

## STOCHASTIC GREEDY ALGORITHMS FOR MULTIPLE MEASUREMENT VECTORS

JING QIN\*

Department of Mathematics  
University of Kentucky, Lexington, KY 40506

SHUANG LI AND DEANNA NEEDELL

Department of Mathematics  
University of California, Los Angeles, CA 90095

ANNA MA

Department of Mathematics  
University of California, Irvine, CA 92697

RACHEL GROTHEER

Department of Mathematics  
Wofford College, Spartanburg, SC 29303

CHENXI HUANG

Center for Outcomes Research and Evaluation  
Yale University, New Haven, CT 06511

NATALIE DURGIN

Spiceworks, Austin, TX 78746

**ABSTRACT.** Sparse representation of a single measurement vector (SMV) has been explored in a variety of compressive sensing applications. Recently, SMV models have been extended to solve multiple measurement vectors (MMV) problems, where the underlying signal is assumed to have joint sparse structures. To circumvent the NP-hardness of the  $\ell_0$  minimization problem, many deterministic MMV algorithms solve the convex relaxed models with limited efficiency. In this paper, we develop stochastic greedy algorithms for solving the joint sparse MMV reconstruction problem. In particular, we propose the MMV Stochastic Iterative Hard Thresholding (MStoIHT) and MMV Stochastic Gradient Matching Pursuit (MStoGradMP) algorithms, and we also utilize the mini-batching technique to further improve their performance. Convergence analysis indicates that the proposed algorithms are able to converge faster than their SMV counterparts, i.e., concatenated StoIHT and StoGradMP, under certain conditions. Numerical experiments have illustrated the superior effectiveness of the proposed algorithms over their SMV counterparts.

---

2020 *Mathematics Subject Classification.* Primary: 68W20, 94A12; Secondary: 47N10.

*Key words and phrases.* Compressive sensing, multiple measurement vectors, stochastic optimization, joint sparsity, signal recovery, video recovery.

\* Corresponding author: Jing Qin.

**1. Introduction.** Reconstruction of sparse signals from limited measurements has been studied extensively with a variety of applications in various imaging sciences, machine learning, computer vision and so on. The major problem is to reconstruct a signal which is sparse by itself or in some transformed domain from a small number of measurements (or observations). Let  $\mathbf{x} \in \mathbb{R}^n$  be the signal to be reconstructed. Then the sparse signal reconstruction problem can be formulated as an  $\ell_0$  constrained minimization problem

$$(1) \quad \min_{\mathbf{x} \in \mathbb{R}^n} F(\mathbf{x}), \quad \text{s.t.} \quad \|\mathbf{x}\|_0 \leq k,$$

where the sparsity  $\|\mathbf{x}\|_0$  counts the number of nonzero elements in  $\mathbf{x}$ . Here  $F(\mathbf{x})$  is a loss function measuring the discrepancy between the acquired measurements  $\mathbf{y} \in \mathbb{R}^m$  ( $m \ll n$ ) and the measurements predicted by the estimated solution. In particular, if the measurements are linearly related to the underlying signal, i.e., there exists a sensing matrix  $A \in \mathbb{R}^{m \times n}$  such that  $\mathbf{y} = A\mathbf{x} + \mathbf{n}$  where  $\mathbf{n}$  is the Gaussian noise, then the least squares loss function is widely used:  $F(\mathbf{x}) = \frac{1}{2} \|\mathbf{y} - A\mathbf{x}\|_2^2$ . Then (1) becomes a single measurement vector (SMV) sparse signal reconstruction problem [11]. The choice of  $F$  depends on the generation mechanism of the measurements. Since the measurements are often generated in real time in many imaging techniques, it becomes significantly important in practice to reconstruct a collection of sparse signals, expressed as a signal matrix, from multiple measurement vectors (MMV). More precisely, the signal matrix  $X \in \mathbb{R}^{n \times L}$  with  $k$  ( $k \leq n$ ) nonzero rows can be obtained by solving the following MMV model

$$(2) \quad \min_{X \in \mathbb{R}^{n \times L}} F(X), \quad \text{s.t.} \quad \|X\|_{r,0} \leq k,$$

where  $\|X\|_{r,0}$  stands for the row-sparsity of  $X$  which counts nonzero rows in  $X$ . In the MMV literature, the column size  $L$  of  $X$  typically represents the number of snapshots of a signal. Note that it is possible that certain columns of  $X$  have more zero components than zero rows of  $X$ . The MMV sparse reconstruction problem was first introduced in magnetoencephalography (MEG) imaging [11], and has been extended to other applications [21, 4, 29, 30, 28, 13, 25].

Many SMV algorithms can be applied to solve MMV problems. The most straightforward way is to use SMV algorithms to reconstruct each signal vector sequentially or simultaneously via parallel computing, and then concatenate all resultant signals to form the estimated signal matrix. We call these types of algorithms, *concatenated SMV* algorithms. On the other hand, the MMV problem can be converted to an SMV one by column-wise stacking the unknown signal matrix  $X$  as a vector and introducing a block diagonal matrix as the new sensing matrix  $A$ . However, neither approach fully takes advantage of the joint sparse structure of the underlying signal matrix, and lack computational efficiency as well. In this paper, we develop MMV algorithms without concatenation of the SMV results or vectorization of the unknown signal matrix.

Since the  $\ell_0$  term in (1) and (2) is non-convex and non-differentiable, many classical convex optimization algorithms fail to produce a satisfactory solution. To handle the NP-hardness of the problem, many convex relaxation methods and their MMV extensions have been developed, e.g., the  $\ell_2$ -regularized M-FOCUSS [11] and the  $\ell_1$ -regularized MMV extensions of the alternating direction method of multipliers [26, 22]. By exploiting the relationship between the measurements and the correct atoms, multiple signal classification (MUSIC) [38, 17] and its improved variants [23, 24] have been developed. However, in the rank defective cases when the rank

of the measurement matrix is much smaller than the desired row-sparsity level, the MUSIC type of methods will mostly fail to identify the correct atoms. The third category of algorithms for solving the  $\ell_0$  constrained problem is the class of greedy algorithms that seek the sparsest solution by updating the support iteratively. SMV greedy algorithms include Orthogonal Matching Pursuit (OMP) [37, 39] and its RIP condition discussion [42], simultaneous OMP (S-OMP) [10, 40], Compressive Sampling Matching Pursuit (CoSaMP) [32], Regularized OMP (ROMP) [33], Subspace-Pursuit (SP) [12], Iterative Hard-Thresholding (IHT) [7], Normalized IHT [8], Hard Thresholding Pursuit (HTP) [18], gradient hard thresholding [41], and conjugate gradient IHT (CGIHT) [6]. It has been shown that CoSaMP and IHT are more efficient than the convex relaxation methods with strong recovery guarantees [32]. Greedy algorithms have also been extended to cosparse analysis to find the nearest cosparse subspace to a vector [20]. However, most of these algorithms work for compressive sensing applications where  $F$  is a least squares loss function. Recently, the Gradient Matching Pursuit (GradMP) [36] has been proposed to extend CoSaMP to handle more general loss functions. To further improve efficiency and consider the non-convex objective function case, Stochastic IHT (StoIHT) and Stochastic GradMP (StoGradMP) have been proposed [35]. Nevertheless, concatenated extension of the aforementioned SMV greedy algorithms to the MMV versions will result in limited performance especially for large data sets. Using the row sparsity, MMV extensions of SMV greedy algorithms have been developed, e.g., M-OMP, M-SP, M-CoSaMP and M-HTP in [5] and references therein. Based on countable and uncountable set of measurements, a class of models consider infinite measurements vectors (IMV) broader than MMV which have been proposed in [31].

In this paper, we propose the MMV Stochastic IHT (MStoIHT) and the MMV Stochastic GradMP (MStoGradMP) algorithms for solving the general MMV joint sparse recovery problem (2). To accelerate convergence, the mini-batching technique is applied to the proposed algorithms. We also theoretically show that under certain conditions the proposed algorithms converge faster than their SMV concatenated counterparts, i.e., CStoIHT and CStoGradMP. Note that CStoIHT and CStoGradMP may fail to converge to a joint sparse solution without the columnwise separability of the objective function. The naming rule for the algorithms follows: initial C - concatenated, initial M - MMV, initial B - batched. A large variety of numerical experiments on joint sparse matrix recovery and video sequence recovery have demonstrated the superior performance of the proposed algorithms over their SMV counterparts in terms of running time and accuracy. Please refer to [15] for our more recent work on MStoGradMP applied to solve the hyperspectral diffuse optical tomography imaging problem, which has shown the great potential of stochastic greedy algorithms on recovering large-scale data with joint sparsity.

**Organization.** The rest of the paper is organized as follows. Preliminary knowledge and notation clarifications are provided in Section 2. Section 3 presents the concatenated SMV algorithms, and the proposed stochastic greedy algorithms, i.e., MStoIHT and MStoGradMP, in detail. Section 4 discusses how to apply the mini-batching technique to accelerate the proposed algorithms. Convergence analysis is provided in Section 5. By choosing the widely used least squares loss function as  $F$ , joint sparse signal recovery in distributed compressive sensing is discussed in Section 6. Extensive numerical results are shown in Section 7. Finally, some concluding remarks are made in Section 8.

**2. Preliminaries.** To make the paper self-contained, we first introduce some useful notation and definitions, and then briefly describe the related algorithms, i.e., StoIHT and StoGradMP. Let  $[m] = \{1, 2, \dots, m\}$  and  $|\Omega|$  be the number of elements in the set  $\Omega$ . Consider a finite atom set  $\mathcal{D} = \{\mathbf{d}_1, \dots, \mathbf{d}_N\}$  (a.k.a. the dictionary) with each atom  $\mathbf{d}_i \in \mathbb{R}^n$ .

**2.1. Vector and matrix sparsity.** Assume that a vector  $\mathbf{x} \in \mathbb{R}^n$  can be written as a linear combination of  $\mathbf{d}_i$ 's, i.e.,  $\mathbf{x} = \sum_{i=1}^N \alpha_i \mathbf{d}_i = D\alpha$  with  $D = [\mathbf{d}_1, \dots, \mathbf{d}_N]$  and  $\alpha = (\alpha_1, \dots, \alpha_N)^T$ . Then the support of  $\mathbf{x}$  with respect to  $\alpha$  and  $\mathcal{D}$  is defined by

$$\text{supp}_{\alpha, \mathcal{D}}(\mathbf{x}) = \{i \in [N] : \alpha_i \neq 0\} := \text{supp}(\alpha).$$

The  $\ell_0$ -norm of  $\mathbf{x}$  with respect to  $\mathcal{D}$  is defined as the minimal support size

$$\|\mathbf{x}\|_{0, \mathcal{D}} = \min_{\alpha} \{ |T| : \mathbf{x} = \sum_{i \in T} \alpha_i \mathbf{d}_i, T \subseteq [N] \}.$$

Since absolute homogeneity does not hold in general, i.e.,  $\|\gamma \mathbf{x}\|_{0, \mathcal{D}} = |\gamma| \|\mathbf{x}\|_{0, \mathcal{D}}$  holds if and only if  $|\gamma| = 1$ , this  $\ell_0$ -norm is not a norm. Here the smallest support  $\text{supp}_{\alpha, \mathcal{D}}(\mathbf{x})$  is called the support of  $\mathbf{x}$  with respect to  $\mathcal{D}$ , denoted by  $\text{supp}_{\mathcal{D}}(\mathbf{x})$ . Thus

$$|\text{supp}_{\mathcal{D}}(\mathbf{x})| = \|\mathbf{x}\|_{0, \mathcal{D}}.$$

Note that the support may not be unique if  $\mathcal{D}$  is over-complete in that there could be multiple representations of  $\mathbf{x}$  with respect to the atom set  $\mathcal{D}$  due to the linear dependence of the atoms in  $\mathcal{D}$ . In this paper, we assume that  $\mathcal{D}$  is chosen so that the support of any arbitrary vector is unique, e.g.,  $\mathcal{D}$  is chosen as the standard basis in a Euclidean space or a learned basis. Furthermore, Donoho defines a general class of *unconditional basis* in [14], e.g., the wavelet basis, which requires the unique representation and thereby is also applicable to our MMV algorithms. Without this uniqueness representation guarantee, we could either restrict the solution set or take a transformation of the unknown variable.

The vector  $\mathbf{x}$  is called *k-sparse* with respect to  $\mathcal{D}$  if  $\|\mathbf{x}\|_{0, \mathcal{D}} \leq k$ . For a subset  $\Gamma \subseteq [N]$ , we denote the set of atoms from  $\mathcal{D}$  with indices restricted to  $\Gamma$  by  $\mathcal{D}_{\Gamma}$ , and denote the subspace of  $\mathbb{R}^n$  spanned by the atoms in  $\mathcal{D}_{\Gamma}$  by  $\mathcal{R}_{\mathcal{D}_{\Gamma}}$ . Given a vector  $\mathbf{w} \in \mathbb{R}^n$ , we use  $\mathcal{P}_{\Gamma} \mathbf{w}$  to denote the orthogonal projection of  $\mathbf{w}$  onto the subspace  $\mathcal{R}(\mathcal{D}_{\Gamma})$ .

We extend vector sparsity and define the row sparsity for a matrix  $X \in \mathbb{R}^{n \times L}$  as follows

$$\|X\|_{r, 0, \mathcal{D}} = \min_{\Omega} \{ |\Omega| : \Omega = \bigcup_{j=1}^L \text{supp}_{\mathcal{D}}(X_{\cdot, j}) \},$$

where  $X_{\cdot, j}$  is the  $j$ -th column of  $X$ . Here the minimal common support  $\Omega$  is called the (row-wise) *joint support* of  $X$  with respect to  $\mathcal{D}$ , denoted by  $\text{supp}_{\mathcal{D}}^r(X)$ , which satisfies  $|\text{supp}_{\mathcal{D}}^r(X)| = \|X\|_{r, 0, \mathcal{D}}$ . The matrix  $X$  is called *k*-row sparse with respect to  $\mathcal{D}$  if all columns of  $X$  share a joint support of size at most  $k$  with respect to  $\mathcal{D}$ , i.e.,  $\|X\|_{r, 0, \mathcal{D}} \leq k$ .

**2.2. Functions defined on a matrix space.** Given a function  $f : \mathbb{R}^{n \times L} \rightarrow \mathbb{R}$ , the matrix derivative is defined by concatenating gradients [27]

$$(3) \quad \frac{\partial f}{\partial X} = \left[ \frac{\partial f}{\partial X_{i, j}} \right]_{n \times L} = [\nabla_{X_{\cdot, 1}} f \quad \cdots \quad \nabla_{X_{\cdot, L}} f],$$

where  $X_{i,j}$  is the  $(i,j)$ -th entry of  $X$ . Notice that

$$\|X\|_F^2 = \sum_{i=1}^n \|X_{i,\cdot}\|_2^2 = \sum_{j=1}^L \|X_{\cdot,j}\|_2^2 = \text{Tr}(X^T X),$$

where  $X_{i,\cdot}$  is the  $i$ -th row vector of  $X$ , and  $\text{Tr}(\cdot)$  is the trace operator to add up all the diagonal entries of a matrix. The inner product for any two matrices  $U, V \in \mathbb{R}^{n \times L}$  is defined as  $\langle U, V \rangle = \text{Tr}(U^T V)$ . Note that the equality

$$\|U + V\|_F^2 = \|U\|_F^2 + \|V\|_F^2 + 2\langle U, V \rangle$$

and the Cauchy-Schwartz inequality

$$\langle U, V \rangle \leq \|U\|_F \|V\|_F$$

still hold. By generalizing the concepts in [35], we define the  $\mathcal{D}$ -restricted strong convexity property and the strong smoothness property (a.k.a. the Lipschitz condition on the gradient) for the functions defined on a matrix space. It is worth noting the variants of these two concepts have been used to study the convergence of the projected gradient descent algorithm [1].

**Definition 2.1.** The function  $f : \mathbb{R}^{n \times L} \rightarrow \mathbb{R}$  satisfies the  $\mathcal{D}$ -restricted strong convexity ( $\mathcal{D}$ -RSC) if there exists  $\rho_k^- > 0$  such that

$$(4) \quad f(X') - f(X) - \left\langle \frac{\partial f}{\partial X}(X), X' - X \right\rangle \geq \frac{\rho_k^-}{2} \|X' - X\|_F^2$$

for  $X', X \in \mathbb{R}^{n \times L}$  with  $|\text{supp}_{\mathcal{D}}^r(X) \cup \text{supp}_{\mathcal{D}}^r(X')| \leq k$ .

**Definition 2.2.** The function  $f : \mathbb{R}^{n \times L} \rightarrow \mathbb{R}$  satisfies the  $\mathcal{D}$ -restricted strong smoothness ( $\mathcal{D}$ -RSS) if there exists  $\rho_k^+ > 0$  such that

$$(5) \quad \left\| \frac{\partial f}{\partial X}(X) - \frac{\partial f}{\partial X}(X') \right\|_F \leq \rho_k^+ \|X - X'\|_F$$

for  $X', X \in \mathbb{R}^{n \times L}$  with  $|\text{supp}_{\mathcal{D}}^r(X) \cup \text{supp}_{\mathcal{D}}^r(X')| \leq k$ .

**2.3. Related work.** Many stochastic greedy algorithms based on matching pursuit and hard thresholding have been proposed to solve the SMV problem. In particular, GradMP generalizes CoSaMP to handle general convex objective functions [36]. Motivated by the success of stochastic convex optimization, StoIHT (see Algorithm 1) and StoGradMP (see Algorithm 2) have been later proposed to solve the  $\ell_0$  constrained SMV problem [35].

$$(6) \quad \min_{\mathbf{x} \in \mathbb{R}^n} \frac{1}{M} \sum_{i=1}^M \tilde{f}_i(\mathbf{x}), \quad \text{subject to} \quad \|\mathbf{x}\|_{0,\mathcal{D}} \leq k,$$

At each iteration of StoIHT, one component function  $\tilde{f}_i : \mathbb{R}^n \rightarrow \mathbb{R}$  is first randomly selected with probability  $p(i)$ . Here the input discrete probability distribution  $p(i)$ 's satisfy  $\sum_{i=1}^M p(i) = 1$ , and  $p(i) \geq 0$ ,  $i = 1, \dots, M$ . Next, in the “Proxy” step, gradient descent along the selected component is performed. Then the last two steps, i.e., “Identify” and “Estimate”, essentially project the gradient descent result to its best  $k$ -sparse approximation. Given  $\mathbf{w} = (w_1, \dots, w_n)^T$  and  $\eta \geq 1$ , the best

$k$ -sparse approximation operator acted on  $\mathbf{w}$  and  $\eta$ , denoted by  $\text{approx}_k(\mathbf{w}, \eta)$ , constructs an index set  $\Gamma$  with  $|\Gamma| = k$  such that

$$(7) \quad \|\mathcal{P}_\Gamma \mathbf{w} - \mathbf{w}\|_2 \leq \eta \|\mathbf{w} - \mathbf{w}_{(k)}\|_2, \quad \mathbf{w}_{(k)} = \operatorname*{argmin}_{\substack{\mathbf{y} \in \mathcal{R}(\mathcal{D}_\Gamma) \\ |\Gamma| \leq k}} \|\mathbf{w} - \mathbf{y}\|_2.$$

Thus, if  $\Gamma^* = \operatorname*{argmin}_{\Gamma: |\Gamma| \leq k} \|\mathbf{w} - \mathcal{P}_\Gamma \mathbf{w}\|_2$ , we have  $\|\mathbf{w} - \mathcal{P}_{\Gamma^*} \mathbf{w}\|_2 \leq \eta \|\mathbf{w} - \mathcal{P}_{\Gamma^*} \mathbf{w}\|_2$ , and  $\mathbf{w}_{(k)}$  defined in (7) becomes the best  $k$ -sparse approximation of  $\mathbf{w}$  in the subspace  $\mathcal{R}(\mathcal{D}_{\Gamma^*})$ . In particular, if  $\eta \geq 1$  and  $\mathcal{D} = \{\mathbf{e}_i : i = 1, 2, \dots, n\}$  with  $\mathbf{e}_i = [0, \dots, \underset{(i)}{1}, \dots, 0]^T$ , then  $\text{approx}_k(\mathbf{w}, \eta)$  returns the index set of the first  $k$  largest entries of  $\mathbf{w}$  in absolute value, i.e.,

$$\text{approx}_k(\mathbf{w}, 1) = \{i_1, \dots, i_k : |w_{i_1}| \geq \dots \geq |w_{i_k}| \geq \dots \geq |w_{i_n}|\} := \widehat{\Gamma}.$$

Then the projection  $\mathcal{P}_\Gamma \mathbf{w}$  reads as in componentwise form

$$(\mathcal{P}_{\widehat{\Gamma}}(\mathbf{w}))_j = \begin{cases} w_j & \text{if } j \in \widehat{\Gamma}, \\ 0 & \text{if } j \notin \widehat{\Gamma}. \end{cases}$$

There are two widely used stopping criteria:  $\frac{\|\mathbf{x}^{t+1} - \mathbf{x}^t\|_2}{\|\mathbf{x}^t\|_2} < \varepsilon$ , and  $\frac{1}{M} \sum_{i=1}^M \tilde{f}_i(\mathbf{x}^t) < \varepsilon$ , where  $\varepsilon > 0$  is a small tolerance. It is well known that the first stopping criteria is more robust in practice [9].

---

**Algorithm 1** Stochastic Iterative Hard Thresholding (StoIHT)

---

**Input:**  $k, \gamma, \eta, p(i), \varepsilon$ .

**Output:**  $\widehat{\mathbf{x}} = \mathbf{x}^t$ .

**Initialize:**  $\mathbf{x}^0 = \mathbf{0}$ .

**for**  $t = 0, 1, 2, \dots, T$  **do**

    Randomly select an index  $i_t \in \{1, 2, \dots, M\}$  with probability  $p(i_t)$

    Proxy:  $\mathbf{b}^t = \mathbf{x}^t - \frac{\gamma}{Mp(i_t)} \nabla \tilde{f}_{i_t}(\mathbf{x}^t)$

    Identify:  $\Gamma^t = \text{approx}_k(\mathbf{b}^t, \eta)$ .

    Estimate:  $\mathbf{x}^{t+1} = \mathcal{P}_{\Gamma^t}(\mathbf{b}^t)$ .

    If the stopping criteria are met, exit.

**end for**

---

Different from StoIHT, StoGradMP involves gradient matching, i.e., seeking the best  $k$ -sparse approximation of the gradient rather than the solution. At the solution estimation step, the original problem is restricted to the components from the estimated support. It has been empirically shown that StoGradMP converges faster than StoIHT due to the more accurate estimation of the support [35]. But StoGradMP requires that the sparsity level  $k$  is no more than  $n/2$ .

**3. Proposed stochastic greedy algorithms.** In this section, we present concatenated SMV algorithms, and develop stochastic greedy algorithms for MMV problems based on StoIHT and StoGradMP. Suppose that there are  $M$  differentiable and convex functions  $f_i : \mathbb{R}^{n \times L} \rightarrow \mathbb{R}$  that satisfy the  $\mathcal{D}$ -restricted strong smoothness property (see Definition 2.2), and their mean

$$(8) \quad F(X) = \frac{1}{M} \sum_{i=1}^M f_i(X)$$

**Algorithm 2** Stochastic Gradient Matching Pursuit (StoGradMP)

---

**Input:**  $k, \eta_1, \eta_2, p(i), \varepsilon$ .  
**Output:**  $\hat{\mathbf{x}} = \mathbf{x}^t$ .  
**Initialize:**  $\mathbf{x}^0 = \mathbf{0}, \Lambda = \emptyset$ .  
**for**  $t = 0, 1, 2, \dots, T$  **do**  
    Randomly select an index  $i_t \in \{1, 2, \dots, M\}$  with probability  $p(i_t)$   
    Calculate the gradient  $\mathbf{r}^t = \nabla \tilde{f}_{i_t}(\mathbf{x}^t)$   
     $\Gamma = \text{approx}_{2k}(\mathbf{r}^t, \eta_1)$   
     $\hat{\Gamma} = \Gamma \cup \Lambda$   
     $\mathbf{b}^t = \text{argmin}_{\mathbf{x}} \frac{1}{M} \sum_{i=1}^M \tilde{f}_i(\mathbf{x}), \quad \mathbf{x} \in \mathcal{R}(\mathcal{D}_{\hat{\Gamma}})$   
     $\Lambda = \text{approx}_k(\mathbf{b}^t, \eta_2)$   
     $\mathbf{x}^{t+1} = \mathcal{P}_{\Lambda}(\mathbf{b}^t)$   
    If the stopping criteria are met, exit.  
**end for**

---

satisfies the  $\mathcal{D}$ -restricted strong convexity property (see Definition 2.1). These assumptions will be used extensively throughout the entire paper. Consider the row-sparsity constrained MMV problem

$$(9) \quad \min_{X \in \mathbb{R}^{n \times L}} \frac{1}{M} \sum_{i=1}^M f_i(X), \quad \text{subject to} \quad \|X\|_{r,0,\mathcal{D}} \leq k.$$

By vectorizing  $X$ , i.e., rewriting  $X$  as a vector  $\mathbf{x} \in \mathbb{R}^{nL}$  by columnwise stacking, we can *relax* (9) to a sparsity constrained SMV problem of the form (6) where the sparsity level  $k$  is replaced by  $kL$ . Since  $\|\mathbf{x}\|_{0,\mathcal{D}} \leq kL$  does not necessarily guarantee  $\|X\|_{r,0,\mathcal{D}} \leq k$ , the solution to the relaxed problem may not be the same as the vectorization of the solution to (9). On the other hand, the iterative stochastic algorithms such as StoIHT and StoGradMP, can be developed to the *concatenated* versions for solving (9) under the following assumption on the objective function  $f_i$ 's.

**Definition 3.1.** Each component  $f_i$  of the objective function  $F$  in (9) is columnwise separable, in the sense that a collection of functions  $g_{i,j} : \mathbb{R}^n \rightarrow \mathbb{R}$  exist with

$$(10) \quad f_i(X) = \sum_{j=1}^L g_{i,j}(X_{\cdot,j}), \quad i = 1, \dots, M.$$

In this case, the objective function can be rewritten as

$$(11) \quad F(X) = \sum_{j=1}^L \left( \frac{1}{M} \sum_{i=1}^M g_{i,j}(X_{\cdot,j}) \right) := \sum_{j=1}^L \widehat{F}_j(X_{\cdot,j}),$$

which implies that minimizing  $F$  with respect to  $X$  can be reduced to minimizing  $\widehat{F}_j$  for  $j = 1, \dots, L$  individually.

Under this assumption, the concatenated algorithms, i.e., CStoIHT in Algorithm 3 and CStoGradMP in Algorithm 4, can be applied to solve (9), which essentially reconstruct each column of  $X$  by solving the SMV problem (6). Notice that the outer loops of CStoIHT and CStoGradMP can be executed in a parallel manner on a multi-core computer, when the order of the inner loop and the outer loop in Algorithm 3 can be swapped. Note that although a joint support is not

enforced at each iteration, the paralleled version of CStoIHT and CStoGradMP can still yield joint sparse results with guaranteed convergence (see Theorem 5.5 and Theorem 5.9). However, if the sparsity level  $k$  is very large, then the support sets of  $X_{\cdot,j}$ 's are prone to overlap less initially which results in the less accurate estimation of the joint support and larger errors in the initial iterates. In addition, for some nonlinear function  $f_i(X)$  which can not be separated as a sum of functions for columns of  $X$ , it will be challenging to find an appropriate objective function for the corresponding SMV problem.

---

**Algorithm 3** Concatenated Stochastic Iterative Hard Thresholding (CStoIHT)

---

**Input:**  $k, \gamma, \eta, p(i), \varepsilon$ .  
**Output:**  $\hat{X} = X^t$ .  
**Initialize:**  $X^0 = \mathbf{0} \in \mathbb{R}^{n \times L}$ .  
**for**  $j = 1, \dots, L$  **do**  
  **for**  $t = 0, 1, 2, \dots, T$  **do**  
    Randomly select an index  $i_t \in \{1, 2, \dots, M\}$  with probability  $p(i_t)$   
    Proxy:  $\mathbf{b}^t = X_{\cdot,j}^t - \frac{\gamma}{Mp(i_t)} \nabla g_{i_t,j}(X_{\cdot,j}^t)$   
    Identify:  $\Gamma^t = \text{approx}_k(\mathbf{b}^t, \eta)$ .  
    Estimate:  $X_{\cdot,j}^{t+1} = \mathcal{P}_{\Gamma^t}(\mathbf{b}^t)$ .  
    If the stopping criteria are met, exit.  
  **end for**  
**end for**

---

**Algorithm 4** Concatenated Stochastic Gradient Matching Pursuit (CStoGradMP)

---

**Input:**  $k, \eta_1, \eta_2, p(i), \varepsilon$ .  
**Output:**  $\hat{X} = X^t$ .  
**Initialize:**  $X^0 = \mathbf{0} \in \mathbb{R}^{n \times L}, \Lambda = \emptyset$ .  
**for**  $j = 1, \dots, L$  **do**  
  **for**  $t = 0, 1, 2, \dots, T$  **do**  
    Randomly select an index  $i_t \in \{1, 2, \dots, M\}$  with probability  $p(i_t)$   
    Calculate the gradient  $\mathbf{r}^t = \nabla g_{i_t,j}(X_{\cdot,j}^t)$   
     $\Gamma = \text{approx}_{2k}(\mathbf{r}^t, \eta_1)$   
     $\hat{\Gamma} = \Gamma \cup \Lambda$   
     $\mathbf{b}^t = \text{argmin}_{\mathbf{x}} \frac{1}{M} \sum_{i=1}^M g_{i,j}(\mathbf{x}), \quad \mathbf{x} \in \mathcal{R}(\mathcal{D}_{\hat{\Gamma}})$   
     $\Lambda = \text{approx}_k(\mathbf{b}^t, \eta_2)$   
     $X_{\cdot,j}^{t+1} = \mathcal{P}_{\Lambda}(\mathbf{b}^t)$   
    If the stopping criteria are met, exit.  
  **end for**  
**end for**

---

To circumvent the aforementioned issues, we first propose the MMV Stochastic Iterative Hard Thresholding algorithm (MStoIHT) detailed in Algorithm 5. Compared to StoIHT, MStoIHT replaces the gradient by the matrix derivative (3). The second significant difference lies in the “Identify” and “Estimate” steps, especially the operator  $\text{approx}_k(\cdot, \cdot)$ . Now we extend the operator  $\text{approx}_k(\cdot, \cdot)$  from sparse vectors to row-sparse matrices. Given  $X \in \mathbb{R}^{n \times L}$  and  $\eta \geq 1$ , the best  $k$ -row sparse

approximation operator acted on  $X$  and  $\eta$ , denoted by  $\text{approx}_k^r(X, \eta)$ , constructs a *row* index set  $\Gamma$  such that

$$\|\mathcal{P}_\Gamma X_{\cdot, j} - X_{\cdot, j}\|_2 \leq \eta \|X_{\cdot, j} - (X_{\cdot, j})_k\|_2, \quad j = 1, \dots, L,$$

where  $(X_{\cdot, j})_k$  is the best  $k$ -sparse approximation of the column vector  $X_{\cdot, j}$  with respect to  $\mathcal{D}$ . Note that we can choose  $\Gamma$  as the union of all  $\Gamma_j$ 's if  $\Gamma_j$  is the index set satisfying the inequality with a specified index  $j$ . In particular, if  $\mathcal{D} = \{\mathbf{e}_i : i = 1, \dots, n\}$  and  $\eta = 1$ ,  $\text{approx}_k^r(X, 1)$  returns the row index set of the first  $k$  largest  $\ell_2$  row norms in  $X$ , i.e.,

$$\text{approx}_k^r(X, 1) = \{i_1, \dots, i_k : \|X_{i_1, \cdot}\|_2 \geq \dots \geq \|X_{i_n, \cdot}\|_2\} := \tilde{\Gamma}.$$

By abusing the notation, we define  $\mathcal{P}_{\tilde{\Gamma}}(X)$  to be the projection of  $X$  onto the subspace of all row-sparse matrices with row indices restricted to  $\tilde{\Gamma}$ . Therefore, we have  $\mathcal{P}_{\tilde{\Gamma}}X = [\mathcal{P}_{\tilde{\Gamma}}X_{\cdot, 1} \ \dots \ \mathcal{P}_{\tilde{\Gamma}}X_{\cdot, L}]$ . Due to the common support  $\tilde{\Gamma}$ , the projection  $\mathcal{P}_{\tilde{\Gamma}}(X)$  can also be written as in row-wise form

$$(12) \quad (\mathcal{P}_{\tilde{\Gamma}}(X))_{j, \cdot} = \begin{cases} X_{j, \cdot} & \text{if } j \in \tilde{\Gamma}, \\ \mathbf{0} & \text{if } j \notin \tilde{\Gamma}. \end{cases}$$

Here  $\mathcal{P}_{\tilde{\Gamma}}(X)$  returns a  $k$ -row sparse matrix, whose nonzero rows correspond to the  $k$  rows of  $X$  with largest  $\ell_2$  row norms.

---

**Algorithm 5** MMV Stochastic Iterative Hard Thresholding (MStoIHT)

---

**Input:**  $k, \gamma, \eta, p(i), \varepsilon$ .  
**Output:**  $\hat{X} = X^t$ .  
**Initialize:**  $X^0 = \mathbf{0}$ .  
**for**  $t = 0, 1, 2, \dots, T$  **do**  
    Randomly select an index  $i_t \in \{1, 2, \dots, M\}$  with probability  $p(i_t)$   
    Proxy:  $B^t = X^t - \frac{\gamma}{Mp(i_t)} \frac{\partial f_{i_t}(X^t)}{\partial X}$   
    Identify:  $\Gamma^t = \text{approx}_k^r(B^t, \eta)$ .  
    Estimate:  $X^{t+1} = \mathcal{P}_{\Gamma^t}(B^t)$ .  
    If the stopping criteria are met, exit.  
**end for**

---

Next, we propose the MMV Stochastic Gradient Matching Pursuit (MStoGrad MP) detailed in Algorithm 6, where the two gradient matching steps involve the operator  $\text{approx}_k^r(\cdot, \cdot)$ . The stopping criteria in all proposed algorithms can be set as the same as those in Algorithm 1 and Algorithm 2. Here we choose the stopping criterion:  $\|X^{t+1} - X^t\|_F / \|X^t\|_F < \varepsilon$ , where  $\varepsilon > 0$  is a small tolerance.

**4. Batched acceleration.** To accelerate computations and improve performance, we apply the mini-batching technique to obtain batched variants of Algorithms 5 and 6. We first partition the index set  $\{1, 2, \dots, M\}$  into a collection of equal-sized batches  $\tau_1, \dots, \tau_d$  where the batch size  $|\tau_i| = b$  for all  $i = 1, 2, \dots, \lceil M/b \rceil := d$ . For simplicity, we assume that  $M/b$  is an integer. Similar to the approach in [34], we reformulate (8) as

$$(13) \quad F(X) = \frac{1}{d} \sum_{i=1}^d \left( \frac{1}{b} \sum_{j \in \tau_i} f_j(X) \right) := \frac{1}{d} \sum_{i=1}^d \hat{f}_i(X).$$

**Algorithm 6** MMV Stochastic Gradient Matching Pursuit (MStoGradMP)

---

**Input:**  $k, \eta_1, \eta_2, p(i), \varepsilon$ .  
**Output:**  $\hat{X} = X^t$ .  
**Initialize:**  $X^0 = \mathbf{0}$ ,  $\Lambda = \emptyset$ .  
**for**  $t = 0, 1, 2, \dots, T$  **do**  
    Randomly select an index  $i_t \in \{1, 2, \dots, M\}$  with probability  $p(i_t)$   
    Calculate the generalized gradient  $R^t = \frac{\partial f_{i_t}(X^t)}{\partial X}$   
     $\Gamma = \text{approx}_{2k}^r(R^t, \eta_1)$   
     $\hat{\Gamma} = \Gamma \cup \Lambda$   
     $B^t = \arg\min_X F(X)$ ,  $X \in \mathcal{R}(\mathcal{D}_{\hat{\Gamma}})$   
     $\Lambda = \text{approx}_k^r(B^t, \eta_2)$   
     $X^{t+1} = \mathcal{P}_{\Lambda}(B^t)$   
    If the stopping criteria are met, exit.  
**end for**

---

That is,  $\hat{f}_i$  is the average of the  $i$ -th batch of component functions  $\{f_j\}_{j \in \tau_i}$ . Based on this new formulation, we get the batched version of Algorithm 5, which is termed as BMStoIHT, described in Algorithm 7. Here the input probability  $p(i)$  satisfies

$$\frac{1}{d} \sum_{i=1}^d p(i) = 1 \quad \text{and} \quad p(i) \geq 0, \quad i = 1, \dots, d.$$

Likewise, we get a batched version of MStoGradMP, termed as BMStoGradMP, which is detailed in Algorithm 8. Note that  $\hat{f}_{\tau_t}$  in Algorithms 7 and 8 is the  $\tau_t$ -th component function defined in (13), and  $\frac{\partial \hat{f}_{\tau_t}(X^t)}{\partial X}$  is the derivative of  $\hat{f}_{\tau_t}$  defined in (3). It is empirically shown in Section 7 that the increase of the batch size greatly speeds up the convergence of both algorithms, which is more obvious in BMStoIHT. As a by-product, mini-batching can also improve the recovery accuracy based on our experiments. However, there is a trade-off between the batch size and the performance improvement as mentioned in [34].

**Algorithm 7** Batched MMV Stochastic Iterative Hard Thresholding (BMStoIHT)

---

**Input:**  $k, \gamma, \eta, b$  and  $p(i)$ .  
**Output:**  $\hat{X} = X^t$ .  
**Initialize:**  $X^0 = \mathbf{0}$ .  
**for**  $t = 0, 1, 2, \dots, T$  **do**  
    Randomly select a batch index  $\tau_t \subseteq \{1, 2, \dots, d\}$  of size  $b$  with probability  $p(\tau_t)$   
    Proxy:  $B^t = X^t - \frac{\gamma}{dp(\tau_t)} \frac{\partial \hat{f}_{\tau_t}(X^t)}{\partial X}$   
    Identify  $\Gamma^t = \text{approx}_k^r(B^t, \eta)$ .  
    Estimate  $X^{t+1} = \mathcal{P}_{\Gamma^t}(B^t)$ .  
    If the stopping criteria are met, exit.  
**end for**

---

**5. Convergence analysis.** In this section, we provide the convergence guarantees for the proposed MStoIHT and MStoGradMP, together with their SMV counterparts, i.e., CStoIHT and CStoGradMP. To simplify the discussion, the result at the

---

**Algorithm 8** Batched MMV Stochastic Gradient Matching Pursuit (BM-StoGradMP)

---

**Input:**  $k, \eta_1, \eta_2, b$  and  $p(i)$ .  
**Output:**  $\hat{X} = X^t$ .  
**Initialize:**  $X^0 = \mathbf{0}, \Lambda = \emptyset$ .  
**for**  $t = 0, 1, 2, \dots, T$  **do**  
    Randomly select a batch index  $\tau_t \subseteq \{1, 2, \dots, d\}$  of size  $b$  with probability  $p(\tau_t)$   
    Calculate the generalized gradient  $R^t = \frac{\partial \hat{f}_{\tau_t}(X^t)}{\partial X}$   
     $\Gamma = \text{approx}_{2k}^r(R^t, \eta_1)$   
     $\hat{\Gamma} = \Gamma \cup \Lambda$   
     $B^t = \arg\min_X F(X), \quad X \in \text{span}(D_{\hat{\Gamma}})$   
     $\Lambda = \text{approx}_k^r(B^t, \eta_2)$   
     $X^{t+1} = \mathcal{P}_{\Lambda}(B^t)$   
    If the stopping criteria are met, exit.  
**end for**

---

$t$ -th iteration of CStoIHT/CStoGradMP refers to the result obtained after  $t$  inner iterations and  $L$  outer iterations of Algorithm 3/Algorithm 4, or equivalently the maximum number of inner iterations is set as  $t$ . Furthermore, all convergence results can be extended to their batched versions, i.e., BStoIHT and BMStoGradMP. Comparison of contraction coefficients shows that the proposed algorithms have faster convergence under the  $\mathcal{D}$ -RSC,  $\mathcal{D}$ -RSS and columnwise separability of the objective function (see Definition 3.1). Similar to Section 3, we consider the model (9) under the following assumptions.

**Assumption 1.** The objective function  $F(X) = \frac{1}{M} \sum_{i=1}^M f_i(X) : \mathbb{R}^{n \times L} \rightarrow \mathbb{R}$  satisfies

(a)  $f_i$  is columnwise separable with differentiable and convex functions  $g_{i,j}$  that satisfy (10) and  $\mathcal{D}$ -RSS in vector form, i.e., there exists  $\rho_k^+(i, j) > 0$  such that

$$(14) \quad \|\nabla g_{i,j}(\mathbf{w}) - \nabla g_{i,j}(\hat{\mathbf{w}})\|_2 \leq \rho_k^+(i, j) \|\mathbf{w} - \hat{\mathbf{w}}\|_2$$

for all vectors  $\mathbf{w}, \hat{\mathbf{w}} \in \mathbb{R}^n$  with  $|\text{supp}_{\mathcal{D}}(\mathbf{w}) \cup \text{supp}_{\mathcal{D}}(\hat{\mathbf{w}})| \leq k$ .

(b)  $\widehat{F}_j = \sum_i g_{i,j}$  satisfies  $\mathcal{D}$ -RSC in vector form, i.e., there exists  $\rho_{k,j}^- > 0$  such that

$$(15) \quad \widehat{F}_j(\hat{\mathbf{w}}) - \widehat{F}_j(\mathbf{w}) - \langle \nabla \widehat{F}_j(\mathbf{w}), \hat{\mathbf{w}} - \mathbf{w} \rangle \geq \frac{\rho_{k,j}^-}{2} \|\hat{\mathbf{w}} - \mathbf{w}\|_2^2$$

for all vectors  $\mathbf{w}, \hat{\mathbf{w}} \in \mathbb{R}^n$  with  $|\text{supp}_{\mathcal{D}}(\mathbf{w}) \cup \text{supp}_{\mathcal{D}}(\hat{\mathbf{w}})| \leq k$ .

Similar to [35], the  $\mathcal{D}$ -RSS property (a) and  $\mathcal{D}$ -RSC property (b) in this assumption can guarantee the uniqueness of the solution to (9).

**Lemma 5.1.** If Assumption (a) is satisfied, then  $f_i$  satisfies  $\mathcal{D}$ -RSS with constant  $\mu_k^+(i) = \max_{1 \leq j \leq L} \rho_k^+(i, j)$  for  $i = 1, \dots, M$ .

**Lemma 5.2.** If Assumption (b) is satisfied, then  $F(X)$  satisfies  $\mathcal{D}$ -RSC with constant  $\mu_k^- = \min_{1 \leq j \leq L} \rho_{k,j}^-$ .

The proofs of Lemma 5.1 and Lemma 5.2 are straightforward, which can be found in the Appendix. Throughout the paper, we denote

$$(16) \quad \begin{aligned} \alpha_{k,j} &= \max_{1 \leq i \leq M} \frac{\rho_k^+(i,j)}{Mp(i)}, \\ \rho_{k,j}^+ &= \max_{1 \leq i \leq M} \rho_k^+(i,j), \\ \bar{\rho}_{k,j}^+ &= \frac{1}{M} \sum_{i=1}^M \rho_k^+(i,j), \end{aligned}$$

where  $\rho_k^+(i,j)$  is the  $\mathcal{D}$ -RSS constant for  $g_{i,j}$  defined in Assumption (a).

**5.1. MStoIHT.** By replacing the  $\ell_2$ -norm and vector inner product with the Frobenius norm and the matrix inner product respectively and using the properties of inner product in Section 2, we get similar convergence results for MStoIHT in [35] in the following theorem. Here we skip the proof which is a straightforward matrix extension of [35, Appendix B].

**Theorem 5.3** (MStoIHT). *Let  $X^*$  be a feasible solution of (9) and  $X^0$  be the initial solution. Assume that  $F(X) = \frac{1}{M} \sum_{i=1}^M f_i(X)$  satisfies the  $\mathcal{D}$ -RSC with constant  $\rho_k^-$  in (4) and  $f_i$  satisfies  $\mathcal{D}$ -RSS with constant  $\rho_k^+(i)$  in (5). Then at the  $(t+1)$ -th iteration, the expectation of the recovery error of Algorithm 5 is bounded by*

$$(17) \quad E \|X^{t+1} - X^*\|_F \leq \kappa^{t+1} \|X^0 - X^*\|_F + \frac{\sigma_{X^*}}{1 - \kappa},$$

where the contraction coefficient  $\kappa$  and the tolerance parameter  $\sigma_{X^*}$  are

$$(18) \quad \begin{aligned} \kappa &= 2\sqrt{1 - \gamma(2 - \gamma\alpha_{3k})\rho_{3k}^-} + \sqrt{(\eta^2 - 1)\left(1 + \gamma^2\alpha_{3k}\sum_{i=1}^M \rho_{3k}^+(i) - 2\gamma\rho_{3k}^-\right)}, \\ \sigma_{X^*} &= \frac{\gamma}{\min_{1 \leq i \leq M} Mp(i)} \left(2E_i \max_{|\Omega| \leq 3k} \left\| \mathcal{P}_\Omega \frac{\partial f_i}{\partial X}(X^*) \right\|_F + \sqrt{\eta^2 - 1}E_i \left\| \frac{\partial f_i}{\partial X}(X^*) \right\|_F\right). \end{aligned}$$

Here  $\mathcal{P}_\Omega$  is defined in (12) and

$$(19) \quad \alpha_k = \max_{1 \leq i \leq M} \frac{\rho_k^+(i)}{Mp(i)}, \quad \rho_k^+ = \max_{1 \leq i \leq M} \rho_k^+(i), \quad \bar{\rho}_k^+ = \frac{1}{M} \sum_{i=1}^M \rho_k^+(i).$$

Thus Algorithm 7 converges linearly if  $\kappa < 1$ . In particular, if  $\eta = \gamma = 1$ , then

$$(20) \quad \kappa = 2\sqrt{1 - 2\rho_{3k}^- + \alpha_{3k}\rho_{3k}^-}.$$

**Remark.** From (18), one can see that the tolerance parameter  $\sigma_{X^*}$  increases as the step size  $\gamma$  grows which implies that the step size cannot be too large. The contraction coefficient  $\kappa$  which controls the convergence speed depends on the  $\mathcal{D}$ -RSC and  $\mathcal{D}$ -RSS constants. In addition, in the simplest case when  $\eta = \gamma = 1$  and  $p(i)$  is a uniform distribution, one can see that  $\kappa$  decreases if either the  $\mathcal{D}$ -RSC constant  $\rho_{3k}^-$  increases or the maximum of all  $\mathcal{D}$ -RSS constants  $\rho_{3k}^+(i)$  decreases according to (20). That implies that the proposed algorithms will converge fast when each component function of the objective has either a small  $\mathcal{D}$ -RSS constant or a large  $\mathcal{D}$ -RSC constant in this case.

**5.2. CStoIHT.** To solve the problem (9), CStoIHT uses StoIHT to reconstruct each column of  $X$  separately and then concatenate all column vectors to form a matrix. To analyze the convergence of CStoIHT, we first derive an upper bound for  $E \|X_{\cdot,j} - X_{\cdot,j}^*\|_2^2$  following the proof outline in [35, Appendix B]. Note that in our proof we look for the contraction coefficient of  $E \|X_{\cdot,j} - X_{\cdot,j}^*\|_2^2$  rather than  $E \|X_{\cdot,j} - X_{\cdot,j}^*\|_2$  as in [35].

**Lemma 5.4.** *Let  $X^*$  be a feasible solution of (9) and  $X^0$  be the initial solution. Under the Assumptions, there exist  $\kappa_j, \sigma_j > 0$  such that the expectation of the recovery error squared at the  $t$ -th iteration of Algorithm 3 for estimating the  $j$ -th column of  $X^*$  is bounded by*

$$(21) \quad E_{I_t} \|X_{\cdot,j}^{t+1} - X_{\cdot,j}^*\|_2^2 \leq \kappa_j^{t+1} \|X_{\cdot,j}^0 - X_{\cdot,j}^*\|_2^2 + \frac{\sigma_j}{1 - \kappa_j},$$

where  $X_{\cdot,j}^t$  is the approximation of  $X_{\cdot,j}^*$  at the  $t$ -th iteration of StoIHT with the initial guess  $X_{\cdot,j}^0$ , i.e., the result at the  $t$ -th inner iteration and  $j$ -th outer iteration of Algorithm 3 with the initial guess  $X^0$ . Here  $I_t$  is the set of all indices  $i_1, \dots, i_t$  randomly selected at or before the  $t$ -th step of the algorithm.

*Proof.* Due to the separable form of  $f_i$  in (10), we consider  $L$  problems of the form

$$(22) \quad \min_{\mathbf{w}} \frac{1}{M} \sum_{i=1}^M g_{i,j}(\mathbf{w}), \quad \|\mathbf{w}\|_{0,\mathcal{D}} \leq k, \quad j = 1, \dots, L,$$

where  $g_{i,j}$  are given in (10). This relaxation is valid since  $X_{\cdot,j}^*$  is also a feasible solution of (22) if  $X^*$  is a feasible solution of (9). Let  $\mathbf{w}^t = X_{\cdot,j}^t$ ,  $\mathbf{w}^* = X_{\cdot,j}^*$ ,  $x = \|\mathbf{w}^{t+1} - \mathbf{w}^*\|_2$ ,

$$u = \left\| \mathbf{w}^t - \mathbf{w}^* - \frac{\gamma}{Mp(i_t)} \mathcal{P}_\Omega(\nabla g_{i_t,j}(\mathbf{w}^t) - \nabla g_{i_t,j}(\mathbf{w}^*)) \right\|_2 + \left\| \frac{\gamma}{Mp(i_t)} \mathcal{P}_\Omega \nabla g_{i_t,j}(\mathbf{w}^*) \right\|_2 \\ := u_1 + u_2,$$

and

$$v = (\eta^2 - 1) \left\| \mathbf{w}^t - \mathbf{w}^* - \frac{\gamma}{Mp(i_t)} \nabla g_{i_t,j}(\mathbf{w}^t) \right\|_2^2.$$

Let

$$v_1 = (\eta^2 - 1) \left\| \mathbf{w}^t - \mathbf{w}^* - \frac{\gamma}{Mp(i_t)} (\nabla g_{i_t,j}(\mathbf{w}^t) - \nabla g_{i_t,j}(\mathbf{w}^*)) \right\|_2^2, \\ v_2 = (\eta^2 - 1) \left\| \frac{\gamma}{Mp(i_t)} \nabla g_{i_t,j}(\mathbf{w}^*) \right\|_2^2.$$

The inequality  $(a + b)^2 \leq 2a^2 + 2b^2$  yields that  $v \leq 2v_1 + 2v_2$ . Similar to [35, Appendix B], we can show that solving  $x^2 - 2ux - v \leq 0$  leads to  $x \leq u + \sqrt{u^2 + v}$ . Thus the inequality  $a + b \leq \sqrt{2a^2 + 2b^2}$  yields

$$x^2 \leq 2u^2 + 2(u^2 + v) = 4u^2 + 2v \leq 8(u_1^2 + u_2^2) + 4v_1 + 4v_2.$$

By taking the conditional expectation  $E_{i_t|I_{t-1}}$  on both sides, we obtain

$$E_{i_t|I_{t-1}} x^2 \leq 8E_{i_t|I_{t-1}} u_1^2 + 8E_{i_t|I_{t-1}} u_2^2 + 4E_{i_t|I_{t-1}} (v_1 + v_2) \\ \leq 8(1 - (2\gamma - \gamma^2 \alpha_{3k,j}) \rho_{3k,j}^-) \|\mathbf{w}^t - \mathbf{w}^*\|_2^2$$

$$\begin{aligned}
& + \frac{8\gamma^2}{\min_{i_t} M^2(p(i_t))^2} E_{i_t|I_{t-1}} \|\mathcal{P}_\Omega \nabla g_{i_t,j}(\mathbf{w}^*)\|_2^2 \\
& + 4(\eta^2 - 1)(1 + \gamma^2 \alpha_{3k,j} \bar{\rho}_{3k,j}^+ - 2\gamma \rho_{3k,j}^-) \|\mathbf{w}^t - \mathbf{w}^*\|_2^2 \\
& + \frac{4\gamma^2(\eta^2 - 1)}{\min_{i_t} M^2(p(i_t))^2} E_{i_t} \|\nabla g_{i_t,j}(\mathbf{w}^*)\|_2^2 \\
& = \left( 8(1 - (2\gamma - \gamma^2 \alpha_{3k,j}) \rho_{3k,j}^-) + 4(\eta^2 - 1)(1 + \gamma^2 \alpha_{3k,j} \bar{\rho}_{3k,j}^+ - 2\gamma \rho_{3k,j}^-) \right) \|\mathbf{w}^t - \mathbf{w}^*\|_2^2 \\
& + \frac{4\gamma^2}{\min_{i_t} M^2(p(i_t))^2} \left( 2E_{i_t} \|\mathcal{P}_\Omega \nabla g_{i_t,j}(\mathbf{w}^*)\|_2^2 + (\eta^2 - 1)E_{i_t} \|\nabla g_{i_t,j}(\mathbf{w}^*)\|_2^2 \right) \\
& := \kappa_j \|\mathbf{w}^t - \mathbf{w}^*\|_2^2 + \sigma_j.
\end{aligned}$$

In the second inequality, the respective upper bounds for  $E_{i_t|I_{t-1}} u_1^2$  and  $E_{i_t|I_{t-1}} u_2^2$  are obtained by using the Corollary 8 of [35]. Here  $\rho_{k,j}^-$  is defined in (15),  $\alpha_{k,j}$  and  $\bar{\rho}_{k,j}^+$  are defined in (16). Therefore

$$E_{I_t} \|\mathbf{w}^{t+1} - \mathbf{w}^*\|_2^2 \leq \kappa_j E_{I_{t-1}} \|\mathbf{w}^t - \mathbf{w}^*\|_2^2 + \sigma_j.$$

which implies that

$$E_{I_t} \|X_{\cdot,j}^{t+1} - X_{\cdot,j}^*\|_2^2 \leq \kappa_j^{t+1} \|X_{\cdot,j}^0 - X_{\cdot,j}^*\|_2^2 + \frac{\sigma_j}{1 - \kappa_j},$$

for  $j = 1, \dots, L$ . Here the contraction coefficient is

$$(23) \quad \kappa_j = 8(1 - (2\gamma - \gamma^2 \alpha_{3k,j}) \rho_{3k,j}^-) + 4(\eta^2 - 1)(1 + \gamma^2 \alpha_{3k,j} \bar{\rho}_{3k,j}^+ - 2\gamma \rho_{3k,j}^-),$$

and the tolerance parameter is

$$(24) \quad \sigma_j = \frac{4\gamma^2}{\min_{1 \leq i \leq M} M^2(p(i_t))^2} \left( 2E_{i_t} \|\mathcal{P}_\Omega \nabla g_{i_t,j}(\mathbf{w}^*)\|_2^2 + (\eta^2 - 1)E_{i_t} \|\nabla g_{i_t,j}(\mathbf{w}^*)\|_2^2 \right).$$

In particular, if  $\gamma = \eta = 1$  and  $p(i) = 1/M$ , then

$$\begin{aligned}
\kappa_j &= 8(1 - 2\rho_{3k,j}^- + \alpha_{3k,j} \rho_{3k,j}^-), \\
\sigma_j &= \frac{8}{M} \sum_{i=1}^M \|\mathcal{P}_\Omega \nabla g_{i,j}(X_{\cdot,j}^*)\|_2^2.
\end{aligned}$$

□

**Theorem 5.5** (CStoIHT). *Let  $X^*$  be a feasible solution of (9) and  $X^0$  be the initial solution. Under Assumptions, the expectation of the recovery error at the  $(t+1)$ -th iteration of Algorithm 3 is bounded by*

$$(25) \quad E \|X^{t+1} - X^*\|_F \leq \hat{\kappa}^{t+1} \|X^0 - X^*\|_F + \hat{\sigma},$$

where  $X^t$  is the approximation of  $X^*$  at the  $t$ -th iteration of Algorithm 3 with the initial guess  $X_{\cdot,j}^0$ . Here the contraction coefficient  $\hat{\kappa}$  and the tolerance parameter  $\hat{\sigma}$  are defined as

$$\hat{\kappa} = \sqrt{\max_{1 \leq j \leq L} \kappa_j}, \quad \hat{\sigma} = \sqrt{\frac{\sum_{j=1}^L \sigma_j}{1 - \max_{1 \leq j \leq L} \kappa_j}},$$

where  $\kappa_j$  is the contraction coefficient for each StoIHT defined in (23).

*Proof.* For each  $j = 1, 2, \dots, L$ , StoIHT with the initial guess  $X_{\cdot,j}^0$  generates  $X_{\cdot,j}^{t+1}$  after  $t$  iterations, i.e., the result at the  $(t+1)$ -th inner iteration and  $j$ -th outer iteration of Algorithm 3. Then the expectation of the recovery error squared is bounded by

$$E \|X_{\cdot,j}^{t+1} - X_{\cdot,j}^*\|_2^2 \leq \kappa_j^{t+1} \|X_{\cdot,j}^0 - X_{\cdot,j}^*\|_2^2 + \frac{\sigma_j}{1 - \kappa_j},$$

where  $\kappa_j$  and  $\sigma_j$  are defined in Lemma 5.4. Note that  $\kappa_j$  depends only on the constants in the  $\mathcal{D}$ -RSC and  $\mathcal{D}$ -RSS properties of the objective function or its component function  $g_{i,j}$  while  $\sigma_j$  depends on the feasible solution  $X_{\cdot,j}^*$ . By combining all  $L$  components of  $X^{t+1}$ , we get

$$\begin{aligned} E \|X^{t+1} - X^*\|_F &\leq \sqrt{E \|X^{t+1} - X^*\|_F^2} \\ &= \sqrt{\sum_{j=1}^L E \|X_{\cdot,j}^{t+1} - X_{\cdot,j}^*\|_2^2} \\ &\leq \sqrt{\sum_{j=1}^L \left( \kappa_j^{t+1} \|X_{\cdot,j}^0 - X_{\cdot,j}^*\|_2^2 + \frac{\sigma_j}{1 - \kappa_j} \right)} \\ &\leq \sqrt{(\max_{1 \leq j \leq L} \kappa_j)^{t+1} \|X^0 - X^*\|_F^2 + \frac{\sum_{j=1}^L \sigma_j}{1 - \max_{1 \leq j \leq L} \kappa_j}} \\ &\leq \left( \sqrt{\max_{1 \leq j \leq L} \kappa_j} \right)^{t+1} \|X^0 - X^*\|_F + \sqrt{\frac{\sum_{j=1}^L \sigma_j}{1 - \max_{1 \leq j \leq L} \kappa_j}} \\ &:= \hat{\kappa}^{t+1} \|X^0 - X^*\|_F + \hat{\sigma}. \end{aligned}$$

In particular, if  $\gamma = \eta = 1$ , then

$$\begin{aligned} \hat{\kappa} &= \sqrt{\max_{1 \leq j \leq L} \kappa_j} = \sqrt{\max_{1 \leq j \leq L} 8(1 - 2\rho_{3k,j}^- + \alpha_{3k,j}\rho_{3k,j}^-)} \\ &= 2\sqrt{2} \sqrt{\max_{1 \leq j \leq L} (1 - 2\rho_{3k,j}^- + \alpha_{3k,j}\rho_{3k,j}^-)}. \end{aligned}$$

□

**Remark.** Firstly, MStoIHT can converge without the Assumptions while CStoIHT may not converge to a joint sparse solution without Assumptions, since the joint sparsity is not enforced at each iteration of CStoIHT. Secondly, under Assumptions, if  $\rho_{3k,1}^- = \dots = \rho_{3k,L}^-$  and  $\rho_{3k,1}^+ = \dots = \rho_{3k,L}^+$ , Lemmas 5.1 and 5.2 yield  $\hat{\kappa} = \sqrt{2}\kappa$ , where  $\kappa$  is defined Theorem 5.3.

**5.3. MStoGradMP.** By using the same proof techniques as in Theorem 5.3, we can get the following convergence result for MStoGradMP.

**Theorem 5.6** (MStoGradMP). *Let  $X^*$  be a feasible solution of (9) and  $X^0$  be the initial solution. Assume that  $F(X) = \frac{1}{M} \sum_{i=1}^M f_i(X)$  satisfies the  $\mathcal{D}$ -RSC with constant  $\rho_k^-$  in (4) and  $f_i$  satisfies  $\mathcal{D}$ -RSS with constant  $\rho_k^+(i)$  in (5). At the  $(t+1)$ -th iteration of Algorithm 6, the expectation of the recovery error is bounded by*

$$(26) \quad E \|X^{t+1} - X^*\|_F \leq \kappa^{t+1} \|X^0 - X^*\|_F + \frac{\sigma_{X^*}}{1 - \kappa},$$

where

$$\kappa = (1 + \eta_2) \sqrt{\frac{\alpha_{4k}}{\rho_{4k}^-}} \left( \sqrt{\hat{p}} \sqrt{\frac{\rho_{4k}^+(2\eta_1^2 - 1)}{\rho_{4k}^-\eta_2^2}} - 1 + \frac{\sqrt{\eta_1^2 - 1}}{\eta_1} \right),$$

$$\sigma_{X^*} = \frac{(1 + \eta_2)}{\rho_{4k}^-\hat{p}} \left( 2\hat{p} \sqrt{\frac{\alpha_{4k}}{\rho_{4k}^-}} + 3 \right) \max_{\substack{|\Omega| \leq 4k \\ 1 \leq i \leq M}} \left\| \mathcal{P}_\Omega \frac{\partial f_i}{\partial X}(X^*) \right\|_F.$$

Here  $\hat{p} = \max_{1 \leq i \leq M} Mp(i)$ ,  $\tilde{p} = \min_{1 \leq i \leq M} Mp(i)$  and  $\alpha_k, \rho_k^+, \rho_k^-$  are defined in (19). Thus Algorithm 6 converges linearly if  $\kappa < 1$ . In particular, if  $\eta_1 = \eta_2 = 1$  and  $p(i) = 1/M$ , then

$$(27) \quad \kappa = \frac{2\sqrt{\alpha_{4k}(\rho_{4k}^+ - \rho_{4k}^-)}}{\rho_{4k}^-}.$$

**5.4. CStoGradMP.** Similar to CStoIHT, we start the convergence analysis for CStoGradMP by finding the contraction coefficient for the expectation of recovery error squared at each iteration of CStoGradMP.

**Lemma 5.7.** *Let  $X^*$  be a feasible solution of (13) and  $X^0$  be the initial solution. Under Assumptions, the expectation of the recovery error squared at the  $t$ -th iteration of Algorithm 4 for estimating the  $j$ -th column of  $X^*$  is bounded by*

$$(28) \quad E_{I_t} \left\| \mathbf{b}^t - X_{\cdot,j}^* \right\|_2^2 \leq \beta_1 E_{I_t} \left\| \mathcal{P}_{\widehat{\Gamma}}(\mathbf{b}^t - X_{\cdot,j}^*) \right\|_2^2 + \xi_1,$$

where

$$(29) \quad \begin{aligned} \beta_1 &= \frac{\alpha_{4k,j}}{2\rho_{4k,j}^- - \alpha_{4k,j}}, \\ \xi_1 &= \frac{2E_{I_t} E_i \left\| \mathcal{P}_{\widehat{\Gamma}} \nabla g_{i,j}(X_{\cdot,j}) \right\|_2^2}{\alpha_{4k,j} (2\rho_{4k,j}^- - \alpha_{4k,j}) \min_{1 \leq i \leq M} M^2(p(i))^2}. \end{aligned}$$

Here  $\alpha_{k,j}$  and  $\rho_{k,j}^-$  are defined in (16), and  $I_t$  is the set of all indices  $i_1, \dots, i_t$  randomly selected at or before the  $t$ -th step of the algorithm.

*Proof.* Consider the problem (22). Following the proof in [35, Appendix C], we can get

$$\begin{aligned} \left\| \mathcal{P}_{\widehat{\Gamma}}(\mathbf{b}^t - X_{\cdot,j}^*) \right\|_2^2 &\leq 2(1 - (2\gamma - \gamma^2 \alpha_{4k,j}) \rho_{4k,j}^-) \left\| \mathbf{b}^t - X_{\cdot,j}^* \right\|_2^2 \\ &\quad + \frac{2\gamma^2}{\min_{1 \leq i \leq M} M^2(p(i))^2} E_i \left\| \mathcal{P}_{\widehat{\Gamma}} \nabla g_{i,j}(X_{\cdot,j}) \right\|_2^2. \end{aligned}$$

Here we use the inequality  $(a + b)^2 \leq 2a^2 + 2b^2$  for  $a, b \in \mathbb{R}$  and the expectation inequality  $(EX)^2 \leq E(X^2)$ . Then we have

$$\begin{aligned} \left\| \mathbf{b}^t - X_{\cdot,j}^* \right\|_2^2 &= \left\| \mathcal{P}_{\widehat{\Gamma}}(\mathbf{b}^t - X_{\cdot,j}^*) \right\|_2^2 + \left\| \mathcal{P}_{\widehat{\Gamma}^c}(\mathbf{b}^t - X_{\cdot,j}^*) \right\|_2^2 \\ &\leq 2(1 - (2\gamma - \gamma^2 \alpha_{4k,j}) \rho_{4k,j}^-) \left\| \mathbf{b}^t - X_{\cdot,j}^* \right\|_2^2 \\ &\quad + \frac{2\gamma^2}{\min_{1 \leq i \leq M} M^2(p(i))^2} E_i \left\| \mathcal{P}_{\widehat{\Gamma}} \nabla g_{i,j}(X_{\cdot,j}^*) \right\|_2^2 + \left\| \mathcal{P}_{\widehat{\Gamma}^c}(\mathbf{b}^t - X_{\cdot,j}^*) \right\|_2^2. \end{aligned}$$

Moving the first term on the right hand side to the left hand side leads to

$$\left\| \mathbf{b}^t - X_{\cdot,j}^* \right\|_2^2 \leq \frac{2\gamma^2}{\phi \min_{1 \leq i \leq M} M^2(p(i))^2} E_i \left\| \mathcal{P}_{\widehat{\Gamma}} \nabla g_{i,j}(X_{\cdot,j}) \right\|_2^2 + \frac{1}{\phi} \left\| \mathcal{P}_{\widehat{\Gamma}^c}(\mathbf{b}^t - X_{\cdot,j}^*) \right\|_2^2,$$

where  $\phi = 2\rho_{4k,j}^-(2\gamma - \gamma^2\alpha_{4k,j}) - 1$ . Maximizing  $\phi$  with respect to  $\gamma$  yields  $\gamma = 1/\alpha_{4k,j}$  and  $\phi_{\max} = (2\rho_{4k,j}^- - \alpha_{4k,j})/\alpha_{4k,j}$ . By choosing the optimal value of  $\gamma$  and taking the expectation with respect to  $I_t$  on the both sides of the above inequality, we get (28).  $\square$

Similarly, using the inequality  $EX \leq \sqrt{E(X^2)}$  and the fact that  $a \leq b + c$  yields  $a^2 \leq 2b^2 + 2c^2$ , we are able to get the following result, which is different from Lemma 3 in [35] in that we consider the expectation for the  $\ell_2$ -norm squared here rather than that for the  $\ell_2$ -norm.

**Lemma 5.8.** *Let  $X^*$  be a feasible solution of (13) and  $X^0$  be the initial solution. Under Assumptions, the expectation of the recovery error squared at the  $t$ -th iteration of Algorithm 4 for estimating the  $j$ -th column of  $X^*$  is bounded by*

$$(30) \quad E_{i_t} \|\mathcal{P}_{\hat{\Gamma}^c}(\mathbf{b}^t - X_{\cdot,j}^*)\|_2^2 \leq \beta_2 \|X_{\cdot,j}^t - X_{\cdot,j}^*\|_2^2 + \xi_2,$$

where  $i_t$  is the index randomly selected at the  $t$ -th iteration of the CStoGradMP and

$$(31) \quad \begin{aligned} \beta_2 &= 4 \max_{1 \leq i \leq M} M p(i) \frac{(2\eta_1^2 - 1)\rho_{4k,j}^+ - \eta_1^2 \rho_{4k,j}^-}{\eta_1^2 \rho_{4k,j}^-} + \frac{2(\eta_1^2 - 1)}{\eta_1^2}, \\ \xi_2 &= 8 \left( \frac{\max_{1 \leq i \leq M} p(i)}{\rho_{4k,j}^- \min_{1 \leq i \leq M} p(i)} \right)^2 \max_{\substack{|\Omega| \leq 4k \\ 1 \leq i \leq M}} \|\mathcal{P}_\Omega \nabla g_{i,j}(X_{\cdot,j}^*)\|_2^2. \end{aligned}$$

**Theorem 5.9** (CStoGradMP). *Let  $X^*$  be a feasible solution of (13) and  $X^0$  be the initial solution. Under Assumptions, at the  $(t+1)$ -th iteration of Algorithm 4, there exist  $\tilde{\kappa}, \tilde{\sigma} > 0$  such that the expectation of the recovery error is bounded by*

$$(32) \quad E \|X^{t+1} - X^*\|_F \leq \tilde{\kappa}^{t+1} \|X^0 - X^*\|_F + \tilde{\sigma},$$

where  $X_{\cdot,j}^t$  is the approximation of  $X_{\cdot,j}^*$  at the  $t$ -th iteration of CStoGradMP with the initial guess  $X_{\cdot,j}^0$ .

*Proof.* At the  $t$ -th iteration of Algorithm 4, we have

$$\|X_{\cdot,j}^{t+1} - \mathbf{b}^t\|_2^2 \leq \eta_2^2 \|\mathbf{b}_{(k)}^t - \mathbf{b}^t\|_2^2 \leq \eta_2^2 \|X_{\cdot,j}^* - \mathbf{b}^t\|_2^2,$$

where  $\mathbf{b}_{(k)}^t$  is the best  $k$ -sparse approximation of  $\mathbf{b}^t$  with respect to the atom set  $\mathcal{D}$ . Therefore, we get

$$\begin{aligned} \|X_{\cdot,j}^{t+1} - X_{\cdot,j}^*\|_2^2 &\leq \|X_{\cdot,j}^{t+1} - \mathbf{b}^t + \mathbf{b}^t - X_{\cdot,j}^*\|_2^2 \\ &\leq 2 \|X_{\cdot,j}^{t+1} - \mathbf{b}^t\|_2^2 + 2 \|\mathbf{b}^t - X_{\cdot,j}^*\|_2^2 \\ &\leq (2 + 2\eta_2^2) \|\mathbf{b}^t - X_{\cdot,j}^*\|_2^2. \end{aligned}$$

Next we establish the relationships among various expectations

$$\begin{aligned} E_{I_t} \|X_{\cdot,j}^{t+1} - X_{\cdot,j}^*\|_2^2 &\leq (2 + 2\eta_2^2) E_{I_t} \|\mathbf{b}^t - X_{\cdot,j}^*\|_2^2 \\ &\leq (2 + 2\eta_2^2) \left( \beta_1 E_{I_t} \|\mathcal{P}_{\hat{\Gamma}}(\mathbf{b}^t - X_{\cdot,j}^*)\|_2^2 + \xi_1 \right) \\ &\leq (2 + 2\eta_2^2) \beta_1 \left( \beta_2 E_{I_t} \|X_{\cdot,j}^t - X_{\cdot,j}^*\|_2^2 + \xi_2 \right) + (2 + 2\eta_2^2) \xi_1 \\ &:= \kappa_j \|X_{\cdot,j}^t - X_{\cdot,j}^*\|_2^2 + \sigma_j, \end{aligned}$$

where the first inequality is guaranteed by Lemma 5.7 and the second inequality is guaranteed by Lemma 5.8. Here the contraction coefficient  $\kappa_j$  and the tolerance parameter  $\sigma_j$  are defined by

$$\kappa_j = (2 + 2\eta_2^2)\beta_1\beta_2, \quad \sigma_j = (2 + 2\eta_2^2)\beta_1\xi_2 + (2 + 2\eta_2^2)\xi_1,$$

where  $\beta_1, \xi_1$  are defined in (29) and  $\beta_2, \xi_2$  are defined in (31). Then similar to the proof of Theorem 5.5, we can derive that

$$E \|X^{t+1} - X^*\|_F \leq \tilde{\kappa}^{t+1} \|X^0 - X^*\|_F + \tilde{\sigma}$$

where

$$\tilde{\kappa} = \sqrt{\max_{1 \leq j \leq L} \kappa_j}, \quad \text{and} \quad \tilde{\sigma} = \sqrt{\frac{\sum_{j=1}^L \sigma_j}{1 - \max_{1 \leq j \leq L} \kappa_j}}.$$

In particular, if  $\eta_1 = \eta_2 = 1$ ,  $p(i) = 1/M$ ,  $\alpha_{4k,j} = \alpha_{4k}$ ,  $\rho_{4k,j}^+ = \rho_{4k}^+$  and  $\rho_{4k,j}^- = \rho_{4k}^-$  for  $j = 1, \dots, L$ , then we have

$$\tilde{\kappa} = 4 \sqrt{\frac{\alpha_{4k}(\rho_{4k}^+ - \rho_{4k}^-)}{\rho_{4k}^-(2\rho_{4k}^- - \alpha_{4k})}}.$$

If, in addition,  $\alpha_{4k} = \rho_{4k}^-$ , then we have  $\tilde{\kappa} = 2\kappa$ , where the contraction coefficient  $\kappa$  for MStoGradMP is given in (27), which implies that MStoGradMP converges faster than CStoGradMP in this case due to the smaller contraction coefficient. Compared with MStoIHT, MStoGradMP has even larger convergence improvement in terms of recovery accuracy and running time.  $\square$

**Remark.** Convergence analysis in this section provides useful theoretical tools to compare the proposed MMV stochastic greedy algorithms and their SMV counterparts. First, we provide general guarantees via contraction coefficients for all proposed algorithms in Theorems 5.3, 5.5, 5.6 and 5.9. Since the contraction coefficient controls the convergence speed of each algorithm, it can provide theoretical guidance on parameter selection in such type of stochastic greedy algorithms. Second, the complicated form of  $\kappa$  and  $\hat{\kappa}$  can be simplified in some special cases, e.g., the case when  $\eta = 1$ . Lastly, the relationship  $\kappa = \hat{\kappa}/\sqrt{2}$  indicates why MStoIHT/MStoGradMP performs better than CStoIHT/CStoGradMP in expectation.

**6. Distributed compressive sensing application.** In this section, we show that the objective function commonly used in the distributed compressive sensing problem satisfies the  $\mathcal{D}$ -RSC and  $\mathcal{D}$ -RSS properties, which paves the theoretical foundation for using the proposed algorithms in this application. Suppose that there are  $L$  underlying signals  $\mathbf{x}_j \in \mathbb{R}^n$  for  $j = 1, 2, \dots, L$ , and their measurements are generated by

$$\mathbf{y}_j = A^{(j)}\mathbf{x}_j, \quad j = 1, 2, \dots, L$$

where  $A^{(j)} \in \mathbb{R}^{m \times n}$  ( $m \ll n$ ) is the measurement matrix (a.k.a. the sensing matrix). For discussion simplicity, we assume all measurement matrices are the same, i.e.,  $A^{(j)} = A = [A_{\cdot,1}, \dots, A_{\cdot,n}]$ . By concatenating all vectors as a matrix, we rewrite the above equation as  $Y = AX$  where  $Y = [\mathbf{y}_1, \dots, \mathbf{y}_L] \in \mathbb{R}^{m \times L}$ , and  $X = [\mathbf{x}_1, \dots, \mathbf{x}_L] \in \mathbb{R}^{n \times L}$ .

Now assume that the atom set is finite and denote  $\mathcal{D} = \{\mathbf{d}_1, \dots, \mathbf{d}_N\}$  with the corresponding dictionary  $D = [\mathbf{d}_1, \dots, \mathbf{d}_N]$ . Consider the following distributed compressive sensing model with common sparse supports [3]

$$(33) \quad \begin{aligned} \min_X \frac{1}{2m} \sum_{j=1}^L \|\mathbf{y}_j - A\mathbf{x}_j\|_2^2 \\ \text{s.t. } \mathbf{x}_j = D\theta_j \quad \text{supp}(\theta_j) = \Omega \subseteq \{1, 2, \dots, N\}. \end{aligned}$$

Here the objective function has the form

$$(34) \quad F(X) = \frac{1}{2m} \|Y - AX\|_F^2.$$

Then  $F(X)$  can be written as  $F(X) = \frac{1}{M} \sum_{i=1}^M f_i(X)$ , where  $M = m/b$  and

$$(35) \quad f_i(X) = \frac{1}{2b} \sum_{j=1}^L (Y_{i,j} - \sum_{k=1}^n A_{i,k} X_{k,j})^2 = \frac{1}{2b} \|Y_{i,\cdot} - A_{i,\cdot} X\|_2^2.$$

The above expression shows that  $f_i$ 's satisfy the Assumption (a) and thereby the concatenated algorithms in Section 3 can be applied. We first compute the partial derivative. For  $s = 1, 2, \dots, n$  and  $t = 1, 2, \dots, L$ , we have

$$\begin{aligned} \frac{\partial f_i(X)}{\partial X_{s,t}} &= \frac{1}{2b} \sum_{j=1}^L 2 \left( \sum_{k=1}^n A_{i,k} X_{k,j} - Y_{i,j} \right) \sum_{k=1}^n A_{i,k} \frac{\partial X_{k,j}}{\partial X_{s,t}} \\ &= \frac{1}{b} \sum_{j=1}^L \left( \sum_{k=1}^n A_{i,k} X_{k,j} - Y_{i,j} \right) \sum_{k=1}^n A_{i,k} \delta_{k,s} \delta_{j,t} \\ &= \frac{1}{b} \left( \sum_{k=1}^n A_{i,k} X_{k,t} - Y_{i,t} \right) A_{i,s}. \end{aligned}$$

Here  $\delta_{i,j} = 1$  if  $i = j$  and zero otherwise. Thus the generalized gradient of  $f_i(X)$  with respect to  $X$  has the form

$$\frac{\partial f_i(X)}{\partial X} = \frac{1}{b} \left[ \frac{\partial f_i(X)}{\partial X_{s,t}} \right]_{st} = \frac{1}{b} A_{i,\cdot}^T (A_{i,\cdot} X - Y_{i,\cdot}).$$

**Lemma 6.1.** *If the sensing matrix  $A \in \mathbb{R}^{m \times n}$  satisfies the Restricted Isometry Property (RIP), i.e., there exists  $\delta_k > 0$  such that*

$$(1 - \delta_k) \|\mathbf{x}\|_2^2 \leq \|A\mathbf{x}\|_2^2 \leq (1 + \delta_k) \|\mathbf{x}\|_2^2$$

for any  $k$ -sparse vector  $\mathbf{x} \in \mathbb{R}^n$ , then the function  $F(X)$  defined in (34) satisfies the  $\mathcal{D}$ -restricted strong convexity property.

*Proof.* Let  $X \in \mathbb{R}^{n \times L}$  with  $k$  nonzero rows, which implies that each column of  $X$  has at most  $k$  nonzero components. By the RIP of  $A$ , we have

$$(1 - \delta_k) \|X_{\cdot,j}\|_2^2 \leq \|AX_{\cdot,j}\|_2^2 \leq (1 + \delta_k) \|X_{\cdot,j}\|_2^2,$$

for  $j = 1, \dots, L$ . Note that  $\|X\|_F^2 = \sum_{j=1}^L \|X_{\cdot,j}\|_2^2$ . Thus we get

$$(1 - \delta_k) \|X\|_F^2 \leq \|AX\|_F^2 \leq (1 + \delta_k) \|X\|_F^2.$$

For any two  $X, X' \in \mathbb{R}^{n \times L}$  with  $|\text{supp}_{\mathcal{D}}^r(X) \cup \text{supp}_{\mathcal{D}}^r(X')| \leq k$ , we have

$$\begin{aligned} & F(X') - F(X) - \left\langle \frac{\partial F(X)}{\partial X}, X' - X \right\rangle \\ &= \frac{1}{2m} \left( \|Y - AX'\|_F^2 - \|Y - AX\|_F^2 \right) - \left\langle \frac{1}{m} A^T(AX - Y), X' - X \right\rangle \\ &= \frac{1}{2m} \|A(X' - X)\|_F^2 \geq \frac{1 - \delta_k}{2m} \|X' - X\|_F^2. \end{aligned}$$

Thus  $F(X)$  satisfies the  $\mathcal{D}$ -restricted strong convexity property with  $\rho_k^- = \frac{1 - \delta_k}{2m}$ .  $\square$

**Lemma 6.2.** *If the sensing matrix  $A \in \mathbb{R}^{m \times n}$  satisfies the following property: for any  $k$ -sparse vector  $\mathbf{x} \in \mathbb{R}^n$ , there exists  $\delta_k > 0$  such that*

$$\frac{1}{b} \|A_{\tau_i, \cdot}^T A_{\tau_i, \cdot} \mathbf{x}\|_2 \leq (1 + \delta_k) \|\mathbf{x}\|_2$$

where  $A_{\tau_i, \cdot}$  is formed by extracting rows of  $A$  with row indices in the  $i$ -th batch index set  $\tau_i$ . Then the function  $f_i(X)$  defined in (35) satisfies the  $\mathcal{D}$ -restricted strong smoothness property.

*Proof.* Let  $X \in \mathbb{R}^{n \times L}$  have  $k$  nonzero rows. Then for  $j = 1, \dots, L$ , we have

$$\frac{1}{b} \|A_{\tau_i, \cdot}^T A_{\tau_i, \cdot} X_{\cdot, j}\|_2 \leq (1 + \delta_k) \|X_{\cdot, j}\|_2,$$

which implies that

$$\frac{1}{b} \|A_{\tau_i, \cdot}^T A_{\tau_i, \cdot} X\|_F \leq (1 + \delta_k) \|X\|_F.$$

For any two  $X, X' \in \mathbb{R}^{n \times L}$  with  $|\text{supp}_{\mathcal{D}}^r(X) \cup \text{supp}_{\mathcal{D}}^r(X')| \leq k$ , we have

$$\begin{aligned} & \left\| \frac{\partial f_i(X)}{\partial X} - \frac{\partial f_i(X')}{\partial X} \right\|_F \\ &= \frac{1}{b} \|A_{\tau_i, \cdot}^T (A_{\tau_i, \cdot} X - Y_{\tau_i, \cdot}) - A_{\tau_i, \cdot}^T (A_{\tau_i, \cdot} X' - Y_{\tau_i, \cdot})\|_F \\ &= \frac{1}{b} \|A_{\tau_i, \cdot}^T A_{\tau_i, \cdot} (X - X')\|_F \leq (1 + \delta_k) \|X - X'\|_F. \end{aligned}$$

Therefore  $f_i(X)$  satisfies the  $\mathcal{D}$ -restricted strong convexity with  $\rho_k^+(i) = 1 + \delta_k$ .  $\square$

By Lemma 6.1, Lemma 6.2 and the convergence analysis in Section 5, the contraction coefficient in the proposed algorithms depends on the coefficient in the RIP condition, whose infimum for some special type of matrices are available [19]. In addition, the convergence guarantees of all proposed algorithms in terms of the RIP constant for the distributed compressive sensing can be obtained by plugging the expression of  $\rho_k^+$  and  $\rho_k^-$  in the proofs of Lemmas 6.1 and 6.2 into  $\kappa$ ,  $\hat{\kappa}$  and  $\tilde{\kappa}$  in Theorems 5.3, 5.5, 5.6, and 5.9.

**7. Numerical experiments.** In this section, we conduct a variety of numerical experiments to validate the effectiveness of the proposed algorithms. More specifically, our tests include reconstruction of row sparse signals from a linear system and joint sparse video sequence recovery. To compare different results quantitatively, we use the relative error defined as  $\text{ReErr} = \|X^t - X^*\|_F / \|X^*\|_F$ , where  $X^*$  is the ground truth and  $X^t$  is the estimation of  $X^*$  at the  $t$ -th iteration. Regarding the computational efficiency, we also record the running time which counts all the computation time over a specified number of iterations excluding data loading or generation. Here we record the running time by using the commands `tic`

and `toc` in Matlab. To assess the concatenated SMV algorithms, i.e., CStoIHT and CStoGradMP, we apply the SMV algorithm sequentially to the same sensing matrix and all columns of the measurement matrix  $Y$ , and save all intermediate approximations of each column of  $X$  for further computation of the relative error. In all tests, we use the discrete uniform distribution, i.e.,  $p(i) = 1/M$  for  $i = 1, 2, \dots, M$  in the non-batched versions and  $p(i) = 1/d$  for  $i = 1, 2, \dots, d$  in the batched versions. The parameter  $\eta$  is fixed as 1. By default, each algorithm is stopped when either the relative error between two subsequent approximations of  $X^*$  reaches the tolerance or the maximum number of iterations is achieved. All our experiments are performed in a desktop with an Intel® Xeon® CPU E5-2650 v4 @ 2.2GHz and 64GB RAM in double precision. The algorithms are implemented in Matlab 2016a running on Windows 10.

**7.1. Joint sparse matrix recovery.** In the first set of experiments, we compare the proposed algorithms and their concatenated SMV counterparts in terms of reconstruction error and running time. In particular, we investigate the impact of the sparsity level  $k$ , and the number of underlying signals  $L$  to be reconstructed on the performance of BCStoIHT, BMStoIHT, BCStoGradMP and BMStoGradMP, in terms of relative error and the running time. To reduce randomness in the results, we run 50 trials for each test with fixed parameters and then take the average over the number of trials. Note that since different trials may take different number of iterations to reach the desired accuracy, we only count the common iterations when comparing the reconstruction error and the runtime.

First, we compare CStoIHT and MStoIHT in both non-batched and batched versions, and fix the maximum number of iterations as 1000, the stopping criteria tolerance  $\varepsilon = 10^{-6}$  and  $\gamma = 1$  in both algorithms. To start with, we create a sensing matrix  $A \in \mathbb{R}^{100 \times 200}$  where each entry follows the normal distribution with zero mean and variance of 1/100 and each column of  $A$  is normalized by dividing its  $\ell_2$ -norm. In this way, it can be shown that the spark of  $A$ , i.e., the smallest number of linearly dependent columns of  $A$ , is 100 with probability one [16]. To create a signal matrix  $X^* \in \mathbb{R}^{200 \times 40}$ , we first generate a Gaussian distributed random matrix of size  $200 \times 40$ , and then randomly zero out  $(200 - k)$  rows where  $k$  is the row sparsity of  $X^*$ . The measurement matrix  $Y$  is created by  $AX$  for the noise-free cases. By choosing the sparsity level  $k \in \{10, 20\}$  and the batch size  $b \in \{1, 10\}$ , we obtain the results shown in Figure 1. Since the initial guess for the signal matrix is set a zero matrix, all the error curves start with the point  $(0, 1)$ . Notice that to show the computational efficiency, we use the running time in seconds as the horizontal axis rather than the number of iterations. It can be seen that as the sparsity level grows, i.e., the signal matrix is less joint sparse, more running time (or iterations) is required to achieve the provided tolerance in terms of the relative reconstruction error. Meanwhile, as the batch size increases, BMStoIHT performs better than the sequential BCStoIHT. With large sparsity levels, the inaccurate joint support obtained in the concatenated SMV algorithms cause large relative errors in the first few iterations (see Figure 1).

Next, we fix the sparsity level  $k$  as 5 and choose the number of signals as  $L \in \{40, 80\}$ . Figure 2 compares the results obtained by BCStoIHT and BMStoIHT when the batch size is 1 and 10. In general, BMStoIHT takes less running time than its sequential concatenated SMV counterpart. We can see that mini-batching significantly improves the reconstruction accuracy and reduces the running time of BMStoIHT. After a large number of tests, we also find that the computational

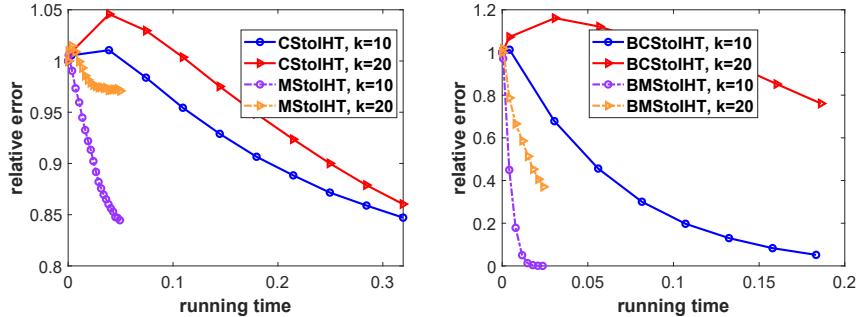


FIGURE 1. Comparison of BCStoIHT and BMStoIHT for various sparsity levels of the signal matrix. From left to right: batch sizes are 1 and 10.

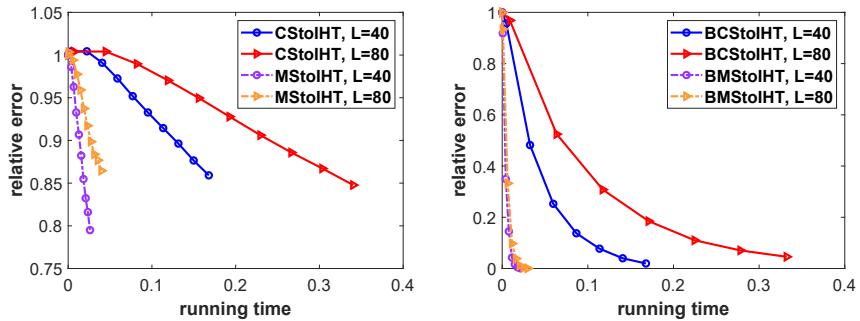


FIGURE 2. Comparison of BCStoIHT and BMStoIHT for various numbers of signals to be reconstructed. From left to right: batch sizes are 1 and 10.

time of BMStoIHT is almost linear with respect to the number of signals to be reconstructed. More numerical experiments can be found in our more recent work on the hyperspectral diffuse optical imaging [15]. Lastly, to test the robustness to noise, we add the Gaussian noise with zero mean and standard deviation (a.k.a. noise level)  $\sigma \in \{0.04, 0.08\}$  to the measurement matrix  $Y$ . The relative errors for all BCStoIHT and BMStoIHT results versus running time are shown in Figure 3. It is worth noting that the change of sparsity and noise levels have insignificant impact on the running time, which explains that the curve corresponding to the same algorithm stops almost at the same horizontal coordinate in Figure 1 and Figure 3. By contrast, the running time grows as the number of signals to be recovered increases which suggests that the endpoint of each curve has different horizontal coordinates in Figure 2.

In the second set of tests, we compare CStoGradMP and MStoGradMP in non-batched and batched versions. It is known that StoGradMP usually converges much faster than StoIHT, which is also true for the developed MMV versions. We fix the maximum number of iterations as 30, the stopping criteria tolerance  $\varepsilon = 10^{-5}$ ,  $\gamma = 1$ , and the batch size as 1 (non-batched version). Similar to the previous tests, we create a  $100 \times 200$  random matrix whose columns are normalized, and fix the

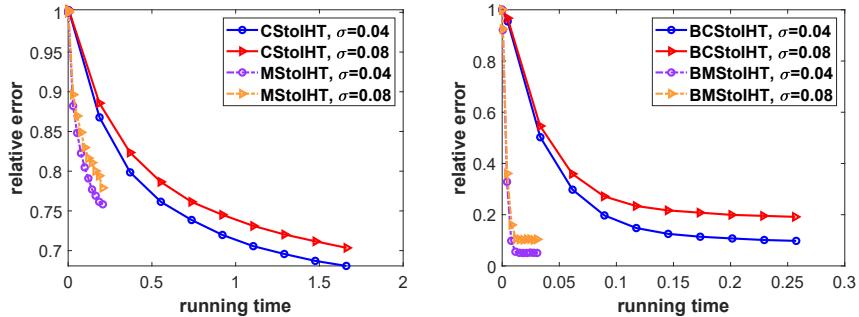


FIGURE 3. Comparison of BCStoIHT and BMStoIHT for various noise levels to the measurement matrix. From left to right: batch sizes are 1 and 10.

number of signals  $L = 40$ . Figure 4 shows the results obtained by CStoGradMP and MStoGradMP with sparsity level  $k \in \{70, 90\}$ . Note that for better visualization, we skip the starting point  $(0,1)$  for all relative error plots, and use the base 10 logarithmic scale for the horizontal axis of running time since the MStoGradMP takes much less running time than CStoGradMP after the same number of iterations. Unlike StoIHT and MStoIHT, both CStoGradMP and MStoGradMP require that the sparsity level is no more than  $n/2$ , i.e., 100 in our case. As the sparsity  $k$  increases, the operator `pinv` for computing the pseudo-inverse matrix becomes more computationally expensive for matrices with more columns than their rank, which results in the significant growth of running time. For sparse signal matrices, StoGradMP performs better than StoIHT in terms of convergence. Next, we set the number of signals as 20, 80, and get the results shown in Figure 5. It can be seen that MStoGradMP always takes less running time with even higher accuracy than the sequential CStoGradMP. We also discovered that the computation speedup of MStoIHT is almost constant with respect to the number of signals to be reconstructed. The robustness comparison is shown in Figure 6, where the noise level ranges in  $\{0.04, 0.08\}$ . Furthermore, it is empirically shown that the BMStoGradMP performs much better than BMStoIHT considering their respective convergence behavior and robustness.

Lastly, we compare our proposed BMStoIHT and BMStoGradMP with the M-FOCUSS [11], and state-of-the-art MMV greedy algorithms [5] including M-OMP, M-SP, M-CoSaMP and M-HTP, using the previous noise-free data set with  $k = 10$ ,  $L = 40$ , the maximum number of iterations as 100 and the stopping criteria tolerance  $\varepsilon = 10^{-5}$ . The batch size for both BMStoIHT and BMStoGradMP is fixed as 20. The reconstruction errors and running times for all methods are listed in Table 1, which shows that BMStoGradMP can achieve comparable reconstruction accuracy as M-OMP, M-SP, M-CoSaMP and M-HTP but with the least running time.

**7.2. Joint Sparse Video Sequence Recovery.** In this set of experiments, we compare the proposed Algorithm 8, the split Bregman algorithm for constrained MMV problem (SBC) [22, Algorithm 2], M-FOCUSS, M-OMP, M-SP, M-CoSaMP and M-HTP, on joint sparse video sequence reconstruction. Note that SBC is based on  $\ell_1$ -minimization rather than  $\ell_0$ -norm constrained least-squares. We first download a candle video consisting of 75 frames from the Dynamic Texture Toolbox

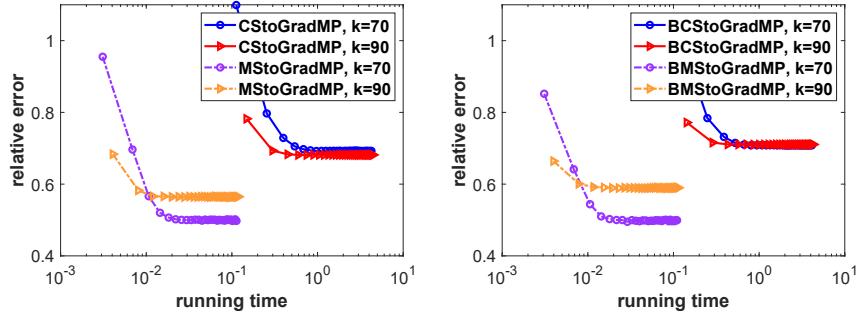


FIGURE 4. Comparison of BCStoGradMP and BMStoGradMP with various sparsity levels. From left to right: batch sizes are 1 and 10.

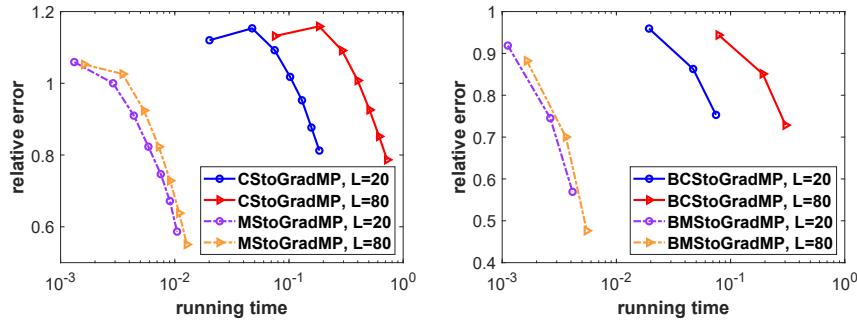


FIGURE 5. Comparison of BCStoGradMP and BMStoGradMP with various numbers of signals. From left to right: batch sizes are 1 and 10.

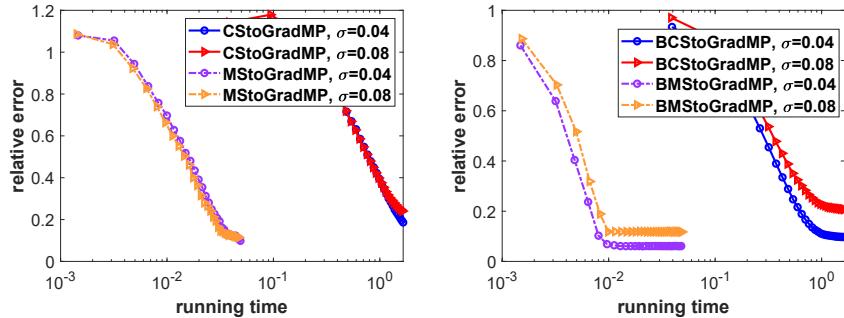


FIGURE 6. Comparison of BCStoGradMP and BMStoGradMP with various noise levels. From left to right: batch sizes are 1 and 10.

in <http://www.vision.jhu.edu/code/>. In order to make the test video sequence possess a joint sparse structure, we extract 11 frames of the original data, i.e., frames 1 to 7, 29, 37, 69 and 70, each of which is of size  $80 \times 30$ . Then we create a

Method	Error	Runtime (sec.)
BMStoIHT	$5.64 \times 10^{-7}$	0.0184
BMStoGradMP	$1.05 \times 10^{-15}$	0.0023
M-FOCUSS	$3.80 \times 10^{-3}$	0.0190
M-OMP	$7.45 \times 10^{-16}$	0.0492
M-SP	$6.17 \times 10^{-16}$	0.0217
M-CoSaMP	$8.22 \times 10^{-16}$	0.0218
M-HTP	$5.33 \times 10^{-16}$	0.0912

TABLE 1. Comparison of the proposed algorithms and other MMV algorithms.

Method	Error	Runtime (sec.)
BMStoGradMP	$3.09 \times 10^{-15}$	$2.56 \times 10^{-4}$
M-FOCUSS	$3.21 \times 10^{-3}$	0.2452
M-OMP	$2.51 \times 10^{-16}$	0.0743
M-SP	$2.06 \times 10^{-16}$	0.1297
M-CoSaMP	$4.07 \times 10^{-16}$	0.0115
M-HTP	$5.72 \times 10^{-2}$	0.0134
SBC	$1.19 \times 10^{-8}$	$6.28 \times 10^{-4}$

TABLE 2. Performance comparison on joint sparse video recovery.

data matrix  $X \in \mathbb{R}^{2400 \times 11}$ , whose columns are a vectorization of all video frames. To further obtain a sparse representation of  $X$ , K-SVD [2] is applied to obtain a dictionary  $\Psi \in \mathbb{R}^{2400 \times 50}$  for  $X$ . The K-SVD dictionary  $\Psi$  and the support of the corresponding coefficient matrix  $\Theta$  for the extracted 11 frames are shown in Figure 7 (a) and (b). Some selected columns of the dictionary  $\Psi$ , namely *atoms*, are reshaped as an image of size  $80 \times 30$  illustrated in Figure 7 (c) and (d). It can be seen that these 11 frames are nearly joint sparse under the learned dictionary. The relative error of using this K-SVD dictionary  $\Psi$  to represent the data matrix  $X$  is  $\frac{\|X - \Psi\Theta\|_F}{\|X\|_F} = 0.0870$ . A Gaussian random matrix  $\Phi \in \mathbb{R}^{60 \times 2400}$  with zero mean and unit variance is set as a sensing matrix, which is used to measure this data matrix. That is, the measurements  $Y \in \mathbb{R}^{60 \times 11}$  are generated via  $Y = \Phi X$ . Given the measurements  $Y$  and the new sparse representation dictionary  $A = \Phi \Psi$ , we then apply Algorithm 8 and SBC to recover the joint sparse coefficient matrix  $\hat{\Theta}$ . In Algorithm 8, the sparsity level  $k$  is set as 10, and the block size  $b$  is set as 3, which implies that there are  $d = 20$  blocks. In addition, we set  $\eta_1 = \eta_2 = 1$ . Both StoGradMP and SBC stop when the residual error reaches a tolerance threshold  $\tau = 10^{-6}$ , i.e.,  $\|Y - A\hat{\Theta}\|_F \leq \tau$ . In Table 2, we compare the reconstruction error, i.e.,  $\|Y - A\hat{\Theta}\|_F / \|Y\|_F$ , and running time for all methods being compared. One can see that the proposed Algorithm 8 is faster than all other comparing methods with relatively high accuracy.

**8. Conclusions.** In this paper, we study the MMV joint sparse signal reconstruction problem, which is of great importance in a large amount of signal processing applications. One straightforward way is to columnwise concatenate results obtained by solving the SMV sparse signal recovery problem. However, concatenated SMV algorithms may not enforce joint sparsity of the approximated solutions. To

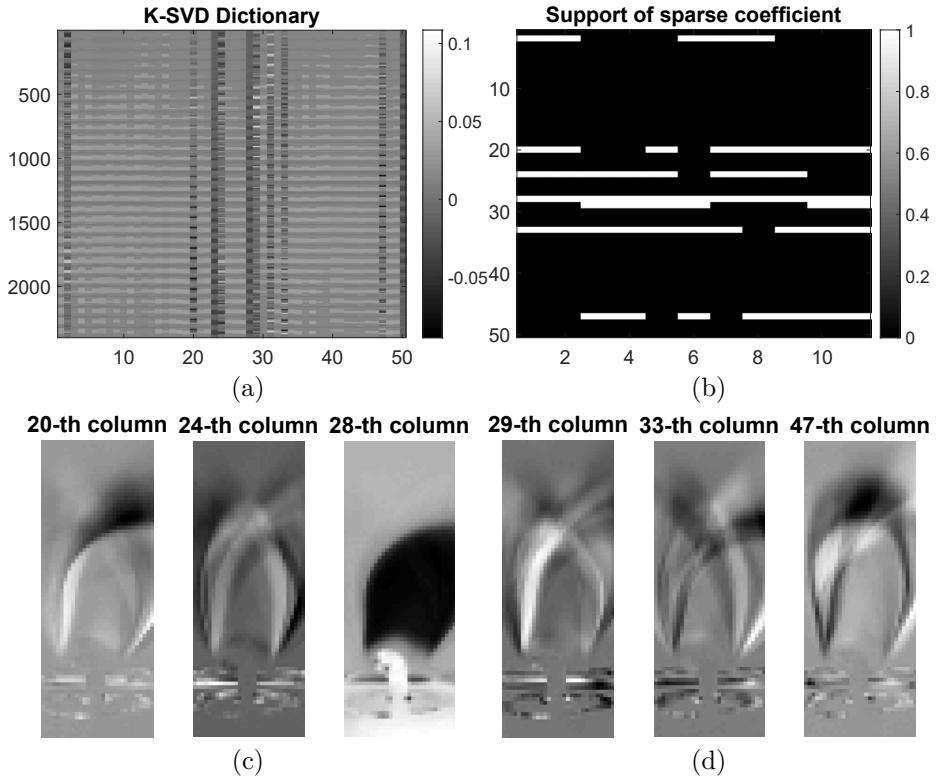


FIGURE 7. (a) K-SVD dictionary learned from the total 75 frames from the candle video. (b) Support of sparse coefficient matrix for extracted 11 frames. The sparse coefficient matrix has non-zero entries on white area. (c-d) Some columns of the learned K-SVD dictionary.

address this issue, we propose two stochastic greedy algorithms, MStoIHT and MStoGradMP, together with their respective accelerated versions by applying the mini-batching technique. Our convergence analysis has shown that the proposed algorithms converge faster than their concatenated SMV counterparts. Moreover, we theoretically justify the applicability of the proposed algorithms to the distributed compressive sensing problem. To the best of our knowledge, this is the first work to use the term ‘‘concatenated SMV algorithms’’ with theoretical discussions on their convergence. A variety of numerical experiments on linear systems and video recovery have demonstrated that the proposed algorithms outperform the concatenated SMV algorithms in terms of efficiency, accuracy and robustness to the noise.

## Appendix.

### A. Proof of lemma 5.1.

*Proof.* For any  $X', X \in \mathbb{R}^{n \times L}$  with  $|\text{supp}_{\mathcal{D}}^r(X) \cup \text{supp}_{\mathcal{D}}^r(X')| \leq k$ , all columns of  $X'$  and  $X$  are  $k$ -sparse and  $|\text{supp}_{\mathcal{D}}(X_{\cdot,j}) \cup \text{supp}_{\mathcal{D}}^r(X'_{\cdot,j})| \leq k$ . By (3) and  $\mathcal{D}$ -RSS of

each  $g_{i,j}$ , we have for  $i = 1, \dots, M$

$$\begin{aligned} \left\| \frac{\partial}{\partial X} f_i(X) - \frac{\partial}{\partial X} f_i(X') \right\|_F &= \sqrt{\sum_{j=1}^L \|\nabla g_{i,j}(X_{\cdot,j}) - \nabla g_{i,j}(X'_{\cdot,j})\|_2^2} \\ &\leq \max_{1 \leq j \leq L} \rho_k^+(j) \|X - X'\|_F := \mu_k^+(i) \|X - X'\|_F. \end{aligned}$$

□

### B. Proof of lemma 5.2.

*Proof.* Similar to the previous lemma, for any  $X', X \in \mathbb{R}^{n \times L}$  with  $|\text{supp}_{\mathcal{D}}^r(X) \cup \text{supp}_{\mathcal{D}}^r(X')| \leq k$ , we have

$$\begin{aligned} F(X') - F(X) - \langle \frac{\partial}{\partial X} F(X), X' - X \rangle &= \frac{1}{L} \sum_{j=1}^L \widehat{F}_j(X'_{\cdot,j}) - \widehat{F}_j(X_{\cdot,j}) - \frac{1}{L} \sum_{j=1}^L \langle \nabla \widehat{F}_j(X_{\cdot,j}), X'_{\cdot,j} - X_{\cdot,j} \rangle \\ &\geq \frac{1}{L} \sum_{j=1}^L \frac{\rho_{k,j}^-}{2} \|X_{\cdot,j} - X'_{\cdot,j}\|_2^2 \\ &\geq \frac{\mu_k^-}{2} \|X - X'\|_F^2, \end{aligned}$$

where  $\mu_k^- = \min_{1 \leq j \leq L} \rho_{k,j}^-$ . □

**Acknowledgments.** The initial research for this effort was conducted at the Research Collaboration Workshop for Women in Data Science and Mathematics, July 17-21 held at ICERM. Funding for the workshop was provided by ICERM, AWM and DIMACS (NSF grant CCF-1144502). Qin is supported by NSF DMS-1941197. Li was supported by the NSF grants CCF-1409258, CCF-1704204 and the DARPA Lagrange Program under ONR/SPAWAR contract N660011824020. Needell is supported by NSF CAREER DMS-1348721 and NSF BIGDATA DMS-1740325.

### REFERENCES

- [1] A. Agarwal, S. Negahban and M. J. Wainwright, [Fast global convergence rates of gradient methods for high-dimensional statistical recovery](#), *Ann. Statist.*, **40** (2012), 2452–2482.
- [2] M. Aharon, M. Elad and A. Bruckstein, [K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation](#), *IEEE Transactions on Signal Processing*, **54** (2006), 4311–4322.
- [3] D. Baron, M. B. Wakin, M. F. Duarte, S. Sarvotham and R. G. Baraniuk, [Distributed compressed sensing](#), *IEEE Transactions on Information Theory*, **52** (2006), 5406–5425.
- [4] J. A. Bazerque and G. B. Giannakis, [Distributed spectrum sensing for cognitive radio networks by exploiting sparsity](#), *IEEE Transactions on Signal Processing*, **58** (2010), 1847–1862.
- [5] J. D. Blanchard, M. Cermak, D. Hanle and Y. Jing, [Greedy algorithms for joint sparse recovery](#), *IEEE Transactions on Signal Processing*, **62** (2014), 1694–1704.
- [6] J. D. Blanchard, J. Tanner and K. Wei, [CGIHT: conjugate gradient iterative hard thresholding for compressed sensing and matrix completion](#), *Information and Inference: A Journal of the IMA*, **4** (2015), 289–327.
- [7] T. Blumensath and M. E. Davies, [Iterative hard thresholding for compressed sensing](#), *Applied and Computational Harmonic Analysis*, **27** (2009), 265–274.
- [8] T. Blumensath and M. E. Davies, [Normalized iterative hard thresholding: Guaranteed stability and performance](#), *IEEE Journal of Selected Topics in Signal Processing*, **4** (2010), 298–309.

- [9] R. Burden and J. Faires, *Numerical Analysis*, Cengage Learning, 2004.
- [10] J. Chen and X. Huo, *Theoretical results on sparse representations of multiple-measurement vectors*, *IEEE Transactions on Signal Processing*, **54** (2006), 4634–4643.
- [11] S. F. Cotter, B. D. Rao, K. Engan and K. Kreutz-Delgado, *Sparse solutions to linear inverse problems with multiple measurement vectors*, *IEEE Transactions on Signal Processing*, **53** (2005), 2477–2488.
- [12] W. Dai and O. Milenkovic, *Subspace pursuit for compressive sensing signal reconstruction*, *IEEE Transactions on Information Theory*, **55** (2009), 2230–2249.
- [13] M. E. Davies and Y. C. Eldar, *Rank awareness in joint sparse recovery*, *IEEE Transactions on Information Theory*, **58** (2012), 1135–1146.
- [14] D. L. Donoho, *Unconditional bases are optimal bases for data compression and for statistical estimation*, *Applied and Computational Harmonic Analysis*, **1** (1993), 100–115.
- [15] N. Durgin, C. Huang, R. Grotheer, S. Li, A. Ma, D. Needell and J. Qin, *Fast hyperspectral diffuse optical imaging method with joint sparsity*, in *41st Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2019.
- [16] M. Elad, *Sparse and Redundant Representations From Theory to Applications in Signal and Image Processing*, Springer, 2010.
- [17] P. Feng and Y. Bresler, *Spectrum-blind minimum-rate sampling and reconstruction of multi-band signals*, in *Acoustics, Speech and Signal Processing (ICASSP), 1996 IEEE International Conference on*, vol. 3, IEEE, 1996, 1688–1691.
- [18] S. Foucart, *Hard thresholding pursuit: an algorithm for compressive sensing*, *SIAM Journal on Numerical Analysis*, **49** (2011), 2543–2563.
- [19] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*, vol. 1, Birkhäuser/Springer, New York, , 2013.
- [20] R. Giryes, S. Nam, M. Elad, R. Gribonval and M. E. Davies, *Greedy-like algorithms for the cosparse analysis model*, *Linear Algebra and its Applications*, **441** (2014), 22–60.
- [21] Z. He, A. Cichocki, R. Zdunek and J. Cao, *CG-M-FOCUSS and its application to distributed compressed sensing*, in *International Symposium on Neural Networks*, Springer, 2008, 237–245.
- [22] Z. Jian, F. Yuli, Z. Qiheng and L. Haifeng, Split bregman algorithms for multiple measurement vector problem, *Multidim Syst Sign Process*.
- [23] J. M. Kim, O. K. Lee and J. C. Ye, *Compressive MUSIC: Revisiting the link between compressive sensing and array signal processing*, *IEEE Transactions on Information Theory*, **58** (2012), 278–301.
- [24] K. Lee, Y. Bresler and M. Junge, *Subspace methods for joint sparse recovery*, *IEEE Transactions on Information Theory*, **58** (2012), 3613–3641.
- [25] S. Li, D. Yang, G. Tang and M. B. Wakin, *Atomic norm minimization for modal analysis from random and compressed samples*, *IEEE Transactions on Signal Processing*, **66** (2018), 1817–1831.
- [26] H. Lu, X. Long and J. Lv, A fast algorithm for recovery of jointly sparse vectors based on the alternating direction methods, in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, 2011, 461–469.
- [27] J. R. Magnus, *On the concept of matrix derivative*, *Journal of Multivariate Analysis*, **101** (2010), 2200–2206.
- [28] A. Majumdar and R. Ward, *Rank awareness in group-sparse recovery of multi-echo MR images*, *Sensors*, **13** (2013), 3902–3921.
- [29] A. Majumdar and R. K. Ward, *Joint reconstruction of multiecho mr images using correlated sparsity*, *Magnetic Resonance Imaging*, **29** (2011), 899–906.
- [30] A. Majumdar and R. K. Ward, *Face recognition from video: An MMV recovery approach*, in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, IEEE, 2012, 2221–2224.
- [31] M. Mishali and Y. C. Eldar, *Reduce and boost: Recovering arbitrary sets of jointly sparse vectors*, *IEEE Transactions on Signal Processing*, **56** (2008), 4692–4702.
- [32] D. Needell and J. A. Tropp, *CoSaMP: Iterative signal recovery from incomplete and inaccurate samples*, *Applied and Computational Harmonic Analysis*, **26** (2009), 301–321.
- [33] D. Needell and R. Vershynin, *Signal recovery from incomplete and inaccurate measurements via regularized orthogonal matching pursuit*, *IEEE Journal of Selected Topics in Signal Processing*, **4** (2010), 310–316.

- [34] D. Needell and R. Ward, [Batched stochastic gradient descent with weighted sampling](#), in *International Conference Approximation Theory*, Springer, 2016, 279–306.
- [35] N. Nguyen, D. Needell and T. Woolf, [Linear convergence of stochastic iterative greedy algorithms with sparse constraints](#), *IEEE Transactions on Information Theory*, **63** (2017), 6869–6895.
- [36] N. Nguyen, S. Chin and T. Tran, [A Unified Iterative Greedy Algorithm for Sparsity-Constrained Optimization](#), 2012.
- [37] Y. C. Pati, R. Rezaifar and P. S. Krishnaprasad, [Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition](#), in *Proceedings of 27th Asilomar Conference on Signals, Systems and Computers*, IEEE, 1993, 40–44.
- [38] R. Schmidt, [Multiple emitter location and signal parameter estimation](#), *IEEE Transactions on Antennas and Propagation*, **34** (1986), 276–280.
- [39] J. A. Tropp, [Greed is good: Algorithmic results for sparse approximation](#), *IEEE Transactions on Information Theory*, **50** (2004), 2231–2242.
- [40] J. A. Tropp, A. C. Gilbert and M. J. Strauss, [Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit](#), *Signal Processing*, **86** (2006), 572–588.
- [41] X.-T. Yuan, P. Li and T. Zhang, [Gradient hard thresholding pursuit for sparsity-constrained optimization](#), in *International Conference on Machine Learning*, 2014, 127–135.
- [42] T. Zhang, [Sparse recovery with orthogonal matching pursuit under RIP](#), *IEEE Transactions on Information Theory*, **57** (2011), 6215–6221.

Received December 2019; 1st revision May 2020; 2nd revision July 2020.

*E-mail address:* jing.qin@uky.edu

*E-mail address:* shuangli@mymail.mines.edu

*E-mail address:* deanna@math.ucla.edu

*E-mail address:* anna.ma@uci.edu

*E-mail address:* grotheerre@wofford.edu

*E-mail address:* chenxi.huang@yale.edu

*E-mail address:* njdurgin@gmail.com