EEMCS, MS c. Applied Mathematics, Optimization
to obtain the degree of Master of Science at the Delft University of Technology,
to be defended on the 26th of September 2022

Convergence of the mixing method, an iterative algorithm for solving diagonally constrained semidefinite programs

Supervisor:
David de Laat

Author:
Dominic Eelkema, 5415071

Committee:
Dion Gijswijt (Chair)
Alexander Heinlein

Abstract

This thesis explores the convergence of the mixing method, an iterative algorithm for solving diagonally constrained semidefinite programs [13]. In this paper we first give an exposition of the convergence proof for the mixing method based on the proof by Wang, Chang, and Kolter [13], where we restructure some parts of the proof and provide extra details. Then we construct an example where the linear convergence rate of the mixing method is close to one when near the optimal solution. The mixing method is then compared for convergence speed to a semidefinite programming solver and gradient descent on random max-cut instances. For instances of the max-cut, it is found that the mixing method outperforms other methods.

Contents

1 Introduction

As one of the more generally formulated convex problems, semidefinite problems (SDPs) are used to solve a great variety of different optimization problems. One of the most important foundations for semidefinite programming was the relaxation bound from Goemans and Williamson for approximating the maximum cut problem (max-cut) [6]. In his paper [6], he showed that the max-cut problem, which in general is an NP-hard problem, can be formulated as an SDP using a relaxation. Goemans et al. showed that the optimal objective value of this relaxation has a 0.878 approximation guarantee. This in turn spurred on more interest and research into finding tight approximations to NP-hard problems. Goeman's relaxation of the max-cut problem belongs to a subclass of SDPs, namely diagonally constrained SDPs. SDPs are part of the set of convex problems and are desirable as convex problems can be solved in polynomial time since locally optimal points of a convex problem are globally optimal. So convex formulations can be solved efficiently. However, there are instances in which a non-convex formulation of an optimization problem can be solved more efficiently than a convex one. This paper presents such a case for diagonally constrained semidefinite problems. Diagonally constrained SDPs are SDPs of the form

$$\begin{aligned} &\text{minimize } \langle C, X\rangle \\ &\text{subject to } \langle X, E_{i,i}\rangle = b_i \ \forall i \in [n], \\ &\qquad X \succeq 0, b_i \in \mathbb{R} \end{aligned} \tag{1}$$

where $E_{i,i}$ is the zero matrix except for a 1 at index (i,i). Pataki [10] showed that when $X \in F$, where F is a face of the feasible region, then the rank k of X can be bounded as follows:

$$\frac{k(k+1)}{2} \leq p + \dim(F),$$

where p is the number of constraints. Hence, by taking a face of dimension 0 in the optimal face, we obtain that an SDP always admits an optimal solution of rank

$$k(k+1) \leq 2p.$$

Since this problem contains n constraints, it is evident that that whenever $k(k+1) \geq 2n$, there is an optimal solution X^* that can be written as $X^* = V^T V$ with $V \in R^{k \times n}$. This allows us to rewrite the SDP as

$$\begin{aligned} &\text{minimize } \langle C, V^T V\rangle \\ &\text{subject to } \|v_i\|_2^2 = b_i \ \forall i \in [n], \\ &\qquad V \in \mathbb{R}^{k \times n}, b_i \in \mathbb{R} \end{aligned}$$

where v_i are the columns of V. Since the diagonal of X is fixed we can remove the diagonal of C out of the optimization problem as this adds no further value. Also without loss of generality we can pull in the right-hand constant value b_i into the norm on the other side by dividing by b_i. Therefore throughout the rest of the paper we assume that the diagonal of C is zero and we write the problem as

$$\begin{aligned} &\text{minimize } \langle C, V^T V \rangle \\ &\text{subject to } \|v_i\|^2 = 1 \ \forall i \in [n], \\ &V \in \mathbb{R}^{k \times n} \end{aligned}$$

This problem can then be solved heuristically with a coordinate descent approach. This brings us to the 'mixing method'.

1.1 The Mixing Method

To solve this optimization problem we optimize over a single column v_i, where we fix the rest of the columns of V. When only optimizing over a single vector, the SDP objective value becomes

$$\langle C, V^T V \rangle = 2 \sum_{j \neq i} C_{i,j} v_i^T v_j + E,$$

where $E \in \mathbb{R}$ is a constant value and there a no quadratic terms for v_i as $C_{i,i} = 0$. Let c_i denote the i-th column of C. Then the minimization problem can be simplified to

$$\begin{aligned} &\text{minimize } 2 v_i^T V c_i \\ &\text{subject to } \|v_i\|_2^2 = 1, \\ &v_i \in \mathbb{R}^k \end{aligned} \tag{2}$$

We now have linear optimization problem with a single quadratic equality constraint to solve. Given that Vc_i is a constant as it does not depend on v_i ($C_{i,i} = 0$), we can solve this optimization problem by noting that

$$\left\| v_i + \frac{V c_i}{\|V c_i\|_2} \right\|_2^2 \geq 0,$$

which implies that

$$2 v^T V c_i \geq -2 \|V c_i\|_2 ,$$

since $v^T v = 1$ and $\frac{(Vc_i)^T V c_i}{\|V c_i\|_2^2} = 1$. Now if we let

$$v_i = -\frac{V c_i}{\|V c_i\|_2},$$

then we have found a feasible solution ($v_i^T v_i = 1$) which achieves equality for this lower bound of our objective function. Hence this feasible solution found for the optimization problem when optimizing over a single column must be optimal. This process can be done iteratively, where in each iteration the next column is updated as above. This leads to an algorithm that we call the 'mixing method'. A sketch of such an algorithm is given by

Data: $C \in \mathbb{R}^{n \times n}$ symmetric and zero on the diagonal, $k \in \mathbb{Z}_{>0}$
Result: Optimal V
Randomly initialize V;
while *V not yet converged* **do**
  for $i = 1 : n$ **do**
    $v_i = -Vc_i$
    $v_i = \frac{1}{\|v_i\|} v_i$
  end
end

Algorithm 1: The mixing method

Linear convergence for the mixing method has been proven in Wang, Chang and Kolter's paper [13] for any random starting point V. They then showed that the mixing method performs better than other state of the art SDP solvers for the maximum cut problem. They also showed that for the maximum satisfibility (Max-SAT) problem there exists a similar approximation algorithm. Although the bounds were not fully proven, numerically the mixing method outperformed state of the art competitors from the 2016 Max-SAT competition for certain instances[1]. This thesis aims to build on their result, refining some of the proofs and provide key insights into the convergence dependencies of this algorithm. In chapter 3 and 4 of this thesis we provide the proof regarding convergence of the mixing method. These chapters provide more of a restructuring of the proof from Wang, Chang and Kolter's paper, adding extra details to the proof in some areas. In chapter 5 the convergence rate of the mixing method is compared to its theoretical bound numerically. Chapter 6 compares the mixing method's speed of convergence to gradient descent and a state of the art SDP solvers from Leijenhorst et al [4], which is specialized in solving SDPs with sparse constraint matrices. We end the thesis with a conclusion and future study recommendations. Before starting the proof let us first cover some of the basic notation that is used throughout the paper.

2 Preliminaries

2.1 Notation

Here we introduce some notation that will be used throughout the paper. Let S^n denote the set of $n \times n$ symmetric matrices and S^n_+ to be the set of $n \times n$ positive semidefinite matrices. For a vector $y \in \mathbb{R}^n$ let $\mathrm{Diag}(y)$ denote the zero matrix with y on the diagonal. Let $\sigma_{min}(A)$ denote the smallest singular value of the matrix A and $\sigma_{\min-\mathrm{nz}}(A)$ to be the smallest nonzero singular value of A.

The commonly used norms in this paper are defined as follows

Definition 2.1. *For $x \in \mathbb{R}^n$, define the Euclidean norm as*

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$

Definition 2.2. *For $X \in \mathbb{R}^{n\times n}$, define the Frobenius norm as*

$$\|X\|_F = \sqrt{\mathrm{tr}(X^T X)} = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} X_{i,j}^2}$$

Definition 2.3. *For $X \in \mathbb{R}^{n\times n}$, define the spectral norm as*

$$\|A\|_2 = \max_{x\neq 0,\, x\in\mathbb{R}^n} \frac{\|Ax\|_2}{\|x\|_2}$$

2.2 Norm cookbook

Below we provide lemmas with proofs for some of the most commonly used tricks when it comes to bounding the Frobenius matrix norm $\|\cdot\|_F$, spectral norm and Euclidean vector norm $\|\cdot\|_2$

Lemma 2.1. *Let $A \in S^n$. Then,*

$$\|A\|_2 = \rho(A)$$

Proof. Squaring the norm allows us to write the norm as

$$\begin{aligned}
\|A\|_2^2 &= \max_{x\neq 0,\, x\in\mathbb{R}^n} \frac{\|Ax\|_2^2}{\|x\|_2^2} \\
&= \max_{x\neq 0,\, x\in\mathbb{R}^n} \frac{x^T A^T A x}{x^T x} \\
&= \lambda_{\max}(A^T A) \\
&= \rho(A^T A),
\end{aligned}$$

where the third equality follows from the Courant–Fischer–Weyl min–max principle. See e.g. [12, Chapter 12]. □

As another consequence of this lemma we can bound the vector two norm

Lemma 2.2. *For a symmetric matrix $S \in S^n$ and $x \in \mathbb{R}^n$ we have*

$$\|x\|_2^2 \geq \frac{|x^T S x|}{\rho(S)}$$

Lemma 2.3. *For $A \in \mathbb{R}^{m\times n}$ and $x \in \mathbb{R}^n$,*

$$\|Ax\|_2 \leq \|A\|_F \|x\|_2$$

Proof. We have

$$\begin{aligned}\|Ax\|_2^2 &= \sum_{i=1}^{m}\sum_{j=1}^{n} A_{i,j}^2 x_j^2 \\ &\leq \sum_{i=1}^{m}\sum_{j=1}^{n} A_{i,j}^2 \|x\|_2^2 \\ &= \|A\|_F^2 \|x\|_2^2\end{aligned}$$

Taking the square root on both sides gives the desired inequality. □

Lemma 2.4. *Let $A \in S^n$ and $x \in \mathbb{R}^n$. Then,*

$$\|Ax\|_2 \geq \sigma_{\min}(S) \|x\|_2 .$$

Furthermore, when $x \in \text{range}(S)$

$$\|Ax\|_2 \geq \sigma_{\min-\text{nz}}(S) \|x\|_2 .$$

Proof. First note that A can be decomposed in its singular value decomposition $A = U\Sigma V^T$. Also note that for orthogonal matrices X it holds that $\|Xv\|_2 = \|v\|_2$ for any $v \in \mathbb{R}^n$ Then

$$\begin{aligned}\|Ax\|_2 &= \left\|U\Sigma V^T x\right\|_2 \\ &= \left\|\Sigma V^T x\right\|_2 \\ &= \left\|V^T \Sigma x\right\|_2 \\ &= \|\Sigma x\|_2 \\ &= \sqrt{\sigma_1^2 x_1^2 + \cdots + \sigma_n^2 x_n^2} \\ &\geq \sqrt{\sigma_{\min}^2 (x_1^2 + \cdots + x_n^2)} \\ &= \sigma_{\min} \|x\| .\end{aligned}$$

When $x \in \text{range}(S)$ it is not hard to see that then

$$\begin{aligned}\|Ax\|_2 &= \sqrt{\sigma_1^2 x_1^2 + \cdots + \sigma_n^2 x_n^2} \\ &\geq \sqrt{\sigma_{\min-\text{nz}}^2 (x_1^2 + \cdots + x_n^2)} \\ &= \sigma_{\min-\text{nz}} \|x\| .\end{aligned}$$

□

Lemma 2.5. *Let $A, B \in S_+^n$. Then,*

$$\text{tr}(AB) \geq 0.$$

Proof. Note that since $A \succeq 0$ there is a cholesky decomposition L such that $L^T L = A$. Then

$$\begin{aligned}\operatorname{tr}(AB) &= \operatorname{tr}(L^T BL) \\ &= \sum_{i=1}^{n} e_i^T L^T BLe_i \ \ (e_i \text{ being the } i\text{-th unit vector}) \\ &\geq 0,\end{aligned}$$

since $B \succeq 0$. □

This theorem is often used in the proof by noting that for any $A \in S_+^n$ and $C \in S^n$, when we pick $B = C - \lambda_{min}(C)I_n$, then $B \succeq 0$. So we obtain the inequality

$$\operatorname{tr}(AC) \geq \lambda_{\min}(C)\operatorname{tr}(A).$$

Similarly setting $B = \lambda_{max}(C)I_n - C \succeq 0$ which results in the inequality

$$\operatorname{tr}(AC) \leq \lambda_{\max}(C)\operatorname{tr}(A).$$

Another consequence of this lemma is the following

Lemma 2.6. *Let $A, B \in S_+^n$. Then,*

$$\operatorname{tr}(ABB) \geq \operatorname{tr}(AB)\lambda_{\min-\mathrm{nz}}(B)$$

Proof. Take any eigenvector v of B corresponding to the eigenvalue α. Then

$$B(B - \lambda_{\min-\mathrm{nz}}(B))v = \alpha^2 v - \lambda_{\min-\mathrm{nz}}(B)\alpha v.$$

Since $B \succeq 0$, either $\alpha = 0$ or $\alpha > 0$. For $\alpha = 0$ the above sum equals zero. For $\alpha > 0$ we know that $\alpha \geq \lambda_{\min-\mathrm{nz}}(B)$ hence

$$\alpha^2 - \lambda_{\min-\mathrm{nz}}(B)\alpha \geq 0.$$

Since this holds for any eigenvector and eigenvalue pair of B we have shown $B(B - \lambda_{\min-\mathrm{nz}}(B)) \succeq 0$. So by Lemma 2.5 we find

$$\operatorname{tr}(ABB) \geq \operatorname{tr}(AB)\lambda_{\min-\mathrm{nz}}(B),$$

completing the proof. □

2.3 Lagrange dual function

To gain full understanding of the paper, the reader must also be partially understanding of the Lagrangian dual function. Suppose we have a nonlinear optimization problem of the form

$$\begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } g_i(x) \leq 0, \quad i = 1, \ldots, m \\ &\qquad\qquad h_j(x) = 0, \quad j = 1, \ldots, p \\ &\qquad\qquad x \in \mathbb{R}^n \end{aligned} \tag{3}$$

Then the Lagrangian function is defined as

Definition 2.4. *The Langragian of 3 is the function* $L : \mathbb{R}^n \times \mathbb{R}^m$ *given by*

$$L(x, \lambda, \mu) = f(x) + \sum_{i=1}^{m} \lambda_i g_i(x) + \sum_{j=1}^{p} \mu h_j(x)$$

The Langrangian function is used to set up the langrange dual problem which is defined as

Definition 2.5. *The Langrage dual problem of 3 is the optimization problem*

$$\begin{aligned} &\textit{maximize} \quad \inf_{x \in \mathbb{R}^n} L(x, \lambda, \mu) \\ &\textit{subject to } \lambda \in \mathbb{R}^m_{\geq 0}, \mu \in \mathbb{R}^p \end{aligned} \tag{4}$$

For minimization problems, feasible solutions of the Langrange dual function (4) always yield a lower bound for feasible solutions of the initial minimization problem (3). The gap between the optimal objective value of the original problem and the dual problem is known as the duality gap. When this duality gap is 0, we have that strong duality holds. Strong duality means that both the optimization problem (3) and its respective dual problem (4) have the same optimal objective value. When strong duality holds, the KKT conditions provide criteria to which any optimal solution must hold. For the general optimization problem (3), the KKT conditions look as follows for a feasible point x^*:

1. Feasibility: All constraints are satisfied

$$\begin{aligned} g(x^*) \leq 0, \\ h(x^*) = 0. \end{aligned}$$

2. Dual feasibility: The penalization is towards feasibility

$$\mu \geq 0.$$

3. Complementary slackness: Either μ_i=0, or $g_i(x^*) = 0$

$$\mu \cdot g = 0.$$

4. Stationarity: The cost function lies tangent to each active constraint

$$\nabla f(x^*) - \sum_i \mu_i \nabla g_i(x^*) - \sum_i \lambda_i \nabla h_i(x^*) = 0.$$

These KKT conditions are often vital for finding optimal solutions when strong duality holds. For example, the system of equations (2) in the introduction can also be solved by noting that for this optimization problem strong duality holds. Then combining the feasibility and stationarity yields the same closed form solution.

3 The Gauss-Seidel method

Before starting with the main convergence proof of the mixing method let us first shortly discuss the Gauss-Seidel method (see e.g., [7, page 512]) as it possesses some similar qualities to the mixing method and will be useful to prove some of the theorems in the main proof. This chapter is a restructuring of Wang, Chang and Kolter's lemma 3.10 [13]. The Gauss-Seidel method is a basic iterative solution method aimed to solve linear systems of the form

$$Ax = b,$$

where we assume A to be invertible. The method is named after Carl Friedrich Gauss who invented the method in 1820. Later the iterative algorithm was rediscovered by Philipp Ludwig von Seidel in 1874 and is therefore known as the 'Gauss-Seidel' method. For a system of the form $Ax = b$, write $A = L + D + U$, where L is the strictly lower triangular component of A, D is the diagonal, and U is the strictly upper triangular component of A. Then the Gauss-Seidel method is an iterative method that performs the operation

$$x_{k+1} = (L + D)^{-1}(b - Ux_k).$$

Interest for basic iterative methods grew in the 1950s as large sparse systems of the form $Ax = b$ had to be solved for instance in weather prediction and nuclear diffusion calculations. For example very difficult partial differential equations can be approximated in an area by finitely discretizing the problem into a chosen mesh. That results in a system $Ax = b$ which can be solved with a basic iterative method.

For this paper, the Gauss-Seidel method is relevant as it allows us to prove a theorem that aids us in proving the convergence of the mixing method. First consider the lemma

Lemma 3.1. *Let $S = C + \mathrm{Diag}(y)$, where $C \in \mathbb{R}^{n\times n}$ is a symmetric matrix zero on the diagonal and $y \in \mathbb{R}^n_{>0}$ is entrywise strictly positive. Suppose there exists a vector x_k with $x_k^T S x_k < 0$. Let $z_0 = x_k$ and let $z_1, ..., z_n$ be the vectors obtained through n coordinate updates of the Gauss-Seidel method. Then there exists $\omega \in (0, \frac{1}{2n})$ and an index $j \leq n$ such that*

$$\frac{1}{y_j}|e_j^T S z_j| \geq \omega \|z_j\|$$

Proof. Assume for contradiction that

$$\frac{1}{y_i}|e_j^T S z_j| < \omega \|z_j\| \ \ \forall j \text{ and } \forall \omega \in \left(0, \frac{1}{2n}\right).$$

First we show by induction that for $j = 1, \ldots, n$

$$\|x_k - z_j\| < 2j\omega \|x_k\|$$

For $j = 1$,

$$\|x_k - z_1\| = \|z_0 - z_1\| = \frac{1}{y_1}|e_1^T S z_1| < \omega \|z_1\| < 2\omega \|x_k\|.$$

Now suppose the inequality holds for a $0 \leq j \leq n-1$. Then for $j+1$

$$\begin{aligned}
\|x_k - z_{j+1}\| &= \|x_k - z_j + z_j - z_{j+1}\| \\
&\geq \|x_k - z_j\| + \|z_j - z_{j+1}\| \\
&< 2j\omega + \frac{1}{y_j}|e_j^T S z_j| \\
&< 2j\omega + \omega(2(j-1)\omega + 1)\|x_k\| \\
&= 2j\omega \|x_k\| + \omega(2j\omega - 1)\|x_k\|.
\end{aligned}$$

Pick $\omega \in \left(0, \frac{1}{2n}\right)$, then

$$\begin{aligned}
2j\omega - 1 &< 2j\frac{1}{2n} - 1 \\
&= \frac{j}{n} - 1 < 0.
\end{aligned}$$

Hence, when $\omega \in \left(0, \frac{1}{2n}\right)$

$$\|x_k - z_{j+1}\| \leq 2j\omega \|x_k\|.$$

Then for each $j = 1, \ldots, n$

$$\begin{aligned}
\frac{1}{y_{max}}|e_j^T S x_k| &\leq \frac{1}{y_k}|e_j^T S x_k| \\
&= \frac{1}{y_k}|e_j^T S(x_k + z_j - z_j)| \\
&\leq \frac{1}{y_k}(|e_j^T S z_j| + |e_j^T S(x_k - z_j)|) \\
&< \omega \|z_j\| + \frac{1}{y_k}|e_j^T S(x_k - z_j)|
\end{aligned}$$

$$
\begin{aligned}
&\leq \omega \|z_j\| + \frac{1}{y_k} \|Se_j\| \|x_k - z_j\| \text{ by the Cauchy inequality} \\
&< \omega(1 + 2j\omega) \|x_k\| + \frac{1}{y_k} 2j\omega \|Se_j\| \|x_k\| \\
&\leq \omega \left(1 + 2n\omega + \frac{2n \|Se_j\|}{y_{\min}}\right) \|x_k\| .
\end{aligned}
$$

Squaring both sides and summing over all $j = 1, \ldots, n$ we obtain the inequality

$$
n\omega^2 \left(1 + 2n\omega + \frac{2n \|Se_j\|}{y_{\min}}\right)^2 \|x_k\|^2 > \frac{1}{y_{\max}^2} (\sum_{j=1}^{n} |e_j^T S x_k|^2) = \frac{1}{y_{\max}^2} \|Sx_k\|^2 .
$$

$$
\begin{aligned}
\sqrt{n}\omega \left(1 + 2n\omega + \frac{2n \|Se_j\|}{y_{\min}}\right) \|x_k\| &> \frac{1}{y_{\max}} \|Sx_k\| \\
&\geq \frac{\sigma_{\min-\mathrm{nz}}(S)}{y_{max}} \|x_k\| ,
\end{aligned}
$$

by lemma 2.4 as $x_k \in \text{range}(S)$. So we found the inequality

$$
\sqrt{n}\omega \left(1 + 2n\omega + \frac{2n \|Se_j\|}{y_{\min}}\right) \geq \frac{\sigma_{\min-\mathrm{nz}}(S)}{y_{max}}.
$$

Note that the right side stays constant whereas the left hand side converges to 0 as $\omega \to 0$. Hence we have arrived at a contradiction. □

Theorem 3.2. *Let $C \in \mathbb{R}^{n \times n}$ be a symmetric matrix, zero on the diagonal and let L be the lower triangular part of C. Let $y \in \mathbb{R}^n_{>0}$. Suppose*

$$
J_{GS} = -(L + \text{Diag}(y))^{-1} L^T.
$$

Then $\rho(J_{GS}) > 1$ if the matrix $S = C + \text{Diag}(y)$ is not positive semidefinite.

Proof. Consider the minimization problem

$$
\begin{aligned}
&\text{minimize } f(x) = x^T (C + \text{Diag}(y)) x, \\
&\text{subject to } x \in \mathbb{R}^n,
\end{aligned}
$$

The gradient of the cost function is.

$$
\nabla f(x) = 2(C + \text{Diag}(y)) x.
$$

As this is a convex optimisation problem, we can apply the KKT conditions to this minimization problem which yields the optimality criteria

$$
(C + \text{Diag}(y)) x^* = 0.
$$

Then, applying the the Gauss-Seidel method to this problem generates the iterative sequence

$$x_{k+1} = -(L + \text{Diag}(y))^{-1} L^T x_k.$$

Equivalently the Gaus-Seidel method provides a coordinate update of the form

$$x_{k+1,i} = \frac{-1}{y_i}(\sum_{j=1}^{i-1} c_{i,j} x_{k+1,j} + \sum_{j=i+1}^{n} c_{i,j} x_{k,j}).$$

We can also show that the difference between Gauss-Seidel steps is

$$\begin{aligned}
f(x_k) - f(x_{k+1}) =& x_k^T (C + \text{Diag}(y)) x_k - x_{k+1}^T (C + \text{Diag}(y)) x_{k+1} \\
=& \sum_{i=1}^{n} y_i (x_{k,i}^2 - x_{k+1,i}^2) + \sum_{i=1}^{n} \sum_{j=1}^{n} c_{i,j} x_{k,i} x_{k,j} \\
&+ \sum_{i=1}^{n} \sum_{j=1}^{n} c_{i,j} x_{k+1,i} x_{k+1,j} \\
&\vdots \\
=& \sum_{i=1}^{n} y_i (x_{k,i}^2 - x_{k+1,i}^2) \\
&+ 2 \sum_{i=1}^{n} (\sum_{j=1}^{i-1} c_{i,j} x_{k+1,j} + \sum_{j=i+1}^{n} c_{i,j} x_{k,j})(x_{k,i} - x_{k+1,i}) \\
=& \sum_{i=1}^{n} y_i (x_{k,i}^2 - 2 x_{k,i} x_{k+1,i} + x_{k+1,i}^2) \\
=& \sum_{i=1}^{n} y_i \left\| x_{k,i} - x_{k+1,i} \right\|_2^2,
\end{aligned}$$

where the proof runs parallel to the proof of Lemma 4.2 (Substitute V with x_k). Suppose now that $S \not\succeq 0$. This implies that there exists an eigenvector $q \in \mathbb{R}^n$ such that $q^T S q < 0$. Let $x_0 = q$ such that $f(q) < 0$. From the decreasing property of the Gauss-Seidel method showed above we can identify two cases of convergence: Either the sequence converges to a cost of $-\infty$ or it converges to a number below 0.

Case 1: First let's assume that the Gauss-Seidel method converges to a constant value in cost. Then there is a subsequence that converges. Denote the limit of this sequence by $\bar{x}$. This means that $\bar{x}$ is a fixed point however that also means that $\bar{x}^T S \bar{x} = 0$ which contradicts the monotonic decreasing property of Gauss-Seidel and the fact that $f(x_0) < 0$.

Case 2: So $x_k^T S x_k$ must converge to $-\infty$. Hence suppose $x_r^T S x_r$ converges to $-\infty$. Note that one iteration of the Gaus-Seidel method can be subdivided into smaller steps z_i^k where instead only 1 coordinate is updated at a time

$$x_k = z_1^k \to z_2^k \to \cdots \to z_n^k \to z_{n+1}^k = x_{k+1}$$

where in step i the i-th coordinate of $z_{i,i}^k$ is updated to

$$z_{i+1,i}^r = \frac{-1}{y_i} \sum_{j=1}^{n} c_{i,j} z_{i,j}.$$

Then the cost function can be rewritten to

$$\begin{aligned} f(x_k) - f(x_{k+1}) &= \sum_{i=1}^{n} y_i \left\| x_i^k - x_{k+1}^r \right\|_2^2 \\ &= \sum_{i=1}^{n} y_i \left\| z_i - z_{i+1} \right\|^2. \end{aligned}$$

The cost difference of a single step between z_i and z_j can be written explicitly as

$$\begin{aligned} f(z_i) - f(z_{i+1}) &= y_i \left\| z_i - z_{i+1} \right\|^2 \\ &= y_i \left(z_{i,i} + \frac{1}{y_i} \sum_{j=1}^{n} c_{i,j} z_{i,j} \right)^2 \\ &= \frac{1}{y_i} \left(y_i z_{i,i} + \sum_{j=1}^{n} c_{i,j} z_{i,j} \right)^2 \\ &= \frac{|e_i^T S z_i|^2}{y_i}. \end{aligned}$$

Where $e_i \in \mathbb{R}^n$ is the i-th standard basis vector. Now by lemma 3.1, as x_k is in the range of S because of our initial starting value $x_0 = q$, we have

$$\begin{aligned} f(x_k) - f(x_{k+1}) =& f(z_1) - f(z_{n+1}) \\ \geq& \frac{|e_j^T S z_j|^2}{y_j} \\ \geq& y_j \omega^2 \left\| z_j \right\|^2, \end{aligned}$$

for some $\omega \in (0, \frac{1}{2n})$. Then by lemma 2.2

$$f(x_{k+1}) \leq f(x_k) - y_j \omega^2 \left\| z_j \right\|^2$$

$$\leq f(x_k) + \frac{y_j \omega^2}{\rho(S)} z_j^T S z_j$$
$$\leq (1 + \frac{y_{\min} \omega^2}{\rho(S)}) f(x_k).$$

Applying these steps recursively yields

$$f(x_{k+1}) \leq \left(1 + \frac{y_{\min} \omega^2}{\rho(S)}\right)^{k+1} f(x_0)$$

Now by lemma 2.5, since $S - \lambda_{\min} I_n \succeq 0$ we can upper bound $f(x_{k+1})$ as

$$\begin{aligned} f(x_{k+1}) &= x_k^T S x_k \\ &= \operatorname{tr}(x_k x_k^T S) \\ &\geq \lambda_{\min}(S) \operatorname{tr}(x_k x_k^T) \text{ by lemma 2.5} \\ &= \lambda_{\min}(S) \left\|x_k\right\|_2^2 \\ &= \lambda_{\min}(S) \left\|(J_{GS})^{k+1} x_0\right\|_2^2 \\ &\geq \lambda_{\min}(S) \left\|(J_{GS})^{k+1}\right\|_F^2 \left\|x_0\right\|_2^2 \text{ by lemma 2.3.} \end{aligned}$$

Combining the previously lower and upper bound we obtain

$$\lambda_{\min}(S) \left\|(J_{GS})^k\right\|_F^2 \left\|x_0\right\|_2^2 \leq \left(1 + \frac{y_{\min} \omega^2}{\rho(S)}\right)^k f(x_0).$$

Taking the equation to the power $\frac{2}{k}$ yields

$$\lambda_{\min}(S)^{\frac{2}{k}} \left\|(J_{GS})^k\right\|_F^{\frac{1}{k}} \left\|x_0\right\|_2^{\frac{1}{k}} \leq \sqrt{1 + \frac{y_{\min} \omega^2}{\rho(S)}} f(x_0)^{\frac{2}{k}}.$$

As k goes to infinity, we obtain

$$\lim_{k \to \infty} \left\|(J_{GS})^k\right\|^{\frac{1}{k}} \geq \sqrt{1 + \frac{y_{\min} \omega^2}{\rho(S)}}.$$

However the left hand side is exactly the spectral norm of J_{GS} by Gelfand's formula [2] (Which holds for any Banach algebra) so we obtain

$$\rho(J_{GS}) = \lim_{k \to \infty} \left\|(J_{GS})^k\right\|^{\frac{1}{k}} \geq \sqrt{1 + \frac{y_{\min} \omega^2}{\rho(S)}}.$$

Hence since $\rho(S), \omega, y_{\min} > 0$, we can conclude that $\rho(J_{GS}) > 1$. □

4 Convergence of the mixing method

Now that the preliminaries have been covered, the convergence proof of the mixing method can be given. The proof closely follows the proof of Wang, Chang and Kolter's paper [13]. Proofs where a significant details have been added, have been made without reference. For proofs without or only slight changes we reference the original lemma or theorem from Wang, Chang and Kolter's paper. Before we can analyze the convergence of the mixing method, we must first write the mixing method in matrix notation. First consider the following notation that will be used throughout the paper. As shown in section 1.1, the mixing method updates every column v_i of V as

$$v_i = -\frac{Vc_i}{\|Vc_i\|},$$

where c_i is the i-th column of C. Denote Vc_i by g_i and define $y \in \mathbb{R}^n$ by $y_i = \|g_i\|$. Let L be the matrix containing the strictly lower triangular part of C. Then one full iteration of the mixing method such that each column of V is updated once, can be written as

Lemma 4.1. *One iteration of the mixing method $M : \mathbb{R}^{k\times n} \to \mathbb{R}^{k\times n}$ can be written as*

$$M(V) = -VL(L^T + \mathrm{Diag}(y))^{-1} \tag{5}$$

Proof. To work out this proof we begin by writing the claim as

$$\begin{aligned} M(V) = -VL(L^T + \mathrm{Diag}(y))^{-1} = \hat{V} \\ VL = -\hat{V}(L^T + \mathrm{Diag}(y)) \\ VL + \hat{V}L^T = -\hat{V}\mathrm{Diag}(y). \end{aligned}$$

For the last equation, every row can be written out as

$$\sum_{j<i} c_{ij}\hat{v_j}^T + \sum_{j>i} c_{ij}v_j^T = g_i^T.$$

So indeed equality (1) holds as for one iteration of the mixing method we have exactly that

$$g_i = \sum_{j<i} c_{ij}\hat{v_j} + \sum_{j>i} c_{ij}v_j,$$

completing the proof. □

The strategy of the convergence proof will be to first prove that iterations of the mixing method $M(V)$ converge to a critical (fixed) point of the optimization problem. Then, we prove that this critical point is optimal by showing that non-optimal critical points are unstable. Finally the convergence speed is considered by proving linear convergence with a given convergence rate.

4.1 Convergence to a first order critical point

First we prove that the mixing method is monotonically nonincreasing. We can show the difference in the objective function between iterations of the mixing method is equal to

Lemma 4.2. *Let $\hat{V} = M(V)$ and $y_i = \|g_i\|_2$, where $g_i = Vc_i$. Then,*

$$f(V) - f(\hat{V}) = \sum_{i=1}^{n} y_i \|v_i - \hat{v}_i\|_2^2 .$$

Proof. The difference in the objective function between iterations of the mixing method is

$$\begin{aligned}
f(V) - f(\hat{V}) =& \operatorname{tr}(V^T V C) - \operatorname{tr}(\hat{V}^T \hat{V} C) \\
=& \sum_{i=1}^{n}\sum_{j=1}^{n} c_{i,j} v_i^T v_j - \sum_{i=1}^{n}\sum_{j=1}^{n} c_{i,j} \hat{v}_i^T \hat{v}_j \\
=& \sum_{i=1}^{n} v_i^T (\sum_{j=1}^{n} c_{i,j} v_j + \sum_{j<i} c_{i,j} \hat{v}_j) - \sum_{i=1}^{n} \hat{v}_i^T (\sum_{j=1}^{n} c_{i,j} \hat{v}_j + \sum_{j>i} c_{i,j} v_j) \\
&+ \sum_{i=1}^{n} \hat{v}_i^T \sum_{j>i} c_{i,j} v_j - \sum_{i=1}^{n} v_i^T \sum_{j<i} c_{i,j} \hat{v}_j \\
=& \sum_{i=1}^{n} v_i^T g_i - \sum_{i=1}^{n} \hat{v}_i^T g_i + (\sum_{i=1}^{n} \hat{v}_i^T \sum_{j>i} c_{i,j} v_j + \sum_{i=1}^{n} v_i^T \sum_{j<i} c_{i,j} v_j) \\
&- (\sum_{i=1}^{n} v_i^T \sum_{j<i} c_{i,j} \hat{v}_j + \sum_{i=1}^{n} \hat{v}_i^T \sum_{j>i} c_{i,j} \hat{v}_j),
\end{aligned}$$

since $c_{i,i} = 0$ for $i = 1, \ldots, n$. Now note that

$$\begin{aligned}
\sum_{i=1}^{n} v_i^T \sum_{j<i} c_{i,j} \hat{v}_j &= \sum_{i=1}^{n}\sum_{j<i} c_{i,j} \hat{v}_j^T v_i \\
&= \sum_{j=1}^{n} \hat{v}_j^T \sum_{j<i} c_{i,j} v_i \\
&= \sum_{i=1}^{n} \hat{v}_i^T \sum_{j>i} c_{i,j} v_j,
\end{aligned}$$

where the last equality holds provided that C is symmetric. So,

$$\begin{aligned}
f(V) - f(\hat{V}) =& \sum_{i=1}^{n} (v_i - \hat{v}_i)^T g_i + \sum_{i=1}^{n} v_i^T (\sum_{j<i} c_{i,j} \hat{v}_j + \sum_{j>i} c_{i,j} v_j) \\
&- \sum_{i=1}^{n} \hat{v}_i (\sum_{j<i} c_{i,j} \hat{v}_j + \sum_{j>i} c_{i,j} v_j)
\end{aligned}$$

$$
\begin{aligned}
&= 2\sum_{i=1}^{n}(v_i - \hat{v}_i)^T g_i \\
&= -2\sum_{i=1}^{n} y_i (v_i - \hat{v}_i)^T \hat{v}_i \\
&= \sum_{i=1}^{n} y_i (2 - 2\hat{v}_i^T v_i) \\
&= \sum_{i=1}^{n} y_i \left\| v_i - \hat{v}_i \right\|_2^2 .
\end{aligned}
$$

□

To ensure that the objective value does not just monotonically decrease but strictly decreases until reaching a fixed point we need to assert that each $y_i = \|Vc_i\|_2$ never becomes zero. Following this lemma the next theorem is proven.

Theorem 4.3. *Suppose that there exists $\delta > 0$ such that $\|M^k(V)c_i\|_2 > \delta$ for all $k \geq 0$ and $i = 1, \ldots, n$. Then the mixing method on the given SDP problem is strictly decreasing and always converges to a first order critical point.*

Proof. Under the fact that $f(V) - f(\hat{V}) = \sum_{i=1}^{n} y_i \|v_i - \hat{v}_i\|_2^2 \geq 0$ from lemma 4.2 we have that the sequence $f(M^k(V))$ is a cauchy sequence. By the definition of a cauchy sequence, note that for every $\epsilon > 0$, for sufficiently large $n, m \in \mathbb{N}_{>0}$ we have

$$
|f(M^n(V)) - f(M^m(V))| = \sum y_i \|v_i^{(n)} - v_i^{(m)}\|_2^2 < \epsilon.
$$

As $y_i > \delta > 0$ by our assumption, we have that for each column of V, v_i, that

$$
\|v_i^{(n)} - v_i^{(m)}\|_2^2 < \frac{\epsilon}{\delta}.
$$

So each sequence $v_i^{(k)}$ is also a cauchy sequence. Hence, the sequence $M^k(V)$ converges to unique limit (critical) point $\bar{V}$ such that $M(\bar{V}) = \bar{V}$. This equation can be rewritten to

$$
\begin{aligned}
-(L + \mathrm{Diag}(\bar{y}))^{-1} L^T \bar{V}^T &= \bar{V}^T \\
-L^T \bar{V}^T &= -(L + \mathrm{Diag}(\bar{y}))\bar{V}^T.
\end{aligned}
$$

Transposing the equation and subtracting the right hand side from both sides yields

$$
\begin{aligned}
-\bar{V}L &= \bar{V}(L^T + \mathrm{Diag}(\bar{y})) \\
\bar{V}(C + \mathrm{Diag}(\bar{y})) &= 0.
\end{aligned}
$$

Now let's compute the projected gradient of our cost function $f(V) = \langle C, V^T V \rangle$. First note that the euclidean gradient of $f(V)$ is

$$
\nabla f(V) = 2VC,
$$

where this identity is obtained from the the matrix cookbook [11]. Following the proof of Boumel et al. [3], the Riemannian gradient of f on the manifold is related to this gradient as

$$\text{grad}f(V) = \text{Proj}_{T_V M} \nabla f(V),$$

which is the projection onto $T_V M$, where M is the manifold $S^{k-1} \times \cdots \times S^{k-1}$ n-times. This manifold M can be represented by

$$M = \{V \in \mathbb{R}^{k \times n} : Ve_i \text{ is a unit vector for } i = 1, \ldots, n\}.$$

The tangent space then becomes

$$T_V M = \{X : (X^T V)_{i.i} = 0 \text{ for } i = 1, \ldots, n\}.$$

This means that there exists $Z \in (T_V M)^{\perp}$ such that

$$\text{grad}f(V) = \nabla f(V) + Z.$$

The space $(T_V M)^{\perp}$ is simply the orthogonal space of TvM and can be represented by:

$$(TvM)^{\perp} = \{V\text{Diag}(\mu) : \mu \in \mathbb{R}^n\}.$$

So if we let $\mu \in \mathbb{R}^n$ then the projected gradient is of the form

$$\begin{aligned} \text{grad}f(V) &= \nabla f(V) + 2V\text{Diag}(\mu) \\ &= 2VC + 2V\text{Diag}(\mu). \end{aligned}$$

Multiplying by V^T on the left provides us with the expression

$$V^T \text{grad}f(V) = 2V^T VC + 2V^T V\text{Diag}(\mu).$$

Now μ can be obtained by relating this expression to the constraints of our semidefinite program ($\langle E_{i,i}, V^T V \rangle = 1$) and equating them to zero

$$\begin{aligned} \langle E_{i,i}, 2V^T VC + 2V^T V\text{Diag}(\mu) \rangle &= 0 \\ \langle E_{i,i}, 2V^T V\text{Diag}(\mu) \rangle &= -\langle E_{i,i}, 2V^T VC \rangle. \end{aligned}$$

For each constraint $\langle E_{i,i}, V^T V \rangle$ $i = 1, .., n$, we find that this equation equates to

$$\mu_i v_i^T v_i = \sum_{j \neq i} c_{i,j} v_j^T v_i,$$

which for the converged mixing method exactly equates to

$$\begin{aligned}\mu_i \bar{v}_i^T \bar{v}_i &= g_i^T \bar{v}_i \\ \mu_i \bar{v}_i^T \bar{v}_i &= -\|g_i\|_2 \bar{v}_i^T \bar{v}_i \\ \mu_i &= -\|g_i\|_2 .\end{aligned}$$

Hence the projected gradient of the cost function for the converged mixing method is

$$\text{grad} f(\bar{V}) = 2\bar{V}C + 2\bar{V}\text{Diag}(\bar{y}).$$

From the equation $\bar{V}(C + \text{Diag}(\bar{y})) = 0$ it becomes clear that the projected gradient of our cost function $\langle C, V^T V\rangle$ is zero for the converged mixing method which implies that the mixing method converges to a critical point. □

4.2 Instability of non-optimal first order critical points

So the mixing method indeed converges to a critical point. However, this critical point need not be optimal. Let S denote the matrix $C+\text{Diag}(y)$. For the mixing method, we can identify two cases. Either $S \succeq 0$ or $S \not\succeq 0$. When S is positive semidefinite, then the mixing method converges to an optimal point. Consider the following theorem.

Lemma 4.4. *For a critical solution V, let $S = C + \text{diag}(y)$, where $y_i = \|Vc_i\|$. Then*

$$S \succeq 0 \Rightarrow V \text{ is optimal}$$

Proof. Following Wang, Chang and Kolter [13, lemma 3.12], the strategy for this proof is to first write the dual of the original SDP 1

$$\begin{aligned}&\text{maximize } -1_n^T y \\ &\text{subject to } C + \text{Diag}(y) \succeq 0 \\ &y \in \mathbb{R}^n\end{aligned}$$

Then one can show that for a fixed point V, where $S \succeq 0$, that the duality gap between the optimization problem and the dual is 0. Theorem 4.3 showed that for fixed points V of the mixing method we have

$$V(C + \text{Diag}(V)) = 0$$

Multiplying by V^T, then bringing $V^T V\text{Diag}(V)$ to the other side and taking the trace on each side we obtain

$$\text{tr}(V^T VC) = -1_n^T y.$$

So, taking y as the solution for the dual problem leads to a feasible solution as $C + \text{Diag}(y) \succeq 0$. Hence the solution is indeed optimal, thereby completing the proof. □

Now we only have to show that the mixing method does not converge to a fixed V with $S \not\succeq 0$. To show the mixing method converges to an optimal feasible point first note that the Jacobian of the mixing method can explicitly be written as the following.

Lemma 4.5. *The Jacobian of the mixing method is:*

$$J(V) = -((L + \mathrm{Diag}(y))^{-1} \otimes I_k)P(L^T \otimes I_k),$$

where P is the matrix

$$P = \mathrm{diag}(P_1, ..., P_n) \in \mathbb{R}^{nk \times nk} \quad \textit{where } P_i = I_k - \hat{v}_i \hat{v}_i^T \in \mathbb{R}^{k \times k}.$$

Proof. The strategy of this proof will be to utilize the implicit function theorem to obtain the Jacobian from the partial derivatives of the updated columns of V, since the derivatives of the columns of V are quite easy to compute. The implicit function theorem provides a set of conditions, under which a system of equations can be solved for certain dependent variables. In our case the dependent variables are the updated variables $\hat{v}_i$ whereas the vectors v_i are independent. The theorem states that if the $\hat{v}_i$ can be represented by v_i in some vicinity $(v_0, \hat{v}_0)$, where $(v_0, \hat{v}_0)$ satisfies our system of equations, $M(V)$, then if the Jacobian with respect to $\hat{v}$ $J_{M(V),\hat{v}} = [\frac{\partial M(V)}{\partial \hat{v}_{i,i}}]$ is invertible we can calculate the Jacobian of $M(V)$ as

$$J_{M(V)} = J^{-1}_{M(V),\hat{v}} J_{M(V),v}$$

The following is a slightly adjusted proof from Wang, Chang and Kolter's paper [13, Lemma 3.7]. First consider each column of $M(V) = \hat{v_i}$ separately. The update of the mixing method for a $\hat{v}_i$ is

$$\|g_i\|_2 \hat{v}_i = -g_i.$$

Applying implicit differentiation and the product rule we obtain

$$\begin{aligned}
\|g_i\|_2 \partial \hat{v}_i + \frac{g_i}{\|g_i\|_2} \hat{v}_i^T \partial g_i &= -\partial g_i \\
\|g_i\|_2 \partial \hat{v}_i &= -P_i \partial g_i (= -P_i(\sum_{j<i} c_{i,j} \partial \hat{v_j} + \sum_{j>i} c_{i,j} \partial v_j)) \\
\|g_i\|_2 \partial \hat{v}_i + P_i \sum_{j<i} c_{i,j} \partial \hat{v_j} &= -P_i \sum_{j>i} c_{i,j} \partial v_j.
\end{aligned}$$

To apply the implicit function theorem we first write the derivatives of $\partial \hat{v}_i$ and ∂v_i as one vector $\partial \hat{v}$ and ∂v respectively. Stacking the vectors obtained by the

previous equation yields

$$\begin{pmatrix} y_1 I_k & 0 & & 0 \\ c_{1,2}P_2 & y_2 I_k & \ddots & \\ \vdots & \vdots & \ddots & 0 \\ c_{1,n}P_n & c_{2,n}P_n & \cdots & y_n I_k \end{pmatrix} \partial \hat{v} = \begin{pmatrix} 0 & c_{1,2}P_1 & \cdots & c_{1,n}P_1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & c_{n-1,n}P_{n-1} \\ 0 & 0 & \cdots & 0 \end{pmatrix} \partial v$$

The left hand side is always invertible so by the implicit function theorem the Jacobian of the mixing method is

$$J(V) = - \begin{pmatrix} y_1 I_k & 0 & & 0 \\ c_{1,2}P_2 & y_2 I_k & \ddots & \\ \vdots & \vdots & \ddots & 0 \\ c_{1,n}P_n & c_{2,n}P_n & \cdots & y_n I_k \end{pmatrix}^{-1} \begin{pmatrix} 0 & c_{1,2}P_1 & \cdots & c_{1,n}P_1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & c_{n-1,n}P_{n-1} \\ 0 & 0 & \cdots & 0 \end{pmatrix}.$$

This equation for $J(V)$ is equivalent to

$$J(V) = -(P(L \otimes I_k) + \mathrm{Diag}(y) \otimes I_k)^{-1} P(L^T \otimes I_k),$$

where P is the zero matrix with the projection matrices P_i on the diagonal. Therefore P is also a projection matrix, meaning $P = P^T$ and that the Moore-Penrose inverse $P^\dagger = P$. So the Jacobian $J(V)$ can be written as

$$\begin{aligned} (P(L \otimes I_k) + \mathrm{Diag}(y) \otimes I_k)^{-1} P &= (P(L \otimes I_k) + \mathrm{Diag}(y) \otimes I_k)^{-1} P^\dagger \\ &= (P^2(L \otimes I_k) + P(\mathrm{Diag}(y) \otimes I_k))^\dagger \\ &= (P(L \otimes I_k) + P(\mathrm{Diag}(y) \otimes I_k))^\dagger \\ &= (P((L + \mathrm{Diag}(y)) \otimes I_k))^\dagger \\ &= ((L + \mathrm{Diag}(y)) \otimes I_k)^{-1} P \\ &= ((L + \mathrm{Diag}(y))^{-1} \otimes I_k) P. \end{aligned}$$

Hence,

$$J(V) = -((L + \mathrm{Diag}(y))^{-1} \otimes I_k) P (L^T \otimes I_k). \qquad \square$$

Although we have now obtained a formulation of the Jacobian, from this it is still challenging to evaluate the eigenvalues. Fortunately we can analyze a smaller Jacobian instead. Consider the following lemma.

Lemma 4.6. *Assume $V \in \mathbb{R}^{k\times n}$ has rank$(V) < k$. Let $P = \text{diag}(P_1, ..., P_n)$, where $P_i = I_k - v_i v_i^T$. Then for any $A, B \in \mathbb{R}^{n\times n}$, any eigenvalue of AB is also an eigenvalue of*

$$J = (A \otimes I_k)P(B \otimes I_k).$$

Proof. The following is a slightly adjusted proof from Wang, Chang and Kolter's paper [13, Lemma 3.9]. Since rank$(V) < k$ there is an orthogonal vector r such that $r^T v_i = 0$ for all columns of V v_i. Now suppose $\lambda \in \mathbb{C}$ is an eigenvalue of AB with eigenvector $q \in \mathbb{C}^n$. Then note that,

$$\begin{aligned}
J\text{vec}(Z) &= (A \otimes I_k)P(B \otimes I_k)\text{vec}(rq^T) \\
&= (A \otimes I_k)P\text{vec}(rq^T B^T)^* \\
&= (A \otimes I_k)P\text{vec}(r(Bq)^T) \\
&= (A \otimes I_k)\text{vec}(P_1 r(Bq)_1^T, \ldots, P_n r(Bq)_n^T) \\
&= (A \otimes I_k)\text{vec}(r(Bq)_1^T, \ldots, r(Bq)_n^T) \quad (P_i r = r) \\
&= (A \otimes I_k)\text{vec}r(Bq)^T) \\
&= \text{vec}(r(Bq)^T A^T)^* \\
&= \text{vec}(r(ABq)^T \\
&= \text{vec}(r(\lambda q)^T) \\
&= \lambda\text{vec}(Z),
\end{aligned}$$

where at * we use the identity $(B^T \otimes A)\text{vec}(X) = \text{vec}(AXB)$ (See the matrix cookbook [11] for this identity). Hence we showed that $\text{vec}(Z) = \text{vec}(rq^T)$ is the corresponding eigenvector for J with eigenvalue λ, completing the proof. □

Hence if we take the matrices from the lemma to be $A = -(L + \text{Diag}(y))^{-1}$ and $B = L^T$, then analyzing the eigenvalues of $-(L + \text{Diag}(y))^{-1}L^T$ is sufficient for proving divergence. Thus we need to show that for the matrix $-(L + \text{Diag}(y))^{-1}L^T$, there is an eigenvalue $\lambda > 1$. This brings us to the next lemma that covers the criteria for which the mixing method diverges.

Lemma 4.7. *Let L be the lower triangular part of C and let $y \in \mathbb{R}^n_{>0}$. Let*

$$J_{GS} = -(L + \text{Diag}(y))^{-1}L^T.$$

Then $\rho(J_{GS}) > 1$ if the matrix $S = C + \text{Diag}(y))$ is not positive semidefinite.

Proof. This lemma follows directly from theorem 3.2. □

Note that for lemma 4.6, a necessary condition for J_{GS} to contain the same eigenvalues as $J(V)$ is that V cannot be of full column rank. This leads to the following necessary lemma.

Lemma 4.8. *Let $\frac{k(k+1)}{4} > n$. Then, for almost all $C \in \mathbb{R}^{n\times n}/\{0$ on the diagonal$\}$, all first order critical points $V \in \mathbb{R}^{k\times n}$ have rank smaller than k*

Proof. Proven in Wang, Chang and Kolter's paper [13, Appendix E]. There is a slight difference in the proof as we assume here that the values on the diagonal of C are zero. The strategy for this proof involves bounding the rank of V by noting that for critical points

$$V(C + \text{Diag}(y)) = 0.$$

Therefore, the rank of V is upper bounded by

$$\text{rank}(V) \leq \text{null}(C + \text{Diag}(y)).$$

Now suppose that V has full row rank $rank(V) = k$. We can show that C is part of a set $\mathscr{S} \subseteq S^n$ with the dimension of $\mathscr{S}$ bounded by

$$\text{Dim}(\mathscr{S}) \leq \frac{n(n+3)}{2} - \frac{k(k+1)}{2}.$$

If we take k such that $\frac{k(k+1)}{4} > n$, then

$$\text{Dim}(\mathscr{S}) < \frac{n(n-1)}{2}.$$

Since almost no C fulfills this, as $C \in \{X \in S^n | X_{i,i} = 0 \ \forall i\}$ where $\text{Dim}(\{X \in S^n | X_{i,i} = 0 \, \forall i\} = \frac{n(n-1)}{2}$ so for almost no C, $rank(V) = k$. Hence we have that for almost all C, $rank(V) < k$, completing the proof. □

This is the final detail needed to show that fixed points V with $S \not\succeq 0$ are unstable. With 'unstable' we mean a point of which the Jacobian of the mixing method at that point has eigenvalues higher than 1. This is indicative of a point that the mixing method diverges from in at least 1 coordinate. So no matter how close the mixing method gets to this unstable point, it will always diverge away from it in the next iteration. Hence, the only way the mixing method ends on a unstable fixed point is when it immediately 'jumps' to it, which in practice is extremely unlikely. Combining the previous lemmas yields the critical point theorem

Theorem 4.9. *Let $\frac{k(k+1)}{4} > n$ and suppose that there exists $\delta > 0$ such that $\|M^k(V)c_i\|_2 > \delta$ for all $k \geq 0$ and $i = 1, \ldots, n$. Then, for almost all C, all non-optimal first order critical points are unstable fixed points for the Mixing method.*

Proof. From theorem 4.3 it is clear that the mixing method always converges to a fixed point V. Lemma 4.4 proves that for this fixed point V, if $S = C + \text{Diag}(y) \succeq 0$ then V must be optimal. Therefore we only need to consider fixed points V with $S \not\succeq 0$. Lemma 4.8 shows that when when we take k such that $\frac{k(k+1)}{4} > n$, then V is not of full column rank. Hence by lemma 4.6

the eigenvalues of the Jacobian of the mixing method $J(V)$ can be partially evaluated with the eigenvalues of

$$J_{GS} = -(L + \text{Diag}(y)))^{-1}L^T.$$

Lemma 4.7 proves that $\rho(J_{GS}) > 1$ when $S \not\succeq 0$. So $\rho(J(V))) > 1$ meaning that the mixing method will never converge to points where S is not semipositive definite, completing the proof. □

4.3 Proof of Linear convergence

Now that it is shown that the mixing method only allows for optimal critical points we can show that the method converges asymptotically linear to a critical point under certain conditions. First we need the following three lemmas. First it can be shown that the mixing method is Lipschitz continuous.

Lemma 4.10. *Suppose that there exists $\delta > 0$ such that $\|M^k(V)c_i\|_2 > \delta$ for all $k \geq 0$ and $i = 1, \ldots, n$. Then, the mixing method is $\frac{2\sqrt{\sum_{i=1}^n \|c_i\|^2}}{\delta}$-Lipschitz continuous.*

Proof. First note that for each updated column v_i that

$$\begin{aligned}
\left\| \frac{-Vc_i}{\|Vc_i\|_2} - \frac{-V^*c_i}{\|V^*c_i\|_2} \right\|_2 &= \left\| \frac{-Vc_i}{\|Vc_i\|_2} - \frac{-V^*c_i}{\|Vc_i\|_2} - \frac{V^*c_i}{\|Vc_i\|_2} + \frac{V^*c_i}{\|V^*c_i\|_2} \right\|_2 \\
&= \left\| \frac{-Vc_i}{\|Vc_i\|_2} - \frac{-V^*c_i}{\|Vc_i\|_2} - \left(\frac{\|V^*c_i\|_2}{\|V^*c_i\|_2} - \frac{\|V^*c_i\|_2}{\|Vc_i\|_2} \right) * \frac{-V^*c_i}{\|V^*c_i\|_2} \right\|_2 \\
&= \left\| \frac{-Vc_i}{\|Vc_i\|_2} - \frac{-V^*c_i}{\|Vc_i\|_2} - (\frac{\|Vc_i\|_2}{\|Vc_i\|_2} - \frac{\|V^*c_i\|_2}{\|Vc_i\|_2}) * \frac{-V^*c_i}{\|V^*c_i\|_2} \right\|_2 \\
&= \frac{1}{\|Vc_i\|_2} \left\| -Vc_i + V^*c_i - (\|Vc_i\|_2 - \|V^*c_i\|_2) \frac{-V^*c_i}{\|V^*c_i\|_2} \right\|_2 \\
&\leq \frac{1}{\|Vc_i\|_2} (\|-Vc_i + V^*c_i\|_2 + \|\|Vc_i\|_2 - \|V^*c_i\|_2\|_2 \left\| \frac{-V^*c_i}{\|V^*c_i\|_2} \right\|_2) \\
&\leq \frac{1}{\delta} (\|V - V^*\|_F \|c_i\|_2 + \|Vc_i - V^*c_i\|_2) \\
&\leq \frac{2\|c_i\|_2}{\delta} \|V - V^*\|_F .
\end{aligned}$$

Now note that

$$M(V) = (Vc_1 \ldots Vc_n).$$

From this it becomes clear that

$$\begin{aligned}
\|M(V) - M(V^*)\|_F^2 &= \sum_{j=1}^{k} \sum_{i=1}^{n} ((\hat{v}_i - \hat{v}_i^*)_j)^2 \\
&= \sum_{j=1}^{k} \sum_{i=1}^{n} ((\frac{-Vc_i}{\|Vc_i\|_2} - \frac{-V^*c_i}{\|V^*c_i\|_2})_j)^2
\end{aligned}$$

$$= \sum_{i=1}^{n} \left\| \frac{-Vc_i}{\|Vc_i\|_2} - \frac{-V^*c_i}{\|V^*c_i\|_2} \right\|_2^2$$
$$\leq \frac{4\sum_{i=1}^{n} \|c_i\|^2}{\delta^2} \|V - V^*\|_2^2 .$$

Hence,

$$\|M(V) - M(V^*)\|_F \leq \frac{2\sqrt{\sum_{i=1}^{n} \|c_i\|^2}}{\delta} \|V - V^*\|_F . \qquad \square$$

Now note that for any iteration of the mixing method, the difference between two iterations can actually be bounded by the difference in objective value between an optimal solution V^*.

Lemma 4.11. *Suppose V^* is an optimal solution and let $S^* = C + \mathrm{Diag}(y^*)$. For an iteration of the mixing method we obtain the upper bound*

$$\|V - M(V)\|_F^2 \geq \left(\frac{\lambda_{\min-\mathrm{nz}}(S^*)}{y_{\max}^2} - \frac{2\|y - y^*\|_2}{y_{\min}^2}\right)(f(V) - f(V^*)).$$

Proof. Let $R = (L + \mathrm{Diag}(y))^{-1}$ and $S = C + \mathrm{Diag}(y)$. Then

$$\begin{aligned}
\|V - M(V)\|_F^2 &= \left\|V(L^T + \mathrm{Diag}(y))(L^T + \mathrm{Diag}(y))^{-1} + VL(L^T + \mathrm{Diag}(y))^{-1}\right\|_F^2 \\
&= \left\|VS(L^T + \mathrm{Diag}(y))^{-1}\right\|_F^2 \\
&= \|V(C + \mathrm{Diag}(y) - \mathrm{Diag}(y^*) + \mathrm{Diag}(y^*))R\|_F^2 \\
&= \|VS^*R\|_F^2 + 2\mathrm{tr}(VS^*RR^T(\mathrm{Diag}(y) - \mathrm{Diag}(y^*))) \\
&\quad + \|V(\mathrm{Diag}(y) - \mathrm{Diag}(y^*))R\|_F^2 \\
&\geq \|VS^*R\|_F^2 + 2\mathrm{tr}(VS^*RR^T(\mathrm{Diag}(y) - \mathrm{Diag}(y^*))) \\
&= \mathrm{tr}(S^*V^TVS^*RR^T) + 2\mathrm{tr}(VS^*RR^T(\mathrm{Diag}(y) - \mathrm{Diag}(y^*))).
\end{aligned}$$

As S and $V^TV \succeq 0$, evidently $S^*V^TVS^* \succeq 0$. We also have that $RR^T - \lambda_{\min}(RR^T)I_n \succeq 0$. Then applying lemma 2.5 yields

$$\mathrm{tr}(S^*V^TVS^*RR^T) \geq \lambda_{\min}(RR^T)\mathrm{tr}(S^*V^TVS^*).$$

Similarly, noting that $\|y - y^*\| I_n + (\mathrm{Diag}(y) - \mathrm{Diag}(y^*)) \succeq 0$ and $VS^*RR^T \succeq 0$ we obtain

$$\mathrm{tr}(VS^*RR^T(\mathrm{Diag}(y) - \mathrm{Diag}(y^*))) \leq \|y - y^*\|_2 \,\mathrm{tr}(VS^*RR^T).$$

Hence,

$$\begin{aligned}
\|V - M(V)\|_F^2 &\geq \lambda_{\min}(RR^T)\mathrm{tr}(S^*V^TVS^*) + 2\mathrm{tr}(VS^*RR^T(\mathrm{Diag}(y) - \mathrm{Diag}(y^*))) \\
&\geq \lambda_{\min}(RR^T)\mathrm{tr}(S^*V^TVS^*) - 2\|y - y^*\|_2 \,\mathrm{tr}(V^TVS^*RR^T) \\
&\geq \sigma_{\min}^2(R)\mathrm{tr}(S^*V^TVS^*) - 2\sigma_{\max}^2(R)\|y - y^*\|_2 \,\mathrm{tr}(V^TVS^*)
\end{aligned}$$

$$
\begin{aligned}
&\geq \sigma^2_{\min}(R)\lambda_{\min-\mathrm{nz}}(S^*)\mathrm{tr}(V^TVS^*) - 2\sigma^2_{\max}(R)\left\|y-y^*\right\|_2 \mathrm{tr}(V^TVS^*)\\
&= (\frac{\lambda_{\min-\mathrm{nz}}(S^*)}{y^2_{\max}} - \frac{2\left\|y-y^*\right\|_2}{y^2_{\min}})\mathrm{tr}(V^TVS^*)\\
&= (\frac{\lambda_{\min-\mathrm{nz}}(S^*)}{y^2_{\max}} - \frac{2\left\|y-y^*\right\|_2}{y^2_{\min}})(f(V)-f(V^*)),
\end{aligned}
$$

where the 4th inequality follows from lemma 2.6. □

Lemma 4.12. *Suppose that there exists $\delta > 0$ such that $\|M^k(V)c_i\|_2 > \delta$ for all $k \geq 0$ and $i = 1, \ldots, n$. Then, there is a constant $\tau > 0$ such that for any optimal solution V^* we have*

$$\|y - y^*\|_2^2 \leq \tau(f(V) - f(V^*)).$$

Proof. Recall that one iteration of the mixing method can be subdivided into n steps where each column v_i gets updated individually to $\hat{v_i}$ in step i

$$V = Z_1 \to Z_2 \to \cdots \to Z_n \to Z_{n+1} = M(V).$$

Let $S^* = C + \mathrm{Diag}(y^*)$ for an optimum V^* and let s_i denote the i-th column of S^*. Then first observe that

$$
\begin{aligned}
(y_i - y_i^*)\hat{v_i} &= \|Vc_i\|_2 \hat{v_i} - y_i^*\hat{v_i}\\
&= \sqrt{c_i^TV^TVc_i}\hat{v_i} - y_i^*\hat{v_i}\\
&= \sum_{j<i} c_{i,j}\hat{v_j} + \sum_{j>i} c_{i,j}v_j - y_i^*\hat{v_i}\\
&= -Z_ic_i - y_i^*\hat{v_i}\\
&= -Z_i(s_i^* + (0\ldots y_i \ldots 0)^T)\\
&= -Z_is_i^* + y_i^*(v_i - v_i^*).
\end{aligned}
$$

To bound this we first note that

$$
\begin{aligned}
f(V) - f(V^*) &\geq f(Z_i) - f(V^*)\\
&= \mathrm{tr}(Z_i^TZ_iC^T) - \mathrm{tr}(C^TV^*TV^*)\\
&= \mathrm{tr}(Z_i^TZ_iS^*)\\
&\geq \frac{1}{\lambda_{max}(S^*)}\left\|Z_iS^*\right\|_2^2\\
&\geq \frac{1}{\lambda_{max}(S^*)}\left\|Z_is_i^*\right\|_2^2.
\end{aligned}
$$

Similarly we find,

$$
\begin{aligned}
f(V) - f(V^*) &\geq f(Z_i) - f(Z_{i+1})\\
&\geq \delta\left\|v_i - \hat{v_i}\right\|_2^2.
\end{aligned}
$$

Taking the norm squared on both sides for the equation $(y_i - y_i^*)\hat{v_i} = -Z_i s_i^* + y_i^*(v_i - v_i^*)$ yields

$$\begin{aligned}
\|(y_i - y_i^*)\hat{v_i}\|_2^2 &= \|-Z_i s_i^* + y_i^*(v_i - v_i^*)\|_2^2 \\
(y_i - y_i^*)^2 &\leq (\|-Z_i s_i^*\|_2 + \|y_i^*(v_i - v_i^*)\|_2)^2 \\
&= \|Z_i s_i^*\|_2^2 + \|Z_i s_i^*\|_2 \|y_i^*(v_i - v_i^*)\|_2 + \|y_i^*(v_i - v_i^*)\|_2^2 \\
&\leq 2\|Z_i s_i^*\|_2^2 + 2y_i^* \|v_i - v_i^*\|_2^2 \ \ (\|a-b\|_2^2 \geq 0) \\
&\leq (2\lambda_{\max}(S^*) + \frac{2y_i^*}{\delta})(f(V) - f(V^*)).
\end{aligned}$$

Hence when we sum over all columns

$$\|y - y^*\|_2^2 \leq (2n\lambda_{\max}(S^*) + \frac{2\sum_{i=1}^n y_i^*}{\delta})(f(V) - f(V^*)). \qquad \square$$

This lemma actually shows a very interesting geometrical property of the mixing method. Namely that for any two optimal solutions V_1^* and V_2^*, we have that the respective y_1^* and y_2^* are equal. Using the previous two lemmas, we can now prove linear convergence for the mixing method.

Theorem 4.13. *Suppose V is sufficiently close to an optimal solution and there exists $\delta > 0$ such that $\|M^k(V)c_i\|_2 > \delta$ for all $k \geq 0$ and $i = 1, \ldots, n$. Then the mixing method with as initial point V converges linearly with a convergence rate μ bounded by*

$$\mu \leq 1 - \delta\kappa,$$

where $0 < \kappa < \frac{\lambda_{\min-\text{nz}}(S^*)}{y_{max}} - \frac{2\|y-y^*\|_2}{y_{min}^2}$.

Proof. By Lemma 4.11, for any optimal solution V^* we have

$$\|V - M(V)\|_F^2 \geq \left(\frac{\lambda_{\min-\text{nz}}(S^*)}{y_{\max}} - \frac{2\|y-y^*\|_2}{y_{\min}^2}\right)(f(V) - f(V^*)).$$

By Lemma 4.12, since $\|y-y^*\|^2 \leq \tau(f(V) - f(V^*))$, we can take $f(V)$ and $f(V^*)$ close enough such that

$$\frac{\lambda_{\min-\text{nz}}(S^*)}{y_{max}} > \frac{2\|y-y^*\|_2}{y_{min}^2}.$$

So there exists $\kappa > 0$ with

$$\frac{\lambda_{\min-\text{nz}}(S^*)}{y_{max}} - \frac{2\|y-y^*\|_2}{y_{min}^2} > \kappa.$$

Hence,

$$f(V) - f(M(V)) = \sum_{i=1}^n y_i \|v_i - \hat{v_i}\|_2^2 \ \text{ By Lemma 4.2}$$

$$
\begin{aligned}
&\geq \min_{i=1,\dots n} y_i \left\|V - M(V)\right\|_F^2 \\
&\geq \delta \left\|V - M(V)\right\|_F^2 \\
&\geq \delta \left\|V - M(V)\right\|_F^2 . \\
&\geq \kappa\delta(f(V) - f(V^*))
\end{aligned}
$$

This implies

$$(1 - \delta\kappa)(f(V) - f(V^*)) \geq f(M(V)) - f(V^*). \qquad \square$$

5 Numerical convergence analysis of the mixing method

In section 4.3 we proved that the mixing method converges linearly with rate $1 - \delta\kappa$. We noted that

$$\delta \leq \min_{i=1,\dots n} y_i,$$

and

$$\kappa < \frac{\lambda_{\min-\mathrm{nz}}(S^*)}{y_{max}} - \frac{2\left\|y - y^*\right\|_2}{y_{min}^2}.$$

Note that lemma 4.12 guarantees that for any two optimal solutions V_1^*, V_2^* we have that $y_1^* = y_2^*$. Hence also $\lambda_{\min-\mathrm{nz}}(S_1^*) = \lambda_{\min-\mathrm{nz}}(S_2^*)$. This allows us to provide a theoretical convergence rate for the mixing method and compare that with the practically achieved convergence rate. Note that the linear convergence factor of the mixing method is bound by the smallest non-zero eigenvalue of S^*. Hence if the optimal solution space of our problem only admits solutions with small eigenvalues, convergence can be slow with the convergence rate being close to 1. Take for example the problem of maximizing the sum of squared pairwise distances.

5.1 Maximizing the sum of squared pairwise distances on the unit sphere

The problem of maximizing the sum of squared pairwise distances can be formulated as follows

$$
\begin{aligned}
\text{maximize } & \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \left\|x_i - x_j\right\|_2^2 \\
\text{subject to } & \left\|x_i\right\|_2 = 1 \ \forall i \in [n], \\
& x_i \in \mathbb{R}^k.
\end{aligned}
$$

The optimal solution of maximizing the sum of squared pairwise distances is trivially obtained by the following formula

Lemma 5.1. *Let $n \in \mathbb{Z}_{>0}$ denote the amount of points on the sphere and $k \in \mathbb{Z}_{>1}$ the dimension. Then one of the optimal solutions of maximizing the sum of squared pairwise distances is given by:*

When n is odd, $\frac{n-1}{2}$ points of the form

$$\left(\frac{1}{n-1}, \sqrt{1-\frac{1}{(n-1)^2}}, 0, \ldots, 0\right),$$

$\frac{n-1}{2}$ *points of the form*

$$\left(\frac{1}{n-1}, -\sqrt{1-\frac{1}{(n-1)^2}}, 0, \ldots, 0\right),$$

1 *point of the form*

$$(-1, 0, 0, \ldots, 0),$$

with objective value $f(x) = n^2$.

When n is even, an optimal solution is given by $\frac{n}{2}$ points of the form

$$(1, 0, 0, \ldots, 0),$$

$\frac{n}{2}$ *points of the form*

$$(-1, 0, 0, \ldots, 0),$$

with objective value $f(x) = n^2$.

Proof. Consider the problem statement of this problem as before

$$\begin{aligned} \text{maximize } & \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \|x_i - x_j\|_2^2 \\ \text{subject to } & \|x_i\|_2 = 1 \ \forall i \in [n], \\ & x_i \in \mathbb{R}^k. \end{aligned}$$

Note that

$$\sum_{i=1}^{n}\sum_{j=1}^{n} x_i^T x_j = \left\|\sum_{i=1}^{n} x_i\right\|_2^2 \geq 0,$$

hence,

$$\sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} x_i^T x_j \geq -n.$$

Trivially, $\|x_i - x_j\|^2 = 2 - 2\langle x_i, x_j\rangle$ for $x_i, x_j \in \mathbb{R}^n$ such that $\|x_i\| = \|x_j\| = 1$. So this lower bound is equivalent to the upper bound

$$\sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \|x_i - x_j\|^2 \leq 2n^2.$$

Now suppose n is even. Then, by setting half the points to $(1, 0, \ldots, 0)$ and the other half to $(-1, 0, \ldots, 0)$ we obtain

$$\sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \|x_i - x_j\|_2^2 = 2n^2,$$

which means that this solution is optimal. Now suppose n is odd. Then setting 1 point to $(-1, 0, 0, \ldots, 0)$, $\frac{n-1}{2}$ points to $(\frac{1}{n-1}, \sqrt{1 - \frac{1}{(n-1)^2}}, 0, \ldots, 0)$ and $\frac{n-1}{2}$ points to $(\frac{1}{n-1}, \sqrt{1 - \frac{1}{(n-1)^2}}, 0, \ldots, 0)$, we similarly obtain

$$\sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \|x_i - x_j\|_2^2 = 2n^2. \qquad \square$$

This problem can actually be written as a diagonally constrained SDP problem for which the mixing method is applicable. First note that we can write the norm $\|x_i - x_j\|^2$ as

$$\begin{aligned} \|x_i - x_j\|_2^2 &= \langle x_i - x_j, x_i - x_j \rangle \\ &= \langle x_i, x_i \rangle + \langle x_j, x_j \rangle - 2\langle x_i, x_j \rangle \\ &= 2 - 2\langle x_i, x_j \rangle. \end{aligned}$$

Hence solving the problem of maximizing the sum of squared pairwise distances is equivalent to solving

$$\begin{aligned} \text{minimize} \quad & \sum_{i,j \in [n]\ i \neq j} \langle x_i, x_j \rangle \\ \text{subject to} \quad & \|x_i\|_2 = 1 \ \ \forall i \in [N]. \end{aligned}$$

This can be rewritten to a rank constrained semi definite program, by taking the matrix $X \in S_+^n$ with $X_{i,j} = \langle x_i, x_j \rangle$. Then the problem becomes

$$\begin{aligned} \text{minimize} \quad & \sum_{i,j \in [n]\ i \neq j} X_{i,j} \\ \text{subject to} \quad & X_{i,i} = 1, \ \ \forall i \in [n] \\ & rank(X) \leq k. \end{aligned}$$

which can be written formerly as:

$$\begin{aligned} \text{minimize} \quad & \langle J_n - I_n, X \rangle \\ \text{subject to} \quad & X_{i,i} = 1, \ \ \forall i \in [n], \\ & rank(X) \leq k, \\ & X \succeq 0 \end{aligned}$$

where J_n is the $n \times n$ matrix consisting of only 1's

5.2 Applying the mixing method to maximizing the sum of squared pairwise distances on the unit sphere

In section 5.1 we showed that the linear convergence rate of the mixing method is dependent on the smallest eigenvalue of S^*. Recall that S^* is given by

$$S^* = C + \text{Diag}(y^*),$$

where $y_i^* = \left\| \sum_{j<i} c_{i,j} v_j + \sum_{j>i} c_{i,j} v_j \right\|$. From the previous construction of the solution for maximizing the sum of squared pairwise distances it is trivial to show that for the even and uneven case, the optimal solution provides a y^* with

$$y_i^* = 1 \text{ for } i = 1, \ldots, n.$$

Hence that means that the matrix S^* is

$$S^* = J_n,$$

with J_n being the all one matrix. This matrix S^* therefore has a smallest non-zero eigenvalue of $\lambda_{\min-\text{nz}}(S^*) = 1$. However to qualify for the mixing method we need a generic enough C for lemma 4.8 to hold. Hence we add a random perturbation $\epsilon_{i,j} \in Unif(-\frac{1}{2}10^{-m}, \frac{1}{2}10^{-m})$ to every index except for the diagonal. The smallest nonzero eigenvalue of S^* is then almost surely in the range of $[10^{-(m+1)}, 10^{-m}]$. What this means for the convergence of the mixing method is that the convergence rate μ is

$$\begin{aligned} \mu = 1 - \delta\kappa &\leq 1 - \frac{\sigma_{min}(S^*)}{y_{max}} + \frac{2\left\|y - y^*\right\|_2}{y_{min}^2} \\ &= 1 - \sigma_{min}(S^*) + 2\left\|y - y^*\right\|_2, \end{aligned}$$

as y_i was determined to converge to 1. We assume to already be near the optimal solution so $\|y - y^*\| \approx 0$. Hence we obtain

$$\begin{aligned} \mu &\leq 1 - \sigma_{min}(S^*) \\ &= 1 - c\epsilon, \end{aligned}$$

with $c \in [\frac{1}{10}, 1]$. Hence convergence can be as slow as how small we make the perturbation. We have no guarantee in this case on fast linear convergence as the convergence rate converges to 1 as we decrease ϵ. To test practical convergence, we implemented the mixing method in Julia. To each non diagonal entry of the matrix C a perturbation $\epsilon_{i,j} \in Unif(-\frac{1}{2}10^{-m}, \frac{1}{2}10^{-m})$ is added for $m = 6, 8, 10, 12, 14, 16$. To compute the convergence rate we randomly generate V and run the mixing method in which with each iteration we approximate the convergence rate with

$$\mu \approx \frac{|f(V^*) - f(V_{k+1})|}{|f(V^*) - f(V_k)|}.$$

Here the optimal solution V^* is obtained by running the SDP solver from Leijenhorst et al. [4]. This SDP solver is very efficient for solving SDPs with sparse constraint matrices, and is not dependent on the eigenvalues of S^*. As every constraint in our SDP problem consists of $\langle X, E_{i,i} \rangle = 1$ with only 1 digit of $E_{i,i}$ the matrix nonzero, this SDP solver runs very efficiently and calculates solutions with high precision.

For comparison of the convergence rate in a general case, the mixing method is first run on a random symmetric C matrix. As it rarely occurs that the smallest nonzero eigenvalue of S^* is near 0 we expect the mixing method's solution to converge to the manually set machine precision. Figure 1 shows the average random convergence rate for the mixing method on a random symmetric matrix C.

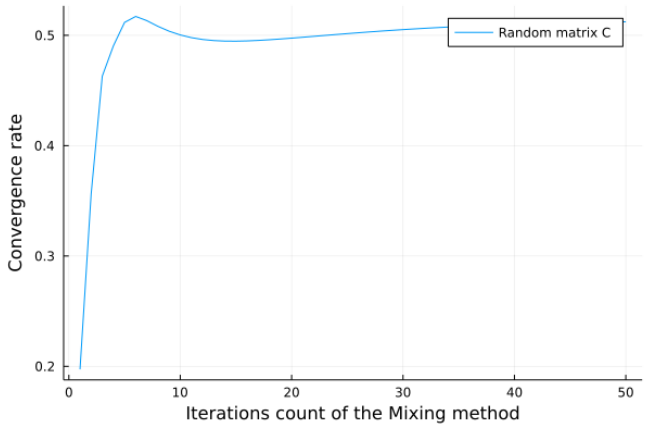

Figure 1: Convergence rate of the mixing method for a random symmetric matrix C

The converge rate is sitting at $\mu = 0.5$ meaning that the difference between subsequent iterations gets halved every iteration. For the random example, the mixing method continuously decreases until the difference between it and the optimal solution is 10^{-80}, which is our machine precision.

Now consider running the mixing method on maximizing the sum of squared pairwise distances on the unit sphere with a variable perturbation ϵ. The first 15 iterations of the mixing method's convergence rate are shown in figure 2. The figure is made on a logarithmic plot where $1 - \mu$ is plotted on the y-axis with the iteration count on the x-axis. After 15 iterations, the convergence rate stays constant (For at least 10^5 more iterations).

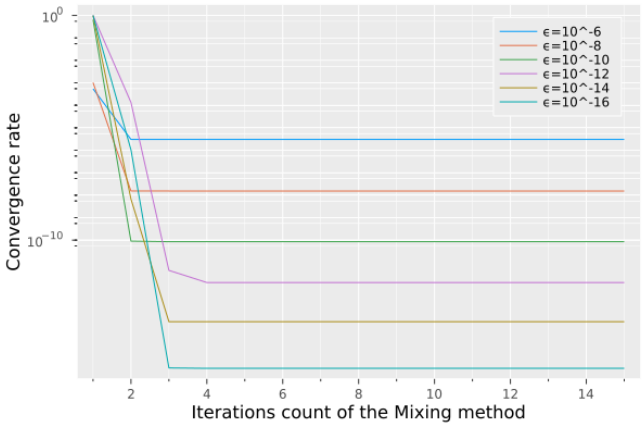

Figure 2: Convergence rate of the mixing method for a variable perturbation ϵ

For any ϵ the mixing method initially converges very fast converging almost superlinearly ($\mu = 0$). However, once the mixing method draws closer, the convergence slows down comparatively with the perturbation. This can be seen in figure 3, where the absolute difference between the current and optimal cost is plotted against the iterations.

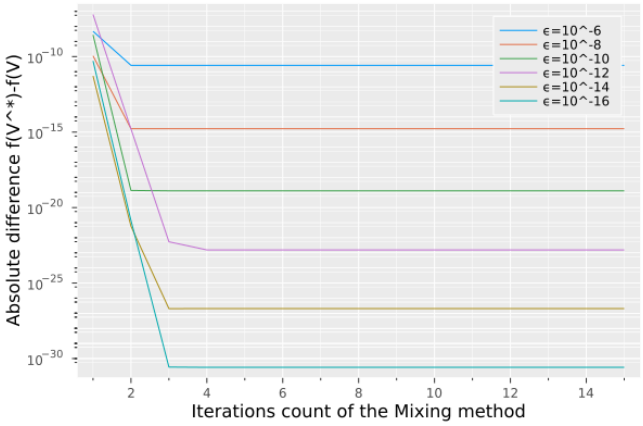

Figure 3: Difference in objective value between the optimal objective value and the mixing method's objective value for a variable perturbation ϵ

The convergence rate is in practice very close to the theoretical slowest convergence. The average convergence rates after 15 iterations is shown in table 1.

Perturbation ϵ	**Convergence rate μ**
10^{-6}	$1 - 3.09683 * 10^{-6}$
10^{-8}	$1 - 1.53072 * 10^{-8}$
10^{-10}	$1 - 8.51954 * 10^{-10}$
10^{-12}	$1 - 1.28965 * 10^{-12}$
10^{-14}	$1 - 2.33799 * 10^{-14}$
10^{-16}	$1 - 1.95822 * 10^{-16}$

Table 1: Average convergence rate for the mixing method when applied to maximizing the sum of squared pairwise distances for variable perturbation ϵ

It is interesting to note that had we not added any perturbation to our matrix C, the mixing method does converge super linearly to the optimal solution, as the smallest eigenvalue is 1. The mixing method however does not guarantee this as our C is not generic. After running the mixing method on the matrix $C = J_n - I_n$ numerous times it also shows that the obtained optimal solution is not rank deficient meaning that the theorems in place do not guarantee convergence for this C.

6 Comparing the convergence speed of the Mixing method to gradient descent and SDP solvers

The previous section has given insight into when the mixing method has slow convergence. These examples however, are rare. To evaluate the speed of convergence of the mixing method more generally, we test the method against a general SDP solver and gradient descent for a practical problem. Note that all computations in this section have been performed in high precision. This is because with high precision it is easier to determine whether a method has truly converged or not. In the previous section we have seen that the mixing method can be close to the optimal solution but also not have converged yet so this is a valid concern. High precision also reduces issues with memory locality.

We compare these three methods in their efficiency in solving a relaxation of the 'max-cut' problem problem, specifically the Goemans Williamson relaxation [6]. The max-cut problem is an NP-hard optimisation problem which tasks to find the greatest cut $\delta_{G(V,E)}(S)$ in weight for a given undirected weighted graph $G(V, E)$; see the book by Laurent et al. [9, Chapter 5] for a full explanation. Each vertex is assigned a binary variable $v_i \in \{-1, 1\}$ denoting whether the vertex is in S or not. Let $w_{i,j} \in \mathbb{R}_>$ denote the weight between vertices i and

j. If there is no edge between the two vertices, let $w_{i,j} = 0$. Then the max-cut problem can be written as the following optimization problem.

$$\begin{aligned} &\text{minimize } \frac{1}{2} \sum_{i,j \in [n]} w_{i,j} \frac{1 - v_i v_j}{2} \\ &\text{subject to } v_i \in \{-1, 1\}, \ \forall i \in [n] \end{aligned}$$

The Goemans and Williamson algorithm is an approximation algorithm that turns this binary optimization problem into an SDP. Instead of optimizing over binary variables, the variables v_i are transformed into $v_i \in \mathbb{R}^n$ such that $\|v_i\| = 1$. This relaxation is the following minimization problem.

$$\begin{aligned} &\text{minimize } \frac{1}{2} \sum_{i,j \in [n]} w_{i,j} \frac{1 - v_i v_j}{2} \\ &\text{subject to } \|v_i\| = 1 \ \forall i \in [n] \\ &v_i \in \mathbb{R}^n \end{aligned}$$

This is and SDP with only constraints on the diagonal and can hence be solved by the mixing method. The Goemans and Williamson algorithm then employs a rounding algorithm to obtain a feasible solution of the max-cut problem. Goemens et al.[6] proved that when all the weights $w_{i,j}$ are positive, the optimal objective value of the max-cut problem $f(S^*)$ is bounded by

$$f(\hat{S}^*) \geq f(S^*) \geq 0.878 f(\hat{S}^*),$$

where $\hat{S}^*$ is the solution found by the Goemans and Williamson algorithm. This practical problem can be used to test the convergence speed of the mixing method against other solvers. For this we generate a random graph. To generate a random graph we employ the well-documented Erdös-Rényi model popularized by Erdös et al. [5]. Given the amount of vertices n and a probability p, the model adds an edge between two vertices with probability p.

The SDP solver that will be used is once again the SDP solver from Leijenhorst et al. [4]. For gradient descent we employ a backtracking line search algorithm (See e.g. Kochender et al. [8, Section 4.3]). Each of the three methods will be tested against a random instance of the Goemans and Williamson max-cut relaxation. Instances are randomly generated using the Erdös-Rényi model with $p = 0.7$. The weights $w_{i,j}$ between edges are uniformly picked from the interval $[0, 100]$. 5 random instances are run for each algorithm to calculate an average speed of convergence for the node sizes $n = [5, 40]$. A method is considered to have completely converged once the objective value is within 10^{-22} of the the optimal objective function. Figure 4 shows the average computation time for each of the three methods.

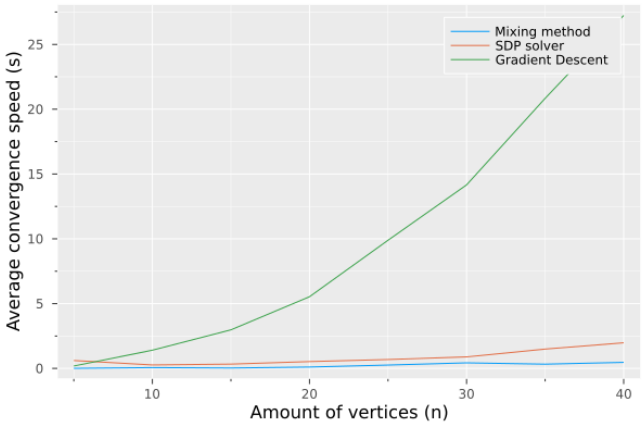

Figure 4: Convergence speed for variable node sizes of the max-cut problem. The figure represents the time it takes for each of the three algorithms to reach a distance of 10^{-22} in the neighbourhood of the optimal objective value.

From figure 4 it is evident that the gradient descent method is significantly slower than the other two methods. Gradient descent also doesn't always converge to the optimal solution. In fact, in only 80% percent of cases did gradient descent converge to the set standards, when averaged over all n. The figure does not yet show a strong difference in convergence between the mixing method implementation and the SDP solver from Leijenhorst et al. To test the scalability of these two methods, the convergence speed is again computed for node sizes $n = [50, 250]$ and presented in figure 5.

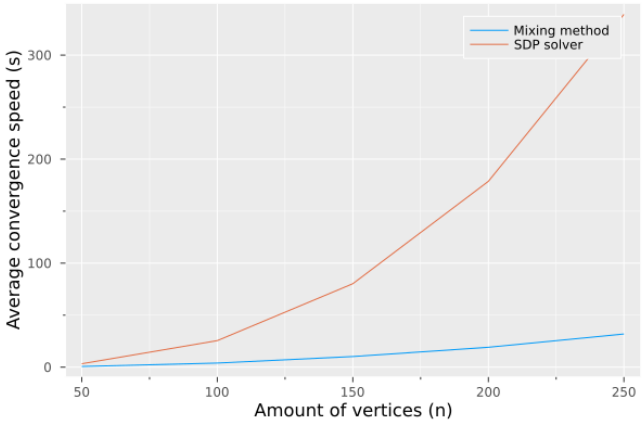

Figure 5: Convergence speed for variable node sizes of the max-cut problem. The figure compares the time it takes for the mixing method and the SDP solver to reach a distance of 10^{-22} in the neighbourhood of the optimal objective value.

From figure 5 it becomes more apparent that the mixing method is faster and scales better with an increase of the node size n, than the SDP solver does. Previously in section 5, the mixing method converged very slow for specific constructions of C. The SDP is not constrained by the smallest nonzero eigenvalue of $C + \text{Diag}(y)$ whereas the mixing method is. In general however, the mixing method converges faster than any other state of the art SDP solver.

7 Conclusion

In this thesis we have closely examined the mixing method; an iterative optimization algorithm that optimizes one column at a time. Surprisingly, even though the mixing method turns a convex problem into a non-convex problem, it becomes easier to solve reaching faster solving speeds than conventional SDP solvers. This is because one property this class of problems has is that when optimizing over only a single vector v_i, we can find a closed form solution. Hence the optimal solution is immediately found for this heuristic allowing an algorithm such as the mixing method to still have fast convergence even when the problem is non-convex.

We also saw that there is a class of symmetric matrices C, for which the convergence rate of the mixing method rate is linear but can be arbitrarily slow as the rate is bounded by the smallest eigenvalue of $C + \text{Diag}(y)$. These examples were generated from a problem of finding the maximal sum of squared pairwise distances for points on the unit sphere. This is one example of finding optimal spherical configurations, but there are a whole range of formulations for maximizing the distance between points on an k-dimensional sphere. For example the class

$$\begin{aligned} \text{minimize } & \sum \|v_i - v_j\|_2^m, \\ \text{subject to } & \|v_i\|_2 = 1 \ \forall i \in [n] \\ & v_i \in \mathbb{R}^n, \end{aligned}$$

where $m \in \mathbb{Z}_{\neq 0}$. This class of optimization problems also contains the problem discussed in section 5. However, if we attempt to find a similar closed form solution for $m \leq 1$, we find that no such optimal solution can be derived. Hence we see that the mixing method and other coordinate descent methods with similar qualities, still only have some niche applications.

It could therefore be interesting to find different applications for the mixing method. Wang, Chang and Kolter's paper already discussed other areas in which the mixing method can be applied such as relaxations for the Max-SAT problem as well as machine learning problems discussed in the Appendix. The promising numerical results they found for these examples can perhaps be consolidated with a proof of an approximation bound to bring more insight into the qualities of such SDP problems and spark further research to find real world applications for the mixing method.

References

[1] J. Argelich, C. Min Li, F. Manya, and J. Planes. Max-sat-2016 eleventh max-sat evaluation. *http://maxsat.ia.udl.cat/*, 2016.

[2] M. S. Birman and M. Z. Solomjak. *Spectral theory of self-adjoint operators in Hilbert space*, volume 5. Springer Science & Business Media, 2012.

[3] N. Boumal, V. Voroninski, and A. Bandeira. The non-convex burer-monteiro approach works on smooth semidefinite programs. *Advances in Neural Information Processing Systems*, 29, 2016.

[4] D. de Laat and N. Leijenhorst. Solving clustered low-rank semidefinite programs arising from polynomial optimization. *arXiv preprint arXiv:2202.12077*, 2022.

[5] P. Erdos, A. Rényi, et al. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci*, 5(1):17–60, 1960.

[6] M. Goemans and D. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming, submitted to j. *ACM.(contact goemans@ math. mit. edu for copies)*, 1994.

[7] G. H. Golub and C. F. Van Loan. *Matrix computations*. JHU press, 2013.

[8] M. J. Kochenderfer and T. A. Wheeler. *Algorithms for optimization*. Mit Press, 2019.

[9] M. Laurent and F. Vallentin. Semidefinite optimization. *Lecture Notes, available at http://page. mi. fu-berlin. de/fmario/sdp/laurentv. pdf*, 2012.

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

[11] K. B. Petersen, M. S. Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.

[12] K. Schmüdgen. Discrete spectra of self-adjoint operators. In *Unbounded self-adjoint operators on Hilbert space*, pages 265–280. Springer, 2012.

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