On approximations of the PSD cone by a polynomial number of smaller-sized PSD cones

On approximations of the PSD cone by a polynomial number of smaller-sized PSD cones

Noname manuscript No.
(will be inserted by the editor)

On Approximations of the PSD Cone by a Polynomial Number of Smaller-sized PSD Cones

Dogyoon Song · Pablo A. Parrilo

Received: date / Accepted: date

Abstract We study the problem of approximating the cone of positive semidefinite (PSD) matrices with a cone that can be described by smaller-sized PSD constraints. Specifically, we ask the question: "how closely can we approximate the set of unit-trace $n \times n$ PSD matrices, denoted by D, using at most N number of $k \times k$ PSD constraints?" In this paper, we prove lower bounds on N to achieve a good approximation of D by considering two constructions of an approximating set. First, we consider the unit-trace $n \times n$ symmetric matrices that are PSD when restricted to a fixed set of k-dimensional subspaces in $\mathbb{R}^n$. We prove that if this set is a good approximation of D, then the number of subspaces must be at least exponentially large in n for any $k = o(n)$. Second, we show that any set S that approximates D within a constant approximation ratio must have superpolynomial $\boldsymbol{S}_+^k$-extension complexity. To be more precise, if S is a constant factor approximation of D, then S must have $\boldsymbol{S}_+^k$-extension complexity at least $\exp(C \cdot \min\{\sqrt{n}, n/k\})$ where C is some absolute constant. In addition, we show that any set S such that $D \subseteq S$ and the Gaussian width of S is at most a constant times larger than the Gaussian width of D must have $\boldsymbol{S}_+^k$-extension complexity at least $\exp(C \cdot \min\{n^{1/3}, \sqrt{n/k}\})$. These results imply that the cone of $n \times n$ PSD matrices cannot be approximated by a polynomial number of $k \times k$ PSD constraints for any $k = o(n/\log^2 n)$. These results generalize the recent work of Fawzi [11] on the hardness of polyhedral approximations of $\boldsymbol{S}_+^n$, which corresponds to the special case with $k = 1$.

1 Introduction

Semidefinite programming (SDP) is a branch of convex optimization that considers problems of the form

$$\begin{aligned} \text{maximize} \quad & \langle C, X \rangle \\ \text{subject to} \quad & \langle A_i, X \rangle = b_i, \quad i = 1, \dots, m, \\ & X \in \boldsymbol{S}_+^n, \end{aligned} \tag{1}$$

where C, A_i's are $n \times n$ symmetric matrices and $\boldsymbol{S}_+^n$ denotes the cone of $n \times n$ positive semidefinite (PSD) matrices. SDP has attracted great interest in many fields as a powerful tool to provide theoretical guarantees as well as practical algorithms. Although current SDP solvers utilizing interior-point methods can solve an SDP up to arbitrary accuracy, they suffer from large computational cost and memory requirement when n is large. Indeed, such scalability issues remain as major challenges for researchers in the field.

To deal with these problems, one can seek an alternative formulation of the problem (1) that is computationally more tractable. For instance, one can try to replace the PSD cone with a computationally more tractable convex cone $\mathcal{K}$ to approximate the feasible set. If $\mathcal{K}$ is a polyhedral cone, we obtain a linear programming (LP) approximation of (1), and if $\mathcal{K}$ is a second-order cone, then we get a second-order cone programming (SOCP) approximation [3], etc. These approximate conic programs can be solved potentially much faster than the original SDP, but possibly at the expense of the quality of the solution.

Dogyoon Song
University of Michigan, Ann Arbor
E-mail: dogyoons@umich.edu

Pablo A. Parrilo
Massachusetts Institute of Technology
E-mail: parrilo@mit.edu

There arises an inevitable question: "how much error is incurred in the optimal value of (1) when $\boldsymbol{S}_+^n$ is replaced by $\mathcal{K}$?" We study this problem by asking the following question:

"How closely can we approximate $\boldsymbol{S}_+^n$ with a cone $\mathcal{K}$ that can be described using at most N number of $k \times k$ PSD constraints?"

In this work, we focus on global, non-adaptive approximations of $\boldsymbol{S}_+^n$ that do not make use of the problem data $C, (A_i, b_i)_{i=1}^m$. We remark that there have been proposed some adaptive approaches to locally (only in the direction of C) approximate $\boldsymbol{S}_+^n$, e.g. in [1, 2], but their analysis is beyond the scope of this paper.

1.1 Contributions

To formally state the question above, we need to specify the notion of approximation as well as what 'a cone $\mathcal{K}$ that can be described using at most N number of $k \times k$ PSD constraints' means.

Notions of Approximation. First, we specify the notions of approximation for cones as follows. Let $H = \{X \in \boldsymbol{S}^n : \operatorname{Tr} X = 1\}$ and let $B_H(\mathcal{K}) = (\mathcal{K} \cap H) - \frac{1}{n} I_n$ for any cone $\mathcal{K}$, where I_n is the $n \times n$ identity matrix. That is, $B_H(\mathcal{K})$ is the unit-trace affine section of $\mathcal{K}$ translated by $-\frac{1}{n} I_n$; note that $0 \in B_H(\boldsymbol{S}_+^n)$. For $\epsilon > 0$, we say $\mathcal{K}$ is an *ϵ-approximation* of $\boldsymbol{S}_+^n$ if $B_H(\boldsymbol{S}_+^n) \subseteq B_H(\mathcal{K}) \subseteq (1+\epsilon) B_H(\boldsymbol{S}_+^n)$.

This notion of approximation is natural and closely related to quantifying the difference in the optimal value (optimality gap) induced by relaxing $\boldsymbol{S}_+^n$ to $\mathcal{K}$. Suppose that we are given a problem of the form (1) with $m = 1$, $A_1 = I_n$, and $b_1 = 1$, and we relax the problem by replacing $\boldsymbol{S}_+^n$ with a cone $\mathcal{K} \supseteq \boldsymbol{S}_+^n$. If $\mathcal{K}$ is an ϵ-approximation of $\boldsymbol{S}_+^n$, then the relative optimality gap is at most ϵ for all $C \in \boldsymbol{S}^n$.

We also define two auxiliary notions of approximation for the convenience of our analysis. Observe that the notion of ϵ-approximation requires $B_H(\mathcal{K})$ to approximate $B_H(\boldsymbol{S}_+^n)$ well in all directions in the ambient space. We introduce more lenient notions of approximation by requiring the relative optimality gap to be small only on average for randomized C with standard Gaussian distribution. Specifically, $\mathcal{K}$ is called an ϵ-approximation of $\boldsymbol{S}_+^k$ in the *average sense* if $B_H(\boldsymbol{S}_+^n) \subseteq B_H(\mathcal{K})$ and $w_G\big(B_H(\mathcal{K})\big) \leq (1+\epsilon) \cdot w_G\big(B_H(\boldsymbol{S}_+^n)\big)$ where $w_G(S) = \mathbb{E}_g[\sup_{x \in S} \langle g, x \rangle]$ denotes the Gaussian width of S. Likewise, $\mathcal{K}$ is called an ϵ-approximation of $\boldsymbol{S}_+^k$ in the *dual-average sense* if $B_H(\boldsymbol{S}_+^n) \subseteq B_H(\mathcal{K})$ and $w_G\big(B_H(\mathcal{K})^\circ\big) \geq \frac{1}{1+\epsilon} \cdot w_G\big(B_H(\boldsymbol{S}_+^n)^\circ\big)$. More details about these notions can be found in Section 3.

k-PSD Approximations of $\boldsymbol{S}_+^n$. In Section 4, we consider approximating $\boldsymbol{S}_+^n$ by enforcing PSD constraints on a fixed set of k-dimensional subspaces in $\mathbb{R}^n$. We begin by formally defining the notion of k-PSD approximations to $\boldsymbol{S}_+^n$.

Definition 1 (k-PSD approximation) Let $\mathcal{V} = \{V_1, \ldots, V_N\}$ be a set of k-dimensional subspaces of $\mathbb{R}^n$. The k-PSD approximation of $\boldsymbol{S}_+^n$ induced by $\mathcal{V}$ is the convex cone

$$\boldsymbol{S}_+^{n,k}(\mathcal{V}) := \big\{X \in \boldsymbol{S}^n : v^T X v \geq 0,\ \forall v \in V_i,\ \forall i = 1, \ldots, N\big\}.$$

Note that $\boldsymbol{S}_+^n \subseteq \boldsymbol{S}_+^{n,k}(\mathcal{V})$ and that $\boldsymbol{S}_+^{n,k}(\mathcal{V})$ is characterized using at most $N = |\mathcal{V}|$ number of $k \times k$ PSD constraints. A prominent example is the so-called sparse k-PSD approximation, denoted by $\boldsymbol{S}_+^{n,k}$, which is a k-PSD approximation of $\boldsymbol{S}_+^n$ induced by the collection of $\binom{n}{k}$ subspaces of k-sparse vectors in $\mathbb{R}^n$.

Our first main results (Theorem 1 and Corollary 1) state that if $\boldsymbol{S}_+^{n,k}(\mathcal{V})$ is a dual-average ϵ-approximation of $\boldsymbol{S}_+^n$, then $N \geq \exp\big(n \cdot \max\{1/(1+\epsilon) - \sqrt{k/n},\ 0\}^2\big)$ is necessary, regardless of the choice of the subspaces $V_1, \ldots, V_N$; see Remark 8 in Section 4.1. For instance, Corollary 1 implies that for any $\epsilon > 0$, $\boldsymbol{S}_+^{n,k}$ cannot be a dual-average ϵ-approximation of $\boldsymbol{S}_+^n$ unless $k = \Omega_n(n)$. We draw a similar conclusion for approximating $\boldsymbol{S}_+^n$ with $\boldsymbol{S}_+^{n,k}$ in the average sense (Corollary 2).

We remark that the conclusion of Theorem 1 (and Corollaries 1 & 2) is possibly too conservative, especially when the subspaces have overlaps. It is because Theorem 1 only takes the number of subspaces into consideration, and is oblivious to the configuration of the subspaces in $\mathcal{V}$. In Section 4.2, we elaborate on this point with an example of the sparse k-PSD approximation. Although Corollary 1 already suggests that k must scale at least linearly as n in order for $\boldsymbol{S}_+^{n,k}$ to approximate $\boldsymbol{S}_+^n$, it becomes uninformative once k/n exceeds a certain threshold (approximately 0.137); see Section 4.2.1 and Figure 4b.

In Section 4.2.2, a tailored analysis for the sparse k-PSD approximation is provided. To be specific, we consider a carefully designed matrix in $\boldsymbol{S}_+^{n,k} \setminus \boldsymbol{S}_+^n$ to show $\epsilon^*(\boldsymbol{S}_+^n, \boldsymbol{S}_+^{n,k}) \geq \frac{n-k}{k-1}$ (Proposition 1) where $\epsilon^*(\boldsymbol{S}_+^n, \boldsymbol{S}_+^{n,k}) := \inf\{\epsilon > 0 : \boldsymbol{S}_+^{n,k}$ is an ϵ-approximation of $\boldsymbol{S}_+^n\}$. Furthermore, we prove a sharper lower bound for $\epsilon^*_{\text{dual-avg}}(\boldsymbol{S}_+^n, \boldsymbol{S}_+^{n,k})$ that is strictly positive for all $1 \leq k < n$, using the duality between $\boldsymbol{S}_+^{n,k}$ and the cone of factor width at most k (Proposition 2). See Figure 1a for comparison between these tailored results and the weak bound obtained from Corollary 1.

Table 1: Overview of our results about the hardness of approximating $\boldsymbol{S}^n_+$ with $\boldsymbol{S}^k_+$, presented in terms of the number, N, of the $k \times k$ PSD constraints needed to construct an ϵ-approximation of $\boldsymbol{S}^n_+$. Here, $C_1, C_2 > 0$ are some universal constants and $\gtrsim$ indicates that the inequality holds in the limit $n \to \infty$.

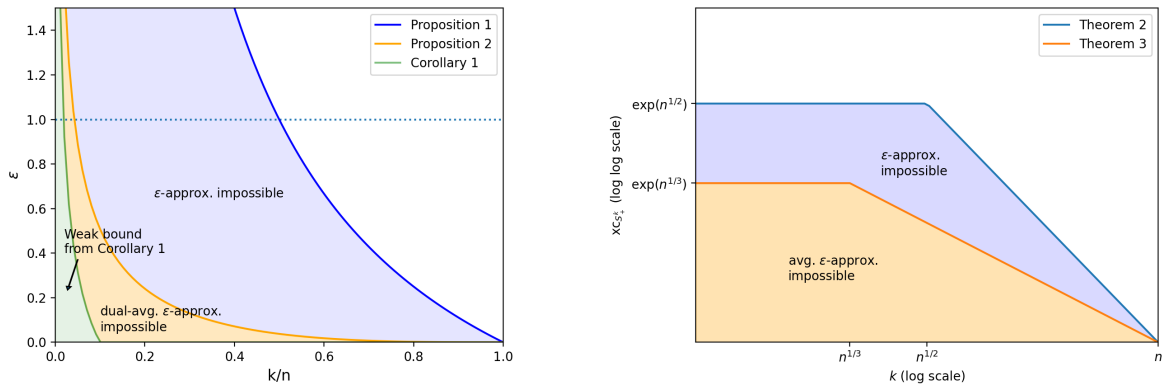
Notion of Approx.	k-PSD Approximations of $\boldsymbol{S}^n_+$	Approximate Extended Formulations of $\boldsymbol{S}^n_+$
ϵ-approx. (Definition 9)	$N \gtrsim \exp\left(n \cdot \max\left\{\frac{1}{1+\epsilon} - \sqrt{\frac{k}{n}},\ 0\right\}^2\right)$ (Theorem 1 & Corollary 1)	$N \geq \exp\left(C_1 \cdot \min\left\{\sqrt{\frac{n}{1+\epsilon}},\ \frac{1}{1+\epsilon}\ \frac{n}{k}\right\}\right)$ (Theorem 2)
avg. ϵ-approx. (Definition 10)	$N \gtrsim \exp\left(n \cdot \max\left\{\frac{1}{4(1+\epsilon)} - \sqrt{\frac{k}{n}},\ 0\right\}^2\right)$ (Theorem 1 & Corollary 2)	$N \geq \exp\left(C_2 \cdot \min\left\{\left(\frac{n}{(1+\epsilon)^2}\right)^{1/3},\ \frac{1}{1+\epsilon}\sqrt{\frac{n}{k}}\right\}\right)$ (Theorem 3)

Approximate Extended Formulations of $\boldsymbol{S}^n_+$. In Section 5, we prove a construction-independent lower bound on N, the number of $k \times k$ PSD constraints needed. Recall that a k-PSD approximation of $\boldsymbol{S}^n_+$ is the intersection of sets in $\boldsymbol{S}^n$, each of which is described with a $k \times k$ PSD constraint. Instead of directly intersecting sets in $\boldsymbol{S}^n$, we may introduce additional variables in pursuit of a more compact description. To be precise, we can lift $\boldsymbol{S}^n$ to a higher-dimensional space by embedding, intersect the lifted space with $k \times k$ PSD constraints, and then project the intersection back to describe a set in $\boldsymbol{S}^n$. The resulting description is called an extended formulation of the set, and the preimage of the projection is called the lifted representation (or PSD lift) of the set. The $\boldsymbol{S}^k_+$-extension complexity of a set S, denoted by $\mathrm{xc}_{\boldsymbol{S}^k_+}(S)$, counts the minimum number of $k \times k$ PSD constraints required to describe S using extended formulation (i.e., with an arbitrary number of additional variables allowed in the description).

In Section 5, we argue that any set that well approximates $B_H\left(\boldsymbol{S}^n_+\right)$ must have $\boldsymbol{S}^k_+$-extension complexity at least superpolynomially large in n for all k much smaller than n. That is, it is impossible to approximate $B_H\left(\boldsymbol{S}^n_+\right)$ using only polynomially many $k \times k$ PSD constraints, for any construction of the approximating set. To be precise, if S is an ϵ-approximation of $B_H\left(\boldsymbol{S}^n_+\right)$, then $\mathrm{xc}_{\boldsymbol{S}^k_+}(S) \geq \exp\left(C \cdot \min\left\{\left(\frac{n}{1+\epsilon}\right)^{1/2},\ \frac{1}{1+\epsilon}\ \frac{n}{k}\right\}\right)$ (Theorem 2); and if S is an average ϵ-approximation of $B_H\left(\boldsymbol{S}^n_+\right)$, then $\mathrm{xc}_{\boldsymbol{S}^k_+}(S) \geq \exp\left(C \cdot \min\left\{\left(\frac{n}{(1+\epsilon)^2}\right)^{1/3},\ \frac{1}{1+\epsilon}\left(\frac{n}{k}\right)^{1/2}\right\}\right)$ (Theorem 3). These results are illustrated in Figure 1b. We remark that these results extend [11, Theorems 1 & 2] beyond the special case $k = 1$.

Nevertheless, we do not know whether our extension complexity lower bounds are tight. It might be possible to achieve stronger extension complexity lower bounds (i.e., move the curves in Figure 1b upward) by means of a more sophisticated analysis. We leave this as an interesting open problem.

Summary of Results. Table 1 summarizes the results in this paper. The lower bounds in the table imply the hardness of approximating $\boldsymbol{S}^n_+$ by using only a small number of $k \times k$ PSD constraints when $k = o(n)$.



(a) Hardness results for sparse k-PSD approximations. Generic (Cor. 1) and tailored bounds (Props. 1 & 2).

(b) Impossibility of approximating $B_H(\boldsymbol{S}^n_+)$ with a polynomial number of $k \times k$ PSD constraints.

Fig. 1: Summary of our results about the hardness of approximating $\boldsymbol{S}^n_+$.

1.2 Discussion and Related Work

Here we make a few comments on our results and some related work.

- Blekherman et al. [7] also investigated the question of how well $\boldsymbol{S}_+^{n,k}$ approximates $\boldsymbol{S}_+^n$. They use the quantity $\overline{\mathrm{dist}}_F(\boldsymbol{S}_+^{n,k}, \boldsymbol{S}_+^n) := \sup_{X\in\boldsymbol{S}_+^{n,k}, \|X\|_F=1} \inf_{Y\in\boldsymbol{S}_+^n} \|X-Y\|_F$ to measure the quality of approximation, and thus, their result has a connection with our result on ϵ-approximation. In this work, we extend the scope of the question in two directions: first, we consider the 'average' distance with the notion of average ϵ-approximation as well as the maximal distance; second, our result (Theorem 1) applies to not only $\boldsymbol{S}_+^{n,k}$, but also $\boldsymbol{S}_+^{n,k}(\mathcal{V})$ with an arbitrary collection of k-dimensional subspaces $\mathcal{V}$.
- Fawzi [11] showed that any polytope that well approximates $B_H\left(\boldsymbol{S}_+^n\right)$ must have LP extension complexity at least exponentially large in n. Our Theorems 2 and 3 generalize their results beyond the special case of $k=1$. Our proof refines and adapts the ideas from [11] to prove a lower bound for arbitrary k. Specifically, we devise a different way of decomposing the component functions of the $\boldsymbol{S}_+^k$-factorization of the slack matrix into their sharp and flat parts, which enable us to apply Fawzi's argument even when $k>1$. In addition, we compare the variance of two representations of the slack matrix instead of their tail probabilities to obtain a nontrivial $\boldsymbol{S}_+^k$-extension complexity lower bound even when $k=\Omega_n(\sqrt{n})$. See the proof of Theorem 2 in Section 5.2 for details.
- Here we compare our Theorem 2 (extension complexity lower bound for an ϵ-approximation of $B_H\left(\boldsymbol{S}_+^n\right)$) with a back-of-the-envelope calculation based on known results about the LP extension complexity of $B_H\left(\boldsymbol{S}_+^n\right)$. Assume that there is a set S such that $\mathrm{xc}_{\boldsymbol{S}_+^k}(S)=N$ and S is an ϵ-approximation of $B_H\left(\boldsymbol{S}_+^n\right)$. On the one hand, each of the N cones of $k\times k$ PSD matrices can be approximated by $\exp(ck)$ facets of linear inequalities, where $c>0$ is an absolute constant; see Aubrun and Szarek [4, Proposition 10]. Thus, the LP extension complexity of S is at most $N\exp(ck)$. On the other hand, the LP extension complexity of S is at least $\exp(c'\sqrt{n})$; see Fawzi [11, Theorem 1]. Therefore, we get $N\geq\exp(c'\sqrt{n}-ck)$. This lower bound becomes trivial when $k=\Omega_n(\sqrt{n})$. In contrast, the lower bound from Theorem 2 remains superpolynomial as long as $k=o_n(n/\log n)$.
- We also remark some earlier works that studied lower bounds on the semidefinite extension complexity of *polytopes* associated with NP-Hard combinatorial problems. Fawzi and Parrilo [13] showed that the $\boldsymbol{S}_+^k$-extension complexity of the correlation polytope, $\mathrm{COR}(n) := \mathrm{conv}\{vv^T : v\in\{0,1\}^n\}$, is exponentially large in n for any fixed constant k. Their proof relies on a combinatorial argument that counts possible sparsity patterns of certain matrices with small PSD rank. Lee, Raghavendra, and Steurer [16] proved a lower bound on the semidefinite extension complexity[1] of $\mathrm{COR}(n)$, based on the notion of low-degree sum-of-squares proof. While these works consider a similar problem to ours, the object of study is different; we are interested in the approximate semidefinite extension complexity of the *spectrahedron* $B_H(\boldsymbol{S}_+^n)$. Specifically, the slack matrix of $\mathrm{COR}(n)$ is not a submatrix of the slack operator of $B_H(\boldsymbol{S}_+^n)$, and neither of the earlier results imply the results in this work.
- Let $D_k=(\boldsymbol{S}_+^{n,k})^*\cap H$. In our analysis, $w_G(D_k)$ turns out to be the expectation of the largest k-sparse eigenvalue of the Gaussian Orthogonal Ensemble (GOE) (divided by $\sqrt{2}$). In this work, we only provide an asymptotic upper bound for $w_G(D_k)$ (Proposition 2), however, it might be possible to prove a lower bound for $w_G(D_k)$ using tools from random matrix theory.
- We do not know whether our lower bounds in Theorems 2 and 3 are tight. We remark that our proof techniques utilize information from the slack matrix only up to degree 2, i.e., up to the second moment. It may be possible to achieve a stronger lower bound by exploiting higher-order moments.
- In this work, we consider the question of approximating $\boldsymbol{S}_+^n\cap H$ and show that at least superpolynomially many $k\times k$ PSD constraints are needed when $k\ll n$. However, if one is allowed to exploit the problem data – C, A_i, b_i – it could be still possible to construct a good approximation F' of the feasible set $F=\{X\in\boldsymbol{S}_+^n : \langle A_i, X\rangle = b_i\}$ with a smaller number of $\boldsymbol{S}_+^k$ so that the optimality gap $\sup_{X\in F'}\langle C,X\rangle - \sup_{X\in F}\langle C,X\rangle$ is small as empirically evidenced in [3].

1.3 Organization and Notation

In Section 2, we review some background materials. In Section 3, we define the notions of approximation that will be used in this paper. In Section 4, we consider the k-PSD approximations of $\boldsymbol{S}_+^n$. Specifically, Section 4.1 discusses a generic lower bound on the number of subspaces required to approximate $\boldsymbol{S}_+^n$,

[1] The smallest integer m such that $\mathrm{COR}(n)$ admits a $\boldsymbol{S}_+^m$-lift; see Section 2.2. This is related to the $\boldsymbol{S}_+^k$-extension complexity, but they are different notions of complexity.

and Section 4.2 provides a more refined analysis tailored to the so-called sparse k-PSD approximation of $\boldsymbol{S}^n_+$. In Section 5, we consider the approximate extended formulations of $\boldsymbol{S}^n_+$. In Section 5.1, we present two main theorems about the hardness of approximating $\boldsymbol{S}^n_+$. Section 5.2 and Section 5.3 are dedicated to their proofs.

Notation. For $x \in \mathbb{R}$, $[x]_+ := \max\{x, 0\}$. For a positive integer n, we let $[n] := \{1, 2, \ldots, n\}$. $\mathbb{R}^n$ denotes the n-dimensional real Euclidean space and $\mathbb{S}^{n-1}$ is the unit sphere in $\mathbb{R}^n$. We also let $\boldsymbol{S}^n$ denote the set of $n \times n$ real symmetric matrices. Given $X \in \boldsymbol{S}^n$ and $I \subset [n]$, let $X_I \in \boldsymbol{S}^{|I|}$ denote the principal submatrix of X with row/column indices in I. For a matrix $X \in \boldsymbol{S}^n$, $\lambda_1(X) \geq \cdots \geq \lambda_n(X)$ are the eigenvalues of X in descending order. A matrix $X \in \boldsymbol{S}^n$ is positive semidefinite, denoted by $X \succeq 0$, if $v^T X v \geq 0$ for all $v \in \mathbb{R}^n$. We let $\boldsymbol{S}^n_+ := \{X \in \boldsymbol{S}^n : X \succeq 0\}$. The letter H is reserved to indicate the subspace of unit trace: $H = \{X \in \boldsymbol{S}^n : \operatorname{Tr} X = 1\}$, and I_n denotes the $n \times n$ identity matrix. For a cone $\mathcal{K} \subseteq \boldsymbol{S}^n$, its base (translated by $-\frac{1}{n} I_n$) is the compact set defined to be $B_H(\mathcal{K}) := \left(\mathcal{K} \cap H\right) - \frac{1}{n} I_n = \{X - \frac{1}{n} I_n \in \boldsymbol{S}^n : X \in \mathcal{K} \cap H\}$, and we define $B^*_H(\mathcal{K}) := B_H(\mathcal{K}^*)$ for notational convenience. Given a set S, we let $\operatorname{cl}(S)$, $\operatorname{conv}(S)$, and $\operatorname{cone}(S)$ denote the closure, the convex hull, and the conical hull of S, respectively. Lastly, we let $N(\mu, \Sigma)$ denote the multivariate Gaussian distribution with mean μ and covariance Σ.

2 Background

In this section, we review some mathematical preliminaries that are used in our proof of the main theorems. Expert readers may want to skip this section and continue reading from Section 3.

2.1 Primer on Convex Geometry

We recall some basic concepts and results in convex analysis. The materials in this section are standard and can be found in classic references; we refer the interested readers to [20] and [5] for more details.

Duality If $S \subseteq \mathbb{R}^d$, the polar of S (in $\mathbb{R}^d$) is the closed convex set

$$S^\circ := \left\{y \in \mathbb{R}^d : \langle x, y\rangle \leq 1 \text{ for all } x \in S\right\}. \tag{2}$$

We observe a few properties involving polars. First of all, if $S \subseteq T$, then $S^\circ \supseteq T^\circ$. Next, it is useful to note that $(S \cup T)^\circ = S^\circ \cap T^\circ$ for any $S, T \subseteq \mathbb{R}^d$, and that $(S \cap T)^\circ = \operatorname{cl}\operatorname{conv}(S^\circ \cup T^\circ)$ if S, T are closed, convex, and contain the origin. Lastly, if S is a closed, convex set that contains the origin, then $(S^\circ)^\circ = S$; this is known as the bipolar theorem in convex analysis (see Lemma 2).

A nonempty closed convex set $\mathcal{K} \subset \mathbb{R}^d$ is called a cone if $\mathcal{K}$ is invariant under positive scaling, i.e., whenever $x \in \mathcal{K}$ and $t \geq 0$, then $tx \in \mathcal{K}$. Given a cone $\mathcal{K}$, its dual cone $\mathcal{K}^*$ (in $\mathbb{R}^d$) is defined via

$$\mathcal{K}^* := \left\{y \in \mathbb{R}^d : \langle x, y\rangle \geq 0 \text{ for all } x \in \mathcal{K}\right\}. \tag{3}$$

Note that $\mathcal{K}^* = -\mathcal{K}^\circ$. Thus, it follows from the properties of polars that (i) $(\mathcal{K}^*)^* = \mathcal{K}$; (ii) if $\mathcal{K}_1 \subseteq \mathcal{K}_2$, then $\mathcal{K}_1^* \supseteq \mathcal{K}_2^*$; and (iii) $\left(\mathcal{K}_1 \cap \mathcal{K}_2\right)^* = \operatorname{cl}\operatorname{cone}(\mathcal{K}_1 \cup \mathcal{K}_2)$ for two cones $\mathcal{K}_1, \mathcal{K}_2$.

The notion of cone duality is closely related to that of set polarity. To clarify the link, we first define a base of a closed convex cone $\mathcal{K}$. Fix a nonzero vector $e \in \mathbb{R}^d$ and the corresponding affine hyperplane

$$H_e := \{x \in \mathbb{R}^d : \langle e, x\rangle = \langle e, e\rangle\}.$$

If $e \in \mathcal{K}^* \setminus \mathcal{K}^\perp$ where $\mathcal{K}^\perp = \{v \in \mathbb{R}^d : \langle v, x\rangle = 0, \forall x \in \mathcal{K}\}$, then we call the set $\mathcal{K}^b_e := \mathcal{K} \cap H_e$ the *base of* $\mathcal{K}$ *with respect to* e. The duality of cones carries over to a duality of bases as follows.

Lemma 1 ([5], Lemma 1.6) *Let $\mathcal{K} \subset \mathbb{R}^d$ be a closed convex cone such that $\mathcal{K} \cap -\mathcal{K} = \{0\}$ and let $e \in \mathcal{K} \cap \mathcal{K}^*$ be a nonzero vector. Then*

$$(\mathcal{K}^*)^b_e = \left\{y \in H_e : \langle -(y - e), x - e\rangle \leq \langle e, e\rangle \text{ for all } x \in \mathcal{K}^b_e\right\}.$$

In other words, if we translate H_e so that e becomes the origin, and consider $\mathcal{K}^b_e$ and $(\mathcal{K}^*)^b_e$ as subsets of that vector space, then $(\mathcal{K}^*)^b_e = -\langle e, e\rangle(\mathcal{K}^b_e)^\circ$.

Remark 1 In this paper, we are concerned with cones $\mathcal{K}$ such that $\boldsymbol{S}^n_+ \subseteq \mathcal{K} \subseteq \boldsymbol{S}^n$ and the unit-trace subspace H. Note that $H = H_e$ with $e = \frac{1}{n} I_n$. We let $B_H(\mathcal{K}) = \mathcal{K}^b_e - \frac{1}{n} I_n$ denote the base of $\mathcal{K}$ with respect to $e = \frac{1}{n} I_n$, translated by $-\frac{1}{n} I_n$ to contain 0. Also, we let $B^*_H(\mathcal{K}) := B_H(\mathcal{K}^*)$ for notational convenience.

Minkowski Functional and support function Let S be a nonempty subset of $\mathbb{R}^d$. The Minkowski functional (or gauge function) of S is defined to be the function $p_S : \mathbb{R}^d \to [0, \infty]$ valued in the extended real numbers such that

$$p_S(x) := \inf\{\lambda \in \mathbb{R} : \lambda > 0 \text{ and } x \in \lambda S\}. \tag{4}$$

We follow the convention that the infimum of the empty set is positive infinity ∞. The support function of S is defined as $h_S : \mathbb{R}^d \to [0, \infty]$ such that

$$h_S(x) := \sup_{z \in S} \langle x, z \rangle. \tag{5}$$

There is a duality between the gauge function and the support function. In words, the gauge function of a convex set is the support function of its polar, and vice versa.

Lemma 2 ([20], Theorem 14.5) *Let S be a closed convex set containing the origin. Then the polar S° is another closed convex set containing the origin, and $(S^\circ)^\circ = S$. Moreover,*

$$p_S(x) = h_{S^\circ}(x) \qquad \text{and} \qquad p_{S^\circ}(x) = h_S(x).$$

Mean Width Given a nonempty, bounded set $S \subset \mathbb{R}^d$, we define the mean width of S as the average of $h_S(u)$ with u distributed uniformly over the unit sphere in the ambient space:

$$w(S) := \int_{\mathbb{S}^{d-1}} h_S(u) d\sigma(u),$$

where $\mathbb{S}^{d-1}$ is the unit sphere in $\mathbb{R}^d$ and σ is the normalized Haar measure on $\mathbb{S}^{d-1}$ (uniform probability measure on $\mathbb{S}^{d-1}$). It is often convenient to consider the Gaussian variant of the mean width because its value does not depend on the ambient dimension.

Definition 2 (Gaussian width) For any nonempty bounded set $S \subset \mathbb{R}^d$, the Gaussian (mean) width of S is defined as

$$w_G(S) := \mathbb{E}_g h_S(g) = \mathbb{E}_g \left[\sup_{x \in S} \langle g,\ x \rangle \right] = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \sup_{x \in S} \langle z,\ x \rangle \exp(-\|z\|^2/2) dz.$$

where g is a standard Gaussian random vector in $\mathbb{R}^d$.

Remark 2 It is easy to verify that $w_G(S) = \kappa_d \cdot w(S)$ where $\kappa_d := \mathbb{E}_g \|g\|_2 = \frac{\sqrt{2}\Gamma((d+1)/2)}{\Gamma(d/2)}$. Note that κ_d depends only on d and is of order $\sqrt{d}$; it is known that $\sqrt{d - 1/2} \le \kappa_d \le \sqrt{d - d/(2d+1)}$.

The Gaussian width has many nice properties. Here we list a few of them that we use in later sections.

1. The Gaussian width does not depend on the ambient dimension.
2. The Gaussian width is invariant under translation and rotation.
3. If $S \subseteq S'$, then $w_G(S) \le w_G(S')$.

Urysohn's Inequality Given a bounded measurable set $S \subset \mathbb{R}^d$, its volume radius is defined as

$$\mathrm{vrad}(S) := \left(\frac{\mathrm{vol}(S)}{\mathrm{vol}(B_2^d)} \right)^{1/d}$$

where B_2^d is the unit d-dimensional Euclidean ball. The volume radius of S is the radius of the Euclidean ball that has the same volume as S.

A set $K \subset \mathbb{R}^d$ is a convex body if it is a convex, compact set with nonempty interior. The following inequality, known as Urysohn's inequality, states that the mean width is minimized for Euclidean balls, among the sets that have the same volume.

Lemma 3 (Urysohn's inequality; [5], Propositions 4.15 & 4.16) *Let $K \subset \mathbb{R}^d$ be a convex body containing the origin in its interior. Then*

$$\frac{1}{w(K^\circ)} \le \mathrm{vrad}(K) \le w(K).$$

2.2 Lifts, Extension Complexity and Slack Operator

Here we briefly review the $\mathcal{K}$-extension complexity of a convex body and its connection to the $\mathcal{K}$-rank of its slack operator. We refer interested readers to [14] and [12] for more details.

Let $\mathcal{K}$ be a closed convex cone. Given a positive integer r, let $\mathcal{K}^r = \mathcal{K} \times \cdots \times \mathcal{K}$ (r times) denote the Cartesian product of r copies of $\mathcal{K}$. We say that a set $S \subset \mathbb{R}^d$ admits a $\mathcal{K}^r$-lift if S can be expressed as

$$S = \pi\big(\mathcal{K}^r \cap L\big)$$

where π is a linear map and L is an affine subspace. The convex set $\mathcal{K}^r \cap L$ is called a $\mathcal{K}^r$-lift of S. The $\mathcal{K}$-extension complexity of S, denoted by $\mathrm{xc}_{\mathcal{K}}(S)$, is defined as the smallest r such that S admits a $\mathcal{K}^r$-lift.

Let P, Q be two convex bodies such that $P \subseteq Q \subseteq \mathbb{R}^d$ and the origin is contained in the interior of P. Let Q° be the polar of Q; see (2). We let $\mathrm{ext}(P)$ denote the set of extreme points of P and define the slack operator $s_{P,Q}$ for (P, Q) as follows.

Definition 3 (Slack operator) For a pair of convex bodies $P \subseteq Q$ with 0 in the interior of P, the map $s_{P,Q} : \mathrm{ext}(P) \times \mathrm{ext}(Q^\circ) \to \mathbb{R}$ such that $s_{P,Q}(x, y) = 1 - \langle x, y\rangle$ is called its associated slack operator. The slack operator $s_{P,Q}$ admits a $\mathcal{K}$-factorization if there exists a pair of maps $A : \mathrm{ext}(P) \to \mathcal{K}$ and $B : \mathrm{ext}(Q^\circ) \to \mathcal{K}^*$ such that $s_{P,Q}(x, y) = \langle A(x),\ B(y)\rangle$ for all $x \in \mathrm{ext}(P)$ and $y \in \mathrm{ext}(Q^\circ)$.

Note that $s_{P,Q}(x, y) \geq 0$ for all $(x, y) \in \mathrm{ext}(P) \times \mathrm{ext}(Q^\circ)$ because $P \subseteq Q$ and therefore $\langle x, y\rangle \leq 1$ for all $(x, y) \in P \times Q^\circ$ by definition of the polar.

The existence of a $\mathcal{K}$-lift of a convex body S is closely related to that of a $\mathcal{K}$-factorization of $s_{P,Q}$ for some convex bodies P, Q such that $P \subseteq S \subseteq Q$. This connection is originally established by Yannakakis [22] for the special case with $\mathcal{K} = \mathbb{R}_+$, motivated by computational considerations about linear programming (LP). This special case of the $\mathbb{R}_+$-extension complexity is widely known as the LP extension complexity (or the extension complexity of polytopes), which counts the minimum number of linear inequalities required to describe S. If $\mathrm{xc}_{\mathbb{R}_+}(S) = N$, then one can optimize a linear function on S by solving a linear program with N inequality constraints. Note that a polytope is generated by a finite number of extreme points, and thus its slack operator is a nonnegative matrix (the so-called *slack matrix*). Yannakakis' theorem states that the LP extension complexity of a polytope is equal to the nonnegative rank of its slack matrix. The Yannakakis' theorem is later generalized in [14].

In this paper, we are interested in the case where $\mathcal{K}$ is a Cartesian product of small PSD cones, $\boldsymbol{S}^k_+$ where $k \geq 1$ is a fixed constant. We state in the next lemma a version of Yannakakis theorem that is generalized to such semidefinite cones (cf. [12, Proposition 3.12]), which immediately follows from [14, Theorem 1 & Corollary 1].

Lemma 4 *Let $\mathcal{K}$ be a semidefinite cone such that $\mathcal{K} = (\boldsymbol{S}^k_+)^r$ for some positive integers k, r. Let P, Q be a pair of convex bodies such that $P \subseteq Q$. Then the slack operator $s_{P,Q}$ has a $\mathcal{K}$-factorization if and only if there is a convex body S such that S admits a $\mathcal{K}$-lift and $P \subseteq S \subseteq Q$.*

We define the $\boldsymbol{S}^k_+$-extension complexity of S, denoted by $\mathrm{xc}_{\boldsymbol{S}^k_+}(S)$, to be the smallest integer r such that S admits a $(\boldsymbol{S}^k_+)^r$-lift. Given a nonnegative operator s, we define $\mathrm{rank}_{\boldsymbol{S}^k_+}(s)$ to be the least r such that s admits a $(\boldsymbol{S}^k_+)^r$-factorization. As a consequence of Lemma 4, we obtain

$$\inf_{S : P \subseteq S \subseteq Q} \mathrm{xc}_{\boldsymbol{S}^k_+}(S) = \mathrm{rank}_{\boldsymbol{S}^k_+}(s_{P,Q}). \tag{6}$$

Note that if $\mathrm{xc}_{\boldsymbol{S}^k_+}(S) = N$, then one can optimize a linear function on S by solving an SDP involving N variables in $\boldsymbol{S}^k_+$.

2.3 Fourier Analysis on the Hypercube and Hypercontractivity

Later in the proof of Theorems 2 and 3, we consider a certain slack operator restricted on the n-dimensional hypercube and use its degree-2 Fourier component to prove extension complexity lower bounds. Specifically, we will need to control the norm of the degree-2 Fourier component. We review the necessary notions here and refer the interested readers to a more comprehensive reference, e.g., [17].

Let $H_n = \{-1, 1\}^n$ denote the vertex set of the n-dimensional hypercube. Every function $f : H_n \to \mathbb{R}$ has a unique Fourier expansion

$$f = f_0 + f_1 + \cdots + f_n \tag{7}$$

where each f_k is a homogeneous multilinear polynomial of degree k. We call f_k the k-th harmonic[2] component of f and let $\mathrm{proj}_k : f \mapsto f_k$ denote the projection onto the degree-k harmonic subspace (the subspace of homogeneous multilinear polynomials of degree k).

The following operator plays an important role in the analysis of Boolean functions.

Definition 4 (Noise operator) Let $\rho \in [0,1]$. For a fixed $x \in H_n$, we write $y \sim N_\rho(x)$ to denote that the random vector $y \in H_n$ is drawn as follows: for each $i \in [n]$, y_i is independently drawn as

$$y_i = \begin{cases} x_i & \text{with probability } \rho, \\ \text{uniformly random in } \{\pm 1\} & \text{with probability } 1-\rho. \end{cases}$$

The noise operator with parameter ρ is the linear operator T_ρ on functions $f : H_n \to \mathbb{R}$ such that

$$T_\rho f(x) = \mathbb{E}_{y \sim N_\rho(x)}[f(y)].$$

It is known that the noise operator T_ρ smooths $f : H_n \to \mathbb{R}$, by attenuating its high-frequency modes, e.g., [17, Proposition 2.47]. To be precise, T_ρ acts on f multiplying the k-th Fourier coefficient of f by ρ^k, i.e.,

$$T_\rho f = \sum_{k=0}^{n} \rho^k f_k.$$

For $\rho < 1$, $T_\rho f$ is 'smoother' than f as the high-frequency terms of f are diminished. In one extreme, $T_\rho f$ is constant equal to $\mathbb{E} f$ when $\rho = 0$; in the other extreme where $\rho = 1$, there is no smoothing effect and $T_\rho f = f$.

Next, we recall that the p-norm ($p \geq 1$) of $f : H_n \to \mathbb{R}$ is defined as

$$\|f\|_p = \left(\mathbb{E}_{x \sim \mu(H_n)}\left[|f(x)|^p\right]\right)^{\frac{1}{p}}. \tag{8}$$

where $\mu(H_n)$ denotes the uniform probability measure over H_n. Note that $\|f\|_p \leq \|f\|_q$ for $p \leq q$. When $p < q$, there is no general way to control $\|f\|_q$ with $\|f\|_p$, and the ratio $\|f\|_q / \|f\|_p$ can be arbitrarily large; the ratio becomes larger as f fluctuates more wildly.

The hypercontractive inequality for T_ρ due to Bonami and Beckner [9, 6] provides an upper bound on $\|T_\rho f\|_q$ in terms of $\|f\|_p$ with $p < q$, thereby giving an estimate for how much smoother $T_\rho f$ is, when compared to f. It can be stated as follows.

Lemma 5 (Hypercontractivity) *Given $f : H_n \to \mathbb{R}$, for any $0 < \rho \leq 1$ and $p \geq 1$, we have $\|T_\rho f\|_q \leq \|f\|_p$ where $q = 1 + \frac{1}{\rho^2}(p-1)$.*

We use Lemma 5 to control the norm of the degree-2 harmonic component of a bounded nonnegative function as stated below in Lemma 6, following [19, Lemma 2.3] and [11, Lemma 3]. Its proof is included in Appendix A for completeness.

Lemma 6 *Let $f : H_n \to \mathbb{R}$ satisfy (i) $0 \leq f(x) \leq \Lambda$ for all $x \in H_n$; and (ii) $\mathbb{E}_{x \sim \mu(H_n)}[f(x)] \leq 1$. Then*

$$\|\mathrm{proj}_2 f\|_2 \leq \begin{cases} \Lambda & \text{if } \Lambda < e, \\ e \log(\Lambda) & \text{if } \Lambda \geq e. \end{cases}$$

2.4 Some Useful Facts about (Sub-)Gaussians

Here we collect a few facts about Gaussians that are useful to control the fluctuation of Gaussian processes. These are standard results and more details can be found in references such as [10, 21, 5].

[2] Note that f_k is square-free (because f_k is multilinear), and thus, $\nabla^2 f_k = 0$.

2.4.1 Gaussian Random Matrices and Sub-gaussian Random Variables

Standard Gaussian Distribution in $\boldsymbol{S}^n$ Recall that the space $\boldsymbol{S}^n$ of real symmetric $n \times n$ matrices can be viewed as real Euclidean space of dimension $\binom{n+1}{2}$ equipped with the trace inner product $\langle A, B\rangle = \mathrm{Tr}(AB)$. We define the standard Gaussian distribution in $\boldsymbol{S}^n$ via the natural isomorphism between $\boldsymbol{S}^n$ and $\mathbb{R}^{\binom{n+1}{2}}$.

Definition 5 A random matrix $A \in \boldsymbol{S}^n$ has the standard Gaussian distribution if the random variables $(a_{ij})_{1\le i\le j\le n}$ are independent, with $a_{ii} \sim N(0,1)$ and $a_{ij} \sim N(0,1/2)$ for $i < j$.

Note that A is a standard Gaussian vector in the space $\boldsymbol{S}^n$ if and only if $\sqrt{2}A$ is a GOE(n) (Gaussian Orthogonal Ensemble) matrix, cf. [5, Section 6.2]. The GOE has the property of orthogonal invariance, i.e., if $A \in \boldsymbol{S}^n$ is a GOE(n) matrix, then for any fixed orthogonal matrix $U \in O(n)$, the random matrix UAU^T is also a GOE(n) matrix.

Sub-Gaussian and Sub-exponential Random Variables Many interesting properties of Gaussian random variables are due to the fast decaying tail probabilities. Such properties are shared by some of non-Gaussian random variables, so called the class of sub-Gaussian random variables. This notion can be formalized based on the moment-generating function $\mathbb{E}[e^{\lambda X}]$:

Definition 6 A random variable X is sub-Gaussian with parameter $v > 0$ if $\mathbb{E}[X] = 0$ and

$$\mathbb{E}[e^{\lambda X}] \le \exp\left(\frac{\lambda^2}{2}v\right), \quad \forall \lambda \in \mathbb{R}.$$

Definition 7 A random variable X is sub-exponential with parameters $v, c > 0$ if $\mathbb{E}[X] = 0$ and

$$\mathbb{E}[e^{\lambda X}] \le \exp\left(\frac{\lambda^2}{2}v\right), \quad \forall \lambda \text{ such that } |\lambda| \le \frac{1}{c}.$$

For example, exponential and chi-squared random variables (with centering) are sub-exponential. Informally, a sub-gaussian random variable can be viewed as a sub-exponential random variable with c tending to 0.

A sub-exponential random variable exhibits sub-Gaussian tail behavior around its center and have exponentially decaying tail probabilities far away from 0. More precisely, the following tail probability bounds can be obtained by the Cramér-Chernoff method (see e.g. [10, Section 2.2]): if X is a sub-exponential random variable with parameters (v, c), then for every $t > 0$,

$$\max\left\{\Pr[X > t],\ \Pr[X < -t]\right\} \le \begin{cases} e^{-t^2/2v} & \text{if } 0 \le t \le \frac{v}{c}, \\ e^{-t/2c} & \text{if } t > \frac{v}{c}. \end{cases}$$

2.4.2 Useful Inequalities

Gaussian Comparison Inequality The following fundamental inequality, known as Sudakov-Fernique inequality, expresses that a Gaussian process can get farther (i.e., has a larger supremum) when it has weaker correlations. We refer the interested readers to [21, Theorem 7.2.11] for more details.

Definition 8 A random process $(X_t)_{t\in T}$ is a Gaussian process if the random vector $(X_t)_{t\in T_0}$ has normal distribution for all finite subsets $T_0 \subset T$.

Lemma 7 (Sudakov-Fernique inequality) *Let $(X_t)_{t\in T}$ and $(Y_t)_{t\in T}$ be Gaussian processes. Suppose that for all $t, s \in T$, the following two conditions hold: (i) $\mathbb{E}X_t = \mathbb{E}Y_t = 0$; and (ii) $\mathbb{E}(X_t - X_s)^2 \le \mathbb{E}(Y_t - Y_s)^2$. Then*

$$\mathbb{E}\sup_{t\in T} X_t \le \mathbb{E}\sup_{t\in T} Y_t.$$

There is a well known upper bound for the expectation of the largest eigenvalue of a standard Gaussian random matrix in $\boldsymbol{S}^n$. Its proof is based on Sudakov-Fernique inequality (Lemma 7) and standard; see Appendix A for the proof.

Lemma 8 *If a random matrix $G \in \boldsymbol{S}^n$ has the standard Gaussian distribution, then*

$$\mathbb{E}_G\big[\lambda_1(G)\big] = \mathbb{E}_G\Big[\sup_{v\in\mathbb{S}^{n-1}} \langle v, Gv\rangle\Big] \le \sqrt{2n}.$$

Remark 3 It is known that $\lim_{n\to\infty} \mathbb{E}_G\big[\lambda_1(G)\big]/\sqrt{2n} = 1$. Indeed, not only its expected value, but also its limiting distribution is known in the literature. The quantity $\lambda_1(G) - \sqrt{2n}$ is of order $n^{-1/6}$ and its distribution converges to the Tracy-Widom distribution after normalization.

Gaussian Concentration A smooth function of independent Gaussian random variables is sub-Gaussian. The following result is widely known as the Gaussian concentration inequality; see [10, Theorem 5.5] for example. Note that the sub-Gaussian parameter L^2 depends only on the smoothness of the function, and not on the number of Gaussian random variables.

Lemma 9 (Gaussian concentration) *Let $X = (X_1, \dots, X_n)$ be a vector of n independent standard Gaussian random variables. If $f : \mathbb{R}^d \to \mathbb{R}$ is L-Lipschitz (with respect to the ℓ_2 norm), then $f(X) - \mathbb{E}f(X)$ is sub-Gaussian with sub-Gaussian parameter L^2.*

The following lemma states that the support function of a convex set concentrates around its mean. It can be proved applying Lemma 9 to the support function, which is Lipschitz with the Lipschitz constant being the diameter of the set. We provide a proof in Appendix A for readers' convenience.

Lemma 10 *Let $K \subset \mathbb{R}^d$ be a convex set containing 0. Let $w_G(K)$ denote the Gaussian width of K. Then for any $\alpha \geq 0$,*

$$\max\left\{ \Pr_{g\sim N(0,I_d)}\Big[\max_{x\in K}\langle g, x\rangle < (1-\alpha)w_G(K)\Big],\ \Pr_{g\sim N(0,I_d)}\Big[\max_{x\in K}\langle g, x\rangle > (1+\alpha)w_G(K)\Big]\right\} \leq \exp\left(-\frac{\alpha^2}{4\pi}\right).$$

MGF of Sub-Gaussian Chaos of Order 2 We review the concentration of quadratic forms of the type

$$\sum_{i,j=1}^{n} a_{ij}X_iX_j = X^TAX$$

where $A = (a_{ij})$ is an $n \times n$ matrix of coefficients, and $X = (X_1, \dots, X_n)$ is a random vector with independent coordinates. Such a quadratic form is known as a chaos (of order 2) in probability theory.

When X_i's are sub-Gaussian random variables (e.g., Gaussian or Rademacher), the quadratic form X^TAX is sub-exponential. The following upper bound is well known, and can be used to derive a Bernstein-type exponential concentration results (e.g., Hanson-Wright inequality) for X^TAX. Its proof is based on standard techniques such as decoupling and comparison to Gaussian chaos. We sketch the proof of Lemma 11 in Appendix A, omitting the details of the proof. We refer the interested readers to [21, Sections 6.1 & 6.2] for more details.

Lemma 11 (MGF of sub-Gaussian chaos of order 2) *Let $X = (X_1, \dots, X_n) \in \mathbb{R}^n$ be a random vector with independent sub-Gaussian coordinates with sub-Gaussian parameter v, and let A be an $n \times n$ matrix with zero diagonal. Then X^TAX is sub-exponential with parameters $(32v^2\|A\|_F^2,\ 4\sqrt{2}v\|A\|_{op})$, i.e.,*

$$\mathbb{E}\exp\left(\lambda X^TAX\right) \leq \exp\left(16v^2\|A\|_F^2\lambda^2\right), \qquad \text{for all } \lambda \text{ s.t. } |\lambda| \leq \frac{1}{4\sqrt{2}v\|A\|_{op}}.$$

Observe that for any function $f : H_n \to \mathbb{R}$, its degree-2 projection, $\mathrm{proj}_2(f)$, is a multilinear quadratic form on H_n. That is, there exists some matrix A with zero diagonal[3] such that $\mathrm{proj}_2(f)(x) = x^TAx$ for all $x \in H_n$. Therefore, the random variable $\mathrm{proj}_2(f)(X)$ derived from the uniform random vector $X \sim \mu(H_n)$ is sub-exponential by Lemma 11. We formally state this observation in the following lemma to use later in the proof of Theorem 2; see Appendix A for its proof.

Lemma 12 *Let X be a random vector uniformly distributed over H_n. For any function $f : H_n \to \mathbb{R}$, the derived random variable $\mathrm{proj}_2(f)(X)$ is sub-exponential with parameters $(16M_f^2,\ 4M_f)$ where $M_f = \|\mathrm{proj}_2 f\|_2$, i.e.,*

$$\mathbb{E}_{X\sim\mu(H_n)}\exp\left(\lambda\ \mathrm{proj}_2(f)(X)\right) \leq \exp\left(\frac{\lambda^2}{2}\cdot 16M_f^2\right), \qquad \text{for all } \lambda \text{ s.t. } |\lambda| \leq \frac{1}{4M_f}.$$

Maximal Inequalities The following simple maximal inequality is well known, and it is asymptotically sharp if the random variables are i.i.d. Gaussian. Its proof can be found in Appendix A.

Lemma 13 *Let $X_1, \dots, X_N$ be sub-exponential random variables with parameters (v, c). Then*

$$\mathbb{E}\Big[\max_{i\in[N]} X_i\Big] \leq \max\left\{\sqrt{2v\log N},\ 2c\log N\right\}.$$

[3] More precisely, $A_{ij} = \frac{1}{2}\mathbb{E}_{X\sim\mu(H_n)}[X_iX_jf(X)]$ for $i, j \in [n]$ such that $i \neq j$.

3 Three Notions of Approximation

Recall that we want to approximate the positive semidefinite cone $\boldsymbol{S}_+^n$ with a convex cone $\mathcal{K} \supseteq \boldsymbol{S}_+^n$ so that the feasible set $B_H(\mathcal{K}) = (\mathcal{K} \cap H) - \frac{1}{n} I_n$ (cf. Remark 1) well approximates $B_H(\boldsymbol{S}_+^n)$. In Section 3.1, we introduce three notions of approximation for sets. In Section 3.2, we extend these notions to cones to assess the quality of $\mathcal{K}$ as an approximation of $\boldsymbol{S}_+^n$.

Specifically, we first define a natural notion of ϵ-approximation for sets that contain the origin (Definition 9). Then, we additionally describe two auxiliary notions of approximation for the convenience of our analysis, namely, the average ϵ-approximation (Definition 10) and the dual-average ϵ-approximation (Definition 11). These two auxiliary notions can be obtained by relaxing a quantifier in the definition of ϵ-approximation. These relaxed notions are closely related, but incomparable to each other. They will be respectively used in Section 4 and Section 5 to prove the hardness of approximating $\boldsymbol{S}_+^n$ with a small number of $k \times k$ PSD constraints.

3.1 Notions of Approximation for Sets

To begin with, we define the notion of ϵ-approximation for sets containing the origin.

Definition 9 (ϵ-approximation) Let P be a set containing 0. For $\epsilon > 0$, a set S is an *ϵ-approximation* of P if $P \subseteq S \subseteq (1+\epsilon)P$. Given two sets P, S that contain 0, we let

$$\epsilon^*(P, S) := \inf\{\epsilon > 0 : S \text{ is an } \epsilon\text{-approximation of } P\}.$$

This is a natural notion to quantify how tightly a set P containing 0 can be approximated by another set $S \supseteq P$. Recall the definition of the support function $h_S(x) := \sup_{z \in S} \langle x, z \rangle$, cf. (5). We observe that if S is an ϵ-approximation of P, then

$$h_P(x) \le h_S(x) \le (1+\epsilon) h_P(x) \quad \text{for all } x. \tag{9}$$

That is, if S is an ϵ-approximation of P, then for every direction in the ambient space, the distance from the supporting hyperplane of S to the origin is at most $(1+\epsilon)$ times the distance from the supporting hyperplane of P to the origin. Moreover, when P and S are convex, the converse is also true.

Next, we define a more lenient notion of approximation by relaxing the quantifier 'for all x' in (9) by taking average over random direction x. To this end, recall the notion of Gaussian width from Definition 2 that $w_G(S) = \mathbb{E}_G[h_S(G)]$ for any nonempty bounded set $S \subset \mathbb{R}^d$, where G is a standard Gaussian random vector in $\mathbb{R}^d$.

Definition 10 (average ϵ-approximation) Let P be a set containing 0. For $\epsilon > 0$, a set S is an *average ϵ-approximation* of P, or *ϵ-approximation of P in the average sense*, if $P \subseteq S$ and $w_G(S) \le (1+\epsilon) w_G(P)$. Given two sets P, S that contain 0, we let

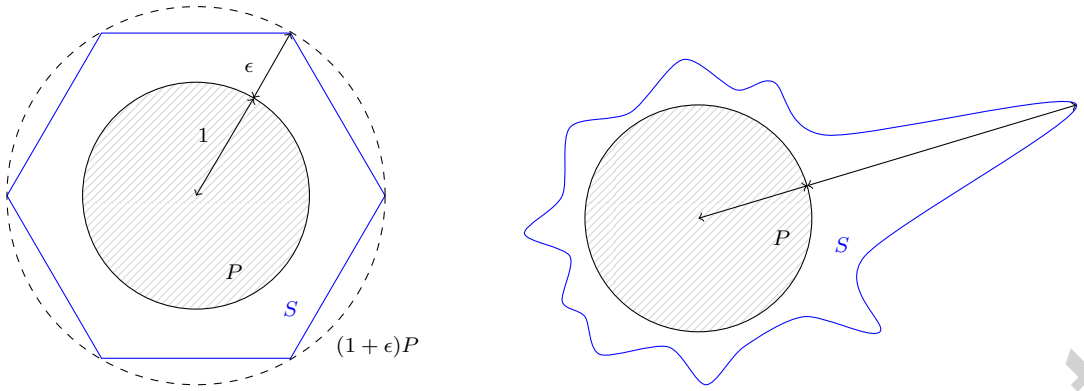
$$\epsilon^*_{\text{avg}}(P, S) := \inf\{\epsilon > 0 : S \text{ is an average } \epsilon\text{-approximation of } P\}.$$

By definition, S is an average ϵ-approximation of P if and only if $\mathbb{E}_G[h_S(G) - h_P(G)] \le \epsilon \cdot \mathbb{E}_G[h_P(G)]$ where G is a standard Gaussian random vector.

Note that average ϵ-approximation is a weaker notion than ϵ-approximation because $\epsilon^*(P, S) \ge \epsilon^*_{\text{avg}}(P, S)$. That is, for a fixed $\epsilon > 0$, if S is an ϵ-approximation of P, then S is also an average ϵ-approximation of P. As a matter of fact, average ϵ-approximation is a strictly weaker notion because there exists a pair of sets (P, S) such that $\epsilon^*(P, S) > \epsilon^*_{\text{avg}}(P, S)$, i.e., there exists some $\epsilon > 0$ for which S is not an ϵ-approximation of P whereas S is an average ϵ-approximation of P. We illustrate this point with the following two examples.

Example 1 Let $P = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \le 1\}$ and $S = \{(x, y) \in \mathbb{R}^2 : x^2/4 + y^2 \le 1\}$. Then $\epsilon^*(P, S) = 1$. On the other hand, $\epsilon^*_{\text{avg}}(P, S) = \frac{4}{\pi} E(3/4) - 1 \approx 0.54196$ where $E(m)$ is the complete elliptic integral of the second kind with parameter $m = k^2$. The value of $\epsilon^*_{\text{avg}}(P, S)$ can be computed by observing that $w_G(P) = \mathbb{E}_{G \in N(0, I_2)} \|g\|_2$ and $w_G(S) = \mathbb{E}_{g \in N(0, I_2)} \|g\|_2 \cdot \frac{1}{2\pi} \int_0^{2\pi} \sqrt{4\cos^2\theta + \sin^2\theta}\, d\theta = \frac{4}{\pi} E(3/4) \mathbb{E}_{g \in N(0, I_2)} \|g\|_2$.

Example 2 Let $P = \{x \in \mathbb{R}^n : \|x\|_2 \le 1\}$ and $S = \{x \in \mathbb{R}^n : \|x\|_1 \le \sqrt{n}\}$ where $\|\cdot\|_p$ denotes the ℓ_p-norm. Then $\epsilon^*(P, S) = \sqrt{n} - 1$. On the other hand, $\epsilon^*_{\text{avg}}(P, S) = f(n)$ where $f(n)$ is a function of n such that $f(n) \approx \sqrt{2 \log n}$ for sufficiently large n. This is because $w_G(P) = \mathbb{E}_{g \sim N(0, I_n)} \|g\|_2 \approx \sqrt{n}$ and $w_G(S) = \sqrt{n} \cdot \mathbb{E}_{g \sim N(0, I_n)} \max_{i \in [n]} |g_i| \approx \sqrt{2n \log n}$. See [21, Section 7.5.3] for more details.



(a) S is an ϵ-approximation of P (b) S 'poorly' approximates P only in some directions

Fig. 2: Cartoons illustrating the difference between ϵ-approximation and average ϵ-approximation.

In the two examples above, we observed there exists $\epsilon > 0$ such that S is an ϵ-approximation of P in the average sense, while it is not an ϵ-approximation. This happens because $h_S(G) - h_P(G)$ is small on average, but the difference can be potentially large for some G. In other words, S approximates P well on average, but poorly for certain 'bad' directions in the ambient space, as illustrated in Figure 2. Nevertheless, the set of 'bad' directions might have only a small measure as in Example 2, and the notion of ϵ-approximation as in Definition 9 can be overly conservative. That is why we additionally consider the notion of average ϵ-approximation, which is more lenient with the shape of the approximating set S.

One drawback of evaluating the quality of approximation with the notion of average ϵ-approximation is that it only measures the difference averaged over an ensemble of random objectives. Thus, we cannot control the gap $h_S(x) - h_P(x)$ for any specific x, however, we can still establish a probabilistic upper bound on $h_S(G) - h_P(G)$ when G is randomly drawn from the standard Gaussian distribution.

Lemma 14 *Let S be an average ϵ-approximation of P for some $\epsilon > 0$. Then for all $\tau > 0$,*

$$\Pr_{G \sim \text{std Gaussian}} \left[h_S(G) - h_P(G) > \tau \right] \leq \epsilon \frac{w_G(P)}{\tau}.$$

Lemma 14 operationally means that if S is an average ϵ-approximation of P for small ϵ, then $h_S(x) - h_P(x)$ can be large only for x in a set that has small measure. In particular, the probability upper bound converges to 0 as $\epsilon \to 0$. That is, $h_S(x) - h_P(x)$ converges to 0 for all x (but those in a set of measure-zero) as $\epsilon \to 0$.

Proof (Proof of Lemma 14) Note that $h_S(G) - h_P(G) \geq 0$ for all g because $P \subseteq S$. The conclusion follows from Markov's inequality and the observation that $w_G(S) - w_G(P) \leq \epsilon \cdot w_G(P)$. □

Lastly, we revisit Definition 9 to introduce an alternative relaxation of ϵ-approximation, namely, the 'dual' version of average ϵ-approximation. Recall from (4) that the gauge function of S is defined as $p_S(x) := \inf\{\lambda \in \mathbb{R} : \lambda > 0 \text{ and } x \in \lambda S\}$. Observe that $P \subseteq S \subseteq (1+\epsilon)P$ if and only if $\frac{1}{1+\epsilon} p_P(x) \leq p_S(x) \leq p_P(x)$ for all x. When P and S are closed convex sets, $p_P(x) = h_{P^\circ}(x)$ and $p_S(x) = h_{S^\circ}(x)$ by Lemma 2. Therefore, S is an ϵ-approximation of P if and only if $\frac{1}{1+\epsilon} h_{P^\circ}(x) \leq p_{S^\circ}(x) \leq p_{P^\circ}(x)$ for all x. As before, we ease the condition "$\frac{1}{1+\epsilon} h_{P^\circ}(x) \leq p_{S^\circ}(x)$ for all x" by averaging over x to reach at the following definition.

Definition 11 (dual-average ϵ-approximation) Let P be a set containing 0. For $\epsilon > 0$, a set S is a *dual-average ϵ-approximation* of P, or *ϵ-approximation of P in the dual-average sense*, if $P \subseteq S$ and $w_G(S^\circ) \geq \frac{1}{1+\epsilon} w_G(P^\circ)$. Given two sets P, S that contain 0, we define

$$\epsilon^*_{\text{dual-avg}}(P, S) := \inf\{\epsilon > 0 : S \text{ is a dual-average } \epsilon\text{-approximation of } P\}.$$

Note that dual-average ϵ-approximation is also a weaker notion than ϵ-approximation. That is, for a fixed $\epsilon > 0$, if S is an ϵ-approximation of P, then S is also a dual-average ϵ-approximation of P. In Section 4, we use the notion of dual-average ϵ-approximation as a technical tool to prove the hardness of k-PSD approximations of $\boldsymbol{S}^n_+$.

The notion of dual-average ϵ-approximation is closely related to the notion of average ϵ-approximation; they are dual to each other. However, they are not equivalent notions of approximation, i.e., there exist convex sets P, S such that S is a good approximation of P in the average sense, but not in the dual average sense. The opposite is also possible. See Remark 4 and Example 3.

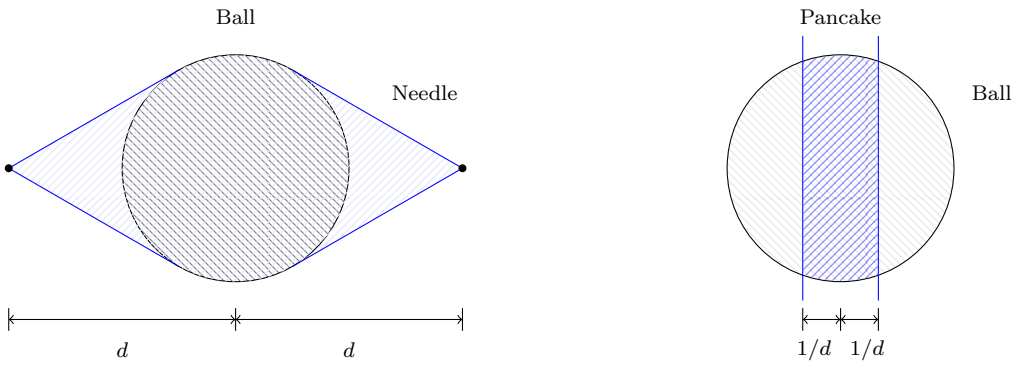


Fig. 3: The sets described in Example 3.

Remark 4 For $\epsilon > 0$, S is an average ϵ-approximation of P if and only if P° is a dual-average ϵ-approximation of S°. In other words, $\epsilon^*_{\text{avg}}(P, S) = \epsilon^*_{\text{dual-avg}}(S^\circ, P^\circ)$. In this sense, the notion of dual-average ϵ-approximation is the dual of the notion of average ϵ-approximation.

Example 3 (Ball, Needle, and Pancake) Consider a d-dimensional unit ℓ_2-ball and a 'needle' obtained by taking the convex hull of the union of the ball and two points that are located on the opposite side of the origin at distance d. The polar of this 'needle' is the 'pancake' obtained by intersecting the unit ball with a slab of thickness $2/d$ along its equator. These three sets are illustrated in Figure 3. We observe that the Gaussian width of the ball, the needle, and the pancake are approximately $\sqrt{d-1/2}, d\sqrt{2/\pi}$, and $\sqrt{d-3/2}$, respectively. Thus, the ball is a good approximation of the pancake in the average sense, but not in the dual-average sense. Likewise, the needle is a good approximation of the ball in the dual-average sense, but not in the average sense.

3.2 Notions of Approximation for Cones

Recall that our primary motivation for introducing the notions of approximation is to quantify the optimality gap that arises from a conic programming relaxation of the problem in (1). Suppose that we are to relax the problem (1) by replacing the PSD cone $\boldsymbol{S}^n_+$ with a larger cone $\mathcal{K} \supseteq \boldsymbol{S}^n_+$. Letting $P = \{X \in \boldsymbol{S}^n_+ : \ \langle A_i, X\rangle = b_i, \ i = 1, \ldots, m\}$ and $S = \{X \in \mathcal{K} : \ \langle A_i, X\rangle = b_i, \ i = 1, \ldots, m\}$ denote the feasible sets of the original and the relaxed problems, we can see that $S \supseteq P$ and there arises an increase in the optimal value, $\Gamma_{P,S}(C) := h_S(C) - h_P(C)$, as a result of the relaxation.

We extend the notions of approximation for sets, defined in Section 3.1, to the notions for cones by fixing a certain affine constraint. Recall that for a cone $\mathcal{K} \subseteq \boldsymbol{S}^n$, we let $B_H(\mathcal{K}) := \left(\mathcal{K} \cap H\right) - \frac{1}{n} I_n = \left\{X - \frac{1}{n} I_n \in \boldsymbol{S}^n : X \in \mathcal{K} \cap H\right\}$ where $H = \{X \in \boldsymbol{S}^n : \operatorname{Tr} X = 1\}$ and I_n denotes the $n \times n$ identity matrix. Note that $B_H(\mathcal{K})$ is the feasible set of the problem (1), translated by $-\frac{1}{n} I_n$, when the affine constraint in (1) is the unit trace constraint. We define the notions of approximation for cones as follows.

Definition 12 (ϵ-approximation for cones in $\boldsymbol{S}^n$) A cone $\mathcal{K} \subseteq \boldsymbol{S}^n$ is an ϵ-approximation (average ϵ-approximation / dual-average ϵ-approximation, resp.) of $\boldsymbol{S}^n_+$ if $B_H(\mathcal{K})$ is an ϵ-approximation (average ϵ-approximation / dual-average ϵ-approximation, resp.) of $B_H\left(\boldsymbol{S}^n_+\right)$. Also, we let

$$\epsilon^*(\boldsymbol{S}^n_+, \mathcal{K}) := \epsilon^*\left(B_H\left(\boldsymbol{S}^n_+\right), B_H(\mathcal{K})\right)$$

and define $\epsilon^*_{\text{avg}}(\boldsymbol{S}^n_+, \mathcal{K})$ and $\epsilon^*_{\text{dual-avg}}(\boldsymbol{S}^n_+, \mathcal{K})$ in a similar manner.

Remark 5 For later use, we remark here that $w_G\left(B_H\left(\boldsymbol{S}^n_+\right)\right) \leq \sqrt{2n}$ and that $\lim_{n\to\infty} \frac{w_G\left(B_H\left(\boldsymbol{S}^n_+\right)\right)}{\sqrt{2n}} = 1$ because $\boldsymbol{S}^n_+ \cap H = \operatorname{conv}\{vv^T : v \in \mathbb{S}^{n-1}\}$, cf. Lemma 8 and Remark 3.

4 k-PSD Approximations of $\boldsymbol{S}_+^n$

One option to relax the PSD constraint $X \in \boldsymbol{S}_+^n$ in (1) is to enforce the PSD constraints only on the smaller $k \times k$ principal submatrices of X, which leads to the following relaxation:

$$\begin{aligned} \text{maximize} \quad & \langle C, X \rangle \\ \text{subject to} \quad & \langle A_i, X \rangle = b_i, \quad i = 1, \dots, m, \\ & k \times k \text{ principal submatrices of } X \in \boldsymbol{S}_+^k. \end{aligned} \tag{10}$$

Note that the PSD cone $\boldsymbol{S}_+^n$ in (1) is replaced with a relaxed cone that is defined using $(k \times k)$-sized PSD constraints, and (10) can be solved more efficiently when $k \ll n$. For example, $k = 1$ yields a linear programming (LP) approximation and $k = 2$ produces a second-order cone programming (SOCP) approximation of the original SDP [3].

In this section, we consider a scheme to approximate $\boldsymbol{S}_+^n$ by enforcing $k \times k$ PSD constraints on particular subspaces. To be precise, we choose a fixed set of k-dimensional subspaces in $\mathbb{R}^n$ and define a cone of $n \times n$ symmetric matrices that are PSD when restricted to these subspaces. The cone associated with (10) is an example of this construction that is obtained by imposing PSD constraints on the $\binom{n}{k}$ subspaces of k-sparse vectors in $\mathbb{R}^n$, and will be referred to as the sparse k-PSD approximation of $\boldsymbol{S}_+^n$.

In Section 4.1, we formalize the definition of the k-PSD approximation and prove a lower bound on the number of $k \times k$ PSD constraints required. We show that when k is much smaller than n, it is necessary to impose PSD constraints on at least exponentially many subspaces to produce a cone that approximates $\boldsymbol{S}_+^n$ well. In Section 4.2, we discuss the sparse k-PSD approximation in more detail.

4.1 Lower Bound for k-PSD Approximations of $\boldsymbol{S}_+^n$

We recall the definition of the k-PSD approximation of $\boldsymbol{S}_+^n$ from Definition 1.

Definition 13 (k-PSD approximation of $\boldsymbol{S}_+^n$; restatement of Definition 1) Let $\mathcal{V} = \{V_1, \dots, V_N\}$ be a set of k-dimensional subspaces of $\mathbb{R}^n$. The k-PSD approximation of $\boldsymbol{S}_+^n$ induced by $\mathcal{V}$ is the convex cone

$$\boldsymbol{S}_+^{n,k}(\mathcal{V}) := \{X \in \boldsymbol{S}^n : v^T X v \geq 0, \ \forall v \in V_i, \ \forall i \in [N]\}.$$

Note that $\boldsymbol{S}_+^{n,k}(\mathcal{V}) \supseteq \boldsymbol{S}_+^n$ is the set of $n \times n$ symmetric matrices whose associated quadratic forms are positive semidefinite when restricted to $\bigcup_{i=1}^N V_i$. Thus, if $U_i \in \mathbb{R}^{n \times k}$ is a matrix whose columns form a basis of $V_i \in \mathcal{V}$, then $\boldsymbol{S}_+^{n,k}(\mathcal{V}) = \{X \in \boldsymbol{S}^n : U_i^T X U_i \in \boldsymbol{S}_+^k, \ \forall i \in [N]\}$.

Our first theorem presents upper bounds on the Gaussian width of the base of the dual cone of $\boldsymbol{S}_+^{n,k}(\mathcal{V})$ as a function of k and $N = |\mathcal{V}|$. Recall that $B_H^*(\mathcal{K}) := (\mathcal{K}^* \cap H) - \frac{1}{n} I_n$ for a cone $\mathcal{K} \subseteq \boldsymbol{S}^n$ (Remark 1).

Theorem 1 *Let n, k be positive integers such that $1 \leq k \leq n$ and $\mathcal{V} = \{V_1, \dots, V_N\}$ be a set of k-dimensional subspaces of $\mathbb{R}^n$. Let $S_{\mathcal{V}} = \{uu^T : u \in T_{\mathcal{V}}\} \subset \boldsymbol{S}^n$ where $T_{\mathcal{V}} := \left(\bigcup_{i=1}^N V_i\right) \cap \mathbb{S}^{n-1} \subset \mathbb{R}^n$. Then*

$$w_G\Big(B_H^*\Big(\boldsymbol{S}_+^{n,k}(\mathcal{V})\Big)\Big) = w_G(S_{\mathcal{V}}).$$

In addition, $w_G(S_{\mathcal{V}})$ satisfies

1. *$w_G(S_{\mathcal{V}}) \leq \sqrt{2} \cdot w_G(T_{\mathcal{V}})$, and*
2. *$w_G(S_{\mathcal{V}}) \leq \sqrt{2k} + \sqrt{2 \log N}$ for any configurations of subspaces $V_1, \dots, V_N$ in $\mathcal{V}$.*

Recall that $w_G\big(B_H^*\left(\boldsymbol{S}_+^n\right)\big) = w_G\big(B_H\left(\boldsymbol{S}_+^n\right)\big) \approx \sqrt{2n}$ in the sense that $\lim_{n\to\infty} \frac{w_G(B_H(\boldsymbol{S}_+^n))}{\sqrt{2n}} = 1$, cf. Remark 5. Comparing the upper bounds in Theorem 1 against $\sqrt{2n}$, we can contrast the size of $B_H^*\left(\boldsymbol{S}_+^{n,k}(\mathcal{V})\right)$ relative to $B_H^*\left(\boldsymbol{S}_+^n\right)$. For example, when k and N are small, $\sqrt{2k} + \sqrt{2 \log N} \ll \sqrt{2n}$, and we can intuitively see that the dual of the cone $\boldsymbol{S}_+^{n,k}(\mathcal{V})$ is much smaller than the original PSD cone $\boldsymbol{S}_+^n$. Therefore, the primal cone $\boldsymbol{S}_+^{n,k}(\mathcal{V})$ is too big to well approximate $\boldsymbol{S}_+^n$ in such a case.

Remark 6 Theorem 1 suggests that if some of the k-dimensional subspaces, $V_1, \dots, V_N$, are closely aligned to each other, then the effective size of $\mathcal{V}$ is reduced (as $w_G(T_{\mathcal{V}})$ decreases). Intuitively, we would want $V_1, \dots, V_N$ to be 'well spread' in the Grassmannian $\mathrm{Gr}(k, \mathbb{R}^n)$ to approximate $\boldsymbol{S}_+^n$ well with $\boldsymbol{S}_+^{n,k}(\mathcal{V})$, especially when N is small.

Remark 7 Note that the upper bound $w_G(S_{\mathcal{V}}) \le \sqrt{2k} + \sqrt{2\log N}$ in Theorem 1 holds regardless of the subspaces $V_1, \dots, V_N$, i.e., it is oblivious to the configuration of the subspaces in $\mathcal{V}$. That is, this upper bound is valid even for the "best" possible configuration of subspaces to imitate the expressive power of the full-sized PSD cone. We also note that this upper bound could conceivably be too conservative, especially when N is large, because it hinges on the union bound (through the use of Lemma 13).

Proof (Proof of Theorem 1) First of all, observe that

$$w_G\Big(B_H^*\left(\boldsymbol{S}_+^{n,k}(\mathcal{V})\right)\Big) = w_G\left(\boldsymbol{S}_+^{n,k}(\mathcal{V})^* \cap H - \frac{1}{n} I_n\right) = w_G\left(\boldsymbol{S}_+^{n,k}(\mathcal{V})^* \cap H\right)$$

due to the translation invariance of the Gaussian width. Letting $U_i \in \mathbb{R}^{n\times k}$ be a matrix whose columns form an orthonormal basis of V_i for each $V_i \in \mathcal{V}$, we observe that

$$\boldsymbol{S}_+^{n,k}(\mathcal{V})^* = \operatorname{cl}\operatorname{cone}\left(\bigcup_{i\in[N]} \{U_i Z U_i^T : Z \in \boldsymbol{S}_+^k\}\right)$$

because $(\boldsymbol{S}_+^k)^* = \boldsymbol{S}_+^k$ and $(C_1 \cap C_2)^* = \operatorname{cl}\operatorname{cone}(C_1^* \cup C_2^*)$, cf. Section 2.1. Thus, $\boldsymbol{S}_+^{n,k}(\mathcal{V})^* \cap H = \operatorname{conv}\Big(\bigcup_{i\in[N]}\{U_i v v^T U_i^T : v \in \mathbb{S}^{k-1}\}\Big)$, and therefore,

$$w_G\Big(B_H^*\left(\boldsymbol{S}_+^{n,k}(\mathcal{V})\right)\Big) = \mathbb{E}_G\left[\sup_{\substack{i\in[N]\\ v\in\mathbb{S}^{k-1}}} \left\langle G,\ U_i v v^T U_i^T\right\rangle\right] = \mathbb{E}_G\left[\sup_{u\in\left(\bigcup_{i=1}^N V_i\right)\cap\mathbb{S}^{n-1}} \left\langle G,\ uu^T\right\rangle\right] = w_G(S_{\mathcal{V}}).$$

Now, it remains to prove the two upper bounds on $w_G(S_{\mathcal{V}})$.

Upper Bound 1: Proof of $w_G(S_{\mathcal{V}}) \le \sqrt{2}\cdot w_G(T_{\mathcal{V}})$. Consider a Gaussian process $(X_u)_{u\in T_{\mathcal{V}}}$ such that $X_u = u^T G u + \gamma$ with G being standard Gaussian in $\boldsymbol{S}^n$ and $\gamma \sim N(0,1)$ independent of G. It is easy to verify that

$$w_G(S_{\mathcal{V}}) = \mathbb{E}_G\left[\sup_{u\in T_{\mathcal{V}}} \left\langle G,\ uu^T\right\rangle\right] = \mathbb{E}_{G,\gamma}\left[\sup_{u\in T_{\mathcal{V}}} \left\{u^T G u + \gamma\right\}\right] = \mathbb{E}_{G,\gamma}\left[\sup_{u\in T_{\mathcal{V}}} X_u\right].$$

Next, we introduce an instrumental Gaussian process $(Y_u)_{u\in T_{\mathcal{V}}}$ such that $Y_u = g^T u$ with $g \sim N(0, 2I_n)$. It is easy to check that (1) $\mathbb{E}X_u = \mathbb{E}Y_u = 0$ for all $u \in T_{\mathcal{V}}$; and (2) $\mathbb{E}(X_u - X_v)^2 \le \mathbb{E}(Y_u - Y_v)^2$ for all $u, v \in T_{\mathcal{V}}$ because $\mathbb{E}X_u^2 = \mathbb{E}Y_u^2 = 2$ and $\mathbb{E}X_u X_v - \mathbb{E}Y_u Y_v = (1 - u^T v)^2 \ge 0$. Now we can apply Sudakov-Fernique inequality (Lemma 7) to obtain $\mathbb{E}_{G,\gamma}\left[\sup_{u\in T_{\mathcal{V}}} X_u\right] \le \mathbb{E}_{g\sim N(0,2I_n)}\left[\sup_{u\in T_{\mathcal{V}}} Y_u\right] = \sqrt{2}\cdot w_G(T_{\mathcal{V}})$.

Upper Bound 2: Proof of $w_G(S_{\mathcal{V}}) \le \sqrt{2k} + \sqrt{2\log N}$. Observe that for each $i \in [N]$,

$$\sup_{v\in\mathbb{S}^{k-1}} \left\langle G,\ U_i v v^T U_i^T\right\rangle = \sup_{v\in\mathbb{S}^{k-1}} v^T (U_i^T G U_i) v = \lambda_1(U_i^T G U_i),$$

and that the random matrix $U_i^T G U_i \in \boldsymbol{S}^k$ has the standard Gaussian distribution in $\boldsymbol{S}^k$. Thus,

$$\begin{aligned} w_G(S_{\mathcal{V}}) &= \mathbb{E}_G\left[\sup_{i\in[N]} \lambda_1(U_i^T G U_i)\right] \\ &\le \sup_{i\in[N]} \mathbb{E}_G\left[\lambda_1(U_i^T G U_i)\right] + \mathbb{E}_G\left[\sup_{i\in[N]} \left(\lambda_1(U_i^T G U_i) - \mathbb{E}_G\left[\lambda_1(U_i^T G U_i)\right]\right)\right]. \end{aligned}$$

Note that $\mathbb{E}_G\left[\lambda_1(U_i^T G U_i)\right] \le \sqrt{2k}$ for all $i \in [N]$ by Lemma 8. For each $i \in [N]$, the function $G \mapsto \lambda_1(U_i^T G U_i)$ is 1-Lipschitz, and therefore, the random variable $\lambda_1(U_i^T G U_i) - \mathbb{E}_G\left[\lambda_1(U_i^T G U_i)\right]$ is sub-Gaussian with sub-Gaussian parameter 1 by Lemma 9. Then it follows from Lemma 13 that $\mathbb{E}_G\left[\sup_{i\in[N]}\left(\lambda_1(U_i^T G U_i) - \mathbb{E}_G\left[\lambda_1(U_i^T G U_i)\right]\right)\right] \le \sqrt{2\log N}$. □

Now we discuss how Theorem 1 implies the hardness of approximating $\boldsymbol{S}_+^n$ with a small number of $k \times k$ PSD constraints. In the next corollary, we show that if $N = |\mathcal{V}|$ is below a certain threshold determined by n, k, ϵ, then $\boldsymbol{S}_+^{n,k}(\mathcal{V})$ cannot be a dual-average ϵ-approximation of $\boldsymbol{S}_+^n$. Thus, it cannot be an ϵ-approximation of $\boldsymbol{S}_+^n$, either.

Corollary 1 *Let n, k be positive integers such that $1 \le k \le n$, and $\epsilon > 0$. If $\boldsymbol{S}_+^{n,k}(\mathcal{V})$ is a dual-average ϵ-approximation of $\boldsymbol{S}_+^n$, then $|\mathcal{V}| \ge \exp\big(n \cdot \varphi_{\text{dual-avg}}(n,k,\epsilon)\big)$ where*

$$\varphi_{\text{dual-avg}}(n,k,\epsilon) = \left[\frac{1}{1+\epsilon} \frac{w_G\big(B_H\left(\boldsymbol{S}_+^n\right)\big)}{\sqrt{2n}} - \sqrt{\frac{k}{n}} \right]_+^2.$$

Proof (Proof of Corollary 1) Suppose that $\boldsymbol{S}_+^{n,k}(\mathcal{V})$ is a dual-average ϵ-approximation of $\boldsymbol{S}_+^n$. Then by definition of the dual-average approximation (see Definitions 11 and 12),

$$w_G\Big(B_H\left(\boldsymbol{S}_+^{n,k}(\mathcal{V})\right)^\circ\Big) \ge \frac{1}{1+\epsilon} w_G\Big(B_H\left(\boldsymbol{S}_+^n\right)^\circ\Big). \tag{11}$$

By Lemma 1, we have $B_H\left(\boldsymbol{S}_+^{n,k}(\mathcal{V})\right)^\circ = -nB_H^*\left(\boldsymbol{S}_+^{n,k}(\mathcal{V})\right)$ and $B_H\left(\boldsymbol{S}_+^n \cap H\right)^\circ = -nB_H^*\left(\boldsymbol{S}_+^n\right) = -nB_H\left(\boldsymbol{S}_+^n\right)$ because $\boldsymbol{S}_+^n$ is self-dual. Thus, Theorem 1, combined with the inequality (11), implies

$$\frac{1}{1+\epsilon} w_G\Big(B_H\left(\boldsymbol{S}_+^n\right)\Big) \le w_G\Big(B_H\left(\boldsymbol{S}_+^{n,k}(\mathcal{V})^*\right)\Big) \le \sqrt{2k} + \sqrt{2\log|\mathcal{V}|}.$$

Note that this inequality holds if and only if

$$\sqrt{\log|\mathcal{V}|} \ge \frac{1}{\sqrt{2}(1+\epsilon)} w_G\Big(B_H\left(\boldsymbol{S}_+^n\right)\Big) - \sqrt{k},$$

which is again equivalent to

$$|\mathcal{V}| \ge \exp\left(n \cdot \left[\frac{1}{1+\epsilon} \frac{w_G\big(B_H\left(\boldsymbol{S}_+^n\right)\big)}{\sqrt{2n}} - \sqrt{\frac{k}{n}} \right]_+^2\right) = \exp\big(n \cdot \varphi_{\text{dual-avg}}(n,k,\epsilon)\big). \quad \square$$

Remark 8 Recall from Remark 3 that $\lim_{n\to\infty} w_G\big(B_H\left(\boldsymbol{S}_+^n\right)\big)/\sqrt{2n} = 1$. With $k = \lfloor \delta n \rfloor$ for $0 < \delta < 1$,

$$\lim_{n\to\infty} \varphi_{\text{dual-avg}}\big(n, \lfloor \delta n \rfloor, \epsilon\big) = \left[\frac{1}{1+\epsilon} - \sqrt{\delta} \right]_+^2.$$

That is, when n is sufficiently large, $|\mathcal{V}| \ge \exp\big(n\big[1/(1+\epsilon) - \sqrt{\delta}\big]_+^2\big)$ is necessary for the cone $\boldsymbol{S}_+^{n,k}(\mathcal{V})$ to be a dual-average ϵ-approximation of $\boldsymbol{S}_+^n$.

As discussed in Remark 7, our lower bound in Corollary 1 can be conservative due to the union bound. In fact, we do not know whether our lower bound is tight. Thus, it is possible that even when $N \ge \exp\big(n \cdot \varphi_{\text{dual-avg}}(n,k,\epsilon)\big)$, there does not exist any $\mathcal{V}$ such that $|\mathcal{V}| = N$ and $\boldsymbol{S}_+^{n,k}(\mathcal{V})$ is a dual-average ϵ-approximation of $\boldsymbol{S}_+^n$.

The next corollary states that if $N = |\mathcal{V}|$ is below a certain threshold, then $\boldsymbol{S}_+^{n,k}(\mathcal{V})$ cannot be an ϵ-approximation of $\boldsymbol{S}_+^n$ in the average sense, either.

Corollary 2 *Let n, k be positive integers such that $n \ge 3$ and $1 \le k \le n$, and $\epsilon > 0$. If $\boldsymbol{S}_+^{n,k}(\mathcal{V})$ is an average ϵ-approximation of $\boldsymbol{S}_+^n$, then $|\mathcal{V}| \ge \exp\big(n \cdot \varphi_{\text{avg}}(n,k,\epsilon)\big)$ where*

$$\varphi_{\text{avg}}(n,k,\epsilon) = \left[\frac{1}{4(1+\epsilon)} \frac{\sqrt{2n}}{w_G\big(B_H(\boldsymbol{S}_+^n)\big)} - \sqrt{\frac{k}{n}} \right]_+^2.$$

Proof (Proof of Corollary 2) Suppose that $\boldsymbol{S}_+^{n,k}(\mathcal{V})$ is an average ϵ-approximation of $\boldsymbol{S}_+^n$. Then by definition of the average approximation (see Definitions 10 and 12),

$$w_G\big(B_H(\boldsymbol{S}_+^{n,k}(\mathcal{V}))\big) \le (1+\epsilon) \cdot w_G\big(B_H(\boldsymbol{S}_+^n)\big). \tag{12}$$

By Lemma 1, we observe that $B_H\big(\boldsymbol{S}_+^{n,k}(\mathcal{V})\big) = -\frac{1}{n} B_H^*\big(\boldsymbol{S}_+^{n,k}(\mathcal{V})\big)^\circ$. It follows from Urysohn's inequality (Lemma 3) that $w\big(B_H(\boldsymbol{S}_+^{n,k}(\mathcal{V}))\big) \ge \frac{1}{n \cdot w\left(B_H^*(\boldsymbol{S}_+^{n,k}(\mathcal{V}))\right)}$, and therefore,

$$w_G\big(B_H(\boldsymbol{S}_+^{n,k}(\mathcal{V}))\big) \ge \frac{\kappa_d^2}{n \cdot w_G\big(B_H^*(\boldsymbol{S}_+^{n,k}(\mathcal{V}))\big)}$$

where $d = \binom{n+1}{2} - 1$ is the dimension of the affine subspace H and $\kappa_d = \frac{\sqrt{2}\Gamma((d+1)/2)}{\Gamma(d/2)}$. Because $\kappa_d^2 \geq d - \frac{1}{2}$ (see Remark 2) and $\binom{n+1}{2} - \frac{3}{2} \geq \frac{n^2}{2}$ for all $n \geq 3$, we obtain that

$$\begin{aligned}\frac{w_G\big(B_H\big(\boldsymbol{S}_+^{n,k}(\mathcal{V})\big)\big)}{w_G\big(B_H\big(\boldsymbol{S}_+^n\big)\big)} &\geq \frac{\sqrt{2n}}{w_G\big(B_H\big(\boldsymbol{S}_+^n\big)\big)} \frac{\sqrt{2n}}{w_G\big(B_H^*\big(\boldsymbol{S}_+^{n,k}(\mathcal{V})\big)\big)} \cdot \frac{1}{2n^2}\left\{\binom{n+1}{2} - \frac{3}{2}\right\} \\ &\geq \frac{1}{4}\frac{\sqrt{2n}}{w_G\big(B_H\big(\boldsymbol{S}_+^n\big)\big)} \frac{\sqrt{2n}}{w_G\big(B_H^*\big(\boldsymbol{S}_+^{n,k}(\mathcal{V})\big)\big)}.\end{aligned}$$

Recall from Theorem 1 that $w_G\big(B_H^*\big(\boldsymbol{S}_+^{n,k}(\mathcal{V})\big)\big) \leq \sqrt{2k} + \sqrt{2\log|\mathcal{V}|}$. Thus, if $\frac{w_G\left(B_H\left(\boldsymbol{S}_+^{n,k}(\mathcal{V})\right)\right)}{w_G\left(B_H\left(\boldsymbol{S}_+^n\right)\right)} \leq 1+\epsilon$, then

$$\sqrt{\log|\mathcal{V}|} \geq \frac{\sqrt{n}}{4(1+\epsilon)} \frac{\sqrt{2n}}{w_G\big(B_H\big(\boldsymbol{S}_+^n\big)\big)} - \sqrt{k},$$

which is equivalent to

$$|\mathcal{V}| \geq \exp\left(n \cdot \left[\, \frac{1}{4(1+\epsilon)} \frac{\sqrt{2n}}{w_G\big(B_H\big(\boldsymbol{S}_+^n\big)\big)} - \sqrt{\frac{k}{n}} \,\right]_+^2\right) = \exp\big(n \cdot \varphi_{\text{avg}}(n,k,\epsilon)\big). \quad \square$$

Remark 9 Note that the cardinality lower bound in Corollary 2 is weaker than that in Corollary 1, due to the additional 1/4 factor in the expression of φ_{avg} (in comparison with $\varphi_{\text{dual-avg}}$ in the limit $n \to \infty$) that is introduced as a result of applying Urysohn's inequality (Lemma 3). Thus, establishing a direct lower bound on $w_G\big(B_H \boldsymbol{S}_+^{n,k}(\mathcal{V})\big)$ can possibly accomplish a more refined analysis of average ϵ-approximability of $\boldsymbol{S}_+^n$ by $\boldsymbol{S}_+^{n,k}(\mathcal{V})$, yielding an improved cardinality lower bound than that in Corollary 2.

4.2 Example: the Sparse k-PSD Approximation of $\boldsymbol{S}_+^n$

In this section, we consider the sparse k-PSD approximation, which is a concrete example of the k-PSD approximation of $\boldsymbol{S}_+^n$ (Definition 13) discussed in the previous section.

Definition 14 (Sparse k-PSD approximation of $\boldsymbol{S}_+^n$) Given positive integers n and $1 \leq k \leq n$, the sparse k-PSD approximation of $\boldsymbol{S}_+^n$ is the set

$$\boldsymbol{S}_+^{n,k} := \big\{X \in \boldsymbol{S}^n : X_I \succeq 0, \quad \forall I \subset [n] \text{ with } |I| \leq k\big\}.$$

We observe that the sparse k-PSD approximation is an instance of the k-PSD approximation $\boldsymbol{S}_+^{n,k}(\mathcal{V})$ such that $\mathcal{V} = \{V_I : I \in [n] \text{ with } |I| = k\}$ where $V_I = \{v \in \mathbb{R}^n : v_i = 0, \ \forall i \notin I\}$. Note that $|\mathcal{V}| = \binom{n}{k}$.

In Section 4.2.1, we examine the implications of Corollary 1 for the sparse k-PSD approximation of $\boldsymbol{S}_+^n$. In Section 4.2.2, we provide a more refined analysis that is tailored to $\boldsymbol{S}_+^{n,k}$, based on properties that are specific to $\boldsymbol{S}_+^{n,k}$. It turns out that we can derive stronger hardness results from the tailored approach.

4.2.1 A Weak Bound Using Corollary 1

First of all, we inspect what the lower bound obtained in Section 4.1 implies for the sparse k-PSD approximation of $\boldsymbol{S}_+^n$. According to the contrapositive of Corollary 1, when n and $\epsilon > 0$ are fixed, $\boldsymbol{S}_+^{n,k}$ cannot be a dual-average ϵ-approximation of $\boldsymbol{S}_+^n$ if k satisfies the following inequality:

$$\binom{n}{k} < \exp\left(n \cdot \left[\, \frac{1}{1+\epsilon} \frac{w_G\big(B_H\left(\boldsymbol{S}_+^n\right)\big)}{\sqrt{2n}} - \sqrt{\frac{k}{n}} \,\right]_+^2\right). \tag{13}$$

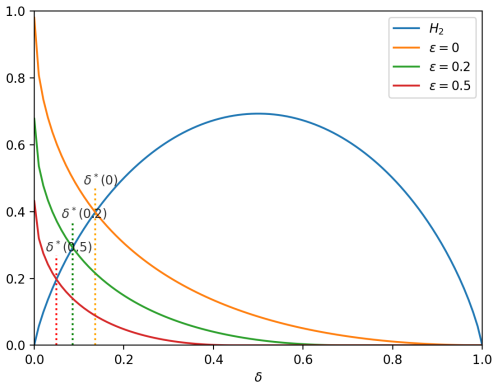
Let's assume $k = \delta n$ for some $0 < \delta < 1$ and n tends to infinity. By Stirling's approximation,

$$\log\binom{n}{k} = \big(1 + o_n(1)\big) H_2\left(\frac{k}{n}\right) n,$$

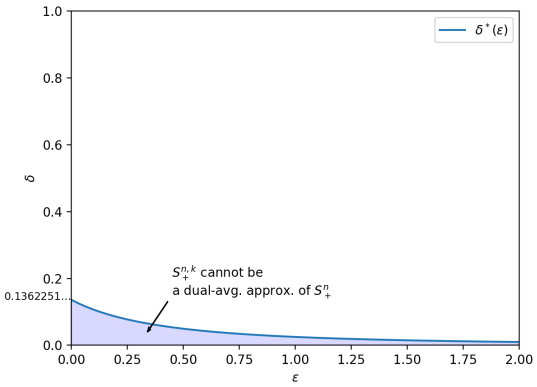
where $H_2(p) = -p\log p - (1-p)\log(1-p)$ is the binary entropy function defined for $p \in [0,1]$. With this asymptotic approximation and the observation that $w_G\big(B_H\left(\boldsymbol{S}_+^n\right)\big)/\sqrt{2n} \leq 1$, we take logarithm of both sides of (13) to obtain the inequality (in the limit $n \to \infty$),

$$H_2(\delta) < \left[\, \frac{1}{1+\epsilon} - \sqrt{\delta} \,\right]_+^2. \tag{14}$$

Given $\epsilon \geq 0$, let $g_\epsilon(\delta) := \left[\frac{1}{1+\epsilon} - \sqrt{\delta} \right]_+^2 - H_2(\delta)$. Note that g_ϵ is strictly convex on the interval $\delta \in [0,1]$ and $g_\epsilon(0) > 0$. Moreover, if $\epsilon > 0$, then $g_\epsilon(1/(1+\epsilon)^2) < 0$. By the intermediate value theorem, there exists a unique $0 < \delta^*(\epsilon) < 1/(1+\epsilon)^2$ such that $g_\epsilon(\delta^*(\epsilon)) = 0$ and $g_\epsilon(\delta) > 0$ for all $0 \leq \delta < \delta^*(\epsilon)$. As a result, if $k/n < \delta^*(\epsilon)$, then $\boldsymbol{S}_+^{n,k}$ cannot be a dual-average ϵ-approximation of $\boldsymbol{S}_+^n$. The expressions on both sides of Eq. (14) are illustrated in Figure 4a for a few values of ϵ; the plot of $\delta^*(\epsilon)$ vs ϵ is depicted in Figure 4b.



(a) Plot of the expressions in Eq. (14): $H_2(\delta)$ (entropy) vs $[\, 1/(1+\epsilon) - \sqrt{\delta}\,]_+^2$ for $\epsilon = 0$, 0.2, and 0.5. The location of $\delta^*(\epsilon)$ are also annotated.



(b) Plot of $\delta^*(\epsilon)$ vs ϵ. For a fixed $\epsilon > 0$, if k/n is contained in the blue region, $\boldsymbol{S}_+^{n,k}$ cannot be a dual-average ϵ-approximation of $\boldsymbol{S}_+^n$.

Fig. 4: Illustration of the hardness results obtained by applying Corollary 1 to the sparse k-PSD.

Recall the definition of $\epsilon^*_{\text{dual-avg}}(P, S) = \inf\{\epsilon > 0 : S \text{ is a dual-average } \epsilon \text{ approximation of } P\}$, which indicates the 'best possible' (i.e., the smallest) $\epsilon > 0$ for which S is a dual-average ϵ-approximation of P. For fixed n and k, the preceding discussion leads to a lower bound on $\epsilon^*_{\text{dual-avg}}(\boldsymbol{S}_+^n, \boldsymbol{S}_+^{n,k})$ as

$$\epsilon^*_{\text{dual-avg}}(\boldsymbol{S}_+^n, \boldsymbol{S}_+^{n,k}) \geq \sup\left\{ \epsilon > 0 : H_2\Big(\frac{k}{n}\Big) < \left[\frac{1}{1+\epsilon} - \sqrt{\frac{k}{n}} \right]_+^2 \right\} =: \xi(k/n). \tag{15}$$

On the one hand, we can already see from the above discussion that for any fixed $\epsilon > 0$, $\boldsymbol{S}_+^{n,k}$ with $k = o_n(n)$ cannot be an ϵ-approximation of $\boldsymbol{S}_+^n$ (in the dual-average sense). That is, k must scale linearly with respect to n for $\boldsymbol{S}_+^{n,k}$ to be a good approximation of $\boldsymbol{S}_+^n$. On the other hand, the lower bound on k from the discussion above – $k/n \geq \delta^*(\epsilon)$ – becomes uninformative once k increases beyond a certain threshold because $\delta^*(\epsilon) < \delta^*(0) \approx 0.137$ for all $\epsilon > 0$. In other words, if $k/n > \delta^*(0)$, then we can only get a trivial lower bound $\epsilon^*_{\text{dual-avg}}(\boldsymbol{S}_+^n, \boldsymbol{S}_+^{n,k}) > -\infty$, and do not know whether $\boldsymbol{S}_+^{n,k}$ approximates $\boldsymbol{S}_+^n$ well or not.

We remark that this is possibly due to the conservative nature of inequality (13), which is inherited from Corollary 1. Recall that the cardinality lower bound from Corollary 1 is oblivious to the configuration of the subspaces $V_1, \ldots, V_N$ in $\mathcal{V}$. That is, it is valid even for the "best" possible configuration of subspaces to imitate the expressive power of the full-sized PSD cone. Nevertheless, the subspaces of k-sparse vectors have overlaps, and some of them could be redundant. Thus, the general lower bound from Corollary 1 can be excessively conservative to apply to the sparse k-PSD approximation of $\boldsymbol{S}_+^n$.

Indeed, we can acquire a tighter lower bound for $\epsilon^*_{\text{dual-avg}}(\boldsymbol{S}_+^n, \boldsymbol{S}_+^{n,k})$ by using the knowledge about the subspaces of $\boldsymbol{S}_+^{n,k}$. This is the topic that will be discussed in Section 4.2.2.

4.2.2 A More Refined Analysis Tailored to $\boldsymbol{S}_+^{n,k}$

In this section, we derive lower bounds on $\epsilon^*(\boldsymbol{S}_+^n, \boldsymbol{S}_+^{n,k})$ and $\epsilon^*_{\text{dual-avg}}(\boldsymbol{S}_+^n, \boldsymbol{S}_+^{n,k})$ with an analysis that exploits specific properties of $\boldsymbol{S}_+^{n,k}$. More precisely, we construct a matrix on the boundary of $B_H(\boldsymbol{S}_+^{n,k})$ to argue a lower bound on $\epsilon^*(\boldsymbol{S}_+^n, \boldsymbol{S}_+^{n,k})$ (see Proposition 1), and characterize $\epsilon^*_{\text{dual-avg}}(\boldsymbol{S}_+^n, \boldsymbol{S}_+^{n,k})$ by observing that the Gaussian width of $B_H^*(\boldsymbol{S}_+^{n,k})$ is the expectation of the largest k-sparse eigenvalue

of a standard Gaussian random matrix (see Proposition 2). The resulting lower bounds imply stronger hardness results for approximating $\boldsymbol{S}_+^n$ with $\boldsymbol{S}_+^{n,k}$ than those discussed in Section 4.2.1. In addition, we discuss a lower bound on $\epsilon^*_{\text{avg}}(\boldsymbol{S}_+^n, \boldsymbol{S}_+^{n,k})$ obtained as an immediate corollary (Corollary 3).

Hardness of ϵ-approximation First of all, we discuss how hard it is to approximate $\boldsymbol{S}_+^n$ with $\boldsymbol{S}_+^{n,k}$ in the ϵ-approximation sense (see Definition 9) when k is small. For that purpose, we consider a specific matrix on the line segment connecting $\frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$ and $\frac{1}{n}I_n$ where $\mathbf{1}_n \in \mathbb{R}^{n\times 1}$ denotes the $n\times 1$ column matrix with all entries equal to 1. Specifically, we construct a matrix $M \in B_H(\boldsymbol{S}_+^{n,k})$ that is far away from $B_H\left(\boldsymbol{S}_+^n\right)$, and prove a lower bound for $\epsilon > 0$ as a necessary condition for $M \in (1+\epsilon)\cdot B_H\left(\boldsymbol{S}_+^n\right)$.

Proposition 1 *If $\boldsymbol{S}_+^{n,k}$ is an ϵ-approximation of $\boldsymbol{S}_+^n$, then $k > \frac{n-1}{1+\epsilon}$.*

Proof Let $P_1(n) := \mathbf{1}_n\mathbf{1}_n^T/n$ and $P_2(n) := I_n - P_1(n)$. Note that $P_1(n)$ and $P_2(n)$ are projection matrices. For $a, b \in \mathbb{R}$, we define

$$G(a,b;n) := aP_1(n) + bP_2(n).$$

It is easy to verify that the eigenvalues of $G(a,b;n)$ are a with multiplicity 1, and b with multiplicity $n-1$.

Next, recall from Definition 14 that $G(a,b;n) \in \boldsymbol{S}_+^{n,k}$ if and only if $G(a,b;n)_{[k]} \succeq 0$. Observe that $G(a,b;n)_{[k]} = \frac{ka+(n-k)b}{n}P_1(k) + bP_2(k) \succeq 0$ if and only if $ka+(n-k)b \geq 0$ and $b \geq 0$. Letting $a = \frac{k-n}{n(k-1)}$ and $b = \frac{k}{n(k-1)}$, we observe that (1) $G(a,b;n) \in \boldsymbol{S}_+^{n,k}$ because $ka+(n-k)b = 0$ and $b \geq 0$; and (2) $G(a,b;n) \in H$ because $\operatorname{Tr} G(a,b;n) = a + b(n-1) = 1$. Next, we can also verify that $G(a,b;n) - \frac{1}{n}I_n \in (1+\epsilon)\cdot B_H\left(\boldsymbol{S}_+^n\right)$ if and only if $\epsilon \geq \frac{n-k}{k-1}$. It is because $G(a,b;n) + \frac{\epsilon}{n}I_n = G\left(a+\frac{\epsilon}{n}, b+\frac{\epsilon}{n}; n\right) \in \boldsymbol{S}_+^n$ if and only if $a + \frac{\epsilon}{n} \geq 0$. Rewriting $\epsilon \geq \frac{n-k}{k-1}$ as a condition for k in terms of ϵ, we obtain $k \geq \frac{n-1}{1+\epsilon} + 1$. □

Alternatively, when k is fixed, Proposition 1 implies that

$$\epsilon^*(\boldsymbol{S}_+^n, \boldsymbol{S}_+^{n,k}) \geq \frac{n-k}{k-1} \geq \frac{1-k/n}{k/n} =: \zeta(k/n). \tag{16}$$

Hardness of dual average ϵ-approximation Next, we re-examine how well $\boldsymbol{S}_+^{n,k}$ can approximate $\boldsymbol{S}_+^n$ in the dual-average sense (Definition 11) to find a better lower bound on $\epsilon^*_{\text{dual-avg}}(\boldsymbol{S}_+^n, \boldsymbol{S}_+^{n,k})$. We use the duality between $\boldsymbol{S}_+^{n,k}$ and its dual cone, $(\boldsymbol{S}_+^{n,k})^* = \operatorname{cone}\{vv^T : v \in \mathbb{R}^n \text{ with } \|v\|_0 \leq k\}$, which is the cone of matrices that have factor width at most k [8].

Observe that $\boldsymbol{S}_+^n \cap H = \{X \in \boldsymbol{S}_+^n : \operatorname{Tr}(X) = 1\} = \operatorname{conv}\{vv^T : x \in \mathbb{R}^n,\ \|v\|_2 = 1\}$. For any $G \in \boldsymbol{S}^n$, $\max_{X\in\boldsymbol{S}_+^n\cap H}\langle G, X\rangle = \lambda_1(G)$ and thus, $w_G\left(\boldsymbol{S}_+^n \cap H\right)$ is equal to the expectation of the largest eigenvalue of a random matrix that has the standard Gaussian distribution in $\boldsymbol{S}^n$ (Definition 5). Likewise, $(\boldsymbol{S}_+^{n,k})^* \cap H = \operatorname{conv}\{vv^T : x \in \mathbb{R}^n,\ \|v\|_2 = 1,\ \|v\|_0 \leq k\}$, and $\max_{X\in(\boldsymbol{S}_+^{n,k})^*\cap H}\langle G, X\rangle$ is the largest k-sparse eigenvalue of G. Based on these observations, we show an asymptotic upper bound on the ratio $w_G\left(B_H^*(\boldsymbol{S}_+^{n,k})\right)/w_G(B_H^*(\boldsymbol{S}_+^n))$ in Proposition 2 that subsequently leads to a tighter lower bound on $\epsilon^*_{\text{dual-avg}}(\boldsymbol{S}_+^n, \boldsymbol{S}_+^{n,k})$ in (19).

Proposition 2 *Fix $0 < \delta < 1$ and let $k = \lfloor \delta n \rfloor$. Then*

$$\lim_{n\to\infty} \frac{w_G\left(B_H^*(\boldsymbol{S}_+^{n,k})\right)}{w_G\left(B_H^*(\boldsymbol{S}_+^n)\right)} \leq \left(\int_0^\delta Q_{\chi^2}(1-s)ds\right)^{1/2}, \tag{17}$$

where Q_{χ^2} denotes the quantile function[4] of the χ^2-distribution with one degree of freedom. Moreover,

$$\int_0^\delta Q_{\chi^2}(1-s)ds = \delta + \sqrt{\frac{2}{\pi}}\Phi^{-1}\left(1-\frac{\delta}{2}\right)\exp\left(-\frac{1}{2}\left[\Phi^{-1}\left(1-\frac{\delta}{2}\right)\right]^2\right) \tag{18}$$

where $\Phi(x)$ is the cumulative distribution function of the standard normal distribution.

[4] That is, $Q_{\chi^2}(s) := \inf\{x \in \mathbb{R} : F_{\chi^2}(x) \geq s\}$ for $0 < s \leq 1$ where F_{χ^2} be the cumulative distribution function of the χ^2-distribution with one degree of freedom.

Before we prove Proposition 2, we note that it implies the following lower bound in the asymptotic limit $n \to \infty$:

$$\epsilon^*_{\text{dual-avg}}(\boldsymbol{S}^n_+, \boldsymbol{S}^{n,k}_+) \geq \left(\int_0^\delta Q_{\chi^2}(1-s)ds \right)^{-1/2} - 1 =: \psi(k/n). \tag{19}$$

See Figure 1 (left) in Section 1 to compare the three lower bounds, ξ (Corollary 1 and (15)), ζ (Proposition 1 and (16)), and ψ (Proposition 2 and (19)). We make two remarks: one on the advantage of tailored analysis for $\boldsymbol{S}^{n,k}_+$; and the other on comparing the rate of convergence for ζ vs ψ.

- (Generic vs tailored) The lower bound ψ gives a sharper lower bound than ξ. In particular, $\psi(\delta) > 0$ for all $0 < \delta < 1$ and ψ gracefully converges to 0 as $k/n \to 1$, whereas $\xi(\delta) = 0$ for all $\delta \geq \delta^*(0)$.
- (ϵ-approx. vs dual-avg. ϵ-approx.) We can see from the expression in (18) that $\psi(1-\delta) = \Theta_\delta(\delta^3)$ as $\delta \to 0$. This sharply contrasts with $\varphi(1-\delta) = \Theta_\delta(\delta)$. That is, $\boldsymbol{S}^{n,k}_+$ gets harder to approximate $\boldsymbol{S}^n_+$ in both senses as k diminishes from n, but at a much slower rate in the dual-average sense.

Proof (Proof of Proposition 2) Fix $k \in \{0, 1, \ldots, n\}$. Let $T = \{u \in \mathbb{R}^n : \|u\|_2 \leq 1, \|u\|_0 \leq k\}$ and observe that

$$w_G\big(B^*_H(\boldsymbol{S}^{n,k}_+)\big) = w_G\big((\boldsymbol{S}^{n,k}_+)^* \cap H\big) = \mathbb{E}_G\left[\sup_{u \in T} \langle G,\ uu^T \rangle \right].$$

We consider a Gaussian process $(X_u)_{u \in T}$ such that $X_u = u^T G u + \gamma$ with G being standard Gaussian in $\boldsymbol{S}^n$ and $\gamma \sim N(0,1)$ independent of G. It is easy to verify that

$$\mathbb{E}_G\left[\sup_{u \in T} \langle G,\ uu^T \rangle \right] = \mathbb{E}_{G,\gamma}\left[\sup_{u \in T} \{u^T G u + \gamma\} \right] = \mathbb{E}_{G,\gamma}\left[\sup_{u \in T} X_u \right].$$

Next, we introduce an instrumental Gaussian process $(Y_u)_{u \in T}$ such that $Y_u = g^T u$ with $g \sim N(0, 2I_n)$. It is easy to check that for all $u, v \in T$, (1) $\mathbb{E}X_u = \mathbb{E}Y_u = 0$; and (2) $\mathbb{E}(X_u - X_v)^2 \leq \mathbb{E}(Y_u - Y_v)^2$ because $\mathbb{E}X_u^2 = \mathbb{E}Y_u^2 = 2$ and $\mathbb{E}X_u X_v - \mathbb{E}Y_u Y_v = (1 - u^T v)^2 \geq 0$. Now we can apply Sudakov-Fernique inequality (Lemma 7) to obtain $\mathbb{E}_{G,\gamma}\big[\sup_{u \in T} X_u\big] \leq \mathbb{E}_{g \sim N(0,2I_n)}\big[\sup_{u \in T} Y_u\big]$. Then it follows that

$$w_G\big(B^*_H(\boldsymbol{S}^{n,k}_+)\big) = \mathbb{E}_{G,\gamma}\left[\sup_{u \in T} X_u \right] \leq \mathbb{E}_{g \sim N(0,2I_n)}\big[\sup_{u \in T} Y_u \big] = \mathbb{E}_{g \sim N(0,2I_n)} \sup_{\substack{u \in \mathbb{R}^n \\ \|u\|_2 \leq 1, \|u\|_0 \leq k}} g^T u.$$

Therefore,

$$\frac{1}{\sqrt{2n}} w_G\big(B^*_H(\boldsymbol{S}^{n,k}_+)\big) \leq \mathbb{E}_{g \sim N(0,2I_n)}\left[\frac{\|g\|_2}{\sqrt{2n}} \sup_{\substack{u \in \mathbb{R}^n \\ \|u\|_2 \leq 1, \|u\|_0 \leq k}} \frac{g^T u}{\|g\|_2} \right].$$

Note that when $g \sim N(0, 2I_n)$, $\frac{\|g\|_2}{\sqrt{2n}} \to 1$ in probability as $n \to \infty$. Thus, it suffices to identify the limit of $\sup_{\substack{u \in \mathbb{R}^n \\ \|u\|_2 \leq 1, \|u\|_0 \leq k}} \frac{g^T u}{\|g\|_2}$ (in probability) to compute the expectation on the right-hand side.

Given $x \in \mathbb{R}^n$, we let $(x_i^2)^\downarrow$ denote the i-th largest element in the set $\{x_1^2, x_2^2, \ldots, x_n^2\}$. Observe that

$$\sup_{\substack{u \in \mathbb{R}^n \\ \|u\|_2 \leq 1, \|u\|_0 \leq k}} \frac{g^T u}{\|g\|_2} = \frac{1}{\|g\|_2} \frac{\sum_{i=1}^k (g_i^2)^\downarrow}{\sqrt{\sum_{i=1}^k (g_i^2)^\downarrow}} = \left(\frac{1}{\|g\|_2^2} \sum_{i=1}^k (g_i^2)^\downarrow \right)^{1/2}$$

and that $(g_1^2)^\downarrow \geq (g_2^2)^\downarrow \geq \cdots \geq (g_n^2)^\downarrow$ are χ^2 order statistics of degree 1, multiplied by a factor of 2. It is well known from literature on extreme order statistics (e.g., [18, Theorem 2.7]) that for any fixed $0 < \delta < 1$,

$$\frac{1}{\|g\|_2^2} \sum_{i=1}^{\lfloor \delta n \rfloor} (g_i^2)^\downarrow \longrightarrow \int_0^\delta Q_{\chi^2}(1-s)ds \qquad \text{in probability as } n \to \infty.$$

Combining these observations and the well-known fact that $\lim_{n \to \infty} \frac{w_G(B^*_H(\boldsymbol{S}^n_+))}{\sqrt{2n}} = 1$, cf. Remark 5, we obtain the desired inequality:

$$\lim_{n \to \infty} \frac{w_G\big(B^*_H(\boldsymbol{S}^{n,k}_+)\big)}{w_G\big(B^*_H(\boldsymbol{S}^n_+)\big)} = \lim_{n \to \infty} \frac{\sqrt{2n}}{w_G\big(B^*_H(\boldsymbol{S}^n_+)\big)} \lim_{n \to \infty} \frac{w_G\big(B^*_H(\boldsymbol{S}^{n,k}_+)\big)}{\sqrt{2n}} \leq \left(\int_0^\delta Q_{\chi^2}(1-s)ds \right)^{1/2}.$$

We conclude the proof by computing the integral in the upper bound. An explicit formula for the integral is well known; see [18, Remark 2.8], for example.

$$\int_0^\delta Q_{\chi^2}(1-s)ds = 2 \int_{\Phi^{-1}(1-\frac{\delta}{2})}^\infty s^2 \Phi'(s)ds = \delta + \sqrt{\frac{2}{\pi}} \Phi^{-1}\left(1 - \frac{\delta}{2}\right) \exp\left(-\frac{1}{2}\left[\Phi^{-1}\left(1 - \frac{\delta}{2}\right) \right]^2 \right). \quad \square$$

Hardness of average ϵ-approximation As a matter of fact, we can derive the following corollary from Proposition 2 by applying Urysohn's inequality (Lemma 3), thereby obtaining an asymptotic lower bound on $\epsilon^*_{\text{avg}}(\boldsymbol{S}^n_+, \boldsymbol{S}^{n,k}_+)$ (see Definition 10).

Corollary 3 *Fix $0 < \delta < 1$ and let $k = \lfloor \delta n \rfloor$. Then*

$$\lim_{n\to\infty} \frac{w_G\big(B_H(\boldsymbol{S}^{n,k}_+)\big)}{w_G\big(B_H(\boldsymbol{S}^n_+)\big)} \geq \frac{1}{4}\left(\int_0^\delta Q_{\chi^2}(1-s)ds \right)^{-1/2}.$$

Corollary 3 implies that $\epsilon^*_{\text{avg}}(\boldsymbol{S}^n_+, \boldsymbol{S}^{n,k}_+) \geq \frac{1}{4}\big(\int_0^\delta Q_{\chi^2}(1-s)ds\big)^{-1/2} - 1$. Note that this lower bound is more conservative than the lower bound for $\epsilon^*_{\text{dual-avg}}(\boldsymbol{S}^n_+, \boldsymbol{S}^{n,k}_+)$ in (19), due to the additional multiplier $1/4$ that arises from the use of Urysohn's inequality. It might be possible to derive a better lower bound for $\epsilon^*_{\text{avg}}(\boldsymbol{S}^n_+, \boldsymbol{S}^{n,k}_+)$, which is beyond the scope of this paper.

Proof (Proof of Corollary 3) By Lemma 1, we observe that $B_H(\boldsymbol{S}^{n,k}_+) = -\frac{1}{n} B^*_H(\boldsymbol{S}^{n,k}_+)^\circ$. It follows from Lemma 3 that $w\big(B_H(\boldsymbol{S}^{n,k}_+)\big) \geq \frac{1}{n \cdot w\left(B^*_H(\boldsymbol{S}^{n,k}_+)\right)}$, and therefore,

$$w_G\big(B_H(\boldsymbol{S}^{n,k}_+)\big) \geq \frac{\kappa_d^2}{n \cdot w_G\big(B^*_H(\boldsymbol{S}^{n,k}_+)\big)}$$

where $d = \binom{n+1}{2} - 1$ is the dimension of H. Since $\kappa_d^2 \geq d - \frac{1}{2}$, we obtain for any $0 < \delta < 1$,

$$\lim_{n\to\infty} \frac{w_G\big(B_H(\boldsymbol{S}^{n,k}_+)\big)}{\sqrt{2n}} \geq \lim_{n\to\infty} \frac{\sqrt{2n}}{w_G\big(B^*_H(\boldsymbol{S}^{n,k}_+)\big)} \frac{1}{2n^2}\left\{ \binom{n+1}{2} - \frac{3}{2} \right\} = \frac{1}{4f(\delta)}. \quad \square$$

5 Approximate Extended Formulations of $\boldsymbol{S}^n_+$

Now we further extend our discussion beyond the k-PSD approximation. Specifically, we consider an arbitrary approximation of $\boldsymbol{S}^n_+$ through extended formulations. This defines a much broader class of approximations as we are allowed to introduce as many new variables as we want. However, even in this case, at least superpolynomially many $k \times k$ PSD constraints are required to approximate $\boldsymbol{S}^n_+$ when $k \ll n$. In Section 5.1, we present two main theorems about the extension complexity lower bounds that hold for any ϵ-approximation of $B_H(\boldsymbol{S}^n_+)$. Sections 5.2 and 5.3 are dedicated to the proof of the theorems.

5.1 Theorem Statements

Recall that $B_H\left(\boldsymbol{S}^n_+\right) = \boldsymbol{S}^n_+ \cap H - \frac{1}{n} I_n$. In this section, we present two main theorems on the hardness of approximating $B_H\left(\boldsymbol{S}^n_+\right)$ with a small number of $k \times k$ PSD constraints. Our first theorem is about an $\boldsymbol{S}^k_+$-extension complexity lower bound that holds for any ϵ-approximation of $B_H\left(\boldsymbol{S}^n_+\right)$.

Theorem 2 *There exists a constant $C > 0$ such that if S is an ϵ-approximation of $B_H\left(\boldsymbol{S}^n_+\right)$, then*

$$\text{xc}_{\boldsymbol{S}^k_+}(S) \geq \exp\left(C \cdot \min\left\{ \sqrt{\frac{n}{1+\epsilon}},\ \frac{1}{1+\epsilon}\ \frac{n}{k} \right\} \right)$$

for all positive integers n and k satisfying $1 \leq k \leq n$.

Theorem 2 suggests that at least $\Omega_n(\exp(\sqrt{n}))$ copies of $\boldsymbol{S}^k_+$ are required to approximate $\boldsymbol{S}^n_+$ when $k = O_n(\sqrt{n})$. When $k = \Omega_n(\sqrt{n})$, this extension complexity lower bound gracefully decreases to 1 as k increases to n. We remark that Theorem 2 holds for arbitrary k, and thus, extends the result of Fawzi [11, Theorem 1] beyond the special case of $k = 1$. A more formal version of Theorem 2 and its proof are deferred until Section 5.2.

Next, we consider the $\boldsymbol{S}^k_+$-extension complexity of a set that is an average ϵ-approximation of $B_H\left(\boldsymbol{S}^n_+\right)$.

Theorem 3 *There exists a constant $C > 0$ such that if S is an average ϵ-approximation of $B_H\left(\boldsymbol{S}_+^n\right)$, then*

$$\mathrm{xc}_{\boldsymbol{S}_+^k}(S) \geq \exp\left(C \cdot \min\left\{\left(\frac{n}{(1+\epsilon)^2}\right)^{1/3}, \ \frac{1}{1+\epsilon}\sqrt{\frac{n}{k}}\right\}\right)$$

for all positive integers n and k satisfying $1 \leq k \leq n$.

Theorem 3 is a stronger result than Theorem 2 because it provides an extension complexity lower bound for a broader range of sets that approximate $B_H\left(\boldsymbol{S}_+^n\right)$. Again, this result subsumes [11, Theorem 2] as a special case for $k = 1$. Specifically, Theorem 3 states that even if we relax the notion of approximation, we still need at least superpolynomially many number of $k \times k$ PSD constraints to approximate $\boldsymbol{S}_+^n$ when k is small, namely, when k is smaller than $\frac{1}{(1+\epsilon)^2}\frac{n}{\log^2 n}$. A more formal version of Theorem 3 and its proof can be found in Section 5.3.

Theorem 2 and Theorem 3 imply that any set that well approximates $B_H\left(\boldsymbol{S}_+^n\right)$ must have $\boldsymbol{S}_+^k$-extension complexity at least superpolynomially large in n for all k much smaller than n. Thus, we conclude that it is impossible to approximate $B_H\left(\boldsymbol{S}_+^n\right)$ using only polynomially many $k \times k$ PSD constraints, for *any* construction of the approximating set. Note that these are stronger hardness results than those discussed in Section 4, which only apply to the k-PSD approximations. Lastly, we mention that we do not know whether our lower bounds are tight. Thus, it might be possible to achieve even stronger lower bounds by means of a more sophisticated analysis.

5.2 Proof of Theorem 2

We state a full version of Theorem 2 as follows.

Theorem 4 *Let n, k be positive integers such that $1 \leq k \leq n$. If S is an ϵ-approximation of $B_H\left(\boldsymbol{S}_+^n\right)$, then*

$$\log \mathrm{xc}_{\boldsymbol{S}_+^k}(S) \geq -\frac{\alpha+\beta}{2} + \sqrt{\left(\frac{\alpha-\beta}{2}\right)^2 + \gamma}$$

where

$$\alpha = 2k\log 3, \qquad \beta = \log\left(\frac{n^3}{64\log 3}\right), \qquad \gamma = \frac{1}{16e}\frac{n-1}{(1+\epsilon)}.$$

Now we discuss how Theorem 2 can be derived from Theorem 4. Suppose that n is sufficiently large, tending to infinity.

- When $k = o_n(\sqrt{\frac{n}{1+\epsilon}})$, observe that $\gamma \gg \max\{\alpha^2, \beta^2\}$, and therefore, $-\frac{\alpha+\beta}{2} + \left\{(\frac{\alpha-\beta}{2})^2 + \gamma\right\}^{1/2} \approx \sqrt{\gamma}$.
- When $k = \omega_n(\sqrt{\frac{n}{1+\epsilon}})$, $\alpha \gg \max\{\beta, \sqrt{\gamma}\}$. Thus, $\left\{(\frac{\alpha-\beta}{2})^2 + \gamma\right\}^{1/2} \approx \frac{\alpha}{2}\left(1 + \frac{4\gamma}{\alpha^2}\right)^{1/2} \approx \frac{\alpha}{2}\left(1 + \frac{2\gamma}{\alpha^2}\right)$. As a result, $-\frac{\alpha+\beta}{2} + \left\{(\frac{\alpha-\beta}{2})^2 + \gamma\right\}^{1/2} \approx \frac{\gamma}{\alpha}$.

In the rest of this section, we prove Theorem 4. Our proof is based on similar arguments to those in the proof of [11, Theorem 1], but with appropriate adaptations. Indeed, our results can be seen as an extension of Fawzi's beyond the special case with $k = 1$, which is made possible by introducing different notions of normalization, (24), and decomposition of $\boldsymbol{S}_+^k$-factors into sharp and flat components, (25).

Proof (Proof of Theorem 4) We begin with a rough sketch of the main ideas used in the proof. First, we consider the generalized slack matrix s of the pair $\left(B_H\left(\boldsymbol{S}_+^n\right), \ (1+\epsilon)B_H\left(\boldsymbol{S}_+^n\right)\right)$ restricted to the hypercube H_n. In light of the generalized Yannakakis theorem (Lemma 4), the $\boldsymbol{S}_+^k$-extension complexity of S is bounded from below by the $\boldsymbol{S}_+^k$-rank of the slack matrix s, cf. (6). Thus, it suffices to prove a lower bound for $\mathrm{rank}_{\boldsymbol{S}_+^k}(s)$.

To this end, we express the slack matrix s in two equivalent ways: one obtained from the knowledge about the extreme points of $B_H\left(\boldsymbol{S}_+^n\right)$, and the other obtained by assuming that s admits a $\boldsymbol{S}_+^k$-factorization having N factors. Interpreting the extreme points of $B_H\left(\boldsymbol{S}_+^n\right)$ and $(1+\epsilon)B_H\left(\boldsymbol{S}_+^n\right)$ as formal variables, x and y, we may view the two expressions of the slack matrix as bivariate polynomials. Next, we 'smooth out' the two expressions with respect to one variable, x, by taking projection onto the harmonic subspace of degree 2 (i.e., the subspace of homogeneous multilinear polynomials of degree 2), cf. (7); and then take expectation with respect to the other variable, y, with conditioning on $x = y$[5].

[5] In the language of matrix operations, given a slack matrix $s(x, y)$ whose rows are indexed by x and columns are indexed by y, we left-multiply the degree-2 projection matrix to $s(x, y)$ and then take the (scaled) trace of the resulting matrix.

Comparing the two resulting expressions, we derive a lower bound on the number of factors N, which implies a lower bound on the $\boldsymbol{S}_+^k$-extension complexity of S.

Here we remark that we consider the degree-2 projection (with respect to x) of the slack operator s to control the variability of the $\boldsymbol{S}_+^k$-factors in an arbitrary $\boldsymbol{S}_+^k$-factorization, motivated by the observation that s can be expressed as a biquadratic (even) polynomial in x and y (quadratic in x and y, respectively); see (20). Thus, when viewed as a univariate polynomial in $x \in H_n$, there are only two nonzero harmonic components in the Fourier expansion of $s(x,y)$, namely, the harmonic components of degree 0 and 2. Thus, it suffices to examine the degree-2 harmonic component of s (with respect to x) for our purpose, because the degree-0 component (which is constant) does not contain much information about N, the number of factors required to express the variability of $s(x,y)$. Eventually, we attempt to control the 2-norm[6] of the degree-2 projection of the individual $\boldsymbol{S}_+^k$-factors in an arbitrary factorization of s. This may seem an elusive task at first glance because there is no a priori upper bound available for the norm of the arbitrary individual $\boldsymbol{S}_+^k$-factors. To overcome this challenge, we introduce a technical sharp/flat decomposition of the $\boldsymbol{S}_+^k$-factors, and establish an upper bound on the 2-norm of the truncated flat components of the $\boldsymbol{S}_+^k$-factors using hypercontractivity (Lemma 6); see Step 3 for details.

Step 1. Slack Matrix and $\boldsymbol{S}_+^k$-Factorization We consider the (generalized) slack operator associated to the pair $\left(B_H\left(\boldsymbol{S}_+^n\right),\ (1+\epsilon)B_H\left(\boldsymbol{S}_+^n\right)\right)$. Let $\mathbb{S}^{n-1} = \{x \in \mathbb{R}^n : \|x\|_2 = 1\}$. Observe that the extreme points of $B_H\left(\boldsymbol{S}_+^n\right)$ are $\tilde{x}\tilde{x}^T - \frac{1}{n}I_n$ for $\tilde{x} \in \mathbb{S}^{n-1}$, and that $\left((1+\epsilon)B_H\left(\boldsymbol{S}_+^n\right)\right)^\circ = -\frac{n}{1+\epsilon}B_H\left(\boldsymbol{S}_+^n\right)$. Thus, we are led to study the following infinite matrix:

$$(\tilde{x},\tilde{y}) \in \mathbb{S}^{n-1}\times\mathbb{S}^{n-1} \mapsto 1 - \left\langle \tilde{x}\tilde{x}^T - \frac{1}{n}I_n,\ -\frac{n}{1+\epsilon}\left(\tilde{y}\tilde{y}^T - \frac{1}{n}I_n\right)\right\rangle = \frac{n}{1+\epsilon}(\tilde{x}^T\tilde{y})^2 + \frac{\epsilon}{1+\epsilon}.$$

We consider the PSD rank ($\boldsymbol{S}_+^k$-rank) of the finite submatrix restricted to $\tilde{x},\tilde{y} \in \left\{-\frac{1}{\sqrt{n}},\frac{1}{\sqrt{n}}\right\}^n \subset \mathbb{S}^{n-1}$. Specifically, we consider the following matrix s defined on the n-dimensional hypercube, with a proper reparametrization ($x = \sqrt{n}\tilde{x}$ and $y = \sqrt{n}\tilde{y}$):

$$s : (x,y) \in \{-1,1\}^n \times \{-1,1\}^n \mapsto \frac{1}{1+\epsilon}\left(\frac{1}{n}(x^Ty)^2 + \epsilon\right). \tag{20}$$

Assuming that we can write the matrix (20) as a sum of N trace inner products of $\boldsymbol{S}_+^k$ factors, we have

$$\frac{1}{1+\epsilon}\left(\frac{1}{n}(x^Ty)^2 + \epsilon\right) = s(x,y) = \sum_{i=1}^N \langle A_i(x),\ B_i(y)\rangle, \qquad \forall x,y \in H_n \tag{21}$$

where $A_i, B_i : H_n \to \boldsymbol{S}_+^k$ are some matrix-valued functions on H_n.

In this proof, we use the two expressions of $s(x,y)$ in (21) to derive a lower bound on N. First, we fix $y \in H_n$ and 'smooth out' the expressions on both sides of (21) with respect to x by taking projection onto the space of harmonic polynomials of degree 2. Then we plug $x = y$ and consider the expectation of the smoothed functions with respect to $y \in H_n$.

More precisely, for each fixed $y \in H_n$, we let $q_y(x) = (x^Ty)^2 - n$. Also, let μ denote the uniform probability measure on H_n. The inner product of any two functions $f,g : H_n \to \mathbb{R}$ is defined as $\langle f,g\rangle_\mu = \mathbb{E}_{x\sim\mu}\big[f(x)g(x)\big]$. We observe that $\langle f(x), q_y(x)\rangle_\mu = 2\ \mathrm{proj}_2 f(y)$.

Taking the inner product of both sides of (21) with $q_y(x)$, we obtain

$$\mathbb{E}_{x\sim\mu}\left[\frac{1}{(1+\epsilon)n}q_y(x)^2 + q_y(x)\right] = \sum_{i=1}^N \left\langle \mathbb{E}_{x\sim\mu}\big[q_y(x)A_i(x)\big],\ B_i(y)\right\rangle.$$

Subsequently, we get the following equation by taking expectation over $y \sim \mu$:

$$\underbrace{\mathbb{E}_{y\sim\mu}\mathbb{E}_{x\sim\mu}\left[\frac{1}{(1+\epsilon)n}q_y(x)^2 + q_y(x)\right]}_{=:LHS} = \underbrace{\mathbb{E}_{y\sim\mu}\sum_{i=1}^N \left\langle \mathbb{E}_{x\sim\mu}\big[q_y(x)A_i(x)\big],\ B_i(y)\right\rangle}_{=:RHS}. \tag{22}$$

The rest of the proof is organized as follows. In Step 2, we compute the expectation on the left-hand side exactly. In Step 3, we derive an upper bound on the expectation on the right-hand side as a function of N. In the end, we obtain the desired lower bound on N in Step 4 by comparing these two quantities.

[6] See (8) for the definition of p-norm in Step 3 of the proof. Note that $\|\mathrm{proj}_2 f(x)\|_2^2$ for $f : H_n \to \mathbb{R}$ can be interpreted as the variance of the random variable $\mathrm{proj}_2 f(X)$ where $X \sim \mu(H_n)$.

Step 2. The Left-hand Side of (22). Observe that when $x \sim \mu$ and $y \in H_n$ is fixed, the random variable $x^T y$ is identically distributed as $\sum_{i=1}^n x_i$ for all $y \in H_n$ due to symmetry. Then it follows that[7]

$$\begin{aligned}\mathbb{E}_{x\sim\mu}\big[(x^Ty)^2\big] &= \mathbb{E}_{x\sim\mu}\left[\left(\sum_{i=1}^n x_i\right)^2\right] = \sum_{i,j=1}^n \mathbb{E}_{x\sim\mu}[x_ix_j] = \sum_{i=1}^n \mathbb{E}_{x\sim\mu}[x_i^2] \\ &= n, \\ \mathbb{E}_{x\sim\mu}\Big[(x^Ty)^4\Big] &= \mathbb{E}_{x\sim\mu}\left[\left(\sum_{i=1}^n x_i\right)^4\right] = \sum_{i=1}^n \mathbb{E}_{x\sim\mu}[x_i^4] + 3\sum_{\substack{i=1\\ j\neq i}}^n \mathbb{E}_{x\sim\mu}[x_i^2]\cdot\mathbb{E}_{x\sim\mu}[x_j^2] \\ &= n + 3n(n-1).\end{aligned}$$

Therefore, $\mathbb{E}_{x\sim\mu}[q_y(x)] = \mathbb{E}_{x\sim\mu}\big[(x^Ty)^2 - n\big] = 0$ and $\mathbb{E}_{x\sim\mu}[q_y(x)^2] = \mathbb{E}_{x\sim\mu}\big[(x^Ty)^4 - 2n(x^Ty)^2 + n^2\big] = 2n(n-1)$. It follows that for any $y \in H_n$,

$$\mathbb{E}_{x\sim\mu}\left[\frac{1}{(1+\epsilon)n}q_y(x)^2 + q_y(x)\right] = \frac{2}{1+\epsilon}(n-1). \tag{23}$$

This does not depend on y, and therefore, LHS in (22) $= \frac{2}{1+\epsilon}(n-1)$.

Step 3. An Upper Bound for the Right-hand Side of (22). Now, we prove an upper bound on the right-hand side of (22), which has the form of an increasing function of N. This step is the most technical part of the proof, and is composed of four mini-steps.

First of all, in Step 3-A, we claim that we may assume without loss of generality that the factor functions A_i, B_i satisfy

$$\big\|\mathbb{E}_{x\sim\mu}[A_i(x)]\big\|_{op} = 1, \quad \forall i \in [N] \qquad \text{and} \qquad \sum_{i=1}^N \mathrm{Tr}\big(B_i(y)\big) = 1, \quad \forall y \in H_n. \tag{24}$$

Next, in Step 3-B, we decompose each A_i into its sharp component $A_i^\sharp$ and flat component $A_i^\flat$ with a fixed threshold $\Lambda \geq e$ whose value will be determined later in Step 4 of the proof; see (25). Then due to linearity of expectation, we observe that RHS in (22) $= \big[\mathbb{E}_y \sum_{i=1}^N \langle \mathrm{proj}_2 A_i^\sharp(y), B_i(y)\rangle + \mathbb{E}_y \sum_{i=1}^N \langle \mathrm{proj}_2 A_i^\flat(y), B_i(y)\rangle\big]$. Lastly, we prove upper bounds for the two terms separately in Step 3-C and Step 3-D. The key idea is that for all $i \in [N]$, $A_i^\sharp$ is supported only on a set of small measure due to the normalization, (24), and $\|\mathrm{proj}_2(v^T A_i^\flat v)\|_2 \leq e\log\Lambda$ for all $v \in \mathbb{S}^{k-1}$ due to hypercontractivity (Lemma 6).

Step 3-A: Normalization of Factor Functions A_i, B_i. We claim that if $s(x,y)$ admits a $(\boldsymbol{S}_+^k)^N$-factorization, then we may assume (24) without loss of generality. More precisely, we show it is possible to normalize arbitrary factor functions $\{(\tilde{A}_i, \tilde{B}_i)\}_{i=1}^N$ to $\{(A_i, B_i)\}_{i=1}^N$ so that the conditions in (24) are satisfied.

Suppose that $s(x,y)$, defined in (20), admits a factorization $\{(\tilde{A}_i, \tilde{B}_i)\}_{i=1}^N$ such that $\tilde{A}_i, \tilde{B}_i : H_n \to \boldsymbol{S}_+^k$ and $s(x,y) = \sum_{i=1}^N \langle \tilde{A}_i(x), \tilde{B}_i(y)\rangle$ for all $x, y \in H_n$. For each $i \in [N]$, we can see that $\mathbb{E}_x[\tilde{A}_i(x)] \in \boldsymbol{S}_+^k$, and therefore, we may define $W_i = \mathbb{E}_x[\tilde{A}_i(x)]^{1/2}$ to be the principal square root of $\mathbb{E}_x[\tilde{A}_i(x)]$. Let $W_i^\dagger$ denote the Moore-Penrose pseudoinverse of W_i. Note that the column space of $\tilde{A}_i(x)$ is contained in the column space of W_i for all $x \in H_n$. Thus, $W_iW_i^\dagger \tilde{A}_i(x) W_i^\dagger W_i = \tilde{A}_i(x)$ for all $x \in H_n$.

For each $i \in [N]$, let $A_i(x) = W_i^\dagger \tilde{A}_i(x) W_i^\dagger$ and $B_i(y) = W_i \tilde{B}_i(y) W_i$. It is easy to verify that $\langle A_i(x), B_i(y)\rangle = \mathrm{Tr}\big(W_i^\dagger \tilde{A}_i(x) W_i^\dagger W_i \tilde{B}_i(y) W_i\big) = \mathrm{Tr}\big(W_iW_i^\dagger \tilde{A}_i(x) W_i^\dagger W_i \tilde{B}_i(y)\big) = \mathrm{Tr}\big(\tilde{A}_i(x)\tilde{B}_i(y)\big) = \big\langle \tilde{A}_i(x), \tilde{B}_i(y)\big\rangle$ for all $x, y \in H_n$. Thus, $\{(A_i, B_i)\}_{i=1}^N$ also constitutes a valid $\boldsymbol{S}_+^k$-factorization of $s(x,y)$.

Now it remains to check if $\{(A_i, B_i)\}_{i=1}^N$ satisfies (24). First, we can easily observe that

$$\mathbb{E}_{x\sim\mu}[A_i(x)] = W_i^\dagger \mathbb{E}_{x\sim\mu}[\tilde{A}_i(x)] W_i^\dagger = W_i^\dagger W_i^2 W_i^\dagger = \Pi_{\mathcal{R}(W_i)}$$

where $\mathcal{R}(W_i)$ is the range of W_i and $\Pi_{\mathcal{R}(W_i)}$ is the projection matrix onto $\mathcal{R}(W_i)$. Thus, $\big\|\mathbb{E}_{x\sim\mu}[A_i(x)]\big\|_{op} = \|\Pi_{\mathcal{R}(W_i)}\|_{op} = 1$. Next, we revisit (21), fix any $y \in H_n$, and take expectation with respect to $x \sim \mu$. On

[7] Notice that we are computing the second and the fourth moments of the sum of the entries of a random vector uniformly distributed over the n-dimensional hypercube. Thus, we already expect to obtain by the central limit theorem that $\lim_{n\to\infty} \frac{1}{n}\mathbb{E}_{x\sim\mu}\big[(x^Ty)^2\big] = 1$ and $\lim_{n\to\infty} \frac{1}{n^2}\mathbb{E}_{x\sim\mu}\big[(x^Ty)^4\big] = 3$, which correspond to the second and the fourth moments of the standard Gaussian distribution.

the left-hand side, we obtain $\mathbb{E}_{x\sim\mu}\left[\frac{1}{1+\epsilon}\left(\frac{1}{n}(x^Ty)^2+\epsilon\right)\right]=1$ because $\mathbb{E}_{x\sim\mu}\left[(x^Ty)^2\right]=n$ (see Step 2). On the right-hand side, we have

$$\mathbb{E}_{x\sim\mu}\sum_{i=1}^{N}\langle A_i(x),\ B_i(y)\rangle=\sum_{i=1}^{N}\langle\mathbb{E}_{x\sim\mu}[A_i(x)],\ B_i(y)\rangle=\sum_{i=1}^{N}\left\langle \Pi_{\mathcal{R}(W_i)},\ B_i(y)\right\rangle=\sum_{i=1}^{N}\operatorname{Tr}\ B_i(y)$$

because $\Pi_{\mathcal{R}(W_i)}B_i(y)=B_i(y)$ for all $y\in H_n$, by definition of B_i. Therefore, $\sum_{i=1}^{N}\operatorname{Tr}\ B_i(y)=1$ for all $y\in H_n$.

Step 3-B: Decomposition of A_i. We decompose each A_i into its 'sharp' (spiky) component $A_i^\sharp$ and the 'flat' component $A_i^\flat$ using a fixed threshold Λ whose value will be determined later in Step 4 of the proof. To be specific, for each $i\in[N]$, we define the component functions $A_i^\sharp, A_i^\flat: H_n\to\boldsymbol{S}_+^k$ as follows. Given $x\in H_n$, let $A_i(x)=\sum_{a=1}^{k}\lambda_a u_a u_a^T$ be the eigendecomposition of $A_i(x)$. Then we let

$$A_i^\sharp(x)=\sum_{a=1}^{k}\lambda_a\mathbf{1}_{\{\lambda_a>\Lambda\}}u_au_a^T,\qquad\text{and}\qquad A_i^\flat(x)=\sum_{a=1}^{k}\lambda_a\mathbf{1}_{\{\lambda_a\le\Lambda\}}u_au_a^T. \tag{25}$$

Observe that $A_i=A_i^\sharp+A_i^\flat$ and $\left\langle A_i^\sharp(x),\ A_i^\flat(x)\right\rangle=\operatorname{Tr}\big(A_i^\sharp(x)A_i^\flat(x)\big)=0$ for all $x\in H_n$. From now on, we may refer to $A_i^\sharp$ ($A_i^\flat$, resp.) as the sharp component (flat component, resp.) of A_i.

By linearity of expectation, we can decompose the expression on the right-hand side of (22) as follows:

$$RHS\text{ in (22)}=\mathbb{E}_{y\sim\mu}\sum_{i=1}^{N}\left\langle\mathbb{E}_{x\sim\mu}[q_y(x)A_i^\sharp(x)],\ B_i(y)\right\rangle+\mathbb{E}_{y\sim\mu}\sum_{i=1}^{N}\left\langle\mathbb{E}_{x\sim\mu}[q_y(x)A_i^\flat(x)],\ B_i(y)\right\rangle. \tag{26}$$

Step 3-C. Upper Bound on the Contribution of Sharp Components in (26). In this paragraph, we argue that the first term on the right hand side of (26) is bounded from above by $N\frac{k}{\Lambda}n^3$. Our argument is based on the following three observations.

- Let $\operatorname{supp}\ A_i^\sharp=\{x\in H_n:A_i^\sharp(x)\neq 0\}$. Then $|\operatorname{supp}\ A_i^\sharp|<\frac{k}{\Lambda}\,2^n$ for all $i\in[N]$. It is because (i) $\operatorname{Tr}\ \mathbb{E}_{x\sim\mu}[A_i^\sharp(x)]\le\operatorname{Tr}\ \mathbb{E}_{x\sim\mu}[A_i(x)]\le k\left\|\mathbb{E}_{x\sim\mu}[A_i(x)]\right\|_{op}=k$, cf. (24); (ii) $\operatorname{Tr}\ \mathbb{E}_{x\sim\mu}[A_i^\sharp(x)]=\frac{1}{2^n}\sum_{x\in H_n}\operatorname{Tr}\ A_i^\sharp(x)$; and (iii) $\operatorname{Tr}\ A_i^\sharp(x)>\Lambda$ for all $x\in\operatorname{supp}\ A_i^\sharp$ by definition of $A_i^\sharp$.
- For each $i\in[N]$, $\left\langle A_i^\sharp(x),\ B_i(y)\right\rangle\le n$ for all $x,y\in H_n$. This follows from Eq. (21) because

$$\left\langle A_i^\sharp(x),B_i(y)\right\rangle\le\sum_{i=1}^{N}\langle A_i(x),\ B_i(y)\rangle=\frac{1}{1+\epsilon}\left(\frac{1}{n}(x^Ty)^2+\epsilon\right)\le\frac{1}{1+\epsilon}\big(n+\epsilon\big)\le n.$$

- Lastly, $|q_y(x)|\le n(n-1)$ for all $x,y\in H_n$ because $q_y(x)=(x^Ty)^2-n\le n(n-1)$ and $q_y(x)\ge -n$.

Combining the three observations above, we can see that for any $y\in H_n$,

$$\begin{aligned}\sum_{i=1}^{N}\left\langle\mathbb{E}_{x\sim\mu}[q_y(x)A_i^\sharp(x)],\ B_i(y)\right\rangle&=\sum_{i=1}^{N}\mathbb{E}_{x\sim\mu}\left[q_y(x)\left\langle A_i^\sharp(x),\ B_i(y)\right\rangle\right]\\&=\sum_{i=1}^{N}\sum_{x\in\operatorname{supp}\ A_i^\sharp}\frac{1}{2^n}q_y(x)\left\langle A_i^\sharp(x),\ B_i(y)\right\rangle\\&\le\sum_{i=1}^{N}\frac{|\operatorname{supp}\ A_i^\sharp|}{2^n}\Big(\max_{x,y\in H_n}|q_y(x)|\Big)\Big(\max_{x,y\in H_n}\left\langle A_i^\sharp(x),\ B_i(y)\right\rangle\Big)\\&\le\frac{k}{\Lambda}n^2(n-1)N.\end{aligned}$$

Taking expectation with respect to $y\sim\mu$, we obtain

$$\mathbb{E}_{y\sim\mu}\sum_{i=1}^{N}\left\langle\mathbb{E}_{x\sim\mu}[q_y(x)A_i^\sharp(x)],\ B_i(y)\right\rangle\le\frac{k}{\Lambda}n^2(n-1)N\le\frac{k}{\Lambda}n^3N. \tag{27}$$

Step 3-D. Upper Bound on the Contribution of Flat Components in (26). Next, we prove an upper bound for the second term on the right hand side of (26). Our proof is based on the concentration of the degree-2 harmonic components of bounded functions and the usual ϵ-net argument.

First, we reduce the matrix-valued function $A_i^\flat$'s to the supremum of multiple scalar-valued functions indexed over a finite set. Given $\epsilon_{\text{net}} > 0$, let $\mathcal{N}$ be an ϵ_{net}-net of $\mathbb{S}^{k-1}$ with the smallest possible cardinality. Note that $|\mathcal{N}| \le \left(1 + \frac{2}{\epsilon_{\text{net}}}\right)^k$ by the well-known upper bound on the ϵ_{net}-covering number of $\mathbb{S}^{k-1}$, e.g. [21, Corollary 4.2.13]. Then

$$\begin{aligned}
\mathbb{E}_{y\sim\mu}\left[\sum_{i=1}^N \left\langle \mathbb{E}_{x\sim\mu}[q_y(x)A_i^\flat(x)],\ B_i(y)\right\rangle\right] &\overset{(a)}{\le} \mathbb{E}_{y\sim\mu}\left[\sum_{i=1}^N \left\|\mathbb{E}_{x\sim\mu}[q_y(x)A_i^\flat(x)]\right\|_{op} \operatorname{Tr}\ B_i(y)\right] \\
&\overset{(b)}{\le} \mathbb{E}_{y\sim\mu}\left[\max_{i\in[N]} \left\|\mathbb{E}_{x\sim\mu}[q_y(x)A_i^\flat(x)]\right\|_{op}\right] \\
&\overset{(c)}{\le} \frac{1}{1-2\epsilon_{\text{net}}}\mathbb{E}_{y\sim\mu}\left[\max_{\substack{i\in[N]\\ v\in\mathcal{N}}} \left|v^T\mathbb{E}_{x\sim\mu}\left[q_y(x)A_i^\flat(x)\right]v\right|\right] \\
&= \frac{1}{1-2\epsilon_{\text{net}}}\mathbb{E}_{y\sim\mu}\left[\max_{\substack{i\in[N]\\ v\in\mathcal{N}}} \left|\mathbb{E}_{x\sim\mu}\left[q_y(x)\ v^TA_i^\flat(x)v\right]\right|\right].
\end{aligned}$$

In the above lines, (a) follows from Hölder's inequality for Schatten norms; (b) is due to the normalization $\sum_{i=1}^N \operatorname{Tr}\ B_i(y) \equiv 1$; and (c) is obtained by the ϵ-net argument[8], i.e., if $\mathcal{N}$ is an ϵ_{net}-net of $\mathbb{S}^{k-1}$, then for any $M \in \boldsymbol{S}_+^k$, $\|M\|_{op} \le \frac{1}{1-2\epsilon_{\text{net}}} \sup_{v\in\mathcal{N}} v^TMv$. It remains to evaluate the expectation in the last line.

Recall that $\mathbb{E}_{x\sim\mu}\left[q_y(x)v^TA_i^\flat(x)v\right] = \left\langle q_y(x),\ v^TA_i^\flat(x)v\right\rangle_\mu = 2\text{proj}_2\left(v^TA_i^\flat v\right)(y)$. We observe that for each $(i,v) \in [N]\times\mathcal{N}$, the derived random variable $\text{proj}_2\left(v^TA_i^\flat v\right)(y)$ is sub-exponential with parameters $(16\|\text{proj}_2(v^TA_i^\flat v)\|_2^2,\ 4\|\text{proj}_2(v^TA_i^\flat v)\|_2)$, due to Lemma 12.

Next, we find a common upper bound on $\|\text{proj}_2(v^TA_i^\flat v)\|_2$ that holds for all (i,v). Note that for all (i,v), $\mathbb{E}_{y\in\mu}[v^TA_i^\flat(y)v] \le \mathbb{E}_{y\in\mu}\|A_i^\flat(y)\|_{op} \le \mathbb{E}_{y\in\mu}\|A_i(y)\|_{op} = 1$ due to the normalization in (24), and $0 \le v^TA_i^\flat v \le \Lambda$ by definition of $A_i^\flat$. Thus, we can apply Lemma 6 to get $\|\text{proj}_2(v^TA_i^\flat v)\|_2 \le e\log\Lambda$ for all (i,v), provided that we will choose the threshold $\Lambda \ge e$.

Now we can use a result on the expected maximum of $N|\mathcal{N}|$ sub-exponential random variables (Lemma 13) to obtain

$$\begin{aligned}
\mathbb{E}_{y\sim\mu}\left[\max_{\substack{i\in[N]\\ v\in\mathcal{N}}} \left|\mathbb{E}_{x\sim\mu}\left[q_y(x)\ v^TA_i^\flat(x)v\right]\right|\right] &= 2\cdot\mathbb{E}_{y\sim\mu}\left[\max_{\substack{i\in[N]\\ v\in\mathcal{N}}} \left|\text{proj}_2\left(v^TA_i^\flat v\right)(y)\right|\right] \\
&\le 2\cdot\|\text{proj}_2(v^TA_i^\flat v)\|_2 \cdot \max\left\{4\sqrt{2\log(N|\mathcal{N}|)},\ 8\log(N|\mathcal{N}|)\right\} \\
&\le 8\sqrt{2}e\log\Lambda\cdot\max\left\{\sqrt{\log(N|\mathcal{N}|)},\ \sqrt{2}\log(N|\mathcal{N}|)\right\}.
\end{aligned}$$

Collecting the pieces in this step, we obtain the following upper bound:

$$\begin{aligned}
&\mathbb{E}_{y\sim\mu}\sum_{i=1}^N \left\langle \mathbb{E}_{x\in H_n}[q_y(x)A_i^\flat(x)],\ g_i(y)\right\rangle \\
&\le \frac{8\sqrt{2}e\log\Lambda}{1-2\epsilon_{\text{net}}}\max\left\{\sqrt{\log\left[N\left(1+\frac{2}{\epsilon_{\text{net}}}\right)^k\right]},\ \sqrt{2}\log\left[N\left(1+\frac{2}{\epsilon_{\text{net}}}\right)^k\right]\right\}.
\end{aligned} \tag{28}$$

Step 4. Concluding the Proof Lastly, we revisit Eq. (22) to conclude the proof. Recall that we obtained the value of the left-hand side in Step 2, cf. (23), and derived an upper bound for the right-hand side in Step 3, cf. (26), (27), and (28). Putting these together, we have the following inequality that holds for any choice of parameters ϵ_{net}, Λ such that $0 < \epsilon_{\text{net}} < \frac{1}{2}$ and $\Lambda \ge e$:

$$\frac{2}{1+\epsilon}(n-1) \le \frac{k}{\Lambda}n^3N + \frac{8\sqrt{2}e\log\Lambda}{1-2\epsilon_{\text{net}}}\max\left\{\sqrt{\log\left[N\left(1+\frac{2}{\epsilon_{\text{net}}}\right)^k\right]},\ \sqrt{2}\log\left[N\left(1+\frac{2}{\epsilon_{\text{net}}}\right)^k\right]\right\}. \tag{29}$$

We choose $\epsilon_{\text{net}} = 1/4$ for simplicity because optimizing ϵ_{net} does not make much difference. Observe that $\log(9^kN) \ge 1$ for all $k, N \ge 1$. Thus, $\sqrt{\log(9^kN)} \le \sqrt{2}\log(9^kN)$ for all $k, N \ge 1$.

[8] See [21, Lemma 4.4.1] for example.

Then, we select Λ that minimizes the right-hand side of (29). It is easy to see that the upper bound is minimized (w.r.t. Λ) at $\Lambda^* = \frac{kn^3N}{32e\log(9^kN)}$. Noticing that $\Lambda^* \le \frac{kn^3N}{32e\log(9^k)}$ (because $N \ge 1$), we get the following quadratic inequality in $\log N$ as a necessary condition for (29):

$$\begin{aligned}\frac{2}{1+\epsilon}(n-1) &\le 32e\cdot\log(9^kN)\big(1+\log\Lambda^*\big)\\ &\le 32e\cdot\big[\log N + 2k\log 3\big]\left[\log N + \log\left(\frac{n^3}{64\log 3}\right)\right]. \end{aligned} \tag{30}$$

Letting $z = \log N \ge 0$, we note that (30) is a quadratic inequality of the form $(z+\alpha)(z+\beta)\ge\gamma$ where

$$\alpha = 2k\log 3, \qquad \beta = \log\left(\frac{n^3}{64\log 3}\right), \qquad \gamma = \frac{1}{16e}\frac{n-1}{(1+\epsilon)}.$$

We want to solve this quadratic inequality with an implicit constraint $z\ge 0$ because $N\ge 1$. Observe that its discriminant $D = (\alpha-\beta)^2+4\gamma > 0$, regardless of n,k,ϵ. Therefore, the set of solutions is given as $\big\{z\in\mathbb{R}:\ (z+\alpha)(z+\beta)\ge\gamma,\ z\ge 0\big\} = \big\{z\in\mathbb{R}:\ z\ge\big[\frac{-(\alpha+\beta)+\sqrt{D}}{2}\big]_+\big\}$ where $[x]_+ = \max\{x,0\}$. □

5.3 Proof of Theorem 3

The following is a formal version of Theorem 3, which will be proved later in this section.

Theorem 5 *Let n,k be positive integers such that $1\le k\le n$. If S is an average ϵ-approximation of $B_H\left(\boldsymbol{S}^n_+\right)$, then*

$$\log \mathrm{xc}_{\boldsymbol{S}^k_+}(S) \ge \left\{\left(\alpha+\sqrt{\alpha^2+\beta^3}\right)^{1/3} + \left(\alpha-\sqrt{\alpha^2+\beta^3}\right)^{1/3}\right\}^2 - 2k\log 3$$

where

$$\alpha = \frac{\sqrt{n}}{22000e(1+\epsilon)} \qquad \textit{and} \qquad \beta = \frac{1}{3}\left\{\log\left(\frac{16(1+\epsilon)k^{1/2}n^{3/2}}{5\sqrt{2\log 3}}\right) - 2k\log 3\right\}.$$

Now we discuss how Theorem 3 can be derived from Theorem 5. For notational brevity in our derivation, we let $T_+ = \left(\alpha+\sqrt{\alpha^2+\beta^3}\right)^{1/3}$ and $T_- = \left(\alpha-\sqrt{\alpha^2+\beta^3}\right)^{1/3}$. Suppose that n is sufficiently large, tending to infinity.

- When $k = o_n\big(\big(\frac{n}{(1+\epsilon)^2}\big)^{1/3}\big)$, we can see that $\alpha^2 \gg |\beta|^3$ and thus, $\sqrt{\alpha^2+\beta^3}\approx\alpha$. Therefore, $T_+ + T_- \approx (2\alpha)^{1/3}$, and in the end, $(T_++T_-)^2 - 2k\log 3 \approx (2\alpha)^{2/3} \approx C\cdot\frac{n^{1/3}}{(1+\epsilon)^{2/3}}$ for some constant C.
- When $k = \omega_n\big(\big(\frac{n}{(1+\epsilon)^2}\big)^{1/3}\big)$, note that $\beta<0$ and $\alpha^2 \ll |\beta|^3$. Let $\gamma := \sqrt{\alpha^2+\beta^3}$. We observe that $\gamma\approx|\beta|^{3/2}i$, and thus, $|\gamma|\approx|\beta|^{3/2}\gg\alpha$. Then, we can see that $T_+ = (\alpha+\gamma)^{1/3}\approx\gamma^{1/3}\big(1+\frac{\alpha}{3\gamma}\big)$, and likewise, $T_-\approx\bar\gamma^{1/3}\big(1+\frac{\alpha}{3\bar\gamma}\big)$ where $\bar\gamma$ is the complex conjugate of γ. Then it follows that

$$\begin{aligned}T_+ + T_- &\approx \gamma^{1/3}+\bar\gamma^{1/3} + \frac{\alpha}{3}\big(\frac{1}{\gamma^{2/3}}+\frac{1}{\bar\gamma^{2/3}}\big) \approx |\gamma|^{1/3}\big(e^{i\frac{\pi}{6}}+e^{-i\frac{\pi}{6}}\big) + \frac{\alpha}{3|\gamma|^{2/3}}\big(e^{-i\frac{\pi}{3}}+e^{i\frac{\pi}{3}}\big)\\ &= \sqrt{3}|\gamma|^{1/3}\left(1+\frac{\alpha}{3\sqrt{3}|\gamma|}\right).\end{aligned}$$

  Therefore, $(T_++T_-)^2 \approx 3|\gamma|^{2/3}\big(1+\frac{2\alpha}{3\sqrt{3}|\gamma|}\big) = 3|\gamma|^{2/3} + \frac{2}{\sqrt{3}}\frac{\alpha}{|\gamma|^{1/3}}$. Lastly, noticing that $|\gamma|^{2/3}\approx|\beta|\approx\frac{2}{3}k\log 3$, we can conclude that $(T_++T_-)^2 - 2k\log 3 \approx C\cdot\frac{1}{1+\epsilon}\sqrt{\frac{n}{k}}$ for some constant C.

Proof (Proof of Theorem 5) We follow a similar strategy to that of Theorem 4 with some modifications. Here, we assume S is ϵ-approximation of $B_H\left(\boldsymbol{S}^n_+\right)$ only in the average sense, and thus, S can be arbitrarily shaped and $S\subseteq(1+\epsilon)B_H\left(\boldsymbol{S}^n_+\right)$ is not necessarily true. Instead, we define a set Q – to be precise, we let $Q = 10(1+\epsilon)\sqrt{n}\mathbb{G}^\circ_S$ for $\mathbb{G}_S$ to be defined in (31) – that contains S in an adaptive manner. Then we consider the generalized slack matrix of the pair $\left(B_H\left(\boldsymbol{S}^n_+\right),\ Q\right)$. We express the slack matrix in two equivalent ways: one is obtained from the knowledge about the extreme points of $B_H\left(\boldsymbol{S}^n_+\right)$, and the other is obtained by assuming the existence of a $\boldsymbol{S}^k_+$-factorization having N factors. Interpreting the extreme points of $B_H\left(\boldsymbol{S}^n_+\right)$ and Q° as formal variables, x and G, we may view the two expressions of the slack matrix as bivariate polynomials. As already done in the proof of Theorem 4, we 'smooth out' the two expressions with respect to one variable, x; and then take expectation with respect to the other variable, G. Comparing the two resulting expressions, we derive a lower bound on the number of factors N, which implies a lower bound on the $\boldsymbol{S}^k_+$-extension complexity of S.

Step 1. Gaussian Surrogate for S° and the Associated Slack Matrix Let $\boldsymbol{S}_0^n$ denote the set of $n \times n$ symmetric matrices with trace zero, endowed with the trace inner product. Let $\mathcal{N}_0$ denote the standard Gaussian distribution associated to $\boldsymbol{S}_0^n$, i.e., $G_0 \sim \mathcal{N}_0$ if $G_0 = G - \frac{\mathrm{Tr}G}{n} I_n$ where G has the standard Gaussian distribution in $\boldsymbol{S}^n$. Then we define a set

$$\mathbb{G}_S = \left\{ G \in \boldsymbol{S}_0^n : \left| \langle G,\ X \rangle \right| \le 5\sqrt{2} w_G(S),\ \forall X \in S \right\}. \tag{31}$$

The number $5\sqrt{2}$ is chosen for the convenience of our analysis, and has no special meaning. Observe that $w_G(S) \le (1+\epsilon) \cdot w_G\big(B_H\left(\boldsymbol{S}_+^n\right)\big) \le (1+\epsilon)\sqrt{2n}$, cf. Remark 5.

Then, we can see that

$$-10(1+\epsilon)\sqrt{n} \le \langle G,\ X \rangle \le 10(1+\epsilon)\sqrt{n}, \qquad \forall (X, G) \in S \times \mathbb{G}_S.$$

This implies that $\frac{1}{10(1+\epsilon)\sqrt{n}} \mathbb{G}_S \subseteq S^\circ$, or equivalently, $S \subseteq 10(1+\epsilon)\sqrt{n}\mathbb{G}_S^\circ$.

Now we consider the slack operator associated to the pair $\left(B_H\left(\boldsymbol{S}_+^n\right),\ 10(1+\epsilon)\sqrt{n}\mathbb{G}_S^\circ\right)$, treating $\frac{1}{10(1+\epsilon)\sqrt{n}}\mathbb{G}_S$ as a surrogate for S°. Specifically, we are led to study the following infinite matrix:

$$(\tilde{x}, G) \in \mathbb{S}^{n-1} \times \mathbb{G}_S \mapsto 1 - \left\langle \tilde{x}\tilde{x}^T - \frac{1}{n} I_n,\ \frac{1}{10(1+\epsilon)\sqrt{n}} G \right\rangle = 1 - \frac{1}{10(1+\epsilon)\sqrt{n}} \tilde{x}^T G \tilde{x}.$$

We consider the PSD rank ($\boldsymbol{S}_+^k$-rank) of the submatrix restricted to $\tilde{x} \in \left\{ -\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}} \right\}^n \subset \mathbb{S}^{n-1}$, with a proper reparametrization ($x = \sqrt{n}\tilde{x}$), namely,

$$s : (x, G) \in H_n \times \mathbb{G}_S \mapsto 1 - \frac{1}{10(1+\epsilon)n\sqrt{n}} x^T G x. \tag{32}$$

Assuming that we can write the matrix (32) as a sum of N trace inner products of $\boldsymbol{S}_+^k$ factors, we have

$$1 - \frac{1}{10(1+\epsilon)n\sqrt{n}} x^T G x = s(x, G) = \sum_{i=1}^N \langle A_i(x),\ B_i(G) \rangle, \qquad \forall (x, G) \in H_n \times \mathbb{G}_S \tag{33}$$

where $A_i : H_n \to \boldsymbol{S}_+^k$ and $B_i : \mathbb{G}_S \to \boldsymbol{S}_+^k$ are some matrix-valued functions.

Again, we 'smooth out' the two expressions of $s(x, G)$ in (33) and compare them to derive a lower bound for N. To be precise, for each fixed $G \in \mathbb{G}_S$, we let $q_G(x) = -x^T G x$. Recall that we let μ denote the uniform probability measure on H_n, and observe that for any function $f : H_n \to \mathbb{R}$, the inner product, $\langle f, q_G(x) \rangle_\mu = \mathbb{E}_{x \sim \mu}[f(x) q_G(x)]$ is a centered Gaussian random variable.

Taking the inner product of both sides of (33) with $q_G(x)$, we get

$$\mathbb{E}_{x \sim \mu}\left[q_G(x) + \frac{1}{10(1+\epsilon)n\sqrt{n}} q_G(x)^2 \right] = \sum_{i=1}^N \langle \mathbb{E}_{x \sim \mu}[q_G(x) \cdot A_i(x)],\ B_i(G) \rangle.$$

Letting $\mathbb{E}_{G \sim \mathcal{N}_0 | \mathbb{G}_S}[\ \cdot\ ]$ denote the conditional expectation with respect to $G \sim \mathcal{N}_0$ given $G \in \mathbb{G}_S$, we can see that

$$\underbrace{\frac{1}{10(1+\epsilon)n\sqrt{n}} \cdot \mathbb{E}_{G \sim \mathcal{N}_0 | \mathbb{G}_S} \mathbb{E}_{x \sim \mu}\left[q_G(x)^2 \right]}_{=:LHS} = \underbrace{\mathbb{E}_{G \sim \mathcal{N}_0 | \mathbb{G}_S} \sum_{i=1}^N \langle \mathbb{E}_{x \sim \mu}[q_G(x) \cdot A_i(x)],\ B_i(G) \rangle}_{=:RHS}. \tag{34}$$

The rest of the proof is organized as follows. In Step 2, we prove a lower bound for the expectation on the left-hand side. In Step 3, we derive an upper bound on the expectation on the right-hand side as a function of N. In the end, we obtain the desired lower bound on N in Step 4 by comparing these bounds.

Step 2. A Lower Bound for the Left-hand side of (34). We additionally define a set

$$\mathbb{G}_{1/2} = \left\{ G \in \boldsymbol{S}_0^n : \mathbb{E}_{x\sim\mu}\left[q_G(x)^2\right] \geq \frac{1}{5}n(n-1) \right\}.$$

The constant $1/5$ is chosen for the convenience of analysis, and has no special meaning. By the law of total probability, we can see that

$$\mathbb{E}_{G\sim\mathcal{N}_0|\mathbb{G}_S}\mathbb{E}_{x\sim\mu}\left[q_G(x)^2\right] \geq \mathbb{E}_{G\sim\mathcal{N}_0|\mathbb{G}_S\cap\mathbb{G}_{1/2}}\mathbb{E}_{x\sim\mu}\left[q_G(x)^2\right] \cdot \Pr[G \in \mathbb{G}_S \cap \mathbb{G}_{1/2} \mid G \in \mathbb{G}_S].$$

Note that $\mathbb{E}_{G\sim\mathcal{N}_0|\mathbb{G}_S\cap\mathbb{G}_{1/2}}\mathbb{E}_{x\sim\mu}\left[q_G(x)^2\right] \geq \frac{1}{5}n(n-1)$ by definition of $\mathbb{G}_{1/2}$. Thus, it suffices to find a lower bound for the conditional probability, $\Pr[G \in \mathbb{G}_S \cap \mathbb{G}_{1/2} \mid G \in \mathbb{G}_S]$.

It is easy to see that

$$\Pr[G \in \mathbb{G}_S \cap \mathbb{G}_{1/2} \mid G \in \mathbb{G}_S] = \frac{\Pr[G \in \mathbb{G}_S \cap \mathbb{G}_{1/2}]}{\Pr[G \in \mathbb{G}_S]} \geq \frac{\Pr[G \in \mathbb{G}_S] - \Pr[G \notin \mathbb{G}_{1/2}]}{\Pr[G \in \mathbb{G}_S]}.$$

Observe that $\Pr[G \in \mathbb{G}_S] \geq 1 - \exp\left(-\frac{(5\sqrt{2}-1)^2}{4\pi}\right) > 0.893$ by Lemma 10. Now it remains to show an upper bound for $\Pr[G \notin \mathbb{G}_{1/2}]$.

We use standard concentration results for the chi-square distribution. Note that if $G \sim \mathcal{N}_0$, then $q_G(x) = -\mathrm{Tr}\, G - 2\sum_{i<j} G_{ij}x_ix_j = -2\sum_{i<j} G_{ij}x_ix_j$, and therefore, $\mathbb{E}_{x\sim\mu}\left[q_G(x)^2\right] = 4\sum_{i<j} G_{ij}^2$. Thus, we have $\mathbb{E}_{G\sim\mathcal{N}_0}\mathbb{E}_{x\sim\mu}[q_G(x)^2] = 4\binom{n}{2}\frac{1}{2} = n(n-1)$. Using an exponential inequality for chi-square distribution (e.g., [15, Lemma 1]), we obtain $\Pr[G \notin \mathbb{G}_{1/2}] \leq \exp\left(-\frac{2}{25}n(n-1)\right) \leq 0.8522$ for all $n \geq 1$.

All in all, we obtain

$$LHS \text{ in } (34) \geq \frac{1}{10(1+\epsilon)n\sqrt{n}} \cdot \frac{1}{5}n(n-1) \cdot \frac{\Pr[G \in \mathbb{G}_S] - \Pr[G \notin \mathbb{G}_{1/2}]}{\Pr[G \in \mathbb{G}_S]} \geq \frac{\sqrt{n}}{2200(1+\epsilon)} \tag{35}$$

because $\frac{0.893-0.8522}{0.893} \geq 1/22$ and $n - 1 \geq n/2$ for all $n \geq 1$.

Step 3. An Upper Bound for the Right-hand side of (34). Next, we prove an upper bound on the right-hand side of (34), which is a function of N. Note that for the same reason as discussed in Step 3-A of the proof of Theorem 4, we may assume without loss of generality that the factor functions A_i, B_i satisfy

$$\left\|\mathbb{E}_{x\sim\mu}[A_i(x)]\right\|_{op} = 1, \quad \forall i \in [N] \qquad \text{and} \qquad \sum_{i=1}^{n} \mathrm{Tr}\big(B_i(G)\big) = 1, \quad \forall G \in \mathbb{G}_S. \tag{36}$$

For each $i \in [N]$, we define the component functions $A_i^\sharp, A_i^\flat : H_n \to \boldsymbol{S}_+^k$ in the same way as in (25), using a fixed threshold Λ whose value will be determined later in this proof, cf. Step 3-B of the proof of Theorem 4.

By linearity of expectation, we can decompose the expression on the right-hand side of (34) as

$$RHS \text{ in } (34) = \mathbb{E}_{G\sim\mathcal{N}_0|\mathbb{G}_S} \sum_{i=1}^{N} \Big\langle \mathbb{E}_{x\sim\mu}[q_G(x) \cdot A_i^\sharp(x)],\ B_i(G) \Big\rangle + \mathbb{E}_{G\sim\mathcal{N}_0|\mathbb{G}_S} \sum_{i=1}^{N} \Big\langle \mathbb{E}_{x\sim\mu}[q_G(x) \cdot A_i^\flat(x)],\ B_i(G) \Big\rangle. \tag{37}$$

In the two sub-steps below, we prove upper bounds for the two terms on the right hand side separately.

Step 3-A. Upper Bound on the Contribution of Sharp Components in (37). Here we argue that the first term on the right hand side of (37) is bounded from above by $\frac{16(1+\epsilon)}{\Lambda}kn\sqrt{n}N$. Our argument is based on the following three observations.

- Let $\mathrm{supp}\, A_i^\sharp = \{x \in H_n : A_i^\sharp(x) \neq 0\}$. Then $|\mathrm{supp}\, A_i^\sharp| < \frac{k}{\Lambda}\, 2^n$ for all $i \in [N]$, cf. Step 3-C of the proof of Theorem 4.
- Observe that $\langle A_i^\sharp(x),\ B_i(G)\rangle \leq \langle A_i(x),\ B_i(G)\rangle \leq s(x,G) \leq 2$ for all $i \in [N]$ and for all $(x,G) \in H_n \times \mathbb{G}_S$.
- $q_G(x) = -x^TGx = 8(1+\epsilon)n\sqrt{n}\big(s(x,G) - 1\big) \leq 8(1+\epsilon)n\sqrt{n}$ for all $(x,G) \in H_n \times \mathbb{G}_S$.

Combining these observations, we can see that for every $G \in \mathbb{G}_S$,

$$\sum_{i=1}^N \Big\langle \mathbb{E}_{x\sim\mu}[q_G(x)A_i^\sharp(x)],\ B_i(G)\Big\rangle = \sum_{i=1}^N \mathbb{E}_{x\sim\mu}\Big[q_G(x)\Big\langle A_i^\sharp(x),\ B_i(G)\Big\rangle\Big] \le \sum_{i=1}^N \frac{|\mathrm{supp}\ A_i^\sharp|}{2^n}\cdot 16(1+\epsilon)n\sqrt{n}$$
$$\le \frac{16(1+\epsilon)}{\Lambda} kn\sqrt{n}N.$$

This upper bound is independent of G, and thus, we get

$$\mathbb{E}_{G\sim\mathcal{N}_0|\mathbb{G}_S}\sum_{i=1}^N \Big\langle \mathbb{E}_{x\sim\mu}[q_G(x)\cdot A_i^\sharp(x)],\ B_i(G)\Big\rangle \le \frac{16(1+\epsilon)}{\Lambda} kn\sqrt{n}N. \tag{38}$$

Step 3-B. Upper Bound on the Contribution of Flat Components in (37). Here we prove an upper bound for the second term in (37). First of all, we observe that for every $G \in \mathbb{G}_S$,

$$\sum_{i=1}^N \Big\langle \mathbb{E}_{x\sim\mu}[q_G(x)\cdot A_i^\flat(x)],\ B_i(G)\Big\rangle \le \sum_{i=1}^N \Big\|\mathbb{E}_{x\sim\mu}[q_G(x)\cdot A_i^\flat(x)]\Big\|_{op} \mathrm{Tr}\ B_i(G) \le \max_{i\in[N]}\Big\|\mathbb{E}_{x\sim\mu}[q_G(x)\cdot A_i^\flat(x)]\Big\|_{op}$$

due to Hölder's inequality for Schatten norms and the normalization assumption that $\sum_{i=1}^N \mathrm{Tr}\ B_i(G) = 1,\ \forall G \in \mathbb{G}_S$.

Given $\epsilon_{\mathrm{net}} > 0$, let $\mathcal{N}$ be an ϵ_{net}-net of $\mathbb{S}^{k-1}$ with the smallest possible cardinality. It follows from the standard ϵ-net argument that for each $i \in [N]$,

$$\Big\|\mathbb{E}_{x\sim\mu}[q_G(x)\cdot A_i^\flat(x)]\Big\|_{op} = \sup_{v\in\mathbb{S}^{k-1}} v^T\mathbb{E}_{x\sim\mu}\big[q_G(x)\cdot A_i^\flat(x)\big]v \le \frac{1}{1-2\epsilon_{\mathrm{net}}}\max_{v\in\mathcal{N}}\mathbb{E}_{x\sim\mu}\big[q_G(x)\cdot v^T A_i^\flat(x)v\big].$$

Next, we observe that if $G \sim \mathcal{N}_0$, then for every function $f : H_n \to \mathbb{R}$, the derived random variable $\langle f,\ q_G(x)\rangle_\mu$ is a centered Gaussian random variable with variance

$$\mathbb{E}_{G\sim\mathcal{N}_0}\Big[\langle f,\ q_G(x)\rangle_\mu^2\Big] = \mathbb{E}_{G\sim\mathcal{N}_0}\Big[\mathbb{E}_{x\sim\mu}\big[f(x)\cdot x^TGx\big]^2\Big] = \mathbb{E}_{G\sim\mathcal{N}_0}\left[\mathbb{E}_{x\sim\mu}\left[f(x)\cdot\Big(\mathrm{Tr}\ G + 2\sum_{i<j}G_{ij}x_ix_j\Big)\right]^2\right]$$
$$= 4\sum_{i<j}\mathbb{E}_{G\sim\mathcal{N}_0}[G_{ij}^2]\cdot\mathbb{E}_{x\sim\mu}[f(x)x_ix_j]^2 = 2\sum_{i<j}\langle f(x),\ x_ix_j\rangle_\mu^2 = 2\|\mathrm{proj}_2 f\|_2^2.$$

Then we use Lemma 13 to obtain the following inequalities:

$$\mathbb{E}_{G\sim\mathcal{N}_0|\mathbb{G}_S}\sum_{i=1}^N \Big\langle \mathbb{E}_{x\sim\mu}[q_G(x)\cdot A_i^\flat(x)],\ B_i(G)\Big\rangle \le \frac{1}{\Pr[G\in\mathbb{G}_S]}\mathbb{E}_{G\sim\mathcal{N}_0}\sum_{i=1}^N \Big\langle \mathbb{E}_{x\sim\mu}[q_G(x)\cdot A_i^\flat(x)],\ B_i(G)\Big\rangle$$
$$\le \frac{1}{\Pr[G\in\mathbb{G}_S]}\frac{1}{1-2\epsilon_{\mathrm{net}}}\mathbb{E}_{G\sim\mathcal{N}_0}\left[\max_{\substack{i\in[N]\\ v\in\mathcal{N}}}\mathbb{E}_{x\sim\mu}[q_G(x)\ v^TA_i^\flat(x)v]\right]$$
$$\le \frac{1}{\Pr[G\in\mathbb{G}_S]}\frac{2}{1-2\epsilon_{\mathrm{net}}}\left(\max_{\substack{i\in[N]\\ v\in\mathcal{N}}}\|\mathrm{proj}_2(v^TA_i^\flat v)\|_2\right)\sqrt{\log(N|\mathcal{N}|)}.$$

We have seen in Step 2 that $\Pr[G \in \mathbb{G}_S] \ge 1 - \exp\big(-\frac{(5\sqrt{2}-1)^2}{4\pi}\big) \ge 4/5$. Also, Lemma 6 ensures that $\|\mathrm{proj}_2(v^TA_i^\flat v)\|_2 \le e\log\Lambda$ for all (i,v), provided that we will choose the threshold $\Lambda \ge e$. Lastly, it is well known that $|\mathcal{N}| \le \big(1+\frac{2}{\epsilon_{\mathrm{net}}}\big)^k$. In conclusion, we obtain

$$\mathbb{E}_{G\sim\mathcal{N}_0|\mathbb{G}_S}\sum_{i=1}^N \Big\langle \mathbb{E}_{x\sim\mu}[q_G(x)\cdot A_i^\flat(x)],\ B_i(G)\Big\rangle \le \frac{5e\log\Lambda}{2(1-2\epsilon_{\mathrm{net}})}\sqrt{\log\left[N\Big(1+\frac{2}{\epsilon_{\mathrm{net}}}\Big)^k\right]}. \tag{39}$$

Step 4. Concluding the Proof Lastly, we revisit Eq. (34) to conclude the proof. Recall that we obtained a lower bound for the left-hand side in Step 2, cf. (35), and derived an upper bound for the right-hand side in Step 3, cf. (37), (38), and (39). Putting these together, we obtain the following inequality that holds for any choice of parameters ϵ_{net}, Λ such that $0 < \epsilon_{\text{net}} < \frac{1}{2}$ and $\Lambda \geq e$:

$$\frac{1}{2200(1+\epsilon)}\sqrt{n} \leq \frac{16(1+\epsilon)}{\Lambda} kn\sqrt{n}N + \frac{5e\log\Lambda}{2(1-2\epsilon_{\text{net}})}\sqrt{\log\left[N\Big(1+\frac{2}{\epsilon_{\text{net}}}\Big)^k\right]}. \tag{40}$$

We choose $\epsilon_{\text{net}} = 1/4$ for simplicity because optimizing ϵ_{net} does not make much difference. Next, we find Λ that minimizes the right-hand side of (40). It is easy to see that the upper bound is minimized (w.r.t. Λ) at $\Lambda^* = \frac{16(1+\epsilon)kn\sqrt{n}N}{5e\sqrt{\log(9^k N)}}$. As a result, we get the following inequality from (40) by choosing $\Lambda = \Lambda^*$ and noticing $N \geq 1$:

$$\begin{aligned}\frac{1}{11000e(1+\epsilon)}\sqrt{n} &\leq \sqrt{\log(9^k N)} \cdot \log\left(\frac{16(1+\epsilon)kn\sqrt{n}N}{5\sqrt{\log(9^k N)}}\right) \\ &\leq \sqrt{\log(9^k N)} \cdot \left[\log N + \log\left(\frac{16(1+\epsilon)kn\sqrt{n}}{5\sqrt{\log(9^k)}}\right)\right]. \end{aligned} \tag{41}$$

Letting $z = \sqrt{\log(9^k N)}$, we can see that (41) is a cubic inequality of the form $z^3 + 3\beta z \geq 2\alpha$ where

$$\alpha = \frac{\sqrt{n}}{22000e(1+\epsilon)} \quad \text{and} \quad \beta = \frac{1}{3}\log\left(\frac{16(1+\epsilon)kn\sqrt{n}}{5\cdot 9^k\sqrt{\log(9^k)}}\right).$$

We want to solve the cubic inequality with an implicit constraint $z > 0$ because $\log(9^k N) > 0$ for all $k, N \geq 1$.

Note that $\alpha > 0$ for all $\epsilon \geq 0$, $n \geq 1$. Observe that the cubic equation $z^3 + 3\beta z - 2\alpha = 0$ always has a unique positive real root when $\alpha > 0$, regardless of the value of β. Letting z_* denote the positive real root, we can see that $\{z \in \mathbb{R} : \ z^3 + 3\beta z \geq 2\alpha, \ z > 0\} = \{z \in \mathbb{R} : \ z \geq z_*\}$. Indeed, we can explicitly write z_* as $z_* = \left(\alpha + \sqrt{\alpha^2+\beta^3}\right)^{1/3} + \left(\alpha - \sqrt{\alpha^2+\beta^3}\right)^{1/3}$, due to the general cubic formula, commonly referred to as Cardano's formula. See Appendix C for more details.

Consequently, we obtain the following lower bound for N by solving (30):

$$\log N \geq \left\{\left(\alpha + \sqrt{\alpha^2+\beta^3}\right)^{1/3} + \left(\alpha - \sqrt{\alpha^2+\beta^3}\right)^{1/3}\right\}^2 - 2k\log 3$$

because $\sqrt{\log(9^k N)} \geq z_*$ if and only if $\log N \geq z_*^2 - 2k\log 3$. □

References

1. Amir Ali Ahmadi, Sanjeeb Dash, and Georgina Hall. Optimization over structured subsets of positive semidefinite matrices via column generation. *Discrete Optimization*, 24:129–151, 2017.
2. Amir Ali Ahmadi and Georgina Hall. Sum of squares basis pursuit with linear and second order cone programming. *Algebraic and Geometric Methods in Discrete Mathematics*, 685:27–53, 2017.
3. Amir Ali Ahmadi and Anirudha Majumdar. DSOS and SDSOS optimization: More tractable alternatives to sum of squares and semidefinite optimization. *SIAM Journal on Applied Algebra and Geometry*, 3(2):193–230, 2019.
4. Guillaume Aubrun and Stanislaw Szarek. Dvoretzky's theorem and the complexity of entanglement detection. *Discrete Analysis*, page 1242, 2017.
5. Guillaume Aubrun and Stanisław J Szarek. *Alice and Bob meet Banach*, volume 223. American Mathematical Soc., 2017.
6. William Beckner. Inequalities in Fourier analysis. *Annals of Mathematics*, pages 159–182, 1975.
7. Grigoriy Blekherman, Santanu S Dey, Marco Molinaro, and Shengding Sun. Sparse PSD approximation of the PSD cone. *Mathematical Programming*, pages 1–24, 2020.
8. Erik G Boman, Doron Chen, Ojas Parekh, and Sivan Toledo. On factor width and symmetric H-matrices. *Linear Algebra and Its Applications*, 405:239–248, 2005.
9. Aline Bonami. Étude des coefficients de Fourier des fonctions de $L^p(G)$. In *Annales de l'Institut Fourier*, volume 20, pages 335–402, 1970.

10. Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press, 2013.
11. Hamza Fawzi. On polyhedral approximations of the positive semidefinite cone. *Mathematics of Operations Research*, 2021.
12. Hamza Fawzi, João Gouveia, Pablo A Parrilo, James Saunderson, and Rekha R Thomas. Lifting for simplicity: Concise descriptions of convex sets. *arXiv preprint arXiv:2002.09788*, 2020.
13. Hamza Fawzi and Pablo A Parrilo. Exponential lower bounds on fixed-size psd rank and semidefinite extension complexity. *arXiv preprint arXiv:1311.2571*, 2013.
14. Joao Gouveia, Pablo A Parrilo, and Rekha R Thomas. Lifts of convex sets and cone factorizations. *Mathematics of Operations Research*, 38(2):248–264, 2013.
15. Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.
16. James R Lee, Prasad Raghavendra, and David Steurer. Lower bounds on the size of semidefinite programming relaxations. In *Proceedings of the Forty-seventh Annual ACM Symposium on Theory of Computing*, pages 567–576, 2015.
17. Ryan O'Donnell. *Analysis of boolean functions*. Cambridge University Press, 2014.
18. Sean O'Rourke, Van Vu, and Ke Wang. Eigenvectors of random matrices: A survey. *Journal of Combinatorial Theory, Series A*, 144:361–442, 2016.
19. Oded Regev and Bo'az Klartag. Quantum one-way communication can be exponentially stronger than classical communication. In *Proceedings of the forty-third annual ACM Symposium on Theory of Computing*, pages 31–40, 2011.
20. R Tyrrell Rockafellar. *Convex analysis*. Princeton University Press, 1970.
21. Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.
22. Mihalis Yannakakis. Expressing combinatorial optimization problems by linear programs. *Journal of Computer and System Sciences*, 43(3):441–466, 1991.

A Proof of Some Lemmas from Section 2

A.1 Proof of Lemma 6

Proof Let $f = f_0 + f_1 + f_2 + \cdots + f_n$ be the Fourier expansion of f. Then for $0 \le \rho \le 1$,

$$\|\mathrm{proj}_2 f\|_2^2 = \|f_2\|_2^2 = \frac{1}{\rho^4}\left(\rho^2\|f_2\|_2\right)^2 \le \frac{1}{\rho^4}\sum_{k=0}^{n}\rho^{2k}\|f_k\|_2^2 = \frac{1}{\rho^4}\|T_\rho f\|_2^2.$$

With $\rho = \sqrt{p-1}$ for $1 \le p \le 2$, we have $\|T_\rho f\|_2 \le \|f\|_p$ by hypercontractivity. Then it follows that

$$\|\mathrm{proj}_2 f\|_2 \le \frac{1}{\rho^2}\|T_\rho f\|_2 \le \frac{1}{p-1}\|f\|_p \le \frac{1}{p-1}\Lambda^{p-1}$$

because $\|f\|_p = \mathbb{E}[f^p]^{\frac{1}{p}} \le \Lambda^{\frac{p-1}{p}}\mathbb{E}[f]^{\frac{1}{p}} \le \Lambda^{\frac{p-1}{p}} \le \Lambda^{p-1}$. If $\Lambda < e$, we choose $p = 2$ to get $\|\mathrm{proj}_2 f\|_2 \le \Lambda$. Otherwise, we choose $p = 1 + \frac{1}{\log \Lambda}$ to obtain $\|\mathrm{proj}_2 f\|_2 \le e\log(\Lambda)$. □

A.2 Proof of Lemma 8

Proof We consider a Gaussian process $(X_v)_{v\in\mathbb{S}^{n-1}}$ defined over $\mathbb{S}^{n-1}$ such that $X_v = v^T G v + \gamma$ with G being standard Gaussian in $\boldsymbol{S}^n$ and $\gamma \sim N(0,1)$ independent of G. It is easy to verify that $\mathbb{E}\left[\sup_{v\in\mathbb{S}^{n-1}}\langle v, Gv\rangle\right] = \mathbb{E}_{G,\gamma}\left[\sup_{v\in\mathbb{S}^{n-1}} X_v\right]$. Now we introduce an auxiliary Gaussian process $(Y_v)_{v\in\mathbb{S}^{n-1}}$ such that $Y_v = g^T v$ with $g \sim N(0, 2I_n)$. Observe that for all $u, v \in \mathbb{S}^{n-1}$, (1) $\mathbb{E}X_v = \mathbb{E}Y_v = 0$; and (2) $\mathbb{E}(X_u - X_v)^2 \le \mathbb{E}(Y_u - Y_v)^2$ because $\mathbb{E}X_v^2 = \mathbb{E}Y_v^2 = 2$ and $\mathbb{E}X_u X_v - \mathbb{E}Y_u Y_v = (1 - u^T v)^2 \ge 0$. Thus, we can apply Sudakov-Fernique inequality (Lemma 7) to obtain $\mathbb{E}_{G,\gamma}\left[\sup_{v\in\mathbb{S}^{n-1}} X_v\right] \le \mathbb{E}_{g\sim N(0,2I_n)}\left[\sup_{v\in\mathbb{S}^{n-1}} Y_v\right] = \mathbb{E}_{g\sim N(0,2I_n)}\|g\|_2 \le \left(\mathbb{E}_{g\sim N(0,2I_n)}\|g\|_2^2\right)^{1/2} = \sqrt{2n}$. □

A.3 Proof of Lemma 10

Proof Let $h_K(u) := \max_{x\in K}\langle u,\, x\rangle = \|u\|_{K^\circ}$ denote the support function of K. The function h_K is L-Lipschitz with $L = \sup_{x\in K}\|x\|_2$, the diameter of K, because for any $u, v \in \mathbb{R}^d$,

$$\left|h_K(u) - h_K(v)\right| = \left|\|u\|_{K^\circ} - \|v\|_{K^\circ}\right| \le \|u - v\|_{K^\circ} \le \sup_{x\in K}\|x\|_2\|u - v\|_2.$$

Moreover, we can show that $\sup_{x\in K}\|x\|_2 \le \sqrt{2\pi}w_G(K)$. To see this, let $B(0,R)$ denote the Euclidean ball centered at 0 with radius R. It follows from [21, Proposition 7.5.2-(e)] that $\sup_{x,y\in K}\|x - y\|_2 \le \sqrt{2\pi}w_G(K)$. Since $0 \in K$, this implies $K \subseteq B(0, \sqrt{2\pi}w_G(K))$. Applying Lemma 9 with $f = h_K$ and $\tau = \alpha w_G(K)$ completes the proof. □

A.4 Proof Sketch of Lemma 11

An Auxiliary Lemma The MGF of a decoupled Gaussian chaos is bounded as in Lemma 15.

Lemma 15 (MGF of Gaussian chaos) *Let $X, X' \sim N(0, I_n)$ be independent Gaussian random vectors and let $A \in \mathbb{R}^{n\times n}$. Then*

$$\mathbb{E}\exp(\lambda X^T A X') \le \exp(\lambda^2 \|A\|_F^2)$$

for all λ satisfying $|\lambda| \le \frac{1}{\sqrt{2}\|A\|_{op}}$.

Proof Let $A = U\Sigma V^T$ be a singular value decomposition of A, and let $g = U^T X$, $g' = V^T X'$. Observe that g, g' are independent standard Gaussian random vectors in $\mathbb{R}^n$, and that $X^T A X' = \sum_i s_i g_i g_i'$ where $\{s_i\}_{i=1}^n$ are the singular values of A (i.e., the diagonal elements of the nonnegative diagonal matrix Σ). As this is a sum of n independent random variables, we have

$$\mathbb{E}\exp(\lambda X^T A X') = \prod_{i=1}^n \mathbb{E}\exp(\lambda s_i g_i g_i').$$

Now, for each $i \in [n]$, we use the MGF formulas for the Gaussian and the chi-squared random variables to get

$$\mathbb{E}_{g_i, g_i'} \exp(\lambda s_i g_i g_i') = \mathbb{E}_{g_i} \exp(\lambda^2 s_i^2 g_i^2/2) = (1-\lambda^2 s_i^2)^{-1/2} \quad \text{for } \lambda \text{ such that } \lambda^2 s_i^2 < 1.$$

Since $(1-t)^{-1/2} \le e^t$ for all t satisfying $0 \le t \le 0.7968$, we have

$$\mathbb{E}\exp(\lambda X^T A X') \le \prod_{i=1}^n \exp(\lambda^2 s_i^2) = \exp(\lambda^2 \|A\|_F^2)$$

for all λ such that $\lambda^2 \|A\|_{op}^2 \le 1/2 < 0.7968$. □

Proof Sketch of Lemma 11

Proof (Sketch) Let X' be an independent copy of X. Then $\mathbb{E}\exp(\lambda X^T A X) \le \mathbb{E}\exp(4\lambda X^T A X')$ for all $\lambda \in \mathbb{R}$ by decoupling lemma [21, Theorem 6.1.1]. Next, let $g, g' \sim N(0, I_n)$ be independent Gaussian random vectors. Then $\mathbb{E}\exp(\lambda X^T A X') \le \mathbb{E}\exp(\lambda v g^T A g')$ for all $\lambda \in \mathbb{R}$ due to comparison lemma[9] [21, Lemma 6.2.3]. Lastly, we apply Lemma 15 the MGF upper bound for Gaussian chaos to conclude the proof. □

A.5 Proof of Lemma 12

Proof Let A be a symmetric $n \times n$ matrix such that $A_{ii} = 0$, $\forall i$ and $A_{ij} = \frac{1}{2}\mathbb{E}_{Y\sim\mu(H_n)}[Y_i Y_j f(Y)]$ for $i \ne j$. Then we observe that for all $X \in H_n$,

$$\mathrm{proj}_2(f)(X) = \sum_{\substack{i=1\\ j>i}}^n X_i X_j \mathbb{E}_{Y\sim\mu(H_n)}[Y_i Y_j f(Y)] = X^T A X.$$

Note that X_i is sub-Gaussian with sub-Gaussian parameter 1 for all i because $\mathbb{E}[e^{\lambda X_i}] = \frac{1}{2}(e^\lambda + e^{-\lambda}) \le e^{\lambda^2/2}$. To conclude the proof, it suffices to observe that $\|A\|_F^2 = \sum_{\substack{i=1\\ j\ne i}}^n \left(\frac{1}{2}\mathbb{E}_{X\sim\mu(H_n)}[X_i X_j f(X)]\right)^2 = \frac{1}{2}\|\mathrm{proj}_2 f\|_2^2$ and $\|A\|_{op} \le \|A\|_F$, and then apply Lemma 11. □

A.6 Proof of Lemma 13

Proof For any $\lambda \in (0, 1/c]$,

$$\begin{aligned}
\mathbb{E}\Big[\max_{i\in[N]} X_i\Big] &= \frac{1}{\lambda}\mathbb{E}\Big[\log\exp\Big(\lambda \max_{i\in[N]} X_i\Big)\Big] \le \frac{1}{\lambda}\log\mathbb{E}\Big[\exp\Big(\lambda \max_{i\in[N]} X_i\Big)\Big] \qquad \because \text{Jensen's inequality}\\
&= \frac{1}{\lambda}\log\mathbb{E}\Big[\max_{i\in[N]}\exp\Big(\lambda X_i\Big)\Big] \le \frac{1}{\lambda}\log\Big(\sum_{i=1}^N \mathbb{E}\Big[\exp\big(\lambda X_i\big)\Big]\Big)\\
&\le \frac{1}{\lambda}\log\Big(\sum_{i=1}^N e^{\frac{\lambda^2 v}{2}}\Big) \qquad \because \text{sub-exponential}\\
&= \frac{\log N}{\lambda} + \frac{\lambda v}{2}.
\end{aligned}$$

It remains to choose λ in the interval $(0, 1/c]$ to optimize the upper bound. If $\sqrt{2\log N/v} \le 1/c$, then we choose $\lambda = \sqrt{2\log N/v}$ to get $\mathbb{E}\big[\max_{i\in[N]} X_i\big] \le \sqrt{2v\log N}$. On the other hand, if $\sqrt{2\log N/v} \le 1/c$, then we choose $\lambda = 1/c$ to get $\mathbb{E}\big[\max_{i\in[N]} X_i\big] \le 2c\log N$ since $v/2c \le \sqrt{2\log N/v} \le c\log N$. □

[9] The lemma in the reference is stated with an unspecified constant C, but one can verify the inequality stated here by carefully following the proof of [21, Lemma 6.2.3].

B More on Example 3 (Ball, Needle, and Pancake)

Let $B_2^d := \{x \in \mathbb{R}^d : \|x\|_2 \le 1\}$ denote the d-dimensional unit ℓ_2-ball, and let $B = B_2^d$. Fix $0 < \delta < 1$, and let $N = \mathrm{conv}\big\{B_2^d(0,1) \cup \{\pm\frac{1}{\delta}e_1\}\big\}$ be the 'needle' where $e_1 = (1, 0, \dots, 0) \in \mathbb{R}^d$. Lastly, we define the 'pancake' $P = \{x \in B : -\delta \le x_1 \le \delta\}$ where x_1 is the first coordinate of $x \in \mathbb{R}^d$. Observe that N and P are the polars of each other, and B is the polar of itself.

First of all, $w_G(B) = \mathbb{E}_g\|g\|_2 = \kappa_d$ and it is known that $\sqrt{d-1/2} \le \kappa_d \le \sqrt{d - d/(2d+1)}$, cf. the paragraph below Definition 2. Next, we can see that $w_G(N) \ge \frac{1}{\delta}\sqrt{2/\pi}$ because $\{\pm\frac{1}{\delta}e_1\} \subseteq N$ and thus, $w_G(N) \ge w_G\big(\{\pm\frac{1}{\delta}e_1\}\big) = \frac{1}{\delta}\mathbb{E}_{g\sim\mathcal{N}(0,1)}|g| = \frac{1}{\delta}\sqrt{2/\pi}$. Lastly, observe that $w_G(P) \ge \kappa_{d-1} \ge \sqrt{d-3/2}$ because $\{0\} \times B_2^{d-1}(0,1) \subseteq P$ and $w_G(P) \ge w_G\big(\{0\} \times B_2^{d-1}(0,1)\big) = w_G\big(B_2^{d-1}(0,1)\big) = \kappa_{d-1}$.

It follows that B is an ϵ-approximation of P in the average sense for $\epsilon = \kappa_d/\kappa_{d-1} - 1 \le 3/(2d-3)$. Nevertheless, B is not an ϵ'-approximation of P in the dual-average sense unless $\epsilon' \ge \frac{1}{\delta}\sqrt{2/\pi}/\kappa_d - 1 \ge \frac{2}{\delta\sqrt{\pi(2d-1)}} - 1$, which can be made arbitrarily large by choosing small δ. For example, if we choose $\delta \le 1/\sqrt{\pi(2d-1)}$, then $\epsilon^*_{\text{dual-avg}}(P,S) \ge 1$ whereas $\epsilon^*_{\text{avg}}(P,S) \le 3/(2d-3)$ regardless of δ.

C Solving the Cubic Inequality $z^3 + \alpha z \ge \beta$ with $\beta > 0$

Consider a cubic equation of the form $z^3 + \alpha z - \beta = 0$, which is commonly referred to as a depressed cubic. Note that when $\beta > 0$, this cubic equation always has a positive real root. The other two roots can be either negative real roots (when $D \le 0$), or a pair of complex conjugate roots (when $D > 0$), depending on the sign of its discriminant, $D = (\alpha/3)^3 + (\beta/2)^2$.

Indeed, we can find the roots with a generic cubic formula, known as Cardano's formula. Let $i = \sqrt{-1}$ denote the imaginary unit, $\omega = \frac{-1+\sqrt{3}i}{2}$ be a primitive 3rd of unity, and

$$T_+ = \sqrt[3]{\frac{\beta}{2} + \sqrt{\Big(\frac{\beta}{2}\Big)^2 + \Big(\frac{\alpha}{3}\Big)^3}} \quad \text{and} \quad T_- = \sqrt[3]{\frac{\beta}{2} - \sqrt{\Big(\frac{\beta}{2}\Big)^2 + \Big(\frac{\alpha}{3}\Big)^3}}. \tag{42}$$

Case 1: $D > 0$. When $D > 0$, the cubic equation $z^3 + \alpha z - \beta = 0$ with $\beta > 0$ has only one real root, $z^* = T_+ + T_-$, which turns out to be positive. Thus, the set of real solutions for the cubic inequality $z^3 + \alpha z \ge \beta$ is $\{z \in \mathbb{R} : z \ge T_+ + T_-\}$.

Case 2: $D \le 0$. There are three real roots for the cubic equation $z^3 + \alpha z - \beta = 0$, which can be written as

$$z_1 = T_+ + T_-, \qquad z_2 = \omega T_+ + \omega^2 T_-, \qquad z_3 = \omega^2 T_+ + \omega T_-.$$

One of these three real roots is positive, and the other two are negative.

Note that (42) now involves complex roots, and the choice of branches might affect the order of the roots, z_1, z_2, z_3, however, the choice will not change the values of the roots. To avoid any ambiguity in our description, we choose the principal branch so that $\mathrm{Arg}\big(\sqrt[m]{z}\big) \in (-\frac{\pi}{m}, \frac{\pi}{m}]$ for any complex number z and any positive integer m.

Observe that $T_+ = \sqrt[3]{\beta/2 + \sqrt{|D|}i}$ and $\mathrm{Arg}\,(T_+) \in [0, \pi/3)$. Similarly, we can see that $\mathrm{Arg}\,(T_-) \in (-\pi/3, 0]$. It follows that $T_+ + T_-$ is a positive real number, and thus, the largest real root. Thus, the set of real solutions for the cubic inequality $z^3 + \alpha z \ge \beta$ is $\{z \in \mathbb{R} : z \ge T_+ + T_-\}$.