matrix. *arXiv preprint arXiv:1303.5647*, 2013.

Cai, T Tony, Liang, Tengyuan, and Rakhlin, Alexander. Computational and statistical boundaries for submatrix localization in a large noisy matrix. *arXiv preprint arXiv:1502.01988*, 2015.

Chandrasekaran, Venkat and Jordan, Michael I. Computational and statistical tradeoffs via convex relaxation. *Proceedings of the National Academy of Sciences*, 110 (13):1181–1190, 2013.

Chandrasekaran, Venkat, Recht, Benjamin, Parrilo, Pablo A, and Willsky, Alan S. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

Chen, Yudong and Xu, Jiaming. Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *arXiv preprint arXiv:1402.1267*, 2014.

Chlamtac, Eden and Tulsiani, Madhur. Convex relaxations and integrality gaps. In *Handbook on semidefinite, conic and polynomial optimization*, pp. 139–169. Springer, 2012.

Coja-Oghlan, Amin. Graph partitioning via adaptive spectral techniques. *Combinatorics, Probability and Computing*, 19(02):227–284, 2010.

Decelle, Aurelien, Krzakala, Florent, Moore, Cristopher, and Zdeborová, Lenka. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6):066106, 2011.

Deshpande, Yash and Montanari, Andrea. Improved sum-of-squares lower bounds for hidden clique and hidden submatrix problems. *arXiv preprint arXiv:1502.06590*, 2015.

Fortunato, Santo. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010.

Gao, Chao, Ma, Zongming, and Zhou, Harrison H. Sparse CCA: Adaptive estimation and computational barriers. *arXiv preprint arXiv:1409.8565*, 2014.

Hajek, Bruce, Wu, Yihong, and Xu, Jiaming. Computational lower bounds for community detection on random graphs. *arXiv preprint arXiv:1406.6625*, 2014.

Kolar, Mladen, Balakrishnan, Sivaraman, Rinaldo, Alessandro, and Singh, Aarti. Minimax localization of structural information in large noisy matrices. In *Advances in Neural Information Processing Systems*, pp. 909–917, 2011.

Krauthgamer, Robert, Nadler, Boaz, and Vilenchik, Dan. Do semidefinite relaxations really solve sparse PCA? *arXiv preprint arXiv:1306.3690*, 2013.

Kučera, Luděk. Expected complexity of graph partitioning problems. *Discrete Applied Mathematics*, 57(2):193–212, 1995.

Lasserre, Jean B. Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization*, 11(3):796–817, 2001.

Laurent, Monique. A comparison of the Sherali-Adams, Lovász-Schrijver, and Lasserre relaxations for 0-1 programming. *Mathematics of Operations Research*, 28(3): 470–496, 2003.

Lovász, László and Schrijver, Alexander. Cones of matrices and set-functions and 0-1 optimization. *SIAM Journal on Optimization*, 1(2):166–190, 1991.

Ma, Tengyu and Wigderson, Avi. Sum-of-squares lower bounds for sparse PCA. In *Advances in Neural Information Processing Systems*, pp. 1603–1611, 2015.

Ma, Zongming and Wu, Yihong. Computational barriers in minimax submatrix detection. *arXiv preprint arXiv:1309.5914*, 2013.

Massoulié, Laurent. Community detection thresholds and the weak Ramanujan property. In *ACM Symposium on Theory of Computing*, pp. 694–703, 2014.

Meka, Raghu, Potechin, Aaron, and Wigderson, Avi. Sum-of-squares lower bounds for the planted clique problem. In *ACM Symposium on Theory of Computing*, 2015.

Mossel, Elchanan, Neeman, Joe, and Sly, Allan. Stochastic block models and reconstruction. *arXiv preprint arXiv:1202.1499*, 2012.

Mossel, Elchanan, Neeman, Joe, and Sly, Allan. A proof of the block model threshold conjecture. *arXiv preprint arXiv:1311.4115*, 2013.

Parrilo, Pablo A. *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization*. PhD thesis, California Institute of Technology, 2000.

Parrilo, Pablo A. Semidefinite programming relaxations for semialgebraic problems. *Mathematical Programming*, 96 (2):293–320, 2003.

Shabalin, Andrey A, Weigman, Victor J, Perou, Charles M, and Nobel, Andrew B. Finding large average submatrices in high dimensional data. *The Annals of Applied Statistics*, pp. 985–1012, 2009.

Sherali, Hanif D and Adams, Warren P. A hierarchy of relaxations between the continuous and convex hull representations for zero-one programming problems. *SIAM Journal on Discrete Mathematics*, 3(3):411–430, 1990.

Sun, Xing and Nobel, Andrew B. On the maximal size of large-average and ANOVA-fit submatrices in a Gaussian random matrix. *Bernoulli*, 19(1):275–294, 2013.

Verzelen, Nicolas and Arias-Castro, Ery. Community detection in sparse random networks. *arXiv preprint arXiv:1308.2955*, 2013.

Wang, Tengyao, Berthet, Quentin, and Samworth, Richard J. Statistical and computational trade-offs in estimation of sparse principal components. *arXiv preprint arXiv:1408.5369*, 2014.

Zhang, Yuchen, Wainwright, Martin J, and Jordan, Michael I. Lower bounds on the performance of polynomial-time algorithms for sparse linear regression. *arXiv preprint arXiv:1402.1918*, 2014.