On Collaborative Compressive Sensing Systems:
The Framework, Design and Algorithm

Zhihui Zhu, Gang Li, Jiajun Ding, Qiuwei Li, and Xiongxiong He *

September 21, 2017

Abstract

We propose a collaborative compressive sensing (CCS) framework consisting of a bank of \( K \) compressive sensing (CS) systems that share the same sensing matrix but have different sparsifying dictionaries. This CCS system is guaranteed to yield better performance than each individual CS system in a statistical sense, while with the parallel computing strategy, it requires the same time as that needed for each individual CS system to conduct compression and signal recovery. We then provide an approach to designing optimal CCS systems by utilizing a measure that involves both the sensing matrix and dictionaries and hence allows us to simultaneously optimize the sensing matrix and all the \( K \) dictionaries under the same scheme. An alternating minimization-based algorithm is derived for solving the corresponding optimal design problem. We provide a rigorous convergence analysis to show that the proposed algorithm is convergent. Experiments with real images are carried out and show that the proposed CCS system significantly improves on existing CS systems in terms of the signal recovery accuracy.

1 Introduction

Compressive (or compressed) sensing (CS), aiming to sample signals beyond Shannon-Nyquist limit \([13, 14, 16, 19]\), is a mathematical framework that efficiently compresses a signal vector \( x \in \mathbb{R}^N \) by a measurement vector \( y \in \mathbb{R}^M \) of the form

\[
y = \Phi x
\]

where \( \Phi \in \mathbb{R}^{M \times N} \) is a carefully chosen sensing matrix capturing the information contained in the signal vector \( x \).

As \( M \ll N \), we have to exploit additional constraints or structures on the signal vector \( x \) in order to recover it from the measurement \( y \) and sparsity is such a structure. In CS, it assumed that the original signal \( x \) can be expressed as a linear combination of few elements/atoms from a \( l_2 \)-normalized set \( \{\psi_\ell\} \), i.e., \( \|\psi_\ell\|_2 = 1, \ \forall \ell \):

\[
x = \sum_{\ell=1}^{L} \psi_\ell s_\ell = \Psi s
\]

where \( \Psi := [\psi_1 \ \psi_2 \ \cdots \ \psi_L] \in \mathbb{R}^{N \times L} \) is called the dictionary and the entries of \( s \in \mathbb{R}^L \) are referred to as coefficients. \( x \) is said to be \( \kappa \)-sparse in \( \Psi \) if \( \|s\|_0 = \kappa \), where \( \|s\|_0 \) is the number of non-zero elements of vector \( s \). Under the framework of compressed sensing, the original signal \( x \) can be recovered from \( x = \Psi \hat{s} \) with \( \hat{s} \) the solution of the following problem when the sparsity level \( \kappa \) is small:

\[
y = As \quad \text{s.t.} \quad \|s\|_0 \leq \kappa
\]

*This work was supported by the Grants of NSFCs 61273195, 61304124, & 61473262. Zhihui Zhu and Qiuwe Li (e-mail: {zzhu,qiuli}@mines.edu) are with the Department of Electrical Engineering at the Colorado School of Mines, 1500 Illinois St., Golden, CO 80401, USA. Gang Li (e-mail: ieligang@zjut.edu.cn) is with the College of Automation & Electrical Engineering, Zhejiang University of Science & Technology, Hangzhou, 310023, Zhejiang, P.R. China. Jiajun Ding and Xiongxiong He are with the College of Information Engineering, Zhejiang University of Technology, Hangzhou, 310014, Zhejiang, P.R. China.
\( A := \Phi\Psi \) is called equivalent dictionary. A CS system is referred to equations (1) and (2) plus an algorithm used to solve (3) and its ultimate goal is to reconstruct the original signal \( x \) from the low-dimensional measurement \( y \). The latter depends strongly on the properties of \( \Psi \) and \( \Phi \).

In general, the choice of dictionary \( \Psi \) depends on the signal model. Based on whether the dictionary is given in a closed form or learned from data, the previous works on designing dictionary can be roughly classified into two categories. In the first one, one attempts to concisely capture the structure contained in the signals of interest by a well-designed dictionary, like the wavelet dictionary \( [30] \) for piecewise regular signals, the Fourier matrix for frequency-sparse signals, and a multiband modulated Discrete Prolate Spheroidal Sequences (DPSS’s) dictionary for sampled multiband signals \( [12] \). The second category is to learn the dictionary in an adaptive way to sparsely represent a set of representative signals (called training data). Typical algorithms for solving this sparsifying dictionary learning problem include the method of optimal directions (MOD) \( [24] \), K-singular value decomposition (K-SVD) \( [3] \) based algorithms, and the method for designing incoherent sparsifying dictionary \( [25] \). Learning a sparsifying dictionary has proved to be extremely useful and has achieved state-of-the-art performance in many signal and image processing applications, like image denoising, inpainting, deblurring and compression \( [2,23] \).

To recover the signal \( x \) from its low dimensional measurement \( y \), another important factor in a CS system is to select an appropriate sensing matrix \( \Phi \) that preserves the useful information contained in \( x \). It has been shown that a sparse signal \( x \) can be exactly reconstructed from its measurement \( y \) with greedy algorithms such as those orthogonal matching pursuit (OMP)-based ones \( [11,31,32,34,37] \) or methods based on convex optimization \( [13,17,20] \), if the equivalent dictionary \( A = \Phi\Psi \) satisfies the restricted isometry property (RIP) \( [9,13,14] \), or the sensing matrix \( \Phi \) satisfies the so-called \( \Psi \)-RIP \( [15] \). However, despite the fact that a random matrix \( A \) (or \( \Phi \)) with a specific distribution satisfies the RIP (or \( \Psi \)-RIP) with high probability \( [9] \), it is hard to certify the RIP (or \( \Psi \)-RIP) for a given sensing matrix that is utilized in practical applications \( [4] \). Thus, Elad \( [22] \) proposed to design a sensing matrix via minimizing the mutual coherence, another property of the equivalent dictionary that is more tractable and much easier to verify. Since then, it has led to a class of approaches for designing CS systems \( [1,6,21,25,29,39,40] \). When the signal is not exactly sparse, which is true for most natural signals such as image signals even with a learned dictionary, it is observed that the sensing matrix obtained via only minimizing the mutual coherence yields poor performance. Recently, a modified approach considering the sparse representation error was proposed in \( [27] \) where the designed sensing matrix is robust against the sparse representation error and achieves state-of-the-art performance for image compression.

The ultimate objective of dictionary learning is to determine a \( \Psi \in \mathbb{R}^{N \times L} \) such that the signals of interest (like image patches extracted from natural images) can be well represented with a sparsity level \( \kappa \) given. To that end, a large \( L \) is preferred as increasing \( L \) can enrich the components/atoms that are used for sparse representation. One notes, however, that mutual coherence of \( \Psi \), formally defined as

\[
\mu(\Psi) = \max_{i \neq j} \frac{||\psi_i^T \psi_j||}{||\psi_i||_2 \cdot ||\psi_j||_2}
\]

does not satisfy \( 0 < \mu(\Psi) < 1 \) ([38])

\[
\sqrt{\frac{L - N}{N(L - 1)}} \leq \mu(\Psi) \leq 1
\]

which implies that when \( L \) is very large, the dictionary \( \Psi \) will have large mutual coherence. Since the equivalent dictionary \( A = \Phi\Psi \), where \( \Phi \) is of dimension \( M \times N \) with \( M \ll N \), \( \mu(A) \) gets in general larger than \( \mu(\Psi) \). So, increasing \( L \) may affect sparse signal recovery dramatically. Besides, a large scale training data set is required in order to learn richer atoms (or features) and thus it needs a long time for the existing algorithms to learn a dictionary, though it is not a big issue for off-line design.

These motivate us to develop an alternative CS framework that yields a high performance in terms of reconstruction accuracy as if a high dimensional dictionary were used in the traditional CS framework (see Fig. 1(a)), but gets rid of the drawbacks of a high dimensional dictionary as mentioned above. In particular, our main argument is that given different training data, we can learn different dictionaries containing different atoms or features for the signals of interest.

The main contributions of this paper are stated as follows:

- The first contribution is to propose a new framework named collaborative CS (CCS) system based on the maximum likelihood estimation (MLE) principle. Such a framework is actually a bank of \( K \) traditional CS systems that have an identical sensing matrix but different dictionaries of dimension
$N \times L_0$. As shown in Fig. 1(b), the estimate of the original signal provided by the proposed collaborative CS system is to fuse those obtained by all the individual CS systems and can be interpreted as a linear $\kappa K$-sparse representation in a huge dictionary of dimension $N \times (L_0K)$:

- The second contribution is to provide an approach to simultaneously designing the sensing matrix and sparsifying dictionaries for the proposed CCS system. Unlike [6,21], where the sensing matrix and dictionary are designed independently (though the sensing matrix is involved when the dictionary is updated), we use a measure that is a function of sensing matrix, dictionaries and sparse coefficients and hence allows us to consider both of the sensing matrix and dictionaries under the one and the same framework. By doing so, it is possible to enhance the system performance and to come up with an algorithm with guaranteed convergence for the optimal CCS design;

- The third contribution is to provide an alternating minimization-based algorithm for optimal design of the proposed CCS system. An algorithm is proposed for updating the dictionaries column by column. Despite the sparse coding stage, the problems involving sensing matrix updating and dictionary updating are solved in such a way that ensures that the cost function decreases at each iteration. It should be pointed out that the alternating minimization-based approach has been popularly used in designing sensing matrix and sparsifying dictionary [3,21,22,27,29], in which the convergence of the algorithms is usually neither ensured nor seriously considered. As one of the important results in this paper, a rigorous convergence analysis is provided and hence the proposed algorithm is ensured to be convergent.

Notations: Throughout this paper, finite-dimensional vectors and matrices are indicated by bold characters. The symbols $\mathbf{I}$ and $\mathbf{0}$ respectively represent the identity and zero matrices with appropriate sizes. In particular, $\mathbf{I}_N$ denotes the $N \times N$ identity matrix. $\mathcal{U}_N := \{ \mathbf{h} \in \mathbb{R}^N : \| \mathbf{h} \|_2 = 1 \}$ is the unit sphere. Also, $\mathcal{O}_{M,N} := \{ \mathbf{R} \in \mathbb{R}^{M \times N} : \mathbf{RR}^T = \mathbf{I}_M \}$ with $M \leq N$ and when $M = N$, $\mathcal{O}_{N,N}$ is simply denoted as $\mathcal{O}_N$. For any natural number $N$, we let $[N]$ denote the set $\{1,2,...,N\}$. We adopt MATLAB notations for matrix indexing; that is, for a matrix $\mathbf{Q}$, its $(i,j)$-th element is denoted by $\mathbf{Q}(i,j)$, its $i$-th row (or column) is denoted by $\mathbf{Q}(i,:)$ (or $\mathbf{Q}(:,i)$). When it is clear from the context, we also use $q_i$, $q_{i,j}$ to denote the $i$-th column and $(i,j)$-th element of $\mathbf{Q}$. Similarly, we use $q_i$ or $q(i)$ to denote the $i$-th element of the vector $\mathbf{q} \in \mathbb{R}^N$.

The outline of this paper is given as follows. In Section 2, based on the MLE principle it is shown that a sensing matrix $\Phi \in \mathcal{O}_{M,N}$ is preferred, when dealing with sparse representation errors, and a collaborative estimator is derived, which leads to the CCS scheme. A framework is proposed for designing an optimal CCS system and an algorithm is derived for finding it in Section 3. Section 4 is devoted to convergence analysis for the proposed algorithm. To demonstrate the performance of the proposed CCS systems, experiments are carried out in Section 5. The paper is concluded in Section 6.

2 A collaborative compressive sensing scheme

Consider the following more general signal model than (2):

$$\mathbf{x} = \Psi \mathbf{s} + \mathbf{e} \tag{5}$$

where $\mathbf{e}$ is the sparse representation noise (a.k.a signal noise). The measurement $\mathbf{y} = \Phi \mathbf{x}$ is then of form

$$\mathbf{y} = \mathbf{A} \mathbf{s} + \mathbf{v}$$

where $\mathbf{v} := \mathbf{y} - \mathbf{A} \mathbf{s} = \Phi \mathbf{e}$ and $\mathbf{A} = \Phi \Psi$, as defined before, is the equivalent dictionary of the CS system.

Suppose $\mathbf{e}$ is of normal distribution with a zero-mean and

$$\mathbb{E}[\mathbf{ee}^T] = \sigma^2_\mathbf{e} \mathbf{I}_N,$$

where $\mathbb{E}[\cdot]$ denotes the statistical averaging operator. Then, $\mathbf{v}$ has the multivariate normal distribution with $\mathcal{N}(\mathbf{0}, \sigma^2_\mathbf{e} \Phi \Phi^T)$; that is its probability density function (PDF) is given by

$$f_\mathbf{v}(\mathbf{\xi}) = \frac{e^{-\frac{1}{2} \mathbf{\xi}^T \Sigma_\mathbf{v}^{-1} \mathbf{\xi}}}{\sqrt{2\pi|\Sigma_\mathbf{v}|}} \tag{6}$$
where \(|\cdot|\) represents the determinant of a matrix and \(\Sigma_w := \sigma_w^2 \Phi \Phi^T\), where the sensing matrix \(\Phi\) is assumed of full row rank. According to the MLE principle [26] along with the sparsity assumption on \(s\), the best estimate \(\hat{s}\) of the sparse coefficient vector \(s\) using the measurement \(y\) is the one that maximizes the likelihood function in \(s\); that is \(f_\ell(\xi)\) with \(\xi = y - As\). This leads to

\[
\hat{s} \triangleq \arg \min_{s} \left\| (\Phi \Phi^T)^{-1/2} (y - As) \right\|_2^2
\]

\[
\text{s.t. } \|s\|_0 \leq \kappa
\]

(7)

which is different from [4], where the signal noise \(e\) is assumed nil.

2.1 Structure of sensing matrices

Note that the objective function in (4) can be rewritten as

\[
\| (\Phi \Phi^T)^{-1/2} (y - As) \|_2^2 = \| \bar{y} - \Phi \Psi s \|_2^2
\]

(8)

where \(\bar{y} := (\Phi \Phi^T)^{-1/2} y\) and \(\Phi := (\Phi \Phi^T)^{-1/2}\). This implies that the MLE estimator (7) is equivalent to solving the classical sparse recovery problem if we fold both the measurement \(y\) and the sensing matrix \(\Phi\) by \((\Phi \Phi^T)^{-1/2}\).

A matrix \(Q\) is said to be RIP of order \(\kappa\) if there exists a constant \(\gamma \in (0, 1)\) such that

\[
\sqrt{1 - \gamma} \leq \frac{\|Qs\|_2}{\|s\|_2} \leq \sqrt{1 + \gamma}
\]

(9)

for all \(\|s\|_0 \leq \kappa\). It is shown in [18, Lemma 4.1] that the matrix \(\Phi\) satisfies the RIP if so does the sensing matrix \(\Phi\). A generalization of the RIP is \(\Psi\)-RIP [15]: a matrix \(\Phi\) is said to be \(\Psi\)-RIP of order \(\kappa\) if there exists a constant \(\delta \in (0, 1)\) such that

\[
\sqrt{1 - \delta} \leq \frac{\|\Phi \Psi s\|_2}{\|\Psi s\|_2} \leq \sqrt{1 + \delta}
\]

(10)

holds for all \(s\) with \(\|s\|_0 \leq \kappa\).

We have the following interesting results.

**Lemma 1.** Let \((\Phi, \Psi)\) be a CS system. Suppose that \(\Psi\) is RIP of order \(\kappa\) with constant \(\gamma \in (0, 1)\) and that \(\Phi \in \mathbb{R}^{M \times N}\) satisfies the \(\Psi\)-RIP of order \(\kappa\) with constant \(\delta < 1\) and \(x = \Phi s\) with \(\|s\| \leq \kappa\). Denote \(\bar{\Phi} = (\Phi \Phi^T)^{-1/2} \Phi\), \(A = \Phi \Psi\) and \(\bar{A} = \Phi \Psi\). Then,

\[
\sqrt{(1 - \delta)(1 - \gamma)} \leq \frac{\|As\|_2}{\|s\|_2} \leq \sqrt{(1 + \delta)(1 + \gamma)}
\]

(11)

and

\[
\frac{\sqrt{(1 - \delta)(1 - \gamma)}}{\pi_{\text{max}}(\Phi)} \leq \frac{\|As\|_2}{\|s\|_2} \leq \frac{\sqrt{(1 + \delta)(1 + \gamma)}}{\pi_{\text{min}}(\Phi)}
\]

(12)

where \(\pi_{\text{max}}(\Phi)\) and \(\pi_{\text{min}}(\Phi)\) denotes the largest and smallest nonzero singular values of \(\Phi\) respectively.

**Proof of Lemma 1.** Since \(\Psi\) is RIP, (9) holds with \(Q = \Psi\), combining which with (10) we have (11) straightforward.

Let \(\bar{T} = T^{-1}\Phi\) for any non-singular matrix \(T\). Then, \(\Phi = T \bar{\Phi}\). It follows from matrix analysis that

\[
\frac{\|\bar{\Phi} x\|_2}{\|x\|_2} \leq \frac{\|T^{-1}\|_2 \|\Phi x\|_2}{\|x\|_2}
\]

\[
\frac{\|\Phi x\|_2}{\|x\|_2} \leq \frac{\|T\|_2 \|\bar{\Phi} x\|_2}{\|x\|_2}
\]

which implies \(\|T\|_2^{-1} \|\Phi x\|_2 \leq \|\bar{\Phi} x\|_2 \leq \|T\|_2 \|\Phi x\|_2\). With \(T = (\Phi \Phi^T)^{1/2}\) and \(\Phi\) satisfying \(\Psi\)-RIP, we have

\[
\frac{\sqrt{1 - \delta}}{\pi_{\text{max}}(\Phi)} \leq \frac{\|\Phi x\|_2}{\|x\|_2} \leq \frac{\sqrt{1 + \delta}}{\pi_{\text{min}}(\Phi)}
\]

(13)

Thus (12) follows from (13) and (11). This completes the proof.

\[\square\]
Remark 2.1:
- It was shown in [15] that a random matrix $\Phi$ with a specific distribution satisfies (10) with high probability. The 2nd part of this lemma implies that the equivalent dictionaries $A$ and $\overline{A}$ have RIP of order $\kappa$ if so is the dictionary $\Psi$.
- Both (11) and (12) imply that $\overline{\Phi}$ serves as an isometric operator similar to $\Phi$ for all $\kappa$-sparse signal, but the constants on the left and right hand sides of them depend on the condition number of $\Phi$. Specifically, for a given $\Psi$, $\overline{\Phi}$ makes $\overline{A}$ achieve the tightest RIP constants if $\Phi$ is a tight frame.
- More importantly, (7) and (8) imply that the CS system $(\Phi, \Psi)$ (with a sparse recovery procedure) is totally equivalent to the one $(\overline{\Phi}, \Psi)$ in terms of signal reconstruction accuracy. Note that for any full row rank sensing matrix $\Phi$, $\overline{\Phi}$ is a unit tight frame as
  $$\overline{\Phi}^T = (\Phi^T)^{-1/2} \Phi^T (\Phi^T)^{-1/2} = I_M$$

Also if the sensing matrix $\Phi$ has orthonormal rows (i.e., $\Phi^T \Phi = I_M$), then the folded sensing matrix $\overline{\Phi}$ is equivalent to $\Phi$.

Based on the equivalence mentioned in the 2nd point of Remark 2.1, it is assumed in the sequel that the sensing matrix under consideration is constrained with $\Phi^T \Phi = I_M$, that is
  $$\Phi \in O_{M,N}$$

2.2 Collaborative compressed sensing

Let us have a set of estimators, all used for estimating the identical signal $x \in \mathbb{R}^{N \times 1}$. Assume that the output of the $i$th estimator is given by
  $$x_i = x + e_i, \quad i = 1, \ldots, K$$
where $e_i$ is the estimation error of the $i$th estimator. A better estimate $\hat{x} = f(x_1, \ldots, x_K)$ can be obtained by choosing an appropriate function $f$ to fuse all the estimates $\{x_i\}$. In fact, Lemma 2 presented below yields such an estimate of $x$.

Lemma 2. Let $\{e_i\}$ be the set of the estimation errors defined in $e_i$ in [15]. Under the assumption that all the errors $\{e_i\}$ are statistically independent and that each $e_i$ obeys a normal distribution $\mathcal{N}(0, \epsilon_i^2 I_N)$ for all $i$, the best estimate $\hat{x}$ in the MLE sense of $x$ which can be achieved from the observations $\{x_i\}$ is given by
  $$\hat{x} = \sum_{i=1}^{K} \omega_i x_i$$
where $\omega_i := \epsilon_i^{-2} / \sum_{\ell=1}^{K} \epsilon_{\ell}^{-2}$, $\forall \ i$, and
  $$\mathbb{E}[\|\hat{x} - x\|^2] = \sum_{i=1}^{K} \omega_i^2 \mathbb{E}[\|x_i - x\|^2]$$

Proof of Lemma 2. Under the assumptions made in the lemma, the joint PDF of $(e_1, \ldots, e_i, \ldots, e_K)$ is given by $\prod_{i=1}^{K} f_{e_i}(\xi_i)$, where $f_{e_i}(\xi_i)$ is defined in (6) with $\Sigma_{e_i} = \epsilon_i^2 I_N$ and $\xi_i = x_i - x$ for all $i$. According to the MLE principle, the best estimate $\hat{x}$ of $x$ from $\{x_i\}$ is the one that maximizes this joint PDF. This can be achieved by setting the derivative of $\sum_{i=1}^{K} \frac{1}{2} \xi_i^T \Sigma_{e_i}^{-1} \xi_i$ w.r.t. $x$ to zero, which leads to (16).

Note that all $e_i = x_i - x$, $\forall \ i$ are statistically independent, (17) follows directly from (16) and the fact $\sum_{i=1}^{K} \omega_i = 1$. \qed

\footnote{We say $\Phi \in \mathbb{R}^{M \times N}$ a frame of $\mathbb{R}^M$ if there are positive constants $c_1$ and $c_2$ such that $0 < c_1 \leq c_2 < \infty$ and for any $y \in \mathbb{R}^M$, $c_1 \|y\|^2 \leq \|\Phi^T y\|^2 \leq c_2 \|y\|^2$; A frame is a tight frame if $c_1 = c_2 = c$, which implies $\Phi^T \Phi = cI_M$. A unit tight frame corresponds to $c = 1$.}
The estimator given by (16) is called a collaborative estimator for $x$. Now, suppose $E[\|e_i\|^2] = \epsilon_0^2$, $\forall i$ with $\epsilon_0^2$ a constant, independent of $i$. Then $\omega_i = \frac{1}{K}$, $\forall i$ and hence (16) becomes

$$\hat{x} = \frac{1}{K} \sum_{i=1}^{K} x_i \tag{18}$$

It is observed that by Cauchy-Schwartz inequality

$$\|\hat{x} - x\|^2 = \|\frac{1}{K} \sum_{i=1}^{K} (x_i - x)\|^2 \leq \frac{1}{K} \sum_{i=1}^{K} \|x_i - x\|^2$$

which implies that the reconstruction error variance of the collaborative estimator is smaller than the averaged variance of the bank of $K$ estimators. Furthermore, with $E[\|e_i\|^2] = \epsilon_0^2$, $\forall i$ (17) yields

$$E[\|\hat{x} - x\|^2] = \frac{\epsilon_0^2}{K}$$

This implies that the collaborative estimator yields a better estimate than a single estimator under the assumption that the estimation errors of all the individual estimators are statistically independent with an identical 2nd moment.

Now, we specify the $K$ estimators with a bank of $K$ CS systems $\{\Phi_i, \Psi_i\}$ and the corresponding collaborative estimator is then named collaborative compressive sensing - the CCS.

For the same measurement $y$ - the compressed version of $x$ via (1), the $i$th estimate $x_i$ of $x$ is given by the CS system $(\Phi, \Psi_i)$ with

$$x_i = \Psi_i \hat{s}_i, \forall i \tag{19}$$

where $\Psi_i \in \mathbb{R}^{N \times L_0}$, $\forall i$ all are designed using different training samples and $\hat{s}_i$ is the solution of (7) with $A = \Phi \Psi_i$. The whole procedure of the proposed CCS system is depicted in Fig. 1(b) for image compression.

**Remark 2.2:**

- Combining (18) and (19) yields

$$\hat{x} = \begin{bmatrix} \Psi_1 & \cdots & \Psi_K \end{bmatrix} \begin{bmatrix} \hat{s}_1/K \\ \vdots \\ \hat{s}_K/K \end{bmatrix} = \Psi_c s_c \tag{20}$$

which can be viewed as a linear sparse approximation of $x$ using a larger dictionary $\Psi_c$, having $KL_0$ atoms, with a sparsity level of $\kappa K$. A signal $x$ can be represented in many different ways. In fact, in the proposed CCS system $x$ is $\kappa$-sparsely represented in each of $K$ different dictionaries and the collaborative estimator (18) is equivalent to a representation in a high dimensional dictionary $\Psi_c \in \mathbb{R}^{N \times KL_0}$ and a constrained block sparse coefficient vector $s_c$. It is due to this equivalent high dimensional dictionary that enriches the signal representation ability and hence makes the proposed CCS scheme yield an improved performance;

- We finally note that the collaborative CS system has an identical compression stage as the classical CS system (as shown in Fig. 1). With respect to the signal recovery stage, in the collaborative CS system the estimators $\{x_i\}$ are independent to each other and can be obtained simultaneously by solving (19) in parallel.

3 Design of An Optimal CCS System

In this section, we propose an approach to designing optimal CCS systems by jointly optimizing the sensing matrix $\Phi$ and the bank of sparsifying dictionaries $\{\Psi_i\}$. An alternating minimization strategy is then utilized to solve the corresponding problem.
3.1 Framework for designing optimal CCS systems

Let 

\[ X = [X_1 \cdots X_i \cdots X_K] \]

be the matrix of training samples as defined before.

Our objective here is to jointly design the sensing matrix and dictionaries for CCS systems with training samples \( X \). To this end, a proper measure is needed.

First of all, it follows from (21) that

\[ X_i = \tilde{\Psi}_i \tilde{S}_i + E_i \]

where \( \tilde{\Psi}_i \) is the underlying dictionary and \( \| \tilde{S}_i(:,j) \|_0 \leq \kappa \). A widely utilized objective function for dictionary learning is

\[ g_1(\Psi_i, S_i, X_i) \triangleq \| X_i - \Psi_i S_i \|_F^2 \]

which represents the variance of the sparse representation error \( E_i \) and is utilized in state-of-the-art dictionary learning algorithms like MOD and K-SVD. Similarly, denote

\[ g_2(\Phi, \Psi_i, S_i, X_i) \triangleq \| \Phi(X_i - \Psi_i S_i) \|_F^2 \]

as the variance of the projected signal noise \( \Phi E_i \).

It follows from (21) that the measurements are of the form \( Y_i = \Phi X_i = \Phi \tilde{\Psi}_i \tilde{S}_i + \Phi E_i \). Thus besides reducing the projected signal noise variance \( g_2 \), the sensing matrix \( \Phi \) is expected to sense most of the key ingredients \( \tilde{\Psi}_i \tilde{S}_i \). This can be done by choosing the sensing matrix \( \Phi \) such that \( \| \Phi \tilde{\Psi}_i \tilde{S}_i \|_F^2 \) is maximized. As \( \Phi^T \Phi = I_M \) is assumed (see (14)), maximizing \( \| \Phi \Psi S_i \|_F^2 \) in terms of \( \Phi \) is equivalent to minimizing

\[ g_3(\Phi, \Psi_i, S_i) \triangleq \| (I - \Phi^T \Phi) \Psi S_i \|_F^2 \]

Figure 1: Illustration of (a) classical CS system, and (b) collaborative CS system. In this paper, we simply utilize \( f(x_1, \ldots, x_K) = \frac{1}{K} \sum_{i=1}^{K} x_i \).
Therefore, the following measure is proposed for designing optimal CCS systems:

\[
g(\Phi, \Psi, S, X_i) = g_1(\Psi, S, X_i) + \alpha g_2(\Phi, \Psi, S, X_i) \\
+ \beta g_3(\Phi, \Psi, S, X_i)
\] (25)

where \(g_1(\cdot), \ g_2(\cdot)\) and \(g_3(\cdot)\) are the measures defined in \([22] - [24]\), respectively, and \(\alpha\) and \(\beta\) are weighting factors to balance the importance of the three terms.

The problem of designing optimal CCS systems is then formulated as

\[
(\hat{\Phi}, \hat{\Psi}, \hat{S}) \triangleq \arg \min_{\Phi(\Psi, S, i)} \sum_{i=1}^{K} g(\Phi, \Psi, S, X_i)
\]

s.t. \(\Phi \in \mathcal{O}_{M,N}; \ \|S_i(\cdot;j)\|_0 \leq \kappa, \ \forall \ j\) \(\Psi_i(\cdot;l) \in \mathcal{U}_N, \ \forall \ l\)

(26)

in which both the sensing matrix \(\Phi\) and the bank of dictionaries \(\{\Psi_i\}\) are jointly optimized using the identical measure.

### 3.2 Algorithms for designing optimal CCS systems

We now propose an algorithm based on alternating minimization for solving (26). The basic idea is to construct a sequence \(\{\Phi_k, \{\Psi_{i,k}\}, \{S_i,k\}\}\) such that \(\sum_{i=1}^{K} g(\Phi_k, \Psi_{i,k}, S_i,k, X_i)\) is a descent sequence.

First of all, note that \(g(\Phi, \Psi, S, X_i)\) can be rewritten as

\[
g(\Phi, \Psi, S, X_i) = \|C(\Phi, X_i) - B(\Phi)\Psi S\|_F^2
\]

(27)

where

\[
C(\Phi, X_i) \triangleq \begin{bmatrix} X_i \\
\sqrt{\alpha} \Phi X_i \\
0 \end{bmatrix}, \ \ B(\Phi) \triangleq \begin{bmatrix} I_N \\
\sqrt{\alpha} \Phi \\\n\sqrt{\beta}(I_N - \Phi^T \Phi) \end{bmatrix}
\]

(28)

The proposed algorithm is then outlined as below:

**Algorithm 1** - To design an optimal CCS system

*Input*: the training data \(\{X_i\}\), the dimension of the projection space \(M\), the number of atoms \(L_0\), and the number of iterations \(N_{ite}\).

**Initialization**: set initial dictionaries\(^2\) \(\{\Psi_{i,0}\}\) and an initial sensing matrix\(^3\) \(\Phi_0\).

**Begin** \(k = 1, 2, \cdots, N_{ite}\)

- Update \(S_i\) with \(\Phi = \Phi_{k-1}, \Psi_i = \Psi_{i,k-1}, \ \forall i\): Note that (26) becomes

\[
S_{i,k} = \arg \min_S \|C(\Phi, X_i) - B(\Phi)\Psi S\|_F^2
\]

s.t. \(\|S(\cdot;j)\|_0 \leq \kappa, \ \forall \ j\)

(29)

which can be addressed using an OMP-based algorithm.

- Update \(\Psi_i\) with \(\Phi = \Phi_{k-1}, S_i = S_{i,k}, \ \forall i\): (26) leads to

\[
\Psi_{i,k} = \arg \min_{\psi_i \in \mathcal{U}_N, \forall l} \|C(\Phi, X_i) - B(\Phi)\Psi S\|_F^2
\]

\(i = 1, 2, \cdots, K\)

(30)

An algorithm will be given later for solving such a problem.

\(\text{2Each dictionary can be chosen as a DCT matrix or randomly selected from the data.}\)

\(\text{3It can be chosen a random one.}\)
• Update $\Phi$ with $\Psi_i = \Psi_i, S_i = S_i, \forall i$: The design problem (26) turns to

$$\Phi_k \triangleq \text{argmin}_{\Phi \in O_{M,N}} \sum_{i=1}^K \rho(\Phi, \Psi_i, S_i)$$ (31)

As to be seen later, this problem can be solved analytically.

End

Output: $\tilde{\Phi} = \Phi_{N_{ite}}$ and $\{\tilde{\Psi}_i = \Psi_{i, N_{ite}}\}$.

As seen from the outline of the algorithm proposed above, all the dictionaries $\{\Psi_i\}$ (and the coefficients $\{S_i\}$) can be updated concurrently. Thus, utilizing parallel computing strategy, it takes the same time to design a CCS system as that to design a traditional CS system.

The proposed Algorithm 1 consists of three minimization problems, specified by (29), (30) and (31), respectively. The first one can be addressed using an OMP-based algorithm and an iterative algorithm, denoted as Alg$\Phi$, for (30) is postponed to Section 3.4, while the solution to (31) is given by the following lemma:

**Lemma 3.** Define $\tilde{G} \triangleq \sum_{i=1}^K G(\Psi_i, S_i)$, where

$$G(\Psi_i, S_i) \triangleq \beta \Psi_i S_i (\Psi_i S_i)^T - \alpha (X_i - \Psi_i S_i)(X_i - \Psi_i S_i)^T$$ (32)

Let $\tilde{G} = V_G \Pi_G V_G^T$ be an eigen-decomposition (ED) of $N \times N$ symmetric $\tilde{G}$, where $\Pi$ is diagonal and its diagonal elements $\{\pi_n\}$ are assumed to satisfy $\pi_1 \geq \pi_2 \geq \cdots \geq \pi_N$. Then the solution to (31) is given by

$$\Phi_k = U \left[ \begin{array}{c} I_M \\ 0 \end{array} \right] V_G^T$$ (33)

with $U \in O_N$ arbitrary.

The proof is given in Appendix A.

### 3.3 Quadratic programming on the unit sphere

As to be seen in Section 3.4, the iterative algorithm Alg$\Phi$ for (30) requires to solve a set of constrained quadratic programmings of form

$$h_{opt} \triangleq \min_{h \in U_N} \{ \xi(h) := h^T D h - 2d^T h \}$$ (34)

where $D \in \mathbb{R}^{N \times N}$ is a positive semi-definite (PSD) matrix and $d \in \mathbb{R}^N$, both are independent of $h$. Let $D = U \Sigma U^T$ be an ED of $D$ where $\Sigma$ is an $N \times N$ diagonal matrix with diagonals ordered as $\sigma_1 \geq \cdots \sigma_N \geq 0$, and $U \in O_N$. Note that when $d = 0$, (34) is equivalent to finding the smallest eigenvector of $\Sigma$ and hence the solution to (34) is $h = U(:,; N)$. Thus $d \neq 0$ is assumed in the sequel. Unlike the unconstrained one, due to the unit-norm constraint, in general there is no closed-form solution to (34).

Denote the Lagrange function $\mathcal{L}(h, \lambda)$ of problem (34) as

$$\mathcal{L}(h, \lambda) = \xi(h) - \lambda(h^T h - 1)$$

Then, it follows from the Karush-Kuhn-Tucker (KKT) conditions that any solution to (34) should satisfy

$$\nabla h \mathcal{L}(h, \lambda) = 2((D - \lambda I_N)h - d) = 0$$

$$h^T h - 1 = 0$$ (35)

Note that the KKT condition in (35) is equivalent to

$$\vec{\Sigma} \vec{h} = \vec{d}$$

$$\vec{h}^T \vec{h} - 1 = 0$$ (36)

Note that to simplify the notations, we omit the subscript $k$ in $\Psi_{i,k}$ and $S_{i,k}$.
where $h = U\mathbf{h}$ and $d = U\mathbf{d}$.

With some manipulations, one can show that for any solution $(\mathbf{h}, \lambda)$ to (36), the objective function of (34) is given by

$$
\xi(h) = \sum_{n=1}^{N} \frac{\bar{d}_n^2}{\lambda - \sigma_n} + \lambda \hat{=} \xi(\lambda)
$$

for $\lambda \in \mathcal{S}_\lambda$, where $\mathcal{S}_\lambda$ is the solution set of

$$
g(\lambda) \hat{=} \|\mathbf{h}\|^2 = \sum_{n=1}^{N} \frac{\bar{d}_n^2}{(\sigma_n - \lambda)^2} = 1
$$

Therefore, the optimal $\lambda$ that corresponds to the solution of (34) is given by

$$
\lambda_{opt} \hat{=} \arg\min_{\lambda \in \mathcal{S}_\lambda} \xi(\lambda)
$$

with which the solution of (34) can then be obtained from

$$(D - \lambda_{opt} \mathbf{1}_N) \mathbf{h}_{opt} = \mathbf{d}
$$

Some properties of the set $\mathcal{S}_\lambda$ are summarized in the following theorem, which are given in [41].

**Theorem 1.** Let $g(\lambda)$ be defined in (38) for all $\lambda \in \mathbb{R}$ such that $\lambda \neq \sigma_l \forall l$ unless $\bar{d}_l = 0$. Then

- $g(\lambda)$ is convex on each of the interval $(-\infty, \sigma_N)$, $(\sigma_N, \sigma_{N-1})$, ..., $(\sigma_2, \sigma_1)$, and $(\sigma_1, \infty)$.
- there exist $N$ solutions $\{\lambda_n\}$ for $g(\lambda) = 1$ with $2 \leq N \leq 2N$. Suppose $\lambda_1 > \lambda_2 > \cdots > \lambda_N$. Then

$$
\lambda_1 \hat{=} \max_l \{\sigma_l \pm \bar{d}_l\} \leq \lambda_2 \leq \cdots \leq \lambda_N
$$

and $g(\lambda)$ is monotonically increasing within $[\lambda_1, \lambda_1]$.
- $\hat{\xi}(\lambda)$ (defined in (37)) is increasing on $\lambda_n$:

$$
\hat{\xi}(\lambda_1) \leq \hat{\xi}(\lambda_2) \leq \cdots \leq \hat{\xi}(\lambda_N)
$$

which implies $\lambda_{opt} = \lambda_1$.

As guaranteed by (40), $\lambda_{opt} = \lambda_1$. As $g(\lambda)$ is monotonically increasing within $[\lambda_1, \lambda_1]$, we can find $\lambda_1$ easily via a standard algorithm, say a bi-section-based algorithm. For convenience, the whole procedure for solving the unit-norm constrained quadratic programming (43) is denoted as Alg$_{CQP}$.

### 3.4 An iterative algorithm for (30)

With Alg$_{CQP}$ for solving the quadratic programming with unit-norm constraint (34), we now provide an iterative algorithm for (30). Note that (30) consists of a set of constrained minimizations of form

$$
\tilde{\Psi} = \arg\min_{\psi_i \in \mathcal{H}_N \cap \ell} \left\{ f(\Psi) := \| C - B \Psi S \|^2_F \right\}
$$

where $C, B$ and $S$ are independent of $\Psi$.

The idea behind the algorithm can be explained as follows. Assume that the first $l - 1$ columns of $\Psi$ have been updated. Rewrite the objective function as

$$
f(\Psi) = \| C - B \sum_{j \neq l} (\psi_j s_j^T) - B \psi_l s_l^T \|^2_F
$$

Let $E_l := C - B \sum_{j \neq l} (\psi_j s_j^T)$. Then minimizing $f$ in terms of $\psi_l$ is equivalent to solving

$$
\tilde{\psi}_l = \arg\min_{\psi \in \mathcal{H}_N} \| E_l - B \psi s_l^T \|^2_F
$$

This completes the iterative algorithm for solving (30), which is denoted as Alg$_{CQP}$. The whole procedure is given in Algorithm 1.
With some manipulations, it can be shown that
\[ ||E_l - B \psi s_i ||_F^2 = \psi^T s_i^T s_i B^T B \psi - 2s_i^T E_l^T B \psi + ||E_l||_F^2 \]
which implies (43) is equivalent to
\[ \tilde{\psi}_l = \arg\min_{\psi \in \mathcal{U}_N} \{ \psi^T D_l \psi - 2d_l^T \psi \} \]  
(44)
where \( D_l = ||s_l||_2^2 B^T B \) is a PSD matrix and \( d_l = B^T E_l s_l \). Obviously, the above is exactly of the same form as (34).

Based on the developments given above, the following iterative algorithm for addressing (30) is proposed:

**Alg.** - To update dictionary \( \Psi_i \) via solving (30)

*Initialization:* set \( C = C(\Phi_{k-1}, X_i), B = B(\Psi_{k-1}), S = S_{i,k}, \) and \( \Psi = \Psi_{i,k-1} \).

*Begin* \( l = 1, 2, \cdots, L_d \), update the \( l \)-th column of \( \Psi \) by
- Compute the overall representation error by
  \[ E_l = C - B \sum_{j \neq l} \psi_j s_j^T \]
  and \( D_l \) and \( d_l \) as
  \[ D_l = ||s_l||_2^2 B^T B, \ d_l = B^T E_l s_l \]
- Update the \( l \)-th column by solving (44) with \( \text{Alg}_{CQP} \) illustrated in Section 3.3 and setting \( \Psi(\cdot, l) = \tilde{\psi}_l \).

*End*

*Output:* \( \Psi_{i,k} = \Psi \).

**Remark 3.1:** Our proposed Algorithm 1 is based on the basic idea of alternating minimization that has led to a class of iterative algorithms for designing sensing matrices and dictionaries. The key difference that makes one algorithm different from another lies in the ways how the iterates are updated. It should be pointed out that though the alternating minimization-based algorithms practically work well for nonconvex problems, there is still lack of rigorous analysis for algorithm properties such as convergence. To the best of our knowledge, a few of results on this issue have been reported 35.

### 4 Convergence Analysis

In this section, we provide a rigorous convergence analysis for Algorithm 1. In particular, we show that the objective function monotonically decreases and more importantly, any accumulation point of the sequence (or iterates) is a stationary point of (26).

Let \( \mathcal{S}_{\kappa} \) denote the set of \( \kappa \)-sparse signals:
\[ \mathcal{S}_\kappa = \left\{ s \in \mathbb{R}^{L_0} : ||s||_0 \leq \kappa \right\}. \]
and \( \Psi = \{ \Psi_1, \ldots, \Psi_K \} \) and \( \mathcal{S} = \{ S_1, \ldots, S_K \} \). We also replace \( g(\Phi, \Psi, S_i, X_i) \) in (26) by \( g(\Phi, \Psi, S_i) \) since \( X_i \) is fixed during the learning process of the CCS system. With these notations, define \( g_i(\Phi, \Psi, S) = \sum_{i=1}^{K} g(\Phi, \Psi, S_i) \).

Furthermore, let \( \psi_{i,l} \) be the \( l \)-th column of \( \Psi_i \) and \( \Psi \setminus \psi_{i,l} / \psi \) (or \( \Psi_i \setminus \psi_{i,l} / \psi \)) be the matrix whose columns equal to those of \( \Psi \) (or \( \Psi_i \)) except the \( \psi_{i,l} \) that is replaced with \( \psi \).

Let \( \{(\Phi_k, \Psi_k, S_k)\} \) be the sequence generated by our proposed Algorithm 1. Roughly speaking, convergence analysis of Algorithm 1 is to study the behaviors of the sequence \( \{(\Phi_k, \Psi_k, S_k)\} \).

One notes that the cost function \( g_i(\Phi, \Psi, S) \) is nonnegative and both the dictionary update and the sensing matrix update can ensure the cost function decrease. Therefore, the proposed algorithm is
convergent in terms of cost function as \( \lim_{k \to \infty} \varrho_c(\Phi_k, \Psi_k, \mathcal{S}_k) \) exists as long as the sparse coding stage (see (29)) can be performed perfectly. The latter is not true in general if the sparse coding is carried out with the formulation (29) that is usually addressed using an OMP-based algorithm. One can get rid of this requirement if the \( \| \cdot \|_0 \)-based sparsity measure is replaced with \( \ell_1 \)-norm-based one since the sparse coding stage then is convex and can be solved exactly. The OMP-based algorithms are widely utilized in sparse coding problem due to its excellent computational efficiency. Simulations showed that this type of algorithms can provide solutions with a satisfying accuracy, especially when signals are very sparse [3,21,27]. In the sequel, it is assumed that the sparse coding stage can be performed perfectly.

Theorem 2 yields more results on the convergence behaviors of Algorithm 1.

**Theorem 2.** (Convergence of Algorithm 1) Suppose that the sparse coding stage can be performed perfectly such that the best \( k \)-sparse approximations to the signals \( C(\Phi_{k-1}, X_i) \) represented in the dictionary \( B(\Phi_{k-1}) \) can be obtained. Let \( \{(\Phi_k, \Psi_k, \mathcal{S}_k)\} \) be the sequence generated by Algorithm 1.

1. The sequence \( \{(\Phi_k, \Psi_k, \mathcal{S}_k)\} \) is bounded, implying it has at least one accumulation point.
2. Every accumulation point lies in \( \mathcal{O}_{M,N} \times \mathcal{U}_{N}^{\ell_0K} \times \mathcal{S}_M^{\ell_0K} \).
3. Every accumulation point \( (\Phi^*, \Psi^*, \mathcal{S}^*) \) satisfies
   \[
   \varrho_c(\Phi^*, \Psi^*, \mathcal{S}^*) = \lim_{k \to \infty} \varrho_c(\Phi_k, \Psi_k, \mathcal{S}_k)
   \]
4. Every accumulation point \( (\Phi^*, \Psi^*, \mathcal{S}^*) \) satisfies
   \[
   \varrho_c(\Phi^*, \Psi^*, \mathcal{S}^*) = \min_{\Phi \in \mathcal{O}_{M,N}} \varrho_c(\Phi, \Psi^*, \mathcal{S}^*)
   = \min_{\mathcal{S} \in \mathcal{S}_M^{\ell_0K}} \varrho_c(\Phi^*, \Psi^*, \mathcal{S})
   = \min_{\Psi \in \mathcal{U}_{N}} \varrho_c(\Phi^*, \Psi^* \setminus \psi_i, \mathcal{S}^*)
   
   for all \( l \in [L_0] \) and \( i \in [K] \).
5. Any accumulation point of \( \{(\Phi_k, \Psi_k, \mathcal{S}_k)\} \) is a stationary point of \( \varrho_c(\Phi, \Psi, \mathcal{S}) \).

**Remark 4.1:**

- We note that Theorem 2 implies the sequence generated by Algorithm 1 has at least one accumulation point and each accumulation point is a stationary point of \( \varrho_c(\Phi, \Psi, \mathcal{S}) \). However, there is nothing to guarantee that the sequence itself is convergent. Recently, the convergence analysis of proximal methods for solving a class of non-smooth and non-convex problems are presented in [10,12]. In particular, by adding proximal terms, one can show that the sequence is a Cauchy sequence and converges to a stationary point. This technique has been utilized in [8] for analyzing the convergence of a particular algorithm for learning tight frame with orthogonality constraint. We defer this direction to future research;

- Though the proposed Algorithm 1 is not guaranteed to make the sequence \( \{(\Phi_k, \Psi_k, \mathcal{S}_k)\} \) convergent, the 3rd and 5th points in Theorem 2 imply that \( (\Phi_k, \Psi_k, \mathcal{S}_k) \) with \( k \) very big can yield the same performance as that by a stationary point of \( \varrho_c(\Phi, \Psi, \mathcal{S}) \) corresponding to the point \( (\Phi^*, \Psi^*, \mathcal{S}^*) \). Also the 4th point in Theorem 2 ensures that every accumulation point is the minimum of \( \varrho_c \) with respect to one variable (like in terms of \( \Phi \) when the dictionary \( \Psi^* \) and sparse coefficients \( \mathcal{S}^* \) are fixed). From a practical point of view, such \( \{(\Phi_k, \Psi_k, \mathcal{S}_k)\} \) is as good as \( (\Phi^*, \Psi^*, \mathcal{S}^*) \) as the CCS system \( (\Phi_k, \Psi_k) \) yields the same performance as the CCS system \( (\Phi^*, \Psi^*) \) does according to the design criterion.

It should be pointed out that our strategy for simultaneously learning the sensing matrix and dictionary for a CS system differs from the one in [21] in that we provide a unified and identical measure for jointly optimizing the sensing matrix and dictionary. This framework ensures that it is possible to provide an algorithm with guaranteed convergence.
5 Experiments

In this section, we will present a series of experiments to examine the performance of the proposed CCS scheme and approach for designing optimal CCS systems.

5.1 Demonstration of the Collaborative CS Scheme

In this section, we will demonstrate the performance of the notion of collaborative CS scheme (as shown in Fig. 1) and compare it with that of traditional ones.

Through the experiments for this part, we set the number of dictionaries $K = 5$. Each training data $X_i \in \mathbb{R}^{N_i \times J_0} (i \in [K])$ is obtained by 1) randomly extracting 15 non-overlapping patches (the dimension of each patch is $8 \times 8$) from each of 400 images in the LabelMe \cite{33} training data set, and 2) arranging each patch of $8 \times 8$ as a vector of $64 \times 1$. Such a setting implies $N = 64$ and $J_0 = 15 \times 400 = 6000$. For each training data set $X_i$, we apply the K-SVD algorithm to obtain a sparsifying dictionary $\Psi_{KSVD_i} \in \mathbb{R}^{N_i \times L_0}$ with $L_0 = 100$ and a given sparsity level $\kappa = 4$. We also apply the K-SVD algorithm for the entire training data $X = [X_1 \cdots X_i \cdots X_K]$ to obtain a sparsifying dictionary $\Psi_{KSVD} \in \mathbb{R}^{N \times L}$ with $L = L_0$ and $\Psi_{KSVD} \in \mathbb{R}^{N \times L'}$ with $L' = 256$. We then generate a random $M \times N$ sensing matrix $\Phi$. The CS systems with $(\Phi, \Psi_{KSVD_i}), (\Phi, \Psi_{KSVD})$ and $(\Phi, \Psi_{KSVD_i})$ are then respectively denoted by $CS_{Rdm_i}$, $CS_{Rdm}$ and $CS_{Rdm}$. For any image $H$, we apply CS system $CS_{Rdm}$, to compress it and use $\hat{H}_i$ to denote the output of each dictionary system. The collaborative CS system (denoted by $CCS_{Rdm}$) has the output $\tilde{H} = \frac{1}{K} \sum_{i=1}^{K} H_i$. For convenience, we utilize $CCS_{Rdm_m}$ to denote the collaborative CS system that fuses the estimates from the first $m$ CS systems, i.e., with the output $\tilde{H}_{m} = \frac{1}{m} \sum_{i=1}^{m} H_i$. Clearly, $\tilde{H} = \tilde{H}_K$.

The reconstruction accuracy is evaluated in terms of peak signal-to-noise ratio (PSNR), defined as

$$\sigma_{psnr} = 10 \times \log_{10} \left( \frac{(2^r - 1)^2}{\sigma_{mse}} \right)$$

where $r = 8$ bits per pixel and $\sigma_{mse}$ for images $H, \tilde{H} \in \mathbb{R}^{N_1 \times N_2}$ (where $\tilde{H}$ is an estimate of $H$) defined as

$$\sigma_{mse} = \frac{1}{N_1 \times N_2} \sum_{n_1=1}^{N_1} \sum_{n_2=1}^{N_2} |H(n_1, n_2) - \tilde{H}(n_1, n_2)|^2$$

Fig. 2 shows the PSNR $\sigma_{psnr}$ of the CS systems $CS_{Rdm_i}$ and $CS_{Rdm}$, when applied to the image ‘Plane’. Table 1 provides $\sigma_{psnr}$ for the CS systems $CS_{Rdm_i}$, $CS_{Rdm}$, $CS_{Rdm}$ and $CS_{Rdm}$, tested with the twelve images. Fig. 3 displays the visual effects of image ‘Plane’ for the CS systems $CS_{Rdm_i}$ and $CS_{Rdm}$.

Remark 5.1:

- It is observed from Fig. 2 that the $\sigma_{psnr}$ increases when $k$ increases for $CS_{Rdm_k}$. This implies that by fusing more estimators from the classical CS systems $\{CS_{Rdm_i}\}$, the collaborative CS system has better performance. This coincides with Lemma 17 that $\sigma_{mse}$ is expected to decrease as $K$ increases for the collaborative CS systems;
- Table 1 also demonstrates the advantages of $CS_{Rdm}$. It is of great interest to note that $CS_{Rdm}$ has still much better performance than $CS_{Rdm}$ whose dictionary is learned with all the training data $X$;
- As observed from Fig. 3 different CS systems $\{CS_{Rdm_i}\}$ yield different estimates of the image ‘Plane’. As each dictionary has limited capacity of representation, the corresponding CS system may yield excellent performance for some of the patches (of the image), but very poor one for the others. As PSNR is evaluated over all the patches, from a statistical point of view the PSNR of a single CS system is around the averaged PSNR and hence most likely lower than the one by the proposed CCS scheme;
- Finally, it is observed from Table 1 that $CS_{Rdm}$, though with a higher dimensional dictionary $\Psi_{KSVD}$, does not outperform $CS_{Rdm}$. This, as conjectured before and one of the arguments for the proposed CCS framework, is due to the fact that the mutual coherence of the dictionary $\Psi$ and hence the equivalent one $A$ gets higher when the dimension increases and consequently, the signal reconstruction accuracy decreases.
Figure 2: PSNR $\sigma_{psnr}$ of CS systems and collaborative CS systems for image ‘Plane’.

![Figure 3: The original 'Plane' and reconstructed images from their CS samples with $M = 20$ and $\kappa = 4$. (a) The original; (b) CS$_{Rdm}$; (c) CS$_{Rdm}$; (d) CCS$_{Rdm}$.](image)

5.2 Performance of the Optimized Collaborative CS Systems

With the obtained training data $X$, we now examine the performance of the collaborative CS system with sensing matrix and dictionaries simultaneously learned by solving (26) with $\alpha = 0.2$, $\beta = 1$. We run the proposed Algorithm 1 with $N_{iters} = 30$ iterations to solve (26). In this section, this learned collaborative CS system is simply denoted by CCS. Fig. 4 shows the

---

5 Though there is no systematic way to find the best $\alpha$ and $\beta$ (which depend on specific applications and are also not the main focus of this paper), we provide a rough guide to choose these parameters. With respect to $\alpha$ which weights the importance of the projected signal noise, similar to what is suggested in [21], $\alpha < 1$ is preferred since we need to highlight the sparse representation error $E$. In terms of $\beta$, $\beta > \alpha$ is suggested since in dictionary learning for image processing, the sparse representation error $E$ is not very small and capturing most of the key information in $\Psi S$ is more important for a sensing matrix.
Table 1: Statistics of $\sigma_{psnr}$ (dB) for images processed with $M = 20, N = 64, L_0 = 100, \kappa = 4$, and $K = 5$

|                | Baboon | Boat | Child | Couple | Crowd | Elanie | Finger | Lake | Lax | Lena | Man | Plane |
|----------------|--------|------|-------|--------|-------|--------|--------|------|-----|------|-----|-------|
| $CS_{Rdm_1}$  | 22.11  | 26.79| 31.03 | 26.90  | 27.51 | 29.45  | 23.38  | 25.83| 22.49| 27.51| 29.45| 25.83 |
| $CS_{Rdm_2}$  | 22.06  | 26.63| 31.08 | 26.95  | 27.55 | 29.37  | 23.17  | 25.87| 22.51| 29.43| 27.37| 28.49 |
| $CS_{Rdm_3}$  | 22.21  | 26.81| 31.05 | 26.98  | 27.53 | 29.39  | 23.81  | 26.00| 22.55| 29.74| 27.45| 28.54 |
| $CS_{Rdm_4}$  | 22.10  | 26.67| 31.10 | 26.92  | 27.59 | 29.56  | 23.51  | 25.92| 22.54| 29.61| 27.41| 28.02 |
| $CS_{Rdm_5}$  | 22.39  | 27.24| 31.70 | 27.51  | 28.31 | 29.78  | 24.31  | 26.46| 22.77| 30.32| 27.89| 29.22 |

We now compare the CCS with other CS systems. The CS systems $CS_{Elad}$, $CS_{Classic}$, $CS_{TKK}$, $CS_{LZYCB}$, and $CS_f$ are the ones with the learned dictionary $\Psi_{KSVD}$ and the sensing matrix $\Phi \in \mathbb{R}^{M \times N}$ designed via the method in [22], [39], [36], [29], and [27] (by minimizing the average coherence of the equivalent dictionary), respectively. The CS system $CS_{DCS}$ is the one with the sensing matrix and dictionary that are simultaneously designed with the training data $X$ using the method cin [21].

Table 2 provides the performance of these CS systems for the same twelve images. Fig. 5 displays the visual effects of image ‘Plane’ for these CS systems.

Table 2: Statistics of $\sigma_{psnr}$ (dB) for images processed with $M = 20, N = 64, L_0 = 100, \kappa = 4$, and $K = 5$

|                | Baboon | Boat | Child | Couple | Crowd | Elanie | Finger | Lake | Lax | Lena | Man | Plane |
|----------------|--------|------|-------|--------|-------|--------|--------|------|-----|------|-----|-------|
| $CS_{Elad}$   | 12.11  | 16.55| 25.38 | 19.40  | 22.67 | 14.51  | 20.94  | 14.98| 10.86| 21.52| 18.68| 23.69 |
| $CS_{Classic}$| 9.39   | 13.78| 22.79 | 16.69  | 20.27 | 11.83  | 18.91  | 12.26| 8.11 | 18.78| 15.88| 21.26 |
| $CS_{TKK}$    | 11.93  | 16.31| 25.03 | 19.02  | 22.28 | 14.32  | 20.39  | 14.60| 10.68| 21.19| 18.34| 23.32 |
| $CS_{LZYCB}$  | 8.15   | 12.54| 21.64 | 15.52  | 19.22 | 10.73  | 17.98  | 11.08| 6.90 | 17.67| 14.82| 20.11 |
| $CS_{DCS}$    | 25.35  | 30.39| 34.74 | 30.58  | 31.45 | 32.62  | 27.39  | 29.59| 25.74| 33.38| 30.87| 32.50 |
| $CS_f$        | 24.85  | 30.04| 34.70 | 30.16  | 31.46 | 32.18  | 27.44  | 29.42| 25.27| 33.01| 30.61| 32.45 |
| CCS           | 26.27  | 31.50| 36.41 | 31.98  | 33.28 | 33.36  | 29.96  | 31.08| 26.43| 34.70| 32.03| 34.20 |
| CCS8          | 26.33  | 31.76| 36.70 | 32.04  | 33.65 | 33.45  | 30.40  | 31.25| 26.56| 35.03| 32.26| 34.45 |

--

We set the coupling factor in $CS_{DCS}$ to 0.5, which is suggested in [21].
Figure 5: The reconstructed 'Plane' images from their CS samples with $M = 20$ and $\kappa = 4$. (a) $\text{CS}_{\text{DCS}}$; (b) $\text{CS}_f$; (c) $\text{CS}_{\text{etf}}$; (d) $\text{CS}$. 

Remark 5.2: 

- It is observed from Table 1 and Table 2 that the CS systems using a random sensing matrix outperform $\text{CS}_{\text{Elad}}$, $\text{CS}_{\text{Classic}}$, $\text{CS}_{\text{TKK}}$, and $\text{CS}_{\text{LZYCB}}$. This is due to the fact that the sensing matrix in each of those CS system is optimized with the assumption that the sparse representation error of signals is nil, which is not the case for images at all; 

- As the sensing matrices used in $\text{CS}_f$ and $\text{CS}_{\text{DCS}}$ are optimized with the sparse representation error taken into account, they perform much better than those using a random sensing matrix; 

- The superiority of our proposed CCS system is again demonstrated clearly. Compared with $\text{CS}_{\text{Rdm}}$, it is basically 2.5 dB better in terms of $\sigma_{\text{psnr}}$. Also, it yields an improvement over $\text{CS}_{\text{DCS}}$ by more than 1 dB. The amount of improvement can be further increased if the number of CS systems in the CCS system gets bigger. With the $5 \times 6000 = 30,000$ samples in the obtained training data matrix $X$, a CCS system with $K = 8$, denoted as $\text{CCS}_8$, is also obtained and its performance is better than the CCS (corresponding to $K = 5$) by an amount of 0.15 dB on average. See Table 2.

6 Conclusion 

Based on the MLE principle, a collaborative compressed sensing framework has been raised and analyzed. Such a system consists of a bank of $K$ CS systems that share the same sensing matrix. It has been shown that the CCS system yields a better performance than a sub CS system in a statistical sense. A measure has been proposed, which allows us to simultaneously optimize the sensing matrix and all the $K$ dictionaries under the same scheme. An algorithm has been derived for solving the corresponding optimal CCS system problem. Experiments were carried out using real images and confirmed our theoretical results. The superiority of the optimized CCS system over the traditional CS systems has been clearly demonstrated.
A Proof of Lemma 3

Proof of Lemma 3. First of all, let \( g(\Phi, \Psi_s, S_i, X_i) \) and \( G(\Psi_i, S_i) \) be the same as defined before in (25) and (32), respectively. With \( \Phi \in \mathcal{O}_{M,N} \), that is \( \Phi \Phi^T = I_M \), it can be shown that

\[
g(\Phi, \Psi_s, S_i, X_i) = -\text{trace} \left( \Phi^T \Phi G(\Psi_i, S_i) \right) + ct
\]

where \( ct \) is independent of \( \Phi \). Thus, (31) is equivalent to

\[
\Phi_k = \arg\max_{\Phi \in \mathcal{O}_{M,N}} \left\{ \sum_{i=1}^K \text{trace} \left( \Phi^T \Phi G(\Psi_i, S_i) \right) \right\}
\]

It then follows from the ED of \( \sum_{i=1}^K G(\Psi_i, S_i) = V_G \Pi V_G^T \) and \( \Phi = U \left[ I_M \ 0 \right] V^T \) - an SVD of \( \Phi \) due to \( \Phi \in \mathcal{O}_{M,N} \) that

\[
\eta = \text{trace} \left( \hat{V}^T \hat{\Pi} \hat{V} \left[ I_M \ 0 \right] \right) = \sum_{n=1}^M Q(n,n)
\]

where \( \hat{V} \triangleq V_G \hat{V} \) and \( \hat{\Pi} \triangleq \hat{V}^T \hat{\Pi} \hat{V} \).

Without loss of generality, assume that the diagonal elements \( \{\pi_n\} \) of \( \Pi \) satisfy \( \pi_n \geq \pi_{n+1}, \forall n \).

According to [29] (see the proof of Theorem 3 there), one has

\[
\eta = \sum_{n=1}^M Q(n,n) \leq \sum_{n=1}^M \pi_n
\]

Thus, \( \eta \) is maximized if \( \hat{V} \) is of the following form

\[
\hat{V} = \left[ \begin{array}{cc} \hat{V}_{11} & 0 \\ 0 & \hat{V}_{22} \end{array} \right]
\]

(45)

where both \( \hat{V}_{11} \in \mathcal{O}_M \) and \( \hat{V}_{22} \in \mathcal{O}_{N-M} \) are arbitrary. Consequently, the optimal sensing matrix is given by

\[
\Phi_k = \left[ I_M \ 0 \right] \hat{V}^T V_G^T
\]

where \( \hat{V} \) is of form given by (45) and \( U \in \mathcal{O}_M \) is arbitrary. Clearly, \( \hat{V} = I_N \) yields one of the solutions, that is (33).

B Proof of Theorem 2

B.1 Supporting results:

We list some definitions and important results here.

A subsequence of \( \{x_k\} \) is a sequence \( \{x_{k_m}\} \), where \( k_1 < k_2 < \cdots \) is an increasing sequence of indices. Suppose \( x_k \in \mathbb{R}^N \).

Definition 1. The set of subsequence limits \( \Omega \) of a sequence \( \{x_k\} \) is the set of limits of convergent subsequences, i.e.

\[
\Omega = \left\{ x \in \mathbb{R}^N : \exists \{x_{k_m}\} \text{ s.t. } x_{k_m} \to x \right\}.
\]

Theorem 3 (Bolzano-Weierstrass Theorem [10]). The Bolzano-Weierstrass Theorem states that any bounded sequence in \( \mathbb{R}^N \) has a convergent subsequence.

Definition 2. Let \( f : \mathbb{R}^N \to \mathbb{R} \cup \{\infty\} \) be a proper lower semi-continuous function.

- The domain of \( f \) is defined by \( \text{dom} f := \{x \in \mathbb{R}^N : f(x) < \infty\} \).
- For each \( x \in \text{dom} f \), \( x \) is called the coordinate-wise minimum of \( f \) if it satisfies

\[
f(x + [0,\ldots,d_n,\ldots,0]^T) \leq f(x), \ \forall d_n, n \in [N].
\]
• The Fréchet subdifferential \( \partial_f f \) is defined by
\[
\partial_f f(x) = \left\{ z : \lim_{{y \to x}} \inf \frac{f(y) - f(x) - \langle z, y - x \rangle}{\|y - x\|} \geq 0 \right\}
\]
for any \( x \in \text{dom} f \) and \( \partial_f f(x) = \emptyset \) if \( x \notin \text{dom} f \).

• For each \( x \in \text{dom} f \), \( x \) is called the **stationary point** of \( f \) if it satisfies \( 0 \in \partial f \).

• Other definition of stationary point includes: \( x \) is called the stationary point of \( f \) if it satisfies \( 0 \in \partial f \), where \( \partial f \) is the limiting subdifferential given by
\[
\partial f = \{ z : \exists x_k \to x, f(x_k) \to f(x), z_k \in \partial_f f(x_k) \to z \}
\]

### B.2 Main proofs

Define
\[
f(\Phi, \Psi, S) \triangleq \varrho_c(\Phi, \Psi, S) + I_{\mathcal{O}_{M,N}}(\Phi) + I_{\mathcal{U}_{K}^{L_0N}}(\Psi) + I_{\mathcal{S}_{L_{0}N}}(S)
\]
where \( I_S(x) = 0 \), if \( x \in S \) and \( \infty \) otherwise, for \( S = \mathcal{O}_{M,N}, \mathcal{U}_{K}^{L_0N}, \mathcal{S}_{L_{0}N} \) with \( \mathcal{O}_{M,N}, \mathcal{U}_{K} \) and \( \mathcal{S}_{L_{0}N} \) defined before.

Then, the minimization [26] is equivalent to the following unconstrained problem
\[
\min_{\Phi, \Psi, S} f(\Phi, \Psi, S)
\]

**Lemma 4.** The sequence \( \{Z_k \triangleq (\Phi_k, \Psi_k, S_k)\} \) generated by Algorithm 1 is a bounded sequence. For any convergent sub-sequence \( \{Z_{k_m}\} \) with limit point \( Z^* = (\Phi^*, \Psi^*, S^*) \), we have
\[
\Phi^* \in \mathcal{O}_{M,N}, \quad \Psi^* \in \mathcal{U}_{K}^{L_0N}, \quad S^* \in \mathcal{S}_{L_{0}N}
\]
and
\[
f(k) = \lim_{k \to \infty} f(Z_{k_m}) = \lim_{k \to \infty} \varrho_c(Z_{k_m}) = \varrho_c(Z^*)
\]

**Proof of Lemma 4.** It is clear that \( \Phi_k \) and \( \Psi_k \) are bounded for all \( k \) since both \( \mathcal{O}_{M,N} \) and \( \mathcal{U}_{K} \) are compact sets. This further implies \( C(\Phi_k, X_k) \) and \( B(\Phi_k)\Psi_k \) are bounded and hence \( S_{k+1} \) is also bounded for all \( k \) and \( i \in [K] \). Thus \( S_k \) is also bounded. Since \( \Phi_{k_m} \in \mathcal{O}_{M,N}, \Psi_{k_m} \in \mathcal{U}_{K}^{L_0N}, S_{k_m} \in \mathcal{S}_{L_{0}N} \) and all of them are bounded, we conclude \( \Phi^* \in \mathcal{O}_{M,N}, \Psi^* \in \mathcal{U}_{K}^{L_0N}, S^* \in \mathcal{S}_{L_{0}N} \).

Noting that \( \varrho_c \) is a continuous function, we have \( \varrho_c(Z_{k_m}) \to \varrho_c(Z^*) \) as \( k_m \to \infty \).

By the definition of \( \Phi_k, \Psi_k \) and \( S_k \), we have
\[
f(Z_0) \geq \cdots \geq f(Z_{k-1}) \geq f(\Phi_{k-1}, \Psi_{k-1}, S_k) \\
\geq f(\Phi_{k-1}, \Psi_{k-1}, S_k) \geq f(Z_k) \geq \cdots
\]

which implies that \( f(Z_k) \) is decreasing and also \( f(Z_k) \geq 0 \) by its definition. Thus \( f(Z_k) \) is a convergent sequence (see Theorem 3). Consequently, we have
\[
\lim_{k_m \to \infty} f(Z_{k_m}) = \lim_{k_m \to \infty} \{ \varrho_c(Z_{k_m}) + I_{\mathcal{O}_{M,N}}(\Phi_{k_m}) + I_{\mathcal{U}_{K}^{L_0N}}(\Psi_{k_m}) + I_{\mathcal{S}_{L_{0}N}}(S_{k_m}) \} = \varrho_c(Z^*)
\]

**Lemma 5.** Let \( \{Z_k = (\Phi_k, \Psi_k, S_k)\} \) denote the sequence generated by Algorithm 1 and let \( \Omega \) denote the set of subsequence limits. Then, \( \Omega \) is not empty and
\[
f(Z^*) = \inf_k f(Z_k), \quad \forall(Z^*) \in \Omega.
\]
Proof of Lemma \[\text{Lemma 5}\] Lemma \[\text{Lemma 4}\] establishes that \(\{Z_k\}\) is a bounded sequence. Thus, Bolzano-Weierstrass Theorem (see Theorem \[\text{Theorem 3}\]) tells us that the set \(\Omega\) is not empty. Notice that \(\{f(Z_k)\}\) is a decreasing sequence and \(f(Z_k) \geq 0\). By monotone convergence theorem \[\text{Theorem 10}\], there exists some constant \(c\) such that \(\inf_k f(Z_k) = c\). For any \(Z^* \in \Omega\), assume \(Z_{k_m} \to Z^*\) as \(k_m \to \infty\). We have that \(\lim_{k_m \to \infty} f(Z_{k_m}) = f(Z^*) = c\).

Lemma \[\text{Lemma 5}\] tells that the sequence \(\{Z_k\}\) has at least one limit point. Consider any convergent subsequence \(\{Z_{k_m}\}\) with limit point \(Z^* = (\Phi^*, \Psi^*, S^*)\). Since by definition \(\Phi_{k_m} = \arg\min_{\Phi \in O_{M,N}} \varrho_c(\Phi, \Psi_{k_m}, S_{k_m})\), we have

\[
\varrho_c(\Phi_{k_m}, \Psi_{k_m}, S_{k_m}) \leq \varrho_c(\Phi, \Psi_{k_m}, S_{k_m}), \quad \forall \Phi \in O_{M,N}
\]

Taking \(k_m \to \infty\), we have

\[
\varrho_c(Z^*) \leq \varrho_c(\Phi, \Psi^*, S^*), \quad \forall \Phi \in O_{M,N}
\]

which implies

\[
\varrho_c(Z^*) + I_{O_{M,N}}(\Phi^*) \\
\leq \varrho_c(\Phi + \Phi^*, \Psi^*, S^*) + I_{O_{M,N}}(\Phi + \Phi^*), \quad \forall \Phi \in \mathbb{R}^{M \times N}
\]

since \(I_{O_{M,N}}(\Phi + \Phi^*) = 0\) if \(\Phi + \Phi^* \notin O_{M,N}\). Thus, we obtain

\[
f(Z^*) \leq f(\Phi + \Phi^*, \Psi^*, S^*), \quad \forall \Phi \in \mathbb{R}^{M \times N} \quad (47)
\]

Let \(\psi_{i,k,l}\) denote the \(l\)-th column of \(\Psi_{i,k}\). By the definition of \(Z_{k_{m+1}}\), we similarly have

\[
\varrho(\Phi_{k_m}, \psi_{i,k_m+1,1}, \psi_{i,k_m,2}, \ldots, \psi_{i,k_m,L}, S_{i,k_m+1}) \\
\leq \varrho(\Phi_{k_m}, \psi_{i,k_m,1}, \ldots, \psi_{i,k_m,L}, S_{i,k_m+1}), \quad \forall \psi \in U_N
\]

for all \(i \in K\). The summation of the first inequality, the second inequality with \(\psi = \psi_{i,k_m,2}, \ldots, \) and the last inequality with \(\psi = \psi_{i,k_m,L}\) gives

\[
\varrho(\Phi_{k_m}, \psi_{i,k_m+1,1}, S_{i,k_m+1}) \\
\leq \varrho(\Phi_{k_m}, \psi_{i,k_m+1,1} \psi_{i,k_m,1}/\psi_{i,k_m+1}, S_{i,k_m+1}), \quad \forall \psi \in U_N
\]

for all \(i \in K\). Thus we have

\[
\varrho_c(\Phi_{k_m}, \psi_{i,k_m+1,1}, S_{k_m+1}) = \sum_{i=1}^{K} \varrho(\Phi_{k_m}, \psi_{i,k_m+1,1}, S_{i,k_m+1}) \\
\leq \varrho_c(\Phi_{k_m}, \psi_{i,k_m+1,1} \psi_{i,k_m,1}/\psi_{i,k_m+1}, S_{k_m+1}), \quad \forall \psi \in U_N
\]

for any \(i \in K\). This along with

\[
\varrho_c(\Phi_{k_m+1}, \psi_{k_m+1,1}, S_{k_m+1}) \leq \varrho_c(\Phi_{k_m+1}, \psi_{k_m+1,1}, S_{k_m+1})
\]

for all \(\Phi \in O_{M,N}\) gives

\[
\varrho_c(Z_{k_m+1}) \leq \varrho_c(\Phi_{k_m}, \psi_{i,k_m,1}/\psi_{i,k_m+1}), \quad \forall \psi \in U_N
\]

for any \(i \in K\). Taking \(k_m \to \infty\), we have

\[
\varrho_c(Z^*) \leq \varrho_c(\Phi^*, \psi^*/\psi^*, \Sigma^*), \quad \forall \psi \in U_N
\]

for any \(i \in K\).
To simplify the notation, let
\[ \Xi(Z) = \Xi(\Phi, \Psi, S) = I_{\mathcal{O}_{M,N}}(\Phi) + I_{U_N^0K}(\Psi) + I_{S_N^0K}(S) \]

Utilizing the fact that \( \Phi^* \in \mathcal{O}_{M,N}, \Psi^* \in U_N^0K, S^* \in S_N^0K, \) we have
\[ g_\epsilon(Z^*) + \Xi(Z^*) \leq g_\epsilon(\Phi^*, \Psi^* \setminus \psi_{i,1}^*/(\psi_{i,1}^* + \psi), S^*) + \Xi(\Phi^*, \Psi^* \setminus \psi_{i,1}^*/(\psi_{i,1}^* + \psi), S^*) \]

since \( I_{U_N}(\psi_{i,1}^* + \psi) = \infty \) if \( \psi_{i,1}^* + \psi \notin U_N. \) This implies
\[ f(Z^*) \leq f(\Phi^*, \Psi^* \setminus \psi_{i,1}^*/(\psi_{i,1}^* + \psi), S^*), \forall \psi \in \mathbb{R}^N \]
for all \( i \in [K]. \) Similarly, we obtain
\[ f(Z^*) \leq f(\Phi^*, \Psi^* \setminus \psi_{i,l}^*/(\psi_{i,l}^* + \psi), S^*), \forall \psi \in \mathbb{R}^N \]
for all \( l \in [L], i \in [K]. \)

With similar argument, we also have
\[ g_\epsilon(Z^*) \leq g_\epsilon(\Phi^*, \Psi^* \setminus \psi_{i,l}^*/(\psi_{i,l}^* + \psi), S^*) \]
\[ \forall S \in \mathbb{R}^{L_0 \times J_0 K} \]

Thus, we have shown that the point \( (\Phi^*, \Psi^*, S^*) \) is a coordinate-wise minimum point.

Now, for any \( \delta z = (\delta \Phi, \delta \Psi, \delta S), \) we have
\[
\liminf_{\|\delta z\|_F \to 0} \frac{f(Z^* + \delta z) - f(Z^*)}{\|\delta z\|_F} \\
= \liminf_{\|\delta z\|_F \to 0} \frac{g_\epsilon(Z^* + \delta z) - g_\epsilon(Z^*) + \Xi(Z^* + \delta z) - \Xi(Z^*)}{\|\delta z\|_F} \\
= \liminf_{\|\delta z\|_F \to 0} \frac{\epsilon^* - o(\|\delta z\|_F) + \Xi(Z^* + \delta z) - \Xi(Z^*)}{\|\delta z\|_F} \\
= \liminf_{\|\delta z\|_F \to 0} \left( \frac{f(\Phi^* + \delta \Phi, \Psi^* \setminus \psi_{i,l}^*/(\psi_{i,l}^* + \psi), S^*) - f(Z^*) - o(\|\delta \Phi\|_F)}{\|\delta z\|_F} \\
+ \sum_{i,l} \frac{f(\Phi^* \setminus \psi_{i,l}^*/(\psi_{i,l}^* + \delta \psi_{i,l}), S^*) - f(Z^*) - o(\|\delta \psi_{i,l}\|)}{\|\delta z\|_F} \right) \\
\geq \liminf_{\|\delta z\|_F \to 0} \frac{-o(\|\delta \Phi\|_F) - o(\|\delta \psi_{i,l}\|) - o(\|\delta S\|_F)}{\|\delta z\|_F} = 0.
\]

where \( \epsilon^* = \langle \nabla g_\epsilon(Z^*), \delta z \rangle, \) the “little-o” notation \( o(c) \) is defined by \( \lim_{c \to 0} \frac{o(c)}{c} = 0, \) while the second equality follows from Taylor’s theorem expanding \( g_\epsilon(Z^* + \delta z) \) at \( Z^* \), and the inequality follows from (47) - (49). Thus, the point \( Z^* \) is a stationary point. This completes the proof of Theorem 2.

References
[1] V. Abolghasemi, S. Ferdowsi, and S. Sanei. A gradient-based alternating minimization approach for optimization of the measurement matrix in compressive sensing. Signal Process., 92(4):999–1009, 2012.
[2] M. Aharon and M. Elad. Sparse and redundant modeling of image content using an image-signature-dictionary. SIAM J. Imag. Sci., 1(3):228–247, 2008.
[3] M. Aharon, M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. IEEE Trans. Signal Process., 54(11):4311–4322, 2006.
[4] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-lojasiewicz inequality. *Math. Operations Research*, 35(2):438–457, 2010.

[5] H. Attouch, J. Bolte, and B. F. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized gauss–seidel methods. *Math. Program.*

[6] H. Bai, G. Li, S. Li, Q. Li, Q. Jiang, and L. Chang. Alternating optimization of sensing matrix and sparsifying dictionary for compressed sensing. *IEEE Trans. Signal Process.*, 63(6):1581–1594, 2015.

[7] A. S. Bandeira, E. Dobriban, D. G. Mixon, and W. F. Sawin. Certifying the restricted isometry property is hard. *IEEE Trans. Inf. Theory*, 59(6):3448–3450, 2013.

[8] C. Bao, H. Ji, and Z. Shen. Convergence analysis for iterative data-driven tight frame construction scheme. *Appl. Comput. Harmon. Anal.*, 38(3):510–523, 2015.

[9] A. S. Bandeira, E. Dobriban, D. G. Mixon, and W. F. Sawin. Certifying the restricted isometry property is hard. *IEEE Trans. Inf. Theory*, 59(6):3448–3450, 2013.

[10] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constr. Approx.,* 28(3):253–263, 2008.

[11] T. Blumensath and M. E. Davies. Iterative thresholding for sparse approximations. *J. Fourier Anal. Appl.*, 14(5-6):629–654, 2008.

[12] J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Math. Program.*, 146(1-2):459–494, 2014.

[13] E. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory*, 52(2):489–509, 2006.

[14] E. Candès and M. B. Wakin. An introduction to compressive sampling. *IEEE Signal Process. Mag.*, 25(2):21–30, 2008.

[15] E. J. Candès, Y. C. Eldar, D. Needell, and P. Randall. Compressed sensing with coherent and redundant dictionaries. *Appl. Comput. Harmon. Anal.*, 31(1):59–73, 2011.

[16] E. J. Candès, J. K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Commun. Pure Appl. Math.*, 59(8):1207–1223, 2006.

[17] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20(1):33–61, 1998.

[18] M. A. Davenport, J. N. Laska, J. R. Treichler, and R. G. Baraniuk. The pros and cons of compressive sensing for wideband signal acquisition: Noise folding versus dynamic range. *IEEE Trans. Signal Process.*, 60(9):4628–4642, 2012.

[19] D. L. Donoho. Compressed sensing. *IEEE Trans. Inf. Theory*, 52(4):1289–1306, 2006.

[20] D. L. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via l1 minimization. *Proc. Natl. Acad. Sci.*, 100(5):2197–2202, 2003.

[21] J. M. Duarte-Carvajalino and G. Sapiro. Learning to sense sparse signals: Simultaneous sensing matrix and sparsifying dictionary optimization. *IEEE Trans. Image Process.*, 18(7):1395–1408, 2009.

[22] M. Elad. Optimized projections for compressed sensing. *IEEE Trans. Signal Process.*, 55(12):5695–5702, 2007.

[23] M. Elad. *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Springer Publishing Company, Incorporated, 1st edition, 2010.

[24] K. Engan, S. O. Aase, and J. Hakon Husoy. Method of optimal directions for frame design. In *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP)*, volume 5, pages 2443–2446, 1999.

[25] T. Hong, H. Bai, S. Li, and Z. Zhu. An efficient algorithm for designing projection matrix in compressive sensing based on alternating optimization. *Signal Process.*, 125:9–20, 2016.

[26] S. M. Kay. *Fundamentals of Statistical Signal Processing*. Prentice Hall PTR, 1993.

[27] G. Li, X. Li, S. Li, H. Bai, Q. Jiang, and X. He. Designing robust sensing matrix for image compression. *IEEE Trans. Image Processing*, 24(12):5389–5400, 2015.
[28] G. Li, Z. Zhu, H. Bai, and A. Yu. A new framework for designing incoherent sparsifying dictionaries. In Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing (ICASSP), pages 4416–4420. IEEE, 2017.

[29] G. Li, Z. Zhu, D. Yang, L. Chang, and H. Bai. On projection matrix optimization for compressive sensing systems. IEEE Trans. Signal Process., 61(11):2887–2898, 2013.

[30] S. Mallat. A Wavelet Tour of Signal Processing. Academic Press, 1999.

[31] S. G. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. IEEE Trans. Signal Process., 41(12):3397–3415, 1993.

[32] D. Needell and J. A. Tropp. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. Appl. Comput. Harmonic Anal., 26(3):301–321, 2009.

[33] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: a database and web-based tool for image annotation. Int. J. Computer Vision, 77(1):157–173, 2008.

[34] J. A. Tropp. Greed is good: Algorithmic results for sparse approximation. IEEE Trans. Inf. Theory, 50(10):2231–2242, 2004.

[35] J. A. Tropp, I. S. Dhillon, R. W. Heath, and T. Strohmer. Designing structured tight frames via an alternating projection method. IEEE Trans. Inform. Theory, 51(1):188–209, 2005.

[36] E. V. Tsiligianni, L. P. Kondi, and A. K. Katsaggelos. Construction of incoherent unit norm tight frames with application to compressed sensing. IEEE Trans. Inform. Theory, 60(4):2319–2330, 2014.

[37] Y. Wang and W. Yin. Sparse signal reconstruction via iterative support detection. SIAM J. Imag. Sci., 3(3):462–491, 2010.

[38] L. Welch. Lower bounds on the maximum cross correlation of signals (corresp.). IEEE Transactions on Information Theory, 20(3):397–399, 1974.

[39] J. Xu, Y. Pi, and Z. Cao. Optimized projection matrix for compressive sensing. EURASIP J. Adv. Signal Process., 2010(1):560549, 2010.

[40] L. Zelnik-Manor, K. Rosenblum, and Y. C. Eldar. Sensing matrix optimization for block-sparse decoding. IEEE Trans. Signal Process., 59(9):4300–4312, 2011.

[41] Z. Zhu, G. Li, T. Hong, and B. Hou. On designing incoherent unit norm frames and robust compressed sensing systems. submitted to IEEE Trans. Inf. Theory.

[42] Z. Zhu and M. B. Wakin. Approximating sampled sinusoids and multiband signals using multiband modulated DPSS dictionaries. to appear in J. Fourier Anal. Appl.