Learning Fast Sparsifying Transforms

Cristian Rusu and John Thompson

***Abstract*—Given a dataset, the task of learning a transform that allows sparse representations of the data bears the name of dictionary learning. In many applications, these learned dictionaries represent the data much better than the static well-known transforms (Fourier, Hadamard etc.). The main downside of learned transforms is that they lack structure and therefore they are not computationally efficient, unlike their classical counterparts. This posses several difficulties especially when using power limited hardware such as mobile devices, therefore discouraging the application of sparsity techniques in such scenarios. In this paper we construct orthonormal and non-orthonormal dictionaries that are factorized as a product of a few basic transformations. In the orthonormal case, we solve exactly the dictionary update problem for one basic transformation, which can be viewed as a generalized Givens rotation, and then propose to construct orthonormal dictionaries that are a product of these transformations, guaranteeing their fast manipulation. We also propose a method to construct fast square but non-orthonormal dictionaries that are factorized as a product of few transforms that can be viewed as a further generalization of Givens rotations to the non-orthonormal setting. We show how the proposed transforms can balance very well data representation performance and computational complexity. We also compare with classical fast and learned general and orthonormal transforms.**

I. Introduction

Dictionary learning methods [1] represent a new class of algorithms that have seen many applications in signal processing [2], image processing [3], wireless communications [4] and machine learning [5]. The key idea of this approach is not to use an off-the-shelf transform like the Fourier, Hadamard or wavelet but to learn a new transform, often called an overcomplete dictionary, for a particular task (like coding and classification) from the data itself. While the dictionary learning problem is NP-hard [6] in general, it has been extensively studied and several good algorithms to tackle it exist. Alternating minimization methods like the method of optimal directions (MOD) [7], K–SVD [8], [9] and direct optimization [10] have been shown to work well in practice and also enjoy some theoretical performance guarantees. While learning a dictionary we need to construct two objects: the dictionary and the representation of the data in the dictionary.

One problem that arises in general when using learned dictionaries is the fact that they lack any structure. This is to be compared with the previously mentioned off-the-shelf transforms that have a rich structure. This is reflected in their low computational complexity, i.e., they can be applied directly using $O(n \log n)$ computations for example [11]. Our goal in this paper is to provide a solution to the problem of constructing fast transforms, based upon the structure of Givens rotations, learned from training data.

We first choose to study orthonormal structures since sparse reconstruction is computationally cheaper in such a dictionary: we project the data onto the column space of the dictionary and keep the largest s coefficients in magnitude to obtain the provable best s-term approximation. Working in an n dimensional feature space, this operation has complexity $O(n^2)$. In a general non-orthonormal (and even overcomplete) dictionary, special non-linear reconstruction methods such as ℓ_1 minimization [12] or greedy approaches like orthogonal matching pursuit (OMP) [13] need to be applied. Aside from the fact that in general these methods cannot guarantee to produce best s-term approximations they are also computationally expensive. For example, the classical OMP has complexity $O(sn^2)$ [14] and, assuming that we are looking for sparse approximations with $s \ll n$, it is in general computationally cheaper than ℓ_1 optimization. Therefore, considering a square orthonormal dictionary is a first step in the direction of constructing a fast transform. Still, notice that computing sparse representations in such a dictionary has complexity $O(n^2)$ and therefore, our goal of constructing a fast transform cannot be reached with just a general orthonormal dictionary. We make the case that our fundamental goal is to actually build a structured orthonormal dictionary such that matrix-vector multiplications with this dictionary can be achieved with less than $O(n^2)$ operations, preferably $O(n \log n)$. This connects our paper to previous work on approximating orthonormal (and symmetric) matrices [15] such that matrix-vector multiplications are computationally efficient.

Previous work [16] [17] [18] [19] [20] in the literature has already proposed various structured dictionaries to cope with the high computational complexity of learned transforms. Previous work also dealt with the construction of structured orthonormal dictionaries. Specifically, [21] proposed to build an orthonormal dictionary composed of a product of only a few Householder reflectors. In this fashion, the computational complexity of the dictionary can be controlled and a trade-off between representation performance and computational complexity was shown.

Learned dictionaries with low computational complexity can bridge the gap between the classical transforms that are preferred especially in power limited hardware (or battery operated devices) and the overcomplete, computationally cumbersome, learned dictionaries that provide state-of-the-art performance in many machine learning tasks. The contribution of this paper is twofold.

First, we consider the problem of constructing an orthonormal dictionary as a product of a given number of generalized

The authors are with the Institute for Digital Communications, School of Engineering, The University of Edinburgh, Scotland. Email: {c.rusu, john.thompson}@ed.ac.uk. Demo source code available online at https://udrc.eng.ed.ac.uk/sites/udrc.eng.ed.ac.uk/files/attachments/demo.zip.

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) Grant number EP/K014277/1 and the MOD University Defence Research Collaboration (UDRC) in Signal Processing.

Givens rotations. We start by showing the optimum solution to the dictionary learning problem when the dictionary is a single generalized Givens rotation and then move to expand on this result and propose an algorithm that sequentially builds a product of generalized Givens rotations to act as a dictionary for sparse representations. Each step of the algorithm solves exactly the proposed optimization problem and therefore we can guarantee that it monotonically converges to a local minimum. We show numerically that the fast dictionaries proposed in this paper outperform those based on Householder reflectors [21] in terms of representation error, for the same computational complexity.

Second, based on a structure similar to the generalized Givens rotation we then propose a learning method that constructs square, non-orthonormal, computationally efficient dictionaries. In order to construct the dictionary we again solve exactly a series of optimization problems. Unfortunately we cannot prove the monotonic converge of the algorithm since the sparse reconstruction step, based in this paper on OMP, cannot guarantee in general a monotonic reduction in our objective function. Still, we are able to show that these fast non-orthonormal transforms perform very well, better than their orthonormal counterparts.

In the results section we compare the proposed methods among each other and to previously proposed dictionary learning methods in the literature. We show that the methods proposed in this paper provide a clear trade-off between representation performance and computational complexity. Interestingly, we are able to provide numerical examples where the proposed fast orthonormal dictionaries have higher computational efficiency and provide better representation performance than the well-known discrete cosine transform (DCT), the transform at the heart of the jpeg compression standard [22].

The paper is organized as follows. Section II outlines the dictionary learning problem in general and with orthonormal constraints. Sections III and IV describe the proposed learning methods. Section V discusses the computational complexities of using different learned dictionaries while Section IV presents numerical evidence on image data that the proposed dictionaries can provide a balance between computational complexity and representation performance.

II. A brief description of dictionary learning optimization problems

Given a real dataset $\mathbf{Y} \in \mathbb{R}^{n\times N}$ and sparsity level s, the general dictionary learning problem is to produce the factorization $\mathbf{Y} \approx \mathbf{D}\mathbf{X}$ given by the optimization problem:

$$\begin{aligned} \underset{\mathbf{D},\ \mathbf{X}}{\text{minimize}} \quad & \|\mathbf{Y}-\mathbf{D}\mathbf{X}\|_F^2 \\ \text{subject to} \quad & \|\mathbf{x}_i\|_0 \leq s,\ 1\leq i\leq N \\ & \|\mathbf{d}_j\|_2 = 1,\ 1\leq j\leq n, \end{aligned} \tag{1}$$

where the objective function describes the Frobenius norm representation error achieved by the square dictionary $\mathbf{D}\in\mathbb{R}^{n\times n}$ with the sparse representations $\mathbf{X}\in\mathbb{R}^{n\times N}$ whose columns are subject to the ℓ_0 pseudo-norm $\|\mathbf{x}_i\|_0$ (the number of non-zero elements of columns $\mathbf{x}_i$). To avoid trivial solutions, the dimensions obey $s \ll n \ll N$. Several algorithms that work very well in practice exist [7] [8] [14] to solve this factorization problem. Their approach, and the one we also adopt in this paper, is to keep the dictionary fixed and update the representations and then reverse the roles by updating the dictionary with the representations fixed. This alternating minimization approach proves to work very well experimentally [7], [8] and allows some theoretical insights [23].

In this paper we also consider the dictionary learning problem (1) with an orthonormal dictionary $\mathbf{Q}\in\mathbb{R}^{n\times n}$ [24] [25] [26] [27]. The orthonormal dictionary learning problem (which we call in this paper Q–DLA) [28] is formulated as:

$$\begin{aligned} \underset{\mathbf{Q},\ \mathbf{X};\ \mathbf{Q}\mathbf{Q}^T=\mathbf{Q}^T\mathbf{Q}=\mathbf{I}}{\text{minimize}} \quad & \|\mathbf{Y}-\mathbf{Q}\mathbf{X}\|_F^2 \\ \text{subject to} \quad & \|\mathbf{x}_i\|_0 \leq s,\ 1\leq i\leq N. \end{aligned} \tag{2}$$

Since the dictionary $\mathbf{Q}$ is orthonormal, the construction of $\mathbf{X}$ no longer involves ℓ_1 [12] or OMP [13] approaches as in (1), but reduces to $\mathbf{X} = \mathcal{T}_s(\mathbf{Q}^T\mathbf{Y})$, where $\mathcal{T}_s()$ is an operator that given an input vector zeros all entries except the largest s in magnitude and given an input matrix applies the same operation on each column in turn. To solve (2) for variable $\mathbf{Q}$ and fixed $\mathbf{X}$, a problem also known as the orthonormal Procrustes problem [29], a closed form solution $\mathbf{Q} = \mathbf{U}\mathbf{V}^T$ is given by the singular value decomposition of $\mathbf{Y}\mathbf{X}^T = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$.

We now move to introduce learning methods that create computationally efficient orthonormal and non-orthonormal dictionaries.

III. A method for designing fast orthonormal transforms: G_m–DLA

In this section we propose a method called G_m–DLA to learn orthonormal dictionaries that are factorized as a product of m building blocks, matrices that we call G-transforms.

A. An overview of G-transforms

We call a G-transform an orthonormal matrix parameterized by $c, d \in \mathbb{R}$ and the indices (i, j), with $i \neq j$ as

$$\mathbf{G}_{ij} = \begin{bmatrix} \mathbf{I}_{i-1} & & & & \\ & * & & * & \\ & & \mathbf{I}_{j-i-1} & & \\ & * & & * & \\ & & & & \mathbf{I}_{n-j} \end{bmatrix} \in \mathbb{R}^{n\times n}, \tag{3}$$

where we have denoted $\mathbf{I}_i$ as the identity matrix of size i and $*$ stands for a non-zero entry. For simplicity, we denote the non-trivial part of $\mathbf{G}_{ij}$ as

$$\tilde{\mathbf{G}}_{ij} = \left\{ \begin{bmatrix} c & d \\ -d & c \end{bmatrix}, \begin{bmatrix} c & d \\ d & -c \end{bmatrix} \right\} \in \mathbb{R}^{2\times 2},\ c^2+d^2=1. \tag{4}$$

Classically, a Givens rotation is a matrix as in (3) with $\tilde{\mathbf{G}}_{ij} = \begin{bmatrix} c & d \\ -d & c \end{bmatrix}$ such that $\det(\mathbf{G}_{ij}) = 1$, i.e., proper rotation matrices are orthonormal matrices with determinant one. These rotations are important since any orthonormal dictionary of size $n \times n$ can be factorized in a product of $\binom{n}{2}$ Givens rotations [30]. In this paper, since we are interested in the computational complexity of these structures, we allow both

options in (4) that fully characterize all 2×2 real orthonormal matrices. With $\tilde{\mathbf{G}}_{ij} = \begin{bmatrix} c & d \\ d & -c \end{bmatrix}$ the G-transform in (3) is in fact a Householder reflector $\mathbf{G}_{ij} = \mathbf{I} - 2\mathbf{g}_{ij}\mathbf{g}_{ij}^T$ where $\mathbf{g}_{ij} \in \mathbb{R}^n$, $\|\mathbf{g}_{ij}\|_2 = 1$, has all entries equal to zero except for the i^{th} and j^{th} entries that are $\sqrt{0.5(1-c)}$ and $-\text{sign}(d)\sqrt{0.5(1+c)}$, respectively. A right side multiplication between a G-transform with $\tilde{\mathbf{G}}_{ij} = \begin{bmatrix} c & d \\ -d & c \end{bmatrix}$ and a matrix $\mathbf{X} \in \mathbb{R}^{n\times N}$ updates only rows i and j as

$$\mathbf{G}_{ij}\mathbf{X} = \begin{bmatrix} \dots & c\mathbf{x}_i + d\mathbf{x}_j & \dots & -d\mathbf{x}_i + c\mathbf{x}_j & \dots \end{bmatrix}^T, \quad (5)$$

where $\mathbf{x}_i^T$ is the i^{th} row of $\mathbf{X}$. The number of operations needed for this task is only $6N$. Left and right multiplications with a G-transform (or its transpose) are therefore computationally efficient. Givens rotations have been previously used with great success in several matrix factorization applications [31], [32].

B. One G-transform as a dictionary

Consider now the dictionary learning problem in (2). Let us keep the sparse representations $\mathbf{X}$ fixed and consider a single G-transform as a dictionary. We reach the following

$$\underset{(i,j),\ \tilde{\mathbf{G}}_{ij}}{\text{minimize}} \ \|\mathbf{Y} - \mathbf{G}_{ij}\mathbf{X}\|_F^2. \quad (6)$$

When indices (i,j) are fixed, the problem reduces to constructing $\tilde{\mathbf{G}}_{ij}$, a constrained two dimensional optimization problem. To select the indices (i,j), among the $\binom{n}{2}$ possibilities, an appropriate strategy needs to be defined. We detail next how to deal with these two problems to provide an overall solution for (6). For simplicity of exposition we define

$$\mathbf{Z} = \mathbf{Y}\mathbf{X}^T, \mathbf{Z}_{\{i,j\}} = \begin{bmatrix} Z_{ii} & Z_{ij} \\ Z_{ji} & Z_{jj} \end{bmatrix} \in \mathbb{R}^{2\times 2}, Z_{ij} = \mathbf{y}_i^T\mathbf{x}_j, \quad (7)$$

where $\mathbf{y}_i^T$ and $\mathbf{x}_i^T$ are the i^{th} rows of $\mathbf{Y}$ and $\mathbf{X}$, respectively. **To solve** (6) **for fixed** (i,j) we start by noticing that the operations occur only to rows i and j, and therefore the dictionary learning problem with a G-transform structure is equivalent to

$$\underset{\tilde{\mathbf{G}}_{ij};\ \tilde{\mathbf{G}}_{ij}^T\tilde{\mathbf{G}}_{ij}=\tilde{\mathbf{G}}_{ij}\tilde{\mathbf{G}}_{ij}^T=\mathbf{I}}{\text{minimize}} \left\| \begin{bmatrix} \mathbf{y}_i^T \\ \mathbf{y}_j^T \end{bmatrix} - \tilde{\mathbf{G}}_{ij} \begin{bmatrix} \mathbf{x}_i^T \\ \mathbf{x}_j^T \end{bmatrix} \right\|_F^2. \quad (8)$$

This is a two dimensional Procrustes problem [29] whose optimum solution is $\tilde{\mathbf{G}}_{ij} = \mathbf{U}\mathbf{V}^T$ where $\mathbf{Z}_{\{i,j\}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$. **Choosing** (i,j) **in** (6) requires a closer look at its objective function. If we consider the identity dictionary (itself a G-transform) then

$$\|\mathbf{Y} - \mathbf{X}\|_F^2 = \|\mathbf{Y}\|_F^2 + \|\mathbf{X}\|_F^2 - 2\text{tr}(\mathbf{Z}), \quad (9)$$

where $\text{tr}(\mathbf{Z})$ is the trace of $\mathbf{Z}$, i.e., the sum of its eigenvalues. When using a G-transform, for $(n-2)$ indices (all indices except the selected i and j) this is effectively the computation that takes place. It has been shown in [21] that the reduction in the objective function of the dictionary learning problem when considering an orthonormal dictionary $\mathbf{Q}$ is given by the Procrustes solution $\mathbf{Q} = \mathbf{U}\mathbf{V}^T$ [21], with $\mathbf{Z} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, is

$$\begin{aligned} \|\mathbf{Y} - \mathbf{Q}\mathbf{X}\|_F^2 =& \|\mathbf{Y}\|_F^2 + \|\mathbf{X}\|_F^2 - 2\text{tr}(\mathbf{Q}^T\mathbf{Y}\mathbf{X}^T) \\ =& \|\mathbf{Y}\|_F^2 + \|\mathbf{X}\|_F^2 - 2\|\mathbf{Z}\|_*, \end{aligned} \quad (10)$$

where $\|\mathbf{Z}\|_*$ is the nuclear norm of $\mathbf{Z}$, i.e., the sum of its singular values.

Using (9) and (10) we can state a result in the special case of a G-transform. We need both because for any indices (i,j) the reduction in the objective function invokes the nuclear norm, while for the other indices the reduction invokes the trace. We can analyze the two objective function values separately because the Frobenius norm is elementwise and as such also blockwise. Therefore, the objective function of (6) is

$$\begin{aligned} &\|\mathbf{Y} - \mathbf{G}_{ij}\mathbf{X}\|_F^2 = \|\mathbf{Y}\|_F^2 + \|\mathbf{X}\|_F^2 - 2\text{tr}(\mathbf{Z}) - 2C_{ij}, \\ &\text{where } C_{ij} = \|\mathbf{Z}_{\{i,j\}}\|_* - \text{tr}(\mathbf{Z}_{\{i,j\}}). \end{aligned} \quad (11)$$

Since we want to minimize this quantity, the choice of indices needs to be made as follows

$$(i^\star, j^\star) = \underset{(i,j),\ j>i}{\arg\max}\ C_{ij}, \quad (12)$$

and then solve a Procrustes problem to construct $\tilde{\mathbf{G}}_{i^\star j^\star}$.

These $(i^\star, j^\star)$ values are the optimum indices that lead to the maximal reduction in the objective function of (6). The expression in (12) is computationally cheap given that $\mathbf{Z}_{\{i,j\}}$ is a 2×2 real matrix. Its trace is trivial to compute $\text{tr}(\mathbf{Z}_{\{i,j\}}) = Z_{ii} + Z_{jj}$ (one addition operation) while the singular values of $\mathbf{Z}_{\{i,j\}}$ can be explicitly computed as

$$\sigma_{1,2} = \sqrt{\frac{1}{2}\left(\|\mathbf{Z}_{\{i,j\}}\|_F^2 \pm \sqrt{\|\mathbf{Z}_{\{i,j\}}\|_F^4 - 4\det(\mathbf{Z}_{\{i,j\}})^2}\right)}. \quad (13)$$

Therefore, the full singular value decomposition can be avoided and the sum of the singular values from (13) can be computed in only 23 operations (three of which are taking square roots). Across all indices (i,j) the cost of computing all C_{ij} is $25\frac{n(n-1)}{2}$ operations. The computational burden is still dominated by constructing $\mathbf{Z} = \mathbf{Y}\mathbf{X}^T$ which takes $2snN$ operations.

Remark 1. Notice that $C_{ij} \geq 0$ always. In general, this is because the sum of the singular values of any matrix $\mathbf{Z}$ of size $n \times n$ is always greater than the sum of its eigenvalues. To see this, use the singular value decomposition of $\mathbf{Z} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \mathbf{\Sigma} = \text{diag}(\sigma)$, and develop:

$$\begin{aligned} \text{tr}(\mathbf{Z}) =& \text{tr}(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T) = \text{tr}(\mathbf{\Sigma}\mathbf{V}^T\mathbf{U}) \\ =& \text{tr}(\mathbf{\Sigma}\mathbf{\Delta}) = \sum_{k=1}^n \sigma_k\Delta_{kk} \leq \sum_{k=1}^n \sigma_k = \|\mathbf{Z}\|_*, \end{aligned} \quad (14)$$

where we have use the circular property of the trace and $\mathbf{\Delta} = \mathbf{V}^T\mathbf{U}$ where Δ_{kk} are its diagonal entries which obey $|\Delta_{kk}| \leq 1$ since both $\mathbf{U}$ and $\mathbf{V}$ are orthonormal and their entries are sub-unitary in magnitude. Therefore, in our particular case, we have that $C_{ij} = 0$ when $\mathbf{Z}_{\{i,j\}}$ is symmetric and positive semidefinite (we have that $\mathbf{\Delta} = \mathbf{I}$ in (14) and therefore $\text{tr}(\mathbf{Z}_{\{i,j\}}) = \|\mathbf{Z}_{\{i,j\}}\|_*$). If we have that $C_{ij} = 0$ for all i and j then no G-transform can reduce the objective function in (6) and therefore the solution is $\mathbf{G}_{ij} = \mathbf{I}$. ■

This concludes our discussion for the single G-transform case. Notice that the solution outlined in this section solves (6) exactly, i.e., it finds the optimum G-transform.

C. A method for designing fast orthonormal transforms: G_m–DLA

In this paper we propose to construct an orthonormal dictionary $\mathbf{U} \in \mathbb{R}^{n\times n}$ with the following structure:

$$\mathbf{U} = \mathbf{G}_{i_m j_m} \dots \mathbf{G}_{i_2 j_2} \mathbf{G}_{i_1 j_1}. \tag{15}$$

The value of m is a choice of the user. For example, if we choose m to be $O(n \log n)$ the dictionary $\mathbf{U}$ can be computed in $O(n \log n)$ computational complexity – similar to the classical fast transforms. The goal of this section is to propose a learning method that constructs such an orthonormal dictionary.

We fix the representations $\mathbf{X}$ and all G-transforms in (15) except for the k^{th}, denoted as $\mathbf{G}_{i_k j_k}$. To optimize the dictionary $\mathbf{U}$ only for this transform we reach the objective function

$$\begin{aligned}\|\mathbf{Y} - \mathbf{U}\mathbf{X}\|_F^2 &= \|\mathbf{Y} - \mathbf{G}_{i_m j_m} \dots \mathbf{G}_{i_1 j_1} \mathbf{X}\|_F^2 \\ &= \|\mathbf{G}_{i_{k+1} j_{k+1}}^T \dots \mathbf{G}_{i_m j_m}^T \mathbf{Y} - \mathbf{G}_{i_k j_k} \dots \mathbf{G}_{i_1 j_1} \mathbf{X}\|_F^2 \\ &= \|\mathbf{Y}_k - \mathbf{G}_{i_k j_k} \mathbf{X}_k\|_F^2,\end{aligned} \tag{16}$$

where we have used the fact that multiplication by any orthonormal transform preserves the Frobenius norm. Matrices $\mathbf{Y}_k$ and $\mathbf{X}_k$ contain the accumulations of the G-transforms.

Notice that we have reduced the problem to the formulation in (6) whose full solution is outlined in the previous section. We can apply this procedure for all G-transforms in the product of $\mathbf{U}$ and therefore a full update procedure presents itself: we will sequentially update each transform and then the sparse representations until convergence. The full procedure we propose, called G_m–DLA, is detailed in Algorithm 1.

The initialization of G_m–DLA uses a known construction. It has been shown experimentally in the past [33], that a good initial orthonormal dictionary is to choose $\mathbf{U}$ from the singular value decomposition of the dataset $\mathbf{Y} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$. We can also provide a theoretical argument for this choice. Consider that

$$\mathbf{X} = \mathcal{T}_s(\mathbf{U}^T\mathbf{Y}) = \mathcal{T}_s(\mathbf{U}^T\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T) = \mathcal{T}_s(\mathbf{\Sigma}\mathbf{V}^T). \tag{17}$$

A sub-optimal choice is to assume that the operator $\mathcal{T}_s$ keeps only the first s rows of $\mathbf{\Sigma}\mathbf{V}^T$, i.e., $\mathbf{X} = \mathbf{\Sigma}_s\mathbf{V}^T$ where $\mathbf{\Sigma}_s$ is the $\mathbf{\Sigma}$ matrix where we keep only the leading principal submatrix of size $s \times s$ and set to zero everything else. This is a good choice since the positive diagonal elements of $\mathbf{\Sigma}$ are sorted in decreasing order of their values and therefore we expect $\mathbf{X}$ to keep entries with large magnitude. In fact, $\|\mathbf{X}\|_F^2 = \sum_{k=1}^s \sigma_k^2$, where the σ_k's are the diagonal elements of $\mathbf{\Sigma}$, due to the fact that the rows of $\mathbf{V}^T$ have unit magnitude. Furthermore, with the same $\mathbf{X} = \mathbf{\Sigma}_s\mathbf{V}^T$ we have

$$\|\mathbf{Y} - \mathbf{U}\mathbf{X}\|_F^2 = \|\mathbf{U}(\mathbf{\Sigma} - \mathbf{\Sigma}_s)\mathbf{V}^T\|_F^2 = \sum_{k=s+1}^{n} \sigma_k^2. \tag{18}$$

We expect this error term to be relatively small since we sum over the smallest squared singular values of $\mathbf{Y}$. Therefore, with this choice of $\mathbf{U}$ and the optimal $\mathbf{X} = \mathcal{T}_s(\mathbf{U}^T\mathbf{Y})$ we have that $\|\mathbf{Y} - \mathbf{U}\mathbf{X}\|_F^2 \leq \sum_{k=s+1}^{n} \sigma_k^2$, i.e., the representation error is always smaller than the error given by the best s-rank approximation of $\mathbf{Y}$.

Algorithm 1 – G_m–DLA.
Fast Orthonormal Transform Learning.
Input: The dataset $\mathbf{Y} \in \mathbb{R}^{n\times N}$, the number of G-transforms m, the target sparsity s and the number of iterations K.
Output: The sparsifying square orthonormal transform $\mathbf{U} = \mathbf{G}_{i_m j_m} \dots \mathbf{G}_{i_2 j_2} \mathbf{G}_{i_1 j_1}$ and sparse representations $\mathbf{X}$ such that $\|\mathbf{Y} - \mathbf{U}\mathbf{X}\|_F^2$ is reduced.

Initialization:

1) Perform the economy size singular value decomposition of the dataset $\mathbf{Y} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$.
2) Compute sparse representations $\mathbf{X} = \mathcal{T}_s(\mathbf{U}^T\mathbf{Y})$.
3) For $k = 1, \dots, m$: with all previous $k-1$ G-transforms fixed, construct the new $\mathbf{G}_{i_k j_k}$ such that

$$\|\mathbf{Y} - \mathbf{G}_{i_k j_k} \mathbf{G}_{i_{k-1} j_{k-1}} \dots \mathbf{G}_{i_1 j_1} \mathbf{X}\|_F^2 \tag{19}$$

is minimized.

Iterations $1, \dots, K$:

1) For $k = 1, \dots, m$: update the new $\mathbf{G}_{i_k j_k}$, with all other transforms fixed, such that (16) is minimized.
2) Compute sparse representations $\mathbf{X} = \mathcal{T}_s(\mathbf{U}^T\mathbf{Y})$, where $\mathbf{U}$ is given by (15).

In G_m–DLA, with the sparse representations $\mathbf{X} = \mathcal{T}_s(\mathbf{U}^T\mathbf{Y})$ we proceed to iteratively construct each G-transform. At step k, the problem to be solved is similar to (16) but all transforms indexed above k are currently the identity (not initialized) and will be computed in the following steps.

Notice that $\mathbf{Z} = \mathbf{Y}\mathbf{X}^T$, necessary to compute all the values C_{ij}, is computed fully only once. After each transform is constructed notice that only two columns of $\mathbf{Z}$ need to be recomputed for the next step. The update of $\mathbf{Z}$ is trivial since it involves the linear combinations of two columns according to a G-transform multiplication (5).

The iterations of G_m–DLA update each G-transform sequentially, keeping all other constant, in order to minimize the current error term.

The algorithm is fast since the matrices involved in all computations can be updated from previous iterations. For example, at step $k+1$, notice from (16) that $\mathbf{Y}_{k+1} = \mathbf{G}_{i_{k+1} j_{k+1}} \mathbf{Y}_k$ and $\mathbf{X}_{k+1} = \mathbf{G}_{i_k j_k} \mathbf{X}_k$. The same observation holds for $\mathbf{Z}_{k+1} = \mathbf{G}_{i_{k+1} j_{k+1}} \mathbf{Y}_k \mathbf{X}_k^T \mathbf{G}_{i_k j_k}^T = \mathbf{G}_{i_{k+1} j_{k+1}} \mathbf{Z}_k \mathbf{G}_{i_k j_k}^T$. Of course, $\mathbf{Y}_1 = \mathbf{G}_{i_2 j_2}^T \dots \mathbf{G}_{i_m j_m}^T \mathbf{Y}$ and $\mathbf{X}_1 = \mathbf{X}$. We always need to construct $\mathbf{Z}_1$ from scratch since $\mathbf{X}$ has been fully updated in the sparse reconstruction step.

After all transforms are computed, the dictionary $\mathbf{U}$ is never explicitly constructed. We always remember its factorization (15) and apply it (directly or inversely) by sequentially applying the G-transforms in its composition. The total computational complexity of applying this dictionary for $\mathbf{U}^T\mathbf{Y}$ is $O(mN)$ which is $O(n \log n N)$ for sufficiently large m (of order $n \log n$). This is to be compared with the $O(n^2 N)$ of a general orthonormal dictionary. Additionally, when consecutive G-transforms operate on different indices they can be applied in parallel, reducing the running time of G_m–DLA and that of manipulating the resulting dictionary.

The number of transforms m could be decided during the

runtime of G_m–DLA based on the magnitude of the largest value C_{ij}. Since this magnitude decides the reduction in the objective function of our problem, a threshold can be introduced to decide on the fly if a new transform is worth adding to the factorization.

These observations are important from a computational perspective since the number of transforms is relatively high, $O(n \log n)$, and therefore their manipulation should be performed efficiently when learning the dictionary to keep the running time of G_m–DLA low.

Since each G-transform computed in our method maximally reduces the objective function and because the sparse reconstruction step is exact when using an orthonormal dictionary, we can guarantee that the proposed method monotonically converges to a local optimum point.

Remark 2. Currently, at each iteration of the proposed algorithm we update a single G-transform according to the maximum value C_{ij}. We have in fact the opportunity to update a maximum of $\lfloor n/2 \rfloor$ transforms simultaneously. We could for example partition the set $\{1, 2, \ldots, n\}$ in pairs of two and construct the corresponding G-transforms such that the sum of their C_{ij} is maximized. With such a strategy fewer iterations are necessary but the problem of partitioning the indices such that the error is maximally reduced can be computationally demanding (all possible unique combinations of indices associations need to be generated). We expect G_m–DLA, as it is, to produce better results (lower representation error) due to the one-by-one transform update mechanism.

Even so, there are several options regarding the ordering. We can process the G-transforms in the order of their indices or in a random order for example in an effort to try to avoid local minimum points. ■

Remark 3. After indices (i, j) are selected we have that $C_{ij} = 0$ and therefore this pair cannot be selected again until either index i or j participates in the construction of a future G-transform. To see why this is, imagine we have just constructed $\mathbf{G}_{ij}$ to minimize (8). We have that $\mathbf{Z}_{\{i,j\}}$ is updated to $\mathbf{Z}_{\{i,j\}}\tilde{\mathbf{G}}_{ij}^T = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\mathbf{V}\mathbf{U}^T = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^T$ which is symmetric and positive definite due to the solution $\tilde{\mathbf{G}}_{ij} = \mathbf{U}\mathbf{V}^T$. ■

Remark 4. As previously discussed, the Procrustes solution $\mathbf{Q}$ is the best orthonormal minimizer of (10). It has been shown in [21] that with this $\mathbf{Q}$ we have that $\mathbf{T} = \mathbf{Y}\mathbf{X}^T\mathbf{Q}^T = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^T$ is symmetric positive semidefinite. Since $\mathbf{Q}$ is the global minimizer, there cannot be a G-transform $\mathbf{G}_{ij}$ such that $\mathbf{G}_{ij}\mathbf{Q}$ further reduces the error. This means that all symmetric 2×2 sub-matrices $\mathbf{T}_{\{i,j\}} = \begin{bmatrix} T_{ii} & T_{ij} \\ T_{ij} & T_{jj} \end{bmatrix}$ of $\mathbf{T}$ are positive semidefinite, i.e., $T_{ii} + T_{jj} \geq 0$ and $T_{ii}T_{jj} \geq T_{ij}^2$ for all pairs (i, j). This observation needs to hold for any symmetric positive definite matrix $\mathbf{T}$. Unfortunately, the converse is not true in general.

This means that even with an appropriately large $m \sim O(n^2)$, G_m–DLA might not always be able to match the performance of Q–DLA. This is not a major concern since in this paper we explore fast transforms and therefore $m \ll n^2$.■

This concludes the presentation of the proposed G_m–DLA method. Based on similar principles next we provide a learning method for fast square but non-orthonormal dictionaries.

IV. A method for designing fast general, non-orthonormal transforms: R_m–DLA

In the case of orthonormal dictionaries, the fundamental building blocks like Householder reflectors and Givens rotations are readily available. This is not the case for general dictionaries. In this section we propose a building block for non-orthonormal structures in subsection A and then show how this can used in a similar fashion to the G-transform to learn computationally efficient square non-orthonormal dictionaries by deriving the R_m–DLA method in subsection B.

A. A building block for fast non-orthonormal transforms

Let us consider a transform similar to a G-transform (3). For indices (i, j) and variables $p, q, r, t \in \mathbb{R}$ let us define

$$\mathbf{R}_{ij} = \begin{bmatrix} \mathbf{I}_{i-1} & & & & \\ & p & & r & \\ & & \mathbf{I}_{j-i-1} & & \\ & q & & t & \\ & & & & \mathbf{I}_{n-j} \end{bmatrix} \in \mathbb{R}^{n \times n}. \tag{20}$$

We assume no constraints on the variables p, q, r, t and therefore $\mathbf{R}_{ij}$ is no longer orthonormal in general. For simplicity, we denote the non-trivial part of $\mathbf{R}_{ij}$ as

$$\tilde{\mathbf{R}}_{ij} = \begin{bmatrix} p & r \\ q & t \end{bmatrix} \in \mathbb{R}^{2 \times 2}. \tag{21}$$

Notice that, as with the G-transforms, left or right matrix-vector multiplications with $\mathbf{R}_{ij}$ (or its transpose) take 6 operations, similarly to (5).

We propose to solve the following optimization problem

$$\underset{(i,j),\ \tilde{\mathbf{R}}_{ij}}{\text{minimize}} \ \|\mathbf{Y} - \mathbf{R}_{ij}\mathbf{X}\|_F^2 . \tag{22}$$

As in the G-transform case, we proceed with analyzing how indices (i, j) are selected and then how to solve the optimization problem (22), with the indices fixed. We define

$$\mathbf{Z} = \mathbf{Y}\mathbf{X}^T, \ \mathbf{W} = \mathbf{X}\mathbf{X}^T, \tag{23}$$

with entries Z_{ij} and W_{ij} respectively.

Solving (22) **for fixed** (i, j) leads to a least squares optimization problem as

$$\underset{\tilde{\mathbf{R}}_{ij}}{\text{minimize}} \ \left\| \begin{bmatrix} \mathbf{y}_i^T \\ \mathbf{y}_j^T \end{bmatrix} - \tilde{\mathbf{R}}_{ij} \begin{bmatrix} \mathbf{x}_i^T \\ \mathbf{x}_j^T \end{bmatrix} \right\|_F^2 , \tag{24}$$

where $\mathbf{y}_i^T, \mathbf{x}_i^T$ are i^{th} rows of $\mathbf{Y}$ and $\mathbf{X}$ respectively and whose solution is $\tilde{\mathbf{R}}_{ij} = \begin{bmatrix} Z_{ii} & Z_{ij} \\ Z_{ji} & Z_{jj} \end{bmatrix} \begin{bmatrix} W_{ii} & W_{ij} \\ W_{ji} & W_{jj} \end{bmatrix}^{-1}$.

Choosing (i, j) **in** (22) depends upon two situations. First notice that with $\mathbf{R}_{ij} = \mathbf{I}$ we have the same development as in (9). Second, the objective function in (24) develops into

$$\begin{aligned} \left\| \begin{bmatrix} \mathbf{y}_i^T \\ \mathbf{y}_j^T \end{bmatrix} - \tilde{\mathbf{R}}_{ij} \begin{bmatrix} \mathbf{x}_i^T \\ \mathbf{x}_j^T \end{bmatrix} \right\|_F^2 &= \left\| \begin{bmatrix} \mathbf{y}_i^T \\ \mathbf{y}_j^T \end{bmatrix} \right\|_F^2 \\ - \text{tr} &\left(\begin{bmatrix} Z_{ii} & Z_{ij} \\ Z_{ji} & Z_{jj} \end{bmatrix}^T \begin{bmatrix} Z_{ii} & Z_{ij} \\ Z_{ji} & Z_{jj} \end{bmatrix} \begin{bmatrix} W_{ii} & W_{ij} \\ W_{ji} & W_{jj} \end{bmatrix}^{-1} \right). \end{aligned} \tag{25}$$

Therefore, in (24) for $(n-2)$ coordinates the objective function computes (9) while for the two indices selected (i,j) the objective function computes (25). Overall we reach that

$$\begin{aligned}\|\mathbf{Y}-\mathbf{R}_{ij}\mathbf{X}\|_F^2 &= \|\mathbf{Y}\|_F^2+\|\mathbf{X}\|_F^2-2\text{tr}(\mathbf{Z})-C_{ij},\\ \text{with } C_{ij} &= \left\|\begin{bmatrix}\mathbf{x}_i^T\\ \mathbf{x}_j^T\end{bmatrix}\right\|_F^2-2\text{tr}\left(\begin{bmatrix}Z_{ii} & Z_{ij}\\ Z_{ji} & Z_{jj}\end{bmatrix}\right)\\ &+\text{tr}\left(\begin{bmatrix}Z_{ii} & Z_{ij}\\ Z_{ji} & Z_{jj}\end{bmatrix}^T\begin{bmatrix}Z_{ii} & Z_{ij}\\ Z_{ji} & Z_{jj}\end{bmatrix}\begin{bmatrix}W_{ii} & W_{ij}\\ W_{ji} & W_{jj}\end{bmatrix}^{-1}\right).\end{aligned} \tag{26}$$

Since the matrices involved in the computation of C_{ij} are 2×2 we can use the trace formula and the inversion of a 2×2 matrix formula to explicitly calculate

$$\begin{aligned}C_{ij} = W_{ii}+W_{jj}-2(Z_{ii}+Z_{jj})\\ +\frac{W_{ii}(Z_{ij}^2+Z_{jj}^2)+W_{jj}(Z_{ii}^2+Z_{ji}^2)}{W_{ii}W_{jj}-W_{ij}W_{ji}}\\ -\frac{(Z_{ii}Z_{ij}+Z_{ji}Z_{jj})(W_{ij}+W_{ji})}{W_{ii}W_{jj}-W_{ij}W_{ji}}.\end{aligned} \tag{27}$$

Finally, to solve (22) we select the indices

$$(i^\star,j^\star)=\underset{(i,j),\ j>i}{\arg\max}\ C_{ij}, \tag{28}$$

and then solve a least square problem to construct $\tilde{\mathbf{R}}_{i^\star j^\star}$. The C_{ij} are computed only when $W_{ii}W_{jj}-W_{ij}W_{ji}\neq 0$, otherwise $C_{ij}=-\infty$. To compute each C_{ij} in (27) we need 24 operations and there are $\frac{n(n-1)}{2}$ such C_{ij}. The computational burden is still dominated by constructing $\mathbf{Z}=\mathbf{Y}\mathbf{X}^T, \mathbf{W}=\mathbf{X}\mathbf{X}^T$ which take $2snN$ and approximately snN operations, respectively.

Remark 5. A necessary condition for a dictionary $\mathbf{D}\in\mathbb{R}^{n\times n}$ to be a local minimum point for the dictionary learning problem is that all $C_{ij}=0$ for $\mathbf{Z}=\mathbf{Y}\mathbf{X}^T\mathbf{D}^T, \mathbf{W}=\mathbf{D}\mathbf{X}\mathbf{X}^T\mathbf{D}^T$.■

This concludes our discussion for one transform $\mathbf{R}_{ij}$. Notice that just like in the case of one G-transform, the solution given here finds the optimum $\mathbf{R}_{ij}$ to minimize (22).

B. A method for designing fast general transforms: R_m–DLA

Similarly to G_m–DLA, we now propose to construct a general dictionary $\mathbf{D}\in\mathbb{R}^{n\times n}$ with the following structure:

$$\mathbf{D}=\mathbf{R}_{i_mj_m}\dots\mathbf{R}_{i_2j_2}\mathbf{R}_{i_1j_1}\mathbf{\Delta}. \tag{29}$$

The value of m is a choice of the user. For example, if we choose m to be $O(n\log n)$ the dictionary $\mathbf{D}$ can be applied in $O(n\log n)$ computational complexity – similar to the classical fast transforms. The goal of this section is to propose a learning method that constructs such a general dictionary. As the transformations $\mathbf{R}_{ij}$ are general, the diagonal matrix $\mathbf{\Delta}\in\mathbf{R}^{n\times n}$ is there to ensure that all columns $\mathbf{d}_j$ of $\mathbf{D}$ are normalized $\|\mathbf{d}_j\|_2=1$ (as in the formulation (1)). This normalization does not affect the performance of the method since $\mathbf{D}\mathbf{X}$ is equivalent to $(\mathbf{D}\mathbf{\Delta})(\mathbf{\Delta}^{-1}\mathbf{X})$.

We fix the representations $\mathbf{X}$ and all transforms in (29) except for the k^{th} transform $\mathbf{R}_{i_kj_k}$. Moreover, all transforms $\mathbf{R}_{i_{k+1}j_{k+1}},\dots,\mathbf{R}_{i_mj_m}$ are set to $\mathbf{I}$. Because the transforms $\mathbf{R}_{ij}$ are not orthonormal we cannot access directly any transform $\mathbf{R}_{i_kj_k}$ in (29), but only the left most one $\mathbf{R}_{i_mj_m}$. In this case, to optimize the dictionary $\mathbf{D}$ only for this $\mathbf{R}_{i_kj_k}$ transform we reach the objective

$$\|\mathbf{Y}-\mathbf{R}_{i_kj_k}\dots\mathbf{R}_{i_2j_2}\mathbf{R}_{i_1j_1}\mathbf{X}\|_F^2=\|\mathbf{Y}-\mathbf{R}_{i_kj_k}\mathbf{X}_k\|_F^2. \tag{30}$$

The matrix $\mathbf{X}_k$ contain the accumulations of the transforms. Therefore, our goal is to solve

$$\underset{\mathbf{R}_{i_kj_k}}{\text{minimize}}\ \|\mathbf{Y}-\mathbf{R}_{i_kj_k}\mathbf{X}_k\|_F^2. \tag{31}$$

Notice that we have reduced the problem to the formulation in (22) whose full solution is outlined in the previous section. We can apply this procedure for all G-transforms in the product of $\mathbf{D}$ and therefore a full update procedure presents itself: we will sequentially update each transform in (29), from the right to the left, and then the sparse representations until convergence or for a total number of iterations K.

Once these iterations terminate we can refine the result. As previously mentioned, we cannot arbitrarily update a transform $\tilde{\mathbf{R}}_{i_kj_k}$ because this transform is not orthonormal. But we can update its non-trivial part $\tilde{\mathbf{R}}_{i_kj_k}$. Consider the development:

$$\begin{aligned}&\|\mathbf{Y}-\mathbf{R}_{i_mj_m}\dots\mathbf{R}_{i_{k+1}j_{k+1}}\mathbf{R}_{i_kj_k}\mathbf{R}_{i_{k-1}j_{k-1}}\dots\mathbf{R}_{i_1j_1}\mathbf{X}\|_F^2\\ &=\|\mathbf{Y}-\mathbf{B}_k\mathbf{R}_{i_kj_k}\mathbf{X}_k\|_F^2\\ &=\|\text{vec}(\mathbf{Y})-(\mathbf{X}_k^T\otimes\mathbf{B}_k)\text{vec}(\mathbf{R}_{i_kj_k})\|_F^2\\ &=\left\|\text{vec}(\mathbf{Y})-\sum_{t\in\{1,\dots,n\}\setminus\{i_k,j_k\}}((\mathbf{X}_k^T)_t\otimes(\mathbf{B}_k)_t)-\mathbf{C}\mathbf{x}\right\|_F^2\\ &=\|\mathbf{w}-\mathbf{C}\mathbf{x}\|_F^2,\end{aligned} \tag{32}$$

where $\mathbf{x}=\text{vec}(\tilde{\mathbf{R}}_{i_kj_k})\in\mathbb{R}^4$ and $\mathbf{C}=[(\mathbf{X}_k^T)_{i_k}\otimes(\mathbf{B}_k)_{i_k}\ \ (\mathbf{X}_k^T)_{i_k}\otimes(\mathbf{B}_k)_{j_k}\ \ (\mathbf{X}_k^T)_{j_k}\otimes(\mathbf{B}_k)_{i_k}\ \ (\mathbf{X}_k^T)_{j_k}\otimes(\mathbf{B}_k)_{j_k}]\in\mathbb{R}^{nN\times 4}$. We have denoted by $(\mathbf{B}_k)_{i_k}$ the i_k^{th} column

---

Algorithm 2 – $\mathbf{R}_m$–DLA.
Fast Non-orthonormal Transform Learning.
Input: The dataset $\mathbf{Y}\in\mathbb{R}^{n\times N}$, the number of $\mathbf{R}_{ij}$ transforms m, the target sparsity s and the number of iterations K.
Output: The sparsifying square non-orthonormal transform $\mathbf{D}=\mathbf{R}_{i_mj_m}\dots\mathbf{R}_{i_2j_2}\mathbf{R}_{i_1j_1}\mathbf{\Delta}$ and sparse representations $\mathbf{X}$ such that $\|\mathbf{Y}-\mathbf{D}\mathbf{X}\|_F^2$ is reduced.

Initialization:

1) Perform the economy size singular value decomposition of the dataset $\mathbf{Y}=\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$.
2) Compute sparse representations $\mathbf{X}=\mathcal{T}_s(\mathbf{U}^T\mathbf{Y})$.

Iterations $1,\dots,K$:

1) For $k=1,\dots,m$: with all previous $k-1$ transforms fixed and all next $(m-k)$ transforms set to $\mathbf{I}$, construct the new $\mathbf{R}_{i_kj_k}$ such that (30) is minimized.
2) Choose an appropriate $\mathbf{\Delta}$ in (29).
3) Compute sparse representations $\mathbf{X}=\text{OMP}(\mathbf{D},\mathbf{Y})$.

Iterations $1,\dots,K$:

1) For $k=1,\dots,m$: with all transforms except the k^{th} fixed, update the non-trivial part of $\mathbf{R}_{i_kj_k}$, denoted $\tilde{\mathbf{R}}_{i_kj_k}$, such that (32) is minimized.
2) Choose an appropriate $\mathbf{\Delta}$ in (29).
3) Compute sparse representations $\mathbf{X}=\text{OMP}(\mathbf{D},\mathbf{Y})$.

of $\mathbf{B}_k$ and $\otimes$ is the Kronecker product. To develop (32) we have used the fact that the Frobenius norm is an elementwise operator, the structure of $\mathbf{R}_{i_k j_k}$ and the fact that

$$\text{vec}(\mathbf{B}_k \mathbf{R}_{i_k j_k} \mathbf{X}_k) = (\mathbf{X}_k^T \otimes \mathbf{B}_k)\text{vec}(\mathbf{R}_{i_k j_k}). \quad (33)$$

The $\mathbf{x}$ that minimizes (32) is given by the least squares solution $\mathbf{x} = (\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T\mathbf{w}$. Therefore, once the product of the m transforms is constructed we can update the non-trivial part of any transform to further reduce the objective function. What we cannot do is update the indices (i_k, j_k) on which the calculation takes place, these stay the same.

Therefore, we propose a learning procedure that has two sets of iterations: the first constructs the transforms $\mathbf{R}_{i_k j_k}$ in a rigid manner, ordered from right to left most, and the second only updates the non-trivial parts $\tilde{\mathbf{R}}_{i_k j_k}$ of all the transforms without changing the coordinates (i_k, j_k). The full procedure we propose, called R_m–DLA, is detailed in Algorithm 2. This algorithm has two main parts which we will now describe.

The initialization of R_m–DLA has the goal to construct the sparse representation matrix $\mathbf{X} \in \mathbb{R}^{n \times N}$. We have several options in this step. We can construct $\mathbf{X}$ in the same way as for G_m–DLA from the singular value decomposition of the dataset or by running another dictionary learning algorithm (like the K–SVD [8] for example) and use the $\mathbf{X}$ it constructs.

The iterations of R_m–DLA are divided into two sets. The goal of the first set of iterations is to decide upon all the indices (i_k, j_k) while the second set of iterations optimizes over the non-trivial components of all the transforms in the factorization with the fixed indices previously decided.

The proposed R_m–DLA is can be itself efficiently implemented. When iteratively solving problems as (30) we have that $\mathbf{X}_{k+1} = \mathbf{R}_{i_k j_k}\mathbf{X}_k$ with $\mathbf{X}_1 = \mathbf{X}$ while when iteratively solving problems as (32) we have that $\mathbf{X}_{k+1} = \mathbf{R}_{i_k j_k}\mathbf{X}_k$ and $\mathbf{B}_{k+1} = \mathbf{B}_k\mathbf{R}_{i_k j_k}^{-1}$ with $\mathbf{X}_1 = \mathbf{X}$ and $\mathbf{B}_1 = \mathbf{R}_{i_m j_m} \dots \mathbf{R}_{i_2 j_2}$. The explicit inverse $\mathbf{R}_{i_k j_k}^{-1}$ is not computed, instead the equivalent linear system for 2 variables can be efficiently solved.

The updates of all the transforms $\mathbf{R}_{i_k j_k}$ monotonically decrease our objective function since we solve exactly the optimization problems in these variables. Unfortunately, normalizing to unit ℓ_2 norm the columns of the transform and constructing the sparse approximations via OMP, which is not an exact optimization step, may cause increases in the objective function. For these reasons, monotonic convergence of R_m–DLA to a local minimum point cannot be guaranteed. For this reason, at all times we keep track of the best solution pair $(\mathbf{D}, \mathbf{X})$ and return it at the end of each iterative process.

This concludes our discussion of R_m–DLA. We now move to discuss the computational complexity of the transforms created by the proposed methods and to show experimentally their representation capabilities.

V. The computational complexity of using learned transforms

In this section we look at the computational complexity of using the learned dictionaries to create the sparse representations on a dataset $\mathbf{Y}$ of size $n \times N$. We are in a computational regime where we assume dimensions obey

$$s \ll n \ll N. \quad (34)$$

The computational complexity of using a general non-orthonormal dictionary $\mathbf{A}$ of size $n \times n$ in sparse recovery problems with Batch–OMP [14] is

$$N_{\mathbf{A}} \approx (2n^2 + s^2 n + 3sn + s^3)N + n^3. \quad (35)$$

The cost of n^3 is associated with the construction of the Gram matrix of the dictionary and it does not depend on the number of samples N in the data. The total number of operations is dominated by constructing the projections in the dictionary column space which takes $2n^2$ operations per sample. The other operations dependent on the sparsity s expresses the cost of iteratively finding the support of the sparse approximation.

The computational complexity of using an orthonormal dictionary $\mathbf{Q}$ designed via Q–DLA is

$$N_{\mathbf{Q}} \approx (2n^2 + ns)N. \quad (36)$$

As in the general case, the cost is dominated by constructing the projections $\mathbf{Q}^T\mathbf{Y}$ which takes $2n^2$ operations. The cost of ns expresses the approximate work done to identify the largest s entries in magnitude in the representation of each data sample. This can be performed in an efficient manner by keeping the s largest components in magnitude while the projections are computed for each data sample. Compared with (35), the iterative steps of constructing the support of the OMP solution for each sample and the construction of the Gram matrix (which is the identity matrix in this case) is no longer needed.

The same operation with a dictionary $\mathbf{U}$ as (15) computed via G_{m_1}–DLA takes

$$N_{\mathbf{U}} \approx (6m_1 + ns)N. \quad (37)$$

The result is similar to (36) but now the cost of constructing the projections $\mathbf{U}^T\mathbf{Y}$ takes now only $6m_1$ operations per data sample. Here is where the G-transform factorization is used explicitly to reduce the computational complexity.

Finally, with a dictionary $\mathbf{D}$ as (29) computed via R_{m_2}–DLA the sparse approximation step via Batch–OMP [14] takes

$$N_{\mathbf{D}} \approx (6m_2 + n + s^2 n + 3sn + s^3)N + 6m_2 n. \quad (38)$$

In this case, the cost of building the projections $\mathbf{D}^T\mathbf{Y}$ takes $6m_2$ operations to apply each $\mathbf{R}_{ij}$ transform and then n operations to apply the scaling of the diagonal $\mathbf{\Delta}$. Simplifications occur also for the construction of the symmetric Gram matrix $\mathbf{D}^T\mathbf{D}$ which now takes $6m_2 n$ operations, instead of the regular n^3 operations. This later simplification might not be significant since it is not dependent on the size of the dataset N.

A dictionary $\mathbf{U}$ designed via G_{m_1}–DLA has approximately the same computationally complexity as a general orthonormal dictionary $\mathbf{Q}$ designed via Q–DLA when

$$m_1 = \left\lfloor \frac{n^2}{3} \right\rfloor. \quad (39)$$

Because any orthonormal matrix can be factorized as a product of $\frac{n(n-1)}{2}$ G-transforms and because of the upper limit imposed in (39) it is clear that we cannot express any orthonormal dictionary as an efficient transform for sparse representations. In some cases, the full orthonormal dictionary

$\mathbf{Q}$ might be more efficient than its factorization with G-transforms. In general, the representation error achieved by general orthonormal dictionaries designed via Q–DLA is a performance limit for G-transform based dictionaries.

A similar comparison can be made between the computational complexity of a general dictionary $\mathbf{A}$ and that of a dictionary $\mathbf{D}$ composed of m_2 transformations $\mathbf{R}_{ij}$. Their complexities approximately match when

$$m_2 = \left\lfloor \frac{(2n^2 + s^2 n + 3sn + s^3)N + n^3}{6(N+n)} \right\rfloor \overset{N\to\infty}{\approx} \left\lfloor \frac{n^2}{3} \right\rfloor. \tag{40}$$

This shows that for both G_m–DLA and R_m–DLA the computationally efficient regimes are when $m \sim O(n)$ or in general $m \ll n^2$.

A last comment regards the comparison between dictionaries created with G_{m_1}–DLA and R_{m_2}–DLA. When $m_1 = m_2$ we expect R_{m_2}–DLA to perform better but at a higher computational cost. Assuming large datasets $N \to \infty$ and low sparsity $s \ll n$, computational complexities approximately match when

$$m_1 \approx \left\lfloor m_2 + \frac{(s^2 + 3s + 1)n}{6} \right\rfloor. \tag{41}$$

Due to the use of the OMP procedure for non-orthogonal dictionaries to create the sparse approximations, dictionaries designed via R_m–DLA are much more computationally complex than the orthonormal dictionaries designed via G_m–DLA. Otherwise, as depicted in (41), for the same representation performance the orthonormal dictionaries may contain many more G-transforms in their factorization than $\mathbf{R}_{ij}$ transforms contained in the factorization of a non-orthonormal dictionary. As a consequence, it may be that orthonormal dictionaries are always more computationally efficient than general dictionaries for approximately equal representation capabilities. A definite advantage of R_m–DLA is that it has the potential to create dictionaries that go below representation errors given by orthonormal dictionaries designed via Q–DLA, the performance limit of G_m–DLA.

Using these approximate complexities, we discuss in the results section the representation performance versus the computational complexity trade-off that the dictionaries constructed via the proposed methods display.

VI. Experimental results

In this section we provide experimental results that show how transforms designed via the proposed methods G_m–DLA and R_m–DLA behave on image data.

The input data that we consider are taken from popular test images from the image processing literature (pirate, peppers, boat etc.). The test datasets $\mathbf{Y} \in \mathbb{R}^{n \times N}$ consist of 8×8 non-overlapping patches with their means removed and normalized $\mathbf{Y} = \mathbf{Y}/255$. We choose to compare the proposed methods on image data since in this setting fast transforms that perform very well, like the Discrete Cosine Transform (DCT) [35] for example, are available. Our goal is to provide dictionaries based on factorizations like (15) and (29) that perform well in terms of representation error with a small number m of basic transforms in their composition. All algorithms run for $K = 150$ iterations and there are $N = 12288$ image patches in the dataset $\mathbf{Y}$ each of size $n = 64$.

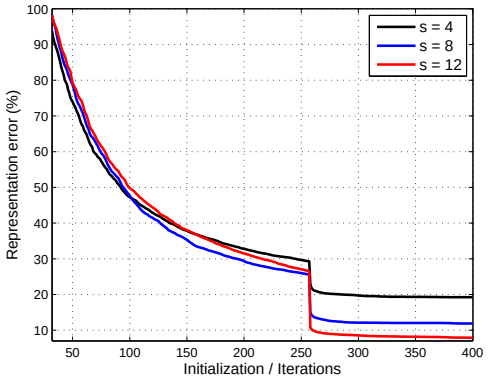

Fig. 1. For the proposed G_{256}–DLA we show the evolution of the relative representation error (42) for the dataset $\mathbf{Y}$ created from the patches of the images couple, peppers and boat with sparsity $s \in \{4, 8, 12\}$. The first 256 points in the plot are due to the initialization step ($m = 256$ transforms are initialized) and the other $K = 150$ are the regular iterations of G_{256}–DLA.

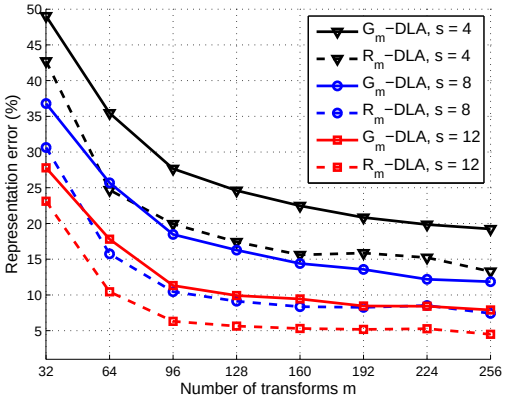

Fig. 2. Performance of G_m–DLA and R_m–DLA in terms of the relative representation error (42) for different sparsity levels $s \in \{4, 8, 12\}$.

To measure the quality of a dictionary $\mathbf{D}$ we choose to evaluate the relative representation error

$$\epsilon = \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 \|\mathbf{Y}\|_F^{-2} \ (\%). \tag{42}$$

Figure 1 shows the evolution of G_{256}–DLA for $K = 150$ iterations (including the initialization procedure, i.e., the first 256 steps of the algorithm). The figure shows how effective the initialization is in reducing the representation error for any sparsity level. Notice that the initialization procedure provides similar results regardless of the sparsity level s. The $K = 150$ iterations of G_{256}–DLA further lower the representation error providing better results with larger sparsity level. As previously discussed, each step of the algorithm monotonically decreases the objective function of the dictionary learning problem until convergence.

Figure 2 shows how G_m–DLA evolves with the number of transforms m and the sparsity s. As expected, increasing the number of transforms $\mathbf{G}_{ij}$ and $\mathbf{R}_{ij}$ in the factorization always lowers the representation error but with diminishing returns as m increases. This figure helps choose the number of transforms m while balancing between the computational

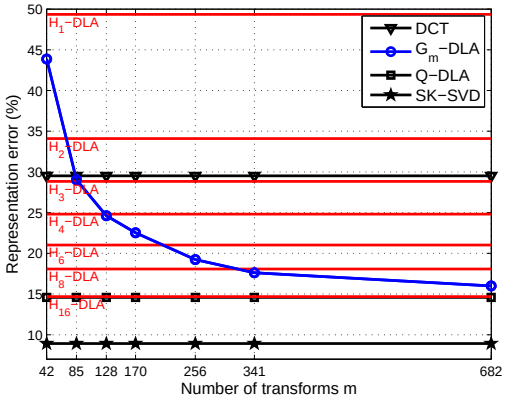

Fig. 3. Comparisons, in terms of relative representation errors (42), of G_m–DLA against the DCT, Q–DLA [28], SK–SVD [34] and Householder based orthonormal dictionaries [21] denoted here H_p–DLA where p is the number of reflectors in the factorization of the dictionary. The number of transforms m is chosen so that computational complexity comparisons against H_p–DLA is possible. Computational complexity approximately match between: H_1–DLA and G_{42}–DLA, H_2–DLA and G_{85}–DLA, H_3–DLA and G_{128}–DLA, H_4–DLA and G_{170}–DLA, H_6–DLA and G_{256}–DLA, H_8–DLA and G_{341}–DLA, H_{16}–DLA and G_{682}–DLA. The sparsity level is set to $s = 4$ for all methods.

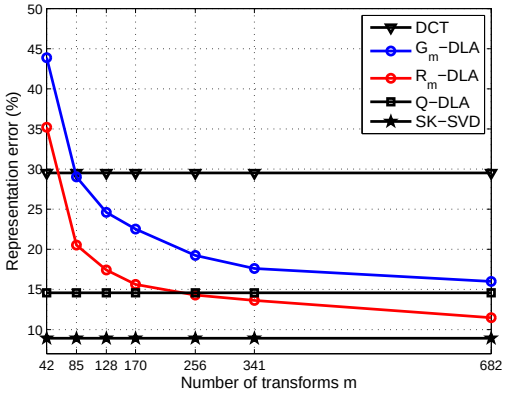

Fig. 4. For the same experimental setup as in Figure 3 we compare G_m–DLA against R_m–DLA.

complexity and representation performance. Large decreases in the representation error are seen up to $m = 96$ or $m = 128$ while thereafter increasing m brings smaller benefits. Also, with higher sparsity levels the number of transforms m becomes less relevant. We notice that with $s = 12$ the representation performance hits a plateau after $m \geq 128$ transforms.

An interesting point of comparison is between the dictionaries constructed via G_m–DLA and H_p–DLA [21]. Figure 3 provides a detailed comparison between the two. Each Householder reflector has n degrees of freedom while each G-transform has only 2. A matrix-vector multiplication takes $4n$ operations for a reflector and only 6 operations for a G-transform. If we compare the computational complexities of the dictionaries constructed by the two methods we find approximate equality between H_p–DLA and $G_{\lfloor \frac{2}{3} np \rfloor}$–DLA. Notice from this figure that for a low m the G-transform approach provides better results than the Householder approach while also enjoying lower computational complexity. As the complexity of the dictionaries increases (larger number of G-transforms or reflectors) the gap between the two approaches decreases. The most complex dictionaries are designed via H_{16}–DLA and G_{682}–DLA and they closely match the performance of the general orthonormal dictionary learning approach Q–DLA while still keeping a computational advantage. In this case, the Householder approach keeps a slight edge in representation performance. Since the proposed approach updates the G-transforms sequentially the probability of getting stuck in local minimum points is more likely with large m. The difficulties that G_m–DLA encounters for large m are also discussed in Remark 4.

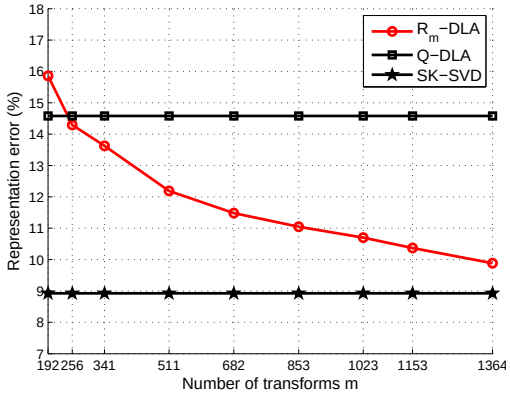

Fig. 5. The evolution of R_m–DLA for m large enough to outperform any orthonormal dictionary.

It is also interesting to see that the representation performance of the DCT is matched by H_3–DLA and G_{85}–DLA. The computational complexity of H_3–DLA approximately matches that of the DCT [35] (based on the FFT) while G_{85}–DLA is actually computationally simpler than the DCT. In fact, any dictionary constructed by G_m–DLA for $85 \leq m \leq 128$ is faster and provides better representations than the DCT.

Figure 4 compares the G_m–DLA against the R_m–DLA for the same number of transforms m in their factorizations. R_m–DLA always outperforms G_m–DLA since the $\mathbf{G}_{ij}$ is a constrained version of $\mathbf{R}_{ij}$. Unfortunately, the non-orthonormal transforms also have much higher computational complexity than their orthonormal counterparts in the sparse approximation step. For example, the computational complexity is approximately equivalent between dictionaries designed via R_{42}–DLA and G_{351}–DLA. The main benefit of non-orthonormal transforms is that ultimately, for large enough m, their performance surpasses that of general (computationally inefficient) orthonormal transforms designed via Q–DLA. In our case this happens for $m \geq 256$. The performance of the DCT is approximately matched by R_{50}–DLA. Surprisingly, less than n factors in the product of the transform suffice to match the performance of the classical DCT transform for sparse recovery.

Figure 5 shows the performance of R_m–DLA in a regime close to the results of the SK–SVD dictionary learning method [34]. The complexity of the dictionary designed via R_{1364}–DLA matches that of the dictionary designed via SK–SVD while there is a small performance gap between the two.

We notice experimentally that the iterative procedure of R_m–DLA improves performance always when increasing m but the probability of getting stuck in local minimum points increases. Therefore, just as G_m–DLA has some trouble matching the performance of Q–DLA, R_m–DLA has some trouble exactly matching the performance of SK–SVD.

When designing a very fast orthonormal transform (whose complexity let us say is order n or $n \log n$) then G_m–DLA provides very good results while achieving the lowest computational complexity. For improved performance, more complex orthonormal transforms perform better when designed via H_p–DLA. If representation capabilities is the only performance metric then the non-orthonormal transforms designed by R_m–DLA are the weapon of choice. For large m both G_m–DLA and R_m–DLA can suffer from long running times. For example, G_{128}–DLA takes several minutes to terminate while R_{128}–DLA's running time is close to ten minutes on a modern Intel i7 computer system. We note that the algorithms are implemented in Matlab®. A careful implementation in a lower level compiled programming language will drive these running times much lower and reduce the memory footprint.

VII. Conclusions

In this paper we present practical procedures to learn square orthonormal and non-orthonormal dictionaries already factored into a fixed number of computationally efficient blocks. We show how effective the dictionaries constructed via the proposed methods are on image data where we compare against the fast cosine transform on one side and general non-orthonormal and orthonormal dictionaries on the other. We also show comparisons with a recently proposed method that constructs Householder based orthonormal dictionaries. We show empirically that the proposed methods construct transforms that provide an improved trade-off between computational complexity and representation performance among the methods we consider. We are able to construct transforms that exhibit lower computational efficiency and lower representation error than the fast cosine transform for image data. We expect the current work to extend the use of learned transforms in time critical scenarios and to devices where, due to power limitations, only low complexity algorithms can be deployed.

References

[1] I. Tosic and P. Frossard, "Dictionary learning," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 27–38, 2011.

[2] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM Review*, vol. 51, pp. 34–81, 2009.

[3] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. Image Proc.*, vol. 15, no. 12, pp. 3736–3745, 2006.

[4] C. Rusu, R. Mendez-Rial, N. Gonzalez-Prelcic, and R. W. Heath, "Low complexity hybrid sparse precoding and combining in millimeter wave MIMO systems," in *Proc. IEEE ICC*, 2015.

[5] J. Mairal, F. Bach, and J. Ponce, "Task-driven dictionary learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 791–804, 2012.

[6] A. M. Tillmann, "On the computational intractability of exact and approximate dictionary learning," *IEEE Signal Processing Letters*, vol. 22, no. 1, pp. 45–49, 2014.

[7] K. Engan, S. O. Aase, and J. H. Husøy, "Method of optimal directions for frame design," in *Proc. IEEE ICASSP*, 1999, pp. 2443–2446.

[8] M. Aharon, M. Elad, and A. M. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Sig. Proc.*, vol. 54, no. 11, pp. 4311–4322, 2006.

[9] ——, "On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them," *Linear Algebra and Applications*, vol. 416, pp. 48–67, 2006.

[10] A. Rakotomamonjy, "Direct optimization of the dictionary learning problem," *IEEE Trans. Sig. Proc.*, vol. 61, no. 22, pp. 5495–5506, 2013.

[11] J. Cooley and J. Tukey, "An algorithm for the machine calculation of complex Fourier series," *Mathematics of Computation*, vol. 19, no. 90, pp. 297–301, 1965.

[12] J. A. Tropp, "Just relax: Convex programming methods for subset selection and sparse approximation," *IEEE Trans. Inf. Theory*, vol. 52, pp. 1030–1051, 2006.

[13] ——, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Inf. Theory*, vol. 50, pp. 2231–2242, 2004.

[14] R. Rubinstein, M. Zibulevsky, and M. Elad, "Efficient implementation of the K-SVD algorithm using batch orthogonal matching pursuit," *CS Technion*, 2008.

[15] M. Mathieu and Y. LeCun, "Fast approximation of rotations and Hessians matrices," *arXiv:1404.7195*, 2014.

[16] R. Rubinstein, M. Zibulevsky, and M. Elad, "Double sparsity: learning sparse dictionaries for sparse signal approximation," *IEEE Trans. Sig. Proc.*, vol. 58, no. 3, pp. 1553–1564, 2010.

[17] L. L. Magoarou and R. Gribonval, "Learning computationally efficient dictionaries and their implementation as fast transforms," *arXiv:1406.5388*, 2014.

[18] O. Chabiron, F. Malgouyres, J.-Y. Tourneret, and N. Dobigeon, "Toward fast transform learning," *Technical report*, 2013.

[19] C. Rusu, B. Dumitrescu, and S. A. Tsaftaris, "Explicit shift-invariant dictionary learning," *IEEE Signal Proc. Let.*, vol. 21, no. 1, pp. 6–9, 2014.

[20] C. Rusu and B. Dumitrescu, "Block orthonormal overcomplete dictionary learning," in *21st European Sig. Proc. Conf.*, 2013, pp. 1–5.

[21] C. Rusu, N. Gonzalez-Prelcic, and R. Heath, "Fast orthonormal sparsifying transforms based on Householder reflectors," *IEEE Trans. Sig. Process.*, vol. 64, no. 24, pp. 6589–6599, 2016.

[22] G. K. Wallace, "The JPEG still picture compression standard," *IEEE Trans. Consumer Electronics*, vol. 38, no. 1, 1992.

[23] A. Agarwal, A. Anandkumar, P. Jain, P. Netrapalli, and R. Tandon, "Learning sparsely used overcomplete dictionaries," in *JMLR Workshop and Conference Proceedings*, vol. 35, 2014, pp. 1–15.

[24] O. G. Sezer, O. Harmanci, and O. G. Guleryuz, "Sparse orthonormal transforms for image compression," in *Proc. IEEE ICIP*, 2008, pp. 149–152.

[25] O. G. Sezer, O. G. Guleryuz, and Y. Altunbasak, "Approximation and compression with sparse orthonormal transforms," *IEEE Trans. Image Proc.*, vol. 24, no. 8, pp. 2328–2343, 2015.

[26] A. Dremeau and C. Herzet, "An EM-algorithm approach for the design of orthonormal bases adapted to sparse representations," in *Proc. IEEE ICASSP*, 2010, pp. 2046–2049.

[27] H. Schutze, E. Barth, and T. Martinetz, "Learning efficient data representations with orthogonal sparse coding," *IEEE Trans. Comp. Imaging*, vol. 2, no. 3, pp. 177–189, 2016.

[28] S. Lesage, R. Gribonval, F. Bimbot, and L. Benaroya, "Learning unions of orthonormal bases with thresholded singular value decompositon," in *Proc. IEEE ICASSP*, 2005, pp. 293–296.

[29] P. Schonemann, "A generalized solution of the orthogonal Procrustes problem," *Psychometrika*, vol. 31, no. 1, pp. 1–10, 1966.

[30] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Johns Hopkins University Press, 1996.

[31] C. Guangzhi, L. R. Bachega, and C. A. Bouman, "The sparse matrix transform for covariance estimation and analysis of high dimensional signals," *IEEE Trans. Image Processing*, vol. 20, no. 3, pp. 625–640, 2011.

[32] R. Kondor, N. Teneva, and V. Garg, "Multiresolution matrix factorization," in *Proc. of the 31st International Conference on Machine Learning*, 2014, pp. 1620–1628.

[33] C. Rusu and B. Dumitrescu, "An initialization strategy for the dictionary learning problem," in *Proc. IEEE ICASSP*, 2014, pp. 6731–6735.

[34] ——, "Stagewise K-SVD to design efficient dictionaries for sparse representations," *IEEE Signal Processing Letters*, vol. 19, no. 10, pp. 631–634, 2012.

[35] W.-H. Chen, C. H. Smith, and S. C. Fralick, "A fast computational algorithm for the discrete cosine transform," *IEEE Trans. Communications*, vol. 25, no. 9, pp. 1004–1009, 1977.