Fast Orthonormal Sparsifying Transforms Based on Householder Reflectors

Citation for published version:
Rusu, C, Gonzalez-Prelcic, N & Heath, R 2016, 'Fast Orthonormal Sparsifying Transforms Based on Householder Reflectors', *IEEE Transactions on Signal Processing*, vol. 64, no. 24, pp. 6589 - 6599. https://doi.org/10.1109/TSP.2016.2612168

Digital Object Identifier (DOI):
10.1109/TSP.2016.2612168

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
IEEE Transactions on Signal Processing

Fast Orthonormal Sparsifying Transforms Based on Householder Reflectors

Cristian Rusu, Nuria González-Prelcic and Robert W. Heath Jr.

Abstract—**Dictionary learning is the task of determining a data-dependent transform that yields a sparse representation of some observed data. The dictionary learning problem is non-convex, and usually solved via computationally complex iterative algorithms. Furthermore, the resulting transforms obtained generally lack structure that permits their fast application to data. To address this issue, this paper develops a framework for learning orthonormal dictionaries which are built from products of a few Householder reflectors. Two algorithms are proposed to learn the reflector coefficients: one that considers a sequential update of the reflectors and one with a simultaneous update of all reflectors that imposes an additional internal orthogonal constraint. The proposed methods have low computational complexity and are shown to converge to local minimum points which can be described in terms of the spectral properties of the matrices involved. The resulting dictionaries balance between the computational complexity and the quality of the sparse representations by controlling the number of Householder reflectors in their product. Simulations of the proposed algorithms are shown in the image processing setting where well-known fast transforms are available for comparisons. The proposed algorithms have favorable reconstruction error and the advantage of a fast implementation relative to the classical, unstructured, dictionaries.**

Index Terms—**sparsifying transforms, fast transforms, dictionary learning, compressed sensing.**

I. Introduction

Sparsifying transforms [1] allow efficient representation of data when a data-dependent overcomplete dictionary is available. Overcomplete dictionaries are useful in image processing [2], [3], [4], speech processing [5] and wireless communications [6], [7]. Unfortunately, the selection of a sparsifying transform involves solving a non-convex optimization problem for a dictionary matrix $\mathbf{D}$ such that a real data set can be represented with a sparse representation matrix $\mathbf{X}$ whose sparsity level is constrained. Because direct solution of the optimization method is difficult [8], [9], proposed algorithms seek a suboptimal solution via alternating minimization.

Most prior work considers alternating minimization for dictionaries that are overcomplete. Algorithms like the method of optimal directions (MOD) [10], K-SVD [11] and algorithms based on direct optimization [12] all perform alternating updates, differing in the ways they actually perform the update of the dictionary and of the sparse representations. Unfortunately, most general solutions to the dictionary learning problem are relatively slow in computing the dictionary and they lack a formal analysis of performance. Some of these difficulties stem from the fact that the proposed algorithms produce non-orthonormal, even overcomplete, dictionaries. Furthermore, overcomplete transforms themselves present some disadvantages when compared to the classical, fixed and fast, transforms. In any application, one general drawback of these computed dictionaries is that they need to be stored (or transmitted) along with the encoded/compressed data. Another, and more important, drawback is that representing vectors in a non-orthonormal or overcomplete dictionary involves a non-linear, computationally expensive, procedure [13], [14].

Fast transforms allow more efficient application of the dictionary to compute the sparse representation. For example, the discrete cosine, Fourier, Hadamard or wavelet transforms all have computationally efficient implementations, i.e., for example $O(n \log n)$ computational complexity [15]. These fast transforms are widely used in signal and image processing but unfortunately are not the best sparsifying transforms in every situation.

Recent work has devised fast sparsifying dictionaries that are built from fast transforms. For example, one of the first proposed algorithm called sparse K-SVD [16], considers constructing a dictionary by using sparse linear combinations of the components of a fast transform. These dictionaries are efficient to apply since a linear combination of just a few components (which themselves are computed fast) can be done efficiently. The second, more recent, approach [17] considers factorizing the dictionary as a product of a few very sparse matrices that can be easily manipulated. This is in the spirit of several fixed sparsifying transform that have this property, like the aforementioned Hadamard case which enjoys a factorization as a product of sparse matrices. Other approaches, like the one in [18] treats each atom of the dictionary as the composition of several circular convolutions so that the overall dictionary can be manipulated quickly, via Fourier transforms. The approach in [19] is to construct an overcomplete dictionary from concatenations of several orthonormal sub-dictionaries and partition the sparse representations such that they belong exclusively to only one sub-dictionary. Tree structures have been used to quickly constructing sparse approximations [20]. The learning algorithms proposed in [16]–[20] are slow in general, lack performance analysis or guarantees and usually involve relatively complex

C. Rusu is with the Institute for Digital Communications, School of Engineering, The University of Edinburgh. N. González-Prelcic is with the Atlantic Research Center for Information and Communication Technologies, University of Vigo. R. Heath is with The University of Texas at Austin.

C. Rusu (c.rusu@ed.ac.uk) is the corresponding author. Demo source code available online at https://udrc.eng.ed.ac.uk/sites/udrc.eng.ed.ac.uk/files/attachments/demo.zip.

C. Rusu acknowledges support from the Engineering and Physical Sciences Research Council (EPSRC) [EP/K014277/1] and the MOD University Defence Research Centre (UDRC) on Signal Processing. N. González-Prelcic acknowledges support from the Spanish Government and the European Regional Development Fund (ERDF) under project TACTICA. R. Heath acknowledges support from the National Science Foundation under Grant No. NSF-CCF-1319556.

algorithms and extra data structures for the description of the dictionary. The approach in [21], provides a fast procedure for learning circulant dictionaries but, unfortunately, these dictionaries are not a general solution due to their low number of degrees of freedom.

In this paper we develop algorithms for finding orthonormal dictionaries that can be used directly and inversely faster than the unconstrained, general, orthonormal dictionaries. We reduce the computational complexity of manipulating the dictionaries by considering that they are products of only a few Householder reflectors [22]. While any orthonormal dictionary of size $n \times n$ can be factorized into n reflectors, in this paper we use $m \ll n$ reflectors in the structure of the dictionary. This way, by applying the reflectors sequentially, low complexity dictionary manipulation is achieved. We choose to use Householder reflectors as the building blocks of our dictionaries since they enjoy low complexity manipulation, e.g., the reflector-vector product is computed in $O(n)$. By using fewer reflectors than needed, our algorithms cannot explore the entire space of orthonormal dictionaries rather only a subset of these. The main advantage though is the low complexity manipulation of the dictionaries designed this way. In general, an open question is if all orthonormal and Hessian matrices can be well represented and approximated with low complexity [23] (factored into $(1/2)n \log n$ Givens rotations).

In this paper, we propose two algorithms that compute the coefficients of the Householder reflectors. The first approach builds an orthonormal dictionary composed of just a few reflectors by updating all the coefficients of each reflector sequentially, keeping the other ones fixed. The main advantages of this approach are: (i) each reflector update is done efficiently by solving an eigenvalue problem and (ii) the overall performance of this method approaches the performance of general orthonormal dictionary learning when the number of reflectors increases. Since each reflector is updated individually, this approach is relatively slow due to the large number of matrix manipulations that need to be performed. A natural question is if it possible to decouple the problem such that all reflectors can be updated simultaneously. This idea, which is realized by adding an additional orthogonal constraint of the reflector coefficients, is at the core of the second proposed method. The main benefit of this second approach is that it outperforms the first in terms of running time due to fewer manipulations required, but is slightly inferior in terms of representation quality. Additionally, for this second approach, we are able to perform a detailed performance analysis. While the dictionaries designed by both proposed methods enjoy fast (controllable computational complexity) manipulation the first, slower, approach provides better representation results.

We compare the proposed algorithms in image processing applications, a classical scenario for the evaluation/comparison of sparsifying transforms. We show that the proposed methods cover the full performance range of computational complexity and representation error. Adjusting the number of reflectors in the transform, we can construct anything from dictionaries as fast as the well-known, fixed bases used in image compression with similar representation performance to slower dictionaries that have representation errors matching those of general orthonormal dictionaries. We provide insight into ways of choosing the number of reflectors thus allowing full flexibility to the proposed solutions. Furthermore, we show that in our experimental runs we are always able to construct a fast dictionary that matches the performance of the general orthonormal dictionary with a relative low number of reflectors. Based on these results we conclude that the proposed algorithms are well suited to produce solutions that balance the computational complexity and representation quality of learned dictionaries.

The paper is organized as follows. Section II reviews the concept of orthonormal dictionary learning, Section III presents the proposed algorithms, Section IV provides performance insights into the proposed methods while Section V shows experimentally their effectiveness.

II. General orthonormal dictionary learning

In this section, we review prior work on learning general orthonormal dictionaries and provide some new insights. The objective is to describe the mathematical foundations of the dictionary learning problem, introduce the main notation, formulation and previously proposed solutions.

Given a real dataset $\mathbf{Y} \in \mathbb{R}^{n \times N}$ and sparsity level s, the orthonormal dictionary learning algorithm (which we will call Q–DLA) [24] is formulated as:

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

where the objective function describes the representation error achieved by the orthonormal dictionary $\mathbf{Q} \in \mathbb{R}^{n \times n}$ 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 $s \ll n \ll N$. The problem described in (1) has been extensively studied and used in many applications especially in image processing for compression [25], [26], [27]. Optimizations similar to (1) have been proposed in the past to learn incoherent dictionaries [28] or to build initial dictionaries for the general dictionary learning problem [29].

The solution to (1) proposed in [24] alternates between computing $\mathbf{X}$ and $\mathbf{Q}$ with one of them fixed, just like in the general dictionary learning case [10]. We detail the steps next.

Since the dictionary $\mathbf{Q}$ is orthonormal the sparse representation step 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 columnwise. To select the largest entries, per signal, a fast partial sorting algorithm [30] can be used whose complexity is only $O(n)$.

To solve (1) for variable $\mathbf{Q}$ and fixed $\mathbf{X}$, a problem also known as the orthonormal Procrustes problem [31], a closed form solution $\mathbf{Q} = \mathbf{U}\mathbf{V}^T$ is given by the singular value decomposition (SVD) of $\mathbf{Y}\mathbf{X}^T = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$. Notice that with the representations $\mathbf{X}$ fixed, the reduction in the objective function of (1) achieved by a general orthonormal dictionary $\mathbf{Q}$ is given by:

$$\begin{array}{l} \|\mathbf{Y} - \mathbf{Q}\mathbf{X}\|_F^2 = \|\mathbf{Y}\|_F^2 + \|\mathbf{X}\|_F^2 + C, \\ \text{with } C = -2\text{tr}(\mathbf{Q}^T\mathbf{Y}\mathbf{X}^T). \end{array} \tag{2}$$

Develop further to reach

$$\text{tr}(\mathbf{Q}^T\mathbf{Y}\mathbf{X}^T) = \text{tr}(\mathbf{V}\mathbf{U}^T\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T) = \text{tr}(\mathbf{\Sigma}) = \|\mathbf{Y}\mathbf{X}^T\|_*. \quad (3)$$

Thus, the reduction in the objective function is $2\|\mathbf{Y}\mathbf{X}^T\|_*$. This shows that when considering orthonormal dictionaries, the learning problem can be seen as a nuclear norm maximization with sparsity constraints (and with $\|\mathbf{X}\|_F^2 \le \|\mathbf{Y}\|_F^2$ to avoid trivial unbounded solutions). Also, notice that at the optimum we have the symmetric positive semidefinite

$$\mathbf{Q}^T\mathbf{Y}\mathbf{X}^T = \mathbf{V}\mathbf{\Sigma}\mathbf{V}^T,\ \mathbf{Q}\mathbf{X}\mathbf{Y}^T = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^T. \quad (4)$$

The two are identical since $\mathbf{Q}^T(\mathbf{Q}\mathbf{X}\mathbf{Y}^T)\mathbf{Q} = \mathbf{X}\mathbf{Y}^T\mathbf{Q} = (\mathbf{Q}^T\mathbf{Y}\mathbf{X}^T)^T = \mathbf{Q}^T\mathbf{Y}\mathbf{X}^T$, i.e, $\mathbf{V} = \mathbf{Q}^T\mathbf{U}$.

Remark 1. A positive semidefinite condition for the symmetric $\mathbf{X}(\mathbf{Q}^T\mathbf{Y})^T$, based on the Gershgorin disk theorem, can be stated. Starting from the positive semidefinite condition (4), the focus falls on the spectral properties of the symmetric $\mathbf{R} = \mathbf{X}(\mathbf{Q}^T\mathbf{Y})^T = \mathcal{T}_s(\mathbf{Q}^T\mathbf{Y})(\mathbf{Q}^T\mathbf{Y})^T$. The diagonal elements of this matrix are positive since they are the squared ℓ_2 norms of the rows of $\mathcal{T}_s(\mathbf{Q}^T\mathbf{Y})$ and moreover they have relative large magnitude since the sparse representation step keeps only the largest s entries $\big($in fact $\text{tr}(\mathbf{R}) = \|\mathcal{T}_s(\mathbf{Q}^T\mathbf{Y})\|_F^2 = \|\mathbf{X}\|_F^2\big)$. Therefore, we can assume that $\mathbf{R}$ is diagonally dominant. We also assume that we eliminate zero rows or rows with very few non-zero entries from $\mathbf{R}$, which corresponds to having atoms in the dictionary that are never/rarely used in the representations. To be more precise, let us denote by ϕ_i^T the i^{th} row of $\mathbf{Q}^T\mathbf{Y}$ and with ψ_j^T the j^{th} row of $\mathcal{T}_s(\mathbf{Q}^T\mathbf{Y})$ and we have that $R_{jj} = \psi_j^T\phi_j = \psi_j^T\psi_j$ and $R_{ij} = \psi_i^T\phi_j$ and, by Gershgorin's disk theorem, the conditions for a positive semidefinite $\mathbf{R}$ are:

$$\psi_j^T\psi_j \le \sum_{i=1, i\neq j}^{n} |\psi_j^T\phi_i| \le (n-1)\mu, \quad (5)$$

for $j = 1, \dots, n$ and where $\mu = \max_{i\neq j} |\psi_j^T\phi_i|$.

The result states that if rows of $\mathbf{Q}^T\mathbf{Y}$ are weakly correlated with the rows of $\mathcal{T}_s(\mathbf{Q}^T\mathbf{Y})$, except for the rows with the same indices, then the pair $(\mathbf{Q}, \mathbf{X})$ is a local minimum of the orthonormal dictionary learning problem. ■

Remark 2. Given a dataset $\mathbf{Y}$ and its factorization in a general dictionary $\mathbf{D}$ with sparse representations $\mathbf{X}$, there is no orthonormal transformation $\mathbf{Q}$ such that $\mathbf{Q}\mathbf{D}$ achieves better representation than $\mathbf{D}$ if $\mathbf{Y}\mathbf{D}^T\mathbf{X}^T$ is symmetric.

Proof. Consider the Procrustes optimization problem in variable $\mathbf{Q}$:

$$\underset{\mathbf{Q};\ \mathbf{Q}\mathbf{Q}^T=\mathbf{Q}^T\mathbf{Q}=\mathbf{I}}{\text{minimize}} \|\mathbf{Y} - \mathbf{Q}\mathbf{D}\mathbf{X}\|_F^2, \quad (6)$$

and notice that the minimizer is $\mathbf{Q} = \mathbf{U}\mathbf{U}^T = \mathbf{I}$ given that $\mathbf{Y}\mathbf{D}^T\mathbf{X}^T = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^T$ is symmetric. ■

Remark 3. A necessary condition that a general orthonormal dictionary $\mathbf{Q}$ with representations $\mathbf{X}$ is a local minimum of the dictionary learning problem is that $\|\mathbf{X}\|_F^2 = \|\mathbf{Y}\mathbf{X}^T\|_*$. For a general overcomplete dictionary $\mathbf{D}$ with representations $\mathbf{X}$, the necessary condition reads $\text{tr}(\mathbf{Y}\mathbf{X}^T\mathbf{D}^T) = \|\mathbf{Y}\mathbf{X}^T\mathbf{D}^T\|_*$.

Proof. With the optimum choice of $\mathbf{Q}$ from the Procrustes result, $\mathbf{Q}\mathbf{X}\mathbf{Y}^T$ and $\mathbf{Q}^T\mathbf{Y}\mathbf{X}^T$ are symmetric by (4). By matching the objective function value from (2) with the performance of the orthonormal dictionary from (3), for fixed $\mathbf{X}$ and $\mathbf{Y}$ there is no general orthonormal dictionary $\mathbf{Q}$ that provides better representation performance than the identity dictionary $\mathbf{I}$ if

$$\text{tr}(\mathbf{Y}\mathbf{X}^T) = \|\mathbf{Y}\mathbf{X}^T\|_*, \quad (7)$$

which holds for example whenever $\mathbf{Y}\mathbf{X}^T$ is normal (orthogonal, symmetric or skew-symmetric in general) and positive semidefinite – notice that orthogonal and positive definite just mean that the solution to the Procrustes problem is $\mathbf{Q} = \mathbf{I}$ because $\mathbf{Y}\mathbf{X}^T = \mathbf{I}$. In general, following the same reasoning, we know from (7) that general orthonormal $\mathbf{Q}$ with representations $\mathbf{X} = \mathcal{T}_s(\mathbf{Q}^T\mathbf{Y})$ is a local minimum when

$$\text{tr}(\mathbf{Y}\mathbf{X}^T\mathbf{Q}^T) = \|\mathbf{Y}\mathbf{X}^T\mathbf{Q}^T\|_*. \quad (8)$$

Finally, using the fact that $\|\mathbf{Y}\mathbf{X}^T\mathbf{Q}^T\|_* = \|\mathbf{Y}\mathbf{X}^T\|_*$ and that $\text{tr}(\mathbf{Y}\mathbf{X}^T\mathbf{Q}^T) = \text{tr}(\mathbf{Q}^T\mathbf{Y}\mathbf{X}^T) = \|\mathbf{X}\|_F^2$ we reach

$$\|\mathbf{X}\|_F^2 = \|\mathbf{Y}\mathbf{X}^T\|_*, \quad (9)$$

and therefore the objective function in (2) takes the value $\|\mathbf{Y}\|_F^2 - \|\mathbf{X}\|_F^2$. Equation (8) is also a necessary condition for the local optimality of a general overcomplete dictionary $\mathbf{D}$ with representations $\mathbf{X}$, i.e., $\text{tr}(\mathbf{Y}\mathbf{X}^T\mathbf{D}^T) = \|\mathbf{Y}\mathbf{X}^T\mathbf{D}^T\|_*$ meaning that there is no orthonormal transformation that improves the representation performance of $\mathbf{D}$. ■

Previous work in the literature deals with the description of local minimum $(\mathbf{D}, \mathbf{X})$ of general dictionary learning schemes [32], [33], while other work is concerned with the sample complexity of recovering a dictionary [34], [35], [36], [37] under various statistical assumptions and dictionary dimensions. The general analysis in [38] provides sample complexity estimates to control how much the empirical average deviates from the expected objective functions of matrix factorization problems.

As with any alternating minimization solution, the initialization procedure plays an important role. For Q–DLA, our experimental findings show that a very good initial point is the orthonormal basis $\mathbf{Q}$ created from the SVD of the dataset: $\mathbf{Y} = \mathbf{Q}\mathbf{\Sigma}\mathbf{V}^T$. This choice is also intuitive [29]. A full factorization of $\mathbf{Y}$ is not necessary since we are interested only in the basis $\mathbf{Q}$. As such, a reduced or so called economy size SVD can be performed. Still, depending on the size of the dataset N, this step can become expensive in terms of running time. In this paper we propose to approximate $\mathbf{Q}$ with a new orthonormal basis $\bar{\mathbf{Q}}$ obtained by:

1) Approximate first $\bar{n} \ll n$ principal components in by using iterative methods [39].
2) Complete the partial structure with random components to obtain the full basis. Finalize by QR orthogonalization to get $\bar{\mathbf{Q}}$.

This initialization works well because typically the lowest singular values of a dataset consisting of real world data have low magnitude.

There are several limitations associated with conventional orthonormal dictionaries. Although the sparse representation step is fast when using an orthonormal dictionary, i.e., no matching [13] or basis pursuit [14] is necessary and only correlations need to be computed, the representation performance is inferior to that of general dictionaries while the computational complexity is comparable to these dictionaries. For this reason we now move to explore transform structures that allow for a computationally cheaper orthonormal dictionary without destroying the sparsifying properties.

III. A Householder approach to orthonormal dictionary learning

In this section, we describe our new approach for dictionary learning based on Householder reflectors. We use the same alternative optimization procedure generally used for dictionary learning and described in Section II. Since we are using orthonormal dictionaries, the sparse approximation step is the same, and thus the focus falls on the dictionary update step which is detailed in this section.

Therefore, we start by analyzing the properties of Householder reflectors and then introduce two dictionary learning procedures that build orthonormal dictionaries directly factorized into a product of reflectors. We finish the discussion by making some considerations on the initialization of the proposed methods.

A. Householder reflectors for dictionary learning

Let $\mathbf{u}_1 \in \mathbb{R}^n$ be a normalized vector, i.e., $\|\mathbf{u}_1\|_2 = 1$. We define the orthonormal symmetric Householder reflector $\mathbf{U}_1 \in \mathbb{R}^{n\times n}$ as

$$\mathbf{U}_1 = \mathbf{I} - 2\mathbf{u}_1\mathbf{u}_1^T. \tag{10}$$

The reflector $\mathbf{U}_1$ is completely defined by the vector $\mathbf{u}_1$ and as such they may be used equivalently to refer to the reflector. Given a Householder reflector $\mathbf{U}_1 \in \mathbb{R}^{n\times n}$ and a vector $\mathbf{x} \in \mathbb{R}^n$, the product

$$\mathbf{U}_1\mathbf{x} = \left(\mathbf{I} - 2\mathbf{u}_1\mathbf{u}_1^T\right)\mathbf{x} = \mathbf{x} - 2\mathbf{u}_1(\mathbf{u}_1^T\mathbf{x}) = \mathbf{x} - \nu\mathbf{u}_1, \tag{11}$$

where $\nu = 2\mathbf{u}_1^T\mathbf{x}$. The computational complexity of (11) is $N_{\text{op}} = 4n$, an order of magnitude lower than the general matrix-vector multiplication complexity of $N_{\text{op}} = n(2n-1)$. Given $\mathbf{X} \in \mathbb{R}^{n\times N}$ a result similar to (11) also holds for matrix-matrix multiplication

$$\mathbf{U}_1\mathbf{X} = \left(\mathbf{I} - 2\mathbf{u}_1\mathbf{u}_1^T\right)\mathbf{X} = \mathbf{X} - \mathbf{u}_1\mathbf{v}_1^T, \tag{12}$$

where $\mathbf{v}_1 = 2\mathbf{X}^T\mathbf{u}_1$.

Householder reflectors are often used to introduce zeros in the entries of vectors and to reduce full matrices to upper (or lower) triangular forms with applications to computing least square solutions and QR decompositions. Given a general orthonormal basis $\mathbf{Q} \in \mathbb{R}^{n\times n}$, there exists a sequence of $n-1$ Householder reflectors $\mathbf{U}_j$ such that the following factorization holds:

$$\mathbf{Q} = \mathbf{U}_{n-1}\mathbf{U}_{n-2}\cdots\mathbf{U}_1\mathbf{D}, \tag{13}$$

where $\mathbf{D}$ is a diagonal matrix of size $n\times n$ with entries $D_{ii} = \{\pm 1\}, i = 1,\dots,n$. This result follows from the QR factorization of a unitary matrix with Householder reflectors, and from the facts that an orthonormal upper (or lower) triangular matrix is actually diagonal and a product of unitary matrices is itself orthonormal. In this case the reflectors enjoy additional sparse structure since the reflectors vectors $\mathbf{u}_j$ have the first $j-1$ entries set to zero. In the following section we will consider general reflector vectors without any sparsity assumptions. Furthermore we will consider products of m Householder reflectors with $m \ll n$ which will open the way to orthonormal dictionaries that can be manipulated fast. Related work explores the ways of representing an orthonormal basis [40].

In this section we describe algorithms to learn an orthonormal dictionary $\mathbf{U} \in \mathbb{R}^{n\times n}$ that is a product of a few Householder reflectors, balancing performance and computational complexity. We consider dictionaries with the following structure:

$$\mathbf{U} = \mathbf{U}_m\mathbf{U}_{m-1}\cdots\mathbf{U}_2\mathbf{U}_1, \tag{14}$$

where all $\mathbf{U}_j$ are Householder reflectors and the number m is on the order $O(\log n)$. Of course, we have that all $\|\mathbf{u}_j\|_2 = 1$. For brevity we do not copy these constraints, but consider them imposed.

B. Learning products of Householder reflectors: an extra orthonormal constraint

We first explore matrix structures that allow for the simultaneous update of all reflectors in the product $\mathbf{U}$. We keep the same overall dictionary formulation as in (14) but with the additional constraint that the reflector vectors obey $\mathbf{u}_i^T\mathbf{u}_j = 0$ for all $i \neq j$. With this orthogonal constraint the new overall orthonormal symmetric dictionary is

$$\mathbf{U} = \mathbf{U}_m\mathbf{U}_{m-1}\cdots\mathbf{U}_2\mathbf{U}_1 = \mathbf{I} - 2\sum_{j=1}^{m}\mathbf{u}_j\mathbf{u}_j^T. \tag{15}$$

Using the fact that the reflector vectors $\mathbf{u}_j$ are orthogonal, the objective function simplifies as

$$\begin{aligned}\|\mathbf{Y} - \mathbf{U}\mathbf{X}\|_F^2 &= \left\|\mathbf{Y} - \mathbf{X} + 2\sum_{j=1}^{m}\mathbf{u}_j\mathbf{u}_j^T\mathbf{X}\right\|_F^2 \\ &= \|\mathbf{Y} - \mathbf{X}\|_F^2 + \sum_{j=1}^{m}\mathbf{u}_j^T\mathbf{Z}\mathbf{u}_j,\end{aligned} \tag{16}$$

where we have defined

$$\mathbf{Z} = 2(\mathbf{X}\mathbf{Y}^T + \mathbf{Y}\mathbf{X}^T) = 2\tilde{\mathbf{Z}}. \tag{17}$$

To minimize (16), the reflector vectors $\mathbf{u}_j$ are chosen to be the eigenvectors associated with the lowest m negative eigenvalues of $\mathbf{Z}$ (assuming that m negative eigenvalues of $\mathbf{Z}$ exist). Since $\mathbf{Z}$ is symmetric, its eigenvectors are orthonormal and thus obey the constraint that we consider on the reflector vectors $\mathbf{u}_j$. If $\mathbf{Z}$ does not possess m negative eigenvectors then fewer than m reflectors should be constructed, the rest up to m can be set to the zero vector (the reflector becomes the identity).

The full proposed learning procedure, which we call QH$_m$–DLA, is detailed in Algorithm 1. Notice that the product $\mathbf{U}^T\mathbf{Y}$ in the computation of $\mathbf{X}$, step 3) of the iterative process, can be efficiently carried out by using the Householder factorization of $\mathbf{U}$ (complexity $O(nN\log n)$ instead of $O(n^2N)$). This is due to the numerical efficiency of the dictionary $\mathbf{U}$.

The updates of the reflectors in $\mathbf{U}$ and of the sparse representations $\mathbf{X}$ are done exactly at each alternating step

Algorithm 1 – QH$_m$–DLA (Orthogonal Householder Dictionary Learning Algorithm).

Input: The dataset $\mathbf{Y} \in \mathbb{R}^{n \times N}$, the number of Householder reflectors in the transform m, the target sparsity s and the maximum number of iterations K.

Output: The sparsifying transform $\mathbf{U} = \mathbf{U}_m \cdots \mathbf{U}_1$ with $\mathbf{u}_i^T \mathbf{u}_j = 0,\ i \neq j$ 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 size $m+1$ of the dataset $\mathbf{Y} = \mathbf{Q}\mathbf{\Sigma}\mathbf{V}^T$.
    2) Reduce $\mathbf{Q} \in \mathbb{R}^{n \times (m+1)}$ to an upper triangular form with Householder reflectors defined by $\mathbf{u}_1, \ldots, \mathbf{u}_m$. The reflector that introduces zeros in the first column is $\mathbf{u}_m$.
    3) Orthogonalize $\mathbf{u}_1, \ldots, \mathbf{u}_m$ by the QR algorithm.
    4) Compute sparse representations $\mathbf{X} = \mathcal{T}_s(\mathbf{U}^T\mathbf{Y})$.
Iterations $1, \ldots, K$:
    1) Construct the matrix: $\tilde{\mathbf{Z}} = \mathbf{X}\mathbf{Y}^T + \mathbf{Y}\mathbf{X}^T$.
    2) Compute the $m$ lowest eigenvalue/eigenvector pairs of $\tilde{\mathbf{Z}}$. Set to $\mathbf{0}$ the eigenvectors associated to nonnegative eigenvalues. Update reflector vectors $\mathbf{u}_j$ with the eigenvectors just computed. The eigenvector of the lowest negative eigenvalue goes to $\mathbf{u}_m$.
    3) Compute sparse representations $\mathbf{X} = \mathcal{T}_s(\mathbf{U}^T\mathbf{Y})$.
```

of the algorithm and thus the objective function decreases monotonically to a local optimum.

Additionally, QH$_m$–DLA is satisfactory from a theoretically perspective since, as we will see, it allows performance analysis and comparison with Q–DLA. Furthermore, notice that the orthonormal dictionaries created by QH$_m$–DLA are also symmetric.

If we consider a general dictionary $\mathbf{D}$ for sparse representations, then the pair dictionary/representations $(\mathbf{D}, \mathbf{X})$ is equivalent to the pair $(-\mathbf{D}, -\mathbf{X})$ [41]. In our setup, notice that if $\mathbf{U}_1$ is a Householder reflector then $-\mathbf{U}_1$ cannot be constructed by (10), as $\mathbf{U}_1$ is. Now assume that the matrix $\mathbf{T} = \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \ldots & \mathbf{u}_n \end{bmatrix}$ contains all n eigenvectors of $\tilde{\mathbf{Z}}$ ordered in increased order of their corresponding eigenvalues. Let $\mathbf{T}_{i:j}$ denote a matrix consisting of all the reflector vectors from the i^{th} to the j^{th} column of $\mathbf{T}$. Then, due to $\mathbf{T}_{1:i}\mathbf{T}_{1:i}^T + \mathbf{T}_{i+1:n}\mathbf{T}_{i+1:n}^T = \mathbf{I}$, we have that:

$$-(\mathbf{I} - 2\mathbf{T}_{1:i}\mathbf{T}_{1:i}^T) = \mathbf{I} - 2\mathbf{T}_{i+1:n}\mathbf{T}_{i+1:n}^T. \tag{18}$$

This shows that there is a correspondence in performance according to the number of reflectors that are selected: with the first m reflectors we have the dictionary $\mathbf{U}$ (and representations $\mathbf{X}$) while with the other $n - m$ reflectors we have the dictionary $-\mathbf{U}$ (and representations $-\mathbf{X}$).

C. Learning products of Householder reflectors: the unconstrained case

We again consider the case where the dictionary $\mathbf{U}$ has the structure from (14) but now no additional constraints are assumed on the reflectors. This time we update each reflector sequentially. The new objective function becomes $\|\mathbf{Y} - \mathbf{U}_m\mathbf{U}_{m-1}\cdots\mathbf{U}_2\mathbf{U}_1\mathbf{X}\|_F^2$. To optimize the j^{th} Householder reflector, we write the objective function as

$$\|\left(\mathbf{U}_{j+1}\cdots\mathbf{U}_m\right)\mathbf{Y} - \mathbf{U}_j\left(\mathbf{U}_{j-1}\cdots\mathbf{U}_1\right)\mathbf{X}\|_F^2, \tag{19}$$

where we have used that all unitary matrices, and thus Householder reflectors, preserve the Frobenius norm and the fact that the reflectors are symmetric:

$$\|\mathbf{Y} - \mathbf{U}_1\mathbf{X}\|_F^2 = \|\mathbf{U}_1^T\mathbf{Y} - \mathbf{X}\|_F^2 = \|\mathbf{U}_1\mathbf{Y} - \mathbf{X}\|_F^2. \tag{20}$$

We have now reduced the problem to the QH$_1$–DLA case for the updated dataset $\left(\mathbf{U}_{j+1}\cdots\mathbf{U}_m\right)\mathbf{Y}$ and the updated representations $\left(\mathbf{U}_{j-1}\cdots\mathbf{U}_1\right)\mathbf{X}$. Following the same computation that leads to (17), we now reach that the best update for the fixed $\mathbf{u}_j$ is the eigenvector associated with the lowest negative eigenvalue of

$$\begin{aligned}\mathbf{Z} =& 2\left(\mathbf{U}_{j-1}\cdots\mathbf{U}_1\right)\mathbf{X}\mathbf{Y}^T\left(\mathbf{U}_{j+1}\cdots\mathbf{U}_m\right)^T + \\ & 2\left(\mathbf{U}_{j+1}\cdots\mathbf{U}_m\right)\mathbf{Y}\mathbf{X}^T\left(\mathbf{U}_{j-1}\cdots\mathbf{U}_1\right)^T.\end{aligned} \tag{21}$$

Each reflector in the product of $\mathbf{U}$ is updated sequentially in this manner. The full procedure, which we call H$_m$–DLA, is detailed in Algorithm 2. We expect the performance of this algorithm to be in general inferior to that of Q–DLA in terms of representation error, approaching it as m approaches n, and to be superior to that of QH$_m$–DLA, due to the missing additional orthogonal constraints. Still, since all reflectors are computed together and no extensive matrix manipulation is required QH$_m$–DLA runs faster than H$_m$–DLA. This opens the possibility of using QH$_m$–DLA as an initialization procedure for H$_m$–DLA. Finally, QH$_1$–DLA and H$_1$–DLA are equivalent. Also notice that the computation of $\tilde{\mathbf{Z}}$ can be optimized across the iterative process in step 1a): denote $\mathbf{R}_j = \left(\mathbf{U}_{j-1}\cdots\mathbf{U}_1\right)\mathbf{X}\mathbf{Y}^T\left(\mathbf{U}_m\cdots\mathbf{U}_{j+1}\right)$ from the j^{th} iteration, the for the next iteration when computing $\mathbf{U}_{j+1}$ we simply have that $\mathbf{R}_{j+1} = \mathbf{U}_j\mathbf{R}_j\mathbf{U}_{j+1}^T$ – which can be done efficiently by left and right reflector multiplication formulas.

Just as in the case of QH$_m$–DLA, the update of each reflector $\mathbf{U}_j$ and of the representations $\mathbf{X}$ are done by solving exactly the optimization problems (with the other variables fixed) and thus the objective function monotonically decreases to a local minimum point.

D. The initializations of H$_m$–DLA and QH$_m$–DLA

Initialization is important for any alternating minimization algorithm. In principle, the proposed methods can be initialized with random reflectors $\mathbf{u}_j$ but the idea is to provide an initialization such that the methods converge in few iterations. The computational complexity of the initialization should be much lower than that of the learning algorithms.

In both the cases of H$_m$–DLA and QH$_m$–DLA, the initialization procedures start by computing the reduced singular value decomposition of size m of the dataset $\mathbf{Y} = \mathbf{Q}\mathbf{\Sigma}\mathbf{V}^T$. Then $\mathbf{Q}$ is diagonalized by Householder reflectors thus providing the n reflectors. Among these we choose m reflectors to initialize our algorithms. In the case of QH$_m$–DLA the reflectors previously obtained are further orthogonalized by the QR algorithm thus ensuring compliance with all the constraints of the method.

Algorithm 2 – H$_m$–DLA (Householder Dictionary Learning Algorithm).

Input: The dataset $\mathbf{Y} \in \mathbb{R}^{n \times N}$, the number of Householder reflectors in the transform m, the target sparsity s and the maximum number of iterations K.

Output: The sparsifying transform $\mathbf{U} = \mathbf{U}_m \cdots \mathbf{U}_1$ 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 size $m+1$ of the dataset $\mathbf{Y} = \mathbf{Q}\mathbf{\Sigma}\mathbf{V}^T$.
    2) Reduce $\mathbf{Q} \in \mathbb{R}^{n \times (m+1)}$ to an upper triangular form by Householder reflectors defined by $\mathbf{u}_1, \dots, \mathbf{u}_m$. The reflector that introduces zeros in the first column is $\mathbf{u}_m$.
    3) Compute sparse representations $\mathbf{X} = \mathcal{T}_s(\mathbf{U}^T\mathbf{Y})$.
Iterations $1, \dots, K$:
    1) For $j = 1, \dots, m$:
        a) Construct the matrix:
           $\tilde{\mathbf{Z}} = (\mathbf{U}_{j-1} \cdots \mathbf{U}_1)\mathbf{X}\mathbf{Y}^T(\mathbf{U}_{j+1} \cdots \mathbf{U}_m)^T$,
           $\tilde{\mathbf{Z}} = \tilde{\mathbf{Z}} + \tilde{\mathbf{Z}}^T$.
        b) Compute lowest eigenvalue $\lambda_{\text{min}}$ of $\tilde{\mathbf{Z}}$ with eigenvector $\mathbf{v}$. If $\lambda_{\text{min}} \geq 0$ then set $\mathbf{v} = \mathbf{0}$. Update reflector vector $\mathbf{u}_j = \mathbf{v}$.
    2) Compute sparse representations $\mathbf{X} = \mathcal{T}_s(\mathbf{U}^T\mathbf{Y})$.
```

IV. Comments on the proposed algorithms and connections to previous work

Now that the main algorithms have been described, in this section we examine the achievable representation performance of Householder based dictionaries. First, we analyze the simple case of a single Householder reflector dictionary (analysis that is pertinent also to each step of the H$_m$–DLA) and then consider the QH$_m$–DLA. Finally, we show the similarities between the representation error achievable by our proposed dictionaries and that of general orthonormal dictionaries.

A. Performance of a single Householder reflector dictionary

Considering a dictionary composed of a single Householder reflector, the objective function in (16) reduces to

$$\begin{aligned} &\|\mathbf{Y} - \mathbf{U}_1\mathbf{X}\|_F^2 = \|\mathbf{Y}\|_F^2 + \|\mathbf{X}\|_F^2 + C, \\ &\text{with } C = -2\text{tr}(\mathbf{X}\mathbf{Y}^T) + 2\mathbf{u}_1^T(\mathbf{X}\mathbf{Y}^T + \mathbf{Y}\mathbf{X}^T)\mathbf{u}_1. \end{aligned} \tag{22}$$

Assuming some normalization of the dataset like mean subtraction and ℓ_2 normalization of the columns, it is reasonable to consider $\|\mathbf{Y}\|_F^2 = N$. The norm $\|\mathbf{X}\|_F^2$ is maximized in the sparse reconstruction step, where we keep the largest absolute value entries in the representations. The goal is twofold:

- Maximize the trace of $\mathbf{X}\mathbf{Y}^T$.
- Minimize the lowest eigenvalue of $\tilde{\mathbf{Z}} = \mathbf{X}\mathbf{Y}^T + \mathbf{Y}\mathbf{X}^T$.

The two goals are related since tr$(\tilde{\mathbf{Z}}) = 2$tr$(\mathbf{X}\mathbf{Y}^T)$. Therefore, the performance of our algorithms depends on the spectral properties of $\tilde{\mathbf{Z}}$. In an ideal situation, the lowest, negative, eigenvalue of this matrix should be maximally reduced while the rest of the eigenvalues remain positive and their sum is maximized. An ideal case would be that the spectrum obeys $\Lambda(\tilde{\mathbf{Z}}) = \{-\alpha_1, \beta_1, \dots, \beta_{n-1}\}$, one negative eigenvalue and $n-1$ non-negative. Now the cost in (22) is maximally reduced by the sum of the singular values of $\tilde{\mathbf{Z}}$ also known as its nuclear norm, i.e., $C = -\|\tilde{\mathbf{Z}}\|_* = -\left(\alpha_1 + \sum_{i=1}^{n-1}\beta_i\right)$.

B. Performance of Householder based dictionaries

We now analyze the dictionaries created by QH$_m$–DLA. In the case of H$_m$–DLA, since the reflectors are updated sequentially, we defer to the discussion for QH$_1$–DLA.

The case that can be more easily approached from an analysis perspective is the one of QH$_m$–DLA, where all reflectors are updated simultaneously. In this case, the objective function (16) reduces to

$$\begin{aligned} &\|\mathbf{Y} - \mathbf{U}\mathbf{X}\|_F^2 = \|\mathbf{Y}\|_F^2 + \|\mathbf{X}\|_F^2 + C, \\ &\text{with } C = -2\text{tr}(\mathbf{X}\mathbf{Y}^T) + 2\sum_{j=1}^{m}\mathbf{u}_j^T(\mathbf{X}\mathbf{Y}^T + \mathbf{Y}\mathbf{X}^T)\mathbf{u}_j. \end{aligned} \tag{23}$$

Similar to the single Householder reflector case, the performance depends on the spectrum $\Lambda(\tilde{\mathbf{Z}}) = \{-\alpha_1, \dots, -\alpha_m, \beta_1, \dots, \beta_{n-m}\}$. To minimize the objective function in (23), we need to choose m Householder reflectors corresponding to the m negative eigenvalues in $\Lambda(\tilde{\mathbf{Z}})$. In this way, (23) is maximally reduced by the nuclear norm of $\tilde{\mathbf{Z}}$, i.e., $C = -\|\tilde{\mathbf{Z}}\|_* = -\left(\sum_{i=1}^{m}\alpha_i + \sum_{i=1}^{n-m}\beta_i\right)$.

If the spectrum of $\tilde{\mathbf{Z}}$ is non-negative, then no reflector can decrease the objective function and the dictionary is set to $\mathbf{U} = \mathbf{I}$; with the given $\mathbf{Y}$ and $\mathbf{X}$ there is no Householder reflector that can improve upon the representation performance. Equally bad, if the spectrum is non-positive then all n eigenvectors are selected and by (15) it follows that the dictionary is $\mathbf{U} = -\mathbf{I}$. In practice, depending on the magnitude of all the m negative eigenvalues of $\tilde{\mathbf{Z}}$ we may choose a smaller number of reflectors to construct $\mathbf{U}$. Of course, the representation performance is slightly inferior this way but the benefit is a faster transform. The trade-off can be balanced based on application specific requirements.

A situation that is of interest is when the sparse factorization can be done exactly, i.e., $\mathbf{Y} = \mathbf{U}\mathbf{X}$. Considering that some normalization has taken place for the dataset such that $\|\mathbf{Y}\|_F^2 = N$ and because orthonormal transformations preserve ℓ_2 norms we have that $\|\mathbf{X}\|_F^2 = N$, i.e., we have in effect exactly $\mathbf{X} = \mathbf{U}^T\mathbf{Y}$. The objective function of the optimization problem reaches zero and thus the nuclear norm of $\tilde{\mathbf{Z}}$ is maximized to $2N$.

A last comment concerns the addition to the reflector $\mathbf{u}_i$ of the sparse structure typical of QR decompositions, i.e., consider $\mathbf{u}_i = \begin{bmatrix}\mathbf{0}; & \tilde{\mathbf{u}}_i\end{bmatrix}$. With this new structure the minimizer $\tilde{\mathbf{u}}_i$ of the expression in (22) is given by the eigenvector associated with the smallest, negative, eigenvalue of the lower right-hand side square sub-matrix of size $(n - i + 1)$ from $\tilde{\mathbf{Z}}$. This structure appears during the initialization step discussed in Section III.

C. Connections between Householder based dictionaries and general orthonormal dictionaries

The proposed algorithms are closely connected to the task of learning a general orthonormal dictionary. Increasing m

for H$_m$–DLA and QH$_m$–DLA will reduce the performance gap between dictionaries designed by these methods and the orthonormal dictionaries designed via Q–DLA, of course at the cost of higher computational demand. We discuss now some properties and connections between the various dictionary learning procedures.

Remark 4. Given a dataset $\mathbf{Y}$ represented in the general dictionary $\mathbf{D}$ with the sparse representations $\mathbf{X}$, there is no reflector $\mathbf{U}_1$ such that $\mathbf{U}_1\mathbf{D}$ achieves lower representation error than $\mathbf{D}$ if $\mathbf{D}\mathbf{X}\mathbf{Y}^T + \mathbf{Y}(\mathbf{D}\mathbf{X})^T$ is positive semidefinite.

Proof. Check for the existence of a reflector $\mathbf{U}_1$ such that no left dictionary update improves the representation

$$\|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 > \|\mathbf{Y} - \mathbf{U}_1\mathbf{D}\mathbf{X}\|_F^2. \tag{24}$$

If such a reflector does not exist then $\mathbf{D}$ may be viewed as a local minimum (this is a necessary condition). This is equivalent to considering an updated dictionary $\mathbf{U}_1\mathbf{D}$. Therefore, if the symmetric matrix

$$\mathbf{Z}_1 = \mathbf{D}\mathbf{X}\mathbf{Y}^T + \mathbf{Y}(\mathbf{D}\mathbf{X})^T, \tag{25}$$

is positive semidefinite then $\mathbf{D}$ is a local minimum of, i.e., there is no reflector $\mathbf{U}_1$ such that $\mathbf{U}_1\mathbf{D}$ is able to achieve a lower objective function value in than $\mathbf{D}$. Compare this with Remark 2. ■

As we have seen in the previous sections, the positive semidefinite condition is necessary and sufficient when describing local minima of the Householder based dictionaries. In the case of general orthonormal and, due to (24) and (25), also general (even overcomplete) dictionaries the condition is necessary, but not sufficient.

Remark 5. Q–DLA always performs better than QH$_m$–DLA, the performance matches when $\mathbf{Y}\mathbf{X}^T$ is symmetric. We have shown by (3) that the objective function reduction possible by a general orthogonal dictionary is $2\|\mathbf{Y}\mathbf{X}^T\|_*$. Due to the triangle inequality which is obeyed by the nuclear norm, this quantity is larger or equal at worse to the reduction achievable when using a symmetric dictionary designed via QH$_m$–DLA, which is $\|\mathbf{X}\mathbf{Y}^T + \mathbf{Y}\mathbf{X}^T\|_*$. As expected, due to its additional constraints, QH$_m$–DLA performs worse than the Q–DLA. In general, only H$_m$–DLA, with a sufficiently large m, has the capability to match the Q–DLA. ■

A simple example in $\mathbb{R}^2$. To illustrate the previous results, consider a dataset $\mathbf{Y} \in \mathbb{R}^{2\times N}$ and the initial dictionary $\mathbf{Q} = \mathbf{I}$. With target sparsity $s = 1$ we have, under a permutation of columns to highlight the row structure, the representations

$$\mathbf{X} = \begin{bmatrix} \mathbf{y}_{11}^T & \mathbf{0}^T \\ \mathbf{0}^T & \mathbf{y}_{22}^T \end{bmatrix} \text{ where } \mathbf{Y} = \begin{bmatrix} \mathbf{y}_{11}^T & \mathbf{y}_{12}^T \\ \mathbf{y}_{21}^T & \mathbf{y}_{22}^T \end{bmatrix}, \tag{26}$$

and therefore

$$\mathbf{Y}\mathbf{X}^T = \begin{bmatrix} \|\mathbf{y}_{11}\|_2^2 & \mathbf{y}_{12}^T\mathbf{y}_{22} \\ \mathbf{y}_{11}^T\mathbf{y}_{21} & \|\mathbf{y}_{22}\|_2^2 \end{bmatrix}, \ \tilde{\mathbf{Z}} = \mathbf{Y}\mathbf{X}^T + \mathbf{X}\mathbf{Y}^T. \tag{27}$$

By (5) and with (27) we see that there is no Householder based dictionary that improves the presentation error if $2\|\mathbf{y}_{11}\|_2^2 < \mathbf{y}_{11}^T\mathbf{y}_{21} + \mathbf{y}_{22}^T\mathbf{y}_{12}$ and $2\|\mathbf{y}_{22}\|_2^2 < \mathbf{y}_{11}^T\mathbf{y}_{21} + \mathbf{y}_{22}^T\mathbf{y}_{12}$, i.e., $\tilde{\mathbf{Z}}$ is positive semidefinite. Since $\text{tr}(\tilde{\mathbf{Z}}) = 2\|\mathbf{X}\|_F^2 > 0$ one of the eigenvalues is necessarily positive and therefore the previous two conditions lead to $2\|\mathbf{y}_{11}\|_2\|\mathbf{y}_{22}\|_2 < \mathbf{y}_{11}^T\mathbf{y}_{21} + \mathbf{y}_{22}^T\mathbf{y}_{12}$. Therefore, with a Householder based dictionary the possible reduction in the representation error is $\|\tilde{\mathbf{Z}}\|_* = 2\sqrt{\text{tr}(\tilde{\mathbf{Z}})^2/4 - \det(\tilde{\mathbf{Z}})} = 2\sqrt{\|\mathbf{X}\|_F^4 - \det(\tilde{\mathbf{Z}})}$. The Frobenius norm of the representations is maximized in the sparse approximation step while $-\det(\tilde{\mathbf{Z}})$, which is always positive, is increased when maximizing $\mathbf{y}_{11}^T\mathbf{y}_{21} + \mathbf{y}_{22}^T\mathbf{y}_{12}$.

Assuming $\mathbf{Y}\mathbf{X}^T$ is positive semidefinite then we know there is no Householder based dictionary that can improve the representations. If we consider now general orthonormal dictionaries with (27) we know from (7) that if $\mathbf{y}_{12}^T\mathbf{y}_{22} \approx \pm\mathbf{y}_{11}^T\mathbf{y}_{21}$ (i.e., $\mathbf{Y}\mathbf{X}^T$ is approximately symmetric or skew-symmetric) there is also no orthonormal dictionary that can perform much better in terms of representation than the identity. ■

D. Householder reflectors vs. Givens rotations for learning fast dictionaries

Householder reflectors are not the only elementary building blocks for orthonormal structures. Any orthonormal dictionary of size $n \times n$ can also be factorized in a product of Givens rotations [22] parameterized by c, s and the indices (i, j) like

$$\mathbf{G}_{ij} = \begin{bmatrix} \mathbf{I}_{i-1} & & & & \\ & c & & s & \\ & & \mathbf{I}_{j-i-1} & & \\ & -s & & c & \\ & & & & \mathbf{I}_{n-j} \end{bmatrix}, \ c^2 + s^2 = 1.$$

Givens rotations have been previously used with great success in several matrix factorization applications [42], [43], [44].

Consider using a single Givens rotation as a dictionary. We reach the optimization problem $\underset{c,s,(i,j);\ c^2+s^2=1}{\text{minimize}} \|\mathbf{Y} - \mathbf{G}_{ij}\mathbf{X}\|_F^2$, which is equivalent to

$$\underset{c,s,(i,j);\ c^2+s^2=1}{\text{minimize}} \left\| \begin{bmatrix} \mathbf{y}_i^T \\ \mathbf{y}_j^T \end{bmatrix} - \begin{bmatrix} c & s \\ -s & c \end{bmatrix} \begin{bmatrix} \mathbf{x}_i^T \\ \mathbf{x}_j^T \end{bmatrix} \right\|_F^2,$$

where $\mathbf{y}_i^T$ and $\mathbf{x}_i^T$ are the i^{th} rows of $\mathbf{Y}$ and $\mathbf{X}$, respectively.

When indices (i, j) are fixed, the optimization reduces to a two dimensional orthogonal Procrustes problem. While to select the indices (i, j), among the $\binom{n}{2}$ possibilities, an appropriate strategy needs to be defined. Indeed, Givens rotations also seem an appropriate tool to approach the fast dictionary learning problem, but it is beyond the scope of this paper to analyze it in detail.

V. Results

In this section we provide experimental results to illustrate the representation capabilities of the proposed methods.

A. Sparsely representing data

The input data that we consider is taken from popular test images from the image processing literature (pirate, peppers, boat etc.). The test datasets $\mathbf{Y} \in \mathbb{R}^{64\times N}$ 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) for example, are available. Our goal is to provide Householder

Table I: RMSE in the case of several dictionaries computed from known test images. Sparsity level is $s = 4$ and the dataset is $\mathbf{Y} \in \mathbb{R}^{64\times 4096}$ in each case. The learning procedures run after mean extraction and normalization $\mathbf{Y} = \mathbf{Y}/255$. The best results of the fast dictionaries are shown in bold font.

	peppers	boat	pollen	mri	cameraman	pirate	barb	baboon	hill	couple	house	fingerprint
DCT	0.0395	0.0419	0.0461	0.0721	0.0619	0.0507	**0.0435**	0.0694	0.0361	0.0432	0.0374	0.0765
H_6–DLA	0.0294	0.0371	0.0421	0.0649	0.0568	0.0453	0.0508	0.0738	0.0331	0.0405	0.0298	0.0536
H_{12}–DLA	**0.0261**	**0.0324**	**0.0376**	**0.0611**	**0.0512**	**0.0421**	0.0436	**0.0691**	**0.0302**	**0.0353**	**0.0255**	**0.0497**
QH_6–DLA	0.0306	0.0375	0.0425	0.0656	0.0575	0.0457	0.0508	0.0739	0.0334	0.0411	0.0302	0.0542
QH_{12}–DLA	0.0278	0.0336	0.0388	0.0626	0.0533	0.0434	0.0444	0.0702	0.0313	0.0366	0.0275	0.0512
H_{32}–DLA	0.0253	0.0310	0.0371	0.0594	0.0472	0.0407	0.0348	0.0649	0.0288	0.0336	0.0234	0.0492
QH_{32}–DLA	0.0278	0.0332	0.0385	0.0617	0.0519	0.0428	0.0397	0.0681	0.0305	0.0364	0.0265	0.0511
Q–DLA [24]	0.0256	0.0312	0.0372	0.0596	0.0473	0.0409	0.0361	0.0654	0.0292	0.0339	0.0241	0.0496
SK–SVD [45]	0.0191	0.0231	0.0275	0.0462	0.0311	0.0328	0.0266	0.0561	0.0235	0.0266	0.0143	0.0344

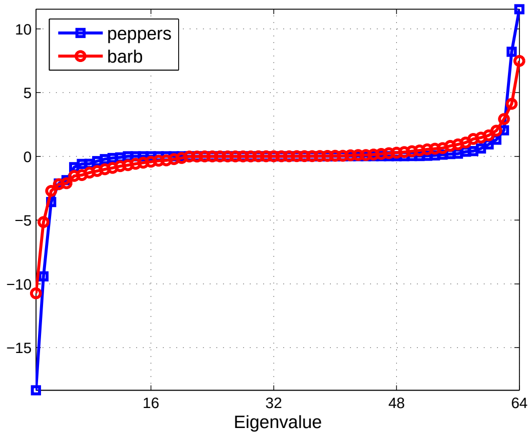

Figure 1: Normalized eigenvalues of $\tilde{\mathbf{Z}}$ after convergence of QH_m–DLA for images peppers and barb with sparsity $s = 4$ and $m = 12$ reflectors.

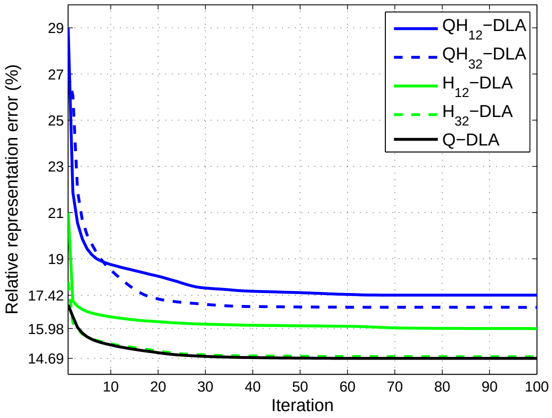

Figure 2: For the proposed methods we show the evolution of the relative representation error $\|\mathbf{Y} - \mathbf{DX}\|_F^2 \|\mathbf{Y}\|_F^{-2}$ for the dataset $\mathbf{Y}$ created from the patches of the images couple, peppers and boat with sparsity $s = 4$ and for $m \in \{12, 32\}$ reflectors. For reference we show Q–DLA [24].

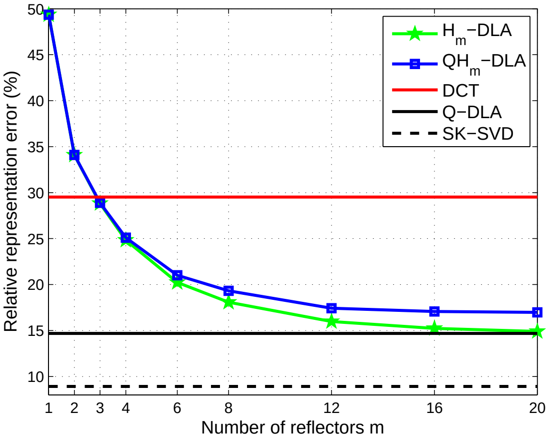

Figure 3: Relative representation error $\|\mathbf{Y} - \mathbf{DX}\|_F^2 \|\mathbf{Y}\|_F^{-2}$, in percent, for the proposed algorithms with the dataset composed of all patches from the images couple, peppers and boat for sparsity $s = 4$. For reference we show the DCT, Q–DLA [24] and SK–SVD [45].

based dictionaries that perform well in terms of representation error with a small number of reflectors m in their composition.

Table I shows the root mean squared error (RMSE) achieved by dictionaries trained on each test image separately and then used to sparsely represent those particular images. We show the performances of HQ_m–DLA and H_m–DLA for $m = 6$ and $m = 12$ reflectors. For perspective, we also show the performance achieved by the DCT on one hand and general (orthonormal and unconstrained) dictionary learning on the other – we use Q–DLA and Stagewise K–SVD (SK–SVD) [45]. For non-orthonormal dictionaries we use the OMP algorithm [46] in the sparse reconstruction step. As expected, increasing the number of reflectors decreases RMSE in all cases. The best performing method of the ones proposed in this paper and shown in the table is H_{12}–DLA. The worse performance of this approach is achieved for the barb test image. To understand why we can see in Figure 1 the eigenvalue distribution of the matrix $\tilde{\mathbf{Z}}$ from (17) for barb and peppers. As shown, most of the eigenvalues are close to (or exactly) zero. The difference comes when analyzing the negative eigenvalues which in the case of peppers are fewer and have larger magnitude than those of barb. We mention that for the barb test image the performance of Q–DLA is matched only by H_{24}–DLA. Table I shows on top the reference DCT and the proposed *fast* dictionaries performance while the bottom shows the *slower*

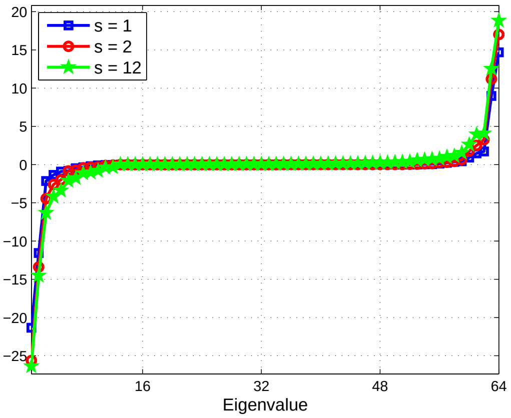

Figure 4: Normalized eigenvalues of $\tilde{\mathbf{Z}}$ after convergence of QH_{12}–DLA with various sparsity levels for the dataset in Figure 3.

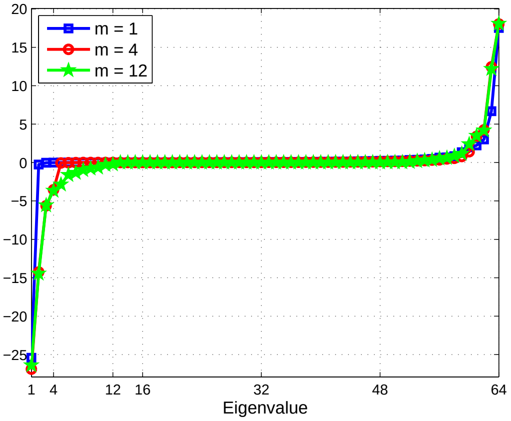

Figure 5: Normalized eigenvalues of $\tilde{\mathbf{Z}}$ after convergence of QH_m–DLA with sparsity $s = 4$ for various number of reflectors for the dataset in Figure 3.

dictionaries, including the H_{32}–DLA which generally performs slightly better even than Q–DLA. We would like to note here that the general dictionaries designed via K–SVD or SK–SVD do exhibit high mutual coherence in general, even though we do not construct overcomplete dictionaries. For example, the dictionary designed via SK–SVD and that reaches the best performance in terms of RMSE for the image peppers has mutual coherence over 0.9, very high.

In the case of H_m–DLA we have tested two strategies to update the reflectors: sequential (in order of their index) and random. Since the difference between the two is negligible, the results shown use sequential update.

In Figure 2 we show the representation error evolution of the proposed algorithms and of Q–DLA with each iteration. The plot shows the effectiveness of the initialization procedures and the monotonically decrease in the objective function value. As expected, Q–DLA and SK–SVD perform best while QH_m–DLA the worse. Still, for the number of reflectors considered $m \in \{12, 32\}$ the differences are not large. When we consider a larger number of reflectors like, $m = 32$, we see that in all cases the RMSE is only slightly higher than that of Q–DLA.

In Figure 3 we show the representation error for a dataset $\mathbf{Y}$ consisting of $N = 12288$ patches from several test images. For reference, we show again the DCT and Q–DLA representation performance. It is easy to see from the plot that the performance of the fixed transform is reached with a small number of reflectors m (3 in both the cases of the proposed methods). When we increase the number of reflectors, H_m–DLA reaches the performance of Q–DLA for $m = 20$ while QH_m–DLA converges to a slightly worse result. As discussed in Section IV, we did expect QH_m–DLA to always perform worse than Q–DLA. Notice that for a small number of reflectors the performance of H_m–DLA and QH_m–DLA are very close suggesting that the extra orthogonal constraint is natural in this regime. The results are interesting when comparing with the references: it is clear that the dictionaries based on reflectors match the performance of Q–DLA for $m < n/2$ while they outperform the fixed DCT transform for $m \ll n$. This shows that a full orthonormal dictionary can be avoided without sacrificing performance.

In Figures 4 and 5 we show the eigenvalues of $\tilde{\mathbf{Z}}$ for Householder dictionaries created by H_m–DLA using the dataset described in Figure 3. The eigenvalues are distributed similarly, independent of the choice of sparsity s and number of reflectors m. In Figure 5 notice that the choice of m determines the number of negative eigenvalues with large magnitudes. As explained in Section IV this drives the reduction in the objective function of the Householder dictionary learning problem.

As seen, QH_m–DLA and H_m–DLA perform similarly. For best performance H_m–DLA is preferred but when the dictionary learning procedure is time critical QH_m–DLA is a better choice given the small loss in performance.

In Table II we show the speed-ups provided by the Householder based dictionaries as a function of the number of reflectors. We show the comparative computational complexity of using the dictionaries, not their training. We compare against the complexity of using a general orthonormal dictionary and against that of the DCT (we compare against an efficient implementation, the fast cosine transform). The cost of finding the largest entries in magnitude is the same for all methods and thus it is not accounted for. Since computing the correlations between the dictionary and a target signal takes $4nm$ for a Householder based dictionary with m reflectors, the speed-ups are computed as

$$\rho_{\text{Q–DLA}} = \frac{(2n-1)n}{4nm}, \ \rho_{\text{FCT}} = \frac{5/2n\log n - 3n + 6}{4nm}. \quad (28)$$

The computational complexity of FCT is taken from [47]. The latter is for perspective since it does not seem reasonable to assume that for image data we can construct a dictionary faster than the FCT that achieves the performance of Q–DLA. Still, notice that a Householder based dictionary with $m = 3$ components closely matches the performance of the FCT both in terms of speed and in terms of performance (see Figure 3). An important observation is that with $m = 20$ reflectors we match closely the performance of Q–DLA while we still keep a computational advantage. From (28) it is clear that the proposed methods have lower computational complexity than

Table II: Speed-up provided by Householder based dictionaries as compared to the general orthonormal dictionaries and DCT – in this case a fast implementation, the Fast Cosine Transform (FCT) [47], is considered. We count the number of operations necessary to apply the dictionary as a direct and inverse operator, i.e., the computation of the correlations $\mathbf{D}^T\mathbf{y}$. We do not compare with the general sparse approximation methods like OMP since they are much slower – they are at least s times slower than an orthonormal dictionary, by (29). The number of reflectors m for which the complexity of the proposed dictionaries approximately coincides with Q–DLA and FCT is $m = 32$ and $m = 3$ respectively.

Number of reflectors m	1	2	3	4	6	8	12	16	20	24
speed-up $\rho_{\text{Q–DLA}}$ (28)	32×	16×	11×	8×	5×	4×	3×	2×	1.6×	1.3×
speed-up ρ_{FCT} (28)	3×	1.5×	1×	0.8×	0.5×	0.4×	0.3×	0.2×	0.2×	0.1×

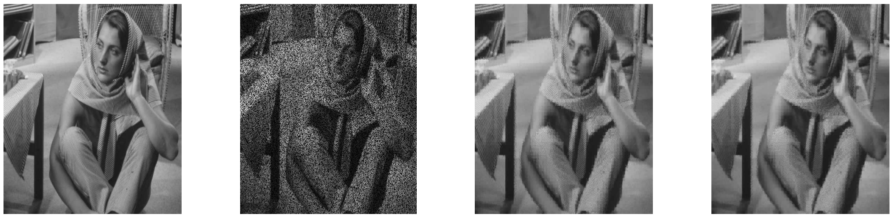

Figure 6: The figure contains from left to right: the original image, the corrupted image missing 40% of the pixels chosen uniformly at random, the reconstruction using the orthonormal dictionary $(\text{MAE} = 0.0305, \text{MSE} = 0.0492)$ and the reconstruction using the Householder based dictionary with $m = 14$ reflectors $(\text{MAE} = 0.0321, \text{MSE} = 0.0512)$. We always have that $s = 6$.

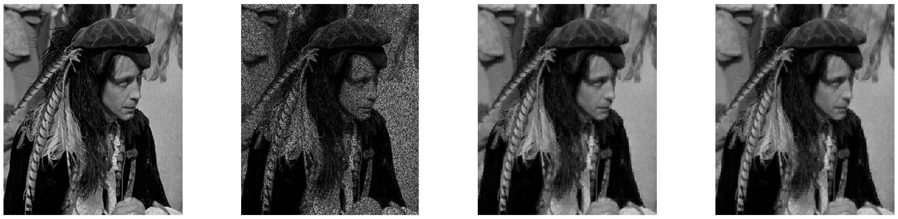

Figure 7: Analogous with Figure 6. The orthonormal dictionary reaches $\text{MAE} = 0.0333, \text{MSE} = 0.0548$ and the Householder dictionary reaches $\text{MAE} = 0.0334, \text{MSE} = 0.0549$.

general orthonormal dictionaries whenever $m \ll n$. We do not compare with the computational complexity of iterative methods since they are in general much slower than the methods discussed in this paper; for example, a batch variant of OMP called OMP–Cholesky [46] needs

$$N_{\text{OMP–Cholesky}} = 2sn^2 + 2s^2n + 4sn + s^3, \qquad (29)$$

operations. Since in general we do assume that we are dealing with sparse representations, i.e., $s \ll n$, the computational complexity of OMP–Cholesky is dominated by the first term which expresses the complexity of the explicit dictionary operator, the term that dominates also the computational complexity of using an orthonormal dictionary.

The final advantage of the proposed methods is the space requirement. As stated, in the case of dictionary learning the entries of the dictionaries need to be stored (or transmitted) together with the encoded data. With the proposed methods only the reflector vectors need to be stored, i.e., mn entries.

In terms of the computational complexity of the learning procedures themselves we report that in constructing the dictionaries for Figure 3 we have the approximate running times of 15 seconds for Q–DLA, 13 seconds for H$_8$–DLA, 7 seconds for QH$_8$–DLA all running for $K = 100$ iterations while SK–SVD took over one minute. All running times include the initialization procedures. The simulations were conducted in the Mathworks Matlab® 2014 environment, using a modern laptop computer i7 processor, 16 GB RAM running Windows®. As such, more efficient implementations are possible and the purpose of reporting the running times here is to provide a sense of the complexity of the learning procedure itself.

B. Application: denoising images

We also choose to test the trained dictionaries in reconstructions scenarios to fill in missing pixels from an image [11]. The experimental environment is as follows. We train a general orthonormal dictionary and one based on Householder reflectors on uncorrupted data (non-overlapping 8×8 image patches). We then blank a fixed percentage of the pixels in the images and perform the reconstruction using the previously trained dictionaries. Performance is measured in mean absolute error (MAE) and mean squared error (MSE) and the results are shown in Figures 6 and 7.

We compare Q–DLA and H$_{14}$–DLA to show that there are no large performance drawbacks when using dictionaries that are computationally efficient.

VI. Conclusions

In this manuscript we describe algorithms for the orthonormal dictionary learning task based on Householder reflectors. We are able to construct dictionaries that can be efficiently manipulated and that also perform very well in terms of representation capabilities where we compare with the fast, fixed transforms and general orthonormal, learned dictionaries. We are also able to provide local minimum conditions for the Householder based and general orthonormal dictionary learning problems.

Acknowledgment

The authors would like to thank anonymous reviewers and Bogdan Dumitrescu whose comments greatly improved the clarity of this manuscript.

References

[1] 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.

[2] 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.

[3] S. Beckouche, J. L. Starck, and J. Fadili, "Astronomical image denoising using dictionary learning," *Astron. Astrophys.*, vol. 556, no. A132, 2013.

[4] F. Couzinie-Devy, J. Mairal, F. Bach, and J. Ponce, "Dictionary learning for deblurring and digital zoom," *arXiv:1110.0957*, 2011.

[5] C. D. Sigg, T. Dikk, and J. M. Buhmann, "Speech enhancement using generative dictionary learning," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1698–1712, 2012.

[6] O. El Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. Heath, "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 13, no. 3, pp. 1499–1513, 2014.

[7] 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.

[8] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Comput.*, vol. 24, no. 2, pp. 227–234, 1995.

[9] 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.

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

[11] 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.

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

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

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

[15] 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.

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

[20] K. Gregor and Y. LeCun, "Learning fast approximations of sparse coding," in *Proc. International Conference on Machine Learning*, 2010, pp. 399–406.

[21] 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.

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

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

[24] 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.

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

[27] 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.

[28] D. Barchiesi and M. D. Plumbley, "Learning incoherent dictionaries for sparse approximation using iterative projections and rotations," *IEEE Trans. Sig. Proc.*, vol. 61, no. 8, pp. 2055–2065, 2013.

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

[30] J. M. Chambers, "Partial sorting," *CACM*, vol. 14, no. 5, pp. 357–358, 1971.

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

[32] R. Gribonval and K. Schnass, "Dictionary identification sparse matrix-factorization via ℓ_1-minimization," *IEEE Trans. Inf. Theory*, vol. 56, no. 7, pp. 3523–3539, 2010.

[33] Q. Geng and J. Wright, "On the local correctness of ℓ^1-minimization for dictionary learning," in *IEEE International Symposium on Information Theory*, 2014, pp. 3180–3184.

[34] D. Spielman, H. Wang, and J. Wright, "Exact recovery of sparsely-used dictionaries," *arXiv:1206.5882*, 2012.

[35] D. Vainsencher, S. Mannor, and A. M. Bruckstein, "The sample complexity of dictionary learning," *J. Machine Learning Research*, vol. 12, pp. 3259–3281, 2011.

[36] S. Arora, A. Bhaskara, R. Ge, and T. Ma, "More algorithms for provable dictionary learning," *arXiv:1401.0579*, 2014.

[37] A. Agarwal, A. Anandkumar, P. Jain, P. Netrapalli, and R. Tandon, "Learning sparsely used overcomplete dictionaries via alternating minimization," *arXiv:1310.7991*, 2013.

[38] R. Gribonval, R. Jenatton, F. Bach, M. Kleinsteuber, and M. Seibert, "Sample complexity of dictionary learning and other matrix factorizations," *IEEE Trans. Inf. Theory*, vol. 61, no. 6, pp. 3469–3486, 2015.

[39] Y. Saad, "Numerical methods for large eigenvalue problems," *Manchester University Press*, 1992.

[40] X. Sun and C. Bischof, "A basis-kernel representation of orthogonal matrices," *SIAM J. Matrix Anal. Appl*, vol. 16, no. 4, pp. 1184–1196, 1995.

[41] M. Aharon, M. Elad, and A. Bruckstein, "On the uniqueness of overcomplete dictionaries, and a practical way to retrieve them," *Linear Algebra Appl.*, vol. 416, pp. 48–67, 2006.

[42] A. B. Lee, B. Nadler, and L. Wasserman, "Treelets - an adaptive multiscale basis for sparse unordered data," *The Annals of Applied Statistics*, vol. 2, no. 2, pp. 435–471, 2008.

[43] 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.

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

[45] C. Rusu and B. Dumitrescu, "Stagewise K-SVD to design efficient dictionaries for sparse representations," *IEEE Signal Processing Letters*, vol. 19, no. 10, pp. 631–634, 2012.

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

[47] 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.