MEASURES OF SCALABILITY

XUEMEI CHEN, GITTA KUTYNIOK, KASSO A. OKOUDJOU, FRIEDRICH PHILIPP,
AND RONGRONG WANG

ABSTRACT. Scalable frames are frames with the property that the frame vectors
can be rescaled resulting in tight frames. However, if a frame is not scalable, one
has to aim for an approximate procedure. For this, in this paper we introduce
three novel quantitative measures of the closeness to scalability for frames in finite
dimensional real Euclidean spaces. Besides the natural measure of scalability given
by the distance of a frame to the set of scalable frames, another measure is obtained
by optimizing a quadratic functional, while the third is given by the volume of
the ellipsoid of minimal volume containing the symmetrized frame. After proving
that these measures are equivalent in a certain sense, we establish bounds on the
probability of a randomly selected frame to be scalable. In the process, we also
derive new necessary and sufficient conditions for a frame to be scalable.

1. INTRODUCTION

During the last years, frames have had a tremendous impact on applications due
to their unique ability to deliver redundant, yet stable expansions. The redundancy
of a frame is typically utilized by applications which either require robustness of the
frame coefficients to noise, erasures, quantization, etc. or require sparse expansions
in the frame. More precisely, letting $\Phi = \{\varphi_i\}_{i=1}^{M} \subset \mathbb{R}^N$ be a frame, either *decompositions* into a sequence of frame coefficients of a signal $x \in \mathbb{R}^N$, which is the image
of x under the analysis operator $T : \mathbb{R}^N \to \mathbb{R}^M$, $x \mapsto (\langle x, \varphi_i \rangle)_{i=1}^{M}$, are exploited
by applications such as telecommunications and imaging sciences, or *expansions* in
terms of the frame, i.e., $x = \sum_{i=1}^{M} c_i \varphi_i$ with suitable choice of coefficients $(c_i)_{i=1}^{M}$,
are required by applications such as efficient PDE solvers. Intriguingly, the novel
area of compressed sensing is based on the fact that typically signals exhibit a sparse
expansion in a frame, which is nowadays considered the standard paradigm in data
processing. Some compressed sensing applications also 'hope' that the sequence of
frame coefficients itself is sparse; a connection deeply studied in a series of papers on
cosparsity (cf. [18]).

The discussed applications certainly require stability, numerically as well as theoretically. For instance, notice that most results in compressed sensing are stated for
tight frames, i.e., for optimal stability. It is known that such frames – in the case of
normalized vectors – can be characterized by the frame potential (see, e.g., [2, 6, 11])
and construction methods have been derived (cf. [5] and [21] for an algebro-geometric

point of view). However, a crucial question remains: Given a frame with desirable properties, can we turn it into a tight frame? The immediate answer is yes, since it can easily be shown that applying $S^{-1/2}$ to each frame element, $S : \mathbb{R}^N \to \mathbb{R}^N$ denoting the frame operator $Sx = \sum_{i=1}^{M} \langle x, \varphi_i \rangle \varphi_i$, produces a Parseval frame. Thinking further one however realizes a serious problem with this seemingly elegant approach; it typically completely destroys any properties of the frame for which it was carefully designed before. Thus, unless we are merely interested in theoretical considerations, this approach is unacceptable.

Trying to be as careful as possible, the most noninvasive approach seems to merely scale each frame vector, i.e., multiply it by a scalar. And, indeed, almost all frame properties one can think of such as erasure resilience or sparse expansions are left untouched by this modification. In fact, this approach is currently extensively studied under the theme of scalable frames.

1.1. **Scalability of Frames.** The notion of a *scalable frame* was first introduced in [17] as a frame whose frame vectors can be rescaled to yield a tight frame. Recall that a sequence $\Phi = \{\varphi_i\}_{i=1}^{M} \subset \mathbb{R}^N$ forms a *frame* provided that

$$A\|x\|^2 \leq \sum_{i=1}^{M} |\langle x, \varphi_i \rangle|^2 \leq B\|x\|^2$$

for all $x \in \mathbb{R}^N$, where A and B are called the *frame bounds*. One often also writes Φ for the $N \times M$ matrix whose ith column is the vector φ_i. When $A = B$, the frame is called a *tight frame*. Furthermore, $A = B = 1$ produces a *Parseval frame*. In the sequel, the set of frames with M vectors in $\mathbb{R}^N$ will be denoted by $\mathcal{F}(M, N)$. We refer to [9] for an introduction to frame theory and to [7] for an overview of the current research in the field.

A frame $\Phi = \{\varphi_i\}_{i=1}^{M}$ for $\mathbb{R}^N$ is called (*strictly*) *scalable* if there exist nonnegative (positive, respectively) scalars $\{s_i\}_{i=1}^{M}$ such that $\{s_i\varphi_i\}_{i=1}^{M}$ is a tight frame for $\mathbb{R}^N$. The set of (strictly) scalable frames is denoted by $\mathcal{SC}(M, N)$ ($\mathcal{SC}_+(M, N)$, respectively). This definition obviously allows one to restrict the study to the class of unit norm frames

$$\mathcal{F}_u(M, N) := \left\{ \{\varphi_i\}_{i=1}^{M} \in \mathcal{F}(M, N) : \|\varphi_i\|_2 = 1 \text{ for } i = 1, \ldots, M \right\},$$

and further to substitute tight frame by Parseval frame in the above definition. Therefore a frame $\Phi = \{\varphi_i\}_{i=1}^{M}$ is scalable if and only if there exist non-negative scalars $\{c_i\}_{i=1}^{M}$ such that

$$(1.1) \qquad \Phi C \Phi^T = \sum_{i=1}^{M} c_i \varphi_1 \varphi_i^T = I, \quad \text{where } C = \text{diag}(c_i).$$

In [17], characterizations of $\mathcal{SC}(M, N)$ and $\mathcal{SC}_+(M, N)$, both of functional analytic and geometric type were derived in the infinite as well as finite dimensional setting. As a sequel, using topological considerations, it was proved in [16] that the set of scalable frames, $\mathcal{SC}(M, N)$, is a 'small' subset of $\mathcal{F}(M, N)$ when M is relatively small and a yet different characterization using a particular mapping was derived. This last

mapping is closely related to the so-called diagram vectors/mapping in [10]. In [4], arbitrary scalars in $\mathbb{C}$ were allowed, and it was shown that in this case most frames are either not scalable or scalable in a unique way and, if uniqueness is not given, the set of all possible sequences of scalars is studied.

1.2. **How Scalable is a Frame?** However, in the applied world, scalability seems too idealistic, in particular, if our frame at hand is not scalable. This calls for a measure of 'closeness to being scalable'. It is though not obvious how to define such a measure, and one can easily justify different points of view of what 'closeness' shall mean. Let us discuss the following three viewpoints:

- *Distance to $\mathcal{SC}(M, N)$.* Maybe the most straightforward approach is to measure the distance of a frame $\Phi \in \mathcal{F}_u(M, N)$ to the set of scalable frames:

$$d_\Phi := \inf_{\Psi \in \mathcal{SC}(M,N)} \|\Phi - \Psi\|_F.$$

  This notion seems natural if we anticipate efficient algorithmic approaches for computing the closest scalable frame by projections onto $\mathcal{SC}(M, N)$.

- *Conical Viewpoint.* Inspired by (1.1), we observe that Φ is scalable if and only if the identity operator I lies in the cone generated by the vectors $\varphi_i \varphi_i^T$, $i = 1, \ldots, M$, which is $\{\sum_{i=1}^{M} c_i \varphi_i \varphi_i^T : c_i \geq 0\}$. Thus the distance of I to this cone seems to be another suitable measure for scalability of $\Phi \in \mathcal{F}_u(M, N)$, and we define

$$D_\Phi := \min_{C \geq 0 \text{ diagonal}} \left\| \Phi C \Phi^T - I \right\|_F,$$

  where $\| \cdot \|_F$ denotes the Frobenius norm. Note that the minimum is attained because this polyhedral cone is closed. This conical viewpoint leads to a computationally efficient algorithm, since we can recast the problem as a quadratic program (see Section 3.2).

- *Ellipsoidal Viewpoint.* Finally, one can consider the ellipsoid of minimal volume (also known as the Löwner ellipsoid) circumscribing the convex hull of the symmetrized frame of $\Phi \in \mathcal{F}_u(M, N)$:

$$\Phi_{\text{Sym}} := \{\varphi_i\}_{i=1}^{M} \cup \{-\varphi_i\}_{i=1}^{M},$$

  which in the sequel we denote by E_Φ and refer to as the *minimal ellipsoid of* Φ. Its 'normalized' volume is defined by

$$V_\Phi := \frac{\text{Vol}(E_\Phi)}{\omega_N},$$

  where ω_N is the volume of the unit ball in $\mathbb{R}^N$. By definition, we have $V_\Phi \leq 1$, and we will later show (Theorem 2.11) that $V_\Phi = 1$ holds if and only if the frame Φ is scalable. Hence, yet another conceivably useful measure for scalability is the closeness of V_Φ to 1. This ellipsoidal viewpoint establishes a novel link to convex geometry. Moreover, it will turn out that this measure is of particular use when estimating the probability of a random frame being scalable.

Each notion seems justified from a different perspective, and hence there is no 'general truth' for what the best measure is.

1.3. **Our Contributions.** Our contributions are three-fold: First, we introduce the scalability measures d_Φ, D_Φ, and V_Φ, derive estimates for their values, and study their relations in Theorems 3.3 and 3.4. Second, with Theorems 2.11 and 4.1 we provide new necessary and sufficient conditions for scalability based on the ellipsoidal viewpoint. And, third, we estimate the probability of a frame being scalable when each frame vector is drawn independently and uniformly from the unit sphere (see Theorem 4.9).

1.4. **Expected Impact.** We anticipate our results to have the following impacts:

- *Constructions of Scalable Frames:* One construction procedure which is a byproduct of our analysis is to consider random frames with the probability of scalability being explicitly given. However, certainly, there is the need for more sophisticated efficient algorithmic approaches. But with the measures provided in our work, the groundwork is laid for analyzing their accuracy.
- *Extensions of Scalability:* One might also imagine other methodological approaches to modify a frame to become tight. If sparse approximation properties is what one seeks, another possibility is to be allowed to take linear combinations of 'few' frame vectors in the spirit of the 'double sparsity' approach in [20]. The introduced three quality measures provide an understanding of scalability which we hope might allow an attack on analyzing those approaches as well.
- *ϵ-Scalability:* One key question even more important to applications than scalability is that of what is typically loosely coined ϵ-scalability, meaning a frame which is scalable 'up to an ϵ', but which was not precisely defined before. The scalability measures now immediately provide even three definitions of ϵ-scalability in a very natural way, opening three doors to approaching this problem.
- *Convex Geometry:* The ellipsoidal viewpoint of scalability provides a very interesting link between frame theory and convex geometry. Theorem 4.1 and Theorem 4.9 are results about frames using convex geometry tools; Theorem 2.13 is a result about minimal ellipsoids exploiting frame theory. We strongly expect the link established in this paper to bear further fruits in frame theory, in particular the approach of regarding frames from a convex geometric viewpoint by analyzing the convex hull of a (symmetrized) frame.

1.5. **Outline.** This paper is organized as follows. In Section, 2, the three measures of closeness of a given frame to be scalable are introduced in three respective subsections and some basic properties are studied. This is followed by a comparison of the measures both theoretically and numerically (Section 3). Finally, in Section 4 we exploit those results to analyze the probability of a frame to be scalable. Interestingly, along the way we derive necessary and sufficient (deterministic) conditions for a frame to be scalable (see Subsection 4.1).

2. Properties of the measures of Scalability

In this section, we explore some basic properties of the three measures of scalibility which we introduced in the previous section. As mentioned before, we consider only unit norm frames.

2.1. **Distance to the Set of Scalable Frames.** Recall that the measure d_Φ was defined as the distance of Φ to the set of scalable frames:

$$(2.1) \qquad d_\Phi = \inf_{\Psi \in \mathcal{SC}(M,N)} \|\Phi - \Psi\|_F.$$

Since the set $\mathcal{SC}(M, N)$ is not closed (choose $\Phi \in \mathcal{SC}(M, N)$, then $(\frac{1}{n}\Phi)_{n\in\mathbb{N}}$ is a sequence in $\mathcal{SC}(M, N)$ which converges to the zero matrix), it is not clear whether the infimum in (2.1) is attained. The following proposition, however, shows that this is the case if $d_\Phi < 1$.

Proposition 2.1. *If $\Phi \in \mathcal{F}_u(M, N)$ such that $d_\Phi < 1$ then there exists $\hat{\Phi} \in \mathcal{SC}(M, N)$ such that $\|\Phi - \hat{\Phi}\|_F = d_\Phi$.*

Proof. Let $\varepsilon = \frac{1-d_\Phi}{2}$, and $\{\Phi_n\}_{n\in\mathbb{N}} \subset \mathcal{SC}(M, N)$ be a sequence of scalable frames such that $\|\Phi - \Phi_n\|_F \leq d_\Phi + \varepsilon/n$. The sequence $\{\Phi_n\}_{n\in\mathbb{N}}$ is bounded as

$$\|\Phi_n\|_F \leq \|\Phi\|_F + \|\Phi - \Phi_n\|_F \leq \sqrt{M} + d_\Phi + \frac{1 - d_\Phi}{2},$$

so without loss of generality, we assume that $\{\Phi_n\}_{n\in\mathbb{N}}$ converges to some $\hat{\Phi} \in \mathbb{R}^{N\times M}$. It remains to prove that $\hat{\Phi}$ is scalable. For this, denote by $\varphi_{i,n}$ the i-th column of Φ_n. Then

$$\|\varphi_{i,n}\|_2 \geq \|\varphi_i\|_2 - \|\varphi_i - \varphi_{i,n}\|_2 \geq 1 - d_\Phi - \varepsilon = \varepsilon$$

for all $n \geq 0$ and all $i \in \{1, \ldots, M\}$. Let $C_n = \text{diag}(c_{1,n}, \ldots, c_{M,n})$ be a non-negative diagonal matrix such that $\Phi_n C_n \Phi_n^T = I$. Now, for each $j \in \{1, \ldots, M\}$ and each $n \geq 0$ we have

$$N = \text{Tr}(I) = \text{Tr}\left(\sum_{i=1}^{M} c_{i,n}\varphi_{i,n}\varphi_{i,n}^T\right) = \sum_{i=1}^{M} c_{i,n}\|\varphi_{i,n}\|^2 \geq \varepsilon^2 c_{j,n}.$$

Therefore, each sequence $(c_{i,n})_{n\in\mathbb{N}}$ is bounded. Thus, we find an index sequence $(n_k)_{k\in\mathbb{N}}$ such that

$$c_i := \lim_{k\to\infty} c_{i,n_k}$$

exists for each $i \in \{1, \ldots, M\}$. Now, it is easy to see that $\Phi_{n_k} C_{n_k} \Phi_{n_k}^T$ converges to $\hat{\Phi} C \hat{\Phi}^T$ as $k \to \infty$, where $C := \text{diag}(c_1, \ldots, c_m)$. Hence, $\hat{\Phi} C \hat{\Phi}^T = I$, and $\hat{\Phi}$ is a scalable frame. $\square$

Remark 2.2. The proof of Proposition 2.1 also yields that the frame vectors of any minimizer of (2.1) are non-zero if $d_\Phi < 1$.

Lemma 2.3. *Assume that $d_\Phi < 1$, and let $\hat{\Phi} = \{\hat{\varphi}_i\}_{i=1}^{M}$ be a minimizer of (2.1). Then for every $i = 1, \ldots, M$,*
   (i) $\langle \varphi_i, \hat{\varphi}_i \rangle = \|\hat{\varphi}_i\|_2^2$.

(ii) $\|\hat{\varphi}_i\|_2 \leq 1$, and equality holds if and only if $\hat{\varphi}_i = \varphi_i$.

(iii) $\|\hat{\Phi}\|_F^2 = M - d_\Phi^2$.

Proof. (i). Fix $j \in \{1, \ldots, M\}$ and $\alpha \in \mathbb{R} \setminus \{0\}$ be arbitrary. Define the frame $\Psi = \{\psi_i\}_{i=1}^M$ as

$$\psi_i = \begin{cases} \hat{\varphi}_i & \text{if } i \neq j \\ \alpha\hat{\varphi}_j & \text{if } i = j \end{cases},$$

which is scalable. Hence, we have

$$\|\Phi - \hat{\Phi}\|_F^2 \leq \|\Phi - \Psi\|_F^2 = \sum_{i=1}^M \|\varphi_i - \psi_i\|_2^2 = \sum_{i \neq j} \|\varphi_i - \hat{\varphi}_i\|_2^2 + \|\varphi_j - \alpha\hat{\varphi}_j\|_2^2$$

$$= \sum_{i=1}^M \|\varphi_i - \hat{\varphi}_i\|_2^2 + \left(\|\varphi_j - \alpha\hat{\varphi}_j\|_2^2 - \|\varphi_j - \hat{\varphi}_j\|_2^2 \right)$$

$$= \|\Phi - \hat{\Phi}\|_F^2 + \left(\|\varphi_j - \alpha\hat{\varphi}_j\|_2^2 - \|\varphi_j - \hat{\varphi}_j\|_2^2 \right).$$

This implies

$$\|\varphi_j - \alpha\hat{\varphi}_j\|_2^2 \geq \|\varphi_j - \hat{\varphi}_j\|_2^2$$

or, equivalently,

(2.2) $$-2\alpha \langle \varphi_j, \hat{\varphi}_j \rangle + \alpha^2 \|\hat{\varphi}_j\|_2^2 \geq -2\langle \varphi_j, \hat{\varphi}_j \rangle + \|\hat{\varphi}_j\|_2^2$$

for all $\alpha \in \mathbb{R} \setminus \{0\}$ and all $j \in \{1, \ldots, M\}$. Putting $\alpha = \frac{\langle \varphi_j, \hat{\varphi}_j \rangle}{\|\hat{\varphi}_j\|_2^2}$ in (2.2) gives

$$-\frac{\langle \varphi_j, \hat{\varphi}_j \rangle^2}{\|\hat{\varphi}_j\|_2^2} \geq -2\langle \varphi_j, \hat{\varphi}_j \rangle + \|\hat{\varphi}_j\|_2^2,$$

which is equivalent to

$$0 \geq \left(\frac{\langle \varphi_j, \hat{\varphi}_j \rangle}{\|\hat{\varphi}_j\|_2} - \|\hat{\varphi}_j\|_2 \right)^2,$$

which leads to the conclusion.

(ii). By (i) we have

$$\|\hat{\varphi}_j\|_2^2 = \langle \varphi_j, \hat{\varphi}_j \rangle \leq \|\varphi_j\|_2 \|\hat{\varphi}_j\|_2 = \|\hat{\varphi}_j\|_2$$

This proves $\|\hat{\varphi}_j\|_2 \leq 1$ and that $\|\hat{\varphi}_j\|_2 = 1$ holds if and only if $\hat{\varphi}_j = \lambda\varphi_j$ for some $\lambda \in \mathbb{R}$. In the latter case, as both vectors are normalized, we have $\lambda = \pm 1$. But $\hat{\varphi}_j = -\varphi_j$ is impossible due to (i). Thus, $\varphi_j = \hat{\varphi}_j$ follows.

(iii). By (i),

$$M - d_\Phi^2 = M - \|\Phi - \hat{\Phi}\|_F^2 = M - \sum_{i=1}^M \|\varphi_i - \hat{\varphi}_i\|_2^2$$

$$= M - \sum_{i=1}^M \left(1 - 2\langle \varphi_i, \hat{\varphi}_i \rangle + \|\hat{\varphi}_i\|_2^2 \right) = \sum_{i=1}^M \|\hat{\varphi}_i\|_2^2 = \|\hat{\Phi}\|_F^2.$$

This proves the claim.   □

Since we do not yet have a complete understanding of the set $\mathcal{SC}(M, N)$, we do not have an algorithm for calculating the infimum d_Φ in (2.1). For this reason, we introduce two other measures of scalability in the remainder of this section which are more accessible in practice. We will relate these measures to each other and to d_Φ in Section 3.

2.2. **Distance of the Identity to a Cone.** As mentioned in the introduction, the measure D_Φ for the scalability of $\Phi \in \mathcal{F}_u(M, N)$ is the distance of the identity operator on $\mathbb{R}^N$ to the cone generated by $\{\varphi_i \varphi_i^T\}$. Let us recall its definition:

$$(2.3) \qquad D_\Phi := \min_{c_i \geq 0} \left\| \sum_{i=1}^{M} c_i \varphi_i \varphi_i^T - I \right\|_F = \min_{C \geq 0 \text{ diagonal}} \left\| \Phi C \Phi^T - I \right\|_F.$$

For the following, it is convenient to represent the function to be minimized in (2.3) in another form:

$$\left\| \sum_{i=1}^{M} c_i \varphi_i \varphi_i^T - I \right\|_F^2 = \mathrm{Tr} \left(\sum_{i,j=1}^{M} c_i c_j \varphi_i \varphi_i^T \varphi_j \varphi_j^T - 2 \sum_{i=1}^{M} c_i \varphi_i \varphi_i^T + I \right)$$

$$(2.4) \qquad = \sum_{i,j=1}^{M} c_i c_j |\langle \varphi_i, \varphi_j \rangle|^2 - 2 \sum_{i=1}^{M} c_i + N.$$

If we now put $\mathbf{1} := (1, \ldots, 1)^T \in \mathbb{R}^M$, $f_{ij} := |\langle \varphi_i, \varphi_j \rangle|^2$, $i, j = 1, \ldots, M$, $F := (f_{ij})_{i,j=1}^M$, and $c := (c_1, \ldots, c_m)^T$, we obtain

$$(2.5) \qquad g(c) := \left\| \sum_{i=1}^{M} c_i \varphi_i \varphi_i^T - I \right\|_F^2 = c^T F c - 2 \cdot \mathbf{1}^T c + N.$$

First of all, we can associate D_Φ with the frame potential (see, e.g., [2]):

$$\mathbb{FP}(\Phi) := \sum_{i,j=1}^{M} |\langle \varphi_i, \varphi_j \rangle|^2.$$

By plugging in $\alpha \mathbf{1}$ into g with $\alpha > 0$:

$$g(\alpha \mathbf{1}) = \alpha^2 \mathbb{FP}(\Phi) - 2M\alpha + N.$$

So,

$$D_\Phi^2 \leq \min_{\alpha \geq 0} g(\alpha \mathbf{1}) = N - \frac{M^2}{\mathbb{FP}(\Phi)}.$$

We summarize the above discussion in a proposition.

Proposition 2.4. *For $\Phi \in \mathcal{F}_u(M, N)$ we have*

$$(2.6) \qquad D_\Phi^2 \leq N - \frac{M^2}{\mathbb{FP}(\Phi)}.$$

Remark 2.5. Since $\mathbb{FP}(\Phi) < M^2$, the inequality (2.6) implies that $D_\Phi < \sqrt{N-1}$. It is worth noting that this upper bound is sharp in the sense that for each $\varepsilon > 0$ there exists $\Phi \in \mathcal{F}_u(M, N)$ such that $D_\Phi > \sqrt{N-1} - \varepsilon$. This can be proved by essentially choosing the frame vectors of Φ very close to each other.

The following proposition can be thought of as an analog to Lemma 2.3 (iii).

Proposition 2.6. *Let the non-negative diagonal matrix* $\hat{C} = \mathrm{diag}(\hat{c}_1, \ldots, \hat{c}_M) \in \mathbb{R}^{M \times M}$ *be a minimizer of* (2.3). *Then*

$$(2.7) \qquad Tr(\Phi \hat{C} \Phi^T) = \sum_{i=1}^{M} \hat{c}_i = N - D_\Phi^2.$$

Proof. The first equality in (2.7) is due to the fact that the φ_i's are normalized. Define

$$f(\alpha) := g(\alpha\hat{c}) = \alpha^2 \hat{c}^T F \hat{c} - 2\alpha \mathbf{1}^T \hat{c} + N.$$

for $\alpha > 0$. The function $f(\alpha)$ has a local minimum at $\alpha = 1$, therefore

$$\left.\frac{df}{d\alpha}\right|_{\alpha=1} = 0 \quad \Longrightarrow \quad \hat{c}^T F \hat{c} = \mathbf{1}^T c.$$

So,

$$D_\Phi^2 = f(1) = \hat{c}^T F \hat{c} - 2 \cdot \mathbf{1}^T \hat{c} + N = N - \mathbf{1}^T \hat{c} = N - \sum_{i=1}^{M} \hat{c}_i,$$

which proves the proposition. $\qquad\square$

2.3. **Volume of the Smallest Ellipsoid Enclosing the Symmetrized Frame.** In the following, we shall examine the properties of the measure V_Φ. We will have to recall a few facts from convex geometry, especially results dealing with the ellipsoid of a convex polytope first. An N-dimensional ellipsoid centered at c is defined as

$$E(X, c) := c + X^{-1/2}(B) = \{v : \langle X(v - c), (v - c)\rangle \leq 1\},$$

where X is an $N \times N$ positive definite matrix, and B is the unit ball in $\mathbb{R}^N$. It is easy to see that

$$(2.8) \qquad \mathrm{Vol}(E(X, c)) = \det(X^{-1/2})\omega_N.$$

Here, as already mentioned in the introduction, ω_N denotes the volume of the unit ball in $\mathbb{R}^N$.

A *convex body* in $\mathbb{R}^N$ is a nonempty compact convex subset of $\mathbb{R}^N$. It is well-known that for any convex body K in $\mathbb{R}^N$ with nonempty interior there is a unique ellipsoid of minimal volume containing K and a unique ellipsoid of maximal volume contained in K; see, e.g., [22, Chapter 3]. We refer to [1, 12, 22] for more on these extremal ellipsoids.

In what follows, we only consider the ellipsoid of minimal volume that encloses a given convex body, and this ellipsoid will be called the *minimal ellipsoid* of that convex body. The following theorem is a generalization of John's ellipsoid theorem [13], which will be referred as John's theorem in this paper.

Theorem 2.7. [12, Theorem 12.9] *Let $K \subset \mathbb{R}^N$ be a convex body and let X be an $N \times N$ positive definite matrix. Then the following are equivalent:*

(i) *$E(X, c)$ is the minimal ellipsoid of K.*

(ii) *$K \subset E(X, c)$, and there exist positive multipliers $\{\lambda_i\}_{i=1}^{k}$, and contact points $\{u_i\}_{i=1}^{k}$ in K such that*

$$(2.9) \qquad X^{-1} = \sum_{i=1}^{k} \lambda_i (u_i - c)(u_i - c)^T,$$

$$(2.10) \qquad 0 = \sum_{i=1}^{k} \lambda_i (u_i - c),$$

$$(2.11) \qquad u_i \in \partial K \cap \partial E(X, c), \quad i = 1, \ldots, k.$$

Given a frame $\Phi = \{\varphi_i\}_{i=1}^{M} \in \mathcal{F}_u(M, N)$, we will apply John's theorem to the convex hull of the symmetrized frame $\Phi_{\mathrm{Sym}} = \{\varphi_i\}_{i=1}^{M} \cup \{-\varphi_i\}_{i=1}^{M}$. By E_Φ we will denote the minimal ellipsoid of the convex hull of Φ_{Sym}. We shall also call this ellipsoid the *minimal ellipsoid of* Φ. This is not in conflict with the notion of the minimal ellipsoid of a convex body since the finite set Φ is not a convex body. The next lemma says that the center of E_Φ is always 0.

Lemma 2.8. *Let K be a convex body which is symmetric about the origin. Then the center of the minimal ellipsoid of K is 0.*

Proof. Let $E(X, c)$ denote the minimal ellipsoid of K. By definition, if $u \in K$ we also have $-u \in K$, which implies

$$\langle X(u - c), u - c \rangle \leq 1 \quad \text{and} \quad \langle X(-u - c), -u - c \rangle \leq 1.$$

Adding those inequalities, we obtain

$$2\langle Xu, u \rangle + 2\langle Xc, c \rangle \leq 2.$$

Since $X \in \mathbb{R}^{N \times N}$ is positive definite, the above equation implies $\langle Xu, u \rangle \leq 1$ or, equivalently, $u \in E(X, 0)$. This proves $K \subset E(X, 0)$. And as $E(X, 0)$ has the same volume as $E(X, c)$, it follows from the uniqueness of minimal ellipsoids that $c = 0$. $\square$

In the following, we write $E(X)$ instead of $E(X, 0)$. For completeness, we now state a version of Theorem 2.7 that is specifically taylored to our situation.

Corollary 2.9. *Let $\Phi = \{\varphi_i\}_{i=1}^{M} \in \mathcal{F}_u(M, N)$, and let X be an $N \times N$ positive definite matrix. Then the following are equivalent:*

(i) *$E(X)$ is the minimal ellipsoid of Φ.*

(ii) *There exist nonnegative scalars $\{\rho_i\}_{i=1}^{M}$ such that*

$$(2.12) \qquad X^{-1} = \sum_{i=1}^{M} \rho_i \varphi_i \varphi_i^T,$$

$$(2.13) \qquad \langle X\varphi_i, \varphi_i \rangle \leq 1 \quad \text{for all } i = 1, 2, \ldots, M,$$

(2.14) $$\langle X\varphi_i, \varphi_i \rangle = 1 \ \text{if} \ \rho_i > 0.$$

Proof. (i)$\Rightarrow$(ii). By John's theorem, the contact points must be points in the set Φ_{Sym}. Since $\varphi_i\varphi_i^T = (-\varphi_i)(-\varphi_i)^T$, equation (2.9) with the center $c = 0$ implies that there exists $I \subset \{1, \ldots, M\}$ such that

$$X^{-1} = \sum_{i \in I} \lambda_i \varphi_i \varphi_i^T.$$

Setting $\rho_i = \lambda_i$ for $i \in I$ and $\rho_i = 0$ for $i \notin I$, we get (2.12). Equation (2.13) follows from the fact that $\varphi_i \in E(X)$ for each $i = 1, \ldots, M$, and equation (2.14) is implied by (2.11).

(ii)$\Rightarrow$(i). Let $I = \{i : \rho_i > 0\}$. Then the assumptions imply conditions (2.9) and (2.11) with $\{u_i\}_{i \in I} = \{\varphi_i\}_{i \in I}$, and $\{\lambda_i\}_{i \in I} = \{\rho_i\}_{i \in I}$. We just need to slightly modify $\{u_i\}, \{\lambda_i\}$ to make it satisfy (2.10) as well. Indeed, we replace u_i by the pair $\pm u_i$ each with half the weight of the original λ_i. Finally, (2.13) implies that the convex hull of Φ_{Sym} is contained in $E(X)$. Now, (i) follows from the application of John's theorem. $\square$

Remark 2.10. It is convenient to view (2.12) as saying that $\{X^{1/2}\varphi_i\}_{i=1}^M$ is scalable with scalars $\{\sqrt{\rho_i}\}_{i=1}^M$. Therefore by [16, Remark 3.12] (see also [4, Corollary 3.4], since the dimension of $\mathrm{span}\{\varphi_i\varphi_i\}_{i=1}^M$ is at most $\frac{N(N+1)}{2}$), we can always pick a set of ρ_i's as in (ii) above such that the number of non-zero (i.e., positive) ρ_i's does not exceed $\frac{N(N+1)}{2}$.

Recall that in the introduction we defined a third measure of scalability V_Φ as follows:

(2.15) $$V_\Phi = \frac{\mathrm{Vol}(E_\Phi)}{\omega_N} = \det\left(X^{-1/2}\right).$$

The second equality is due to (2.8).

Let us now see how V_Φ relates to scalability of Φ. If $\Phi \in \mathcal{F}_u(M, N)$ is scalable then (2.12)–(2.14) hold with $X = I$. Therefore, $E_\Phi = E(I)$ is the unit ball which implies $V_\Phi = 1$. Conversely, if $V_\Phi = 1$ then E_Φ must be the unit ball since the ellipsoid of minimal volume is unique. Hence, $E_\Phi = E(I)$, and (2.12) implies that Φ is scalable. This quickly provides another characterization of scalability.

Theorem 2.11. *A frame $\Phi \in \mathcal{F}_u(M, N)$ is scalable if and only if its minimal ellipsoid is the N-dimensional unit ball, in which case $V_\Phi = 1$.*

We can now prove an important property of the minimal ellipsoid E_Φ of a unit norm frame Φ.

Lemma 2.12. *Given $\Phi \in \mathcal{F}_u(M, N)$, let $E(X)$ be the minimal ellipsoid of Φ where $X^{-1} = \sum_{i=1}^M \rho_i \varphi_i \varphi_i^T$, and let $\{\lambda_i\}_{i=1}^N$ be the eigenvalues of X^{-1}. Then*

(2.16) $$V_\Phi = \prod_{i=1}^N \lambda_i^{1/2},$$

$$(2.17) \qquad Tr\left(X^{-1}\right) = \sum_{i=1}^{M} \rho_i = \sum_{i=1}^{N} \lambda_i = N.$$

Proof. The relation (2.16) immediately follows from (2.15). To prove (2.17), we set $u_i = X^{1/2}\varphi_i$. Then

$$(2.18) \qquad I = X^{1/2}X^{-1}X^{1/2} = X^{1/2}\left(\sum_{i=1}^{M} \rho_i\varphi_i\varphi_i^T\right)X^{1/2} = \sum_{i=1}^{M} \rho_i u_i u_i^T.$$

In addition, we know that whenever $\rho_i > 0$, we have $\langle \varphi_i, X\varphi_i \rangle = 1$, or equivalently $\|u_i\|_2 = 1$. Using this fact as well as (2.18), we deduce

$$\sum_{i=1}^{M} \rho_i = \sum_{\rho_i > 0} \rho_i \mathrm{Tr}(u_i u_i^T) = \mathrm{Tr}\left(\sum_{i=1}^{M} \rho_i u_i u_i^T\right) = \mathrm{Tr}(I) = N.$$

The lemma is proved. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ □

Given a frame Φ with minimal ellipsoid $E_\Phi = E(X)$, we have shown in (2.17) that the trace of X^{-1} is always fixed. This naturally raises the question whether any ellipsoid $E(X)$ with $\mathrm{Tr}(X^{-1}) = N$ is necessarily the minimal ellipsoid of some unit norm frame. The next theorem answers this question in the affirmative.

Theorem 2.13. *Every ellipsoid $E(X)$ with $Tr(X^{-1}) = N$ is the minimal ellipsoid of some frame $\Phi \in \mathcal{F}_u(M, N)$.*

Proof. Given any invertible positive definite matrix X^{-1} whose trace is N, there exists $\Phi' = \{\varphi_i\}_{i=1}^{N} \in \mathcal{F}_u(N, N)$ such that

$$(2.19) \qquad X^{-1} = \sum_{i=1}^{N} \varphi_i\varphi_i^T.$$

This is a direct result of Corollary 3.1 in [8].

Next, we show that $\langle X\varphi_i, \varphi_i \rangle = 1$ for all $i = 1, \dots, N$. For this, fix $j \in \{1, \dots, N\}$ and choose $x \in \{\varphi_i : i \neq j\}^{\perp}$ with $\langle x, \varphi_j \rangle = 1$. Then

$$1 = \langle x, \varphi_j \rangle = \left\langle \sum_{i=1}^{N} X\varphi_i\varphi_i^T x, \varphi_j \right\rangle = \langle X\varphi_j\varphi_j^T x, \varphi_j \rangle = \langle X\varphi_j, \varphi_j \rangle.$$

Now, it follows from Corollary 2.9 that $E(X)$ is the minimal ellipsoid of Φ'. Construct $\Phi \in \mathcal{F}_u(M, N)$ by adding $M - N$ unit norm vectors inside $E(X)$ to Φ'. Then $E(X)$ is also the minimal ellipsoid of Φ since (2.19) still holds with N replaced by M and $\rho_i = 0$ for $i = N + 1, \dots, M$. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ □

Remark 2.14. It is possible using the geometric characterization of scalable frames by V_Φ to define an equivalence relation on $\mathcal{F}_u(M, N)$. Indeed, $\Phi, \Psi \in \mathcal{F}_u(M, N)$ can be defined to be equivalent if $V_\Phi = V_\Psi$. We denote each equivalence class by the unique volume for all its members. Specifically, for any $0 < a \leq 1$, the class $P[M, N, a]$

consists of all $\Phi \in \mathcal{F}_u(M, N)$ with $V_\Phi = a$. Then, $\mathcal{SC}(M, N) = P[M, N, 1]$. This also allows a parametrization of $\mathcal{F}_u(M, N)$:

$$\mathcal{F}_u(M, N) = \bigcup_{a \in (0,1]} P[M, N, a].$$

3. Comparison of the Measures

In this section, we relate the three measures d_Φ, D_Φ, and V_Φ of scalability to each other. Hereby, we will frequently make use of the standard inequalities in the following lemma, in particular the arithmetic geometric means inequality.

Lemma 3.1. *Given $a_i > 0$, $i = 1, \ldots, N$, we have*

$$(3.1) \qquad \frac{N}{\sum_{i=1}^{N} a_i^{-1}} \leq \prod_{i=1}^{N} a_i^{\frac{1}{N}} \leq \frac{\sum_{i=1}^{N} a_i}{N},$$

$$(3.2) \qquad \sum_{i<j} a_i a_j \geq \frac{N(N-1)}{2} \prod_{i=1}^{N} a_i^{\frac{2}{N}}.$$

The inequality (3.2) *is a special case of the right hand side inequality of* (3.1).

3.1. **Comparison of D_Φ and V_Φ.** Given a frame $\Phi \in \mathcal{F}_u(M, N)$, by definition $V_\Phi \leq 1$. Moreover, by Theorem 2.11, we have $V_\Phi = 1$ if and only if the frame is scalable. Intuitively, when a frame is scalable, the frame vectors spread out in the space, which makes its minimal ellipsoid to be the unit ball. But when a frame gets more and more non-scalable, the frame vectors tend to bundle in one place, and thus produce a very "flat" ellipsoid with small volume. In this section, we formalize this intuition, and establish that V_Φ is just as suitable as D_Φ in quantifying how scalable a frame is.

We first consider the 2-dimensional case, where there is a straightforward characterization of scalability: $\Phi = \{\varphi_i\}_{i=1}^{M}$ is a scalable frame of $\mathbb{R}^2$ if and only if the smallest double cone (with apex at origin) containing all the frame vectors of Φ_{Sym} has an apex angle greater than or equal to $\pi/2$. This is essentially proved in [17, Corollary 3.8]; See also Remark 4.2 (b).

Example 3.2. Given $\Phi \in \mathcal{F}_u(M, 2)$, suppose $\varphi_1, \varphi_2 \in \Phi_{\mathrm{Sym}}$ generate the smallest cone containing Φ_{Sym}. Without loss of generality, we assume $\varphi_1 = (\cos\theta, \sin\theta)$ and $\varphi_2 = (\cos\theta, -\sin\theta)$, where 2θ is the apex angle. We have $E_\Phi = E_{\{\varphi_1, \varphi_2\}}$, and this ellipsoid is determined by the solution of the following problem:

$$\min_{a,b} ab \qquad \text{s.t.} \qquad \frac{\cos^2\theta}{a^2} + \frac{\sin^2\theta}{b^2} = 1.$$

The solution is $a = \sqrt{2}\cos\theta$, $b = \sqrt{2}\sin\theta$. So in this case,

$$X^{-1} = \begin{pmatrix} 2\cos^2\theta & 0 \\ 0 & 2\sin^2\theta \end{pmatrix} = \varphi_1\varphi_1^T + \varphi_2\varphi_2^T,$$

and $V_\Phi = \det(X^{-1/2}) = \sin 2\theta$.

Now let us calculate D_Φ. Since all vectors of Φ_{Sym} are contained in the cone $\{\pm(a\varphi_1 + b\varphi_2), a, b \geq 0\}$, any φ_i can be represented as $\varphi_i = c\varphi_1 + d\varphi_2$ with $cd \geq 0$. Thus $\varphi_i\varphi_i^T = c^2\varphi_1\varphi_1^T + d^2\varphi_2\varphi_2^T + cd(\varphi_1\varphi_2^T + \varphi_2\varphi_1^T)$. Therefore, the Frobenius norm minimization problem becomes

$$\min_{a,b,c\geq 0} \left\| a\varphi_1\varphi_1^T + b\varphi_2\varphi_2^T + c(\varphi_1\varphi_2^T + \varphi_2\varphi_1^T) - I \right\|_F.$$

The solution of this problem is $a = b = \frac{2}{3+\cos 4\theta}$, $c = 0$, and thus

$$D_\Phi^2 = 2 - 2a = 2 - \frac{2}{2 - V_\Phi^2}.$$

So, as V_Φ is approaching 1, D_Φ is approaching 0, and vice versa.

In Example 3.2 it is shown that in the 2-dimensional case, V_Φ is a function of D_Φ. However, in general V_Φ is no longer uniquely determined by D_Φ but falls into a range defined by D_Φ as the following theorem indicates. But the key point here is that it still remains true that D_Φ approaches zero if and only if the volume ratio tends to one.

Theorem 3.3. *Let $\Phi = \{\varphi_i\}_{i=1}^M \in \mathcal{F}_u(M, N)$, then*

$$(3.3) \qquad \frac{N(1 - D_\Phi^2)}{N - D_\Phi^2} \leq V_\Phi^{4/N} \leq \frac{N(N - 1 - D_\Phi^2)}{(N - 1)(N - D_\Phi^2)} \leq 1,$$

where the leftmost inequality requires $D_\Phi < 1$. Consequently, $V_\Phi \to 1$ is equivalent to $D_\Phi \to 0$.

Proof. The rightmost inequality is clear. Let us prove the upper bound on $V_\Phi^{4/N}$ in (3.3). For this, let $E_\Phi = E(X)$ be the minimal ellipsoid of Φ, and let $\{\lambda_i\}_{i=1}^N$ be the eigenvalues of $X^{-1} = \sum_{i=1}^M \rho_i\varphi_i\varphi_i^T$. For any $\alpha > 0$, we have

$$D_\Phi^2 \leq \left\| \sum_{i=1}^M \alpha\rho_i\varphi_i\varphi_i^T - I \right\|_F^2 = \|\alpha X^{-1} - I\|_F^2 = \sum_{i=1}^N (\alpha\lambda_i - 1)^2 = \alpha^2 \sum_{i=1}^N \lambda_i^2 - 2\alpha \sum_{i=1}^N \lambda_i + N.$$

Therefore, by (2.17),

$$(3.4) \qquad D_\Phi^2 \leq \min_{\alpha>0}\left(\alpha^2 \sum_{i=1}^N \lambda_i^2 - 2\alpha \sum_{i=1}^N \lambda_i + N \right) = N - \frac{N^2}{\sum_{i=1}^N \lambda_i^2}.$$

We use (2.16) and (3.2) to estimate $\sum_{i=1}^N \lambda_i^2$:

$$\sum_{i=1}^N \lambda_i^2 = \left(\sum_{i=1}^N \lambda_i \right)^2 - 2\sum_{i<j} \lambda_i\lambda_j = N^2 - 2\sum_{i<j} \lambda_i\lambda_j$$

$$(3.5) \qquad \leq N^2 - N(N-1)\prod_{i=1}^N \lambda_i^{2/N} = N^2 - N(N-1)V_\Phi^{4/N}.$$

Plugging (3.5) in (3.4) and solving it for $V_\Phi^{4/N}$ yields the upper bound in (3.3).

For the lower bound, let $\hat{C} = \text{diag}\{c_i\}_{i=1}^M$ be a minimizer of (2.3). Then $D_\Phi = \|\Phi\hat{C}\Phi^T - I\|_F$. Moreover,

$$\text{Tr}(\Phi\hat{C}\Phi^T X) = \sum_{i=1}^M c_i \varphi_i^T X \varphi_i \leq \sum_{i=1}^M c_i.$$

The last inequality holds due to (2.13). Therefore,

$$
\begin{aligned}
\text{Tr}(X) &= \text{Tr}(\Phi\hat{C}\Phi^T X) - \text{Tr}((\Phi\hat{C}\Phi^T - I)X) \\
&= \text{Tr}(\Phi\hat{C}\Phi^T X) - \text{Tr}(\Phi\hat{C}\Phi^T - I) - \text{Tr}((\Phi\hat{C}\Phi^T - I)(X - I)) \\
&\leq \sum_{i=1}^M c_i - \left(\sum_{i=1}^M c_i - N\right) - \text{Tr}((\Phi\hat{C}\Phi^T - I)(X - I)) \\
&\leq N + \|\Phi\hat{C}\Phi^T - I\|_F \|X - I\|_F \\
&= N + D_\Phi \sqrt{\sum_{i=1}^N \left(\lambda_i^{-1} - 1\right)^2} \\
&= N + D_\Phi \sqrt{\left(\sum_{i=1}^N \lambda_i^{-1}\right)^2 - 2\sum_{i<j} \lambda_i^{-1}\lambda_j^{-1} - 2\sum_{i=1}^N \lambda_i^{-1} + N} \\
&\leq N + D_\Phi \sqrt{\text{Tr}^2(X) - N(N-1)V_\Phi^{-4/N} - 2\text{Tr}(X) + N},
\end{aligned}
$$

(3.6)

where the last inequality is due to (3.2) with $a_i = \lambda_i^{-1}$, and (2.16). By (3.1),

$$\text{(3.7)} \qquad\qquad \text{Tr}(X) = \sum_{i=1}^N \lambda_i^{-1} \geq \frac{N}{\prod_{i=1}^N \lambda_i^{1/N}} = NV_\Phi^{-2/N} \geq N.$$

Now, we subtract N on both sides of (3.6), square both sides, and obtain

$$(\text{Tr}(X) - N)^2 \leq D_\Phi^2 \left(\text{Tr}^2(X) - 2\text{Tr}(X) + N - N(N-1)V_\Phi^{-4/N}\right).$$

The latter inequality is equivalent to

$$\left(\text{Tr}(X) - \frac{N - D_\Phi^2}{1 - D_\Phi^2}\right)^2 \leq \frac{D_\Phi^2(N-1)}{(1 - D_\Phi^2)^2} \left(N - D_\Phi^2 - (1 - D_\Phi^2)NV_\Phi^{-4/N}\right).$$

This proves that

$$N - D_\Phi^2 - (1 - D_\Phi^2)NV_\Phi^{-4/N} \geq 0,$$

which is equivalent to the leftmost inequality in (3.3).

$\square$

3.2. **Algorithms and Numerical Experiments.** The computation in (2.4) shows that D_Φ can be computed via Quadratic Programming (QP). As is well known, this problem can be solved by many well developed methods, e.g., Active-Set, Conjugate Gradient, Interior point.

The minimal ellipsoid problem has been studied for half a century. For a given convex body K and a small quantity $\eta > 0$, a fast algorithm to compute an ellipsoid $E \supseteq K$ with

$$\text{Vol}(E) \le (1 + \eta)\,\text{Vol}(\text{Minimal ellipsoid}(K))$$

is the Khachiyan's barycentric coordinate descent algorithm [14], which needs a total of $O(M^{3.5}\ln(M\eta^{-1}))$ operations. For the case $N \ll M$, Kumar and Yildirim [15] improved this algorithm using core sets and reduced the complexity to $O(MN^3\eta^{-1})$.

For all numerical simulations in this paper, we use Khachiyan's method to compute minimal ellipsoids and the active set method to solve the quadratic programming in (2.3). As expected, we have observed a much faster computational speed of the latter, especially when the problem grows large in size.

Figure 1 shows the values of D_Φ and V_Φ for randomly generated frames in $\mathcal{F}_u(M, 4)$ with $M = 6, 11, 15$, and 20. In each plot, we generated 1000 frames, where each column of the frame is chosen uniformly at random from the unit sphere, and calculated both V_Φ and D_Φ.

As expected, for a fixed D_Φ, we see a range of V_Φ. For a direct comparison, we plotted the two bounds from (3.3). The lower bound from (3.3) is quite optimal based on the figure.

On the other hand, as M increases, we observe a change of concentration of the points from scattering around to being heavily distributed around $D_\Phi = 0$: "the scalable region". Indeed, as shown by Theorem 4.9 in Section 4, the threshold for having positive probability of scalable frames in dimension $N = 4$ is $N(N+1)/2 = 10$. Therefore, we have considerably many points achieving $D_\Phi = 0$ for $M = 11, 15, 20$. In fact, about 60% of these 1000 frames in $\mathcal{F}(4, 20)$ are scalable (up to a machine error).

This suggests that the two measures of scalability, the distance between D_Φ and 0 and the distance between V_Φ and 1, though closely related, are indeed different in the sense that there is no one-to-one correspondence between them. An advantage of using D_Φ to measure scalability lies in the fact it is more naturally related to the notion of $m-$scalability (defined in [16]) and is more efficient to compute. By contrast, V_Φ is a more intuitive measure of scalability from a geometric point of view.

3.3. **Comparison of the Measures D_Φ and V_Φ with d_Φ.** The distance of a frame $\Phi \in \mathcal{F}_u(M, N)$ to the set of scalable frames is the most intuitive and natural measure of scalability. The next theorem shows that the practically more accessible measures D_Φ and V_Φ are equivalent to d_Φ in the sense that d_Φ tends to zero if and only if the same holds for D_Φ or $1 - V_\Phi$.
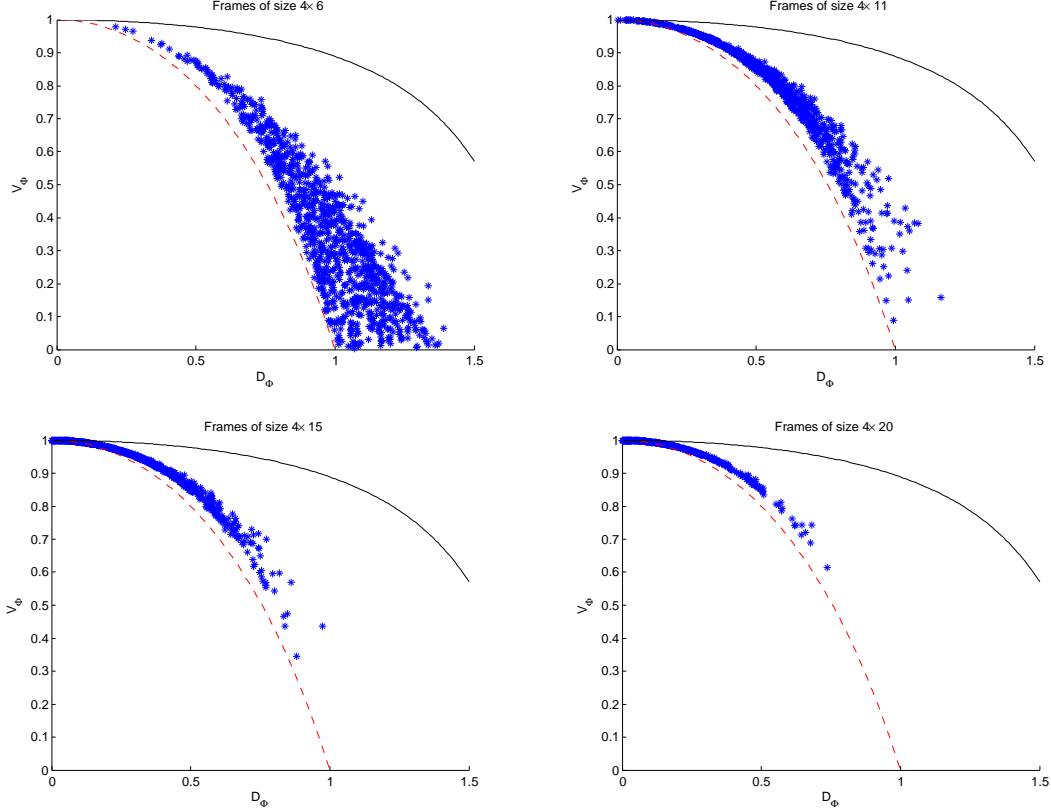
FIGURE 1. Relation between V_Φ and D_Φ for $\Phi \in \mathcal{F}_u(M, 4)$ with $M = 6, 11, 15, 20$. The solid line indicates the upper bound in (3.3), while the dash line indicates the lower bound.

Theorem 3.4. *Let* $\Phi \in \mathcal{F}_u(M, N)$ *and assume that* $d_\Phi < 1$. *Then with* $K := \min\{M, \frac{N(N+1)}{2}\}$ *and* $\omega := D_\Phi + \sqrt{K}$ *we have*

$$(3.8) \qquad \frac{D_\Phi}{\omega + \sqrt{\omega^2 - D_\Phi^2}} \le d_\Phi \le \sqrt{KN\left(1 - V_\Phi^{2/N}\right)}.$$

Consequently, with the help of Theorem 3.3, *we can bound* d_Φ *below and above by expressions of* D_Φ *or expressions of* V_Φ.

Proof. Following the same notation as in the proof of Theorem 3.3, let $\lambda_1, \ldots, \lambda_N$ be the eigenvalues of $X^{-1} = \sum_{i=1}^{M} \rho_i \varphi_i \varphi_i^T$. Furthermore, let $J = \{i : \rho_i > 0\}$. By Remark 2.10, $|J| \le K$. Define a frame $\widetilde{\Phi} = \{\tilde{\varphi}_i\}_{i=1}^M$ by

$$(3.9) \qquad \tilde{\varphi}_i := \begin{cases} V_\Phi^{1/N} X^{1/2} \varphi_i & \text{if } i \in J \\ \varphi_i & \text{if } i \notin J. \end{cases}$$

Note that $\widetilde{\Phi}$ is scalable, and, moreover, $\|X^{1/2}\varphi_i\|_2 = 1$ for $i \in J$ by (2.14). So,

$$\|\Phi - \widetilde{\Phi}\|_F^2 = \sum_{i \in J} \|\varphi_i - V_\Phi^{1/N} X^{1/2} \varphi_i\|_2^2 = \sum_{i \in J} \|(X^{-1/2} - V_\Phi^{1/N} I) X^{1/2} \varphi_i\|_2^2$$

$$\leq \|X^{-1/2} - V_\Phi^{1/N} I\|_F^2 \sum_{i \in J} \|X^{1/2} \varphi_i\|_2^2 \leq K \sum_{j=1}^{N} \left(\lambda_j^{1/2} - V_\Phi^{1/N}\right)^2$$

$$= K \left(N + V_\Phi^{2/N} N - 2 V_\Phi^{1/N} \sum_{j=1}^{N} \lambda_j^{1/2}\right)$$

$$= K N \left(1 + V_\Phi^{2/N} - 2 V_\Phi^{1/N} \frac{1}{N} \sum_{j=1}^{N} \lambda_j^{1/2}\right)$$

(3.10)
$$\leq K N \left(1 + V_\Phi^{2/N} - 2 V_\Phi^{2/N}\right) = K N \left(1 - V_\Phi^{2/N}\right)^2.$$

As $d_\Phi \leq \|\Phi - \widetilde{\Phi}\|_F$, this proves the right-hand side of (3.8).

Let $\hat{\Phi}$ be a minimizer of (2.1) (which exists due to Proposition 2.1 and has non-zero columns by Remark 2.2). Since $\hat{\Phi}$ is scalable, there exists a non-negative diagonal matrix $S = \mathrm{diag}(s_i)_{i=1}^{M}$ such that $\hat{\Phi} S \hat{\Phi}^T = I$. Again by Remark 2.10, we may assume that at most K of the s_i are non-zero. We then have

$$\Phi S \Phi^T - I = \Phi S \Phi^T - \hat{\Phi} S \hat{\Phi}^T = \sum_{i=1}^{M} s_i \left[\varphi_i(\varphi_i^T - \hat{\varphi}_i^T) + (\varphi_i - \hat{\varphi}_i)\hat{\varphi}_i^T\right],$$

and therefore, as $\|\hat{\varphi}_i\|_2 \leq 1$ (see Lemma 2.3 (ii)),

$$D_\Phi \leq \|\Phi S \Phi^T - I\|_F \leq \sum_{i=1}^{M} s_i \left(\|\varphi_i - \hat{\varphi}_i\|_2 + \|\varphi_i - \hat{\varphi}_i\|_2 \|\hat{\varphi}_i\|_2\right)$$

$$\leq 2 \sum_{i=1}^{M} s_i \|\varphi_i - \hat{\varphi}_i\|_2 \leq 2 \left(\sum_{i=1}^{M} s_i^2\right)^{1/2} \left(\sum_{i=1}^{M} \|\varphi_i - \hat{\varphi}_i\|_2^2\right)^{1/2}$$

$$\leq 2 \left(K \max_i s_i^2\right)^{1/2} \|\Phi - \hat{\Phi}\|_F = 2\sqrt{K} \left(\max_i s_i\right) d_\Phi$$

Now, for each $i \in \{1, \ldots, M\}$ we have

$$s_i = \frac{\hat{\varphi}_i^T s_i \hat{\varphi}_i \hat{\varphi}_i^T \hat{\varphi}_i}{\|\hat{\varphi}_i\|_2^4} \leq \sum_{k=1}^{M} \frac{\hat{\varphi}_i^T s_k \hat{\varphi}_k \hat{\varphi}_k^T \hat{\varphi}_i}{\|\hat{\varphi}_i\|_2^4} = \frac{\hat{\varphi}_i^T \hat{\Phi} S \hat{\Phi}^T \hat{\varphi}_i}{\|\hat{\varphi}_i\|_2^4} = \frac{1}{\|\hat{\varphi}_i\|_2^2} \leq \frac{1}{(1 - d_\Phi)^2},$$

where the last inequality follows from the triangle inequality. This gives

(3.11)
$$D_\Phi \leq \frac{2\sqrt{K} d_\Phi}{(1 - d_\Phi)^2}.$$

Solving for d_Φ in the last inequality leads to the left hand side of (3.8).

□

We conclude this section by a theorem on approximating unit norm frames by scalable frames.

Theorem 3.5 (Approximation by scalable frames)**.** *Let $\Phi \in \mathcal{F}_u(M, N)$ and assume that $d_\Phi \leq \frac{1}{2}(1 + \sqrt{K})^{-1}$. Let $\hat{\Phi}$ be a minimizer of (2.1), and let $E_\Phi = E(X)$ be the minimal ellipsoid of Φ, where $X^{-1} = \sum_{i=1}^M \rho_i \varphi_i \varphi_i^T$. Then the scalable frame $\widetilde{\Phi} = \{\tilde{\varphi}_i\}_{i=1}^M$ defined in (3.9) is a good approximation to Φ in the following sense:*

$$(3.12) \quad \|\widetilde{\Phi} - \Phi\|_F \leq \sqrt{KN} \left(1 - \sqrt{N \frac{(1 - d_\Phi)^4 - 4Kd_\Phi^2}{N(1 - d_\Phi)^4 - 4Kd_\Phi^2}}\right)^{1/2} = K\sqrt{N} O(d_\Phi),$$

where $K = \min\{M, \frac{N(N+1)}{2}\}$.

Proof. We extend the estimate (3.10) with the help of the leftmost inequality of (3.3):

$$(3.13) \qquad \|\Phi - \widetilde{\Phi}\|_F^2 \leq KN \left(1 - \sqrt{N \frac{1 - D_\Phi^2}{N - D_\Phi^2}}\right).$$

Since the right-hand side of (3.13) is an increasing function of D_Φ on $[0, 1]$, we substitute (3.11) into (3.13) and obtain the left hand side of (3.12), where we need the requirement on d_Φ so that $D_\Phi < 1$. $\qquad \square$

4. Probability of having scalable frames

This section aims to estimate the probability $P_{M,N}$ of unit norm frames to be scalable when the frame vectors are drawn independently and uniformly from the unit sphere $\mathbb{S}^{N-1} \subset \mathbb{R}^N$. This is in a sense equivalent to estimating the "size" of $\mathcal{SC}(M, N)$ in $\mathcal{F}_u(M, N)$.

The basic idea is to use the characterization of scalability in terms of the minimum volume ellipsoids through John's theorem, see Theorem 2.11. From this geometric point of view, we derive new and relatively simple conditions for scalability and non-scalability (Theorem 4.1). These conditions are the key tools we use to estimate the probability $P_{M,N}$.

4.1. **Necessary and Sufficient Conditions for Scalability.** The following theorem plays a crucial role in the proof of our main theorem on the probability of having scalable frames in Subsection 4.2. However, it is also of independent interest.

Theorem 4.1. *Let $\Phi \in \mathcal{F}_u(M, N)$. Then the following hold:*

(a) (*A necessary condition for scalability*) *If Φ is scalable, then*

$$(4.1) \qquad \min_{\|d\|_2 = 1} \max_i |\langle d, \varphi_i \rangle| \geq \frac{1}{\sqrt{N}}.$$

(b) (*A sufficient condition for scalability*) If

$$\min_{\|d\|_2=1} \max_i |\langle d, \varphi_i \rangle| \geq \sqrt{\frac{N-1}{N}}, \tag{4.2}$$

then Φ is scalable.

Proof. (a). We will use the following fact: if E_K is the minimal ellipsoid of a convex body $K \subset \mathbb{R}^N$ which is symmetric about the origin, then $\frac{1}{\sqrt{N}} E_K \subset K$, see [12, Theorem 12.11]. If Φ is scalable, then the unit ball is the minimal ellipsoid of the convex hull $\mathrm{co}(\Phi_{\mathrm{Sym}})$ of Φ_{Sym}. Therefore, $\frac{1}{\sqrt{N}} B \subset \mathrm{co}(\Phi_{\mathrm{Sym}})$. And as a continuous convex function on a compact convex set attains its maximum at an extreme point of this set (see, e.g., [19, Theorem 3.4.7]), we conclude that for each $d \in \mathbb{S}^{N-1}$ we have

$$\frac{1}{\sqrt{N}} = \max_{x \in \frac{1}{\sqrt{N}} B} |\langle d, x \rangle| \leq \max_{x \in \mathrm{co}(\Phi_{\mathrm{Sym}})} |\langle d, x \rangle| \leq \max_i |\langle d, \varphi_i \rangle|.$$

(b). Let $E_\Phi = E(X)$ be the minimal ellipsoid of Φ. With a unitary transformation, we can assume $X^{-1/2} = \mathrm{diag}(a_i)_{i=1}^N$. Towards a contradiction, suppose that (4.2) holds, but that Φ is not scalable. Then, by Theorem 2.11, $a_1 \leq a_2 \leq \ldots \leq a_N$ with $a_1 < a_N$. Take any frame vector $\varphi = (x_1, x_2, \ldots, x_N)^T$ from Φ. It satisfies $\sum_{i=1}^N \frac{x_i^2}{a_i^2} = \langle X\varphi, \varphi \rangle \leq 1$ and $\sum_{i=1}^N x_i^2 = 1$, which implies

$$\sum_{i=1}^{N-1} x_i^2 \left(\frac{1}{a_i^2} - \frac{1}{a_N^2} \right) \leq 1 - \frac{1}{a_N^2}.$$

Hence, setting $\rho = (1 - \frac{1}{a_N^2})/(\frac{1}{a_1^2} - \frac{1}{a_N^2})$, we have $x_1^2 \leq \rho$. We claim that

$$\rho < \frac{N-1}{N}. \tag{4.3}$$

Then we choose $d = (1, 0, \ldots, 0)^T$ and find that $|\langle d, \varphi \rangle| = |x_1| < \sqrt{\frac{N-1}{N}}$ for each $\varphi \in \Phi$ which contradicts the assumption.

Proving (4.3) is equivalent to proving $\frac{1}{a_N^2} + \frac{N-1}{a_1^2} > N$, which is true because

$$\frac{1}{a_N^2} + \frac{N-1}{a_1^2} \geq \sum_{i=1}^N \frac{1}{a_i^2} > \frac{N^2}{\sum_{i=1}^N a_i^2} = N,$$

where we have used (2.17) and (3.1) (in which equality holds if and only if $a_1 = \ldots = a_N$). $\qquad\square$

Remark 4.2. (a) Another necessary condition for scalability was proved in [10, Theorem 3.1]. We wish to point out that this necessary condition is unrelated to the one given in part (a) of the previous theorem in the sense that neither of these conditions implies the other.

(b) When the dimension $N = 2$, Theorem 4.1 gives a necessary and sufficient condition for a frame to be scalable. This condition can be easily interpreted in terms of cones as already mentioned before: $\{\varphi_i\}_{i=1}^M$ is a scalable frame for $\mathbb{R}^2$ if and

only if every double cone with apex at origin and containing Φ_{Sym} has an apex angle greater than or equal to $\pi/2$.

(c) For a general N, the gap between these two conditions is large. However, this gap cannot be improved. Theorem 4.1(a) is tight in the sense that we cannot replace $1/\sqrt{N}$ by a bigger constant. This is because an orthonormal basis reaches this constant. The sufficient condition is also optimal in the sense that $\sqrt{(N-1)/N}$ cannot be replaced by a smaller number. This requires some more analysis as shown below.

Proposition 4.3. *For any small $\varepsilon > 0$ and any $N \in \mathbb{N}$, there exists a unit norm frame Φ for $\mathbb{R}^N$, such that*

$$\min_{\|d\|_2 = 1} \max_i |\langle d, \varphi_i \rangle| \geq \sqrt{\frac{N-1}{N}} - 2\varepsilon,$$

but Φ is not scalable.

Proof. Pick an ellipsoid $E(X)$ with $X^{-1} = \text{diag}(a_1^2, a_2^2, \ldots, a_{N-1}^2, a_N^2)$, where $a_1^2 = a_2^2 = \ldots = a_{N-1}^2 = \frac{N-1-\varepsilon}{N-1}$, and $a_N^2 = 1 + \varepsilon$. By Theorem 2.13, there exists a (non-scalable) frame $\Phi \in \mathcal{F}_u(M, N)$ whose minimal ellipsoid is $E(X)$.

Then for any $x \in E(X) \cap \mathbb{S}^{N-1}$, we have

$$1 \geq \sum_{i=1}^{N-1} \frac{x_i^2}{a_i^2} + \frac{x_N^2}{a_N^2} = \frac{(N-1)(1 - x_N^2)}{N-1-\varepsilon} + \frac{x_N^2}{1+\varepsilon},$$

which implies that

$$x_N^2 \geq \frac{1+\varepsilon}{N}.$$

Now for any $d = (d_1, d_2, \ldots, d_N) \in \mathbb{S}^{N-1}$, if $d_N^2 < \frac{1+\varepsilon}{N}$, then let

$$x_0 = \sqrt{1 - \frac{1+\varepsilon}{N}} \frac{\tilde{d}}{\|\tilde{d}\|} + \text{sign}(d_N) \left(0, 0, ..., 0, \sqrt{\frac{1+\varepsilon}{N}} \right),$$

where $\tilde{d} = (d_1, d_2, ..., d_{N-1}, 0)$. It is easy to verify that $x_0 \in E(X) \cap \mathbb{S}^{N-1}$ and that $\langle x_0, d \rangle \geq \sqrt{\frac{N-1-\varepsilon}{N}}$. If $d_N^2 \geq \frac{1+\varepsilon}{N}$, then let $x_0 = d$. It is again easy to check $x_0 \in E(X) \cap \mathbb{S}^{N-1}$ and $\langle x_0, d \rangle = 1$. In summary, for any $d \in \mathbb{S}^{N-1}$, there exists an $x_0 \in E(X) \cap \mathbb{S}^{N-1}$, such that $\langle x_0, d \rangle \geq \sqrt{\frac{N-1-\varepsilon}{N}}$.

We add vectors from the set $E(X) \cap \mathbb{S}^{N-1}$ to Φ such that the frame vectors are dense enough to form an ε-ball of $E(X) \cap \mathbb{S}^{N-1}$, i.e., for any $x \in E(X) \cap \mathbb{S}^{N-1}$, there exists a $\varphi_i \in E(X) \cap \mathbb{S}^{N-1}$, such that $\|\varphi_i - x\|_2 \leq \varepsilon$. Notice this new frame has the same minimal ellipsoid. With this construction, for any $d \in \mathbb{S}^{N-1}$, we can find a frame vector φ_i such that $\langle \varphi_i, d \rangle = \langle x, d \rangle + \langle \varphi_i - x, d \rangle \geq \sqrt{\frac{N-1-\varepsilon}{N}} - \varepsilon \geq \sqrt{\frac{N-1}{N}} - 2\varepsilon$ provided that ε is small enough. $\square$

In Remark 4.2(b), we mentioned that (4.1) is necessary and sufficient for scalability if $N = 2$. In the following, we shall show that the same holds if $M = N$:

Theorem 4.4. *For $\Phi \in \mathcal{F}_u(N, N)$, the following statements are equivalent.*

(i) Φ *is scalable.*

(ii) Φ *is unitary.*

(iii) $\min_{\|d\|_2=1} \max_i |\langle d, \varphi_i \rangle| \geq \frac{1}{\sqrt{N}}.$

In order to prove Theorem 4.4, we need the following lemma.

Lemma 4.5. *Let $\Phi \in \mathbb{R}^{N \times N}$ be a non-unitary invertible matrix with unit norm columns. Then there exists a vector $d \in \mathbb{R}^N$ with $\|d\|_2 > 1$ and a vector $a \in \mathbb{R}^N$ with $|a_i| = 1/\sqrt{N}$ for all $i = 1, \ldots, N$, such that $\Phi^T d = a$.*

Proof. Let $\{b_i\}_{i=1}^N$ be a sequence with each entry being a Bernoulli random variable, $\Psi = \mathrm{diag}(b_i)_{i=1}^N$, and $g = \frac{1}{\sqrt{N}}(1, \ldots, 1)^T$. Suppose d_Ψ is the solution to

(4.4) $$\Phi^T d_\Psi = \Psi g.$$

Let $\Phi^T = U\Sigma V^T$ be the singular value decomposition of Φ^T, where $\Sigma = \mathrm{diag}(\sigma_i)$. Observe that

$$\sqrt{N} = \|\Phi\|_F = \|\Sigma\|_F = \sqrt{\sum_{i=1}^N \sigma_i^2}.$$

Hence, from (3.1) we obtain

$$\sum_{i=1}^N \sigma_i^{-2} \geq \frac{N^2}{\sum_{i=1}^N \sigma_i^2} = N.$$

On the other hand, from (4.4) we have

$$V^T d_\Psi = \Sigma^{-1} U^T \Psi g.$$

Next, we calculate the expectation $\mathbb{E}\|d_\Psi\|^2$. If it is greater than 1, then there must exist one instance of d_Ψ with norm greater than 1, which makes the lemma hold. As $\mathbb{E}(b_i b_j) = \delta_{ij}$, we obtain

$$\mathbb{E}\|d_\Psi\|^2 = \mathbb{E}\|V^T d_\Psi\|^2 = \mathbb{E}\|\Sigma^{-1} U^T \Psi g\|^2$$

$$= \frac{1}{N} \mathbb{E}\left(\sum_i \sigma_i^{-2} \left(\sum_j u_{ji} b_j \right)^2 \right)$$

$$= \frac{1}{N} \sum_i \sigma_i^{-2} \mathbb{E}\left(\sum_j u_{ji}^2 + \sum_j \sum_{k, k \neq j} u_{ji} u_{ki} b_j b_k \right)$$

$$= \frac{1}{N} \sum_i \sigma_i^{-2} \sum_j u_{ji}^2 = \frac{1}{N} \sum_i \sigma_i^{-2} \geq 1,$$

while for the last inequality, equality holds only when all σ_i are equal, i.e., Φ is unitary, which is ruled out by our assumption. Therefore the last inequality is strict. $\square$

Proof of Theorem 4.4. The equivalence (i)$\Leftrightarrow$(ii) is easy to see and follows from, e.g., [17, Corollary 2.8]. Moreover, (i)$\Rightarrow$(iii) is a direct consequence of Theorem 4.1(a). It remains to prove that (iii) implies (i). For this, we prove the contraposition. Suppose that Φ is not scalable. Then Φ is not unitary, and Lemma 4.5 implies the existence of $d' \in \mathbb{R}^N$, $\|d'\|_2 > 1$, such that $|\langle d', \varphi_i \rangle| = 1/\sqrt{N}$ for all $i = 1, \ldots, N$. Hence, with $d = d'/\|d'\|_2$ we have $|\langle d, \varphi_i \rangle| = (\|d'\|_2\sqrt{N})^{-1} < 1/\sqrt{N}$ for all $i = 1, \ldots, N$. That is, (iii) does not hold, and the theorem is proved. $\square$

4.2. **Estimation of the probability.** With the help of Theorem 4.1, we now estimate the probability for a frame to be scalable when its vectors are drawn independently and uniformly from $\mathbb{S}^{N-1}$. First of all, it is easy to see the probability strictly increases as M increases. Secondly, $\varphi_i \varphi_i^T \in \mathrm{Sym}_N$, where

$$\mathrm{Sym}_N := \left\{ A \in \mathbb{R}^{N \times N} : A = A^T \right\},$$

which is a vector space of dimension $\frac{N(N+1)}{2}$. By (1.1), being scalable requires I to be in the positive cone generated by $\{\varphi_i \varphi_i^T\}_{i=1}^M$. If $M < \frac{N(N+1)}{2}$, then this set cannot be a basis of Sym_N, so the chance for any symmetric matrix to be in the span of $\{\varphi_i \varphi_i^T\}_{i=1}^M$ is minimal, which makes it even more difficult for I to be in positive cone generated by this set. Therefore we expect the probability to be 0 when $M < \frac{N(N+1)}{2}$. Finally, as $M \to \infty$, we expect the probability of frames to be scalable to approach 1.

Let us first consider the case $N = 2$ for which the probability $P_{2,M}$ can be explicitly computed.

Example 4.6. If vectors $\varphi_1, \ldots, \varphi_M$ are drawn independently and uniformly from $\mathbb{S}^1$, then the probability of $\{\varphi_i\}_{i=1}^M$ to be a scalable frame in $\mathcal{F}_u(M, 2)$ is given by

$$P_{M,2} = 1 - \frac{M}{2^{M-1}}, \quad M \geq 2.$$

Proof. First of all, define the angle of a vector v as the angle between v and positive x-axis, counterclockwise. Among all the double cones that cover all the vectors in Φ_{Sym}, let P_Φ be the one with the smallest apex angle α. It is known that Φ is scalable if and only if $\alpha \geq \pi/2$. Let φ_Φ be the "right boundary" of P_Φ. To be rigorous, Let φ_Φ be the vector with angle $\beta_0 \in [0, \pi)$ such that for β in some neighborhood of β_0 we have $(\cos\beta, \sin\beta)^T \in P_\Phi$ if $\beta > \beta_0$ and $(\cos\beta, \sin\beta)^T \notin P_\Phi$ if $\beta < \beta_0$. For fixed $i \in \{1, \ldots, M\}$ we then have

$$\Pr(\Phi \text{ not scalable and } \varphi_\Phi = \pm\varphi_i)$$

$$= \frac{1}{2\pi} \int_0^{2\pi} \Pr\left(\Phi \text{ not scalable and } \varphi_\Phi = \pm\varphi_i \mid \angle\varphi_i = \beta\right) d\beta$$

$$= \frac{1}{2\pi} \int_0^{2\pi} \frac{1}{2^{M-1}} d\beta = \frac{1}{2^{M-1}}.$$

Now, it follows that $\Pr(\Phi \text{ is not scalable}) = \sum_i \Pr(\Phi \text{ not scalable and } \varphi_\Phi = \pm\varphi_i) = M/2^{M-1}$. $\square$

We can see in $\mathbb{R}^2$, as the number of frame vectors increases, the probability $P_{M,2}$ increases as well, starting from zero and eventually approaching 1. The critical point where the probability turns from zero to positive is $M = 3 = \frac{N(N+1)}{2}$, which meets our expectation. We will show that this is true for arbitrary dimension, and provide an estimate for the probability of frames being scalable. The following lemma completes the series of preparatory statements for the proof of our main theorem.

Lemma 4.7. *If* $\Phi = \{\varphi_i\}_{i=1}^M$ *is a strictly scalable frame for* $\mathbb{R}^N$ *and* $\{\varphi_i\varphi_i^T\}_{i=1}^M$ *is a frame for* Sym_N, *then there exists* $\varepsilon > 0$ *such that any frame* Ψ *satisfying* $\|\Psi - \Phi\|_F < \varepsilon$ *is strictly scalable.*

Proof. Let A be the lower frame bound of $\{\varphi_i\varphi_i^T\}_{i=1}^M$, where Sym_N is endowed with the Frobenius norm. Moreover, by $F : \mathrm{Diag}_M \to \mathrm{Sym}_N$ denote the synthesis operator of $\{\varphi_i\varphi_i^T\}_{i=1}^M$, where Diag_M denotes the space of all diagonal matrices in Sym_M. Then $FD = \Phi D \Phi^T$, $D \in \mathrm{Diag}_M$. Since Φ is strictly scalable, there exists a positive definite $D \in \mathrm{Diag}_M$ such that $FD = I$.

Let $\delta > 0$ be so small that whenever $\Delta \in \mathrm{Diag}_M$ with $\|\Delta\|_F \leq \delta$, we have that $D + \Delta$ remains positive definite. Moreover, let $\varepsilon > 0$ be so small that

$$\tau := (\varepsilon + 2\|\Phi\|_F)\varepsilon \leq \max\left\{ \frac{\sqrt{A}}{2}, \frac{\delta A}{2(\sqrt{A} + 2\|F\|_{\mathrm{op}})\|D\|_F} \right\}.$$

Now, let $\Psi = \{\psi_i\}_{i=1}^M \subset \mathbb{R}^N$ be such that $\|\Phi - \Psi\|_F < \varepsilon$. By $G : \mathrm{Diag}_M \to \mathrm{Sym}_N$ denote the synthesis operator of $\{\psi_i\psi_i^T\}_{i=1}^M$. We can see that $\|F - G\|_{\mathrm{op}} \leq \tau$, since for any diagonal matrix C,

$$\|FC - GC\|_F = \|\Phi C \Phi^T - \Psi C \Psi^T\|_F \leq \|\Phi C(\Phi^T - \Psi^T)\|_F + \|(\Phi - \Psi)C\Psi^T\|_F$$
$$\leq \epsilon(\|\Phi\|_F + \|\Psi\|_F)\|C\|_F \leq \epsilon(\epsilon + 2\|\Phi\|_F)\|C\|_F.$$

Hence, for $X \in \mathrm{Sym}_N$ we have

$$\|G^*X\|_F \geq \|F^*X\|_F - \|(F - G)^*X\|_F \geq (\sqrt{A} - \tau)\|X\|_F \geq (\sqrt{A}/2)\|X\|_F.$$

In particular, this implies that $\{\psi_i\psi_i^T\}_{i=1}^M$ is a frame for Sym_N, and $\langle GG^*X, X\rangle_F = \|G^*X\|_F^2 \geq (A/4)\|X\|_F^2$ yields $\|(GG^*)^{-1}\|_{\mathrm{op}} \leq 4/A$. Now, we define

$$\Delta := G^*(GG^*)^{-1}(F - G)D \in \mathrm{Diag}_M.$$

Then $G(D + \Delta) = GD + (F - G)D = FD = I$. Moreover,

$$\|\Delta\|_F \leq \|G\|_{\mathrm{op}}\|(GG^*)^{-1}\|_{\mathrm{op}}\|F - G\|_{\mathrm{op}}\|D\|_F \leq ((\sqrt{A}/2) + \|F\|_{\mathrm{op}})(4/A)\tau\|D\|_F \leq \delta,$$

so that $D + \Delta$ is positive definite. Consequently, Ψ is strictly scalable. $\square$

Remark 4.8. We mention that Lemma 4.7 implies that the set $\{\Phi \in SC_+(M, N) : \{\varphi_i\varphi_i^T\}_{i=1}^M$ is a frame$\}$ is open.

The statement and proof of the main theorem use the notion of spherical caps. We define $R_a^N(C)$ to be the spherical cap in $\mathbb{S}^N$ with angular radius a, centered at C, i.e.

$$R_a^N(C) = \left\{ x \in \mathbb{S}^N : \langle x, C\rangle \geq \cos(a) \right\}.$$
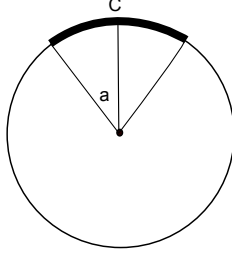
By A_a^N we denote the relative area of $R_a^N(C)$ (ratio of area of $R_a^N(C)$ and area of $\mathbb{S}^N$).

Theorem 4.9. *Given* $\Phi = \{\varphi_i\}_{i=1}^M \subset \mathbb{R}^N$, *where each vector* φ_i *is drawn independently and uniformly from* $\mathbb{S}^{N-1}$, *let* $P_{M,N}$ *denote the probability that* Φ *is scalable. Then the following holds:*

(i) *When* $M < \frac{N(N+1)}{2}$, $P_{M,N} = 0$

(ii) *When* $M \geq \frac{N(N+1)}{2}$, $P_{M,N} > 0$ *and*

$$C_N \left(1 - A_\alpha^{N-1}\right)^M \geq 1 - P_{M,N} \geq \left(1 - A_a^{N-1}\right)^{M-N},$$

where

$$\alpha = \frac{1}{2}\arccos\sqrt{\frac{N-1}{N}}, \quad a = \arccos\frac{1}{\sqrt{N}},$$

and where C_N *is the number of caps with angular radius* α *needed to cover* $\mathbb{S}^{N-1}$. *Consequently,* $\lim_{M\to\infty} P_{M,N} = 1$.

Proof. By μ_u we denote the uniform measure on $\mathbb{S}^{N-1}$ and by μ_G the Gaussian measure on $\mathbb{R}^N$. Furthermore, on $(\mathbb{S}^{N-1})^M$ and $(\mathbb{R}^N)^M$ define the product measures

$$\mu_u^k := \bigotimes_{j=1}^k \mu_u \quad \text{and} \quad \mu_G^k := \bigotimes_{j=1}^k \mu_G,$$

respectively. For a set $B \subset (\mathbb{S}^{N-1})^k$, $k \in \mathbb{N}$, we define

$$B' := \left\{(x_1, \ldots, x_k) \in (\mathbb{R}^N \setminus \{0\})^k : \left(\frac{x_1}{\|x_1\|_2}, \ldots, \frac{x_k}{\|x_k\|_2}\right) \in B\right\}.$$

Since $\mu_u(A) = \mu_G(A')$ for any $A \subset \mathbb{S}^{N-1}$, we have

(4.5) $$\mu_u^k(B) = \mu_G^k(B') \quad \text{for any } B \subset (\mathbb{S}^{N-1})^k.$$

(i). Set $K = N(N+1)/2$. It suffices to show $P_{M,N} = 0$ only for $M = K - 1$. For this, let

$$B := \left\{(\varphi_1, \ldots, \varphi_M) \in (\mathbb{S}^{N-1})^M : \{\varphi_1\varphi_1^T, \ldots, \varphi_M\varphi_M^T, I\} \text{ is linearly dependent}\right\}.$$

Then

$$B' = \left\{(\varphi_1, \ldots, \varphi_M) \in (\mathbb{R}^N \setminus \{0\})^M : \{\varphi_1\varphi_1^T, \ldots, \varphi_M\varphi_M^T, I\} \text{ is linearly dependent}\right\}.$$

This set, seen as a subset of $\mathbb{R}^{NM}$, is contained in the zero locus of a polynomial in the entries of the φ_i's. Therefore, the Lebesgue measure of B' is zero. But this shows

that $\mu_G^M(B') = 0$ since μ_G^M is absolutely continuous with respect to the Lebesgue measure. Consequently, we obtain

$$P_{M,N} = \mu_u^M(\{\Phi \in \mathcal{F}_u(M, N) : \Phi \text{ scalable}\}) \leq \mu_u^M(B) = \mu_G^M(B') = 0.$$

(ii). With Lemma 4.7, we only need to prove the existence of a strictly scalable unit norm frame Φ such that $\{\varphi_i \varphi_i^T\}_{i=1}^M$ spans Sym_N. For this, we note that by [4, Theorem 2.1], there exists a frame $V = \{v_i\}_{i=1}^M$ such that $\{v_i v_i^T\}_{i=1}^M$ spans Sym_N. Let S be its frame operator, and $\varphi_i = S^{-1/2} v_i$. Therefore $\Phi = \{\varphi_i\}$ is a tight frame, thus strictly scalable. It is also easy to check that the linear map $T : \text{Sym}_N \to \text{Sym}_N$, defined by $T(A) := S^{-1/2} A S^{-1/2}$, $A \in \text{Sym}_N$, is invertible and maps $v_i v_i^T$ to $\varphi_i \varphi_i^T$. Therefore, $\{\varphi_i \varphi_i^T\}_{i=1}^M$ also spans Sym_N. Finally, we normalize Φ to attain the desired frame.

For the estimate on $1 - P_{M,N}$, we first prove the right hand side inequality. For this, we put $\Psi := \{\varphi_i\}_{i=1}^N$ and $\Upsilon := \{\varphi_i\}_{i=N+1}^M$. If Ψ is not unitary, by Theorem 4.4 there exists $d_\Psi \in \mathbb{S}^{N-1}$ such that $|\langle d_\Psi, \varphi_i \rangle| < 1/\sqrt{N}$ and hence $\varphi_i \notin R_a^{N-1}(d_\Psi)$ for $i = 1, \ldots, N$. Therefore, if Ψ is not unitary and $\varphi_{N+1}, \ldots, \varphi_M \notin R_a^{N-1}(d_\Psi)$ then Φ is not scalable by Theorem 4.1(a). This yields

$$1 - P_{M,N} \geq \mu_u^M\left(\{\Phi : \Psi \notin \mathcal{SC}(N, N), \forall \varphi \in \Upsilon : \varphi \notin R_a^{N-1}(d_\Psi)\}\right)$$

$$= \int_{\mathcal{SC}(N,N)^c} \mu_u^{M-N}\left(\{\Upsilon \in (\mathbb{S}^{N-1})^{M-N} : \forall \varphi \in \Upsilon : \varphi \notin R_a^{N-1}(d_\Psi)\}\right) d\mu_u^N(\Psi)$$

$$= \int_{\mathcal{SC}(N,N)^c} \mu_u^{M-N}\left(\left([R_a^{N-1}(d_\Psi)]^c\right)^{M-N}\right) d\mu_u^N(\Psi)$$

$$= \left(1 - A_a^{N-1}\right)^{M-N} \int_{\mathcal{SC}(N,N)^c} d\mu_u^N(\Psi)$$

$$= \left(1 - A_a^{N-1}\right)^{M-N}\left(1 - \mu_u^N(\mathcal{SC}(N, N))\right).$$

But $\mu_u^N(\mathcal{SC}(N, N)) = 0$ by (i), and hence the inequality follows.

For the left hand side inequality, let $\{R_j\}_{j=1}^C$ be a cover of $\mathbb{S}^{N-1}$ with spherical caps of angular radius α. Define the event $E := \{\forall j \in \{1, 2, \cdots, C\} \exists i \text{ such that } \varphi_i \in R_j\}$. If event E holds, whenever $d \in \mathbb{S}^{N-1}$, there exists j such that $d \in R_j$. Thus, there also exists i such that d and φ_i are in the same spherical cap, which means $\langle d, \varphi_i \rangle \geq \sqrt{\frac{N-1}{N}}$. Therefore, Theorem 4.1(b) yields that Φ is scalable. So, we have

$$P_{M,N} \geq \Pr(E) = 1 - \Pr\left(\exists j \, \forall i : \varphi_i \in R_j^c\right)$$

$$= 1 - \Pr\left(\bigcup_j \{\forall i : \varphi_i \in \mathbb{R}_j^c\}\right)$$

$$\geq 1 - \sum_j \Pr\left(\{\forall i : \varphi_i \in \mathbb{R}_j^c\}\right)$$

$$= 1 - \sum_j \left(1 - A_\alpha^{N-1}\right)^M = 1 - C \left(1 - A_\alpha^{N-1}\right)^M.$$

This finishes the proof of the theorem. $\qquad\square$

Remark 4.10. An upper bound on C_N can be found in [3, Theorem 1.2] as

$$C_N \le 3N + 2 + \sqrt{N}(N+1)\cos(a)(A_a^{N-1})^{-2} \left(\frac{1}{2A_a^{N-1}}\right)^N.$$

5. Acknowledgments

References

[1] K. Ball, *An elementary introduction to modern convex geometry,* Flavors of geometry, 1–58, Math. Sci. Res. Inst. Publ., **31**, Cambridge Univ. Press, Cambridge, 1997.

[2] J. J. Benedetto and M. Fickus, *Finite Normalized Tight Frames,* Adv. Comput. Math., **18** (2003), 357–385.

[3] P. Bürgisser, F. Cucker, and M. Lotz, *Coverage processes on spheres and condition numbers for linear programming,* The Annals of Probability, **38.2** (2010): 570–604.

[4] J. Cahill and X. Chen, *A note on scalable frames*, Proceedings of the 10th International Conference on Sampling Theory and Applications, pp. 93–96.

[5] J. Cahill, M. Fickus, D. G. Mixon, M. J. Poteet, and N. Strawn, *Constructing finite frames of a given spectrum and set of lengths,* Appl. Comput. Harmon. Anal., **35** (2013), no. 1, 52–73.

[6] P. G. Casazza, M. Fickus, and D. G. Mixon, *Auto-tuning unit norm frames,* Appl. Comput. Harmon. Anal., **32** (2012), no. 1, 1-15.

[7] P. G. Casazza and G. Kutyniok, Finite Frame Theory, Eds., Birkhäuser, Boston (2012).

[8] P. G. Casazza and M. Leon. *Existence and construction of finite frames with a given frame operator.* Int. J. Pure Appl. Math, **63** (2010), 149–158.

[9] O. Christensen, An introduction to frames and Riesz bases, Applied and Numerical Harmonic Analysis. Birkhäuser Boston, Inc., Boston, MA, 2003.

[10] M. S. Copenhaver, Y. H. Kim, C. Logan, K. Mayfield, S. K. Narayan, and J. Sheperd, *Diagram vectors and tight frame scaling in finite dimensions,* Operators and Matrices, **8**, no.1 (2014), 73 – 88.

[11] M. Fickus, B. D. Johnson, K. Kornelson, and K. A. Okoudjou, *Convolutional frames and the frame potential,* Appl. Comput. Harmon. Anal., **19** (2005), 77–91.

[12] O. Güler, Foundations of Optimization, Graduate Texts in Mathematics, **258** Springer, New York, 2010.

[13] F. John, *Extremum problems with inequalities as subsidiary conditions,* Studies and Essays Presented to R. Courant on his 60th Birthday, January 8, 1948, 187–204. Interscience Publishers, Inc., New York, N. Y., 1948.

[14] L. G. Khachiyan, *Rounding of polytopes in the real number model of computation,* Math. Oper. Res., **21**, 1996, 307–320.

[15] P. Kumar, E. A. Yildirim, *Minimum volume enclosing ellipsoids and core sets*, J. Optim. Theory Appl., **126** (2005), 1–21.

[16] G. Kutyniok, K. A. Okoudjou, and F. Philipp, *Scalable frames and convex geometry,* Spectra of Wavelets, Tilings, and Frames (Boulder, CO, 2012), Contemp. Math. 345, Amer. Math. Soc., Providence, RI (2013), to appear.

[17] G. Kutyniok, K. A. Okoudjou, F. Philipp, and E. K. Tuley, *Scalable frames*, Linear Algebra and its Applications **438** (2013), 2225–2238.

[18] S. Nam, M. E. Davies, M. Elad, and R. Gribonval, *The Cosparse Analysis Model and Algorithms*, Appl. Comput. Harmon. Anal., **34** (2013), 30–56.

[19] C P. Niculescu and L.-E. Persson, Convex Functions and Their Applications – A Contemporary Approach, Canadian Mathematical Society, Springer, New York, 2006.

[20] R. Rubinstein, M. Zibulevsky, and M. Elad, *Double Sparsity: Learning Sparse Dictionaries for Sparse Signal Approximation*, IEEE Trans. Signal Process., **58** (2010), 1553–1564.

[21] N. Strawn, *Optimization over finite frame varieties and structured dictionary design,* Appl. Comput. Harmon. Anal., **32** (2012), 413–434.

[22] N. Tomczak-Jaegermann, Banach-Mazur Distances and Finite-Dimensional Operator Ideals, Pitman Monographs and Surveys in Pure and Applied Mathematics, 38 Longman Scientific & Technical, Harlow; copublished in the United States with John Wiley & Sons, Inc., New York, 1989.

XUEMEI CHEN, DEPARTMENT OF MATHEMATICS, UNIVERSITY OF MARYLAND, COLLEGE PARK, MD 20742 USA

E-mail address: xuemeic@math.umd.edu

GITTA KUTYNIOK, INSTITUT FÜR MATHEMATIK, TECHNISCHE UNIVERSITÄT BERLIN, STRASSE DES 17. JUNI 136, 10623 BERLIN, GERMANY

E-mail address: kutyniok@math.tu-berlin.de

KASSO A. OKOUDJOU, DEPARTMENT OF MATHEMATICS, UNIVERSITY OF MARYLAND, COLLEGE PARK, MD 20742 USA

E-mail address: kasso@math.umd.edu

FRIEDRICH PHILIPP, INSTITUT FÜR MATHEMATIK, TECHNISCHE UNIVERSITÄT BERLIN, STRASSE DES 17. JUNI 136, D 10623 BERLIN, GERMANY

E-mail address: philipp@math.tu-berlin.de

RONGRONG WANG, DEPARTMENT OF MATHEMATICS, UNIVERSITY OF BRITISH COLUMBIA, VANCOUVER, BC V6T1Z2 CANADA

E-mail address: rongwang@math.ubc.ca