Labeled Graph Clustering via Projected Gradient Descent

Shiau Hong Lim
IBM Research
shonglim@sg.ibm.com

Gregory Calvez
Ecole Polytechnique
gregory.calvez@polytechnique.edu

Abstract

Advances in recovering low-rank matrices from noisy observations have led to tractable algorithms for clustering from general pairwise labels with provable performance guarantees. Based on convex relaxation, it has been shown that the ground truth clusters can be recovered with high probability under a generalized stochastic block model by solving a semidefinite program. Although tractable, the algorithm is typically too slow for sufficiently large problems in practice. Inspired by recent advances in non-convex approaches to low-rank recovery problems, we propose an algorithm based on projected gradient descent that enjoys similar provable guarantees as the convex counterpart, but can be orders of magnitude faster. Our theoretical results are further supported by encouraging empirical results.

1 Introduction

In the labeled graph clustering problem, the input is a set of pairwise observations among a finite set of n nodes. Each observation between a pair of nodes is in the form of a label. For a binary graph, the label is simply the indicator for the presence or absence of an edge, or "unknown" for an unobserved pair. From such pairwise observations, one seeks to partition the nodes into disjoint clusters such that within-cluster pairs are more closely related than between-cluster pairs. The use of general labels allows a very rich encoding of pairwise interactions and covers a wide range of graph clustering settings, including the standard binary graphs, partially observed graphs, weighted graphs as well as time-varying graphs.

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84.

Under a generalized stochastic block model, it has been shown (Lim et al., 2014) that the underlying ground truth clustering can be exactly recovered with high probability by solving a semidefinite program, which is a convex relaxation of the maximum-likelihood problem. In certain case, this approach is shown to be order-wise optimal, in the sense that the necessary and the sufficient conditions for recovery match up to a constant factor.

The convex approach, while tractable, is computationally demanding in practice and does not scale well to large data sets. Recently, non-convex optimization approachs have been shown to successfully solve many low-rank recovery problems (Gu et al., 2016; Chen and Wainwright, 2015; Sun and Luo, 2015; Zheng and Lafferty, 2015; Zhao et al., 2015; Yi et al., 2016). In particular, Chen and Wainwright (2015) outlined a general framework for a family of low-rank recovery problems and proposed an approach based on projected gradient descent. The closest setting to ours in (Chen and Wainwright, 2015) is a planted densest subgraph problem, which can be considered a single-cluster discovery problem. It is, however, unclear whether the setup can be directly generalized to the full-fledged graph clustering problem.

Inspired by results from Chen and Wainwright (2015) and Lim et al. (2014), we propose and analyze a projected gradient approach for the general labeled graph clustering problem. Our main contribution is in showing that the proposed approach enjoys similar theoretical performance guarantees in terms of recovering the underlying ground truth but can be orders of magnitude faster than the convex counterpart.

The rest of the paper is organized as follows. We provide references to some closely related works in Section 1.1. We describe our problem setup and present the algorithm in Section 2. Our main results, with most of the proofs, are presented in Section 3. Experiment results are presented in Section 4. Finally, Section 5 discusses some potential future works.

1.1 Related Works

The stochastic block model, also known as the planted partition model (Holland et al., 1983; Condon and Karp, 2001) have been widely used to analyze graph clustering algorithms. These include partially observed graphs (Oymak and Hassibi, 2011; Chen et al., 2014a), weighted graphs (Chen et al., 2014b), and the more general labeled graphs (Heimlicher et al., 2012; Lelarge et al., 2013; Lim et al., 2014). Approaches for graph clustering include spectral clustering (McSherry, 2001; Chaudhuri et al., 2012) and convex optimization (Mathieu and Schudy, 2010; Ames and Vavasis, 2011; Lim et al., 2014). The problem setting in (Saade et al., 2016) is very similar to ours. Their works focus on the 2-cluster case, where they derive necessary and sufficient conditions for partial recovery of the ground truth.

Also related are stochastic block models that allow overlapping clusters, or mixed memberships (Airoldi et al., 2008; Latouche et al., 2011; Kaufmann et al., 2016). There are generally fewer approaches with provable performance guarantees in this setting. An interesting recent work is by Kaufmann et al. (2016), where they propose and show the consistency of a spectral algorithm for the overlapping stochastic block models.

Gradient descent approach to low-rank recovery problems have been proposed by Burer and Monteiro (2005); Chen and Wainwright (2015); Sun and Luo (2015); Zheng and Lafferty (2015); Zhao et al. (2015); Yi et al. (2016). Other non-convex algorithms with provable guarantees include phase retrieval (Cands et al., 2015), EM (Wang et al., 2015) and tensor decompositions (Anandkumar et al., 2014). We refer the reader to these works and the references therein for further related literature.

2 Problem Setup and Algorithm

We assume that a ground-truth clustering exists, where n nodes are partitioned into r disjoint clusters of size $K_1, \dots, K_r$ respectively. Without loss of generality, we assume $K_1 \geq K_2 \geq \dots \geq K_r =: K$. where K is the smallest cluster size. Define membership index set $\mathcal{I}_k$ such that $i \in \mathcal{I}_k$ iff node i belongs to cluster k, for $k = 1 \dots r$. We use the shorthand $k(i)$ to refer to the cluster index of node i. Define $F^* \in \mathbb{R}^{n \times r}$ the cluster membership matrix where $F^*_{ik} = 1$ if $i \in \mathcal{I}_k$ and 0 otherwise. Let $Y^* = F^* F^{*\top}$ be the corresponding cluster matrix. Note that $\text{rank}(Y^*) = r$ and $K_1 \dots K_r$ are its singular values. We define $\kappa = \frac{K_1}{K_r}$.

We use the generalized stochastic block model as proposed in (Lim et al., 2014), where each pair (i, j) of nodes independently generates an observed label $\Omega_{ij} \in \mathfrak{O}$ where $\mathfrak{O}$ is a label set. The labels are generated according to distributions μ and ν such that $\Omega_{ij} \sim \mu$ if $k(i) = k(j)$ and $\Omega_{ij} \sim \nu$ if $k(i) \neq k(j)$. To illustrate, the standard stochastic block model would have $\mathfrak{O} = \{+1, -1\}$ that corresponds to "edge" or "no-edge" in the graph, with $\mu(+1) = p$ and $\nu(+1) = q$.

Our objective is to recover Y^* (or F^*) given the observations Ω.

2.1 Convex Approach

Let $w : \mathfrak{O} \to \mathbb{R}$ be a weight function and the matrix $W \in \mathbb{R}^{n \times n}$ be a weight matrix where $W_{ji} = W_{ij} = w(\Omega_{ij})$ for all $i, j \in \{1 \dots n\}$. The convex approach to recovering Y^* involves solving the following semidefinite program:

$$\begin{aligned} \max_{Y} \quad & \langle W, Y \rangle \\ \text{s.t.} \quad & Y \in \mathcal{S}^n_+ , \\ & 0 \leq Y_{ij} \leq 1, \forall (i,j), \end{aligned} \tag{1}$$

where $\langle W, Y \rangle = \text{trace}(W^\top Y)$ and $\mathcal{S}^n_+$ is the space of $n \times n$ real symmetric positive semidefinite matrices.

Intuitively, one would want to assign larger weights to more informative labels. The following result states the conditions sufficient for exact recovery of Y^* from Ω:

Theorem 1 (Lim et al. 2014)**.** *Suppose b is any number that satisfies $|w(l)| \leq b$ almost everywhere (a.e.) over $\mathfrak{O}$ with respect to μ and ν. There exists a universal constant $c > 0$ such that if*

$$\begin{aligned} & \min\{\mathbb{E}_\mu w, -\mathbb{E}_\nu w\} \\ & \geq c \frac{b \log n + \sqrt{n \log n} \sqrt{\max\{\text{Var}_\mu w, \text{Var}_\nu w\}}}{K}, \end{aligned} \tag{2}$$

then Y^* is the unique solution to program (1) *with high probability.*[1]

We shall focus on weight functions where

$$E := \mathbb{E}_\mu w = -\mathbb{E}_\nu w > 0. \tag{3}$$

Also, we define $\text{Var } w := \max\{\text{Var}_\mu w, \text{Var}_\nu w\}$.

In particular, we propose to use the "linear" weights

$$w^{\text{LIN}}(l) := \frac{\mu(l) - \nu(l)}{\mu(l) + \nu(l)}.$$

It is easy to see that $-1 \leq w(l) \leq 1$ for all $l \in \mathfrak{O}$ and that (3) holds for w^{LIN}. The following corollary is immediate:

[1] By high probability we mean $> 1 - n^{-d}$ where d only affects the constant c linearly.

Corollary 1. *Suppose the weight function w^{LIN} is used. There exists a universal constant c such that Y^* is the unique solution to program* (1) *with high probability if*

$$\int_{l\in\mathfrak{D}} \frac{(\mu(l)-\nu(l))^2}{\mu(l)+\nu(l)}\,\mathrm{d}\lambda \geq c\frac{n\log n}{K^2}. \tag{4}$$

It has been shown (Lim et al., 2017) that this is also the necessary condition (up to a constant factor) for recovery in the case of two equal-size clusters.

2.2 Projected Gradient Descent

Chen and Wainwright (2015) proposed a general framework for solving low-rank recovery problems based on projected gradient descent. In particular, it solves the following semidefinite problem:

$$\min_{Y} \mathcal{L}(Y) \quad \text{s.t.} \quad Y \in \mathcal{S}_+^n \,\text{and}\, Y \in \mathcal{Y}$$

via a factorized formulation:

$$\min_{F\in\mathbb{R}^{n\times r}} \mathcal{L}(FF^\top) \quad \text{s.t.} \quad F \in \mathcal{F} \tag{5}$$

where $\mathcal{L}: \mathcal{S}_+^n \to \mathbb{R}$ is a loss function.

Program (5) is in general non-convex even if the original program is convex. It is solved using a sequence of projected gradient descent updates as follows:

$$F^{t+1} = \Pi_{\mathcal{F}}\left(F^t - \eta^t \nabla\mathcal{L}(FF^\top)\right)$$

where $\eta^t > 0$ is a step size parameter, and $\Pi_{\mathcal{F}}$ denotes the Euclidean projection onto the set $\mathcal{F}$.

Chen and Wainwright (2015) provide a set of conditions under which the projected gradient descent converges to a (near)-optimal solution. They have demonstrated that the algorithm works for the planted densest subgraph problem, which can be considered a single-cluster (rank-1) version of the more general planted partition problem. The algorithm uses $\mathcal{F} = \{F : F \in [0,1]^{n\times 1}, \sum_{i=1}^n F_i = K\}$, which assumes that the cluster size K is known. It is not clear that the result can be generalized to two or more clusters.

We extend the results to solving the general labeled graph clustering problem. In particular, we propose to use

$$\mathcal{F} = \{F : F \in [0,1]^{n\times r}, \sum_{k=1}^{r} F_{ik} = 1,\, \forall i = 1\ldots n\}. \tag{6}$$

Under this choice of $\mathcal{F}$, the conditions in Chen and Wainwright (2015) do not immediately follow. In the next section, with a much simplified proof, we show that the projected gradient descent indeed converges at linear rate under the right condition.

Algorithm 1 provides the pseudocode. The projection $\Pi_{\mathcal{F}}$ can be done efficiently in $O(nr\log r)$ using the algorithm described in (Wang and Carreira-Perpiñán, 2013). Due to the linear convergence rate, the stopping condition can simply be a fixed iteration count or a threshold on the change in the loss function.

Algorithm 1 Projected Gradient Descent

Input: Weight matrix $W \in \mathbb{R}^{n\times n}$ symmetric, rank r, step size η
Output: $F \in [0,1]^{n\times r}$

1. $t \leftarrow 0$, F^0 initialized as in Section 2.3.
2. $F^{t+1} = \Pi_{\mathcal{F}}\left(F^t + \eta W F^t\right)$
3. If stopping condition satisfied, then stop and output $F = F^{t+1}$.
4. $t \leftarrow t+1$, go to step 2.

2.3 Initialization

To initialize Algorithm 1, we propose to perform spectral clustering on the normalized matrix

$$W' = \left(\frac{1}{\mathbb{E}_\mu w - \mathbb{E}_\nu w}\right)(W - \mathbb{E}_\nu w). \tag{7}$$

In particular, let U' be $n \times r$ matrix whose columns are eigenvectors of W' corresponding to its r largest absolute eigenvalues. Next, run k-means on the rows of U' to obtain an initial clustering. Finally, set F^0 according to this initial clustering.

3 Main Results

We derive our main results in this section. For any matrix A, We use $\|A\|_{\mathrm{F}}$ for the Frobenius norm, $\|A\|_{\mathrm{op}}$ for the operator norm, $\|A\|_1 = \sum_{ij}|A_{ij}|$ and $\|A\|_\infty = \max_{ij}|A_{ij}|$ respectively.

First, note that the factorization $Y^* = \bar{F}\bar{F}^\top$ is not unique. We define the equivalent class of valid solutions as follows

$$\mathcal{E}(Y^*) := \{\bar{F} \in \mathbb{R}^{n\times r} : \bar{F}\bar{F}^\top = Y^*, \bar{F} \in \mathcal{F}\}$$

where $\mathcal{F}$ is given by (6). For any $F \in \mathcal{F}$, define

$$d(F, F^*) = \min_{\bar{F}\in\mathcal{E}(Y^*)} \|F - \bar{F}\|_F.$$

For $t = 0, 1\ldots$, let

$$\bar{F}^t := \arg\min_{\bar{F}\in\mathcal{E}(Y^*)} \|F^t - \bar{F}\|_F,$$

where F^t are as defined in Algorithm 1. Define $\Lambda^t := F^{t+1} - F^t$ and $\Delta^t = F^t - \bar{F}^t$.

The following lemma characterizes the set $\mathcal{E}(Y^*)$:

Lemma 1. *Any $\bar{F} \in \mathcal{E}(Y^*)$ satisfies $\bar{F} = F^*P$ for some permutation matrix $P \in \mathbb{R}^{r\times r}$. Furthermore, for any $\bar{F}, \bar{F}' \in \mathcal{E}(Y^*)$, $\bar{F} \neq \bar{F}'$, we have $\|\bar{F} - \bar{F}'\|_F \geq 2\sqrt{K}$.*

Proof. Fix any $\bar{F} \in \mathcal{E}(Y^*)$. Since $\bar{F}\bar{F}^\top = F^*F^{*\top}$, $\bar{F}$ and F^* have the same singular values. Let $\bar{F} = U\Sigma V^\top$ and $F^* = U_0\Sigma V_0^\top$ be the rank-r SVD for $\bar{F}$ and F^* respectively. By the spectral theorem, $U = U_0Q$ for some orthogonal matrix Q, and Q is block-diagonal such that $Q\Sigma = \Sigma Q$. Let $P = V_0QV^\top$. Then P is orthogonal and $F^*P = U_0\Sigma V_0^\top V_0QV^\top = U_0Q\Sigma V^\top = \bar{F}$. Since both $\bar{F}$ and $\bar{F}_0$ are entrywise non-negative, P must be a permutation matrix. To prove the last statement, note that since $\bar{F} \neq \bar{F}'$, they must differ in at least two columns. Since all columns have disjoint support, $\|\bar{F}_{\cdot j} - \bar{F}'_{\cdot j}\|_F^2 \geq 2K$ if $\bar{F}_{\cdot j} \neq \bar{F}'_{\cdot j}$. □

The following corollary is immediate:

Corollary 2. *If $d(F, F^*) < \sqrt{K}$, then there is a unique $\bar{F} \in \mathcal{E}(Y^*)$ such that $\|F - \bar{F}\|_F < \sqrt{K}$.*

Our main result is stated as follows:

Theorem 2. *Assume that the weight function satisfies (2) and (3). Suppose $d(F^0, F^*) \leq \frac{1}{10}\sqrt{K}$. Then, with high probability, Algorithm 1 with step size $\eta = \frac{K}{35E\sqrt{\kappa}n^2}$ generates a sequence of F^t satisfying*

$$d(F^t, F^*)^2 \leq \left(1 - \frac{1}{350\kappa^3 r^2}\right)^t d(F^0, F^*)^2.$$

In other words, when F^0 is initialized sufficiently close to $\mathcal{E}(Y^*)$, the projected gradient descent converges exponentially fast toward the exact clustering, with high probability. Here, the randomness is over all possible label matrices generated from the ground truth.

3.1 Initialization

Theorem 2 requires that we initialize F^0 sufficiently close to any $\bar{F} \in \mathcal{E}(Y^*)$. As described in Section 2.3, we propose to perform spectral clustering to obtain F^0. The following theorem shows that it can provide a suitable F^0:

Theorem 3. *Assume that the weight function satisfies (3). There exists a universal constant c such that if*

$$\frac{E}{\sqrt{(2+\epsilon)\kappa r}} \geq c\frac{b\log n + \sqrt{n(\text{Var } w)\log n}}{K}, \qquad (8)$$

then the initialization by spectral clustering using $(1+\epsilon)$-approximate k-means on W' as defined by (7) outputs F^0 with $d(F^0, F^*) \leq \frac{1}{10}\sqrt{K}$ with high probability.

Proof. By matrix Bernstein inequality(Tropp, 2012), we have that with high probability,

$$\begin{aligned}\|W - \mathbb{E}W\|_{\text{op}} &\leq C\left(b\log n + \sqrt{n(\text{Var } w)\log n}\right)\\ &\overset{(a)}{\leq} \frac{EK}{\sqrt{3200(2+\epsilon)\kappa r}}\end{aligned}$$

where we use (8) in (a).

Let $\hat{U}$ and $\tilde{U}$ be the leading eigenvectors of W' and $\mathbb{E}W'$ respectively. By Lemma 5.1 of (Lei and Rinaldo, 2015), there exists an orthogonal matrix Q such that

$$\begin{aligned}\|\hat{U} - \tilde{U}Q\|_{\text{F}} &\leq \frac{2\sqrt{2r}}{K}\|W' - \mathbb{E}W'\|_{\text{op}}\\ &= \frac{\sqrt{2r}}{EK}\|W - \mathbb{E}W\|_{\text{op}} \leq \frac{1}{\sqrt{1600(2+\epsilon)\kappa}} \qquad (9)\end{aligned}$$

where we use the fact that $\mathbb{E}W' = Y^*$ whose smallest non-zero singular value is K.

By Lemma 5.3 of (Lei and Rinaldo, 2015), with $\delta_k := \sqrt{\frac{1}{K_k} + \frac{1}{\max_{l\neq k} K_l}}$ and $U := \tilde{U}Q$, we have that the number of mis-classified members in cluster k is bounded by $|S_k|$ for $k = 1\dots r$ and $\sum_{k=1}^r |S_k|\delta_k^2 \leq 8(2+\epsilon)\|\hat{U} - U\|_{\text{F}}^2$. Therefore,

$$\begin{aligned}d(F^0, F^*)^2 &\leq 2\sum_{k=1}^r |S_k|\\ &\leq 2\sum_{k=1}^r |S_k|K_2\left(\frac{1}{K_k} + \frac{1}{\max_{l\neq k} K_l}\right)\\ &\leq 16\kappa K(2+\epsilon)\|\hat{U} - U\|_{\text{F}}^2\\ &\overset{(b)}{\leq} \frac{1}{100}K\end{aligned}$$

where we use (9) in (b). □

Notice that if we hold the number of clusters r constant, then the condition (8) of Theorem 3 is identical to the condition (2) of both Theorem 1 and 2. On the other hand, if we allow the number of clusters to grow with n, then Theorem 3 admits a slower rate of growth for r than $\sqrt{n}$ (ignoring log factors), which is achievable by (2). Improving this rate is left for future work.

3.2 Preliminary Lemmas

To prove Theorem 2, we first need a few technical lemmas.

Lemma 2. *There exists a universal constant c such that if condition (2) holds, then with probability at least $1 - n^{-3}$,*

$$\|W - \mathbb{E}W\|_{op} \leq \frac{EK}{4} \quad \textit{and} \quad \|(W - \mathbb{E}W)\bar{F}\|_\infty \leq \frac{EK}{4}$$

for any $\bar{F} \in \mathcal{E}(Y^*)$.

Proof. The operator norm is bounded by matrix Bernstein inequality (Tropp, 2012). For the second inequality, note that each entry of $(W - \mathbb{E}W)\bar{F}$ is the sum of at most K_1 independent zero-mean random variables, and we can apply the standard Bernstein inequality. The bounds then follow from condition (2). We omit the details. □

Lemma 3. *Under the conditions of Theorem 2, if* $d(F^t, F^*) \leq \frac{1}{10}\sqrt{K}$*, then* $d(F^{t+1}, F^*) \leq \frac{7}{50}\sqrt{K}$*. Furthermore,* $\bar{F}^t = \bar{F}^{t+1}$.

Proof. By Algorithm 1, $F^{t+1} = \Pi_{\mathcal{F}}(\tilde{F}^{t+1})$ where $\tilde{F}^{t+1} = F^t + \eta W F^t$. The projection $\Pi_{\mathcal{F}}$ is a convex optimization problem

$$\Pi_{\mathcal{F}}(\tilde{F}^{t+1}) = \arg\min_{F \in \mathcal{F}} \|F - \tilde{F}^{t+1}\|_F^2.$$

By first-order conditions for optimality, we have that

$$\langle F^{t+1} - \tilde{F}^{t+1}, F - F^{t+1} \rangle \geq 0, \quad \forall F \in \mathcal{F}. \tag{10}$$

Taking $F = F^t$ in (10) and re-arranging, we therefore have

$$\begin{aligned}
&\|\Lambda^t\|_F^2 \\
&\leq \eta \langle W F^t, \Lambda^t \rangle \\
&= \eta \langle W, F^t (\Lambda^t)^\top \rangle \\
&= \eta \langle \mathbb{E}W, F^t (\Lambda^t)^\top \rangle + \eta \langle W - \mathbb{E}W, F^t (\Lambda^t)^\top \rangle \\
&\leq \eta \|\mathbb{E}W\|_{\mathrm{F}} \|F^t\|_{\mathrm{op}} \|\Lambda^t\|_{\mathrm{F}} + \eta \|W - \mathbb{E}W\|_{\mathrm{op}} \|F^t\|_{\mathrm{F}} \|\Lambda^t\|_{\mathrm{F}} \\
&\overset{(a)}{\leq} \eta (En) \left(\frac{11}{10}\sqrt{\kappa K}\right) \|\Lambda^t\|_{\mathrm{F}} + \eta \left(\frac{EK}{4}\right) \left(\frac{11}{10}\sqrt{n}\right) \|\Lambda^t\|_{\mathrm{F}} \\
&\leq \frac{1}{25}\sqrt{K} \|\Lambda^t\|_{\mathrm{F}}
\end{aligned}$$

where in (a), we apply Lemma 2 for $\|W - \mathbb{E}W\|_{\mathrm{op}}$ and

$$\begin{aligned}
\|F^t\|_{\mathrm{op}} \leq \|F^t - \bar{F}^t\|_{\mathrm{op}} + \|\bar{F}^t\|_{\mathrm{op}} &\leq \frac{1}{10}\sqrt{K} + \sqrt{\kappa K} \\
&\leq \frac{11}{10}\sqrt{\kappa K},
\end{aligned}$$

and

$$\|F^t\|_{\mathrm{F}} \leq \|F^t - \bar{F}^t\|_{\mathrm{F}} + \|\bar{F}^t\|_{\mathrm{F}} \leq \frac{1}{10}\sqrt{K} + \sqrt{n} \leq \frac{11}{10}\sqrt{n}.$$

Therefore,

$$\|F^{t+1} - \bar{F}^t\|_{\mathrm{F}} \leq \|\Lambda^t\|_{\mathrm{F}} + \|F^t - \bar{F}^t\|_{\mathrm{F}} \leq \frac{7}{50}\sqrt{K}.$$

The last statement follows from Corollary 2. □

Lemma 3 by itself is too weak to establish convergence, but it ensures that $d(F^{t+1}, F^*) < \sqrt{K}$ and $\bar{F}^{t+1} = \bar{F}^t$. The next lemma, which corresponds to the local descent condition in (Chen and Wainwright, 2015), is crucial.

Lemma 4. *Under the conditions of Theorem 2, suppose that* $d(F^{t+1}, F^*) \leq \tau\sqrt{K}$*. Then*

$$\langle -WF^{t+1}, \Delta^{t+1} \rangle \geq \left(\frac{1}{2} - 2\tau\right) EK \|\Delta^{t+1}\|_F^2.$$

Proof. To reduce clutter, we use $F = F^{t+1}$, $\Delta = \Delta^{t+1}$ and $\bar{F} = \bar{F}^{t+1}$ for the rest of the proof.

We have that

$$\begin{aligned}
\langle -WF, \Delta \rangle &= \langle -W, F\Delta^\top \rangle \\
&= \langle -\mathbb{E}W, F\Delta^\top \rangle + \langle -(W - \mathbb{E}W), F\Delta^\top \rangle.
\end{aligned} \tag{11}$$

We now bound the first term on the RHS. Let $k(i)$ denote the column index of node i according to $\bar{F}$. Note that $\mathbb{E}W = ES$ where S is block diagonal with $S_{ij} = 1$ if $k(i) = k(j)$ and $S_{ij} = -1$ if $k(i) \neq k(j)$. For any i, j such that $k(i) = k(j) = c$, we have

$$\begin{aligned}
-S_{ij}(F\Delta^\top)_{ij} &= F_{ic}(1 - F_{jc}) - \sum_{k \neq c} F_{ik} F_{jk} \\
&\overset{(a)}{\geq} F_{ic}(1 - F_{jc}) - \sum_{k \neq c} (1 - F_{ic}) F_{jk} \\
&= F_{ic}(1 - F_{jc}) - (1 - F_{ic})(1 - F_{jc}) \\
&= (1 - F_{jc})[1 - 2(1 - F_{ic})]
\end{aligned}$$

where in (a) we use the fact that, for any $F \in \mathcal{F}$, each row sums to 1, and therefore for any $k' \neq c$,

$$F_{ik'} \leq \sum_{k \neq c} F_{ik} = 1 - F_{ic}.$$

We therefore have that for each $c = 1 \dots r$,

$$\begin{aligned}
&\sum_{ij:k(i)=k(j)=c} -S_{ij}(F\Delta^\top)_{ij} \\
&\geq \left(\sum_{i:k(i)=c} 1 - F_{ic}\right) \left(\sum_{i:k(i)=c} 1 - 2(1 - F_{ic})\right) \\
&= \left(\sum_{i:k(i)=c} |\Delta_{ic}|\right) \left(K_c - 2 \sum_{i:k(i)=c} (1 - F_{ic})\right) \\
&\overset{(b)}{\geq} \left(\sum_{i:k(i)=c} |\Delta_{ic}|\right) \left(K_c - 2\sqrt{K_c \sum_{i:k(i)=c} (1 - F_{ic})^2}\right) \\
&\geq \left(\sum_{i:k(i)=c} |\Delta_{ic}|\right) \left(K_c - 2\sqrt{K_c}(\tau\sqrt{K})\right) \\
&\geq (1 - 2\tau) K \left(\sum_{i:k(i)=c} |\Delta_{ic}|\right)
\end{aligned} \tag{12}$$

where (b) is by Jensen's inequality.

Similarly, for any i, j such that $k(i) = c \neq d = k(j)$, we have

$$\begin{aligned}
&- S_{ij}(F\Delta^\top)_{ij} - S_{ji}(F\Delta^\top)_{ji} \\
&= \left(F_{id}(F_{jd} - 1) + \sum_{k\neq d} F_{ik}F_{jk}\right) + \\
&\quad \left(F_{jc}(F_{ic} - 1) + \sum_{k\neq c} F_{ik}F_{jk}\right) \\
&\geq (F_{ic}F_{jc} - F_{id}(1 - F_{jd})) + (F_{jd}F_{id} - F_{jc}(1 - F_{ic})) \\
&= F_{jc}(1 - 2(1 - F_{ic})) + F_{id}(1 - 2(1 - F_{jd}))
\end{aligned}$$

and therefore for each pair (c, d), $c \neq d$,

$$\begin{aligned}
&\sum_{ij:k(i)=c,k(j)=d} -S_{ij}(F\Delta^\top)_{ij} + \sum_{ij:k(i)=d,k(j)=c} -S_{ij}(F\Delta^\top)_{ij} \\
&\geq \sum_{ij:k(i)=c,k(j)=d} F_{jc}(1 - 2(1 - F_{ic})) + F_{id}(1 - 2(1 - F_{jd})) \\
&= \left(\sum_{i:k(i)=d} F_{ic}\right)\left(\sum_{i:k(i)=c} 1 - 2(1 - F_{ic})\right) + \\
&\quad \left(\sum_{i:k(i)=c} F_{id}\right)\left(\sum_{i:k(i)=d} 1 - 2(1 - F_{id})\right) \\
&\geq (1 - 2\tau)K\left(\sum_{i:k(i)=d} |\Delta_{ic}|\right) + (1 - 2\tau)K\left(\sum_{i:k(i)=c} |\Delta_{id}|\right).
\end{aligned} \tag{13}$$

Combining (12) and (13), we have

$$\langle -\mathbb{E}W, F\Delta^\top\rangle \geq (1 - 2\tau)EK\|\Delta\|_1. \tag{14}$$

We are now ready to finish the proof. Continuing from (11),

$$\begin{aligned}
&\langle -WF, \Delta\rangle \\
&= \langle -\mathbb{E}W, F\Delta^\top\rangle + \langle -(W - \mathbb{E}W), F\Delta^\top\rangle \\
&= \langle -\mathbb{E}W, F\Delta^\top\rangle + \langle -(W - \mathbb{E}W), \bar{F}\Delta^\top\rangle + \\
&\quad \langle -(W - \mathbb{E}W), \Delta\Delta^\top\rangle \\
&\geq (1 - 2\tau)EK\|\Delta\|_1 - \|(W - \mathbb{E}W)F\|_\infty\|\Delta\|_1 \\
&\quad - \|W - \mathbb{E}W\|_{\mathrm{op}}\|\Delta\|_{\mathrm{F}}^2 \\
&\overset{(c)}{\geq} (1 - 2\tau)EK\|\Delta\|_1 - \frac{EK}{2}\|\Delta\|_1 \\
&\geq \left(\frac{1}{2} - 2\tau\right)EK\|\Delta\|_{\mathrm{F}}^2
\end{aligned}$$

where in (c) we apply Lemma 2 and use the fact that $\|\Delta\|_{\mathrm{F}}^2 \leq \|\Delta\|_\infty\|\Delta\|_1 \leq \|\Delta\|_1$. □

Lemma 4 is used to derive the following key bound for the proof of the main theorem.

Lemma 5. *Under the conditions of Theorem 2, suppose that $d(F^{t+1}, F^*) \leq \frac{7}{50}\sqrt{K}$. Then*

$$\frac{1}{\eta}\langle \Lambda^t, -\Delta^{t+1}\rangle \geq \frac{1}{5}EK\|\Delta^{t+1}\|_F^2 - \frac{25^2}{32}\frac{En^2}{K}\|\Lambda^t\|_F^2$$

Proof. Using Lemma 4 with $\tau = \frac{7}{50}$, we have that

$$\begin{aligned}
&\langle -WF^t, \Delta^{t+1}\rangle \\
&= \langle -WF^{t+1} + W\Lambda^t, \Delta^{t+1}\rangle \\
&\geq \frac{11}{50}EK\|\Delta^{t+1}\|_{\mathrm{F}}^2 - |\langle (\mathbb{E}W)\Lambda^t, \Delta^{t+1}\rangle| \\
&\quad - |\langle (W - \mathbb{E}W)\Lambda^t, \Delta^{t+1}\rangle| \\
&\geq \frac{11}{50}EK\|\Delta^{t+1}\|_{\mathrm{F}}^2 - \|\mathbb{E}W\|_{\mathrm{F}}\|\Lambda^t\|_{\mathrm{F}}\|\Delta^{t+1}\|_{\mathrm{F}} \\
&\quad - \|W - \mathbb{E}W\|_{\mathrm{op}}\|\Lambda^t\|_{\mathrm{F}}\|\Delta^{t+1}\|_{\mathrm{F}} \\
&\overset{(a)}{\geq} \frac{11}{50}EK\|\Delta^{t+1}\|_{\mathrm{F}}^2 - \frac{5En}{4}\|\Lambda^t\|_{\mathrm{F}}\|\Delta^{t+1}\|_{\mathrm{F}} \\
&\overset{(b)}{\geq} \frac{11}{50}EK\|\Delta^{t+1}\|_{\mathrm{F}}^2 - \frac{EK}{50}\|\Delta^{t+1}\|_{\mathrm{F}}^2 - \frac{25^2En^2}{32K}\|\Lambda^t\|_{\mathrm{F}}^2 \\
&\geq \frac{1}{5}EK\|\Delta^{t+1}\|_{\mathrm{F}}^2 - \frac{25^2En^2}{32K}\|\Lambda^t\|_{\mathrm{F}}^2
\end{aligned} \tag{15}$$

where in (a) we apply Lemma 2 and in (b) we use the fact that $ab \leq \frac{a^2+b^2}{2}$.

Using (10) with $F = \bar{F}^t$, we have that

$$\langle WF^t - \frac{\Lambda^t}{\eta}, \Delta^{t+1}\rangle \geq 0.$$

Adding (15) to the above completes the proof. □

3.3 Proof of Theorem 2

Following Chen and Wainwright (2015), we prove by induction. By the condition of the theorem, we have $d(F^0, F^*) \leq \frac{1}{10}\sqrt{K}$. We will show that for $t = 0, 1\ldots$, if $d(F^t, F^*) \leq \frac{1}{10}\sqrt{K}$, then

$$d(F^{t+1}, F^*)^2 \leq \left(1 - \frac{1}{350\kappa^3 r^2}\right)d(F^t, F^*)^2.$$

Lemma 3 ensures that $\bar{F}^{t+1} = \bar{F}^t$. Considering the triangle formed by $\bar{F}^t$, F^t and F^{t+1}, we have that

$$\begin{aligned}
\|\Delta^{t+1}\|_{\mathrm{F}}^2 &= \|\Delta^t\|_{\mathrm{F}}^2 - \|\Lambda^t\|_{\mathrm{F}}^2 + 2\langle \Lambda^t, \Delta^{t+1}\rangle \\
&\overset{(a)}{\leq} \|\Delta^t\|_{\mathrm{F}}^2 - \|\Lambda^t\|_{\mathrm{F}}^2 \\
&\quad - \eta\frac{1}{5}EK\|\Delta^{t+1}\|_{\mathrm{F}}^2 + \eta\frac{25^2}{32}\frac{En^2}{K}\|\Lambda^t\|_{\mathrm{F}}^2 \\
&\leq \left(1 + \frac{K^2}{175\sqrt{\kappa}n^2}\right)^{-1}\|\Delta^t\|_{\mathrm{F}}^2 \\
&\leq \left(1 - \frac{K^2}{350\sqrt{\kappa}n^2}\right)\|\Delta^t\|_{\mathrm{F}}^2 \\
&\leq \left(1 - \frac{1}{350\kappa^3 r^2}\right)\|\Delta^t\|_{\mathrm{F}}^2
\end{aligned}$$

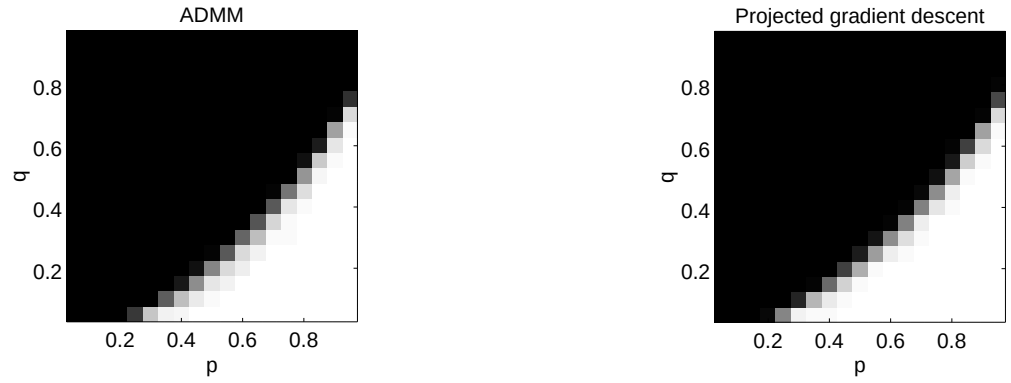

Figure 1: Ground truth recovery rate. White regions denote success while dark regions are failures.

where we apply Lemma 5 in (a). Recursively applying the above completes the proof.

4 Experiments

We evaluate the proposed algorithm on graphs generated by the stochastic block model under various parameter settings. In all our experiments, we use a constant step size of 0.01 and stopping threshold $\epsilon = 10^{-6}$ for the projected gradient descent. All k-means steps use k-means++ with 10 repetitions. We focus on comparing the performance between the convex and the proposed non-convex approach. For more extensive empirical results on the convex approach in the labeled graph clustering setting, including results on real data sets, we refer the reader to (Lim et al., 2017).

A first feasible check is performed on standard binary graphs with within-cluster edge probability p and between-cluster edge probability q, across a wide range of p and q. We measure the performance in terms of the rate of exact recovery of the underlying ground truth and compare this with the convex approach using the ADMM solver as proposed in (Lim et al., 2014).

Figure 1 shows the results over various combination of (p, q) pairs, for $n = 200$ and 4 equal-size clusters. Since we assume $p > q$, only the bottom-right half of the plot is meaningful. We clearly observe that the region of success is comparable between the two approaches. In fact, we observe a slightly wider coverage by the projected gradient approach – we attribute this to the differences in stopping conditions used in each approach.

To go beyond the binary graph, Figure 2 shows the performance for partially observable graphs, where a pair of nodes produces the label "unknown" with probability 0.5. Here, we include the results from spectral clustering, which we use for initialization. Also included are results for "vanilla" k-means on entire rows of W. The horizontal axis shows KL-divergence between the within-class label distribution μ and between-class label distribution ν. We observe that both the projected gradient descent and ADMM outperform spectral clustering and vanilla k-means, in terms of both full recovery rate as well as pairwise error rate. Figure 3 shows the same settings but with two independent observations per pair. Such observations can be obtained from, for example, multiple independent snapshots of a graph. We observe overall improvement in clustering performance as expected, with the same qualitative difference between the algorithms.

The real difference between the convex and non-convex approaches is illustrated in Figure 4, where we plot the computation time required for convergence as n grows. It is clear that the gap in terms of computation time grows dramatically with respect to n. Considering that both approaches have similar theoretical performance guarantees, we believe that the proposed approach would be more relevant in practical scenarios.

5 Discussion and Future Work

We have presented a non-convex approach based on projected gradient descent and showed that it enjoys similar theoretical performance guarantees as in the convex approach. While the bound in Theorem 2 is encouraging, we believe that it is not optimal, at least based on empirical evidence so far. In particular, the rate of convergence has a high-degree dependence on the number of clusters, making the (naive) overall time complexity $O(r^3 n^2 \log \frac{1}{\varepsilon})$. While this can be alleviated to a certain extend in the sparse case, the current rate seems suboptimal, considering that $O(rn^2 \log \frac{1}{\varepsilon})$ has been demonstrated in the case of robust PCA (Yi et al., 2016), also with projected gradient descent.

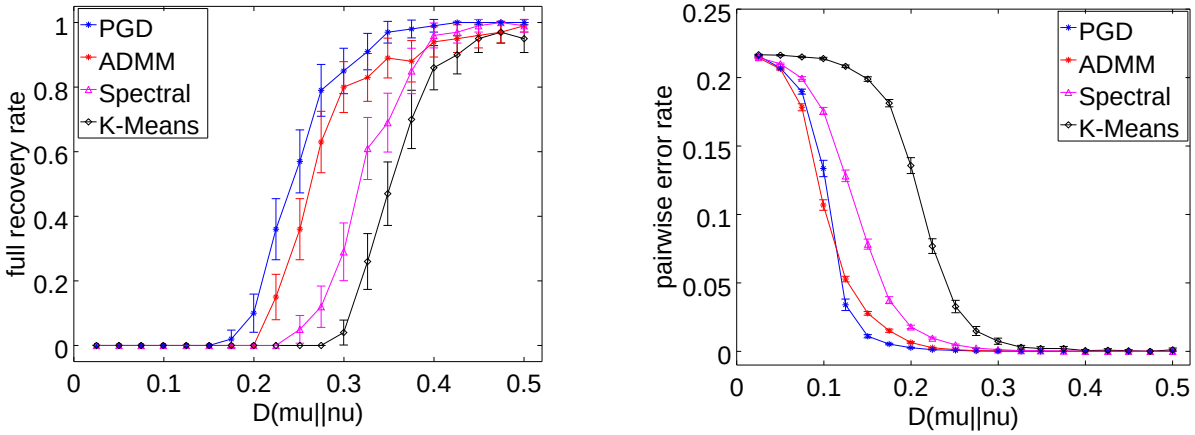

Figure 2: $n = 400$, 8 clusters

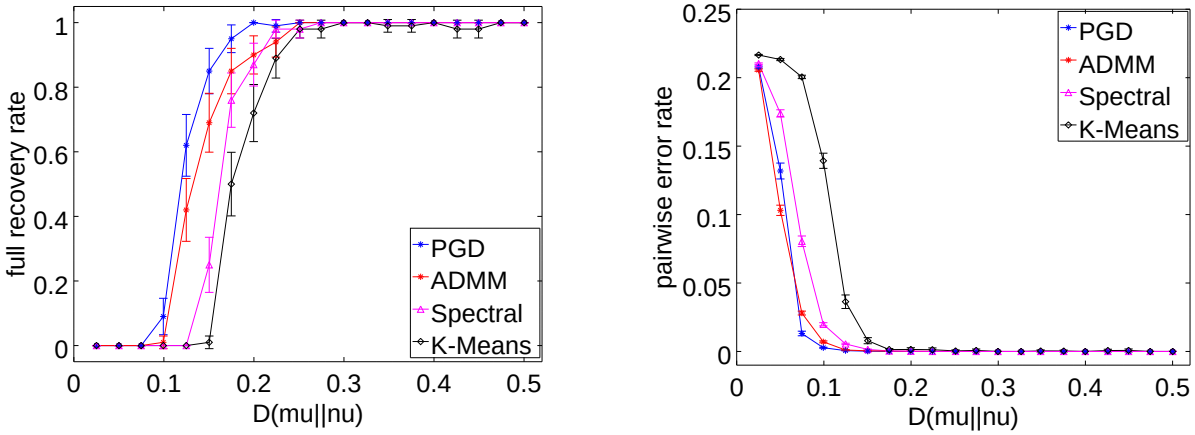

Figure 3: $n = 400$, 8 clusters, two snapshots

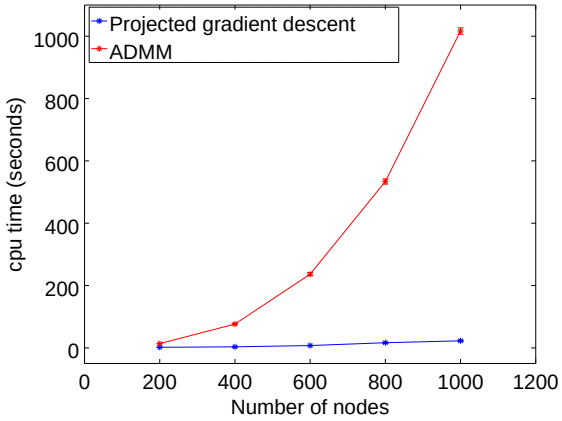

Figure 4: Computation time as n grows

Improving this rate is therefore an interesting future work.

Another direction would be to adapt the approach to handle overlapping clusters, a scenario of great practical relevance. In this case, we are currently not aware of a clear convex counterpart in terms of approaches with provable performance guarantees, making this an interesting direction for future work.

References

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008). Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.*, 9:1981–2014.

Ames, B. P. W. and Vavasis, S. (2011). Nuclear norm minimization for the planted clique and biclique problems. *Mathematical Programming*, 129(1):69–89.

Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. (2014). Tensor decompositions for learning latent variable models. *J. Mach. Learn. Res.*, 15(1):2773–2832.

Burer, S. and Monteiro, R. D. (2005). Local minima and convergence in low-rank semidefinite programming. *Mathematical Programming*, 103(3):427–444.

Cands, E. J., Li, X., and Soltanolkotabi, M. (2015). Phase retrieval via wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007.

Chaudhuri, K., Chung, F., and Tsiatas, A. (2012). Spectral clustering of graphs with general degrees in the extended planted partition model. In *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, pages 35.1–35.23.

Chen, Y., Jalali, A., Sanghavi, S., and Xu, H. (2014a). Clustering partially observed graphs via convex optimization. *Journal of Machine Learning Research*, 15:2213–2238.

Chen, Y., Lim, S. H., and Xu, H. (2014b). Weighted graph clustering with non-uniform uncertainties. In *International Conference on Machine Learning*.

Chen, Y. and Wainwright, M. J. (2015). Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *ArXiv e-prints*.

Condon, A. and Karp, R. M. (2001). Algorithms for graph partitioning on the planted partition model. *Random Structures and Algorithms*, 18(2):116–140.

Gu, Q., Wang, Z. W., and Liu, H. (2016). Low-rank and sparse structure pursuit via alternating minimization. In Gretton, A. and Robert, C. C., editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 600–609, Cadiz, Spain. PMLR.

Heimlicher, S., Lelarge, M., and Massoulié, L. (2012). Community detection in the labelled stochastic block model. In *NIPS Workshop on Algorithmic and Statistical Approaches for Large Social Networks*.

Holland, P. W., Laskey, K. B., and Leinhardt, S. (1983). Stochastic blockmodels: Some first steps. *Social Networks*, 5:109–137.

Kaufmann, E., Bonald, T., and Lelarge, M. (2016). A spectral algorithm with additive clustering for the recovery of overlapping communities in networks. In *Algorithmic Learning Theory*.

Latouche, P., Birmel, E., and Ambroise, C. (2011). Overlapping stochastic block models with application to the french political blogosphere. *Ann. Appl. Stat.*, 5(1):309–336.

Lei, J. and Rinaldo, A. (2015). Consistency of spectral clustering in stochastic block models. *Ann. Statist.*, 43(1):215–237.

Lelarge, M., Massoulié, L., and Xu, J. (2013). Reconstruction in the Labeled Stochastic Block Model. In *IEEE Information Theory Workshop*, Seville, Spain.

Lim, S. H., Chen, Y., and Xu, H. (2014). Clustering from labels and time-varying graphs. In *Advances in Neural Information Processing Systems*, pages 1188–1196.

Lim, S. H., Chen, Y., and Xu, H. (2017). Clustering from general pairwise observations with applications to time-varying graphs. *Journal of Machine Learning Research*, 18(49):1–47.

Mathieu, C. and Schudy, W. (2010). Correlation clustering with noisy input. In *Proceedings of the 21st Annual ACM-SIAM Symposium on Discrete Algorithms*, page 712.

McSherry, F. (2001). Spectral partitioning of random graphs. In *Proceedings of 42nd IEEE Symposium on Foundations of Computer Science*, pages 529–537.

Oymak, S. and Hassibi, B. (2011). Finding dense clusters via low rank + sparse decomposition. arXiv:1104.5186v1.

Saade, A., Lelarge, M., Krzakala, F., and Zdeborov, L. (2016). Clustering from sparse pairwise measurements. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pages 780–784.

Sun, R. and Luo, Z. Q. (2015). Guaranteed matrix completion via nonconvex factorization. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 270–289.

Tropp, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434.

Wang, W. and Carreira-Perpiñán, M. Á. (2013). Projection onto the probability simplex: An efficient algorithm with a simple proof, and an application. *ArXiv e-prints*.

Wang, Z., Gu, Q., Ning, Y., and Liu, H. (2015). High dimensional em algorithm: Statistical optimization and asymptotic normality. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 2521–2529. Curran Associates, Inc.

Yi, X., Park, D., Chen, Y., and Caramanis, C. (2016). Fast algorithms for robust pca via gradient descent. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, pages 4152–4160. Curran Associates, Inc.

Zhao, T., Wang, Z., and Liu, H. (2015). A nonconvex optimization framework for low rank matrix estimation. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 559–567. Curran Associates, Inc.

Zheng, Q. and Lafferty, J. (2015). A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, NIPS'15, pages 109–117, Cambridge, MA, USA. MIT Press.