Signal Processing 92 (2012) 999–1009

Contents lists available at SciVerse ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

A gradient-based alternating minimization approach for optimization of the measurement matrix in compressive sensing

Vahid Abolghasemi*, Saideh Ferdowsi, Saeid Sanei

Faculty of Engineering and Physical Sciences, University of Surrey, Guildford GU2 7XH, Surrey, UK

ARTICLE INFO

Article history:
Received 7 December 2010
Received in revised form
27 July 2011
Accepted 14 October 2011
Available online 25 October 2011

Keywords:
Basis pursuit (BP)
Compressive sensing (CS)
Gradient descent
Orthogonal matching pursuit (OMP)
Random Gaussian matrices
Sparse representation

ABSTRACT

In this paper the problem of optimization of the measurement matrix in compressive (also called compressed) sensing framework is addressed. In compressed sensing a measurement matrix that has a small *coherence* with the sparsifying dictionary (or basis) is of interest. Random measurement matrices have been used so far since they present small *coherence* with almost any sparsifying dictionary. However, it has been recently shown that optimizing the measurement matrix toward decreasing the *coherence* is possible and can improve the performance. Based on this conclusion, we propose here an alternating minimization approach for this purpose which is a variant of Grassmannian frame design modified by a gradient-based technique. The objective is to optimize an initially random measurement matrix to a matrix which presents a smaller *coherence* than the initial one. We established several experiments to measure the performance of the proposed method and compare it with those of the existing approaches. The results are encouraging and indicate improved reconstruction quality, when utilizing the proposed method.

1. Introduction

Sparse signal recovery problem has been widely researched within signal and image processing communities in recent years. Sparse signals/images have few non-zero components. A signal or image with an underlying sparse structure can be more efficiently and advantageously compressed, stored or transmitted than normal non-sparse signals. Such an optimum compression requires a novel sampling scheme, followed by an appropriate reconstruction method, different from conventional techniques. Given the compressed measurements $\mathbf{y}$ and the sampling pattern $\mathbf{D}$, the main challenge is to solve the following inverse problem:

$$\min\|\boldsymbol{\alpha}\|_0 \quad \text{s.t. } \mathbf{y}=\mathbf{D}\boldsymbol{\alpha}, \tag{1}$$

where $\mathbf{y}\in\mathbb{R}^p$ and the rectangular matrix $\mathbf{D}\in\mathbb{R}^{p\times m}$ with $p<m$ are known and the sparse signal $\boldsymbol{\alpha}\in\mathbb{R}^m$ should be found. ℓ_0-norm ($\|\cdot\|_0$) counts the number of non-zero components and is a measure of sparsity degree in the signal. Note that in most cases the original signal is sparse in a different domain (e.g. Fourier, wavelet, or discrete cosine transform (DCT)). In such cases the notation changes to $\mathbf{y}=\boldsymbol{\Phi}\mathbf{x}=\boldsymbol{\Phi}\boldsymbol{\Psi}\boldsymbol{\alpha}=\mathbf{D}\boldsymbol{\alpha}$, where $\boldsymbol{\Phi}\in\mathbb{R}^{p\times n}$ and $\boldsymbol{\Psi}\in\mathbb{R}^{n\times m}$ are called measurement (sensing) matrix and sparsifying transform, respectively. Note that $\boldsymbol{\Phi}$ is always an overcomplete matrix ($p\ll n$) and $\boldsymbol{\Psi}$ can either be overcomplete ($n<m$), called dictionary, or complete ($n=m$) which is called basis.

One of the recent emerging fields having direct connection to sparse recovery problem is compressed sensing (CS) [1,2]. In CS we not only attempt to find a sparse solution for (1), but also discuss the possible ways to generate $\boldsymbol{\Phi}$ to compressively acquire the input signal and yet not to violate the uniqueness conditions for sparse recovery in (1). In other words, CS targets Shannon's theorem [3] and takes the advantages of sparsity to decrease the sampling rate.

The major difficulty in solving (1) is its non-convexity. Its solution requires a combinatorial search through all

* Corresponding author. Tel.: +44 7529343857.
E-mail addresses: v.abolghasemi@surrey.ac.uk, vabolghasemi@ieee.org (V. Abolghasemi).

0165-1684/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.sigpro.2011.10.012

possible sparse $\boldsymbol{\alpha}$ which is not feasible. However, there have been reported several approaches providing the solution for the sparse recovery problem, by either relaxing it to

$$\min\|\boldsymbol{\alpha}\|_1 \quad \text{s.t. } \mathbf{y}=\mathbf{D}\boldsymbol{\alpha}, \tag{2}$$

where $\|\boldsymbol{\alpha}\|_1=\sum_i|\alpha_i|$ refers to ℓ_1-norm, or using a greedy strategy. Note that in addition to (1) and (2), there are other formulations as well which are not presented here. Some of the well known recovery methods are basis pursuit (BP) [4], orthogonal matching pursuit (OMP) [5], least angle regression (LARS) [6], and least absolute shrinkage and selection operator (LASSO) [7].

Although *sparsity* is an essential requirement for the signals in the CS framework, it is not sufficient for a signal to be successfully reconstructed. Thus far, it is realized that in order to be able to attain a reasonably small p and yet have a stable reconstruction, $\boldsymbol{\Phi}$ requires to have specific properties [8], the one which is focused on here, is *mutual coherence*. In CS theory, a suitable sensing matrix $\boldsymbol{\Phi}$ is desired to be as incoherent as possible with sparsifying matrix $\boldsymbol{\Psi}$. In other words, the correlation between any distinct pair of columns in equivalent matrix $\mathbf{D}$ should be very small, and that means to have a nearly orthogonal matrix $\mathbf{D}$. It is shown that random matrices with Gaussian or Bernoulli distributions are appropriate choices for $\boldsymbol{\Phi}$ [9], as they satisfy this property with high probability and can be generated non-adaptively. Although $\boldsymbol{\Phi}$ is normally selected to be random, some recent works have been reported in which the authors attempt to find an optimal structure for $\boldsymbol{\Phi}$ (explicitly or implicitly) with the hope of increasing the reconstruction quality and to take fewer measurements [10–15]. Elad [10] attempts to iteratively decrease the average mutual coherence using a shrinkage operation followed by a singular value decomposition (SVD) step. In [11] the authors apply a kind of non-uniform sampling by segmenting the input signal and taking samples with different rates from each segment. In [12], which is an application to magnetic resonance imaging (MRI), the authors define an incoherence criterion based on point spread function (PSF) and propose a Monte Carlo scheme for random incoherent sampling of this type of data. Wang et al. [13] propose a variable density sampling strategy by exploiting the prior information about the statistical distributions of natural images in the wavelet domain. Their proposed method is computationally efficient and can be applied to several transform domains. In another work in [14], Wang et al. propose to generate colored random projections using an adaptive scheme. Duarte-Carvajalino et al. [15] take the advantage of an eigenvalue decomposition process followed by a KSVD-based algorithm [15] (see also [16]) to optimize $\boldsymbol{\Phi}$ and learn dictionary $\boldsymbol{\Psi}$, respectively. The results of all previous methods reveal the improved performance which is an evidence of the benefits that optimal sampling can provide for this framework.

Our contribution in this paper is to develop a novel approach for optimizing the measurement matrix. We propose a gradient-based method to optimize a randomly selected measurement matrix to decrease the coherence with the given sparsifying matrix. The proposed algorithm, which is related to the design of Grassmannian frames [17–19], aims at minimizing a cost function in an alternating scheme. The results of our experiments and simulations confirm the ability of the proposed method to decrease the coherence leading to higher reconstruction quality.

The rest of the paper is organized as follows. In the next section, we first define the criteria that can be used for measuring the coherence of a matrix. Then, the proposed approach for optimization of the measurement matrix is described. In Section 3 an adaptive scheme to choose the optimum stepsize is proposed. In Section 4, some simulations are given to examine the proposed method from the practical perspectives. The discussion and conclusion are presented in Sections 5 and 6, respectively.

2. Measurement matrix optimization

2.1. Problem formulation

Following the notations stated earlier, assume $\boldsymbol{\alpha}$ has $s \ll m$ non-zero samples and called s-sparse. A successful CS requires $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ to be incoherent. A good measure of coherence between these two matrices (or equivalently columns of $\mathbf{D}$) can be obtained by referring to the definition of *mutual coherence* [20,21]. Here, we quote this definition as that presented by Elad [10]:

Definition 1. For a matrix $\mathbf{D}$, the mutual coherence is defined as the maximum absolute value and normalized inner product between all columns in $\mathbf{D}$ that can be described as

$$\mu_{mx}(\mathbf{D})=\max_{i\neq j,\,1\le i,j\le m}\left\{\frac{|\mathbf{d}_i^T\mathbf{d}_j|}{\|\mathbf{d}_i\|_2\cdot\|\mathbf{d}_j\|_2}\right\}. \tag{3}$$

A desired $\boldsymbol{\Phi}$ has a small μ_{mx} with respect to $\boldsymbol{\Psi}$. An alternative description for μ_{mx} is obtained by referring to the corresponding Gram matrix $\tilde{\mathbf{G}}=\tilde{\mathbf{D}}^T\tilde{\mathbf{D}}$, in which $\tilde{\mathbf{D}}$ is column-normalized version of $\mathbf{D}$. We set two coherence measures based on this matrix, which are used to study the performance of different methods later in the Results section. These parameters are the maximum and average of off-diagonal elements in $\tilde{\mathbf{G}}$, denoted respectively as

$$\mu_{mx}=\max_{i\neq j,\,1\le i,j\le m}|\tilde{g}_{ij}|, \tag{4}$$

$$\mu_{av}=\frac{\sum_{j\neq i}\sum_{i\neq j}|\tilde{g}_{ij}|}{m(m-1)}. \tag{5}$$

There are important reasons which motivate us to search for measurement matrices with small coherence. The first reason is the influence that the mutual coherence has on the achievable sparsity bound for the given signal. A detailed discussion about this fact has been documented in [21,22]. In a brief statement, if we suppose that the following inequality holds in a noiseless setting:

$$\|\boldsymbol{\alpha}\|_0<\frac{1}{2}\left(1+\frac{1}{\mu_{mx}(\mathbf{D})}\right) \tag{6}$$

then, $\boldsymbol{\alpha}$ is necessarily the sparsest signal when $\mathbf{y}$ and $\mathbf{D}$ are known. More importantly, both BP and OMP are guaranteed to successfully recover it [21,22]. This implies that as long as

the mutual coherence is small, a successful reconstruction is achievable for a wider range of sparsity degree.

Another significant requirement is the *restricted isometry property* (RIP) [9,20], which is defined with respect to an isometry constant $0<\delta\le 1$. For a s-sparse signal $\boldsymbol{\alpha}$ and any integer $\tau\le s$, the isometry constant of $\mathbf{D}$ is the smallest δ_s that satisfies

$$(1-\delta_s)\|\boldsymbol{\alpha}_\tau\|_2^2\le\|\mathbf{D}_\tau\boldsymbol{\alpha}_\tau\|_2^2\le(1+\delta_s)\|\boldsymbol{\alpha}_\tau\|_2^2. \tag{7}$$

This property implies that for a proper isometry constant (the ideal δ is 0), RIP ensures that any τ subset of columns in $\mathbf{D}$ with cardinality less than s is nearly orthogonal. This orthogonality can be related to the coherence of columns; less coherence corresponds to higher degree of orthogonality. This results in a better CS behavior and guarantees the identifiability of the original signal by both OMP and BP [4]. Note that in noisy circumstances more severe conditions are applied to ensure the stability of reconstruction, which are not dealt here.

2.2. The proposed method

Recall from the previous section that an incoherent $\mathbf{D}$ has a small mutual coherence. In other words, the corresponding Gram matrix has its absolute off-diagonal elements close to zero and diagonal elements equal to one. This, indicates a Gram matrix which is equal to identity matrix. In a previous work [23], we proposed a method to minimize the difference between Gram matrix and identity matrix in the form of Frobenius norm:

$$\min_{\mathbf{G}}\|\mathbf{G}-\mathbf{I}\|_F^2, \tag{8}$$

where $\|\cdot\|_F^2$ denotes the Frobenius norm and $\mathbf{I}$ is identity matrix. The proposed method in [23] is able to optimize an initial $\boldsymbol{\Phi}$ to a less coherent matrix. However, further study revealed that choosing identity matrix in the above objective function is a very strict constraint, and a Gram matrix equal to $\mathbf{I}$ is only achievable for $p=m$, which is not the case here. We believe that we are able to optimize $\boldsymbol{\Phi}$ not with respect to an identity matrix, but with a matrix which is updated inside the algorithm, accordingly. We shortly propose a method in this paper which follows this idea.

Another interesting description of matrices with small coherence is inferred from the equiangular tight frame (ETF) design, which is a special type of *Grassmannian frames* [18,19]. A tight frame is a generalization of an orthonormal basis. In orthonormal bases all vectors have unit norm and they are perpendicular to each other. In a more general term, every orthonormal basis is *equiangular* [18] and that causes each pair of distinct vectors to have zero inner product. An extended and more general type of orthonormal basis is called *tight frame* which contains vectors maximally partitioned in space. Now, we study ETFs in detail. Consider m unit-norm vectors in a p-dimensional space, where $p\le m$. ETF is an arrangement of such vectors that have the same—normally minimum—absolute inner product with respect to each other (in orthonormal basis this value is zero, equivalent to 90° Euclidean angle). Suppose we arrange these vectors as columns of matrix $\tilde{\mathbf{D}}$. It is known that inner product between all vectors are simply obtained by multiplying $\tilde{\mathbf{D}}$ by its transposed: $\tilde{\mathbf{D}}^T\tilde{\mathbf{D}}$. This is equivalent to the Gram matrix $\tilde{\mathbf{G}}$ which has already been described. Note that diagonal elements of $\tilde{\mathbf{G}}$ are not of our interest since they indicate the Euclidean norms of columns of $\tilde{\mathbf{D}}$. In particular, we are only interested in the off-diagonal elements of $\tilde{\mathbf{G}}$ which correspond to inner products between any two distinct vectors defined as

$$\tilde{g}_{ij}=\cos(\theta_{ij}):=\langle\tilde{\mathbf{d}}_i,\tilde{\mathbf{d}}_j\rangle=\tilde{\mathbf{d}}_i^T\tilde{\mathbf{d}}_j\quad\text{for } i\ne j, \tag{9}$$

where $\langle\cdot,\cdot\rangle$ denotes the vector inner product. It is important to note that minimum possible absolute off-diagonals of $\tilde{\mathbf{G}}$ are zero and can occur only if $\tilde{\mathbf{D}}$ is orthonormal with $p=m$. This is not the case here, since we only deal with $p<m$. It can be shown that minimum absolute inner product for each vector pair in an equiangular tight frame is [18]:

$$\mu_E=\sqrt{\frac{m-p}{p(m-1)}}. \tag{10}$$

In fact, (10) is the minimum achievable mutual coherence of $\mathbf{D}$, too. We now summarize our discussion in the following definition [18]:

Definition 2. A matrix $\mathbf{D}$ is called an equiangular tight frame (ETF) if its corresponding Gram matrix has unit diagonal elements and off-diagonals equal to $\pm\sqrt{(m-p)/p(m-1)}$:

$$\tilde{g}_{ij}=\begin{cases}\pm\sqrt{\dfrac{m-p}{p(m-1)}} & \text{for } i\ne j,\\ 1 & \text{otherwise.}\end{cases} \tag{11}$$

In general, designing an exact ETF is a complicated procedure [18] and such frames do not exist for any arbitrary p and m—it is known that a real equiangular tight frame can exist only if $m\le p(p+1)/2$ [18]. Nonetheless, there are methods established to find ETFs such as in [17,18]. The problem we face here is more challenging, where we require to find a measurement matrix $\boldsymbol{\Phi}$ for any arbitrary $p<m$, having small coherence with a known fixed dictionary/basis $\boldsymbol{\Psi}$. In other words, $\mathbf{D}=\boldsymbol{\Phi}\boldsymbol{\Psi}$ is desired to be as close as possible (and not exactly equal) to an ETF, where $\boldsymbol{\Phi}$, and potentially $\boldsymbol{\Psi}$, are overcomplete matrices. What follows is the proposed approach to find such a measurement matrix.

Consider a convex set $\mathcal{H}_{\mu_E}$ which contains the ideal ETFs [18]:

$$\mathcal{H}_{\mu_E}=\left\{\mathbf{H}\in\mathbb{R}^{m\times m}:\mathbf{H}=\mathbf{H}^T,\operatorname{diag}\mathbf{H}=\mathbf{1},\max_{i\ne j}|h_{ij}|\le\mu_E\right\}, \tag{12}$$

where $\mathbf{1}$ is a column vector of all 'ones'. As mentioned earlier, we cannot guarantee to find a $\boldsymbol{\Phi}$ leading to a $\mathbf{D}$ with coherence exactly equal to μ_E. In other words, it is not always possible to find a $\boldsymbol{\Phi}$ with a Gram matrix lying in the set $\mathcal{H}_{\mu_E}$, because of two reasons: (1) dealing with arbitrary dimensions and not necessarily $m\le p(p+1)/2$, and (2) being constrained by the fixed matrix $\boldsymbol{\Psi}$. As a solution, we relax this situation and attempt to optimize a random $\boldsymbol{\Phi}$ to have a Gram matrix lying in a set $\mathcal{H}_\mu$ with $\mu\ge\mu_E$. In order to find such a matrix, we define the following minimization problem which needs to be solved,

alternately:

$$\min_{\mathbf{\Phi},\mathbf{H}\in\mathcal{H}_\mu} \|\mathbf{\Psi}^T\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{\Psi}-\mathbf{H}\|_F^2. \tag{13}$$

The reason for introducing μ is to emphasize that we can use the proposed method to find a matrix with a coherence close to any arbitrary μ and not necessarily μ_E. In the following subsections, we propose an alternating minimization algorithm, which iteratively minimizes (13) to find the desired $\mathbf{\Phi}$.[1] The proposed algorithm attempts to force an upper limit on off-diagonal elements of $\mathbf{\Psi}^T\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{\Psi}$, and find the desired $\mathbf{\Phi}$, accordingly.

2.2.1. Updating $\mathbf{\Phi}$: a gradient descent approach

Starting with a $\mathbf{\Phi}$, constructed from random i.i.d (independent and identically distributed) columns, we apply a gradient descent method to minimize (13), while keeping $\mathbf{H}$ fixed. We denote the ij-th element of $\mathbf{\Phi}$ by ϕ_{ij} and define the cost function as

$$\mathcal{J} = \|\mathbf{\Psi}^T\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{\Psi}-\mathbf{H}\|_F^2. \tag{14}$$

The minimization method can be described as an iterative process such that $\phi_{ij}\leftarrow\phi_{ij}-\eta\partial\mathcal{J}/(\partial\phi_{ij})$, where $\eta>0$ is the iteration stepsize. In order to apply the proposed method to all elements of matrix $\mathbf{\Phi}$, we compute the gradient of $\mathcal{J}$ with respect to $\mathbf{\Phi}$ (denoted as $\nabla_{\mathbf{\Phi}}\mathcal{J}=\partial\mathcal{J}/\partial\mathbf{\Phi}$), where $\mathbf{H}$ is considered fixed:

$$\frac{\partial\mathcal{J}}{\partial\mathbf{\Phi}} = \frac{\partial}{\partial\mathbf{\Phi}}\mathrm{Tr}\{(\mathbf{\Psi}^T\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{\Psi}-\mathbf{H})^T(\mathbf{\Psi}^T\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{\Psi}-\mathbf{H})\}. \tag{15}$$

Here, $\mathrm{Tr}\{\cdot\}$ denotes the matrix trace operation. Considering the matrix derivative rules in matrix computation [25] we simplify (15) to

$$\frac{\partial\mathcal{J}}{\partial\mathbf{\Phi}} = 4\mathbf{\Phi}\mathbf{\Psi}(\mathbf{\Psi}^T\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{\Psi}-\mathbf{H})\mathbf{\Psi}^T. \tag{16}$$

The full update equation would then be

$$\mathbf{\Phi}_{(k+1)} = \mathbf{\Phi}_{(k)}-\beta\mathbf{\Phi}_{(k)}\mathbf{\Psi}(\mathbf{\Psi}^T\mathbf{\Phi}_{(k)}^T\mathbf{\Phi}_{(k)}\mathbf{\Psi}-\mathbf{H})\mathbf{\Psi}^T, \tag{17}$$

where k is the iteration index and $\beta=4\eta$. The scalar stepsize β can either be a fixed value or be adaptively updated during the iterations. We will discuss about such adaptive stepsize shortly. At k-th iteration and after updating the measurement matrix using (17), we keep $\mathbf{\Phi}$ fixed and update $\mathbf{H}$ to minimize (13). We next describe the procedure for updating $\mathbf{H}$.

2.2.2. Updating $\mathbf{H}$: a component-wise approach

In order to have no high-coherent element between $\mathbf{\Phi}$ and $\mathbf{\Psi}$, an upper limit on off-diagonal elements of $\mathbf{\Psi}^T\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{\Psi}$ is imposed. This is done while updating $\mathbf{H}$, based on a given threshold μ. Minimization of (13) with respect to $\mathbf{H}$, which is a Hermitian matrix, is known as matrix nearness problem and is solved element-wise [17,18]. In fact, the closest $\mathbf{H}$ to $\mathbf{\Psi}^T\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{\Psi}$, in terms of Frobenius norm, has similar sign pattern. On the other hand, it is known that the nearest element of a set to a point (in terms of Frobenius norm) can be obtained by projecting that point onto the set, which is a unique solution [24]. Here also $\mathcal{H}_\mu$ is a convex set and therefore, the solution is straightforward which can be simply described as a matrix having unit diagonal elements and off-diagonals satisfying

$$\forall i,j \quad i\neq j : h_{ij} = \begin{cases} g_{ij} & \text{if } |g_{ij}|\leq\mu, \\ \mu\cdot\mathrm{sign}(g_{ij}) & \text{otherwise,} \end{cases} \tag{18}$$

where g_{ij} is the ij-th element of $\mathbf{G}=\mathbf{\Psi}^T\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{\Psi}$ and μ is the desired mutual coherence defined by the user. Such an element-wise scheme in updating $\mathbf{H}$ assures that in the next update of $\mathbf{\Phi}$, the elements with higher coherence will be intensively constrained. The pseudocode of the algorithm appears in Algorithm 1.

Algorithm 1. The proposed optimization method.

Input: Sparse representation matrix $\mathbf{\Psi}$, stepsize β, threshold μ, and number of iterations: L, K.
Output: Measurement matrix $\mathbf{\Phi}$.

```
begin
  Initialize Φ to a random matrix.
  Initialize H to an identity matrix.
  for l = 1 to L do
    G ← Ψ^T Φ^T Φ Ψ
    Update H:
    for all i ≠ j do
      h_ij = { g_ij            if |g_ij| ≤ μ
             { μ · sign(g_ij)  otherwise
    end
    Update Φ:
    for k = 1 to K do
      Φ ← Φ − βΦΨ(Ψ^T Φ^T Φ Ψ − H)Ψ^T
    end
  end
end
```

We also note that having unit diagonal elements in $\mathbf{H}$ can be thought as a normalization constraint on the columns of $\mathbf{D}=\mathbf{\Phi}\mathbf{\Psi}$, when solving (13). Therefore, normalizing the columns of $\mathbf{D}$ in Algorithm 1 is not applied. This has also the advantage of less computation cost. However, in order to be consistent in representing the results, we always compute the coherence based on the column-normalized $\mathbf{D}$, denoted by $\tilde{\mathbf{D}}$, and accordingly $\tilde{\mathbf{G}}=\tilde{\mathbf{D}}^T\tilde{\mathbf{D}}$, which is also seen in (4) and (5).

The proposed algorithm is repeated for a number of iterations, where $\mathbf{\Phi}$ and $\mathbf{H}$ are alternately updated to reduce (13). Regarding the convergence of the proposed method we mention that constructing ETFs using alternating minimization methods have a global minimum subject to an appropriate initialization [17,18]. It has been proven in [17,18] that there exist at least one accumulation point for $\min_{\mathbf{G},\mathbf{H}}\|\mathbf{G}-\mathbf{H}\|_F^2$, and the minimization leads to a global minimum after infinite number of iterations (more discussion on convergence of alternating minimization methods can be found in [17,18]). The proposed method, however, offers a gradient descent method to update the measurement matrix which guarantees gradual reduction of (13) and convergence to a *local* minimum. Regarding the stopping criterion, we used a fixed number of iterations. However, other stopping criteria can also be used. For instance, one may continue

[1] A partially similar and independent work in parametric dictionary design has also been reported in [24].

the algorithm until reaching to a desired coherence, or until the value of cost function for two successive iterations does not change significantly.

Next, we describe a procedure for choosing an adaptive stepsize which improves the convergence behavior of the proposed method compared with using a fixed stepsize.

3. Adaptive stepsize selection

Although we mentioned that a fixed stepsize can be used in the proposed algorithm, it may not be an optimum choice. Hence, we describe the procedure for selecting an adaptive stepsize to achieve a higher performance and faster initial adaptation. An adaptive stepsize varies during the iterations of the algorithm based on some criteria. *Stochastic gradient* stepsize has been shown to be advantageous over the fixed one [26–28]. In order to find such stepsize at each iteration of the algorithm we can use the steepest descent method as follows:

$$\beta_{(k+1)} = \beta_{(k)} - \xi \frac{\partial \mathcal{J}(\mathbf{\Phi}_{(k+1)})}{\partial \beta_{(k)}}, \tag{19}$$

where ξ is a constant. It is seen from (19) that the stepsize is adapted based on the gradient of the total cost function and the stepsize of the previous iteration. In other words, it is adapted in a way to minimize (14). In order to calculate the differentiation in (19) we replace (17) into (14) and obtain the following equation:

$$\mathcal{J}(\mathbf{\Phi}_{(k+1)}) = \|\mathbf{\Psi}^T(\mathbf{\Phi}_{(k)} - \beta_{(k)}\mathbf{\Phi}_{(k)}\mathbf{\Psi}(\mathbf{\Psi}^T\mathbf{\Phi}_{(k)}^T\mathbf{\Phi}_{(k)}\mathbf{\Psi} - \mathbf{H}_{(k)})\mathbf{\Psi}^T)^T$$
$$(\mathbf{\Phi}_{(k)} - \beta_{(k)}\mathbf{\Phi}_{(k)}\mathbf{\Psi}(\mathbf{\Psi}^T\mathbf{\Phi}_{(k)}^T\mathbf{\Phi}_{(k)}\mathbf{\Psi} - \mathbf{H}_{(k)})\mathbf{\Psi}^T)\mathbf{\Psi} - \mathbf{H}_{(k+1)}\|_F^2. \tag{20}$$

For notational simplicity and avoiding very long equations we define

$$\mathbf{M}_{(k)} \triangleq \mathbf{\Phi}_{(k)}\mathbf{\Psi}(\mathbf{\Psi}^T\mathbf{\Phi}_{(k)}^T\mathbf{\Phi}_{(k)}\mathbf{\Psi} - \mathbf{H}_{(k)})\mathbf{\Psi}^T \tag{21}$$

and then simplify (20) to

$$\mathcal{J}(\mathbf{\Phi}_{(k+1)}) = \|\mathbf{\Psi}^T(\mathbf{\Phi}_{(k)} - \beta_{(k)}\mathbf{M}_{(k)})^T(\mathbf{\Phi}_{(k)} - \beta_{(k)}\mathbf{M}_{(k)})\mathbf{\Psi} - \mathbf{H}_{(k+1)}\|_F^2. \tag{22}$$

In order to calculate the derivative of (22) with respect to $\beta_{(k)}$ we expand $\|\cdot\|_F^2$ using the trace operator and obtain:

$$\begin{aligned}\frac{\partial \mathcal{J}(\mathbf{\Phi}_{(k+1)})}{\partial \beta_{(k)}} &= -2\mathrm{Tr}\{(\mathbf{A}_{(k)} - \mathbf{H}_{(k+1)})^T(\mathbf{B}_{(k)} + \mathbf{B}_{(k)}^T)\} \\ &\quad + 2\,\mathrm{Tr}\{2\mathbf{C}_{(k)}(\mathbf{A}_{(k)} - \mathbf{H}_{(k+1)}) + (\mathbf{B}_{(k)} + \mathbf{B}_{(k)}^T)^2\}\beta_{(k)} \\ &\quad - 6\,\mathrm{Tr}\{\mathbf{C}_{(k)}(\mathbf{B}_{(k)} + \mathbf{B}_{(k)}^T)\}\beta_{(k)}^2 + 4\,\mathrm{Tr}\{\mathbf{C}_{(k)}^2\}\beta_{(k)}^3,\end{aligned} \tag{23}$$

where

$$\mathbf{A}_{(k)} \triangleq \mathbf{\Psi}^T\mathbf{\Phi}_{(k)}^T\mathbf{\Phi}_{(k)}\mathbf{\Psi},$$

$$\mathbf{B}_{(k)} \triangleq \mathbf{\Psi}^T\mathbf{\Phi}_{(k)}^T\mathbf{M}_{(k)}\mathbf{\Psi},$$

$$\mathbf{C}_{(k)} \triangleq \mathbf{\Psi}^T\mathbf{M}_{(k)}^T\mathbf{M}_{(k)}\mathbf{\Psi}. \tag{24}$$

Now, the stepsize value for the $(k+1)$-th iteration can be found from (19). This adaptive update of stepsize should be executed at each iteration of the inner loop in Algorithm 1. We still need to manually choose ξ (called stepsize of stepsize). However, we observed that the overall performance is relatively insensitive to its value; the fact that has been also indicated in the corresponding literature [28].

4. Results

In this section we examine the performance of the proposed method and compare it with well-established similar methods by presenting the empirical results obtained from an extensive set of experiments. In the first experiment, we applied the proposed algorithm to a random $\mathbf{\Phi}_{(30\times200)}$ taken from a Gaussian distribution where a random dictionary $\mathbf{\Psi}$ of size 200×400 was used. Then, Elad's algorithm [10], accessible via SparseLab toolbox [29], was applied to the same matrices. The distributions of absolute off-diagonal elements in $\tilde{\mathbf{G}}$ for the proposed method with $\mu=0.2$, $L=100$, $K=50$ and a fixed $\beta=0.01$, and for Elad's [10] with $t=0.2$, and 1000 iterations, for three different γ: $\gamma_1=0.5$, $\gamma_2=0.7$, $\gamma_3=0.9$ are depicted in Fig. 1(a) and (b). We also added the distribution graph of our previous work in [23] for comparison. It is seen in Fig. 1(a) and (b) that the proposed method has a better response in decreasing both μ_{av} and μ_{mx}. Also, Fig. 1(b) shows that the proposed method is superior to [23] in reducing μ_{mx}, but the method in [23] is more powerful in pulling the center of distribution toward zero (equivalent to reducing μ_{av}). Further, we repeated this experiment with the same settings but when an overcomplete DCT dictionary was used for $\mathbf{\Psi}$. The results are shown in Fig. 1(c). The results in Fig. 1(b) and (c) indicate that the proposed method performs almost similarly for both random and DCT dictionaries. This implies that the performance of the proposed method does not highly depend on the type of dictionary used, which indicates universality of the obtained measurement matrix. The numerical values of μ_{mx} and μ_{av} obtained from this experiment are given in Table 1, for better illustration.

In order to observe the convergence behavior of the proposed method and also evaluate the influence of using an adaptive stepsize we calculated the objective function (14) and also evolution of μ_{mx}, while the iterations were proceeding. In this experiment both $\mathbf{\Phi}$ and $\mathbf{\Psi}$ were generated randomly and other parameters were as follows: $p=50$, $n=80$, $m=100$, $\mu=\mu_E=0.082$, $\xi=10^{-5}$, and $\beta_{(0)}=0.05$. We also applied the methods in [10,23] to this data. The resulting graphs appear in Fig. 2. It is seen from Fig. 2(a) that the proposed method (for both cases of fixed and adaptive stepsize) achieves a lower mutual coherence compared with other methods. Moreover, the achieved mutual coherence when using adaptive stepsize is even smaller. Similar superiority is observed by inspecting Fig. 2(b) where the objective function error is depicted.[2] It is seen that an adaptive stepsize leads to a smaller steady state error and a faster convergence. This behavior is due to the adaptation of stepsize β and reach for an optimum value which is clearly seen from the evolution graph in Fig. 2(c).

[2] Note that since there exist no explicit objective function in [10], we were unable to plot such a graph for it in Fig. 2(b).

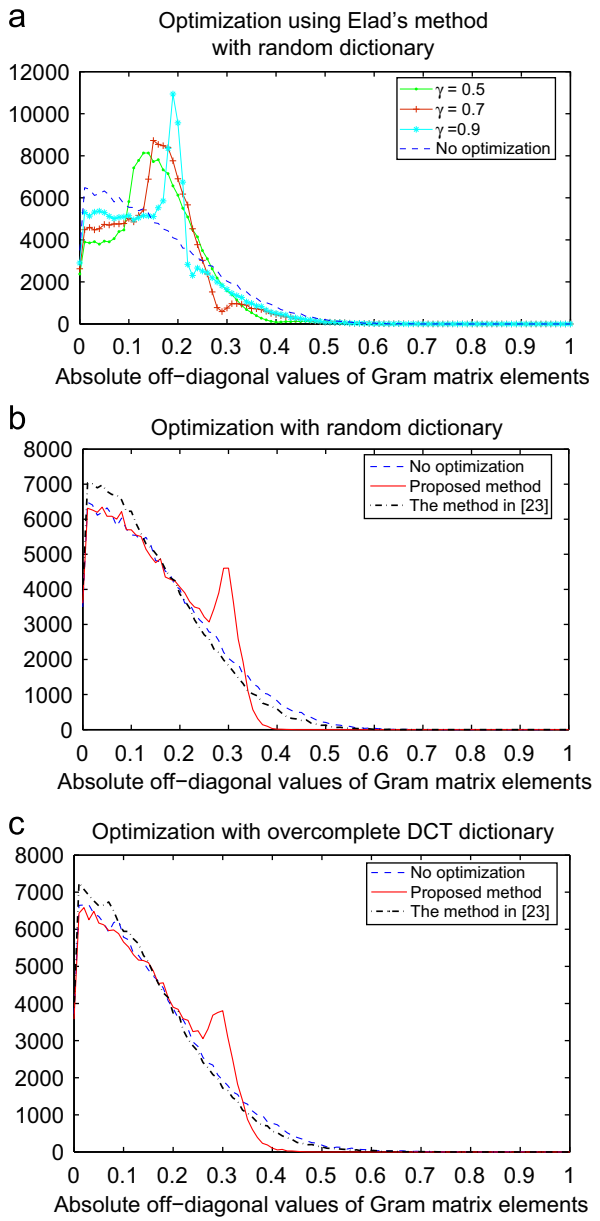

Fig. 1. Comparing the distributions of off-diagonal elements of $\tilde{\mathbf{G}}$: (a), (b) using random dictionary, and (c) with overcomplete DCT dictionary.

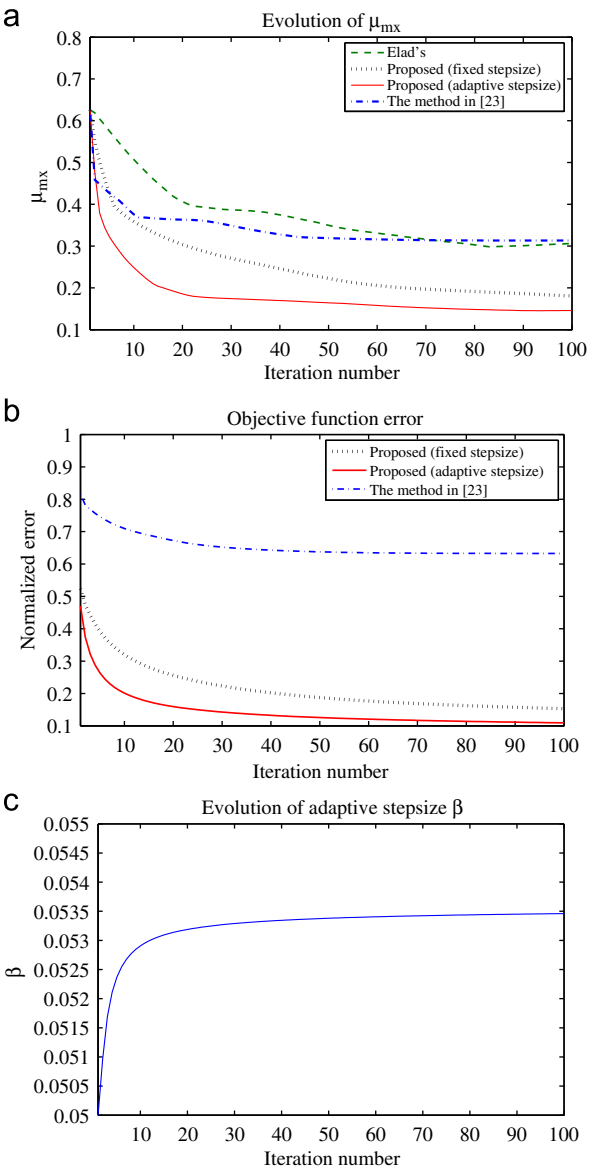

Fig. 2. The convergence results: (a) evolution of μ_{mx}, (b) the objective function, and (c) the evolution of adaptive stepsize β, all versus iterations number.

Table 1
Effects of different methods on mutual coherence when using two types of dictionaries: random and DCT.

Dictionary type	Before optimization	Method in [23]	Proposed method	Elad's method		
				γ_1	γ_2	γ_3
Random dictionary:						
μ_{mx}	0.7326	0.6803	0.4298	0.7831	0.8552	0.9281
μ_{av}	0.1557	0.1436	0.1475	0.1554	0.1548	0.1556
DCT dictionary:						
μ_{mx}	0.9334	0.7429	0.4855	0.9593	0.9146	0.8932
μ_{av}	0.1527	0.1430	0.1471	0.1551	0.1538	0.1552

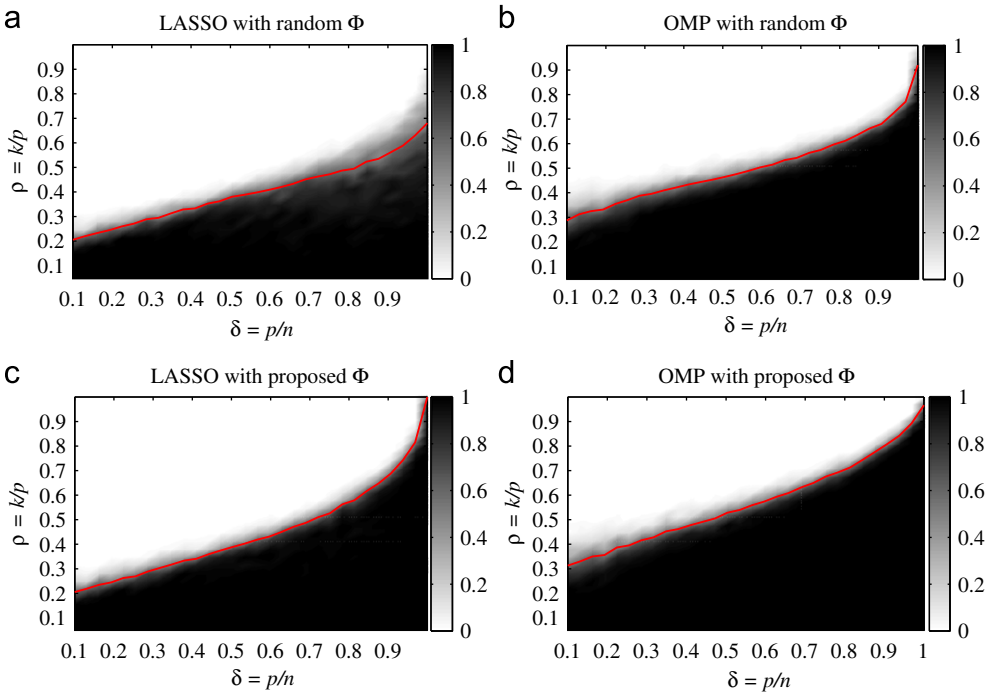

Fig. 3. Empirical phase transition in the (ρ,δ) plane, when (a) LASSO and (b) OMP with random $\mathbf{\Phi}$, (c) LASSO and (d) OMP with proposed method was used.

In the next part of the experiments we investigated the advantages of measurement matrix optimization in the sparse signal recovery. We selected OMP, a greedy algorithm, which iteratively builds up an approximation of the sparse signal and is widely used as a reconstruction algorithm in this framework. Further, LASSO (a continuation of BP [4]) was selected as a minimization approach which attempts to fit a regression model while imposing ℓ_1-norm constraint on the regression coefficients and solving:

$$\min_{\boldsymbol{\alpha}} \|\mathbf{y}-\mathbf{D}\boldsymbol{\alpha}\|_2^2 \quad \text{s.t. } \|\boldsymbol{\alpha}\|_1 \le \kappa, \tag{25}$$

where κ is a nonnegative constant [7,30]. We present the results of applying these methods as they are well known in sparse recovery problems. Extensive simulations were established to sketch the *phase transition* diagram[3] in a noiseless setting. One hundred trials were generated at each point on a grid of $\delta = p/n$ and $\rho = s/p$ axes, for $n = m = 500$, $\delta \in [0.1\ 1]$, and $\rho \in [0.05\ 1]$. The (ρ,δ) plane is made of 30×30 equispaced points in total. $\mathbf{\Psi}$ was chosen to be a random matrix of size 500×500. Non-zeros of the sparse signals in each trial were drawn from random Gaussian distribution. The resulting diagrams are demonstrated in Fig. 3 for two different recovery algorithms, namely OMP [5] and LASSO [7].[4] Shaded attributes are the proportion of successful outcomes to all trials. An outcome is considered to be successful if the reconstruction error $\|\mathbf{x}-\hat{\mathbf{x}}\|_2/\|\mathbf{x}\|_2$ is less than 0.01. The overlaid curves show the estimate of $\boldsymbol{\alpha}$ for each n where the reconstruction is successful with probability $1-\gamma$. γ is the estimator's threshold and was set to $\gamma = 0.25$ (see [29] for more details about the estimator's threshold). Preliminary inspection of the results, especially those in Fig. 4, where the curves for different methods are depicted on the same graph, reveals the success of the proposed approach. It is seen that the proposed approach does give improvements over random matrices, and Elad's method. However, it has been observed that these advantages decrease as the problems size grows (as n increases from around 800–1000 and beyond). Nevertheless, the proposed method outperforms both Elad's and pure random matrices at large dimensions, though not very noticeably.

Following the previous test in analysis of the reconstruction performance, we set up a new experiment where a redundant DCT dictionary $\mathbf{\Psi}_{(80\times120)}$ was used. We conducted two separate CS experiments, first by fixing $p=25$ and varying s from 1 to 10 and second by fixing $s=4$ and varying p from 15 to 40 (see also [10,29]). Each experiment was performed for 50,000 random sparse ensembles and the average reconstruction error was recorded. The corresponding graphs shown in Fig. 5 indicate the improved performance when using an optimized $\mathbf{\Phi}$ instead of a random $\mathbf{\Phi}$. Moreover, it is seen that the proposed method performs better, especially when OMP is used as the recovery method. The reason can be inferred from the fact that OMP works based on orthogonality, which is improved when the proposed method is applied.

More importantly, Fig. 5(a) reveals that by using the optimized measurement matrices one can achieve the same performance as that obtained using pure random matrices, but with fewer number of measurements.

[3] Details about transition diagram can be found in [29,30].

[4] Other methods such as BP [4] and LARS [6] were also tested and similar behaviors were observed.

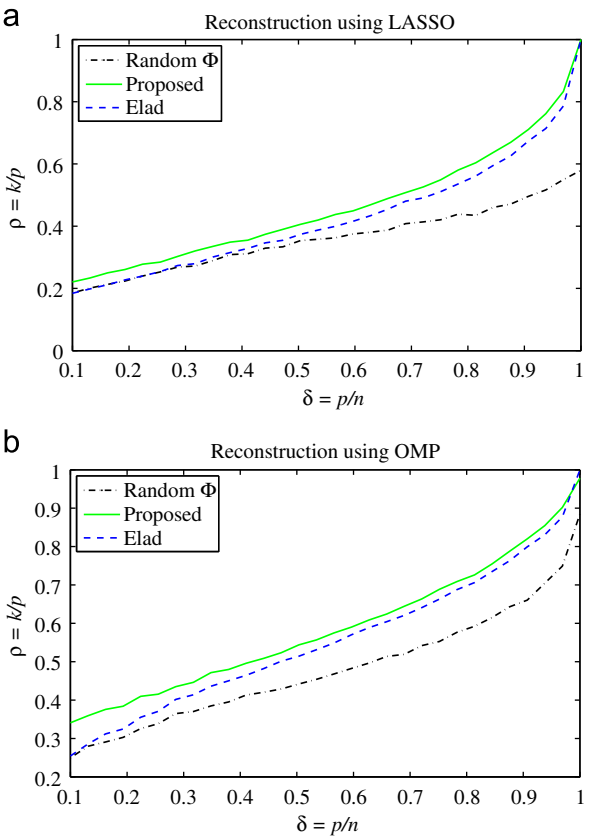

Fig. 4. (a) Empirical phase transition curves of (a) LASSO and (b) OMP when different measurement matrices were used.

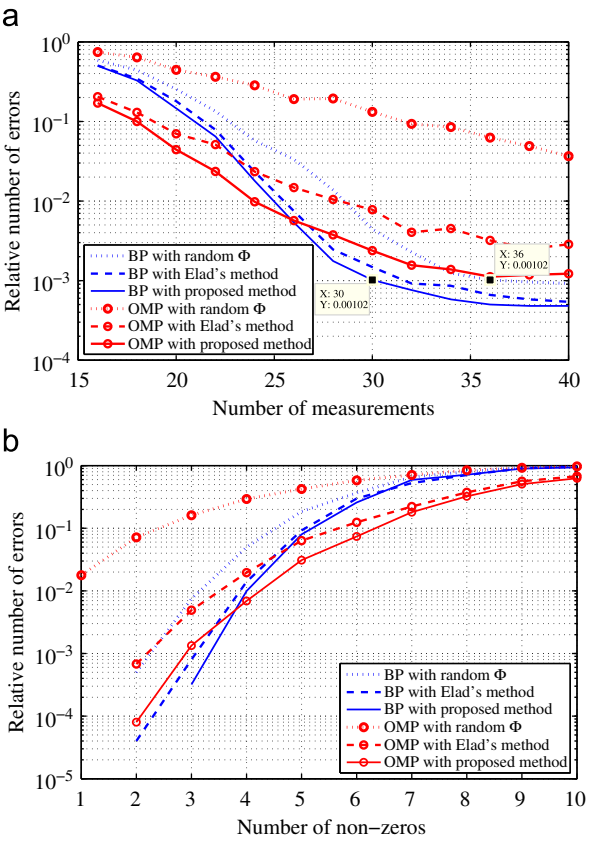

Fig. 5. Relative number of errors as a function of (a) measurement numbers p and (b) number of non-zeros.

For instance, notice the recovery error of BP using the optimized measurement matrix at $p=30$ in Fig. 5(a). It is almost as small as the recovery rate at $p=36$ for a pure random $\mathbf{\Phi}$. The same behavior can also be noticed for the recovery using OMP.

In this part of the simulations the possibility of using the proposed approach for optimizing the Bernoulli projections are studied. The main advantages of using random Bernoulli matrices are reduction of the hardware complexity and storage due to their binary nature. Assuming the same settings in the previous experiment we first generated a measurement matrix with random Bernoulli distribution. Then, the proposed method was applied to optimize it. The optimized matrix would not obviously have ± 1 values and includes values with any amplitude. One preliminary way to quantize the optimized elements back to the Bernoulli type, i.e. ± 1, is to apply the nonlinear sign$(\cdot)$ operation. The recovery error rates using this scheme are given in Fig. 6, which indicates its superiority over pure random Bernoulli matrices. However, applying such nonlinear operation is not optimum and further work is required to better optimize random Bernoulli matrices.

As previously mentioned, the parameter μ in Algorithm 1 can be theoretically chosen as $\mu \geq \mu_E$. However, we empirically observed that the best performance is achieved when $\mu \simeq \mu_E$. This is a reasonable choice since μ_E is the optimum

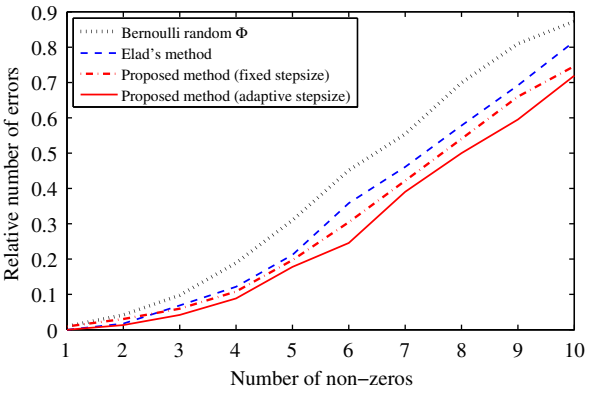

Fig. 6. Relative reconstruction error using OMP versus number of non-zeros.

mutual coherence and the aim is to approach this value. In order to support the effectiveness of this choice and demonstrate the influence of choosing different μ we measured the reconstruction error of 10,000 signal ensembles while optimizing $\mathbf{\Phi}$ with different values of μ. In this experiment, the signal ensembles were of length $m=100$ with five non-zeros, $\mathbf{\Psi}$ was an overcomplete DCT dictionary, and OMP was used as the recovery method. Other parameters of the algorithm were $p=30$, $n=64$, $L=100$ and $K=20$. Based on

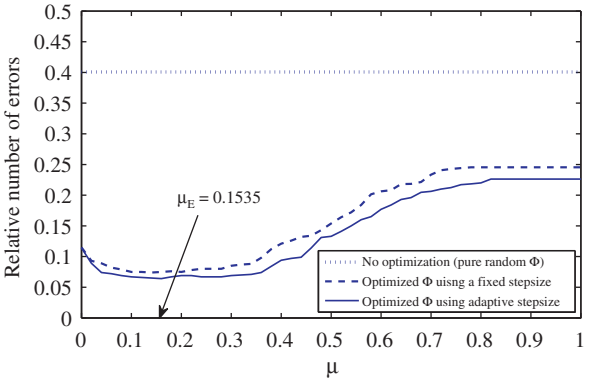

Fig. 7. Relative reconstruction error versus μ. The parameter μ is used in Algorithm 1 for optimization of the measurement matrix.

the given parameters, the optimum mutual coherence can be calculated as $\mu_E=\sqrt{(m-d)/d(m-1)}=0.1535$. Fig. 7 represents the relative reconstruction error as a function of μ for three cases of no optimization, optimization with a fixed stepsize $\beta=0.01$ and optimization with adaptive stepsize of $\xi=10^{-6}$ and $\beta_{(0)}=0.02$. It is seen from the graphs in Fig. 7 that the minimum error is achieved when $\mu\simeq 0.1\sim 0.3$, and hence, $\mu\simeq\mu_E$ is a suitable choice. It is also noticeable that for $\mu<\mu_E$ still the error is relatively small but it increases as it approaches to $\mu=0$. It is worth to note that the choice of $\mu=0$ is equivalent to using the method in [23] which means replacing $\mathbf{H}$ with the identity matrix $\mathbf{I}$. We have also observed that as μ decreases, the steady state error of objective function (14) becomes higher and vice versa.

All the aforementioned experiments were conducted under the noiseless settings. In order to show the robustness of the proposed method in noisy situations we considered the noisy model $\mathbf{y}=\mathbf{\Phi\Psi\alpha}+\mathbf{v}$, where $\mathbf{v}\in\mathbb{R}^p$ is the vector of additive Gaussian noise with zero mean. The main parameter settings for this simulation were: $p=25$, $n=64$, $m=100$, number of non-zeros $s=5$, and $\mu=0.17$. First, the proposed method was applied to a random measurement matrix, where a DCT dictionary was used. Then, the noisy measurements were generated with different signal to noise ratios (SNRs) from 0 to 60 dB. We applied OMP and LASSO to these noisy measurements to recover the sparse signals. The relative reconstruction error under different noise levels are given in Fig. 8. It is seen from Fig. 8 that the reconstruction performance improves as the SNR increases and vice versa. However, the optimized measurement matrix is more robust to noise compared with the pure random matrix.

Finally, we compare the computation time of the proposed method with other methods. We set up a simulation to optimize measurement matrices with different dimensions. A desktop computer with a Core 2 Duo CPU of 3.00 GHz, and 2 GB of RAM was used for this experiment. Table 2 represents the running time per iteration with respect to different dimensions for $\mathbf{\Phi}$ and $\mathbf{\Psi}$. As it is seen from the table, the computation time increases with increasing the dimensions. The methods in [23,10] present lowest and highest computation times, respectively. For large dimensions the computation of the Elad's method

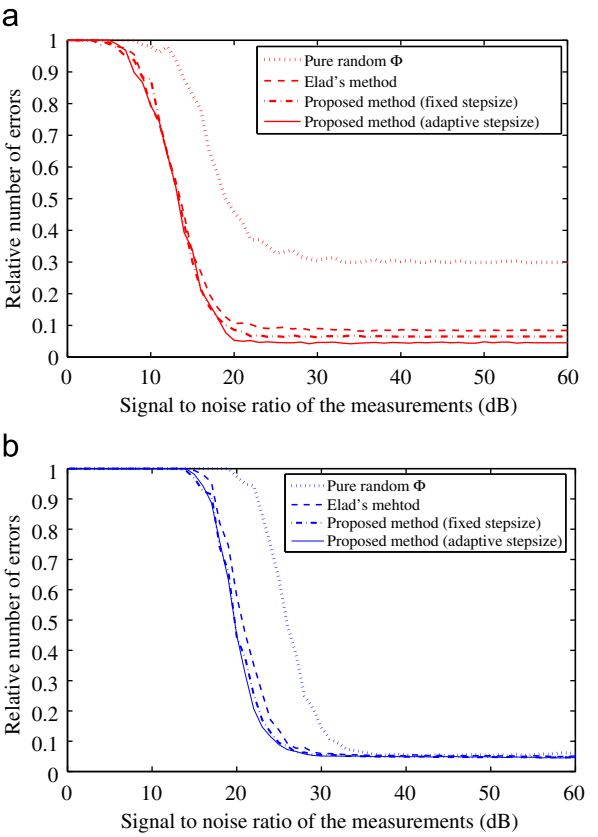

Fig. 8. Relative reconstruction error versus the SNR of the measurements in dB using (a) OMP and (b) LASSO.

[10] is almost twice as that for the proposed method with fixed stepsize (notice the last column in Table 2). That can be because of using SVD in Elad's method which is computationally expensive, especially for high dimensions. In addition, it is seen that using adaptive stepsize for the proposed method incurs more complexity and hence increases the computation time. In general, high computation time in very-large scale problems is a drawback of such optimization techniques.

5. Discussion

In this short section we discuss the characteristics of the proposed method with respect to a given class of signals or images. It might have been realized that the proposed method is non-adaptive and only requires the knowledge about sparsifying dictionary. In contrast to some methods like [14], it does not exploit any prior knowledge (e.g. the frequency spectrum, location of non-zeros) about the signal of interest into the optimization process. Due to such independency from the original signal, the optimized measurement matrix is not affected by the properties (whether known or unknown) of the sparse signal. In particular, the power spectrum density of the optimized measurement matrix, which is originally white for a pure random matrix, does not change significantly and remains white, even if the signal to be

Table 2
The computation time (in seconds) per iteration for different methods.

Optimization technique	Matrix dimensions				
	Φ: 60×100	120×200	180×300	240×400	300×500
	Ψ: 100×200	200×400	300×600	400×800	500×1000
Elad's method [10]	0.0772	0.4533	1.2673	2.8761	5.2674
The method in [23]	0.0079	0.0591	0.2453	0.6048	1.1691
The proposed method (fixed stepsize)	0.0565	0.2562	0.7069	1.4385	2.5095
The proposed method (adaptive stepsize)	0.0673	0.3526	0.9854	2.0480	3.6545

sampled has a particular spectrum. We have observed such behavior in other non-adaptive methods too, such as [10,23]. This non-adaptivity has the advantage that the proposed method can be applied to a wide range of signals and images without any restriction. On the other hand, the disadvantage is that it is not able to incorporate the prior knowledge into the optimization process in cases where such *a priori* is available. However, in the case of available knowledge about spectrum of the sparse signal (e.g. the signal is bandpass or highpass), one reasonable approach can be to first generate a colored random measurement matrix using the method described in [14], then apply the proposed method in this paper to optimize it, and decrease the mutual coherence. This is the context that has to be further studied in the future.

6. Conclusion

In this paper optimization of the measurement matrix in compressed sensing was addressed. Although random sampling has been widely used so far in this framework, it has been recently shown that optimized projections with the property of smaller mutual coherence with respect to sparsifying matrix can significantly improve the performance. We proposed a gradient-based alternating optimization method which iteratively optimizes the measurement matrix to decrease the mutual coherence. The advantages are higher performance in sparse recovery, and also possibility to take fewer measurements and yet achieve the same performance level when using pure random matrices. We also described a steepest descent method for obtaining an adaptive stepsize to improve the performance. The presented numerical results verify the success of the proposed approach. In addition, the experimental results indicate that the proposed method is computationally less expensive than some current methods in the literature.

Acknowledgments

The authors would like to thank the anonymous reviewers for their thorough remarks and fruitful suggestions.

References

[1] D. Donoho, Compressed sensing, IEEE Transactions on Information Theory 52 (4) (2006) 1289–1306.
[2] E.J. Candes, J. Romberg, T. Tao, Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information, IEEE Transactions on Information Theory 52 (2) (2006) 489–509.
[3] C.E. Shannon, Communication in the presence of noise, Proceedings of the IRE 37 (1) (1949) 10–21.
[4] S. Chen, D.L. Donoho, M.A. Saunders, Atomic decomposition by basis pursuit, SIAM Journal on Scientific Computing 20 (1) (1999) 33–61.
[5] J.A. Tropp, A.C. Gilbert, Signal recovery from random measurements via orthogonal matching pursuit, IEEE Transactions on Information Theory 53 (12) (2007) 4655–4666.
[6] B. Efron, T. Hastie, L. Johnstone, R. Tibshirani, Least angle regression, Annals of Statistics 32 (2004) 407–499.
[7] R. Tibshirani, Regression shrinkage and selection via the LASSO, Journal of the Royal Statistical Society (Series B) 58 (1) (1996) 267–288.
[8] E.J. Candès, M.B. Wakin, An introduction to compressive sampling, IEEE Signal Processing Magazine 25 (2) (2008) 21–30.
[9] E.J. Candès, J.K. Romberg, T. Tao, Stable signal recovery from incomplete and inaccurate measurements, Communications on Pure and Applied Mathematics 59 (8) (2006) 1207–1223.
[10] M. Elad, Optimized projections for compressed sensing, IEEE Transactions on Signal Processing 55 (12) (2007) 5695–5702.
[11] V. Abolghasemi, S. Sanei, S. Ferdowsi, F. Ghaderi, A. Belcher, Segmented compressive sensing, in: IEEE/SP 15th Workshop on Statistical Signal Processing SSP 2009, September 2009.
[12] M. Lustig, D. Donoho, J.M. Pauly, Sparse MRI: the application of compressed sensing for rapid MR imaging, Magnetic Resonance in Medicine 58 (6) (2007) 1182–1195.
[13] Z. Wang, G.R. Arce, Variable density compressed image sampling, IEEE Transactions on Image Processing 19 (1) (2010) 264–270.
[14] Z. Wang, G.R. Arce, J.L. Paredes, Colored random projections for compressed sensing, IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2007, vol. 3, April 2007, pp. III-873–III-876.
[15] J.M. Duarte-Carvajalino, G. Sapiro, Learning to sense sparse signals: simultaneous sensing matrix and sparsifying dictionary optimization, IEEE Transactions on Image Processing 18 (7) (2009) 1395–1408.
[16] M. Aharon, M. Elad, A. Bruckstein, K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation, IEEE Transactions on Signal Processing 54 (11) (2006) 4311–4322.
[17] I.S. Dhillon, R.W. Heath, T. Strohmer, J.A. Tropp, Constructing packings in Grassmannian manifolds via alternating projection, Experimental Mathematics 17 (1) (2008) 9–35.
[18] J.A. Tropp, I.S. Dhillon, R.W. Heath, T. Strohmer, Designing structured tight frames via an alternating projection method, IEEE Transactions on Information Theory 51 (1) (2005) 188–209.
[19] T. Strohmer, R.W. Heath, Grassmannian frames with applications to coding and communication, Applied and Computational Harmonic Analysis 14 (3) (2003) 257–275.
[20] D.L. Donoho, P.B. Stark, Uncertainty principles and signal recovery, SIAM Journal on Applied Mathematics 49 (3) (1989) 906–931.
[21] D.L. Donoho, M. Elad, Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1-minimization, Proceedings of the National Academy of Sciences 100 (5) (2003) 2197–2202.
[22] J.A. Tropp, Greed is good: algorithmic results for sparse approximation, IEEE Transactions Information Theory 50 (2004) 2231–2242.
[23] V. Abolghasemi, S. Ferdowsi, B. Makkiabadi, S. Sanei, On optimization of the measurement matrix for compressive sensing, in: European Signal Processing Conference EUSIPCO, Aalborg, Denmark, August 2010.

[24] M. Yaghoobi, L. Daudet, M.E. Davies, Parametric dictionary design for sparse coding, IEEE Transactions on Signal Processing 57 (12) (2009) 4800–4810.
[25] K.B. Petersen, M.S. Pedersen, The Matrix Cookbook, Version 20081110, October 2008.
[26] S. Amari, A theory of adaptive pattern classifiers, IEEE Transactions on Electronic Computers EC EC-16 (3) (1967) 299–307.
[27] V.J. Mathews, Z. Xie, A stochastic gradient adaptive filter with gradient adaptive step size, IEEE Transactions on Signal Processing 41 (6) (1993) 2075–2087.
[28] S.C. Douglas, Generalized gradient adaptive step sizes for stochastic gradient adaptive filters, in: International Conference on Acoustics, Speech, and Signal Processing, ICASSP 1995, vol. 2, pp. 1396–1399, May 1995.
[29] D.L. Donoho, I. Drori, V. Stodden, Y. Tsaig, Sparselab ⟨http://sparselab.stanford.edu/⟩, December 2005.
[30] D.L. Donoho, Y. Tsaig, Fast solution of ℓ_1-norm minimization problems when the solution may be sparse, IEEE Transactions on Information Theory 54 (11) (2008) 4789–4812.