SIAM J. OPTIM.
Vol. 26, No. 4, pp. 2775–2799

© 2016 Society for Industrial and Applied Mathematics

Downloaded 09/15/19 to 75.102.135.243. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

LEARNING SPARSELY USED OVERCOMPLETE DICTIONARIES VIA ALTERNATING MINIMIZATION*

ALEKH AGARWAL[†], ANIMASHREE ANANDKUMAR[‡], PRATEEK JAIN[§], AND PRANEETH NETRAPALLI[§]

Abstract. We consider the problem of sparse coding, where each sample consists of a sparse linear combination of a set of dictionary atoms, and the task is to learn both the dictionary elements and the mixing coefficients. Alternating minimization is a popular heuristic for sparse coding, where the dictionary and the coefficients are estimated in alternate steps, keeping the other fixed. Typically, the coefficients are estimated via ℓ_1 minimization, keeping the dictionary fixed, and the dictionary is estimated through least squares, keeping the coefficients fixed. In this paper, we establish local linear convergence for this variant of alternating minimization and establish that the basin of attraction for the global optimum (corresponding to the true dictionary and the coefficients) is $\mathcal{O}(1/s^2)$, where s is the sparsity level in each sample and the dictionary satisfies restricted isometry property. Combined with the recent results of approximate dictionary estimation, this yields provable guarantees for exact recovery of both the dictionary elements and the coefficients, when the dictionary elements are incoherent.

Key words. dictionary learning, sparse coding, alternating minimization, RIP, incoherence, lasso

AMS subject classifications. 90C26, 68T10

DOI. 10.1137/140979861

1. Introduction.

A sparse code encodes each sample with a sparse set of elements, termed dictionary atoms. Specifically, given a set of samples $Y \in \mathbb{R}^{d\times n}$, the generative model is

$$Y = A^*X^*, \qquad A^* \in \mathbb{R}^{d\times r}, X^* \in \mathbb{R}^{r\times n},$$

and additionally, each column of X^* has at most s nonzero entries. The columns of A^* correspond to the dictionary atoms, and the columns of X^* correspond to the mixing coefficients of each sample. Each sample is a combination of at most s dictionary atoms. Sparse codes can thus succinctly represent high dimensional observed data. The problem of sparse coding consists of unsupervised learning of the dictionary and the coefficient matrices. Thus, given only unlabeled data, we aim to learn the set of dictionary atoms or basis functions that provide a good fit to the observed data. Sparse coding is applied in a variety of domains. Sparse coding of natural images has yielded dictionary atoms which resemble the receptive fields of neurons in the visual cortex [26, 27] and has also yielded localized dictionary elements on speech and video data [19, 25].

*Received by the editors July 29, 2014; accepted for publication (in revised form) September 12, 2016; published electronically December 8, 2016. Part of this work was done when P. Netrapalli was a student at UT Austin and A. Anandkumar and P. Netrapalli were visiting Microsoft Research. An extended abstract containing an earlier version of these results appears in *Proceedings of COLT* 2014.

http://www.siam.org/journals/siopt/26-4/97986.html

Funding: The second author is supported in part by Microsoft Faculty Fellowship, Google Faculty Award, NSF Career Award CCF-1254106, ONR Award N00014-14-1-0665, and AFOSR YIP FA9550-15-1-0221.

[†] Microsoft Research, New York, NY 10023 (alekha@microsoft.com).

[‡] Center for Pervasive Communications and Computing, Electrical Engineering and Computer Science Department, University of California, Irvine, CA 92697 (a.anandkumar@uci.edu).

[§] Microsoft Research, Bangalore, India (prajain@microsoft.com, praneeth@microsoft.com).

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 09/15/19 to 75.102.135.243. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

An important strength of sparse coding is that it can incorporate overcomplete dictionaries, where the number of dictionary atoms r can exceed the observed dimensionality d. It has been argued that having overcomplete representation provides greater flexibility in modeling and more robustness to noise [19], which is crucial for encoding complex signals present in images, speech, and video. It has been shown that the performance of most machine learning methods employed downstream is critically dependent on the choice of data representations, and overcomplete representations are the key to obtaining state-of-the-art prediction results [6].

On the downside, the problem of learning sparse codes or the underlying dictionary is computationally challenging and is, in general, NP-hard [9, 32]. In practice, heuristics are employed based on alternating minimization. At a high level, this consists of alternating steps, where the dictionary is kept fixed and the coefficients are updated and vice versa. Such alternating minimization methods have enjoyed empirical success in a number of settings [18, 10, 2, 20, 37]. In this paper, we carry out a theoretical analysis of the alternating minimization procedure for sparse coding.

1.1. Summary of results.

We consider the alternating minimization procedure where we employ an initial estimate of the dictionary and then use ℓ_1 based minimization for estimating the coefficient matrix, given the dictionary estimate. The dictionary is subsequently reestimated given the coefficient estimates. We establish local convergence to the true dictionary A^* and coefficient matrix X^* for this procedure whenever A^* satisfies restricted isometry property (RIP) for $2s$-sparse vectors. In other words, we characterize the "basin of attraction" for the true solution (A^*, X^*) and establish that alternating minimization succeeds in its recovery when a dictionary is initialized with an error of at most $\mathcal{O}(1/s^2)$, where s is the sparsity level. More precisely, the initial dictionary estimate $A(0)$ is required to satisfy

$$\epsilon_0 := \max_{i\in[r]} \min_{z\in\{-1,+1\}} \left\|zA_i^* - A(0)_i\right\|_2 = \mathcal{O}\left(\frac{1}{s^2}\right),$$

where A_i^* represents ith column of A^*.

Further, when the sparsity level satisfies $s = \mathcal{O}(d^{1/6})$ and the number of samples satisfies $n = \mathcal{O}(r^2)$, we establish a linear rate of convergence for the alternating minimization procedure to the true dictionary even when the dictionary is overcomplete $(r \geq d)$.

Note that our results assume RIP as compared to most other results for this problem, which assume incoherence. Though all incoherent matrices are s-RIP for $s < \mathcal{O}(\sqrt{d})$, there are many s-RIP matrices that are not incoherent. Consider, for example, a dictionary where each row of the dictionary matrix is sampled from a Gaussian distribution $\mathcal{N}(0, \Sigma)$, where $\Sigma_{ii} = 1 \forall i$, $\Sigma_{12} = \Sigma_{21} = \delta/4$, and all other elements of Σ are zero. The resulting dictionary is not incoherent when δ is larger than $\mathcal{O}(1/\sqrt{d})$, but it is always 2-RIP with RIP constant δ. More generally, if we have nonzero off-diagonal elements, the resulting dictionary will not be incoherent while still satisfying RIP. The key difficulty is that a constant value suffices in RIP, while incoherence requires the inner products of dictionary elements to go down as $1/\sqrt{d}$. See Lemma 24 in Appendix A for more details.

For the case of incoherent dictionaries, by combining the above result with recent results on approximate dictionary estimation by Agarwal, Anandkumar, and Netrapalli [1] or Arora et al. [3], we guarantee exact recovery of the true solution (A^*, X^*) when the alternating procedure is initialized with the output of [1] or [3]. If we employ the procedure of Agarwal, Anandkumar, and Netrapalli [1], the overall

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 09/15/19 to 75.102.135.243. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

requirements are as follows: the sparsity level is required to be $s = \mathcal{O}(\frac{d^{1/4}}{\sqrt{\mu_0}}, d^{1/9}, r^{1/8})$, where μ_0 is the incoherence parameter and the number of samples $n = \mathcal{O}(r^2)$ to guarantee exact recovery of the true solution. If we employ the procedure of Arora et al. [3] (in particular their OverlappingAverage procedure), we can establish exact recovery assuming $s = \mathcal{O}(r^{1/6}, \sqrt{d}/\mu_0)$.

Our results are also stable to the presence of noise. Indeed a simple perturbation argument shows that as long as the noise in each sample has ℓ_2-norm at most ϵ, our results apply and yield an error of at most ϵ in the estimation of the dictionary. More concretely, we can use noise robust results for compressed sensing, e.g., [11], and a robust version of Lemma 6 and the lemmas within that to obtain such a guarantee.

1.2. Related work.

Analysis of local optima of nonconvex programs for sparse coding. Gribonval and Schnass [14], Geng, Wang, and Wright [12] and Gribonval, Jenatton, and Bach [13] carry out a theoretical analysis and study the conditions under which the true solution turns out to be a local optimum of a nonconvex optimization problem for dictionary recovery. Gribonval and Schnass [14] and Geng, Wang, and Wright [12] both consider the noiseless setting and analyze the following nonconvex program:

$$\min \|X\|_1 \qquad \text{s.t. } Y = AX,\ \|A_i\|_2 = 1,\ \forall\, i \in [r]. \tag{1}$$

Since A and X are both unknown, the constraint $Y = AX$ is nonconvex. It is natural to expect the true solution (A^*, X^*) to be a local optimum for (1) under fairly mild conditions, but this turns out to be nontrivial to establish. The difficulties arise from the nonconvexity of the problem and the presence of sign-permutation ambiguity which leads to exponentially many equivalent solutions obtained via sign change and permutation. Gribonval and Schnass [14] established that (A^*, X^*) is a local optimum for (1) but limited to the case where the dictionary matrix A is square and, hence, did not incorporate the overcomplete setting. Geng, Wang, and Wright [12] extend the analysis to the overcomplete setting and establish that the true solution is a local optimum of (1) with hight probability (w.h.p.) for incoherent dictionaries, when the number of samples n and sparsity level s scale as

$$n = \Omega\left(\|A\|_2^4 r^3 s\right), \quad s = \mathcal{O}(\sqrt{d}/\mu_0). \tag{2}$$

In our setting, where the spectral norm is assumed to be $\|A\|_2 < \mu_1\sqrt{r/d}$, for some constant $\mu_1 > 0$, the sample complexity simplifies as $n = \Omega\left(r^5 s/d^2\right)$. Gribonval, Jenatton, and Bach [13] consider the noisy setting and analyze the modified nonconvex program involving ℓ_1 penalty for the coefficient matrix and ℓ_2 penalty for the loss in fitting the samples, and establish that the true solution is in the neighborhood of a local optimum of the modified nonconvex program w.h.p. when the number of samples scales as $n = \Omega\left(\|A\|_2^2 r^3 d s^2\right)$. In our setting, this reduces to $n = \Omega\left(r^4 s^2\right)$. A similar local analysis is carried out by Schnass [29] for the K-SVD algorithm. There are significant differences between the above works and ours. While these works establish that (A^*, X^*) is a local optimum of a nonconvex program, they do not provide a tractable algorithm to reach this particular solution as opposed to another local optimum. In contrast, we establish guarantees for a simple alternating minimization algorithm and explicitly characterize the "basin of attraction" for the true solution (A^*, X^*). This provides precise initialization conditions for the alternating minimization to succeed. Moreover, our sample complexity requirements are much weaker and we require only $n = \mathcal{O}(r^2)$ samples for our guarantees to hold.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 09/15/19 to 75.102.135.243. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Alternating minimization for sparse coding. Our analysis in this paper provides a theoretical explanation for the empirical success of alternating minimization, observed in a number of works [18, 10, 2, 20, 37]. These methods are all based on alternating minimization but differ mostly in how they update the dictionary elements. For instance, Lee et al. carry out least squares for updating the dictionary [18] similar to the method of optimal directions [10], while the K-SVD procedure [2] updates the dictionary estimate using a spectral procedure on the residual. However, none of the previous works provide theoretical guarantees on the success of the alternating minimization procedure for sparse coding.

Guaranteed dictionary estimation. Some of the recent works provide theoretical guarantees on the estimation of the true dictionary. Spielman, Wang, and Wright [30] establish exact recovery under ℓ_1 based optimization when the true dictionary A^* is a basis, which rules out the overcomplete setting. Agarwal, Anandkumar, and Netrapalli [1] and Arora et al. [3] propose methods for approximate dictionary estimation in the overcomplete setting. At a high level, both their methods involve a clustering-based approach for finding samples which share a dictionary element and then using the subset of samples to estimate a dictionary element. Agarwal, Anandkumar, and Netrapalli [1] establish exact recovery of the true solution (A^*, X^*) under a "one-shot" lasso procedure, when the nonzero coefficients are Bernoulli $\{-1, +1\}$ (or more generally discrete). On the other hand, we assume only mild conditions on the nonzero elements. Arora et al. [3] consider an alternating minimization procedure. However, a key distinction is that their analysis requires *fresh* samples in each iteration, while we consider the same samples for all the iterations. We show *exact* recovery using $n = \Omega(r^2)$ samples, while [3] can only establish that the error is bounded by $\exp[-O(n/r^2)]$. Furthermore, both the above papers [3, 1] assume that the dictionary elements are mutually incoherent, allowing the use of simpler procedures than ℓ_1 minimization for dictionary estimation. Our local convergence result in this paper assumes only that the dictionary matrix satisfies RIP (which is strictly weaker than incoherence). For the case of incoherent dictionaries, we can employ the procedures of [1] or [3] for initializing the alternating procedure and obtain overall guarantees in such scenarios.

Other works on sparse coding. Some of the other recent works are only tangentially related to this paper. For instance, the works [34, 22, 21, 31] provide generalization bounds for predictive sparse coding, without computational considerations, which differs from our generative setting here and algorithmic considerations. Parametric dictionary learning is considered in [36], where the data is fitted to dictionaries with small coherence. Note that we provide guarantees when the underlying dictionary is incoherent but do not constrain our method to produce an incoherent dictionary. The problem of sparse coding is also closely related to the problem of blind source separation, and we refer the reader to [1] for an extended survey of these works.

Majorization-minimization algorithms for biconvex optimization. Beyond the specific problem of sparse coding, alternating optimization procedures more generally are a natural fit for biconvex optimization problems, where the objective is individually convex in two sets of variables but not jointly convex. Perhaps the most general study of these problems has been carried out in the framework of majorization-minimization schemes [17], or under the name of the EM algorithm in statistics literature. In this generality, the strongest result one can typically provide is a convergence guarantee to a local optimum of the problem. When the biconvex objective is defined over probability measures, Csiszar presents a fairly general set of conditions on the objective function, under which linear convergence to the global optimum is guaranteed (see,

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 09/15/19 to 75.102.135.243. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Algorithm 1. AltMinDict$(Y, A(0), \epsilon_0)$: Alternating minimization for dictionary learning.

```
Input: Samples $Y$, initial dictionary estimate $A(0)$, accuracy sequence $\epsilon_t$, and sparsity level $s$. Thresholding function $\mathrm{T}_\rho(a) = a$ if $|a| > \rho$ and 0 otherwise.
1: for iterations $t = 0, 1, 2, \ldots, T-1$ do
2:     for samples $i = 1, 2, \ldots, n$ do
3:         $X(t+1)_i = \arg\min_{x \in \mathbb{R}^r} \|x\|_1$   such that, $\|Y_i - A(t)x\|_2 \le \epsilon_t$
4:     end for
5:     Threshold: $X(t+1) = X(t+1) \cdot *(\mathbb{I}[X(t+1) > 9s\epsilon_t])$
6:     Estimate $A(t+1) = YX(t+1)^+$
7:     Normalize: $A(t+1)_i = \frac{A(t+1)_i}{\|A(t+1)_i\|_2}$
8: end for
Output: $A(T)$
```

e.g., the recent tutorial [8] for an excellent overview). However, these conditions do not seem to easily hold in the context of dictionary learning. Alternating optimization in related contexts has also been studied in a variety of matrix factorization problems such as low-rank matrix completion and nonnegative matrix factorization. Perhaps the most related to our work are similar results for low-rank matrix completion problems by Jain, Netrapalli, and Sanghavi [15].

Notation. Let $[n] := \{1, 2, \ldots, n\}$. For a vector v or a matrix W, we will use the shorthand $\mathrm{Supp}(v)$ and $\mathrm{Supp}(W)$ to denote the set of nonzero entries of v and W, respectively. $\|w\|_p$ denotes the ℓ_p norm of vector w; by default, $\|w\|$ denotes the ℓ_2 norm of w. $\|W\|_2$ denotes the spectral norm (largest singular value) of matrix W. $\|W\|_\infty$ denotes the largest element (in magnitude) of W. For a matrix X, X^i, X_i and X^i_j denote the ith row, ith column, and (i, j)th element of X, respectively. Using the above notation, for a square matrix M, $M_i^{\setminus i}$ denotes the ith column of the restriction of M to its off-diagonal entries. We will abuse notation to refer to it as the off-diagonal elements of the ith column of M.

2. Algorithm.

Given an initial estimate of the dictionary, we alternate between two procedures, viz., a sparse recovery step for estimating the coefficients given a dictionary, and a least squares step for a dictionary given the estimates of the coefficients. The details of this approach are presented in Algorithm 1.

The sparse recovery step of Algorithm 1 is based on ℓ_1-regularization, followed by thresholding. The thresholding is required for us to guarantee that the support set of our coefficient estimate $X(t)$ is a *subset* of the true support w.h.p. Once we have an estimate of the coefficients, the dictionary is reestimated through least squares. The overall algorithmic scheme is popular for dictionary learning, and there are a number of variants of the basic method. For instance, the ℓ_1-regularized problem in step 3 can also be replaced by other robust sparse recovery procedures such as OMP [33] or GraDeS [11]. More generally the exact lasso and least squares steps may be replaced with other optimization methods for computational efficiency, e.g., [16].

3. Main results and their proofs.

In this section, we provide our local convergence result for alternating minimization and also clearly specify all the required assumptions on A^* and X^*. We provide a brief sketch of our proof for each of the steps in section 3.4.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 09/15/19 to 75.102.135.243. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

3.1. Assumptions. We start by formally describing the assumptions needed for the main recovery result of this paper. Without loss of generality, assume that all the elements are normalized: $\|A^*_i\|_2 = 1$ for $i \in [r]$. This is because we can always rescale the dictionary elements and the corresponding coefficients and obtain the same observations.

Assumptions.

(A1) Dictionary matrix satisfying RIP. The dictionary matrix A^* has a $2s$-RIP constant of $\delta_{2s} < 0.1$.

(A2) Spectral condition on dictionary elements. The dictionary matrix has bounded spectral norm for some constant $\mu_1 > 0$, $\|A^*\|_2 < \mu_1\sqrt{\frac{r}{d}}$.

(A3) Nonzero entries in coefficient matrix. We assume that the nonzero entries of X^* are drawn independent and identically distributed (i.i.d.) from a distribution such that $\mathbb{E}\left[({X^*}^i_j)^2\right] = 1$ and satisfy the following a.s.: $|{X^*}^i_j| \leq M \forall i, j$.

(A4) Sparse coefficient matrix. The columns of coefficient matrix have s nonzero entries which are selected uniformly at random from the set of all s-sized subsets of $[r]$, i.e., $|\operatorname{Supp}(X^*_i)| = s \;\forall\, i \in [n]$. We require s to satisfy $s < \frac{d^{1/6}}{c_2 \mu_1^{1/3}}$ for some universal constant c_2.

(A5) Sample complexity. For some universal constant $c > 0$ and a given failure parameter $\delta > 0$, the number of samples n needs to satisfy

$$n \geq c_3\, r^2 M^2 \log \frac{2r}{\delta},$$

where $c_3 > 0$ is a universal constant.

(A6) Initial dictionary with guaranteed error bound. We assume that we have access to an initial dictionary estimate $A(0)$ such that

$$\widehat{\epsilon_0} := \max_{i \in [r]} \min_{z \in \{-1,+1\}} \|zA_i(0) - A^*_i\|_2 < \frac{1}{2592s^2}.$$

(A7) Choice of parameters for alternating minimization. Algorithm 1 uses a sequence of accuracy parameters $\epsilon_0 = 1/2592s^2$ and

$$\epsilon_{t+1} = \frac{25050\mu_1 s^3}{\sqrt{d}}\epsilon_t. \tag{3}$$

Assumption (A1) regarding the RIP assumption is crucial in establishing our guarantees, since it is critical for analyzing the performance of the compressed sensing subroutine in Algorithm 1 (steps 2–5). It is possible to further weaken this assumption to a restricted eigenvalue (RE) condition which is often used in the sparse regression literature as well [28, 23]. We will present a more detailed discussion of this condition in the proof sketch. In order to keep the results with cleaner constants, we will continue with the RIP assumption for the rest of the analysis, while mentioning how the result can be extended easily under a more general RE assumption.

The assumption (A2) provides a bound on the spectral norm of A^*. Note that the RIP and spectral assumptions are satisfied w.h.p. when the dictionary elements are randomly drawn from a mean-zero sub-Gaussian distribution.

Assumption (A3) imposes some natural constraints on the nonzero entries of X^*. Some distributional assumption like assumption (A4) on sparsity in the coefficient matrix is crucial for identifiability of the dictionary learning problem. While we assume completely random supports and i.i.d. coefficients in this analysis, qualitatively

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 09/15/19 to 75.102.135.243. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

similar results continue to hold as long as the first and second moments of the coefficient distribution have upper and lower bounds of the same order (that is, $\mathcal{O}(s/r)$ and $\mathcal{O}(s^2/r^2)$, respectively). This is intuitive since without such assumptions, two dictionary elements can be so correlated in their occurrence that they cannot be disentangled, or one dictionary element might almost never occur. However, we will continue to make the completely random support assumption here since it captures most of the essential intuition and allows us to easily combine with the initialization results in what follows.

Assumption (A5) provides a bound on sample complexity. Assumption (A6) specifies the accuracy of the initial estimate required by Algorithm 1. Recent works [4, 1] provide provable ways of obtaining such an estimate. See section 3.3 for more details.

Assumption (A7) specifies the choice of accuracy parameters used by the alternating method in algorithm 1. Due to assumption (A4) on sparsity level s, we have that $\frac{25050\mu_1 s^3}{\sqrt{d}} < 1/2$ and the accuracy parameters in (3) form a decreasing sequence. This implies that in Algorithm 1, the accuracy constraint becomes more stringent with the iterations of the alternating method.

3.2. Guarantees for alternating minimization.

We now prove a local convergence result for alternating minimization. We assume that we have access to a good initial estimate of the dictionary:

THEOREM 1 (local linear convergence). *Under assumptions* (A1)–(A7)*, with probability at least* $1 - 2\delta$ *the iterate* $A(t)$ *of Algorithm* 1 *satisfies the following* $\forall\ t \geq 1$*:*

$$\min_{z\in\{-1,1\}} \|zA_i(t) - A_i^*\|_2 \leq \epsilon_t, 1 \leq i \leq r.$$

Remarks. Note that we have a sign ambiguity in recovery of the dictionary elements, since we can exchange the signs of the dictionary elements and the coefficients to obtain the same observations.

Theorem 1 guarantees that we can recover the dictionary A^* to an arbitrary precision ϵ (based on the number of iterations T of Algorithm 1), given $n = \mathcal{O}(r^2)$ samples. We contrast this with the results of [4], which also provides recovery guarantees to an arbitrary accuracy ϵ, but only if the number of samples is allowed to increase as $\mathcal{O}(r^2 \log 1/\epsilon)$.

The consequences of Theorem 1 are powerful combined with our assumption (A4) and the recurrence (3) (since (A4) ensures that ϵ_t forms a decreasing sequence). In particular, it is implied that w.h.p we obtain

$$\min_{z\in\{-1,1\}} \|zA_i(t) - A^*{}_i\|_2 \leq \widehat{\epsilon_0} 2^{-t}.$$

Given the above bound, we need at most $\mathcal{O}(\log_2 \frac{\widehat{\epsilon_0}}{\epsilon})$ iterations in order to ensure $\|zA_i(T) - A^*{}_i\|_2 \leq \epsilon\ \forall$ the dictionary elements $i = 1, 2, \ldots, r$. In the convex optimization parlance, the result demonstrates a local linear convergence of Algorithm 1 to the globally optimal solution under an initialization condition. Another way of interpreting our result is that the global optimum has a *basin of attraction* of size $\mathcal{O}(1/s^2)$ for our alternating minimization procedure under these assumptions (since we require $\widehat{\epsilon_0} \leq \mathcal{O}(1/s^2)$).

We also recall that the lasso step in Algorithm 1 can be replaced with a different robust sparse recovery procedure, with qualitatively similar theorems.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 09/15/19 to 75.102.135.243. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

3.3. Using local convergence for complete recovery. In the above section, we showed a local convergence result for Algorithm 1. In particular, assumption (A6) requires that the initial dictionary estimate be at most $\mathcal{O}(\frac{1}{s^2})$ away from A^*. In this section, we use the recent result of [1] to obtain an initialization which satisfies assumption (A6), and thus we obtain a full recovery result for the sparsely used dictionary problem with assumptions only on the model parameters. In order to obtain the initialization from the method of [1], we require the following assumptions:

(B1) Incoherent dictionary Elements. Without loss of generality, assume that all the elements are normalized: $\|A_i^*\|_2 = 1$ for $i \in [r]$. We assume a pairwise incoherence condition on the dictionary elements for some constant $\mu_0 > 0$, $|\langle A_i^*, A_j^* \rangle| < \frac{\mu_0}{\sqrt{d}}$.

(B3) Nonzero entries in coefficient matrix. We assume that the nonzero entries of X^* are drawn i.i.d. from a zero-mean distribution such that $\mathbb{E}\left[({X^*}^i_j)^2\right] = 1$ and satisfy the following a.s.: $m \le |{X^*}^i_j| \le M \forall i, j$.

(B4) Sparse coefficient matrix. The columns of the coefficient matrix have a bounded number of nonzero entries s which are selected randomly, i.e.,

$$|\operatorname{Supp}(x_i)| = s \quad \forall\, i \in [n]. \tag{4}$$

We require s to be

$$s < c_1 \min\left(\frac{m}{M} \frac{d^{1/4}}{\sqrt{\mu_0}}, \left(\frac{d}{\mu_1^2} \frac{m^4}{M^4} \right)^{1/9}, r^{1/8} \left(\frac{m}{M} \right)^{1/4} \right)$$

for universal constants $c_1 > 0$. Constants m, M are as specified above.

(B5) Sample complexity. Given universal constant $c_2 > 0$, choose $\delta > 0$ and the number of samples n such that

$$n := n(d, r, s, \delta) \ge c_2\, r^2 \frac{M^2}{m^2} \log \frac{2r}{\delta}.$$

We note that assumption (B3) also requires the entries in the coefficient matrix to be lower bounded and mean-zero, in addition to (A3).

Theorem 2 (specialization of Theorem 2.1 from [1]). *Under assumptions* (B1), (A2)*,* (B3)−−(B5)*, and* (A7)*, there exists an algorithm which given* Y *outputs* $A(0)$*, such that assumption* (A6) *holds with probability greater than* $1 - 2n^2\delta$*.*

The restatement follows by setting $\alpha = s^{-9/2} \frac{m^2}{M^2}$ in that result which ensures that the error in the initialization is at most $1/s^2$ as required by assumption (A6). Combining the above theorem with Theorem 1 gives the following powerful corollary.

Corollary 3 (exact recovery). *Suppose assumptions* (B1), (A2)-(A5), (B3)-(B5)*, and* (A7) *hold. If we start Algorithm* 1 *with the output of Algorithm* 1 *of* [1]*, then the following holds* $\forall\ t \ge 1$*:*

$$\min_{z \in \{-1,1\}} \|z A_i(t) - A_i^*\|_2 \le \epsilon_t, 1 \le i \le r.$$

The above result makes use of Lemma 21 in the appendix, which shows that assumptions (B1) and (B4) imply (A1). Note that the above corollary gives an exact recovery result with the only assumptions being those on the model parameters. We also note that the conclusion of Corollary 3 does not crucially rely on initialization specifically by the output of Algorithm 1 of [1] and admits any other initialization

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 09/15/19 to 75.102.135.243. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

satisfying assumption (A6). As remarked earlier, the recent work of [4] provides an alternative initialization strategy for our alternating minimization procedure. Indeed, under our sample complexity assumption, their OverlappingAverage method provides a solution with $\widehat{\epsilon_0} = \mathcal{O}(s/\sqrt{r})$ assuming $s = \mathcal{O}(\max(r^{2/5}, \sqrt{d}))$. In particular, if $s = \mathcal{O}(r^{1/6})$, we obtain the desired initial error of $1/s^2$ using that algorithm. The sample complexity of the entire procedure remains identical to that in assumption (A5).

3.4. Overview of proof. In this section we outline the key steps in proving Theorem 1.

For ease of notation, let us consider just one iteration of Algorithm 1 and denote $X(t+1)$ as X, YX^+ as A, and $A(t)$ as $\widetilde{A}$. Note that A is $A(t+1)$ before column normalization. Then we have the least squares update

$$\begin{aligned} A - A^* &= YX^+ - A^* \\ &= A^*X^*X^+ - A^*XX^+ = A^*\triangle XX^+, \end{aligned}$$

where $\triangle X = X^* - X$. This means that we can understand the error in dictionary recovery by the error in the least squares operator $\triangle XX^+$. In particular, we can further expand the error in a column p as

$$A_p - A^*{}_p = A^*{}_p(\triangle XX^+)_p^p + A^*{}_{\backslash p}(\triangle XX^+)_p^{\backslash p},$$

where the notation $\backslash p$ represents the collection of all indices apart from p, i.e., $A^*{}_{\backslash p}$ denote all the columns of A except the pth column and $(\triangle XX^+)_p^{\backslash p}$ denotes the off-diagonal elements of the pth column of $(\triangle XX^+)$. The above equation indicates that there are two sources of error in our dictionary estimate. The element $(\triangle XX^+)_p^p$ causes the rescaling of A_p relative to $A^*{}_p$. However, this is a minor issue and we will show that renormalization will correct it.

More serious is the contribution from the off-diagonal terms $(\triangle XX^+)_{\backslash p}^p$, which corrupt our estimate A_p with other dictionary elements beyond $A^*{}_p$. Indeed, a crucial argument in our proof is controlling the contribution of these terms at an appropriately small level. In order to do that, we start by controlling the magnitude of $\triangle X$.

Lemma 4 (error in sparse recovery). *Let $\triangle X := X(t) - X^*$. Assume that $2\mu_0 s/\sqrt{d} \leq 0.1$ and $\sqrt{s\epsilon_t} \leq 0.1$. Then, we have* $\mathrm{Supp}(\triangle X) \subseteq \mathrm{Supp}(X^*)$ *and the error bound $\|\triangle X\|_\infty \leq 9s\epsilon_t$.*

In particular, our assumptions (A6) and (A7) ensure that $9s\epsilon_t \leq \frac{1}{288s}$ for every t which will be used in what follows.

More general RE conditions. The above lemma is the only part of our proof where we require the RIP assumption. This is in order to invoke the result of Candes [7] regarding the error in compressed sensing with (bounded) deterministic noise. Such results can also be typically established under weaker RE assumptions. Given a vector $v \in \mathbb{R}^r$, these RE conditions study the norms $\|Av\|_2^2$. A particular form of RE condition for these approximately sparse vectors then posits (see, e.g., [28, 23])

$$\|Av\|_2 \geq \gamma\|v\|_2 - \tau\|v\|_1. \tag{5}$$

Under such a condition, it can be readily shown that Lemma 4 continues to hold with an error bound which is $\mathcal{O}(s\epsilon_t/(\gamma - s\tau))$. For many random matrix ensembles for A,

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 09/15/19 to 75.102.135.243. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

it is known that $\tau = \mathcal{O}(\sqrt{(\log r)/d})$, which means that $\gamma - s\tau$ will be bounded away from zero under our assumptions on the sparsity level s. These RE conditions are the weakest known conditions under which compressed sensing using efficient procedures is possible, and we employ this as a subroutine in our alternating minimization procedure.

The next lemma is very useful in our error analysis, since we establish that any matrix W satisfying $\mathrm{Supp}(W) \subseteq \mathrm{Supp}(X^*)$ has a good bound on its spectral norm (even if the entries depend on A^*, X^*).

LEMMA 5. *With probability at least* $1 - r\exp\left(-\frac{Cn}{rs}\right)$, *for every* $r \times n$ *matrix* W *s.t.* $\mathrm{Supp}(W) \subseteq \mathrm{Supp}(X^*)$, *we have*

$$\|W\|_2 \le 2\|W\|_\infty \sqrt{\frac{s^2 n}{r}}.$$

A particular consequence of this lemma is that it guarantees the invertibility of the matrix $XX^\top$, so that the pseudoinverse X^+ is well-defined for subsequent least squares updates. Next, we present the most crucial step, which is controlling the off-diagonal terms $(\triangle XX^+)^p_{\backslash p}$.

LEMMA 6 (off-diagonal error bound). *With probability at least* $1 - Cr^2 \exp\left(-\frac{Cn}{r^2M^2}\right)$, *we have uniformly for every* $p \in [r]$ *and every* $\triangle X$ *such that* $\|\triangle X\|_\infty < \frac{1}{288s}$,

$$\left\| \left(\triangle XX^+\right)^{\backslash p}_p \right\|_2 = \left\| \left(X^* X^+\right)^{\backslash p}_p \right\|_2 \le \frac{1968 s^2 \|\triangle X\|_\infty}{\sqrt{r}}.$$

The lemma uses the earlier two lemmas along with a few other auxiliary results. Given these lemmas, the proof of the main theorem follows using basic linear algebra arguments. Specifically, for any unit vector w such that $w \perp {A^*}_p$, we can bound the normalized inner product $\langle w, A_p\rangle / \|A_p\|_2$ which suffices to obtain the result of the theorem.

3.5. Detailed proof of Theorem 1. We now provide a proof of the theorem using the above given lemmas. The proofs of the lemmas are deferred to the appendix. Recall that we denote $A(t)$ as $\widetilde{A}$ and $A(t+1)$ as A. Similarly we denote $X(t)$ and $X(t+1)$ as $\widetilde{X}$ and X, respectively. Then the goal is to show that A is closer to A^* than $\widetilde{A}$. For the purposes of our analysis, we will find it more convenient to directly work with dot products instead of ℓ_2-distances (and hence avoid sign ambiguities). With this motivation, we define the following notion of distance between two vectors.

DEFINITION 7. *For any two vectors* $z, w \in \mathbb{R}^d$, *we define the distance between them as follows:*

$$dist(z, w) := \sup_{v \perp w} \frac{\langle v, z\rangle}{\|v\|_2 \|z\|_2} = \sup_{v \perp z} \frac{\langle v, w\rangle}{\|v\|_2 \|w\|_2}.$$

This definition of distance suffices for our purposes due to the following simple lemma.

LEMMA 8. *For any two unit vectors* $u, v \in \mathbb{R}^d$, *we have*

$$dist(u, v) \le \min_{z \in \{-1,1\}} \|zu - v\|_2 \le \sqrt{2}\, dist(u, v).$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 09/15/19 to 75.102.135.243. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Proof. The proof is rather straightforward. Suppose that $\langle u, v\rangle > 0$ so that the minimum happens at $z = 1$. The other case is identical. We can easily rewrite

$$\|u - v\|_2^2 = (2 - 2\langle u, v\rangle) \le 2(1 - \langle u, v\rangle^2),$$

where the final inequality follows since $0 \le \langle u, v\rangle \le 1$. Writing $u = \langle u, v\rangle v + v_\perp$, where $\langle v_\perp, v\rangle = 0$, we see that

$$1 = \|u\|_2^2 = \langle u, v\rangle^2 + \|v_\perp\|^2 = \langle u, v\rangle^2 + dist(u, v)^2.$$

Substituting this into our earlier bound, we obtain the upper bound. For the lower bound, we note that

$$\begin{aligned} dist(u, v)^2 &= 1 - \left(1 - \frac{\|u - v\|^2}{2}\right)^2 \\ &\le \|u - v\|^2. \end{aligned}$$

□

The distance is naturally extended to matrices for our purposes by applying it columnwise.

Definition 9. *For any two $d \times r$ matrices Z and W, we define the distance between them as follows:*

$$dist(Z, W) := \sup_{p\in[r]} dist(Z_p, W_p).$$

Note that the normalization in the definition of $dist(z, w)$ ensures that we can apply the distance directly to the result of the least squares step without worrying about the effects of normalization. This allows us to work with a closed-form expression for A

$$A = Y X^+ = A^* X^* X^+. \tag{6}$$

Since $dist(\cdot, \cdot)$ is invariant to scaling, any bound we obtain for A also applies to $A(t+1)$ after normalization. We are now in a position to prove Theorem 1.

Proof of Theorem 1. As an induction hypothesis, we have $dist(\widetilde{A}, A^*) < \frac{\epsilon_t}{\sqrt{2}}$, where we recall the definition (3). We will show that for every $p \in [r]$, we will have

$$dist(A, A^*) \le \frac{23616\mu_1 s^3}{\sqrt{d}}\epsilon_t \le \frac{\epsilon_{t+1}}{\sqrt{2}}. \tag{7}$$

This suffices to prove the theorem by appealing to Lemma 8.

Furthermore, under the inductive hypothesis, Lemma 4 guarantees that $\|\triangle X\|_\infty \le 9s\epsilon_t$, which is at most $\frac{1}{288s}$ by assumption (A6). Consequently Lemma 6 can be applied under the inductive hypothesis. Fix any $w \perp {A^*}_p$ such that $\|w\|_2 = 1$. We first provide a bound on $\langle w, A_p\rangle$. We have w.h.p.

$$\begin{aligned} \langle w, A_p\rangle = w^\top A^* X^* {X^+}_p &\overset{(\zeta_1)}{\le} \left\|w^\top A^*\right\|_2 \left\|\left(X^* X^+\right)_p^{\setminus p}\right\|_2 \\ &\overset{(\zeta_2)}{\le} \mu_1 \sqrt{\frac{r}{d}} \cdot \frac{1968 s^2 \|\triangle X\|_\infty}{\sqrt{r}} \\ &= \frac{17712\mu_1 s^3}{\sqrt{d}}\epsilon_t, \end{aligned} \tag{8}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 09/15/19 to 75.102.135.243. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

where (ζ_1) follows from the fact that $w^\top A^*{}_p = 0$ and (ζ_2) follows from assumption (A2) and Lemma 6.

In order to bound $dist(A, A^*)$, it remains to show a lower bound on $\|A_p\|_2$. This again follows using basic algebra, given our main lemmas.

$$\begin{aligned} \|A_p\|_2 = \left\|A^* X^* X^+{}_p\right\|_2 &= \left\|A^* \left(X - \triangle X\right) X^+{}_p\right\|_2 \\ &\overset{(\zeta_1)}{=} \left\|A^*{}_p - A^* \triangle X X^+{}_p\right\|_2 \\ &\geq \|A^*{}_p\|_2 - \left\|A^* \left(\triangle X X^+\right)_p\right\|_2, \end{aligned}$$

where (ζ_1) follows from the fact that $XX^+ = \mathbb{I}$. We decompose the second term into diagonal and off-diagonal terms of $\triangle XX^+$, followed by the triangle inequality, and obtain

$$\begin{aligned} \|A_p\|_2 &\geq 1 - \left\|A^*{}_p \left(\triangle X X^+\right)_p^p + A^*{}_{\backslash p} \left(\triangle X X^+\right)_p^{\backslash p}\right\|_2 \\ &\geq 1 - \|A^*{}_p\|_2 \left|\left(\triangle X X^+\right)_p^p\right| - \|A^*{}_{\backslash p}\|_2 \left\|\left(\triangle X X^+\right)_p^{\backslash p}\right\|_2 \\ &\geq 1 - 1 \cdot \left\|\triangle X X^\top \left(X X^\top\right)^{-1}\right\|_2 - \|A^*{}_{\backslash p}\|_2 \left\|\left(\triangle X X^+\right)_p^{\backslash p}\right\|_2 \\ &\geq 1 - \underbrace{\|\triangle X\|_2 \left\|X^\top\right\|_2 \left\|\left(X X^\top\right)^{-1}\right\|_2}_{\mathcal{T}_1} - \underbrace{\|A^*{}_{\backslash p}\|_2 \left\|\left(\triangle X X^+\right)_p^{\backslash p}\right\|_2}_{\mathcal{T}_2}. \end{aligned}$$

It remains to control $\mathcal{T}_1$ and $\mathcal{T}_2$ at an appropriate level. We start from $\mathcal{T}_1$. Note that $\|\triangle X\|_2$ is bounded by Lemmas 4 and 5, while $\left\|X^\top\right\|_2$ is controlled by Lemma 17 (recall $\|\triangle X\|_\infty \leq 1/(288s)$). Invoking Lemma 17 to control $\|(XX^\top)^{-1}\|$, we obtain the following bound on $\mathcal{T}_1$ with probability at least $1 - r\exp\left(-\frac{Cn}{rM^2 s}\right)$:

$$\mathcal{T}_1 \leq 18\epsilon_t s^2 \sqrt{\frac{n}{r}} \cdot 3s\sqrt{\frac{n}{r}} \cdot \frac{8r}{sn} = 432 s^2 \epsilon_t.$$

The second term $\mathcal{T}_2$ is directly controlled by Lemma 6, yielding the following (with probability at least $1 - Cr^2 \exp(-\frac{Cn}{r^2 M^2})$):

$$\mathcal{T}_2 \leq \mu_1 \sqrt{\frac{r}{d}} \frac{1968 s^3 \epsilon_t}{\sqrt{r}}.$$

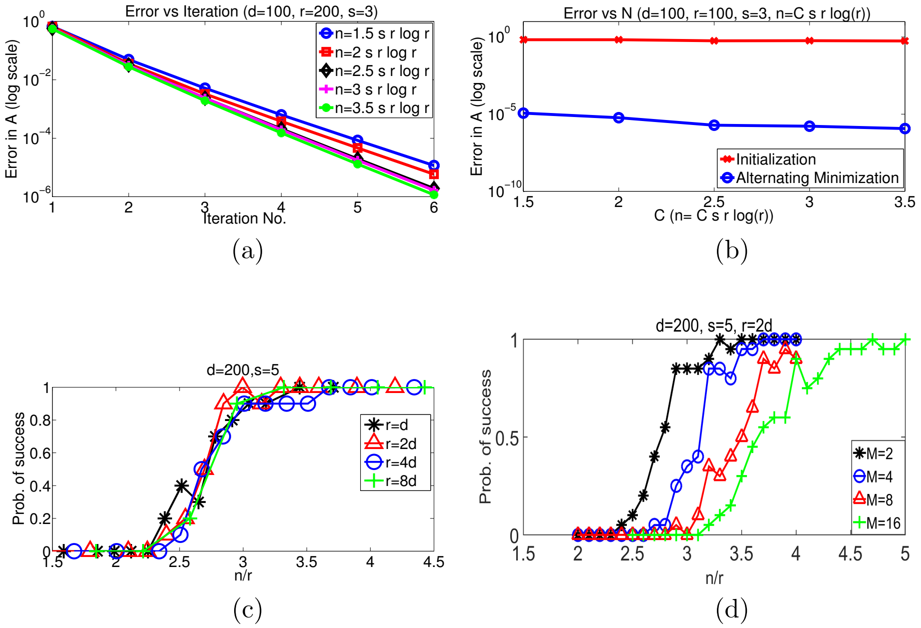
Putting all the terms together, we get

$$\|A_p\|_2 \geq 1 - 9s^2 \left(48 + \frac{1968 s\mu_1}{\sqrt{d}}\right) \epsilon_t \geq \frac{3}{4}, \tag{9}$$

where the inequality follows since $9s^2(48 + \frac{1968s\mu_1}{\sqrt{d}})\epsilon_t \leq 9s^2(48 + \frac{1968s\mu_1}{\sqrt{d}})\epsilon_0 \leq 1/4$ by our assumption (A6) on ϵ_0. Combining the bounds (8) and (9) yields the desired recursion (7). Appealing to Lemma 8 along with our setting of ϵ_t (3) completes the proof of the claim (7). Finally note that the error probability in the theorem is obtained by using the fact that $M \geq 1$ and that the failure probability is purely incurred from the structure of the nonzero entries of X^*, so that it is incurred only once and not at each round. This avoids the need of a union bound over all the rounds, yielding the result. ∎

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 09/15/19 to 75.102.135.243. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php



(a)

(b)

(c)

(d)

FIG. 1. (a) *Average error after each alternating minimization step of Algorithm* 1 *on log-scale.* (b) *Average error after the initialization procedure (Algorithm* 1 *of* [1]*) and after* 5 *alternating minimization steps of Algorithm* 1. (c) *Sample complexity requirement of the alternating minimization algorithm. For ease of experiments, we initialize the dictionary using a random perturbation of the true dictionary rather than using Algorithm* 1 *of* [1] *which should in fact give better initial point with smaller error.* (d) *Sample complexity of alternating minimization for various values of M. In order to understand the effect of M on local convergence, similar to the setting in* (c)*, we initialize the dictionary using a random perturbation of the true dictionary. We can see that the sample complexity increases with M but sublinearly.*

4. Experiments. Alternating minimization/descent approaches have been widely used for dictionary learning and several existing works show effectiveness of these methods on real-world/synthetic datasets [5, 31]. Hence, instead of replicating those results, in this section we focus on illustrating the following three key properties of our algorithms via experiments in a controlled setting: (a) advantage of alternating minimization over one-shot initialization, (b) linear convergence of alternating minimization, (c) sample complexity of alternating minimization.

Data generation model. Each entry of the dictionary matrix A is chosen i.i.d. from $\mathcal{N}(0, 1/\sqrt{d})$. Note that random Gaussian matrices are known to satisfy incoherence and the spectral norm bound [35]. The support of each column of X was chosen independently and uniformly from the set of all s-subsets of $[r]$. Similarly, each nonzero element of X was chosen independently from the uniform distribution on $[-2, -1] \cup [1, 2]$. We use the GraDeS algorithm of [11] to solve the sparse recovery step, as it is faster than lasso. We measure error in the recovery of dictionary by $error(A) = \max_i \sqrt{1 - \frac{\langle A_i, A_i^* \rangle^2}{\|A_i\|_2^2 \|A_i^*\|_2^2}}$. The first two plots are for a typical run and the third plot averages over 10 runs. The implementation is in Matlab.

Linear convergence. In the first set of experiments, we fixed $d = 100$, $r = 200$ and measured error after each step of our algorithm for increasing values of n. Figure 1(a) plots error observed after each iteration of alternating minimization; the first data point refers to the error incurred by the initialization method (Algorithm 1 of [1]). As expected due to Theorem 1, we observe a geometric decay in the error.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 09/15/19 to 75.102.135.243. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

One-shot versus iterative algorithm. It is conceivable that a good initialization procedure itself is sufficient to obtain an estimate of the dictionary up to reasonable accuracy, without recourse to the alternating minimization procedure of Algorithm 1. Figure 1(b) shows that this is not the case. The figure plots the error in recovery versus the number of samples used for both Algorithm 1 of [1] and Algorithm 1. It is clear that the recovery error of the alternating minimization procedure is significantly smaller than that of the initialization procedure. For example, for $n = 2.5sr \log r$ with $s = 3, r = 200, d = 100$, initialization incurs error of .56, while alternating minimization incurs error of 10^{-6}. Note, however, that the recovery accuracy of the initialization procedure is nontrivial and also crucial to the success of alternating minimization—a random vector in $\mathbb{R}^d$ would give an error of $1 - \frac{1}{d} = 0.99$ (since the inner product is concentrated around $1/\sqrt{d}$), whereas the error after the initialization procedure is ≈ 0.55.

Sample complexity versus $\boldsymbol{r}$. We study the sample complexity requirement of the alternating minimization algorithm which is $n = \mathcal{O}(r^2 \log r)$ according to assumption (A5), assuming good enough initialization. Figure 1(c) suggests that in fact only $\mathcal{O}(r)$ samples are sufficient for success of alternating minimization. The figure plots the probability of success with respect to $\frac{n}{r}$ for various values of r. A trial is said to succeed if at the end of 25 iterations, the error is smaller than 10^{-6}. Since we focus only on the sample complexity of alternating minimization, we use a faster initialization procedure: we initialize the dictionary by randomly perturbing the true dictionary as $A(0) = A^* + Z$, where each element of Z is an $\mathcal{N}(0, 0.5/\sqrt{d})$ random variable. Figure 1(c) shows that the success probability transitions at nearly the same value for various values of r, suggesting that the sample complexity of the alternating minimization procedure in this regime of $r = \mathcal{O}(d)$ is just $O(r)$.

Sample complexity versus $\boldsymbol{M}$. Finally, we look at the dependence of sample complexity on M. Since our focus is on alternating minimization, we initialize the dictionary by randomly perturbing the true dictionary just like in the setting of Figure 1(c). The figure plots the probability of success with respect to $\frac{n}{r}$ for various values of M.

5. Conclusions.

In this paper we provide the first analysis for the local linear convergence of the popular alternating minimization heuristic commonly used for solving dictionary learning problems in practice. Combined with some recent results, this also provides an efficient method for global and exact recovery of the unknown overcomplete dictionary under favorable assumptions. The results are of interest from both theoretical and practical standpoints. From a theoretical standpoint, this is one of the very few results that provides guarantees on a dictionary learned using an efficient algorithm, and one of the first for the overcomplete setting. From a practical standpoint, there is a tremendous interest in the problem, and we believe that an understanding of the theoretical properties of existing methods is critical in designing better methods. Indeed, our work provides some such hints toward designing a better algorithm. For instance, the sparse recovery step in our method decodes the coefficients individually for each sample. We believe that a better method can be designed by jointly decoding all the samples, which allows one to force consistency across samples (for instance, in our random coefficient model, the number of samples per dictionary element is also controlled in addition to the number of dictionary elements per sample).

More interestingly, our work extends a growing body of recent literature on analysis of alternating minimization methods for a variety of nonconvex factorization problems [15, 24], where global in addition to local results are being established with

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 09/15/19 to 75.102.135.243. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

appropriate initialization strategies. Of course, results on alternating minimization go much further back, even in nonconvex optimization to Csiszar's seminal works (see, e.g., the recent tutorial [8] for an overview), as well as in convex minimization and projection problems. However, the recent work has been largely motivated by applications of nonconvex optimization arising in machine learning. We believe that the emergence of these newer results indicates the possibility of a more general theory of alternating optimization procedures for a broad class of factorization-style nonconvex problems and should be an exciting question for future research

Appendix A. Proofs for alternating minimization. In this section, we will present our proof for the results on alternating minimization. We present the proofs for Theorem 1 and the other main lemmas in section A.1. In section A.2, we present the auxiliary lemmas and their proofs.

A.1. Proofs of main lemmas. For the reader's convenience, we recall Lemmas 4, 5, and 6 from section 3.4 along with their proofs. The more technical lemmas are deferred to the next section.

We first recall some notation and define additional abbreviations before proving the lemmas. Recall the modeling assumption $Y = A^*X^*$ and that we denote $X(t+1)$ as X, YX^+ as A, $A(t)$ as $\widetilde{A}$, and $X(t)$ as $\widetilde{X}$. Denote ${X^*}_i^p = \chi_i^p M_i^p$ $\forall 1 \leq p \leq r$, $\forall 1 \leq i \leq n$, where $\chi_i^p = 1$ if $p \in \mathrm{Supp}(X_i^*)$ and 0 otherwise and M_i^p are i.i.d. random variables with $\mathbb{E}\left[M_i^p\right] = \mu$ and $\mathbb{E}\left[(M_i^p)^2\right] = \sigma^2 + \mu^2$. Assumption $(A3)$ gives us

1. $\mu^2 + \sigma^2 = 1$, and
2. $|M_i^p| \leq M$ a.s.

LEMMA 10 (restatement of 4). *Let $\triangle X := \widetilde{X} - X^*$. Assume that $2\mu_0 s/\sqrt{d} \leq 0.1$ and $\sqrt{s\epsilon_t} \leq 0.1$ Then, we have*

1. $\mathrm{Supp}(\triangle X) \subseteq \mathrm{Supp}(X^*)$,
2. $\left\|\triangle X\right\|_\infty \leq 9s \cdot dist(\widetilde{A}, A^*) \leq 9s\epsilon_t$.

Proof. In order to establish the lemma, we use a result of Candes regarding the lasso estimator with deterministic noise for the recovery procedure:

$$\widehat{x}_i = \arg\min_{x \in \mathbb{R}^r} \left\|x\right\|_1 \quad \text{such that} \quad \left\|Y_i - Ax\right\|_2 \leq \epsilon. \tag{10}$$

THEOREM 11 (Theorem 1.2 from [7]). *Suppose $Y_i = Ax_i + z_i$, where x_i is s-sparse and $\left\|z_i\right\|_2 \leq \epsilon$. Assume further that $\delta_{2s} \leq \sqrt{2} - 1$. Then the solution to* (10) *obeys the following, for a universal constant C_1:*

$$\left\|\widehat{x}_i - x_i\right\|_2 \leq C_1\epsilon.$$

In particular, $C_1 = 8.5$ suffices for $\delta_{2s} \leq 0.2$.

In order to apply the theorem, we need to demonstrate that the RIP condition holds on $\widetilde{A}$. Consider any $2s$-sparse subset S of $[r]$. We have

$$\sigma_{\min}(\widetilde{A}_S) \geq \sigma_{\min}(A_S^*) - \left\|A_S^* - \widetilde{A}_S\right\|_2 \overset{(\zeta_1)}{\geq} 1 - \delta_{2s} - \left\|A_S^* - \widetilde{A}_S\right\|_F \quad \text{and}$$
$$\sigma_{\max}(\widetilde{A}_S) \leq \sigma_{\max}(A_S^*) + \left\|A_S^* - \widetilde{A}_S\right\|_2 \overset{(\zeta_2)}{\leq} 1 + \delta_{2s} + \left\|A_S^* - \widetilde{A}_S\right\|_F,$$

where ζ_1 and ζ_2 follow from assumption (A1). Recalling the assumption $\sqrt{s\epsilon_t} < 0.1$, and that $\delta_{2s} < 0.1$, we see that the maximum and minimum singular values of $\widetilde{A}_S$ are at least 4/5 and at most 6/5, respectively. Appealing to Theorem 11, we see that

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 09/15/19 to 75.102.135.243. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

this guarantees $\|\triangle X_i\|_2 \leq 9s\epsilon_t$. Since this is also an infinity norm error bound, we obtain the second part of the lemma. The proof of the first part is further implied by the choice of our threshold at a level of $9s\epsilon_t$, which ensures that any nonzero element in X has $\left|{X^*}_p^i\right| > 0$ (since we would have $\left|X_p^i\right| \leq 9s\epsilon_t$ by our infinity norm bound otherwise). □

We now move on to the proof of Lemma 5. We point out that the lemma applies uniformly to all matrices W satisfying $\mathrm{Supp}(W) \subseteq \mathrm{Supp}(X^*)$, irrespective of the values of these entries. This might be surprising at first but is a rather straightforward consequence of random matrix concentration theory.

LEMMA 12 (restatement of lem:W-spectralnorm). *With probability at least* $1 - r\exp\left(-\frac{Cn}{rs}\right)$*, for every* $r \times n$ *matrix* W *s.t.* $\mathrm{Supp}(W) \subseteq \mathrm{Supp}(X^*)$*, we have*

$$\|W\|_2 \leq 2\|W\|_\infty \sqrt{\frac{s^2 n}{r}}.$$

Proof. Since the support of W is a subset of the support of X^*, $W_i^p = \chi_i^p W_i^p$. Now,

$$\begin{aligned}\|W\|_2 &= \max_{u,v\|u\|_2=1,\|v\|_2=1} \sum_{ip} W_i^p u^i v^p = \max_{u,v\|u\|_2=1,\|v\|_2=1} \sum_{ip} \chi_i^p W_i^p u^i v^p \\ &\leq \|W\|_\infty \cdot \max_{u,v\|u\|_2=1,\|v\|_2=1} \cdot \sum_{ip} \chi_i^p u^i v^p,\end{aligned}$$

where the inequality holds since the maximum inner product over all pairs (u, v) from the unit sphere is larger than that over pairs with $u^i v^p \geq 0 \ \forall \ i, p$. Note that the last expression is equal to $\|W\|_\infty u^\top \chi v$, where we use χ to denote the matrix with the nonzero pattern of the matrix X^*. In order to bound $\|W\|_2$ uniformly $\forall$ W satisfying the sparsity pattern, it suffices to control the operator norm of χ. This can indeed be done by applying Lemma 16 with $\mu = M = 1$ and $\sigma = 0$. Doing so yields with probability at least $1 - r\exp\left(-\frac{Cn}{rs}\right)$

$$\|W\|_2 \leq 2\|W\|_\infty \sqrt{\frac{s^2 n}{r}},$$

which completes the proof. □

We now finally prove Lemma 6, which is our main lemma on the structure of $X^* X^+$. Specifically, the lemma will show how to control the off-diagonal elements of this matrix carefully.

LEMMA 13 (restatement of 6). *With probability at least* $1 - Cr^2 \exp\left(-\frac{Cn}{r^2 M^2}\right)$*, we have uniformly for every* $p \in [r]$ *and every* $\triangle X$ *such that* $\|\triangle X\|_\infty < \frac{1}{288s}$*,*

$$\left\|\left(\triangle X X^+\right)_p^{\setminus p}\right\|_2 = \left\|\left(X^* X^+\right)_p^{\setminus p}\right\|_2 \leq \frac{1968 s^2 \|\triangle X\|_\infty}{\sqrt{r}}.$$

Proof. For simplicity, we will prove the statement for $p = 1$. We first relate $X^* X^+$ to $\triangle X X^+$.

$$\begin{aligned}\left(X^* X^+\right)_1^{\setminus 1} &= \left(\left(X^* - X\right) X^+\right)_1^{\setminus 1} \\ &= -\left(\triangle X X^+\right)_1^{\setminus 1} \\ &= -\left(\triangle X X^\top \left(X X^\top\right)^{-1}\right)_1^{\setminus 1},\end{aligned}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 09/15/19 to 75.102.135.243. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

where the first step follows from the fact that $XX^{+} = \mathbb{I}$. This proves the first part of the lemma. We now expand the above as follows:

$$\left(\triangle XX^\top \left(XX^\top\right)^{-1}\right)_1^{\setminus 1} = \left(\triangle XX^\top\right)_1^{\setminus 1}\left(\left(XX^\top\right)^{-1}\right)_1^{1} + \left(\triangle XX^\top\right)_{\setminus 1}^{\setminus 1}\left(\left(XX^\top\right)^{-1}\right)_1^{\setminus 1}.$$

Using the triangle inequality, we have

$$\begin{aligned}\left\|\left(\triangle XX^\top \left(XX^\top\right)^{-1}\right)_1^{\setminus 1}\right\|_2 \le{}& \underbrace{\left|\left(\left(XX^\top\right)^{-1}\right)_1^{1}\right|}_{\mathcal{T}_1}\underbrace{\left\|\left(\triangle XX^\top\right)_1^{\setminus 1}\right\|_2}_{\mathcal{T}_2}\\ &+ \underbrace{\left\|\left(\triangle XX^\top\right)_{\setminus 1}^{\setminus 1}\right\|_2}_{\mathcal{T}_3}\underbrace{\left\|\left(\left(XX^\top\right)^{-1}\right)_1^{\setminus 1}\right\|_2}_{\mathcal{T}_4}.\end{aligned} \tag{11}$$

We now bound each of the above four quantities. We can easily bound $\mathcal{T}_1$ via a spectral norm bound on $\left(XX^\top\right)^{-1}$. Doing so, we obtain with probability at least $1 - r\exp(-\frac{Cn}{rM^2 s})$

$$\mathcal{T}_1 = \left|\left(\left(XX^\top\right)^{-1}\right)_1^{1}\right| \le \left\|\left(XX^\top\right)^{-1}\right\|_2 \overset{(\zeta_1)}{\le} \frac{8r}{ns}, \tag{12}$$

where (ζ_1) follows from Lemma 16. To bound $\mathcal{T}_2$, we use Lemma 20 and obtain with probability at least $1 - Cr^2\exp\left(-\frac{Cn}{r^2M^2}\right)$

$$\mathcal{T}_2 = \left\|\left(\triangle XX^\top\right)_1^{\setminus 1}\right\|_2 \le \frac{6\left\|\triangle X\right\|_\infty s^2 n}{r^{\frac{3}{2}}}, \tag{13}$$

where we recall the assumption $\left\|\triangle X\right\|_\infty \le 1/(64s)$. We now bound $\mathcal{T}_3$ as follows:

$$\begin{aligned}\mathcal{T}_3 = \left\|\left(\triangle XX^\top\right)_{\setminus 1}^{\setminus 1}\right\|_2 &\le \left\|\left(\triangle X\right)_{\setminus 1}^{\setminus 1}\right\|_2\left\|\left(X\right)_{\setminus 1}^{\setminus 1}\right\|_2\\ &\overset{(\zeta_1)}{\le} 2\left\|\triangle X\right\|_\infty s\sqrt{\frac{n}{r}}\cdot 2(1+\left\|\triangle X\right\|_\infty)s\sqrt{\frac{n}{r}}\\ &< \frac{6\left\|\triangle X\right\|_\infty s^2 n}{r},\end{aligned} \tag{14}$$

where (ζ_1) follows from Lemmas 5 and 17 (since $\mathrm{Supp}(\triangle X) \subseteq \mathrm{Supp}(X) \cup \mathrm{Supp}(X^*) = \mathrm{Supp}(X^*)$). Finally, to bound $\mathcal{T}_4$, we start by noting the following block decomposition of the matrix $XX^\top$:

$$XX^\top = \begin{bmatrix} X^1\left(X^1\right)^\top & X^1\left(X^{\setminus 1}\right)^\top \\ X^{\setminus 1}X^{1^\top} & X^{\setminus 1}\left(X^{\setminus 1}\right)^\top \end{bmatrix}.$$

Given this block-structure, we can now invoke Lemma 22 (Schur complement lemma) to obtain

$$\left(\left(XX^\top\right)^{-1}\right)_1^{\setminus 1} = -\frac{1}{X^1\left(X^1\right)^\top}BX^{\setminus 1}\left(X^1\right)^\top,$$

where

$$B := \left(\left(XX^\top\right)^{-1}\right)_{\setminus 1}^{\setminus 1}. \tag{15}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 09/15/19 to 75.102.135.243. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Here we recall that B^{-1} is the Schur complement of $X_1 {X_1}^\top$. Using Lemma 20 and equation (21) we have with probability at least $1 - Cr^2 \exp\left(-\frac{Cn}{r^2 M^2}\right)$

$$\left\| \left(\left(XX^\top \right)^{-1} \right)_1^{\backslash 1} \right\|_2 \le \frac{1}{\left| X^1 (X^1)^\top \right|} \left\| B \right\|_2 \left\| X^{\backslash 1} (X^1)^\top \right\|_2 \le \frac{8r}{sn} \cdot \left\| B \right\|_2 \cdot \frac{5s^2 n}{r^{\frac{3}{2}}} = \frac{40s}{\sqrt{r}} \left\| B \right\|_2 . \tag{16}$$

Using the expression (15) and the lower bound on $\sigma_{\min}(X)$ from Lemma 17, we also have the following bound for $\left\| B \right\|_2$ with probability at least $1 - r\exp\left(-\frac{Cn}{rM^2 s}\right)$:

$$\left\| B \right\|_2 = \left\| \left(\left(XX^\top \right)^{-1} \right)_{\backslash 1}^{\backslash 1} \right\|_2 \le \left\| \left(XX^\top \right)^{-1} \right\|_2 \le \frac{8r}{ns}.$$

Plugging the above into (16) gives us

$$\left\| \left(\left(XX^\top \right)^{-1} \right)_1^{\backslash 1} \right\|_2 \le \frac{40s}{\sqrt{r}} \cdot \frac{8r}{ns} \le \frac{320\sqrt{r}}{n}. \tag{17}$$

Combining (12), (13), (5), and (17), we obtain with probability at least $1 - Cr^2$ $\exp(-\frac{Cn}{r^2 M^2})$,

$$\begin{aligned} \left\| \left(XX^{*+} \right)_p^{\backslash p} \right\|_2 &\le \frac{48 \left\| \triangle X \right\|_\infty s}{\sqrt{r}} + \frac{1920 \left\| \triangle X \right\|_\infty s^2}{\sqrt{r}} \\ &\le \frac{1968 s^2 \left\| \triangle X \right\|_\infty}{\sqrt{r}}. \end{aligned}$$

□

A.2. Main technical lemmas. In this section, we state and prove the main technical lemmas used in our results.

Lemma 14. *Under assumptions* (A3) *and* (A4), *we have*

$$\Sigma := \mathbb{E}\left[{X^*}_i {X^*}_i^\top \right] = \left(\frac{s}{r} - \frac{s(s-1)\mu^2}{r(r-1)} \right) \mathbb{I} + \frac{s(s-1)\mu^2}{r(r-1)} 11^\top .$$

Proof. Note that $\chi_i^p, 1 \le p \le r$, all have the same distribution. Hence, by symmetry and linearity of expectation, $\mathbb{E}\left[\chi_i^p\right] = \frac{1}{r}\left[\sum_{q=1}^r \chi_i^q\right] = \frac{s}{r}$. Similarly, $\mathbb{E}\left[(\chi_i^p)^2\right] = \frac{1}{r}\left[\sum_{q=1}^r (\chi_i^q)^2\right] = \frac{s}{r}$. Also, since $\sum_{q=1}^r \chi_i^q = s$, we have

$$s^2 = \mathbb{E}\left[\left(\sum_{q=1}^r (\chi_i^q) \right)^2 \right] = \mathbb{E}\left[\sum_{p,q} \chi_i^p \chi_i^q \right] = r\mathbb{E}\left[(\chi_i^p)^2\right] + (r^2 - r)\mathbb{E}\left[\chi_i^p \chi_i^q\right].$$

Hence, $\mathbb{E}\left[\chi_i^p \chi_i^q\right] = \frac{s(s-1)}{r(r-1)}$ for $p \ne q$.

Now, recall that ${X^*}_i^p = \chi_i^p M_i^p$. Now, we first consider diagonal terms of Σ:

$$\Sigma_p^p = \mathbb{E}\left[({X^*}_i^p)^2\right] = \mathbb{E}\left[(\chi_i^p)^2\right] \mathbb{E}\left[(M_i^p)^2\right] = \frac{s}{r}(\mu^2 + \sigma^2) = \frac{s}{r}. \tag{18}$$

Similarly, using independence of M_i^p and M_i^q, off-diagonal terms of Σ are given by

$$\Sigma_p^q = \mathbb{E}\left[\chi_i^p \chi_i^q\right] \mathbb{E}\left[M_i^p\right] \mathbb{E}\left[M_i^q\right] = \frac{s(s-1)}{r(r-1)}\mu^2. \tag{19}$$

Lemma 14 now follows by using (18) and (19). □

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 09/15/19 to 75.102.135.243. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

In particular, two consequences of the lemma which will be particularly useful are about the extreme singular values of Σ. Recalling that $2s \le r$ and $\mu^2 \le 1$ by assumption, we obtain

$$\sigma_{\min}(\Sigma) \ge \frac{s}{2r}, \quad \text{and} \quad \sigma_{\max}(\Sigma) \le \frac{2s^2}{r}. \tag{20}$$

We next establish some results on the spectrum of the empirical covariance matrix, using a standard result from random matrix theory. For the convenience of the reader, we recall the following theorem from [35].

THEOREM 15 (restatement of Theorem 5.44 from [35]). *Consider an $r \times n$ matrix W where each column w_i of W is an independent random vector with covariance matrix Σ. Suppose further that $\|w_i\|_2 \le \sqrt{u}$ a.s. $\forall$ i. Then for any $t \ge 0$, the following inequality holds with probability at least $1 - r\exp\left(-ct^2\right)$:*

$$\left\| \frac{1}{n} W W^T - \Sigma \right\|_2 \le \max\left(\|\Sigma\|_2^{1/2} \gamma, \gamma^2 \right), \ \textit{where } \gamma = t\sqrt{\frac{u}{n}}.$$

Here $c > 0$ is an absolute numerical constant. In particular, this inequality yields

$$\|W\|_2 \le \|\Sigma\|_2^{\frac{1}{2}} \sqrt{n} + t\sqrt{u}.$$

Using the theorem, we can establish the following results on concentration of empirical covariance matrices. Hereafter, C will be a universal constant that can change from line to line.

LEMMA 16. *There exists a universal constant C such that with probability at least $1 - r\exp(-\frac{C\delta^2 ns}{rM^2})$, we have*

$$\left\| \frac{1}{n} X^* {X^*}^\top - \Sigma \right\|_2 \le \max\left(\sqrt{2}\delta, \delta^2 \right) \frac{s^2}{r}.$$

In particular, with probability at least $1 - r\exp(-\frac{Cn}{rM^2 s})$, we have the bounds

$$\|X^*\|_2 \le 2\sqrt{\frac{ns^2}{r}} \quad \textit{and} \quad \sigma_{\min}(X^* {X^*}^\top) \ge \frac{ns}{3r}.$$

Proof. Note that $\|X_i^*\|_2 \le \sqrt{s}M$. Also, by (20), $\|\Sigma\|_2 \le \frac{s}{r} + \frac{s(s-1)\mu^2}{r-1} \le \frac{2s^2}{r}$. Using Theorem 15 with $t = \delta\sqrt{\frac{ns}{rM^2}}$, we obtain

$$\left\| \frac{1}{n} X^* {X^*}^\top - \Sigma \right\|_2 \le \max\left(\sqrt{2}\delta, \delta^2 \right) \frac{s^2}{r},$$

w.p. with probability greater than $1 - r\exp\left(-\frac{C\delta^2 ns}{rM^2}\right)$. In order to obtain the second part, we apply the first part of the lemma with $\delta = 1/6\sqrt{2}s$ as well as Lemma 14 to bound the largest and smallest singular values of $XX^\top/n$. Taking square roots completes the proof. □

The next lemma we state is a specialization of Lemma 5 to obtain bounds on the spectral norm of our iterates X.

LEMMA 17. *With probability at least $1 - r\exp\left(-\frac{Cn}{rs}\right) - r\exp\left(-\frac{Cn}{rM^2 s}\right)$, for every $r \times n$ matrix X s.t. $\mathrm{Supp}(X) \subseteq \mathrm{Supp}(X^*)$, we have*

$$\|X\|_2 \le 2 \cdot (1 + \|X - X^*\|_\infty) \cdot s\sqrt{\frac{n}{r}}.$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 09/15/19 to 75.102.135.243. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Proof. Let $X = X^* + E_{X^*}$, where $\mathrm{Supp}(E_{X^*}) \subseteq \mathrm{Supp}(X^*)$. Hence, $\|X\|_2 \le \|X^*\|_2 + \|X - X^*\|_2$. Lemma 17 follows directly using Lemmas 16 and 5. ◻

The above lemma can be used to obtain a singular value perturbation bound for matrices of the form $XX^\top$. We will need control over the upper and lower singular values of such matrices for our proofs, which we next provide.

LEMMA 18. *With probability at least* $1 - r\exp\left(-\frac{Cn}{rs}\right) - r\exp\left(-\frac{Cn}{rM^2 s}\right)$, *for every* $r \times n$ *matrix* X *s.t.* $\mathrm{Supp}(X) \subseteq \mathrm{Supp}(X^*)$, *we have*

$$\left\|XX^\top - X^*{X^*}^\top\right\|_2 \le 4\left(2\|X - X^*\|_\infty + \|X - X^*\|_\infty^2\right) \cdot \frac{s^2 n}{r}.$$

Further assuming $\|X - X^*\|_\infty \le 1/(64s)$, *we have with the same probability*

$$\sigma_{\min}(XX^\top) \ge \frac{ns}{8r}.$$

Proof. Let $X = X^* + E_{X^*}$. Note that $\mathrm{Supp}(E_{X^*}) \subseteq \mathrm{Supp}(X^*)$. Now,

$$\|XX^\top - X^*{X^*}^\top\|_2 \le \|E_{X^*}\|_2(\|E_{X^*}\|_2 + 2\|X^*\|_2).$$

By Lemma 5, $\|E_{X^*}\|_2 \le 2s\sqrt{\frac{n}{r}}\|E_{X^*}\|_\infty$ with probability at least $\ge 1 - r\exp\left(-\frac{Cn}{rs}\right)$. Combining this with the bound on $\|X^*\|_2$ from Lemma 16 completes the proof. The second statement now follows by combining the result with our earlier lower bound on the minimum singular value of X^* in Lemma 16. ◻

A particular consequence of this lemma which will be useful is a lower bound on the diagonal entries of the matrix $XX^\top$. Indeed, we see that under the assumption $\|X - X^*\|_\infty \le 1/(64s)$, with probability at least $1 - r\exp\left(-\frac{Cn}{rs}\right) - r\exp\left(-\frac{Cn}{rM^2 s}\right)$ we have the lower bound uniformly $\forall\ p = 1, 2, \ldots, r$,

$$X^p {X^p}^\top \ge \frac{ns}{8r}. \tag{21}$$

We finally have the following concentration lemma, which is a simple consequence of the Bernstein concentration bound (Lemma 23).

LEMMA 19. *Let* χ_i^p *be as defined in section* A.1. *Then,*

1. *with probability at least* $1 - 2r\exp\left(-\frac{\delta^2 ns}{4r}\right)$: $(1-\delta)\frac{sn}{r} \le \sum_{i=1}^n \chi_i^p \le (1+\delta)\frac{sn}{r}\ \forall\, p \in [r]$, *and*
2. *with probability at least* $1 - 2r\exp\left(-\frac{\delta^2 ns}{4rM^2}\right)$: $(1-\delta)\frac{sn}{r} \le \|{X^*}_p\|_2^2 = \sum_{i=1}^n \chi_i^p (M_i^p)^2 \le (1+\delta)\frac{sn}{r}\ \forall\, p \in [r]$.

Proof. We start with the proof of the second part, noting that the first part then immediately follows by setting $(M_i^p)^2 \equiv 1$. The second part will follow from a straightforward use of Bernstein's inequality (Lemma 23). Note that $|M_i^p| \le M$ and $\mathbb{E}[(M_i^p)^2] = 1$. As a result, $\forall\ i = 1, 2, \ldots, n$ we have $|\chi_i^p (M_i^p)^2| \le M^2$, and

$$\mathrm{Var}[\chi_i^p (M_i^p)^2] \le \mathbb{E}[\chi_i^p (M_i^p)^4] \le M^2 \mathbb{E}[\chi_i^p (M_i^p)^2] = M^2 \frac{s}{r}.$$

The last step follows since $\mathbb{E}[\chi_i^p (M_i^p)^2] = \mathbb{E}[\chi_i^p] = s/r$. Consequently, we obtain that with probability at least $1 - 2\exp(-ns\delta^2/(2rM^2(1+\delta/3)))$ we have

$$\left|\sum_{i=1}^n \chi_i^p (M_i^p)^2 - \frac{ns}{r}\right| \le \frac{\delta ns}{r}.$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 09/15/19 to 75.102.135.243. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

To complete the proof, note that $1 \geq \delta/3$, which yields the stated error probability after a union bound over $p \in [r]$. Finally, as stated before, we can recover the first part by setting $(M_i^p)^2 \equiv 1$. ∎

LEMMA 20. *With probability at least* $1 - Cr^2 \exp\left(-\frac{Cn}{r^2 M^2}\right)$, *for every* $r \times n$ *matrix* X *s.t.* $\mathrm{Supp}(X) \subseteq \mathrm{Supp}(X^*)$, *and* $\|X - X^*\|_\infty \leq \theta$, *we have the following bounds uniformly* $\forall\ p = 1, 2, \ldots, r$*:*

1. $\|\left(\triangle X X^\top\right)_p^{\setminus p}\|_2 \leq \left(1 + \sqrt{2s}\theta\right) \frac{4\sqrt{2}\theta s^2 n}{r^{\frac{3}{2}}}$, *and*
2. $\|X^{\setminus p}(X^p)^\top\|_2 \leq (1 + \sqrt{s}\theta)^2 \frac{4s^2 n}{r^{\frac{3}{2}}}$,

where $\triangle X := X - X^*$.

Proof. We start by proving the first part of the lemma.

Proof of part 1. Let D denote the $n \times n$ diagonal matrix with $D_i^i = \chi_i^p$. Using this notation, we have $\left(\triangle X X^\top\right)_p^{\setminus p} = \left(\triangle X D X^\top\right)_p^{\setminus p}$. So, we have

$$\begin{aligned}
\left\|\left(\triangle X X^\top\right)_p^{\setminus p}\right\|_2 &= \left\|\left(\triangle X D X^\top\right)_p^{\setminus p}\right\|_2 \\
&\leq \left\|(\triangle X D)^{\setminus p}\right\|_2 \left\|\left(X^\top\right)_p\right\|_2 \\
&\leq \left\|(\triangle X D)^{\setminus p}\right\|_2 \left\|\left({X^*}^\top\right)_p + \left(\triangle X^\top\right)_p\right\|_2 \\
&\overset{(\zeta_1)}{\leq} \left\|(\triangle X D)^{\setminus p}\right\|_2 \cdot \left(\sqrt{\frac{2sn}{r}} + \|\triangle X\|_\infty 2s\sqrt{\frac{n}{r}}\right)
\end{aligned}$$

with probability at least $1 - Cr\exp\left(-Cn/rs\right)$ uniformly for every $p \in [r]$. Note that the first term in (ζ_1) follows from the second part of Lemma 19 and the second is a consequence of Lemma 5. For the rest of the proof, we choose $p = 1$ and then apply a union bound to obtain the result for every $p \in [r]$. In order to control $\|(\triangle X D)^{\setminus 1}\|_2$, we observe that it is a matrix with a random number of columns selected by the matrix D. In particular, conditioned on $\left\{i : D_i^i = 1\right\}$, the support of ${X^*}_i^{\setminus 1}$ is uniform over all subsets of $\{2, \ldots, d\}$ of size $s-1$ (and the support of $\triangle X$ is a subset of the support of X^*). Hence we can easily see that

$$\begin{aligned}
\mathbb{P}\left[\left\|(\triangle X D)^{\setminus 1}\right\|_2 > t\right] &\leq \mathbb{P}\left[\left\{\left\|(\triangle X D)^{\setminus 1}\right\|_2 > t\right\} \cap \left\{\frac{sn}{2r} < \left|\left\{i : D_i^i = 1\right\}\right| < \frac{2sn}{r}\right\}\right] \\
&\quad + \mathbb{P}\left[\left\{\left|\left\{i : D_i^i = 1\right\}\right| \leq \frac{sn}{2r}\right\} \cup \left\{\left|\left\{i : D_i^i = 1\right\}\right| \geq \frac{2sn}{r}\right\}\right]. \qquad (22)
\end{aligned}$$

In order to control the first probability, note that

$$\begin{aligned}
&\mathbb{P}\left[\left\{\left\|(\triangle X D)^{\setminus 1}\right\|_2 > t\right\} \cap \left\{\frac{sn}{2r} < \left|\left\{i : D_i^i = 1\right\}\right| < \frac{2sn}{r}\right\}\right] \\
&\leq \sum_{k=\lceil sn/2r \rceil}^{\lfloor 2sn/r \rfloor} \mathbb{P}\left[\left\{\left\|(\triangle X D)^{\setminus 1}\right\|_2 > t\right\} \cap \left\{\left|\left\{i : D_i^i = 1\right\}\right| = k\right\}\right] \\
&= \sum_{k=\lceil sn/2r \rceil}^{\lfloor 2sn/r \rfloor} \mathbb{P}\left[\left\|(\triangle X D)^{\setminus 1}\right\|_2 > t \mid \left|\left\{i : D_i^i = 1\right\}\right| = k\right] \mathbb{P}\left[\left|\left\{i : D_i^i = 1\right\}\right| = k\right].
\end{aligned}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 09/15/19 to 75.102.135.243. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

Setting $t = 2\theta\sqrt{\frac{s^2}{r} \cdot \frac{2sn}{r}} = 2\theta\sqrt{\frac{2s^3n}{r^2}}$, we obtain as a consequence of Lemma 5

$$\mathbb{P}\left[\left\{\left\|(\triangle XD)^{\backslash 1}\right\|_2 > t\right\} \cap \left\{\frac{sn}{2r} < \left|\left\{i : D_i^i = 1\right\}\right| < \frac{2sn}{r}\right\}\right]$$

$$\leq \sum_{k=\lceil sn/2r\rceil}^{\lfloor 2sn/r \rfloor} r\exp(-C\lceil sn/2r\rceil/rs)\mathbb{P}\left[\left|\left\{i : D_i^i = 1\right\}\right| = k\right] \leq r\exp\left(-\frac{Cn}{2r^2}\right). \tag{23}$$

For the second probability in (22), note that $\left|\left\{i : D_i^i = 1\right\}\right| = \sum_{i=1}^n \chi_i^1$, and use part 1 of Lemma 19 with $\delta = 1/2$. In particular, this gives us

$$\mathbb{P}\left[\left\{\left|\left\{i : D_i^i = 1\right\}\right| \leq \frac{sn}{2r}\right\} \cup \left\{\left|\left\{i : D_i^i = 1\right\}\right| \geq \frac{2sn}{r}\right\}\right] \leq 2r\exp\left(-\frac{Cns}{r}\right). \tag{24}$$

Plugging (23) and (24) into (22) and using union bound over $p \in [r]$, we obtain with probability at least $1 - Cr^2\exp\left(-\frac{Cn}{r^2}\right)$,

$$\left\|(\triangle XD)^{\backslash p}\right\|_2 \leq 2\theta\sqrt{\frac{2s^3n}{r^2}} \quad \forall\, p \in [r].$$

Proof of Part 2. The proof of this is similar to that of part 1. We have

$$\begin{aligned}\left\|X^{\backslash p}(X^p)^\top\right\|_2 &= \left\|X^{\backslash p}D(X^p)^\top\right\|_2 \\ &\leq \left\|(XD)^{\backslash p}\right\|_2 \left\|\left(X^\top\right)_p\right\|_2 \\ &\leq \left\|(XD)^{\backslash p}\right\|_2 \cdot 2\left(1 + \sqrt{s}\left\|\triangle X\right\|_\infty\right)\sqrt{\frac{sn}{r}}.\end{aligned}$$

For the first term above, we have

$$\left\|(XD)^{\backslash p}\right\|_2 \leq \left\|(X^*D)^{\backslash p}\right\|_2 + \left\|(\triangle XD)^{\backslash p}\right\|_2.$$

The second term in this decomposition was controlled above and the first one can be similarly bounded. We briefly describe how this is accomplished. As noted in the proof of part 1, conditioned on $\left\{i : D_i^i = 1\right\}$, the support of ${X^*}_i^{\backslash p}$ is uniform over all subsets of $[d] \setminus \{p\}$ of size $s-1$. Letting $t = 2\sqrt{\frac{s^2}{r} \cdot \frac{2sn}{r}}$, we use Lemma 16 to conclude that

$$\begin{aligned}\mathbb{P}\left[\left\{\left\|(X^*D)^{\backslash 1}\right\|_2 > t\right\} \cap \left\{\frac{sn}{2r} < \left|\left\{i : D_i^i = 1\right\}\right| < \frac{2sn}{r}\right\}\right] &\leq r\exp\left(-C\lfloor ns/2r\rfloor/rsM^2\right) \\ &\leq r\exp\left(-Cn/r^2M^2\right).\end{aligned}$$

Using a decomposition similar to that in (22) with the above bound, and finally apply a union bound over all $p \in [r]$, we conclude that

$$\left\|(XD)^{\backslash p}\right\|_2 \leq 2s\left(1 + \sqrt{2s}\theta\right)\sqrt{\frac{2sn}{r^2}}\ \forall\, p \in [r]$$

with probability greater than $1 - Cr^2\exp(-Cn/r^2M^2)$. This proves the lemma. □

We begin with an auxiliary result on the RIP constant of an incoherent matrix.

Downloaded 09/15/19 to 75.102.135.243. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

LEMMA 21. *Suppose A^* satisfies assumption* (B1)*. Then, the $2s$-RIP constant of A^*, δ_{2s} satisfies $\delta_{2s} < \frac{2\mu_0 s}{\sqrt{d}}$.*

Proof. Consider a $2s$-sparse unit vector $w \in \mathbb{R}^r$ with $\operatorname{Supp}(w) = S$. We have

$$\begin{aligned}\|Aw\|^2 = \left(\sum_{j\in S} w_j {A^*}_j\right)^2 &= \sum_j w_j^2 \|{A^*}_j\|^2 + \sum_{j,l\in S, j\neq l} w_j w_l \langle {A^*}_j, {A^*}_l\rangle \\ &\geq 1 - \sum_{j,l\in S, j\neq l} |w_j w_l|\, |\langle {A^*}_j, {A^*}_l\rangle| \\ &\geq 1 - \sum_{j,l\in S, j\neq l} |w_j w_l|\, \frac{\mu_0}{\sqrt{d}} \\ &\geq 1 - \frac{\mu_0}{\sqrt{d}} {\|w\|_1}^2 \\ &\geq 1 - \frac{\mu_0}{\sqrt{d}} 2s \cdot \|w\|_2^2 = 1 - \frac{2\mu_0 s}{\sqrt{d}}.\end{aligned}$$

Similarly, we have

$$\|A^* w\|_2^2 \leq 1 + \frac{2\mu_0 s}{\sqrt{d}}.$$

This proves the lemma. □

LEMMA 22 (Schur complement formula). *We have the following formula for matrix inversion:*

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} + A^{-1}BMCA^{-1} & -A^{-1}BM \\ -MCA^{-1} & M \end{bmatrix},$$

where $M^{-1} := \left(D - CA^{-1}B\right)$ is the Schur complement of A in the above matrix.

LEMMA 23 (Bernstein's concentration inequality). *Let $W_1, W_2, \ldots, W_q$ be independent zero-mean random variables. Suppose that $|W_i| \leq R$, almost surely $\forall\ i$. Then, for all positive t, we have*

$$\mathbb{P}\left[\sum_{i=1}^q W_i > t\right] \leq \exp\left(\frac{-t^2/2}{\sum_i \mathbb{E}\left[W_i^2\right] + Rt/3}\right).$$

LEMMA 24. *Let A be a $d\times d$ dictionary matrix such that each row $A^i \sim \frac{1}{\sqrt{d}}\mathcal{N}(0,\Sigma)$ where $\Sigma_{ii} = 1\ \forall\ i$, $\Sigma_{12} = \Sigma_{21} = \delta/4$ and $\Sigma_{ij} = 0$ otherwise. Then, with probability greater than $1 - d^2\exp(-c\delta^2 d)$, we have*

- *A is not incoherent, and*
- *A is 2-RIP with RIP constant δ.*

Here c is a universal constant.

Proof. We first note that by definition, $\mathbb{E}\left[\langle A_i, A_j\rangle\right] = \Sigma_{ij}$. Using the Chernoff bound for the sum of product of two Guassian random variables, we know that with probability greater than $1 - \exp(-c\delta^2 d)$, we have

$$|\langle A_i, A_j\rangle - \Sigma_{ij}| < \frac{\delta}{8}.$$

Using the union bound, we have that the above statement holds uniformly $\forall\ i, j \in [d]$ with probability greater than $1 - d^2\exp(-c\delta^2 d)$. The first claim now follows easily

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 09/15/19 to 75.102.135.243. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

since $\langle A_1, A_2\rangle \geq \Sigma_{12} - \frac{\delta}{8} = \frac{\delta}{8}$. When δ is not $\mathcal{O}(1/\sqrt{d})$, the dictionary A clearly violates the incoherence condition.

For the second claim, note that we are interested in bounding eigenvalues of the matrices $\left[\begin{smallmatrix} \|A_i\|_2^2 & \langle A_i, A_j\rangle \\ \langle A_j, A_i\rangle & \|A_j\|_2^2 \end{smallmatrix}\right]$. Under the high probability event we established above, these matrices in the worst case look like $\left[\begin{smallmatrix} 1-\delta/4 & \delta/2 \\ \delta/2 & 1-\delta/4 \end{smallmatrix}\right]$, and eigenvalues of all such matrices are at least $1 - \delta$. This proves the lemma. □

REFERENCES

[1] A. Agarwal, A. Anandkumar, and P. Netrapalli, *A Clustering Approach to Learn Sparsely-Used Overcomplete Dictionaries*, preprint, arXiv:1309.1952, 2013.

[2] M. Aharon, M. Elad, and A. Bruckstein, *K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation*, IEEE Trans. Signal Process., 54 (2006), pp. 4311–4322.

[3] S. Arora, R. Ge, Y. Halpern, D. Mimno, A. Moitra, D. Sontag, Y. Wu, and M. Zhu, *A practical algorithm for topic modeling with provable guarantees*, in Proceedings of the 30th International Conference on Machine Learning, ACM, 2013, pp. 280–288.

[4] S. Arora, R. Ge, and A. Moitra, *New Algorithms for Learning Incoherent and Overcomplete Dictionaries*, arXiv:1308.6273, 2013.

[5] K. Balasubramanian, K. Yu, and G. Lebanon, *Smooth sparse coding via marginal regression for learning sparse representations*, in Proceedings of ICML, 2013.

[6] Y. Bengio, A. Courville, and P. Vincent, *Representation learning: A review and new perspectives*, IEEE Trans. Pattern Anal. Mach. Intel., 35 (2013), pp. 1798–1828.

[7] E. J. Candes, *The restricted isometry property and its implications for compressed sensing*, C. R. Acad. Sci. Paris Ser. I, 346 (2008), pp. 589–592.

[8] I. Csiszar and P. Shields, *Information theory and statistics: A tutorial*, Found. Trends Commun. Inform. Theory, 1 (2004), pp. 417–528, http://dx.doi.org/10.1561/0100000004.

[9] G. Davis, *Adaptive Nonlinear Approximations*, Ph.D. thesis, New York University, New York, 1994.

[10] K. Engan, S. O. Aase, and J. Hakon Husoy, *Method of optimal directions for frame design*, in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 5, IEEE, 1999, pp. 2443–2446.

[11] R. Garg and R. Khandekar, *Gradient descent with sparsification: An iterative algorithm for sparse recovery with restricted isometry property*, in Proceedings of ICML, 2009.

[12] Q. Geng, H. Wang, and J. Wright, *On the Local Correctness of ℓ_1 Minimization for Dictionary Learning*, preprint, arXiv:1101:5672, 2011.

[13] R. Gribonval, R. Jenatton, and F. Bach, *Sparse and spurious: Dictionary learning with noise and outliers*, IEEE Trans. Inform. Theory, 61 (2015), pp. 6298–6319.

[14] R. Gribonval and K. Schnass, *Dictionary identification: Sparse matrix-factorization via* l_1 minimization, IEEE Trans. inform theory, 56 (2010), pp. 3523–3539.

[15] P. Jain, P. Netrapalli, and S. Sanghavi, *Low-rank matrix completion using alternating minimization*, in Proceedings of the 45th Annual ACM Symposium on Theory of Computing, 2013, pp. 665–674.

[16] R. Jenatton, J. Mairal, F. R. Bach, and G. R. Obozinski, *Proximal methods for sparse hierarchical dictionary learning*, in Proceedings of the 27th International Conference on Machine Learning, 2010, pp. 487–494.

[17] K. Lange, D. R. Hunter, and I. Yang, *Optimization transfer using surrogate objective functions*, J. comput. Graphi. Statisti., 9 (2000), pp. 1–20.

[18] H. Lee, A. Battle, R. Raina, and A. Ng, *Efficient sparse coding algorithms*, in Advances in Neural Information Processing Systems, 2006, pp. 801–808.

[19] M. S. Lewicki and T. J. Sejnowski, *Learning overcomplete representations*, Neural Comput., 12 (2000), pp. 337–365.

[20] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, *Discriminative learned dictionaries for local image analysis*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2008, pp. 1–8.

[21] A. Maurer, M. Pontil, and B. Romera-Paredes, *Sparse Coding for Multitask and Transfer Learning*, preprint, arXiv:1209.0738, 2012.

[22] N. Mehta and A. G. Gray, *Sparsity-based generalization bounds for predictive sparse coding*, in Proceedings of the 30th International Conference on Machine Learning, 2013, pp. 36–44.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 09/15/19 to 75.102.135.243. Redistribution subject to SIAM license or copyright; see http://www.siam.org/journals/ojsa.php

[23] S. N. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, *A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers*, Statist. Sci., 27 (2012), pp. 538–557.

[24] P. Netrapalli, P. Jain, and S. Sanghavi, *Phase retrieval using alternating minimization*, in Advances in Neural Information Processing Systems, 2013, pp. 2796–2804.

[25] B. A. Olshausen, *Sparse coding of time-varying natural images*, in Proceedings of the International confevence on Independent Component Analysis and Blind Source Separation, 2000, pp. 603–608.

[26] B. A. Olshausen et al., *Emergence of simple-cell receptive field properties by learning a sparse code for natural images*, Nature, 381 (1996), pp. 607–609.

[27] B. A. Olshausen and D. J. Field, *Sparse coding with an overcomplete basis set: A strategy employed by V1?*, Vision Research, 37 (1997), pp. 3311–3325.

[28] G. Raskutti, M. J. Wainwright, and B. Yu, *Restricted eigenvalue properties for correlated gaussian designs*, J. Mach. Learn. Res., 11 (2010), pp. 2241–2259.

[29] K. Schnass, *On the identifiability of overcomplete dictionaries via the minimisation principle underlying k-svd*, Appl. Comput. Harmon. Analy., 37 (2014), pp. 464–491.

[30] D. A. Spielman, H. Wang, and J. Wright, *Exact recovery of sparsely-used dictionaries*, in Proceedings of the conference on Learning Theory, 2012.

[31] J. J. Thiagarajan, K. N. Ramamurthy, and A. Spanias, *Learning stable multilevel dictionaries for sparse representations*, IEEE Trans. Neural Networks Learning Systems, 26 (2015), pp. 1913–1926, doi:10.1109/TNNLS.2014.2361052.

[32] A. M. Tillmann, *On the computational intractability of exact and approximate dictionary learning*, IEEE Signal Process. Lett., 22 (2015), pp. 45–49.

[33] J. Tropp and A. Gilbert, *Signal recovery from random measurements via orthogonal matching pursuit*, IEEE Trans. Inform. Theory, 53 (2007), pp. 4655–4666.

[34] D. Vainsencher, S. Mannor, and A. M. Bruckstein, *The sample complexity of dictionary learning*, J. Mach. Learn. Res., 12 (2011), pp. 3259–3281.

[35] R. Vershynin, *Introduction to the non-asymptotic analysis of random matrices*, in Compressed sensing: Theory and applications, Y. C. Eldar and G. Kutyniok, eds., Cambridge University Press, Cambridge, UK, 2012.

[36] M. Yaghoobi, L. Daudet, and M. E. Davies, *Parametric dictionary design for sparse coding*, IEEE Trans. Signal Process., 57 (2009), pp. 4800–4810.

[37] J. Yang, J. Wright, T. S. Huang, and Y. Ma, *Image super-resolution via sparse representation*, IEEE Trans. Image Process., 19 (2010), pp. 2861–2873.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.