Online Sparsifying Transform Learning - Part II: Convergence Analysis

Saiprasad Ravishankar, *Student Member, IEEE,* and Yoram Bresler, *Fellow, IEEE*

***Abstract*—Sparsity based techniques have been widely popular in signal processing applications such as compression, denoising, and compressed sensing. Recently, the learning of sparsifying transforms for data has received interest. The advantage of the transform model is that it enables cheap and exact computations. In Part I of this work, efficient methods for online learning of square sparsifying transforms were introduced and investigated (by numerical experiments). The online schemes process signals sequentially, and can be especially useful when dealing with big data, and for real-time, or limited latency signal processing applications. In this paper, we prove that although the associated optimization problems are non-convex, the online transform learning algorithms are guaranteed to converge to the set of stationary points of the learning problem. The guarantee relies on a few simple assumptions. In practice, the algorithms work well, as demonstrated by examples of applications to representing and denoising signals.**

***Index Terms*—Sparse representations, Sparsifying transforms, Convergence guarantees, Online learning, Big data, Dictionary learning, Machine learning.**

I. Introduction

This paper, the theoretical counterpart to the work in Part I [1] on data-driven online learning of sparsifying transforms, provides a convergence analysis of the algorithms proposed in [1]. We start with a brief review of the background and motivation for the work. More detailed discussions and the relevant references can be found in Part I [1].

A. Background and Contributions

Techniques exploiting the sparsity of natural signals and images in a transform domain or dictionary have been widely popular in various applications. Various sparse models have been studied such as the synthesis, analysis, and transform models. The data-driven learning of such models has been shown to be useful in various applications such as denoising, and compressed sensing. In this work, we focus our attention on the classical transform model, which suggests that a signal $y \in \mathbb{R}^n$ is approximately sparsifiable using a transform $W \in \mathbb{R}^{m \times n}$, that is $Wy = x + e$, where $x \in \mathbb{R}^m$ is sparse in some sense, and e is a small residual in the transform domain. The learning of transform models has been shown to be much cheaper than synthesis, or analysis dictionary learning [2], [3]. Adaptive transforms also provide competitive, or useful signal reconstruction quality in applications (cf. [2], [3], and [1] and the references therein).

Prior work on transform learning focused on batch learning [2], [4], where the sparsifying transform is learnt using all the training data simultaneously. In Part I [1] of this work as well as here, the focus is instead on the online learning of sparsifying transforms. Various formulations and algorithms for online sparsifying transform learning have been proposed in Part I [1]. In this paper, we focus exclusively on the convergence properties of the algorithms in [1]. We prove that although the associated optimization problems are non-convex, the online transform learning algorithms are guaranteed to converge to the set of stationary points of the learning problem. The guarantee relies on a few simple assumptions. In practice, the algorithms work well, as demonstrated by sample applications to representing and denoising signals [1].

While the online learning of synthesis dictionaries has been studied previously, the online adaptation of the transform model allows for much cheaper computations [1]. Furthermore, the proof by Mairal et al. [5] of the convergence of online synthesis dictionary learning requires various restrictive assumptions (see Section II for details). In contrast, our analysis relies on simpler assumptions. Another feature distinguishing the online transform learning formulation is that in the previous work [5], the objective is biconvex, so that the non-convexity in the problem vanishes when a particular variable is kept fixed. This is not the case in the formulation here, in which the non-convexity is due to the ℓ_0 "norm" and the log determinant terms. The transform learning formulation remains non-convex even when one of the variables is fixed.

We now briefly review the problem formulations and algorithms for online sparsifying transform learning. The review serves to aid the understanding of our convergence results. The complete details of the formulations and algorithms can be found in Part I of this work [1].

B. Problem Formulations

The goal of online transform learning is to adapt the sparsifying transform $W \in \mathbb{R}^{n \times n}$ and sparse codes $\{x_t\}$ to data $\{y_t\}$ that arrive, or are processed sequentially in time. For time $t = 1, 2, 3, ...$, the optimization problem is as follows

$$\text{(P1)} \ \left\{\hat{W}_t, \hat{x}_t\right\} = \underset{W,\, x_t}{\arg\min} \frac{1}{t} \sum_{j=1}^{t} \left\{ \|Wy_j - x_j\|_2^2 + \lambda_j v(W) \right\} \quad s.t. \ \ \|x_t\|_0 \leq s, \ x_j = \hat{x}_j, \ 1 \leq j \leq t-1$$

This work was supported in part by the National Science Foundation (NSF) under grants CCF-1018660 and CCF-1320953.

S. Ravishankar and Y. Bresler are with the Department of Electrical and Computer Engineering and the Coordinated Science Laboratory, University of Illinois, Urbana-Champaign, IL, 61801 USA e-mail: (ravisha3, ybresler)@illinois.edu.

where $v(W) = -\log|\det W| + \|W\|_F^2$ is a regularizer, and the weight $\lambda_j = \lambda_0 \|y_j\|_2^2 \ \forall\, j$. Matrix $\hat{W}_t$ is the optimal transform at time t, and $\hat{x}_t$ is the optimal sparse code for y_t using $\hat{W}_t$. The sparsity is measured using the ℓ_0 "norm", which counts the number of non-zeros in a vector. Note that only the latest sparse code is updated at time t in Problem (P1). The condition $x_j = \hat{x}_j,\ 1 \leq j \leq t-1$, is therefore assumed. For brevity, we will not explicitly restate this condition (or, its appropriate variant) in the formulations in the rest of this paper. On the other hand, at each time t, the transform $\hat{W}_t$ is optimized using all the data y_j and sparse codes x_j up to time t.

The first term in the objective of (P1) is the sparsification error, which is the modeling error in the transform model. The regularizer $v(W)$ controls the condition number and scaling of W [1], [2]. As $\lambda_0 \to \infty$, the condition number of the optimal transform in (P1) tends to 1, and the spectral norm (or, scaling) tends to $1/\sqrt{2}$ [2]. The objective in (P1) is lower bounded by $\frac{n\lambda}{2} + \frac{n\lambda}{2}\log(2)$, which is positive [2].

A very useful variation of (P1) is *mini-batch* transform learning, where we process more than one signal at a time. Here, assuming a fixed block size of M, the J^{th} ($J \geq 1$) block of signals is $Y_J = \left[y_{JM-M+1} \mid y_{JM-M+2} \mid \ \ldots \ \mid y_{JM}\right]$. For $J = 1, 2, 3, ...$, the mini-batch sparsifying transform learning problem is

$$\left\{\hat{W}_J, \hat{X}_J\right\} = \underset{W,\, X_J}{\arg\min} \frac{1}{JM} \sum_{j=1}^{J} \left\{\|WY_j - X_j\|_F^2 + \Lambda_j v(W)\right\}$$
$$s.t. \ \ \|x_{JM-M+i}\|_0 \leq s \ \ \forall\, i \in \{1, .., M\} \quad \text{(P2)}$$

where $\Lambda_j = \lambda_0 \|Y_j\|_F^2$, and the sparse code matrix $X_J = \left[x_{JM-M+1} \mid x_{JM-M+2} \mid \ \ldots \ \mid x_{JM}\right]$ contains the block of sparse codes corresponding to Y_J.

In [1], variations of Problem (P1) involving a forgetting factor for dynamically changing data, or cycling (i.e., multiple passes through a data set) for fixed data sets, have been proposed but we do not consider these variations in this paper.

C. Algorithms

We now briefly discuss the algorithms [1] for (P1) and (P2). These algorithms alternate (only once) between a *sparse coding step* (where W is fixed and the sparse code(s) are updated), and a *transform update step* (where W is updated with fixed sparse code(s)) for each t or J.

The sparse coding step is similar for both (P1) and (P2). We discuss it for (P2) here. The result for (P1) follows by setting $M = 1$, replacing the index J with t, and replacing the capital letters Y and X (for matrices) with y and x (for vectors), respectively below.

The sparse coding step solves for X_J in (P2), with a fixed W ($= \hat{W}_{J-1}$, i.e., warm start) as follows

$$\min_{X_J} \|WY_J - X_J\|_F^2 \ s.t. \ \|x_{JM-M+i}\|_0 \leq s, \forall\, 1 \leq i \leq M$$

The optimal solution to the above problem is obtained as $\hat{x}_{JM-M+i} = H_s(Wy_{JM-M+i}) \ \forall\ i \in \{1, .., M\}$, where the operator $H_s(\cdot)$ zeros out all but the s coefficients of largest magnitude in a vector. If there is more than one choice for the s coefficients of largest magnitude in a vector z (which can occur when z has multiple entries of identical magnitude), then we choose $H_s(z)$ as the solution for which the s largest magnitude elements (in z) have lowest possible indices.

In the transform update step, (P2) is solved with respect to W, with fixed sparse codes $X_j = \hat{X}_j$, $1 \leq j \leq J$. Hence, the transform update step solves

$$\min_{W} \ \frac{1}{JM} \sum_{j=1}^{J} \left\{\|WY_j - X_j\|_F^2 + \Lambda_j\, Q(W)\right\} \qquad (1)$$

Problem (1) has an exact analytical solution presented in Part I of this work [1]. This solution is

$$\hat{W}_J = 0.5 R_J \left(\Sigma_J + \left(\Sigma_J^2 + 2\beta_J I\right)^{\frac{1}{2}}\right) Q_J^T L_J^{-1} \qquad (2)$$

where $(\cdot)^{\frac{1}{2}}$ denotes the positive definite square root of a symmetric matrix, I is the identity matrix, $\beta_J = (\lambda_0/JM) \sum_{j=1}^{J} \|Y_j\|_F^2$, and $Q_J \Sigma_J R_J^T$ denotes a full singular value decomposition (SVD) of $L_J^{-1}\Theta_J$, with $\Theta_J = (1/JM) \sum_{j=1}^{J} Y_j \hat{X}_j^T$ and L_J the positive definite square root of $(1/JM) \sum_{j=1}^{J} (Y_j Y_j^T + \lambda_0 \|Y_j\|_F^2\ I)$.

The $\hat{W}_J$ in (2) is an exact solution to the non-convex problem (1). The update (2) can be performed efficiently over time (see the algorithm in Fig. 2 of [1], where various matrices and scalars are updated sequentially, without requiring to store all the Y_j's and $\hat{X}_j$'s). Similar to the sparse coding step, the transform update problem and (exact) solution for Problem (P1) are obtained by setting $M = 1$, replacing J with t, and replacing Y and X with y and x, respectively, within the above mini-batch equations (1) and (2). In Part I of this work [1], an approximate transform update method for Problem (P1) was also presented (see the algorithm in Fig. 1 of [1]), which provides speedups over the exact one above, and works equally well in practice. For Problem (P2), a similar approximate transform update method is used for small $M \ll n$, whereas the exact update strategy above (i.e., by (2)) is more efficient at larger M [1]. For our convergence analysis, we will work with the exact transform update methods.

The rest of this paper is devoted to the convergence analysis of the algorithms. We will mostly focus on the convergence behavior for (P1), and briefly mention corresponding results for the (similar) algorithm for the block-based (P2). The organization of the rest of the paper is as follows. In Section II, we first present some notations and assumptions for our convergence analysis. Section III presents the main convergence results. The proof of convergence is detailed in Section IV. Finally, in Section V, we conclude.

II. Notations and Assumptions

A. Notations

The objective in Problem (P1) at time t is denoted as

$$\tilde{g}_t(W, x_t) \triangleq \frac{1}{t} \sum_{j=1}^{t} \left\{\|Wy_j - x_j\|_2^2 + \lambda_j v(W)\right\} \qquad (3)$$

where $\{x_j = \hat{x}_j\}_{j<t}$ have been computed at previous t values. The algorithm for (P1) [1] finds the sparse code as $\hat{x}_t = H_s(Wy_t)$, with $W = \hat{W}_{t-1}$. This is followed by a transform

update step. Let us denote the objective of the transform update step as

$$\hat{g}_t(W) \triangleq \tilde{g}_t(W, \hat{x}_t) \tag{4}$$

For a signal y, transform W, and vector x, we define

$$\tilde{u}(y, W, x) \triangleq \|Wy - x\|_2^2 + \lambda_0 \|y\|_2^2 v(W) \tag{5}$$

Then, we define the signal-wise loss function $u(y, W)$ as

$$\begin{aligned} u(y, W) &\triangleq \min_{x:\|x\|_0 \leq s} \tilde{u}(y, W, x) \\ &= \|Wy - H_s(Wy)\|_2^2 + \lambda_0 \|y\|_2^2 v(W) \end{aligned} \tag{6}$$

Thus, $u(y, W)$ is small for signals (assuming signals of similar scaling) that are sparsified well by W. We use the operation $\tilde{H}_s(b)$ to denote the *set* of all optimal projections of $b \in \mathbb{R}^n$ onto the s-ℓ_0 ball defined as $\{x \in \mathbb{R}^n : \|x\|_0 \leq s\}$. When $\tilde{H}_s(b)$ is a unique element, it satisfies $\tilde{H}_s(b) = H_s(b)$, for $H_s(\cdot)$ defined as in Section I-C.

We also define the *empirical objective function*

$$g_t(W) \triangleq \frac{1}{t} \sum_{j=1}^{t} u(y_j, W) \tag{7}$$

The empirical objective function involves the optimal sparse code (in W) for each y_j, and it is the objective that is minimized by batch transform learning algorithms [2], [4]. Note however that in an online setting, the sparse codes of past signals cannot be optimally set at future times t.

For convenience, we split the various functions ($\tilde{g}_t(W, x_t)$, $\hat{g}_t(W)$, $u(y, W)$, $g_t(W)$) that have to do with the objective into the sum of two terms: the first, with a superscript of (1) will be used to denote the sparsification error term; the second will denote the regularizer term. For example, $u(y, W) = u^{(1)}(y, W) + u^{(2)}(y, W)$, where $u^{(1)}(y, W) = \|Wy - H_s(Wy)\|_2^2$ and $u^{(2)}(y, W) = \lambda_0 \|y\|_2^2 v(W)$. Finally, we use the abbreviations wp.1 and a.s for "with probability 1" and "almost surely", respectively, and use the notations $\overset{wp.1}{=}$ and $\overset{a.s}{=}$ to denote equality wp.1 or a.s.

B. Assumptions

In order to derive the convergence results, we will make the following few assumptions.

(A1) Signal Normalization. First, we assume that the input signals y_t are normalized, i.e., $\|y_t\|_2 = 1$. (Any input that is 0 can always be dropped, or processed trivially.) This assumption eliminates the dependence on y of the regularizer weighting, for the various functions in Section II-A.

(A2) Exact Computation. The transform update step of the algorithm(s) is assumed to be performed exactly (referred to as "exact", since there is a simple closed-form solution involving the SVD [1]). This is always the case for the mini-batch algorithm in the large M ($M \sim O(n)$ or larger) case [1]. For the algorithm for (P1), the exact transform update method in Section I-C is slower than the approximate one [2] in [1], and has an $O(n^3)$ rather than $O(n^2 \log^2 n)$ computational cost per signal, but will be still assumed to be the one used, for the purpose of theoretical analysis.

(A3) Nondegenerate SVD. We assume (for each $t > n$) that $L_t^{-1}\Theta_t = t^{-1} \sum_{j=1}^{t} L_t^{-1} y_j \hat{x}_j^T$ has non-degenerate (distinct, non-zero) singular values, i.e., there is a minimum separation $\hat{\eta} > 0$ between any two singular values as well as between the smallest singular value and zero. We observed this assumption to hold in numerical experiments. One could also simply monitor the singular values of $L_t^{-1}\Theta_t$ (over t), and drop (i.e., ignore from the formulation/algorithm) the signals y_t for those time instances (t), when the assumption is violated. Such signals could be treated as "outliers" and processed separately [3]. Assumption A3 is not required for showing the convergence of the objective function $\hat{g}_t$ in the algorithms.

(A4) Random Signals. The signals y_t are assumed to be independently and identically distributed over the unit sphere $\{y \in \mathbb{R}^n : \|y\|_2 = 1\}$, according to an absolutely continuous probability density function $p(y)$.

Our assumptions are less restrictive (and also easier to verify) than the ones in [5]. There, the authors assume the uniqueness of the synthesis sparse coding solution. However, such a uniqueness assumption may not hold in general. Moreover, the proof in [5] assumes that the synthesis sparse coding problem is solved exactly at each t (a similar assumption is also made for the dictionary update step in [5]), which is typically impractical [4] in general. Another assumption in [5] is the positive definiteness of the Hessian of the dictionary learning objective (this is similar to our assumption on $L_t^{-1}\Theta_t$). We would also like to emphasize that as opposed to the prior work [5], we work with an optimization problem that is not simply biconvex. Specifically, our problem involves the ℓ_0 "norm" for sparsity, and a non-convex log-determinant penalty.

C. Expected Transform Learning Cost

Given the statistical assumptions about the signals, we follow the standard approach in the analysis of online algorithms (cf. [5]–[9]) and consider the minimization of the expected cost

$$g(W) \triangleq \mathbb{E}_y \left[u(y, W) \right] \tag{8}$$

where the expectation is with respect to the (unknown) probability distribution $p(y)$ of the data. It follows from Assumption A4 that for any fixed W $\lim_{t\to\infty} g_t(W) = g(W)$ a.s (almost sure convergence). In particular, given a specific training set, it may be unnecessary to minimize the empirical objective function $g_t(W)$ to high precision, since it is only an approximation to the expected cost. In fact, even an inaccurate minimizer of $g_t(W)$ could (potentially) provide the same, or better value of the expected cost than a fully optimized one. Although we cannot directly minimize the expected cost, we will show

[1] Although in practice the SVD is computed using iterative methods, the methods are guaranteed to quickly provide machine precision accuracy.

[2] The approximate transform update method performs equally well (as the exact one) in practice [1]. The convergence results in this work may therefore be also relevant to the approximate method.

[3] Assuming that such outliers occur infrequently, they could for example be processed (sparse coded) using a fixed analytical sparsifying transform such as Wavelets.

[4] In general, the iterative optimization algorithms [5] (for either synthesis sparse coding, or dictionary update) may take a large number of iterations to reach machine precision accuracy. Moreover, these algorithms do not provide a method to determine the accuracy of the computed solution at any given iteration.

interesting asymptotic properties for the algorithms in [1] with respect to the expected cost.

III. Main Results

The main convergence results in this work are briefly stated as follows. We assume some particular (non-singular) initialization $\hat{W}_0$ for the algorithms [1]. For simplicity, we state results for the online algorithm for (P1). Similar results can be easily shown to hold for the mini-batch scheme (for (P2)). For the sequence $\left\{\hat{W}_t\right\}$ generated by the online scheme, we have

(i) As $t \to \infty$, $\hat{g}_t(\hat{W}_t)$, $g_t(\hat{W}_t)$, and $g(\hat{W}_t)$ all converge a.s to a common limit, say g^*.
(ii) The sequence $\left\{\hat{W}_t\right\}$ is bounded. Every accumulation point $\hat{W}_\infty$ of $\left\{\hat{W}_t\right\}$ is a stationary point of the expected cost $g(W)$ satisfying $\nabla g(\hat{W}_\infty) = 0$, wp.1.
(iii) Every accumulation point $\hat{W}_\infty$ of $\left\{\hat{W}_t\right\}$ achieves the same value (i.e., g^*) of the expected cost g wp.1.
(iv) The distance between $\hat{W}_t$ and the set of stationary points of the expected cost $g(W)$ converges to 0 almost surely as $t \to \infty$.

Statement (i) above shows convergence of the objective function sequences. It is interesting that $\hat{g}_t(\hat{W}_t)$ and $g_t(\hat{W}_t)$ converge almost surely to the same limit. Since the definition of $g_t(\hat{W}_t)$ involves the optimal sparse codes computed using the common $\hat{W}_t$ for all signals y_j $1 \leq j \leq t$, whereas $\hat{g}_t$ uses the sequentially computed $\hat{W}_{j-1}$ for (sparse coding) signal y_j $(1 \leq j \leq t)$, the convergence result (i) means that we do not lose (asymptotically) by sparse coding the signals only sequentially.

Statement (ii) above says that every accumulation point $\hat{W}_\infty$ of the iterate sequence is a stationary point of the expected cost $g(W)$ with probability 1. Furthermore, Statement (iii) shows that every accumulation point $\hat{W}_\infty$ of $\left\{\hat{W}_t\right\}$ satisfies $g(\hat{W}_\infty) = g^*$ wp.1. In other words, every accumulation point is equally good in terms of its expected cost (i.e., $g(\cdot)$) value.

Statement (iv) indicates that the iterate sequence $\left\{\hat{W}_t\right\}$ converges to the set of stationary points of $g(W)$ wp.1.

Finally, we also show the following other interesting results, which hold for every realization of the $\{y_t\}$ such that Assumptions A1-A3 hold.

(v) $\hat{g}_{t+1}(\hat{W}_{t+1}) - \hat{g}_t(\hat{W}_t)$ decays as $O(1/t)$.
(vi) $\hat{g}_{t+1}(\hat{W}_{t+1}) - \hat{g}_{t+1}(\hat{W}_t)$ decays as $O(1/t^2)$.
(vii) $\hat{W}_{t+1} - \hat{W}_t$ decays (in norm) as $O(1/t)$.

The above results (v)-(vii) show the convergence rate of the difference between successive iterates or objective values. Statement (vi) indicates that the objective decreases at a $O(1/t^2)$ rate within the transform update step of the algorithm. However, when the sparse coding step is included in the calculation (i.e., statement (v) above), the rate of decrease is only $O(1/t)$, due to the uncertainty introduced by a newly added signal. Note that Statements (v)-(vii) do not by themselves indicate convergence of the objective, or iterate sequences, but will be used to prove such convergence.

IV. Proof of Convergence

We now prove the convergence properties of the online algorithm for (P1). The various results (leading up to our main convergence results) are proved here in the following order.

(i) The iterate sequence $\left\{\hat{x}_t, \hat{W}_t\right\}$ generated by the algorithm is bounded.
(ii) The objective sequence $\left\{\hat{g}_t(\hat{W}_t)\right\}$ is also bounded.
(iii) The objective and iterate sequences each have at least one convergent subsequence.
(iv) $\hat{g}_{t+1}(\hat{W}_{t+1}) - \hat{g}_t(\hat{W}_t)$ decays as $O(1/t)$.
(v) $\hat{W}_{t+1} - \hat{W}_t$ also decays (in norm) as $O(1/t)$.
(vi) $\hat{g}_{t+1}(\hat{W}_{t+1}) - \hat{g}_{t+1}(\hat{W}_t)$ decays as $O(1/t^2)$.
(vii) As $t \to \infty$, $\hat{g}_t(\hat{W}_t)$, $g_t(\hat{W}_t)$, and $g(\hat{W}_t)$ converge a.s to a common limit.
(viii) Each accumulation point of $\left\{\hat{W}_t\right\}$ is a stationary point of the expected cost $g(W)$ wp.1. Moreover, every accumulation point achieves the same value (g^*) of the expected cost g wp.1.
(ix) The distance between $\hat{W}_t$ and the set of stationary points of the expected cost $g(W)$ converges to 0 almost surely as $t \to \infty$.

Result (i) above is given by the following lemma. To simplify the proofs, we assume that $\lambda_0 \geq 2$ in this section. This condition leads to a simple bound of unity for the norms of the iterates in our proofs.

Lemma 1: Under Assumptions A1 and A2, for any $\lambda_0 \geq 2$, the iterate sequence $\left\{\hat{x}_t, \hat{W}_t\right\}$ generated by the online algorithm is bounded $\forall\ t$ as $\left\|\hat{x}_t\right\|_2 \leq 1$, and $\left\|\hat{W}_t\right\|_2 \leq 1$. Furthermore, we have that

$$\left\|\Sigma_t\right\|_2 \leq \left\|L_t^{-1}\right\|_2 \leq \frac{1}{\sqrt{\lambda_0}} \ \forall\, t \tag{9}$$

Proof: Assuming without loss of generality that the initialization is scaled so that $\left\|\hat{W}_0\right\|_2 \leq 1$, we have for $t = 1$

$$\left\|\hat{x}_1\right\|_2 = \left\|H_s(\hat{W}_0 y_1)\right\|_2 \leq \left\|\hat{W}_0 y_1\right\|_2 \leq \left\|\hat{W}_0\right\|_2 \left\|y_1\right\|_2 \leq 1 \tag{10}$$

Furthermore, $\hat{W}_t = 0.5 R_t \left(\Sigma_t + \left(\Sigma_t^2 + 2\beta_t I\right)^{1/2}\right) Q_t^T L_t^{-1}$ for any t, where L_t is the positive definite square root of $t^{-1}\sum_{j=1}^t (y_j y_j^T + \lambda_j I)$, and $Q_t \Sigma_t R_t^T$ is the full SVD of $L_t^{-1}\Theta_t$ with $\Theta_t = (1/t)\sum_{j=1}^t y_j \hat{x}_j^T$, and $\beta_t = \sum_{j=1}^t (t)^{-1}\lambda_j$. Therefore,

$$\left\|\hat{W}_1\right\|_2 \leq 0.5 \left\|\Sigma_1 + \left(\Sigma_1^2 + 2\beta_1 I\right)^{\frac{1}{2}}\right\|_2 \left\|L_1^{-1}\right\|_2 \tag{11}$$

by the sub-multiplicativity of the matrix spectral norm. Since, $\left\|y_t\right\|_2 = 1\ \forall\ t$, we get $\beta_t = \lambda_0$ for all t. Moreover, for every t,

$$\left\|L_t^{-1}\right\|_2 = \left\|\left(\tfrac{1}{t}\sum_{j=1}^t y_j y_j^T + \lambda_0 I\right)^{-1/2}\right\|_2 \leq \frac{1}{\sqrt{\lambda_0}} \tag{12}$$

We also have

$$\left\|\Sigma_t\right\|_2 = \left\|L_t^{-1}\Theta_t\right\|_2 \leq t^{-1}\left\|L_t^{-1}\right\|_2 \sum_{j=1}^t \left\|y_j\right\|_2 \left\|\hat{x}_j\right\|_2 \tag{13}$$

It is then obvious using (10) that $\left\|\Sigma_1\right\|_2 \leq \left\|L_1^{-1}\right\|_2$. Substituting this into (11) and using (12), we easily get

$$\left\|\hat{W}_1\right\|_2 \leq 0.5\lambda_0^{-1} + 0.5\left(\lambda_0^{-2} + 2\right)^{1/2} \tag{14}$$

It follows that $\left\|\hat{W}_1\right\|_2 \leq 1$, whenever $\lambda_0 \geq 2$ holds. Then, upon repeating the aforementioned arguments for $t = 2, 3$, etc. (equivalently, by induction), we obtain that $\hat{x}_t$ and $\hat{W}_t$ satisfy $\left\|\hat{W}_t\right\|_2 \leq 1$ and $\left\|\hat{x}_t\right\|_2 \leq 1$ for each and every t. Then, (9) also holds for every t, just as shown above for the $t = 1$ case. ■

Next, we show that the objective sequence $\left\{\hat{g}_t(\hat{W}_t)\right\}$ is bounded.

Lemma 2: Under Assumptions A1 and A2, the objective sequence $\left\{\hat{g}_t(\hat{W}_t)\right\}$ is bounded.

Proof: The transform $\hat{W}_t$ is a minimizer of the transform update objective for fixed $x_t = \hat{x}_t = H_s(\hat{W}_{t-1}y_t)$ (i.e., $\hat{W}_t$ minimizes $\hat{g}_t(W)$). Therefore, we have

$$\hat{g}_t(\hat{W}_t) \leq \hat{g}_t(\hat{W}_0) = \lambda_0 v(\hat{W}_0) + \frac{1}{t}\sum_{j=1}^{t}\left\|\hat{W}_0 y_j - \hat{x}_j\right\|_2^2 \quad (15)$$

By Lemma 1, we know that the $\hat{x}_j$'s in (15) are all bounded. Therefore, $\hat{g}_t(\hat{W}_t)$ is also upper bounded (by a constant independent of t). We also have (see Section I-B) that $\hat{g}_t(\hat{W}_t) \geq 0$. Therefore, $\hat{g}_t(\hat{W}_t)$ is bounded for each t. ■

Proposition 1: The objective and iterate sequences in the online algorithm, each have at least one convergent subsequence.

Proof: Since the objective and the iterate sequences are bounded, the existence of a convergent subsequence (for a bounded sequence) is a standard result. ■

The next two propositions show the $O(1/t)$ decay of the difference between successive elements of the objective and iterate sequences.

Proposition 2: Let Assumptions A1 and A2 hold. Then, the objective sequence $\left\{\hat{g}_t(\hat{W}_t)\right\}$ satisfies

$$\left|\hat{g}_{t+1}(\hat{W}_{t+1}) - \hat{g}_t(\hat{W}_t)\right| \leq \frac{C}{t+1} \quad (16)$$

where $C > 0$ is a constant independent of t.

Proof: First, we have by the definition of $\hat{g}_{t+1}$ (4) that

$$\hat{g}_{t+1}(\hat{W}_{t+1}) = \frac{t\hat{g}_t(\hat{W}_{t+1})}{t+1} + \frac{\tilde{u}(y_{t+1}, \hat{W}_{t+1}, \hat{x}_{t+1})}{t+1} \quad (17)$$

Since $\hat{g}_t(\hat{W}_{t+1}) \geq \hat{g}_t(\hat{W}_t)$, and $\tilde{u}(y_{t+1}, \hat{W}_{t+1}, \hat{x}_{t+1}) \geq 0$, we get

$$\hat{g}_{t+1}(\hat{W}_{t+1}) - \hat{g}_t(\hat{W}_t) \geq -\frac{\hat{g}_t(\hat{W}_t)}{t+1} \geq -\frac{C}{t+1} \quad (18)$$

where the last inequality above follows from Lemma 2.

Second, since $\hat{g}_{t+1}(\hat{W}_{t+1}) \leq \hat{g}_{t+1}(\hat{W}_t)$, we also have that $\hat{g}_{t+1}(\hat{W}_{t+1}) - \hat{g}_t(\hat{W}_t) \leq \hat{g}_{t+1}(\hat{W}_t) - \hat{g}_t(\hat{W}_t)$. Combining this with (17) (with $\hat{W}_{t+1}$ replaced by $\hat{W}_t$ in (17)), we get

$$\hat{g}_{t+1}(\hat{W}_{t+1}) - \hat{g}_t(\hat{W}_t) \leq -\frac{\hat{g}_t(\hat{W}_t)}{t+1} + \frac{\tilde{u}(y_{t+1}, \hat{W}_t, \hat{x}_{t+1})}{t+1} \quad (19)$$

Since $-\hat{g}_t(\hat{W}_t) + \tilde{u}(y_{t+1}, \hat{W}_t, \hat{x}_{t+1}) \leq \left\|\hat{W}_t y_{t+1} - \hat{x}_{t+1}\right\|_2^2$ (by algebraic manipulations), we have

$$\hat{g}_{t+1}(\hat{W}_{t+1}) - \hat{g}_t(\hat{W}_t) \leq \frac{\left\|\hat{W}_t y_{t+1} - \hat{x}_{t+1}\right\|_2^2}{t+1} \leq \frac{C}{t+1} \quad (20)$$

where the last inequality follows by Lemma 1 and Assumption A1. Combining (20) and (18), we have the desired result. ■

Proposition 3: Let Assumptions A1-A3 hold. Then, the sequence $\left\{\hat{W}_t\right\}$ satisfies

$$\left\|\hat{W}_{t+1} - \hat{W}_t\right\|_F \leq \frac{C}{t} \quad \forall\, t \quad (21)$$

where C is a constant independent of t.

Proof: Let $A_t \triangleq 0.5R_t\left(\Sigma_t + \left(\Sigma_t^2 + 2\beta_t I\right)^{1/2}\right)Q_t^T$, where $Q_t\Sigma_t R_t^T$ is the full SVD of $L_t^{-1}\Theta_t$ (cf. Section I-C), and $\beta_t = \lambda_0$. Then, we have $\hat{W}_{t+1} - \hat{W}_t = (A_{t+1} - A_t)\,L_t^{-1} + A_{t+1}\left(L_{t+1}^{-1} - L_t^{-1}\right)$. Therefore,

$$\begin{aligned}\left\|\hat{W}_{t+1} - \hat{W}_t\right\|_F \leq{}& \left\|L_t^{-1}\right\|_2\left\|A_{t+1} - A_t\right\|_F \\ &+ \left\|A_{t+1}\right\|_2\left\|L_{t+1}^{-1} - L_t^{-1}\right\|_F \end{aligned} \quad (22)$$

By Lemma 1, $\left\|L_t^{-1}\right\|_2 \leq 1/\sqrt{\lambda_0}$ and $\left\|A_{t+1}\right\|_2 \leq (0.5/\sqrt{\lambda_0}) + 0.5\left(\lambda_0^{-1} + 2\lambda_0\right)^{1/2}$.

We now bound $\left\|L_{t+1}^{-1} - L_t^{-1}\right\|_F$ in (22). Defining $K_t \triangleq t^{-1}\sum_{j=1}^{t} y_j y_j^T + \lambda_0 I$, we have that $L_t^{-1} = K_t^{-1/2}$. First, it is easy to show that

$$\left\|K_{t+1} - K_t\right\|_F \leq \frac{2}{t+1} \quad (23)$$

i.e., $K_{t+1} - K_t = O(1/t)$. Second, using Taylor series expansions of the matrix inverse [5] and square root [6], it can be shown that for large t, $K_{t+1}^{-1/2} = K_t^{-1/2} + E_t$, with $E_t = O(1/t)$. Since the result holds for all sufficiently large (determined by λ_0) t, we can always find a constant c_0 such that $\left\|L_{t+1}^{-1} - L_t^{-1}\right\|_F \leq c_0/t$, $\forall\, t$. Alternatively, we can drop a finite number of sequence elements, so that the result holds for all remaining t.

Next, we bound $\left\|A_{t+1} - A_t\right\|_F$ in (22). Since A_t is a function of the SVD of $L_t^{-1}\Theta_t$, we first need a bound on

$$\begin{aligned}\left\|L_{t+1}^{-1}\Theta_{t+1} - L_t^{-1}\Theta_t\right\|_F \leq{}& \left\|L_{t+1}^{-1}\right\|_2\left\|\Theta_{t+1} - \Theta_t\right\|_F \\ &+ \left\|\Theta_t\right\|_2\left\|L_{t+1}^{-1} - L_t^{-1}\right\|_F \end{aligned} \quad (24)$$

Since $\left\|\Theta_{t+1} - \Theta_t\right\|_F \leq 2/(t+1)$, and $\left\|\Theta_t\right\|_2 \leq 1$ (by Lemma 1 and Assumption A1), it is clear that $\left\|L_{t+1}^{-1}\Theta_{t+1} - L_t^{-1}\Theta_t\right\|_F \leq c_1/t$ (with $c_1 = c_0 + (2/\sqrt{\lambda_0})$). Now, we apply Theorem 1 from Appendix C here to conclude that the singular values of $L_{t+1}^{-1}\Theta_{t+1}$ differ from the corresponding singular values of $L_t^{-1}\Theta_t$ only by $O(1/t)$. Moreover, each left and right singular vector pair of $L_{t+1}^{-1}\Theta_{t+1}$ differs (under our Assumption A3 that $L_t^{-1}\Theta_t$ has non-degenerate singular values $\forall\ t > n$) from the corresponding pair of $L_t^{-1}\Theta_t$ only by $O(1/t)$ (a particular pair may be scaled by -1 for the result to hold). Using these perturbation bounds for the singular values and singular vectors (and a simple scalar Taylor series expansion for the diagonal elements of $\left(\Sigma_{t+1}^2 + 2\lambda_0 I\right)^{1/2}$), it is easy to show that $\left\|A_{t+1} - A_t\right\|_F \leq c_2/t$, with c_2 a constant independent of t.

Finally, substituting the bounds on $\left\|A_{t+1} - A_t\right\|_F$ and $\left\|L_{t+1}^{-1} - L_t^{-1}\right\|_F$ into (22), equation (21) follows. ■

[5] Assuming $A \in \mathbb{R}^{n\times n}$ is invertible, and B is small, $(A+B)^{-1} = A^{-1} - A^{-1}BA^{-1} + A^{-1}BA^{-1}BA^{-1} - A^{-1}BA^{-1}BA^{-1}BA^{-1} + \ldots$

[6] For small $B \in \mathbb{R}^{n\times n}$, $(I+B)^{1/2} = I + \frac{1}{2}B - \frac{1}{8}B^2 + \frac{1}{16}B^3 - \ldots$

As mentioned before, the fact that the difference between successive iterates, or objective values decays as $O(1/t)$, although interesting, does not by itself indicate convergence of the respective sequences. In order to prove such convergence, we also need the following Lemma. The lemma shows that the objective decreases as $O(1/t^2)$ within the transform update step at time t.

Lemma 3: Let Assumptions A1 and A2 hold. Then, there exists a constant $c > 0$ independent of t such that

$$\left|\hat{g}_{t+1}(\hat{W}_{t+1}) - \hat{g}_{t+1}(\hat{W}_t)\right| \leq \frac{c}{t^2} \tag{25}$$

Proof: First, due to the optimality condition (in transform update step) for $\hat{W}_{t+1}$, we have $\hat{g}_{t+1}(\hat{W}_{t+1}) - \hat{g}_{t+1}(\hat{W}_t) \leq 0$. Second, we have that

$$\begin{aligned}\hat{g}_{t+1}(\hat{W}_{t+1}) - \hat{g}_{t+1}(\hat{W}_t) &= \hat{g}_{t+1}(\hat{W}_{t+1}) - \hat{g}_t(\hat{W}_{t+1}) \\ &+ \hat{g}_t(\hat{W}_{t+1}) - \hat{g}_t(\hat{W}_t) + \hat{g}_t(\hat{W}_t) - \hat{g}_{t+1}(\hat{W}_t)\end{aligned} \tag{26}$$

Since $\hat{g}_t(\hat{W}_{t+1}) - \hat{g}_t(\hat{W}_t) \geq 0$, we get

$$\begin{aligned}\hat{g}_{t+1}(\hat{W}_{t+1}) - \hat{g}_{t+1}(\hat{W}_t) &\geq \hat{g}_{t+1}(\hat{W}_{t+1}) - \hat{g}_t(\hat{W}_{t+1}) \\ &+ \hat{g}_t(\hat{W}_t) - \hat{g}_{t+1}(\hat{W}_t)\end{aligned} \tag{27}$$

We now consider the function $\hat{g}_{t+1} - \hat{g}_t$. This function does not depend on the transform learning regularizer $v(W)$ (it cancels out). It is in fact a quadratic in W. Importantly, since the transforms in our case belong to a bounded set (by Lemma 1), it is easy to show that the function $\hat{g}_{t+1} - \hat{g}_t$ is Lipschitz with respect to W, with a Lipschitz constant $\leq \frac{C_1}{t+1}$. Using this fact in (27), we get that

$$0 \geq \hat{g}_{t+1}(\hat{W}_{t+1}) - \hat{g}_{t+1}(\hat{W}_t) \geq \frac{-C_1}{t+1}\left\|\hat{W}_{t+1} - \hat{W}_t\right\|_F$$

Combining the above result with the result (21) of Proposition 3, the required result (25) follows. ■

The almost sure convergence of the objective sequence is provided by the following proposition.

Proposition 4: Let Assumptions A1-A4 hold. Then, as $t \to \infty$, $\hat{g}_t(\hat{W}_t)$, $g_t(\hat{W}_t)$, and $g(\hat{W}_t)$ all converge almost surely to a common limit g^*.

Proof: Our proof methodology here is similar to that in [5], [8], but differs in the details, and required conditions/assumptions. Define $r_t = \hat{g}_t(\hat{W}_t)$. Then, $r_t \geq 0$. We will use Theorem 3 in Appendix C to show that r_t is a quasi-martingale and converges almost surely. To apply Theorem 3, we first need to investigate $r_{t+1} - r_t$, which is given as

$$\begin{aligned}r_{t+1} - r_t &= \hat{g}_{t+1}(\hat{W}_{t+1}) - \hat{g}_{t+1}(\hat{W}_t) + \hat{g}_{t+1}(\hat{W}_t) - \hat{g}_t(\hat{W}_t) \\ &\leq \frac{t\hat{g}_t(\hat{W}_t)}{t+1} + \frac{u(y_{t+1}, \hat{W}_t)}{t+1} - \hat{g}_t(\hat{W}_t) \qquad (28) \\ &= \frac{g_t(\hat{W}_t) - \hat{g}_t(\hat{W}_t)}{t+1} + \frac{u(y_{t+1}, \hat{W}_t) - g_t(\hat{W}_t)}{t+1} \qquad (29) \\ &\leq \frac{u(y_{t+1}, \hat{W}_t) - g_t(\hat{W}_t)}{t+1} \qquad (30)\end{aligned}$$

where $\hat{g}_{t+1}(\hat{W}_{t+1}) \leq \hat{g}_{t+1}(\hat{W}_t)$ is used to arrive at (28), and $g_t(\hat{W}_t) \leq \hat{g}_t(\hat{W}_t)$ is used to arrive at (30). The inequality $g_t(\hat{W}_t) \leq \hat{g}_t(\hat{W}_t)$ is true because the definition of $g_t(\hat{W}_t)$ in (7) involves computing the optimal sparse codes using the common $\hat{W}_t$ for all y_j, $1 \leq j \leq t$, whereas $\hat{g}_t(\hat{W}_t)$ uses sub-optimal values (computed sequentially in the algorithm) for those sparse codes.

Now, we consider the filtration $\mathcal{F}_t$ (cf. Theorem 3) determined by the past information up to time t as follows

$$\begin{aligned}\mathbb{E}[r_{t+1} - r_t|\mathcal{F}_t] &\leq \frac{\mathbb{E}[u(y_{t+1}, \hat{W}_t)|\mathcal{F}_t] - g_t(\hat{W}_t)}{t+1} \qquad (31) \\ = \frac{g(\hat{W}_t) - g_t(\hat{W}_t)}{t+1} &= \frac{g^{(1)}(\hat{W}_t) - g_t^{(1)}(\hat{W}_t)}{t+1} \leq \frac{\left\|g^{(1)} - g_t^{(1)}\right\|_\infty}{t+1}\end{aligned}$$

where $\left\|g^{(1)} - g_t^{(1)}\right\|_\infty = \sup_{W \in S}\left|g^{(1)}(W) - g_t^{(1)}(W)\right|$, with the set S as defined in Lemma 5 of Appendix B, and $g^{(1)}(W) = \mathbb{E}_y\left[u^{(1)}(y, W)\right]$.

In order to satisfy the requirements of Theorem 3 (Appendix C), we will first bound $\mathbb{E}\left[\sqrt{t}\left\|g^{(1)} - g_t^{(1)}\right\|_\infty\right]$, for which we use Theorem 2 in Appendix C. Note that by Lemma 5 in Appendix B $u^{(1)}(y, W)$ is Lipschitz with respect to W on the bounded set S. Moreover, $u^{(1)}(y, W)$ is bounded and $\mathbb{E}_y\left[\left\{u^{(1)}(y, W)\right\}^2\right]$ is also (uniformly) bounded. Therefore, directly applying Theorem 2, we conclude that $\mathbb{E}\left[\sqrt{t}\left\|g^{(1)} - g_t^{(1)}\right\|_\infty\right] = O(1)$. Thus, defining δ_t as in Theorem 3 (Appendix C), we get

$$\mathbb{E}\left[\delta_t(r_{t+1} - r_t)\right] = \mathbb{E}\left[\mathbb{E}\left[r_{t+1} - r_t|\mathcal{F}_t\right]^+\right] \leq \frac{c}{t^{\frac{3}{2}}} \tag{32}$$

where c is a constant, and the $(\cdot)^+$ operation zeros out negative numbers. Equation (32) immediately implies that the requirement $\sum_{t=1}^{\infty}\mathbb{E}[\delta_t(r_{t+1} - r_t)] < \infty$ is met for Theorem 3 (Appendix C). Therefore, as $t \to \infty$, r_t converges a.s.

The rest of the results are simple to derive. We briefly mention the steps here for completeness. We will now prove that $g(\hat{W}_t)$ converges almost surely. First, we have from Theorem 3 (Appendix C) that

$$\sum_{t=1}^{\infty}\left|\mathbb{E}[r_{t+1} - r_t|\mathcal{F}_t]\right| < \infty \;\; a.s \tag{33}$$

We now use the fact that

$$r_{t+1} - r_t = \hat{g}_{t+1}(\hat{W}_{t+1}) - \hat{g}_{t+1}(\hat{W}_t) + \frac{u(y_{t+1}, \hat{W}_t) - \hat{g}_t(\hat{W}_t)}{t+1}$$

to get the following result

$$\begin{aligned}\sum_{t=1}^{\infty}\frac{\left|g(\hat{W}_t) - \hat{g}_t(\hat{W}_t)\right|}{t+1} &\leq \sum_{t=1}^{\infty}\left|\mathbb{E}[r_{t+1} - r_t|\mathcal{F}_t]\right| \\ &+ \sum_{t=1}^{\infty}\left|\mathbb{E}\left[\hat{g}_{t+1}(\hat{W}_{t+1}) - \hat{g}_{t+1}(\hat{W}_t)|\mathcal{F}_t\right]\right| \qquad (34)\end{aligned}$$

We know from Lemma 3 that

$$\sum_{t=1}^{\infty}\mathbb{E}\left[\left|\hat{g}_{t+1}(\hat{W}_{t+1}) - \hat{g}_{t+1}(\hat{W}_t)\right||\mathcal{F}_t\right] \leq \sum_{t=1}^{\infty}\frac{c}{t^2} < \infty$$

Using this result and (33) in (34), we immediately get the almost sure convergence of the following sum

$$\sum_{t=1}^{\infty}\frac{\left|g(\hat{W}_t) - \hat{g}_t(\hat{W}_t)\right|}{t+1} = \sum_{t=1}^{\infty}\frac{\left|g^{(1)}(\hat{W}_t) - \hat{g}_t^{(1)}(\hat{W}_t)\right|}{t+1} \tag{35}$$

We now use (35), and apply Theorem 4 in Appendix C to show that $g(\hat{W}_t) - \hat{g}_t(\hat{W}_t) \rightarrow 0$ a.s. Let $b_t = \frac{1}{t+1}$, $a_t = g(\hat{W}_t) - \hat{g}_t(\hat{W}_t) = g^{(1)}(\hat{W}_t) - \hat{g}_t^{(1)}(\hat{W}_t)$, and $d_t = |a_t|$ (using similar notations as in Theorem 4 of Appendix C). Then, it is easy to show that

$$\left|d_{t+1} - d_t\right| \leq \left|a_{t+1} - a_t\right| \leq \left|\hat{g}_t^{(1)}(\hat{W}_{t+1}) - \hat{g}_t^{(1)}(\hat{W}_t)\right| + \left|g^{(1)}(\hat{W}_{t+1}) - g^{(1)}(\hat{W}_t)\right| + \frac{\nu}{t+1} \quad (36)$$

where ν is a constant that upper bounds $\left|\hat{g}_t^{(1)}(\hat{W}_{t+1})\right| + \left|\tilde{u}^{(1)}(y_{t+1}, \hat{W}_{t+1}, \hat{x}_{t+1})\right|$. It is also easy to see that both $\hat{g}_t^{(1)}(W)$ (a quadratic) and $g^{(1)}(W)$ (follows using Lemma 5 here) are Lipschitz over W in a compact set. Combining this with the result of Proposition 3, we can easily see that (36) implies $\left|d_{t+1} - d_t\right| < \hat{c}/(t+1)$ for some $\hat{c} > 0$ that does not depend on t. Now all the conditions for Theorem 4 in Appendix C are satisfied (i.e., $b_t \geq 0$, $d_t \geq 0$, $\sum_{t=1}^{\infty} b_t d_t < \infty$, $\sum_{t=1}^{\infty} b_t = \infty$, and $|d_{t+1} - d_t| < \hat{c}\, b_t \ \forall\ t$), thereby guaranteeing that $\lim_{t\rightarrow\infty} d_t = 0$ a.s, i.e.,

$$g(\hat{W}_t) - \hat{g}_t(\hat{W}_t) \rightarrow 0 \ \ a.s \quad (37)$$

Since we have already proved that $\hat{g}_t(\hat{W}_t)$ converges almost surely, (37) implies the almost sure convergence of $g(\hat{W}_t)$.

We also have by the generalized Glivenko-Cantelli Theorem (for example, see Theorem 19.4 and the result in Example 19.7 in [10]) that $\left\|g_t - g\right\|_{\infty} = \left\|g_t^{(1)} - g^{(1)}\right\|_{\infty} \rightarrow 0$ almost surely as $t \rightarrow \infty$. Combining this with the almost sure convergence of $g(\hat{W}_t)$, we finally obtain that $g_t(\hat{W}_t)$ converges almost surely as well. Moreover, with probability 1, the limits of $g(\hat{W}_t)$, $\hat{g}_t(\hat{W}_t)$ and $g_t(\hat{W}_t)$ are identical. ■

The fact that $g(\hat{W}_t) - \hat{g}_t(\hat{W}_t) \rightarrow 0$ a.s (in the above proof) also directly implies that $g^{(1)}(\hat{W}_t) - \hat{g}_t^{(1)}(\hat{W}_t) \rightarrow 0$ a.s.

Our final two results discuss the convergence of the iterates.

Proposition 5: Let Assumptions A1-A4 hold. Then, every accumulation point $\hat{W}_{\infty}$ of $\left\{\hat{W}_t\right\}$ is a stationary point of the expected cost $g(W)$ wp.1. Moreover, every accumulation point achieves the same value g^* of the expected cost wp.1.

Proof: Since the sequence $\left\{\hat{W}_t\right\}$ in the online algorithm is bounded, each of its accumulation points is bounded as well. Consider a convergent subsequence $\left\{\hat{W}_{l_t}\right\}$ of the bounded sequence $\left\{\hat{W}_t\right\}$, with accumulation point $\hat{W}_{\infty}$. The matrix $\hat{W}_{\infty}$ is non-singular – otherwise the boundedness of the objective sequence $\left\{\hat{g}_{l_t}(\hat{W}_{l_t})\right\}$ would be violated.

Now, consider a matrix $W \in \mathbb{R}^{n\times n}$. By the definition of g_t, we have that

$$\hat{g}_t(W) \geq g_t(W) \ \forall\ t \quad (38)$$

This implies that $\hat{g}_t^{(1)}(W) \geq g_t^{(1)}(W)$. We know that

$$\hat{g}_t^{(1)}(W) = tr\left\{W\Gamma_t W^T - 2W\Theta_t + C_t\right\} \quad (39)$$

where $\Gamma_t = t^{-1}\sum_{j=1}^{t} y_j y_j^T$, $\Theta_t = t^{-1}\sum_{j=1}^{t} y_j \hat{x}_j^T$, $C_t = t^{-1}\sum_{j=1}^{t} \hat{x}_j \hat{x}_j^T$, and $tr(\cdot)$ denotes the matrix trace. Similarly,

$$g_t^{(1)}(W) = tr\left\{W\Gamma_t W^T - 2W\hat{\Theta}_t + \hat{C}_t\right\} \quad (40)$$

where the matrix $\hat{\Theta}_t = t^{-1}\sum_{j=1}^{t} y_j \left(H_s(Wy_j)\right)^T$, and $\hat{C}_t = t^{-1}\sum_{j=1}^{t} H_s(Wy_j)\left(H_s(Wy_j)\right)^T$. Assuming a fixed W, we have that Γ_t, Θ_t, C_t, $\hat{\Theta}_t$, and $\hat{C}_t$ are all bounded. Therefore, we can find a convergent subsequence of $\left\{\Gamma_{l_t}, \Theta_{l_t}, C_{l_t}, \hat{\Theta}_{l_t}, \hat{C}_{l_t}\right\}$. Let the subsequence be indexed by l_{q_t} ($t \geq 1$), and let the accumulation point be $\left\{\Gamma_{\infty}, \Theta_{\infty}, C_{\infty}, \hat{\Theta}_{\infty}, \hat{C}_{\infty}\right\}$. Taking the limit $t \rightarrow \infty$ on both sides of the inequality $\hat{g}_{l_{q_t}}^{(1)}(W) \geq g_{l_{q_t}}^{(1)}(W)$ immediately yields

$$\hat{g}_{\infty}^{(1)}(W) \geq g_{\infty}^{(1)}(W) \quad (41)$$

where the subscript ∞ denotes the corresponding accumulation point (of subsequence). Since $g_{\infty}^{(1)}(W) = g^{(1)}(W)$ almost surely, we have that

$$\hat{g}_{\infty}^{(1)}(W) \geq g^{(1)}(W) \ \ a.s \quad (42)$$

Since $\left\{\hat{W}_{l_t}\right\}$ converges to $\hat{W}_{\infty}$, it means that the subsequence $\left\{\hat{W}_{l_{q_t}}\right\}$ converges to the same limit $\hat{W}_{\infty}$. To prove that $\hat{W}_{\infty}$ is a stationary point of $g(W)$, we set $W = \hat{W}_{\infty} + \alpha B$ in (42) for $B \in \mathbb{R}^{n\times n}$ and small $\alpha > 0$, and consider expansions of the functions in (42). By (39), taking the limit in t of $\hat{g}_{l_{q_t}}^{(1)}(W)$, the function $\hat{g}_{\infty}^{(1)}(W)$ is (convex) quadratic with Lipschitz gradient and Lipschitz constant L, and has a quadratic upper bound, i.e., it satisfies the property $\hat{g}_{\infty}^{(1)}(W_2) \leq \hat{g}_{\infty}^{(1)}(W_1) + tr\left\{(W_2 - W_1)^T \nabla \hat{g}_{\infty}^{(1)}(W_1)\right\} + \frac{L}{2}\left\|W_2 - W_1\right\|_F^2 \ \forall\ W_1, W_2$. In particular, we set $W_2 = \hat{W}_{\infty} + \alpha B$ and $W_1 = \hat{W}_{\infty}$ in the preceding equation, which gives an upper bound for $\hat{g}_{\infty}^{(1)}(\hat{W}_{\infty} + \alpha B)$. We also use a first order Taylor series expansion for $g^{(1)}(\hat{W}_{\infty} + \alpha B)$ (the existence of $\nabla g^{(1)}(\hat{W}_{\infty})$ is proved in Appendix D) in (42) to get

$$\hat{g}_{\infty}^{(1)}(\hat{W}_{\infty}) + tr\left\{\alpha B^T \nabla \hat{g}_{\infty}^{(1)}(\hat{W}_{\infty})\right\} + \frac{L}{2}\left\|\alpha B\right\|^2 \geq g^{(1)}(\hat{W}_{\infty}) + tr\left\{\alpha B^T \nabla g^{(1)}(\hat{W}_{\infty})\right\} + \epsilon \ \ a.s \quad (43)$$

where ϵ is the Taylor series remainder for the right hand side.

We now show that $\hat{g}_{\infty}^{(1)}(\hat{W}_{\infty}) \stackrel{a.s}{=} g^{(1)}(\hat{W}_{\infty})$ in (43). First, we use the fact that $\hat{g}_{l_{q_t}}^{(1)}(\hat{W}_{l_{q_t}})$ converges (by using (39)) to $\hat{g}_{\infty}^{(1)}(\hat{W}_{\infty})$. Furthermore, similar to the function $u^{(1)}(y, W)$ in Lemma 5 of Appendix B, $g^{(1)}(W)$ is also Lipschitz on a compact set. Since $\hat{W}_{l_{q_t}}$ converges to $\hat{W}_{\infty}$, we therefore have that $g^{(1)}(\hat{W}_{l_{q_t}})$ converges to $g^{(1)}(\hat{W}_{\infty})$. Therefore, using Proposition 4 (note that $\hat{g}_{l_{q_t}}^{(2)}(\hat{W}_{l_{q_t}})$ and $g^{(2)}(\hat{W}_{l_{q_t}})$ both converge to $\lambda_0\, v(\hat{W}_{\infty})$, where $\hat{W}_{\infty}$ has full rank), it is clear that $\hat{g}_{\infty}^{(1)}(\hat{W}_{\infty}) \stackrel{a.s}{=} g^{(1)}(\hat{W}_{\infty})$ in (43).

Based on the above arguments, it is also clear that $g(\hat{W}_{l_{q_t}}) = g^{(1)}(\hat{W}_{l_{q_t}}) + g^{(2)}(\hat{W}_{l_{q_t}})$ converges (as $t \rightarrow \infty$) to $g(\hat{W}_{\infty})$. Combining this with the result of Proposition 4, it is clear that $g(\hat{W}_{\infty}) \stackrel{wp.1}{=} g^*$, where g^* is the limit of $g(\hat{W}_t)$ defined in Proposition 4. Thus, the accumulation point $\hat{W}_{\infty}$ achieves the value g^* of the expected cost wp.1.

Now, upon substituting $\hat{g}_{\infty}^{(1)}(\hat{W}_{\infty}) \stackrel{a.s}{=} g^{(1)}(\hat{W}_{\infty})$ into (43), then dividing (43) by $\left\|\alpha B\right\|$ ($B \neq 0$), and letting $\alpha \rightarrow 0$, we get that

$$tr\left\{B^T \nabla \hat{g}_{\infty}^{(1)}(\hat{W}_{\infty})\right\} \geq tr\left\{B^T \nabla g^{(1)}(\hat{W}_{\infty})\right\} \ \ a.s \quad (44)$$

where we used the property of the Taylor series remainder that $\epsilon = o(\|\alpha B\|)$, so that $\lim_{\alpha \to 0}(\epsilon / \|\alpha B\|) = 0$. Since (44) holds wp.1 for any B, we must have $\nabla g^{(1)}(\hat{W}_\infty) \overset{wp.1}{=} \nabla \hat{g}_\infty^{(1)}(\hat{W}_\infty)$. This also implies that $\nabla g(\hat{W}_\infty) \overset{wp.1}{=} \nabla \hat{g}_\infty(\hat{W}_\infty)$.

Now, since $\hat{W}_{l_{q_t}}$ is a non-singular global minimizer of $\hat{g}_{l_{q_t}}(W)$ (for every t), we have that $\nabla \hat{g}_{l_{q_t}}(\hat{W}_{l_{q_t}})$ exists and $\nabla \hat{g}_{l_{q_t}}(\hat{W}_{l_{q_t}}) = 0$ for each t. Moreover, $\nabla \hat{g}_{l_{q_t}}(\hat{W}_{l_{q_t}})$ converges to $\nabla \hat{g}_\infty(\hat{W}_\infty)$ as $t \to \infty$. Therefore, $\nabla \hat{g}_\infty(\hat{W}_\infty) = 0$. Since $\nabla \hat{g}_\infty(\hat{W}_\infty) = \nabla g(\hat{W}_\infty)$ with probability 1, we therefore have that $\nabla \hat{g}_\infty(\hat{W}_\infty) = \nabla g(\hat{W}_\infty) = 0$ wp.1, or in other words, $\hat{W}_\infty$ is a stationary point of the function $g(W)$ wp.1.

Since we worked with an arbitrary convergent subsequence $\left\{\hat{W}_{l_t}\right\}$ of the bounded sequence $\left\{\hat{W}_t\right\}$ in the above derivation, the result indicates that every accumulation point of $\left\{\hat{W}_t\right\}$ is a stationary point of g wp.1, i.e., it satisfies first-order optimality conditions. Moreover, by the aforementioned arguments, every accumulation point achieves the same value (i.e., g^*) of the expected cost wp.1. ■

Proposition 6: Let Assumptions A1-A4 hold. Then, the distance between $\hat{W}_t$ and the set of stationary points of the expected cost $g(W)$ converges to 0 almost surely as $t \to \infty$.

Proof: From Proposition 5, we know that every accumulation point of the bounded sequence $\left\{\hat{W}_t\right\}$ is a stationary point of the expected cost $g(W)$ wp.1. Now, consider a convergent subsequence $\left\{\hat{W}_{l_t}\right\}$ of $\left\{\hat{W}_t\right\}$, that converges to a (nonsingular) limit $\hat{W}_\infty$. Then, $\nabla g(\hat{W}_\infty) = 0$ wp.1.

Now, by Lemma 7 of Appendix D, $\nabla g^{(1)}(\cdot)$ exists and is continuous at $\hat{W}_\infty$. Moreover, $\nabla g^{(2)}(\cdot)$ is also easily continuous at the nonsingular matrix $\hat{W}_\infty$. Combining these two results, it follows that $\nabla g(\cdot)$ is continuous at $\hat{W}_\infty$. Since $\left\{\hat{W}_{l_t}\right\}$ converges to $\hat{W}_\infty$, therefore, due to continuity $\left\{\nabla g(\hat{W}_{l_t})\right\}$ converges to $\nabla g(\hat{W}_\infty)$, which is 0 wp.1. The above results in fact imply that for any convergent subsequence of $\left\{\nabla g(\hat{W}_t)\right\}$, the limit is 0 wp.1.

We now show that the sequence $\left\{\nabla g(\hat{W}_t)\right\}$ is bounded. The gradient $\nabla g(\hat{W}_t)$ exists for all t because each $\hat{W}_t$ is nonsingular. We can use the expressions for $\nabla g^{(1)}(\hat{W}_t)$ in Appendix D along with the fact that $\left\{\hat{W}_t\right\}$ is bounded to conclude that $\left\{\nabla g^{(1)}(\hat{W}_t)\right\}$ is a bounded sequence. Next, because by Lemma 2, $\left\{\hat{g}_t(\hat{W}_t)\right\}$ is a bounded sequence, we have the following inequality where $C > 0$ is a bound on $\hat{g}_t(\hat{W}_t)\ \forall t$, $\beta_n(\hat{W}_t)$ denotes the smallest singular value of $\hat{W}_t$, and we use the non-negativity of the various terms in the function $\hat{g}_t(\hat{W}_t)$ to arrive at the first inequality below.

$$\lambda_0 \left(\beta_n^2(\hat{W}_t) - \log\, \beta_n(\hat{W}_t)\right) \leq \hat{g}_t(\hat{W}_t) \leq C\ \forall\, t \qquad (45)$$

The above result implies that $-\lambda_0 \log \beta_n(\hat{W}_t) \leq C\ \forall\ t$, or that $\beta_n(\hat{W}_t)$ is bounded from below. This can be used to easily show that $\nabla g^{(2)}(\hat{W}_t) = 2\lambda_0 \hat{W}_t - \lambda_0 \hat{W}_t^{-T}$ is bounded (in norm) for all t. The preceding arguments imply that $\left\{\nabla g(\hat{W}_t)\right\}$ is a bounded sequence. This implies that the limit inferior and limit superior of the scalar sequence formed over t using any specific entry of $\nabla g(\hat{W}_t)$, are finite.

Since wp.1, 0 is the only limit for any convergent subsequence of the bounded sequence $\left\{\nabla g(\hat{W}_t)\right\}$, this implies that the limit inferior and limit superior of the scalar sequence formed over t using any specific entry of $\nabla g(\hat{W}_t)$, are zero wp.1. Thus, $\left\{\nabla g(\hat{W}_t)\right\}$ itself converges to 0 a.s.

From the above arguments, we can conclude that the distance between $\hat{W}_t$ and the set of stationary points of the expected cost $g(W)$ converges to 0 almost surely as $t \to \infty$ (*all* the accumulation points of $\left\{\hat{W}_t\right\}$ are stationary points wp.1). ■

This completes the proof of the results stated in Section III.

V. Conclusions

In this paper, we analyzed the convergence behavior of the newly proposed online sparsifying transform learning algorithms. We showed that the online transform learning algorithms are guaranteed to converge (almost surely) to the set of stationary points of the learning problem. Unlike prior work on online synthesis dictionary learning [5], our guarantee relies on only a few simple assumptions.

Appendix A
Support of a Thresholded Perturbed Vector

For a vector h, we let $\beta_j(h)$ denote the magnitude of the j^{th} largest element (magnitude-wise) of h.

Lemma 4: Consider $\alpha \in \mathbb{R}^n$, and a $\delta \in \mathbb{R}^n$. Let s be a given sparsity level. Then, there exists an $\epsilon > 0$ such that the support of $H_s(\alpha + \delta)$ contains the support of one of the optimal codes in $\tilde{H}_s(\alpha)$, whenever $\|\delta\|_2 < \epsilon$. Furthermore, except in the (degenerate) case when $\beta_s(\alpha) = 0$, the support of $H_s(\alpha + \delta)$ coincides with the support of one of the optimal codes in $\tilde{H}_s(\alpha)$, whenever $\|\delta\|_2 < \epsilon$.

Proof: First, suppose that $\tilde{H}_s(\alpha)$ (the set of optimal projections of α onto the s-ℓ_0 ball) is a singleton, and $\beta_s(\alpha) > 0$, so that $\beta_s(\alpha) - \beta_{s+1}(\alpha) > 0$. Then, whenever $\|\delta\|_\infty < (\beta_s(\alpha) - \beta_{s+1}(\alpha))/2$ holds, we have that $H_s(\alpha + \delta)$ has the same support set (non-zero locations) as $H_s(\alpha) = \tilde{H}_s(\alpha)$.

Next, when $\tilde{H}_s(\alpha)$ is a singleton, but $\beta_s(\alpha) = 0$ (and $\alpha \neq 0$), let γ be the magnitude of the non-zero element of α of smallest magnitude. Then, whenever $\|\delta\|_\infty < \gamma/2$ holds, we have that the support of $H_s(\alpha) = \tilde{H}_s(\alpha)$ is contained in the support of $H_s(\alpha + \delta)$.

Finally, suppose $\tilde{H}_s(\alpha)$ is not a singleton (there are ties). Let us define $a \in \mathbb{R}^{n-1}$ as follows

$$a_j \triangleq \frac{\beta_j(\alpha) - \beta_{j+1}(\alpha)}{2}\ \ 1 \leq j \leq n-1 \qquad (46)$$

Set ϵ to be the smallest non-zero element of a. Then, it is easy to show that the support of $H_s(\alpha + \delta)$ (the support can vary with δ) coincides with the support of one of the optimal codes in $\tilde{H}_s(\alpha)$ when $\|\delta\|_\infty < \epsilon$.

If the a computed using α in (46) is a zero vector, then all elements of α must have identical magnitude. In this case, the lemma trivially holds for any $\epsilon > 0$. ■

Appendix B
Lipschitz Continuity of $u^{(1)}(y, W)$

We denote $u^{(1)}(y, W)$ by $\phi(y, W)$ in this Appendix for simplicity.

Lemma 5: Let Assumption A1 hold. Then, the function $\phi(y,W) = \|Wy - H_s(Wy)\|_2^2$ is uniformly Lipschitz with respect to W on the bounded set $S \triangleq \{W \in \mathbb{R}^{n\times n} : \|W\|_2 \leq 1\}$.

Proof: Consider a bounded open set A containing S, such that $W \in A$ satisfies $\|W\|_2 \leq c$ for some finite $c > 1$. Let us choose two matrices W_1 and W_2 from A. To simplify notation in this proof, we denote $\phi(y,W_i) = a_i^2$, $a_i \geq 0$, $i = 1, 2$. Then,

$$\left|\phi(y,W_1) - \phi(y,W_2)\right| = |a_1 + a_2| \cdot |a_1 - a_2| \tag{47}$$

Since $|a_1 + a_2| \leq 4c$, we need to only (Lipschitz) bound $|a_1 - a_2|$. Let us define $\Delta \triangleq W_1 - W_2$. Then, we have

$$a_1 = \left\|W_2 y + \Delta y - H_s(W_2 y + \Delta y)\right\|_2 \tag{48}$$

Let Γ be a diagonal matrix, with entries either 0 or 1. The 1's are at locations corresponding to the support (indices of the non-zero locations) of $H_s(W_2 y + \Delta y)$. Then, applying the triangle inequality to (48), we have

$$a_1 \leq \left\|W_2 y - \Gamma W_2 y\right\|_2 + \left\|\Delta y - \Gamma \Delta y\right\|_2$$

$$a_1 \geq \left\|W_2 y - \Gamma W_2 y\right\|_2 - \left\|\Delta y - \Gamma \Delta y\right\|_2$$

Now, by Lemma 4 in Appendix A, it is clear that there exists an $\epsilon > 0$ depending on $W_2 y$ (or, depending on W_2, for fixed y) such that whenever $\left\|\Delta\right\|_2 < \epsilon$ (i.e., $\left\|\Delta y\right\|_2 < \epsilon$), then the support of $H_s(W_2 y + \Delta y)$ either coincides with, or contains the support of one of the optimal sparse codes in $\tilde{H}_s(W_2 y)$.

It follows that for each Δ with $\left\|\Delta\right\|_2 < \epsilon$, we have that the support of the diagonal of Γ (depends on Δ) is also the support of some sparse code in $\tilde{H}_s(W_2 y)$ [7], and because all the codes in $\tilde{H}_s(W_2 y)$ provide the same sparsification error, we have $\left\|W_2 y - \Gamma W_2 y\right\|_2 = \left\|W_2 y - H_s(W_2 y)\right\|_2$. Furthermore, $\left\|\Delta y - \Gamma \Delta y\right\|_2 \leq \left\|\Delta y\right\|_2 \leq \left\|\Delta\right\|_2$. Combining the inequalities for a_1 with the above results, we obtain

$$-\left\|\Delta\right\|_2 \leq a_1 - a_2 \leq \left\|\Delta\right\|_2 = \left\|W_1 - W_2\right\|_2 \tag{49}$$

Combining (49) and (47), we obtain

$$\left|\phi(y,W_1) - \phi(y,W_2)\right| \leq 4c \left\|W_1 - W_2\right\|_2 \tag{50}$$

whenever $\left\|W_1 - W_2\right\|_2 < \epsilon$. Since W_2 was arbitrarily chosen, and ϵ is a function of W_2 only, (50) implies that $\phi(y,W)$ is locally Lipschitz on the open set A. Finally, since S is a compact subset of A ($\phi(y,W)$ is also bounded on S), it follows from standard results [11] that $\phi(y,W)$ is Lipschitz on S. ■

APPENDIX C
USEFUL THEOREMS

Here, we list some theorems relevant to the convergence analysis of the online algorithm. The first theorem is from the perturbation theory of singular value decompositions (cf. [12], [13], and Chapter 15 of [14] and references therein).

Theorem 1: Let matrices $A \in \mathbb{R}^{n\times n}$ and $\tilde{A} \in \mathbb{R}^{n\times n}$, with $\tilde{A} = A + E$, where E is a perturbation matrix. Let $A = U\Sigma V^T$ and $\tilde{A} = \tilde{U}\tilde{\Sigma}\tilde{V}^T$ be the respective SVDs with the respective singular values σ_i and $\tilde{\sigma}_i$ arranged in decreasing order for $1 \leq i \leq n$. Then, we have

$$|\tilde{\sigma}_i - \sigma_i| \leq \|E\|_2 , \quad \forall\, i \tag{51}$$

Furthermore, for any index i, define

$$\eta_i = \min\left\{\left(\min_{j:j\neq i} |\tilde{\sigma}_i - \sigma_j|\right), \tilde{\sigma}_i + \sigma_i\right\} \tag{52}$$

If $\eta_i > 0$, then

$$\min_{\tilde{\alpha}\in\{-1,+1\}} \sqrt{\|u_i\tilde{\alpha} - \tilde{u}_i\|_2^2 + \|v_i\tilde{\alpha} - \tilde{v}_i\|_2^2} \leq \sqrt{2}\frac{\sqrt{\|r\|_2^2 + \|s\|_2^2}}{\eta_i} \tag{53}$$

where u_i, v_i, $\tilde{u}_i$, and $\tilde{v}_i$ denote the i^{th} columns of U, V, $\tilde{U}$, and $\tilde{V}$, respectively, and $r = A\tilde{v}_i - \tilde{\sigma}_i\tilde{u}_i$, $s = A^T\tilde{u}_i - \tilde{\sigma}_i\tilde{v}_i$. Specifically, $\|r\|_2 \leq \|E\|_2$ and $\|s\|_2 \leq \|E\|_2$.

When the perturbation E is small, Theorem 1 indicates that corresponding singular values of A and $\tilde{A} = A + E$ are close. In order for the corresponding singular vectors of A and $\tilde{A}$ to be close to each other as well (up to ± 1 scaling), the constant η_i in (52) needs to be positive for each i. This is true, for example, when A has full rank, and has distinct singular values, and E is sufficiently small.

The next two Theorems (cf. [5], [8], [10] and references therein) are used in the proof of proposition 4. In theorem 2 (see Chapter 19.2, Lemma 19.36, and Example 19.7 in [10]), the expectation $\mathbb{E}_Z[\cdot]$ is calculated with respect to the probability measure on χ, and Z_1, Z_2,.. are independent random vectors distributed according to this probability measure.

Theorem 2: Let $\mathcal{G} = \{g_\theta : \chi \mapsto \mathbb{R}, \theta \in \Xi\}$ be a set of measurable functions indexed by a bounded subset Ξ of $\mathbb{R}^d$. Suppose that there exists a constant M such that $|g_{\theta_1}(z) - g_{\theta_2}(z)| \leq M\|\theta_1 - \theta_2\|_2$ for every $\theta_1, \theta_2 \in \Xi$ and $z \in \chi$. Then, $\mathcal{G}$ is P-Donsker (see [10]). For any g in $\mathcal{G}$, define $\mathbb{P}_t g \triangleq \frac{1}{t}\sum_{j=1}^t g(Z_j)$, $\mathbb{P}g \triangleq \mathbb{E}_Z[g(Z)]$, and $\mathbb{H}_t g \triangleq \sqrt{t}(\mathbb{P}_t g - \mathbb{P}g)$. Suppose also that for all g, $\mathbb{P}g^2 < \delta^2$ and $\|g\|_\infty < C$ for some $\delta, C < \infty$, and that the random elements Z_1, Z_2,... are Borel-measurable. Then, we have

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}} |\mathbb{H}_t g|\right] = O(1) \tag{54}$$

The following theorem [5], [8], [15] is on the convergence of a stochastic sequence.

Theorem 3: Let $(\Omega, \mathcal{F}, P)$ be a measurable probability space, r_t, for $t \geq 0$, be a realization of a stochastic process and $\mathcal{F}_t$ be the filtration determined by the past information at time t. Define

$$\delta_t = \begin{cases} 1 & \text{if } \mathbb{E}[r_{t+1} - r_t|\mathcal{F}_t] > 0 \\ 0 & \text{otherwise} \end{cases}$$

If for all t, $r_t \geq 0$, and $\sum_{t=1}^\infty \mathbb{E}[\delta_t \cdot (r_{t+1} - r_t)] < \infty$, then r_t is a quasi-martingale and converges almost surely. Moreover,

$$\sum_{t=1}^\infty \left|\mathbb{E}[r_{t+1} - r_t|\mathcal{F}_t]\right| < \infty \quad a.s. \tag{55}$$

The next result is on the limit of a real sequence [5].

[7] In the trivial case that the s^{th} largest magnitude element of $W_2 y$ is zero, the support of the diagonal of Γ contains the entire support of the singleton $\tilde{H}_s(W_2 y) = W_2 y$ (Appendix A).

Theorem 4: Let $\{b_k\}$ and $\{d_k\}$ be two sequences of real numbers such that $\forall\ k$, we have $b_k \geq 0$, $d_k \geq 0$, $\sum_{k=1}^{\infty} b_k d_k < \infty$, and $\sum_{k=1}^{\infty} b_k = \infty$. Furthermore, $\exists\ C > 0$ such that $|d_{k+1} - d_k| < C b_k\ \forall\ k$. Then, $\lim_{k\to\infty} d_k = 0$.

Finally, the following result is on the directional differentiability of optimal value functions [16], [17].

Theorem 5: Let $h : \mathbb{R}^p \times \mathbb{R}^q \mapsto \mathbb{R}$. Suppose that for all $x \in \mathbb{R}^p$, the function $h(x,\cdot)$ is differentiable, and that h and $\nabla_u h(x,u)$, the derivative of $h(x,\cdot)$, are continuous on $\mathbb{R}^p \times \mathbb{R}^q$. Let $l(u)$ be the optimal value function defined as $l(u) = \min_{x\in A} h(x,u)$, where A is a compact subset of $\mathbb{R}^p$. Then, we have that $l(u)$ is directionally differentiable. Furthermore, if for $u_0 \in \mathbb{R}^q$, $h(\cdot,u_0)$ has a unique minimizer x_0, then $l(u)$ is differentiable at u_0 and $\nabla_u l(u_0) = \nabla_u h(x_0,u_0)$.

Appendix D
On the Existence of $\nabla g^{(1)}(W)$

In the following Lemma 7, we establish the existence and continuity of $\nabla g^{(1)}(\cdot)$. Let the Assumptions A1-A4 stated in Section II hold. First, the following lemma establishes conditions under which the operation $\tilde{H}_s(\cdot)$ produces a singleton.

Lemma 6: Let $y \in \mathbb{R}^n$ be distributed over the n-dimensional unit sphere $\{y \in \mathbb{R}^n : \|y\|_2 = 1\}$, with an absolutely continuous probability measure. Suppose matrix $W \in \mathbb{R}^{n\times n}$ is nonsingular. Then $\tilde{H}_s(Wy)$ is a singleton wp.1.

Proof: In order for $\tilde{H}_s(Wy)$ to not be a singleton (i.e., non-unique transform sparse coding solution), a necessary condition is that at least two entries of the vector Wy have the same magnitude, i.e., $w_i y = w_j y$ or $w_i y = -w_j y$, where w_i and w_j are distinct rows of W. This event corresponds to $(w_i \pm w_j)y = 0$. Since W is full rank, $(w_i \pm w_j) \neq 0$. Therefore, the non-singleton event occurs when y lies on one of the two $n-1$ dimensional hyperplanes characterized by $(w_i \pm w_j)z = 0$. However, the intersection of each these $n-1$ dimensional hyperplanes with the n-dimensional unit sphere is a set of zero Lebesgue measure, and hence, because of the absolute continuity of the distribution of y, $\mathbb{P}\{(w_i \pm w_j)y = 0\} = 0$. This easily implies that the probability that any two entries (any $i \neq j$) of the vector Wy have the same magnitude is zero, or that $\tilde{H}_s(Wy)$ is a singleton with probability 1. ■

Lemma 7: Consider a matrix $W \in \mathbb{R}^{n\times n}$ that has full rank, and is bounded ($\|W\|_2 \leq 1$). Then, $\nabla g^{(1)}(\cdot)$ exists, and is continuous at W.

Proof: For $y \in \mathbb{R}^n$ with $\|y\|_2 = 1$, we have

$$u^{(1)}(y,W) = \min_{x:\|x\|_0 \leq s} \|Wy - x\|_2^2 \tag{56}$$

The set of optimal $\hat{x}$ above is characterized by $\tilde{H}_s(Wy)$. Similar to the bound in (10), we can conclude that any optimal $\hat{x}$ satisfies $\|\hat{x}\|_2 \leq 1$. Therefore, without loss of generality, we can consider the minimization in (56) over the set $A \triangleq \{x \in \mathbb{R}^n : \|x\|_0 \leq s, \|x\|_2 \leq 1\}$. It is easy to verify that A is compact.

Now, we apply Theorem 5 in Appendix C to prove that $u^{(1)}(y,\cdot)$ is continuously differentiable at (full rank, bounded) W with probability 1. For any y, for which $\tilde{H}_s(Wy)$ is a singleton, it is easy to check that all conditions in Theorem 5 are satisfied with the function h in Theorem 5 defined as $\|Wy - x\|_2^2$, and the compact set A defined as above. Therefore, $\nabla_W u^{(1)}(y,W)$ exists (at the full rank, bounded W) and is equal [8] to $2Wyy^T - 2H_s(Wy)y^T$. It can be easily verified (using similar arguments as in the proofs of Lemma 5 and Lemma 4) that the gradient $2Wyy^T - 2H_s(Wy)y^T$ (which is bounded in Frobenius norm by 2) is continuous at W, when $\tilde{H}_s(Wy)$ is a singleton. For our full rank and bounded W, since $\tilde{H}_s(Wy)$ is a singleton wp.1 (by Lemma 6), we therefore have that the function $u^{(1)}(y,\cdot)$ is continuously differentiable at W with probability 1.

Based on the above results, the following equation holds.

$$\nabla g^{(1)}(W) = \nabla \mathbb{E}_y[u^{(1)}(y,W)] = \mathbb{E}_y\left[\nabla_W u^{(1)}(y,W)\right] \tag{57}$$

Thus, $\nabla g^{(1)}(W)$ exists for any $W \in \mathbb{R}^{n\times n}$ that is full rank with $\|W\|_2 \leq 1$. It is easy to show that it is continuous (since the derivative of $u^{(1)}(y,\cdot)$ is continuous at W with probability 1) at each such W. ■

References

[1] S. Ravishankar, B. Wen, and Y. Bresler, "Online sparsifying transform learning - part I: Algorithms," *IEEE Journal of Selected Topics in Signal Process.*, 2015, to appear.

[2] S. Ravishankar and Y. Bresler, "Learning sparsifying transforms," *IEEE Trans. Signal Process.*, vol. 61, no. 5, pp. 1072–1086, 2013.

[3] ——, "Learning doubly sparse transforms for images," *IEEE Trans. Image Process.*, vol. 22, no. 12, pp. 4598–4612, 2013.

[4] ——, "Closed-form solutions within sparsifying transform learning," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 5378–5382.

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

[6] Y. Z. Tsypkin, *Adaptation and Learning in automatic systems*. New York, NY: Academic Press, 1971.

[7] ——, *Foundations of the theory of learning systems*. New York, NY: Academic Press, 1973.

[8] L. Bottou, "Online algorithms and stochastic approximations," in *Online Learning and Neural Networks*, D. Saad, Ed. Cambridge, UK: Cambridge University Press, 1998.

[9] L. Bottou and O. Bousquet, "The tradeoffs of large scale learning," in *Advances in Neural Information Processing Systems*, vol. 20. MIT Press, 2008, pp. 161–168.

[10] A. W. van der Vaart, *Asymptotic Statistics*. New York, NY: Cambridge University Press, 1998.

[11] N. G. Markley, *Principles of Differential Equations*. Hoboken, New Jersey: John Wiley & Sons, Inc., 2004.

[12] G. W. Stewart, "Perturbation theory for the singular value decomposition," in *SVD and Signal Processing, II: Algorithms, Analysis and Applications*. Elsevier, 1990, pp. 99–109.

[13] F. M. Dopico, "A note on $\sin\theta$ theorems for singular subspace variations," *BIT Numerical Mathematics*, vol. 40, no. 2, pp. 395–403, 2000.

[14] L. Hogben, *Handbook of Linear Algebra*. Boca Raton, FL: CRC Press, 2006.

[15] D. L. Fisk, "Quasi-Martingales," *Transactions of the American Mathematical Society*, vol. 120, pp. 369–389, 1965.

[16] J. M. Danskin, *The Theory of Max-Min and Its Applications to Weapons Allocation Problems*. Heidelberg, Germany: Springer, 1967.

[17] J. F. Bonnans and A. Shapiro, "Optimization problems with perturbations: A guided tour," *SIAM Review*, vol. 40, no. 2, pp. 202–227, 1998.

The authors' biographies and photos appear in Part I [1].

[8] It is also easy to derive $\nabla_W u^{(1)}(y,W)$ (since $\tilde{H}_s(Wy)$ is a singleton) starting from the definition of the derivative as a limit, even without using Theorem 5.