Research Article

Preconditioning Filter Bank Decomposition Using Structured Normalized Tight Frames

Martin Ehler

Faculty of Mathematics, University of Vienna, Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria

Correspondence should be addressed to Martin Ehler; martin.ehler@univie.ac.at

We turn a given filter bank into a filtering scheme that provides perfect reconstruction, synthesis is the adjoint of the analysis part (so-called unitary filter banks), all filters have equal norm, and the essential features of the original filter bank are preserved. Unitary filter banks providing perfect reconstruction are induced by tight generalized frames, which enable signal decomposition using a set of linear operators. If, in addition, frame elements have equal norm, then the signal energy is spread through the various filter bank channels in some uniform fashion, which is often more suitable for further signal processing. We start with a given generalized frame whose elements allow for fast matrix vector multiplication, as, for instance, convolution operators, and compute a normalized tight frame, for which signal analysis and synthesis still preserve those fast algorithmic schemes.

1. Introduction

Increasingly detailed data are acquired in all sorts of measurements nowadays, so that fast algorithms are an important factor for successful signal processing. The concept of generalized frames has a long tradition in signal processing and many unitary filter bank schemes with the perfect reconstruction property are induced by tight generalized frames. Frames themselves are basis-like systems that span a vector space but allow for linear dependency. The inherent redundancy of frames can yield advantageous features unavailable within the basis concept [1–4]. If the frame is *tight* and its elements have *unit norm*, then it resembles the concept of an orthonormal basis, with the add-on of useful redundancy, and frame coefficients measure the signal energy in a uniform fashion. Generalized frames were introduced in [5] as a tool for signal decomposition using a set of linear operators. In [6], collections of orthogonal projectors were considered under the name fusion frames, fusion frame filter banks have been considered in [7], and the concept of tight *p*-fusion frames was developed in [8]. As convolution operators are linear, most filter banks can be thought of as pairs of generalized frames, one for analysis and the other for synthesis. Hence, in view of filter banks, it is not sufficient to deal with frames but

we must inevitably consider their generalized counterpart. Tightness of a generalized frame means that the induced unitary filter bank provides perfect reconstruction. As with frames, we seek unit norm tight generalized frames because signal energy is then spread through the various channels in a more uniform fashion. The latter was used in [9] to verify robustness of tight fusion frames against erasures, meaning that it is beneficial to have tight fusion frames with equal norm when dealing with distortions and loss of data. To keep the filter bank perspective, we will focus on generalized frames consisting of convolution operators enabling fast algorithms.

In the present paper we start with a generalized frame, whose elements allow for fast matrix vector multiplications (e.g., convolution operators) and construct a unit norm tight generalized frame that induces a filter bank scheme preserving those fast algorithms. The latter is related to the so-called Paulsen problem for frames, where one is given a unit norm frame and one asks for the closest tight frame with unit norm and for an algorithm to find it. This problem for frames has been partially solved in [10–12]. Note that if we are given a unit norm generalized frame, whose elements allow for fast matrix vector multiplications, then the closest tight generalized frame with unit norm may not provide such fast

algorithmic schemes in general. Here, we aim to find a related tight unit norm generalized frame in such a way that signal analysis and synthesis can still benefit from the underlying fast matrix vector multiplications.

We should point out that we used the term filter bank beyond sets of convolution operators similar to [7], where weighted orthogonal projectors are considered. Nonetheless, if the starting generalized frame consists of convolution operators, then the resulting scheme still represents convolution operators in each channel, but we require one additional linear operator for global pre- and postmultiplication. As this operator has a special structure being the inverse of a convolution frame operator, there are still fast computation schemes available [13].

Our construction is inspired by pseudocovariance estimators of elliptical distributions in [14]; see also [15, 16]. We derive an iterative algorithm on positive definite matrices, for which we prove convergence, so that we obtain a positive definite matrix Γ that enables us to construct the tight unit norm generalized frame.

For related research topics, such as the optimal rescaling of filter banks, we refer to [17, 18]. Preconditioning in the context of Gabor frames is addressed in [19]. In order to assess the benefits of our approach in image analysis, we suggest the use of the structural similarity proposed in [20].

The outline is as follows: In Section 2, we introduce the concept of generalized frames and motivate the construction of unit norm tight generalized frames. In Section 3, we present our iterative algorithm, for which we verify convergence, enabling us to construct tight generalized frames with unit norm that preserve fast analysis and synthesis due to their special structure. In Section 4 we provide few examples of random matrices whose samples satisfy the convergence assumptions needed. We also point out examples for convolution operators and further operators enabling fast matrix vector multiplications. In Section 5, we discuss the structure of our construction when the underlying generalized frame is a sample from an elliptical distribution. Some concluding remarks are contained in Section 6.

2. Generalized Frames

Let $\mathbb{K}$ denote either $\mathbb{R}$ or $\mathbb{C}$. We follow [5] and call a collection $\{T_j\}_{j=1}^n \subset \mathbb{K}^{d \times r}$ a *generalized frame* (or a *g-frame* for short) if there are two constants $0 < A \leq B < \infty$ such that

$$A \|x\|^2 \leq \sum_{j=1}^N \left\| T_j^* x \right\|^2 \leq B \|x\|^2, \quad \forall x \in \mathbb{K}^d. \tag{1}$$

If the constants can be chosen $0 < A = B$, then $\{T_j\}_{j=1}^n$ is called a *tight g-frame* and it is called a *Parseval g*-frame if $0 < A = B = 1$. For $r = 1$, we have $T_j^* x = \langle T_j, x \rangle$, so that we recover the concept of frames (cf. [1]). It turns out that a collection $\{T_j\}_{j=1}^n \subset \mathbb{K}^{d \times r}$ is a *g*-frame if and only if $\bigcup_{j=1}^n \text{range}(T_j)$ spans $\mathbb{K}^d$.

If $\{T_j\}_{j=1}^n$ is a *g*-frame, then the *analysis operator* $\mathcal{F}$: $\mathbb{K}^d \rightarrow \mathbb{K}^{n \times r}$ is given by $x \mapsto (x^* T_j)_{j=1}^n$. Its adjoint is the

synthesis operator $\mathcal{F}^*$: $\mathbb{K}^{n \times r} \rightarrow \mathbb{K}^d$, defined by $(c_j)_{j=1}^n \mapsto \sum_{j=1}^n T_j c_j^*$, such that the generalized frame operator is

$$S = \mathcal{F}^* \mathcal{F} : \mathbb{K}^d \longrightarrow \mathbb{K}^d,$$
$$x \longmapsto \sum_{j=1}^n T_j T_j^* x. \tag{2}$$

The collection $\{S^{-1} T_j\}_{j=1}^n$ is called the *canonical dual g-frame* and yields the expansion

$$x = \sum_{j=1}^n T_j \left(S^{-1} T_j \right)^* x = \sum_{j=1}^n S^{-1} T_j T_j^* x, \quad \forall x \in \mathbb{K}^d, \tag{3}$$

which simply follows from $SS^{-1} = S^{-1}S = I_d$, where I_d denotes the $d \times d$ identity matrix.

Proposition 1. *If $\{T_j\}_{j=1}^\infty \subset \mathbb{K}^{d \times r}$ is a g-frame with g-frame operator S, then $\{S^{-1/2} T_j\}_{j=1}^n$ is a Parseval g-frame and, for any other Parseval g-frame $\{R_j\}_{j=1}^\infty \subset \mathbb{K}^{d \times r}$, one has*

$$\sum_{j=1}^n \left\| T_j - S^{-1/2} T_j \right\|_{HS}^2 \leq \sum_{j=1}^n \left\| T_j - R_j \right\|_{HS}^2. \tag{4}$$

Equality holds if and only if $R_j = S^{-1/2} T_j$, for $j = 1, \ldots, n$.

Here, $\| \cdot \|_{HS}$ denotes the Hilbert-Schmidt (Frobenius) norm. Most parts of the proof of Proposition 1 can follow the lines in [21], where $r = 1$ is considered, so we omit the proof.

Remark 2. We have supposed that the linear operators of a *g*-frame have all the same dimensions, which simplifies notation but is not necessary. The entire paper could also deal with sets of linear operators $\{T_j\}_{j=1}^n$, where $T_j \in \mathbb{K}^{d \times r_j}$, for $j = 1, \ldots, n$. Then $T_j T_j^* \in \mathbb{K}^{d \times d}$ and this is all we need.

Tight frames are desirable because synthesis is simply the adjoint of the analysis part. For signal processing purposes, we are interested in tight *g*-frames that additionally have unit norm, because those more resemble orthonormal bases, and $\{\|T_j^* x\|\}_{j=1}^n$ then has more information about the signal energy in the direction of a particular frame element; see [22] for $r = 1$.

Given some *g*-frame, say with unit norm elements, let us seek a tight *g*-frame with equal norm elements that is nearby. If we give up the equal norm requirement and S is the frame operator of some *g*-frame $\{T_j\}_{j=1}^n$, then the collection $\{S^{-1/2} T_j\}_{j=1}^n$ is a Parseval frame closest to $\{T_j\}_{j=1}^n$; see Proposition 1. In general, however, $\{S^{-1/2} T_j\}_{j=1}^n$ may not have equal norms. The search for the closest Parseval frame with equal norm elements has become known as the Paulsen problem. It is essentially the same problem if we restrict ourselves to the sphere; that is, given a unit norm *g*-frame, we aim for the closest unit norm tight *g*-frame. For $r = 1$, this problem was partially solved in [10–12].
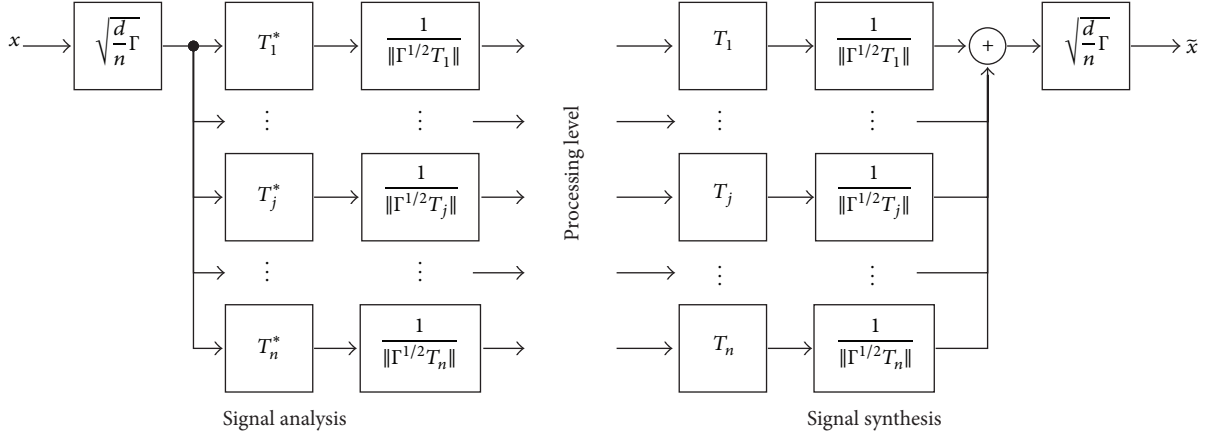
FIGURE 1: Analysis and synthesis scheme, in which fast matrix vector multiplication of T_j^* and T_j can still be used after and before pre- and postmultiplication by $\Gamma^{1/2}$, respectively.

Suppose now that we are given a g-frame $\{T_j\}_{j=1}^n$ that allows for fast matrix vector multiplications for each T_j and T_j^*. The closest equal norm Parseval g-frame may not preserve such features. From a computational point of view, it would be preferable to find an equal norm Parseval g-frame that still allows for fast analysis and synthesis schemes, and this is indeed our topic in the subsequent sections.

3. Constructing Unit Norm Tight g-Frames That Preserve Fast Algorithms

The g-frame operator S of some g-frame $\{T_j\}_{j=1}^n$ and hence also S^{-1} are positive definite and Proposition 1 yields the tight g-frame $\{S^{-1/2}T_j\}_{j=1}^n$ that may not have unit norm, and $\{S^{-1/2}T_j/\|S^{-1/2}T_j\|_{HS}\}_{j=1}^n$ has unit norm but may not be tight. To construct a unit norm tight g-frame that preserves fast matrix vector multiplications, we get inspired by Proposition 1 and aim to find a positive definite matrix Γ such that

$$\left\{ \frac{\Gamma^{1/2}T_j}{\|\Gamma^{1/2}T_j\|_{HS}} \right\}_{j=1}^n \tag{5}$$

is a unit norm tight g-frame. As opposed to Proposition 1, we replace S^{-1} with Γ and normalize. The unit norm g-frame (5) is tight if and only if

$$I_d = \Gamma^{1/2} \frac{d}{n} \sum_{j=1}^n \frac{T_j T_j^*}{\|\Gamma^{1/2}T_j\|_{HS}^2} \Gamma^{1/2}. \tag{6}$$

Signal analysis and synthesis requires pre- and postmultiplication by $\Gamma^{1/2}$ but in-between we can use n times the fast algorithms provided by T_j^* and T_j (cf. Figure 1). Now, $T_j^*\Gamma^{1/2}/\|\Gamma^{1/2}T_j\|_{HS}$ has unit norm, so that the signal energy better relates to the magnitudes of $\{\|T_j^*\Gamma^{1/2}x\|/\|\Gamma^{1/2}T_j\|_{HS}\}_{j=1}^n$. Thus, the special structure (5) can be advantageous over other unit norm tight g-frames that can be closer to the original one.

Remark 3. The filtering scheme in Figure 1 can preserve many properties of the original g-frame $\{T_j\}_{j=1}^n$, which can go beyond fast matrix vector multiplications, such as being orthogonal projectors and sparse matrices, as long as the application of $\Gamma^{1/2}$ is implemented separately and we do not use $\Gamma^{1/2}T_j$ directly.

Remark 4. We point out that structure (5) is different from the approach in [23], where rescalings are sought to derive tight frames. The authors in [24] discuss the setting when a linear operator exists that maps a frame into a unit norm tight frame. We are more general here, because we are joining both and we apply a linear operator and allow for rescaling.

Note that (6) is equivalent to

$$\Gamma = \left(\frac{d}{n} \sum_{j=1}^n \frac{T_j T_j^*}{\text{trace}\left(T_j^*\Gamma T_j\right)} \right)^{-1}. \tag{7}$$

Thus, Γ is the inverse of the generalized frame operator of $\{\sqrt{d/n}\|\Gamma^{1/2}T_j\|_{HS}^2 T_j\}_{j=1}^n$. Since this equation is invariant under scalings, we can look for a solution Γ with $\text{trace}(\Gamma) = 1$. Let $\mathscr{P}$ be the collection of hermitian positive definite matrices in $\mathbb{K}^{d\times d}$ and denote by $\mathscr{P}_1$ the same space with the additional requirement that the trace is 1. The fixed point equation (7) gives rise to an iterative scheme that was already considered in [14, 15] for $r = 1$ to estimate the covariance of elliptical distributions. As initialization we choose $\Gamma_0 := (1/d)I_d \in \mathscr{P}_1$ and define

$$\Gamma_{k+1}$$
$$:= \frac{\left((d/n) \sum_{j=1}^n \left(T_j T_j^* / \text{trace}\left(T_j^*\Gamma_k T_j\right) \right) \right)^{-1}}{\text{trace}\left(\left((d/n) \sum_{j=1}^n \left(T_j T_j^* / \text{trace}\left(T_j^*\Gamma_k T_j\right) \right) \right)^{-1} \right)}. \tag{8}$$

Note that $\Gamma_1 = S^{-1}/\text{trace}(S^{-1})$, and, to verify convergence, we will follow the ideas of the technical procedure used in [14,

15] for $r = 1$. For analysis purposes, we will introduce the mapping

$$M : \mathscr{P} \longrightarrow \mathscr{P},$$

$$M(\Gamma) := \frac{d}{n} \sum_{j=1}^{n} \frac{\Gamma^{1/2} T_j T_j^* \Gamma^{1/2}}{\text{trace}\left(T_j^* \Gamma T_j\right)}, \quad (9)$$

$$M_k := M(\Gamma_k),$$

so that

$$\Gamma_{k+1} = \frac{\Gamma_k^{1/2} M_k^{-1} \Gamma_k^{1/2}}{\text{trace}\left(\Gamma_k M_k^{-1}\right)},$$

$$\Gamma_0 := \frac{1}{d} I_d. \quad (10)$$

We will first check that the mapping M is injective up to scalings, which generalizes [14, Theorem 2.1] from $r = 1$ to the general case.

Lemma 5. *Let* $\{T_j\}_{j=1}^{n} \subset \mathbb{K}^{d \times r}$ *be a g-frame and let* $\Gamma_a, \Gamma_b \in \mathbb{K}^d$ *be positive definite. Then* $M(\Gamma_a) = M(\Gamma_b)$ *if and only if there is a positive constant* $c > 0$ *such that* $\Gamma_a = c\Gamma_b$.

Proof. Without loss of generality, we can assume that $\Gamma_b = I_d$. Otherwise, replace $\{T_j\}_{j=1}^{n}$ with $\{\Gamma_b^{-1/2} T_j\}_{j=1}^{n}$. Let γ_1 be the largest eigenvalue of Γ_a and let P_1 be the associated eigenprojector. Moreover, let $\{P_i\}_{i=2}^{s}$ be one-dimensional eigenprojectors of Γ_a associated with eigenvalues $\{\gamma_i\}_{i=2}^{s}$ such that $\gamma_1 > \gamma_i$, for $i = 2, \ldots, s$, and $\sum_{i=1}^{s} \gamma_i P_i = \Gamma_a$. Note that $\gamma_2, \ldots, \gamma_s$ do not need to be pairwise distinct. Since $\Gamma_a^{1/2} P_1 \Gamma_a^{1/2} = \gamma_1 P_1$ and $\sum_{i=1}^{s} P_i = I_d$, we obtain

$$\text{trace}\left(P_1 M(\Gamma_a)\right) = \frac{d}{n} \sum_{j=1}^{n} \frac{\text{trace}\left(P_1 \Gamma_a^{1/2} T_j T_j^* \Gamma_a^{1/2}\right)}{\text{trace}\left(T_j^* \Gamma_a T_j\right)}$$

$$= \gamma_1 \frac{d}{n} \sum_{j=1}^{n} \frac{\text{trace}\left(T_j^* P_1 T_j\right)}{\text{trace}\left(T_j^* \Gamma_a T_j\right)}$$

$$= \gamma_1 \frac{d}{n} \sum_{j=1}^{n} \frac{\text{trace}\left(T_j^* P_1 T_j\right)}{\sum_{i=1}^{s} \gamma_i \text{trace}\left(T_j^* P_i T_j\right)} \quad (11)$$

$$\geq \gamma_1 \frac{d}{n} \sum_{j=1}^{n} \frac{\text{trace}\left(T_j^* P_1 T_j\right)}{\gamma_1 \text{trace}\left(T_j^* T_j\right)}$$

$$= \text{trace}\left(P_1 M(I_d)\right).$$

Now, $M(\Gamma_a) = M(I_d)$ implies that either $\text{trace}(T_j P_i T_j^*) = 0$, for all $j = 1, \ldots, n$ and $i = 2, \ldots, r$, or $P_1 = I_d$ and, hence, $\Gamma_a = \gamma_1 I_d$. Since $\{T_j\}_{j=1}^{n}$ is a generalized frame, $\text{trace}(T_j^* P_i T_j)$ cannot vanish simultaneously for all $j = 1, \ldots, n$, so that, indeed, Γ_a is a nonzero multiple of the identity. $\quad \square$

The following result says that if we find a proper tight g-frame with unit norm based on (5), then this tight g-frame is unique.

Proposition 6. *Let* $\{T_j\}_{j=1}^{n}$ *be a g-frame and suppose that there are two positive definite matrices* Γ_a *and* Γ_b *such that both* $\{\Gamma_a^{1/2} T_j / \|\Gamma_a^{1/2} T_j\|_{HS}\}_{j=1}^{n}$ *and* $\{\Gamma_b^{1/2} T_j / \|\Gamma_b^{1/2} T_j\|_{HS}\}_{j=1}^{n}$ *are tight. Then those two tight g-frames are identical.*

Proof. The tightness assumptions imply that $M(\Gamma_a) = M(\Gamma_b) = I$. According to Lemma 5, there is a positive constant c such that $\Gamma_a = c\Gamma_b$. Therefore, the two tight g-frames are identical. $\quad \square$

Next, we use scheme (10) to compute a unit norm tight g-frame.

Theorem 7. *Let* $\{T_j\}_{j=1}^{n} \subset \mathbb{K}^{d \times r}$ *satisfy the following points:*

(i) $\{T_j\}_{j=1}^{n}$ *is a g-frame.*

(ii) *If* $\{1, \ldots, n\} = N_1 \cup N_2$ *and* $N_1 \cap N_2 = \emptyset$, *then* $\bigcup_{j \in N_i} \text{range}(T_j)$ *spans* $\mathbb{K}^d$ *for either* $i = 1$ *or* $i = 2$.

(iii) *If* L *is a proper linear subspace of* $\mathbb{K}^d$, *then* $\#\{j : \text{range}(T_j) \subset L\} < n/d$.

(iv) *If* L *is a proper linear subspace of* $\mathbb{K}^d$, *then* $\#\{j : \text{range}(T_j) \subset L\} < \dim(L)/\alpha$, *where* $\alpha := \max_{1 \leq j \leq n} \text{trace}(T_j^* S^{-1} T_j)$ *and* S *is the g-frame operator of* $\{T_j\}_{j=1}^{n}$.

Then the recursive scheme (10) with $\Gamma_0 = (1/d)I_d$ *converges towards a positive definite* Γ *and* $\{R_j\}_{j=1}^{n}$ *defined by*

$$R_j := \frac{\Gamma^{1/2} T_j}{\|\Gamma^{1/2} T_j\|_{HS}} \quad (12)$$

is a tight g-frame.

This theorem generalizes results in [14, 15], where convergence is verified for $r = 1$. The conditions in Theorem 7 are redundant. Condition (ii) clearly implies (i). For $d \geq 2$, (iii) yields (ii). In fact, conditions (i), (ii), and (iii) depend on the range of each T_j but not on their norm. Note that condition (iii) can only be satisfied by some $\{T_j\}_{j=1}^{n}$ if $n/d > \lfloor (d-1)/r \rfloor$, which yields $n > d\lfloor (d-1)/r \rfloor$. Condition (iv) is independent of global scalings since multiplication of all T_j by some constant c means that the inverse frame operator needs to be divided by c^2. It requires $\alpha < r$, which is, in fact, quite weak.

Proposition 8. *Suppose that* $\{T_j\}_{j=1}^{n} \subset \mathbb{K}^{d \times r}$ *is a g-frame. One then has* $\text{trace}(T_j^* S^{-1} T_j) \leq r$, *for all* $j = 1, \ldots, n$, *and if there is* j *with* $\text{trace}(T_j^* S^{-1} T_j) = r$, *then* T_j *does not have any zero columns and* $\text{range}(S^{-1/2} T_k) \perp \text{range}(S^{-1/2} T_j)$, *for all* $k \neq j$.

Proof of Proposition 8. We first choose an orthonormal basis $\{e_i\}_{i=1}^{r}$ for $\mathbb{K}^r$ and define, for some $1 \leq j \leq n$, the index

set $\mathscr{I}_j := \{i : 1 \leq i \leq r, S^{-1/2}T_j e_i \neq 0\}$. We apply Proposition 1 r-times to derive

$$
\begin{aligned}
r &\geq \sum_{i \in \mathscr{I}_j} \sum_{k=1}^n \left\| T_k^* S^{-1/2} \frac{S^{-1/2}T_j e_i}{\|S^{-1/2}T_j e_i\|} \right\|^2 \\
&= \sum_{i \in \mathscr{I}_j} \frac{1}{\|S^{-1/2}T_j e_i\|^2} \sum_{j=1}^n \sum_{l=1}^r \left| \langle S^{-1/2}T_k e_l, S^{-1/2}T_j e_i \rangle \right|^2 \\
&\geq \sum_{i \in \mathscr{I}_j} \frac{1}{\|S^{-1/2}T_j e_i\|^2} \left| \langle S^{-1/2}T_j e_i, S^{-1/2}T_j e_i \rangle \right|^2 \\
&= \sum_{i \in \mathscr{I}_j} \left\| S^{-1/2}T_j e_i \right\|^2 = \left\| S^{-1/2}T_j \right\|_{HS}^2 \\
&= \text{trace}\left(T_j^* S^{-1} T_j \right).
\end{aligned}
\tag{13}
$$

$\square$

The assumption $\text{trace}(T_j^* S^{-1} T_j) = r$ implies that the two above inequalities become equalities, which yield the required statements.

Note that Proposition 8 bounds the worst case scenario. Since $I_d = S^{-1}S$, taking the trace on both sides yields that $d = \sum_{j=1}^n \text{trace}(T_j^* S^{-1} T_j)$. If $\{T_j\}_{j=1}^n$ has unit norm and is close to being tight meaning $S \approx (n/d)I_d$, then $\alpha \approx d/n$. If $\{T_j\}_{j=1}^n$ is sufficiently generic or in sufficient general position, then (i)–(iv) are satisfied for sufficiently large n.

If $\{T_j\}_{j=1}^n$ is a sample of a continuous distribution on $\mathbb{K}^{d \times r}$ and n is sufficiently large, then with probability one all of the assumptions in Theorem 7 are satisfied.

Proof of Theorem 7. Since $\{T_j\}_{j=1}^n$ is a g-frame, the sequence $\{\Gamma_k\}_{k=1}^\infty$ is well defined. It is also clear that Γ_k is hermitian positive definite and $\text{trace}(\Gamma_k) = 1$, for all $k = 0, 1, 2, \ldots$. If we suppose that $\{\Gamma_k\}_{k=1}^\infty$ converges towards Γ, then $M(\Gamma) = I_d$ must hold, and a direct computation yields that $\{R_j\}_{j=1}^n$ is a tight g-frame with $\|R_j\|_{HS} = 1$, for $j = 1, \ldots, n$.

It remains to verify convergence, which we check in two steps.

Step 1 (refers to Lemma 2.1 in [14]). Let $\lambda_{1,k}$ and $\lambda_{d,k}$ be the largest and smallest eigenvalues of M_k, respectively. We observe

$$
I_d = \frac{d}{n} \sum_{j=1}^n \frac{M_k^{-1/2} \Gamma_k^{1/2} T_j T_j^* \Gamma_k^{1/2} M_k^{-1/2}}{\text{trace}\left(T_j^* \Gamma_k T_j \right)}.
\tag{14}
$$

The positive definite square root of $\Gamma_k^{1/2} M_k^{-1} \Gamma_k^{1/2}$ has the form $Q_k M_k^{-1/2} \Gamma_k^{1/2}$, where Q_k is an orthogonal matrix. Since $Q_k Q_k^* = I_d$, we obtain

$$
I_d = \frac{d}{n} \sum_{j=1}^n \frac{Q_k M_k^{-1/2} \Gamma_k^{1/2} T_j T_j^* \Gamma_k^{1/2} M_k^{-1/2} Q_k^*}{\text{trace}\left(T_j^* \Gamma_k T_j \right)}.
\tag{15}
$$

According to (10), we have

$$
\Gamma_{k+1}^{1/2} = \frac{\left(\Gamma_k^{1/2} M_k^{-1} \Gamma_k^{1/2} \right)^{1/2}}{\sqrt{\text{trace}\left(\Gamma_k M_k^{-1} \right)}} = \frac{Q_k M_k^{-1/2} \Gamma_k^{1/2}}{\sqrt{\text{trace}\left(\Gamma_k M_k^{-1} \right)}},
\tag{16}
$$

which yields

$$
\begin{aligned}
I_d &= \frac{d}{n} \sum_{j=1}^n \left(\frac{\Gamma_{k+1}^{1/2} T_j T_j^* \Gamma_{k+1}^{1/2}}{\text{trace}\left(T_j^* \Gamma_k T_j \right)} \right) \text{trace}\left(\Gamma_k M_k^{-1} \right) = \frac{d}{n} \\
&\cdot \sum_{j=1}^n \left(\frac{\Gamma_{k+1}^{1/2} T_j T_j^* \Gamma_{k+1}^{1/2}}{\text{trace}\left(T_j^* \Gamma_{k+1} T_j \right)} \right) \\
&\cdot \frac{\text{trace}\left(T_j^* \Gamma_{k+1} T_j \right) \text{trace}\left(\Gamma_k M_k^{-1} \right)}{\text{trace}\left(T_j^* \Gamma_k T_j \right)}.
\end{aligned}
\tag{17}
$$

According to (10), we have the identity $\text{trace}(T_j^* \Gamma_{k+1} T_j) = \text{trace}(T_j^* \Gamma_k^{1/2} M_k^{-1} \Gamma_k^{1/2} T_j)/\text{trace}(\Gamma_k M_k^{-1})$, so that

$$
\begin{aligned}
I_d &= \frac{d}{n} \sum_{j=1}^n \left(\frac{\Gamma_{k+1}^{1/2} T_j T_j^* \Gamma_{k+1}^{1/2}}{\text{trace}\left(T_j^* \Gamma_{k+1} T_j \right)} \right) \\
&\cdot \frac{\text{trace}\left(T_j^* \Gamma_k^{1/2} M_k^{-1} \Gamma_k^{1/2} T_j \right)}{\text{trace}\left(T_j^* \Gamma_k T_j \right)} = \frac{d}{n} \\
&\cdot \sum_{j=1}^n \left(\frac{\Gamma_{k+1}^{1/2} T_j T_j^* \Gamma_{k+1}^{1/2}}{\text{trace}\left(T_j^* \Gamma_{k+1} T_j \right)} \right) \frac{\left\| M_k^{-1/2} \Gamma_k^{1/2} T_j \right\|_{HS}^2}{\left\| \Gamma_k^{1/2} T_j \right\|_{HS}^2}
\end{aligned}
\tag{18}
$$

holds. Since each of the matrices $\Gamma_{k+1}^{1/2} T_j T_j^* \Gamma_{k+1}^{1/2} / \text{trace}(T_j^* \Gamma_{k+1} T_j)$ is positive semidefinite, the definition of M_{k+1} in (9) with its largest and smallest eigenvalues implies

$$
\lambda_{d,k}^{-1} M_{k+1} \geq I_d \geq \lambda_{1,k}^{-1} M_{k+1}.
\tag{19}
$$

The left inequality yields $\lambda_{d,k+1}/\lambda_{d,k} \geq 1$ and the right inequality implies $\lambda_{1,k+1}/\lambda_{1,k} \leq 1$. Thus, as required, the sequence $(\lambda_{d,k})_{k=1}^\infty$ is increasing, $(\lambda_{1,k})_{k=1}^\infty$ is decreasing, and both converge towards $\lambda_d \leq 1$ and $\lambda_1 \geq 1$, respectively.

Step 2 (refers to Theorem 2.2 and Corollary 2.2 in [14]). Since Γ_k is positive definite with trace 1, there is a subsequence $(\Gamma_{k_m})_{m=1}^\infty$ that converges towards some positive semidefinite matrix Γ. We must now verify that Γ is positive definite and that the entire sequence converges.

If $\Gamma T_j = 0$, then let $\theta_j \in \mathbb{K}^{d \times r}$ be such that

$$
\frac{\Gamma_{k_m}^{1/2} T_j}{\text{trace}\left(T_j^* \Gamma_{k_m} T_j \right)} \longrightarrow \theta_j.
\tag{20}
$$

For $\mathscr{J} := \{j : \Gamma T_j \neq 0\}$, we observe that the subsequence $(M_{k_m})_{m=1}^\infty$ converges to

$$
M = \frac{d}{n} \left(\sum_{j \in \mathscr{J}} \frac{\Gamma^{1/2} T_j T_j^* \Gamma^{1/2}}{\text{trace}\left(T_j^* \Gamma T_j \right)} + \sum_{j \notin \mathscr{J}} \theta_j \theta_j^* \right).
\tag{21}
$$

Step 1 ensures that M is invertible since its smallest eigenvalue is positive. Let Q_{k_m} be the orthogonal matrices in Step 1, so that $Q_{k_m} \to Q$ and $(\Gamma^{1/2} M^{-1} \Gamma^{1/2})^{1/2} = QM^{-1/2}\Gamma^{1/2}$ is satisfied. According to definition (10), $(\Gamma_{k_m+1})_{m=1}^{\infty}$ converges towards

$$\Gamma_0 = \frac{\Gamma^{1/2} M^{-1} \Gamma^{1/2}}{\text{trace}\left(M^{-1}\Gamma\right)} \qquad (22)$$

and a short calculation yields that the sequence $(M_{k_m+1})_{m=1}^{\infty}$ converges to

$$M_0 = \frac{d}{n}\left(\sum_{j \in \mathscr{J}} \frac{\Gamma_0^{1/2} T_j T_j^* \Gamma_0^{1/2}}{\text{trace}\left(T_j^* \Gamma_0 T_j\right)} + \sum_{j \notin \mathscr{J}} \phi_j \phi_j^*\right), \qquad (23)$$

where $\phi_j = QM^{1/2}\theta_j / \sqrt{\text{trace}(\theta_j^* M^{-1} \theta_j)}$.

Manipulations as in Step 1 applied to the formulas for M and M_0 imply that

$$I_d = \frac{d}{n}\sum_{j \in \mathscr{J}} \frac{M^{-1/2}\Gamma^{1/2} T_j T_j^* \Gamma^{1/2} M^{-1/2}}{\text{trace}\left(T_j^* \Gamma T_j\right)} \\ + \sum_{j \notin \mathscr{J}} M^{-1/2}\theta_j \theta_j^* M^{-1/2} \qquad (24)$$

as well as

$$I_d = \frac{1}{\lambda_1} M_0 + \frac{d}{n} \\ \cdot \sum_{j \in \mathscr{J}} \frac{\Gamma_0^{1/2} T_j T_j^* \Gamma_0^{1/2}}{\text{trace}\left(T_j^* \Gamma_0 T_j\right)} \left(\frac{\text{trace}\left(T_j^* \Gamma^{1/2} M^{-1} \Gamma^{1/2} T_j\right)}{\text{trace}\left(T_j^* \Gamma T_j\right)}\right. \\ \left. - \frac{1}{\lambda_1}\right) + \frac{d}{n}\sum_{j \notin \mathscr{J}} \phi_j \phi_j^* \left(\text{trace}\left(\theta_j^* M^{-1}\theta_j - \frac{1}{\lambda_1}\right)\right). \qquad (25)$$

According to Step 1, the largest and smallest eigenvalue of both M and M_0 are λ_1 and λ_d, respectively. Let P and P_0 be the eigenprojectors of M and M_0, respectively, associated with λ_1 and $s = \text{rank}(P)$ and $s_0 = \text{rank}(P_0)$. As in [14], without loss of generality, we can suppose $s_0 \geq s$.

By multiplying both sides from the left and the right by P_0, the relations $P_0^2 = P_0$ and $(1/\lambda_1)P_0 M P_0 = P_0$ yield

$$0 = \sum_{j \in \mathscr{J}} \frac{P_0 \Gamma_0^{1/2} T_j T_j^* \Gamma_0^{1/2} P_0}{\text{trace}\left(T_j^* \Gamma_0 T_j\right)} \left(\frac{\text{trace}\left(T_j^* \Gamma^{1/2} M^{-1} \Gamma^{1/2} T_j\right)}{\text{trace}\left(T_j^* \Gamma T_j\right)}\right. \\ \left. - \frac{1}{\lambda_1}\right) + \sum_{j \notin \mathscr{J}} P_0 \phi_j \phi_j^* P_0 \left(\text{trace}\left(\theta_j^* M^{-1}\theta_j - \frac{1}{\lambda_1}\right)\right). \qquad (26)$$

Thus, for $j \in \mathscr{J}$, we either have

$$P_0 \Gamma_0^{1/2} T_j T_j^* \Gamma_0^{1/2} P_0 = 0$$

$$\text{or } \frac{\text{trace}\left(T_j^* \Gamma^{1/2} M^{-1} \Gamma^{1/2} T_j\right)}{\text{trace}\left(T_j^* \Gamma T_j\right)} = \frac{1}{\lambda_1}. \qquad (27)$$

The first option yields $P_0 Q M^{-1/2} \Gamma^{1/2} T_j = 0$. The second option implies $\langle P, \Gamma^{1/2} T_j T_j^* \Gamma^{1/2}\rangle = \langle I, \Gamma^{1/2} T_j T_j^* \Gamma^{1/2}\rangle$, which yields after some computations $P\Gamma^{1/2} T_j = \Gamma^{1/2} T_j$.

For $j \notin \mathscr{J}$, we obtain

$$P_0 \phi_j \phi_j^* P_0 = 0$$

$$\text{or trace}\left(\theta_j^* M^{-1}\theta_j - \frac{1}{\lambda_1}\right) = 0. \qquad (28)$$

Similar to the above considerations, the first option yields $P_0 Q M^{-1/2}\theta_j = 0$. The second option implies $P\theta_j = \theta_j$.

We now premultiply both sides of (24) by $P_0 Q$ and postmultiply by $I - P$. The above four options and using that P and $M^{-1/2}$ commute imply $P_0 Q(I - P) = 0$, which is equivalent to $Q^* P_0 Q = Q^* P_0 Q P$. Since $s_0 \geq s$, we obtain $P = Q^* P_0 Q$, so that $QP = P_0 Q$. The latter implies with the above that, for $j \in \mathscr{J}$, either

$$P\Gamma^{1/2} T_j = 0$$

$$\text{or } P\Gamma^{1/2} T_j = \Gamma^{1/2} T_j. \qquad (29)$$

Hence, we can split $\{1, \ldots, n\}$ into two disjoint index sets N_1 and N_2 such that

$$P\Gamma^{1/2} T_j = \begin{cases} 0, & j \in N_1 \\ \Gamma^{1/2} T_j, & j \in N_2. \end{cases} \qquad (30)$$

Condition (ii) yields that $\text{span}(\text{range}(T_j) : j \in N_i)$ equals $\mathbb{K}^d$ for either $i = 1$ or $i = 2$. If this holds for $i = 1$, then we must have $P\Gamma^{1/2} = 0$. If it holds for $i = 2$, then we derive $P\Gamma^{1/2} = \Gamma^{1/2}$.

Suppose now that $P\Gamma^{1/2} = 0$ holds. The same arguments as in the previous paragraph yield, for $j \in \mathscr{J}$, that either $P\theta_j = 0$ or $P\theta_j = \theta_j$; see also [14]. Pre- and postmultiplying both sides in (21) by P yields

$$\lambda_1 P = \frac{d}{n}\sum_{j \in \mathscr{J}_0} \theta_j \theta_j^*, \qquad (31)$$

where $\mathscr{J}_0 := \{j : j \notin \mathscr{J}, P\theta_j = \theta_j\}$. Next, we take the trace on both sides and use that $\text{trace}(\theta_j^* \theta_j) = 1$ to derive

$$\lambda_1 s \leq \frac{d}{n} n_1, \qquad (32)$$

where n_1 is the number of T_j whose range is contained in the null space of Γ. Condition (iii) yields $\lambda_1 \leq dn_1/sn < 1$, which is a contradiction to the results of Step 1. Thus, we must have $P\Gamma^{1/2} = \Gamma^{1/2}$, so that

$$P = \frac{d}{n\lambda_1}\sum_{j \in \mathscr{J}} \frac{\Gamma^{1/2} T_j T_j^* \Gamma^{1/2}}{\text{trace}\left(T_j^* \Gamma T_j\right)} + \frac{d}{n\lambda_1}\sum_{j \in \mathscr{J}_0} \theta_j \theta_j^*. \qquad (33)$$

Since the ranks of the two summations in (21) are additive (see also [14]) the ranks of the two summations in (33)

are additive. Hence, the two terms themselves must be orthogonal projections. According to condition (i), the rank of the first term equals rank(Γ). If rank(Γ) < s, then taking the trace of the second term implies with condition (iii) that $\lambda_1 < 1/(s - \text{rank}(\Gamma)) \leq 1$, which is a contradiction to Step 1. Therefore, $\mathcal{J}_0$ is empty and $s = \text{rank}(\Gamma)$. Taking the trace of the first term in (33) yields

$$\lambda_1 = \frac{d(n - n_1)}{ns}. \tag{34}$$

We obtain $d = \text{trace}(M) \geq s\lambda_1 + (d - s)\lambda_d$, so that (34) implies

$$\lambda_d \leq \frac{dn_1}{n(d - s)}. \tag{35}$$

At this point, we claim that assumption (iv) implies for at least one k that

$$\lambda_{d,k} > \frac{d}{n} \frac{\#\{j : \text{range}(T_j) \subset L\}}{\dim(L)} \tag{36}$$

holds, for all proper linear subspaces $L \subset \mathbb{K}^d$, but postpone the verification to the end of this proof.

Since $\lambda_{d,k}$ is an increasing sequence, (36) implies $d(n_1/n(d - s)) < \lambda_d$ if $d > s$. This violates (35), so that M and hence Γ must have full rank and, therefore, Γ is positive definite. Also, P must have full rank implying $\lambda_1 = \lambda_d = 1$ and $M = I_d$. Since the eigenvalues are monotone, the entire sequence $(M_k)_{k=1}^{\infty}$ converges towards I_d. The latter can be used with Banach's fixed point theorem to verify that also $(\Gamma_k)_{k=1}^{\infty}$ must converge, hence, towards Γ. By continuity, we obtain $M(\Gamma) = I_d$.

We still need to verify (36). We observe $\Gamma_1 = S^{-1}/\text{trace}(S^{-1})$ and define $W_j := \Gamma_1^{1/2}T_j$, which yields $\sum_{j=1}^{n} W_j W_j^* = I_d/\text{trace}(S^{-1})$. By using $\alpha = \max_{1 \leq j \leq n}\text{trace}(T_j^* S^{-1}T_j)$ as in (iv), this implies $M_1 \succeq (d/n\alpha)I_d$, so that

$$\lambda_{d,1} \geq \frac{d}{n\alpha} > \frac{d}{n} \frac{\#\{j : \text{range}(T_j) \subset L\}}{\dim(L)}, \tag{37}$$

where the last inequality is due to (iv). This concludes the proof. □

Remark 9. The inversion of M_k in iterative scheme (10) is numerically stable because there is a lower positive bound on the smallest eigenvalues of M_k. Therefore, we can expect that our scheme is quite stable overall.

Let us have a look at few pathological examples first.

Example 10. (1) If $\{T_j\}_{j=1}^{n}$ is already a unit norm tight g-frame, then $\Gamma_1 = \Gamma = (1/d)I_d$ and $\{R_j\}_{j=1}^{n} = \{T_j\}_{j=1}^{n}$.

(2) If the g-frame consists of a single matrix $T \in \mathbb{K}^{d \times r}$, hence, $r = d$ and T is regular, then $\Gamma_1 = \Gamma = TT^*$ and (12) yields $R = (1/\sqrt{d})(TT^*)^{-1/2}T$, which is a unit norm tight g-frame.

Next, we illustrate Theorem 7 with few numerical examples.

Example 11. Let $d = 2$, $r = 1$, and $n = 3$. We pick $\alpha_1, \alpha_2, \alpha_3$ from a uniform distribution on $[0, 2\pi]$ and define $T_j = \begin{pmatrix} \cos(\alpha_j) \\ \sin(\alpha_j) \end{pmatrix}$, $j = 1, 2, 3$. By multiplication by -1 and rotation of all 3 vectors, we can restrict the angles to lie between 0 and $(2/3)\pi$. For each random choice $\{T_j\}_{j=1}^{3}$, we compute a unit norm tight frame $\{R_j\}_{j=1}^{3}$ using our proposed algorithm. Up to rotations $U_\theta = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix}$ and multiplication by -1, there is only one single unit norm tight frame with three elements. We choose $\{Y_j\}_{j=1}^{3}$, where $Y_j = \begin{pmatrix} \cos(\beta_j) \\ \sin(\beta_j) \end{pmatrix}$ and $\beta_1 = 0$, $\beta_2 = 1/3\pi$, and $\beta_3 = 2/3\pi$. Therefore, to find the tight frame with unit norm that is closest to $\{T_j\}_{j=1}^{3}$, we minimize the distance to $\{T_j\}_{j=1}^{3}$ over all rotations, that is,

$$\hat{\theta} := \arg\min_{\theta \in [-2/3\pi, 2/3\pi]} \sum_{j=1}^{3} \|U_\theta Y_j - T_j\|^2, \tag{38}$$

and define the closest tight frame by $\{Z_j\}_{j=1}^{3} := \{U_{\hat{\theta}}Y_j\}_{j=1}^{3}$. Note that we can suppress the multiplication by -1 because the angles of $\{T_j\}_{j=1}^{3}$ only run in $[0, (2/3)\pi]$. The average error of $\sum_{j=1}^{n} \|Z_j - R_j\|^2$ over 1000 realizations is $\approx 0.0016 = 1/625$; see also Figure 2 for a visualization of few examples. In our numerical experiments, we observed that our proposed algorithm finds a tight frame that is almost identical to the closest tight frame if all pairs T_i and T_j, for $i \neq j$, are far enough from each other.

Example 12. For $0 \leq t \leq 1/2$, define

$$T_1(t) = \begin{pmatrix} \sqrt{1-t} & 0 \\ 0 & \sqrt{t} \end{pmatrix},$$

$$T_2(t) = \begin{pmatrix} \sqrt{t} & \sqrt{t} \\ 0 & \sqrt{1-2t} \end{pmatrix}, \tag{39}$$

$$T_3(t) = \begin{pmatrix} \sqrt{\dfrac{1-t}{2}} & 0 \\ 0 & \sqrt{\dfrac{1+t}{2}} \end{pmatrix},$$

and let $S(t)$ denote the associated g-frame operator. Note that $\{T_j(t)\}_{j=1}^{3}$ satisfies the assumptions of Theorem 7, for all $0 \leq t \leq 1/2$. If $t = 0$, then we have a tight generalized frame with unit norms. Our algorithm provides a Parseval g-frame $\{\sqrt{2/3}R_j(t)\}_{j=1}^{3}$ with equal HS-norm; see Figure 3 for the errors $\sum_{j=1}^{3} \|T_j(t) - \sqrt{2/3}R_j(t)\|_{HS}^2$ and $\sum_{j=1}^{3} \|T_j(t) - S(t)^{-1/2}T_j(t)\|_{HS}^2$.

Example 13. We choose each entry of each element in $\{T_j\}_{j=1}^{3} \subset \mathbb{R}^{2 \times 2}$ independently according to a uniform distribution on $[0, 1]$ and normalize so that $\|T_j\|_{HS} = 1$.
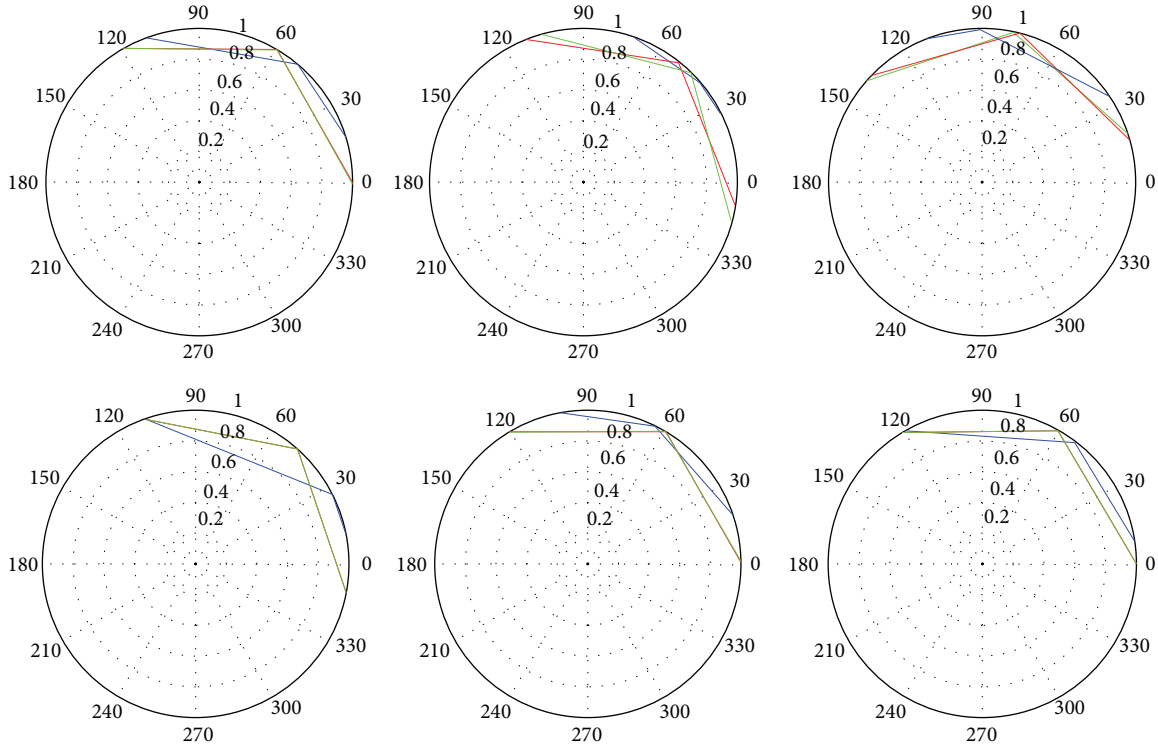
FIGURE 2: Original frame $\{T_j\}_{j=1}^3$ in blue, optimal frame $\{Z_j\}_{j=1}^3 = \{U_{\widehat{\theta}}Y_j\}_{j=1}^3$ in green, and our proposed algorithm finds $\{\Gamma^{1/2}T_j/\|\Gamma^{1/2}T_j\|\}_{j=1}^3$ in red. (Green and red lines are sometimes right on top of each other.)

All numbers in the following are averaged over 10,000 realizations, and let S denote the generalized frame operator of $\{T_j\}_{j=1}^3$. According to Proposition 1, $\{S^{-1/2}T_j\}_{j=1}^3$ is the Parseval generalized frame that is closest to $\{T_j\}_{j=1}^3$, and we compute $\sum_{j=1}^3 \|T_j - S^{-1/2}T_j\|_{HS}^2 \approx 0.61$. However, the elements of $\{S^{-1/2}T_j\}_{j=1}^3$ may not have equal norm. Based on Theorem 7, the collection $\{\widetilde{R}_j\}_{j=1}^3 = \{\sqrt{2/3}R_j\}_{j=1}^3$ is a Parseval generalized frame, its elements have equal norm, and we compute $\sum_{j=1}^3 \|T_j - \widetilde{R}_j\|_{HS}^2 \approx 0.71$. Thus, the additional property of having equal norm costs $\approx 0.10 = 0.71 - 0.61$. It remains open though if there are other Parseval generalized frames whose elements have equal norm and that are closer to $\{T_j\}_{j=1}^3$.

Let us also illustrate when the algorithm fails to converge.

Example 14. For $t = 0$, the collection $\{T_j(t)\}_{j=1}^3$, where

$$T_1(t) = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix},$$

$$T_2(t) = \begin{pmatrix} 1 & t \\ 0 & 0 \end{pmatrix}, \qquad\qquad (40)$$

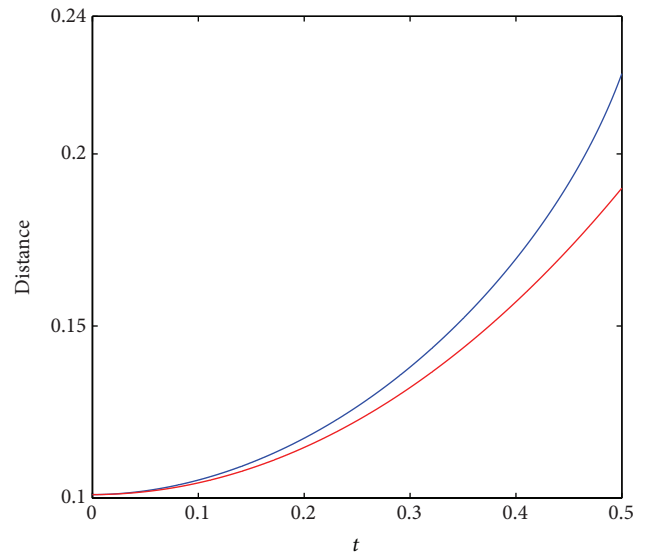$$T_3(t) = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix},$$

FIGURE 3: t versus the distance (blue) $\sum_{j=1}^3 \|T_j(t) - \sqrt{2/3}R_j(t)\|_{HS}^2$ and (red) $\sum_{j=1}^3 \|T_j(t) - S(t)^{-1/2}T_j(t)\|_{HS}^2$ in Example 12. Note that $\{T_j(0)\}_{j=1}^3$ is tight but needs rescaling to become the Parseval generalized frame $\{\sqrt{2/3}T_j(0)\}_{j=1}^3$. It is clear that $R_j(0) = T_j(0)$ and $\sqrt{2/3}T_j(0) = T_j(0)S(0)^{-1/2}$ holds, and we have $\sum_{j=1}^3 \|T_j(0) - \sqrt{2/3}T_j(0)\|_{HS}^2 = 0.1010$. The latter explains why the distance plots do not start at 0.

violates the conditions in Theorem 7, and, indeed, the iterative scheme does not converge towards a positive definite matrix Γ. For $t > 0$, on the other hand, we observe convergence numerically.

The subsequent sections are dedicated to provide some examples of random samples satisfying the assumptions of Theorem 7. We will also provide examples that allow for fast matrix vector multiplications such as convolution operators and we support the intuition that Γ is close to the identity if the sample is close to being tight.

4. Examples of Random Matrices Satisfying the Assumptions for Convergence

We first fuse the concepts of generalized frames and probabilistic frames as developed in [5] and [4], respectively, see also [25].

Definition 15. Let $p \geq 1$ be an integer. One says that a random matrix $T \in \mathbb{K}^{d \times r}$ is a *random g-frame of order p* if there are positive constants A_p and $B_p > 0$ such that

$$A_p \|x\|^{2p} \leq \mathbb{E} \|T^* x\|^{2p} \leq B_p \|x\|^{2p}, \quad \forall x \in \mathbb{K}^d. \quad (41)$$

A random g-frame T of order p is called *tight* if we can choose $A_p = B_p$.

Following the lines of the proof for rank one projectors considered in [2] yields that any random g-frame T of order 1 satisfies $A_1 \leq (1/d)\mathbb{E}\|T(\omega)\|_{HS}^2 \leq B_1$. Similar to finite frames, if T is a random g-frame of order p, then the *random g-frame operator*

$$
\begin{aligned}
S : \mathbb{K}^d &\longrightarrow \mathbb{K}^d, \\
x &\longmapsto \mathbb{E} TT^* x
\end{aligned}
\quad (42)
$$

is positive, self-adjoint, and invertible. Thus, we obtain the reconstruction formula

$$x = \mathbb{E} T \left(S^{-1} T \right)^* x. \quad (43)$$

Moreover, T is a tight random g-frame of order 1 if and only if $S = A I_d$, where $A = (1/d)\mathbb{E}\|T\|_{HS}^2$.

Note that the case $r = 1$ of the following result is already explicitly contained in [26] and see [27] for related results on orthogonal projectors.

Theorem 16. *Let $\{T_j\}_{j=1}^n$ be independent copies of a tight random g-frame $T \in \mathbb{K}^{d \times r}$ of order 1 with $\|T\|_{HS}^2 = R$ for some positive constant R. For fixed $\varepsilon \in (0, 1]$, there are positive constants γ_ε and $c_\varepsilon > 0$ such that, for all $n \geq c_\varepsilon d \ln(d)$, the g-frame operator S_n of the scaled collection $\{\sqrt{d/nR}T_j\}_{j=1}^n$ satisfies $\|I_d - S_n\|_\infty < \varepsilon$ with probability at least $1 - e^{-\gamma_\varepsilon(n/d)}$.*

Proof. Let $\lambda_{\min}(S_n)$ and $\lambda_{\max}(S_n)$ denote the smallest and largest eigenvalue of S_n, respectively. The matrix Chernoff bounds as stated in [28] yield, for all $0 \leq \varepsilon \leq 1$,

$$\mathbb{P}\left(\lambda_{\min}(S_n) \leq 1 - \varepsilon\right) \leq d\left(\frac{e^{-\varepsilon}}{(1-\varepsilon)^{1-\varepsilon}}\right)^{n/d},$$

$$\mathbb{P}\left(\lambda_{\min}(S_n) \geq 1 + \varepsilon\right) \leq d\left(\frac{e^{\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}}\right)^{n/d}. \quad (44)$$

Some calculus yields

$$(1+\varepsilon)^{1+\varepsilon}(1-\varepsilon)^{\varepsilon-1} \leq e^{2\varepsilon}, \quad \forall \varepsilon \in [0,1], \quad (45)$$

so that we derive

$$\mathbb{P}\left(\left\|I_d - \sum_{j=1}^n \frac{d}{nR} T_j^* T_j\right\|_\infty \geq \varepsilon\right) \leq 2d\left(\frac{e^{\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}}\right)^{n/d}. \quad (46)$$

We can further compute

$$
\begin{aligned}
&2d\left(\frac{e^{\varepsilon}}{(1+\varepsilon)^{1+\varepsilon}}\right)^{n/d} \\
&= 2d e^{-(n/d)((1+\varepsilon)\ln(1+\varepsilon)-\varepsilon)} \\
&= e^{-(n/d)((1+\varepsilon)\ln(1+\varepsilon)-\varepsilon-(d/n)\ln(2)\ln(d))}.
\end{aligned}
\quad (47)
$$

Since $(1+\varepsilon)\ln(1+\varepsilon) - \varepsilon > 0$, for all $\varepsilon \in (0, 1]$, we can find a suitable constant $\gamma_\varepsilon > 0$ if n is sufficiently large. $\square$

Remark 17. The constants γ_ε and c_ε in Theorem 16 can be explicitly computed. By using $a_\varepsilon := (1+\varepsilon)\ln(1+\varepsilon) - \varepsilon > 0$, we can choose $c_\varepsilon > \ln(2)/a_\varepsilon$ and $\gamma_\varepsilon = a_\varepsilon - \ln(2)/c_\varepsilon$.

Next, we discuss a few examples.

Example 18 (Gaussian matrices). Let $1 \leq k < d$ and consider the $d \times r$ random matrix T whose entries are i.i.d. Gaussian. Its joint element density is

$$M \longmapsto \frac{1}{(2\pi)^{kd/2}} \exp\left(-\frac{1}{2}\|M\|_{HS}^2\right). \quad (48)$$

The resulting self-adjoint matrix $TT^* \in \mathbb{R}^{d \times d}$ is a singular Wishart-matrix (cf. [29]). According to (48) the distribution of T is invariant under orthogonal transformations, so that T is a tight random g-frame of order p, for all integers p. By using the moments of the chi-squared distribution, we see that the bounds satisfy $A_p = B_p = k(k+2)\cdots(k+2p-2)$.

Example 19 (fusion frames). If the columns of a matrix $T \in \mathbb{K}^{d \times r}, r < d$, have orthonormal columns, then we can identify T with a subspaces $V \in \mathcal{G}_{r,d}(\mathbb{K})$, where $\mathcal{G}_{r,d}(\mathbb{K})$ denotes the Grassmann space, that is, the collection of r-dimensional subspaces of $\mathbb{K}^d$. The Haar measure on $\mathcal{G}_{r,d}(\mathbb{K})$ then induces a random g-frame of order p for all integers p.

Example 20 (Gabor). Time-frequency structured matrices were considered in [30] in relation to compressed sensing, in

which some window vector is modulated and shifted. We use cyclic shifts, which can be performed by applying a matrix C having ones in the lower secondary diagonal, another one in the upper right corner, and zeros anywhere else. The modulation operator on $\mathbb{C}^d$ is given by

$$M = \text{diag}\left(1, e^{2\pi i/d}, \ldots, e^{2\pi i(d-1)/d}\right). \tag{49}$$

For any nonzero $g \in \mathbb{C}^d$, the full Gabor system $\{M^\ell C^k g : \ell, k = 0, \ldots, d-1\}$ has cardinality d^2 and forms a tight frame for $\mathbb{C}^d$ (cf. [31]). We will use the $d^2 \times d$ matrix T, whose rows are formed by the tight frame vectors. A short computation yields that if g is chosen at random as the Rademacher sequence, then T is a tight random g-frame of order 1. Moreover, each TT^* is an orthogonal projector, so that TT^* corresponds to a tight random fusion frame. The same holds when g is the Steinhaus sequence; that is, each entry is uniformly distributed on the complex unit circle.

Next, we have an example that indeed allows for fast matrix vector multiplication.

Example 21 (circulant matrices). Given a vector $x = (x_1, \ldots, x_d)^\top \in \mathbb{R}^d$, the corresponding circulant matrix is

$$\widetilde{T} = \begin{pmatrix} x_1 & x_d & \cdots & x_2 \\ x_2 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & x_d \\ x_d & \cdots & x_2 & x_1 \end{pmatrix}. \tag{50}$$

Each column of $\widetilde{T}$ is a cyclic shift of the previous one. The left $d \times k$ block T of the matrix $\widetilde{T}$ was used as a compressed sensing measurement matrix in [32]. If the entries of x are i.i.d. with zero mean and nonvanishing second moments, then T is a tight random g-frame of order 1 with $A_1 = k\mathbb{E}(x_i^2)$. For instance, if x is the Rademacher sequence, that is, entries are independent and equal to ± 1 with probability $1/2$, then T is tight of order 1 but not of order 2 in general. It is well known that the discrete Fourier matrices diagonalize circulant matrices, so that fast matrix vector multiplications are available. In fact, the terms "filter bank" and "filterning" are usually associated with the application of convolution operators, so that each channel corresponds to a circulant matrix with potentially some subsampling involved.

Remark 22. Samples of all of the above examples satisfy the conditions (i)–(iv) with high probability for sufficiently large sample size. The circulant matrices represent convolution operators and hence correspond to a proper filter bank scheme. They enable fast matrix vector multiplications; hence, the circulant samples in Example 21 are indeed suitable for our construction in Section 3 that preserves this fast algorithmic scheme using the filter bank shown in Figure 1. Each channel corresponds to filtering, but we require one additional linear operator for pre- and postmultiplication.

It must be mentioned that filter banks usually involve some subsampling. Let the matrix $T \in \mathbb{K}^{\widetilde{r} \times r}$, $\widetilde{r} \geq r$, be a random matrix with a single one in each column, whose position is chosen independently at random in a uniform fashion. Then each matrix of the sample $\{T_j\}_{j=1}^n$ corresponds to a sampling operator, so that we derive n samplings of length r. Indeed, T is a tight random g-frame, but it may not satisfy all other conditions in Theorem 7. Nonetheless, subsampling operators in a filter bank are used in combination with more sophisticated filters, say $\{R_j\}_{j=1}^n \subset \mathbb{K}^{d \times \widetilde{r}}$, so that it is possible that the conditions are satisfied by $\{R_j T_j\}_{j=1}^n \subset \mathbb{K}^{d \times r}$.

5. Closeness to the Original g-Frame

To relate the algorithm of the previous section to the Paulsen problem, we would need estimates on the distance between the original and the resulting g-frame. In particular, if the original unit norm g-frame is close to being tight, then we aim to verify that the computed unit norm tight g-frame is nearby. We do not derive any estimates for fixed n but will provide some framework for random samples that supports such intuition.

Theorem 23. *Let T be a random matrix continuously distributed on the set of matrices in $\mathbb{K}^{d \times r}$ and $\{T_j\}_{j=1}^n$ an associated i.i.d. sample with $\Gamma^{(n)}$ being the corresponding limit of the iterative algorithm (10). Then $\Gamma^{(n)}$ converges almost surely towards some positive definite Γ, so that $\Sigma := \Gamma^{-1}$ satisfies*

$$\Sigma = d\mathbb{E}\frac{TT^*}{\text{trace}\left(T^*\Sigma^{-1}T\right)}. \tag{51}$$

Before we shall provide the proof, we have some discussion. As in [14], we observe that results of the previous section applied to a continuously distributed random matrix T yield that (51) has a solution Σ among the symmetric positive definite matrices and is unique up to multiplication by a positive constant.

For elliptical distributions, (51) has a very special meaning. Here, we call a probability distribution on $\mathbb{K}^{d \times r}$ *elliptical* if it has a density f with respect to the standard volume element dT on $\mathbb{K}^{d \times r}$ and

$$\begin{aligned} &f(T) \\ &= |\det(\Sigma)|^{-1/2} g\left(\text{trace}\left(\left(T - \overline{T}\right)^* \Sigma^{-1} \left(T - \overline{T}\right)\right)\right), \end{aligned} \tag{52}$$

where $\Sigma \in \mathbb{K}^{d \times d}$ is hermitian positive definite, $\overline{T} \in \mathbb{K}^{d \times r}$, and g is some nonnegative function not dependent on $\overline{T}$ and Σ with $\int_{\mathbb{K}^{d \times r}} g(\text{trace}(T^*T))dT = 1$. For instance, the Gaussian random matrix in Example 18 is elliptically distributed. A direct computation yields that the matrix Σ of an elliptically distributed random matrix T with $\overline{T} = 0$ satisfies (51). For simplicity, we will restrict ourselves to the case $\overline{T} = 0$ and point out that general $\overline{T}$ can be handled in a similar fashion; see [14] for $r = 1$.

Theorem 23 directly implies the following.

Corollary 24. *If T is elliptically distributed with $\overline{T} = 0$ and Σ is a multiple of the identity, then, for any sample $\{T_j\}_{j=1}^n$, the associated matrices $\Gamma^{(n)}$, for $n \rightarrow \infty$, converge towards $(1/d)I$.*

To verify Theorem 23, we follow the ideas in [14], where $r = 1$ was considered, so we need some notation and two lemmas. Let us define

$$M_n(\Gamma) = \frac{d}{n}\sum_{j=1}^n \frac{\Gamma^{1/2}T_jT_j^*\Gamma^{1/2}}{\text{trace}\left(T_j^*\Gamma T_j\right)},$$

$$M(\Gamma) = d\mathbb{E}\frac{\Gamma^{1/2}TT^*\Gamma^{1/2}}{\text{trace}\left(T^*\Gamma T\right)}, \tag{53}$$

and we denote $h_n(\Gamma) := \text{trace}(M_n(\Gamma)^2)$ and $h(\Gamma) := \text{trace}(M(\Gamma)^2)$.

Lemma 25. *Let $\mathscr{C} \subset \mathbb{K}^{d\times d}$ be a compact set of positive definite matrices with $R \in \mathscr{C}$ implying $\text{trace}(R^{-1}) = 1$ and $\text{trace}(R) \leq K$ with some fixed $K > 1$. Then*

$$\sup_{R\in\mathscr{C}}\left|h_n(R) - h(R)\right| \longrightarrow 0 \tag{54}$$

holds almost surely.

Lemma 26. *The hermitian positive definite matrix $R \in \mathbb{K}^{d\times d}$ is a critical value of h_n if and only if $M_n(R) = I_d$.*

Proof of Theorem 23. Simple arithmetics yield that $\text{trace}(M_n(R)) = \text{trace}(M(R)) = d$ implies $d \leq h_n(R)$, $h(R) \leq d^2$, for all positive definite matrices $R \in \mathbb{K}^{d\times d}$. Furthermore, $h_n(R) = d$ or $h(R) = d$ if and only if $M_n(R) = I$ or $M(R) = I_d$, respectively.

As mentioned above, (51) has a solution Σ among the symmetric positive definite matrices and is unique up to multiplication by a positive constant; see also [14]. Without loss of generality, we can assume that $\Sigma = dI$, which implies $M(\Sigma^{-1}) = M((1/d)I) = I_d$. Choose $\mathscr{C}$ as in Lemma 25 with $(1/d)I_d$ being contained in its interior. For all $R \in \mathscr{C}$ with $R \neq (1/d)I_d$, we must have $d = h((1/d)I_d) < h(R)$. Since h is continuous, Lemma 25 yields that, for any R on the boundary of $\mathscr{C}$, we have $h_n((1/d)I_d) < h_n(R)$ with probability one if n is sufficiently large.

Note that $M_n(\Gamma^{(n)}) = I_d$, so that Lemmas 5 and 26 imply that $\Gamma^{(n)}$ is eventually contained in $\mathscr{C}$ for sufficiently large n. Since $\mathscr{C}$ can be chosen arbitrarily small, it follows that $\Gamma^{(n)} \rightarrow (1/d)I_d = \Sigma^{-1}$ almost surely, which concludes the proof. $\square$

It remains to prove the two Lemmas 25 and 26.

Proof of Lemma 25. We follow [14, Proof of Statement (3.2)]. For $0 \neq T \in \mathbb{K}^{d\times r}$, we define

$$G(T, R) := \frac{R^{1/2}TT^*R^{1/2}}{\text{trace}\left(T^*RT\right)}. \tag{55}$$

As already mentioned in [14] for $r = 1$, G is equicontinuous on $\mathscr{C}$ meaning that, for $\varepsilon > 0$, there is $\delta_\varepsilon > 0$

not dependent on $T \neq 0$ nor on $R_1, R_2 \in \mathscr{C}$, such that $\|R_1 - R_2\| < \delta_\varepsilon$ implies $\|G(T, R_1) - G(T, R_2)\|_{HS} < \varepsilon$. Next, the same covering argument for $\mathscr{C}$ as in [14] used with the equicontinuity and the strong law of large numbers implies (54). We omit the details. $\square$

Proof of Lemma 26. We can simply follow the lines of [14, Proof of Statement (3.3)], where $r = 1$ is discussed. A first order expansion of h_n with the frame property and Kantorovich's inequality yields Lemma 26. No new ideas are involved when dealing with $r > 1$, so we refer to [14] for the details. $\square$

6. Some Concluding Remarks

For some signal processing aspects, the most attractive filter bank schemes are those that provide perfect reconstruction, synthesis is the adjoint of the analysis scheme (so-called unitary filter banks), and filters have equal norm. Tight fusion frames, for instance, correspond to perfect reconstruction filter banks, in which each channel corresponds to an orthogonal projection, and it was verified in [9] that robustness of tight fusion frames against distortions and erasures is maximized when the tight fusion frame has equal norm elements. Our aim was to turn a given filter bank into such more attractive schemes and preserve the essential features of the original filtering process. In terms of frames, we turned a given generalized frame into a tight g-frame with unit norm by rescaling and then applying the inverse square root of the new g-frame operator. Due to our special focus on filter banks, we started with a generalized frame consisting of convolution operators, hence, allowing for fast matrix vector multiplications. Through some iterative scheme, we constructed a generalized tight frame with unit norm, which induced a filter bank that preserved the convolution structure, hence, the fast algorithmic scheme, in each channel. Only one additional global pre- and postmultiplication by $\Gamma^{1/2}$ is necessary. Naturally, the application of $\Gamma^{1/2}$ needs special care because it may be structured but not exactly a convolution operator.

We observed that the assumptions of our algorithm are satisfied by any sufficiently large sample drawn from any continuous distribution or drawn from random convolution operators. Fields of application are filter banks, in which the additional computation costs of the application of $\Gamma^{1/2}$ or Γ, respectively, can be tolerated, as, for instance, when the number of channels is large or when computations are completely offline.

Our findings provide a tool to design new filter banks with improved properties on a theoretical level. Substantial numerical verification goes beyond the scope of the present paper and will be provided in future work. We hope that our theoretical findings can provide the basis for its use in more elaborate signal processing methods.

Conflict of Interests

The author declares that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

References

[1] O. Christensen, *An Introduction to Frames and Riesz Bases*, Birkhäuser, Boston, Mass, USA, 2003.

[2] M. Ehler, "Random tight frames," *The Journal of Fourier Analysis and Applications*, vol. 18, no. 1, pp. 1–20, 2012.

[3] M. Ehler and J. Galanis, "Frame theory in directional statistics," *Statistics & Probability Letters*, vol. 81, no. 8, pp. 1046–1051, 2011.

[4] M. Ehler and K. A. Okoudjou, "Minimization of the probabilistic *p*-frame potential," *Journal of Statistical Planning and Inference*, vol. 142, no. 3, pp. 645–659, 2012.

[5] W. Sun, "G-frames and g-Riesz bases," *Journal of Mathematical Analysis and Applications*, vol. 322, no. 1, pp. 437–452, 2006.

[6] P. G. Casazza, G. Kutyniok, and S. Li, "Fusion frames and distributed processing," *Applied and Computational Harmonic Analysis*, vol. 25, no. 1, pp. 114–132, 2008.

[7] A. Chebira, M. Fickus, and D. G. Mixon, "Filter bank fusion frames," *IEEE Transactions on Signal Processing*, vol. 59, no. 3, pp. 953–963, 2011.

[8] C. Bachoc and M. Ehler, "Tight p-fusion frames," *Applied and Computational Harmonic Analysis*, vol. 35, no. 1, pp. 1–15, 2013.

[9] G. Kutyniok, A. Pezeshki, R. Calderbank, and T. Liu, "Robust dimension reduction, fusion frames, and Grassmannian packings," *Applied and Computational Harmonic Analysis*, vol. 26, no. 1, pp. 64–76, 2009.

[10] B. G. Bodmann and P. G. Casazza, "The road to equal-norm Parseval frames," *Journal of Functional Analysis*, vol. 258, no. 2, pp. 397–420, 2010.

[11] J. Cahill and P. G. Casazza, "The Paulsen problem in operator theory," *Operators and Matrices*, vol. 7, no. 1, pp. 117–130, 2013.

[12] P. G. Casazza, M. Fickus, and D. G. Mixon, "Auto-tuning unit norm frames," *Applied and Computational Harmonic Analysis*, vol. 32, no. 1, pp. 1–15, 2012.

[13] C. Wiesmeyr, N. Holighaus, and P. L. Sondergaard, "Efficient algorithms for discrete Gabor transforms on a nonseparable lattice," *IEEE Transactions on Signal Processing*, vol. 61, no. 20, pp. 5131–5142, 2013.

[14] D. E. Tyler, "A distribution-free *M*-estimator of multivariate scatter," *The Annals of Statistics*, vol. 15, no. 1, pp. 234–251, 1987.

[15] J. T. Kent and D. E. Tyler, "Maximum likelihood estimation for the wrapped Cauchy distribution," *Journal of Applied Statistics*, vol. 15, no. 2, pp. 247–254, 1988.

[16] D. E. Tyler, "Statistical analysis for the angular central Gaussian distribution on the sphere," *Biometrika*, vol. 74, no. 3, pp. 579–589, 1987.

[17] L. Chai, J. Zhang, and Q.-L. Han, "Optimal two-sided diagonal scaling for filter bank frames," *IEEE Transactions on Signal Processing*, vol. 59, no. 12, pp. 5830–5841, 2011.

[18] L. Chai, J. Zhang, C. Zhang, and E. Mosca, "Bound ratio minimization of filter bank frames," *IEEE Transactions on Signal Processing*, vol. 58, no. 1, pp. 209–220, 2010.

[19] P. Balazs, H. G. Feichtinger, M. Hampejs, and G. Kracher, "Double preconditioning for Gabor frames," *IEEE Transactions on Signal Processing*, vol. 54, no. 12, pp. 4597–4610, 2006.

[20] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[21] P. G. Casazza and G. Kutyniok, "A generalization of Gram-Schmidt orthogonalization generating all Parseval frames," *Advances in Computational Mathematics*, vol. 27, no. 1, pp. 65–78, 2007.

[22] J. J. Benedetto and M. Fickus, "Finite normalized tight frames," *Advances in Computational Mathematics*, vol. 18, no. 2–4, pp. 357–385, 2003.

[23] G. Kutyniok, K. A. Okoudjou, F. Philipp, and E. K. Tuley, "Scalable frames," *Linear Algebra and Its Applications*, vol. 438, no. 5, pp. 2225–2238, 2013.

[24] M. Frank, V. I. Paulsen, and T. R. Tiballi, "Symmetric approximation of frames and bases in Hilbert spaces," *Transactions of the American Mathematical Society*, vol. 354, no. 2, pp. 777–793, 2002.

[25] C. Bachoc and M. Ehler, "Signal reconstruction from the magnitude of subspace components," *IEEE Transactions on Information Theory*, vol. 61, no. 7, pp. 1–13, 2015.

[26] R. Vershynin, "How close is the sample covariance matrix to the actual covariance matrix?" *Journal of Theoretical Probability*, vol. 25, no. 3, pp. 655–686, 2012.

[27] B. G. Bodmann, "Random fusion frames are nearly equiangular and tight," *Linear Algebra and its Applications*, vol. 439, no. 5, pp. 1401–1414, 2013.

[28] J. A. Tropp, "User-friendly tail bounds for sums of random matrices," *Foundations of Computational Mathematics*, vol. 12, no. 4, pp. 389–434, 2012.

[29] H. Uhlig, "On singular Wishart and singular multivariate Beta distributions," *The Annals of Statistics*, vol. 22, no. 1, pp. 395–405, 1994.

[30] G. E. Pfander, H. Rauhut, and J. A. Tropp, "The restricted isometry property for time-frequency structured random matrices," *Probability Theory and Related Fields*, vol. 156, no. 3-4, pp. 707–737, 2013.

[31] J. Lawrence, G. E. Pfander, and D. Walnut, "Linear independence of Gabor systems in finite dimensional vector spaces," *The Journal of Fourier Analysis and Applications*, vol. 11, no. 6, pp. 715–726, 2005.

[32] H. Rauhut, J. Romberg, and J. A. Tropp, "Restricted isometries for partial random circulant matrices," *Applied and Computational Harmonic Analysis*, vol. 32, no. 2, pp. 242–254, 2012.