Recovering sparsely used overcomplete dictionaries via alternating minimization

Alekh Agarwal
Microsoft Research
alekha@microsoft.com

Animashree Anandkumar
UC Irvine
a.anandkumar@uci.edu

Prateek Jain
Microsoft Research
prajain@microsoft.com

Praneeth Netrapalli
UT Austin
praneethn@utexas.edu

Rashish Tandon
UT Austin
rashish@cs.utexas.edu

Abstract

We consider the problem of learning sparsely used overcomplete dictionaries, where each observation consists of a sparse combination of the mutually incoherent dictionary elements. We consider an iterative algorithm with the following alternating steps: 1) estimation of the dictionary coefficients for each observation through ℓ_1 minimization, given the dictionary estimate and 2) estimation of the dictionary elements through least squares, given the coefficient estimates. We establish that, under a set of sufficient conditions, our method converges at a linear rate in a local neightborhood of the true dictionary. Combined with recent techniques for initialization within this local neighborhood, our result provides an exact recovery guarantee for overcomplete and incoherent dictionaries.

1 Introduction

The problem of dictionary learning can be stated as follows: given observations $Y \in \mathbb{R}^{d\times n}$, the task is to decompose it as

$$Y = A^* X^*, \quad A^* \in \mathbb{R}^{d\times r}, X^* \in \mathbb{R}^{r\times n}. \tag{1}$$

A^* is referred to as the *dictionary* matrix and X^* is the *coefficient* matrix. r denotes the number of basis elements in this dictionary, and we consider the overcomplete setting where $r \geq d$. Without further constraints, the solution to (1) is not unique. A popular framework is to assume that the coefficient matrix X^* is sparse, and that each observation $Y_i \in \mathbb{R}^d$ is a sparse combination of the dictionary elements (i.e. columns of the dictionary matrix). This problem is known as *sparse coding* and it has been argued that sparse coding can provide a succinct representation of the observed data, given only unlabeled samples [9, 8].

Many practical dictionary learning methods focus on minimizing variants of the objective

$$\min_{A,X} \|Y - AX\|_F^2 + \lambda \|X\|_1. \tag{2}$$

It is challenging to provide guarantees on such procedures owing to the non-convexity of the objective. Indeed, the best results to our knowledge are those concerning the local optimality properties of A^*, X^* in some recent works [5, 6]. There have been other works which consider alternative formulations to solve the underlying dictionary learning problem. Notably, Spielman et al. [11] recently provided a method for guaranteed recovery when the dictionary matrix $A^* \in \mathbb{R}^{d\times r}$ has full

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

Input: Samples Y, initial dictionary estimate $A(0)$, accuracy sequence ϵ_t and sparsity level s. Thresholding function $\mathcal{T}_\rho(a) = a$ if $|a| > \rho$ and 0 o.w.

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, $\|Y_i - A(t)x\|_2 \leq \epsilon_t$.
4: **end for**
5: Threshold: $X(t+1) = \mathcal{T}_{9s\epsilon_t}(X(t+1))$.
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: $A(T)$

column rank. This implies that the number of dictionary elements $r \leq d$, where d is the observed dimension. In the overcomplete setting, the very recent works of Agarwal et al. [1, 2] provide methods for *approximate recovery* of the true dictionary. In this paper, we consider the convergence of alternating minimization procedures for optimizing the objective (2).

Summary of Results: Our main result concerns the convergence to the global optimum of alternating minimization. Our result requires initializatio with a dictionary with an error of at most $O\left(1/s^2\right)$ relative to the true dictionary. Further when $s = O\left(d^{1/6}\right)$ and number of samples satisfies $n = O\left(r^2/s^2\right)$, we establish the linear convergence of the alternating minimization procedure to the true dictionary. Combining our result with that of Agarwal et al. [1], where we initialize the alternating method using their solution as an initialization, we guarantee exact recovery of the true dictionary given that $s = O\left(d^{1/9}, r^{1/8}\right)$, and sufficient number of samples $n = O\left(r^2/s^2\right)$..

2 Algorithm

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 non-zero entries of v and W respectively. Let $\|w\|$ denote the ℓ_2 norm of vector w, and similarly for a matrix W, $\|W\|$ denotes its spectral norm. For a matrix X, X^i, X_i and X_j^i denote the i^{th} row, i^{th} column and $(i,j)^{\text{th}}$ element of X respectively.

As descibed earlier, 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 with high probability. Once we have an estimate of the coefficients, the dictionary is re-estimated 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 [13] or GraDeS [4]. More generally the exact lasso and least-squares steps may be replaced with other optimization methods for computational efficiency, e.g. [7].

3 Guarantees

In this section, we provide our main convergence result and also clearly specify all the required assumptions on A^* and X^*.

3.1 Assumptions and exact recovery result

We start by formally describing the assumptions needed for the main recovery result of this paper.

$(A1)$ **Incoherent Dictionary Elements:** Wlog, assume that all the elements are normalized: $\|A_i^*\| = 1$, for $i \in [r]$. We assume pairwise incoherence condition on the dictionary elements, for some constant $\mu_0 > 0$, $|\langle A_i^*, A_j^* \rangle| < \frac{\mu_0}{\sqrt{d}}$.

- $(A2)$ **Spectral Condition on Dictionary Elements:** The dictionary matrix has bounded spectral norm, for some constant $\mu_1 > 0$, $\|A^*\| < \mu_1 \sqrt{\frac{r}{d}}$.
- $(A3)$ **Non-zero Entries in Coefficient Matrix:** We assume that the non-zero entries of X^* are drawn i.i.d. from a zero-mean unit-variance distribution, 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 non-zero entries which are selected uniformly at random from the set of all s-sized subsets of $[r]$. We require s to satisfy $s < \frac{d^{1/6}}{c_2 \mu_1^{1/3}}$, for universal constant $c_2 > 0$.
- $(A5)$ **Sample Complexity:** For a given failure parameter $\delta > 0$, the number of samples n needs to satisfy $n \geq c_3 \frac{r^2}{s^2} \log \frac{1}{\delta}$, where $c_3 > 0$ is a universal constant.
- $A(6)$ **Initial dictionary with guaranteed error bound:** We assume that we have access to an initial dictionary estimate $A(0)$ such that $\epsilon_0 := \min_{i \in [r]} \min_{z \in \{-1,+1\}} \|zA_i^* - A(0)_i\|_2 < \frac{1}{2592 s^2}$.
- $(A7)$ **Choice of Parameters for Alternating Minimization:** Algorithm 1 uses a sequence of accuracy parameters $\epsilon_0 = 1/2592 s^2$ and $\epsilon_{t+1} = \frac{25050 \mu_1 s^3}{\sqrt{d}} \epsilon_t$.

Assumption $(A1)$ on normalization of dictionary elements is without loss of generality since we can always rescale the dictionary elements and the corresponding coefficients and obtain the same observations. However, the incoherence assumption is crucial in establishing our guarantees. In particular, incoherence also leads to a bound on the restricted isometry property (RIP) constant [10]. The assumption $(A2)$ provides a bound on the spectral norm of A^*. Note that the incoherence and spectral assumptions are satisfied with high probability (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 non-zero entries of X^*. Assumption$(A4)$ on sparsity in the coefficient matrix is crucial for identifiability of the dictionary learning problem. Assumption $(A5)$ provides a bound on sample complexity. Assumption $(A6)$ establishes the radius of the local neighborhood within which we need to initialize Algorithm 1 for our convergence results to hold. Assumption $(A7)$ specifies the choice of accuracy parameters used by alternating method in Algorithm 1. Due to Assumption $(A3)$ on sparsity level s, we have that $\frac{25050 \mu_1 s^3}{\sqrt{d}} < 1$ and the accuracy parameters in $(A7)$ form a decreasing sequence. This implies that in Algorithm 1, the accuracy constraint becomes more stringent with the iterations of the alternating method.

Given these assumptions, we now present our main result regarding the local linear convergence of Algorithm 1.

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 for all* $t \geq 1$*:*

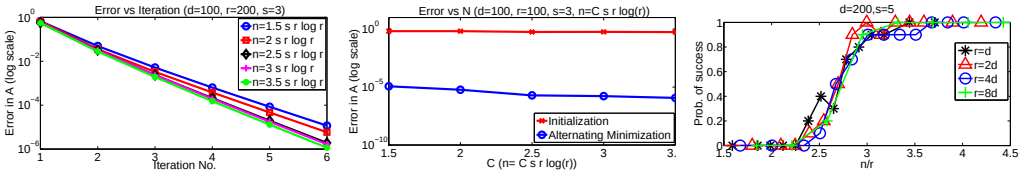
$$\min_{z \in \{-1,1\}} \|zA_i(t) - A_i^*\|_2 \leq \epsilon_t, 1 \leq i \leq r.$$

Remarks: The consequences of Theorem 1 are powerful combined with our Assumption $(A4)$ and the recurrence in $(A7)$ (since $(A4)$ ensures that ϵ_t forms a decreasing sequence). In particular, it is implied that with high probability we obtain,

$$\min_{z \in \{-1,1\}} \|zA(t)_i - A^*{}_i\|_2 \leq \|A(0) - A^*\|_2 2^{-t}.$$

Given the above bound, we need at most $O\left(\log_2 \frac{\epsilon_0}{\epsilon}\right)$ in order to ensure $\|zA(T)_i - A^*{}_i\|_2 \leq \epsilon$ for all 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 $O\left(1/s^2\right)$ for our alternating minimization procedure under these assumptions (since we require $\epsilon_0 \leq O\left(1/s^2\right)$).

We also observe that under the somewhat stronger assumption that $s = O\left(d^{1/9}, r^{1/8}\right)$, and further assuming a lower bound on the non-zero entries of X^*, it is possible to use the results of Agarwal et al. [1] or Arora et al. [2] to guarantee that Assumption $(A6)$ is met. This leads to the *exact recovery*



(a) (b) (c)

Figure 1: (a): Average error after each step alternating minimization step of Algorithm 1 on log-scale. (b): Average error after the initialization procedure of Agarwal et al. [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.

of the true dictionary in the context of the underlying dictionary learning problem. We also recall that the lasso step in Algorithm 1 can be replaced with a different robust sparse recovery procedure, with qualitatively similar results.

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 [3, 12]. 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 $\mathcal{N}(0,1)$. Note that, random Gaussian matrices are known to satisfy incoherence and the spectral norm bound [14]. The support of each column of X was chosen independently and uniformly from the set of all s-subsets of $[r]$. Similarly, each non-zero element of X was chosen independently from the uniform distribution on $[-2,-1] \cup [1,2]$. We use the GraDeS algorithm of [4] 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. As expected due to Theorem 1, we observe a geometric decay in the error. better for higher values of n.

One-shot vs iterative algorithm: It is conceivable that the initialization procedure itself is sufficient to obtain an estimate of the dictionary upto reasonable accuracy. alternatingminimization procedure of Algorithm 1. Figure 1(b) shows that this is not the case. The figure plots the error in recovery vs the number of samples used for both the approach of Agarwal et al. [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 non-trivial and also crucial to the success of alternating minimization- a random vector in $\mathbb{R}^d$ would give an error of $1 - \frac{1}{d} = 0.99$, where as the error after initialization procedure is ≈ 0.55.

Sample complexity: Finally, we study sample complexity requirement of the alternating minimization algorithm which is $n = O\left(r^2 \log r\right)$ according to Theorem 1, assuming good enough initialization. Figure 1(c) suggests that in fact only $O\left(r\right)$ 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 $\mathcal{N}(0, 0.5)$ 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 $r = O\left(d\right)$ is just $O(r)$.

References

[1] Alekh Agarwal, Animashree Anandkumar, and Praneeth Netrapalli. Exact recovery of sparsely used overcomplete dictionaries. *arXiv preprint arXiv:1309.1952*, 2013.

[2] S. Arora, R. Ge, and A. Moitra. New Algorithms for Learning Incoherent and Overcomplete Dictionaries. *ArXiv e-prints*, August 2013.

[3] Krishnakumar Balasubramanian, Kai Yu, and Guy Lebanon. Smooth sparse coding via marginal regression for learning sparse representations. In *ICML*, 2013.

[4] Rahul Garg and Rohit Khandekar. Gradient descent with sparsification: an iterative algorithm for sparse recovery with restricted isometry property. In *ICML*, 2009.

[5] Quan Geng, Huan Wang, and John Wright. On the local correctness of ℓ_1 minimization for dictionary learning. *arXiv preprint arXiv:1101:5672*, 2011. Preprint, URL:http://arxiv.org/abs/1101.5672.

[6] Rodolphe Jenatton, Rémi Gribonval, and Francis Bach. Local stability and robustness of sparse dictionary learning in the presence of noise. *arXiv preprint arXiv:1210.0685*, 2012.

[7] Rodolphe Jenatton, Julien Mairal, Francis R Bach, and Guillaume R Obozinski. Proximal methods for sparse hierarchical dictionary learning. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 487–494, 2010.

[8] Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Ng. Efficient sparse coding algorithms. In *Advances in neural information processing systems*, pages 801–808, 2006.

[9] Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.

[10] Holger Rauhut. Compressive sensing, structured random matrices and recovery of functions in high dimensions. In *Oberwolfach Reports*, volume 7, pages 1990–1993, 2010.

[11] Daniel A Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. In *Proc. of Conf. on Learning Theory*, 2012.

[12] Jayaraman J. Thiagarajan, Karthikeyan Natesan Ramamurthy, and Andreas Spanias. Learning stable multilevel dictionaries for sparse representation of images. *ArXiv 1303.0448*, 2013.

[13] J.A. Tropp and A.C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *Information Theory, IEEE Transactions on*, 53(12):4655–4666, 2007.

[14] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.