

Learning to Sense Sparse Signals: Simultaneous Sensing Matrix and Sparsifying Dictionary Optimization

Julio Martin Duarte-Carvajalino, *Member, IEEE*, and Guillermo Sapiro, *Senior Member, IEEE*

Abstract—Sparse signal representation, analysis, and sensing have received a lot of attention in recent years from the signal processing, optimization, and learning communities. On one hand, learning overcomplete dictionaries that facilitate a sparse representation of the data as a linear combination of a few atoms from such dictionary leads to state-of-the-art results in image and video restoration and classification. On the other hand, the framework of compressed sensing (CS) has shown that sparse signals can be recovered from far less samples than those required by the classical Shannon–Nyquist Theorem. The samples used in CS correspond to linear projections obtained by a sensing projection matrix. **It has been shown that, for example, a nonadaptive random sampling matrix satisfies the fundamental theoretical requirements of CS, enjoying the additional benefit of universality. On the other hand, a projection sensing matrix that is optimally designed for a certain class of signals can further improve the reconstruction accuracy or further reduce the necessary number of samples.** In this paper, we introduce a framework for the joint design and optimization, from a set of training images, of the nonparametric dictionary and the sensing matrix. We show that this joint optimization outperforms both the use of random sensing matrices and those matrices that are optimized independently of the learning of the dictionary. Particular cases of the proposed framework include the optimization of the sensing matrix for a given dictionary as well as the optimization of the dictionary for a predefined sensing environment. The presentation of the framework and its efficient numerical optimization is complemented with numerous examples on classical image datasets.

Index Terms—Compressed sensing, image patches, learning, overcomplete dictionary, sensing projection matrix, sparse representation.

I. INTRODUCTION

IMAGE compression algorithms have been successfully employed to transform a high resolution image into a relatively small set of (quantized) coefficients that efficiently represent the image on an appropriate basis such as DCT or wavelets. This representation is designed to preserve the essential content of the image, while at the same time reducing costs

in storage, processing, and transmission. Since natural images can be compressed on an appropriate basis, sampling the scene into millions of pixels to obtain high resolution images that are then to be compressed before processing, seems often to be wasteful [1]–[11]. The main reason why signals in general, and images in particular, have been traditionally sensed using a large number of samples is the Shannon–Nyquist Theorem: the sampling rate must be at least twice the bandwidth of the signal. Images are not naturally band limited, however, acquisition systems use anti-aliasing low pass filters before sampling, hence, the Shannon–Nyquist Theorem plays an implicit role in images and signals in general.

Compressive sensing (CS) is an emerging framework stating that sparse signals, that is, signals that have a concise (sparse) linear representation on an appropriate basis, can be exactly recovered from a number of linear projections of dimension considerably lower than the number of samples required by the Shannon–Nyquist Theorem (in the order of 2–3 times the sparsity of the signal, regardless of the actual signal bandwidth) [1]–[5], [7]. Moreover, signals that are well approximated by sparse representations, denoted as *compressible*, can also be sensed with linear measurements at a much lower rate than double their actual bandwidth, as required by the Shannon–Nyquist Theorem, and then reconstructed with minimum loss of information [1]–[3]. This includes natural images, which are well known to be compressible, e.g., [12]–[21]. This means that instead of sensing an image using millions of pixels to obtain high resolution, the image can be sensed directly in compressed form, sampling a relatively small number of projections that depends on the actual sparsity (and not bandwidth) of the image [1]–[11].¹

Compressive sensing relies on two fundamental principles, e.g., see the recent review [4].

- **Sparsity:** Let $\mathbf{x} \in \mathbb{R}^N$ be an N -pixels image and $\Psi = [\psi_1 \dots \psi_N]$ an orthonormal basis (dictionary),² with elements also in \mathbb{R}^N , such that

¹There might be many reasons for still sampling following the traditional Shannon–Nyquist requirements, even for very sparse signals, e.g., the simplicity of the reconstruction of band limited signals from their samples as well as the existence of very efficient sampling and reconstruction hardware and software. We consider CS and the concepts of it exploited in this paper as an alternative and addition to the classical Shannon–Nyquist framework, viable and very useful for many signal and image processing scenarios, but not at all as a replacement for it.

²While following the basic theory of CS, we consider for the moment an orthonormal basis, the concept of sparsity is much more general, and best applied in image processing when including overcomplete dictionaries, as exploited in this paper.

Manuscript received May 26, 2008; revised November 08, 2008. First published June 02, 2009; current version published June 12, 2009. This work was supported in part by the NSF, ONR, NGA, DARPA, ARO, and NIH. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Arun Ross.

The authors are with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455-0436 USA (e-mail: guille@umn.edu).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TIP.2009.2022459

$$\mathbf{x} = \sum_{i=1}^N \theta_i \boldsymbol{\psi}_i = \boldsymbol{\Psi} \boldsymbol{\theta} \quad (1)$$

where $\boldsymbol{\theta} = [\theta_1 \dots \theta_N]^T$ is the vector of coefficients that represents \mathbf{x} on the basis $\boldsymbol{\Psi}$. A signal or image is said to be sparse if most of the coefficients of $\boldsymbol{\theta}$ are zero or they can be discarded without much loss of “information.” Let \mathbf{x}_S be the image where only the S largest coefficients of $\boldsymbol{\theta}$ are kept and the rest are set to zero (obtaining $\boldsymbol{\theta}_S$), i.e., $\mathbf{x}_S := \boldsymbol{\Psi} \boldsymbol{\theta}_S$. If the coefficients, sorted in decreasing order of magnitude, decrease quickly, then \mathbf{x} is very well approximated by \mathbf{x}_S , when properly selecting both S and the basis/dictionary $\boldsymbol{\Psi}$. Such signal is said to be approximately S -sparse or compressible. Natural images are known to be sparse or compressible, with S significantly lower than the actual image dimension, when represented on an appropriated basis such as wavelets, sinusoids, or a learned (overcomplete) dictionary. Sparse representations form the basis of many successful image processing and analysis algorithms, from JPEG and JPEG2000 compression [22]–[24], to image enhancement and classification, e.g., [12], [25]–[28].

- *Incoherent Sampling*: Let $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1 \dots \boldsymbol{\phi}_m]^T$ be an $m \times N$ sampling matrix, with $m \ll N$, such that $\mathbf{y} = \boldsymbol{\Phi} \mathbf{x}$ is an $m \times 1$ vector of linear measurements (meaning we no longer observe the image \mathbf{x} but an undercomplete linear projection of it). Compressive Sensing theory requires that $\boldsymbol{\Phi}$, the sensing matrix, and $\boldsymbol{\Psi}$, the sparse representation matrix, be as incoherent (orthogonal) as possible. A possible measure of coherence between $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ is given by

$$\mu(\boldsymbol{\Phi}, \boldsymbol{\Psi}) := \sqrt{N} \cdot \max_{\substack{1 \leq i \leq m, \\ 1 \leq j \leq N}} |\langle \boldsymbol{\phi}_i, \boldsymbol{\psi}_j \rangle|. \quad (2)$$

$\mu(\boldsymbol{\Phi}, \boldsymbol{\Psi}) \in [1, \sqrt{N}]$ measures the maximal correlation between both matrix elements (see also [29] and [30] for the related definition of *mutual coherence* of a dictionary, which plays an important role in the success of *basis pursuit* and the greedy sparsifying *orthogonal matching pursuit* algorithm, as well [29]–[34]; see below).

CS deals with the case of low coherence between the sensing and sparsifying matrices. Intuitively, one can see that $\boldsymbol{\Phi}$ and $\boldsymbol{\Psi}$ must be uncorrelated (incoherent), such that the samples add new information that is not already represented by the known basis $\boldsymbol{\Psi}$. It turns out that random matrices, e.g., Gaussians or ± 1 random matrices, are largely incoherent with any fixed sparsifying basis $\boldsymbol{\Psi}$ with overwhelming probability. This has lead CS, at least at the theoretical level as well as early applications in image processing [10], to strongly rely on random sensing matrices, since they provide universally incoherent sensing-sparsifying pairs and are well conditioned for reconstruction.

Compressed sensing combines both concepts of sparsity and incoherence between the sensing and sparsifying matrix by reconstructing the sparsest possible signal that agrees with the undercomplete ($m \ll N$) set of measurements. Let \mathbf{y} be the vector

of m linear measurements of the sparse signal \mathbf{x} obtained using the sampling/sensing matrix $\boldsymbol{\Phi}$. The retrieval of \mathbf{x} from \mathbf{y} can be done by ℓ_0 -“norm” minimization (note how both the sensing and the sparsifying matrix appear in the formulation)

$$\hat{\boldsymbol{\theta}} := \min_{\boldsymbol{\theta}} \|\boldsymbol{\theta}\|_{\ell_0} \text{ subject to } \mathbf{y} = \boldsymbol{\Phi} \boldsymbol{\Psi} \boldsymbol{\theta}, \quad \mathbf{x} = \boldsymbol{\Psi} \hat{\boldsymbol{\theta}} \quad (3)$$

where the ℓ_0 -“norm” simply counts the number of nonzeros in $\boldsymbol{\theta}$.³ Compressive sensing theory ensures (in one of the many recent fundamental results), that if the number of measurements m satisfies

$$m \leq C \cdot \mu^2(\boldsymbol{\Phi}, \boldsymbol{\Psi}) \cdot S \cdot \log N \quad (4)$$

for some positive constant C , then, with overwhelming probability, the reconstruction is exact (actually using ℓ_1 for sparsity promotion instead of ℓ_0 , making the problem (3) convex; only $O(S)$ samples are needed if using the combinatorial complex ℓ_0 minimization in (3)). While the use of the ℓ_0 pseudo-norm guarantees the computation of the sparsest possible \mathbf{x} , the ℓ_1 norm leads to sparse solutions, but in general not necessarily the sparsest one. One of the fundamental results in compressed sensing, see the above mentioned references, showed under what conditions the use of the ℓ_1 norm guarantees (in probability) the computation of the sparsest \mathbf{x} . Optimizing for the ℓ_1 norm is tractable, while minimizing the ℓ_0 accounts in general to the intractable procedure of trying all possible combinations (unless the conditions for the success of greedy algorithms such as othogonal matching pursuit hold, see below).

As mentioned above, related to $\mu(\boldsymbol{\Phi}, \boldsymbol{\Psi})$ is the notion of *mutual coherence* $\mu(\boldsymbol{\Psi})$ of the sparsifying dictionary (or the equivalent dictionary $\mathbf{D} := \boldsymbol{\Phi} \boldsymbol{\Psi}$), which is the largest absolute normalized inner product between the atoms of the dictionary (see (7) in the next section for the exact definition). If the following inequality holds [29], [30], [33]

$$\|\boldsymbol{\theta}\|_{\ell_0} < \frac{1}{2} \left(1 + \frac{1}{\mu(\boldsymbol{\Psi})} \right) \quad (5)$$

then $\boldsymbol{\theta}$ is necessarily the sparsest solution ($\min_{\boldsymbol{\theta}} \|\boldsymbol{\theta}\|_{\ell_0}$) such that $\mathbf{x} = \boldsymbol{\Psi} \boldsymbol{\theta}$ (or $\mathbf{y} = \boldsymbol{\Phi} \boldsymbol{\Psi} \boldsymbol{\theta}$ when considering sensing and the equivalent dictionary \mathbf{D}), and **Orthogonal Matching Pursuit (OMP), a fast greedy algorithm** used to solve sparse representation problems, is guaranteed to succeed in finding the correct $\boldsymbol{\theta}$ (same for the basis pursuit, which again replaces the ℓ_0 by ℓ_1 ⁴), see, e.g., [29], [30], and [32]–[35]. Note of course that this is a property on the dictionary, and in contrast with (4), does not explicitly consider the number of samples m .

Since, in general, signals of interest are not exactly sparse, but compressible, and also contain noise added by the measurement system, it is imperative for CS to be robust under such nonidealities [4]. A key notion in CS theory that comes to the rescue in this scenario is the *Restricted Isometry Property* (RIP) [1]–[4]

³This is not an actual norm, but it is often referred to as such.

⁴The use of the ℓ_1 norm brings interesting connections with robust statistics and regression.

(and references therein, see also [37] for some results for overcomplete dictionaries).⁵ The S -restricted isometry constant is the smallest $0 < \delta_S < 1$ such that

$$\forall T \leq S : (1 - \delta_S) \|\theta_T\|_{\ell_2}^2 \leq \|\mathbf{D}_T \theta_T\|_{\ell_2}^2 \leq (1 + \delta_S) \|\theta_T\|_{\ell_2}^2$$

where \mathbf{D}_T is a subset of T columns extracted from the equivalent dictionary $\mathbf{D} := \Phi\Psi$, and θ_T are the (sparse) coefficients corresponding to the T selected columns. In words, for proper values of δ_S the RIP ensures that any subset, with cardinality less than S , of columns of the equivalent dictionary $\Phi\Psi$ are nearly orthogonal (the columns cannot be exactly orthogonal since we have more columns than rows), i.e., incoherence between Φ and Ψ is ensured. If the RIP holds (see the above mentioned references for the exact needed values of the isometry constant), greedy algorithms such as *regularized* OMP, [38], and ℓ_1 convex optimization are guaranteed to succeed in finding \mathbf{x} (the results for OMP are weaker). This holds as well, again with the possibility to optimize with the ℓ_1 norm instead of ℓ_0 , in the presence of noise and for signals with nonexact sparsity [1]–[8]. Equation (3) becomes in practice

$$\hat{\theta} := \min_{\theta} \|\theta\|_{\ell_0} \quad \text{subject to} \quad \|\mathbf{y} - \Phi\Psi\hat{\theta}\|_{\ell_2} \leq \epsilon \quad (6)$$

where $\epsilon > 0$ takes into account the possibility of noise in the linear measurements and of nonexact sparsity (again, we can use ℓ_1 instead of ℓ_0 in (6), thereby making the optimization problem convex).

While the sampling projection matrix Φ should, in theory, be independent of the signal, the sparsifying basis Ψ should adapt as much as possible to the image at hand, e.g., to make the representation as sparse as possible. A key result in image processing is that images can be coded and sparsely represented more efficiently using learned overcomplete dictionaries rather than fixed bases, e.g., [21], [36], [39]. Let $\mathbf{x}_p \in \mathbb{R}^n$, $n \ll N$, be a patch, i.e., a square portion of the image \mathbf{x} of size $B \times B = n$ pixels. An overcomplete dictionary is an $n \times K$ matrix $\Psi = [\psi_1 \dots \psi_K]$, $n < K$, such that $\mathbf{x}_p = \Psi\theta_p$, and for the K -dimensional vector of coefficients θ_p , we have $\|\theta_p\|_{\ell_0} \ll n$. Such dictionaries are learned from patches extracted from large datasets of images, and thereby are optimized for the data (they can also be optimized to a particular image class, e.g., [16] and [25]). From the learned nonparametric overcomplete dictionaries, state-of-the-art results in image denoising, inpainting, and demosaicing, have been obtained in the past [12]. This learning framework is also exploited in this work.

Overcomplete learned dictionaries are not orthonormal bases, and, hence, the full extent of the CS theory does not entirely apply here (see, for example, [7] and [37] for some recent results). Recently, Elad [40] (see also [41] and [42]), showed experimentally that a well-designed sampling matrix can significantly improve the performance of CS when compared to random matrices, in terms of improving the incoherence for a given dictionary and the reconstruction accuracy from the

corresponding linear samples. This means that for a specific predefined or prelearned overcomplete dictionary, a sampling projection matrix specifically designed for the dictionary can improve CS over a generic random sampling matrix. Note that the sampling is not adaptive, is just optimized for the signal class. Theoretical studies, with practical implications, regarding the construction of deterministic sampling matrices and their RIP, are starting to appear, e.g., see [43], [44], and references therein.

A. Contributions of this Work

In this paper, we introduce a framework for simultaneously learning, from a set of training images, the overcomplete nonparametric dictionary Ψ and the sensing matrix Φ . That is, in contrast with Elad's work and those briefly discussed in the next section, we do not consider a prelearned or predefined dictionary; we learn it together with the sensing matrix. In contrast with the more standard CS framework, we do not assume the sparsifying basis (or dictionary) is given and consider universal sampling strategies, but simultaneously optimize for both these critical components exploiting image datasets. Contrasting also with earlier work on the learning of overcomplete nonparametric dictionaries, we consider linear projections of the image as the available measurements for reconstruction, and not the image itself (or a noisy version of it). We experimentally show that the proposed framework of simultaneous optimization of the sensing matrix and sparsifying dictionary leads to improvements in the image reconstruction results. We also show that for a given dictionary, the learned sensing matrices have larger incoherence, as requested by the RIP, than random matrices and the ones obtained with the algorithm proposed by Elad, on top of leading to lower image reconstruction errors. Computational improvements, when compared with [40], are obtained as well with our proposed framework when considering a given dictionary.

The learned dictionary and sensing matrix do not need to be communicated to the reconstruction side at the rate of one per image (in case of the use of this framework for transmission), since these are learned for image groups or classes. In the case of the examples presented in this paper, these are learned for the large class of natural images from a standard dataset.

While as shown in this paper, the reconstruction quality obtained with our simultaneously learned dictionary and sensing matrix is far superior to that obtained for example with prelearned dictionaries and random sensing, the actual computational time of the reconstruction can be slower for nonparametric dictionaries, as those here obtained, when compared for example to dictionaries such as wavelets or DCT, see Section V for more discussion on this.

The remainder of this paper is organized as follows. In Section II, we review Elad's approach [40] (as well as [41] and other recent related works), and introduce our proposed new algorithm to learn deterministic sensing matrices when the sparsifying dictionary is given. We show that the algorithm is significantly faster than the one in [40] and leads to improved performance in terms of sensing-dictionary incoherence and accuracy of the reconstructed images. In Section III, we first review the KSVD algorithm for learning overcomplete nonparametric

⁵For practical reasons we are of course interested in the case of compressible signals and nonideal noisy signals. The RIP property applies to the ideal case as well, providing reconstruction guarantees both for the ℓ_1 and OMP techniques, as detailed next and in the mentioned references.

ictionaries from image datasets, [12], [21], [45], and then introduce the novel *coupled-KSVD* as a necessary modification to include the simultaneous learning of both the dictionary and the corresponding sampling projection matrix. In Section IV, we present detailed experimental results indicating the superiority of our new framework to construct deterministic sensing matrices, and the results of simultaneously learning generic sensing matrices and overcomplete dictionaries using datasets of image patches. Finally, concluding remarks and directions for future research are presented in Section V.

While the framework here introduced is applicable to signals in general, from now on, we consider only natural images, and the sparsifying basis will always be dictionaries learned from image datasets.

II. OPTIMIZED PROJECTIONS FOR COMPRESSIVE SAMPLING

In this section, we show how to optimize the sensing matrix given a sparsifying dictionary. We start by reviewing prior related work, following by the presentation of our proposed algorithm and the first results showing the computational and reconstruction advantages of this new approach.

A. Previous Related Work

To the best of our knowledge, only very few recent publications, [40]–[44], explicitly address the idea that nonrandom matrices are important and could be more effective than random projection matrices for sensing sparse signals. In particular, [42] shows that chirp-based sampling matrices can be used instead of random sampling matrices, retaining the same reconstruction accuracy, but with the advantage that the retrieval of the original signal becomes computationally much cheaper than using, for example, orthogonal matching pursuit (OMP). On the other hand, [40], [41] address a different idea: random sampling matrices are not necessarily optimal, in the reconstruction error sense, to sample specific classes of sparse signals, such as natural images.

Considering well-known characteristics and models for the second order statistics of natural images, see, e.g., [46]–[49], Weiss *et al.*, [41], first showed that the Signal to Noise Ratio (SNR) of images projected using (almost any) random sampling matrices goes to zero as the number of pixels increase, while it remains constant for their proposed deterministic sampling matrices. They then introduced the concept of Uncertain Component Analysis (UCA), leading to maximize the posterior probability of the data \mathbf{x} for a given projection matrix Φ and training projection data \mathbf{y} .

Let $\mathbf{x}_1, \dots, \mathbf{x}_P$ be a set of P training patches and $\mathbf{y}_1, \dots, \mathbf{y}_P$ their respective projections, $\mathbf{y}_i = \Phi \mathbf{x}_i$, $1 \leq i \leq P$. Then, the optimal projection matrix, maximizing the probability of retrieving the original patches, is given by $\hat{\Phi} := \arg\max_{\Phi, \|\Phi\|=1} \prod_i P(\mathbf{x}_i | \mathbf{y}_i; \Phi)$. The prior probability is assumed to be i.i.d. Gaussian, $P(\mathbf{y}_i | \mathbf{x}; \Phi) = (1/\sqrt{2\pi\sigma^2})e^{-\|\mathbf{y}_i - \Phi \mathbf{x}\|_{\ell_2}^2 / 2\sigma^2}$. Experimentally, the authors found that UCA works only slightly better than random projection on synthetic signals of fixed sparsity, and it works significantly better than random projections on patches of natural images.

Elad [40] introduced a new algorithm that does not make any assumptions on the statistics of the dataset, but attempts to directly improve the incoherence between the $m \times n$ sampling matrix Φ (m samples of the n dimensional signal),⁶ and the $n \times K$ sparsifying dictionary Ψ , as required by the CS theory, assuming that a given overcomplete dictionary Ψ has been already provided (note that [41] is not explicitly based on a dictionary, which in our proposed work below we directly learn from the data). Experimentally, Elad showed, using synthetic random signals of fixed and exact sparsity, that a well-designed projection matrix that depends on the sparsifying basis Ψ , can reduce the mutual coherence of the equivalent dictionary $\mathbf{D} = \Phi\Psi$ and, hence, reduce the reconstruction error from the projections.

Our work follows in part Elad's idea, since we do not assume any prior knowledge or statistics on the data set, while in contrast with [40], we do not assume a given sparsifying dictionary. In fact, both the overcomplete dictionary, Ψ , and projection matrix, Φ , are learned from the dataset (see next section for the general case of simultaneous learning). Both [40] and [41] provide a way of computing Φ from data or a dictionary, although do not provide any hint on how Φ might help the actual learning of the sparsifying basis Ψ , and vice versa.

Our work starts with presenting an alternative to the framework in [40] for learning the projecting matrix from the previously learned dictionary. This alternative is computationally more efficient and produces significantly better reconstruction results. We, therefore, start by briefly describing the algorithm in [40].

Motivated by the incoherence required by CS, as well as the fundamental conditions for optimal performance of orthogonal matching pursuit and basis pursuit (see previous section), Elad proposed to reduce the *mutual coherence*, $\mu(\mathbf{D})$, of the equivalent dictionary $\mathbf{D} := \Phi\Psi$, $\mathbf{D} = [\mathbf{d}_1 \dots \mathbf{d}_K]$, defined as [40]

$$\mu(\mathbf{D}) := \max_{i \neq j, 1 \leq i, j \leq K} \left\{ \frac{\mathbf{d}_i^T \mathbf{d}_j}{\|\mathbf{d}_i\|_{\ell_2} \|\mathbf{d}_j\|_{\ell_2}} \right\}. \quad (7)$$

If the projection matrix Φ is designed such that $\mu(\Phi\Psi)$ is minimal (recall that Ψ is fix for the moment), a larger number of signals would satisfy (5) and be successfully recovered from their linear projections.

Instead of minimizing (7), Elad proposes to work with

$$\mu_t(\mathbf{D}) := \frac{\sum_{i \neq j, 1 \leq i, j \leq K} (|g_{ij}| > t) \cdot |g_{ij}|}{\sum_{i \neq j, 1 \leq i, j \leq K} (|g_{ij}| > t)} \quad (8)$$

where $g_{ij} = \tilde{\mathbf{d}}_i^T \tilde{\mathbf{d}}_j$, $\tilde{\mathbf{d}}_i = \mathbf{d}_i / \|\mathbf{d}_i\|_{\ell_2}$, and t is a scalar that establishes the minimum value of $\mu_t(\mathbf{D})$ (the sum over values greater than t as represented by the indicator function $|g_{ij}| > t$ in (8)). From (8), then is obvious that $\mu_t \geq t$, since it is the “above value t ” average of the off-diagonal elements of the matrix. Hence, t is the target value Elad proposes to minimize.

An alternative way of looking at the mutual coherence of the equivalent dictionary \mathbf{D} is to consider the Gramm matrix, $\mathbf{G} := \tilde{\mathbf{D}}^T \tilde{\mathbf{D}}$, where $\tilde{\mathbf{D}} = [\tilde{\mathbf{d}}_1 \dots \tilde{\mathbf{d}}_K]$ is the equivalent dictionary \mathbf{D} with all its columns normalized. Elad's idea is to minimize

⁶We use n for the signal dimension since we will work with patches.

the largest absolute values of the off-diagonal elements in the corresponding Gramm matrix, while keeping the rank of the equivalent dictionary equal to $m \ll n$. This in turn minimizes μ_t .

Instead of targeting $\mu_t(\mathbf{D})$, we address the problem of making any subset of columns in \mathbf{D} as orthogonal as possible, or equivalently, making \mathbf{G} as close as possible to the identity matrix. We then directly target what we have learned from the RIP, which also guarantees robustness of CS to noise and nonexact sparsity. As we show below, our proposed approach outperforms the one introduced in [40], especially for real not-exactly sparse signals (images). Computationally, the proposed algorithm is significantly more efficient than the one in [40]. Moreover, after introducing the proposed algorithm to achieve this close to the identity Gramm matrix, we also introduce the novel idea of simultaneously designing Φ and Ψ .

B. Learning the Sensing Matrix

As mentioned above, and considering for the moment that the dictionary Ψ is known, we want to find the sensing matrix Φ such that the corresponding Gramm matrix is as close to the identity as possible, i.e.,

$$\Psi_{Kn}^T \Phi_{nm}^T \Phi_{mn} \Psi_{nK} \approx \mathbf{I}_K.$$

Let us multiply both sides of the previous expression by Ψ on the left and Ψ^T on the right, obtaining

$$\Psi \Psi^T \Phi^T \Phi \Psi \Psi^T \approx \Psi \Psi^T.$$

Let $\mathbf{V} \mathbf{A} \mathbf{V}^T$ be the (known) eigen-decomposition of $\Psi \Psi^T$, then $\mathbf{V} \mathbf{A} \mathbf{V}^T \Phi^T \Phi \mathbf{V} \mathbf{A} \mathbf{V}^T \approx \mathbf{V} \mathbf{A} \mathbf{V}^T$, which is equivalent to $\mathbf{A} \mathbf{V}^T \Phi^T \Phi \mathbf{V} \mathbf{A} \approx \mathbf{A}$. Let us define $\Gamma_{mn} := \Phi_{mn} \mathbf{V}_{nn}$. Then, $\mathbf{A} \Gamma^T \Gamma \mathbf{A} \approx \mathbf{A}$. We want to compute $\Phi(\Gamma)$ in order to minimize

$$\|\mathbf{A} - \mathbf{A} \Gamma^T \Gamma \mathbf{A}\|_{\mathbb{F}}^2. \quad (9)$$

If Ψ is an orthonormal basis and $m = n$ (non standard in CS), then the previous equation would have an exact solution that produces zero error, i.e., $\Gamma = \mathbf{A}^{-1/2}$. However, since the dictionary is often overcomplete, a critical aspect for achieving high sparsity and state-of-the-art image reconstruction [12], [21], [45], and in particular $m \ll n$ (and then $m \ll \text{rank}(\mathbf{A})$), we have to find an approximated solution for minimizing the error in (9). We will achieve this starting from a random sensing matrix Φ (and its corresponding Γ), and progressively improving it in order to reduce this error.⁷ This is detailed next.

Let $\lambda_1, \dots, \lambda_n$ be the eigenvalues of the known diagonal matrix \mathbf{A} , ordered in decreasing order of magnitude, and $\Gamma_{mn} = [\tau_1 \dots \tau_m]^T$. Then, (9) becomes

⁷Alternatively, we can obtain a closed solution to (9) of the form $\Gamma := [\Gamma_1, \Gamma_2]$, where Γ_1 is a diagonal matrix obtained from the top m eigenvectors of \mathbf{A} (elevated to the $-1/2$ power). While this provides a slightly faster algorithm than the m -steps one here proposed, it produces virtually the same reconstruction results. Our proposed approach follows the algorithmic concepts of the KSVD and couple-KSVD described in the next section for dictionary learning, and shows how to progressively improve the classical random matrix of CS. This technique also provides a possible solution to the ambiguity brought by this closed form alternative Γ [we thank D. Goldfarb and S. Ma for pointing out this closed solution and additional comments on the minimization of (9)].

$\|\mathbf{A} - \sum_{i=0}^m \nu_i \nu_i^T\|_{\mathbb{F}}^2$, where $\nu_i = [\lambda_1 \tau_{i,1} \dots \lambda_n \tau_{i,n}]^T$, or equivalently

$$\left\| \mathbf{A} - \sum_{\substack{i=0, \\ i \neq j}} \nu_i \nu_i^T - \nu_j \nu_j^T \right\|_{\mathbb{F}}^2. \quad (10)$$

Let us define $\mathbf{E} := \mathbf{A} - \sum_{i=0}^m \nu_i \nu_i^T$, $\mathbf{E}_j := \mathbf{A} - \sum_{\substack{i=0, \\ i \neq j}}^m \nu_i \nu_i^T$, and let $\mathbf{E}_j = \mathbf{U}_j \mathbf{\Delta} \mathbf{U}_j^T$ be the eigen-decomposition of \mathbf{E}_j . Then, (10) becomes $\|\mathbf{E}_j - \nu_j \nu_j^T\|_{\mathbb{F}}^2 = \left\| \sum_{k=1}^n \xi_{k,j} \mathbf{u}_{k,j} \mathbf{u}_{k,j}^T \nu_j \nu_j^T \right\|_{\mathbb{F}}^2$. If we set $\nu_j = \sqrt{\xi_{1,j}} \mathbf{u}_{1,j}$, $\xi_{1,j}$ being the largest eigenvalue of \mathbf{E}_j and $\mathbf{u}_{1,j}$ its corresponding eigenvector, then the largest error component in \mathbf{E} is eliminated. Replacing ν_j back in terms of τ_j (the rows of the matrix we are optimizing for, $\Gamma = \Phi \mathbf{V}$)

$$[\lambda_1 \tau_{j,1} \dots \lambda_n \tau_{j,n}]^T = \sqrt{\xi_{1,j}} \mathbf{u}_{1,j}. \quad (11)$$

Since the matrix \mathbf{A} is in general not full-rank, then for some $r \geq 0$, $\lambda_{n-r+1}, \dots, \lambda_n$ will be zero, and we can only update the $\tau_{j,1}, \dots, \tau_{j,n-r}$ components of τ_j . This derivation forms the basis of our algorithm for optimizing (9); see below for the exact steps.

Once we obtain $\hat{\Gamma}$, then $\hat{\Phi}$ can be easily computed following the relationship $\hat{\Gamma} = \hat{\Phi} \mathbf{V}$, as $\hat{\Phi} = \hat{\Gamma} \mathbf{V}^T$. Since we can only reduce $m \ll n$ components of the error matrix \mathbf{E} , and the error has a rank lower but close to n (recall that Ψ is almost an orthonormal basis for \mathbb{R}^n), then there is no actual hope to completely eliminate the error in (9). The proposed technique aims at reducing the largest m components of this error matrix \mathbf{E} .

In summary, the following are the steps of the proposed, non-iterative, m -steps algorithm for optimizing the sensing matrix given the dictionary (using the notation defined above).

1. Initialize $\hat{\Phi}$, for example, to a random matrix.
2. Find the eigen-decomposition $\Psi \Psi^T = \mathbf{V} \mathbf{A} \mathbf{V}^T$ and r , the number of nonzero eigenvalues of $\Psi \Psi^T$.
3. Initialize $\hat{\Gamma} := \hat{\Phi} \mathbf{V}$.
4. For $j = 1$ to m :
 Compute \mathbf{E}_j .
 Find the largest eigenvalue and corresponding eigenvector of \mathbf{E}_j , $\xi_{1,j}$ and $\mathbf{u}_{1,j}$.
 Use (11) to update the first r components of $\hat{\tau}_j$ (thereby updating $\hat{\Gamma}$).
5. Compute the optimal $\hat{\Phi} = \hat{\Gamma} \mathbf{V}^T$.

C. Some Preliminary Experimental Results

As we will show in the experimental results, Section IV, this parameter-free algorithm is not only considerably faster than the one introduced in [40], but also significantly improves the reconstruction results and provides a fundamental building block for the simultaneous optimization of the dictionary and sensing matrix. Before these more detailed experimental results, let us now present some illustrative results showing the advantages of the proposed method over the algorithm proposed in [40]. For this, we will also consider below the (average) mutual coherence

TABLE I
COMPARISON OF ELAD'S VERSUS THE NEW PROPOSED ALGORITHM

S	m	Patches	Projection	time(s)	μ_t	σ_μ	μ_g	σ_μ	MSE	σ_{MSE}
4	16	Synthetic	Random	-	0.482	0.005	0.078	0.002	2.957	0.071
			Elad's algorithm	794.53	0.434	0.009	0.069	0.002	1.694	0.071
			New Algorithm	0.23	0.421	2.87E-04	0.059	1.37E-05	1.549	0.024
6	15	Image, fixed sparsity	Random	-	0.582	0.0139	0.115	0.0062	0.152	0.029
			Elad's algorithm	341.55	0.515	0.0146	0.095	0.0052	0.076	0.006
			New Algorithm	0.22	0.456	6.91E-04	0.063	5.09E-05	0.022	0.001
6	15	Real images	Random	-	0.535	0.010	0.097	0.004	0.331	0.028
			Elad's algorithm	426.85	0.487	0.013	0.085	0.004	0.357	0.022
			New Algorithm	0.80	0.442	4.22E-04	0.064	4.69E-05	0.115	0.002

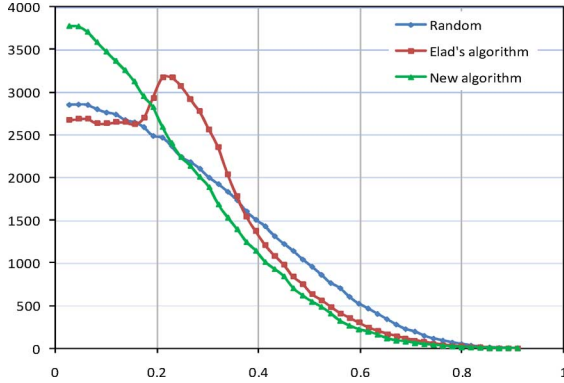


Fig. 1. Distribution of the off-diagonal elements of the Gramm matrix obtained using a random sampling matrix, Elad's algorithm, and the new proposed algorithm.

$$\mu_g(\mathbf{D}) := \frac{\sum_{i \neq j} g_{ij}^2}{K(K-1)}. \quad (12)$$

This is simply the mean square error that accounts for the off-diagonal elements in the Gramm matrix, while (8) only accounts for the larger than t off-diagonal values.

Fig. 1 compares, for three different sensing matrices Φ , the distribution of the absolute value of the off-diagonal elements of the Gramm matrix obtained using a dictionary Ψ learned from 440 natural images (see Section IV for more details on this dictionary). The three sensing matrices considered are: a Gaussian random sampling matrix (as commonly used in CS), the sensing matrix obtained using Elad's algorithm (with parameters optimized to reduce $\mu_t(\mathbf{D})$; $\gamma = 0.6$, $t = 20\%$), and our proposed algorithm.

Both our technique and the one proposed in [40] try to reduce the largest off-diagonal elements in the Gramm matrix, however, Elad's algorithm always presents a consistent artifact (see also [40]), where some off-diagonal elements in the Gramm matrix actually increase their value (notice the peak between 0.2 and 0.4 in Fig. 1). This does not affect μ_t but it negatively affects the RIP (see previous Section). Our proposed algorithm increases the frequency of the off-diagonal elements with low absolute value (between 0 and 0.2, Fig. 1), and reduces the frequency of large absolute values in the Gramm matrix, better enforcing the RIP and attempting to make the columns of \mathbf{D} as close to orthogonal as possible.

This better behavior of the Gramm matrix (recall the RIP) is reflected in increased accuracy in image reconstruction as well. Table I compares Elad's and our new proposed algorithm to design projection matrices, using patches/signals corresponding to synthetic data of fixed sparsity (as in [40]), patches coming from real images preprojected to have a fixed sparsity, and patches coming from real images without restricting their sparsity (compressible patches in contrast to sparse ones). The algorithm in [40] enforces a measure of incoherence, but not directly the RIP and its intuition, and does not actually outperform random sampling, in the mean square error (MSE) sense, when the patches do not have an exact fixed sparsity (see Table I). The first row in Table I corresponds to 10,000 synthetic signals obtained by combining at random $S = 4$ columns of a randomly generated 64×256 dictionary. The second row corresponds to 6,600 patches projected to have a fixed sparsity, $S = 6$, obtained using OMP to reconstruct the real image patches; the dictionary Ψ of size 64×256 was trained using these patches and the KSVD algorithm (see next section). The last row corresponds to 6,600 patches selected at random from 440 real images and a 64×256 dictionary Ψ trained with them using KSVD. The table reports average and variance results.

From this table, we note that $\mu_t(\mathbf{D})$, computed using (8), is similar in Elad's and our algorithm, but $\mu_g(\mathbf{D})$ is lower for the proposed algorithm. The reason is that (8) only takes into account the large off-diagonal values in the Gramm matrix, ignoring the values of the remaining terms and, hence, the artifact introduced by Elad's algorithm (Fig. 1). Finally, note the significant MSE improvements obtained with our proposed algorithm.

For illustration purposes only, Table I also shows the running time of Elad's algorithm versus the proposed new algorithm, on a laptop with a single 1.6-GHz processor and 1.5 Gb of RAM. Even though the time will change from one implementation and computer to another, this illustrates the significant computational advantage of our proposed technique. Elad's algorithm takes about 600 times longer, and clearly not all of this is the consequence of implementation issues. This significant speed-up is expected from a noniterative algorithm as the one here introduced.

Since Elad's algorithm is not robust in the presence of nonidealities such as nonexact sparsity, and its running time becomes prohibitive for practical applications, from now on when we refer to the nonrandom sampling matrix $\Phi(\Psi)$ computed from a dictionary Ψ , we consider the new algorithm here proposed.

III. SIMULTANEOUSLY LEARNING THE DICTIONARY AND THE PROJECTION MATRIX

It is now time to turn to the simultaneous learning of the sparsifying dictionary and the sensing matrix. This will be based on combining the just introduced approach for learning the sensing matrix with the KSVD algorithm for dictionary learning. We start then by briefly introducing this KSVD technique.

A. KSVD Algorithm

Recently, Aharon *et al.*, [21], [45] introduced a novel algorithm for **learning overcomplete dictionaries to sparsely represent images**. Let $\mathbf{X} = [\mathbf{x}_1 \dots \mathbf{x}_P]$ be an $n \times P$ matrix of P training square patches of length n pixels each, used to train an overcomplete dictionary Ψ of size $n \times K$, with $P \gg K$ and $K > n$. The objective of the KSVD algorithm is to solve, for a given sparsity level S

$$\min_{\Psi, \Theta} \|\mathbf{X} - \Psi\Theta\|_F^2 \text{ s.t. } \forall i, \|\theta_i\|_{\ell_0} \leq S \quad (13)$$

where $\Theta = [\theta_1 \dots \theta_P]$, and θ_i is the sparse vector of coefficients representing the i^{th} patch in terms of the columns of the dictionary $\Psi = [\psi_1 \dots \psi_K]$. Starting from an arbitrary Ψ , the KSVD algorithm progressively improves it in order to optimize the above expression, as described next.

Let $\Theta = [\delta_1 \dots \delta_K]^T$, where δ_i^T are the rows of Θ . Then, as in the previous section, the error term in (13) can be decomposed as $\|\mathbf{X} - \sum_i \psi_i \delta_i^T\|_F^2 = \|\mathbf{X} - \sum_{i \neq j} \psi_i \delta_i^T - \psi_j \delta_j^T\|_F^2$. Let us define $\mathbf{E} := \mathbf{X} - \Psi\Theta$ and $\mathbf{E}_j := \mathbf{X} - \sum_{i \neq j} \psi_i \delta_i^T$. Then, (13) can be rewritten as

$$\min_{\Psi, \Theta} \|\mathbf{E}_j - \psi_j \delta_j^T\|_F^2 \text{ s.t. } \forall i, \|\theta_i\|_{\ell_0} \leq S. \quad (14)$$

At this point, **it is very tempting to obtain the SVD decomposition of \mathbf{E}_j and eliminate the largest component of the error matrix \mathbf{E} (see previous section)**. However, (14) requires also satisfying the sparsity constraint. We then have to design a different procedure to solve (14). We show this next.

Let ω_j be the set of all indices corresponding to the training patches that use the atom ψ_j for a given (temporary) dictionary ψ (this is determined simply using OMP or any other sparse representation technique, see below), i.e.,

$$\omega_j := \{p \mid 1 \leq p \leq P, \delta_j^T(p) \neq 0\}, 1 \leq j \leq K. \quad (15)$$

In matrix form, let now Ω_j be a $P \times |\omega_j|$ matrix with ones on the $(\omega_j(i), i)$ entries and zero elsewhere. Then, solving for (14) is equivalent to solving for

$$\min_{\Psi, \Theta} \|\mathbf{E}_j \Omega_j - \psi_j \delta_j^T \Omega_j\|_F^2 \quad (16)$$

where the sparsity (active) set cannot be altered for now. Let $\mathbf{E}_j^R := \mathbf{E}_j \Omega_j$, $\delta_{j,R}^T := \delta_j^T \Omega_j$, thereby, \mathbf{E}_j^R are just the columns of the error \mathbf{E}_j corresponding to atom ψ_j and $\delta_{j,R}^T$ the rows of Θ , where the zeros have been removed. Let $\mathbf{U}\Lambda\mathbf{V}^T$ be the singular value decomposition (SVD) of \mathbf{E}_j^R . Then, (16) becomes

$$\min_{\Psi, \Theta} \|\mathbf{U}\Lambda\mathbf{V}^T - \psi_j \delta_{j,R}^T\|_F^2. \quad (17)$$

We now eliminate the highest component of the error by defining

$$\psi_j := \mathbf{u}_1, \quad \delta_{j,R} := \sigma_1 \mathbf{v}_1 \quad (18)$$

where σ_1 is the largest singular value of \mathbf{E}_j^R and $\mathbf{u}_1, \mathbf{v}_1$ are the corresponding left and right singular vectors. This then improves the dictionary atom ψ_j based on the patches that have used it when considering the temporary dictionary Ψ . This continues in the same fashion for all the other columns (atoms).

In summary, the KSVD algorithm consists of the following key steps:⁸

1. Initialize $\hat{\Psi}$
2. Repeat until convergence:
 - For $\hat{\Psi}$ fixed, solve (13) using OMP to obtain $\hat{\Theta}$,⁹ $\hat{\Theta} = \text{OMP}(\hat{\Psi}, \mathbf{X})$.
 - For $j = 1$ to K
 - Compute ω_j and from there, $\mathbf{E}_j^R, \delta_{j,R}^T$.
 - Obtain the largest singular value of \mathbf{E}_j^R and the corresponding singular vectors.
 - Update $\hat{\Psi}$ and $\hat{\Theta}$ using (18).¹⁰

Experimentally, we found that initializing Ψ with an overcomplete dictionary using the Discrete Cosine Transform (DCT) [21] produces better results than with a zero mean, normalized random matrix. The results reported in Section IV use this initialization method.

B. Coupled-KSVD

Let us consider now the **problem of simultaneously training a dictionary Ψ and the projection matrix Φ** , with the images available from a dataset. We define the following optimization problem:

$$\min_{\Psi, \Phi, \Theta} \{\alpha \|\mathbf{X} - \Psi\Theta\|_F^2 + \|\mathbf{Y} - \Phi\Psi\Theta\|_F^2\} \text{ s.t. } \forall i, \|\theta_i\|_{\ell_0} \leq S \quad (19)$$

where $0 \leq \alpha \leq 1$ is a scalar that controls the weight of the error term $\|\mathbf{X} - \Psi\Theta\|_F^2$ and \mathbf{Y} are the linear samples given by

$$\mathbf{Y} = \Phi\mathbf{X} + \boldsymbol{\eta} \quad (20)$$

considering $\boldsymbol{\eta}$ an additive noise added by the sensing system. Since $\mathbf{Y} = [\mathbf{y}_1 \dots \mathbf{y}_P]$ is an $m \times P$ matrix with $m \ll n, \alpha$ is used in (19) to compensate for the larger value of the reconstruction error given by the term $\|\mathbf{X} - \Psi\Theta\|_F^2$, and also to give more importance to the projection error term, $\|\mathbf{Y} - \Phi\Psi\Theta\|_F^2$,

⁸While the learning of the sensing matrix as presented in the previous section is a sequential and noniterative algorithm, in the KSVD we do have iterations, since both the dictionary and the corresponding sparse coefficients are optimized for.

⁹The original KSVD uses OMP for the sparse coding step, other sparsifying techniques could be used as well.

¹⁰As in a Gauss–Seidel type of approach, both the atom and the corresponding coefficients are updated at this step.

which is what is actually available at the reconstruction stage.¹¹ Equation (19) can be rewritten as

$$\min_{\Psi, \Phi, \Theta} \left\| \begin{pmatrix} \alpha \mathbf{X} \\ \mathbf{Y} \end{pmatrix} - \begin{pmatrix} \alpha \mathbf{I} \\ \Phi \end{pmatrix} \Psi \Theta \right\|_F^2 \quad s.t. \quad \forall i, \|\theta_i\|_{\ell_0} \leq S. \quad (21)$$

A possible way to solve (19), (21) towards a (local) minimum consists in extending the KSVD algorithm to the coupled \mathbf{X} and \mathbf{Y} signals, combining this with the technique to adapt Φ to Ψ introduced in the previous section. As in the KSVD algorithm, we start with an arbitrary dictionary, learn the sensing matrix most appropriate to it following the approach described in the previous section, and then simultaneously improve both of them. Toward achieving this, let us define

$$\mathbf{Z} := \begin{pmatrix} \alpha \mathbf{X} \\ \mathbf{Y} \end{pmatrix}, \quad \mathbf{W} := \begin{pmatrix} \alpha \mathbf{I} \\ \Phi \end{pmatrix}. \quad (22)$$

Then, (19) and (21) can be rewritten as

$$\min_{\Psi, \Phi, \Theta} \|\mathbf{Z} - \mathbf{D}_{\text{eq}} \Theta\|_F^2 \quad s.t. \quad \forall i, \|\theta_i\|_{\ell_0} \leq S \quad (23)$$

where $\mathbf{D}_{\text{eq}} := \mathbf{W}\Psi = [\mathbf{d}_1^{\text{eq}} \quad \dots \quad \mathbf{d}_K^{\text{eq}}]$. As in KSVD, we can write

$$\left\| \mathbf{Z} - \sum_i \mathbf{d}_i^{\text{eq}} \delta_i^T \right\|_F^2 = \left\| \mathbf{Z} - \sum_{i \neq j} \mathbf{d}_i^{\text{eq}} \delta_i^T - \mathbf{d}_j^{\text{eq}} \delta_j^T \right\|_F^2 \quad (24)$$

where δ_i^T are the rows of Θ as defined previously for the standard KSVD algorithm. Let us now define $\mathbf{E} := \mathbf{Z} - \mathbf{D}_{\text{eq}} \Theta$, $\mathbf{E}_j := \mathbf{Z} - \sum_{i \neq j} \mathbf{d}_i^{\text{eq}} \delta_i^T$. Considering also \mathbf{E}_j^R , $\delta_{j,R}^T$, and Ω_j , as defined in the KSVD algorithm, (23) can be rewritten as

$$\min_{\Psi, \Phi, \Theta} \|\mathbf{E}_j^R - \mathbf{d}_j^{\text{eq}} \delta_{j,R}^T\|_F^2. \quad (25)$$

Similarly, let $\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$ be the singular value decomposition (SVD) of \mathbf{E}_j^R , then (25) becomes

$$\min_{\Psi, \Phi, \Theta} \|\mathbf{U}\mathbf{\Lambda}\mathbf{V}^T - \mathbf{d}_j^{\text{eq}} \delta_{j,R}^T\|_F^2 \quad (26)$$

and the highest component of the (coupled) error can be eliminated defining

$$\mathbf{d}_j^{\text{eq}} \delta_{j,R}^T := \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T \quad (27)$$

where σ_1 is the largest singular value of \mathbf{E}_j^R and \mathbf{u}_1 , \mathbf{v}_1 are the corresponding left and right singular vectors.

Now, since $\mathbf{d}_j^{\text{eq}} = \begin{pmatrix} \alpha \mathbf{I} \\ \Phi \end{pmatrix} \psi_j$, (27) is satisfied if

$$\begin{pmatrix} \alpha \mathbf{I} \\ \Phi \end{pmatrix} \psi_j = \mathbf{u}_1, \quad \delta_{j,R} = \sigma_1 \mathbf{v}_1 \quad (28)$$

¹¹While at the sparsifying dictionary and sensing matrix training step, we have available both the images, \mathbf{X} , and their projections, \mathbf{Y} ; at the actual reconstruction step we have only the sensed data, \mathbf{Y} , and the goal is to reconstruct from it the sparsest possible \mathbf{X} , with the learned (Ψ, Φ) (this is the standard CS/sparse-reconstruction scenario, but with our optimized pair (Ψ, Φ)).

where we have $m+n$ equations and n unknowns (the length of the ψ_j atom of dictionary Ψ).

Note the importance of coupling the original images \mathbf{X} and their corresponding sensed data \mathbf{Y} , if we would have not include \mathbf{X} , (28) would have m equations and n unknowns, hence, infinitely many solutions for ψ_j would satisfy (28). In other words, by introducing the regularizing condition that $\|\mathbf{X} - \Psi\Theta\|_F^2$ must also be minimized, we obtain a unique solution to (28) that best fits the training data \mathbf{X} and its projection \mathbf{Y} .

From (28), ψ_j can be computed using the pseudo-inverse as

$$\hat{\psi}_j = (\alpha^2 \mathbf{I} + \Phi^T \Phi)^{-1} (\alpha \mathbf{I} \quad \Phi^T) \mathbf{u}_1. \quad (29)$$

Since $\hat{\psi}_j$ computed using (29) does not necessarily have unit ℓ^2 -norm, and the columns $\hat{\psi}_j$ of the learned dictionary $\hat{\Psi}$ should have unit ℓ^2 -norm [45], we redefine $\hat{\psi}_j$ and $\delta_{j,R}$ as

$$\hat{\psi}_j \leftarrow \frac{\hat{\psi}_j}{\|\hat{\psi}_j\|_{\ell_2}}, \quad \delta_{j,R} \leftarrow \|\hat{\psi}_j\|_{\ell_2} \delta_{j,R} \quad (30)$$

in order to keep the product $\mathbf{d}_j^{\text{eq}} \delta_{j,R}^T$ on (27) unchanged.

We have now updated the dictionary Ψ and the corresponding sparse coefficients Θ (repeating, as in the KSVD, the above procedure for all the atoms), considering Φ fix. The feedback of the sensing matrix into the update of the dictionary is provided using $\Phi(\Psi)$ as defined in Section II, i.e., learning the projection matrix from the just updated dictionary Ψ . In turn, Ψ will be affected by Φ , as indicated on (28), in the next iteration. Additionally, OMP also uses the coupled signal \mathbf{Z} to estimate Θ . Thereby, the whole learning procedure is simultaneous and exploits all the available data.

In summary, the proposed coupled-KSVD algorithm is,¹²

1. Initialize $\hat{\Psi}$.
2. Repeat until convergence:
 - For $\hat{\Psi}$ fixed, compute $\hat{\Phi}(\hat{\Psi})$ using the algorithm given on Section II.
 - For $\hat{\Psi}$, $\hat{\Phi}$ fixed, solve ((21) using OMP to obtain $\hat{\Theta}$, i.e., $\hat{\Theta} = \text{OMP}(\mathbf{D}_{\text{eq}}, \mathbf{Z})$.
 - For $j = 1$ to K
 - Compute ω_j and from there \mathbf{E}_j^R , $\delta_{j,R}^T$ using (22)–(24).
 - Obtain the largest singular value of \mathbf{E}_j^R and the corresponding singular vectors.
 - Update $\hat{\Psi}$ and $\hat{\Theta}$ using (28), (29), and (30).

In the next section, we evaluate the performance of the improved algorithm to compute the sensing projection matrix $\Phi(\Psi)$ from a given sparsifying dictionary Ψ , and of the coupled-KSVD just introduced, and show their advantage over previously reported techniques.

IV. EXPERIMENTAL RESULTS

In this section, we compare different methods to compute the dictionary, Ψ , and sampling matrix, Φ , designed to retrieve the

¹²Note that the projected signal \mathbf{Y} is not explicitly needed, see the algorithm in Section II.

original image patches from their linear measurements.¹³ We compare the retrieval error of testing patches \mathbf{X} extracted from real images, and reconstructed using OMP,¹⁴ with the equivalent dictionary $\mathbf{D} = \Phi\Psi$, from their noisy linear samples, $\mathbf{Y} = \Phi\mathbf{X} + \boldsymbol{\eta}$. The following (Φ, Ψ) strategies are tested and compared.

- A dictionary Ψ , learned from the real image data using the standard KSVD, and a Gaussian random sampling matrix Φ . This represents the standard CS scenario.
- A dictionary Ψ , learned from the real image data using the standard KSVD, and then combined with the optimized sensing matrix $\Phi(\Psi)$ computed from this dictionary, as described in Section II.
- A dictionary Ψ , learned from the data using the coupled-KSVD with a fix Gaussian random sampling matrix Φ .¹⁵
- A dictionary Ψ and sampling matrix $\Phi(\Psi)$, both learned from the data using the full coupled-KSVD introduced in this paper.

The first two strategies are uncoupled, since the dictionary Ψ is learned using the original KSVD, independently of Φ . The third strategy is semi-coupled, since the sampling projection matrix Φ affects the learning process of Ψ through the samples \mathbf{Y} (Section III), but not vice versa. The fourth strategy is completely coupled, since both Φ and Ψ affect each other during the learning process: Φ affects the learning of Ψ through the samples \mathbf{Y} , and in turn, Ψ affects the sampling matrix, since Φ depends on $\Psi\Psi^T$ (Section II). In the following we refer to each strategy as uncoupled random (UR), uncoupled learning (UL), coupled random (CR), and coupled learning (CL), respectively.

The training data consists of 6,600 8×8 patches obtained by extracting at random 15 patches from each one of the 440 images in the training set (250 images from the Berkeley segmentation data set [50] and 190 images from the *Labelme* data set [51]). The testing data consists of 120,000 patches corresponding to all the nonoverlapping patches of size 8×8 extracted from the remaining 50 images in the Berkeley dataset that are not in the training set.

The different strategies (UR, UL, CR, and CL), are all evaluated in terms of the MSE of retrieval, defined as $\text{MSE} := \|\mathbf{X} - \Psi\hat{\Theta}\|_F^2$, where Ψ is the dictionary learned from the training patches, \mathbf{X} is the matrix of testing patches, and $\hat{\Theta}$ is obtained using OMP to solve

$$\min_{\hat{\Theta}} \|\mathbf{Y} - \Phi\Psi\hat{\Theta}\|_F^2 \quad \text{s.t.} \quad \forall i, \|\hat{\Theta}_i\|_{\ell_0} \leq S \quad (31)$$

that is, $\hat{\Theta} = \text{OMP}(\mathbf{D}, \mathbf{Y})$, where, Ψ is a random sampling matrix for the UR and CR strategies and a learned sampling matrix for the UL and CL strategies, and \mathbf{Y} is a noisy version of the projected patches, as defined in (20). The noise added to

¹³Recall that in the real CS-type scenario, once we have already learned the dictionary and sensing matrix, we have to recover the signal only from its linear projections.

¹⁴To be consistent with the KSVD-based dictionary/sensing training, we use OMP at the reconstruction step as well. As previously mentioned, we can replace this by other sparsifying techniques.

¹⁵This means that we incorporate the sensing in the learning of the dictionary, but do not update the sensing matrix and keep it constant during the iterations of the coupled-KSVD algorithm.

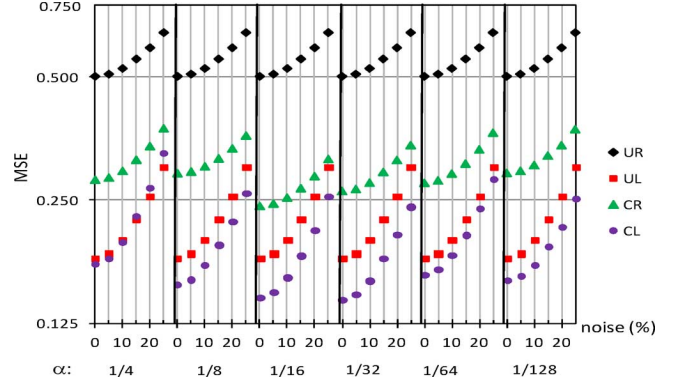


Fig. 2. Retrieval MSE using the four training strategies at different noise levels and values of α (separated by vertical bold lines), for $S = 6$, $m = 12$, $n = 64$, and $K = 256$.

the samples ranges from 0 to 25% in amplitude, with 5% increments. The parameter α in the coupled-KSVD was varied in the set $\{1/2, 1/4, 1/8, 1/16, 1/32, 1/64, 1/128\}$. We use typical values for the other algorithm parameters: sparsity $S = 4$, 5, and 6; sampling dimension $m = 2S, 2.5S$, and $3S$; patch dimension $n = 64$ (8×8); and overcompleteness of $K = 4n$. These values are commonly used in learning overcomplete dictionaries and Compressive Sensing (see, for instance, [1]–[8], [12], [21], and [45]). We include here representative results from this large set of possible parameter combinations, see the supplementary material for numerous additional graphs and tables.

Fig. 2 compares the average MSE of retrieval for the testing patches using $S = 6$, $m = 12$, $n = 64$, and $K = 256$; at different values of α and noise levels, for the four training strategies. We clearly observe the significant advantage of learning the sensing matrix (coupled or uncoupled from the dictionary) over the more standard use of random projections.

Fig. 3(a) shows the retrieval MSE of CL relative to CR. The MSE using coupled learning (CL) is almost 50% (a reduction of ~ 3 Db) of the MSE using semi-coupled learning, with a random projection matrix (as common in standard CS). The difference between CR and CL reduces as the noise level increases. Nevertheless, for noise levels below 20%, the MSE of CL is at least 30% lower than CR for $\alpha > 1/4$. Fig. 3(b) shows the retrieval MSE of CL relative to UL. Here, the advantage of CL over UL is lower than in the previous case [Fig. 3(a)], which indicates that a well-designed projection matrix as introduced in Section II, learned from the dictionary Ψ , can do better than simply coupling the data using a random sampling matrix (see also Fig. 2). However, CL can still reduce the MSE with respect to UL as much as 20% (a reduction of ~ 1 Db), justifying its use.

From Figs. 2 and 3, it is also clear that the retrieval MSE and CL/CR and CL/UL ratios have a minimum at about $\alpha = 1/32$. We comment more on this at the end of this section.

Figs. 4 and 5 show the retrieval MSE and the CL/CR, CL/UL ratios for $K = 64$, i.e., for a dictionary that is also a basis for the vector space of image patches ($K = n$). This is an interesting experiment, since the dictionary now is not overcomplete, being more in agreement with the majority of the theoretical results from the CS framework. These figures show that the proposed

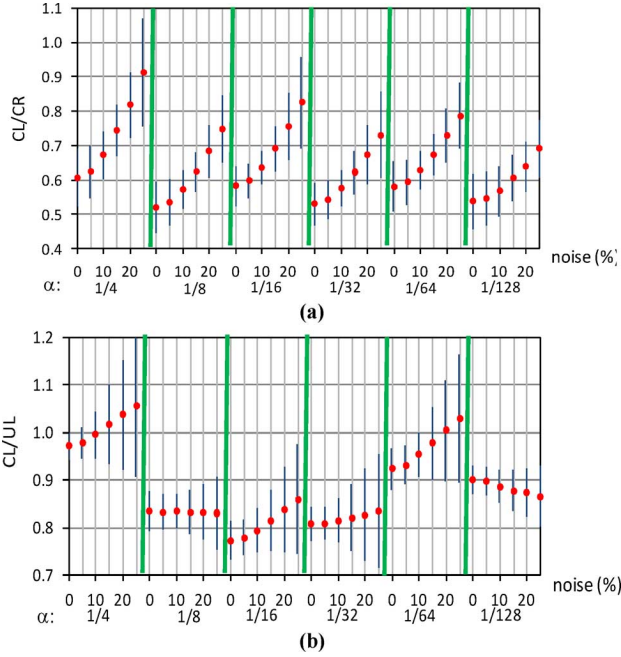


Fig. 3. For $S = 6$, $m = 12$, $n = 64$, and $K = 256$, different noise levels and values of α , (a) ratio between the retrieval MSE for CL and the retrieval MSE for CR, (b) ratio between the retrieval MSE for CL and the retrieval MSE for UL.

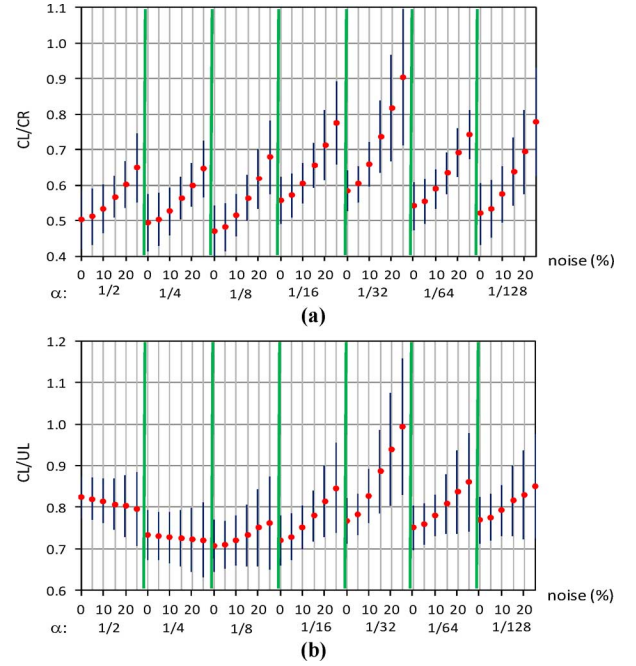


Fig. 5. For $S = 6$, $m = 12$, $n = 64$, and $K = 64$, different noise levels and values of α , (a) ratio between the MSE of retrieval from projections for CL and MSE of retrieval for CR, (b) ratio between the MSE of retrieval from projections for CL and the MSE of retrieval for UL.

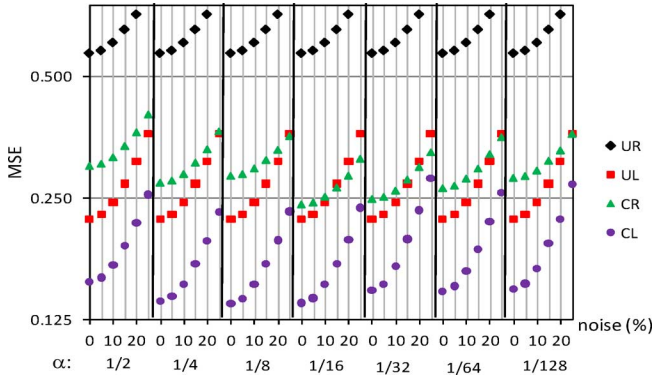


Fig. 4. MSE of retrieval from projections using the four training strategies at different noise levels and values of α , for $S = 6$, $m = 12$, $n = 64$, and $K = 64$.

framework is also valid within this scenario, and as can be appreciated in these figures, the CL/CR and CL/UL ratios are even better than for the overcomplete case (Figs. 2 and 3), indicating that the proposed coupled learning of both Ψ and Φ can be even more critical for nonovercomplete dictionaries.

Due to space limitations, we do not show here the results of all our experiments with all possible parameter variations (again, see supplementary material). However, a set of representative results is shown in Table II, which indicates the best values of α that produced the minimum retrieval MSE and at the same time the best CL/CR and CL/UL ratios, for a representative noise level of 5%. In general, for overcomplete dictionaries with $K = 4n$, we found that the best values for α are $1/32$, $1/16$, and $1/8$ for $m = 2S$, $2.5S$, and $3S$, respectively. This indicates that as the number of samples is reduced, \mathbf{Y} must have greater importance

than \mathbf{X} in the optimization. However, as detailed before, \mathbf{X} has an important role in the coupling algorithm, to limit the number of possible solutions of the under-determined inversion problem with $m \ll n$. Given that in practice $m \geq 2S$, then we must have $\alpha > 0$.

Fig. 6 shows one testing image consisting of nonoverlapping 8×8 patches reconstructed from their noisy projections (5% level of noise) \mathbf{Y} as $\hat{\mathbf{X}} = \Psi\Theta$, where $\Theta = \text{OMP}(\mathbf{D}, \mathbf{Y})$ and Ψ , Φ are obtained using the UR, UL, CR, or CL training strategies. The worst reconstruction case [Fig. 6(a)] is obtained when Ψ is learned using classical KSVD and Φ is simply a random sampling matrix (UR, standard CS scenario), followed by coupled-KSVD using a random sampling matrix [CR, Fig. 6(c)]. Using a well-designed sampling matrix $\Phi(\Psi)$ (UL, Section II) produces a good looking reconstruction of patches from their noisy projections [Fig. 6(b)]. Finally, using the coupled-KSVD (CL, Section III), an even better reconstruction from the noisy projections is obtained [Fig. 6(d)]. The better quality of Fig. 6(d) over Fig. 6(b) can be appreciated by the reduction of the artifacts, especially around the sharp edges on the top of the castle. Additional examples for the UR (standard CS framework) and the proposed CL technique are provided in Fig. 7 (additional details on these figures, including the CR and UL cases, are included with the supplementary material).

We have just presented the improved performance in terms of MSE and visual quality of the reconstructed patches, when using a well designed sampling matrix instead of random projection matrices, and also when exploiting coupled over uncoupled learning. Let us now compare the dictionaries learned from classic KSVD versus coupled-KSVD with a random matrix and with a learned sampling matrix. We also want to compare the

TABLE II
SAMPLE RESULTS FOR THE BEST VALUES OF α FOR A NOISE LEVEL OF 5%

K	S	m	α	MSE				CL/CR		CL/UL	
				UR	UL	CR	CL	mean	std	mean	std
256	4	8	1/32	0.5306	0.2154	0.3872	0.1890	0.4823	0.0880	0.8805	0.0317
		10	1/16	0.4980	0.1981	0.3113	0.1820	0.5746	0.0633	0.9215	0.0285
		12	1/8	0.4754	0.1838	0.2605	0.1731	0.6578	0.0451	0.9402	0.0266
	5	10	1/32	0.5029	0.1926	0.2827	0.1646	0.5720	0.0622	0.8497	0.0308
		12	1/16	0.4781	0.1782	0.2708	0.1614	0.5888	0.0626	0.9193	0.0235
		15	1/8	0.3951	0.1626	0.2303	0.1456	0.6267	0.0381	0.9092	0.0292
	6	12	1/32	0.5094	0.1848	0.2659	0.1465	0.5436	0.0569	0.8090	0.0354
		15	1/16	0.4203	0.1633	0.2254	0.1375	0.6016	0.0452	0.8399	0.0273
		18	1/8	0.3974	0.1477	0.2033	0.1287	0.6294	0.0261	0.8570	0.0266
64	6	12	1/8	0.5801	0.2275	0.2865	0.1409	0.4814	0.0661	0.7107	0.0569
		15	1/4	0.4601	0.1929	0.2241	0.1241	0.5483	0.0499	0.7357	0.0521
		18	1/2	0.3994	0.1838	0.2210	0.1243	0.5568	0.0400	0.8026	0.0314

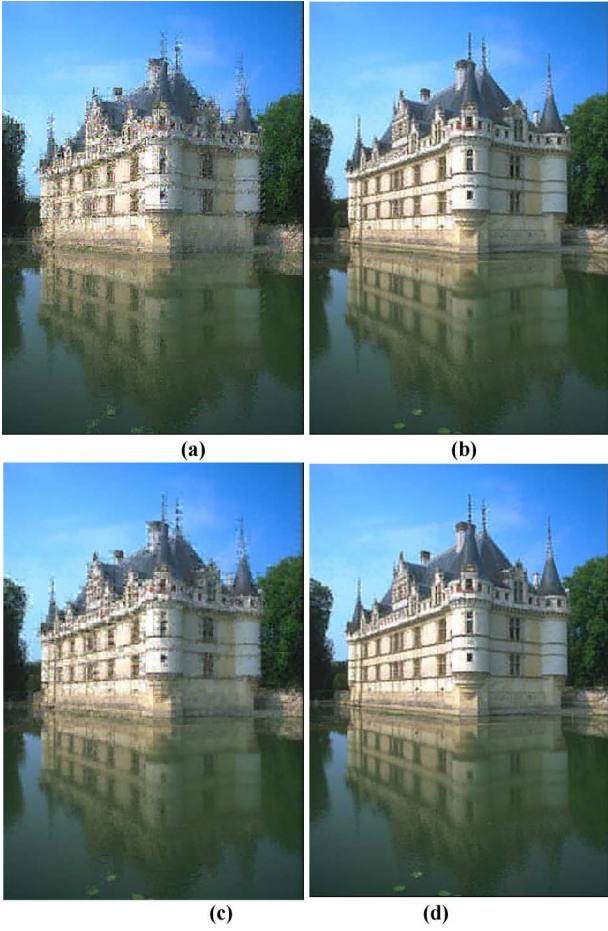


Fig. 6. One test image reconstructed from projected patches, without overlapping, with a noise level of 5%, using (a) uncoupled random, (b) uncoupled learning, (c) coupled random, and (d) coupled learning strategies to learn the dictionary and the projection matrix from the training patches. The retrieval MSE for these images is a) 1.1528, b) 0.4548, c) 0.6721, and d) 0.3769.

random projection matrices versus the sampling matrices computed using the proposed uncoupled and coupled learning techniques. Let us start by comparing the equivalent dictionaries $\mathbf{D} = \Phi\Psi$ for the four strategies considered here, in terms of the closeness of the Gramm matrix $\mathbf{G} = \tilde{\mathbf{D}}^T\tilde{\mathbf{D}}$ to the identity (as inspired by the RIP, see Section II).

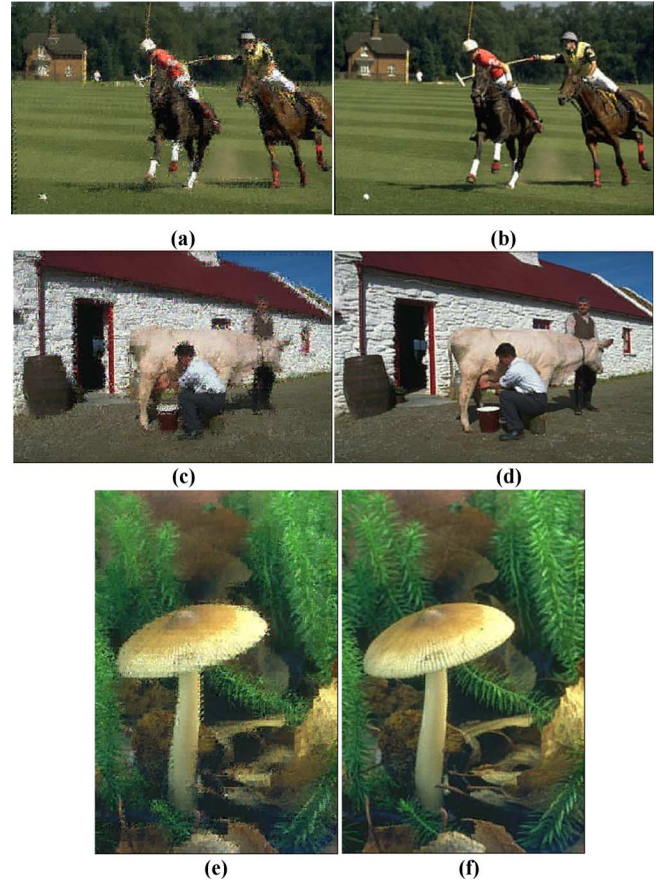


Fig. 7. Additional examples of image recovery. Note the sharp improvement with our proposed CL framework (right images) when compared to the more classical UR scenario from CS (left images). The retrieval MSE for these images is (a) 1.0539, (b) 2.3455, (c) 0.8207, (d) 0.2389, (e) 0.6707, and (f) 0.2204. While here we sample at twice the sparsity rate, even sampling at four times the sparsity, the UR results are far from the CL ones at just twice the sparsity, both in visual quality and MSE (e.g., the MSE for the image in b) becomes 1.1707, while better than the 2.3455, still more than double the 0.6707 MSE obtained with the proposed approach at half the sampling rate).

Fig. 8 shows the distribution of the off-diagonal elements of the Gramm matrix for each one of the four strategies. The Gramm matrix of CR is closer to the identity than the Gramm matrix of UR, and the Gramm matrix of both UL and CL are closer to the identity than UR and UL respectively, but they

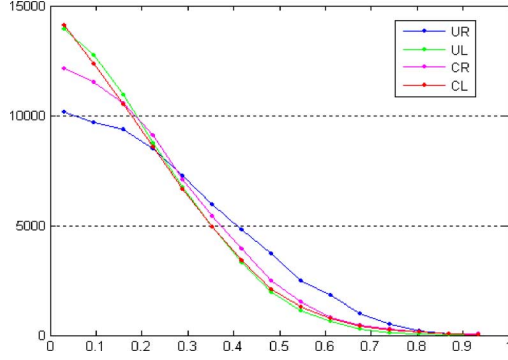


Fig. 8. Distribution of the off-diagonal elements of the Gramm matrix obtained using UR, UL, CR, and CL strategies.

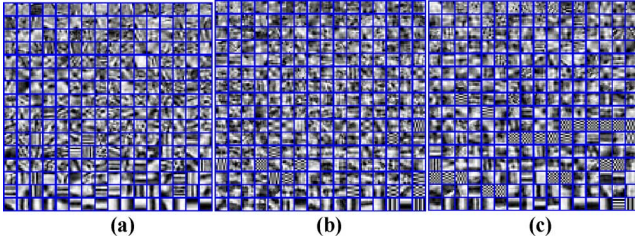


Fig. 9. Visual representation of the learned dictionaries for the different learning strategies.

are almost undistinguishable among themselves in terms of the distribution of the off-diagonal elements. These results are in agreement with the improved performance observed for the UL and CL methods over the other two possible strategies that use random sampling matrices.

Fig. 9 shows the $K = 256$ atoms of the dictionaries learned with KSVD (used for both UR and UL), coupled-KSVD with a random sampling matrix (CR), and coupled KSVD with a learned sampling matrix (CL), respectively. This is presented here as $K = 16 \times 16$ images of size 8×8 . We clearly see that the learned dictionaries are different.

In order to compare the sampling matrices Φ for the different learning strategies, we used the same approach proposed by [41], i.e., good sampling matrices should produce signals that are as spread as possible in \mathbb{R}^m . As in the work of Weiss *et al.* [41], we have also produced scatter plots of the rows of $\mathbf{Y} = \Phi\mathbf{X}$ among themselves (for random, uncoupled following Section II, and coupled following Section III, respectively). We observed that UL produces samples \mathbf{Y} with a larger spread in \mathbb{R}^m than UR, and in turn CL produces samples that are more spread in \mathbb{R}^m than UL (figures omitted due to space limitation).

In addition to making the Gramm matrix closer to the identity, the learned dictionaries should also be able to learn new patterns present only in the projected signals (and not in the signal itself), and coupled-KSVD helps to introduce those new patterns. In addition, a well-designed sampling matrix can improve the spread of the projected signals, when compared to a random sampling matrix, which in turn improves the retrieval of the original signal from projections, thanks to the larger separability of those signals in \mathbb{R}^m . Intuitively, Φ should maximize the mutual information $I(\mathbf{X}, \mathbf{Y})$ between the signal \mathbf{X} and its noisy projection \mathbf{Y} ,

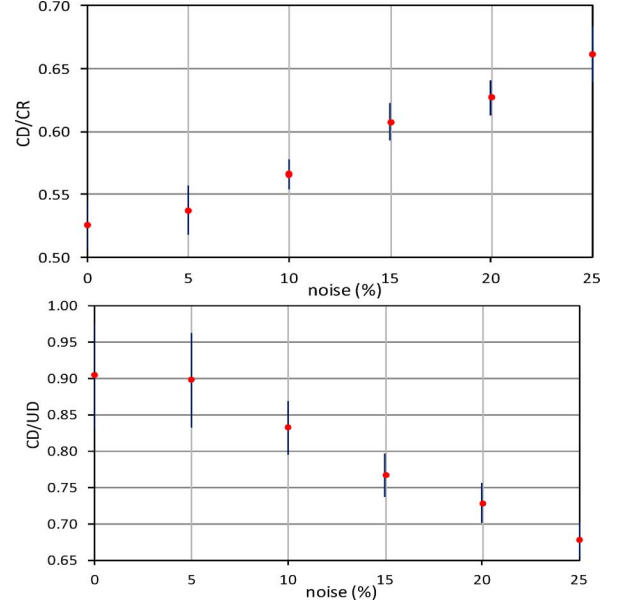


Fig. 10. Sample MSE results for tiny images (top CL/CR, bottom CL/UL), which are consistent with those reported before for patches ($S = 125$, $m = 250$, $\alpha = 1/32$).

which is equivalent to maximize the entropy $H(\mathbf{Y})$ of the output in order to minimize the retrieval error of \mathbf{X} from the noisy projections \mathbf{Y} [41], [52].

V. CONCLUDING REMARKS

A computational framework for learning an optimal sensing matrix for a given sparsifying dictionary was introduced in this paper. This was complemented by a novel approach to simultaneously learning the sensing matrix and sparsifying dictionary from an image database. We showed that such learning leads to significantly improved reconstruction results when compared with more classical compressed sensing scenarios where random sensing matrices are used. The same framework can be used to learn the sparsifying dictionary while keeping the sensing matrix fix (see also [53]).

As mentioned in the introduction, the theoretical results for CS support the use of ℓ_1 optimization, while KSVD-type of algorithms have traditionally been based on OMP (for which the results are weaker). It is important to further improve the results here presented using ℓ_1 -based optimization approaches (see [54]).

The framework here developed is based on image patches, as commonly exploited in image processing. While in principle we could work with entire images, this is computationally unfeasible. For tiny 32×32 images, following [55], we obtain results consistent with the work reported above for the patches, see Fig. 10. Of course, images are much larger than this, and algorithms of the type of KSVD as here developed, or basically any dictionary learning approach, are impractical. In addition, following once again the state of the art results for image enhancement via KSVD, we should work with overlapping patches (e.g., 8×8 or multiscale up to 20×20 , see [56]). Ideally, we would like then to have the dictionary acting on all the overlapping $n \times n$ patches, with a unique sensing matrix globally acting

on the $N \times N$, $n \ll N$, image. This will also permit to naturally include the multiscale framework developed in [56]. Results in this direction, as well as in the adaptation of the sensing to the task following [25], and the incorporation of constraints in the learned dictionary to speed-up the reconstruction, will be reported elsewhere.

ACKNOWLEDGMENT

The authors would like to thank J. Mairal for providing his very efficient code implementing KSVD in C++, which served as our basis to develop the code used in our experiments. They would also like to thank Prof. M. Elad for making publicly available some useful code in Matlab [57], as well as the excellent reviewers that helped to improve the presentation and to clarify a number of points in the paper.

REFERENCES

- [1] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 5, no. 2, pp. 489–509, Feb. 2006.
- [2] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Inf. Theory*, vol. 51, no. 12, pp. 4203–4215, Dec. 2005.
- [3] E. J. Candès, "Compressive sampling," in *Proc. Int. Congr. Math.*, Madrid, Spain, 2006, pp. 1433–1452.
- [4] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, Mar. 2008.
- [5] J. Romberg, "Imaging via compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 14–20, Mar. 2008.
- [6] J. Haupt and R. Nowak, "Compressive sampling vs. conventional imaging," in *IEEE Int. Conf. Image Processing*, Atlanta, GA, 2006, pp. 1269–1272.
- [7] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.
- [8] *Compressive Sensing Resources* [Online]. Available: <http://www.dsp.ece.rice.edu/cs/>
- [9] Compressive Imaging: A New Single Pixel Camera, Compressive Sensing Group [Online]. Available: <http://www.dsp.ece.rice.edu/cs/cscamera/>
- [10] M. F. Duarte, M. A. Davenport, D. Takbar, J. N. Laska, T. Sun, K. F. Kelly, and R. G. Baraniuk, "Single-pixel imaging via compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, Mar. 2008.
- [11] A. Cohen, W. Dahmen, and R. DeVore, *Compressed Sensing and Best k-Term Approximation*, 2006, to be published.
- [12] J. Mairal, M. Elad, and G. Sapiro, "Sparse representation for color image restoration," *IEEE Trans. Image Process.*, vol. 17, no. 1, pp. 53–69, Jan. 2008.
- [13] S. Mallat and E. Le Pennec, "Sparse geometric image representation with bandelets," *IEEE Trans. Image Process.*, vol. 14, no. 4, pp. 423–438, Apr. 2005.
- [14] Y. Weiss and W. T. Freeman, "What makes a good model of natural images?," in *Proc. IEEE Computer Vision and Pattern Recognition*, Minneapolis, MN, 2007, pp. 1–8.
- [15] K. Engan, S. O. Aase, and J. H. Husøy, "Frame based signal compression using method of optimal directions (MOD)," in *Proc. IEEE Int. Symp. Circ. Syst.*, Orlando, FL, 1999, vol. 4, pp. 1–4.
- [16] G. Peyre, Sparse Modeling of Textures 2007–15, Preprint Ceremade [Online]. Available: <http://www.ceremade.dauphine.fr/~peyre/publications/07-Preprint-Peyre-SparseModelingTextures.pdf>
- [17] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?," *Vis. Res.*, vol. 37, no. 23, pp. 3311–3325, Dec. 1997.
- [18] B. D. Rao, "Signal processing with the sparseness constraints," in *Proc. IEEE ICASSP Conf.*, 1998, vol. 3, pp. 1861–1864.
- [19] M. Lewicki and T. Sejnowski, "Learning overcomplete representations," *Neural Comput.*, vol. 12, no. 2, pp. 337–365, 2000.
- [20] K. Kreutz-Delgado, J. F. Murray, B. D. Rao, K. Engan, T.-W. Lee, and T. Sejnowski, "Dictionary learning algorithms for sparse representation," *Neural Comput.*, vol. 15, pp. 349–396, 2003.
- [21] M. Elad and M. Aharon, "Image denoising via learned dictionaries and sparse representation," *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3736–3745, Dec. 2006.
- [22] S. Mallat, *A Wavelet Tour of Signal Processing*, 2nd ed. New York: Academic, 1999.
- [23] M. Do and M. Vetterli, "Framing pyramids," *IEEE Trans. Image Process.*, vol. 51, no. 9, pp. 2329–2342, Sep. 2003.
- [24] M. W. Marcellin, M. J. Gormish, A. Bilgin, and M. P. Boliek, "An overview of JPEG-2000," in *Proc. Data Compression Conf.*, Snowbird, UT, 2000, pp. 523–544.
- [25] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Discriminative learned dictionaries for local image analysis," presented at the IEEE Computer Vision Pattern Recognition, Anchorage, AK, Jun. 2008.
- [26] M. Ranzato, F. Huang, Y. Boureau, and Y. LeCun, "Unsupervised learning of invariant feature hierarchies with applications to object recognition," in *Proc. IEEE Computer Vis. Pattern Recog.*, New York, 2007, pp. 1–8.
- [27] J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.
- [28] A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: Transfer learning from unlabeled data," in *Proc. 24th Int. Conf. Machine Learning*, Corvallis, OR, 2007, pp. 759–766.
- [29] D. L. Donoho and M. Elad, "Optimally sparse representation in general (non-orthogonal) dictionaries via ℓ_1 minimization," *Proc. Nat. Acad. Sci.*, vol. 100, no. 5, pp. 2197–2202, 2003.
- [30] R. Gribonval and M. Nielsen, "Sparse representations in unions of bases," *IEEE Trans. Inf. Theory*, vol. 49, no. 12, pp. 3320–3325, Dec. 2003.
- [31] S. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Trans. Signal Process.*, vol. 41, no. 12, pp. 3397–3415, Dec. 1993.
- [32] Y. C. Pati, R. Rezaifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," in *Proc. 27th Annu. Asilomar Conf. Signals, Systems, and Computers*, Los Alamitos, CA, 1993, vol. 1, pp. 40–44.
- [33] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Inf. Theory*, vol. 50, no. 10, pp. 2231–2242, Oct. 2004.
- [34] 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, 1998.
- [35] J. A. Tropp and A. C. Gilbert, "Signal recovery from partial information via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.
- [36] F. Murray and K. Kreutz-Delgado, "Sparse image coding using learned overcomplete dictionaries," in *Proc. IEEE Machine Learning for Signal*, Sao Luis, Brazil, 2004, pp. 579–588.
- [37] H. Rauhut, K. Schnass, and P. Vanderghynst, "Compressed sensing and redundant dictionaries," *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp. 2210–2219, May 2008.
- [38] D. Needell and R. Vershynin, "Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit," *Foundations Computational Mathematics*, to be published.
- [39] O. Bryt and M. Elad, "Compression of facial images using the K-SVD algorithm," *J. Vis. Commun. Image Represent.*, vol. 19, no. 4, pp. 270–283, May 2008.
- [40] M. Elad, "Optimized projections for compressed sensing," *IEEE Trans. Signal Process.*, vol. 55, no. 12, pp. 5695–5702, Dec. 2007.
- [41] Y. Weiss, H. S. Chang, and W. T. Freeman, "Learning compressed sensing," presented at the Snowbird Learning Workshop, Allerton, CA, 2007.
- [42] L. Applebaum, S. Howard, S. Searle, and R. Calderbank, Chirp Sensing Codes: Deterministic Compressed Sensing Measurements for Fast Recovery 2008 [Online]. Available: <http://www.dsp.ece.rice.edu/cs/>, to be published
- [43] R. A. DeVore, "Deterministic constructions of compressed sensing matrices," *J. Complexity*, vol. 23, no. 4–6, pp. 918–925, Aug. 2007.
- [44] V. Chandar, A Negative Result Concerning Explicit Matrices with the Restricted Isometry Property 2008 [Online]. Available: <http://www.dsp.ece.rice.edu/cs/>, to be published
- [45] M. Aharon, M. Elad, and A. Bruckstein, "The KSVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.
- [46] E. P. Simoncelli, "Statistical models for images: Compression restoration and synthesis," in *Proc. Conf. Signals, Syst., Comput.*, Asilomar, CA, 1997, pp. 673–678.

- [47] D. Mumford and B. Gidas, "Stochastic models for generic images," *Quart. Appl. Math.*, vol. 54, no. 1, pp. 85–111, Mar. 2001.
- [48] A. Srivastava, A. B. Lee, E. P. Simoncelli, and S.-C. Zhu, "On advances in statistical modeling of natural images," *J. Math. Imag. Vis.*, vol. 18, no. 1, pp. 17–33, Jan. 2003.
- [49] J. Huang and D. Mumford, "Statistics of natural images and models," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Fort Collins, CO, 1999, vol. 1, pp. 1541–1547.
- [50] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. IEEE 8th Int. Conf. Computer Vision*, Vancouver, BC, Canada, 2001, vol. 2, pp. 416–423.
- [51] B. Russell, A. Torralba, and W. T. Freeman, Labelme, The Open Annotation Tool MIT, Computer Science and Artificial Intelligence Laboratory [Online]. Available: <http://labelme.csail.mit.edu/>
- [52] J. F. Cardoso, "Infomax and maximum likelihood for blind source separation," *IEEE Signal Processing Lett.*, vol. 4, pp. 112–114, Apr. 1997.
- [53] G. Peyré, Best Basis Compressed Sensing 2007–20, Preprint Ceremade [Online]. Available: <http://www.ceremade.dauphine.fr/~peyre/publications/07-Preprint-Peyre-BestBasisCS.pdf>
- [54] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Supervised dictionary learning," presented at the NIPS Vancouver, BC, Canada, Dec. 2008.
- [55] A. Torralba, R. Fergus, and W. T. Freeman, 80 Million Tiny Images: A Large Data Set for Non-Parametric Object and Scene Recognition Computer Science and Artificial Intelligence Lab, Mass. Inst. Technol., Cambridge, MA, Rep. MIT-CSAIL-TR-2007-024.
- [56] J. Mairal, G. Sapiro, and M. Elad, "Learning multiscale sparse representations for image and video restoration," *SIAM Multiscale Model. Simul.*, vol. 7, no. 1, pp. 214–241, Apr. 2008.
- [57] M. Elad, Personal Page [Online]. Available: <http://www.cs.technion.ac.il/~elad/software/>



Julio Martin Duarte-Carvajalino (S'07–M'07) received the B.S.E.E. degree (cum laude) from the Universidad Industrial de Santander (UIS), Colombia, in 1995, the M.Sc. degree in electric engineering from the University of Puerto Rico at Mayaguez (UPRM), in 2003, and the Ph.D. degree in computing and information sciences and engineering from the UPRM in December 2007.

He was an associated Postdoctoral at the University of Minnesota from January to July 2008, where he worked on optimization methods to learn sparse

signals in Compressed Sensing, under the supervision of Dr. G. Sapiro. Since August 2008, he has been a color scientist at the Kodak Graphic Communications Company, Oakdale, MN. His research interests are in computer vision and color science.



Guillermo Sapiro (M'94–SM'03) was born in Montevideo, Uruguay, on April 3, 1966. He received the B.Sc. (summa cum laude), M.Sc., and Ph.D. degrees from the Department of Electrical Engineering, The Technion—Israel Institute of Technology, Haifa, in 1989, 1991, and 1993 respectively.

After postdoctoral research at MIT, Cambridge, MA, he became a Member of Technical Staff at the research facilities of HP Labs in Palo Alto, CA. He is currently with the Department of Electrical and Computer Engineering, University of Minnesota,

Minneapolis, where he holds the position of Distinguished McKnight University Professor and Vincentine Hermes-Luh Chair in Electrical and Computer Engineering. He works on differential geometry and geometric partial differential equations, both in theory and applications in computer vision, computer graphics, medical imaging, and image analysis. He co-edited a special issue of the IEEE TRANSACTIONS ON IMAGE PROCESSING on this topic and a second one in the *Journal of Visual Communication and Image Representation*. He has authored and coauthored numerous papers in this area and has written a book published by Cambridge University Press, January 2001.

Dr. Sapiro was awarded the Gutwirth Scholarship for Special Excellence in Graduate Studies in 1991, the Ollendorff Fellowship for Excellence in Vision and Image Understanding Work in 1992, the Rothschild Fellowship for Post-Doctoral Studies in 1993, the Office of Naval Research Young Investigator Award in 1998, the Presidential Early Career Awards for Scientist and Engineers (PECASE) in 1998, and the National Science Foundation Career Award in 1999. He is a member of SIAM.

G. Sapiro is the funding Editor-in-Chief of the SIAM Journal on Imaging Sciences.