A Spectral Bundle Method for Sparse Semidefinite Programs

Hesam Mojtahedi, Feng-Yi Liao, and Yang Zheng

***Abstract*— Semidefinite programs (SDPs) have many applications in the field of controls. To improve scalability, it is important to exploit the inherent sparsity when solving SDPs. In this paper, we develop a new spectral bundle algorithm that solves sparse SDPs without introducing additional variables. We first apply chordal decomposition to replace a large positive semidefinite (PSD) constraint with a set of smaller coupled constraints. Then, we move the coupled constraints into the cost function via exact penalty. This leads to an equivalent non-smooth penalized problem, which can be solved by bundle methods. We present a new efficient spectral bundle algorithm, where subgradient information is incorporated to update a lower approximation at each iteration. We further establish sublinear convergences in terms of objective value, primal feasibility, dual feasibility, and duality gap. Under Slater's condition, the algorithm converges with the rate of $\mathcal{O}\left(1/\epsilon^3\right)$, and the rate improves to $\mathcal{O}\left(1/\epsilon\right)$ when strict complementarity holds. Our numerical experiments support the theoretical analysis.**

I. Introduction

Semidefinite programs (SDPs) are an important sub-field of optimization that involves the minimization of a linear objective function over the cone of PSD matrices with linear constraints. The standard primal and dual forms of SDPs are

$$\begin{aligned} \min_{X} \quad & \langle C, X \rangle \\ \text{subject to} \quad & \langle A_i, X \rangle = b_i, \quad i = 1, \ldots, m \\ & X \in \mathbb{S}^n_+, \end{aligned} \tag{1}$$

$$\begin{aligned} \max_{y,Z} \quad & b^\top y \\ \text{subject to} \quad & Z + \sum_{i=1}^{m} A_i y_i = C, \\ & Z \in \mathbb{S}^n_+, \end{aligned} \tag{2}$$

where $A_1, A_2, \ldots, A_m, C \in \mathbb{S}^n$ and $b \in \mathbb{R}^m$ are problem data, $\mathbb{S}^n_+$ stands for the cone of PSD matrices, and $\langle \cdot, \cdot \rangle$ denotes the standard matrix inner product. SDPs have important applications in numerous fields such as combinatorial optimization [1], control theory [2], machine learning [3]. Furthermore, many graph-theoretic problems (such as max cut and graph partitioning) can be addressed using SDPs [4].

In theory, SDPs can be solved to arbitrary accuracy in polynomial time using interior-point methods [5]. However, due to its computational complexity, it is often impractical to solve large-scale SDPs considering memory and time constraints [3], [6]. The state-of-the-art solvers for SDPs, such as MOSEK [7], can only solve medium-sized problems reliably (e.g., $n, m \leq 1000$ in (1)) on regular laptops. Improving the scalability of solving SDPs has received extensive research interest [3], [6], [8]. First-order algorithms are one promising direction for computational scalability when solutions of moderate accuracy are required. For example, a general conic solver based on the alternating direction method of multipliers (ADMM) was developed in [9]. This approach has been extended in [10] to exploit the underlying sparsity in SDPs based on chordal decomposition. We refer interested readers to [6, Section 3] for a recent survey. Despite the efficiency of first-order methods per iteration, obtaining high-accuracy solutions remains challenging and may require an unacceptable number of iterations due to the slow convergence.

Another approach is to apply structured decomposition to decompose a large PSD matrix $X \in \mathbb{S}^n_+$ into structured ones that are easier to impose positivity [6], [11], [12]. For a sparse matrix, we can associate it with a graph, and the principal submatrices can be identified by maximal cliques of the graph (see Section II). If the sparsity graph is *chordal*, which means that all cycles of length greater than three have an edge between nonconsecutive vertices in a cycle, a clique-based decomposition is guaranteed to exist for sparse PSD matrices [13]. In this case, it is possible to equivalently replace a large matrix constraint $X \in \mathbb{S}^n_+$ with a set of smaller and coupled matrix constraints. This chordal decomposition strategy, combined with a dual result on the existence of PSD matrix completions, is promising to significantly reduce the computational complexity of SDPs that involve sparse PSD matrices; see the developments in [6], [14]–[16].

In this paper, we focus on a spectral bundle method proposed in [17], which shows fast practical convergence and enjoys low computation complexity per iteration. In particular, the dual SDP (2) is transformed into an equivalent eigenvalue optimization by exploring the constant trace property in [17]. Very recently, [18] generalized the spectral bundle method to any SDPs and showed convergence in terms of primal feasibility, dual feasibility, and primal-dual duality gap. Furthermore, a linear convergence rate of the spectral bundle method is established under mild assumptions [18]. We refer the interested reader to [19] for a recent comparison.

In this work, inspired by [17]–[19], we propose a first-order spectral bundle method to solve sparse SDPs that are characterized by a chordal graph or chordal extension. Specifically, instead of solving (1) directly, benefiting from chordal sparsity property, we decompose the large semidefinite constraint in (1) into several smaller ones. We emphasize that the smaller PSD constraints are interdependent in general. In many existing methods, such as those outlined in [6, Section 3], different additional consensus constraints have

*This work is supported by NSF ECCS-2154650 and CMMI-2320697.

The authors are with the Department of Electrical and Computer Engineering, University of California San Diego, La Jolla, CA 92093, USA. Emails: hmojtahedi@ucsd.edu; fliao@ucsd.edu; zhengy@ucsd.edu.

been introduced to handle the coupled constraints. Instead, we solve an equivalent penalized problem without introducing extra variables, which is in the form of constrained non-smooth eigenvalue optimization. Similar to [17]–[19], this problem is well-suited to be solved via bundle methods [20]. In particular, we adapt and tailor the techniques in [17]–[19] to solve the resulting non-smooth problem, leading to a new spectral bundle algorithm for sparse SDPs. Assuming Slater's condition, we prove that the algorithm converges as $\mathcal{O}\left(1/\epsilon^3\right)$. If the problem satisfies strict complementarity, the convergence rate is enhanced to $\mathcal{O}\left(1/\epsilon\right)$.

The rest of this paper is structured as follows. We cover some preliminaries on chordal graphs and bundle methods in Section II. In Section III, we introduce an exact penalization for sparse SDPs. This allows us to develop a new spectral bundle algorithm in Section IV. We present numerical results in Section V, and conclude the paper in Section VI. Some technical proofs are provided in Appendix.

II. Preliminaries

In this section, we review graph theory for matrix decomposition and bundle methods for non-smooth optimization.

A. Chordal graphs and matrix decomposition

A graph $\mathcal{G}(\mathcal{V},\mathcal{E})$ is defined by a set of vertices $\mathcal{V}$ and a set of edges $\mathcal{E}\subseteq\mathcal{V}\times\mathcal{V}$. A graph is called *undirected* if the edges do not have orientations, i.e., $(i,j)\in\mathcal{E}\Leftrightarrow(j,i)\in\mathcal{E}$. A subset of vertices $\mathcal{C}\subseteq\mathcal{V}$ is called a clique if every pair of vertices in $\mathcal{C}$ is connected by an edge. A clique is maximal if it is not a subset of any other clique. We use $|\mathcal{C}|$ to denote the number of vertices in the clique. A cycle in a graph is defined as a sequence of vertices and edges that begins and ends at the same vertex. A chord is an edge between two non-consecutive vertices in a cycle. An undirected graph $\mathcal{G}$ is called *chordal* if it contains at least one chord in every cycle of length greater than three.

Given a graph $\mathcal{G}(\mathcal{V},\mathcal{E})$, a matrix $X\in\mathbb{S}^n$ has sparsity pattern $\mathcal{E}$ if $X_{ij}=X_{ji}=0,\ \forall(i,j)\notin\mathcal{E}, i\neq j$. We denote the space of sparse symmetric matrices by

$$\mathbb{S}^n(\mathcal{E},0):=\left\{X\in\mathbb{S}^n\mid X_{ij}=0,\text{ if }(i,j)\notin\mathcal{E},i\neq j\right\}.$$

Given a matrix $X\in\mathbb{S}^n$, let $\mathbb{P}_{\mathbb{S}^n(\mathcal{E},0)}(X)$ be the projection onto $\mathbb{S}^n(\mathcal{E},0)$ with respect to the Frobenius norm, i.e., $M=\mathbb{P}_{\mathbb{S}^n(\mathcal{E},0)}(X)$ with $M_{ij}=0$, if $(i,j)\notin\mathcal{E},i\neq j$ and $M_{ij}=X_{ij}$, otherwise. Then, we define the cone of positive-semidefinite completable matrices as

$$\mathbb{S}^n_+(\mathcal{E},?):=\mathbb{P}_{\mathbb{S}^n(\mathcal{E},0)}\left(\mathbb{S}^n_+\right).$$

In other words, $X\in\mathbb{S}^n_+(\mathcal{E},?)$ if some (or all) of the zero entries X_{ij} with $(i,j)\notin\mathcal{E},i\neq j$ can be replaced with nonzeros to obtain a PSD matrix $\bar{X}\in\mathbb{S}^n_+$. We call $\bar{X}$ the PSD completion of $X\in\mathbb{S}^n_+(\mathcal{E},?)$.

Given a clique $\mathcal{C}_k$ of graph $\mathcal{G}$, we define an index matrix $E_{\mathcal{C}_k}\in\mathbb{R}^{|\mathcal{C}_k|\times n}$ as follow

$$(E_{\mathcal{C}_k})_{ij}=\begin{cases}1, & \text{if } \mathcal{C}_k(i)=j\\ 0, & \text{otherwise.}\end{cases}$$

Given a matrix $X\in\mathbb{S}^n$, the operation $E_{\mathcal{C}_k}XE_{\mathcal{C}_k}^{\mathsf{T}}\in\mathbb{S}^{|\mathcal{C}_k|}$ selects the submatrix indexed by $\mathcal{C}_k$. Alternatively, given $Y\in\mathbb{S}^{|\mathcal{C}_k|}$, the operation $E_{\mathcal{C}_k}^{\mathsf{T}}YE_{\mathcal{C}_k}\in\mathbb{S}^n$ expands Y into a sparse $n\times n$ matrix that contains Y as its principal submatrix indexed $\mathcal{C}_k$, and zero otherwise.

Theorem 1 ([6, Theorem 2.2]): Given a chordal graph $\mathcal{G}(\mathcal{V},\mathcal{E})$ with maximal cliques $\mathcal{C}_1,\ldots,\mathcal{C}_p$, we have $X\in\mathbb{S}^n_+(\mathcal{E},?)$ if and only if $E_{\mathcal{C}_k}XE_{\mathcal{C}_k}^{\mathsf{T}}\in\mathbb{S}^{|\mathcal{C}_k|}_+,\ \forall k=1,\ldots,p$.

This result replaces a large constraint $X\in\mathbb{S}^n_+(\mathcal{E},?)$ with a set of smaller PSD constraints, indexing by the cliques. If the chordal graph has small cliques, we can expect computational improvements, which have been widely used (see [6] for a survey). In this paper, we will exploit Theorem 1 to develop a new spectral bundle method for solving sparse SDPs.

B. Bundle methods

The bundle method [20] is a standard technique to solve a non-smooth convex optimization problem of the form

$$\begin{aligned}\min_{x\in\mathcal{X}_0}\quad & f(x)\\ \text{subject to}\quad & x\in\mathcal{X}_0,\end{aligned}\tag{3}$$

where $f:\mathbb{R}^n\to\mathbb{R}$ is a convex (possibly non-differentiable) function and $\mathcal{X}_0$ is a simple convex set. We refer the interested reader to [20] for a detailed discussion on bundle methods. We only introduce a few key ingredients below.

One key step in the bundle method is to construct a lower approximation $\hat{f}_t(x)$ of the objective function $f(x)$ at each iteration t, i.e., $\hat{f}_t(x)\leq f(x),\forall x\in\mathcal{X}_0$. One standard way is to use a subgradient to form an under-estimator $\hat{f}(x)=f(\hat{x})+\langle g,x-\hat{x}\rangle$ where g is a subgradient of f at point $\hat{x}$, but other methods also exist [17]. At each iteration of the bundle method, we perform the following proximal step

$$y_{t+1}\in\operatorname*{argmin}_{x}\quad\hat{f}_t(x)+\frac{\alpha}{2}\left\|x-x_t\right\|^2,\tag{4}$$

where x_t is the current reference point and $\alpha>0$ penalizes the deviation from x_t. If the candidate point y_{t+1} gives a sufficient descent in the true cost function, i.e. let $\beta\in(0,1)$, we have $\beta\left(f\left(x_t\right)-\hat{f}_t\left(y_{t+1}\right)\right)\leq f\left(x_t\right)-f\left(y_{t+1}\right)$, then we update the current (reference) iterate $x_{t+1}=y_{t+1}$ (*descent step*); otherwise, the reference point does not change, $x_{t+1}=x_t$ (*null step*). In either case, y_{t+1} will be used to update the lower approximation function $\hat{f}_{t+1}(x)$.

The bundle method is guaranteed to return a converging sequence x_t to a minimizer of (3) (if it exists), when $\hat{f}_t$ satisfies three properties [21] and [19, Section 2.3.3]

$$\hat{f}_{t+1}(x)\leq f(x),\ \forall x\in\mathcal{X}_0,\tag{5a}$$

$$\hat{f}_{t+1}(x)\geq f\left(y_{t+1}\right)+\left\langle g_{t+1},x-y_{t+1}\right\rangle,\forall x\in\mathcal{X}_0,\tag{5b}$$

$$\hat{f}_{t+1}(x)\geq\hat{f}_t\left(y_{t+1}\right)+\left\langle s_{t+1},x-y_{t+1}\right\rangle,\forall x\in\mathcal{X}_0,\tag{5c}$$

where $s_{t+1}=\alpha(x_t-y_{t+1})\in\partial\hat{f}(y_{t+1})+\mathcal{N}_{\mathcal{X}_0}(y_{t+1})$, and $g_{t+1}\in\partial f(y_{t+1})$ with the subdifferential defined as

$$\partial f(x)=\{g\in\mathbb{R}^n\mid f(y)\geq f(x)+\langle g,y-x\rangle,\ \forall y\in\mathbb{R}^n\},$$

and the normal cone defined as

$$\mathcal{N}_{\mathcal{X}_0}(y)=\{h\in\mathbb{R}^n\mid\langle h,x-y\rangle\leq 0,\ \forall x\in\mathcal{X}_0\}.$$

III. Exact penalization for sparse SDPs

In this section, we introduce an exact penalization of sparse SDPs (1) into the form of (3). This allows us to develop the spectral bundle method in the next section.

A. Exact penalization for constrained convex optimization

Consider a constrained convex optimization problem:

$$\begin{aligned} \min_{x} \quad & f(x) \\ \text{subject to} \quad & g_i(x) \leq 0, \quad i = 1, \ldots, m, \\ & x \in \mathcal{X}_0, \end{aligned} \tag{6}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $g_i : \mathbb{R}^n \to \mathbb{R}, i = 1, \ldots, m$ are (possibly non-differentiable) convex functions, $\mathcal{X}_0 \subseteq \mathbb{R}^n$ is a convex closed set. The idea of exact penalty methods is to reformulate (6) by introducing an exact penalty function $P(x) = \sum_{i=1}^m \max\{0, g_i(x)\}$. We then consider a penalized problem

$$\begin{aligned} \min_{x} \quad & \Phi_\rho(x) := f(x) + \rho P(x) \\ \text{subject to} \quad & x \in \mathcal{X}_0, \end{aligned} \tag{7}$$

where $\rho > 0$ is a penalty parameter. It is known that when choosing ρ large enough and assuming Slater's condition, problems (6) and (7) are equivalent in the sense that they have the same optimal value and solution set [22, Theorem 7.21]. Therefore, we can transform some nonsmooth constraints that are hard to handle in (6) into the nonsmooth cost in (7). Then, we can apply the bundle method (cf. Section II-B) to solve the nonsmooth problem (7).

B. Non-smooth penalization of sparse SDPs

We consider the standard primal SDP (1). In many practical applications, the matrices $A_1, \ldots, A_m, C$ in problem data are often sparse [6]. If they share a common sparsity pattern $\mathcal{G}(\mathcal{V}, \mathcal{E})$, i.e., $C \in \mathbb{S}^n(\mathcal{E}, 0), A_i \in \mathbb{S}^n(\mathcal{E}, 0), i = 1, \ldots, m$, we refer to this graph as *aggregate sparsity* pattern. It is not difficult to verify that (1) is equivalent to

$$\begin{aligned} \min_{X} \quad & \langle C, X \rangle \\ \text{subject to} \quad & \langle A_i, X \rangle = b_i, \quad i = 1, \ldots, m, \\ & X \in \mathbb{S}^n_+(\mathcal{E}, ?). \end{aligned} \tag{8}$$

Without loss of generality, we assume that the aggregate sparsity pattern $\mathcal{G}(\mathcal{V}, \mathcal{E})$ is a chordal graph with maximal cliques $\mathcal{C}_1, \ldots, \mathcal{C}_p$ (otherwise a suitable chordal extension can be performed). Then, Theorem 1 allows us to reformulate problem (8) into

$$\begin{aligned} \min_{X} \quad & \langle C, X \rangle \\ \text{subject to} \quad & \langle A_i, X \rangle = b_i, \quad i = 1, \ldots, m \\ & E_{\mathcal{C}_k} X E_{\mathcal{C}_k}^{\mathsf{T}} \in \mathbb{S}^{|\mathcal{C}_k|}_+, \quad k = 1, \ldots, p. \end{aligned} \tag{9}$$

The single semidefinite constraint in (1) is replaced by multiple smaller constraints in (9). This decomposition (9) underpins many scalable algorithms for sparse SDPs [6]. The submatrices $E_{\mathcal{C}_k} X E_{\mathcal{C}_k}^{\mathsf{T}}$ induced by maximal cliques may overlap, therefore the semidefinite constraints in (9) are coupled. Previous techniques in [6], [10], [23] introduce a large number of consensus constraints such as $X_k = E_{\mathcal{C}_k} X E_{\mathcal{C}_k}^{\mathsf{T}}$.

In this work, we introduce further reformulations which allow us to solve (9) directly without adding extra variables. It is clear that (9) is equivalent to

$$\begin{aligned} \min_{X} \quad & \langle C, X \rangle \\ \text{subject to} \quad & \langle A_i, X \rangle = b_i, \quad i = 1, \ldots, m, \\ & \lambda_{\max}\left(E_{\mathcal{C}_k}(-X) E_{\mathcal{C}_k}^{\mathsf{T}}\right) \leq 0, \ k = 1, \ldots, p. \end{aligned} \tag{10}$$

The eigenvalue constraints in (10) are non-smooth. Similar to (7), we then get the following formulation via exact penalty

$$\begin{aligned} \min_{X} \quad & \langle C, X \rangle + \rho \sum_{k=1}^p \max\left\{0, \lambda_{\max}\left(E_{\mathcal{C}_k}(-X) E_{\mathcal{C}_k}^{\mathsf{T}}\right)\right\} \\ \text{subject to} \quad & \langle A_i, X \rangle = b_i, \ i = 1, \ldots, m. \end{aligned} \tag{11}$$

Unlike previous results in [6], [10], [23], this formulation (11) only has a single matrix variable X (with no extra variables). It is clear (11) is in the form of (3). We will develop a spectral bundle method to solve (11) in Section IV.

C. Properties and assumptions

Before developing the spectral bundle method, we expect that problems (11) and (1) are equivalent when the penalty parameter ρ is large enough. In particular, let us consider the Lagrange dual problem of (9), which is

$$\begin{aligned} \max_{y, Y_k} \quad & b^{\mathsf{T}} y \\ \text{subject to} \quad & C - \sum_{i=1}^m y_i A_i = \sum_{k=1}^p E_{\mathcal{C}_k}^{\mathsf{T}} Y_k E_{\mathcal{C}_k}, \\ & Y_k \in \mathbb{S}^{|\mathcal{C}_k|}_+, \quad k = 1, \ldots, p. \end{aligned} \tag{12}$$

It can be verified that (12) is also equivalent to (2). We denote the optimal solution set of the decomposed primal SDP (9) and dual SDP (12) by $\mathcal{P}^\star$ and $\mathcal{D}^\star$ respectively. Throughout the paper, we make the following assumptions.

Assumption 1: The matrices $A_i, i = 1, \ldots, m$ in (1) and (2) are linearly independent.

Assumption 2: The primal and dual SDPs (9) and (12) satisfy Slater's condition (i.e., they are strictly feasible) and their solution sets, $\mathcal{P}^\star$ and $\mathcal{D}^\star$, are compact.

We have the following technical result, and its proof is provided in Appendix.

Proposition 1: Under Assumption 2, the non-smooth penalized formulation (11) is equivalent to the original primal SDP (1) if we choose

$$\rho > \mathcal{D}_{\mathcal{Y}^\star} := \max_{(y^\star, \{Y_k^\star\}) \in \mathcal{D}^\star} \left\{\mathbf{tr}(Y_1^\star), \mathbf{tr}(Y_2^\star), \ldots, \mathbf{tr}(Y_p^\star)\right\}.$$

We conclude with a notion of strict complementarity.

Definition 1 (strict complementarity): A pair of optimal solutions $(X^\star, \{y^\star, Y_k^\star\}) \in \mathcal{P}^\star \times \mathcal{D}^\star$ in (9) and (12) satisfies *strict complementarity* if

$$\operatorname{rank}\left(E_{\mathcal{C}_k} X^\star E_{\mathcal{C}_k}^{\mathsf{T}}\right) + \operatorname{rank}\left(Y_k^\star\right) = |\mathcal{C}_k|, \ k = 1, \ldots, p.$$

If such a pair exists, we say the decomposed SDPs (9) and (12) satisfy strict complementarity.

IV. A Spectral Bundle Method

In this section, we introduce a spectral bundle algorithm to solve the penalized nonsmooth problem (11), and establish its convergence guarantees.

A. Constructions of lower-approximation model

For simplicity, we denote the cost function in (11) as

$$G(X) := \langle C, X\rangle + \rho \sum_{k=1}^{p} \max\left\{0, \lambda_{\max}\left(E_{\mathcal{C}_k}(-X)E_{\mathcal{C}_k}^{\mathsf{T}}\right)\right\}.$$

As discussed in Section II-B, one key step in the bundle method is to construct an appropriate lower approximation for $G(X)$ that satisfies (5a) to (5c). Our strategy for constructing a lower approximation is motivated by [18], [19].

For each clique k, we use a matrix $P_k \in \mathbb{R}^{|\mathcal{C}_k|\times r}$ with orthonormal columns, where $r \le \min_k |\mathcal{C}_k|$, such that $P_k^{\mathsf{T}} P_k = I_{r\times r}$, and construct a lower approximation:

$$\begin{aligned}\hat{G}_{\{P_k\}}(X) &= \langle C, X\rangle + \\ &\rho \sum_{k=1}^{p} \max_{\substack{S_k \in \mathbb{S}^r_+,\\ \mathbf{tr}(S_k)\le 1}} \left\langle P_k S_k P_k^{\mathsf{T}}, E_{\mathcal{C}_k}(-X)E_{\mathcal{C}_k}^{\mathsf{T}}\right\rangle.\end{aligned} \tag{13}$$

By definition, we observe $\hat{G}_{\{P_k\}}(X) \le G(X), \forall X \in \mathbb{S}^n$ thanks to the fact

$$\max\{\lambda_{\max}(-X), 0\} = \max_{S\in\mathbb{S}^n_+, \mathbf{tr}(S)\le 1} \langle S, -X\rangle, \quad \forall X \in \mathbb{S}^n,$$

and $\{P_k S P_k^{\mathsf{T}} \in \mathbb{S}^n_+ \mid S \in \mathbb{S}^r_+, \mathbf{tr}(S) \le 1\} \subseteq \{S \in \mathbb{S}^n_+ \mid \mathbf{tr}(S) \le 1\}$. Therefore, $\hat{G}_{\{P_k\}}(X)$ in (13) serves as an underestimator that meets the requirement (5a).

We can also verify the subgradient lower bound condition (5b) when we choose P_k spanning the top eigenvector associated with $E_{\mathcal{C}_k}(-X)E_{\mathcal{C}_k}$. Another modification is required to ensure that the condition (5c) is fulfilled. Along with selecting past and current eigenvectors to generate P_k, the spectral bundle approach in [17] retains a thoughtfully chosen weight to incorporate past information. Notably, we introduce a constant matrix $\bar{W}_k \in \mathbb{S}^{|\mathcal{C}_k|}_+$ for each clique k with $\mathbf{tr}\left(\bar{W}_k\right) = 1$, and define the set, $k = 1, \ldots, p$

$$\begin{aligned}\hat{\mathcal{W}}_k := \{\gamma_k \bar{W}_k + P_k S_k P_k^{\mathsf{T}} \mid S_k \in \mathbb{S}^r_+, \\ \gamma_k \ge 0, \gamma_k + \mathbf{tr}\left(S_k\right) \le 1\}.\end{aligned} \tag{14}$$

We then refine the lower approximation function below

$$\begin{aligned}&\hat{G}_{\{\bar{W}_k, P_k\}}(X) \\ &= \langle C, X\rangle + \rho \sum_{k=1}^{p} \max_{W_k \in \hat{\mathcal{W}}_k} \left\langle W_k, E_{\mathcal{C}_k}(-X)E_{\mathcal{C}_k}^{\mathsf{T}}\right\rangle.\end{aligned} \tag{15}$$

It is evident that the lower approximation model (15) provides a better estimate than (13). Letting $\gamma_k = 0$ reduce (15) to (13); hence (15) satisfies (5a), (5b), as well as (5c) by carefully constructing $\bar{W}_k$ and P_k at each iteration.

B. A spectral bundle algorithm

Following Section II-B, we present a spectral bundle algorithm to solve (11) based on the lower approximation model (15). In this algorithm, we will construct a lower approximation model $\hat{G}_{\{\bar{W}_{t,k}, P_{t,k}\}}(X_t)$ and update the model parameters $\{\bar{W}_{t,k}\}$, $\{P_{t,k}\}$, and the set $\hat{\mathcal{W}}_{t,k} := \{\gamma_{t,k}\bar{W}_{t,k} + P_{t,k}S_{t,k}P_{t,k}^{\mathsf{T}} \mid \gamma_{t,k} \ge 0, S_{t,k} \in \mathbb{S}^r_+, \gamma_{t,k} + \mathbf{tr}(S_{t,k}) \le 1\}$ at each iteration t. The overall algorithm is listed in Algorithm 1, which has the following steps:

Pre-processing: The algorithm starts by extracting aggregate sparsity pattern of the problem data and computing the maximal cliques $\mathcal{C}_1, \ldots, \mathcal{C}_p$. This step can be performed very efficiently; see [6].

Initialization: The algorithm is initiated with a random reference point $\Omega_0 \in \mathbb{S}^n$ and $P_{0,k} \in \mathbb{R}^{|\mathcal{C}_k|\times r}$ by setting the top r eigenvectors of $E_{\mathcal{C}_k}(-\Omega_0)E_{\mathcal{C}_k}^{\mathsf{T}}$ as their columns. We choose $\bar{W}_{0,k} \in \mathbb{S}^{|\mathcal{C}_k|}_+$ with $\mathbf{tr}\left(\bar{W}_{0,k}\right) = 1$, and construct the initial under-estimator $\hat{G}_{\{\bar{W}_{0,k}, P_{0,k}\}}(X)$ as in (15).

Solving the master problem: Similar to (4), our algorithm solves the following problem at iteration $t \ge 0$ to get the next iteration parameters and the candidate reference point

$$\begin{aligned}&\left(X^\star_{t+1}, S^\star_{t,k}, \gamma^\star_{t,k}\right) \\ &= \operatorname*{argmin}_{X\in\mathcal{X}_0} \hat{G}_{\{\bar{W}_{t,k}, P_{t,k}\}}(X) + \frac{\alpha}{2}\left\|X - \Omega_t\right\|_F^2,\end{aligned} \tag{16}$$

where $\mathcal{X}_0 = \{X \in \mathbb{S}^n \mid \langle A_i, X\rangle = b_i, i = 1, 2, \ldots, m\}$, and Ω_t is the reference point at iteration t and $\alpha > 0$ is a parameter which penalizes the deviation from Ω_t.

Update reference point: The algorithm updates the reference point if the following condition holds

$$\begin{aligned}&\beta\left(G\left(\Omega_t\right) - \hat{G}_{\{\bar{W}_{t,k}, P_{t,k}\}}\left(X^\star_{t+1}\right)\right) \\ &\le G\left(\Omega_t\right) - G\left(X^\star_{t+1}\right),\end{aligned} \tag{17}$$

where $\beta \in (0, 1)$. This indicates that if the actual cost reduction $G\left(\Omega_t\right) - G\left(X^\star_{t+1}\right)$ is greater or equal than β portion of the approximate reduction $G\left(\Omega_t\right) - \hat{G}_{\{\bar{W}_{t,k}, P_{t,k}\}}\left(X^\star_{t+1}\right)$, a *decent step* happens and the algorithm updates the reference point, i.e., $\Omega_{t+1} = X^\star_{t+1}$. Otherwise, a *null step* happens and the reference point does not change, i.e., $\Omega_{t+1} = \Omega_t$.

Update the under-estimation model: The algorithm updates the model (15) at each iteration to improve the approximation accuracy. Similar to [18, Section 2.2], to compute $\bar{W}_{t+1,k}, P_{t+1,k}$, we apply eigenvalue decomposition to small matrices $S^\star_{k,t}$ as below

$$S^\star_{t,k} = \begin{bmatrix} Q_{k,1} & Q_{k,2}\end{bmatrix}\begin{bmatrix}\Sigma_{k,1} & 0 \\ 0 & \Sigma_{k,2}\end{bmatrix}\begin{bmatrix} Q_{k,1}^{\mathsf{T}} \\ Q_{k,2}^{\mathsf{T}}\end{bmatrix},$$

where $Q_{k,1} \in \mathbb{R}^{r\times r_{\mathrm{p}}}$ and $Q_{k,2} \in \mathbb{R}^{r\times r_{\mathrm{c}}}$ contain the orthonormal eigenvectors associated with eigenvalues $\Sigma_{k,1}$ and $\Sigma_{k,2}$ respectively, with $\Sigma_{k,1}$ consisting of the largest r_{p} eigenvalues and $\Sigma_{k,2}$ consisting of the remaining eigenvalues. We compute the $V_{t,k} \in \mathbb{R}^{|\mathcal{C}_k|\times r_{\mathrm{c}}}$ with its columns being the top $r_{\mathrm{c}} \ge 1$ orthonormal eigenvectors of $E_{\mathcal{C}_k}(-X^\star_{t+1})E_{\mathcal{C}_k}^{\mathsf{T}}$ which captures the current sub-gradient information. Then,

Algorithm 1 Spectral bundle method for sparse SDPs

```
Require: Problem data $A_1, \ldots, A_m, C \in \mathbb{S}^n$, $b \in \mathbb{R}^n$.
Require: Parameters $r_{\mathrm{p}} \geq 0, r_{\mathrm{c}} \geq 1$, $\alpha > 0$, $\beta \in (0,1)$, $\epsilon \geq 0$. An initial point $\Omega_0 \in \mathbb{S}^n$.
  Pre-processing: Extract aggregate sparsity pattern of the problem data and compute maximal cliques.
  Initialization: Let $r = r_{\mathrm{p}} + r_{\mathrm{c}}$. Initialize $\bar{W}_{0,k} \in \mathbb{S}_+^{|\mathcal{C}_k|}$, with $\mathbf{tr}\left(\bar{W}_{0,k}\right) = 1$, and construct $P_{0,k} \in \mathbb{R}^{|\mathcal{C}_k| \times r}$ with its columns set to the top $r$ orthonormal eigenvectors of $E_{\mathcal{C}_k}(-\Omega_0)E_{\mathcal{C}_k}^T$.
  for $t = 0, \ldots, t_{\max}$ do
      Solve (16) to obtain $X_{t+1}^\star, \gamma_{k,t}^\star$, and $S_{k,t}^\star$.
      \\master problem
      If $G(\Omega_t) - \hat{G}_{\{\bar{W}_{t,k}, P_{t,k}\}}(X_{t+1}^\star) \leq \epsilon$, then stop.
      Set $\Omega_{t+1} = \begin{cases} X_{t+1}^\star, & \text{if (17) holds.} \quad \backslash\backslash \textit{descent step} \\ \Omega_t, & \text{otherwise.} \quad \backslash\backslash \textit{null step} \end{cases}$
      Compute $P_{k,t+1}$ as (18) and $\bar{W}_{k,t+1}$ as (19).
      \\update model
  end for
```

the next parameter $P_{t+1,k}$ is updated as

$$P_{t+1,k} = \operatorname{orth}\left(\begin{bmatrix} V_{t,k}, & P_{t,k}Q_{1,k} \end{bmatrix}\right). \tag{18}$$

The update of the weight matrices $\bar{W}_{t,k}$ captures the remaining past information

$$\bar{W}_{t+1,k} = \frac{\left(\gamma_{t,k}^\star \bar{W}_{t,k} + P_{t,k}Q_{k,2}\Sigma_{k,2}Q_{k,2}^\mathsf{T}P_{t,k}^\mathsf{T}\right)}{\gamma_{t,k}^\star + \mathbf{tr}\left(\Sigma_{k,2}\right)}, \tag{19}$$

where $\bar{W}_{t+1,k}$ is normalized with $\mathbf{tr}(\bar{W}_{t+1,k}) = 1$. If $r_{\mathrm{p}} = 0$, the parameter updates in (18) and (19) become $P_{t+1,k} = V_{t,k} \in \mathbb{R}^{|\mathcal{C}_k| \times r}$ and $\bar{W}_{t+1,k} = \frac{W_{t,k}^\star}{\mathbf{tr}\left(W_{t,k}^\star\right)}$ respectively, where $W_{t,k}^\star$ is the optimal solution of $\gamma_k \bar{W}_{t,k} + P_{t,k}S_kP_{t,k}^\mathsf{T}$ in (16).

C. Computational details

Solving the regularized master problem in (16) is the main computation in Algorithm 1. Therefore, it is crucial to solve the master problem efficiently. We summarize the computation details in Proposition 2. Its proof is provided in the report [24]. For notational simplicity, we define the linear mapping $\mathcal{E}_{\mathcal{C}_k} : \mathbb{S}^n \to \mathbb{S}^{|\mathcal{C}_k|}$ as $\mathcal{E}_{\mathcal{C}_k}(X) = E_{\mathcal{C}_k}XE_{\mathcal{C}_k}^\mathsf{T}$ and $\hat{\mathcal{E}}_{\mathcal{C}_k} : \mathbb{S}^{|\mathcal{C}_k|} \to \mathbb{S}^n$ as $\hat{\mathcal{E}}_{\mathcal{C}_k}(X) = E_{\mathcal{C}_k}^\mathsf{T}XE_{\mathcal{C}_k}$.

Proposition 2: The master problem (16) is equivalent to

$$\begin{aligned} \max_{\substack{W_k \in \hat{\mathcal{W}}_{t,k} \\ y \in \mathbb{R}^m}} \; & \left\langle C - \rho\sum_{k=1}^{p}\hat{\mathcal{E}}_{\mathcal{C}_k}(W_k), \Omega_t \right\rangle + \langle b - \mathcal{A}(\Omega_t), y\rangle \\ & - \frac{1}{2\alpha}\left\|\rho\sum_{k=1}^{p}\hat{\mathcal{E}}_{\mathcal{C}_k}(W_k) + \mathcal{A}^*(y) - C\right\|_{\mathrm{F}}^2. \end{aligned} \tag{20}$$

The optimal solution of X in (16) is recovered by

$$X_{t+1}^\star = \Omega_t + \frac{1}{\alpha}\left(\rho\sum_{k=1}^{p}\hat{\mathcal{E}}_{\mathcal{C}_k}(W_k) + \mathcal{A}^*(y) - C\right). \tag{21}$$

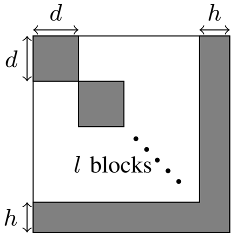

Fig. 1: Block-arrow sparsity pattern [10]: the number of blocks: l; block size: d; the width of the arrow head: h.

D. Convergence guarantees

We present the convergence guarantee for Algorithm 1 when strong duality holds for (9) and (12).

Theorem 2: Suppose strong duality holds for (9) and (12). Given any $\beta \in (0,1)$, $r_{\mathrm{c}} \geq 1$, $r_{\mathrm{p}} \geq 0$, $\alpha > 0$, $r = r_{\mathrm{c}} + r_{\mathrm{p}}$, $\rho > 2\mathcal{D}_{\mathcal{Y}^\star} + 1$, $P_{0,k} \in \mathbb{R}^{n \times r}, \forall 1 \leq k \leq p$, $\Omega_0 \in \mathbb{S}^n$, and target accuracy $\epsilon > 0$, then Algorithm 1 outputs iterates $\left(\Omega_t, \{W_{t,k}^\star\}, y_t^\star\right)$ with

$$G(\Omega_t) - G(X^\star) \leq \epsilon, \tag{22a}$$

$$\left\|\rho\sum_{k=1}^{p}\hat{\mathcal{E}}_{\mathcal{C}_k}\left(W_{t,k}^\star\right) - C + \mathcal{A}^*\left(y_t^\star\right)\right\|_{\mathrm{F}}^2 \leq \epsilon,\; W_{t,k}^\star \succeq 0, \tag{22b}$$

$$\lambda_{\min}\left(E_{\mathcal{C}_k}\Omega_tE_{\mathcal{C}_k}^\mathsf{T}\right) \geq -\epsilon, \quad 1 \leq k \leq p, \tag{22c}$$

$$|\langle C, \Omega_t\rangle - \langle b, y_t^\star\rangle| \leq \sqrt{\epsilon}, \tag{22d}$$

by $t \leq \mathcal{O}\left(1/\epsilon^3\right)$. If the strict complementarity (Definition 1) also holds, then the condition (22) is reached by $t \leq \mathcal{O}(1/\epsilon)$.

Our proof is motivated by [18], [19]; see our report [24] for details.

V. Implementation and Numerical Results

In this section, we present the numerical results of Algorithm 1 to show its efficiency and convergence. All the experiments were executed in MATLAB R2023b on an Ubuntu 22.04 PC 32.0 GB RAM[1]. We consider SDPs with a block-arrow sparsity pattern shown in Figure 1 which has l overlapping maximal cliques of size $d + h$. We randomly generate problem data such that there exists at least one low-rank dual solution; see Section VI-G for further experiment details.

For the implementation, we reformulate the master problem (20) in Algorithm 1 into a quadratic SDP of the form,

$$\begin{aligned} \min_{v} \quad & v^\mathsf{T}Qv + q^\mathsf{T}v + c \\ \text{subject to} \quad & \gamma_k \geq 0, S_k \in \mathbb{S}_+^r, \\ & \gamma_k + \mathbf{tr}(S_k) \leq \rho,\; k = 1, \ldots, p, \end{aligned}$$

where $v = \begin{bmatrix}\gamma_1 & \cdots & \gamma_p & \operatorname{vec}(S_1)^\mathsf{T} & \cdots & \operatorname{vec}(S_p)^\mathsf{T}\end{bmatrix}^\mathsf{T}$, and $\operatorname{vec}(\cdot)$ denotes the vectorization operation, then solve it using MOSEK [7] (See Section VI-F for construction details). The problem above only involves p scalar variables and p small PSD variables, which can be efficiently solved.

[1] Our code and experiments are available at https://github.com/hsmmoj/A-Spectral-Bundle-Method-for-Sparse-SDPs.

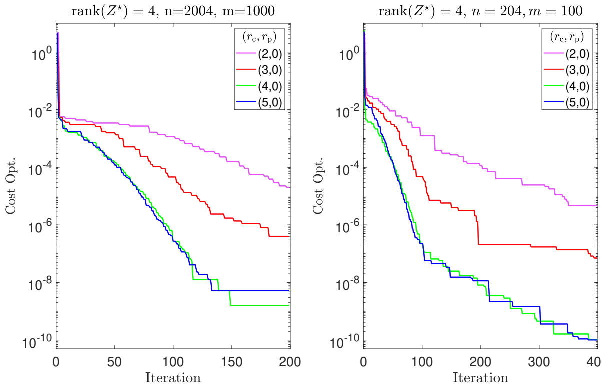

Fig. 2: The relative optimality gap of different choices of (r_c, r_p) for two random SDPs with rank$(X^\star) = 2000$ and rank$(X^\star) = 200$ on the left and right respectively.

TABLE I: Computational results on solving two random SDPs with block-arrow sparsity pattern (we fixed $t_{\max} = 400$ in Algorithm 1).

Dim	(r_p, r_c)	Semi Opt.	Affine Opt.	Dual Gap	Cost Opt.
Small SDP	$(0,2)$	$-0.63e{-2}$	$1.06e{-7}$	$2.08e{-6}$	$4.64e{-6}$
	$(0,3)$	$-3.67e{-4}$	$1.14e{-9}$	$2.36e{-7}$	$7.05e{-8}$
	$(0,4)$	$-3.47e{-8}$	$2.45e{-11}$	$4.54e{-9}$	$1.04e{-10}$
	$(0,5)$	$-1.01e{-10}$	$4.74e{-12}$	$1.61e{-8}$	$1.01e{-10}$
Large SDP	$(0,2)$	$-0.292e{-1}$	$2.05e{-8}$	$9.07e{-8}$	$2.02e{-5}$
	$(0,3)$	$-0.11e{-2}$	$3.83e{-10}$	$1.76e{-8}$	$4.03e{-7}$
	$(0,4)$	$-2.25e{-5}$	$6.30e{-11}$	$9.16e{-11}$	$1.61e{-9}$
	$(0,5)$	$-1.03e{-5}$	$4.17e{-11}$	$9.67e{-10}$	$5.15e{-9}$

We first run Algorithm 1 for two settings:

1) A small-scale problem with dimensions $d = 20, l = 10, h = 4,$ and $m = 100$.
2) A large-scale problem with dimensions $d = 50, l = 40, h = 4,$ and $m = 1000$.

Inspired by [18, Section 5] and [19, Section 6], we choose different configurations of the parameters r_p and r_c. The parameter r_p is fixed to be 0, while different r_c is considered since r_p does not have much influence on the convergence rate as shown in [18, Section 5] and [19, Section 6]. The numerical results are presented in Table I, where ”Semi Opt.”, ”Affine Opt.”, ”Dual Gap”, and ”Cost Opt.” denote the following criteria

$$\lambda_{\min}\left(\Omega_{t+1}\right), \quad \frac{\|C-\mathcal{A}^*(y)-W_t^\star\|_{\mathrm{F}}}{1+\|C\|},$$
$$\frac{|\langle C,\Omega_{t+1}\rangle-\langle b,\omega_t\rangle|}{1+|\langle C,\Omega_{t+1}\rangle|+|\langle b,\omega_t\rangle|}, \quad \frac{|G\left(\Omega_{t+1}\right)-G^\star|}{|G^\star|},$$

with $W_t^\star = \sum_{k=1}^{p} \hat{\mathcal{E}}_{\mathcal{C}_k}(W_{t,k}^\star)$ and $G^\star$ as the true optimal value. In both small-scale and large-scale SDPs, Algorithm 1 returns a solution of high accuracy within 400 iterations.

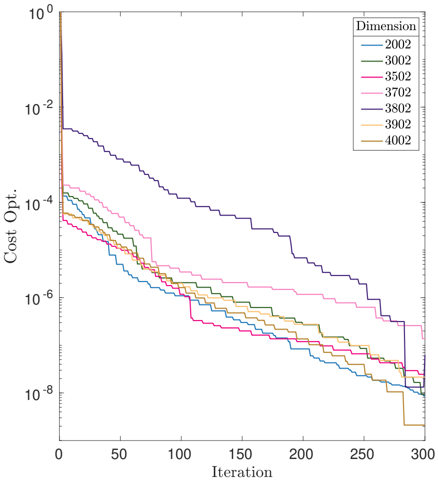

Fig. 3: The relative optimality gap of different problem dimensions with fixed $r_c = 4, r_p = 0$, and $t_{\max} = 300$.

TABLE II: Comparison with different solvers on solving sparse SDPs of different sizes (time in seconds).

	Sedumi [25]		Mosek [7]		SDPNAL+ [26]		Algorithm 1	
Dim	CO	time	CO	time	CO	time	CO	time
4002	$1.7e{-8}$	2243	$9.79e{-12}$	1985	$8.3e{-9}$	887	$2.11e{-9}$	551
3902	$4.8e{-8}$	2035	$1e{-11}$	1790	$2.1e{-8}$	545	$2.14e{-8}$	535
3802	$6.2e{-8}$	2136	$9.8e{-12}$	1557	$6.4e{-8}$	540	$1.2e{-8}$	510
3702	$1.1e{-7}$	1696	$9.09e{-12}$	1510	$3.5e{-7}$	270	$1.37e{-7}$	480
3502	$4.23e{-8}$	1453	$8.7e{-11}$	600	$6.25e{-7}$	310	$2.9e{-7}$	422
3002	$3.21e{-11}$	1206	$8.3e{-11}$	355	$8.7e{-8}$	210	$8.5e{-8}$	360
2002	$1.6e{-9}$	366	$2.3e{-12}$	84	$5.5e{-9}$	108	$6.1e{-9}$	132

CO denotes ”Cost optimality”. We fixed $t_{\max} = 300$ in Algorithm 1.

TABLE III: Peak memory requirement for different solvers on solving sparse SDPs of different sizes (Memory in GB).

Dim	Sedumi	Mosek	SDPNAL+	Algorithm 1
4002	7.7	15.1	2.4	2.5
3902	5.9	13.8	2	2.4
3802	5.8	12.9	1.8	2.2
3702	4.1	12.4	1.7	2.1
3502	5.4	10.8	1.6	1.9
3002	2.7	5.7	1	1.5
2002	1.3	3.8	0.6	0.4

We next conduct experiments comparing our algorithm’s performance with other solves on sparse SDPs, ranging from dimensions 2002 to 4002. The results in Table II and Table III demonstrate our algorithm’s scalability and efficiency to handle large-scale SDPs without excessive resource usage. Our solver outperforms standard interior-point solvers (Sedumi and Mosek) in terms of speed while maintaining comparable accuracy. Also, our preliminary implementation of Algorithm 1 shows comparable (sometimes better) scalability compared to the first-order solver SDPNAL+ [26] on these problem instances.

VI. Conclusions

In this paper, we have developed a new spectral bundle method for sparse SDPs. This approach breaks down a large PSD constraint into several smaller ones using chordal decomposition. We introduce an equivalent non-smooth convex optimization problem by moving the PSD constraints into the objective function. Instead of introducing extra consensus variables as many previous studies [6], [10], we solve the non-smooth problem using a new spectral bundle method, which is shown Algorithm 1. Under a mild condition, the algorithm converges as $\mathcal{O}\left(1/\epsilon^3\right)$. If the problem satisfies strict complementarity, the convergence rate is improved to $\mathcal{O}\left(1/\epsilon\right)$. Our experiments confirm that this new algorithm is promising to efficiently solve large-scale sparse SDPs.

References

[1] F. Alizadeh, "Interior point methods in semidefinite programming with applications to combinatorial optimization," *SIAM Journal on Optimization*, vol. 5, no. 1, pp. 13–51, 1995.

[2] S. Boyd, L. El Ghaoui, E. Feron, and V. Balakrishnan, *Linear matrix inequalities in system and control theory*. SIAM, 1994.

[3] A. Majumdar, G. Hall, and A. Ahmadi, "Recent scalability improvements for semidefinite programming with applications in machine learning, control, and robotics," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 3, pp. 331–360, May 2020.

[4] M. X. Goemans and D. P. Williamson, "Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming," *Journal of the ACM (JACM)*, vol. 42, no. 6, pp. 1115–1145, 1995.

[5] L. Vandenberghe and S. Boyd, "Semidefinite programming," *SIAM Review*, vol. 38, no. 1, pp. 49–95, 1996.

[6] Y. Zheng, G. Fantuzzi, and A. Papachristodoulou, "Chordal and factor-width decompositions for scalable semidefinite and polynomial optimization," *Annual Reviews in Control*, vol. 52, pp. 243–279, 2021.

[7] M. ApS, *The MOSEK optimization toolbox for MATLAB manual. Version 9.0.*, 2019.

[8] Y. Zheng, *Chordal sparsity in control and optimization of large-scale systems*. PhD thesis, University of Oxford, 2019.

[9] B. O'donoghue, E. Chu, N. Parikh, and S. Boyd, "Conic optimization via operator splitting and homogeneous self-dual embedding," *J. Optim. Theory Appl.*, vol. 169, p. 1042–1068, jun 2016.

[10] Y. Zheng, G. Fantuzzi, A. Papachristodoulou, P. Goulart, and A. Wynn, "Chordal decomposition in operator-splitting methods for sparse semidefinite programs," *Mathematical Programming*, vol. 180, no. 1-2, pp. 489–532, 2020.

[11] A. A. Ahmadi and A. Majumdar, "Dsos and sdsos optimization: more tractable alternatives to sum of squares and semidefinite optimization," *SIAM Journal on Applied Algebra and Geometry*, vol. 3, no. 2, pp. 193–230, 2019.

[12] J. Miller, Y. Zheng, M. Sznaier, and A. Papachristodoulou, "Decomposed structured subsets for semidefinite and sum-of-squares optimization," *Automatica*, vol. 137, p. 110125, 2022.

[13] J. Agler, W. Helton, S. McCullough, and L. Rodman, "Positive semidefinite matrices with a given sparsity pattern," *Linear Algebra and its Applications*, vol. 107, pp. 101–149, 1988.

[14] M. Fukuda, M. Kojima, K. Murota, and K. Nakata, "Exploiting sparsity in semidefinite programming via matrix completion i: General framework," *SIAM Journal on Optimization*, vol. 11, no. 3, pp. 647–674, 2001.

[15] L. Vandenberghe and M. S. Andersen, "Chordal graphs and semidefinite optimization," *Foundations and Trends® in Optimization*, vol. 1, no. 4, pp. 241–433, 2015.

[16] Y. Zheng and G. Fantuzzi, "Sum-of-squares chordal decomposition of polynomial matrix inequalities," *Mathematical Programming*, vol. 197, no. 1, pp. 71–108, 2023.

[17] C. Helmberg and F. Rendl, "A spectral bundle method for semidefinite programming," *SIAM Journal on Optimization*, vol. 10, no. 3, pp. 673–696, 2000.

[18] L. Ding and B. Grimmer, "Revisiting spectral bundle methods: Primal-dual (sub) linear convergence rates," *SIAM Journal on Optimization*, vol. 33, no. 2, pp. 1305–1332, 2023.

[19] F.-Y. Liao, L. Ding, and Y. Zheng, "An overview and comparison of spectral bundle methods for primal and dual semidefinite programs," *arXiv preprint arXiv:2307.07651*, 2023.

[20] C. Lemaréchal and J. Zowe, *A Condensed Introduction to Bundle Methods in Nonsmooth Optimization*, pp. 357–382. Dordrecht: Springer Netherlands, 1994.

[21] M. Díaz and B. Grimmer, "Optimal convergence rates for the proximal bundle method," *SIAM Journal on Optimization*, vol. 33, no. 2, pp. 424–454, 2023.

[22] A. Ruszczynski, *Nonlinear Optimization*. Princeton University Press, 2011.

[23] Y. Sun, M. S. Andersen, and L. Vandenberghe, "Decomposition in conic optimization with partially separable structure," *SIAM Journal on Optimization*, vol. 24, no. 2, pp. 873–897, 2014.

[24] M. Hesam, F.-Y. Liao, and Y. Zheng, "A spectral bundle method for sparse semidefinite programs." Technical report `https://hsmmoj.github.io/files/SpecBM-SDPs.pdf`, 2023.

[25] J. F. Sturm, "Using sedumi 1.02, a MATLAB toolbox for optimization over symmetric cones," *Optimization Methods and Software*, vol. 11, no. 1-4, pp. 625–653, 1999.

[26] D. Sun, K.-C. Toh, Y. Yuan, and X.-Y. Zhao, "Sdpnal+: A matlab software for semidefinite programming with bound constraints (version 1.0)," 2019.

[27] M. L. Overton, "Large-scale optimization of eigenvalues," *SIAM Journal on Optimization*, vol. 2, no. 1, pp. 88–120, 1992.

[28] R. T. Rockafellar, *Convex analysis*, vol. 18. Princeton university press, 1970.

[29] K. C. Kiwiel, "Efficiency of proximal bundle methods," *Journal of Optimization Theory and Applications*, vol. 104, no. 3, pp. 589–603, 2000.

[30] D. Drusvyatskiy, H. Wolkowicz, *et al.*, "The many faces of degeneracy in conic optimization," *Foundations and Trends® in Optimization*, vol. 3, no. 2, pp. 77–170, 2017.

[31] J. F. Sturm, "Error bounds for linear matrix inequalities," *SIAM Journal on Optimization*, vol. 10, no. 4, pp. 1228–1248, 2000.

[32] F. Alizadeh, J.-P. A. Haeberly, and M. L. Overton, "Complementarity and nondegeneracy in semidefinite programming," *Mathematical programming*, vol. 77, no. 1, pp. 111–128, 1997.

Appendix

A. *Proof of Proposition 1*

Before proving Proposition 1, we first review two technical lemmas to facilitate the proof.

Theorem 3 ([22, Theorem 7.21]): Suppose that problem (6) satisfies Slater's constraint qualification condition. Then, there exists a constant $\rho_0 \geq 0$ such that for all $\rho > \rho_0$, a point $\hat{x}$ is an optimal solution of problem (6) if and only if it is an optimal solution of problem (7). In particular, we can choose $\rho_0 = \sup_{\mu \in M} \|\mu\|_\infty$, where $M \subseteq \mathbb{R}^m$ is the set of Lagrange multipliers associated with $g_i(x) \leq 0, i = 1, \ldots, m$.

Lemma 1 ([27, Theorem 2]): Given a symmetric matrix $A \in \mathbb{S}^n$. Suppose its maximal eigenvalue $\lambda_{\max}(A)$ has multiplicity t. Then, we have

$$\partial \lambda_{\max}(A) = \left\{ QUQ^\mathsf{T} \mid U \in \mathbb{S}_+^t, \mathbf{tr}(U) = 1 \right\},$$

where the columns of $Q \in \mathbb{R}^{n \times t}$ forms an orthonormal set of the eigenvectors for $\lambda_{\max}(A)$.

We are ready to prove Proposition (1) by applying Theorem (3). Let $M \subseteq \mathbb{R}^p$ be the set of all Lagrange multipliers associated with the inequality constraints $\lambda_{\max}\left(E_{\mathcal{C}_k}(-X^\star)E_{\mathcal{C}_k}^\mathsf{T}\right) \leq 0, k = 1, \ldots, p$, in (10). Consider $\bar{M} = \left\{(\mathrm{tr}(Y_1^\star), \mathrm{tr}(Y_2^\star), \ldots, \mathrm{tr}(Y_p^\star)) \mid (y^\star, Y_k^\star) \in \mathcal{D}^\star\right\}$, where $\mathcal{D}^\star$ is the optimal solution set in (12). It is sufficient to show $M = \bar{M}$. We start the proof by denoting the KKT optimally condition for (10):

$$0 \in C + \sum_{k=1}^{p} \alpha_k \partial \left(\lambda_{\max}\left(E_{\mathcal{C}_k}(-X)E_{\mathcal{C}_k}^\mathsf{T}\right)\right) - \sum_{i=1}^{m} A_i y_i. \quad \text{(23a)}$$

$$\alpha_k \lambda_{\max}\left(E_{\mathcal{C}_k}(-X)E_{\mathcal{C}_k}^\mathsf{T}\right) = 0, \alpha_k \geq 0, k = 1, \ldots, p. \quad \text{(23b)}$$

$$\lambda_{\max}\left(E_{\mathcal{C}_k}(-X)E_{\mathcal{C}_k}^\mathsf{T}\right) \leq 0, k = 1, \ldots, p. \quad \text{(23c)}$$

$$\langle A_i, X\rangle = b_i, i = 1, \ldots, m. \quad \text{(23d)}$$

The proof has two directions. We will show $\bar{M} \subseteq M$, and $M \subseteq \bar{M}$ respectively.

- First, we prove $\bar{M} \subseteq M$. For any primal-dual optimal solution for (9) and (12) denoted by $X^\star$ and $(y^\star, Y_k^\star)$ respectively, it suffices to show $\alpha_k = \mathbf{tr}(Y_k^\star)$ satisfies (23). It is clear that $X^\star$ satisfies (23c) and (23d). Therefore, it remains to prove (23a) and (23b) are also satisfied. For index k such that $\lambda_{\max}\left(E_{\mathcal{C}_k}(-X^\star)E_{\mathcal{C}_k}^\mathsf{T}\right) = 0$, it is trivial that (23b) is satisfied. For index k such that $\lambda_{\max}\left(E_{\mathcal{C}_k}(-X^\star)E_{\mathcal{C}_k}^\mathsf{T}\right) < 0$, we know $Y_k^\star = 0$ due to the complementary slackness $\left(E_{\mathcal{C}_k} X^\star E_{\mathcal{C}_k}^\mathsf{T}\right) Y_k^\star = 0$. Choosing $\alpha_k = \mathbf{tr}(Y_k^\star) = 0$ also satisfies (23b).

  We then move on to prove (23a). From complementary slackness

  $$\begin{aligned}\left(E_{\mathcal{C}_k} X^\star E_{\mathcal{C}_k}^\mathsf{T}\right) Y_k^\star &= Y_k^\star \left(E_{\mathcal{C}_k} X^\star E_{\mathcal{C}_k}^\mathsf{T}\right) \\ &= X^\star \left(E_{\mathcal{C}_k}^\mathsf{T} Y_k^\star E_{\mathcal{C}_k}\right) = 0,\end{aligned}$$

  it is easy to see that

  $$\begin{aligned}\mathrm{range}\left(E_{\mathcal{C}_k}^\mathsf{T} Y_k^\star E_{\mathcal{C}_k}\right) &\subseteq \mathrm{null}\left(X^\star\right), \\ \mathrm{range}\left(E_{\mathcal{C}_k} X^\star E_{\mathcal{C}_k}^\mathsf{T}\right) &\subseteq \mathrm{null}\left(Y_k^\star\right).\end{aligned}$$

  From Lemma 1, we know that

  $$\begin{aligned}&\partial\left(\lambda_{\max}\left(E_{\mathcal{C}_k}(-X^\star)E_{\mathcal{C}_k}^\mathsf{T}\right)\right) \\ &\qquad \in \left\{E_{\mathcal{C}_k}^\mathsf{T} Q U Q^\mathsf{T} E_{\mathcal{C}_k} \mid U \in \mathbb{S}_+^t, \mathbf{tr}(U) = 1\right\},\end{aligned}$$

  where $Q \in \mathbb{R}^{|\mathcal{C}_k| \times t}$ is the orthonormal eigenvectors matrix corresponding to $\lambda_{\max}\left(E_{\mathcal{C}_k}(-X^\star)E_{\mathcal{C}_k}^\mathsf{T}\right)$ (or $\lambda_{\min}\left(E_{\mathcal{C}_k}(X^\star)E_{\mathcal{C}_k}^\mathsf{T}\right)$ equivalently) with t multiplicities. Therefore, it follows that

  $$\begin{aligned}\sum_{i=1}^{p} E_{\mathcal{C}_k}^\mathsf{T} Y_k^\star E_{\mathcal{C}_k} &= C - \sum_{i=1}^{m} A_i y_i^\star \\ &\in -\sum_{k=1}^{p} \mathbf{tr}(Y_k^\star) \partial\left(\lambda_{\max}\left(E_{\mathcal{C}_k}(-X^\star)E_{\mathcal{C}_k}^\mathsf{T}\right)\right).\end{aligned}$$

  Hence, choosing $\alpha_k = \mathbf{tr}(Y_k^\star)$ ensures (23a) is satisfied.
- Second, we prove $M \subseteq \bar{M}$. Note that it is equivalent to prove $\forall \{\alpha_k\} \notin \bar{M}$, we can not find $(X^\star, y^\star)$ that satisfies (23). This is obvious since

  $$\begin{aligned}\sum_{i=1}^{p} E_{\mathcal{C}_k}^\mathsf{T} Y_k^\star E_{\mathcal{C}_k} &= C - \sum_{i=1}^{m} A_i y_i^\star \\ &\notin -\sum_{k=1}^{p} \alpha_k \partial\left(\lambda_{\max}\left(E_{\mathcal{C}_k}(-X^\star)E_{\mathcal{C}_k}^\mathsf{T}\right)\right),\end{aligned}$$

  if $\{\alpha_k\} \notin M$, which completes the proof.

B. *Proof of Proposition 2*

We rewrite the master problem (16) as

$$\min_{X \in \mathcal{X}_0} \max_{W_k \in \hat{\mathcal{W}}_{t,k}} \langle C, X\rangle - \rho \sum_{k=1}^{p} \langle \mathcal{E}_{\mathcal{C}_k}(X), W_k\rangle + \frac{\alpha}{2} \|X - \Omega_t\|_\mathrm{F}^2.$$

Using the fact that $\hat{\mathcal{W}}_{t,k}$ is bounded and strong duality holds for the above problem, we can switch the order of the maximization and minimization [28, Corollary 37.3.2] and recast the problem as

$$\max_{W_k \in \hat{\mathcal{W}}_{t,k}} \min_{X \in \mathcal{X}_0} \langle C, X\rangle - \rho \sum_{k=1}^{p} \langle \mathcal{E}_{\mathcal{C}_k}(X), W_k\rangle + \frac{\alpha}{2} \|X - \Omega_t\|_\mathrm{F}^2.$$

Since the above objective function is a strongly convex function in terms of X, strong duality also holds for the inner minimization. We can derive the dual of the inner problem to remove the affine constraints. To derive the dual problem, we introduce a dual variable $y \in \mathbb{R}^m$ and form the Lagrangian function

$$\begin{aligned}L(X, y) :=& \langle C, X\rangle - \rho \sum_{k=1}^{p} \langle \mathcal{E}_{\mathcal{C}_k}(X), W_k\rangle + \frac{\alpha}{2} \|X - \Omega_t\|_\mathrm{F}^2 \\ &+ y^\mathsf{T}\left(b - \mathcal{A}(X)\right).\end{aligned}$$

The dual function is defined as $g(y) := \min_X L(X, y)$ with the unique minimizer

$$X^\star = \Omega_t + \frac{1}{\alpha}\left(\rho \sum_{i=1}^{p} \hat{\mathcal{E}}_{\mathcal{C}_k}(W_k) + \mathcal{A}^*(y) - C\right).$$

By substituting $X^\star$ in the Lagrangian function, we compute the dual function

$$\begin{aligned} g(y) = &\left\langle C - \rho \sum_{i=1}^{p} \hat{\mathcal{E}}_{\mathcal{C}_k}(W_k), \Omega_t \right\rangle + \langle b - \mathcal{A}(\Omega_t), y\rangle \\ &- \frac{1}{2\alpha}\left\| \rho \sum_{i=1}^{p} \hat{\mathcal{E}}_{\mathcal{C}_k}(W_k) + \mathcal{A}^*(y) - C \right\|_{\mathrm{F}}^2 . \end{aligned}$$

Therefore, the master problem (16) can be simplified as

$$\begin{aligned} \max_{W_k \in \hat{\mathcal{W}}_{t,k}} \max_{y \in \mathbb{R}^m} &\left\langle C - \rho \sum_{i=1}^{p} \hat{\mathcal{E}}_{\mathcal{C}_k}(W_k), \Omega_t \right\rangle + \langle b - \mathcal{A}(\Omega_t), y\rangle \\ &- \frac{1}{2\alpha}\left\| \rho \sum_{i=1}^{p} \hat{\mathcal{E}}_{\mathcal{C}_k}(W_k) + \mathcal{A}^*(y) - C \right\|_{\mathrm{F}}^2 , \end{aligned}$$

which completes the proof.

C. *Proof of Theorem 2*

One main step for the proof of Theorem 2 is to connect the primal feasibility, dual feasibility, and primal-dual optimality to the cost value gap. We summarize those connections in the following lemma.

Lemma 2: Under the parameters in Theorem 2, at each descent step $t > 0$, the following results hold.

- The approximate dual feasibility satisfies
$$\begin{aligned} &\left\| \rho \sum_{k=1}^{p} \hat{\mathcal{E}}_{\mathcal{C}_k}(W_{t,k}^\star) - C + \mathcal{A}^*(y_t^\star) \right\|_{\mathrm{F}}^2 \\ &\leq \frac{2\alpha}{\beta}(G(\Omega_t) - G(X^\star)). \end{aligned}$$
- The approximate primal feasibility satisfies
$$\begin{aligned} &\lambda_{\min}(\mathcal{E}_{\mathcal{C}_k}(\Omega_{t+1})) \geq \frac{-(G(\Omega_t) - G(X^\star))}{\mathcal{D}_{\mathcal{Y}^\star} + 1}, 1 \leq k \leq p, \\ &\text{and} \quad \mathcal{A}(\Omega_{t+1}) = b. \end{aligned}$$
- The approximate primal-dual optimality satisfies
$$\begin{aligned} \langle C, \Omega_{t+1}\rangle - \langle b, y_t^\star\rangle \geq &\left(-(G(\Omega_t) - G(X^\star))\frac{p\rho}{\mathcal{D}_{\mathcal{Y}^\star} + 1} \right) \\ &- \mathcal{D}_{\Omega_0}\sqrt{\frac{2\alpha}{\beta}(F(\Omega_t) - F(X^\star))}, \\ \langle C, \Omega_{t+1}\rangle - \langle b, y_t^\star\rangle \leq &\left(\frac{1-\beta}{\beta}(G(\Omega_t) - G(X^\star)) \right) \\ &+ \mathcal{D}_{\Omega_0}\sqrt{\frac{2\alpha}{\beta}(G(\Omega_t) - G(X^\star))}, \end{aligned}$$

where $\mathcal{D}_{\Omega_0} = \sup_{G(\Omega_t) \leq G(\Omega_0)} \|\Omega_t\|_{\mathrm{F}}$ is maximum norm value over the sublevel set which is bounded.

Lemma 3 (quadratic growth): Under **??** and the selection of parameters in Theorem 2, for any fixed $\epsilon > 0$ and X in the sub-level set $\mathcal{S}_\epsilon := \{X \in \mathbb{S}^n \mid G(X) \leq G(X^\star) + \epsilon, \mathcal{A}(X) = b, \|X\|_{\mathrm{F}} < \infty\}$, there exist some constants $\zeta \geq 1$ and $\mu > 0$ such that

$$G(X) - G(X^\star) \geq \mu \cdot \operatorname{dist}^\zeta(X, \mathcal{P}^\star).$$

Furthermore, if SDPs (9) and (12) satisfy strict complementarity, the exponent term $\zeta = 2$.

The proofs for Lemma 2 and 3 are provided the in the report [24]. With Lemmas 2 and 3, we are ready to prove Theorem 2 by utilizing convergence results in [21], [29].

Proof of Theorem 2. It remains to verify that Algorithm 1 satisfy (5a) to (5c). For notational convenience, we use $\hat{G}_t(X)$ to denote the approximation model $\hat{G}_{\{\bar{W}_{t,k}, P_{t,k}\}}(X)$ at iteration t. First, by the construction of $P_{t,k}$ and $\bar{W}_{t,k}$ in (18) and (19), we have $(P_{t,k})^{\mathsf{T}} P_{t,k} = I$, $\bar{W}_{t,k} \succeq 0$, and $\mathbf{tr}(\bar{W}_{t,k}) = 1$. It follows that

$$\hat{\mathcal{W}}_{t,k} \subset \{W \in \mathbb{S}_+^{|\mathcal{C}_k|} \mid \mathbf{tr}(W) \leq 1\},$$

which implies

$$\begin{aligned} &\max_{W_k \in \hat{\mathcal{W}}_{t,k}} \langle W_k, \mathcal{E}_{\mathcal{C}_k}(-X)\rangle \\ \leq &\max_{\mathbf{tr}(W) \leq 1, W \in \mathbb{S}_+^{|\mathcal{C}_k|}} \langle W, \mathcal{E}_{\mathcal{C}_k}(-X)\rangle \\ = &\max\{\lambda_{\max}(\mathcal{E}_{\mathcal{C}_k}(X)), 0\}, \forall X \in \mathbb{S}^n, 1 \leq k \leq p. \end{aligned}$$

Therefore, the global lower-bound property (5a) is satisfied.

Second, since the update in (18) includes a subgradient information of $\lambda_{\max}(\mathcal{E}_{\mathcal{C}_k}(-X_{t+1}^\star))$, there exists a normalized vector $e_k \in \mathbb{R}^r$ such that $P_{t+1,k} e_k = v_k$, where v_k is a subgradient of $\lambda_{\max}(\mathcal{E}_{\mathcal{C}_k}(-X_{t+1}^\star))$. If $\lambda_{\max}(\mathcal{E}_{\mathcal{C}_k}(-X_{t+1}^\star)) > 0$, we let $\gamma_{t+1,k} = 0$ and $S_{t+1,k} = e_k(e_k)^{\mathsf{T}}$, otherwise, we let $\gamma_{t+1,k} = 0$ and $S_{t+1,k} = 0$. It follows that

$$\begin{aligned} &\hat{G}_{t+1}(X) \\ &\geq \langle C, X\rangle + \rho \sum_{k \in \mathcal{I}} \left\langle P_{t+1,k}(e_k(e_k)^{\mathsf{T}})P_{t+1,k}^{\mathsf{T}}, \mathcal{E}_{\mathcal{C}_k}(-X)\right\rangle \\ &= \langle C, X\rangle + \rho \sum_{k \in \mathcal{I}} \left\langle v_k v_k^{\mathsf{T}}, \mathcal{E}_{\mathcal{C}_k}(-X)\right\rangle \\ &= \langle C, X_{t+1}^\star\rangle + \rho \sum_{k \in \mathcal{I}} \lambda_{\max}(\mathcal{E}_{\mathcal{C}_k}(-X_{t+1}^\star)) \\ &\quad + \langle C - \rho \sum_{k \in \mathcal{I}} \hat{\mathcal{E}}_{\mathcal{C}_k}(v_k v_k^{\mathsf{T}}), X - X_{t+1}^\star\rangle \\ &= G(X_{t+1}^\star) + \langle g_{t+1}, X - X_{t+1}^\star\rangle, \quad \forall X \in \mathcal{X}_0, \end{aligned}$$

where $g_{t+1} = C - \rho \sum_{k \in \mathcal{I}} \hat{\mathcal{E}}_{\mathcal{C}_k}(v_k v_k^{\mathsf{T}}) \in \partial G(X_{t+1}^\star)$ and $\mathcal{I} = \{k \mid \lambda_{\max}\left(\mathcal{E}_{\mathcal{C}_k}(-X_{t+1}^\star)\right) > 0, 1 \leq k \leq p\}$, which verified the second property (5b).

Third, we first argue that the optimal solution at step t is within the feasible region of the next iteration. Recalling the update procedure of the matrix $P_{t+1,k}$ in (18), there exists matrices $\bar{Q}_k \in \mathbb{R}^{|\mathcal{C}_k| \times r}$ with normalized columns such that $P_{t+1,k}\bar{Q}_k = P_{t,k}Q_{1,k}$. By letting $\gamma_{t+1} = \gamma_{t,k}^\star + \mathbf{tr}(\Sigma_{k,2})$ and $S_{t+1,k} = \bar{Q}_k \Sigma_{1,k} \bar{Q}_k^{\mathsf{T}}$, we recover the solution

$$\begin{aligned} W_{t,k}^\star &= \gamma_{t+1}\bar{W}_{t+1,k} + P_{t+1,k} S_{t+1,k} P_{t+1,k}^{\mathsf{T}} \\ &= (\gamma_{t,k}^\star + \mathbf{tr}(\Sigma_{k,2}))\bar{W}_{t,k} + P_{t+1,k}\bar{Q}\Sigma_{1,k}\bar{Q}^{\mathsf{T}} P_{t+1,k}^{\mathsf{T}} \\ &= (\gamma_{t,k}^\star + \mathbf{tr}(\Sigma_{k,2}))\bar{W}_{t,k} + P_{t,k} Q_{1,k}\Sigma_{1,k} Q_{1,k}^{\mathsf{T}} P_{t,k}^{\mathsf{T}}. \end{aligned}$$

Hence, it follows that

$$\begin{aligned}
&\hat{G}_{t+1}(X)\\
&\geq \langle C, X\rangle + \rho\sum_{k=1}^{p}\left\langle W^\star_{t,k}, \mathcal{E}_{\mathcal{C}_k}(-X)\right\rangle\\
&= \left\langle C - \rho\sum_{k=1}^{p} \hat{\mathcal{E}}_{\mathcal{C}_k}(W^\star_{t,k}), X^\star_{t+1}\right\rangle\\
&\quad + \left\langle C - \rho\sum_{k=1}^{p} \hat{\mathcal{E}}_{\mathcal{C}_k}(W^\star_{t,k}) + \mathcal{A}^*(y^\star_t), X - X^\star_{t+1}\right\rangle\\
&= \hat{G}_t(X^\star_{t+1}) + \langle \alpha(X^\star_{t+1} - \Omega_t), X - X^\star_{t+1}\rangle, \forall X \in \mathcal{X}_0,
\end{aligned}$$

where the first equality uses the fact that $\langle \mathcal{A}^*(y^\star_t), X - X^\star_{t+1}\rangle = 0$, and the second equality uses the optimality condition (21). The normal cone of an affine set $\mathcal{X}_0$ is $\mathcal{N}_{\mathcal{X}_0}(X^\star_{t+1}) = \{\mathcal{A}^*(y) \mid y \in \mathbb{R}^m\}$. Hence, we have proved $\alpha(X^\star_{t+1} - \Omega_t) \in \partial \hat{G}_t(X^\star_{t+1}) + \mathcal{N}_{\mathcal{X}_0}(X^\star_{t+1})$ and the third property (5c) is verified. Finally, using the fact that $\max\{0, x\}$ is 1-Lipschitz can show the function $G(X)$ is Lipschitz continuous. By the results in [21] and [19], the convergence rate is $\mathcal{O}(1/\epsilon^3)$ as the objective function is Lipschitz continuous and improved to $\mathcal{O}(1/\epsilon)$ when quadratic growth Lemma 3 is satisfied.

D. Proof of Lemma 2

For notational convenience, we denote $\hat{G}_t$ as the approximation model $\hat{G}_{\{\bar{W}_{t,k}, P_{t,k}\}}$ at iteration t. First, from the definition of the descent step and $X^\star$ being the minimizer, we know that

$$G(\Omega_t) - G(X^\star) \geq \beta\left(G(\Omega_t) - \hat{G}_t(X^\star_{t+1})\right). \tag{24}$$

We also note that $X^\star_{t+1}$ minimizes the master problem

$$\hat{G}_t(X^\star_{t+1}) + \frac{\alpha}{2}\|X^\star - \Omega_t\|_{\mathrm{F}}^2 \leq \hat{G}_t(\Omega_t) \leq G(\Omega_t),$$

which implies

$$\begin{aligned}
&\left(G(\Omega_t) - \hat{G}_t(X^\star_{t+1})\right)\\
&\geq \frac{\alpha}{2}\|X^\star - \Omega_t\|_{\mathrm{F}}^2\\
&= \frac{1}{2\alpha}\left\|\rho\sum_{k=1}^{p}\hat{\mathcal{E}}_{\mathcal{C}_k}(W^\star_{t,C_k}) + \mathcal{A}^*(y) - C\right\|_{\mathrm{F}}^2,
\end{aligned}$$

where the last equality uses the optimality condition in (21) and the construction $W^\star_{t,C_k} = \gamma^\star_{t,k}\bar{W}_{t,k} + P_{t,k}S^\star_{t,k}P_{t,k}^{\mathsf{T}}$. Combining (24) with the above inequality yields

$$G(\Omega_t) - G(X^\star) \geq \frac{\beta}{2\alpha}\left\|\rho\sum_{k=1}^{p}\hat{\mathcal{E}}_{\mathcal{C}_k}(W^\star_{t,C_k}) + \mathcal{A}^*(y) - C\right\|_{\mathrm{F}}^2,$$

which finishes the proof for the approximate dual feasibility.

Second, by the fact that a descent step implies a drop in the cost value, we have

$$\begin{aligned}
&G(\Omega_t) - G(X^\star)\\
&\geq G(\Omega_{t+1}) - G(X^\star)\\
&= \langle C, \Omega_t - X^\star\rangle + \rho\sum_{i=1}^{p}\max\left\{0, \lambda_{\max}\left(\mathcal{E}_{\mathcal{C}_k}(-\Omega_t)\right)\right\}\\
&= \langle C, \Omega_t - X^\star\rangle - \rho\sum_{i=1}^{p}\min\left\{0, \lambda_{\min}\left(\mathcal{E}_{\mathcal{C}_k}(\Omega_t)\right)\right\}.
\end{aligned}$$

The first term can be bounded by strong duality

$$\begin{aligned}
\langle C, \Omega_t - X^\star\rangle &= \left\langle \mathcal{A}^*(y^\star) + \sum_{k=1}^{p}\hat{\mathcal{E}}_{\mathcal{C}_k}(Y^\star_k), \Omega_t - X^\star\right\rangle\\
&= \sum_{k=1}^{p}\langle Y^\star_k, \mathcal{E}_{\mathcal{C}_k}(\Omega_t)\rangle\\
&\geq \sum_{k=1}^{p}\|Y^\star_k\|_* \min\{0, \lambda_{\min}(\mathcal{E}_{\mathcal{C}_k}(\Omega_t))\}\\
&\geq \sum_{k=1}^{p}\mathcal{D}_{\mathcal{Y}^\star}\min\{0, \lambda_{\min}(\mathcal{E}_{\mathcal{C}_k}(\Omega_t))\}
\end{aligned}$$

where the first equality comes from strong duality, the second uses the definition of the adjoint operator, i.e., $\langle \mathcal{A}^*(y), X\rangle = \langle \mathcal{A}(X), y\rangle$, the fact that Ω_t and $X^\star$ both satisfy the affine constraints, complementarity slackness $\langle Y^\star_k, \mathcal{E}_{\mathcal{C}_k}(X^\star)\rangle = 0$, and cyclic property of trace operation. Therefore,

$$\begin{aligned}
&G(\Omega_t) - G(X^\star)\\
&\geq \sum_{k=1}^{p}(\mathcal{D}_{\mathcal{Y}^\star} - \rho)\min\{0, \lambda_{\min}(\mathcal{E}_{\mathcal{C}_k}(\Omega_t)\}\\
&\geq (\mathcal{D}_{\mathcal{Y}^\star} - \rho)\min\{0, \lambda_{\min}\left(\mathcal{E}_{\mathcal{C}_k}(\Omega_t)\right)\}, \forall 1 \leq k \leq p,
\end{aligned}$$

where the last inequality uses the fact that $(\mathcal{D}_{\mathcal{Y}^\star} - \rho)\min\{0, \lambda_{\min}\left(\mathcal{E}_{\mathcal{C}_k}(\Omega_t)\right)\} \geq 0, \forall 1 \leq k \leq p$. Along with the assumption $\rho > (2\mathcal{D}_{\mathcal{Y}^\star} + 1)$, this implies

$$\lambda_{\min}\left(\mathcal{E}_{\mathcal{C}_k}(\Omega_t)\right) \geq \frac{-(G(\Omega_t) - G(X^\star))}{(\mathcal{D}_{\mathcal{Y}^\star} + 1)}, \forall 1 \leq k \leq p. \tag{25}$$

This completes the proof for approximate primal feasibility.

Third, by the feasibility of Ω_{t+1}, the duality gap follows

$$\begin{aligned}
&\langle C, \Omega_{t+1}\rangle - \langle b, y^\star_t\rangle\\
=&\langle C, \Omega_{t+1}\rangle - \langle \mathcal{A}(\Omega_{t+1}), y^\star_t\rangle\\
=&\langle C, \Omega_{t+1}\rangle - \langle \mathcal{A}^*(y^\star_t), \Omega_{t+1}\rangle\\
=&\left\langle \rho\sum_{k=1}^{p}\hat{\mathcal{E}}_{\mathcal{C}_k}(W^\star_{t,k}), \Omega_{t+1}\right\rangle\\
&+\left\langle C - \mathcal{A}^*(y^\star_t) - \rho\sum_{k=1}^{p}\hat{\mathcal{E}}_{\mathcal{C}_k}(W^\star_{t,k}), \Omega_{t+1}\right\rangle.
\end{aligned}$$

The lower bound of the first term follows

$$\begin{aligned}
&\left\langle \sum_{k=1}^{p} \hat{\mathcal{E}}_{\mathcal{C}_k}(W^\star_{t,k}), \Omega_{t+1} \right\rangle \\
\geq& \sum_{k=1}^{p} \|W^\star_{t,k}\|_* \lambda_{\min}\left(\mathcal{E}_{\mathcal{C}_k}(\Omega_{t+1})\right) \\
\geq& \sum_{k=1}^{p} \|W^\star_{t,k}\|_* \frac{-(G(\Omega_t) - G(X^\star))}{\mathcal{D}_{\mathcal{Y}^\star} + 1} \\
\geq& (G(\Omega_t) - G(X^\star)) \sum_{k=1}^{p} \frac{-\rho}{\mathcal{D}_{\mathcal{Y}^\star} + 1} \\
=& -(G(\Omega_t) - G(X^\star)) \frac{p\rho}{\mathcal{D}_{\mathcal{Y}^\star} + 1},
\end{aligned}$$

where the second inequality uses the cyclic property of trace operation and (25) and the third inequality is due to the construction $W^\star_{t,\mathcal{C}_k} \leq \rho$ (14). The upper bound of the first term follows

$$\begin{aligned}
&\left\langle \rho \sum_{k=1}^{p} \hat{\mathcal{E}}_{\mathcal{C}_k}(W^\star_{t,k}), \Omega_{t+1} \right\rangle \\
\leq& \left\langle \rho \sum_{k=1}^{p} \hat{\mathcal{E}}_{\mathcal{C}_k}(W^\star_{t,\mathcal{C}_k}), \Omega_{t+1} \right\rangle \\
&+ \rho \sum_{k=1}^{p} \max\left\{0, \lambda_{\max}(\mathcal{E}_{\mathcal{C}_k}(-\Omega_{t+1}))\right\} \\
=& G(\Omega_{t+1}) - \hat{G}_t(\Omega_{t+1}) \\
\leq& \frac{1-\beta}{\beta}(G(\Omega_t) - G(X^\star)),
\end{aligned}$$

where the last inequality comes from a reformulation of the definition of the descent step (17).

The second term can be bounded by Cauchy inequality

$$\begin{aligned}
&\left| \left\langle C - \mathcal{A}^*(y^\star_t) - \rho \sum_{k=1}^{p} \hat{\mathcal{E}}_{\mathcal{C}_k}(W^\star_{t,\mathcal{C}_k}), \Omega_{t+1} \right\rangle \right| \\
\leq& \left\| C - \mathcal{A}^*(y^\star_t) - \rho \sum_{k=1}^{p} \hat{\mathcal{E}}_{\mathcal{C}_k}(W^\star_{t,\mathcal{C}_k}) \right\|_{\mathrm{F}} \|\Omega_{t+1}\|_{\mathrm{F}} \\
\leq& \mathcal{D}_{\Omega_0} \sqrt{\frac{2\alpha}{\beta}(G(\Omega_t) - G(X^\star))}.
\end{aligned}$$

Combing the lower and upper bounds for both terms finishes the proof for the approximate primal-dual optimality.

E. *Proof of Lemma 3*

We first note that the optimal solution set of (9) and (11) is the same and is compact by **??**. The optimal solution set $\mathcal{P}^\star$ can be considered as $\mathcal{M} \cap \{X \in \mathbb{S}^n \mid \mathcal{A}(X) = b, \langle C, X \rangle = p^\star\}$, where $\mathcal{M} = \{X \in \mathbb{S}^n \mid E_{\mathcal{C}_k} X E_{\mathcal{C}_k} \in \mathbb{S}^{|\mathcal{C}_k|}_+, 1 \leq k \leq p\}$. Let $\mathcal{L} := \{X \in \mathbb{S}^n \mid \langle C, X \rangle = p^\star\}$. The result in [30, Theorem 4.5.1] ensures that there exists a real constant $k_1 > 0$ and $k_2 > 0$ such that $\forall X \in \mathcal{S}_\epsilon$,

$$\begin{aligned}
&\mathrm{dist}^{2^d}(X, \mathcal{P}^\star) \\
\leq& k_1 \left(\mathrm{dist}(X, \mathcal{L}) + \mathrm{dist}(X, \mathcal{M})\right) \\
=& k_1 \left(\mathrm{dist}(X, \mathcal{L}) + \sum_{k=1}^{p} \mathrm{dist}\left(\mathcal{E}_{\mathcal{C}_k}(X), \mathbb{S}^{|\mathcal{C}_k|}_+\right) \right) \\
\leq& k_1 \Bigg(k_2 |\langle C, X \rangle - \langle C, X^\star \rangle| + \\
&\sum_{k=1}^{p} |\mathcal{C}_k| \max\{\lambda_{\max}(-\mathcal{E}_{\mathcal{C}_k}(X)), 0\} \Bigg) \\
\leq& k_1 k_3 \Bigg(|\langle C, X \rangle - \langle C, X^\star \rangle| + \\
&\sum_{k=1}^{p} \max\{\lambda_{\max}(-\mathcal{E}_{\mathcal{C}_k}(X)), 0\} \Bigg),
\end{aligned}$$

where $k_3 = \max\{k_2, |\mathcal{C}_1|, \ldots, |\mathcal{C}_k|\}$ and the second inequality applies the fact that $\mathrm{dist}\left(X, \mathbb{S}^{|\mathcal{C}_k|}_+\right) \leq |\mathcal{C}_k| \max\{\lambda_{\max}(-X), 0\}$ for any $X \in \mathbb{S}^{|\mathcal{C}_k|}$ and $\mathrm{dist}(X, \mathcal{L}) \leq k_2 |\langle C, X \rangle - \langle C, X^\star \rangle|$ since $\mathcal{L}$ is an affine space. To characterize the relationship of $G(X) - G(X^\star)$ and $\mathrm{dist}(X, \mathcal{P}^\star)$, we then proceed to prove that for some large enough $\rho \geq 1$, the following holds

$$\begin{aligned}
&|\langle C, X \rangle - \langle C, X^\star \rangle| + \sum_{k=1}^{p} \max\{\lambda_{\max}\left(-\mathcal{E}_{\mathcal{C}_k}(X)\right), 0\} \\
\leq& G(X) - G(X^\star) \\
=& \langle C, X \rangle - \langle C, X^\star \rangle + \rho \sum_{k=1}^{p} \max\{\lambda_{\max}\left(-\mathcal{E}_{\mathcal{C}_k}(X)\right), 0\}.
\end{aligned}$$

In particular, if $|\langle C, X \rangle - \langle C, X^\star \rangle| = \langle C, X \rangle - \langle C, X^\star \rangle$, the inequality holds naturally. If, however, $|\langle C, X \rangle - \langle C, X^\star \rangle| = \langle C, X^\star \rangle - \langle C, X \rangle$, assuming the above inequality holds, it follows that

$$\begin{aligned}
&2\langle C, X \rangle - 2\langle C, X^\star \rangle \\
&\quad + (\rho - 1) \sum_{k=1}^{p} \max\{\lambda_{\max}(-\mathcal{E}_{\mathcal{C}_k}(X)), 0\} \geq 0,
\end{aligned}$$

which is equivalent to

$$\begin{aligned}
&\langle C, X \rangle - \langle C, X^\star \rangle \\
&\quad + \frac{1}{2}(\rho - 1) \sum_{k=1}^{p} \max\{\lambda_{\max}(-\mathcal{E}_{\mathcal{C}_k}(X)), 0\} \geq 0,
\end{aligned}$$

Indeed, the above inequality holds as long as $\frac{(\rho-1)}{2} > \mathcal{D}_{\mathcal{Y}^\star}$ or, equivalently, $\rho > 2\mathcal{D}_{\mathcal{Y}^\star} + 1$. Therefore, upon choosing $\rho > 2\mathcal{D}_{\mathcal{Y}^\star} + 1$, we show that there exists a constant $\mu > 0$ such that

$$\mathrm{dist}^{2^d}(X, \mathcal{P}^\star) \leq \mu \cdot (G(X) - G(X^\star)).$$

The number d is the singularity degree of (9) and is bounded [31, Lemma 3.6]. Furthermore, if there exists a pair of primal and dual solutions for (9) and (12) that satisfies strict complementarity, the singularity degree d is at most one [31, Section 5].

F. Reformulation of master problem (20)

By eliminating the constant terms in (20) and employing the operator $\mathrm{vec} : \mathbb{S}^n \rightarrow \mathbb{R}^{n^2}$, which vertically stacks the columns of the input matrix, we can rephrase the problem (20) as follows:

$$\begin{aligned} \min_{\gamma, S, y} \; & \begin{bmatrix} \bar{\gamma} & \bar{s}^\top & y^\top \end{bmatrix} \begin{bmatrix} Q_{11} & Q_{12} & Q_{13} \\ Q_{12}^\top & Q_{22} & Q_{23} \\ Q_{13}^\top & Q_{23}^\top & Q_{33} \end{bmatrix} \begin{bmatrix} \bar{\gamma} \\ \bar{s} \\ y \end{bmatrix} \\ & + \begin{bmatrix} q_1^\top & q_2^\top & q_3^\top \end{bmatrix} \begin{bmatrix} \bar{\gamma} \\ \bar{s} \\ y \end{bmatrix} \qquad (26) \\ \text{subject to } & \gamma_k \geq 0, S_k \in \mathbb{S}_+^r, \\ & \gamma_k + \mathrm{tr}(S_k) \leq \rho, \forall k = 1, \ldots, p, \end{aligned}$$

where $y \in \mathbb{R}^m$, $\bar{s}^\top = \left[\mathrm{vec}(S_1)^\top, \ldots, \mathrm{vec}(S_p)^\top\right]$, $\bar{\gamma}^\top = [\gamma_1, \ldots, \gamma_p]$. Let us define the matrices $H_k = E_k \otimes E_k \in \mathbb{R}^{|\mathcal{C}_k|^2 \times n^2}$, $M_k = P_k \otimes P_k \in \mathbb{R}^{|\mathcal{C}_k|^2 \times r^2}$, $w_k = \mathrm{vec}\left(\bar{W}_k\right) \in \mathbb{R}^{|\mathcal{C}_k|^2}$, $A = \begin{bmatrix} \mathrm{vec}\left(A_1\right) & \ldots & \mathrm{vec}\left(A_m\right) \end{bmatrix}^\top \in \mathbb{R}^{m \times n^2}$, $\mathrm{vec}(\Omega_t) = \omega \in \mathbb{R}^{n^2}$, and $\mathrm{vec}(C) = c \in \mathbb{R}^{n^2}$. We construct the matrix Q with the following components

$$\begin{aligned} Q_{11} &= \begin{bmatrix} w_1^\top H_1 H_1^\top w_1 & \ldots & w_1^\top H_1 H_p^\top w_p \\ \vdots & \ddots & \vdots \\ w_p^\top H_p H_1^\top w_1 & \ldots & w_p^\top H_p H_p^\top w_p \end{bmatrix}, \\ Q_{22} &= \begin{bmatrix} M_1^\top H_1 H_1^\top M_1 & \ldots & M_1^\top H_1 H_p^\top M_p \\ \vdots & \ddots & \vdots \\ M_p^\top H_p H_1^\top M_1 & \ldots & M_p^\top H_p H_p^\top M_p \end{bmatrix}, \\ Q_{12} &= \begin{bmatrix} w_1^\top H_1 H_1^\top M_1 & \ldots & w_1^\top H_1 H_p^\top M_p \\ \vdots & \ddots & \vdots \\ w_p^\top H_p H_1^\top M_1 & \ldots & w_p^\top H_p H_p^\top M_p \end{bmatrix}, \\ Q_{13} &= \left[A H_1^\top w_1 \ldots A H_p^\top w_p\right], \\ Q_{23} &= \left[A H_1^\top M_1 \ldots A H_p^\top M_p\right], \\ Q_{33} &= A A^\top. \end{aligned}$$

For the linear term, we have

$$\begin{aligned} q_1^\top &= \left[2\alpha\omega^\top - 2c^\top\right] \left[H_1^\top w_1 \ldots H_p^\top w_p\right], \\ q_2^\top &= \left[2\alpha\omega^\top - 2c^\top\right] \left[H_1^\top M_1 \ldots H_p^\top M_p\right], \\ q_3^\top &= -2\alpha\left(b - \mathcal{A}\left(\Omega_t\right)\right) - 2(Ac)^\top. \end{aligned}$$

While problem (26) can already be solved using standard conic solvers, it's possible to simplify the computation even more by eliminating the variable y. This simplification involves considering the optimality condition for y

$$y = Q_{33}^{-1} \left(\frac{-q_3}{2} - Q_{13}^\top \alpha - Q_{23}^\top \bar{s} \right).$$

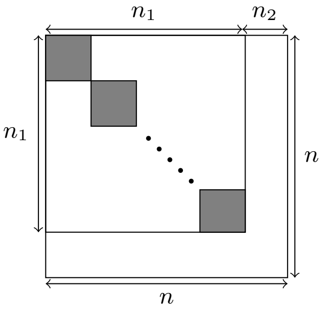

Fig. 4: Sparsity pattern of randomly generated $\hat{X} \in \mathrm{R}^n$ such that $\mathrm{rank}(\hat{X}) = n_1,\ n_1 + n_2 = n$.

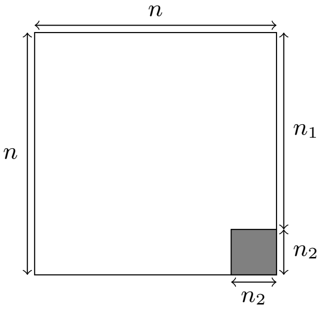

Fig. 5: Sparsity pattern of randomly generated $\hat{Z} \in \mathrm{R}^n$ such that $\mathrm{rank}(\hat{Z}) = n_2,\ n_1 + n_2 = n$.

Then we have

$$\begin{aligned} \operatorname*{argmin}_{\gamma_k, S_k} \; & \begin{bmatrix} \gamma & \bar{s}^\top \end{bmatrix} \begin{bmatrix} M_{11} & M_{12} \\ M_{12}^\top & M_{22} \end{bmatrix} \begin{bmatrix} \gamma \\ \bar{s} \end{bmatrix} \\ & + \begin{bmatrix} m_1^\top & m_2^\top \end{bmatrix} \begin{bmatrix} \gamma \\ \bar{s} \end{bmatrix} \qquad (27) \\ \text{subject to } & \gamma_k \geq 0, S_k \in \mathbb{S}_+^r, \\ & \gamma_k + \mathrm{tr}(S_k) \leq \rho, \forall k = 1, \ldots, p, \end{aligned}$$

where

$$\begin{aligned} & M_{11} = Q_{11} - Q_{13} Q_{33}^{-1} Q_{13}^\top,\ m_1 = q_1 - Q_{13} Q_{33}^{-1} q_3, \\ & M_{12} = Q_{12} - Q_{13} Q_{33}^{-1} Q_{23}^\top,\ m_2 = q_2 - Q_{23} Q_{33}^{-1} q_3, \\ & M_{22} = Q_{22} - Q_{23} Q_{33}^{-1} Q_{23}^\top. \end{aligned}$$

G. Problem data generation details

In this section, we present a detailed account of our data generation process . We commence by introducing pertinent facts related to SDPs.

Lemma 4: [32, Lemma 3] Let X and (y, Z) be respectively primal and dual feasible for (1) and (2). Then they are optimal if and only if there exists $Q \in \mathbb{R}^{n \times n}$, with $Q^\mathrm{T} Q = I$, such that

$$\begin{aligned} & X = Q \operatorname{Diag}\left(\lambda_1, \ldots, \lambda_n\right) Q^\mathrm{T}, \\ & Z = Q \operatorname{Diag}\left(\omega_1, \ldots, \omega_n\right) Q^\mathrm{T}, \\ & \lambda_i \omega_i = 0, \quad i = 1, \ldots, n. \end{aligned}$$

Lemma 4 expresses complementarity in terms of the eigenvalues of X and Z. If X has rank r and Z has rank s, complementarity implies $r + s \leqslant n$. Furthermore, we can formulate an equivalent expression for Lemma 4. Given a pair of optimal solutions $X^\star$ and $(y^\star, Z^\star)$, the complementary slackness condition is synonymous with $Z^\star X^\star = 0$.

To generate random data, we perform the following steps

Pseudocode: Random Data Generation

```
Input: Parameters $m$, $n_1$, $n_2$ and sparsity pattern.
Output: Matrices $\hat{A}_i$, $\hat{X}$, $\hat{b}$, $\hat{Z}$, $\hat{y}$, $\hat{C}$
for $i = 1$ to $m$ do
    Generate random matrix $\hat{A}_i$ with the sparsity pattern as shown in Figure 1 such that $\mathrm{tr}(\hat{A}_i) = 0$
end for
Generate a random sparse matrix $\hat{X} \in \mathbb{S}_+^{n_1+n_2}$ as shown in Figure 4 with $\mathrm{rank}(\hat{X}) = n_1$
for $i = 1$ to $m$ do
    $\hat{b}_i \leftarrow \langle \hat{A}_i, \hat{X} \rangle$
end for
Generate a random sparse dual matrix $\hat{Z} \in \mathbb{S}_+^{n_1+n_2}$ as shown in Figure 5 with $\mathrm{rank}(\hat{Z}) = n_2$
Generate a random vector $\hat{y} \in \mathbb{R}^m$ and set $\hat{C} \leftarrow \hat{Z}$
for $i = 1$ to $m$ do
    $\hat{C} \leftarrow \hat{C} + \hat{A}_i \cdot \hat{y}_i$
end for
```

We note that the feasible primal-dual pair $(\hat{X}, \{\hat{Z}, \hat{y}\})$ exhibits strong duality and $\hat{Z}\hat{X} = 0$, then it is an optimal primal-dual pair with $\mathrm{rank}(\hat{X}) + \mathrm{rank}(\hat{Z}) = n$.