Regularized Low-coherence Overcomplete Dictionary Learning for Sparse Signal Decomposition

Mostafa Sadeghi
Electrical Engineering Department
Sharif University of Technology
Tehran, Iran
Email: m.saadeghii@gmail.com

Massoud Babaie-Zadeh
Electrical Engineering Department
Sharif University of Technology
Tehran, Iran
Email: mbzadeh@yahoo.com

Christian Jutten
GIPSA-Lab, Univ. Grenoble Alpes
and Institut Universitaire de France
France
Email: christian.jutten@gipsa-lab.grenoble-inp.fr

***Abstract*—This paper deals with learning an overcomplete set of atoms that have low mutual coherence. To this aim, we propose a new dictionary learning (DL) problem that enables a control on the amounts of the decomposition error and the mutual coherence of the atoms of the dictionary. Unlike existing methods, our new problem directly incorporates the mutual coherence term into the usual DL problem as a regularizer. We also propose an efficient algorithm to solve the new problem. Our new algorithm uses block coordinate descent, and updates the dictionary atom-by-atom, leading to closed-form solutions. We demonstrate the superiority of our new method over existing approaches in learning low-coherence overcomplete dictionaries for natural image patches.**

I. Introduction

A. Sparse signal decomposition

Sparse signal decomposition [1] provides an efficient signal model that has been effectively exploited in many applications, including image enhancement [1], signal separation [2], signal detection [3], and compressed sensing [4]. Let $\mathbf{y} \in \mathbb{R}^n$ be a target signal and $\mathbf{D} = [\mathbf{d}_1, \cdots, \mathbf{d}_m]$ be a dictionary of m atoms. Then, the decomposition of $\mathbf{y}$ over $\mathbf{d}$ is written as $\mathbf{y} = \sum_{i=1}^{m} x_i \mathbf{d}_i = \mathbf{D}\mathbf{x}$, where $\mathbf{x}$ is the representation vector indicating the contribution of each atom in the decomposition. The dictionary is usually overcomplete, meaning that $m > n$. So, the decomposition problem is ill-posed in general, implying that there exist many solutions for $\mathbf{x}$. Among these numerous solutions, the sparsest one which contains as many as possible zeros is desired in sparse signal decomposition. Fortunately, it has been shown that the sparsest solution is unique under some conditions [1], and many practical algorithms have been proposed to find it [5].

B. Overcomplete dictionary learning

In a sparse decomposition application, an important question is how to choose the dictionary. While there exist some predefined and analytical options for the dictionary, such as discrete cosine transform (DCT), Fourier, and wavelets, it has been shown that adapting the atoms to the structure of the data leads to much better performances [6]–[9]. This process is called *dictionary learning* (DL).

To explain the DL problem more precisely, consider a number of N training signals $\{\mathbf{y}_i\}_{i=1}^{N}$ collected as the columns of a training data matrix $\mathbf{Y} \in \mathbb{R}^{n \times N}$. The DL problem is then formulated as

$$\min_{\mathbf{D},\mathbf{X}} \left\{ \frac{1}{2}\|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 + \lambda_x \|\mathbf{X}\|_0 \right\} \text{ s.t. } \forall i, \|\mathbf{d}_i\|_2 = 1, \quad (1)$$

where, $\|.\|_F$ is the Frobenius norm, $\|.\|_0$ denotes the ℓ_0 pseudo norm which counts the number of nonzeros, and $\lambda_x > 0$ is a regularization parameter. The norm constraint on the atoms is for avoiding scale ambiguity. Algorithms for solving the above DL problem are usually based on alternating minimization [6], [7], [9], consisting of iterative minimization of the cost function over one variable while fixing the other one. The minimization over $\mathbf{X}$ is called the *sparse decomposition stage*, and the minimization over $\mathbf{D}$ is called the *dictionary update stage*.

C. Learning dictionaries with low mutual coherence

To ensure successful performance of sparse decomposition algorithms, the dictionary must satisfy certain conditions. In fact, many previous works have revealed that the uniqueness and the stability of a sparse decomposition problem are directly related to the properties of the dictionary [5], [10], [11]. One measure for the goodness of a dictionary is defined as the maximum (in absolute) amount of the correlations between any two distinct atoms. This is called the mutual coherence (MC) of the dictionary [11], which is mathematically defined as (the dictionary is assumed to have normalized columns)

$$\mu(\mathbf{D}) \triangleq \max_{i \neq j} |\langle \mathbf{d}_i, \mathbf{d}_j \rangle|. \quad (2)$$

For a dictionary of size $n \times m$, the MC value is lower-bounded by $\mu \geq \sqrt{(m-n)/n(m-1)}$, which is known as the Welch bound [12]. The MC has been exploited in many works [13]–[17] to learn dictionaries that provide good decomposition properties in addition to adaptation to the training signals.

This work has been supported by the project CHESS, 2012-ERC-AdG-320684, and by the Center for International Scientific Studies and Collaboration (CISSC).

Previous works on learning dictionaries with low MCs are divided into two categories: regularized and constrained approaches. The first category of the methods, including [13]–[15], are based on the regularization of the DL cost function with an incoherency promoting term as follows[1]

$$\min_{\mathbf{D}\in\mathcal{D}} \ \frac{1}{2}\|\mathbf{Y}-\mathbf{D}\mathbf{X}\|_F^2+\lambda\|\mathbf{D}^T\mathbf{D}-\mathbf{I}\|_F^2, \tag{3}$$

where $\mathcal{D} \triangleq \{\mathbf{D}\in\mathbb{R}^{n\times m} \mid \ \forall i, \|\mathbf{d}_i\|_2=1\}$, and $\lambda>0$. The use of the second term in the above cost function is motivated by the following relation

$$\|\mathbf{D}^T\mathbf{D}-\mathbf{I}\|_F^2=\sum_{i\neq j}|\langle\mathbf{d}_i,\mathbf{d}_j\rangle|^2+\sum_i(\langle\mathbf{d}_i,\mathbf{d}_i\rangle-1)^2, \tag{4}$$

where $\langle .,.\rangle$ denotes the (real) dot-product operation. The first term in the above equation is responsible for minimizing the average correlations between distinct atoms, and the last term encourages the atoms to have unit norms. A number of practical algorithms have been proposed to solve (3), *e.g.*, a limited-memory BFGS (l-BFGS) algorithm [18] proposed in [13], and a proximal algorithm proposed in [15].

The second group of methods including [16], [17] propose the following constrained problem

$$\min_{\mathbf{D}\in\mathcal{D}} \ \frac{1}{2}\|\mathbf{Y}-\mathbf{D}\mathbf{X}\|_F^2 \quad \text{s.t.} \quad \mu(\mathbf{D})\leq\mu_0, \tag{5}$$

where, $\mu_0>0$ is a desired level of MC. To solve this problem, a two-step approach was proposed in [17]. In the first step, the dictionary is updated to minimize the approximation error, ignoring the MC constraint. In the second step, the dictionary returned by the first step is optimized to satisfy the MC constraint. This step is then followed by a dictionary rotation, in which the dictionary is rotated to minimize the approximation error.

In this paper, we focus on regularized approaches, and propose a new regularized low mutual coherence DL problem that unlike existing ones, directly incorporates the MC term defined in (2) into the general DL problem. We also propose an efficient algorithm to solve our new problem, which leads to closed-form solutions. To evaluate the performance of the proposed approach and show its advantage over the previous algorithms, we conducted a sparse image patch decomposition experiment. The results show that our new algorithm makes a much better trade-off between lowering the MC of the dictionary and reducing the decomposition errors of the training signals.

The rest of the paper is organized as follows. Our proposed approach, consisting of our new problem and its proposed solver, is introduced in Section II, and Section III presents the simulation results and discussions.

[1]Note that, the constraint on $\mathbf{D}$ does not affect the sparse decomposition stage, and any algorithm can be used to perform this stage. So, in the remaining of the paper, we focus merely on the dictionary update stage.

II. Proposed approach

A. Problem formulation

In this subsection, we propose our new regularized dictionary update problem to be used within a general DL problem. To this end, first notice the following equivalent definition of MC

$$\mu(\mathbf{D})=\|\mathbf{D}^T\mathbf{D}-\mathbf{I}\|_\infty, \tag{6}$$

where $\|\mathbf{X}\|_\infty \triangleq \max_{i,j}|x_{ij}|$. Our new dictionary update problem is then defined as

$$\min_{\mathbf{D}\in\mathcal{D}} \ \frac{1}{2}\|\mathbf{Y}-\mathbf{D}\mathbf{X}\|_F^2+\lambda\|\mathbf{D}^T\mathbf{D}-\mathbf{I}\|_\infty, \tag{7}$$

Comparing (7) with (3) reveals that the regularization term used in previous work is an approximation of MC, in which the ℓ_∞ norm has been replaced with the Frobenius norm.

B. Proposed algorithm

To solve (7), we use a block coordinate descent approach, and update the atoms one-by-one. Noting the relation $\mathbf{D}\mathbf{X}=\sum_{i=1}^m\mathbf{d}_i\mathbf{z}_i^T$ where $\mathbf{z}_i$ is the ith column of $\mathbf{Z}\triangleq\mathbf{X}^T$, and that $\|\mathbf{d}_i\|_2=1$, the update problem for the ith atom would be

$$\min_{\mathbf{d}} \ \left\{\frac{1}{2}\|\mathbf{E}_i-\mathbf{d}\mathbf{z}_i^T\|_F^2+\lambda\|\mathbf{D}_i^T\mathbf{d}\|_\infty\right\} \ \text{s.t.} \ \|\mathbf{d}_i\|_2=1, \tag{8}$$

where $\mathbf{E}_i\triangleq\mathbf{Y}-\mathbf{D}\mathbf{X}+\mathbf{d}_i\mathbf{z}_i^T$ is the error matrix when the effect of the ith atom is removed, and $\mathbf{D}_i$ consists of the all atoms excluding $\mathbf{d}_i$. The matrices $\mathbf{E}_i$ and $\mathbf{D}_i$ are constructed by using the most recent values of the atoms.

Let us define an auxiliary variable $\mathbf{c}\triangleq\mathbf{D}_i^T\mathbf{d}$ in (8). Doing so, we have the following equivalent problem

$$\min_{\mathbf{d},\mathbf{c}} \ \left\{\frac{1}{2}\|\mathbf{E}_i-\mathbf{d}\mathbf{z}_i^T\|_F^2+\lambda\|\mathbf{c}\|_\infty\right\} \quad \text{s.t.} \ \mathbf{c}=\mathbf{D}_i^T\mathbf{d}, \ \|\mathbf{d}_i\|_2=1, \tag{9}$$

To solve (9), we propose the following alternative problem

$$\min_{\mathbf{d},\mathbf{c}} \ \left\{\frac{1}{2}\|\mathbf{E}_i-\mathbf{d}\mathbf{z}_i^T\|_F^2+\lambda\|\mathbf{c}\|_\infty+\frac{\alpha}{2}\|\mathbf{c}-\mathbf{D}_i^T\mathbf{d}\|_2^2\right\} \quad \text{s.t.} \ \|\mathbf{d}_i\|_2=1, \tag{10}$$

where $\alpha>0$ is a penalty parameter. Note that, when $\alpha\to\infty$, we must have $\mathbf{c}=\mathbf{D}_i^T\mathbf{d}$ in the cost function of (10). So, for a large enough α, problem (10) becomes equivalent to problem (9), and thus (8).

Our strategy for solving (10) is to use alternating minimization. To do so, consider first the update problem for $\mathbf{c}$

$$\min_{\mathbf{c}} \ \left\{\frac{1}{2}\|\mathbf{c}-\mathbf{c}_i\|_2^2+\beta\|\mathbf{c}\|_\infty\right\}, \tag{11}$$

where $\mathbf{c}_i\triangleq\mathbf{D}_i^T\mathbf{d}$ and $\beta\triangleq\lambda/\alpha$. To solve this problem, we use the notion of proximal mapping [19].

Definition 1 ([19]): The proximal mapping of a function $f: \ \text{dom}f\longrightarrow\mathbb{R}$ at a point $\mathbf{x}$ is defined as

$$\text{prox}_f(\mathbf{x})\triangleq\underset{\mathbf{u}\in\text{dom}f}{\text{argmin}} \ \left\{\frac{1}{2}\|\mathbf{u}-\mathbf{x}\|_2^2+f(\mathbf{u})\right\}.$$

So, the solution for problem (11) is recognized as the proximal mapping of $\beta\|\cdot\|_\infty$ at $\mathbf{c}_i$. To derive the associated proximal mapping, we use the Moreau decomposition [19], which states that for any convex function f and its convex conjugate [20] f^* we have

$$\mathbf{x} = \text{prox}_f(\mathbf{x}) + \text{prox}_{f^*}(\mathbf{x}), \tag{12}$$

For βf, it can be verified that

$$\mathbf{x} = \text{prox}_{\beta f}(\mathbf{x}) + \beta\text{prox}_{f^*/\beta}(\mathbf{x}/\beta)\cdot \tag{13}$$

Let $f(\mathbf{x}) \triangleq \|\mathbf{x}\|_\infty$. The convex conjugate of f is $f^* = \mathcal{I}_{\mathcal{B}_1^1}$, that is, the indicator function of the unit ℓ_1 norm ball [20]. Then, using (13) and the fact that the proximal mapping of the indicator function of a set is the projection onto that set [19], it is verified that

$$\text{prox}_{\beta f}(\mathbf{x}) = \mathbf{x} - P_{\mathcal{B}_1^\beta}(\mathbf{x}), \tag{14}$$

where $\mathcal{B}_1^\beta$ is the ℓ_1 norm ball of radius β, and $P_{\mathcal{B}_1^\beta}$ is the projection onto $\mathcal{B}_1^\beta$. To perform this projection, there exist efficient algorithms, such as the one proposed in [21], which is in closed-form. So, the final solution for (11) is obtained as

$$\mathbf{c}^* = \mathbf{c}_i - P_{\mathcal{B}_1^\beta}(\mathbf{c}_i)\cdot \tag{15}$$

Now, consider the update problem for $\mathbf{d}_i$, which is

$$\min_{\mathbf{d}} \left\{\frac{1}{2}\|\mathbf{E}_i - \mathbf{d}\mathbf{z}_i^T\|_F^2 + \frac{\alpha}{2}\|\mathbf{c} - \mathbf{D}_i^T\mathbf{d}\|_2^2\right\} \text{ s.t. } \|\mathbf{d}_i\|_2 = 1\cdot \tag{16}$$

Setting the gradient of the above cost function with respect to $\mathbf{d}$ equal to zero leads to

$$\mathbf{Q}_i\mathbf{d}^* = \mathbf{a}_i, \tag{17}$$

where $\mathbf{Q}_i \triangleq \alpha\mathbf{D}_i\mathbf{D}_i^T + \|\mathbf{z}_i\|_2^2\mathbf{I}$, with $\mathbf{I}$ being the $n\times n$ identity matrix, and $\mathbf{a}_i \triangleq \mathbf{E}_i\mathbf{z}_i + \alpha\mathbf{D}_i\mathbf{c}^*$. Since $\mathbf{Q}_i$ is positive definite, the above system of equations can be efficiently solved using the conjugate gradient algorithm [22]. The solution is then normalized to satisfy the constraint in (16). The final update procedure for the ith atom consists of a few iterations between (15) and (17).

1) Choosing the penalty parameter: As said previously, the penalty parameter α should be chosen large enough to ensure that the solution of (10) well approximates the solution of the original problem (8). This is equivalent to choosing a very small value for the new parameter β. However, considering problem (11) and its solution given in (15), a small value for β decreases the convergence speed of the $\{\mathbf{c}_i\}$ sequence.

To overcome this problem, we propose to solve (10) for a decreasing sequence of β's, starting with a relatively large value. Moreover, the solution of each subproblem, corresponding to a particular value of β, is used as a warm-start for the next one. In this way, the sequence of the solutions gradually approximates the desired solution of problem (8). Here, we propose an exponential decrease for β as follows

$$\beta_j = \gamma^j\cdot\beta_0, \tag{18}$$

where $0<\gamma<1$, and β_j and β_0 denote the jth and the initial value for β, respectively.

Algorithm 1 Proposed Low mutual coherence DL algorithm

Require: $\mathbf{Y}\in\mathbb{R}^{n\times N}$, $\mathbf{D}_0\in\mathbb{R}^{n\times m}$, λ, τ, β_0, γ, J, L
Initialization: $\mathbf{D}=\mathbf{D}_0$
for $k=1,2,\cdots$ **do**
  1. Sparse decomposition: $\mathbf{X}=\text{SD}(\mathbf{Y},\mathbf{D},\tau)$
  2. Dictionary update:
  for $i=1,2,\cdots,m$ **do**
    $\mathbf{E}_i=\mathbf{Y}-\mathbf{D}\mathbf{X}+\mathbf{d}_i\mathbf{z}_i^T$
    for $j=1,2,\cdots,J$ **do**
      $\beta_j=\gamma^j\cdot\beta_0$
      for $l=1,2,\cdots,L$ **do**
        Update $\mathbf{c}_i^*$ using (15) with $\beta=\beta_j$
        Update $\mathbf{d}_i^*$ using (17). Normalize the result.
      end for
    end for
  end for
end for
Output: $\mathbf{D}$

2) Final algorithm: The final algorithm is summarized in Algorithm 1. In this algorithm, $\text{SD}(\mathbf{Y},\mathbf{D},\tau)$ stands for the sparse decompositions of the columns of $\mathbf{Y}$ over $\mathbf{D}$ with parameter τ, which can be, for example, a regularization parameter or maximum allowed number of atoms in the decompositions.

III. Simulations

In this section, we compare the performance of our proposed regularized low mutual coherence DL algorithm with previous ones through a sparse image patch decomposition experiment. From the existing low coherence DL algorithms, we chose the bounded self coherence (BSC) DL algorithm proposed in [13], which showed a better overall performance among the others in this experiment[2].

In the same way as [23], we used a collection of seventeen well-known standard images, including Barbara, Cameraman, Lena, and Peppers. A number of 5000 patches of size 8×8 were extracted from these images, and then converted to equivalent column vectors of length 64. For both DL algorithms, the initial dictionary was chosen as an overcomplete DCT dictionary of size 64×256, whose MC was 0.99. To perform the sparse decomposition stage, the orthogonal matching pursuit (OMP) algorithm [24] was employed, which sequentially selects appropriate atoms one-by-one in a greedy fashion. A number of 5 atoms was allowed to participate in sparse decomposition of every training signal. Also, a number of 100 iterations (alternates) between the sparse decomposition and the dictionary update stages were used for both algorithms. Root mean square error (RMSE), defined as $\|\mathbf{Y}-\mathbf{D}\mathbf{X}\|_F/(n\cdot N)$, was used as a measure of performance.

[2]Here, one may argue that for a fair comparison, the same atom-by-atom updating procedure used for solving the proposed problem (7) should also be used for problem (3), and then compare the two algorithms. We implemented and tested this new solver of problem (3), but the algorithm diverged in most cases.

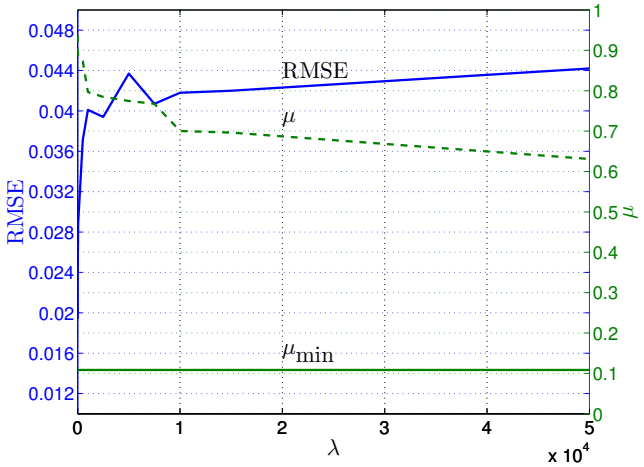


Fig. 1: Final RMSE and the achieved MC versus λ for BSC-DL. The Welch bound is also depicted as μ_{min}.

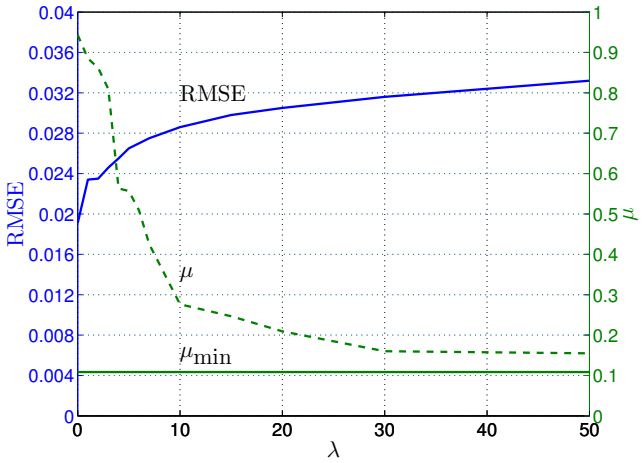


Fig. 2: Final RMSE and the achieved MC versus λ for Algorithm 1. The Welch bound is also depicted as μ_{min}.

The parameters of the l-BFGS algorithm used in BSC-DL were chosen as suggested by the authors of [13]. For our proposed algorithm outlined in Algorithm 1, we set $\gamma = 0.85$, $\beta_0 = 10$, $J = 30$, and $L = 1$, which showed a promising performance.

Final RMSEs and the achieved MCs versus λ are depicted in Figs. 1 and 2. Inspecting the results reveals that, our proposed algorithm makes a much better compromise between minimizing the MC and the RMSE. It is also able to achieve an MC value of 0.15, which is very close to the Welch bound $\mu_{\min} = 0.10$, compared to the minimum MC of 0.63 achieved by BSC-DL.

The evolutions of the MC and the RMSE along DL alternates, for λ's corresponding to the minimum achieved MCs by the two algorithms are shown in Figs. 3 and 4, respectively. As demonstrated, our proposed algorithm shows a smooth behavior, except for the first DL alternate, while BSC-DL has a non-monotonic behavior. Moreover, the final RMSEs for both

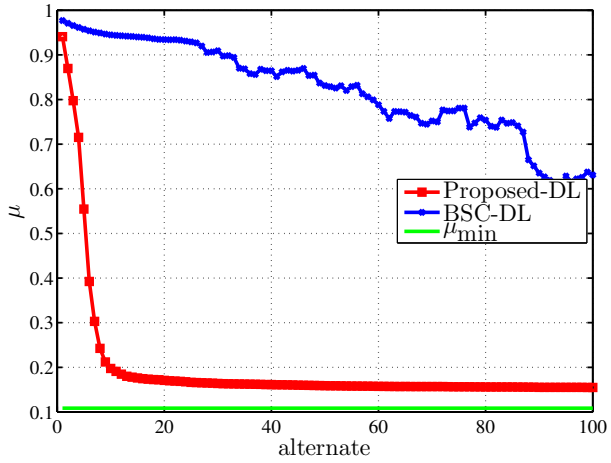


Fig. 3: MC as a function of DL alternates for both algorithms, and for λ's corresponding to the minimum achieved MCs. The Welch bound is also depicted as μ_{min}.

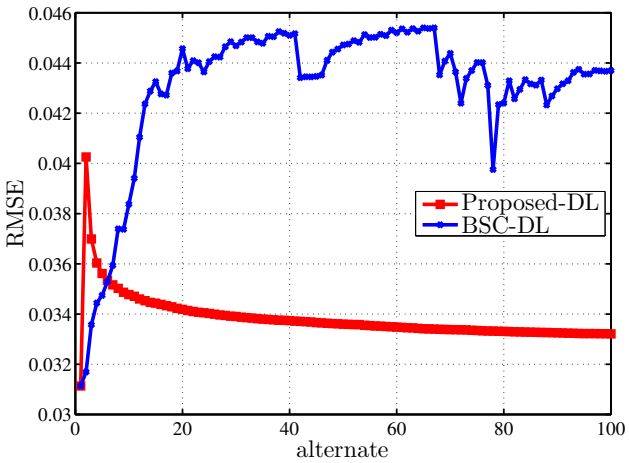
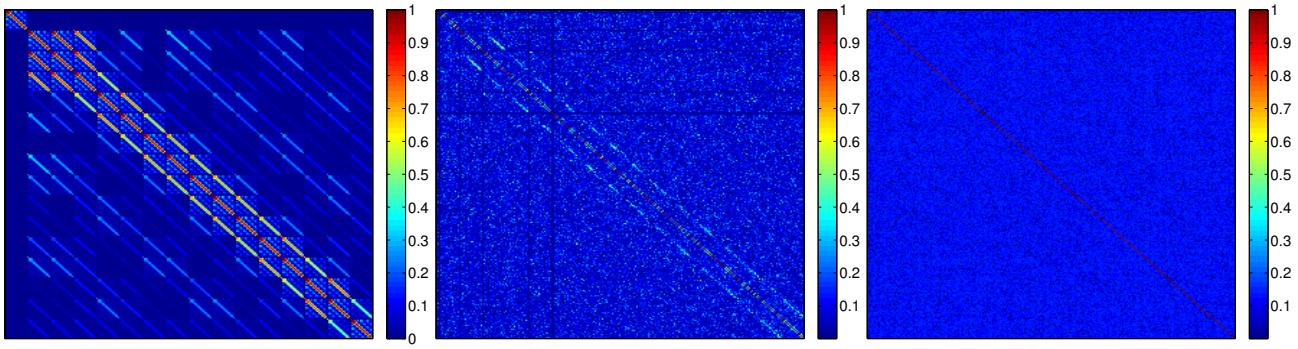


Fig. 4: RMSE as a function of DL alternates for both algorithms, and for λ's corresponding to the minimum achieved MCs.

algorithms are higher that the one corresponding to the DCT dictionary (first alternate). This is because, larger weights are enforced on the incoherency promoting terms in this case, which is more noticeable for our proposed algorithm as it learns a dictionary with a much lower MC than the initial DCT dictionary. From these figures, it is also observed that for our proposed algorithm, while the MC value decreases along the DL alternates, the RMSE value also monotonically decreases. This is in contrast to the BSC-DL's behavior, where proceeding along the DL alternates, the MC of the dictionary decreases while its decomposition error increases, at least in this extreme scenario. To see the distribution of the inner products between every two atoms of the dictionaries corresponding to these minimum MCs, we use the Gram matrix, which is defined as $\mathbf{G} \triangleq \mathbf{D}^T\mathbf{D}$. The off-diagonal entries of $\mathbf{G}$ correspond to the inner products between every two distinct atoms. Magnitudes of the Gram matrices for the initial DCT dictionary and the



(a) Initial dictionary (DCT) (b) BSC-DL (c) Algorithm 1

Fig. 5: Magnitudes of the Gram matrices, defined as $\mathbf{G} \triangleq \mathbf{D}^T\mathbf{D}$, for the initial DCT dictionary, and the final dictionaries learned by BSC-DL and Algorithm 1 that correspond to the smallest possible MC they can reach.

dictionaries learned by BSC-DL and our proposed algorithm are shown in Fig. 5. It can be seen that, compared to the initial dictionary and the final dictionary learned by BSC-DL, our proposed algorithm effectively reduces the correlations between distinct atoms.

To have a rough measure of the computational load of the algorithms, their runtimes are reported. Our simulations were carried out on a 64 bit Windows 7 operating system with 12 GB RAM and an Intel core i7 CPU. The averaged runtimes, in seconds, for BSC-DL and Algorithm 1 were 123 and 178, respectively, indicating that our proposed algorithm has a higher computational complexity.

IV. Conclusion

We addressed learning overcomplete dictionaries with low mutual coherences for sparse signal decomposition. This problem has been studied in some previous works and a number of algorithms have been proposed to this aim. In this paper, we proposed a new dictionary learning problem that unlike previous works, directly solves the mutual coherence regularized DL problem. We empirically demonstrated the superior performance of our new algorithm over the existing ones in a sparse image patch decomposition problem. It was noticed that our proposed algorithm is able to efficiently learn overcomplete dictionaries with very low mutual coherences while being adapted to the training signals.

References

[1] M. Elad, *Sparse and Redundant Representations*, Springer, 2010.

[2] P. Comon and C. Jutten, Eds., *Handbook of Blind Source Separation*, Elsevier, 2010.

[3] M. Joneidi, P. Ahmadi, M. Sadeghi, and N. Rahnavard, "Union of low-rank subspaces detector," *IET Signal Processing*, pp. 55–62, 2016.

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

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

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

[7] R. Rubinstein, A. M. Bruckstein, and M. Elad, "Dictionaries for sparse representation modeling," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1045–1057, 2010.

[8] M. Sadeghi, M. Babaie-Zadeh, and C. Jutten, "Dictionary learning for sparse representation: A novel approach," *IEEE Signal Proc. Letters*, vol. 20, no. 12, pp. 1195–1198, 2013.

[9] M. Sadeghi, M. Babaie-Zadeh, and C. Jutten, "Learning over-complete dictionaries based on atom-by-atom updating," *IEEE Trans. on Si*, vol. 62, no. 4, pp. 883–891, 2014.

[10] E. J. Candès and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies," *IEEE Trans. Info. Theory*, vol. 52, no. 12, pp. 5406–5425, 2006.

[11] D. L. Donoho and X. Huo, "Uncertainty principles and ideal atomic decomposition," *IEEE Trans. Information Theory*, vol. 47, no. 7, pp. 2845–2862, 2001.

[12] T. Strohmer and Heath R. W., "Grassmannian frames with applications to coding and communication," *Applied and Computational Harmonic Analysis*, vol. 14, no. 3, pp. 257–275, 2003.

[13] C. D. Sigg, T. Dikk, and J. M. Buhmann, "Learning dictionaries with bounded self-coherence," *IEEE Signal Proc. Letters*, vol. 19, no. 12, pp. 861–864, 2012.

[14] M. Sadeghi, M. Babaie-Zadeh, and C. Jutten, "A new algorithm for learning overcomplete dictionaries," in *Proceedings of the 21st European Signal Processing Conference (EUSIPCO)*, 2013.

[15] C. Bao, H. Ji, Y. Quan, and Z. Shen, "Dictionary learning for sparse coding: Algorithms and convergence analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015, to be published.

[16] B. Mailhé, D. Barchiesi, and M. D. Plumbley, "INK-SVD: Learning incoherent dictionaries for sparse representations," in *IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, 2012, pp. 3573–3576.

[17] D. Barchiesi and M. D. Plumbley, "Learning incoherent dictionaries for sparse approximation using iterative projections and rotations," *IEEE Trans. on Signal Proc.*, vol. 61, no. 8, pp. 2055–2065, 2013.

[18] D. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Math. Progr.*, vol. 45, no. 1, pp. 503–528, 1989.

[19] N. Parikh and S. Boyd, "Proximal algorithms," *Foundations and Trends in Optimization*, vol. 1, no. 3, pp. 123–231, 2014.

[20] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[21] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, "Efficient projections onto the l1-ball for learning in high dimensions," in *International Conference on Machine Learning (ICML)*, 2008.

[22] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, Belmont, MA, 1999.

[23] L. N. Smith and M. Elad, "Improving dictionary learning: Multiple dictionary updates and coefficient reuse," *IEEE Signal Proc. Letters*, vol. 20, no. 1, pp. 79–82, 2013.

[24] J. A. Tropp and A. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Info. Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.