Exact Sparse Orthogonal Dictionary Learning

Kai Liu * Yongjian Zhao † Hua Wang ‡

Abstract

Over the past decade, learning a dictionary from input images for sparse modeling has been one of the topics which receive most research attention in image processing and compressed sensing. Most existing dictionary learning methods consider an over-complete dictionary, such as the K-SVD method, which may result in high mutual incoherence and therefore has a negative impact in recognition. On the other side, the sparse codes are usually optimized by adding the $\ell_0$ or $\ell_1$-norm penalty, but with no strict sparsity guarantee. In this paper, we propose an orthogonal dictionary learning model which can obtain strictly sparse codes and orthogonal dictionary with global sequence convergence guarantee. We find that our method can result in better denoising results than over-complete dictionary based learning methods, and has the additional advantage of high computation efficiency.

1 Introduction

In recent decades, representing natural images through sparse models have been an important topic in image processing Liu et al. [2018a] and compressed sensing. It is widely acknowledged that sparse models are very useful tools for recognition tasks and image restoration. Generally speaking, sparse image model assumes that local image patches can be represented by a linear combination of basic elements with sparse coefficients Liu et al. [2019c, 2018b], which is so-called atoms while the whole collection of these atoms is named as a dictionary Liu [2019]. Earlier works on designing sparse dictionary learning usually focus on the design of fixed orthogonal dictionaries, e.g. local discrete cosine transform (DCT) Saha [2000], wavelets Chui [2016], Mallat [2008]. These orthogonal dictionaries and their over-complete extensions remain important tools in many image restoration tasks for their simplicity and efficiency Cai et al. [2008, 2012]. Recently, there have been great progresses on constructing dictionaries adaptive to the input image via some learning process Lewicki and Sejnowski [2000], Elad and Aharon [2006], Mairal et al. [2009]. The basic idea is to learn the dictionary adaptive to the target image so as to achieve better sparsity than the fixed ones. As a result, most existing dictionary learning methods consider an over-complete dictionary and formulate the learning process as a minimization problem. Taking the popular K-SVD method Elad and Aharon [2006] for example, the K-SVD method learns an over-complete dictionary from an input image by solving the following minimization model:

$$\min_{U,\{v_i\}} \sum_i \|x_i - U v_i\|_2^2 + \lambda \|v_i\|_0,$$

(1)
where $\|\cdot\|_0$ is the sparsity measure defined as the number of non-zero entries in the input. $X = (x_1, \ldots, x_n)$ is $n$ image patches after vectorization while $U = (u_1, \ldots, u_r)$ denotes the dictionary with $u_i$ represents the atom.

The problem in Eq. (1) is challenging due to its non-convexity. Elad and Aharon [2006] propose an iteration method which alternatively optimizes sparse coding $V = (v_1, \ldots, v_n)$ and dictionary $U$. However, this method is known to be computational demanding. Since then, there have been some methods proposed to modify the model in Eq. (1) by relaxing the nonconvex $\ell_0$-norm to a convex one (e.g. $\ell_1$-norm) [Liu and Wang [2015, 2018], Liu et al. [2019a], Yang et al. [2019], Brand et al. [2020] to obtain approximated solutions faster [Mairal et al. [2009], Wu et al. [2018] or by focusing on reducing the computation complexity [Rubinstein et al. [2008], Gilboa et al. [2018]].

2 Motivation and Our Contributions

The high computational demand of K-SVD method is because dictionary $U$ is redundant with no constraints on the atom’s correlations.

The quality of dictionary for sparse coding is usually measured by mutual incoherence defined by Donoho and Elad [2003]:

$$\mu(U) = \max_{i \neq j} \frac{\langle u_i, u_j \rangle}{\|u_i\|_2 \|u_j\|_2}$$

which measures the correlations between different atoms. Using matching pursuit methods [Tropp [2004]] requires $\mu(U)$ to be sufficiently small to guarantee the performance of sparse coding. In fact, the constant $\mu$ obtained via the K-SVD method usually is not small, since there is no constraints imposed on dictionary $U$.

More recently, there are some works focusing on reducing dictionary redundancy. Among those, orthogonal dictionary learning [Bao et al. [2013]], which learns a strictly orthogonal dictionary, has attracted much attention. Also, there are some other works trying to incorporate the orthogonal constraint as a penalty into the objective [Bao et al. [2014b]], which have yielded promising results. However, there are some concerns in the work mentioned above:

- Incorporating the orthogonal constraint into the objective doesn’t necessarily guarantee the learned dictionary to be strictly orthogonal.
- To guarantee the sparsity of $V$, most works try to minimize the $\ell_0$ or $\ell_1$-norm of $V$ as a penalty term. However, unless the penalty parameter is sufficiently large, the optimized $V$ is not always sparse.
- Though [Bao et al. [2013]] claims their algorithm is fast, it is not the case in practice: there involves a hard threshold to update $V$ in addition to Singular Value Decomposition (SVD) operation in optimizing $U$ within each iteration. Moreover, extensive experiments (as in later section) show that its denoising result is poor.
- Vanilla alternating minimization based methods, such as [Bao et al. [2013]], can only guarantee that the objective decreases monotonically. However, the generated sequence convergence property is not theoretically guaranteed, where oscillation may happen and the case can be found in K-SVD [Elad and Aharon [2006], Bao et al. [2014a]].

In response to the concerns mentioned above, in this paper, we propose a new algorithm to tackle the problems, and we list our contributions explicitly as follows:
• Our algorithm will learn a strictly orthogonal dictionary which avoids the mutual incoherence issue.
• We will obtain a strictly sparse $V$, with each column $\|v\|_0 \leq s$, where $s$ is a hyper-parameter.
• Our method is computationally efficient, and we find the time consumption is only 1/4 as Bao et al. [2014a] and even faster than Bao et al. [2013] in various experimental settings.
• The denoising results of our method are significantly better than Bao et al. [2013] and even better than over-complete dictionary method such as Bao et al. [2014a].
• The objective function decreases monotonically and the generating sequences $U$ and $V$ are global sequence convergent with at least sub-linear convergence rate.

3 Formulation And Algorithm

Vanilla sparse approximation for an input signal can be formulated as the following optimization problem:

$$\min_v \|v\|_0, \text{ s.t. } \|x - Uv\|_2^2 \leq \epsilon.$$  (3)

Since $\ell_0$-norm optimization is NP-hard, most existing methods either use greedy algorithms to iteratively update the solution (e.g. orthogonal matching pursuit (OMP) [Tropp 2004]), or relax to convex norm (e.g. basis pursuit [Chen et al. 2001]). A widely used optimization formulation is:

$$\min_{U,V} \|X - UV\|_F^2 + \theta \|V\|_p,$$  (4)

where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix, which is a trade off between linear approximation loss and sparsity of $V$. When $p = 0$, it turns to be the original sparsity definition; while $p = 1$, it is a relaxation to convex ($\ell_1$-norm) optimization problem. However, it is worth noting that the sparsity property can only be approximately guaranteed when $\theta$ is sufficiently large, which inevitably involves a lot of parameter tuning work.

Different from transferring coding sparsity constraint to the objective function, some research focuses on hard sparse solution via optimizing:

$$\min_{U,V} \|X - UV\|_F^2 \text{ s.t. } \forall i, \|v_i\|_0 \leq s,$$  (5)

and among these includes K-SVD, which is widely used for image denoising. Leveraging alternating optimization between sparse approximation through OMP and dictionary updating via column-wise SVD updates, K-SVD can obtain local minimum while the objective is monotonically decreasing. However, as Bao et al. [2014a] points out, the generated sequence $V = \{V_1, V_2, \ldots, V_k\}$ is not convergent, where the gap between any two consecutive iterations is non-zero.

3.1 Our Formulation

To reduce dictionary redundancy defined in Eq. (2), instead of learning an over-complete dictionary, we set orthogonality constraint such that the mutual incoherence between any two $u$ is 0. In addition, to guarantee the strict sparsity solution, we set hard threshold on $v$. Moreover, the factored nuclear norm regularization ($\|U\|_F^2 + \|V\|_F^2)/2$ (c.f. Recht et al. [2010], Li et al. [2017]) is added to the objective function to further
promote the low-rankness of the both recovered dictionary $U$ and the coefficient matrix $V$. As is shown in [Li et al., 2017], the regularization term $(\|U\|_F^2 + \|V\|_F^2)/2$ plays an equivalent role in promoting the low-rankness as the matrix nuclear norm $\|UV^T\|_*$. Further, since no high-complexity singular value decompositions are required using the factorization nuclear norm term, the resulting algorithm can be much faster and more scalable to large-data situations. As a whole, we formulate our objective function as:

$$\min_{U,V} \frac{1}{2} \|X - UV\|_F^2 + \theta \left(\frac{\|U\|_F^2}{2} + \|V\|_F^2\right)$$

(6)

with the constraints to be:

$$U := \{U|U^TU = I\}, V := \{V||V(:,j)||_0 \leq s, \forall j \in [0, n]\}.$$  

(7)

3.2 Proposed Algorithm

By noticing the orthogonality constraint of $U$, we can reformulate Eq. (6) as:

$$h(U,V) = \|X - UV\|_F^2 + \theta \left(\frac{\|U\|_F^2}{2} + \|V\|_F^2\right), \text{ s.t. } U \in \mathbb{U}, V \in \mathbb{V}.$$  

(8)

For past decades, proximal algorithm has been successfully applied to a wide variety of situations: convex optimization, nonmonotone operators [Combettes and Pennanen, 2004, Kaplan and Tichatschke, 1998] with various applications to nonconvex programming. More recently, proximal alternating linearized minimization (PALM) was introduced by Bolte et al. [2014] as a linearized approximation regularization in optimization. Though only a ‘conceptual’ algorithm was proposed in the paper, still it has attracted a lot of attention and been widely utilized in machine learning research.

Given the fact that objective function in Eq. (8) is nonconvex w.r.t. $U$ and $V$, and that the constraint on $U$ and $V$ are also nonconvex. Therefore, we consider adding a linearized proximal term and optimize the solution as:

$$U_{k+1} = \arg\min_{U^TU = I} \langle U - U_k, \nabla_U h(U_k, V_k) \rangle + \frac{\mu}{2} \|U - U_k\|_F^2$$

$$v_{k+1} = \arg\min_{\|v\|_0 \leq s} \langle v - v_k, \nabla_v h(U_{k+1}, v_k) \rangle + \frac{\lambda}{2} \|v - v_k\|^2.$$  

(9)

where $\mu, \lambda$ are hyper-parameters and the settings will be discuss in later section.

**Remark 1.** To avoid drastic changes, we add the proximal term to make the new updating solution not too far from the previous step. One can see that when the proximal term regularization parameters $\mu, \lambda$ are sufficiently large, they will dominate the objective function. Moreover, we can take the linearized minimization as to minimize the objective with Taylor expansion by making use of the first order (linear) information.

Accordingly, to optimize $U$, we have:

$$U_{k+1} = \arg\min_{U^TU = I} \langle U - U_k, \nabla_U h(U_k, V_k) \rangle + \frac{\mu}{2} \|U - U_k\|_F^2$$

$$= \arg\min_{U^TU = I} \langle U - U_k, \nabla_U h(U_k, V_k) \rangle - \mu \langle U, U_k \rangle$$

$$= \arg\max_{U^TU = I} \text{tr}(U^T M) = YZ^T,$$  

(10)
Algorithm 1 Alternating Linearized Minimization for Problem Eq. (8)

Input: data $X \in \mathbb{R}^{m \times n}$, atom size $r$, regularization parameters $\lambda, \mu, \theta$, number of iterations $K$

Initialization: $U_0 \in \mathbb{R}^{m \times r}$, $V_0 \in \mathbb{R}^{r \times n}$

while $k \leq K$ do
  optimize $U_{k+1}$ by Eq. (10)
  optimize each $v_{k+1}$ by Eq. (11)
end while

Output: $U_K$ and $V_K$

where $M = 2(X - U_k V_k) V_k^T + \mu U_k$ and $Y, Z$ is obtained from $[Y, \Sigma, Z] = \text{SVD}(M)$.

Then we optimize $v_{k+1}$ given $U^T U = I$:

$$v_{k+1} = \arg \min_{\|v\|_0 \leq s} \left\langle v - v_k, \nabla_{v_k} h(U_{k+1}, v_k) \right\rangle + \frac{\lambda}{2} \|v - v_k\|^2_2$$

$$= \arg \min_{\|v\|_0 \leq s} \frac{\lambda}{2} \|v - v_k + \nabla_{v_k} h(U_{k+1}, v_k)\|^2_2$$

$$= \arg \min_{\|v\|_0 \leq s} \|v - q\|^2_2,$$

where $q = v_k - ((\theta + 2)v_k - 2U_{k+1}^T x)/\lambda$.

It is obvious that the optimized $v_{k+1}$ above is obtained by retaining the top $s$ absolute value of $q$ while setting the rest to be 0.

4 Convergence Analysis

In the following case, we let $U$ and $V$ be as defined in Eq. (7), and show the convergence of our proposed algorithm in the last section.

To begin with, we first show that $h(U, V)$ has Lipschitz continuous gradient at $U \in U, V \in V$, which will be very useful for convergence analysis. Before that we need give the following proposition:

Proposition 1. During each iteration in Algorithm 1, $\|V_k\|^2_F$ is bounded.

Proof. By the definition of alternating proximal minimization as well as what later sections will demonstrate, the objective in each iteration is monotonically decreasing, thus we have:

$$h(W_k) \leq h(W_{k-1}) \leq \cdots \leq h(W_1) \leq h(W_0) < \infty \quad (12)$$

where $\{W_k\}_{k \geq 0} = \{(U_k, V_k)\}_{k \geq 0}$, therefore:

$$\frac{\theta}{2} \|V_k\|^2_F \leq \|X - U_k V_k\|^2_F + \frac{\theta}{2} \|V_k\|^2_F = h(W_k) = h(U_k, V_k) \leq h(W_0) = h(U_0, V_0) \quad (13)$$

thus $\|V\|^2_F \leq \frac{2h(U_0, V_0)}{\theta}$. \hfill \Box

Proposition 2. $h(U, V)$ has Lipschitz continuous gradient at $U \in U, V \in V$, where $U$ and $V$ are defined in Eq. (7). That is, there exists a constant $L_c$ such that

$$\|\nabla h(U, V) - \nabla h(U', V')\|_F \leq L_c \|(U, V) - (U', V')\|_F \quad (14)$$
for all \( U, U' \in U \) and \( V, V' \in V \). Here \( L_c > 0 \) is referred to as the Lipschitz constant.

**Proof of Proposition 2** It is equivalent to show \( \| \nabla^2 h(U,V) \|_2 \leq L_c \) for all \( U \in U, V \in V \). Standard computations give the Hessian quadrature form \( [\nabla^2 h(U,V)](\Delta, \Delta) \) for any \( \Delta = \begin{bmatrix} \Delta_U \\ \Delta_V \end{bmatrix} \in \mathbb{R}^{(n+m) \times r} \) (where \( \Delta_U \in \mathbb{R}^{m \times r} \) and \( \Delta_V \in \mathbb{R}^{r \times n} \)) as

\[
[\nabla^2 h(U,V)](\Delta, \Delta) = \| \Delta_U V + U \Delta_V \|_F^2 + 2 \langle UV - X, \Delta_U \Delta_V \rangle,
\]

which gives:

\[
\| \nabla^2 h(U,V) \|_2 = \max_{\| \Delta \|_F = 1} \| \nabla^2 h(U,V)](\Delta, \Delta) \|
\leq \max_{\| \Delta \|_F = 1} \| \Delta_U V + U \Delta_V \|_F^2 + 2 \| \langle UV - X, \Delta_U \Delta_V \rangle \| + \theta
\leq \theta + 2(\| U \|_F^2 + \| V \|_F^2 + \| U \|_F \| V \|_F + \| X \|_F) =: L_c,
\]

where the inequality follows from \( \| A, B \| \leq \| A \|_F \| B \|_F \) and \( \| CD \|_F \leq \| C \|_F \| D \|_F \) (Cauchy–Schwarz inequality). Due to the constraints on \( U \), we have \( \| U \|_F^2 = \text{tr}(U^T U) = \text{tr}(I) = \text{rank} \| V \|_F^2 \leq \frac{2h(U, V)}{\theta} \) according to Proposition 1.

To analyse the convergence, we rewrite Eq. (8) as

\[
\min_{U, V} f(U, V) = h(U, V) + \delta_U(U) + \delta_V(V),
\]

where \( \delta_U(U) = \begin{cases} \begin{array}{ll} 0, & U \in U \\ \infty, & U \notin U \end{array} \end{cases} \) is the indicator function of the set \( U \) and therefore nonsmooth, so is \( \delta_V(V) \).

The following result establishes that the subsequence convergence property of the proposed algorithm, i.e., the sequence generated by Algorithm 1 is bounded and any of its limit point is a critical point of Eq. (17).

**Theorem 1** (Subsequence convergence). Let \( \{W_k\}_{k \geq 0} = \{(U_k, V_k)\}_{k \geq 0} \) be the sequence generated by Algorithm 1 with \( \min \{\lambda, \mu\} > L_c \). Then the sequence \( \{W_k\}_{k \geq 0} \) is bounded and obeys the following properties:

**(P1) sufficient decrease:**

\[
f(W_{k-1}) - f(W_k) \geq \frac{\min(\lambda, \mu) - L_c}{2} \| W_k - W_{k-1} \|_F^2
\]

implying

\[
\lim_{k \to \infty} \| W^{k-1} - W^k \|_F = 0.
\]

**(P2) denote** \( \mathbb{C}(W_0) \) (depending on \( W_0 \)) as the set of all limit points of the iterates \( \{W_k\} \). Then all the limit points \( W^* \) are critical points of \( f \) and have the same function value

\[
f(W^*) = f.
\]

Further, \( \mathbb{C}(W_0) \) is a nonempty, compact and connected set and satisfies

\[
\text{dist}(W_k, \mathbb{C}(W_0)) = 0
\]
Proof of Theorem 1. (P1): First note that for all $k$, according to our alternating minimization method, we always have $\delta_U(U_k) = \delta_{\bar{V}}(V_k) = 0$ and thus $f(W_k) = h(W_k)$.

Since $h(U, V)$ has Lipschitz continuous gradient at $U \in U, V \in V$ with Lipschitz gradient $L_c$ and $\lambda > L_c$, we define $h_{L_c}(U, U', V)$ as proximal regularization of $h(U, V)$ linearized at $U', V$: 

$$h(U', V) + \langle \nabla_U h(U', V), U - U' \rangle + \frac{L_c}{2} \|U - U'\|_F^2,$$

By the definition of Lipschitz continuous gradient and Taylor expansion, we have 

$$h(U, V) \leq h_{L_c}(U, U', V).$$ (22)

Also by the definition of proximal map, we get: 

$$U_k = \arg \min_U \delta_U(U) + \frac{\mu}{2} \|U - U_{k-1}\|_F^2 + \langle \nabla_U h(U_{k-1}, V_{k-1}), U - U_{k-1} \rangle$$ (23)

and hence we take $U_k = U$, which implies that 

$$\delta_U(U_k) + \frac{\mu}{2} \|U_k - U_{k-1}\|_F^2 + \langle \nabla_U h(U_{k-1}, V_{k-1}), U_k - U_{k-1} \rangle \leq \delta_U(U_{k-1})$$ (24)

Combining Eq. (22) to Eq. (24), we have:

$$h(U_k, V_{k-1}) + \delta_{\bar{V}}(U_k) \leq h(U_{k-1}, V_{k-1}) + \langle \nabla_U h(U_{k-1}, V_{k-1}), U_k - U_{k-1} \rangle + \frac{L_c}{2} \|U_k - U_{k-1}\|_F^2 + \delta_{\bar{V}}(U_k)$$

$$\leq h(U_{k-1}, V_{k-1}) + \frac{L_c}{2} \|U_k - U_{k-1}\|_F^2 + \delta_{\bar{V}}(U_{k-1}) - \frac{\mu}{2} \|U_k - U_{k-1}\|_F^2$$

$$= h(U_{k-1}, V_{k-1}) + \delta_{\bar{V}}(U_{k-1}) - \frac{\mu - L_c}{2} \|U_k - U_{k-1}\|_F^2,$$

Similarly, we have 

$$h(U_k, V_k) - h(U_k, V_{k-1}) + \delta_{\bar{V}}(V_k) - \delta_{\bar{V}}(V_{k-1}) \leq -\frac{\lambda - L_c}{2} \|V_k - V_{k-1}\|_F^2$$ (26)

which together with the above equation gives Eq. (18). Now repeating Eq. (18) for all $k$ will give

$$(\min(\lambda, \mu) - L_c) \sum_{k=1}^{\infty} \|W_k - W_{k-1}\|_F^2 \leq 2f(W_0),$$ (27)

which gives Eq. (18).

(P2) Any convergent subsequence $\{W_k' = (U_k', V_k')\}$ from $\{W_k\}$ and denote the limit point of this subsequence as $W^*$. Since $U_k' \in U, V_k' \in V$ for all $k'$ and both of the sets $U$ and $V$ are closed, we have $U^* \in U, V^* \in V$. Since $h$ is continuous, we have

$$\lim_{k' \to \infty} f(W_k') = \lim_{k' \to \infty} h(U_k', V_k') + \delta_{\bar{V}}(U_k') + \delta_{\bar{V}}(V_k')$$

$$= f(W^*),$$

which together with the fact that $\bar{f} = \lim_{k \to \infty} f(W_k)$ gives Eq. (20).
To show $W^*$ is a critical point, we first consider Eq. (9) and the optimality condition yields:

$$\nabla_U h(U_{k-1}, V_{k-1}) + \mu(U_k - U_{k-1}) + \partial \delta_U(U_k) = 0. \quad (28)$$

Similarly, we have

$$\nabla_V h(U_k, V_{k-1}) + \lambda(V_k - V_{k-1}) + \partial \delta_V(V_k) = 0. \quad (29)$$

Now, define

$$\nabla_U h(U_k, V_k) + \partial \delta_U(U_k) \quad \text{and} \quad \nabla_V h(U_k, V_k) + \partial \delta_V(V_k).$$

Thus, we have

$$A_k \in \partial_U f(U_k, V_k), \quad B_k \in \partial_V f(U_k, V_k). \quad (30)$$

Following Eq. (28) and the triangle inequality we have:

$$\lim_{k \to \infty} \|A_k\|_F = \lim_{k \to \infty} \|\nabla_U h(U_k, V_k) - \nabla_U h(U_{k-1}, V_{k-1}) - \mu(U_k - U_{k-1})\|_F$$

$$\leq \lim_{k \to \infty} \|\nabla_U h(U_k, V_k) - \nabla_U h(U_{k-1}, V_{k-1})\|_F + \mu\|U_k - U_{k-1}\|_F$$

$$\leq \lim_{k \to \infty} (L_c + \mu)\|W_k - W_{k-1}\|_F = 0. \quad (31)$$

Similarly, we have:

$$\lim_{k \to \infty} \|B_k\|_F \leq \lim_{k \to \infty} (L_c + \lambda)\|W_k - W_{k-1}\|_F = 0. \quad (32)$$

Then we have:

$$\text{dist}(0, \partial f(W_k)) \leq \mathcal{L}_\theta\|W_k - W_{k-1}\|_F \quad (33)$$

where $\mathcal{L}_\theta := (2L_c + \mu + \lambda)$. Owing to the closedness properties of $\partial f(W_k)$, we finally obtain $0 \in \partial f(W^*)$. Thus, $W^*$ is a critical point of $f$.

**Remark 2.** In our proposed algorithm, since in every update, our solution is closed while satisfying the constraints, thus in fact $\delta_U$ and $\delta_V$ are 0, and $\infty$ is never achieved.

**Theorem 2** (Sequence convergence). The sequence $\{W_k\}_{k \geq 0}$ generated by Algorithm 1 with $\lambda, \mu > L_c$ is global-sequence convergence.

**Remark 3.** Theorem 2 is much stronger than Theorem 1, since we are not guaranteed that the iterates $\{W_k\}$ generated by Algorithm 1 would converge to a limit point only by Theorem 1. Theorem 2 fulfills this gap by directly showing $\{W_k\}$ generated by Algorithm 1 converges to a critical point, and as an consequence, the set of limit points $\bar{C}(W_0)$ becomes a singleton.

**Definition 1** (Kurdyka-Lojasiewicz (KL) property). [Bolte et al., 2007] We say a proper semi-continuous function $h(u)$ satisfies Kurdyka-Lojasiewicz (KL) property, if $\overline{u}$ is a critical point of $h(u)$, then there exists $\delta > 0$, $\theta \in [0, 1)$, $C_1 > 0$, s.t.

$$|h(u) - h(\overline{u})|^\theta \leq C_1 \text{dist}(0, \partial h(u)), \quad \forall u \in B(\overline{u}, \delta) \quad (34)$$

8
Proof of Theorem 2. According to Theorem 1 there exists a positive integer \( k_0 \) so that \( \text{dist}(W_k, \mathcal{C}(W_0)) \leq \delta_0 \) for all \( k \geq k_0 \). Now using KL property, we have:

\[
[f(W_k) - f(W^*)]^{1-\theta_{KL}} - [f(W_{k+1}) - f(W^*)]^{1-\theta_{KL}} \
\geq (1 - \theta_{KL}) \frac{f(W_k) - f(W_{k+1})}{f(W_k) - f(W^*)} \theta_{KL} \\
\geq \frac{\lambda(1 - \theta_{KL})}{2C_{KL}} \frac{\|W_k - W_{k+1}\|_F^2}{\text{dist}(0, \partial f(W_k))} \\
\geq \frac{\lambda(1 - \theta_{KL})}{2C_{KL}L_q} \frac{\|W_k - W_{k+1}\|_F^2}{\|W_k - W_{k-1}\|_F} \\
= \kappa (\|W_k - W_{k+1}\|_F^2 + \|W_k - W_{k-1}\|_F) - \kappa \|W_k - W_{k-1}\|_F \\
\geq \kappa (2\|W_k - W_{k+1}\|_F - \|W_k - W_{k-1}\|_F)
\]

where we have used Eq. (35) in the third line and the fourth line comes from Eq. (33). Accordingly, we have:

\[
2\|W_k - W_{k+1}\|_F - \|W_k - W_{k-1}\|_F \leq \beta \left( [f(W_k) - f(W^*)]^{1-\theta_{KL}} - [f(W_{k+1}) - f(W^*)]^{1-\theta_{KL}} \right)
\]

with \( \kappa := \frac{\lambda(1 - \theta_{KL})}{2C_{KL}L_q} \) and \( \beta := \left( \frac{\lambda(1 - \theta_{KL})}{2C_{KL}L_q} \right)^{-1} \).

Summing up the above inequality from \( \tilde{k} > k_0 \) to infinity:

\[
\sum_{k=\tilde{k}}^{\infty} \|W_k - W_{k+1}\|_F \leq \|W_{\tilde{k}} - W_{\tilde{k}-1}\|_F + \beta [f(W_{\tilde{k}}) - f(W^*)]^{1-\theta_{KL}}
\]

implying \( \sum_{k=\tilde{k}}^{\infty} \|W_k - W_{k+1}\|_F < \infty \),

which indicates sequence \( \{W_k\} \) is Cauchy, and hence convergent. Therefore, the limit point set \( \mathcal{C}(W_0) \) is singleton \( W^* \) [Zhu et al., 2018; Liu et al., 2019b].

\[ \square \]

5 Experiments

Experiments are conducted in MATLAB R2016b (64bit) Linux version on a desktop with an Intel CPU (3.4GHZ) and 64G memory. The initial dictionary is generated by the local DCT transform: either 8 \( \times \) 8 or 16 \( \times \) 16. The image patches for training are uniformly selected from the input image at random. For image size 512 \( \times \) 512, about 4 \( \times \) 10^4 patches are used for training.

Fig. 1 shows that the sparsity of \( \nu \) does play a role in the denoising results: when \( s \) increases, the performance has a trend to degenerate. Therefore in our experiment setting we choose \( s = 10 \).
Figure 1: Denoising results correspond to different values of $s = [5, 10, 20, 40]$ when $\sigma = 25$. We see that generally speaking when $s$ is too small or large, the results are not as good as when $s = 10$, thus in our experiments we set $s = 10$ for comparison.

Figure 2: Time consumption (Upper row) and denoising PSNR comparison (Lower row). The second and third columns are conducted on image ‘House’ with various $\sigma$ and patch sizes. The methods compared with are [Bao et al. 2013], [Bao et al. 2014b] respectively. The atom size, by default, is $16 \times 16$.

### 5.1 Complexity Analysis

We first analyze the complexity comparison between our method and [Bao et al. 2013]. The most computationally intensive steps in our algorithm are computing SVD in optimizing $U$, where the complexity is $O(r^3)$. On the other hand, the complexity of [Bao et al. 2013] mainly depends on two parts: hard threshold computing to optimize $V$ and SVD in optimizing $U$. The former part takes $O(nr)$ and the latter part takes $O(r^3)$. Thus the total complexity is $O(nr + r^3)$. 

10
5.2 Image Denoising

We conducted the experiment on six different pictures shown in Fig. 4. For comparison, we set the parameters for other models as suggested in the original papers while for our method, we set $s = 10$, $\lambda = \mu = \theta = 200$. After the dictionary is learned by training samples, the image is denoised using the coefficients from the OMP method under the learned dictionary in one pass. The results is compared to the DCT-based thresholding method and the K-SVD denoising method Elad and Aharon [2006] with patch size $16 \times 16$. Table 1 shows the list of PSNR values while Fig. 2, Fig. 3, Fig. 6 and 7 show the denoising results. Dictionaries are reported in Fig. 5. We see our proposed method always has the smallest time consumption and achieves the best denoising result in most cases.

| Image    | Lena          | Barbara         | Peppers         |
|----------|---------------|-----------------|-----------------|
| $\sigma$ | 5 10 15 20 25 | 5 10 15 20 25   | 5 10 15 20 25   |
| DCT;     | 38.29 35.25 33.39 32.03 30.96 | 37.16 33.12 31.01 29.65 28.67 | 37.06 34.48 33.02 31.89 30.95 |
| K-SVD Elad and Aharon [2006]; | 38.59 35.47 33.70 32.38 31.32 | 37.61 33.62 31.45 30.13 29.11 | 37.77 34.72 32.37 32.26 31.39 |
| FODL Bao et al. [2013] | 38.00 34.43 32.16 31.67 30.42 | 37.10 32.01 29.80 28.45 27.88 | 34.09 33.87 31.86 30.02 29.76 |
| LNDL Bao et al. [2014b]; 38.49 35.41 33.57 32.25 31.19 | 37.46 33.47 31.43 30.02 29.00 | 37.68 34.64 33.22 32.14 31.18 |
| Ours;    | 39.24 36.26 35.19 32.00 30.98 | 38.66 34.52 31.47 30.01 29.06 | 38.55 35.66 33.64 31.99 30.87 |

| Image    | Boat          | Fingerprint     | Hill            |
|----------|---------------|-----------------|-----------------|
| $\sigma$ | 5 10 15 20 25 | 5 10 15 20 25   | 5 10 15 20 25   |
| DCT;     | 36.79 33.49 31.34 29.96 28.90 | 28.64 28.29 26.85 | 28.54 32.93 31.11 30.02 29.00 |
| K-SVD Elad and Aharon [2006]; | 37.17 33.64 31.73 30.36 29.28 | 36.59 32.39 30.06 28.47 27.26 | 36.99 33.34 31.43 30.17 29.19 |
| FODL Bao et al. [2013]; 36.89 32.46 30.15 29.32 27.65 | 35.19 30.56 28.12 26.55 25.58 | 35.43 31.92 29.87 28.89 27.88 |
| LNDL Bao et al. [2014b]; 37.02 33.53 31.64 30.20 29.16 | 36.59 32.35 29.97 28.28 27.03 | 36.94 33.31 31.29 30.02 29.06 |
| Ours;    | 38.01 34.34 32.21 30.00 28.98 | 37.45 33.31 30.75 28.00 26.45 | 36.87 34.12 31.98 29.39 28.87 |

Table 1: PSNR values of the denoised results
5.3 Face Recognition

Dictionary learning can also be applied to recognition tasks. In this subsection, the performance is evaluated on two face datasets: Extended YaleB dataset \cite{Georghiades2001} and AR face dataset \cite{Martinez1998}. Our approach is compared to three K-SVD based methods: LC-KSVD \cite{Jiang2011}, D-KSVD \cite{Zhang2010}, and KSVD \cite{Elad2006}, in addition to FODL \cite{Bao2012}, CIDL \cite{Bao2013}, LNDL \cite{Bao2014}, WDL \cite{Schmitz2018}, DFEDL \cite{Li2019} and RDCDL \cite{Lin2018}.

**Extended YaleB Database:** The extended YaleB database \cite{Georghiades2001} contains 2,414 images of 38 human frontal faces under about 64 illumination conditions and expressions. There are about 64 images for each person. The original images are cropped to $192 \times 168$ pixels. We follow \cite{Zhang2010}.

---

### Table 2: Training time (seconds) on two face datasets.

| Dataset     | K-SVD | D-KSVD | LC-KSVD | FODL  | CIDL  | LNDL  | RDCDL | WDL  | DFEDL | Ours  |
|-------------|-------|--------|---------|-------|-------|-------|-------|------|-------|-------|
| YaleB       | 48.76 | 72.52  | 201.46  | 9.15  | 13.46 | 25.76 | 20.48 | 52.23| 17.56 | 8.87  |
| AR Face     | 62.33 | 85.42  | 278.89  | 18.47 | 26.96 | 40.09 | 36.77 | 65.80| 34.91 | 17.80 |

### Table 3: Classification accuracies (%) on two face datasets.

| Dataset     | K-SVD | D-KSVD | LC-KSVD | FODL  | CIDL  | LNDL  | RDCDL | WDL  | DFEDL | Ours  |
|-------------|-------|--------|---------|-------|-------|-------|-------|------|-------|-------|
| YaleB       | 93.10 | 94.10  | 95.00   | 93.48 | 95.66 | 94.56 | 92.43 | 95.89| 94.86 | 94.98 |
| AR Face     | 86.50 | 88.80  | 93.70   | 93.28 | 94.41 | 92.77 | 89.80 | 93.41| 93.18 | 94.87 |

---
Figure 5: The dictionaries learned from the ‘Lena’ (Left) ‘Boat’ (Center) and ‘Fingerprint’ (Right) with \( \sigma = 30 \) using FODL Bao et al. [2013] (Upper), LNDL Bao et al. [2014b] (Middle) and ours (Bottom). The atom size is 16 \( \times \) 16. We can see since we set \( U \) to be orthogonal, the dictionaries we learned has no mutual incoherence, which is very different from the second row by LNDL Bao et al. [2014b].

AR Face Database: The AR face dataset Martinez [1998] consists of over 4000 frontal images from 126 individuals. For each individual, 26 pictures were taken in two separate sessions. The main characteristic of the AR database is that it includes frontal views of faces with different facial expressions, lighting conditions and occlusion conditions. Again, we follow Zhang and Li [2010], Bao et al. [2014b] for the experiment setting. We see from Table 2 and Table 3 that our approach performs very competitively in comparison to other methods in terms of accuracy and computational efficiency, which comes from two aspects: first, the alternating minimization method converges very fast; second, we don’t train so many redundant atoms.

6 Conclusion

In this paper, we study the dictionary learning problem where the atom is orthogonal and the code is sparse. We propose a proximal alternating linearized minimization method to obtain the solution with theoretical guarantee. Extensive experiments on various real-world datasets illustrate the superiority of our method over
Figure 6: Various $\sigma$ from $[5, 10, 15, 20, 25, 30]$ correspond to different rows, with columns being: Noisy Images, denoised images by FODL [Bao et al. 2013], LNDL [Bao et al. 2014b] and ours (from left to right). The atom size is $16 \times 16$. 
Figure 7: Various $\sigma$ from $[5, 10, 15, 20, 25, 30]$ correspond to different rows, with columns being: Noisy Images, denoised images by FODL [Bao et al. 2013], LNDL [Bao et al. 2014b] and ours (from left to right). The atom size is $16 \times 16$. 
SOTA in terms of both time consumption and image denoising, indicating its promising applications.

References

Chenglong Bao, Jian-Feng Cai, and Hui Ji. Fast sparsity-based orthogonal dictionary learning for image restoration. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3384–3391, 2013.

Chenglong Bao, Hui Ji, Yuhui Quan, and Zuowei Shen. L0 norm based dictionary learning by proximal methods with global convergence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3858–3865, 2014a.

Chenglong Bao, Yuhui Quan, and Hui Ji. A convergent incoherent dictionary learning algorithm for sparse coding. In *European Conference on Computer Vision*, pages 302–316. Springer, 2014b.

Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007.

Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization or nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1-2):459–494, 2014.

Lodewijk Brand, Xue Yang, Kai Liu, Saad Elbeleidy, Hua Wang, Hao Zhang, and Feiping Nie. Learning robust multilabel sample specific distances for identifying hiv-1 drug resistance. *Journal of Computational Biology*, 27(4):655–672, 2020.

Jian-Feng Cai, Raymond H Chan, and Zuowei Shen. A framelet-based image inpainting algorithm. *Applied and Computational Harmonic Analysis*, 24(2):131–149, 2008.

Jian-Feng Cai, Hui Ji, Chaoqiang Liu, and Zuowei Shen. Framelet-based blind motion deblurring from a single image. *IEEE Transactions on Image Processing*, 21(2):562–572, 2012.

Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001.

Charles K Chui. *An introduction to wavelets*. Elsevier, 2016.

Patrick L Combettes and Teemu Pennanen. Proximal methods for cohypomonotone operators. *SIAM journal on control and optimization*, 43(2):731–742, 2004.

David L Donoho and Michael Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via l1 minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, 2003.

Michael Elad and Michal Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Transactions on Image processing*, 15(12):3736–3745, 2006.

Athinodoros S. Georghiades, Peter N. Belhumeur, and David J. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE transactions on pattern analysis and machine intelligence*, 23(6):643–660, 2001.
Dar Gilboa, Sam Buchanan, and John Wright. Efficient dictionary learning with gradient descent. *arXiv preprint arXiv:1809.10313*, 2018.

Zhuolin Jiang, Zhe Lin, and Larry S Davis. Learning a discriminative dictionary for sparse coding via label consistent k-svd. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1697–1704. IEEE, 2011.

Alexander Kaplan and Rainer Tichatschke. Proximal point methods and nonconvex optimization. *Journal of global Optimization*, 13(4):389–406, 1998.

Michael S Lewicki and Terrence J Sejnowski. Learning overcomplete representations. *Neural computation*, 12(2):337–365, 2000.

Qiuwei Li, Zhihui Zhu, and Gongguo Tang. Geometry of factored nuclear norm regularization. *arXiv preprint arXiv:1704.01265*, 2017.

Zhengming Li, Zheng Zhang, Jie Qin, Zhao Zhang, and Ling Shao. Discriminative fisher embedding dictionary learning algorithm for object recognition. *IEEE transactions on neural networks and learning systems*, 2019.

Guojun Lin, Meng Yang, Jian Yang, Linlin Shen, and Weicheng Xie. Robust, discriminative and comprehensive dictionary learning for face recognition. *Pattern Recognition*, 81:341–356, 2018.

Kai Liu. *Study of non-negative matrix factorizations: foundations, methods, algorithms, and applications*, A. PhD thesis, Colorado School of Mines, 2019.

Kai Liu and Hua Wang. Robust multi-relational clustering via ℓ1-norm symmetric nonnegative matrix factorization-norm symmetric nonnegative matrix factorization. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 397–401, 2015.

Kai Liu and Hua Wang. High-order co-clustering via strictly orthogonal and symmetric l1-norm nonnegative matrix tri-factorization. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 2018.

Kai Liu, Hua Wang, Feiping Nie, and Hao Zhang. Learning multi-instance enriched image representations via non-greedy ratio maximization of the l1-norm distances. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7727–7735, 2018a.

Kai Liu, Hua Wang, Shannon Risacher, Andrew Saykin, and Li Shen. Multiple incomplete views clustering via non-negative matrix factorization with its application in alzheimer’s disease analysis. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1402–1405. IEEE, 2018b.

Kai Liu, Lodewijk Brand, Hua Wang, and Feiping Nie. Learning robust distance metric with side information via ratio minimization of orthogonally constrained l21-norm distances. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019a.

Kai Liu, Qiuwei Li, Hua Wang, and Gongguo Tang. Spherical principal component analysis. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pages 387–395. SIAM, 2019b.
Kai Liu, Hua Wang, Fei Han, and Hao Zhang. Visual place recognition via robust $\ell_2$-norm distance based holism and landmark integration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8034–8041, 2019c.

Julien Mairal, Francis Bach, Jean Ponce, Guillermo Sapiro, and Andrew Zisserman. Non-local sparse models for image restoration. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 2272–2279. IEEE, 2009.

Stephane Mallat. *A wavelet tour of signal processing: the sparse way*. Academic press, 2008.

Aleix M Martinez. *The ar face database*. CVC Technical Report24, 1998.

Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

Ron Rubinstein, Michael Zibulevsky, and Michael Elad. Efficient implementation of the k-svd algorithm using batch orthogonal matching pursuit. Technical report, Computer Science Department, Technion, 2008.

Subhasis Saha. Image compression—from dct to wavelets: a review. *Crossroads*, 6(3):12–21, 2000.

Morgan A Schmitz, Matthieu Heitz, Nicolas Bonneel, Fred Ngole, David Coeurjolly, Marco Cuturi, Gabriel Peyré, and Jean-Luc Starck. Wasserstein dictionary learning: Optimal transport-based unsupervised non-linear dictionary learning. *SIAM Journal on Imaging Sciences*, 11(1):643–678, 2018.

Joel A Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information theory*, 50(10):2231–2242, 2004.

Junfeng Wu, Fang Dai, Gang Hu, and Xuanqin Mou. Low dose ct reconstruction via l 1 norm dictionary learning using alternating minimization algorithm and balancing principle. *Journal of X-ray science and technology*, 26(4):603–622, 2018.

Haoxuan Yang, Kai Liu, Hua Wang, and Feiping Nie. Learning strictly orthogonal p-order nonnegative laplacian embedding via smoothed iterative reweighted method. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, 2019.

Qiang Zhang and Baoxin Li. Discriminative k-svd for dictionary learning in face recognition. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2691–2698. IEEE, 2010.

Zhihui Zhu, Xiao Li, Kai Liu, and Qiuwei Li. Dropping symmetry for fast symmetric nonnegative matrix factorization. *arXiv preprint arXiv:1811.05642*, 2018.