Analysis of Fast Structured Dictionary Learning

Saiprasad Ravishankar[1], Anna Ma[2], and Deanna Needell[3]

[1]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor
[2]Institute of Mathematical Science, Claremont Graduate University
[3]Department of Mathematics, University of California Los Angeles

June 1, 2018

Abstract

Sparsity-based models and techniques have been exploited in many signal processing and imaging applications. Data-driven methods based on dictionary and transform learning enable learning rich image features from data, and can outperform analytical models. In particular, alternating optimization algorithms for dictionary learning have been popular. In this work, we focus on alternating minimization for a specific structured unitary operator learning problem, and provide a convergence analysis. While the algorithm converges to the critical points of the problem generally, our analysis establishes under mild assumptions, the local linear convergence of the algorithm to the underlying generating model of the data. Analysis and numerical simulations show that our assumptions hold well for standard probabilistic data models. In practice, the algorithm is robust to initialization.

1 Introduction

Various models of signals and images are popular in signal processing and imaging applications, including dictionary and sparsifying transform models, tensor models, and manifold models. The learning of such models from training data sets has become increasingly popular. Learned models may outperform analytical or fixed models in various applications. For example, learned dictionaries and sparsifying transforms work well in applications such as image and video denoising, inpainting, and medical image reconstruction [3, 15, 24, 27, 30, 46, 43]. This work focuses on analyzing a structured dictionary learning algorithm. In particular, the ability of the algorithm to recover generative models of data is analyzed. In the following, we present some background in dictionary learning, before discussing the learning problem and algorithm, and our contributions.

1.1 Background

The synthesis dictionary and sparsifying transform models of signals have been quite popular. While the synthesis dictionary model approximates a signal $\mathbf{y} \in \mathbb{R}^n$ as $\mathbf{y} \approx \boldsymbol{D}\mathbf{z}$ with $\mathbf{D} \in \mathbb{R}^{n \times J}$ denoting the synthesizing dictionary and $\mathbf{z} \in \mathbb{R}^J$ denoting the sparse code (i.e., $\|\mathbf{z}\|_0 \ll n$ with the ℓ_0 "norm" counting the number of non-zero vector entries), the sparsifying transform model assumes that $\mathbf{W}\mathbf{y} \approx \mathbf{x}$, where $\mathbf{W} \in \mathbb{R}^{K \times n}$ denotes a sparsifying transform and $\mathbf{x}$ is assumed to have several zeros (corresponding to the transform rows that approximately annihilate the signal). Finding the optimal sparse approximation for a signal in the synthesis dictionary model involves solving the well-known synthesis *sparse coding* problem[1] that is known to be NP-hard in general [18], and numerous algorithms exist for the problem [22, 17, 9, 11, 19, 10] that provide the solution under certain conditions. In the transform model, sparse transform-domain approximations are obtained by simple (e.g., hard or soft) thresholding [27].

The learning of dictionaries and transforms from a collection of signals has been the subject of many works [20, 13, 3, 45, 15, 38, 5, 27, 25, 26, 42], and such learned models provide promising results in applications [12, 16, 24, 14, 28]. Often additional properties are enforced on the dictionary during learning, e.g., incoherence [8, 23], low-rank atoms [31], etc. Algorithms for learning often alternate between a *sparse coding step* (where sparse approximations of training signals are computed) and a *model update step* (where the dictionary or transform are updated) [13, 3, 45, 34, 37, 15, 21, 38, 35, 36, 5, 27, 42].

Dictionary and transform learning problems are often highly non-convex, and many learning algorithms lack proven convergence guarantees or model recovery guarantees. Recent works [39, 1, 4, 44, 5, 6, 2, 32, 29] have studied the convergence of specific learning algorithms. Some of these works [5, 6, 32, 29] demonstrate promising results in applications for efficient algorithms and prove convergence of the learning methods to the critical points (or generalized stationary points [33]) in the problems. Other works [1, 2] prove the recovery of the underlying generative model for specific learning methods, but rely on many restrictive assumptions. Moreover, these schemes are also computationally expensive, and have not been demonstrated to be practically powerful in applications such as inverse problems. Another very recent two-part work [40, 41] focuses on structured, complete dictionaries and studies the geometric properties of the non-convex objective for dictionary learning over a high dimensional sphere. It shows with high probability that there are no spurious local minimizers and provides a specific convergent (to local minimizers) algorithm. In the following section, we outline a *structured* (unitary) dictionary or transform learning approach that involves simple, computationally cheap updates and works well in many applications. We investigate its convergence properties in the rest of the paper.

1.2 Learning Formulation and Algorithm

Given an $n \times N$ training data set $\mathbf{P}$, whose columns represent training signals, our goal is to find an $n \times n$ sparsifying transformation matrix $\mathbf{W}$ and an $n \times N$ sparse coefficients

[1] For example, one may minimize $\|\mathbf{y} - \mathbf{D}\mathbf{z}\|_2^2$ with respect to $\mathbf{z}$ subject to $\|\mathbf{z}\|_0 \leq s$, where s denotes a set sparsity level, or an alternative version of this problem.

(representation) matrix $\mathbf{Z}$ by solving the following constrained optimization problem:

$$\underset{\mathbf{W},\mathbf{Z}}{\arg\min} \|\mathbf{W}\mathbf{P} - \mathbf{Z}\|_F^2 \quad \text{s.t. } \mathbf{W}^T\mathbf{W} = \mathbf{Id}, \left\|\mathbf{Z}_{(.,j)}\right\|_0 \leq s \ \forall j. \tag{1}$$

We focus on the learning of unitary sparsifying operators ($\mathbf{W}^T\mathbf{W} = \mathbf{Id}$ with $\mathbf{Id}$ denoting the identity matrix) that have shown promise in various applications such as denoising and medical image reconstruction [25, 29, 30]. The columns $\mathbf{Z}_{(.,j)}$ of $\mathbf{Z}$ have at most s non-zeros (measured using the ℓ_0 "norm"), where s is a given parameter. Alternatives to Problem (1) involve replacing the column-wise sparsity constraints with a constraint on the total sparsity (aggregate sparsity) of the entire matrix $\mathbf{Z}$, or using a sparsity penalty (e.g., ℓ_p penalties with $0 \leq p \leq 1$). Problem (1) is equivalent to learning a dictionary $\mathbf{W}^T$ for sparsely approximating the training data $\mathbf{P}$ as $\mathbf{W}^T\mathbf{Z}$.

Alternating minimization algorithms are commonly used for dictionary learning [13, 3, 34, 38]. One could alternate between solving for $\mathbf{W}$ and $\mathbf{Z}$ in Problem (1) [25, 29]. In this case, the solution for the tth $\mathbf{Z}$ update (*sparse coding step*) is obtained as $\mathbf{Z}^t_{(.,j)} = H_s(\mathbf{W}^{t-1}\mathbf{P}_{(.,j)}) \ \forall j$, where $\mathbf{P}_{(.,j)}$ and $\mathbf{Z}^t_{(.,j)}$ denote the jth columns of $\mathbf{P}$ and $\mathbf{Z}^t$ respectively, and the operator $H_s(\cdot)$ zeros out all but the s largest magnitude elements of a vector, leaving other entries unchanged (i.e., thresholding to s largest elements). The solution for the subsequent $\mathbf{W}$ update (*operator update step*) is obtained by first computing the full singular value decomposition (SVD) of $\mathbf{Z}^t\mathbf{P}^T$ as $\mathbf{V}\mathbf{\Sigma}\mathbf{U}^T$, and then $\mathbf{W}^t = \mathbf{V}\mathbf{U}^T$. The algorithm repeats these relatively cheap updates until convergence. The overall method is provided in Algorithm 1.

Problem (1) is also interpreted as training an efficient convolutional sparsifying transformation for 2D (or higher dimensional) images. To see this, we observe that if overlapping patches of an image or collection of images of size $\sqrt{n} \times \sqrt{n}$ are (vectorized and) used for training with periodic image boundary condition (so patches at image boundaries wrap around on the opposite side of the image) and a patch stride of 1 pixel in the horizontal and vertical directions (maximal patch overlap), then the transform learned by Problem (1) is applied to sparse code the data by first applying each row to all the image patches via inner products, followed by thresholding operations. The sparse outputs of the transform are thus generated by circularly convolving its reshaped (into 2D patches) rows with the image followed by thresholding. Thus, Problem (1) adapts a collection of orthogonal sparsifying filters for images, and Algorithm 1 can be implemented with filtering-based operations.

1.3 Contributions

In this work, we investigate the convergence properties of the aforementioned alternating minimization algorithm for unitary sparsifying operator learning. Recent works have shown convergence of the algorithm (or its variants) to critical points of the equivalent unconstrained problem [29, 28, 7], where the constraints are replaced with barrier penalties (that take value $+\infty$ when the constraint is violated and 0 otherwise). Here, we further prove the fast local linear convergence of the algorithm to the underlying generative data models. Our results hold under mild assumptions that depend on the properties of the underlying ("generating") sparse coefficients matrix $\mathbf{Z}$. We show analytically and experimentally that these assumptions hold well for well-known probabilistic models of $\mathbf{Z}$ such as when $\mathbf{Z}$ has

columns with random support locations and i.i.d. Gaussian non-zero coefficients. In practice, the learning algorithm is robust or insensitive to initialization.

1.4 Organization

The rest of this paper is organized as follows. Section 2 presents the main convergence results and proofs. Section 3 presents experimental results supporting the statements in Section 2 and illustrating the empirical behavior of the transform learning algorithm. In Section 4, we conclude with proposals for future work.

Algorithm 1 Alternating Optimization for (1)

Input: Training data matrix $\mathbf{P}$, maximum iteration count L, sparsity s
Output: $\mathbf{W}^L$, $\mathbf{Z}^L$
Initialize: $\mathbf{W}^0$ and $t = 1$
for $t \leq L$ **do**
  $\mathbf{Z}^t_{(.,j)} = H_s\left(\mathbf{W}^{t-1}\mathbf{P}_{(.,j)}\right)\ \forall\ j$ ▷ $\mathbf{Z}^t = \arg\min_{\mathbf{Z}:\|\mathbf{Z}_{(.,j)}\|_0 \leq s\ \forall j} \|\mathbf{W}^{t-1}\mathbf{P} - \mathbf{Z}\|_F^2$
  $\mathbf{P}\mathbf{Z}^{tT} = \mathbf{U}^t\mathbf{\Sigma}^t\mathbf{V}^{tT}$ ▷ $\mathbf{W}^t = \arg\min_{\mathbf{W}:\mathbf{W}^T\mathbf{W}=\mathbf{Id}} \|\mathbf{W}\mathbf{P} - \mathbf{Z}^t\|_F^2$
  $\mathbf{W}^t = \mathbf{V}^t\mathbf{U}^{tT}$
  $t = t + 1$
end for

2 Convergence Analysis

The main contribution of this work is the convergence analysis of Algorithm 1. We begin this section outlining notation and the assumptions under which our analysis operates. Following this, we summarize the theoretical guarantees of our work and present the proofs for these results.

2.1 Notation

We adopt the following notation in the rest of the paper. Matrix $\mathbf{Z}$ denotes the $n \times N$ sparse coefficients matrix, $\mathbf{W}$ is the $n \times n$ sparsifying transform, and $\mathbf{P}$ denotes the $n \times N$ data set. The tth approximation of a variable (iterate in the algorithm) is denoted $(\cdot)^t$. Capitalized variables are used for matrices and lowercase variables are used for vectors, with further subscripts denoting the row, column, or entry of the matrix or vector. The ith row, jth column, and the (i,j)th entry of a matrix $\mathbf{M}$ are denoted $\mathbf{M}_{(i,\cdot)}$, $\mathbf{M}_{(\cdot,j)}$, and $\mathbf{M}_{(i,j)}$, respectively. For any vector $\mathbf{v}$, $S(\mathbf{v})$ denotes the function that returns the support, i.e., $S(\mathbf{v}) := \{i : \mathbf{v}_i \neq 0\}$, where $\mathbf{v}_i$ denotes the ith entry (scalar) of $\mathbf{v}$. The operator $H_s(\mathbf{v})$ leaves the s largest magnitude elements of $\mathbf{v}$ unchanged and zeros out all other entries (i.e., thresholding to s largest elements). Matrix $\boldsymbol{D}_k$ denotes an $n \times n$ diagonal matrix of ones and a zero at location (k,k). Additionally, $\tilde{\boldsymbol{D}}_k$ denotes an $N \times N$ diagonal matrix that has ones at entries (i,i) for $i \in S(\mathbf{Z}^*{}_{(k,\cdot)})$ and zeros elsewhere, and matrix $\mathbf{Z}^*$ is defined in Section 2.2 (see assumption (A_1)). The Frobenious norm, denoted $\|\mathbf{X}\|_F^2$, is the the sum of squared

elements of $\mathbf{X}$, and $\|\mathbf{X}\|_2$ denotes the spectral norm. Lastly, $\mathbf{Id}$ denotes the appropriately sized identity matrix.

2.2 Assumptions

We begin with the following assumptions that will be used in various results:

- (A_1) **Generative model:** There exists a $\mathbf{Z}^*$ and unitary $\mathbf{W}^*$ such that $\mathbf{W}^*\mathbf{P} = \mathbf{Z}^*$ and $\|\mathbf{P}\|_2 = 1$ (normalized data).
- (A_2) **Sparsity:** The columns of $\mathbf{Z}^*$ are s-sparse, i.e., $\|\mathbf{Z}^*_{(\cdot,j)}\|_0 = s \ \forall \ j$.
- (A_3) **Spectral property:** The underlying $\mathbf{Z}^*$ satisfies the following bound $\kappa^4(\mathbf{Z}^*) \max_{1\leq k\leq n} \|\boldsymbol{D}_k \mathbf{Z}^* \tilde{\boldsymbol{D}}_k\|_2 < 1$, where $\kappa(\cdot)$ denotes the condition number (ratio of largest to smallest singular value).
- (A_4) **Orthogonal coefficients:** The rows of $\mathbf{Z}^*$ are orthonormal, i.e., $\mathbf{Z}^*\mathbf{Z}^{*T} = \mathbf{Id}$.
- (A_5) **Initialization:** $\|\mathbf{W}^0 - \mathbf{W}^*\|_F \leq \epsilon$ for an appropriate small $\epsilon > 0$. [2]

The first two assumptions are on a generative model for the data, i.e., we would like the algorithm to find an underlying sparsifying transform and representation matrix such that $\mathbf{W}^*\mathbf{P} = \mathbf{Z}^*$ holds, where the columns of $\mathbf{Z}^*$ have at most s nonzeros. The coefficients are assumed "structured" in assumption (A_3), satisfying a spectral property, which will be used to establish our theoretical results. We will present an analysis and empirical results showing that the spectral property holds for well-known probabilistic models. Assumption (A_4) on orthogonality of coefficient matrix (normalized) rows simplifies the condition in assumption (A_3) (since $\kappa(\mathbf{Z}^*) = 1$) and is used in presenting/proving one version of the results, but is omitted in the generalization. In practice, for well-known probabilistic models of the coefficient matrix, we will show empirically that the orthogonality holds asymptotically. Assumption (A_5) on algorithm initialization states that the initial sparsifying transform, $\mathbf{W}^0$ is sufficiently close to the solution $\mathbf{W}^*$. Such an assumption has also been made in other works, where the issue of good initialization is tackled separately [1, 2]. Our main results are stated next.

2.3 Main Results

Theorem 2.1 first presents a convergence result using all the aforementioned assumptions. Then Theorem 2.2 generalizes the result by dropping Assumption (A_4). Proposition 1 states that Assumption (A_3) holds under a general probabilistic model on the sparse representation matrix $\mathbf{Z}^*$. We also later show numerical results illustrating Proposition 1. We also provide a corollary on a special case of Theorems 2.1 and 2.2 and some remarks. In particular, Remark 1 discusses dropping the data normalization assumption in (A_1), and Remark 2 discusses the effect of noise on Theorems 2.1 and 2.2.

[2] Although we do not specify the best (largest permissible) ϵ explicitly, $\epsilon < \frac{1}{2}\min_j \beta\left(\frac{\mathbf{Z}^*_{(\cdot,j)}}{\|\mathbf{Z}^*_{(\cdot,j)}\|_2}\right)$ with $\beta(\cdot)$ denoting the smallest nonzero magnitude in a vector, will arise in one of our proof steps. The actual permissible ϵ is also dictated as per (convergence of) Taylor series expansions discussed in the proof.

Theorem 2.1. *Under Assumptions* $(A_1)-(A_5)$*, the Frobenius error between the iterates generated by Algorithm 1 and the underlying generative model in Assumptions* (A_1) *and* (A_2) *is bounded as follows:*

$$\|\mathbf{Z}^t - \mathbf{Z}^*\|_F \leq q^{t-1}\epsilon, \;\; \|\mathbf{W}^t - \mathbf{W}^*\|_F \leq q^t\epsilon, \tag{2}$$

where $q \triangleq \max_{1\leq k\leq n} \|\boldsymbol{D}_k\mathbf{Z}^*\tilde{\boldsymbol{D}}_k\|_2 < 1$ *and* ϵ *is fixed based on the initialization.*

Theorem 2.2. *Under Assumptions* $(A_1)-(A_3)$ *and* (A_5)*, the iterates in Algorithm 1 converge linearly to the underlying generative model in Assumptions* (A_1) *and* (A_2)*, i.e., the Frobenius error between the iterates and the generative model satisfies*

$$\|\mathbf{Z}^t - \mathbf{Z}^*\|_F \leq q^{t-1}\epsilon, \;\; \|\mathbf{W}^t - \mathbf{W}^*\|_F \leq q^t\epsilon, \tag{3}$$

where $q \triangleq \kappa^4\left(\mathbf{Z}^*\right)\max_{1\leq k\leq n} \|\boldsymbol{D}_k\mathbf{Z}^*\tilde{\boldsymbol{D}}_k\|_2 < 1$ *and* ϵ *is fixed based on the initialization.*

Next, we discuss special cases of Theorem 2.1 and Theorem 2.2 when $s = 2$. In the case of Theorem 2.1, a simple intuitive condition that the supports of no two rows of $\mathbf{Z}^*$ fully overlap ensures linear convergence ($q < 1$), i.e., ensures Assumption (A_3) holds.

Corollary 2.3. *(Case* $s = 2$*) For Theorem 2.1, when* $s = 2$ *and no two rows of* $\mathbf{Z}^*$ *have identical support, then* $q < 1$ *holds in Assumption* (A_3)*. For Theorem 2.2 (without Assumption* (A_4)*), when* $s = 2$*, then* $q < 1$ *holds in Assumption* (A_3) *if* $\|\mathbf{Z^*}_{(i,\cdot)}|_{S(\mathbf{Z^*}_{(i,\cdot)})\cap S(\mathbf{Z^*}_{(k,\cdot)})}\|_2 < \kappa^{-4}(\mathbf{Z}^*)$ *for all* $i \neq k$*, where the norm is computed only with respect to the elements of* $\mathbf{Z^*}_{(i,\cdot)}$ *in the support* $S(\mathbf{Z^*}_{(i,\cdot)}) \cap S(\mathbf{Z^*}_{(k,\cdot)})$*.*

The effect of dropping the data normalization assumption in (A_1) is discussed next. The convergence rate factor q in this case is modified by being normalized by $\|\mathbf{P}\|_2 = \|\mathbf{Z}^*\|_2$, which keeps it invariant to scaling of $\mathbf{Z}^*$.

Remark 1. *When the unit spectral norm condition on* $\mathbf{P}$ *in Assumption* (A_1) *is dropped, the* $\|\mathbf{W}^t - \mathbf{W}^*\|_F$ *bound in Theorem 2.2 holds with* $q \triangleq \left(\kappa^4\left(\mathbf{Z}^*\right)/\|\mathbf{P}\|_2\right)\max_{1\leq k\leq n} \|\boldsymbol{D}_k\mathbf{Z}^*\tilde{\boldsymbol{D}}_k\|_2$ *and the bound on* $\|\mathbf{Z}^t - \mathbf{Z}^*\|_F$ *holds but with* ϵ *replaced by* $\|\mathbf{P}\|_2\epsilon$*.*

As will be clear from the proofs in Section 2.4, when Assumption (A_2) stating $\|\mathbf{Z}^*_{(\cdot,j)}\|_0 = s$ is relaxed to $\|\mathbf{Z}^*_{(\cdot,j)}\|_0 \leq s$, then the (common) linear contraction factor q for the error in each iteration in Theorem 2.2 (with respect to previous iteration's error) is replaced with $q(t) \triangleq \kappa^4\left(\mathbf{Z}^*\right)\max_{1\leq k\leq n} \|\boldsymbol{D}_k\mathbf{Z}^*\tilde{\boldsymbol{D}}^t_k\|_2$, where $\tilde{\boldsymbol{D}}^t_k$ is defined similar to $\tilde{\boldsymbol{D}}_k$ but with respect to $S(\mathbf{Z}^t_{(k,\cdot)})$ (which is shown in Section 2.4 to contain $S(\mathbf{Z^*}_{(k,\cdot)})$). Next, we discuss the scenario and models when the assumption $q < 1$ is generally valid.

Proposition 1. *Suppose the locations of the* s *nonzeros in each column of* $\mathbf{Z}^*$ *are chosen independently and uniformly at random, and the non-zero entries are i.i.d. with mean zero and variance* n/sN*. Then, for fixed* s*,* n*, and* $s < n$*, we have that* $q_N \triangleq \left(\kappa^4\left(\mathbf{Z}^*\right)/\|\mathbf{P}\|_2\right)\max_{1\leq k\leq n} \|\boldsymbol{D}_k\mathbf{Z}^*\tilde{\boldsymbol{D}}_k\|_2 < 1$ *for large enough* N *with high probability. In particular, we have the following limit almost surely.*

$$q^* = \lim_{N\to\infty} q_N = \sqrt{\frac{s-1}{n-1}} \tag{4}$$

Proposition 1 holds for several well-known distributions of $\mathbf{Z}^*$ such as when its column supports are drawn independently and uniformly at random and the nonzero entries are a) i.i.d. with $\mathbf{Z}^*_{(i,j)} \sim \mathcal{N}(0, \frac{n}{sN})$ or b) i.i.d. scaled (by $\sqrt{n/sN}$) random signs with "+" and "-" being equally probable. Section 3 empirically shows the behavior of q with respect to N when $s = O(n)$, a commonly used sparsity criterion in many applications (i.e., with $s = \alpha n$, where $\alpha < 1$ is a small fraction). Finally, we have the following generalization of Theorem 2.2 for noisy generative models (i.e., showing stability to noise).

Remark 2. *When a noisy generative model of the data is used in Assumption* (A_1)*, i.e.,* $\mathbf{W}^*\mathbf{P} = \mathbf{Z}^* + \mathbf{H}$*, where* $\mathbf{H}$ *denotes noise, then for small noise, Theorem 2.2 holds, except that the term* $C\|\mathbf{H}\|_F$*, where* $C > 0$ *is a constant, is added to the right hand side of* (3)*.*

2.4 Proofs of Theorems, Corollary, and Remarks

We first prove Theorem 2.1 and then the proof of Theorem 2.2 is briefly presented highlighting the distinctions arising from the generalization. The proof of Corollary 2.3 is presented for the case of Theorem 2.1 (the proof for the case of Theorem 2.2 is similar). The proof of Remark 2 follows along the same lines as those of the theorems, and is omitted.

To prove Theorem 2.1, we will first prove two supporting lemmas that establish properties of the iterates. First, Lemma 1 shows that the error between the iterate $\mathbf{Z}^1$ and $\mathbf{Z}^*$ is bounded and the bound depends on the approximation error with respect to $\mathbf{W}^*$ for the initial $\mathbf{W}^0$ (bounded by ϵ as in Assumption (A_5)). Lemma 2 and Lemma 3 show that the error between the first $\mathbf{W}$ iterate ($\mathbf{W}^1$) and $\mathbf{W}^*$ is bounded above by $q\epsilon$ for Theorem 2.1 and Theorem 2.2, respectively. Similar bounds are shown to hold for subsequent iterations. Therefore, for Algorithm 1 to converge linearly, one only needs $q < 1$ as in Assumption (A_3) or as established by Proposition 1. The scaling indicated in Remark 1 follows from the proofs of Lemmas 1 and 3.

2.4.1 Proof of Theorem 2.1

For our proofs, we define the sequences $\{\mathbf{E}^t\}$ and $\{\mathbf{\Delta}^t\}$ such that

$$\mathbf{W}^t = \mathbf{W}^* + \mathbf{E}^t, \tag{5}$$
$$\mathbf{Z}^t = \mathbf{Z}^* + \mathbf{\Delta}^t. \tag{6}$$

Lemma 1. *(Approximation error for* $\mathbf{Z}$*) For* $t = 1$ *in Algorithm 1 and under Assumptions* $(A_1) - (A_5)$*, the Frobenious norm of the approximation error of the estimated sparse coefficients with respect to* $\mathbf{Z}^*$ *is bounded by* ϵ *as defined in* (A_5)*. In particular, we have that*

$$\|\mathbf{Z}^1 - \mathbf{Z}^*\|_F \leq \|\mathbf{E}^0\|_F,$$

where $\|\mathbf{E}^0\|_F \leq \epsilon$*.*

Proof. For each column indexed by $j = 1, ..., N$, of the sparse coefficients matrix $\mathbf{Z}^1$, the

following hold:

$$\begin{aligned}\mathbf{Z}^1_{(\cdot,j)} = H_s(\mathbf{W}^0\mathbf{P}_{(\cdot,j)}) &\overset{(Eq.5)}{=} H_s(\mathbf{W}^*\mathbf{P}_{(\cdot,j)} + \mathbf{E}^0\mathbf{P}_{(\cdot,j)}) \\ &\overset{(A_1)}{=} H_s(\mathbf{Z}^*_{(\cdot,j)} + \mathbf{E}^0\mathbf{P}_{(\cdot,j)}) \\ &\overset{(A_5)}{=} \mathbf{Z}^*_{(\cdot,j)} + \mathbf{\Gamma}^1_j\mathbf{E}^0\mathbf{P}_{(\cdot,j)}, \end{aligned} \tag{7}$$

where $\mathbf{\Gamma}^1_j$ is a diagonal matrix with a one in the (i,i)th entry if $i \in S(\mathbf{Z}^1_{(\cdot,j)})$ and zero otherwise and $\mathbf{E}^0$ is as defined in (5). The last equality above follows from the fact that the support of $\mathbf{Z}^1_{(\cdot,j)}$ includes that of $\mathbf{Z}^*_{(\cdot,j)}$, for small enough ϵ (assumption (A_5)). In particular, since $\left\|\mathbf{P}_{(\cdot,j)}\right\|_2 = \left\|\mathbf{Z}^*_{(\cdot,j)}\right\|_2$, we have

$$\left\|\mathbf{E}^0\mathbf{P}_{(\cdot,j)}\right\|_\infty \leq \left\|\mathbf{E}^0\mathbf{P}_{(\cdot,j)}\right\|_2 \leq \left\|\mathbf{E}^0\right\|_F \left\|\mathbf{Z}^*_{(\cdot,j)}\right\|_2.$$

Therefore, whenever $\|\mathbf{E}^0\|_F \leq \epsilon < \frac{1}{2}\min_j \beta\left(\frac{\mathbf{Z}^*_{(\cdot,j)}}{\|\mathbf{Z}^*_{(\cdot,j)}\|_2}\right)$ with $\beta(\cdot)$ being the smallest nonzero magnitude vector entry, the support of $\mathbf{Z}^1_{(\cdot,j)}$ includes[3] that of $\mathbf{Z}^*_{(\cdot,j)}$ (the entries of the perturbation $\mathbf{E}^0\mathbf{P}_{(\cdot,j)}$ are not large enough to change the support). The following results then hold:

$$\begin{aligned}\|\mathbf{Z}^1 - \mathbf{Z}^*\|_F^2 &\overset{(Eq.7)}{=} \|[\mathbf{\Gamma}^1_1\mathbf{E}^0\mathbf{P}_{(\cdot,1)}, ..., \mathbf{\Gamma}^1_N\mathbf{E}^0\mathbf{P}_{(\cdot,N)}]\|_F^2 \\ &\overset{(i)}{\leq} \|\mathbf{E}^0\mathbf{P}\|_F^2 \overset{(ii)}{\leq} \|\mathbf{E}^0\|_F^2\|\mathbf{P}\|_2^2 \overset{(A_1)}{=} \|\mathbf{E}^0\|_F^2.\end{aligned}$$

Here, (i) follows by definition of $\mathbf{\Gamma}^1_j$; step (ii) holds for the Frobenius norm of a matrix-matrix product; and the last equality holds because $\|\mathbf{P}\|_2 = 1$ (Assumption (A_1)). By Assumption (A_5), $\|\mathbf{E}^0\|_F^2 \leq \epsilon$, which completes the proof.

□

Lemma 2. *(Approximation error for $\mathbf{W}$) For $t = 1$ in Algorithm 1 and under Assumptions $(A_1)-(A_5)$, the Frobenious norm of the approximation error of the estimated transform with respect to $\mathbf{W}^*$ is bounded as*

$$\|\mathbf{W}^1 - \mathbf{W}^*\|_F \leq q\epsilon,$$

where q is a scalar coefficient as in Theorem 2.1.

Proof. Denote the SVD of $\mathbf{Z}^*\mathbf{Z}^{1T}$ as $\mathbf{U}^1_z\mathbf{\Sigma}^1_z\mathbf{V}^{1T}_z$. From Algorithm 1, we have

$$\mathbf{W}^1 = \mathbf{V}^1\mathbf{U}^{1T}, \quad \mathbf{P}\mathbf{Z}^{1T} = \mathbf{U}^1\mathbf{\Sigma}^1\mathbf{V}^{1T}.$$

Using the SVD of $\mathbf{Z}^*\mathbf{Z}^{1T}$, we rewrite the above equations as

$$\begin{aligned}\mathbf{P}\mathbf{Z}^{1T} &\overset{(A_1)}{=} \mathbf{W}^{*T}\mathbf{Z}^*\mathbf{Z}^{1T} = \underbrace{\mathbf{W}^{*T}\mathbf{U}^1_z}_{\mathbf{U}^1}\underbrace{\mathbf{\Sigma}^1_z}_{\mathbf{\Sigma}^1}\underbrace{\mathbf{V}^{1T}_z}_{\mathbf{V}^{1T}} \\ \mathbf{W}^1 &= \mathbf{V}^1_z\mathbf{U}^{1T}_z\mathbf{W}^*. \end{aligned} \tag{8}$$

[3]In this case, the support of $\mathbf{Z}^1_{(\cdot,j)}$ in fact coincides with that of $\mathbf{Z}^*_{(\cdot,j)}$. If we relaxed Assumption (A_2) from $\|\mathbf{Z}^*_{(\cdot,j)}\|_0 = s$ to $\|\mathbf{Z}^*_{(\cdot,j)}\|_0 \leq s\ \forall\ j$, then $S\left(\mathbf{Z}^*_{(\cdot,j)}\right) \subseteq S\left(\mathbf{Z}^1_{(\cdot,j)}\right)$ holds, and the lemma still holds.

Now the error between $\mathbf{W}^1$ and $\mathbf{W}^*$ satisfies

$$\|\mathbf{W}^1 - \mathbf{W}^*\|_F^2 \overset{(Eq.8)}{=} \|\mathbf{V}_z^1\mathbf{U}_z^{1T}\mathbf{W}^* - \mathbf{W}^*\|_F^2 = \|(\mathbf{V}_z^1\mathbf{U}_z^{1T} - \mathbf{Id})\mathbf{W}^*\|_F^2 = \|\mathbf{V}_z^1\mathbf{U}_z^{1T} - \mathbf{Id}\|_F^2, \quad (9)$$

where the matrix $\mathbf{V}_z^1\mathbf{U}_z^{1T}$ can be further rewritten as follows:

$$\begin{aligned}\mathbf{V}_z^1\mathbf{U}_z^{1T} &= \mathbf{V}_z^1\left(\mathbf{\Sigma}_z^1\right)^{-1}\mathbf{U}_z^{1T}\mathbf{U}_z^1\mathbf{\Sigma}_z^1\mathbf{U}_z^{1T} \\ &= \underbrace{(\mathbf{Z}^*\mathbf{Z}^{1T})^{-1}}_{(a)}\underbrace{(\mathbf{Z}^*\mathbf{Z}^{1T}\mathbf{Z}^1\mathbf{Z}^{*T})^{\frac{1}{2}}}_{(b)}. \end{aligned} \quad (10)$$

The above equality holds for all $\epsilon < 1$, which suffices to ensure $\mathbf{Z}^*\mathbf{Z}^{1T}$ is invertible.

Using Taylor Series Expansions for the matrix inverse and positive-definite square root along with (6) and the assumption $\mathbf{Z}^*\mathbf{Z}^{*T} = \mathbf{Id}$, we have that

$$\begin{aligned}(a) &= (\mathbf{Z}^*\mathbf{Z}^{1T})^{-1} \overset{(Eq.6)}{=} (\mathbf{Z}^*(\mathbf{Z}^* + \mathbf{\Delta}^1)^T)^{-1} = (\mathbf{Z}^*\mathbf{Z}^{*T} + \mathbf{Z}^*\mathbf{\Delta}^{1T})^{-1} \\ &\overset{(A_4)}{=} (\mathbf{Id} + \mathbf{Z}^*\mathbf{\Delta}^{1T})^{-1} = \mathbf{Id} - \mathbf{Z}^*\mathbf{\Delta}^{1T} + O((\mathbf{\Delta}^1)^2) \\ (b) &= (\mathbf{Z}^*\mathbf{Z}^{1T}\mathbf{Z}^1\mathbf{Z}^{*T})^{\frac{1}{2}} \overset{(A_4)}{=} \mathbf{Id} + \frac{1}{2}(\mathbf{Z}^*\mathbf{\Delta}^{1T} + \mathbf{\Delta}^1\mathbf{Z}^{*T}) + O((\mathbf{\Delta}^1)^2) \\ \mathbf{V}_z^1\mathbf{U}_z^{1T} &\overset{(Eq.10)}{=} (a)(b) \\ &= \left(\mathbf{Id} - \mathbf{Z}^*\mathbf{\Delta}^{1T} + O((\mathbf{\Delta}^1)^2)\right)\left(\mathbf{Id} + \tfrac{1}{2}(\mathbf{Z}^*\mathbf{\Delta}^{1T} + \mathbf{\Delta}^1\mathbf{Z}^{*T}) + O((\mathbf{\Delta}^1)^2)\right) \\ &= \mathbf{Id} + \frac{1}{2}(\mathbf{\Delta}^1\mathbf{Z}^{*T} - \mathbf{Z}^*\mathbf{\Delta}^{1T}) + O((\mathbf{\Delta}^1)^2) \end{aligned} \quad (11)$$

where $O((\mathbf{\Delta}^1)^2)$ denotes corresponding higher order series terms, and is bounded in norm by $C\left\|\mathbf{\Delta}^1\right\|^2$ for some constant C.

Substituting these expressions in (9), the error between the first transform iterate $\mathbf{W}^1$ and $\mathbf{W}^*$ is bounded as

$$\|\mathbf{W}^1 - \mathbf{W}^*\|_F \overset{(Eq.9)}{=} \|\mathbf{V}_z^1\mathbf{U}_z^{1T} - \mathbf{Id}\|_F \approx \frac{1}{2}\|\mathbf{\Delta}^1\mathbf{Z}^{*T} - \mathbf{Z}^*\mathbf{\Delta}^{1T}\|_F. \quad (12)$$

The approximation error above is bounded in norm by $C\epsilon^2$, which is negligible for small ϵ. So we only bound the dominant term $0.5\|\mathbf{\Delta}^1\mathbf{Z}^{*T} - \mathbf{Z}^*\mathbf{\Delta}^{1T}\|_F$ on the right. The matrix $\mathbf{\Delta}^1\mathbf{Z}^{*T} - \mathbf{Z}^*\mathbf{\Delta}^{1T}$ clearly has a zero diagonal (skew-symmetric). Thus, we have the following inequalities:

$$\begin{aligned}\|\mathbf{W}^1 - \mathbf{W}^*\|_F \approx \frac{1}{2}\|\mathbf{\Delta}^1\mathbf{Z}^{*T} - \mathbf{Z}^*\mathbf{\Delta}^{1T}\|_F &\leq \sqrt{\sum_{k=1}^{n}\|\boldsymbol{D}_k\mathbf{Z}^*\tilde{\boldsymbol{D}}_k{\mathbf{\Delta}^1_{(k,\cdot)}}^T\|_2^2} \\ &\leq \sqrt{\sum_{k=1}^{n}\|\boldsymbol{D}_k\mathbf{Z}^*\tilde{\boldsymbol{D}}_k\|_2^2\|\mathbf{\Delta}^1_{(k,\cdot)}\|_2^2} \leq \max_k\|\boldsymbol{D}_k\mathbf{Z}^*\tilde{\boldsymbol{D}}_k\|_2\sqrt{\sum_{k=1}^{n}\|\mathbf{\Delta}^1_{(k,\cdot)}\|_2^2} \\ &= \max_k\|\boldsymbol{D}_k\mathbf{Z}^*\tilde{\boldsymbol{D}}_k\|_2\|\mathbf{\Delta}^1\|_F \overset{Lem.1}{\leq} q\|\mathbf{E}^0\|_F \\ &= q\|\mathbf{W}^0 - \mathbf{W}^*\|_F, \end{aligned} \quad (13)$$

where we define $q := max_k \|\boldsymbol{D}_k \mathbf{Z}^* \tilde{\boldsymbol{D}}_k\|_2$. Since $\|\mathbf{E}^0\|_F \leq \epsilon$ by Assumption (A_5), we obtain the desired result. □

Thus, we have shown the results for the $t = 1$ case. We complete the proof of Theorem 2.1 by observing that for each subsequent iteration $t = \tau + 1$, the same steps as above can be repeated along with the induction hypothesis (IH) to show that

$$\begin{aligned} \|\mathbf{Z}^{\tau+1} - \mathbf{Z}^*\|_F = \|\boldsymbol{\Delta}^{\tau+1}\|_F &\leq \|\mathbf{E}^{\tau}\|_F \\ = \|\mathbf{W}^{\tau} - \mathbf{W}^*\|_F &\overset{(IH)}{\leq} q^{\tau}\epsilon \\ \|\mathbf{W}^{\tau+1} - \mathbf{W}^*\|_F \leq q\|\mathbf{Z}^{\tau+1} - \mathbf{Z}^*\|_F &\leq q(q^{\tau}\epsilon). \end{aligned}$$

□

2.4.2 Proof of Theorem 2.2

Here, we present the distinctions in the proof of Theorem 2.2. When Assumption (A_4) is dropped, Lemma 1 and its proof remain unaffected. The change to Lemma 2 and its proof are outlined next.

Lemma 3. *(Removing Assumption (A_4)) For $t = 1$ in Algorithm 1 and under Assumptions $(A_1) - (A_3)$ and (A_5), the Frobenious norm of the approximation error of the estimated transform with respect to $\mathbf{W}^*$ is bounded as*

$$\|\mathbf{W}^1 - \mathbf{W}^*\|_F \leq q\epsilon,$$

where q is a scalar coefficient as in Theorem 2.2.

Proof. To remove Assumption (A_4) in the proof of Lemma 2, we need to generalize the Taylor Series expansions in (11). Let $\mathbf{G} \triangleq \mathbf{Z}^* {\mathbf{Z}^*}^T$. First, we look at the series expansion of $(\mathbf{Z}^* {\mathbf{Z}^1}^T)^{-1}$, for which the following equalities hold:

$$\begin{aligned} (\mathbf{Z}^* {\mathbf{Z}^1}^T)^{-1} &= (\mathbf{Z}^* {\mathbf{Z}^*}^T + \mathbf{Z}^* (\boldsymbol{\Delta}^1)^T)^{-1} \\ &= (\mathbf{G}(\mathbf{Id} + \mathbf{G}^{-1}\mathbf{Z}^*(\boldsymbol{\Delta}^1)^T))^{-1} = (\mathbf{Id} + \mathbf{G}^{-1}\mathbf{Z}^*(\boldsymbol{\Delta}^1)^T)^{-1}\mathbf{G}^{-1} \\ &= (\mathbf{Id} - \mathbf{G}^{-1}\mathbf{Z}^*(\boldsymbol{\Delta}^1)^T + O((\boldsymbol{\Delta}^1)^2))\mathbf{G}^{-1} \\ &= \mathbf{G}^{-1} - \mathbf{G}^{-1}\mathbf{Z}^*(\boldsymbol{\Delta}^1)^T\mathbf{G}^{-1} + O((\boldsymbol{\Delta}^1)^2), \end{aligned}$$

where we factored out[4] $\mathbf{G}^{-1}$ and then computed the series expansion of a matrix inverse. The result holds for all ϵ with $\kappa^2\left(\mathbf{Z}^*\right)\epsilon < 1$. For the series expansion of the matrix square root in (10), we first observe that

$$\begin{aligned} (\mathbf{Z}^* {\mathbf{Z}^1}^T \mathbf{Z}^1 {\mathbf{Z}^*}^T)^{\frac{1}{2}} &= (\mathbf{Z}^*(\mathbf{Z}^* + \boldsymbol{\Delta}^1)^T(\mathbf{Z}^* + \boldsymbol{\Delta}^1){\mathbf{Z}^*}^T)^{\frac{1}{2}} \\ &= \left(\mathbf{Z}^*{\mathbf{Z}^*}^T\mathbf{Z}^*{\mathbf{Z}^*}^T + (\mathbf{Z}^*{\mathbf{Z}^*}^T\boldsymbol{\Delta}^1{\mathbf{Z}^*}^T + \mathbf{Z}^*{\boldsymbol{\Delta}^1}^T\mathbf{Z}^*{\mathbf{Z}^*}^T + \mathbf{Z}^*{\boldsymbol{\Delta}^1}^T\boldsymbol{\Delta}^1{\mathbf{Z}^*}^T)\right)^{\frac{1}{2}} \\ &= \left(\mathbf{G}^2 + (\mathbf{G}\boldsymbol{\Delta}^1{\mathbf{Z}^*}^T + \mathbf{Z}^*{\boldsymbol{\Delta}^1}^T\mathbf{G} + \mathbf{Z}^*{\boldsymbol{\Delta}^1}^T\boldsymbol{\Delta}^1{\mathbf{Z}^*}^T)\right)^{\frac{1}{2}}. \end{aligned}$$

[4]Matrix $\mathbf{G}$ is invertible for Assumption (A_3) to hold.

Let $F(\mathbf{G}) \triangleq \left(\mathbf{G}^2 + (\mathbf{G}\boldsymbol{\Delta}^1\mathbf{Z}^{*T} + \mathbf{Z}^*\boldsymbol{\Delta}^{1T}\mathbf{G} + \mathbf{Z}^*\boldsymbol{\Delta}^{1T}\boldsymbol{\Delta}^1\mathbf{Z}^{*T})\right)^{\frac{1}{2}} = \left(\mathbf{G}^2 + \tilde{\Delta}\right)^{\frac{1}{2}}$, where $\tilde{\Delta}$ denotes the remainder of terms within the square root. The Taylor Series expansion for $F(\mathbf{G})$ can be written as $F(\mathbf{G}) \approx \mathbf{G} + R^T(\nabla F(\mathbf{G})\text{Vec}(\tilde{\Delta}^T)) + O(\tilde{\Delta}^2)$, where the operator $\text{Vec}(\cdot)$ reshapes a matrix into a vector by stacking the columns, $R(\cdot)$ undoes or inverts the $\text{Vec}(\cdot)$ operation by reshaping a vector into an $n \times n$ matrix, and the gradient of the square root function is obtained as follows, where $\otimes$ denotes the Kronecker product and $\oplus$ denotes the Kronecker sum:

$$\nabla F(\mathbf{G}) = \frac{\partial \text{Vec}(F^T(\mathbf{G}))}{\partial \text{Vec}^T(\mathbf{G}^T)} = (\mathbf{Id} \otimes \mathbf{G} + \mathbf{G} \otimes \mathbf{Id})^{-1} = (\mathbf{G} \oplus \mathbf{G})^{-1}. \tag{14}$$

Using the above expressions, (10) in this case becomes

$$\begin{aligned}
\mathbf{V}_z^1\mathbf{U}_z^{1T} &= (\mathbf{Z}^*\mathbf{Z}^{1T})^{-1}(\mathbf{Z}^*\mathbf{Z}^{1T}\mathbf{Z}^1\mathbf{Z}^{*T})^{\frac{1}{2}} \\
&= \mathbf{Id} - \mathbf{G}^{-1}\mathbf{Z}^*\boldsymbol{\Delta}^{1T} + \mathbf{G}^{-1}R^T((\mathbf{G} \oplus \mathbf{G})^{-1}\text{Vec}(\tilde{\Delta}^T)) + O((\boldsymbol{\Delta}^1)^2), \\
&= \mathbf{Id} - \mathbf{G}^{-1}\mathbf{Z}^*\boldsymbol{\Delta}^{1T} + \mathbf{G}^{-1}R^T((\mathbf{G} \oplus \mathbf{G})^{-1}\text{Vec}(\mathbf{G}\boldsymbol{\Delta}^1\mathbf{Z}^{*T} + \mathbf{Z}^*\boldsymbol{\Delta}^{1T}\mathbf{G})) + O((\boldsymbol{\Delta}^1)^2),
\end{aligned} \tag{15}$$

with $O((\boldsymbol{\Delta}^1)^2)$ denoting corresponding higher order series terms in each step.

Now recall from (12) that $\|\mathbf{W}^1 - \mathbf{W}^*\|_F = \|\mathbf{V}_z^1\mathbf{U}_z^{1T} - \mathbf{Id}\|_F = \|\mathbf{B}\|_F$, where $\mathbf{B} \triangleq \mathbf{V}_z^1\mathbf{U}_z^{1T} - \mathbf{Id} = -\mathbf{G}^{-1}\mathbf{Z}^*(\boldsymbol{\Delta}^1)^T + \mathbf{G}^{-1}R^T((\mathbf{G} \oplus \mathbf{G})^{-1}\text{Vec}(\mathbf{G}\boldsymbol{\Delta}^1\mathbf{Z}^{*T} + \mathbf{Z}^*(\boldsymbol{\Delta}^1)^T\mathbf{G})) + O((\boldsymbol{\Delta}^1)^2)$. First, using the property of the $\text{Vec}(\cdot)$ operator that $\text{Vec}(\mathbf{AXC}) = (\mathbf{C}^T \otimes \mathbf{A})\text{Vec}(\mathbf{X})$, we can easily obtain a simplified expression for $\mathbf{B}$ ignoring the $O((\boldsymbol{\Delta}^1)^2)$ terms (since they are bounded in norm by $C\epsilon^2$, which is negligible for small ϵ and C is a constant) in (15) as follows:

$$\begin{aligned}
\text{Vec}(\mathbf{B}^T) = &- (\mathbf{G}^{-1} \otimes \mathbf{Id})\text{Vec}(\boldsymbol{\Delta}^1\mathbf{Z}^{*T}) + (\mathbf{G}^{-1} \otimes \mathbf{Id})(\mathbf{G} \oplus \mathbf{G})^{-1}(\mathbf{Id} \otimes \mathbf{G})\text{Vec}(\boldsymbol{\Delta}^1\mathbf{Z}^{*T}) \\
&+ (\mathbf{G}^{-1} \otimes \mathbf{Id})(\mathbf{G} \oplus \mathbf{G})^{-1}(\mathbf{G} \otimes \mathbf{Id})\text{Vec}(\mathbf{Z}^*(\boldsymbol{\Delta}^1)^T).
\end{aligned} \tag{16}$$

Denoting the SVD of (positive-definite) $\mathbf{G}$ as $\mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T$, it can be shown that the SVD of the Kronecker sum $\mathbf{G} \oplus \mathbf{G}$ is[5] $(\mathbf{Q} \otimes \mathbf{Q})(\boldsymbol{\Lambda} \oplus \boldsymbol{\Lambda})(\mathbf{Q} \otimes \mathbf{Q})^T$, or that $(\mathbf{G} \oplus \mathbf{G})^{-1} = (\mathbf{Q} \otimes \mathbf{Q})(\boldsymbol{\Lambda} \oplus \boldsymbol{\Lambda})^{-1}(\mathbf{Q} \otimes \mathbf{Q})^T$. Using these SVDs and the standard result that

$$(\mathbf{H}_1 \otimes \mathbf{H}_2)(\mathbf{H}_3 \otimes \mathbf{H}_4) = (\mathbf{H}_1\mathbf{H}_3 \otimes \mathbf{H}_2\mathbf{H}_4), \tag{17}$$

the following results readily hold:

$$\begin{aligned}
(\mathbf{G} \oplus \mathbf{G})^{-1}(\mathbf{Id} \otimes \mathbf{G}) &= (\mathbf{Q} \otimes \mathbf{Q})(\boldsymbol{\Lambda} \oplus \boldsymbol{\Lambda})^{-1}(\mathbf{Q}^T \otimes \mathbf{Q}^T)(\mathbf{Id} \otimes \mathbf{G}) \\
&= (\mathbf{Q} \otimes \mathbf{Q})(\boldsymbol{\Lambda} \oplus \boldsymbol{\Lambda})^{-1}(\mathbf{Q}^T \otimes \mathbf{Q}^T\mathbf{G}) \\
&= (\mathbf{Q} \otimes \mathbf{Q})(\boldsymbol{\Lambda} \oplus \boldsymbol{\Lambda})^{-1}(\mathbf{Q}^T \otimes \boldsymbol{\Lambda}\mathbf{Q}^T) \\
&= (\mathbf{Q} \otimes \mathbf{Q})(\boldsymbol{\Lambda} \oplus \boldsymbol{\Lambda})^{-1}(\mathbf{Id} \otimes \boldsymbol{\Lambda})(\mathbf{Q}^T \otimes \mathbf{Q}^T),
\end{aligned} \tag{18}$$

[5]The SVD of the Kronecker sum is established by the following equalities that use the definitions of the Kronecker sum and SVD of $\mathbf{G}$ and (17): $\mathbf{G} \oplus \mathbf{G} = \mathbf{Id} \otimes \mathbf{G} + \mathbf{G} \otimes \mathbf{Id} = \mathbf{Q}\,\mathbf{Id}\,\mathbf{Q}^T \otimes \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T + \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^T \otimes \mathbf{Q}\,\mathbf{Id}\,\mathbf{Q}^T = (\mathbf{Q} \otimes \mathbf{Q})(\mathbf{Id} \otimes \boldsymbol{\Lambda})(\mathbf{Q}^T \otimes \mathbf{Q}^T) + (\mathbf{Q} \otimes \mathbf{Q})(\boldsymbol{\Lambda} \otimes \mathbf{Id})(\mathbf{Q}^T \otimes \mathbf{Q}^T) = (\mathbf{Q} \otimes \mathbf{Q})(\boldsymbol{\Lambda} \oplus \boldsymbol{\Lambda})(\mathbf{Q}^T \otimes \mathbf{Q}^T)$.

$$
\begin{aligned}
(\mathbf{G}\oplus\mathbf{G})^{-1}(\mathbf{G}\otimes\mathbf{Id}) &= (\mathbf{Q}\otimes\mathbf{Q})(\mathbf{\Lambda}\oplus\mathbf{\Lambda})^{-1}(\mathbf{Q}^T\mathbf{G}\otimes\mathbf{Q}^T)\\
&= (\mathbf{Q}\otimes\mathbf{Q})(\mathbf{\Lambda}\oplus\mathbf{\Lambda})^{-1}(\mathbf{\Lambda}\mathbf{Q}^T\otimes\mathbf{Q}^T)\\
&= (\mathbf{Q}\otimes\mathbf{Q})(\mathbf{\Lambda}\oplus\mathbf{\Lambda})^{-1}(\mathbf{\Lambda}\otimes\mathbf{Id})(\mathbf{Q}^T\otimes\mathbf{Q}^T). \qquad (19)
\end{aligned}
$$

Substituting (18) and (19) in (16) yields

$$
\begin{aligned}
\mathrm{Vec}(\mathbf{B}^T) = (\mathbf{G}^{-1}\otimes\mathbf{Id})(\mathbf{Q}\otimes\mathbf{Q})\Bigg(&\left[(\mathbf{\Lambda}\oplus\mathbf{\Lambda})^{-1}(\mathbf{Id}\otimes\mathbf{\Lambda})-\mathbf{Id}\right](\mathbf{Q}^T\otimes\mathbf{Q}^T)\mathrm{Vec}(\mathbf{\Delta}^1\mathbf{Z}^{*T})\\
&+(\mathbf{\Lambda}\oplus\mathbf{\Lambda})^{-1}(\mathbf{\Lambda}\otimes\mathbf{Id})(\mathbf{Q}^T\otimes\mathbf{Q}^T)\mathrm{Vec}(\mathbf{Z}^*(\mathbf{\Delta}^1)^T)\Bigg). \qquad (20)
\end{aligned}
$$

Moreover, we have that

$$
\begin{aligned}
(\mathbf{\Lambda}\oplus\mathbf{\Lambda})^{-1}(\mathbf{Id}\otimes\mathbf{\Lambda})-\mathbf{Id} &= (\mathbf{\Lambda}\oplus\mathbf{\Lambda})^{-1}\left((\mathbf{Id}\otimes\mathbf{\Lambda})-(\mathbf{\Lambda}\oplus\mathbf{\Lambda})\right)\\
&= (\mathbf{\Lambda}\oplus\mathbf{\Lambda})^{-1}\left((\mathbf{Id}\otimes\mathbf{\Lambda})-(\mathbf{Id}\otimes\mathbf{\Lambda}+\mathbf{\Lambda}\otimes\mathbf{Id})\right) = -(\mathbf{\Lambda}\oplus\mathbf{\Lambda})^{-1}(\mathbf{\Lambda}\otimes\mathbf{Id}). \qquad (21)
\end{aligned}
$$

Thus, equation (20) simplifies as follows:

$$
\mathrm{Vec}(\mathbf{B}^T) = \mathbf{H}\,\mathrm{Vec}\left(\mathbf{Z}^*(\mathbf{\Delta}^1)^T-\mathbf{\Delta}^1\mathbf{Z}^{*T}\right), \qquad (22)
$$

where the matrix $\mathbf{H}$ is defined as

$$
\mathbf{H}\triangleq(\mathbf{G}^{-1}\otimes\mathbf{Id})(\mathbf{Q}\otimes\mathbf{Q})(\mathbf{\Lambda}\oplus\mathbf{\Lambda})^{-1}(\mathbf{\Lambda}\otimes\mathbf{Id})(\mathbf{Q}^T\otimes\mathbf{Q}^T). \qquad (23)
$$

Finally, we use (22) to obtain

$$
\|\mathbf{W}^1-\mathbf{W}^*\|_F = \|\mathbf{V}_z^1\mathbf{U}_z^{1^T}-\mathbf{Id}\|_F \approx \|\mathbf{H}\,\mathrm{Vec}\left(\mathbf{Z}^*(\mathbf{\Delta}^1)^T-\mathbf{\Delta}^1\mathbf{Z}^{*T}\right)\|_2 \le \|\mathbf{H}\|_2\|\mathbf{\Delta}^1\mathbf{Z}^{*T}-\mathbf{Z}^*(\mathbf{\Delta}^1)^T\|_F. \qquad (24)
$$

Here, the submultiplicativity of the spectral norm and the fact that $\|\mathbf{H}_1\otimes\mathbf{H}_2\|_2 = \|\mathbf{H}_1\|_2\,\|\mathbf{H}_2\|_2$ ensures that

$$
\begin{aligned}
\|\mathbf{H}\|_2 &\le \left\|\mathbf{G}^{-1}\otimes\mathbf{Id}\right\|_2\left\|\mathbf{Q}\otimes\mathbf{Q}\right\|_2\left\|(\mathbf{\Lambda}\oplus\mathbf{\Lambda})^{-1}\right\|_2\left\|\mathbf{\Lambda}\otimes\mathbf{Id}\right\|_2\left\|\mathbf{Q}^T\otimes\mathbf{Q}^T\right\|_2\\
&= \left\|\mathbf{G}^{-1}\right\|_2\left\|\mathbf{Q}\right\|_2^2\left\|(\mathbf{\Lambda}\oplus\mathbf{\Lambda})^{-1}\right\|_2\left\|\mathbf{\Lambda}\right\|_2\left\|\mathbf{Q}^T\right\|_2^2 = \frac{\kappa^4(\mathbf{Z}^*)}{2}, \qquad (25)
\end{aligned}
$$

where the last equality follows from the facts that $\left\|\mathbf{Q}\right\|_2 = 1$ (for unitary matrix); $\left\|\mathbf{\Lambda}\right\|_2 = \left\|\mathbf{G}\right\|_2 = \left\|\mathbf{Z}^*\right\|_2^2 = \left\|\mathbf{P}\right\|_2^2 = 1$ (by Assumption (A_1)); $\left\|(\mathbf{\Lambda}\oplus\mathbf{\Lambda})^{-1}\right\|_2 = 0.5\left\|\mathbf{G}^{-1}\right\|_2 = 0.5\sigma_n^{-1}(\mathbf{G}) = 0.5\sigma_n^{-2}(\mathbf{Z}^*)$, where $\sigma_n(\cdot)$ denotes the smallest matrix singular value; and the fact that $\kappa(\mathbf{Z}^*) = \sigma_1(\mathbf{Z}^*)/\sigma_n(\mathbf{Z}^*) = \sigma_n^{-1}(\mathbf{Z}^*)$ (using Assumption (A_1)). Substituting (25) in (24) and using a similar set of inequalities as in (13) to bound the $\|\mathbf{\Delta}^1\mathbf{Z}^{*T}-\mathbf{Z}^*(\mathbf{\Delta}^1)^T\|_F$ term in (24) provides the following bound:

$$
\|\mathbf{W}^1-\mathbf{W}^*\|_F \approx \|\mathbf{H}\,\mathrm{Vec}\left(\mathbf{Z}^*(\mathbf{\Delta}^1)^T-\mathbf{\Delta}^1\mathbf{Z}^{*T}\right)\|_2 \le q\|\mathbf{W}^0-\mathbf{W}^*\|_F, \qquad (26)
$$

where $q=\kappa^4(\mathbf{Z}^*)max_k\|\boldsymbol{D}_k\mathbf{Z}^*\tilde{\boldsymbol{D}}_k\|_2$. Since by Assumption (A_5), $\|\mathbf{E}^0\|_F\le\epsilon$, we obtain the desired result. □

2.4.3 Proof of Corollary 2.3

We have $q = \max_k \|\boldsymbol{D}_k \mathbf{Z}^* \tilde{\boldsymbol{D}}_k\|_2$ with $\mathbf{Z}^* \mathbf{Z}^{*T} = \mathbf{Id}$ by assumptions (A_3) and (A_4), respectively. For brevity in notation, let $\mathbf{M}_k = \boldsymbol{D}_k \mathbf{Z}^* \tilde{\boldsymbol{D}}_k$. Here the matrix $\boldsymbol{D}_k$ zeros out the kth row of $\mathbf{Z}^*$ and $\tilde{\boldsymbol{D}}_k$ zeros out the columns corresponding to the complement of the support of the kth row of $\mathbf{Z}^*$.

The matrix $\mathbf{M}_k \mathbf{M}_k^T$ is then a diagonal matrix where the $(k,k)^{\text{th}}$ entry is 0 and the $(i,i)^{\text{th}}$ entry for $i \neq k$ is $\|{\mathbf{Z}^*}_{(i,\cdot)}|_{S({\mathbf{Z}^*}_{(i,\cdot)}) \cap S({\mathbf{Z}^*}_{(k,\cdot)})}\|_2^2$, where ${\mathbf{Z}^*}_{(i,\cdot)}|_{S({\mathbf{Z}^*}_{(i,\cdot)}) \cap S({\mathbf{Z}^*}_{(k,\cdot)})}$ coincides with ${\mathbf{Z}^*}_{(i,\cdot)}$ on $S({\mathbf{Z}^*}_{(i,\cdot)}) \cap S({\mathbf{Z}^*}_{(k,\cdot)})$ and is zero outside this support. Clearly, the kth row and column of $\mathbf{M}_k \mathbf{M}_k^T$ are zero and its other off-diagonal entries are $\langle {\mathbf{Z}^*}_{(i,\cdot)}|_{S({\mathbf{Z}^*}_{(i,\cdot)}) \cap S({\mathbf{Z}^*}_{(k,\cdot)})}, {\mathbf{Z}^*}_{(j,\cdot)}|_{S({\mathbf{Z}^*}_{(j,\cdot)}) \cap S({\mathbf{Z}^*}_{(k,\cdot)})} \rangle = 0$ because each column of $\mathbf{Z}^*$ has at most $s = 2$ non-zeros and $S({\mathbf{Z}^*}_{(i,\cdot)}) \cap S({\mathbf{Z}^*}_{(j,\cdot)}) \cap S({\mathbf{Z}^*}_{(k,\cdot)}) = \emptyset$ for $i \neq j \neq k$. So, we have that

$$\begin{aligned} q^2 &= \max_k \|\mathbf{M}_k \mathbf{M}_k^T\|_2 \\ &= \max_{1 \leq k \leq n} \max_{i \neq k} \|{\mathbf{Z}^*}_{(i,\cdot)}|_{S({\mathbf{Z}^*}_{(i,\cdot)}) \cap S({\mathbf{Z}^*}_{(k,\cdot)})}\|_2^2 \\ &< 1, \end{aligned}$$

where the last inequality bound follows from the fact that $\|{\mathbf{Z}^*}_{(i,\cdot)}|_{S({\mathbf{Z}^*}_{(i,\cdot)}) \cap S({\mathbf{Z}^*}_{(k,\cdot)})}\|_2^2 < 1$ for all $i \neq k$, which holds because each row of $\mathbf{Z}^*$ has unit ℓ_2 norm (assumption (A_4)) and no two rows have the exact same support. □

2.5 Proof of Proposition 1

Under the conditions stated in Proposition 1, the q factor is expected to be less than 1 given sufficient training signals, i.e., large N. For the proof, we study the asymptotic behavior of the matrices $\mathbf{H} \triangleq \mathbf{Z}^* \mathbf{Z}^{*T}$ and $\mathbf{G} \triangleq \mathbf{M}^k \left(\mathbf{M}^k\right)^T$, where $\mathbf{M}^k \triangleq \boldsymbol{D}_k \mathbf{Z}^* \tilde{\boldsymbol{D}}_k$, which appear in $q_N = \left(\kappa^4\left(\mathbf{Z}^*\right)/\|\mathbf{P}\|_2\right) \max_{1 \leq k \leq n} \|\boldsymbol{D}_k \mathbf{Z}^* \tilde{\boldsymbol{D}}_k\|_2$ as defined in Remark 1. First, we show that $\left(\kappa^4\left(\mathbf{Z}^*\right)/\|\mathbf{P}\|_2\right) \to 1$ almost surely as $N \to \infty$ using $\mathbf{H}$. Then, we will show that $\|\boldsymbol{D}_k \mathbf{Z}^* \tilde{\boldsymbol{D}}_k\|_2 \to 1$ almost surely as $N \to \infty$ using $\mathbf{G}$.

Let $\tilde{\mathbf{Z}}^* = \sqrt{N}\mathbf{Z}^*$. Then the nonzero entries of $\tilde{\mathbf{Z}}^*$ have zero mean and variance of n/s. Let $\mathbb{1}_{\left\{j \in S({\mathbf{Z}^*}_{(.,l)})\right\}}$ denote the indicator function that takes the value 1 when $j \in S({\mathbf{Z}^*}_{(.,l)})$ and is zero otherwise. Since $\mathbf{Z}^* \mathbf{Z}^{*T} = N^{-1} \tilde{\mathbf{Z}}^* \tilde{\mathbf{Z}}^{*T}$, using the law of large numbers, the diagonal entries of $\mathbf{H}$ converge almost surely as follows:

$$\lim_{N \to \infty} \mathbf{H}_{(j,j)} = \lim_{N \to \infty} \frac{1}{N} \sum_{l=1}^{N} \tilde{\mathbf{Z}}^{*2}_{(j,l)} \mathbb{1}_{\left\{j \in S({\mathbf{Z}^*}_{(.,l)})\right\}} = E\left[\tilde{\mathbf{Z}}^{*2}_{(j,l)} \mathbb{1}_{\left\{j \in S({\mathbf{Z}^*}_{(.,l)})\right\}}\right] = 1, \tag{27}$$

where $b \triangleq \tilde{\mathbf{Z}}^{*2}_{(j,l)} \mathbb{1}_{\left\{j \in S({\mathbf{Z}^*}_{(.,l)})\right\}}$ is i.i.d. over the columns l. The random variable b is nonzero (the nonzero part has mean n/s) with probability (w.p.)[6] s/n and is zero w.p. $1 - (s/n)$, implying $E[b] = 1$. Similarly, the off-diagonal entries $\mathbf{H}_{(i,j)}$ for $i \neq j$ converge as follows:

$$\lim_{N \to \infty} \mathbf{H}_{(i,j)} = \lim_{N \to \infty} \frac{1}{N} \sum_{l=1}^{N} \tilde{\mathbf{Z}}^*_{(i,l)} \tilde{\mathbf{Z}}^*_{(j,l)} \mathbb{1}_{\left\{i,j \in S({\mathbf{Z}^*}_{(.,l)})\right\}} = E\left[\tilde{\mathbf{Z}}^*_{(i,l)} \tilde{\mathbf{Z}}^*_{(j,l)} \mathbb{1}_{\left\{i,j \in S({\mathbf{Z}^*}_{(.,l)})\right\}}\right] = 0, \tag{28}$$

[6] The probability that $j \in S\left({\mathbf{Z}^*}_{(.,l)}\right)$ is $\frac{\binom{n-1}{s-1}}{\binom{n}{s}} = \frac{s}{n}$.

where $h \triangleq \tilde{\mathbf{Z}}^*_{(i,l)}\tilde{\mathbf{Z}}^*_{(j,l)}\mathbb{1}_{\left\{i,j\in S(\mathbf{Z}^*{}_{(.,l)})\right\}}$ is nonzero w.p.[7] $r = s(s-1)/n(n-1)$ and zero w.p. $1-r$, implying $E[h] = \left(s(s-1)/n(n-1)\right)E[a] = 0$, where a is the product of two i.i.d. zero mean random variables. Therefore, from (27) and (28), it follows that $\mathbf{H} = \mathbf{Z}^*\mathbf{Z}^{*T}$ converges to $\mathbf{Id}$ almost surely. Thus, as $N \to \infty$, $\kappa^4\left(\mathbf{Z}^*\right)/\|\mathbf{P}\|_2 = \kappa^2\left(\mathbf{H}\right)/\sqrt{\|\mathbf{H}\|_2}$ in the definition of q_N, converges to 1 almost surely.

Now consider $\mathbf{G}$ and note that the kth row and column of the matrix $\mathbf{G}$ are zero. As $N \to \infty$, the diagonal entries of $\mathbf{G}$ have the following limit almost surely:

$$\begin{aligned}\lim_{N\to\infty}\mathbf{G}_{(j,j)} &= \lim_{N\to\infty}\frac{1}{N}\sum_{l=1}^{N}\tilde{\mathbf{Z}}^{*^2}_{(j,l)}\mathbb{1}_{\left\{l\in S(\mathbf{Z}^*{}_{(j,.)})\cap S(\mathbf{Z}^*{}_{(k,.)})\right\}}\\ &= E\left[\tilde{\mathbf{Z}}^{*^2}_{(j,l)}\mathbb{1}_{\left\{l\in S(\mathbf{Z}^*{}_{(j,.)})\cap S(\mathbf{Z}^*{}_{(k,.)})\right\}}\right]\\ &= \frac{s(s-1)}{n(n-1)}\times\frac{n}{s} = \frac{s-1}{n-1},\end{aligned} \tag{29}$$

which holds for all $j \neq k$. The expectation follows from the fact that $\tilde{\mathbf{Z}}^{*^2}_{(j,l)}\mathbb{1}_{\left\{l\in S(\mathbf{Z}^*{}_{(j,.)})\cap S(\mathbf{Z}^*{}_{(k,.)})\right\}}$ is i.i.d. over the columns[8] l, is nonzero (mean n/s for nonzero part) w.p. $s(s-1)/n(n-1)$, and is zero otherwise. The following limit holds almost surely for the off-diagonal entries of $\mathbf{G}$:

$$\begin{aligned}\lim_{N\to\infty}\mathbf{G}_{(i,j)} &= \lim_{N\to\infty}\frac{1}{N}\sum_{l=1}^{N}\tilde{\mathbf{Z}}^*_{(i,l)}\tilde{\mathbf{Z}}^*_{(j,l)}\mathbb{1}_{\left\{l\in S(\mathbf{Z}^*{}_{(i,.)})\cap S(\mathbf{Z}^*{}_{(j,.)})\cap S(\mathbf{Z}^*{}_{(k,.)})\right\}}\\ &= E\left[\tilde{\mathbf{Z}}^*_{(i,l)}\tilde{\mathbf{Z}}^*_{(j,l)}\mathbb{1}_{\left\{l\in S(\mathbf{Z}^*{}_{(i,.)})\cap S(\mathbf{Z}^*{}_{(j,.)})\cap S(\mathbf{Z}^*{}_{(k,.)})\right\}}\right] = 0,\end{aligned} \tag{30}$$

which follows because the indexes i, j, and k all lie in the support of the lth column (to get non-zero indicator function) w.p. $\frac{s(s-1)(s-2)}{n(n-1)(n-2)}$, and the expectation of the product of zero mean i.i.d. random variables is zero. It is obvious from (29) and (30) that

$$\lim_{N\to\infty}\mathbf{M}^k\left(\mathbf{M}^k\right)^T = \frac{s-1}{n-1}\boldsymbol{D}_k \text{ a.s.} \tag{31}$$

Thus, as $N \to \infty$, $\left\|\boldsymbol{D}_k\mathbf{Z}^*\tilde{\boldsymbol{D}}_k\right\|_2 = \left\|\mathbf{M}^k\right\|_2 \to \sqrt{s-1/n-1}$ almost surely, and the same is true for $\max_{1\le k\le n}\|\mathbf{M}^k\|_2$. Combining all the above results, the required result (4) is readily established. □

3 Experiments

In this section, we provide numerical results supporting our findings. We also discuss the empirical behavior of the algorithm with respect to different initializations.

[7] This is the probability that the two indexes i and j both appear in the support of the lth column of $\mathbf{Z}^*$. Thus, $r = \frac{\binom{n-2}{s-2}}{\binom{n}{s}} = \frac{s(s-1)}{n(n-1)}$.

[8] Note that $\mathbb{1}_{\left\{l\in S(\mathbf{Z}^*{}_{(j,.)})\cap S(\mathbf{Z}^*{}_{(k,.)})\right\}} = \mathbb{1}_{\left\{j,k\in S(\mathbf{Z}^*{}_{(.,l)})\right\}}$.

3.1 Empirical Performance of Algorithm

In the first two experiments, we generated the training set $\mathbf{P}$ using randomly generated $\mathbf{W}^*$ and $\mathbf{Z}^*$, and set $n = 50$, $N = 10000$, and $s = \{5, 10\}$. The transform $\mathbf{W}^*$ is generated in each case by applying Matlab's `orth()` function on a standard Gaussian matrix. For generating $\mathbf{Z}^*$, the support of each column is chosen uniformly at random and the nonzero entries are drawn i.i.d. from a Gaussian distribution with mean zero and variance n/sN. Section 2 (Theorems 2.1 and 2.2) established model recovery guarantees for Algorithm 1. Figure 1 shows the empirical evolution of the Frobenius norm of the approximation error of the transform iterates with respect to $\mathbf{W}^*$, for an ϵ initialization ($\epsilon = 0.49 \min_j \beta \left(\mathbf{Z}^*_{(\cdot,j)} / \|\mathbf{Z}^*_{(\cdot,j)}\|_2 \right)$ – see (7)). The plots illustrate the observed linear convergence of the iterates to the underlying generative operator $\mathbf{W}^*$.

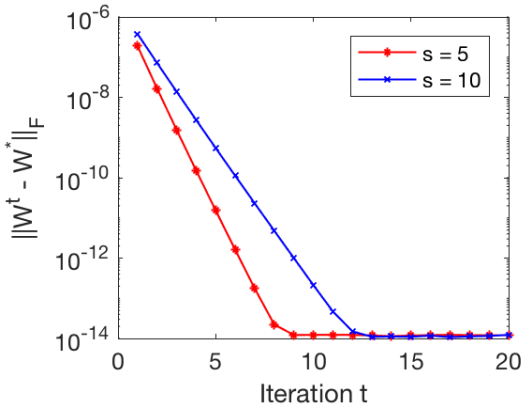

Figure 1: The performance of Algorithm 1 for recovering $\mathbf{W}^*$ for $s = 5$ and $s = 10$.

Figure 2 shows the behavior of Algorithm 1 with different initializations. We plot the evolution of the objective function over iterations and consider six different initializations. The first, labeled 'eps', denotes an initialization as in Fig. 1 with $\epsilon = 0.49 \min_j \beta \left(\mathbf{Z}^*_{(\cdot,j)} / \|\mathbf{Z}^*_{(\cdot,j)}\|_2 \right)$. The other initializations are as follows: entries of $\mathbf{W}^0$ drawn i.i.d. from a standard Gaussian distribution (labeled 'rand'); an $n \times n$ identity matrix $\mathbf{W}^0$ labeled 'id'; a discrete cosine transform (DCT) initialization labeled 'dct'; entries of $\mathbf{W}^0$ drawn i.i.d. from a uniform distribution ranging from 0 to 1 (labeled 'unif'); and $\mathbf{W}^0 = \mathbf{0}^{n \times n}$ labeled 'zero'. Note that the minimum objective value in (1) is 0. For non-epsilon initalizations, we see that the behavior of Algorithm 1 is split into two phases. In the first phase, the iterates slowly decrease the objective. When the iterates are close enough to a solution, the second phase occurs and during this phase, Algorithm 1 enjoys rapid convergence (towards 0). Note that the objective's convergence rate in the second phase is similar to that of the 'eps' case. Additionally, for different initializations, the algorithm converged to a scaled, row permuted version of the predetermined $\mathbf{W}^*$. The behavior of Algorithm 1 is similar for $s = 5$ and $s = 10$, with the latter case taking more iterations to enter the second phase of convergence. This makes sense since there are more coefficients to learn for larger s. This experiment shows that Algorithm 1 is robust to initialization.

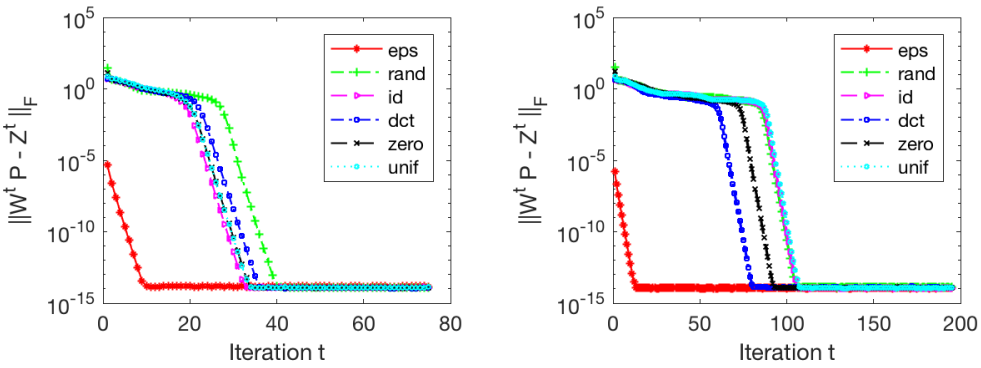

Figure 2: The performance of Algorithm 1 with various initializations for $s = 5$ (left) and $s = 10$ (right).

3.2 The q factor in Proposition 1

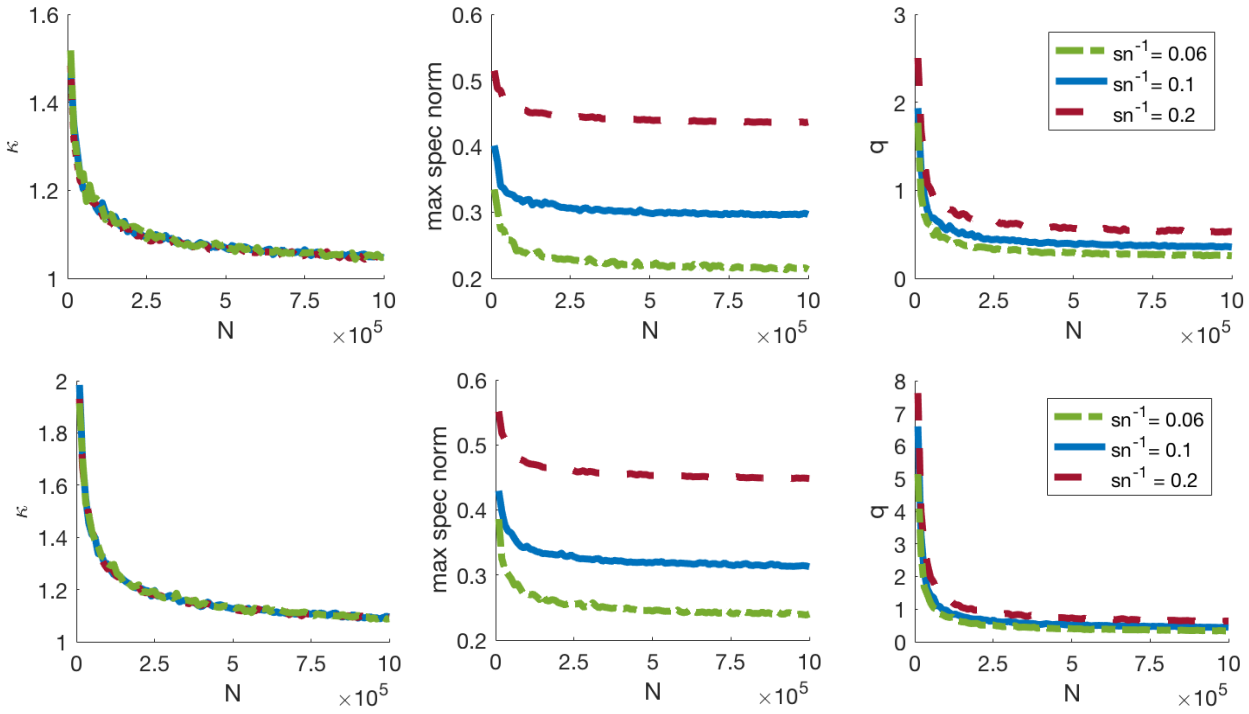

Figure 3: On the x-axis we plot the number of training data points N and on the y-axis, (Left) the condition number $\kappa(\mathbf{Z}^*)$, (Center) the maximum spectral norm over choice k for $\boldsymbol{D}_k\mathbf{Z}^*\tilde{\boldsymbol{D}}_k$, and (Right) the contraction factor $q = (\kappa^4(\mathbf{Z}^*)/\|\mathbf{Z}^*\|_2)\max_k \|\boldsymbol{D}_k\mathbf{Z}^*\tilde{\boldsymbol{D}}_k\|_2$. The top of plots corresponds to $n = 50$ and the bottom row of plots corresponds to $n = 100$.

In our last experiment, we illustrate Proposition 1 empirically. For each trial, we fix the signal dimension to be $n = \{50, 100\}$. In addition to varying N, we vary $s/n = \{0.06, 0.1, 0.2\}$. For each column of the generating $\mathbf{Z}^*$, s indexes are selected uniformly at random to be nonzero i.i.d. Gaussian entries with mean 0 and variance n/sN. We then

compute the following functions of $\mathbf{Z}^*$: the condition number $\kappa(\mathbf{Z}^*)$, the maximum spectral norm over choice k for $\boldsymbol{D}_k \mathbf{Z}^* \tilde{\boldsymbol{D}}_k$, and the contraction factor q that is a function of these quantities. Fig. 3 plots these quantities and clearly shows that $q < 1$ for large N for each n and s/n setting. The maximum spectral norm plots quickly converged close to their expected values of $\sqrt{s - 1/n - 1}$. Moreover, $\kappa(\mathbf{Z}^*)$ approaches close to 1 as N increases as expected, indicating that the probabilistic sparsity model approaches the scenario in Theorem 2.1. We have observed similar empirical behavior for the q factor, when the non-zero entries are drawn from other distributions such as scaled random signs.

4 Conclusion

In this work, we presented a study of the model recovery properties of the alternating minimization algorithm for structured (unitary) sparsifying transform learning. The algorithm converges rapidly to the generative model(s) from local neighborhoods under mild assumptions. The assumptions were shown to hold for well-known probabilistic models. In practice, the sparsifying operator learning method is robust to initialization. Our numerical results showed that the algorithm performs well under various initializations, with similar eventual rates of convergence. We have observed empirically that the algorithm converges to the specific $\mathbf{W}^*$ even with quite large perturbations for the initial $\mathbf{W}^0$ from $\mathbf{W}^*$ (i.e., large ϵ values in Assumption (A_5)). We plan to analyze the effects of initialization in more detail in future work.

Funding

This work was supported by the Office of Naval Research [grant number N00014-15-1-2141 to S.R.]; Defense Advanced Research Projects Agency Young Faculty Award [grant number D14AP00086 to S.R.]; US Army Research Office Multidisciplinary University Research Initiative [grant numbers W911NF-11-1-0391, 2015-05174-05 to S.R.]; National Institutes of Health [grant numbers R01 EB 023618, P01 CA 059827 to S.R.]; and the University of Michigan-Shanghai Jiao Tong University seed grant to S.R. This material was also supported by the National Science Foundation [grant number DMS-1440140 to A.M. and D.N.] while the authors were in residence at the Mathematical Science Research Institute in Berkeley, California, during the Fall 2017 semester, National Science Foundation CAREER Award [grant number 1348721 to A.M. and D.N.]; and National Science Foundation BIGDATA Award [grant number 1740325 to A.M. and D.N.].

References

[1] A. Agarwal, A. Anandkumar, P. Jain, and P. Netrapalli. Learning sparsely used overcomplete dictionaries via alternating minimization. *SIAM Journal on Optimization*, 26(4):2775–2799, 2016.

[2] A. Agarwal, A. Anandkumar, P. Jain, P. Netrapalli, and R. Tandon. Learning sparsely used overcomplete dictionaries. *Journal of Machine Learning Research*, 35:1–15, 2014.

[3] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing*, 54(11):4311–4322, 2006.

[4] S. Arora, R. Ge, and A. Moitra. New algorithms for learning incoherent and overcomplete dictionaries. In *Proceedings of The 27th Conference on Learning Theory*, pages 779–806, 2014.

[5] C. Bao, H. Ji, Y. Quan, and Z. Shen. L0 norm based dictionary learning by proximal methods with global convergence. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3858–3865, 2014.

[6] C. Bao, H. Ji, Y. Quan, and Z. Shen. Dictionary learning for sparse coding: Algorithms and convergence analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(7):1356–1369, July 2016.

[7] Chenglong Bao, Hui Ji, and Zuowei Shen. Convergence analysis for iterative data-driven tight frame construction scheme. *Applied and Computational Harmonic Analysis*, 38(3):510–523, 2015.

[8] D. Barchiesi and M. D. Plumbley. Learning incoherent dictionaries for sparse approximation using iterative projections and rotations. *IEEE Transactions on Signal Processing*, 61(8):2055–2065, 2013.

[9] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20(1):33–61, 1998.

[10] W. Dai and O. Milenkovic. Subspace pursuit for compressive sensing signal reconstruction. *IEEE Trans. Information Theory*, 55(5):2230–2249, 2009.

[11] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32:407–499, 2004.

[12] M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Trans. Image Process.*, 15(12):3736–3745, 2006.

[13] K. Engan, S.O. Aase, and J.H. Hakon-Husoy. Method of optimal directions for frame design. In *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 2443–2446, 1999.

[14] S. Kong and D. Wang. A dictionary learning approach for classification: Separating the particularity and the commonality. In *Proceedings of the 12th European Conference on Computer Vision*, pages 186–199, 2012.

[15] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *J. Mach. Learn. Res.*, 11:19–60, 2010.

[16] J. Mairal, M. Elad, and G. Sapiro. Sparse representation for color image restoration. *IEEE Trans. on Image Processing*, 17(1):53–69, 2008.

[17] S. G. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

[18] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM J. Comput.*, 24(2):227–234, April 1995.

[19] D. Needell and J.A. Tropp. Cosamp: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26(3):301–321, 2009.

[20] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

[21] B. Ophir, M. Lustig, and M. Elad. Multi-scale dictionary learning using wavelets. *IEEE Journal of Selected Topics in Signal Processing*, 5(5):1014–1024, 2011.

[22] Y. Pati, R. Rezaiifar, and P. Krishnaprasad. Orthogonal matching pursuit : recursive function approximation with applications to wavelet decomposition. In *Asilomar Conf. on Signals, Systems and Comput.*, pages 40–44 vol.1, 1993.

[23] I. Ramirez, P. Sprechmann, and G. Sapiro. Classification and clustering via dictionary learning with structured incoherence and shared features. In *Proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) 2010*, pages 3501–3508, 2010.

[24] S. Ravishankar and Y. Bresler. MR image reconstruction from highly undersampled k-space data by dictionary learning. *IEEE Trans. Med. Imag.*, 30(5):1028–1041, 2011.

[25] S. Ravishankar and Y. Bresler. Closed-form solutions within sparsifying transform learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5378–5382, 2013.

[26] S. Ravishankar and Y. Bresler. Learning doubly sparse transforms for images. *IEEE Trans. Image Process.*, 22(12):4598–4612, 2013.

[27] S. Ravishankar and Y. Bresler. Learning sparsifying transforms. *IEEE Trans. Signal Process.*, 61(5):1072–1086, 2013.

[28] S. Ravishankar and Y. Bresler. Efficient blind compressed sensing using sparsifying transforms with convergence guarantees and application to magnetic resonance imaging. *SIAM Journal on Imaging Sciences*, 8(4):2519–2557, 2015.

[29] S. Ravishankar and Y. Bresler. ℓ_0 sparsifying transform learning with efficient optimal updates and convergence guarantees. *IEEE Trans. Signal Process.*, 63(9):2389–2404, May 2015.

[30] S. Ravishankar and Y. Bresler. Data-driven learning of a union of sparsifying transforms model for blind compressed sensing. *IEEE Transactions on Computational Imaging*, 2(3):294–309, 2016.

[31] S. Ravishankar, B. E. Moore, R. R. Nadakuditi, and J. A. Fessler. Low-rank and adaptive sparse signal (LASSI) models for highly accelerated dynamic imaging. *IEEE Transactions on Medical Imaging*, 36(5):1116–1128, 2017.

[32] S. Ravishankar, R. R. Nadakuditi, and J. A. Fessler. Efficient sum of outer products dictionary learning (soup-dil) and its application to inverse problems. *IEEE Transactions on Computational Imaging*, 3(4):694–709, Dec 2017.

[33] R. T. Rockafellar and Roger J.-B. Wets. *Variational Analysis*. Springer-Verlag, Heidelberg, Germany, 1998.

[34] R. Rubinstein, M. Zibulevsky, and M. Elad. Double sparsity: Learning sparse dictionaries for sparse signal approximation. *IEEE Transactions on Signal Processing*, 58(3):1553–1564, 2010.

[35] M. Sadeghi, M. Babaie-Zadeh, and C. Jutten. Dictionary learning for sparse representation: A novel approach. *IEEE Signal Processing Letters*, 20(12):1195–1198, Dec 2013.

[36] A.-K. Seghouane and M. Hanif. A sequential dictionary learning algorithm with enforced sparsity. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3876–3880, 2015.

[37] K. Skretting and K. Engan. Recursive least squares dictionary learning algorithm. *IEEE Transactions on Signal Processing*, 58(4):2121–2130, 2010.

[38] L. N. Smith and M. Elad. Improving dictionary learning: Multiple dictionary updates and coefficient reuse. *IEEE Signal Processing Letters*, 20(1):79–82, Jan 2013.

[39] D. A. Spielman, H. Wang, and J. Wright. Exact recovery of sparsely-used dictionaries. In *Proceedings of the 25th Annual Conference on Learning Theory*, pages 37.1–37.18, 2012.

[40] J. Sun, Q. Qu, and J. Wright. Complete dictionary recovery over the sphere i: Overview and the geometric picture. *IEEE Transactions on Information Theory*, 63(2):853–884, Feb 2017.

[41] J. Sun, Q. Qu, and J. Wright. Complete dictionary recovery over the sphere ii: Recovery by riemannian trust-region method. *IEEE Transactions on Information Theory*, 63(2):885–914, Feb 2017.

[42] B. Wen, S. Ravishankar, and Y. Bresler. Structured overcomplete sparsifying transform learning with convergence guarantees and applications. *International Journal of Computer Vision*, 114(2-3):137–167, 2015.

[43] B. Wen, S. Ravishankar, and Y. Bresler. Video denoising by online 3d sparsifying transform learning. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 118–122, Sept 2015.

[44] Y. Xu and W. Yin. A fast patch-dictionary method for whole image recovery. *Inverse Problems and Imaging*, 10(2):563–583, 2016.

[45] M. Yaghoobi, T. Blumensath, and M. Davies. Dictionary learning for sparse approximations with the majorization method. *IEEE Transaction on Signal Processing*, 57(6):2178–2191, 2009.

[46] X. Zheng, S. Ravishankar, Y. Long, and J. A. Fessier. Union of learned sparsifying transforms based low-dose 3D CT image reconstruction. In *International Conference on Fully 3D Image Reconstruction in Radiology and Nuclear Medicine*, pages 69–72, 2017.