On the Convergence of Block-Coordinate Maximization for Burer-Monteiro Method

Murat A. Erdogdu [1] **Asuman Ozdaglar** [2] **Pablo A. Parrilo** [2] **N. Denizcan Vanli** [2]

Abstract

Semidefinite programming (SDP) with equality constraints arise in many optimization and machine learning problems, such as Max-Cut, community detection and robust PCA. Since generic convex solvers do not scale well with the dimension of the problem, Burer and Monteiro (Burer & Monteiro, 2003) proposed to reduce the dimension of the problem by appealing to a low-rank factorization, and solve the subsequent non-convex problem instead. It is well-understood that the resulting non-convex problem acts as a reliable surrogate to the original SDP, and can be efficiently solved using the block-coordinate maximization method. Despite its simplicity, remarkable success, and wide use in practice, the theoretical understanding of the convergence of this method is limited. We prove that the block-coordinate maximization algorithm applied to the non-convex Burer-Monteiro approach enjoys a global sublinear rate without any assumptions on the problem, and a local linear convergence rate despite no local maxima is locally strongly concave.

1. Introduction

A variety of problems in statistical estimation and machine learning require solving a combinatorial optimization problem, which are often intractable (Vandenberghe & Boyd, 1996). Semidefinite programs (SDP) are commonly used as convex relaxations for these problems, providing efficient algorithms with approximate optimality (Parrilo, 2003).

[1]Microsoft Research, Cambridge, USA [2]Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, USA. Correspondence to: N. Denizcan Vanli <denizcan@mit.edu>.

Workshop on Modern Trends in Nonconvex Optimization for Machine Learning. The 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

A typically used SDP is

$$\begin{aligned} &\text{maximize} && \langle A, X\rangle \\ &\text{subject to} && X_{ii} = 1, \text{ for } i \in [n], \\ &&& X \succeq 0, \end{aligned} \tag{CVX}$$

where $A, X \in \mathbb{R}^{n\times n}$ and $[n] = \{1, 2, ..., n\}$. This problem appears as a convex relaxation to the celebrated Max-Cut problem (Goemans & Williamson, 1995), graphical model inference (Erdogdu et al., 2017), community detection problems (Bandeira et al., 2016), and group synchronization (Mei et al., 2017).

Although SDPs serve as reliable relaxations to many combinatorial problems, the resulting convex problem is still computationally challenging. Interior point methods can solve SDPs to arbitrary accuracy in polynomial-time, but they do not scale well with the problem dimension n. A popular approach to remedy these limitations is to introduce a low-rank factorization $X = \sigma\sigma^\top$, where $\sigma \in \mathbb{R}^{n\times r}$ with r denoting the rank. This reformulation removes the positive semidefinite cone constraint in (CVX) since $X = \sigma\sigma^\top$ is guaranteed to be a positive semidefinite matrix, and choosing $r \ll n$ provides computational efficiency as well as storage benefits. This method is often referred to as Burer-Monteiro approach (Burer & Monteiro, 2003). Denoting i-th row of σ by σ_i, i.e., $\sigma = [\sigma_1, \sigma_2, ..., \sigma_n]^\top$, the resulting non-convex problem can be written as follows

$$\begin{aligned} &\text{maximize} && \langle A, \sigma\sigma^\top\rangle \\ &\text{subject to} && \|\sigma_i\| = 1, \text{ for } i \in [n], \end{aligned} \tag{Non-CVX}$$

where the non-convexity comes from the separable submanifold constraints $\|\sigma_i\| = 1$. In the original Burer-Monteiro approach (Burer & Monteiro, 2003), the authors propose to use an augmented Lagrangian method for a general form SDP. However, it has been recently observed that feasible methods (such as block-coordinate maximization (Javanmard et al., 2016; Wang et al., 2017), Riemannian gradient (Javanmard et al., 2016; Mei et al., 2017) and Riemannian trust-region methods (Absil et al., 2007a; Journee et al., 2010; Boumal et al., 2016b)) provide empirically faster rates since feasibility can be efficiently guaranteed via projection onto the Cartesian product of spheres. Despite many empirical evidence (Javanmard et al., 2016; Mei et al., 2017;

Wang et al., 2017), not much is known on the convergence of these feasible methods (except the Riemannian trust-region method, for which a sublinear convergence rate is shown in (Boumal et al., 2016b) and a local superlinear convergence is shown in (Absil et al., 2007a) with no rate estimate). Among these methods, block-coordinate maximization and projected Riemannian gradient ascent are simpler to implement and have computational complexity of $\mathcal{O}(nr)$ and $\mathcal{O}(n^2r)$, respectively, whereas Riemannian trust-region requires the eigendecomposition of the dual variable (which is usually computed iteratively using the power method, whose each iteration requires $\mathcal{O}(n^2)$ arithmetic operations) and the ascent step requires an additional $\mathcal{O}(n^2r)$ complexity. Furthermore, block-coordinate maximization does not have any step size or tuning parameters, unlike projected Riemannian gradient ascent and Riemannian trust-region methods. Empirical studies further consolidate the use of block-coordinate maximization by presenting excellent results on many problems with often linear convergence. In this paper, we provide the first local and global convergence rate guarantees for the block-coordinate maximization method (applied to Burer-Monteiro approach) in the literature, which are consistent with empirical performance of the algorithm.

1.1. Related Work

There are numerous papers that analyze the landscape of (Non-CVX). In particular, it is known that (CVX) admits a maxima of rank at most $r \leq n(n+1)/2$ (Barvinok, 1995; Pataki, 1998). Using this observation, it has been shown in (Burer & Monteiro, 2003; 2005; Journee et al., 2010) that when $r \geq \sqrt{2n}$, if σ is a rank deficient second-order stationary point, then σ is a global maxima for (Non-CVX) and $X = \sigma\sigma^\top$ is a global maxima for (CVX). The recent paper (Boumal et al., 2018) showed that when $r \geq \sqrt{2n}$, for almost all A, every σ that is a first-order stationary point is rank deficient . For arbitrary rank r, (Montanari, 2016) showed that all local maxima are within a $n\left\|A\right\|_2/\sqrt{r}$ gap from the (CVX) optimum, and (Mei et al., 2017) showed that any ε-approximate concave point is within a Rg(Non-CVX)$/(r-1) + n\varepsilon/2$ gap from the (CVX) optimum, where Rg(Non-CVX) is the range of the problem (Non-CVX).

(Javanmard et al., 2016) presented that when applied to solve (Non-CVX), projected Riemannian gradient ascent and block-coordinate maximization methods provide excellent numerical results, yet no convergence guarantee is provided. Similar experimental results are also observed in (Wang et al., 2017) for the block-coordinate maximization algorithm and (Mei et al., 2017) for the projected Riemannian gradient ascent algorithm. In (Boumal et al., 2016a), the authors provided a global sublinear convergence rate for the Riemannian trust-region method for general non-convex problems and these results have been used in (Boumal et al., 2016b; Mei et al., 2017) for the non-convex Burer-Monteiro approach. Augmented Lagrangian methods have been proposed to solve (Non-CVX) as well (Burer & Monteiro, 2003; 2005), however these methods do not benefit from separability of the manifold constraints, and hence are usually slower (Boumal et al., 2018).

1.2. Notations and Preliminaries

Throughout the paper, all vectors are column vectors. The superscripts are used to denote iteration counters, i.e., σ^k denotes the value of σ at iteration k. For a vector g, $\|g\|$ represents its Euclidean norm. For a matrix A, A_{ij} represents its entry at the i-th row and j-th column, $||A||_{\mathrm{F}}$ represents its Frobenius norm, and $\|A\|_1 = \max_{1\leq j\leq n}\sum_{i=1}^n |A_{ij}|$ represents its 1-norm. For a function h, ∇h and $\mathrm{grad}h$ represent its Euclidean and Riemannian gradient, respectively. Similarly, $\nabla^2 h$ and $\mathrm{Hess}h$ represent its Euclidean and Riemannian Hessian, respectively. We let $\mathbb{S}^{m-1}$ denote the unit sphere in $\mathbb{R}^m$.

Without loss of generality, we assume that A is symmetric and $A_{ii} = 0$, for all $i \in [n]$. Indeed, if A is not a symmetric matrix, then we can replace A by $(A + A^\top)/2$, which is a symmetric matrix, and the objective value (Non-CVX) remains the same for all $\sigma \in \mathbb{R}^{n\times r}$ since $\sigma\sigma^\top$ is symmetric. Similarly, replacing the diagonal entries of A by zeros decreases the objective value by the constant $\mathrm{Tr}\,(A)$, for all $\sigma \in \mathbb{R}^{n\times r}$ since the diagonal entries of $\sigma\sigma^\top$ are equal to $\|\sigma_i\|^2 = 1$.

The rest of the paper is organized as follows. In Section 2, we present the algorithm, discuss its complexity and compare it to the other feasible methods. In Section 3, we prove the global sublinear convergence of the algorithm and provide rate estimates. In Section 4, we show that the algorithm enjoys a local linear convergence rate and provide rate estimates.

2. Block-Coordinate Maximization (BCM) Algorithm

In this section, we discuss the update rule and computational complexity of the BCM algorithm. Given the current iterate σ^k, the BCM algorithm chooses a block σ_{i_k} and maximizes the objective

$$f(\sigma^k) = \sum_{i=1}^n \langle \sigma_i^k, g_i^k\rangle, \quad \text{where} \quad g_i^k := \sum_{j\neq i} A_{ij}\,\sigma_j^k,$$

over $\sigma_{i_k} \in \mathbb{S}^{r-1}$. More formally, we can write the update rule of the algorithm as follows

$$\begin{aligned}\sigma_{i_k}^{k+1} &= \arg\max_{\|\sigma\|=1} f(\sigma_1^k,\dots,\sigma_{i_k-1}^k,\sigma,\sigma_{i_k+1}^k,\dots,\sigma_n^k)\\ &= \arg\max_{\|\sigma\|=1} \langle \sigma, g_{i_k}^k\rangle + \sum_{i\neq i_k}\sum_{j\neq i} A_{ij}\langle\sigma_i^k,\sigma_j^k\rangle\\ &= \arg\max_{\|\sigma\|=1} 2\langle \sigma, g_{i_k}^k\rangle + \sum_{i\neq i_k}\sum_{j\neq i,i_k} A_{ij}\langle\sigma_i^k,\sigma_j^k\rangle\\ &= \arg\max_{\|\sigma\|=1} \langle \sigma, g_{i_k}^k\rangle = \frac{g_{i_k}^k}{\|g_{i_k}^k\|}, \end{aligned} \tag{1}$$

with the convention that $\sigma_{i_k}^{k+1}$ can be chosen arbitrarily when $\|g_{i_k}^k\| = 0$, and where the third equality follows since A is symmetric. Although i_k can be chosen arbitrarily, we focus on uniformly random selection in this paper, i.e., $i_k \overset{\text{i.i.d.}}{\sim} \text{Unif}[n]$, for all $k \geq 0$.

Algorithm 1 Block-Coordinate Maximization (BCM)

Initialize $\sigma^0 \in \mathbb{R}^{n\times r}$.
for $k = 0, 1, 2, \dots$ **do**
  Sample a block $i_k \sim \text{Unif}[n]$.
  $\sigma_{i_k}^{k+1} \leftarrow \frac{g_{i_k}^k}{\|g_{i_k}^k\|}$.
end for

The BCM algorithm with uniform sampling can be implemented in $\mathcal{O}(nr)$ time and space complexity since it only needs to save σ (which is of size nr) and after i_k is chosen $g_{i_k}^k$ can be computed in $2(n-1)r$ floating point operations. However, in many SDP applications (such as Max-Cut and graphical model inference), A is induced by a graph. Therefore, the computational cost of the BCM algorithm can be reduced to $\mathcal{O}(dr)$, where d is the maximum degree of the graph that induces A. In comparison, per iteration computational complexity of the projected Riemannian gradient ascent algorithm is $\mathcal{O}(n^2r)$ for dense A and $\mathcal{O}(d^2r)$ for sparse A. The situation is even worse for Riemannian trust-region algorithm since it requires to perform power method to solve the trust-region subproblem to find an approximate update direction. Hence, per iteration complexity of the BCM algorithm is much smaller than the other feasible methods. Furthermore, projected Riemannian gradient ascent and Riemannian trust-region methods require parameter tuning to guarantee an ascent at each step. On the other hand, the BCM algorithm does not have any tuning parameters and is guaranteed to make an ascent at each iteration as we show in Lemma 3.1.

3. Global Sublinear Convergence Rate

In this section, we prove that the BCM algorithm attains a global sublinear convergence and provide rate estimates. To this end, we first introduce the following ascent lemma, which shows that the sequence of function values $\{f(\sigma^k)\}_{k\geq 0}$ is nondecreasing.

Lemma 3.1. *Each iteration of the BCM algorithm yields the following ascent on the function value:*

$$f(\sigma^{k+1}) - f(\sigma^k) = 2\left(\|g_{i_k}^k\| - \langle \sigma_{i_k}^k, g_{i_k}^k\rangle\right) \geq 0.$$

We emphasize that such an ascent lemma does not necessarily hold for general non-convex functions and algorithms. In particular, in order to guarantee ascent condition, it is often required to use line-search techniques for choosing the step size of first-order methods (e.g., the gradient ascent algorithm) (Schneider & Uschmajew, 2015). On the other hand, the BCM algorithm does not require any parameter tuning and still enjoys the ascent guarantee in Lemma 3.1. This lemma holds a basis for the following theorem, in which we show that the expected functional ascent attained by the BCM algorithm can be related to the expected norm of the Riemannian gradient of the function evaluated at the current iterate. Hence, it is guaranteed that the BCM algorithm returns a solution with arbitrarily small Riemannian gradient as we highlight in the following theorem.

Theorem 3.2. *Let $f^* = \max_{\|\sigma_i\|=1,\forall i\in[n]} f(\sigma)$. Then, in at most $K \geq \left\lceil \left(\|A\|_{1,1}\left(f^* - f(\sigma^0)\right)\right)/\epsilon \right\rceil$ iterations, BCM is guaranteed to return a solution σ^k, for some $k \in [K-1]$, satisfying $\mathbb{E}\left\|\text{grad} f(\sigma^k)\right\|_{\text{F}}^2 \leq \epsilon$. Equivalently, for any $K \geq 1$, BCM yields the following guarantee*

$$\min_{k\in[K-1]} \mathbb{E}\left\|\text{grad} f(\sigma^k)\right\|_{\text{F}}^2 \leq \frac{n\|A\|_1\left(f^* - f(\sigma^0)\right)}{K}. \tag{2}$$

4. Local Linear Convergence Rate

Although the BCM algorithm enjoys the sublinear convergence rates presented in Section 3, it is numerically observed that the rate of convergence is linear when σ^k is close to a local maxima (Javanmard et al., 2016; Wang et al., 2017). A similar conclusion can be made by Figure 1 as well, which illustrates local linear convergence of BCM. In this section, we investigate this behavior and prove that indeed BCM attains a linear convergence rate around a local maxima. In order to prove this result, we require certain tools from manifold optimization (Absil et al., 2007b). We define the following submanifold of $\mathbb{R}^{n\times r}$ that corresponds to the Riemannian geometry induced by the constraints of the problem (Non-CVX) in the Euclidean space:

$$\mathcal{M}_r := \left\{\sigma = (\sigma_1,\dots,\sigma_n)^\top \in \mathbb{R}^{n\times r} : \|\sigma_i\| = 1, \forall i \in [n]\right\}.$$

This manifold represents the Cartesian product of n unit spheres in $\mathbb{R}^r$. For any given point $\sigma \in \mathcal{M}_r$, its tangent space can be found (by taking the differential of the equality constraints) as follows

$$T_\sigma\mathcal{M}_r := \left\{u = (u_1,\dots,u_n)^\top \in \mathbb{R}^{n\times r} : \langle u_i, \sigma_i\rangle = 0, \forall i \in [n]\right\}.$$

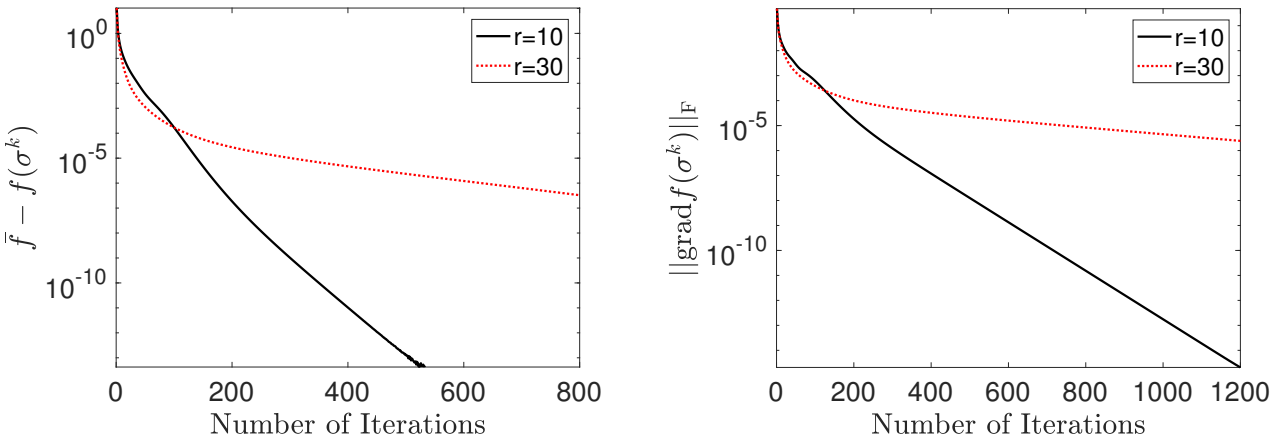

Figure 1: Convergence of the BCM algorithm, where the entries of A are drawn from a normal distribution and $n = 500$.

Using these definitions, the geodesics $t \mapsto \sigma(t)$ (i.e., curves of shortest path with zero acceleration) can be expressed as a function of $\sigma = \sigma(0) \in \mathcal{M}_r$ and $u \in T_\sigma \mathcal{M}_r$ as follows

$$\sigma_i(t) = \sigma_i \cos(\|u_i\| t) + \frac{u_i}{\|u_i\|} \sin(\|u_i\| t). \tag{3}$$

We refer to Section 5.4 of (Absil et al., 2007b) for a more detailed treatment of this topic. The above geodesic can be thought as the set of points on the manifold that are obtained by moving from $\sigma \in \mathcal{M}_r$ towards the direction pointed by $u \in T_\sigma \mathcal{M}_r$. Before understanding the landscape around a local maxima $\sigma \in \mathcal{M}_r$, we first make the following observation. Let $\mathrm{O}(r) = \{Q \in \mathbb{R}^{r \times r} : Q^\top Q = QQ^\top = I\}$ denote the orthogonal group in dimension r. We can observe that $f(\sigma Q) = \langle A, \sigma QQ^\top \sigma^\top \rangle = \langle A, \sigma\sigma^\top \rangle = f(\sigma)$, for any $Q \in \mathrm{O}(r)$. Thus, every local maxima is flat in certain directions in $T_\sigma \mathcal{M}_r$. In order to characterize these directions, we define $\bar{\mathcal{M}}_r = \mathcal{M}_r / \mathrm{O}(r)$ as the quotient of the manifold $\mathcal{M}$ by the orthogonal group $\mathrm{O}(r)$, which can be thought as the set of equivalence classes. We then consider the tangent space $T_\sigma \mathcal{M}_r$ and decompose it into two orthogonal subspaces: the vertical space $\mathcal{V}_\sigma \bar{\mathcal{M}}_r$ and the horizontal space $\mathcal{H}_\sigma \bar{\mathcal{M}}_r$. The vertial space $\mathcal{V}_\sigma \bar{\mathcal{M}}_r$ is the tangent space to equivalence classes, i.e.,

$$\mathcal{V}_\sigma \bar{\mathcal{M}}_r = \{\sigma Q : Q \in \mathbb{R}^{r \times r}, Q^\top = -Q\}.$$

This space contains the tangent vectors along which function value does not change and hence there is no curvature. The horizontal space $\mathcal{H}_\sigma \bar{\mathcal{M}}_r$ is the orthogonal complement of $\mathcal{V}_\sigma \bar{\mathcal{M}}$ in $T_\sigma \mathcal{M}_r$, i.e.,

$$\mathcal{H}_\sigma \bar{\mathcal{M}}_r = \{u \in T_\sigma \mathcal{M}_r : u^\top \sigma = \sigma^\top u\}.$$

In other words, $\mathcal{H}_\sigma \bar{\mathcal{M}}_r$ contains tangent vectors that do not rotate σ at all, which are the directions along which there is curvature. For a more detailed treatment of these definitions, we refer to Chapter 4 of (Journee et al., 2010), where similar equivalence class definitions are introduced to guarantee rotational invariance to design an algorithm, whereas our purpose here is to obtain convergence rate estimates. The main assumption we will use in proving the local linear convergence of the BCM algorithm is that along any direction in $\mathcal{H}_\sigma \bar{\mathcal{M}}_r$, $f(\sigma(t))$ has a negative curvature of at least $\mu > 0$. More formally, we make the following assumption.

Assumption 1. *Let σ be a local maxima of the problem (Non-CVX). Then, $\langle u, \mathrm{Hess} f(\sigma)[u] \rangle \leq -\mu \|u\|_F^2$ holds for all $u \in \mathcal{H}_\sigma \bar{\mathcal{M}}_r$.*

We emphasize that this assumption implies having isolated maximizers on the search space $\bar{\mathcal{M}}$, which is the assumption used in (Journee et al., 2010). In the following theorem, we state the main linear convergence rate result for the BCM algorithm. An informal version of this theorem can be stated as follows. Suppose the BCM algorithm converges to a local maxima, for which Assumption 1 holds with some constant μ. Then, the algorithm attains a local linear convergence rate of $1 - \mu/(n^2 \|A\|_1)$ per iteration and $1 - \mu/(n \|A\|_1)$ per cycle, approximately. In the formal statement of the theorem, we consider the case the sequence $\{\sigma^k\}_{k \geq 0}$ does not converge but instead has distinct limit points, where we emphasize that all limit points have the same function value due to Lemma 3.1.

Theorem 4.1. *Let $\bar{f} = \lim_{k \to \infty} f(\sigma^k)$, suppose Assumption 1 holds and assume that the limit points $\bar{\sigma}$ of the BCM algorithm are local maxima. Then, there exists an integer $K > 0$ such that the iterates generated by the BCM algorithm enjoy the following linear convergence rate*

$$\bar{f} - f(\sigma^{k+1}) \leq \left(1 - \frac{\mu}{n^2 \|A\|_1} + \delta_K\right) \left(\bar{f} - f(\sigma^k)\right), \tag{4}$$

for any $k \geq K$, where δ_K is a constant that goes to 0 as $K \to \infty$.

References

Absil, P.-A., Baker, C., and Gallivan, K. Trust-region methods on riemannian manifolds. *Foundations of Computational Mathematics*, 7(3):303–330, Jul 2007a.

Absil, P.-A., Mahony, R., and Sepulchre, R. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, Princeton, NJ, USA, 2007b.

Bandeira, A. S., Boumal, N., and Voroninski, V. On the low-rank approach for semidefinite programs arising in synchronization and community detection. *ArXiv:1602.04426*, 2016.

Barvinok, A. I. Problems of distance geometry and convex properties of quadratic maps. *Discrete & Computational Geometry*, 13(2):189–202, 1995.

Boumal, N., Absil, P.-A., and Cartis, C. Global rates of convergence for nonconvex optimization on manifolds. *arXiv preprint arXiv:1605.08101*, 2016a.

Boumal, N., Voroninski, V., and Bandeira, A. S. The non-convex Burer-Monteiro approach works on smooth semidefinite programs. In *Advances in Neural Information Processing Systems*, pp. 2757–2765, 2016b.

Boumal, N., Voroninski, V., and Bandeira, A. S. Deterministic guarantees for BurerMonteiro factorizations of smooth semidefinite programs. *arXiv preprint arXiv:1804.02008*, 2018.

Burer, S. and Monteiro, R. D. C. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

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

Erdogdu, M. A., Deshpande, Y., and Montanari, A. Inference in graphical models via semidefinite programming hierarchies. In *Advances in Neural Information Processing Systems*, pp. 416–424, 2017.

Goemans, M. X. and Williamson, D. P. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the ACM (JACM)*, 42(6):1115–1145, 1995.

Grubisic, I. and Pietersz, R. Efficient rank reduction of correlation matrices. *Linear Algebra and its Applications*, 422(2):629 – 653, 2007.

Javanmard, A., Montanari, A., and Ricci-Tersenghi, F. Phase transitions in semidefinite relaxations. *Proceedings of the National Academy of Sciences*, 113(16):E2218–E2223, 2016.

Journee, M., Bach, F., Absil, P.-A., and Sepulchre, R. Low-rank optimization on the cone of positive semidefinite matrices. *SIAM Journal on Optimization*, 20(5):2327–2351, 2010.

Mei, S., Misiakiewicz, T., Montanari, A., and Oliveira, R. I. Solving SDPs for synchronization and MaxCut problems via the Grothendieck inequality. *arXiv preprint arXiv:1703.08729*, 2017.

Montanari, A. A Grothendieck-type inequality for local maxima. *arXiv preprint arXiv:1603.04064*, 2016.

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

Pataki, G. On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues. *Mathematics of operations research*, 23(2):339–358, 1998.

Schneider, R. and Uschmajew, A. Convergence results for projected line-search methods on varieties of low-rank matrices via lojasiewicz inequality. *SIAM Journal on Optimization*, 25(1):622–646, 2015.

Vandenberghe, L. and Boyd, S. Semidefinite programming. *SIAM Review*, 38(1):49–95, 1996.

Wang, P.-W., Chang, W.-C., and Kolter, J. Z. The mixing method: coordinate descent for low-rank semidefinite programming. *arXiv preprint arXiv:1706.00476*, 2017.