Sparse Sums of Positive Semidefinite Matrices

Marcel K. de Carli Silva* Nicholas J. A. Harvey† Cristiane M. Sato‡

Abstract

Recently there has been much interest in "sparsifying" sums of rank one matrices: modifying the coefficients such that only a few are nonzero, while approximately preserving the matrix that results from the sum. Results of this sort have found applications in many different areas, including sparsifying graphs. In this paper we consider the more general problem of sparsifying sums of positive semidefinite matrices. We give several algorithms for solving this problem and describe several applications of these algorithms.

*Department of Combinatorics and Optimization, University of Waterloo. mksilva@uwaterloo.ca. Partially supported by an NSERC Discovery Grant of L. Tunçel.

†Department of Computer Science, University of British Columbia. nickhar@cs.ubc.ca. Supported by an NSERC Discovery Grant.

‡Department of Combinatorics and Optimization, University of Waterloo. cmsato@uwaterloo.ca. Partially supported by an NSERC Discovery Grant of N. Wormald.

1 Introduction

A *sparsifier* of a graph is a subgraph that approximately preserves some structural properties of the graph. The original work in this area studied *cut sparsifiers*, which are weighted subgraphs that approximate every cut arbitrarily well. The celebrated work of Benczúr and Karger [5, 6] proved that every undirected graph with n vertices and m edges (and potentially non-negative weights on its edges) has a subgraph with only $O(n \log n/\varepsilon^2)$ edges (and new weights on those edges) such that, for every cut, the weight of the cut in the original graph and its subgraph agree up to a multiplicative factor of $(1 \pm \varepsilon)$. Benczúr and Karger also gave a randomized algorithm to construct a cut sparsifier in $\tilde{O}(m/\varepsilon^2)$ time. Recent work has extended and improved their algorithm in various ways [9, 10, 11, 13, 14].

Spielman and Teng [33] introduced *spectral sparsifiers*, which are weighted subgraphs such that the quadratic forms defined by the Laplacians of the graph and the sparsifier agree up to a multiplicative factor of $(1 \pm \varepsilon)$. Spectral sparsifiers are also cut sparsifiers, as can be seen by evaluating these quadratic forms at $\{0,1\}$-vectors. They proved that every undirected graph with n vertices and m edges (and potentially non-negative weights on its edges) has a spectral sparsifier with only $n \operatorname{polylog}(n)/\varepsilon^2$ edges (and new weights on those edges). Spielman and Srivastava [32] reduce the graph sparsification problem to the following abstract problem in matrix theory.

Problem 1. *Let $v_1, \ldots, v_m \in \mathbb{R}^n$ be vectors and let $B = \sum_i v_i v_i^T$. Given $\varepsilon \in (0,1)$, find a vector $y \in \mathbb{R}^m$ with small support such that $y \geq 0$ and*

$$B \preceq \sum_i y_i v_i v_i^T \preceq (1+\varepsilon)B. \tag{1}$$

(Here the notation $X \preceq Y$ means that the matrix $Y - X$ is positive semidefinite.)

Spielman and Srivastava [32] observe that Problem 1 can be solved using known concentration bounds on operator-valued random variables, specifically Rudelson's sampling lemma [27, 28]. This approach yields a vector y with support size $O(n \log n/\varepsilon^2)$, and therefore yields a construction of spectral sparsifiers with $O(n \log n/\varepsilon^2)$ edges. Their algorithm relies on the linear system solver of Spielman and Teng [33], which was significantly simplified by Koutis, Miller and Peng [23]. Recent work [21] has improved the space usage of Spielman and Srivastava's algorithm.

In subsequent work, Batson, Spielman and Srivastava [4] give a deterministic algorithm that solves Problem 1 and produces a vector y with support size $O(n/\varepsilon^2)$. Consequently they obtain improved spectral sparsifiers with $O(n/\varepsilon^2)$ edges. This work led to important progress in metric embeddings [24, 29], convex geometry [34] and Banach space theory [31].

In this paper, we focus on a more general problem.

Problem 2. *Let $B_1, \ldots, B_m$ be symmetric, positive semidefinite matrices of size $n \times n$ and let $B = \sum_i B_i$. Given $\varepsilon \in (0,1)$, find a vector $y \in \mathbb{R}^m$ with small support such that $y \geq 0$ and*

$$B \preceq \sum_i y_i B_i \preceq (1+\varepsilon)B. \tag{2}$$

This problem can also be solved by known concentration bounds: Ahlswede and Winter [1] give a method for generalizing Chernoff-like bounds to operator-valued random variables, and one of their theorems [1, Theorem 19] directly yields a solution to Problem 2. (Other expositions of these results also exist [35, 15].) This approach yields a vector y with support size $O(n \log n/\varepsilon^2)$. See Section 3 for more details.

This paper gives two improved solutions to Problem 2. Our interest in this topic is motivated by several applications, such as constructing sparsifiers with certain auxiliary properties. We discuss these applications in Section 1.2.

1.1 Our Results

We give several efficient algorithms for solving Problem 2. Our strongest solution is:

Theorem 3. *Let $B_1, \ldots, B_m$ be symmetric, positive semidefinite matrices of size $n \times n$ and arbitrary rank. Set $B := \sum_i B_i$. For any $\varepsilon \in (0,1)$, there is a deterministic algorithm to construct a vector $y \in \mathbb{R}^m$ with $O(n/\varepsilon^2)$ nonzero entries such that $y \geq 0$ and*

$$B \preceq \sum_i y_i B_i \preceq (1+\varepsilon)B.$$

The algorithm runs in $O(mn^3/\varepsilon^2)$ time.

Our proof of Theorem 3 is quite simple and builds on results of Batson, Spielman and Srivastava [4]. We remark that the assumption that the B_i's are positive semidefinite cannot be removed; see Appendix C.

We also give a second solution to Problem 2 which is quantitatively weaker, although it is based on very general machinery which might prove useful in further applications or generalizations of Problem 2. This second solution is based on the matrix multiplicative weights update method (MMWUM) of Arora and Kale [3, 20]. By a black-box application of their theorems we obtain a deterministic algorithm to construct a vector y with $O(n \log n/\varepsilon^3)$ nonzero entries. By slightly refining their analysis we can improve the number of nonzero entries to $O(n \log n/\varepsilon^2)$. We remark that Orecchia and Vishnoi [25] have used MMWUM for solving the balanced separator problem; this can be used as a subroutine in Spielman and Teng's algorithm for constructing spectral sparsifiers.

Another virtue of our second solution is that it illustrates that the surprising Batson-Spielman-Srivastava (BSS) algorithm is actually closely related to MMWUM. In particular, the algorithms underlying our two solutions are *identical*, except for the use of slightly different potential functions. We explain this connection in Section 7.

1.2 Applications

Approximate Carathéodory theorems. One immediate application for Theorem 3 is an approximate Carathéodory-type theorem. A classic result of this sort is:

Theorem 4 (Althöfer [2])**.** *Let $v_1, \ldots, v_m \in [0,1]^n$ and let $\lambda \in \mathbb{R}^m$ satisfy $\lambda \geq 0$ and $\sum_i \lambda_i = 1$. Then there exists $\mu \in \mathbb{R}^m$ with $\mu \geq 0$, $\sum_i \mu_i = 1$ and only $O(\log n/\varepsilon^2)$ nonzero entries such that $\|\sum_i \lambda_i v_i - \sum_i \mu_i v_i\|_\infty \leq \varepsilon$.*

This theorem has several interesting consequences, including the existence of sparse, low-regret solutions to zero-sum games. Theorem 3 has the following corollary.

Corollary 5. *Let $B_1, \ldots, B_m$ be symmetric, positive semidefinite matrices of size $n \times n$ and let $\lambda \in \mathbb{R}^m$ satisfy $\lambda \geq 0$ and $\sum_i \lambda_i = 1$. Let $B = \sum_i \lambda_i B_i$. For any $\varepsilon \in (0,1)$, there exists $\mu \geq 0$ with $\sum_i \mu_i = 1$ such that μ has $O(n/\varepsilon^2)$ nonzero entries and*

$$(1-\varepsilon)B \preceq \sum_i \mu_i B_i \preceq (1+\varepsilon)B.$$

Although the support size in Corollary 5 is much larger than in Theorem 4, Corollary 5 provides a multiplicative error bound whereas Theorem 4 only provides an additive error bound.

Sparse solutions to semidefinite programs.

Corollary 6. *Let $A_1, \ldots, A_m$ be symmetric, positive semidefinite matrices of size $n \times n$, and let B be a symmetric matrix of size $n \times n$. Let $c \in \mathbb{R}^m$ with $c \geq 0$. Suppose that the semidefinite program (SDP)*

$$\min \left\{ c^T y : \sum_i y_i A_i \succeq B,\ y \in \mathbb{R}^m,\ y \geq 0 \right\} \tag{3}$$

has an optimal solution y^*. Then, for any real $\varepsilon \in (0,1)$, it has a feasible solution $\bar{y}$ with at most $O(n/\varepsilon^2)$ nonzero entries and $c^T \bar{y} \leq (1+\varepsilon) c^T y^*$.

Several important SDPs can be cast as in Corollary 6; see, e.g., [17, 18]. Recently, Jain and Yao [19] gave a parallel approximation algorithm for SDPs in this form with B positive semidefinite.

Sparsifiers with costs.

Corollary 7. *Let $G = (V, E)$ be a graph, let $w \colon E \to \mathbb{R}_+$ be a weight function, and let $c_1, \ldots, c_k \colon E \to \mathbb{R}_+$ be cost functions, with $k = O(n)$. Let $\mathcal{L}_G(w)$ denote the Laplacian matrix for graph G with weight function w. For any real $\varepsilon \in (0,1)$, there is a deterministic polynomial-time algorithm to find a subgraph H of G and a weight function $w_H \colon E(H) \to \mathbb{R}_+$ such that*

$$\begin{aligned} (1-\varepsilon)\mathcal{L}_G(w) \ &\preceq\ \mathcal{L}_H(w_H) \ \preceq\ (1+\varepsilon)\mathcal{L}_G(w), \\ (1-\varepsilon)\sum_{e \in E} w_e c_{i,e} \ &\leq\ \sum_{e \in E(H)} w_{H,e} c_{i,e} \ \leq\ (1+\varepsilon)\sum_{e \in E} w_e c_{i,e} \end{aligned} \tag{4}$$

for all i and $|E(H)| = O(n/\varepsilon^2)$.

The inequalities in (4) are equivalent to the condition that the subgraph H (with weights w_H) is a spectral sparsifier of G (with weights w).

Corollary 7 applied with cost functions which are characteristic vectors of singletons yields sparsifiers with prescribed edges.

Corollary 8. *Let $G = (V, E)$ be a graph, let $w \colon E \to \mathbb{R}_+$ be a weight function, and let $F \subseteq E$ with $|F| = O(n)$. For any real $\varepsilon \in (0,1)$, there is a deterministic polynomial-time algorithm to find a subgraph H of G and a weight function $w_H \colon E(H) \to \mathbb{R}_+$ such that (4) holds, $|E(H)| = O(n/\varepsilon^2)$, and, for every $f \in F$,*

$$(1-\varepsilon) w_f \leq w_{H,f} \leq (1+\varepsilon) w_f.$$

Kolla, Makarychev, Saberi and Teng [22] have also studied sparsifiers for which certain edges are forced to belong to the sparsifier.

Sparsifiers on subgraphs.

Corollary 9. *Let $G = (V, E)$ be a graph, let $w \colon E \to \mathbb{R}_+$ be a weight function, and let $\mathcal{F}$ be a collection of subgraphs of G such that $\sum_{F \in \mathcal{F}} |V(F)| = O(n)$. For any real $\varepsilon \in (0,1)$, there is a deterministic polynomial-time algorithm to find a subgraph H of G and a weight function $w_H \colon E(H) \to \mathbb{R}_+$ such that (4) holds, $|E(H)| = O(n/\varepsilon^2)$, and*

$$(1-\varepsilon)\mathcal{L}_F(w_F) \ \preceq\ \mathcal{L}_{H \cap F}(w_H \restriction_{E(F)}) \ \preceq\ (1+\varepsilon)\mathcal{L}_F(w_F)$$

for all $F \in \mathcal{F}$, where $w_F := w \restriction_{E(F)}$ is the restriction of w to the coordinates $E(F)$ and $H \cap F = \big(V(F), E(F) \cap E(H)\big)$.

Hypergraph sparsifiers. Let $\mathcal{H} = (V, \mathcal{E})$ be a hypergraph, and let $w\colon \mathcal{E} \to \mathbb{R}_+$. We follow the definition of Laplacian for hypergraphs as in [26]. For each hyperedge $E \in \mathcal{E}$, define its Laplacian $\mathcal{L}_E$ as the graph Laplacian of a graph on V whose edge set forms a clique on E. Define the Laplacian for the hypergraph $\mathcal{H}$ with weight function w as the matrix $\mathcal{L}_{\mathcal{H}}(w) := \sum_{E\in\mathcal{E}} w_E \mathcal{L}_E$.

Corollary 10. *Let $\mathcal{H} = (V, \mathcal{E})$ be a hypergraph, let $w\colon \mathcal{E} \to \mathbb{R}_+$ be a weight function. For any real $\varepsilon \in (0, 1)$, there is a deterministic polynomial-time algorithm to find a sub-hypergraph $\mathcal{G}$ of $\mathcal{H}$ and a weight function $w_{\mathcal{G}}\colon \mathcal{E}(\mathcal{G}) \to \mathbb{R}_+$ such that*

$$(1-\varepsilon)\mathcal{L}_{\mathcal{H}}(w) \preceq \mathcal{L}_{\mathcal{G}}(w_{\mathcal{G}}) \preceq (1+\varepsilon)\mathcal{L}_{\mathcal{H}}(w), \tag{5}$$

and $|\mathcal{E}(\mathcal{G})| = O(n/\varepsilon^2)$.

The usual definition of the weight $w(\delta_{\mathcal{H}}(S))$ of a cut induced by a set of vertices S in a hypergraph $\mathcal{H}$ is the sum of the weights of the hyperedges that have vertices both in S and in $\overline{S} := V \setminus S$. An alternative definition of the weight of the cut induced by a set of vertices S would be $w^*(\delta_{\mathcal{H}}(S)) := \sum_{E\in\mathcal{E}} w_E \cdot |S \cap E| \cdot |\overline{S} \cap E|$. Note that $w^*(\delta_{\mathcal{H}}(S))$ is obtained by evaluating the quadratic form $x^T \mathcal{L}_{\mathcal{H}}(w) x$, where x is the characteristic vector of S. Thus, inequality (5) implies that $(1-\varepsilon)w^*(\delta_{\mathcal{H}}(S)) \le w^*(\delta_{\mathcal{G}}(S)) \le (1+\varepsilon)w^*(\delta_{\mathcal{H}}(S))$ for every S.

For r-uniform hypergraphs, we have $(r-1)w(\delta_{\mathcal{H}}(S)) \le w^*(\delta_{\mathcal{H}}(S)) \le \lfloor r/2 \rfloor \lceil r/2 \rceil w(\delta_{\mathcal{H}}(S))$. Together with the inequality (5), this implies that

$$\frac{(1-\varepsilon)(r-1)}{r^2/4} w(\delta_{\mathcal{H}}(S)) \le w_{\mathcal{G}}(\delta_{\mathcal{G}}(S)) \le \frac{(1+\varepsilon)r^2}{4(r-1)} w(\delta_{\mathcal{H}}(S)) \qquad \forall S.$$

In other words, the sparsified hypergraph $\mathcal{G}$ approximates the weight of the cuts in the hypergraph $\mathcal{H}$ to within a factor $\Theta(r^2)$. Moreover, if $r = 3$, then $w^*(\delta_{\mathcal{H}}(S)) = 2w(\delta_{\mathcal{H}}(S))$, and so

$$(1-\varepsilon)w(\delta_{\mathcal{H}}(S)) \le w_{\mathcal{G}}(\delta_{\mathcal{G}}(S)) \le (1+\varepsilon)w(\delta_{\mathcal{H}}(S)) \qquad \forall S.$$

2 Preliminaries

For a nonnegative integer n, we denote $[n] := \{1, \ldots, n\}$. The nonnegative reals are denoted by $\mathbb{R}_+$. The set of $n \times n$ symmetric matrices is denoted by $\mathbb{S}^n$. The set of symmetric, $n \times n$ positive semidefinite (resp., positive definite) matrices is denoted by $\mathbb{S}^n_+$ (resp., $\mathbb{S}^n_{++}$). Recall that $X \in \mathbb{S}^n$ is positive semidefinite if $v^T X v \ge 0$ for all $v \in \mathbb{R}^n$, and X is positive definite if X is positive semidefinite and $v^T X v = 0$ implies $v = 0$. Sometimes we denote $X \in \mathbb{S}^n_+$ by $X \succeq 0$ and the notation $X \succeq Y$ means that $X - Y \succeq 0$. For $X \in \mathbb{S}^n$ and $a, b \in \mathbb{R}$, the notation $X \in [a, b]$ means that $aI \preceq X \preceq bI$, where I is the identity matrix.

For $X \in \mathbb{S}^n$, its trace is $\operatorname{Tr} X := \sum_{i=1}^n X_{ii}$, its largest (resp., smallest) eigenvalue is denoted by $\lambda_{\max}(X)$ (resp., $\lambda_{\min}(X)$). The vector space $\mathbb{S}^n$ can be endowed with the trace inner product $\langle \cdot, \cdot \rangle$ defined by $\langle X, Y \rangle := \operatorname{Tr}(XY) = \sum_{i,j} X_{ij} Y_{ij}$ for every $X, Y \in \mathbb{S}^n$. We shall repeatedly use that $\operatorname{Tr}(XY) = \operatorname{Tr}(YX)$ for any matrices X, Y for which the products XY and YX make sense.

Let $G = (V, E)$ be a graph. The canonical basis vectors of $\mathbb{R}^V$ are $\{ e_i : i \in V \}$, and the canonical basis vectors of $\mathbb{R}^E$ are $\{ e_{\{i,j\}} : \{i, j\} \in E \}$. The Laplacian of G is the linear transformation $\mathcal{L}_G(\cdot)\colon \mathbb{R}^E \to \mathbb{S}^V$ defined by $\mathcal{L}_G(w) = \sum_{\{i,j\}\in E} w_{\{i,j\}} (e_i - e_j)(e_i - e_j)^T$.

When dealing with Problem 2, we may assume that $B = I$. See [4, Proof of Theorem 1.1] for the details of the reduction.

Algorithm 1 A procedure for solving Problem 2 based on the BSS method.

procedure SparsifySumOfMatricesByBSS($B_1, \ldots, B_m$, ε)

input: Matrices $B_1, \ldots, B_m \in \mathbb{S}^n_+$ such that $\sum_i B_i = I$, and a parameter $\varepsilon \in (0,1)$.

output: A vector y with $O(n/\varepsilon^2)$ nonzero entries such that $I \preceq \sum_i y_i B_i \preceq (1+O(\varepsilon))I$.

Initially $A(0) := 0$ and $y(0) := 0$. Set parameters $u_0, \ell_0, \delta_L, \delta_U$ as in (8) and $T := 4n/\varepsilon^2$.

Define the potential functions $\Phi^u(A) := \operatorname{Tr}(uI - A)^{-1}$ and $\Phi_\ell(A) := \operatorname{Tr}(A - \ell I)^{-1}$.

For $t = 1, \ldots, T$

Set $u_t := u_{t-1} + \delta_U$ and $\ell_t := \ell_{t-1} + \delta_L$.

Find a matrix B_j and a value $\alpha > 0$ such that $A(t-1) + \alpha B_j \in [\ell_t, u_t]$, and

$$\Phi^{u_t}(A(t-1) + \alpha B_j) \le \Phi^{u_{t-1}}(A(t-1)) \quad \text{and} \quad \Phi_{\ell_t}(A(t-1) + \alpha B_j) \le \Phi_{\ell_{t-1}}(A(t-1)).$$

Set $A(t) := A(t-1) + \alpha B_j$ and $y(t) := y(t-1) + \alpha e_j$.

Return $y(T)/\lambda_{\min}(A(T))$.

3 Solving Problem 2 by Ahlswede-Winter

As mentioned earlier, Spielman and Srivastava [32] explain how Problem 1 can be solved by Rudelson's sampling lemma. This lemma can be easily generalized to handle matrices of arbitrary rank using the Ahlswede-Winter inequality, yielding a solution to Problem 2.

Let X be a random matrix such that $X = B_i / \operatorname{Tr} B_i$ with probability $p_i := \operatorname{Tr} B_i / \operatorname{Tr} I$. Since $B_i \succeq 0$ and $\sum_i B_i = I$, the p_i's define a probability distribution.

Theorem 11 ([1, Theorem 19]). *Let $X, X_1, \ldots, X_T$ be i.i.d. random variables with values in $\mathbb{S}^n$ such that $X_i \in [0,1]$ for every i and $\mathbf{E}(X) = \mu I$ with $\mu \in [0,1]$. Let $\varepsilon \in (0, 1/2)$. Then*

$$\mathbf{P}\left(\frac{1}{\mu T}\sum_{i=1}^{T} X_i \notin [1-\varepsilon, 1+\varepsilon]\right) \le 2n \cdot \exp\left(-T\frac{\varepsilon^2\mu}{2\ln 2}\right).$$

In our case, $\mathbf{E}(X) = (1/n)I$ and $X \in [0,1]$. So $\mu = 1/n$. Thus, if $T > (2\ln 2) \cdot \frac{\ln n + 2\ln 2}{\varepsilon^2 \mu} = O(n \log n/\varepsilon^2)$, then $\mathbf{P}\left(\frac{1}{\mu T}\sum_{i=1}^{T} X_i \notin [1-\varepsilon, 1+\varepsilon]\right) < 1/2$, as desired.

4 Solving Problem 2 by BSS

In our modification of the BSS algorithm [4], we keep a matrix A of the form $A = \sum_i y_i B_i$ with $y \ge 0$, starting with $A = 0$, and at each iteration we add another term αB_j to A. We enforce the invariant that the eigenvalues of A lie in $[\ell, u]$, where u and ℓ are parameters given by $u = u_0 + t\delta_U$ and $\ell = \ell_0 + t\delta_L$ after t iterations. This procedure is presented in Algorithm 1. The step of the algorithm which finds B_j and α can be done by exhaustive search on j and binary search on α. Instead of the binary search, one could also compare the quantities $U_{A(t-1)}(B_j)$ and $L_{A(t-1)}(B_j)$ defined below.

In the original BSS algorithm, the matrices are rank one: $B_j = v_j v_j^T$ for some vector v_j. Their Lemmas 3.3 and 3.4 give sufficient conditions on the new term $\alpha v_j v_j^T$ so that the invariant on the eigenvalues is maintained; Lemma 3.5 gives sufficient conditions on the remaining parameters so that a suitable new term $\alpha v_j v_j^T$ exists with $\alpha > 0$. In this section we generalize those lemmas to allow B_i matrices of arbitrary rank.

Let $A \in \mathbb{S}^n$. If $u \in \mathbb{R}$ with $\lambda_{\max}(A) < u$, define $\Phi^u(A) := \operatorname{Tr}(uI - A)^{-1}$. If $\ell \in \mathbb{R}$ with $\lambda_{\min}(A) > \ell$, define $\Phi_\ell(A) := \operatorname{Tr}(A - \ell I)^{-1}$. Note that $\Phi_\ell(A) = \sum_i 1/(\lambda_i - \ell)$ and $\Phi^u(A) = \sum_i 1/(u - \lambda_i)$, where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A.

Lemma 12 (Analog of Lemma 3.3 in [4]). *Let $A \in \mathbb{S}^n$ and $X \in \mathbb{S}^n_+$ with $X \neq 0$. Let $u \in \mathbb{R}$ and $\delta_U > 0$. Suppose $\lambda_{\max}(A) < u$. Let $u' := u + \delta_U$ and $M := u'I - A$. If*

$$\frac{1}{\alpha} \geq \frac{\langle M^{-2}, X\rangle}{\Phi^u(A) - \Phi^{u'}(A)} + \langle M^{-1}, X\rangle =: U_A(X),$$

then $\lambda_{\max}(A + \alpha X) < u'$ and $\Phi^{u'}(A + \alpha X) \leq \Phi^u(A)$.

Proof. Clearly $M \succ 0$. Let $V := X^{1/2}$. By the Sherman-Morrison-Woodbury formula [12],

$$\begin{aligned}\Phi^{u'}(A + \alpha X) &= \operatorname{Tr}(M - \alpha VV^T)^{-1} = \operatorname{Tr}\left(M^{-1} + \alpha M^{-1}V(I - \alpha V^T M^{-1}V)^{-1}V^T M^{-1}\right)\\ &= \Phi^{u'}(A) + \operatorname{Tr}\left(\alpha M^{-1}V(I - \alpha V^T M^{-1}V)^{-1}V^T M^{-1}\right).\end{aligned}$$

Since $M^{-1} \succ 0$, $X \neq 0$ and $\Phi^u(A) > \Phi^{u'}(A)$, our hypotheses imply $1/\alpha > \langle M^{-1}, X\rangle = \operatorname{Tr}(V^T M^{-1}V) \geq \lambda_{\max}(V^T M^{-1}V) \geq 0$, so $\beta := \lambda_{\min}(I - \alpha V^T M^{-1}V) = 1 - \alpha\lambda_{\max}(V^T M^{-1}V) > 0$ and by, e.g., [16, Corollary 7.7.4],

$$0 \prec \beta I \preceq I - \alpha V^T M^{-1}V \implies 0 \prec (I - \alpha V^T M^{-1}V)^{-1} \preceq \beta^{-1} I.$$

Thus,

$$\begin{aligned}\Phi^{u'}(A + \alpha X) &\leq \Phi^{u'}(A) + \alpha\beta^{-1}\operatorname{Tr}(V^T M^{-2}V)\\ &= \Phi^u(A) - (\Phi^u(A) - \Phi^{u'}(A)) + \alpha\beta^{-1}\langle M^{-2}, X\rangle\end{aligned}$$

To prove that $\Phi^{u'}(A+\alpha X) \leq \Phi^u(A)$, it suffices to show that $\alpha\beta^{-1}\langle M^{-2}, X\rangle \leq \Phi^u(A) - \Phi^{u'}(A)$. This is equivalent to

$$\frac{\langle M^{-2}, X\rangle}{1/\alpha - \lambda_{\max}(V^T M^{-1}V)} \leq \Phi^u(A) - \Phi^{u'}(A),$$

which follows from $1/\alpha \geq U_A(X)$ since $\lambda_{\max}(V^T M^{-1}V) \leq \operatorname{Tr}(V^T M^{-1}V) = \langle M^{-1}, X\rangle$.

It remains to show that $\lambda_{\max}(A + \alpha X) < u'$. Suppose not. Choose $\varepsilon \in (0, \delta_U)$ such that $1/\varepsilon > \Phi^u(A)$. By continuity, for some $\alpha' \in (0, \alpha)$ we have $\lambda_{\max}(A + \alpha' X) = u' - \varepsilon$. Since $1/\alpha' \geq 1/\alpha \geq U_A(X)$, we get $\Phi^{u'}(A + \alpha' X) \geq 1/\varepsilon > \Phi^u(A) \geq \Phi^{u'}(A + \alpha' X)$, a contradiction. □

Lemma 13 (Analog of Lemma 3.4 in [4]). *Let $A \in \mathbb{S}^n$ and $X \in \mathbb{S}^n_+$, with $n \geq 2$. Let $\ell \in \mathbb{R}$ and $\delta_L > 0$. Suppose $\lambda_{\min}(A) > \ell$ and $\Phi_\ell(A) \leq 1/\delta_L$. Let $\ell' := \ell + \delta_L$ and $N := A - \ell' I$. If*

$$0 < \frac{1}{\alpha} \leq \frac{\langle N^{-2}, X\rangle}{\Phi_{\ell'}(A) - \Phi_\ell(A)} - \langle N^{-1}, X\rangle =: L_A(X),$$

then $\lambda_{\min}(A + \alpha X) > \ell'$ and $\Phi_{\ell'}(A + \alpha X) \leq \Phi_\ell(A)$. Moreover, $N \succ 0$.

Proof. Note that $\lambda_{\min}(A) > \ell$ and $\Phi_\ell(A) \leq 1/\delta_L$ imply that $N \succ 0$, and therefore $\lambda_{\min}(A + \alpha X) > \ell'$. Let $V := X^{1/2}$. By the Sherman-Morrison-Woodbury formula,

$$\begin{aligned}\Phi_{\ell'}(A + \alpha X) &= \operatorname{Tr}(N + \alpha VV^T)^{-1} = \operatorname{Tr}\left(N^{-1} - \alpha N^{-1}V(I + \alpha V^T N^{-1}V)^{-1}V^T N^{-1}\right)\\ &= \Phi_{\ell'}(A) - \operatorname{Tr}\left(\alpha N^{-1}V(I + \alpha V^T N^{-1}V)^{-1}V^T N^{-1}\right).\end{aligned}$$

For $\beta := \lambda_{\max}(I + \alpha V^T N^{-1} V)$, we have

$$0 \prec I + \alpha V^T N^{-1} V \preceq \beta I \implies 0 \prec \beta^{-1} I \preceq (I + \alpha V^T N^{-1} V)^{-1}.$$

Thus,

$$\begin{aligned}\Phi_{\ell'}(A + \alpha X) &\leq \Phi_{\ell'}(A) - \alpha\beta^{-1} \operatorname{Tr}(V^T N^{-2} V) \\ &= \Phi_{\ell}(A) + (\Phi_{\ell'}(A) - \Phi_{\ell}(A)) - \alpha\beta^{-1}\langle N^{-2}, X\rangle\end{aligned}$$

We will be done if we show that $\alpha\beta^{-1}\langle N^{-2}, X\rangle \geq \Phi_{\ell'}(A) - \Phi_{\ell}(A)$. This is equivalent to

$$\frac{\langle N^{-2}, X\rangle}{1/\alpha + \lambda_{\max}(V^T N^{-1} V)} \geq \Phi_{\ell'}(A) - \Phi_{\ell}(A)$$

which follows from $0 < 1/\alpha \leq L_A(X)$, since $\Phi_{\ell'}(A) > \Phi_{\ell}(A)$, $N \succ 0$, and $\lambda_{\max}(V^T N^{-1} V) \leq \operatorname{Tr}(V^T N^{-1} V) = \langle N^{-1}, X\rangle$. □

The next lemma can be proved by a syntactic modification of the proof of Lemma 3.5 in [4].

Lemma 14 (Analog of Lemma 3.5 in [4])**.** *Let $A \in \mathbb{S}^n$ with $n \geq 2$, and let $u, \ell \in \mathbb{R}$ and $\varepsilon_U, \delta_U, \varepsilon_L, \delta_L > 0$ such that $\lambda_{\max}(A) < u$, $\lambda_{\min}(A) > \ell$, $\Phi^u(A) \leq \varepsilon_U$, and $\Phi_\ell(A) \leq \varepsilon_L$. Let $B_1, \ldots, B_m \in \mathbb{S}^n$ such that $\sum_i B_i = I$. If*

$$0 \leq \frac{1}{\delta_U} + \varepsilon_U \leq \frac{1}{\delta_L} - \varepsilon_L \tag{6}$$

then there exists $j \in [m]$ and $\alpha > 0$ for which $L_A(B_j) \geq 1/\alpha \geq U_A(B_j)$.

Proof. As in [4, Lemma 3.5], it suffices to show that $\sum_i L_A(B_i) \geq \sum_i U_A(B_i)$. Let $u' := u + \delta_U$, $M := u'I - A$, $\ell' := \ell + \delta_L$, and $N := A - \ell' I$. It follows from the bilinearity of $\langle \cdot, \cdot \rangle$ and the assumption $\sum_i B_i = I$ that

$$\sum_i U_A(B_i) = \frac{\operatorname{Tr} M^{-2}}{\Phi^u(A) - \Phi^{u'}(A)} + \operatorname{Tr} M^{-1} \tag{7a}$$

$$\sum_i L_A(B_i) = \frac{\operatorname{Tr} N^{-2}}{\Phi_{\ell'}(A) - \Phi_{\ell}(A)} - \operatorname{Tr} N^{-1} \tag{7b}$$

It is shown in [4, Lemma 3.5] that (7a) is at most (7b), completing the proof. □

Now we set the parameters of Lemma 14 similarly as in [4]:

$$\delta_L := 1 \qquad \varepsilon_L := \frac{\varepsilon}{2} \qquad \ell_0 := -\frac{n}{\varepsilon_L} \qquad \delta_U := \frac{2+\varepsilon}{2-\varepsilon} \qquad \varepsilon_U := \frac{\varepsilon}{2\delta_U} \qquad u_0 := \frac{n}{\varepsilon_U}. \tag{8}$$

So (6) holds with equality. If A is the matrix obtained after $T = 4n/\varepsilon^2$ iterations, then

$$\frac{\lambda_{\max}(A)}{\lambda_{\min}(A)} \leq \frac{u_0 + T\delta_U}{\ell_0 + T\delta_L} = \left(\frac{2+\varepsilon}{2-\varepsilon}\right)^2 \leq \frac{1+\varepsilon}{1-\varepsilon}$$

so $A' := A/\lambda_{\min}(A)$ satisfies $I \preceq A' \preceq (1+\varepsilon)I/(1-\varepsilon)$ and A' is a positive linear combination of $O(n/\varepsilon^2)$ of the matrices B_i.

A careful implementation of the algorithm has running time $O(mn^3/\varepsilon^2)$. If the B_i's have $O(1)$ nonzero entries, as in the graph sparsification problem, the algorithm can be made to run in time $O(n^4/\varepsilon^2 + mn/\varepsilon^2)$. This concludes the proof of Theorem 3.

5 Solving Problem 2 by MMWUM

Observe that the set of all vectors y that are feasible for (2) is the feasible region of a semidefinite program (SDP). So solving Problem 2 amounts to finding a *sparse* solution to this SDP. Here "sparse" means that there are few non-zero entries in the solution y; this differs from other notions of "low-complexity" SDP solutions, such as the low-rank solutions studied by So, Ye and Zhang [30].

It has long been known known that the multiplicative weight update method can be used to construct sparse solutions for some linear programs. A prominent example is the construction of sparse, low-regret solutions to zero-sum games [8, 36, 37]. (Another example is the work of Charikar et al. [7] on approximating metrics by few tree metrics.) Building on that idea, one might imagine that Arora and Kale's matrix multiplicative update method (MMWUM) [3] can construct sparse solutions to (2). In this section, we show that this is indeed possible: we obtain a solution y to Problem 2 with $O(n \log n/\varepsilon^3)$ nonzero entries.

5.1 Overview of MMWUM

The MMWUM is an algorithm that helps us approximately solve an SDP feasibility problem. The gist of (a slight modification of) the method is contained in the following result (its proof can be found in Appendix A):

Theorem 15. *Let $T, K, n_1, \ldots, n_K$ be positive integers. Let $C_k, A_{1,k}, \ldots, A_{m,k} \in \mathbb{S}^{n_k}$ for $k \in [K]$. For each $k \in [K]$, let $\eta_k > 0$ and $0 < \beta_k \le 1/2$. Given $X_1, \ldots, X_K \in \mathbb{S}^n$, consider the system*

$$\sum_{i=1}^m y_i \langle A_{i,k}, X_k\rangle \ge \langle C_k, X_k\rangle - \eta_k \operatorname{Tr} X_k, \quad \forall k \in [K], \quad \text{and} \quad y \in \mathbb{R}^m_+. \tag{9}$$

For each $k \in [K]$, let $\{\mathcal{P}_k, \mathcal{N}_k\}$ be a partition of $[T]$, let $0 < \ell_k \le \rho_k$, and let $W_k^{(t)} \in \mathbb{S}^n$ and $\ell_k^{(t)} \in \mathbb{R}$ for $t \in [T+1]$. Let $y^{(t)} \in \mathbb{R}^m$ for $t \in [T]$. Suppose the following properties hold:

$$W_k^{(t+1)} = \exp\left(-\frac{\beta_k}{\ell_k+\rho_k}\sum_{\tau=1}^t\left[\sum_{i=1}^m y_i^{(\tau)} A_{i,k} - C_k + \ell_k^{(\tau)} I\right]\right), \qquad \forall t \in \{0,\ldots,T\},\ \forall k \in [K],$$

$$y = y^{(t)} \text{ is a solution for (9) with } X_k = W_k^{(t)},\ \forall k \in [K], \qquad \forall t \in [T],$$

$$\sum_{i=1}^m y_i^{(t)} A_{i,k} - C_k \in \begin{cases} [-\ell_k, \rho_k], & \text{if } t \in \mathcal{P}_k, \\ [-\rho_k, \ell_k], & \text{if } t \in \mathcal{N}_k, \end{cases} \qquad \forall t \in [T],\ k \in [K],$$

$$\ell_k^{(t)} = \ell_k, \quad \forall t \in \mathcal{P}_k,\ \forall k \in [K], \quad \text{and} \quad \ell_k^{(t)} = -\ell_k, \quad \forall t \in \mathcal{N}_k,\ \forall k \in [K].$$

Define $\bar{y} := \frac{1}{T}\sum_{t=1}^T y^{(t)}$. Then,

$$\sum_{i=1}^m \bar{y}_i A_{i,k} - C_k \succeq -\left[\beta_k \ell_k + \frac{(\rho_k+\ell_k)\ln n}{T\beta_k} + (1+\beta_k)\eta_k\right] I, \qquad \forall k \in [K]. \tag{10}$$

Take $K = 2$, set $C_1 := I$ and $C_2 := -I$, and put $A_{i,1} := B_i$ and $A_{i,2} := -B_i$ for each $i \in [m]$. Then Theorem 15 shows that finding a solution for (2) reduces to constructing an oracle that solves linear systems of the form (9) with a few extra technical properties involving the parameters ℓ_k and ρ_k, and adjusting the other parameters so that the error term on the RHS of (10) is $\le \varepsilon$.

To obtain a feasible solution for (2) that is also sparse, the idea is to design an implementation of the oracle that returns a vector $y^{(t)}$ with only *one* nonzero entry at each iteration t of MMWUM,

and to adjust the parameters so that, after $T = O(n \log n/\varepsilon^3)$ iterations, the smallest and largest eigenvalues of $\sum_{i=1}^m \bar{y}_i B_i$ are ε-close to 1. Since $\bar{y}$ is the average of the $y^{(t)}$'s, the resulting $\bar{y}$ will have at most T nonzero entries.

We set the remaining parameters as follows:

$$\beta := \beta_1 := \beta_2 := \frac{\varepsilon}{4}, \qquad T := \frac{2(\rho+\ell)\ln n}{\beta\varepsilon}, \qquad \eta := \eta_1 := \eta_2 := \frac{\varepsilon}{8},$$
$$\ell := \ell_1 := \ell_2 := 1, \qquad \rho := \rho_1 := \rho_2 := \frac{1+\eta}{\eta} n, \qquad \mathcal{P}_1 := \mathcal{N}_2 := [T], \qquad \mathcal{N}_1 := \mathcal{P}_2 := \emptyset.$$

Then the error term on the RHS of (10) is

$$\beta\ell + \frac{(\rho+\ell)\ln n}{T\beta} + (1+\beta)\eta \;=\; \frac{\varepsilon}{4} + \frac{\varepsilon}{2} + \left(1+\frac{\varepsilon}{4}\right)\frac{\varepsilon}{8} \;=\; \frac{7\varepsilon}{8} + \frac{\varepsilon^2}{32} \;\le\; \varepsilon. \tag{11}$$

Thus, (2) follows from (10) and (11). Moreover, $T = O(n \log n/\varepsilon^3)$, as desired.

5.2 The Oracle

It remains to implement the oracle. Consider an iteration t, and let $X_1 := W_1^{(t)}$ and $X_2 := W_2^{(t)}$ be given. We must find $y := y^{(t)} \in \mathbb{R}_+^m$ with at most one nonzero entry such that

$$\sum_{i=1}^m y_i \langle X_1, B_i\rangle \ge (1-\eta)\operatorname{Tr} X_1, \qquad \sum_{i=1}^m y_i \langle X_2, B_i\rangle \le (1+\eta)\operatorname{Tr} X_2, \quad \text{and} \quad \sum_{i=1}^m y_i B_i \in [0,\rho].$$

Since y should have only one nonzero entry, it suffices to find $j \in [m]$ and $\alpha \in \mathbb{R}_+$ such that

$$\begin{aligned} \alpha\langle X_1, B_j\rangle &\ge (1-\eta)\operatorname{Tr} X_1, \\ \alpha\langle X_2, B_j\rangle &\le (1+\eta)\operatorname{Tr} X_2, \\ \alpha \operatorname{Tr} B_j &\le \rho. \end{aligned} \tag{12}$$

Here we are using the fact that $\lambda_{\max}(B_j) \le \operatorname{Tr} B_j$ since $B_j \succeq 0$. We will show that such j and α exist. Due to the definition of W_1 and W_2, the oracle can assume that X_1 is a scalar multiple of X_2^{-1}, although we will not make use of that fact.

Proposition 16. *Let $B_1, \dots, B_m \in \mathbb{S}_+^n$ such that $\sum_{i=1}^m B_i = I$. Let $\eta > 0$ and $X_1, X_2 \in \mathbb{S}_{++}^n$. Then, for $\rho := (1+\eta)n/\eta$, there exist $j \in [m]$ and $\alpha \ge 0$ such that (12) holds.*

Proof. By possibly dropping some B_i's, we may assume that $B_i \ne 0$ for every $i \in [m]$. Define $p_i := \langle X_1, B_i\rangle / \operatorname{Tr} X_1 > 0$ for every $i \in [m]$. Consider the probability space on $[m]$ where j is sampled from $[m]$ with probability p_j. The fact that $\sum_{j=1}^m p_j = 1$ follows from $\sum_{i=1}^m B_i = I$. Then $\mathbf{E}_j[p_j^{-1} \operatorname{Tr} B_j] = \sum_{i=1}^m \operatorname{Tr} B_i = \operatorname{Tr} I = n$. By Markov's inequality,

$$\mathbf{P}\left(p_j^{-1}\operatorname{Tr} B_j \le \frac{(1+\eta)}{\eta} n\right) = 1 - \mathbf{P}\left(p_j^{-1}\operatorname{Tr} B_j > \frac{(1+\eta)}{\eta} n\right) > 1 - \frac{\eta}{1+\eta} = \frac{1}{1+\eta}. \tag{13}$$

Next note that $\mathbf{E}_j[p_j^{-1}\langle X_2, B_j\rangle] = \sum_{i=1}^m \langle X_2, B_i\rangle = \langle X_2, I\rangle = \operatorname{Tr} X_2$. Together with Markov's inequality, this yields

$$\mathbf{P}\left(p_j^{-1}\langle X_2, B_j\rangle \le (1+\eta)\operatorname{Tr} X_2\right) = 1 - \mathbf{P}\left(p_j^{-1}\langle X_2, B_j\rangle > (1+\eta)\operatorname{Tr} X_2\right) > 1 - \frac{1}{1+\eta}. \tag{14}$$

It follows from (13) and (14) that there exists $j \in [m]$ satisfying

$$p_j^{-1}\langle X_2, B_j\rangle \le (1+\eta)\operatorname{Tr} X_2, \qquad \text{and} \qquad p_j^{-1}\operatorname{Tr} B_j \le \frac{1+\eta}{\eta} n = \rho.$$

Set $\alpha := p_j^{-1}$ and note that

$$\alpha\langle X_1, B_j\rangle = p_j^{-1}\langle X_1, B_j\rangle = \operatorname{Tr} X_1 \ge (1-\eta)\operatorname{Tr} X_1.$$

Hence, j and α satisfy (12). □

The following proposition, proven in Appendix B, shows that the width achieved by Proposition 16 is essentially optimal.

Proposition 17. *Any oracle for satisfying (12) must have $\rho = \Omega(n/\eta)$, even if the B_i matrices have rank one, and even if X_1 is a scalar multiple of X_2^{-1}.*

We also point out that a naive application of MMWUM as stated by Kale in [20] does not work. In his description of MMWUM, the parameter K is fixed as 1. So we must correspondingly adjust our input matrices to be block-diagonal, e.g., C has two blocks: I and $-I$. However, applying Theorem 15 in this manner would lead to a sparsifier with $\Omega(n^2)$ edges. The reason is that the width ρ needs to be $\Omega(n)$, and we must choose $\ell = \rho$ since the spectrum of $\sum_{i=1}^m y_i A_i - C$ is symmetric around zero for any y. Thus, to get the error term on the RHS of (10) to be $\le \varepsilon$, we must take $T = \Omega(n^2)$.

6 Solving Problem 2 by a Tweaked MMWUM

In this section, we modify the method described in the previous section in order to obtain solutions to Problem 2 with only $O(n \log n/\varepsilon^2)$ nonzero entries. This matches the sparsity of the solutions obtained by the Ahlswede-Winter inequality. The main difference from our previous method is a simpler oracle and a refined analysis.

We first state an eigenvalue bound analogous to Theorem 15.

Theorem 18. *Let T be a positive integer. Let $B_1, \dots, B_m \in \mathbb{S}^n_+$ be nonzero. Let $\gamma, \eta, \delta_L, \delta_U > 0$. For any given $X_L, X_U \in \mathbb{S}^n$, consider the system*

$$\begin{aligned} \delta_U &\ge \frac{\exp(\gamma\alpha \operatorname{Tr} B_j) - 1}{\operatorname{Tr} B_j}\langle X_U, B_j\rangle, \\ \delta_L &\le \frac{1 - \exp(-\gamma\alpha \operatorname{Tr} B_j)}{\operatorname{Tr} B_j}\langle X_L, B_j\rangle, \\ \alpha &\in \mathbb{R}_+, \quad j \in [m]. \end{aligned} \tag{15}$$

For each $t \in \{0, \dots, T+1\}$, let $A(t), W_L(t), W_U(t) \in \mathbb{S}^n$, let $\alpha(t) \in \mathbb{R}_+$, and let $j(t) \in [m]$. Suppose the following properties hold:

$$A(t) = \sum_{\tau=1}^{t} \alpha(\tau) B_{j(\tau)}, \qquad \forall t \in \{0, \dots, T\},$$

$$W_U(t+1) = \exp(\gamma A(t)) \qquad \text{and} \qquad W_L(t+1) = \exp(-\gamma A(t)), \qquad \forall t \in \{0, \dots, T\},$$

$$(\alpha, B_j) = (\alpha(t), B_{j(t)}) \text{ is a solution for (15) with } (X_U, X_L) = \left(\frac{W_U(t)}{\operatorname{Tr} W_U(t)}, \frac{W_L(t)}{\operatorname{Tr} W_L(t)}\right), \quad \forall t \in [T].$$

Then

$$\frac{A(T)}{T} \in \left[\frac{\log(1-\delta_L)^{-1}}{\gamma} - \frac{\log n}{T\gamma}, \frac{\log(1+\delta_U)}{\gamma} + \frac{\log n}{T\gamma}\right]. \tag{16}$$

Proof. We will use Golden-Thompson inequality:

$$\operatorname{Tr}(\exp(A+B)) \le \operatorname{Tr}(\exp(A)\exp(B)), \qquad \forall A, B \in \mathbb{S}^n. \tag{17}$$

We will also make use of the following facts. First,

$$\exp(cx) \;\le\; 1 + \frac{\exp(c\cdot b)-1}{b}x \qquad \forall c \in \mathbb{R},\ b>0,\ x \in [0,b].$$

For $X \in \mathbb{S}^n_+$, we have $\lambda_{\max}(X) \le \operatorname{Tr} X$, so $X \in [0, \operatorname{Tr} X]$, and

$$\exp(cX) \;\preceq\; I + \frac{\exp(c\cdot \operatorname{Tr} X)-1}{\operatorname{Tr} X}X. \tag{18}$$

For each $t \in [T+1]$, define $\Phi_L(t) := \operatorname{Tr} W_L(t)$ and $\Phi_U(t) = \operatorname{Tr} W_U(t)$. For each $t \in [T]$,

$$\begin{aligned}
\Phi_U(t+1) &= \operatorname{Tr}\Big(\exp(\gamma A(t))\Big) = \operatorname{Tr}\Big(\exp(\gamma A(t-1) + \gamma\alpha B_j)\Big) \\
&\overset{(17)}{\le} \operatorname{Tr}\Big(\exp(\gamma A(t-1))\exp(\gamma\alpha B_j)\Big) \\
&\overset{(18)}{\le} \operatorname{Tr}\left(\exp(\gamma A(t-1))\Big(\frac{\exp(\gamma\alpha\operatorname{Tr} B_j)-1}{\operatorname{Tr} B_j}B_j + I\Big)\right) \\
&= \frac{\exp(\gamma\alpha\operatorname{Tr} B_j)-1}{\operatorname{Tr} B_j}\operatorname{Tr}(\exp(\gamma A(t-1))B_j) + \operatorname{Tr}(\exp(\gamma A(t-1))) \\
&= \frac{\exp(\gamma\alpha\operatorname{Tr} B_j)-1}{\operatorname{Tr} B_j}\langle W_U(t), B_j\rangle + \Phi_U(t) \\
&\overset{(15)}{\le} (1+\delta_U)\Phi_U(t),
\end{aligned} \tag{19}$$

where we abbreviated $j := j(t)$ and $\alpha := \alpha(t)$.

Since $A(0) = 0$, we have that $\Phi_U(1) = \operatorname{Tr} I = n$. Using (19), after T iterations,

$$\Phi_U(T+1) \le (1+\delta_U)^T n.$$

Thus,

$$\exp(\gamma\lambda_{\max}(A(T))) \le \sum_{i=1}^n \exp(\gamma\lambda_i) = \operatorname{Tr} W_U(T+1) = \Phi_U(T+1) \le (1+\delta_U)^T n,$$

where $\lambda_1, \dots, \lambda_n$ are the eigenvalues of $A(T)$. And so $\gamma\lambda_{\max}(A(T)) \le T\log(1+\delta_U) + \log n$, which implies the upper bound in (16). The proof of the lower bound is analogous. □

Next we establish conditions under which we can construct an oracle for solving the system (15). The proof consists of algebraic manipulations and an averaging argument analogous to the proof of Lemma 3.5 in [4].

Theorem 19. *Let $B_1, \dots, B_m \in \mathbb{S}^n_+$ be nonzero such that $\sum_{i=1}^m B_i = I$. Let $\delta_U, \delta_L > 0$ be such that*

$$\frac{1}{\delta_L} - n \ge \frac{1}{\delta_U}. \tag{20}$$

Then, for any $X_L, X_U \in \mathbb{S}^n_{++}$ with trace one, the system (15) has a solution.

Proof. The first inequality in (15) is equivalent to

$$\frac{\operatorname{Tr} B_j}{\exp(\gamma\alpha \operatorname{Tr} B_j) - 1} \geq \frac{\langle X_U, B_j\rangle}{\delta_U}. \tag{21}$$

Using the identity $\frac{1}{1-1/x} = 1 + \frac{1}{x-1}$, the second inequality in (15) is equivalent to

$$\frac{\operatorname{Tr} B_j}{\exp(\gamma\alpha \operatorname{Tr} B_j) - 1} \leq \frac{\langle X_L, B_j\rangle}{\delta_L} - \operatorname{Tr} B_j. \tag{22}$$

We will choose $j \in [m]$ so that

$$\frac{\langle X_L, B_j\rangle}{\delta_L} - \operatorname{Tr} B_j \geq \frac{\langle X_U, B_j\rangle}{\delta_U} \tag{23}$$

and set α so that (21) holds with equality. Then both (21) and (22) will hold. Note that $\alpha \geq 0$ since $e^{\gamma\alpha \operatorname{Tr} B_j} = 1 + \delta_U \operatorname{Tr} B_j / \langle X_U, B_j\rangle > 1$ and $\gamma \operatorname{Tr} B_j > 0$.

To see that there exists $j \in [m]$ satisfying (23), note that, by (20) and $\sum_{i=1}^m B_i = I$,

$$\sum_{i=1}^m \left[\frac{\langle X_L, B_i\rangle}{\delta_L} - \operatorname{Tr} B_i\right] = \frac{\operatorname{Tr} X_L}{\delta_L} - n = \frac{1}{\delta_L} - n \geq \frac{1}{\delta_U} = \frac{\operatorname{Tr} X_U}{\delta_U} = \sum_{i=1}^m \frac{\langle X_U, B_i\rangle}{\delta_U}.$$

□

Finally, let us show how to set the parameters to get a sparsifier. Given $\varepsilon \in (0, 1)$, set

$$\eta := \varepsilon/2, \qquad \delta_U := \frac{\eta}{n}, \qquad \delta_L := \frac{\eta}{(1+\eta)n}, \qquad T := \frac{n \log n}{\eta^2}. \tag{24}$$

By our choice of δ_L and δ_U, we have $1/\delta_L - n = (1+\eta)n/\eta - n = n/\eta = 1/\delta_U$, so (20) holds with equality. After we run the modified version of MMWUM given by Theorem 18, we obtain a matrix $A(T)$. Set $\bar{A} := A(T)/T$. By Theorem 18,

$$\lambda_{\max}(\bar{A}) \leq \frac{\log(1+\delta_U)}{\gamma} + \frac{\log n}{T\gamma} \leq \left(\delta_U + \frac{\eta^2}{n}\right)/\gamma = \frac{1+\eta}{n\gamma/\eta}.$$

We will use that $-\log(1-x) \geq x$ for $x < 1$. Thus,

$$\lambda_{\min}(\bar{A}) \geq \frac{\log(1-\delta_L)^{-1}}{\gamma} - \frac{\log n}{T\gamma} \geq \left(\delta_L - \frac{\eta^2}{n}\right)/\gamma = \frac{1/(1+\eta) - \eta}{n\gamma/\eta} \geq \frac{1-2\eta}{n\gamma/\eta}.$$

So if we choose $\gamma = \eta/n$ then $(1-\varepsilon)I \preceq \bar{A} \preceq (1+\varepsilon)I$ and $\bar{A}$ is of the form $\sum_i y_i B_i$ with $y \geq 0$ and has at most $T = O(n \log n/\varepsilon^2)$ nonzero entries.

Remark. The choice of γ is actually irrelevant here. We could choose $\gamma > 0$ arbitrarily, then define $\bar{A} = A(T) \cdot (n\gamma/\eta T)$ and the desired conclusion would hold.

7 Comparing our two algorithms

The proof of Theorem 18 defines two potential functions for each iteration t.

$$\begin{aligned}\Phi_U(t) &:= \operatorname{Tr} W_U(t) = \operatorname{Tr}\exp(\gamma A(t))\\ \Phi_L(t) &:= \operatorname{Tr} W_L(t) = \operatorname{Tr}\exp(-\gamma A(t))\end{aligned}$$

The proof shows that, at each iteration, the potentials change as follows:

$$\begin{aligned}\Phi_U(t+1) &\le (1+\delta_U)\Phi_U(t),\\ \Phi_L(t+1) &\le (1-\delta_L)\Phi_L(t).\end{aligned}$$

Due to properties of the exponential function, these inequalities may be equivalently written using potentials that "shift" linearly at each iteration.

We require that

$$\begin{aligned}(1+\delta_U)^{-(t+1)}\cdot\Phi_U(t+1) &\le (1+\delta_U)^{-t}\cdot\Phi_U(t) \qquad \forall t\ge 0,\\ (1-\delta_L)^{-(t+1)}\cdot\Phi_L(t+1) &\le (1-\delta_L)^{-t}\cdot\Phi_L(t) \qquad \forall t\ge 0.\end{aligned}$$

Defining $\Delta_U = \ln(1+\delta_U)$ and $\Delta_L = \ln\left((1-\delta_L)^{-1}\right)$, these inequalities are equivalent to

$$\begin{aligned}\operatorname{Tr}\exp(-(t+1)\Delta_U I + \gamma A(t+1)) &\le \operatorname{Tr}\exp(-t\Delta_U I + \gamma A(t)),\\ \operatorname{Tr}\exp((t+1)\Delta_L I - \gamma A(t+1)) &\le \operatorname{Tr}\exp(t\Delta_L I - \gamma A(t)).\end{aligned} \tag{25}$$

Let us introduce new notation to write these inequalities more succinctly. Define

$$\begin{aligned}\Psi^u(A) &:= \operatorname{Tr}\exp(-uI + \gamma A),\\ \Psi_\ell(A) &:= \operatorname{Tr}\exp(\ell I - \gamma A).\end{aligned}$$

Then, writing $A(t+1) = A(t) + \alpha B_j$, the inequalities in (25) are equivalent to

$$\begin{aligned}\Psi^{(t+1)\Delta_U}(A(t)+\alpha B_j) &\le \Psi^{t\Delta_U}(A(t)),\\ \Psi_{(t+1)\Delta_L}(A(t)+\alpha B_j) &\le \Psi_{t\Delta_L}(A(t)).\end{aligned} \tag{26}$$

Algorithm 2 uses this notation to describe the algorithm of Section 6.

Comparing Algorithms 1 and 2, we notice that they are identical with the exception of different parameters and different potential functions. Thus, we believe that Algorithm 2 sheds some new light on the BSS algorithm — the BSS algorithm can be derived by applying MMWUM to Problem 1, then improving that approach by modifying the potential functions. Indeed, to improve Algorithm 2, one would be tempted to modify the potential functions to more strongly penalize eigenvalues which deviate from the desired range. The natural approach to do this would be to increase the derivatives of the potential function by increasing the parameter γ. However, as remarked at the end of Section 6, the algorithm is actually unaffected by varying γ! Thus, to improve Algorithm 2, one would seek a more substantially different potential function.

Focusing on the upper potential, we consider the question: is there a function $f\colon \mathbb{R}\to\mathbb{R}$ with steeper derivatives than $\exp(u-x)$ and such that, for any matrices A and B, $\operatorname{Tr} f(A+B)$ can be easily related to $\operatorname{Tr} f(A)$? The natural candidates to try are $f(x) = -\log(u-x)$ and $f(x) = (u-x)^{-1}$ since, in both cases, $\operatorname{Tr} f(A+B)$ can be related to $\operatorname{Tr} f(A)$ by the Sherman-Morrison-Woodbury formula. It is not clear whether the choice $f(x) = -\log(u-x)$ can be made to work. However, choosing $f(x) = (u-x)^{-1}$, one arrives at Algorithm 1, our generalization of the BSS algorithm. Nevertheless, even after arriving at this algorithm, the analysis to prove its rapid convergence (namely, Lemma 3.5 in [4]) is rather delicate.

Algorithm 2 A procedure for solving Problem 2 based on the MMWUM method.

procedure SparsifySumOfMatricesByMMWUM($B_1, \ldots, B_m$, ε)

input: Matrices $B_1, \ldots, B_m \in \mathbb{S}^n_+$ such that $\sum_i B_i = I$, and a parameter $\varepsilon \in (0, 1)$.

output: A vector y with $O(n \log n/\varepsilon^2)$ nonzero entries such that $I \preceq \sum_i y_i B_i \preceq (1 + O(\varepsilon))I$.

Initially $A(0) := 0$, and $y(0) := 0$. Set parameters

$$u_0 := 0, \qquad \ell_0 := 0, \qquad \Delta_U := \ln(1 + \delta_U), \qquad \Delta_L := \ln\left((1 - \delta_L)^{-1}\right),$$

where δ_U, δ_L and T are as defined in (24).

Define the potential functions $\Psi^u(A) := \operatorname{Tr}\exp(-uI + \gamma A)$ and $\Psi_\ell(A) := \operatorname{Tr}\exp(\ell I - \gamma A)$.

For $t = 1, \ldots, T$

Set $u_t := u_{t-1} + \Delta_U$ and $\ell_t := \ell_{t-1} + \Delta_L$.

Find a matrix B_j and a value $\alpha > 0$ such that

$$\Psi^{u_t}(A(t-1) + \alpha B_j) \leq \Psi^{u_{t-1}}(A(t-1)) \quad \text{and} \quad \Psi_{\ell_t}(A(t-1) + \alpha B_j) \leq \Psi_{\ell_{t-1}}(A(t-1)).$$

Set $A(t) := A(t-1) + \alpha B_j$ and $y(t) := y(t-1) + \alpha e_j$.

Return $y(T)/\lambda_{\min}(A(T))$.

Acknowledgements

We thank Satyen Kale for helpful discussions.

References

[1] Rudolf Ahlswede and Andreas Winter. Strong converse for identification via quantum channels. *IEEE Transactions on Information Theory*, 48(3):569–579, March 2002. 1, 5

[2] Ingo Althöfer. On sparse approximations to randomized strategies and convex combinations. *Linear Algebra and Applications*, 199:339–355, 1994. 2

[3] Sanjeev Arora and Satyen Kale. A combinatorial, primal-dual approach to semidefinite programs. In *Proceedings of the 39th Annual ACM Symposium on Theory of Computing (STOC)*, 2007. 2, 8

[4] Joshua Batson, Daniel A. Spielman, and Nikhil Srivastava. Twice-Ramanujan sparsifiers. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing (STOC)*, 2009. To appear in SIAM Journal on Scientific Computing. 1, 2, 4, 5, 6, 7, 11, 13

[5] András A. Benczúr and David R. Karger. Approximate s-t min-cuts in $\tilde{O}(n^2)$ time. In *Proceedings of the 28th Annual ACM Symposium on Theory of Computing (STOC)*, 1996. 1

[6] András A. Benczúr and David R. Karger. Randomized approximation schemes for cuts and flows in capacitated graphs, 2002. arXiv:cs/0207078. 1

[7] Moses Charikar, Chandra Chekuri, Ashish Goel, Sudipto Guha, and Serge A. Plotkin. Approximating a finite metric by a small number of tree metrics. In *Proceedings of the 39th Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 1998. 8

[8] Yoav Freund and Robert E. Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29:79–103, 1999. 8

[9] Wai Shing Fung, Ramesh Hariharan, Nicholas J. A. Harvey, and Debmalya Panigrahi. A general framework for graph sparsification. In *Proceedings of the 43rd Annual ACM Symposium on Theory of Computing (STOC)*, 2011. 1

[10] Wai Shing Fung and Nicholas J. A. Harvey. Graph sparsification by edge-connectivity and random spanning trees, May 2010. `http://arxiv.org/abs/1005.0265`. 1

[11] Ashish Goel, Michael Kapralov, and Sanjeev Khanna. Graph sparsification via refinement sampling, April 2010. `http://arxiv.org/abs/1004.4915`. 1

[12] William W. Hager. Updating the inverse of a matrix. *SIAM Review*, 31(2):221–239, 1989. 6

[13] Ramesh Hariharan and Debmalya Panigrahi. A general framework for graph sparsification, April 2010. `http://arxiv.org/abs/1004.4080`. 1

[14] Ramesh Hariharan and Debmalya Panigrahi. A linear-time algorithm for sparsification of unweighted graphs, May 2010. `http://arxiv.org/abs/1005.0670`. 1

[15] Nicholas J. A. Harvey. Lecture notes for C&O 750: Randomized algorithms, 2011. `http://www.math.uwaterloo.ca/~harvey/W11/Lecture11Notes.pdf`. 1

[16] Roger A. Horn and Charles R. Johnson. *Matrix analysis*. Cambridge University Press, Cambridge, 1990. Corrected reprint of the 1985 original. 6

[17] Garud Iyengar, David J. Phillips, and Clifford Stein. Approximation algorithms for semidefinite packing problems with applications to maxcut and graph coloring. In Michael Jünger and Volker Kaibel, editors, *Integer Programming and Combinatorial Optimization*, volume 3509 of *Lecture Notes in Computer Science*, pages 77–90. Springer Berlin / Heidelberg, 2005. 3

[18] Garud Iyengar, David J. Phillips, and Clifford Stein. Approximating semidefinite packing programs. *SIAM Journal on Optimization*, 21(1):231–268, 2011. 3

[19] Rahul Jain and Penghui Yao. A parallel approximation algorithm for positive semidefinite programming. In *The 52nd Annual IEEE Symposium on Foundations of Computer Science (FOCS 2011)*, 2011. (to appear). 3

[20] Satyen Kale. *Efficient Algorithms using the Multiplicative Weights Update Method*. PhD thesis, Princeton University, 2007. Princeton Tech Report TR-804-07. 2, 10, 17

[21] Jonathan A. Kelner and Alex Levin. Spectral sparsification in the semi-streaming setting. In *Proceedings of the 28th International Symposium on Theoretical Aspects of Computer Science (STACS)*, pages 440–451, 2011. 1

[22] Alexandra Kolla, Yury Makarychev, Amin Saberi, and Shang-Hua Teng. Subgraph sparsification and nearly optimal ultrasparsifiers. In *Proceedings of the 42nd Annual ACM Symposium on Theory of Computing (STOC)*, 2010. 3

[23] Ioannis Koutis, Gary L. Miller, and Richard Peng. Approaching optimality for solving SDD systems. In *Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2010. 1

[24] Ilan Newman and Yuri Rabinovich. Finite volume spaces and sparsification, 2010. http://arxiv.org/abs/1002.3541. 1

[25] Lorenzo Orecchia and Nisheeth K. Vishnoi. Towards an SDP-based approach to spectral methods: A nearly-linear time algorithm for graph partitioning and decomposition. In *Proceedings of the 22nd Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 532–545, 2011. 2

[26] Juan A. Rodríguez. On the Laplacian eigenvalues and metric parameters of hypergraphs. *Linear Multilinear Algebra*, 50(1):1–14, 2002. 4

[27] Mark Rudelson. Random vectors in the isotropic position. *J. of Functional Analysis*, 164(1):60–72, 1999. 1

[28] Mark Rudelson and Roman Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *Journal of the ACM*, 54(4), 2007. 1

[29] Gideon Schechtman. Tight embedding of subspaces of L_p in ℓ_p^n for even p. *Proceedings of the AMS*. To appear. 1

[30] Anthony Man-Cho So, Yinyu Ye, and Jiawei Zhang. A unified theorem on SDP rank reduction. *Mathematics of Operations Research*, 33(4):910–920, 2008. 8

[31] Daniel A. Spielman and Nikhil Srivastava. An elementary proof of the restricted invertibility theorem. *Israel J. Math.* To appear. 1

[32] Daniel A. Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing (STOC)*, pages 563–568, 2008. 1, 5

[33] Daniel A. Spielman and Shang-Hua Teng. Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In *Proceedings of the 36th Annual ACM Symposium on Theory of Computing (STOC)*, pages 81–90, 2004. 1

[34] Nikhil Srivastava. On contact points of convex bodies, 2009. http://www.cs.yale.edu/homes/srivastava/papers/contact.pdf. 1

[35] Roman Vershynin. A note on sums of independent random matrices after Ahlswede-Winter, 2008. http://www-personal.umich.edu/~romanv/teaching/reading-group/ahlswede-winter.pdf. 1

[36] Neal Young. Greedy algorithms by derandomizing unknown distributions. Technical Report 1087, Department of ORIE, Cornell University, March 1994. http://hdl.handle.net/1813/8971. 8

[37] Neal Young. Randomized rounding without solving the linear program. In *Proceedings of the 6th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 170–178, 1995. 8

A The MMWUM

In this section we provide some proofs about the MMWUM. These proofs are due to Kale [20]. Our set up and conclusions are slightly different and we modified the proofs accordingly. We reproduce the proofs here for the sake of completeness.

Theorem 15 can be viewed as a block-friendly version of MMWUM. First we show the version with only one block. It is basically the same as [20, Theorem 13 in Chapter 4].

Theorem 20. *Let T be a positive integer. Let $C, A_1, \dots, A_m \in \mathbb{S}^n$. Let $\eta > 0$ and $0 < \beta \le 1/2$. For any given $X \in \mathbb{S}^n$, consider the system*

$$\sum_{i=1}^m y_i \langle A_i, X\rangle \ge \langle C, X\rangle - \eta \operatorname{Tr} X, \quad \text{and} \quad y \in \mathbb{R}^m_+. \tag{27}$$

Let $\{\mathcal{P}, \mathcal{N}\}$ be a partition of $[T]$, let $0 < \ell \le \rho$, and let $W^{(t)} \in \mathbb{S}^n$ and $\ell^{(t)} \in \mathbb{R}$ for $t \in [T+1]$. Let $y^{(t)} \in \mathbb{R}^m$ for $t \in [T]$. Suppose the following properties hold:

$$W^{(t+1)} = \exp\left(-\frac{\beta}{\ell+\rho}\sum_{\tau=1}^t \left[\sum_{i=1}^m y_i^{(\tau)} A_i - C + \ell^{(\tau)} I\right]\right), \qquad \forall t \in \{0, \dots, T\},$$

$$y = y^{(t)} \text{ is a solution for (27) with } X = W^{(t)}, \qquad \forall t \in [T],$$

$$\sum_{i=1}^m y_i^{(t)} A_i - C \in \begin{cases} [-\ell, \rho], & \text{if } t \in \mathcal{P}, \\ [-\rho, \ell], & \text{if } t \in \mathcal{N}, \end{cases} \qquad \forall t \in [T],$$

$$\ell^{(t)} = \ell, \quad \forall t \in \mathcal{P}, \quad \text{and} \quad \ell^{(t)} = -\ell, \quad \forall t \in \mathcal{N}.$$

Define $\bar{y} := \frac{1}{T}\sum_{t=1}^T y^{(t)}$. Then

$$\sum_{i=1}^m \bar{y}_i A_i - C \succeq -\left[\beta\ell + \frac{(\rho+\ell)\ln n}{T\beta} + (1+\beta)\eta\right] I. \tag{28}$$

The main tool for the proof of Theorem 20 is the following result:

Theorem 21 (Kale [20, Corollary 3 in Chapter 3])**.** *Let $0 < \beta \le 1/2$. Let T be a positive integer. Let $\{\mathcal{P}, \mathcal{N}\}$ be a partition of $[T]$, and let $M^{(t)} \in \mathbb{S}^n$ for $t \in [T]$ and $W^{(t)} \in \mathbb{S}^n$ for $t \in [T+1]$ with the following properties:*

$$W^{(t+1)} = \exp\left(-\beta \sum_{\tau=1}^t M^{(\tau)}\right) \quad \forall t = 0, \dots, T,$$

$$0 \preceq M^{(t)} \preceq I, \quad \forall t \in \mathcal{P}, \qquad \text{and} \qquad -I \preceq M^{(t)} \preceq 0, \quad \forall t \in \mathcal{N},$$

Let

$$P^{(t)} := \frac{1}{\operatorname{Tr} W^{(t)}} W^{(t)}, \qquad \forall t \in [T].$$

Then

$$(1-\beta)\sum_{t\in\mathcal{P}} \langle M^{(t)}, P^{(t)}\rangle + (1+\beta)\sum_{t\in\mathcal{N}} \langle M^{(t)}, P^{(t)}\rangle \le \lambda_{\min}\left(\sum_{t=1}^T M^{(t)}\right) + \frac{\ln n}{\beta}. \tag{29}$$

Proof. Set $\Phi^{(t)} := \mathrm{Tr}(W^{(t)})$ for $t \in [T+1]$. Put $\beta_1 := 1 - e^{-\beta}$ and $\beta_2 := e^{\beta} - 1$. Then, for any $t \in [T]$,

$$\begin{aligned}
\Phi^{(t+1)} = \mathrm{Tr}(W^{(t+1)}) &= \mathrm{Tr}\left(\exp\left(-\beta \sum_{\tau=1}^{t} M^{(\tau)}\right)\right) \\
&\le \mathrm{Tr}\left(\exp\left(-\beta \sum_{\tau=1}^{t-1} M^{(\tau)}\right) \exp\left(-\beta M^{(t)}\right)\right) = \mathrm{Tr}\left(W^{(t)} \exp(-\beta M^{(t)})\right) \\
&= \langle W^{(t)}, \exp(-\beta M^{(t)})\rangle,
\end{aligned}$$

where we have used Golden-Thompson's inequality (17).

Using the fact that e^x is convex, one can prove that

$$\begin{aligned}
0 \preceq A \preceq I \implies \exp(-\beta A) &\preceq I - \beta_1 A, \\
-I \preceq A \preceq 0 \implies \exp(-\beta A) &\preceq I - \beta_2 A.
\end{aligned}$$

Suppose that $t \in \mathcal{P}$. Then $\exp(-\beta M^{(t)}) \preceq I - \beta_1 M^{(t)}$, and since $W^{(t)} \succeq 0$, we get

$$\begin{aligned}
\Phi^{(t+1)} &\le \langle W^{(t)}, \exp(-\beta M^{(t)})\rangle \le \langle W^{(t)}, I - \beta_1 M^{(t)}\rangle \\
&= \mathrm{Tr}(W^{(t)}) - \beta_1 \langle W^{(t)}, M^{(t)}\rangle \\
&= \mathrm{Tr}(W^{(t)}) - \mathrm{Tr}(W^{(t)})\beta_1 \langle P^{(t)}, M^{(t)}\rangle \\
&= \mathrm{Tr}(W^{(t)})\Big[1 - \beta_1 \langle P^{(t)}, M^{(t)}\rangle\Big] \\
&= \Phi^{(t)}\Big[1 - \beta_1 \langle P^{(t)}, M^{(t)}\rangle\Big] \\
&\le \Phi^{(t)} \exp(-\beta_1 \langle P^{(t)}, M^{(t)}\rangle).
\end{aligned}$$

Similarly, if $t \in \mathcal{N}$, then

$$\Phi^{(t+1)} \le \Phi^{(t)} \exp(-\beta_2 \langle P^{(t)}, M^{(t)}\rangle).$$

By induction on t, and using $\Phi^{(1)} = \mathrm{Tr}(I) = n$, we get

$$\Phi^{(t+1)} \le n \exp\left(-\beta_1 \sum_{\tau \in \mathcal{P} \cap [t]} \langle M^{(\tau)}, P^{(\tau)}\rangle - \beta_2 \sum_{\tau \in \mathcal{N} \cap [t]} \langle M^{(\tau)}, P^{(\tau)}\rangle\right), \quad \forall t \in [T].$$

For every $A \in \mathbb{S}^n$, we have $\mathrm{Tr}(\exp(A)) = \sum_{i=1}^{n} e^{\lambda_i} \ge e^{\lambda_j}$ for any $j \in [n]$, where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A. Thus,

$$\begin{aligned}
\Phi^{(T+1)} = \mathrm{Tr}(W^{(T+1)}) &= \mathrm{Tr}\left(\exp\left(-\beta \sum_{t=1}^{T} M^{(t)}\right)\right) \\
&\ge \exp\left(\lambda_{\max}\left(-\beta \sum_{t=1}^{T} M^{(t)}\right)\right) = \exp\left(-\beta \lambda_{\min}\left(\sum_{t=1}^{T} M^{(t)}\right)\right).
\end{aligned}$$

Thus,

$$\exp\left[-\beta \lambda_{\min}\left(\sum_{t=1}^{T} M^{(t)}\right)\right] \le n \exp\left[-\beta_1 \sum_{t \in \mathcal{P}} \langle M^{(t)}, P^{(t)}\rangle - \beta_2 \sum_{t \in \mathcal{N}} \langle M^{(t)}, P^{(t)}\rangle\right].$$

By taking $\ln(\cdot)$ on both sides, we get

$$-\beta\lambda_{\min}\left(\sum_{t=1}^{T} M^{(t)}\right) \le \ln n - \left[\beta_1 \sum_{t\in\mathcal{P}} \langle M^{(t)}, P^{(t)}\rangle + \beta_2 \sum_{t\in\mathcal{N}} \langle M^{(t)}, P^{(t)}\rangle\right],$$

so

$$\beta_1 \sum_{t\in\mathcal{P}} \langle M^{(t)}, P^{(t)}\rangle + \beta_2 \sum_{t\in\mathcal{N}} \langle M^{(t)}, P^{(t)}\rangle \le \beta\lambda_{\min}\left(\sum_{t=1}^{T} M^{(t)}\right) + \ln n,$$

and

$$\frac{\beta_1}{\beta} \sum_{t\in\mathcal{P}} \langle M^{(t)}, P^{(t)}\rangle + \frac{\beta_2}{\beta} \sum_{t\in\mathcal{N}} \langle M^{(t)}, P^{(t)}\rangle \le \lambda_{\min}\left(\sum_{t=1}^{T} M^{(t)}\right) + \frac{\ln n}{\beta}.$$

Since $\sum_{t\in\mathcal{P}} \langle M^{(t)}, P^{(t)}\rangle \ge 0$ and $\sum_{t\in\mathcal{N}} \langle M^{(t)}, P^{(t)}\rangle \le 0$, to prove (29) it suffices to show that $1-\beta \le \beta_1/\beta$ and $1+\beta \ge \beta_2/\beta$. It is not hard to prove that

$$1 - e^{-x} \ge x(1-x),\ \forall x \in [0, +\infty) \qquad \text{and} \qquad e^x - 1 \le x(1+x),\ \forall x \in [0, \tfrac{1}{2}]$$

So our choice of β_1 and β_2 ensures that $1-\beta \le \beta_1/\beta$ and $1+\beta \ge \beta_2/\beta$. □

We can now show the proof of Theorem 20.

Proof of Theorem 20. Let $M^{(t)} := \frac{1}{\ell+\rho}\left[\sum_{i=1}^{m} y_i^{(t)} A_i - C + \ell^{(t)} I\right]$ and $P^{(t)} := W^{(t)}/\operatorname{Tr} W^{(t)}$ for every t. For every $t \le T$, using (27),

$$\begin{aligned}\langle M^{(t)}, P^{(t)}\rangle &= \frac{1}{\ell+\rho}\left[\sum_{i=1}^{m} y_i^{(t)} \langle A_i, P^{(t)}\rangle - \langle C, P^{(t)}\rangle + \ell^{(t)} \langle I, P^{(t)}\rangle\right] \\ &= \frac{1}{(\ell+\rho)\operatorname{Tr} W^{(t)}}\left[\sum_{i=1}^{m} y_i^{(t)} \langle A_i, W^{(t)}\rangle - \langle C, W^{(t)}\rangle\right] + \frac{\ell^{(t)}}{\ell+\rho} \ge -\frac{\eta}{\ell+\rho} + \frac{\ell^{(t)}}{\ell+\rho},\end{aligned}$$

since $y^{(t)}$ is a solution for (27) with $X := W^{(t)}$. Thus, by (29),

$$\begin{aligned}\sum_{t\in\mathcal{P}} \frac{(1-\beta)(\ell^{(t)} - \eta)}{\ell+\rho} + \sum_{t\in\mathcal{N}} \frac{(1+\beta)(\ell^{(t)} - \eta)}{\ell+\rho} \\ \le \frac{1}{\rho+\ell}\lambda_{\min}\left(\sum_{t=1}^{T}\left[\left(\sum_{i=1}^{m} y_i^{(t)} A_i\right) - C + \ell^{(t)} I\right]\right) + \frac{\ln n}{\beta}.\end{aligned}$$

Multiply through by $\ell+\rho$ and move $\ell^{(t)} I$ out of $\lambda_{\min}(\cdot)$:

$$\begin{aligned}\sum_{t\in\mathcal{P}} (1-\beta)\ell^{(t)} + \sum_{t\in\mathcal{N}} (1+\beta)\ell^{(t)} - T(1+\beta)\eta \\ \le \lambda_{\min}\left(\sum_{t=1}^{T}\left[\left(\sum_{i=1}^{m} y_i^{(t)} A_i\right) - C\right]\right) + \left(\sum_{t=1}^{T} \ell^{(t)}\right) + \frac{(\rho+\ell)\ln n}{\beta}.\end{aligned}$$

Thus,

$$\sum_{t\in\mathcal{P}} -\beta\ell^{(t)} + \sum_{t\in\mathcal{N}} \beta\ell^{(t)} \le \lambda_{\min}\left(\sum_{t=1}^{T}\left[\left(\sum_{i=1}^{m} y_i^{(t)} A_i\right) - C\right]\right) + \frac{(\rho+\ell)\ln n}{\beta} + T(1+\beta)\eta.$$

Next note that $\sum_{t\in\mathcal{P}} -\ell^{(t)} + \sum_{t\in\mathcal{N}} \ell^{(t)} = \sum_{t\in\mathcal{P}} -\ell + \sum_{t\in\mathcal{N}} -\ell = -T\ell$, so

$$0 \le \lambda_{\min}\left(\sum_{t=1}^{T}\left[\left(\sum_{i=1}^{m} y_i^{(t)} A_i\right) - C\right]\right) + \beta T\ell + \frac{(\rho+\ell)\ln n}{\beta} + T(1+\beta)\eta.$$

and

$$0 \le \lambda_{\min}\left(\frac{1}{T}\sum_{t=1}^{T}\left[\left(\sum_{i=1}^{m} y_i^{(t)} A_i\right) - C\right]\right) + \beta\ell + \frac{(\rho+\ell)\ln n}{T\beta} + (1+\beta)\eta.$$

Thus,

$$\sum_{i=1}^{m} \bar{y}_i A_i - C = \frac{1}{T}\sum_{t=1}^{T}\left[\left(\sum_{i=1}^{m} y_i^{(t)} A_i\right) - C\right] \succeq -\left[\beta\ell + \frac{(\rho+\ell)\ln n}{T\beta} + (1+\beta)\eta\right] I.$$

□

Theorem 15 can be easily proved from Theorem 20. First, we apply Theorem 20 separately for each block. In each iteration, $y^{(t)}$ is a solution for (27) for all blocks simultaneously, and so the conclusion in (28) holds for all blocks with same $\bar{y}$. This new algorithm can be seen as equivalent to running K copies of MMWUM, each with different input data, with the caveat that all copies run for the same number of iterations and the vector $y^{(t)}$ returned from the oracle is the same for all copies at each iteration t.

B Optimality of MMWUM Oracle

Proposition 17. *Any oracle for satisfying (12) must have $\rho = \Omega(n/\eta)$, even if the B_i matrices have rank one, and even if X_1 is a scalar multiple of X_2^{-1}.*

Proof. Let $k = n/3$, let I_k be the identity of size $k \times k$, and let $e_j \in \mathbb{R}^k$ be the jth standard basis vector. Let $\zeta = 3\eta$ and define

$$X_1 = \operatorname{diag}(1, \zeta^3, \zeta) \otimes I_k, \qquad X_2 = \operatorname{diag}(1, 1/\zeta^3, 1/\zeta) \otimes I_k,$$

where $\otimes$ denotes tensor product. For $j = 1, \ldots, k$, define

$$v_{1,j} = [1/\sqrt{2}, -1/\sqrt{2}, 0] \otimes e_j, \qquad v_{2,j} = [1/\sqrt{2}, 1/\sqrt{2}, 0] \otimes e_j, \qquad v_{3,j} = [0, 0, 1] \otimes e_j.$$

Let $B_{i,j} = v_{i,j} v_{i,j}^T$. Note that $\sum_{i,j} B_{i,j} = I$.

The oracle cannot choose a matrix $B_{i,j}$ with $i \in \{1, 2\}$, since satisfying (12) would lead to a contradiction:

$$\frac{\langle X_2, B_i\rangle}{\operatorname{Tr}(X_2)(1+\eta)} \le \frac{1}{\alpha} \le \frac{\langle X_1, B_i\rangle}{\operatorname{Tr}(X_1)(1-\eta)}$$

$$\implies \quad 1 + 3\eta \;=\; 1 + \zeta \;<\; \frac{\langle X_2, B_i\rangle / \operatorname{Tr} X_2}{\langle X_1, B_i\rangle / \operatorname{Tr} X_1} \;\le\; \frac{1+\eta}{1-\eta} \;<\; 1 + 3\eta,$$

for sufficiently small η.

So the oracle must choose a matrix $B_{i,j}$ with $i = 3$. In this case,

$$\frac{\operatorname{Tr} B_{i,j}}{\rho} \le \frac{1}{\alpha} \le \frac{\langle X_1, B_{i,j}\rangle}{\operatorname{Tr}(X_1)(1-\eta)}$$

$$\implies \quad \frac{n}{9\eta} \;=\; \frac{n}{3\zeta} \;\le\; \frac{(1+\zeta^3+\zeta)k}{\zeta} \;=\; \frac{\operatorname{Tr}(B_{i,j})\operatorname{Tr}(X_1)}{\langle X_1, B_{i,j}\rangle} \;\le\; \frac{\rho}{1-\eta}.$$

This shows that $\rho = \Omega(n/\eta)$. □

C The positive semidefiniteness assumption

Proposition 22. *For every positive integer n, there exist matrices $B_1, \dots, B_m \in \mathbb{S}^n$ with $m = \Omega(n^2)$ such that $B := \sum_i B_i$ is positive definite and with the following property: for every $\varepsilon \in (0,1)$ and $y \in \mathbb{R}^m$ such that $(1-\varepsilon)B \preceq \sum_i y_i B_i$, all entries of y are nonzero.*

Proof. Let $\mathcal{P} := \{\, (i,j) : i,j \in [n],\, i < j\}$. For $(i,j) \in \mathcal{P}$, let $E_{ij} := e_i e_j^T + e_j e_i^T$. Let J denote the matrix of all ones. Then $2I + \sum_{(i,j)\in\mathcal{P}} E_{ij} = I + J =: B \succ 0$. Let $\varepsilon \in (0,1)$ and suppose that $(1-\varepsilon)B \preceq 2tI + \sum_{(i,j)\in\mathcal{P}} z_{ij} E_{ij}$ for some $t \in \mathbb{R}$ and $z \in \mathbb{R}^{\mathcal{P}}$. By taking the inner product with E_{ab} on both sides, we see that $0 < 2(1-\varepsilon) \le z_{ab}$ for every $(a,b) \in \mathcal{P}$. Similarly, we find that $0 < 2n(1-\varepsilon) \le 2nt$. □