Sensing Matrix Optimization for Block-Sparse Decoding

Kevin Rosenblum, Lihi Zelnik-Manor, Yonina C. Eldar

Abstract—**Recent work has demonstrated that using a carefully designed sensing matrix rather than a random one, can improve the performance of compressed sensing. In particular, a well-designed sensing matrix can reduce the coherence between the atoms of the equivalent dictionary, and as a consequence, reduce the reconstruction error. In some applications, the signals of interest can be well approximated by a union of a small number of subspaces (e.g., face recognition and motion segmentation). This implies the existence of a dictionary which leads to *block-sparse* representations. In this work, we propose a framework for sensing matrix design that improves the ability of block-sparse approximation techniques to reconstruct and classify signals. This method is based on minimizing a weighted sum of the inter-block coherence and the sub-block coherence of the equivalent dictionary. Our experiments show that the proposed algorithm significantly improves signal recovery and classification ability of the *Block-OMP* algorithm compared to sensing matrix optimization methods that do not employ block structure.**

I. INTRODUCTION

The framework of compressed sensing aims at recovering an unknown vector $x \in R^N$ from an under-determined system of linear equations $y = Ax$, where $A \in R^{M \times N}$ is a sensing matrix, and $y \in R^M$ is an observation vector with $M < N$. Since the system is under-determined, x can not be recovered without additional information. In [1], [2] it was shown that when x is known to have a sufficiently sparse representation, and when A is randomly generated, x can be recovered uniquely with high probability from the measurements y. More specifically, the assumption is that x can be represented as $x = D\theta$ for some orthogonal dictionary $D \in R^{N \times N}$, where $\theta \in R^N$ is sufficiently sparse. The vector x can then be recovered regardless of D and irrespective of the locations of the nonzero entries of θ. This can be achieved by approximating the sparsest representation θ using methods such as *Basis Pursuit* (BP) [3], [1] and *Orthogonal Matching Pursuit* (OMP) [4], [5]. In practice, overcomplete dictionaries $D \in R^{N \times K}$ with $K \geq N$ lead to improved sparse representations and are better suited for most applications. Therefore, we treat the more general case of overcomplete dictionaries in this paper.

A simple way to characterize the recovery ability of sparse approximation algorithms was presented in [4], using the coherence between the columns of the equivalent dictionary $E = AD$. When the coherence is sufficiently low, OMP and BP are guaranteed to recover the sparse vector θ. Accordingly, recent work [6], [7], [8] has demonstrated that designing a sensing matrix such that the coherence of E is low improves

the ability to recover θ. The proposed methods yield good results for general sparse vectors.

In some applications, however, the representations have a unique sparsity structure that can be exploited. Our interest is in the case of signals that are drawn from a union of a small number of subspaces [9], [10], [11], [12]. This occurs naturally, for example, in face recognition [13], [14], motion segmentation [15], multi-band signals [16], [17], [18], measurements of gene expression levels [19], and more. For such signals, sorting the dictionary atoms according to the underlying subspaces leads to sparse representations which exhibit a block-sparse structure, i.e., the nonzero coefficients in θ occur in clusters of varying sizes. Several methods, such as *Block-BP* (BBP) [12], [20], [21] and *Block-OMP* (BOMP) [22], [23] have been proposed to take advantage of this block structure in recovering the block-sparse representations θ. Bounds on the recovery performance were presented in [12] based on the block restricted isometry property (RIP), and in [22] using appropriate coherence measures. In particular, it was shown in [22] that under conditions on the *inter-block coherence* (i.e., the maximal coherence between two blocks) and the *sub-block coherence* (i.e., the maximal coherence between two atoms in the same block) of the equivalent dictionary E, Block-OMP is guaranteed to recover the block-sparse vector θ.

In this paper we propose a method for designing a sensing matrix, assuming that a block-sparsifying dictionary is provided. A method for learning a block-sparsifying dictionary is developed in [24]. Our approach improves the recovery ability of block-sparse approximation algorithms by targeting the Gram matrix of the equivalent dictionary, an approach similar in spirit to that of [7], [8]. While [7] and [8] targeted minimization of the coherence between atoms, our method, which will be referred to as *Weighted Coherence Minimization* (WCM), aims at reducing a weighted sum of the inter-block coherence and the sub-block coherence.

It turns out that the weighted coherence objective is hard to minimize directly. To derive an efficient algorithm, we use the bound-optimization method, and replace our objective with an easier to minimize surrogate function that is updated in each optimization step [25]. We develop a closed form solution for minimizing the surrogate function in each step, and prove that its iterative minimization is guaranteed to converge to a local solution of the original problem.

Our experiments reveal that minimizing the sub-block coherence is more important than minimizing the inter-block coherence. By giving more weight to minimizing the sub-block coherence, the proposed algorithm yields sensing matrices that lead to equivalent dictionaries with nearly orthonormal blocks. Simulations show that such sensing matrices significantly

The authors are with the Technion - Israel Institute of Technology, Haifa, Israel. Email: kevin@tx.technion.ac.il, lihi@ee.technion.ac.il, yonina@ee.technion.ac.il.

improve signal reconstruction and signal classification results compared to previous approaches that do not employ block structure.

We begin by reviewing previous work on sensing matrix design in Section II. In Section III we introduce our definitions of total inter-block coherence and total sub-block coherence. We then present the objective for sensing matrix design, and show that it can be considered as a direct extension of the one used in [8] to the case of blocks. We present the WCM algorithm for minimizing the proposed objective in Section IV and prove its convergence in Appendix A. We evaluate the performance of the proposed algorithm and compare it to previous work in Section V.

Throughout the paper, we denote vectors by lowercase letters, e.g., x, and matrices by uppercase letters, e.g., A. A' is the transpose of A. The jth column of the matrix A is A_j, and the ith row is A^i. The entry of A in the row with index i and the column with index j is A_j^i. We define the Frobenius norm by $\|A\|_F \equiv \sqrt{\sum_j \|A_j\|_2^2}$, and the l_p-norm of a vector x by $\|x\|_p$. The l_0-norm $\|x\|_0$ counts the number of non-zero entries in x. We denote the identity matrix by I or I_s when the dimension is not clear from the context. The largest eigenvalue of the positive-semidefinite matrix $B'B$ is written as $\lambda_{\max}(B)$.

II. PRIOR WORK ON SENSING MATRIX DESIGN

The goal of sensing matrix design is to construct a sensing matrix $A \in R^{M \times N}$ with $M < N$ that improves the recovery ability for a given sparsifying dictionary $D \in R^{N \times K}$ with $K \geq N$. In other words, A is designed to improve the ability of sparse approximation algorithms such as BP and OMP to recover the sparsest representation θ from

$$y = AD\theta = E\theta, \tag{1}$$

where E is the equivalent dictionary. In this section we briefly review the sensing matrix design method introduced by Duarte-Carvajalino and Sapiro [8]. Their algorithm was shown to provide significant improvement in reconstruction success.

The motivation to design sensing matrices stems from the theoretical work of [4], where it was shown that BP and OMP succeed in recovering θ when the following condition holds:

$$\|\theta\|_0 \leq \frac{1}{2}\left(1 + \frac{1}{\mu}\right). \tag{2}$$

Here μ is the coherence defined by:

$$\mu \equiv \max_{i \neq j} \frac{|E_i'E_j|}{\|E_i\|_2\|E_j\|_2}. \tag{3}$$

The smaller μ, the higher the bound on the sparsity of θ. Since E is overcomplete, and as a consequence not orthogonal, μ will always be strictly positive. Condition (2) is a worst-case bound and does not reflect the average recovery ability of sparse approximation methods. However, it does suggest that recovery may be improved when E is as orthogonal as possible.

Motivated by these observations, Duarte-Carvajalino and Sapiro [8] proposed designing a sensing matrix A by mini-

mizing $\|E'E - I\|_F^2$. This problem can be written as:

$$\min_A \|E'E - I\|_F^2 = \min_A \|D'A'AD - I\|_F^2. \tag{4}$$

It is important to note that rather than minimizing μ, (4) minimizes the sum of the squared inner products of all pairs of atoms in E, referred to as the total coherence μ^t:

$$\mu^t = \sum_{j, i \neq j} (E_i'E_j)^2. \tag{5}$$

At the same time, solving (4) keeps the norms of the atoms close to 1.

While an approximate solution to (4) has already been presented in [8], we provide an exact solution that will be of use in the next sections. To solve (4), we rewrite its objective using the well-known relation between the Frobenius norm and the trace, $\|C\|_F^2 = \text{tr}(CC')$:

$$
\begin{aligned}
\|E'E - I_K\|_F^2 &= \text{tr}(E'EE'E - 2E'E + I_K) \\
&= \text{tr}(EE'EE' - 2EE' + I_M) + (K - M) \\
&= \|EE' - I_M\|_F^2 + (K - M) \\
&= \|ADD'A' - I_M\|_F^2 + (K - M). \tag{6}
\end{aligned}
$$

Since the first term in (6) is always positive, the objective of (4) is lower bounded by $\|E'E - I\|_F^2 \geq K - M$.

From (6) it follows that minimizing (4) is equivalent to the minimization of $\|ADD'A' - I_M\|_F^2$. A solution to this problem can be achieved in closed form as follows. Let $U\Lambda U'$ be the eigenvalue decomposition of DD', and let $\Gamma_{M \times N} = AU\Lambda^{1/2}$. Then, (4) is equivalent to:

$$\min_A \|\Gamma\Gamma' - I\|_F^2. \tag{7}$$

This problem is solved by choosing Γ to be any matrix with orthonormal rows, such as $\Gamma = [I_M \ 0]$, leading to $\Gamma\Gamma' = I$. The optimal sensing matrix is then given by $A = \Gamma\Lambda^{-1/2}U'$. Here, and throughout the paper, we assume that D has full row rank, guaranteeing that Λ is invertible. Note that the global minimum of the objective in (4) equals $K - M$. The benefits of using such a sensing matrix were shown empirically in [8].

The same solution is obtained by setting the derivative of (7) equal to zero:

$$\frac{\partial\|\Gamma\Gamma' - I\|_F^2}{\partial\Gamma} = 4(\Gamma\Gamma'\Gamma - \Gamma) = 0 \tag{8}$$

It can be deduced from (8) that for stationary points, the singular values of Γ must be equal to either one or zero. However, only when all the M singular values of Γ equal one, i.e., Γ has full row rank, we have a local minimum (the other stationary points being a local maximum and saddle points). It is important to keep in mind that even though the objective is not convex, every local minimum is a global minimum as well.

III. SENSING MATRIX DESIGN FOR BLOCK-SPARSE DECODING

The design of a sensing matrix according to [8] does not take advantage of block structure in the sparse representations of the data. In this section we formulate the problem of sensing

matrix design for block-sparse decoding. We first introduce the basic concepts of block-sparsity, and then present an objective which can be seen as an extension of (4) to the case of block-sparse decoding.

A. Block-sparse decoding

The framework of block-sparse decoding aims at recovering an unknown vector $x \in R^N$ from an under-determined system of linear equations $y = Ax$, where $A \in R^{M \times N}$ is a sensing matrix, and $y \in R^M$ is an observation vector with $M < N$. The difference with sparse recovery lies in the assumption that x has a sufficiently block-sparse representation $\theta \in R^N$ with respect to some orthogonal block-sparsifying dictionary $D \in R^{N \times N}$. The vector x can then be recovered by approximating the block-sparsest representation corresponding to the measurements y using methods such as Block-BP (BBP) [12], [20], [21] and Block-OMP (BOMP) [22], [23].

A block-sparsifying dictionary D is a dictionary whose atoms are sorted in blocks which enable block-sparse representations for a set of signals. We can represent D as a concatenation of B column-blocks $D[j]$ of size $N \times s_j$, where s_j is the number of atoms belonging to the jth block:

$$D = [D[1] \ D[2] \ \ldots \ D[B]].$$

Similarly, we view the representation θ as a concatenation of B blocks $\theta[j]$ of length s_j:

$$\theta = [\theta[1] \ \theta[2] \ \ldots \ \theta[B]]'.$$

We say that a representation θ is k-block-sparse if its nonzero values are concentrated in k blocks only. This is denoted by $\|\theta\|_{2,0} \leq k$, where

$$\|\theta\|_{2,0} = \sum_{j=1}^{B} I(\|\theta[j]\|_2 > 0).$$

The indicator function $I(\cdot)$ counts the number of blocks in θ with nonzero Euclidean norm.

B. Problem definition

For a given block-sparsifying dictionary $D \in R^{N \times K}$ with $K \geq N$, we wish to design a sensing matrix $A \in R^{M \times N}$ that improves the recovery ability of block-sparse approximation algorithms. Note that we allow D to be overcomplete.

A performance bound on the recovery success of block-sparse signals has been developed in [22] for the case of a dictionary D with blocks of a fixed size s (i.e., $s_i = s_j = s$) and an equivalent dictionary $E = AD$ with normalized columns. The bound is a function of the Gram matrix $G \in R^{K \times K}$ of the equivalent dictionary, defined as $E'E$. The (i,j)th block of G, $E[i]'E[j]$, is denoted by $G[i,j] \in R^{s_i \times s_j}$. The (i,j)th block of any other $K \times K$ matrix will be denoted similarly. It was shown in [22] that BBP and BOMP succeed in recovering the block sparsest representation θ corresponding to the measurements $y = E\theta$ when the following condition holds:

$$\|\theta\|_{2,0} < \frac{1}{2s}\left(\mu_B^{-1} + s - (s-1)\frac{\nu}{\mu_B}\right). \tag{9}$$

Here

$$\mu_B \equiv \max_{j, i \neq j} \frac{1}{s}\sqrt{\lambda_{\max}(G[i,j]'G[i,j])}$$

is the *inter-block coherence* and

$$\nu \equiv \max_j \max_{n, m \neq n} |(G[j,j])_n^m|$$

is the *sub-block coherence*. The inter-block coherence μ_B is a generalization of the coherence μ, and describes the global properties of the equivalent dictionary. More specifically, μ_B measures the cosine of the minimal angle between two blocks in E. The sub-block coherence ν describes the local properties of the dictionary, by measuring the cosine of the minimal angle between two atoms in the same block in E. Note, that when $s = 1$, (9) reduces to the bound in the sparse case (2). The term μ_B^{-1} in (9) suggests that μ_B needs to be reduced in order to loosen the bound. On the other hand, the term $-(s-1)\frac{\nu}{\mu_B}$ implies that the ratio $\frac{\nu}{\mu_B}$ should be small. This leads to a trade-off between minimizing μ_B and minimizing ν to loosen the bound, which is reflected in the sensing matrix design objective presented later in this section.

Condition (9) is a worst case bound and does not represent the average recovery ability of block-sparse approximation methods. It does suggest, however, that in order to improve the average recovery, all pairs of blocks in E should be as orthogonal as possible and also all pairs of atoms within each block should be as orthogonal as possible. Inspired by [8], rather than minimizing the inter-block coherence μ_B and the sub-block coherence ν, we aim at minimizing the *total inter-block coherence* μ_B^t and the *total sub-block coherence* ν^t of the equivalent dictionary E. We define the total inter-block coherence as

$$\mu_B^t = \sum_{j=1}^{B} \sum_{i \neq j} \|G[i,j]\|_F^2, \tag{10}$$

and the total sub-block coherence by

$$\nu^t = \sum_{j=1}^{B} \|G[j,j]\|_F^2 - \sum_{m=1}^{K} (G_m^m)^2, \tag{11}$$

where G_m^m are the diagonal entries of G. The total inter-block coherence μ_B^t equals the sum of the squared entries in G belonging to different blocks (the green entries in Fig. 1). Since this is the sum of Frobenius norms, μ_B^t also equals the sum of the squared singular values of the cross-correlation blocks in G. When E is normalized, μ_B^t is equivalent to the sum of the squared cosines of all the principal angles between all pairs of different blocks. The total sub-block coherence ν^t measures the sum of the squared off-diagonal entries belonging to the same block (the red entries in Fig. 1). When E is normalized, ν^t equals the sum of the squared cosines of all the angles between atoms within the same block. Note that when the size of the blocks equals one, we get $\nu^t = 0$.

Alternatively, one could define the total inter-block coherence as the sum of the squared spectral norms (i.e., the largest singular values) of the cross-correlation blocks in G, and the total sub-block coherence as the sum of the squared maximal off-diagonal entries of the auto-correlation blocks in G. These definitions are closer to the ones used in condition
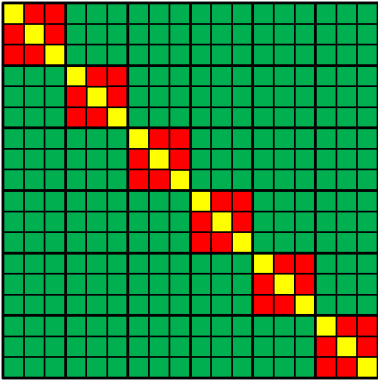
Fig. 1. A graphical depiction of the Gram matrix G of an equivalent dictionary E with 6 blocks of size 3. The entries belonging to different blocks are in green, the off-diagonal entries belonging to the same block are in red, and the diagonal entries are in yellow.

(9). The WCM algorithm presented in the next section can be slightly modified in order to minimize those measures as well. However, besides the increased complexity of the algorithm, the results appear to be inferior compared to minimizing the definitions (10) and (11) of μ_B^t and ν^t. This can be explained by the fact that maximizing only the smallest principal angle between pairs of different blocks in E and maximizing the smallest angle between atoms within the same block, creates a bulk of relatively high singular values and coherence values. While this may improve the worst-case bound in (9), it does not necessarily improve the average recovery ability of block-sparse approximation methods.

When minimizing the total inter-block coherence and the total sub-block coherence, we need to verify that the columns of E are normalized, to avoid the tendency of columns with small norm values to be underused. Rather than enforcing normalization strongly, we penalize for columns with norms that deviate from 1 by defining the *normalization penalty* η:

$$\eta = \sum_{m=1}^{K} (G_m^m - 1)^2. \tag{12}$$

This penalty η measures the sum of the squared distances between the diagonal entries in G (the yellow entries in Fig. 1) and 1.

While [8] did not deal with the block-sparse case, it is straightforward to see that solving (4) is equivalent to minimizing the sum of the normalization penalty, the total inter-block

coherence and the total sub-block coherence:

$$\|E'E - I\|_F^2 = \sum_{j=1}^{B} \sum_{i \neq j} \|E[i]'E[j]\|_F^2 + \sum_{j=1}^{B} \|E[j]'E[j] - I\|_F^2$$
$$= \sum_{j=1}^{B} \sum_{i \neq j} \|G[i,j]\|_F^2 + \sum_{j=1}^{B} \|G[j,j] - I\|_F^2$$
$$= \sum_{j=1}^{B} \sum_{i \neq j} \|G[i,j]\|_F^2 + \sum_{j=1}^{B} \|G[j,j]\|_F^2$$
$$\quad - \sum_{m=1}^{K} (G_m^m)^2 + \sum_{m=1}^{K} (G_m^m - 1)^2$$
$$= \eta + \mu_B^t + \nu^t.$$

We have shown in the previous section that the objective in (4) is bounded below by $K - M$. Therefore,

$$\eta + \mu_B^t + \nu^t \geq K - M. \tag{13}$$

This bound implies a trade-off, and as a consequence, one cannot minimize η, μ_B^t and ν^t freely. Instead, we propose designing a sensing matrix that minimizes the normalization penalty and a weighted sum of the total inter-block coherence and the total sub-block coherence:

$$A = \arg\min_A \frac{1}{2}\eta + (1 - \alpha)\mu_B^t + \alpha\nu^t, \tag{14}$$

where $0 < \alpha < 1$ is a parameter controlling the weight given to the total inter-block coherence and the total sub-block coherence. Note that alternative objectives can be formulated. For example, one could add an additional weighting parameter to the normalization penalty term. While this would allow us to better control the normalization of the atoms in E, we prefer to deal with a single parameter only.

When $\alpha < \frac{1}{2}$, more weight is given to minimizing μ_B^t, and therefore solving (14) leads to lower total inter-block coherence, which is made possible by aligning the atoms within each block (Fig. 2(a)). On the other hand, choosing $\alpha > \frac{1}{2}$ gives more weight to minimizing ν^t. In this case, solving (14) leads to more orthonormal blocks in E at the expense of higher μ_B^t (Fig. 2(c)). Finally, setting $\alpha = \frac{1}{2}$ in (14) gives equal weights to μ_B^t, ν^t and η, and reduces it to (4) (Fig. 2(b)). Therefore, the objective becomes independent of the block structure, which makes $\alpha = \frac{1}{2}$ the correct choice when the signals do not have an underlying block structure. Choosing to ignore the block structure leads to the same conclusion. When an underlying block structure exists, we need to select a value for α. We do that via empirical evaluation in Section V.

In the previous section we have shown that every local minimum of (4), and therefore also of (14) with $\alpha = \frac{1}{2}$, is also a global minimum. Empirical observations reveal that this is not the case when $\alpha \neq \frac{1}{2}$. This is demonstrated in the histograms presented in Fig. 3(a) for $\alpha = 0.01$ with a square dictionary and in Fig. 4(b) for $\alpha = 0.99$ with a highly overcomplete dictionary. Since it is hard to develop a closed form solution for (14), we present an iterative algorithm that converges to a local solution of (14) in the following section.
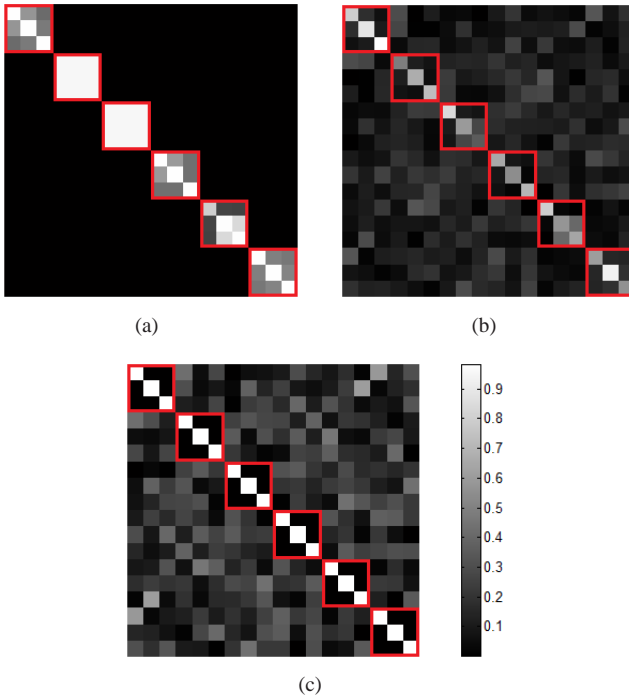
(a)　　　　　　(b)

(c)

Fig. 2. Examples of the absolute value of the Gram matrix of an equivalent dictionary for $\alpha = 0.01$ (a), $\alpha = 0.5$ (b) and $\alpha = 0.99$ (c), where the sensing matrix of size 12×18 was found by solving (14) given a randomly selected square dictionary composed of 6 blocks of size 3. The sub-block entries are highlighted by red squares.
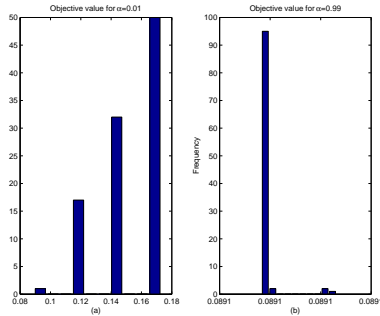

Fig. 3. Histograms of the objective values obtained when solving (14) 100 times with $\alpha = 0.01$ (a) and $\alpha = 0.99$ (b), for a given randomly generated square dictionary composed of 6 blocks of size 3. The sensing matrices of size 12×18 are initialized as matrices with random entries. Note that the distribution is insignificant in (b), indicating that in this specific case, every local minimum is also a global minimum.
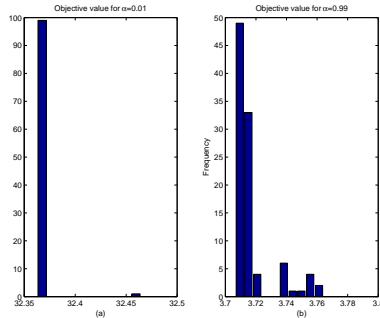

Fig. 4. Histograms of the objective values obtained when solving (14) 100 times with $\alpha = 0.01$ (a) and $\alpha = 0.99$ (b), for a given randomly generated overcomplete dictionary composed of 24 blocks of size 3. The sensing matrices of size 12×18 are initialized as matrices with random entries.

IV. WEIGHTED COHERENCE MINIMIZATION

In this section, we present the *Weighted Coherence Minimization* (WCM) algorithm for minimizing (14), based on the bound-optimization method [25]. This algorithm substitutes the original objective with an easier to minimize surrogate objective that is updated in each optimization step. After defining a surrogate function and showing it can be minimized in closed form, we prove that its iterative minimization is guaranteed to converge to a local solution of the original problem.

A. The Weighted Coherence Minimization Algorithm

To obtain a surrogate function we rewrite the objective of (14), which we denote by $f(G)$, as a function of the Gram matrix of the equivalent dictionary $G = D'A'AD$:

$$f(G) \equiv \frac{1}{2}\eta(G) + (1-\alpha)\mu_B^t(G) + \alpha\nu^t(G)$$
$$= \frac{1}{2}\|u_\eta(G)\|_F^2 + (1-\alpha)\|u_\mu(G)\|_F^2 + \alpha\|u_\nu(G)\|_F^2,$$

where the matrix operators u_μ, u_ν and u_η are defined as:

$$u_\eta(G)[i,j]_n^m = \begin{cases} G[i,j]_n^m - 1, & i=j, m=n; \\ 0, & \text{else} \end{cases},$$

$$u_\mu(G)[i,j]_n^m = \begin{cases} G[i,j]_n^m, & i \neq j; \\ 0, & \text{else} \end{cases},$$

$$u_\nu(G)[i,j]_n^m = \begin{cases} G[i,j]_n^m, & i=j, m \neq n; \\ 0, & \text{else} \end{cases},$$

with $G[i,j]_n^m$ denoting the (m,n)th entry of $G[i,j]$. This equation follows directly from the definitions of η, μ_B^t and ν^t. We can now write:

$$f(G) = \frac{1}{2}\|G - h_\eta(G)\|_F^2 + (1-\alpha)\|G - h_\mu(G)\|_F^2$$
$$+ \alpha\|G - h_\nu(G)\|_F^2, \quad (15)$$

where the matrix operators h_μ, h_ν and h_η are defined as:

$$h_\eta(G)[i,j]_n^m = \begin{cases} 1, & i=j, m=n; \\ G[i,j]_n^m, & \text{else} \end{cases},$$

$$h_\mu(G)[i,j]_n^m = \begin{cases} 0, & i \neq j; \\ G[i,j]_n^m, & \text{else} \end{cases},$$

$$h_\nu(G)[i,j]_n^m = \begin{cases} 0, & i=j, m \neq n; \\ G[i,j]_n^m, & \text{else} \end{cases}.$$

Based on (15), we define a surrogate objective $g(G, G^{(n)})$ at the nth iteration as:

$$g(G, G^{(n)}) \equiv \frac{1}{2}\|G - h_\eta(G^{(n)})\|_F^2 + (1-\alpha)\|G - h_\mu(G^{(n)})\|_F^2$$
$$+ \alpha\|G - h_\nu(G^{(n)})\|_F^2, \quad (16)$$

where $G^{(n)} = D'A^{(n)'}A^{(n)}D$ is the Gram matrix of the equivalent dictionary from the previous iteration. In Appendix B, we prove that $g(G, G^{(n)})$ satisfies the conditions of a surrogate objective for the bound-optimization method. Therefore, iteratively minimizing $g(G, G^{(n)})$ is guaranteed to converge to the minimum of the original objective $f(G)$, i.e., solve (14).

The following proposition describes the closed form solution to minimizing $g(G, G^{(n)})$ at each iteration.

Proposition 1**:** The function $g(G, G^{(n)})$ is minimized by choosing

$$A^{(n+1)} = \Delta_M^{1/2} V_M' \Lambda^{-1/2} U',$$

where $U\Lambda U'$ is the eigenvalue decomposition of DD', Δ_M and V_M are the top M eigenvalues and the corresponding M eigenvectors of $\Lambda^{-1/2} U' D h_t(G^{(n)}) D' U \Lambda^{-1/2}$, and:

$$h_t(\cdot) \equiv \frac{2}{3}\left(\frac{1}{2}h_\eta(\cdot) + (1-\alpha)h_\mu(\cdot) + \alpha h_\nu(\cdot)\right). \quad (17)$$

Proof: See Appendix B. ∎

A summary of the proposed WCM algorithm is given below.

Algorithm 1 Weighted Coherence Minimization

Task: Solve for a given block-sparsifying dictionary $D_{N\times K}$:

$$A = \arg\min_A \frac{1}{2}\eta + (1-\alpha)\mu_B^t + \alpha\nu^t,$$

where $A \in R^{M\times N}$.

Initialization: Calculate the eigenvalue decomposition of $DD' = U\Lambda U'$. Set $A^{(0)}$ as the outcome of (4), i.e., $A^{(0)} = [I_M\ 0]\Lambda^{-1/2}U'$, and $n = 0$.

Repeat until convergence:

1) Set $G^{(n)} = D'A^{(n)'}A^{(n)}D$.
2) Calculate $h_t(G^{(n)})$ as in (17).
3) Find the top M eigenvalues Δ_M and the corresponding M eigenvectors V_M of $\Lambda^{-1/2}U'Dh_t(G^{(n)})D'U\Lambda^{-1/2}$.
4) Set $A^{(n+1)} = \Delta_M^{1/2}V_M'\Lambda^{-1/2}U'$.
5) $n = n + 1$.

V. Experiments

In this section, we evaluate the contribution of the proposed sensing matrix design framework empirically. We compare the recovery and classification abilities of BOMP [22], [23] when using sensing matrices designed by our methods to the outcome of (4), which will be referred to as "Duarte-Sapiro" (DS) [8].

For each simulation, we repeat the following procedure 100 times. We randomly generate a dictionary $D_{N\times K}$ with normally distributed entries and normalize its columns. In order to evaluate WCM on structured dictionaries as well, we repeat the simulations using a dictionary containing N randomly selected rows of the $K \times K$ Discrete Cosine Transform (DCT) matrix. The dictionary is divided into K/s blocks of size s. We then generate $L = 1000$ test signals X of dimension K that have k-block-sparse representations Θ with respect to D. The generating blocks are chosen randomly and independently and the coefficients are i.i.d. uniformly distributed. $A_{M\times N}$ is initialized as the outcome of DS. We find A using the WCM algorithm, and calculate the equivalent dictionary $E = AD$ and the measurements $Y = AX$. Next, we obtain the block-sparsest representations of the measurements, $\hat{\Theta}$, by applying BOMP with a fixed number of k nonzero blocks.

We use two measures to evaluate the success of the simulations based on their outputs A and $\hat{\Theta}$:

- The percentage of recognized generating subspaces of X (i.e., successful classification): $r = \frac{\|\hat{\Theta}\odot\Theta\|_0}{Lks}$ where $\odot$ denotes element-wise multiplication.
- The normalized representation error $e = \frac{\|X - D\hat{\Theta}\|_F}{\|X\|_F}$

To evaluate the performance of the WCM algorithm as a function of α, we choose $s = 3$, $N = 60$ and $K = 2N = 120$. We repeat the experiment for both types of dictionaries, and for $k = 1$ (Fig. 5(a),5(b)), $k = 2$ (Fig. 6(a),6(b)) and $k = 3$ (Fig. 7(a),7(b)) nonzero blocks, with respectively $M = 6$, $M = 14$ and $M = 20$ measurements. To show that the results remain consistent for higher values of k, we add an experiment with $k = 6$, $M = 35$, $N = 180$ and $K = 2N = 360$ (Fig. 8(a),8(b)). We compare the obtained results to randomly set sensing matrices and to the outputs of DS [8], based on the normalized representation error e, the classification success r, and the ratio between the total sub-block coherence and the total inter-block coherence ν^t/μ_B^t. We observe that WCM and DS coincide at $\alpha = 0.5$ for all the three measures, as expected. Note that for $\alpha < 0.5$ we get that ν^t/μ_B^t is high, e is high and r is low. On the other hand, when $\alpha > 0.5$, i.e., when giving more weight to ν^t and less to μ_B^t, the signal reconstruction as well as the signal classification are improved compared to DS. While the improvement for $k = 1$ is more significant, it is maintained for higher values of k as well. Remarkably, for structured dictionaries and for higher values of k, we see that $\alpha < 0.5$ leads to an improvement of r. However, e is compromised in this case. We can conclude that when designing sensing matrices for block sparse decoding, the best results are obtained by choosing α close enough to 1. In other words, the best recovery results are obtained when the equivalent dictionary has nearly orthonormal blocks. This holds for dictionaries containing normally distributed entries as well as for dictionaries containing randomly selected rows of the DCT matrix. As was the case in Fig. 3(b), we observed empirically that for $\alpha > 0.5$, every local minimum is a global minimum as well. This means that the WCM algorithm converges to a global solution of (14) when $\alpha > 0.5$, for all the experiments presented in this section. We emphasize however, that this may not be the case for other sets of parameters.

Fig. 9(a) and Fig. 9(b) show that when using WCM with $\alpha = 0.99$ on dictionaries with normally distributed entries and on structured dictionaries, the improvement in signal recovery using is maintained for a wide range of K, starting from square dictionaries, i.e. $K = N$, to highly overcomplete dictionaries. For this experiment, we chose $s = 3$, $N = 60$, $k = 2$ and $M = 14$. We note that for both types of dictionaries, the improvement of WCM over DS increases as the dictionary becomes more overcomplete.

Finally, we show that WCM improves the results of block-sparse decoding for dictionaries with blocks of varying sizes as well. The generated dictionaries contain 15 blocks of size 4 and 20 blocks of size 3, with $N = 60$ and $K = 2N = 120$. In this example, we set $k = 2$ and $M = 14$. The results are shown as a function of α in Fig. 10(a) for dictionaries with normally distributed entries and in Fig. 10(b) for structured dictionaries.
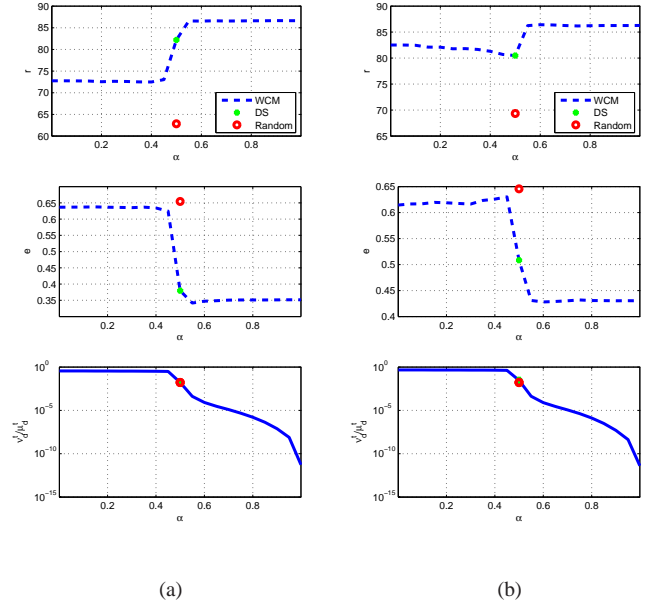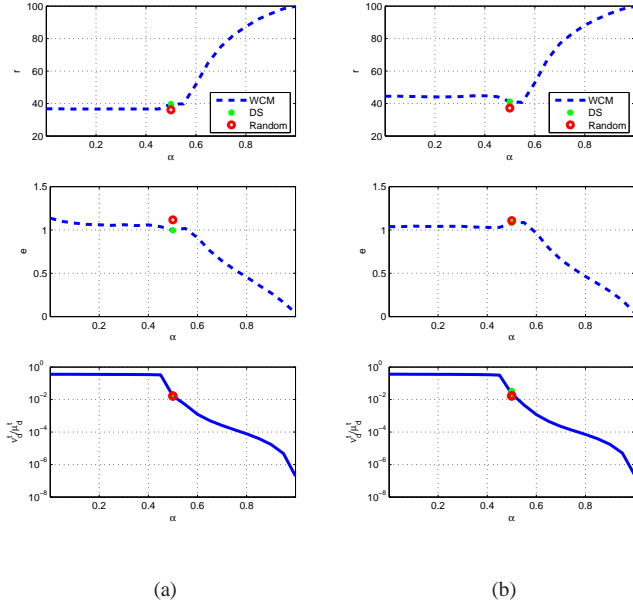
(a)   (b)

Fig. 5. Simulation results of sensing matrix design using the WCM algorithm with $k = 1$ and $M = 6$. The graphs show the normalized representation error e, the classification success r, and the ratio between the total sub-block coherence and the total inter-block coherence ν^t / μ_B^t as a function of α. In (a) the dictionary contains normally distributed entries, and in (b) randomly selected rows of the DCT matrix.
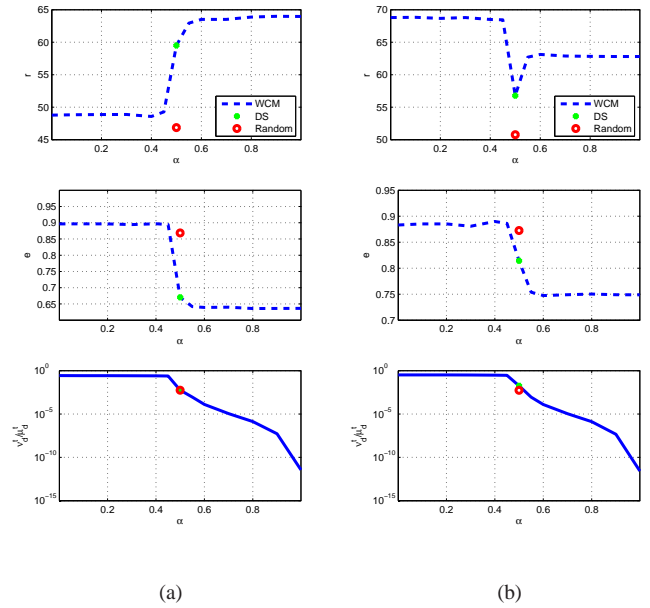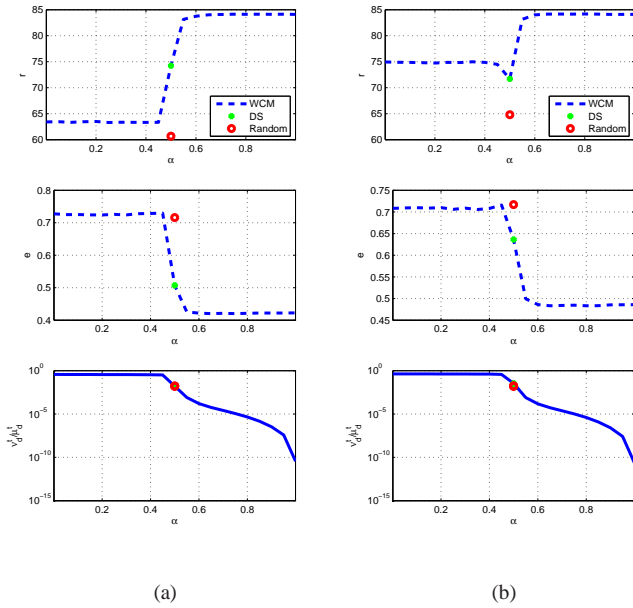
(a)   (b)

Fig. 7. Simulation results of sensing matrix design using the WCM algorithm with $k = 3$ and $M = 20$. The graphs show the normalized representation error e, the classification success r, and the ratio between the total sub-block coherence and the total inter-block coherence ν^t / μ_B^t as a function of α. In (a) the dictionary contains normally distributed entries, and in (b) randomly selected rows of the DCT matrix.

(a)   (b)

Fig. 6. Simulation results of sensing matrix design using the WCM algorithm with $k = 2$ and $M = 14$. The graphs show the normalized representation error e, the classification success r, and the ratio between the total sub-block coherence and the total inter-block coherence ν^t / μ_B^t as a function of α. In (a) the dictionary contains normally distributed entries, and in (b) randomly selected rows of the DCT matrix.

(a)   (b)

Fig. 8. Simulation results of sensing matrix design using the WCM algorithm with $k = 6$ and $M = 35$. The graphs show the normalized representation error e, the classification success r, and the ratio between the total sub-block coherence and the total inter-block coherence ν^t / μ_B^t as a function of α. In (a) the dictionary contains normally distributed entries, and in (b) randomly selected rows of the DCT matrix.
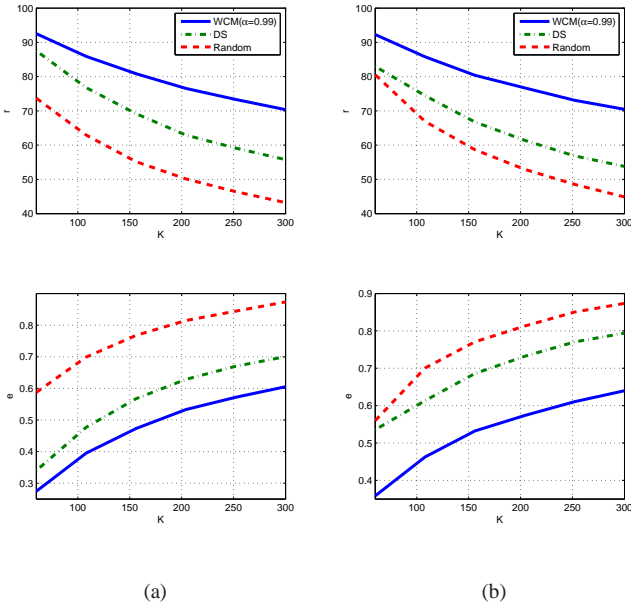
Fig. 9. Simulation results of sensing matrix design using the WCM algorithm with $k = 2$ and $M = 14$. The graphs show the normalized representation error e and the classification success r as a function of K. In (a) the dictionary contains normally distributed entries, and in (b) randomly selected rows of the DCT matrix.
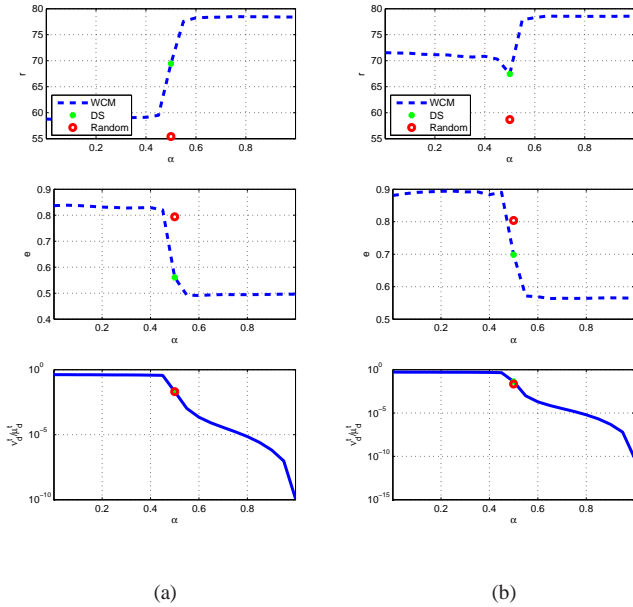

Fig. 10. Simulation results of sensing matrix design using the WCM algorithm on a dictionary containing 15 blocks of size 4 and 20 blocks of size 3, with $k = 2$ and $M = 14$. The graphs show the normalized representation error e, the classification success r, and the ratio between the total sub-block coherence and the total inter-block coherence ν^t/μ_B^t as a function of α. In (a) the dictionary contains normally distributed entries, and in (b) randomly selected rows of the DCT matrix.

VI. CONCLUSIONS

In this paper, we proposed a framework for the design of a sensing matrix, assuming that a block-sparsifying dictionary is provided. We minimize a weighted sum of the total inter-block coherence and the total sub-block coherence, while attempting to keep the atoms in the equivalent dictionary as normalized as possible (see (14)). This objective can be seen as an intuitive extension of (4) to the case of blocks.

While it might be possible to derive a closed form solution to (14), we have presented the Weighted Coherence Minimization algorithm, an elegant iterative solution which is based on the bound-optimization method. In this method, the original objective is replaced with an easier to solve surrogate objective in each step. This algorithm eventually converges to a local solution of (14).

Simulations have shown that the best results are obtained when minimizing mostly the total sub-block coherence. This leads to equivalent dictionaries with nearly orthonormal blocks, at the price of a slightly increased total inter-block coherence. The obtained sensing matrix outperforms the one obtained when using the DS algorithm [8] to solve (4). This improvement manifests itself in lower signal reconstruction errors and higher rates of successful signal classification. When giving equal weight to the total inter-block coherence and to the total sub-block coherence, the results are identical to solving (4). Moreover, both objectives coincide for this specific choice of α, which ignores the existence of a block structure in the sparse representations of the signal data.

APPENDIX A
PROOF OF CONVERGENCE

The surrogate function $g(G, G^{(n)})$ has been chosen in such a way as to bound the original objective $f(G)$ from above for every G, and to coincide at $G = G^{(n)}$. Minimizing $g(G, G^{(n)})$ will then necessarily decrease the value of $f(G)$:

$$\min_G g(G, G^{(n)}) \le g(G^{(n)}, G^{(n)}) = f(G^{(n)}),$$
$$f(G^{(n+1)}) \le g(G^{(n+1)}, G^{(n)}) = \min_G g(G, G^{(n)}).$$

Formally, according to [25], the sequence of solutions generated by iteratively solving

$$G^{(n+1)} = \arg\min_G g(G, G^{(n)}) \qquad (18)$$

is guaranteed to converge to a local minimum of the original objective $f(G)$ when the surrogate objective $g(G, G^{(n)})$ satisfies the following three constraints:

1) **Equality at $G = G^{(n)}$:**

$$g(G^{(n)}, G^{(n)}) = f(G^{(n)}).$$

2) **Upper-bounding the original function:**

$$g(G, G^{(n)}) \ge f(G), \forall G.$$

3) **Equal gradient at $G = G^{(n)}$:**

$$\nabla g(G, G^{(n)})|_{G=G^{(n)}} = \nabla f(G)|_{G=G^{(n)}}.$$

We next prove that the three conditions hold.

Proof: **Equality at** $G = G^{(n)}$**:** This follows from the definition of $g(G, G^{(n)})$.

Upper-bounding the original function: Let us rewrite both functions $g(G, G^{(n)})$ and $f(G)$ using the definition of the Frobenius norm:

$$g(G, G^{(n)}) =$$
$$\sum_{i,j} \sum_{m,n} \left[\frac{1}{2}((G - h_\eta(G^{(n)}))[i,j]_n^m)^2 \right.$$
$$+ (1-\alpha)((G - h_\mu(G^{(n)}))[i,j]_n^m)^2$$
$$\left. + \alpha((G - h_\nu(G^{(n)}))[i,j]_n^m)^2 \right],$$

and

$$f(G) =$$
$$\sum_{i,j} \sum_{m,n} \left[\frac{1}{2}(u_\eta(G)[i,j]_n^m)^2 + (1-\alpha)(u_\mu(G)[i,j]_n^m)^2 \right.$$
$$\left. + \alpha(u_\nu(G)[i,j]_n^m)^2 \right].$$

The following observations prove that each of the terms in $g(G, G^{(n)})$ is larger than or equal to its counterpart in $f(G)$, and therefore $g(G, G^{(n)}) \geq f(G)$:

$$u_\eta(G)[i,j]_n^m = \begin{cases} G[i,j]_n^m - 1, & i = j, m = n; \\ 0, & \text{else.} \end{cases}$$

$$(G - h_\eta(G^{(n)}))[i,j]_n^m = \begin{cases} G[i,j]_n^m - 1, & i = j, m = n; \\ (G - G^{(n)})[i,j]_n^m, & \text{else.} \end{cases}$$

$$u_\mu(G)[i,j]_n^m = \begin{cases} G[i,j]_n^m, & i \neq j; \\ 0, & \text{else.} \end{cases}$$

$$(G - h_\mu(G^{(n)}))[i,j]_n^m = \begin{cases} G[i,j]_n^m, & i \neq j; \\ (G - G^{(n)})[i,j]_n^m, & \text{else.} \end{cases}$$

$$u_\nu(G)[i,j]_n^m = \begin{cases} G[i,j]_n^m, & i = j, m \neq n; \\ 0, & \text{else.} \end{cases}$$

$$(G - h_\nu(G^{(n)}))[i,j]_n^m = \begin{cases} G[i,j]_n^m, & i = j, m \neq n; \\ (G - G^{(n)})[i,j]_n^m, & \text{else.} \end{cases}$$

Equal gradient at $G = G^{(n)}$**:** We calculate the gradient of $g(G, G^{(n)})$ and $f(G)$:

$$\nabla g(G, G^{(n)}) =$$
$$2 \left[\frac{1}{2}(G - h_\eta(G^{(n)})) + (1-\alpha)(G - h_\mu(G^{(n)})) \right.$$
$$\left. + \alpha(G - h_\nu(G^{(n)})) \right],$$
$$\nabla f(G) = 2 \left[\frac{1}{2}u_\eta(G) + (1-\alpha)u_\mu(G) + \alpha u_\nu(G) \right].$$

When substituting $G = G^{(n)}$ we obtain:

$$\nabla g(G, G^{(n)})|_{G=G^{(n)}} = \nabla f(G)|_{G=G^{(n)}}$$
$$= 2(\frac{1}{2}u_\eta(G^{(n)}) + (1-\alpha)u_\mu(G^{(n)}) + \alpha u_\nu(G^{(n)})).$$

Therefore, the gradients of both objectives coincide at $G = G^{(n)}$. This completes the convergence proof. ∎

APPENDIX B
PROOF OF PROPOSITION 1

Proof: In order to minimize $g(G, G^{(n)})$, we rewrite the problem in an alternative form:

$$\min_A g(G, \cdot) =$$
$$\min_A \text{tr} \left(\frac{3}{2}G'G - 2G' \left[\frac{1}{2}h_\eta(\cdot) + (1-\alpha)h_\mu(\cdot) + \alpha h_\nu(\cdot) \right] \right)$$
$$= \min_A \text{tr}(E'EE'E - 2E'Eh_t(\cdot))$$
$$= \min_A \text{tr}(EE'EE' - 2Eh_t(\cdot)E')$$
$$= \min_A \text{tr}(ADD'A'ADD'A' - 2ADh_t(\cdot)D'A'), \quad (19)$$

where $h_t(\cdot)$ is defined in (17). Let $U\Lambda U'$ be the eigenvalue decomposition of DD' and define $\Gamma_{M \times N} = AU\Lambda^{1/2}$. Substituting into (19) yields:

$$\min_A g(G, \cdot) =$$
$$\min_A \text{tr}(\Gamma\Gamma'\Gamma\Gamma' - 2\Gamma\Lambda^{-1/2}U'Dh_t(\cdot)D'U\Lambda^{-1/2}\Gamma')$$
$$= \min_A \|\Gamma'\Gamma - \tilde{h}_t(\cdot)\|_F^2, \quad (20)$$

where $\tilde{h}_t(\cdot) \equiv \Lambda^{-1/2}U'Dh_t(\cdot)D'U\Lambda^{-1/2}$. According to (20), the surrogate objective $g(G, G^{(n)})$ can be minimized in closed form by finding the top M components of $\tilde{h}_t(G^{(n)})$. Let Δ_M be the top M eigenvalues of $\tilde{h}_t(G^{(n)})$ and V_M the corresponding M eigenvectors. Then, (20) is solved by setting $\Gamma = \Delta_M^{1/2}V_M'$. Note that this solution is not unique, since Γ can be multiplied on the left by any unitary matrix. Finally, the optimal sensing matrix is given by $A^{(n+1)} = \Gamma\Lambda^{-1/2}U' = \Delta_M^{1/2}V_M'\Lambda^{-1/2}U'$. The resulting Gram matrix $G^{(n+1)}$ is not influenced by the multiplication of $A^{(n+1)}$ on the left by a unitary matrix. Therefore, the WCM algorithm is not affected by the choice of $A^{(n+1)}$. ∎

References

[1] E. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inform. Theory*, vol. 52, pp. 489–509, Feb. 2006.

[2] D. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[3] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1999.

[4] J. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Inform. Theory*, vol. 50, no. 10, pp. 2231–2242, Oct. 2004.

[5] S. G. Mallat and Z. Zhang, "Matching pursuits and time-frequency dictionaries," *IEEE Trans. Sig. Proc.*, vol. 41, no. 12, pp. 33973415, Dec. 1993.

[6] Y. Weiss, H. S. Chang, and W. T. Freeman, "Learning compressed sensing," *Allerton Conference*, Sep 2007.

[7] M. Elad, "Optimized projections for compressed sensing," *IEEE Trans. Signal Process.*, vol. 55, no. 12, pp. 5695–5702, Dec. 2007.

[8] J. M. Duarte-Carvajalino and G. Sapiro, "Learning to sense sparse signals: Simultaneous sensing matrix and sparsifying dictionary optimization," *IMA Preprint Series*, , no. 2211, May 2008.

[9] Y. M. Lu and M. N. Do, "A theory for sampling signals from a union of subspaces," *IEEE Trans. Sig. Proc.*, vol. 56, no. 6, pp. 2334 – 2345, Jun. 2008.

[10] T. Blumensath and M. E. Davies, "Sampling theorems for signals from the union of finite-dimensional linear subspaces," *IEEE Trans. Inform. Theory*, vol. 55, no. 4, pp. 1872 – 1882, Apr. 2009.

[11] K. Gedalyahu and Y. C. Eldar, "Time delay estimation from low rate samples: A union of subspaces approach," *IEEE Trans. Signal Process.*, vol. 58, no. 6, pp. 3017–3031, Jun. 2010.

[12] Y. C. Eldar and M. Mishali, "Robust recovery of signals from a structured union of subspaces," *IEEE Trans. Inform. Theory*, vol. 55, no. 11, pp. 5302–5316, Nov. 2009.

[13] R. Basri and D. Jacobs, "Lambertian refelectances and linear subspaces," *IEEE Transactions On Pattern Analysis And Machine Intelligence*, vol. 25, no. 2, pp. 383–390, Feb 2003.

[14] A. Y. Yang, J. Wright, Y. Ma, and S. Sastry, "Feature selection in face recognition: A sparse representation perspective," *UC Berkeley Tech Report*, Aug 2007.

[15] R. Vidal and Y. Ma, "A unified algebraic approach to 2-D and 3-D motion segmentation and estimation," *Journal of Mathematical Imaging and Vision*, vol. 25, no. 3, pp. 403–421, Oct. 2006.

[16] M. Mishali and Y. C. Eldar, "Blind multiband signal reconstruction: Compressed sensing for analog signals," *IEEE Trans. Sig. Proc.*, vol. 57, no. 3, pp. 993–1009, Mar. 2009.

[17] M. Mishali and Y. C. Eldar, "From theory to practice: Sub-Nyquist sampling of sparse wideband analog signals," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 2, pp. 375 – 391, Apr. 2010.

[18] H. J. Landau, "Necessary density conditions for sampling and interpolation of certain entire functions," *Acta Math.*, vol. 117, no. 1, pp. 37–52, 1967.

[19] F. Parvaresh, H. Vikalo, S. Misra, and B. Hassibi, "Recovering sparse signals using sparse measurement matrices in compressed dna microarrays," *IEEE Journal of Selected Topics in Signal Processing*, Jun. 2008.

[20] M. Stojnic, F. Parvaresh, and B. Hassibi, "On the reconstruction of block-sparse signals with an optimal number of measurements," *IEEE Trans. Sig. Proc.*, vol. 57, no. 8, pp. 3075–3085, Aug. 2009.

[21] H. Rauhut and Y.C. Eldar, "Average case analysis of multichannel sparse recovery using convex relaxation," *IEEE Trans. Inform. Theory*, vol. 56, no. 1, pp. 505–519, Jan. 2010.

[22] Y. C. Eldar, P. Kuppinger, and H. Bölcskei, "Block-sparse signals: Uncertainty relations and efficient recovery," *IEEE Trans. Sig. Proc.*, Apr. 2010.

[23] Y. C. Eldar and H. Bölcskei, "Block-sparsity: Coherence and efficient recovery," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 0, pp. 2885–2888, 2009.

[24] K. Rosenblum, L. Zelnik-Manor, and Y. C. Eldar, "Dictionary optimization for block-sparse representations," *arXiv.org 1005.0202. submitted to IEEE Trans. Signal Process.*, May 2010.

[25] M.A. Figueiredo, J.M. Bioucas-Dias, and R.D. Nowak, "Majorization-minimization algorithms for wavelet-based image restoration," *IEEE Trans. on Image Process.*, vol. 16, no. 12, pp. 2980 – 2991, Dec. 2007.