Efficient Dictionary Learning via Very Sparse Random Projections

Farhad Pourkamali-Anaraki*‡, Stephen Becker†, Shannon M. Hughes*
*Department of Electrical, Computer, and Energy Engineering
University of Colorado at Boulder, Boulder, Colorado 80309
†Department of Applied Mathematics
University of Colorado at Boulder, Boulder, Colorado 80309
‡Email: Farhad.Pourkamali@colorado.edu

***Abstract*—Performing signal processing tasks on compressive measurements of data has received great attention in recent years. In this paper, we extend previous work on compressive dictionary learning by showing that more general random projections may be used, including sparse ones. More precisely, we examine compressive K-means clustering as a special case of compressive dictionary learning and give theoretical guarantees for its performance for a very general class of random projections. We then propose a memory and computation efficient dictionary learning algorithm, specifically designed for analyzing large volumes of high-dimensional data, which learns the dictionary from very sparse random projections. Experimental results demonstrate that our approach allows for reduction of computational complexity and memory/data access, with controllable loss in accuracy.**

I. Introduction

There are several ways to represent low-dimensional structure of high-dimensional data, the best known being principal component analysis (PCA). However, PCA is based on a linear subspace model that is generally not capable of capturing the geometric structure of real-world datasets [1].

The sparse signal model is a nonlinear generalization of the linear subspace model that has been used in various signal and image processing tasks [2]–[4], as well as compressive sensing [5]. This model assumes that each data sample can be represented as a linear combination of a few elements (atoms) from a dictionary. Data-adaptive dictionary learning can lead to a much more compact representation than predefined dictionaries such as wavelets, and thus a central problem is finding a good data-adaptive dictionary.

Dictionary learning algorithms such as the method of optimal directions (MOD) [6] and the K-SVD algorithm [7] aim to learn a dictionary by minimizing the representation error of data in an iterative procedure involving two steps of sparse coding and dictionary update. The latter often requires ready access to the entire data available at a central processing unit.

Due to increasing sizes of datasets, not only do algorithms take longer to run, but it may not even be feasible or practical to acquire and/or hold every data entry. In applications such as distributed databases, where data is typically distributed over an interconnected set of distributed sites [8], it is important to avoid communicating the entire data.

A promising approach to address these issues is to take a compressive sensing approach, where we only have access to compressive measurements of data. In fact, performing signal processing and data mining tasks on compressive versions of the data has been an important topic in the recent literature. For example, in [9], certain inference problems such as detection and estimation within the compressed domain have been studied. Several lines of work consider recovery of principal components [10]–[13], spectral features [14], and change detection [15] from compressive measurements.

In this paper, we focus on the problem of dictionary learning based on compressive measurements. Our contributions are twofold. First, we show the connection between dictionary learning in the compressed domain and K-means clustering. Most standard dictionary learning algorithms are indeed a generalization of the K-means clustering algorithm [16], where the reference to K-means is a common approach to analyze the performance of these algorithms [7], [17]. This paper takes initial steps towards providing theoretical guarantees for recovery of the true underlying dictionary from compressive measurements. Moreover, our analysis applies to compressive measurements obtained by a general class of random matrices consisting of i.i.d. zero-mean entries and finite first four moments.

Second, we extend the prior work in [18] where compressive dictionary learning for random Gaussian matrices is considered. In particular, we propose a memory and computation efficient dictionary learning algorithm applicable to modern data settings. To do this, we learn a dictionary from very sparse random projections, i.e. projection of the data onto a few very sparse random vectors with Bernoulli-generated nonzero entries. These sparse random projections have been applied in many large-scale applications such as compressive sensing and object tracking [19], [20] and to efficient learning of principal components in the large-scale data setting [13]. To further improve efficiency of our approach, we show how to share the same random matrix across blocks of data samples.

II. Prior Work on Compressive Dictionary Learning

Several attempts have been made to address the problem of dictionary learning from compressive measurements. In three roughly contemporary papers [21], [22], and our work [18], three similar algorithms were presented to learn a dictionary

based on compressive measurements. Each was inspired by the well-known K-SVD algorithm and closely followed its structure, except in that each aimed to minimize the representation error of the compressive measurements instead of that of the original signals. The exact steps of each algorithm have minor differences, but take a similar overall form.

However, none of these works explicitly aimed at *designing* the compressive measurements (sketches) to promote the computational efficiency of the resulting compressive K-SVD, so that it would be maximally practical for dictionary learning on large-scale data. Moreover, none of these works gave theoretical performance analysis for such computationally-efficient sketches.

In this paper, we extend the previous line of work on compressive dictionary learning by analyzing the scheme under assumptions that make it memory and computation efficient. The key to the efficiency of the new scheme is in considering a wider and more general class of random projection matrices for the sketches, including some very sparse ones. We further introduce an initial analysis of the theoretical performance of compressive dictionary learning under these more general random projections.

In this section, we review the general dictionary learning problem and the compressive K-SVD (CK-SVD) algorithm that was introduced in [18] for the case of random Gaussian matrices. (We note that the approaches of [22] and [21] are similar.) Given a set of n training signals $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ in $\mathbb{R}^p$, the dictionary learning problem is to find a dictionary $\mathbf{D} \in \mathbb{R}^{p \times K}$ that leads to the best representation under a strict sparsity constraint for each member in the set, i.e., minimizing

$$\min_{\mathbf{D} \in \mathbb{R}^{p \times K}, \mathbf{C} \in \mathbb{R}^{K \times n}} \sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{D}\mathbf{c}_i\|_2^2 \ s.t.\ \forall i,\ \|\mathbf{c}_i\|_0 \leq T \quad (1)$$

where $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_n]$ is the coefficient matrix and the ℓ_0 pseudo-norm $\|\mathbf{c}_i\|_0$ counts the number of nonzero entries of the coefficient vector $\mathbf{c}_i \in \mathbb{R}^K$. Moreover, the columns of the dictionary $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_K]$ are typically assumed to have unit ℓ_2-norm. Problem (1) is generally intractable so we look for approximate solutions (e.g., via K-SVD [7]).

We then consider compressed measurements (sketches), where each measurement is obtained by taking inner products of the data sample $\mathbf{x}_i \in \mathbb{R}^p$ with the columns of a matrix $\mathbf{R}_i$, i.e., $\mathbf{y}_i = \mathbf{R}_i^T \mathbf{x}_i$ with $\{\mathbf{R}_i\}_{i=1}^n \in \mathbb{R}^{p \times m}$, $m < p$, and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n] \in \mathbb{R}^{m \times n}$. In [18], the entries of $\mathbf{R}_i$ are i.i.d. from a zero-mean Gaussian distribution, which is an assumption we drop in the current paper.

Given access only to the compressed measurements $\mathbf{y}_i$ and not $\mathbf{x}_i$, we attempt to solve the following compressive dictionary learning problem:

$$\min_{\substack{\mathbf{D} \in \mathbb{R}^{p \times K} \\ \mathbf{C} \in \mathbb{R}^{K \times n}}} \sum_{i=1}^{n} \left\|\mathbf{y}_i - \mathbf{R}_i^T \mathbf{D}\mathbf{c}_i\right\|_2^2 \ s.t.\ \begin{cases} \forall i,\ \|\mathbf{c}_i\|_0 \leq T \\ \forall k,\ \|\mathbf{d}_k\|_2 = 1 \end{cases} \quad (2)$$

In the CK-SVD algorithm, the objective function in (2) is minimized in a simple iterative approach that alternates between sparse coding and dictionary update steps.

A. Sparse Coding

In the sparse coding step, the penalty term in (2) is minimized with respect to a fixed $\mathbf{D}$ to find the coefficient matrix $\mathbf{C}$ under the strict sparsity constraint. This can be written as

$$\min_{\mathbf{C} \in \mathbb{R}^{K \times n}} \sum_{i=1}^{n} \|\mathbf{y}_i - \Psi_i \mathbf{c}_i\|_2^2 \ s.t.\ \forall i,\ \|\mathbf{c}_i\|_0 \leq T \quad (3)$$

where $\Psi_i = \mathbf{R}_i^T \mathbf{D} \in \mathbb{R}^{m \times K}$ is a fixed equivalent dictionary for representation of $\mathbf{y}_i$. This optimization problem can be considered as n distinct optimization problems for each compressive measurement. We can then use a variety of algorithms, such as OMP, to find the approximate solution $\mathbf{c}_i$ [23].

B. Dictionary Update

The approach is to update the k^{th} dictionary atom $\mathbf{d}_k$ and its corresponding coefficients while holding $\mathbf{d}_j$ fixed for $j \neq k$, and then repeat for $k+1$, until $k = K$. The penalty term in (2) can be written as

$$\begin{aligned} &\sum_{i=1}^{n} \left\|\mathbf{y}_i - \mathbf{R}_i^T \sum_{j=1}^{K} c_{i,j} \mathbf{d}_j\right\|_2^2 \\ &= \sum_{i=1}^{n} \left\|\left(\mathbf{y}_i - \mathbf{R}_i^T \sum_{j \neq k} c_{i,j} \mathbf{d}_j\right) - c_{i,k} \mathbf{R}_i^T \mathbf{d}_k\right\|_2^2 \\ &= \sum_{i \in \mathcal{I}_k} \left\|\mathbf{e}_{i,k} - c_{i,k} \mathbf{R}_i^T \mathbf{d}_k\right\|_2^2 + \sum_{i \notin \mathcal{I}_k} \left\|\mathbf{e}_{i,k}\right\|_2^2 \end{aligned} \quad (4)$$

where $c_{i,k}$ is the k^{th} element of $\mathbf{c}_i \in \mathbb{R}^K$, $\mathcal{I}_k$ is a set of indices of compressive measurements for which $c_{i,k} \neq 0$, and $\mathbf{e}_{i,k} = \mathbf{y}_i - \mathbf{R}_i^T \sum_{j \neq k} c_{i,j} \mathbf{d}_j \in \mathbb{R}^m$ is the representation error for $\mathbf{y}_i$ when the k^{th} dictionary atom is removed. The penalty term in (4) is a quadratic function of $\mathbf{d}_k$ and the minimizer is obtained by setting the derivative with respect to $\mathbf{d}_k$ equal to zero. Hence,

$$\mathbf{G}_k \mathbf{d}_k = \mathbf{b}_k \quad (5)$$

where $\mathbf{G}_k = \sum_{i \in \mathcal{I}_k} c_{i,k}^2 \mathbf{R}_i \mathbf{R}_i^T$ and $\mathbf{b}_k = \sum_{i \in \mathcal{I}_k} c_{i,k} \mathbf{R}_i \mathbf{e}_{i,k}$. Therefore, we get the closed-form solution $\mathbf{d}_k = \mathbf{G}_k^+ \mathbf{b}_k$, where $\mathbf{G}_k^+$ denotes the Moore-Penrose pseudo-inverse of $\mathbf{G}_k$. Once given the new $\mathbf{d}_k$ (normalized to have unit ℓ_2-norm), the optimal $c_{i,k}$ for each $i \in \mathcal{I}_k$ is given by least squares as $c_{i,k} = \frac{\langle \mathbf{e}_{i,k}, \mathbf{R}_i^T \mathbf{d}_k \rangle}{\|\mathbf{R}_i^T \mathbf{d}_k\|_2^2}$. By design, the support of the coefficient matrix $\mathbf{C}$ is preserved, just as in the K-SVD algorithm.

III. Initial Theoretical Analysis: K-Means Case

In this section, we provide an initial theoretical analysis of the performance of the CK-SVD algorithm by restricting our attention to a special case of dictionary learning: K-means clustering. In this special case, we can provide theoretical guarantees on the performance of CK-SVD at every step, in relation to the steps of K-means. Moreover, these guarantees will hold for a very general class of projection matrices including very sparse random projections.

We consider a statistical framework to establish the connection between CK-SVD and K-means. K-means clustering can be viewed as a special case of dictionary learning in which each data sample is allowed to use one dictionary atom (cluster center), i.e. $T = 1$, and the corresponding coefficient is set to be 1. Therefore, we consider the following generative model

$$\mathbf{x}_i = \overline{\mathbf{d}_k} + \epsilon_i,\ i \in \mathcal{I}_k \tag{6}$$

where $\overline{\mathbf{d}_k}$ is the center of the k^{th} cluster, and $\{\epsilon_i\}_{i=1}^n \in \mathbb{R}^p$ represent residuals in signal approximation and they are drawn i.i.d. from $\mathcal{N}(\mathbf{0}, \frac{\sigma^2}{p}\mathbf{I}_{p\times p})$, so that the approximation error is $\mathbb{E}[\|\epsilon\|_2^2] = \sigma^2$. The set $\{\mathcal{I}_k\}_{k=1}^K$ is an arbitrary partition of $(1, 2, \ldots, n)$, with the condition that $|\mathcal{I}_k| \to \infty$ as $n \to \infty$. The random matrices $\mathbf{R}_i$ are assumed to satisfy the following:

Assumption 1. *Each entry of the random matrices $\{\mathbf{R}_i\}_{i=1}^n \in \mathbb{R}^{p\times m}$ is drawn i.i.d. from a general class of zero-mean distributions with finite first four moments $\{\mu_k\}_{k=1}^4$.*

We will see that the distribution's kurtosis is a key factor in our results. The kurtosis, defined as $\kappa \triangleq \frac{\mu_4}{\mu_2^2} - 3$, is a measure of peakedness and heaviness of tail for a distribution.

We now show how, in this special case of K-means, CK-SVD would update the cluster centers. As mentioned before, in this case, we should set $T = 1$ and the correponding coefficients are set to be 1. This means that for all $i \in \mathcal{I}_k$, we have $c_{i,k} = 1$, and $c_{i,j} = 0$ for $j \neq k$, and it leads to $\mathbf{e}_{i,k} = \mathbf{y}_i, \forall i \in \mathcal{I}_k$. Then, the update formula for the k^{th} dictionary atom of CK-SVD given in (5) reduces to

$$\Big(\sum_{i\in\mathcal{I}_k} \mathbf{R}_i\mathbf{R}_i^T\Big)\mathbf{d}_k = \sum_{i\in\mathcal{I}_k} \mathbf{R}_i\mathbf{y}_i. \tag{7}$$

Hence, similar to K-means, the process of updating K dictionary atoms becomes independent of each other. We can rewrite (7) as $\mathbf{H}_k\mathbf{d}_k = \mathbf{f}_k$, where

$$\mathbf{H}_k \triangleq \frac{1}{m\mu_2}\frac{1}{|\mathcal{I}_k|}\sum_{i\in\mathcal{I}_k} \mathbf{R}_i\mathbf{R}_i^T,\ \mathbf{f}_k \triangleq \frac{1}{m\mu_2}\frac{1}{|\mathcal{I}_k|}\sum_{i\in\mathcal{I}_k} \mathbf{R}_i\mathbf{y}_i. \tag{8}$$

In [13], it is shown that $\mathbb{E}[\mathbf{R}_i\mathbf{R}_i^T] = m\mu_2\mathbf{I}_{p\times p}$. Thus, we see

$$\begin{aligned}\mathbb{E}\left[\mathbf{R}_i\mathbf{y}_i\right] &= \mathbb{E}\left[\mathbf{R}_i\mathbf{R}_i^T\mathbf{x}_i\right] \\ &= \mathbb{E}\left[\mathbf{R}_i\mathbf{R}_i^T\overline{\mathbf{d}_k}\right] + \mathbb{E}\left[\mathbf{R}_i\mathbf{R}_i^T\epsilon_i\right] \\ &= \mathbb{E}\left[\mathbf{R}_i\mathbf{R}_i^T\right]\overline{\mathbf{d}_k} + \mathbb{E}\left[\mathbf{R}_i\mathbf{R}_i^T\right]\mathbb{E}\left[\epsilon_i\right] = m\mu_2\overline{\mathbf{d}_k}.\end{aligned} \tag{9}$$

Therefore, when the number of samples is sufficiently large, using the law of large numbers, $\mathbf{H}_k$ and $\mathbf{f}_k$ converge to $\frac{1}{m\mu_2}\mathbb{E}[\mathbf{R}_i\mathbf{R}_i^T] = \mathbf{I}_{p\times p}$ and $\frac{1}{m\mu_2}\mathbb{E}[\mathbf{R}_i\mathbf{y}_i] = \overline{\mathbf{d}_k}$. Hence, the updated dictionary atom in our CK-SVD is the original center of cluster, i.e. $\mathbf{d}_k = \overline{\mathbf{d}_k}$, exactly as in K-means. Note that in this case even one measurement per signal $m = 1$ is sufficient.

The following theorem characterizes convergence rates for $\mathbf{H}_k$ and $\mathbf{f}_k$ based on various parameters such as the number of samples and the choice of random matrices.

Theorem 2. *Assume Assumption 1. Then, $\mathbf{H}_k$ defined in (8) converges to the identity matrix $\mathbf{I}_{p\times p}$ and for any $\eta > 0$, we have*

$$\mathbb{P}\left(\frac{\|\mathbf{H}_k - \mathbf{I}_{p\times p}\|_F}{\|\mathbf{I}_{p\times p}\|_F} \leq \eta\right) \geq 1 - P_0 \tag{10}$$

where

$$P_0 = \frac{1}{m\,|\mathcal{I}_k|\,\eta^2}\left(\kappa + 1 + p\right). \tag{11}$$

Also, consider the probabilistic model given in (6) and compressive measurements $\mathbf{y}_i = \mathbf{R}_i^T\mathbf{x}_i$. Then, $\mathbf{f}_k$ defined in (8) converges to the center of original data and for any $\eta > 0$, we have

$$\mathbb{P}\left(\frac{\|\mathbf{f}_k - \overline{\mathbf{d}_k}\|_2}{\|\overline{\mathbf{d}_k}\|_2} \leq \eta\right) \geq 1 - P_1 \tag{12}$$

where

$$P_1 = P_0 + \frac{1}{SNR}\Big(P_0 + \frac{1}{|\mathcal{I}_k|\,\eta^2}\Big) \tag{13}$$

and the signal-to-noise ratio is defined as $SNR \triangleq \frac{\|\overline{\mathbf{d}_k}\|_2^2}{\sigma^2}$.

We see that for a fixed error bound η, as $|\mathcal{I}_k|$ increases, the error probability P_0 decreases at rate $\frac{1}{|\mathcal{I}_k|}$. Therefore, for any fixed $\eta > 0$, the error probability P_0 goes to zero as $|\mathcal{I}_k| \to \infty$. Note that the shape of distribution, specified by the kurtosis, is an important factor. For random matrices with heavy-tailed entries, the error probability P_0 increases. However, P_0 gives us an explicit tradeoff between $|\mathcal{I}_k|$, the measurement ratio, and anisotropy in the distribution. For example, the increase in kurtosis can be compensated by increasing $|\mathcal{I}_k|$. The convergence rate analysis for $\mathbf{f}_k$ follows the same path. We further note that P_1 is a decreasing function of the signal-to-noise ratio and as SNR increases, P_1 gets closer to P_0, where for the case that SNR $\to \infty$, then $P_1 \approx P_0$.

Let's consider an example to gain intuition on the choice of random matrices. We are interested in comparing the dense random Gaussian matrices with very sparse random matrices, where each entry is drawn from $\{-1, 0, +1\}$ with probabilities $\{\frac{1}{2s}, 1 - \frac{1}{s}, \frac{1}{2s}\}$ for $s \geq 1$ (we refer to this distribution as a sparse-Bernoulli distribution with parameter s). $\{\mathbf{R}_i\}_{i=1}^n \in \mathbb{R}^{p\times m}$, $p = 100$ and $m/p = 0.3$, are generated with i.i.d. entries both for Gaussian and the sparse-Bernoulli distribution. In Fig. 1, we see that as $|\mathcal{I}_k|$ increases, $\mathbf{H}_k$ gets closer to the identity matrix $\mathbf{I}_{p\times p}$. Also, for a fixed $|\mathcal{I}_k|$, as the sparsity of random matrices increases, the kurtosis $\kappa = s - 3$ increases. Therefore, based on Theorem 2, we expect that the distance between $\mathbf{H}_k$ and $\mathbf{I}_{p\times p}$ increases. Note that for Gaussian and the sparse-Bernoulli with $s = 3$, we have $\kappa = 0$.

As a final note, our theoretical analysis gives us valuable insight about the number of distinct random matrices required. Based on Theorem 2, there is an inherent tradeoff between the accuracy and the number of distinct random matrices used. For example, if we only use one random matrix, we are not

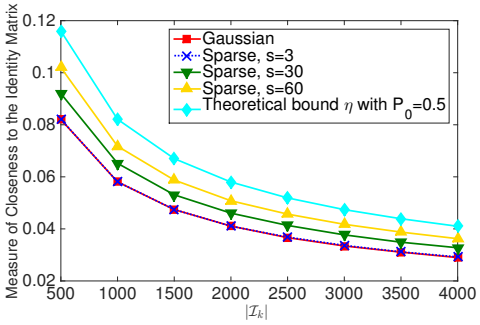


Fig. 1. Closeness of $\mathbf{H}_k$ to $\mathbf{I}_{p\times p}$ defined as $\|\mathbf{H}_k - \mathbf{I}_{p\times p}\|_F / \|\mathbf{I}_{p\times p}\|_F$. $\{\mathbf{R}_i\}_{i=1}^n \in \mathbb{R}^{p\times m}$, $p = 100$ and $m/p = 0.3$, are generated with i.i.d. entries both for Gaussian and the sparse-Bernoulli distribution. We see that as $|\mathcal{I}_k|$ increases, $\mathbf{H}_k$ gets closer to $\mathbf{I}_{p\times p}$. Also, for fixed $|\mathcal{I}_k|$, as the sparsity of random matrices increases, the kurtosis $\kappa = s-3$ increases and consequently the distance between $\mathbf{H}_k$ and $\mathbf{I}_{p\times p}$ increases. For Gaussian and the sparse-Bernoulli with $s = 3$, we have $\kappa = 0$. We also plot the theoretical bound η with $P_0 = 0.5$ for the Gaussian case.

able to recover the true dictionary as observed in [24]. Also, increasing the number of distinct random matrices improves the accuracy, as mentioned in [21]. Hence, we can reduce the number of distinct random matrices in large-scale problems where $n = O(p)$ with controlled loss in accuracy.

IV. Memory and Computation Efficient Dictionary Learning

Now, we return our attention to general dictionary learning. Inspired by the generality of the projection matrices in Theorem 2, we sketch using very sparse random matrices, and furthermore reduce the number of distinct random matrices to increase the efficiency of our approach.

Assume that the original data samples are divided into L blocks $\mathbf{X} = [\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(L)}]$, where $\mathbf{X}^{(l)}$ represents the l^{th} block. Let $\mathbf{R}_l \in \mathbb{R}^{p\times m}$, $m < p$, represent the random matrix used for the l^{th} block. Then, we have

$$\mathbf{Y}^{(l)} = \mathbf{R}_l^T \mathbf{X}^{(l)},\ 1 \le l \le L \tag{14}$$

where $\mathbf{Y}^{(l)}$ is the sketch of $\mathbf{X}^{(l)}$. Each entry of $\{\mathbf{R}_l\}_{l=1}^L$ is distributed on $\{-1, 0, +1\}$ with probabilities $\{\frac{1}{2s}, 1-\frac{1}{s}, \frac{1}{2s}\}$. Here, the parameter s controls the sparsity of random matrices such that each column of $\{\mathbf{R}_l\}_{l=1}^L$ has $\frac{p}{s}$ nonzero entries, on average. We are specifically interested in choosing m and s such that the *compression factor* $\gamma \triangleq \frac{m}{s} < 1$. Thus, the cost to acquire each compressive measurement is $O(\gamma p)$, $\gamma < 1$, vs. the cost for collecting every data entry $O(p)$.

Similarly, we aim to minimize the representation error as

$$\min_{\mathbf{D}\in\mathbb{R}^{p\times K}, \mathbf{C}\in\mathbb{R}^{K\times n}} \sum_{l=1}^{L} \left\|\mathbf{Y}^{(l)} - \mathbf{R}_l^T\mathbf{D}\mathbf{C}^{(l)}\right\|_F^2 s.t. \forall i,\ \left\|\mathbf{c}_i^{(l)}\right\|_0 \le T \tag{15}$$

where $\mathbf{c}_i^{(l)}$ represents the i^{th} sample in the l^{th} block of the coefficient matrix $\mathbf{C}^{(l)}$. As before, the penalty term in (15) is minimized in a simple iterative approach involving two steps. The first step, sparse coding, is the same as the CK-SVD algorithm previously described, except we can take efficient of the block structure and use Batch-OMP [25] in each block which is significantly faster than OMP for each $\mathbf{c}_i$ separately.

A. Dictionary Update

The goal is to update the k^{th} dictionary atom $\mathbf{d}_k$ for $k = 1, \ldots, K$, while assuming that $\mathbf{d}_j$, $j \neq k$, is fixed. The penalty term in (15) can be written as

$$\begin{aligned}
&\sum_{l=1}^{L} \left\|\mathbf{Y}^{(l)} - \mathbf{R}_l^T\mathbf{D}\mathbf{C}^{(l)}\right\|_F^2 = \sum_{l=1}^{L}\sum_{i=1}^{n_l} \left\|\mathbf{y}_i^{(l)} - \mathbf{R}_l^T \sum_{j=1}^{K} c_{i,j}^{(l)}\mathbf{d}_j\right\|_2^2 \\
&= \sum_{l=1}^{L}\sum_{i=1}^{n_l} \left\|\left(\mathbf{y}_i^{(l)} - \mathbf{R}_l^T \sum_{j\neq k} c_{i,j}^{(l)}\mathbf{d}_j\right) - c_{i,k}^{(l)}\mathbf{R}_l^T\mathbf{d}_k\right\|_2^2 \\
&= \sum_{l=1}^{L}\sum_{i=1}^{n_l} \left\|\mathbf{e}_{i,k}^{(l)} - c_{i,k}^{(l)}\mathbf{R}_l^T\mathbf{d}_k\right\|_2^2
\end{aligned} \tag{16}$$

where $c_{i,k}^{(l)}$ is the k^{th} element of $\mathbf{c}_i^{(l)} \in \mathbb{R}^K$, and $\mathbf{e}_{i,k}^{(l)} \in \mathbb{R}^m$ is the representation error for the compressive measurement $\mathbf{y}_i^{(l)}$ when the k^{th} dictionary atom is removed. The objective function in (16) is a quadratic function of $\mathbf{d}_k$ and the minimizer is obtained by setting the derivative of the objective function with respect to $\mathbf{d}_k$ equal to zero. First, let us define $\mathcal{I}_k^{(l)}$ as a set of indices of compressive measurements in the l^{th} block using $\mathbf{d}_k$. Therefore, we get the following expression

$$\mathbf{G}_k\mathbf{d}_k = \mathbf{b}_k,\ \mathbf{G}_k \triangleq \sum_{l=1}^{L} s_k^{(l)}\mathbf{R}_l\mathbf{R}_l^T,\ \mathbf{b}_k \triangleq \sum_{l=1}^{L}\sum_{i\in\mathcal{I}_k^{(l)}} c_{i,k}^{(l)}\mathbf{R}_l\mathbf{e}_{i,k}^{(l)} \tag{17}$$

where $s_k^{(l)}$ is defined as the sum of squares of all the coefficients related to the k^{th} dictionary atom in the l^{th} block, i.e. $s_k^{(l)} \triangleq \sum_{i\in\mathcal{I}_k^{(l)}} (c_{i,k}^{(l)})^2$.

Note that $\mathbf{G}_k$ can be computed efficiently: concatenate $\{\mathbf{R}_l\}_{l=1}^L$ in a matrix $\mathbf{R} \triangleq [\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_L] \in \mathbb{R}^{p\times(mL)}$, and define the diagonal matrix $\mathbf{S}_k$ as

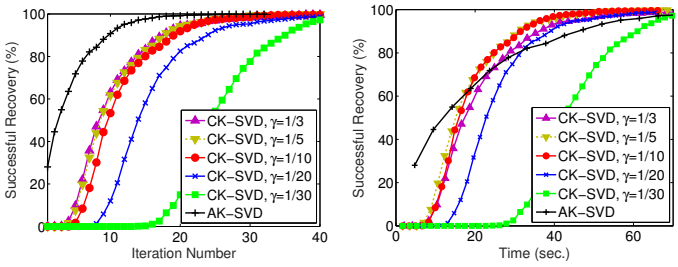
$$\mathbf{S}_k \triangleq \mathrm{diag}\Big(\Big[\underbrace{s_k^{(1)}, \ldots, s_k^{(1)}}_{\text{repeated } m \text{ times}}, \ldots, \underbrace{s_k^{(L)}, \ldots, s_k^{(L)}}_{\text{repeated } m \text{ times}}\Big]\Big) \tag{18}$$

where $\mathrm{diag}(\mathbf{z})$ represents a square diagonal matrix with the elements of vector $\mathbf{z}$ on the main diagonal. Then, we have $\mathbf{G}_k = \mathbf{R}\mathbf{S}_k\mathbf{R}^T$.

Given the updated $\mathbf{d}_k$, the optimal $c_{i,k}^{(l)}$, for all $i \in \mathcal{I}_k^{(l)}$, is given by least squares as $c_{i,k}^{(l)} = \frac{\langle \mathbf{e}_{i,k}^{(l)}, \mathbf{R}_l^T\mathbf{d}_k\rangle}{\|\mathbf{R}_l^T\mathbf{d}_k\|_2^2},\ \forall i \in \mathcal{I}_k^{(l)}$.

V. Experimental Results

We examine the performance of our dictionary learning algorithm on a synthetic dataset. Our proposed method is compared with the fast and efficient implementation of K-SVD known as Approximate K-SVD (AK-SVD) [25] that requires access to the entire data. We generate $K = 15$ dictionary atoms in $\mathbb{R}^p$, $p = 1000$, drawn from the uniform distribution and normalized to have unit norm. A set of data samples $\{\mathbf{x}_i\}_{i=1}^{50,000} \in \mathbb{R}^p$ is generated where each sample



(a) (b)

Fig. 2. Results for synthetic data. Plot of successful recovery vs. (a) iteration number, and (b) time. Our method CK-SVD for varying compression factor γ is compared with AK-SVD. We observe that our method is *both* memory/computation efficient and accurate for $\gamma = \frac{1}{5}$ and $\gamma = \frac{1}{10}$.

is a linear combination of three distinct atoms, i.e. $T = 3$, and the corresponding coefficients are chosen i.i.d. from the Gaussian distribution $\mathcal{N}(0, 100)$. Then, each data is corrupted by Gaussian noise drawn from $\mathcal{N}(0, 0.04\mathbf{I}_{p\times p})$.

CK-SVD is applied on the set of compressive measurements obtained by very sparse random matrices for various values of the compression factor $\gamma = \frac{1}{3}, \frac{1}{5}, \frac{1}{10}, \frac{1}{20}, \frac{1}{30}$. We set the number of blocks $L = 250$ and $m/p = 0.1$. Performance is evaluated by the magnitude of the inner product between learned and true atoms. A value greater than 0.95 is counted as a successful recovery. Fig. 2 shows the results of CK-SVD averaged over 50 independent trials. In practice, when T is small, the updates for $\mathbf{d}_k$ are nearly decoupled, and we may delay updating $c_{i,k}^{(l)}$ until after all K updates of $\mathbf{d}_k$. For $T = 3$, the accuracy results are indistinguishable.

In Fig. 2, we see that our method is able to eventually reach high accuracy even for $\gamma = \frac{1}{30}$, achieving substantial savings in memory/data access. However, there is a tradeoff between memory and computation savings vs. accuracy. Our method is efficient in memory/computation and, at the same time, accurate for $\gamma = \frac{1}{5}$ and $\gamma = \frac{1}{10}$, where it outperforms AK-SVD if the time of each iteration is factored in. We compare with AK-SVD to give an idea of our efficiency, but note that AK-SVD and our CK-SVD are not completely comparable. In our example, both methods reach 100% accuracy eventually but in general they may give different levels of accuracy. The main advantage of CK-SVD appears as the dimensions grow, since then memory/data access is a dominant issue.

ACKNOWLEDGMENT

This material is based upon work supported by the National Science Foundation under Grant CCF-1117775. This work utilized the Janus supercomputer, which is supported by the National Science Foundation (award number CNS-0821794) and the University of Colorado Boulder. The Janus supercomputer is a joint effort of the University of Colorado Boulder, the University of Colorado Denver and the National Center for Atmospheric Research.

REFERENCES

[1] R. Baraniuk, V. Cevher, and M. Wakin, "Low-dimensional models for dimensionality reduction and signal recovery: A geometric perspective," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 959–971, 2010.

[2] M. Elad, M. Figueiredo, and Y. Ma, "On the role of sparse and redundant representations in image processing," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 972–982, 2010.

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

[4] J. Mairal, J. Ponce, G. Sapiro, A. Zisserman, and F. Bach, "Supervised dictionary learning," *Advances in neural information processing systems*, pp. 1033–1040, 2009.

[5] E. Candès and M. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 21–30, 2008.

[6] K. Engan, S. Aase, and J. Husoy, "Method of optimal directions for frame design," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1999, pp. 2443–2446.

[7] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[8] H. Raja and W. Bajwa, "Cloud K-SVD: Computing data-adaptive representations in the cloud," in *51st Annual Allerton Conf. on Communication, Control, and Computing (Allerton)*, 2013, pp. 1474–1481.

[9] M. Davenport, P. Boufounos, M. Wakin, and R. Baraniuk, "Signal processing with compressive measurements," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 445–460, 2010.

[10] J. Fowler, "Compressive-projection principal component analysis," *IEEE Transactions on Image Processing*, vol. 18, no. 10, pp. 2230–2242, 2009.

[11] H. Qi and S. Hughes, "Invariance of principal components under low-dimensional random projection of the data," in *IEEE International Conference on Image Processing (ICIP)*, 2012, pp. 937–940.

[12] F. Pourkamali-Anaraki and S. Hughes, "Efficient recovery of principal components from compressive measurements with application to Gaussian mixture model estimation," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 2314–2318.

[13] ——, "Memory and computation efficient PCA via very sparse random projections," in *Proceedings of the 31st International Conference on Machine Learning (ICML)*, 2014, pp. 1341–1349.

[14] A. Gilbert, J. Park, and M. Wakin, "Sketched SVD: Recovering spectral features from compressive measurements," *arXiv preprint arXiv:1211.0361*, 2012.

[15] G. Atia, "Change detection with compressive measurements," *IEEE Signal Processing Letters*, vol. 22, no. 2, pp. 182–186, 2015.

[16] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning*. Springer, 2009, vol. 2, no. 1.

[17] K. Skretting and K. Engan, "Recursive least squares dictionary learning algorithm," *IEEE Transactions on Signal Processing*, vol. 58, no. 4, pp. 2121–2130, 2010.

[18] F. Pourkamali-Anaraki and S. Hughes, "Compressive K-SVD," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 5469–5473.

[19] T. Do, L. Gan, N. Nguyen, and T. Tran, "Fast and efficient compressive sensing using structurally random matrices," *IEEE Transactions on Signal Processing*, vol. 60, no. 1, pp. 139–154, 2012.

[20] K. Zhang, L. Zhang, and M. Yang, "Fast compressive tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 10, pp. 2002–2015, 2014.

[21] J. Silva, M. Chen, Y. Eldar, G. Sapiro, and L. Carin, "Blind compressed sensing over a structured union of subspaces," arXiv preprint arXiv:1103.2469, Tech. Rep., 2011.

[22] C. Studer and R. Baraniuk, "Dictionary learning from sparsely corrupted or compressed signals," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 3341–3344.

[23] J. Tropp and S. Wright, "Computational methods for sparse solution of linear inverse problems," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 948–958, 2010.

[24] S. Gleichman and Y. Eldar, "Blind compressed sensing," *IEEE Transactions on Information Theory*, vol. 57, no. 10, pp. 6958–6975, 2011.

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