# Nonconvex Regularized Gradient Projection Sparse Reconstruction for Massive MIMO Channel Estimation

Pengxia Wu and Julian Cheng, *Senior Member, IEEE*

*Abstract*—Novel sparse reconstruction algorithms are proposed for beamspace channel estimation in massive multiple-input multiple-output systems. The proposed algorithms minimize a least-squares objective having a nonconvex regularizer. This regularizer removes the penalties on a few large-magnitude elements from the conventional $\ell_1$-norm regularizer, and thus it only forces penalties on the remaining elements that are expected to be zeros. Accurate and fast reconstructions can be achieved by performing gradient projection updates within the framework of difference of convex functions (DC) programming. A double-loop algorithm and a single-loop algorithm are proposed via different DC decompositions, and these two algorithms have distinct computational complexities and convergence rates. An extension algorithm is further proposed by designing new step sizes for the single-loop algorithm. The extension algorithm has a faster convergence rate and can achieve approximately the same level of accuracy as the proposed double-loop algorithm. Numerical results show significant advantages of the proposed algorithms over existing reconstruction algorithms in terms of reconstruction accuracies and runtimes. Compared with the benchmark channel estimation approaches, the proposed algorithms can achieve smaller channel reconstruction error and higher achievable spectral efficiency.

*Index Terms*—DC programming, gradient projection, massive MIMO, nonconvex optimization, sparse channel reconstruction

## I. INTRODUCTION

As a key technology for the fifth-generation and beyond communication systems, massive multiple-input multiple-output (MIMO) heavily relies on accurate knowledge of channel state information (CSI) to reap potential performance benefits from a large number of antennas. However, acquiring downlink CSI in a resource-efficient manner is problematic for frequency division duplex (FDD) systems where uplink and downlink channels operate in different spectrum bands. For such a system, downlink channels are estimated at the user equipments (UEs) and then the UEs send back the estimated CSI to the base station (BS); both the overheads of downlink pilot and CSI feedback are proportional to the number of BS antennas, which can be large for massive MIMO systems. To reduce the overhead of downlink CSI acquisition, one can develop sparse channel estimation methods by exploiting channel sparsity in certain domains. More specifically, the unknown massive MIMO channels can be represented as high-dimensional sparse vectors in certain basis [1]. Based on valid sparse representations of massive

MIMO channels, transmitting pilots over the channel can be regarded as a linear mapping of sparse channel vectors onto a compact subspace to obtain lower-dimensional observations. Then, the full-dimensional channel vectors are expected to be reconstructed from the lower-dimensional observations. In such sparse channel estimation methods, the length of pilot sequence for each antenna is no longer proportional to the number of BS antennas, but depends on the sparsity level of the channel. Thus, the length of downlink pilot sequences can be much less than the number of BS antennas, and the CSI acquisition overhead for FDD massive MIMO will be substantially reduced.

Sparse channel estimation [2] has been proposed for many traditional communication applications such as orthogonal frequency-division multiplexing systems [3], ultra-wideband communications [4], pulse-shaping multicarrier systems [5] and underwater acoustic communications [6]. The channels in these applications typically have sparse impulse responses because the corresponding multipath channel has a large delay spread and a small number of nonzero taps [2]. Nowadays, sparse channel estimation has been increasingly investigated for massive MIMO systems. Due to a limited number of scattering clusters and the small angular spread for each scattering cluster, the massive MIMO channel in angular-domain (beamspace) exhibits a sparse feature [7], i.e., the majority of channel energy is occupied by a few dimensions and most channel coefficients are either zero or nearly zero. Various compressive sensing-based beamspace channel estimation schemes have been developed for massive MIMO [8]–[17] and millimeter-wave (mmWave) massive MIMO [18]–[26]. Several orthogonal matching pursuit (OMP) based algorithms have been proposed for sparse channel estimations of mmWave massive MIMO [20]–[23]. The quantized compressive sensing technique has been applied to the channel estimation for mmWave massive MIMO with few-bit analog-to-digital converters [24]. The image reconstruction technique has been applied to mmWave MIMO channel estimation to develop an SCAMPI [1] algorithm based on the approximate message passing algorithm [26].

Apart from the compressive sensing-based channel estimation scheme, there are other channel estimation schemes such as the codebook-based method [27], the parametric channel estimation [28] and the subspace decomposition-based channel

P. Wu and J. Cheng are with the School of Engineering, The University of British Columbia, Kelowna, BC V1X 1V7, Canada (e-mail:pengxia.wu@alumni.ubc.ca, julian.cheng@ubc.ca). This work was supported by an NSERC Discovery Grant.

---

[1]The SCAMPI algorithm is the sparse noninformative parameter estimator-based cosparse analysis approximate message passing for imaging algorithm [26].

estimation [29], [30]. The codebook-based beamforming [27] only estimates the CSI implicitly; however, explicit CSI is often required for general beamforming tasks. The parametric channel estimation can achieve super-resolutional channel estimation without a discrete angular-domain channel assumption, but it requires the array manifold as a priori. Moreover, real-time CSI acquisition is challenging when using the parametric channel estimation in a high-mobility environment, since it is time-consuming to acquire the angle of arrivals (AoAs), angle of departures (AoDs) and path gains of all channel paths. The subspace decomposition-based channel estimation requires iterative channel sounding operations between the BS and UEs, and this echo process accumulates noise and thus can degrade channel estimation accuracy.

Among various channel estimation methods, the compressive sensing-based channel estimation is attractive since it can obtain instantaneous CSI based on the channel sparsity feature, without requiring statistical priori. Therefore, we focus on the compressive sensing-based scheme and aim to develop efficient algorithms for sparse channel reconstruction in massive MIMO systems. Sparse channel reconstruction requires the reconstruction of high-dimensional sparse channel vectors from lower-dimensional pilot observations. The ultimate sparse reconstruction problem is an NP-hard $\ell_0$-minimization optimization problem [31]. One approach is to replace the $\ell_0$-norm by other simple functions to seek approximate solutions. For this approach, the most popular method is to use the $\ell_1$-norm to approximate the $\ell_0$-norm, and this method is known as $\ell_1$-relaxation or convex relaxation. Various well-known $\ell_1$-relaxation algorithms have been proposed such as the $l1\_ls$ [32], Bregman iterative regularization [33], gradient projection for sparse reconstruction (GPSR) [34], sparse reconstruction by separable approximation (SpaRSA) [35], iterative shrinkage-thresholding algorithm (ISTA) [36] and the fast iterative shrinkage-thresholding algorithm (FISTA) [37]. These $\ell_1$-relaxation algorithms solve convex optimizations to guarantee global optimal solutions; however, an $\ell_1$-norm is a loose approximation of $\ell_0$-norm, leading to biased estimates. To achieve tighter approximations, nonconvex functions have been proposed such as the $\ell_q$-norm for $0 \leq q \leq 1$ [38] and the difference of two continuous functions [39]. Another popular method is the greedy approach, including some representative algorithms such as OMP [40], [41], CoSaMP [42] and the least angle regression (LARS) [43]. These greedy methods work well for sufficiently sparse vectors, but the reconstruction accuracies and speeds degrade severely when the sparsity of reconstructing vectors reduces.

In this paper, we propose to use the classic difference of convex functions (DC) programming and gradient projection descent algorithm to solve the fundamental sparse recovery problem. We develop novel algorithms that are applied to beamspace channel estimation for massive MIMO systems. Instead of approximating the $\ell_0$-norm, we represent exactly the $\ell_0$-norm constraint by introducing the top-$(K, 1)$ norm

[44][2], which is simply the sum of $K$ largest-magnitude elements of a vector in terms of absolute values. Thus, the original $\ell_0$-minimization sparse reconstruction problem can be equivalently transformed into a least squares optimization penalized by a nonconvex regularizer. The nonconvex regularizer is more favorable than the $\ell_1$-regularizer because it removes the penalties on large-magnitude elements. Although the resulting optimization problem is nonconvex, we can solve it by employing the DC programming framework to transform the problem into a list of convex subproblems, and then we elegantly apply a gradient projection descent method to solve these convex subproblems. We propose three different DC gradient projection sparse reconstruction (DC-GPSR) algorithms. The proposed algorithms can achieve similar reconstruction accuracies with different time complexities.

The proposed DC-GPSR algorithms have the following advantages in solving a sparse reconstruction problem compared with conventional sparse recovery algorithms. First, while conventional algorithms approximate the $\ell_0$-constraint by $\ell_1$-relaxation, we can represent exactly the $\ell_0$-norm constraint using a DC function, i.e., subtracting the top-$(K, 1)$ norm from the $\ell_1$-norm of the reconstructing vector. Second, we formulate the sparse recovery as a nonconvex optimization problem whose objective function can be decomposed into DC functions, so that we can naturally adopt the classic DC programming for the remaining optimization tasks. Furthermore, we can perfectly incorporate the gradient projection descent algorithm into the DC programming framework. On one hand, both the gradients of top-$(K, 1)$ norm and $\ell_1$-norm can be easily and quickly calculated. On the other hand, we transform the convex subproblem into a bound-constrained quadratic program (BCQP) with a nonnegativity constraint so that the orthogonal projection operation is simple to compute. Finally, we show that by different DC decompositions and step size strategies we can derive different algorithms that can achieve similar accuracies with different convergence rates. More specifically, the proposed Algorithm 1 has a normal double-loop procedure of the DC programming algorithm, while Algorithm 2 simplifies the double-loop procedure to be a simple single-loop algorithm by using a special DC decomposition. Based on Algorithm 2, which has a fixed step size, we further design new step sizes and propose Algorithm 3 that has a single loop and an accelerated convergence rate.

The contributions of this paper are summarized as follows:

- We propose a double-loop DC-GPSR algorithm (Algorithm 1), which has been reported in our preliminary study [45]. Incorporated by gradient projection descent, the DC programming framework leads to a double-loop DC-GPSR algorithm that shows high accuracy and robustness.
- We propose a single-loop DC-GPSR algorithm (Algorithm 2), which applies a special DC decomposition to the

objective function such that the convex subproblem in the DC programming has a closed-form solution. Compared to Algorithm 1, Algorithm 2 avoids inner loops and solves the DC programming subproblem in a one-step update.

- We make an important observation that an update of the single-loop DC-GPSR algorithm can be interpreted as an update of gradient projection descent with a required step size. Thus, we adopt the Barzilai-Borwein (BB) step size to generalize the basic single-loop DC-GPSR algorithm and propose Algorithm 3 to improve convergence significantly. Compared to Algorithm 1, the runtime of Algorithm 3 is demonstrated to be about $1.5$ times faster in noiseless scenarios and about 20 times faster in noisy scenarios for wide-range signal-to-noise (SNR) values.

## II. SYSTEM MODEL

We consider a downlink massive MIMO system having $N_t$ transmit antennas at the BS and $N_r$ receiver antennas at the UE. Assuming the block-fading narrowband multipath channel [3], the channel can be represented by the Saleh-Valenzuela model [4] as [45]

$$\mathbf{H}_s = \sqrt{N_r N_t} \sum_{l=1}^{N_p} \alpha_l \boldsymbol{\alpha}_r(\theta_{r,l}) \boldsymbol{\alpha}_t^H(\theta_{t,l}) \tag{1}$$

where $\mathbf{H}_s \in \mathbb{C}^{N_r \times N_t}$ is the channel matrix, $N_p$ is the number of paths, $l$ denotes the multipath index, $\alpha_l$ is the complex channel gain of the $l$th path, $\boldsymbol{\alpha}_r(\theta_{r,l})$ and $\boldsymbol{\alpha}_t(\theta_{t,l})$ are respectively the array steering vectors of receiver and transmitter array for the $l$th path, and the corresponding angle of arrival (AoA) and the angle of departure (AoD) are $\theta_{r,l}$ and $\theta_{t,l}$. An array steering vector represents the array phase profile as a function of physical AoA or AoD. For a one-dimensional uniform linear array (ULA) consisting of $N$ elements, the array steering vector can be expressed as $\boldsymbol{\alpha}(\theta) = \frac{1}{\sqrt{N}}[1, e^{-j2\pi\vartheta}, e^{-j4\pi\vartheta}, ..., e^{-j2\pi\vartheta(N-1)}]^T$, where $\vartheta \in [-1, 1]$ is the normalized spatial angle, which is related to the physical angle $\theta \in [-\pi/2, \pi/2]$ by $\vartheta = \frac{d}{\lambda}\sin(\theta)$, and where $d = \lambda/2$ is the antenna spacing and $\lambda$ is the wavelength.

The channel matrix $\mathbf{H}_s$ can be represented using a two-dimensional Fourier transformation of the angular-domain channel matrix $\mathbf{H}_a$ by [7]

$$\begin{aligned} \mathbf{H}_s &= \mathbf{U}_r \mathbf{H}_a \mathbf{U}_t^H \\ &= \sum_{i=0}^{N_r-1} \sum_{j=0}^{N_t-1} H_a(i,j) \boldsymbol{\alpha}_r(\theta_{r,i}) \boldsymbol{\alpha}_t^H(\theta_{t,j}) \end{aligned} \tag{2}$$

[3] In this work, we consider narrowband block-fading channels of massive MIMO. It is worth mentioning that our proposed algorithms are generic sparse reconstruction algorithm that can be conveniently applied to other types of sparse channels as long as the channel has sparse representations in certain domains. For example, our proposed algorithms can be applied to channel estimation of wideband time-varying channels for massive MIMO systems. The only required modification is to exploit the channel sparsity in a virtual angular-delay-Doppler domain.

[4] For simplicity, we adopt the basic Saleh-Valenzuela channel model by assuming that the each scattering cluster consists of a single path. Our proposed approach can also be applied to the extended Saleh-Valenzuela channel model, where each cluster consists of multiple sub-paths.

where $\mathbf{U}_t \in \mathbb{C}^{N_t \times N_t}$ and $\mathbf{U}_r \in \mathbb{C}^{N_r \times N_r}$ are unitary digital Fourier transform (DFT) matrices, and they can be expressed using array steering vectors as $\mathbf{U}_t = [\boldsymbol{\alpha}_t(\theta_{t,0}), \boldsymbol{\alpha}_t(\theta_{t,1}), ..., \boldsymbol{\alpha}_t(\theta_{t,N_t-1})]^T$, $\mathbf{U}_r = [\boldsymbol{\alpha}_r(\theta_{r,0}), \boldsymbol{\alpha}_r(\theta_{r,1}), ..., \boldsymbol{\alpha}_r(\theta_{r,N_r-1})]^T$, where $\{\theta_{t,0}, ..., \theta_{t,N_t-1}\}$ and $\{\theta_{r,0}, ..., \theta_{r,N_r-1}\}$ are the virtual AoDs and AoAs defined by array elements of transmitter and receiver. The beamspace channel $\mathbf{H}_a$ is sparse due to the limited number of multiple paths and the high dimensionality of massive MIMO channels [45].

By transmitting the known pilots $\mathbf{P} \in \mathbb{C}^{N_t \times L}$ from the BS through the channel, we have the pilot signal received at UE as

$$\tilde{\mathbf{R}} = \mathbf{H}_s \mathbf{P} + \tilde{\mathbf{W}} \tag{3}$$

where $\tilde{\mathbf{R}} \in \mathbb{C}^{N_r \times L}$ is the received pilot observations, and where $L$ is the length of training pilot sequence for each antenna; $\mathbf{H}_s \in \mathbb{C}^{N_r \times N_t}$ is the channel matrix in the spatial domain, $\mathbf{P}$ is the transmitted pilot matrix; $\tilde{\mathbf{W}} \in \mathbb{C}^{N_r \times L}$ is the additive white Gaussian noise matrix whose elements are independent identical distributed (i.i.d.) complex Gaussian random variables having a mean of zero and a variance of $\sigma_n^2$. Conventional estimation methods such as the linear minimum mean square error (LMMSE) or the least squares (LS) require $L \geq N_t$ to obtain accurate estimates, whereas the beamspace channel estimation can reconstruct $\hat{\mathbf{H}}_s$ for $L \ll N_t$. From (2), we have $\mathbf{H}_s = \mathbf{U}_r \mathbf{H}_a \mathbf{U}_t^H$, which can be substituted in (3) to have

$$\tilde{\mathbf{R}} = \mathbf{U}_r \mathbf{H}_a \mathbf{U}_t^H \mathbf{P} + \tilde{\mathbf{W}}. \tag{4}$$

By taking the transposes and right multiplications with $\mathbf{U}_r^H$ on both sides of (4), we have

$$\underbrace{\tilde{\mathbf{R}}^T \mathbf{U}_r^H}_{\mathbf{R}} = \underbrace{\mathbf{P}^T \mathbf{U}_t^*}_{\mathbf{S}} \underbrace{\mathbf{H}_a^T \mathbf{U}_r \mathbf{U}_r^H}_{\mathbf{H}} + \underbrace{\tilde{\mathbf{W}}^T \mathbf{U}_r^H}_{\mathbf{W}} \tag{5}$$

We simply write (5) as

$$\mathbf{R} = \mathbf{S}\mathbf{H} + \mathbf{W} \tag{6}$$

where $\mathbf{R} = \tilde{\mathbf{R}}^T \mathbf{U}_r^H \in \mathbb{C}^{L \times N_r}$, $\mathbf{H} = \mathbf{H}_a^T \mathbf{U}_r \mathbf{U}_r^H \in \mathbb{C}^{N_t \times N_r}$, $\mathbf{W} \in \mathbb{C}^{L \times N_r}$, and the matrix $\mathbf{S} = \mathbf{P}^T \mathbf{U}_t^* \in \mathbb{C}^{L \times N_t}$. The beamspace channel estimation $\hat{\mathbf{H}} = \hat{\mathbf{H}}_a^T$ can be obtained by solving the linear equation (6) given the known measurement matrix $\mathbf{S}$ and measurements $\mathbf{R}$. To this end, both an efficient sparse recovery algorithm and a proper measurement matrix are essential to the estimation performances. Eq. (5) suggests that the pilot matrix $\mathbf{P}$ depends on the measurement matrix $\mathbf{S}$. To design the pilot matrix, we can first determine a measurement matrix $\mathbf{S}$, and then obtain the pilot matrix by $\mathbf{P} = \mathbf{S}^T \mathbf{U}_t$, where $\mathbf{U}_t$ is a DFT matrix. Due to $\|\mathbf{S}\|_F^2 = \|\mathbf{P}^T \mathbf{U}_t^*\|_F^2$, the power constraint on a pilot matrix $\|\mathbf{P}\|_F^2 = P$ can be imposed via scaling the measurement matrix by $\mathbf{S} = P\frac{\mathbf{S}'}{\|\mathbf{S}'\|_F^2}$, where $\|\cdot\|_F$ represents the Frobenius norm of a matrix. Various random matrices can be adopted as the measurement matrix. After evaluating the proposed reconstruction algorithms that adopt various random matrices (such as Gaussian, Bernoulli, and partial Fourier matrices), we conclude that these random matrices have similar reconstruction performances.

To solve the multiple measurement vector (MMV) problem (6), we propose the column-wise broadcasting vectorization method to vectorize the channel matrix $\mathbf{H}$. Thus, the MMV problem (6) can be transformed into parallel single measurement vector (SMV) problems. Therefore, although the proposed algorithms are developed for an SMV model, they can be extended to the MMV problem. We express the linear equation (6) as

$$[\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_{N_r}] = \mathbf{S}[\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_{N_r}] + [\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_{N_r}] \quad (7)$$

where $\mathbf{r}_i$ is the $i$th $(1 \leq i \leq N_r)$ column of $\mathbf{R}$; $\mathbf{h}_i$ is the $i$th $(1 \leq i \leq N_r)$ column of $\mathbf{H}$; $\mathbf{w}_i$ is the $i$th $(1 \leq i \leq N_r)$ column of $\mathbf{W}$. For the $i$th column of $\mathbf{R}, \mathbf{H}$ and $\mathbf{W}$ in (6), we have

$$\mathbf{r}_i = \mathbf{S}\mathbf{h}_i + \mathbf{w}_i. \quad (8)$$

We can further express (8) by an equivalent real-valued linear equation as

$$\underbrace{\begin{bmatrix} \Re(\mathbf{r}_i) \\ \Im(\mathbf{r}_i) \end{bmatrix}}_{\mathbf{y}} = \underbrace{\begin{bmatrix} \Re(\mathbf{S}) & -\Im(\mathbf{S}) \\ \Im(\mathbf{S}) & \Re(\mathbf{S}) \end{bmatrix}}_{\mathbf{\Phi}} \underbrace{\begin{bmatrix} \Re(\mathbf{h}_i) \\ \Im(\mathbf{h}_i) \end{bmatrix}}_{\mathbf{x}} + \underbrace{\begin{bmatrix} \Re(\mathbf{w}_i) \\ \Im(\mathbf{w}_i) \end{bmatrix}}_{\mathbf{n}}. \quad (9)$$

Thus, we obtain an SMV problem (9). For the presentational simplicity, we concisely express (9) as

$$\mathbf{y} = \mathbf{\Phi}\mathbf{x} + \mathbf{n}. \quad (10)$$

In the remainder of this paper, we uniquely refer the real-form vector $\mathbf{x} = [\Re(\mathbf{h}_i)^T, \Im(\mathbf{h}_i)^T]^T$ in (10) as the equivalent sparse channel vector, since the vector $\mathbf{x}$ contains the equivalent information of a complex beamspace channel $\mathbf{h}_i$. We omit the subscript $i$ since (10) is valid for arbitrary beamspace channel $\mathbf{h}_i$ for $1 \leq i \leq N_r$.

## III. DC REPRESENTATION FOR SPARSE CONSTRAINT

In this section, we first briefly review the sparse reconstruction optimization problem. Then, we introduce the top-$(K,1)$ norm [44] to represent exactly the $\ell_0$-norm constraint in the sparse reconstruction optimization problem. Also, we provide a threshold to determine the range of the penalty parameter.

According to (10), reconstructing $\mathbf{x}$ from $\mathbf{y}$ and $\mathbf{\Phi}$ using the sparsity as a priori is an NP-hard $\ell_0$-minimization problem, which is defined as [46]

$$\begin{aligned} \min_{\mathbf{x}} \quad & \|\mathbf{x}\|_0 \\ \text{s.t.} \quad & \|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2^2 \leq \tau \end{aligned} \quad (11)$$

where $\tau$ is nonnegative and real. Problem (11) can be rewritten in an equivalent form as [46]

$$\begin{aligned} \min_{\mathbf{x}} \quad & \|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2^2 \\ \text{s.t.} \quad & \|\mathbf{x}\|_0 \leq K \end{aligned} \quad (12)$$

where $K$ is an upper bound for the number of nonzero elements in $\mathbf{x}$, and it is uniquely determined by the parameter $\tau$ in (11).

Instead of using the common $\ell_1$-relaxation, we introduce the top-$(K,1)$ norm to seek an equivalent expression for the

constraint $\|\mathbf{x}\|_0 \leq K$ in the original problem (12). The top-$(K,1)$ norm $\|\mathbf{x}\|_{K,1}$ is defined as the sum of the largest $K$ elements of the vector $\mathbf{x}$ in terms of the absolute values, namely

$$\|\mathbf{x}\|_{K,1} := |x_{(1)}| + |x_{(2)}| + \cdots + |x_{(K)}| \quad (13)$$

where $|x_{(i)}|$ denotes the element whose absolute value is the $i$th-largest among the $N$ elements of the vector $\mathbf{x}$, i.e., $|x_{(1)}| \geq |x_{(2)}| \geq \cdots \geq |x_{(N)}|$. The constraint $\|\mathbf{x}\|_0 \leq K$ is equivalent to the statement that the $(K+1)$th-largest element of the vector $\mathbf{x}$ is zero, i.e., $\|\mathbf{x}\|_{K+1,1} - \|\mathbf{x}\|_{K,1} = 0$. Thus, we have an equivalent relationship between the following two statements [44]

$$\|\mathbf{x}\|_0 \leq K \Leftrightarrow \|\mathbf{x}\|_1 - \|\mathbf{x}\|_{K,1} = 0. \quad (14)$$

Since both $\|\mathbf{x}\|_1$ and $\|\mathbf{x}\|_{K,1}$ are convex, we say the equality $\|\mathbf{x}\|_1 - \|\mathbf{x}\|_{K,1} = 0$ is an exact DC representation for the sparsity constraint. By replacing the sparsity constraint $\|\mathbf{x}\|_0 \leq K$ in (12) using the DC constraint $\|\mathbf{x}\|_1 - \|\mathbf{x}\|_{K,1} = 0$, we rewrite the sparse reconstruction problem as

$$\begin{aligned} \min_{\mathbf{x}} \quad & \|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2^2 \\ \text{s.t.} \quad & \|\mathbf{x}\|_1 - \|\mathbf{x}\|_{K,1} = 0. \end{aligned} \quad (15)$$

Using an appropriate Lagrange multiplier $\rho$, from (15) we obtain the following unconstrained optimization problem

$$\min_{\mathbf{x}} \quad \frac{1}{2}\|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2^2 + \rho(\|\mathbf{x}\|_1 - \|\mathbf{x}\|_{K,1}) := F(\mathbf{x}) \quad (16)$$

where $\rho$ is the regularization parameter that balances the data consistency and the penalty term. Our formulated optimization problem (16) differs from the conventional $\ell_1$-regularized sparse reconstruction [5] only in terms of the subtracted top-$(K,1)$ norm $\|\mathbf{x}\|_{K,1}$ in its penalty term. The regularizer $\rho(\|\mathbf{x}\|_1 - \|\mathbf{x}\|_{K,1})$ is better than an $\ell_1$-norm regularizer because it removes the penalties on the $K$ largest-magnitude elements [6].

To ensure the equivalence between the unconstrained problem (16) and the constrained problem (15), the following theorem specifies the range for the penalty parameter.

***Theorem 1:*** Let $\mathbf{x}_{\rho^*}$ be an optimal solution to (16) with given $\rho^*$. Suppose there exists a constant $q > 0$ such that $\|\mathbf{x}_{\rho^*}\|_2 \leq q$ for any $\rho^* > 0$. Then $\mathbf{x}_{\rho^*}$ is also optimal to (15) if

$$\rho^* \geq \max_i \{q(\|\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{e}_i\|_2 + |(\mathbf{\Phi}^T\mathbf{\Phi})_{ii}|/2) + |(\mathbf{\Phi}^T\mathbf{y})_i|\}$$

where $1 \leq i \leq N_t$; $\mathbf{e}_i$ represents the unit vector in which the $i$th element is one while the other elements are zeros; $(\mathbf{\Phi}^T\mathbf{\Phi})_{ii}$ represents the $i$th diagonal elements of matrix $\mathbf{\Phi}^T\mathbf{\Phi}$; $(\mathbf{\Phi}^T\mathbf{y})_i$ indicates the $i$th element of the vector $\mathbf{\Phi}^T\mathbf{y}$.
*Proof:* See Appendix A.

Theorem 1 indicates that given a penalty parameter $\rho$ having a suitably large value, the optimal solution to (16) is also the optimal solution to (15). To calculate the lower threshold

---

[5] $\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1$

[6] In practice, if the sparsity level of reconstructing vector is already known, we set $K$ as the number of nonzero elements of reconstruction vectors; otherwise, we can use cross validation to determine an approximation of $K$.

$\rho^*$, we can first estimate a constant $q$ such that $\|\mathbf{x}_{\rho^*}\|_2 \leq q$. In practice, to avoid high computational complexity of the inequality in Theorem 1, we can use the cross validation to select a suitable value for the penalty parameter $\rho$.

## IV. DOUBLE-LOOP DC GRADIENT PROJECTION DESCENT FOR SPARSE RECONSTRUCTION

We have formulated the sparse reconstructions into a nonconvex optimization problem (16). In this section, we use DC programming and gradient projection descent to solve (16) and propose a double-loop DC-GPSR algorithm.

### A. DC Programming Framework

For a nonconvex unconstrained optimization problem

$$\min_{\mathbf{x}} f(\mathbf{x}) - g(\mathbf{x}) \tag{17}$$

where $f(\mathbf{x})$ and $g(\mathbf{x})$ are two convex functions. The DC programming method solves the following convex subproblem at the $t$th-iteration,

$$\min_{\mathbf{x}} f(\mathbf{x}) - \mathbf{x}^T \partial g(\mathbf{x}^{t-1}) \tag{18}$$

where the second convex function $g(\mathbf{x})$ in (17) is linearized by $\mathbf{x}^T \partial g(\mathbf{x}^{t-1})$ in (18), and where $\partial g(\mathbf{x}^{t-1})$ represents the gradient (or subgradient) of $g(\mathbf{x}^{t-1})$ with respect to $\mathbf{x}^{t-1}$. The DC algorithm framework can be outlined as follows:

1. **Start:** Given a starting point $\mathbf{x}^0$, and a terminate condition
2. **Repeat:** For $t = 1, 2, ...$
   Compute the gradient (or subgradient) $\partial g(\mathbf{x}^{t-1})$.
   Solve the convex subproblem (18) to obtain $\mathbf{x}^t$.
3. **End:** Until a terminate condition is satisfied.

The DC programming is an iterative algorithm framework that can ensure global convergence, which means the DC algorithm can converge from an arbitrary initial point [7].

### B. A Double-Loop DC-GPSR Algorithm

Following the DC programming framework, we decompose our objective function in (16) as the difference of the two convex functions of $f(\mathbf{x})$ and $g(\mathbf{x})$

$$\min_{\mathbf{x}} \underbrace{\frac{1}{2}\|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2^2 + \rho\|\mathbf{x}\|_1}_{f(\mathbf{x})} - \underbrace{\rho\|\mathbf{x}\|_{K,1}}_{g(\mathbf{x})}. \tag{19}$$

At the $t$th-iteration, we solve the following convex subproblem

$$\min_{\mathbf{x}} \quad f(\mathbf{x}) - \rho\mathbf{x}^T \partial\|\mathbf{x}^{t-1}\|_{K,1} \tag{20}$$

where $\partial\|\mathbf{x}^{t-1}\|_{K,1}$ denotes the subgradient of top-$(K,1)$ norm of $\mathbf{x}^{t-1}$, and where the superscript $t-1$ indicates the $(t-1)$th

update. The subgradient $\partial\|\mathbf{x}\|_{K,1}$ of the top-$(K,1)$ norm of $\mathbf{x}$ is defined as [44]

$$\partial\|\mathbf{x}\|_{K,1} := \arg\max_{\mathbf{w}} \left\{ \sum_{i=1}^{N} x_i w_i \,\Big|\, \sum_{i=1}^{N} |w_i| = K, w_i \in [-1, 1] \right\}. \tag{21}$$

By substituting the $f(\mathbf{x})$ defined in (19) into (20) and denoting $\mathbf{w}_x^{t-1}$ by a feasible value for the subgradient $\partial\|\mathbf{x}^{t-1}\|_{K,1}$, we write the subproblem (20) as

$$\min_{\mathbf{x}} \quad \frac{1}{2}\|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2^2 + \rho\|\mathbf{x}\|_1 - \rho\mathbf{x}^T \mathbf{w}_x^{t-1} \tag{22}$$

where $\mathbf{w}_x^{t-1} \in \partial\|\mathbf{x}^{t-1}\|_{K,1}$. A feasible subgradient $\mathbf{w}_x^{t-1}$ can be simply obtained by setting the signs of the first $K$ largest elements of $|\mathbf{x}^{t-1}|$ to the corresponding elements of $\mathbf{w}_x^{t-1}$, i.e., $(\mathbf{w}_x)_{(i)}^{t-1} = \mathrm{sign}(x_{(i)}^{t-1})$, where the subscript $i$ indicates the $i$th element of a vector, and setting the other elements of $\mathbf{w}_x^{t-1}$ to be zeros.

We obtain a convex subproblem (22). We will turn (22) into a constrained quadratic problem so that we can solve it using the gradient projection descent method. Specifically, we split the positive and negative part of $\mathbf{x}$, and represent $\mathbf{x}$ as the difference of its positive part $\mathbf{u}$ and its negative part $\mathbf{v}$, that is

$$\mathbf{x} = \mathbf{u} - \mathbf{v}, \quad \mathbf{u} \succeq \mathbf{0}, \mathbf{v} \succeq \mathbf{0}. \tag{23}$$

where $\mathbf{u} = (\mathbf{x})_+, \mathbf{v} = (-\mathbf{x})_+$, and where $(\cdot)_+$ is the positive-taking operation that retains the positive elements and sets the other elements to be zeros. More precisely, $(\mathbf{x})_+$ represents for each element $x$ in vector $\mathbf{x}$ we take $(x)_+ = \max\{0, x\}$; $(-\mathbf{x})_+$ represents for each element $-x$ in vector $-\mathbf{x}$ we take $(-x)_+ = \max\{0, -x\}$. Noticing that $\|\mathbf{x}\|_1 = \mathbf{1}^T\mathbf{u} + \mathbf{1}^T\mathbf{v}$, the subproblem (22) can be written as a BCQP problem

$$\begin{aligned} \min_{\mathbf{u},\mathbf{v}} \quad & \frac{1}{2}\|\mathbf{y} - \mathbf{\Phi}(\mathbf{u} - \mathbf{v})\|_2^2 + \rho\mathbf{1}^T\mathbf{u} + \rho\mathbf{1}^T\mathbf{v} \\ & - \rho\mathbf{u}^T\mathbf{w}_u^{t-1} - \rho\mathbf{v}^T\mathbf{w}_v^{t-1} \\ \text{s.t.} \quad & \mathbf{u} \succeq \mathbf{0}, \mathbf{v} \succeq \mathbf{0} \end{aligned} \tag{24}$$

where $\mathbf{w}_u^{t-1}$ and $\mathbf{w}_v^{t-1}$, respectively, represent the positive and negative part of $\mathbf{w}_x^{t-1}$, i.e., $\mathbf{w}_u^{t-1} = (\mathbf{w}_x^{t-1})_+$, $\mathbf{w}_v^{t-1} = (-\mathbf{w}_x^{t-1})_+$. Let $\mathbf{z}$ denote the concatenation of $\mathbf{u}$ and $\mathbf{v}$, i.e., $\mathbf{z} = [\mathbf{u}^T, \mathbf{v}^T]^T$, we rewrite (24) into a compact form

$$\begin{aligned} \min_{\mathbf{z}} \quad & \frac{1}{2}\mathbf{z}^T\mathbf{B}\mathbf{z} + \mathbf{c}^T\mathbf{z} := G(\mathbf{z}), \\ \text{s.t.} \quad & \mathbf{z} \succeq \mathbf{0} \end{aligned} \tag{25}$$

where

$$\mathbf{z} = \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \mathbf{\Phi}^T\mathbf{\Phi} & -\mathbf{\Phi}^T\mathbf{\Phi} \\ -\mathbf{\Phi}^T\mathbf{\Phi} & \mathbf{\Phi}^T\mathbf{\Phi} \end{bmatrix},$$

$$\mathbf{c} = \begin{bmatrix} -\mathbf{\Phi}^T\mathbf{y} \\ \mathbf{\Phi}^T\mathbf{y} \end{bmatrix} + \rho\mathbf{1}^T - \rho\mathbf{w}_z^{t-1}$$

where $\mathbf{1}^T$ represents an all-ones column vector having the same dimension as $\mathbf{z}$, and $\mathbf{w}_z^{t-1} = [(\mathbf{w}_u^{t-1})^T, (\mathbf{w}_v^{t-1})^T]^T$. Note that $\mathbf{w}_z^{t-1}$ is a subgradient of $\|\mathbf{z}^{t-1}\|_{K,1}$, i.e., $\mathbf{w}_z^{t-1} \in \partial\|\mathbf{z}^{t-1}\|_{K,1}$. Since $\mathbf{z}^{t-1} \succeq \mathbf{0}$, a feasible subgradient $\mathbf{w}_z^{t-1}$ can be an indicator vector having either one-valued or zero-valued elements, where the indices for the one-valued elements of

$\mathbf{w}_z^{t-1}$ correspond to the indices of the $K$-largest elements of $\mathbf{z}^{t-1}$. Eq. (25) is an equivalent problem for the subproblem (22). We apply the gradient projection descent method to solve (25), thus the $k$th update is

$$\mathbf{z}^{(k+\frac{1}{2})} = Proj\left(\mathbf{z}^{(k)} - \alpha^k \nabla G(\mathbf{z}^{(k)})\right),$$
$$\mathbf{z}^{(k+1)} = \mathbf{z}^{(k)} + \beta^k(\mathbf{z}^{(k+\frac{1}{2})} - \mathbf{z}^{(k)}) \quad (26)$$

where $\alpha^k > 0$ is the step size and it can be determined by the BB step size, which can be calculated as $\alpha^t = \frac{\|\mathbf{z}^k - \mathbf{z}^{k-1}\|^2}{(\mathbf{z}^k - \mathbf{z}^{k-1})^T(G(\mathbf{z}^k) - G(\mathbf{z}^{k-1}))}$; $\beta^k \in (0,1]$ is another step size to ensure the monotonic-decrease of the objective and it can be calculated in closed-form as $\beta^k = \frac{(\boldsymbol{\delta}^k)^T \nabla G(\mathbf{z}^k)}{(\boldsymbol{\delta}^k)^T \mathbf{B} \boldsymbol{\delta}^k}$, where $\boldsymbol{\delta}^k = \mathbf{z}^{(k+\frac{1}{2})} - \mathbf{z}^{(k)}$ [34]; $Proj(\cdot)$ represents the operation of orthogonal projection that projects the vector onto the nonnegative orthant [8]; $\nabla G(\mathbf{z}^{(k)})$ represents the gradient of $G(\mathbf{z})$ defined in (25) with respect to $\mathbf{z}^{(k)}$. We have $\nabla G(\mathbf{z}^{(k)}) = \mathbf{Bz} + \mathbf{c}$, which can be calculated by

$$\nabla G(\mathbf{z}^{(k)}) = \begin{bmatrix} \boldsymbol{\Phi}^T \boldsymbol{\Phi}(\mathbf{u}^{(k)} - \mathbf{v}^{(k)}) \\ -\boldsymbol{\Phi}^T \boldsymbol{\Phi}(\mathbf{u}^{(k)} - \mathbf{v}^{(k)}) \end{bmatrix} + \begin{bmatrix} -\boldsymbol{\Phi}^T \mathbf{y} \\ \boldsymbol{\Phi}^T \mathbf{y} \end{bmatrix}$$
$$- \rho \mathbf{w}_z^{t-1} + \rho \mathbf{1}^T. \quad (27)$$

In a nutshell, the proposed algorithm computes the following two steps iteratively until convergence:

(a) $\quad \mathbf{w}_z^{t-1} \in \partial \|\mathbf{z}^{t-1}\|_{K,1}$

(b) $\quad \mathbf{z}^t = \underset{\mathbf{z} \succeq 0}{\operatorname{argmin}}\{\frac{1}{2}\mathbf{z}^T \mathbf{B z} + \mathbf{c}^T \mathbf{z}\}$ $\quad (28)$

where $\mathbf{z}$, $\mathbf{B}$ and $\mathbf{c}$ are defined in (25). The step (b) in (28) is calculated by applying the gradient projection descent updates in (26). We summarize this double-loop DC-GPSR algorithm in Algorithm 1.

---

**Algorithm 1** Double-loop DC-GPSR (DlDC-GPSR)

---

**Input:** measurements $\mathbf{y}$, measurement matrix $\boldsymbol{\Phi}$ and a small number $\epsilon$

    **Output:** reconstruction $\hat{\mathbf{x}}$
    **Initialization:** $\mathbf{u}^0, \mathbf{v}^0, \mathbf{z}^0 \leftarrow [(\mathbf{u}^0)^T, (\mathbf{v}^0)^T]^T$

1: **for** $t = 1, 2, \ldots$ **do**
2:     Compute a subgradient $\mathbf{w}_z^{t-1} \in \partial \|\mathbf{z}^{t-1}\|_{K,1}$.
3:     **for** $k = 1, 2, \ldots$ **do**
4:         Compute gradient $\nabla G(\mathbf{z}^{(k)})$ by (27).
5:         Perform gradient projection descent (26) and obtain $\mathbf{z}^{(k+1)}$.
6:         Check convergence, set $\mathbf{z}^* \leftarrow \mathbf{z}^{(k+1)}$ and proceed to Step 7 if convergence is satisfied; otherwise return to Step 3.
7:     **end for**
8:     $\mathbf{z}^t \leftarrow \mathbf{z}^*$
9:     Check terminate condition $\|\mathbf{z}^t - \mathbf{z}^{t-1}\|_2 \leq \epsilon$ and return to Step 1 if not satisfied; otherwise, terminate with $\mathbf{z}^t = [(\mathbf{u}^t)^T, (\mathbf{v}^t)^T]^T$, and obtain the reconstruction $\hat{\mathbf{x}} = \mathbf{u}^t - \mathbf{v}^t$.
10: **end for**

---

[8] Let $\mathbb{R}_+^n = \{\mathbf{x} = (x_1, x_2, ..., x_n) | x_1 \geq 0, x_2 \geq 0, ..., x_n \geq 0\}$ be the nonnegative orthant of $\mathbb{R}^n$.

## V. SINGLE-LOOP DC GRADIENT PROJECTION DESCENT FOR SPARSE RECONSTRUCTION

In the proposed DlDC-GPSR algorithm, at each iteration we solve a nonsmooth convex subproblem (22) using another iterative algorithm, i.e., the gradient projection descent updates. This double-loop procedure can be computationally inefficient. In this section, we derive a closed-form solution to the convex subproblem by performing a special DC decomposition to eliminate the inner iterations. Thus, we propose a basic single-loop DC-GPSR algorithm. Interestingly, we observe that the proposed basic single-loop DC-GPSR algorithm can be interpreted as simple gradient projection descent updates with a required step size. Furthermore, we accelerate the single-loop DC-GPSR algorithm using the BB step size and propose an extension algorithm for the single-loop DC-GPSR.

### A. A Basic Single-Loop DC-GPSR Algorithm

We first rewrite the least squares objective of problem (16) in an equivalent form as

$$\frac{1}{2}\|\mathbf{y} - \boldsymbol{\Phi}\mathbf{x}\|_2^2 = \frac{l}{2}\|\mathbf{x}\|_2^2 - \left(\frac{l}{2}\|\mathbf{x}\|_2^2 - \frac{1}{2}\|\mathbf{y} - \boldsymbol{\Phi}\mathbf{x}\|_2^2\right) (29)$$

where $l \geq 0$ is a Lipschitz constant of the least square objective. By substituting (29) into (16), we write the unconstrained sparse reconstruction problem as

$$\min_{\mathbf{x}} \frac{l}{2}\|\mathbf{x}\|_2^2 - \left(\frac{l}{2}\|\mathbf{x}\|_2^2 - \frac{1}{2}\|\mathbf{y} - \boldsymbol{\Phi}\mathbf{x}\|_2^2\right) + \rho(\|\mathbf{x}\|_1 - \|\mathbf{x}\|_{K,1})$$
$$:= F(\mathbf{x}). \quad (30)$$

Then, we perform the following DC decomposition on the objective $F(\mathbf{x})$ defined in (30), and we have

$$\min_{\mathbf{x}} \underbrace{\frac{l}{2}\|\mathbf{x}\|_2^2 + \rho\|\mathbf{x}\|_1}_{f(\mathbf{x})} - \underbrace{\left(\frac{l}{2}\|\mathbf{x}\|_2^2 - \frac{1}{2}\|\mathbf{y} - \boldsymbol{\Phi}\mathbf{x}\|_2^2 + \rho\|\mathbf{x}\|_{K,1}\right)}_{g(\mathbf{x})}$$
$$(31)$$

where the functions $f(\mathbf{x})$ and $g(\mathbf{x})$ are convex. The convexity of $g(\mathbf{x})$ can be ensured by confirming the convexity of $\frac{l}{2}\|\mathbf{x}\|_2^2 - \frac{1}{2}\|\mathbf{y} - \boldsymbol{\Phi}\mathbf{x}\|_2^2$, which is given in Theorem 2.

*Theorem 2:* The least squares objective $\frac{1}{2}\|\mathbf{y} - \boldsymbol{\Phi}\mathbf{x}\|_2^2$ is smooth and its gradient function is Lipschitz continuous with the Lipschitz constant $l = \lambda_{\max}(\boldsymbol{\Phi}^T \boldsymbol{\Phi})$, where $\lambda_{\max}(\cdot)$ denotes the maximum eigenvalue of a matrix. Thus, the function $h(\mathbf{x}) = \frac{l}{2}\|\mathbf{x}\|_2^2 - \frac{1}{2}\|\mathbf{y} - \boldsymbol{\Phi}\mathbf{x}\|_2^2$ for $l = \lambda_{\max}(\boldsymbol{\Phi}^T \boldsymbol{\Phi})$ is convex. *Proof*: See Appendix B.

Writing the $\ell_2$-square terms in (31) by standard quadratic forms, we have

$$\min_{\mathbf{x}} \quad \underbrace{\frac{l}{2}\mathbf{x}^T \mathbf{x} + \rho\|\mathbf{x}\|_1}_{f(\mathbf{x})}$$
$$- \underbrace{\left(\frac{l}{2}\mathbf{x}^T \mathbf{x} - \frac{1}{2}\mathbf{x}^T \boldsymbol{\Phi}^T \boldsymbol{\Phi}\mathbf{x} + (\boldsymbol{\Phi}^T \mathbf{y})^T \mathbf{x} + \rho\|\mathbf{x}\|_{K,1}\right)}_{g(\mathbf{x})}.$$
$$(32)$$

We split the positive and negative parts of $\mathbf{x}$ by letting $\mathbf{u} = (\mathbf{x})_+, \mathbf{v} = (-\mathbf{x})_+$. By denoting $\mathbf{z} = [\mathbf{u}^T, \mathbf{v}^T]^T$, we express (32) as

$$\min_{\mathbf{z}} \quad \underbrace{\frac{l}{2}\mathbf{z}^T\mathbf{z} + \rho\mathbf{1}^T\mathbf{z}}_{f(\mathbf{z})} - \underbrace{\left(\frac{l}{2}\mathbf{z}^T\mathbf{z} - \frac{1}{2}\mathbf{z}^T\mathbf{B}\mathbf{z} + \mathbf{q}^T\mathbf{z} + \rho\mathbf{1}_K^T\mathbf{z}\right)}_{g(\mathbf{z})}$$

$$\text{s.t.} \quad \mathbf{z} \succeq \mathbf{0} \tag{33}$$

where

$$\mathbf{z} = \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \mathbf{\Phi}^T\mathbf{\Phi} & -\mathbf{\Phi}^T\mathbf{\Phi} \\ -\mathbf{\Phi}^T\mathbf{\Phi} & \mathbf{\Phi}^T\mathbf{\Phi} \end{bmatrix}, \quad \mathbf{q} = \begin{bmatrix} \mathbf{\Phi}^T\mathbf{y} \\ -\mathbf{\Phi}^T\mathbf{y} \end{bmatrix}$$

and $\mathbf{1}_K$ is an indicator vector having either one-valued or zero-valued elements, and the one-valued elements of $\mathbf{1}_K$ indicate the $K$ largest elements of $\mathbf{z}$.

Following the DC algorithm framework to solve the problem (33), we perform the following two steps repeatedly until convergence:

(a) $\quad \partial g(\mathbf{z}^{t-1}) = l\mathbf{z}^{t-1} - \mathbf{B}\mathbf{z}^{t-1} + \mathbf{q} + \rho\mathbf{1}_K^{t-1}$

(b) $\quad \mathbf{z}^t = \underset{\mathbf{z} \geq 0}{\operatorname{argmin}}\{\frac{l}{2}\mathbf{z}^T\mathbf{z} + \rho\mathbf{1}^T\mathbf{z} - \mathbf{z}^T\partial g(\mathbf{z}^{t-1})\}$ (34)

where the superscript $t-1$ and $t$, respectively, indicate the $(t-1)$th and the $t$th update. The subproblem (b) of (34) has a closed-form optimal solution. To derive it, we rewrite the subproblem (b) of (34) as

$$\min_{\mathbf{z}} \quad \frac{l}{2}\mathbf{z}^T\mathbf{z} + \rho\mathbf{z}^T\mathbf{1} - \mathbf{z}^T\partial g(\mathbf{z}^{t-1})$$

$$\text{s.t.} \quad \mathbf{z} \succeq \mathbf{0} \tag{35}$$

where $\partial g(\mathbf{z}^{t-1}) = l\mathbf{z}^{t-1} - \mathbf{B}\mathbf{z}^{t-1} + \mathbf{q} + \rho\mathbf{1}_K^{t-1}$. Next, the problem (35) can be expressed as

$$\min_{\mathbf{z}} \quad \frac{l}{2}\|\mathbf{z} - \frac{1}{l}\left(\partial g(\mathbf{z}^{t-1}) - \rho\mathbf{1}\right)\|_2^2$$

$$\text{s.t.} \quad \mathbf{z} \succeq \mathbf{0}. \tag{36}$$

The objective of problem (36) is to minimize the $\|\mathbf{z} - \frac{1}{l}\left(\partial g(\mathbf{z}^{t-1}) - \rho\mathbf{1}\right)\|^2$ over $\mathbf{z} \succeq \mathbf{0}$ with respect to variable $\mathbf{z}$. The solution is simply the Euclidean projection of $\frac{1}{l}\left(\partial g(\mathbf{z}^{t-1}) - \rho\mathbf{1}\right)$ onto the nonnegative orthant. Thus, the gradient projection descent update has a closed-form optimal solution

$$\mathbf{z}^* = Proj\left(\frac{1}{l}\left(\partial g(\mathbf{z}^{t-1}) - \rho\mathbf{1}\right)\right)$$

$$= \left(\frac{1}{l}\left(\partial g(\mathbf{z}^{t-1}) - \rho\mathbf{1}\right)\right)_+ \tag{37}$$

where the $Proj(\cdot)$ represents the Euclidean projection onto the feasible set $\mathbf{z} \succeq \mathbf{0}$, which is simply the positive-taking operation $(\cdot)_+$. Applying the closed-form solution (37) to the subproblem (b) of (34), the DC programming procedure in (34) is simplified to perform the following two single-step computations repeatedly until convergence:

(a) $\quad \partial g(\mathbf{z}^{t-1}) = l\mathbf{z}^{t-1} - \mathbf{B}\mathbf{z}^{t-1} + \mathbf{q} + \rho\mathbf{1}_K^{t-1}$

(b) $\quad \mathbf{z}^t = \left(\frac{1}{l}(\partial g(\mathbf{z}^{t-1}) - \rho\mathbf{1})\right)_+. \tag{38}$

We summarize the updates (38) as a single-loop DC gradient projection algorithm in Algorithm 2. By avoiding the inner-loop iterations, the computational complexity of Algorithm 2 for each iteration is largely reduced compared with Algorithm 1. However, we comment that if the parameter $l$ is large, Algorithm 2 can be slow to converge. Therefore, we will further propose an accelerated algorithm in the following section.

---

**Algorithm 2** Single-loop DC-GPSR-Basic (SlDC-GPSR-Basic)

---

**Input:** measurements $\mathbf{y}$, measurement matrix $\mathbf{\Phi}$ and a small number $\epsilon$

**Output:** reconstruction $\hat{\mathbf{x}}$

**Initialization:** $\mathbf{u}^0, \mathbf{v}^0, \mathbf{z}^0 \leftarrow [(\mathbf{u}^0)^T, (\mathbf{v}^0)^T]^T$

1: **for** $t = 1, 2, \dots$ **do**
2:      Compute the gradient $\partial g(\mathbf{z}^{t-1}) = l\mathbf{z}^{t-1} - \mathbf{B}\mathbf{z}^{t-1} + \mathbf{q} + \rho\mathbf{1}_K^{t-1}$ as (a) of (38).
3:      Perform the optimal projection operation $\mathbf{z}^t = \left(\frac{1}{l}(\partial g(\mathbf{z}^{t-1}) - \rho\mathbf{1})\right)_+$ as (b) of (38).
4:      Check terminate condition $\|\mathbf{z}^t - \mathbf{z}^{t-1}\|_2 \leq \epsilon$, return to Step 1 if not satisfied; otherwise, terminate with $\mathbf{z}^t = [(\mathbf{u}^t)^T, (\mathbf{v}^t)^T]^T$, and return the reconstruction $\hat{\mathbf{x}} = \mathbf{u}^t - \mathbf{v}^t$.
5: **end for**

---

*B. An Extension Algorithm for Single-loop DC-GPSR Using Monotonic BB Step Size*

A crucial observation on SlDC-GPSR-Basic (Algorithm 2) is that we can interpret it as a simple gradient projection descent method to solve a nonconvex optimization problem with a guarantee of global convergence. This observation can be obtained by substituting $\partial g(\mathbf{z}^{t-1})$ in (a) into (b) of (38) such that Step 2 and Step 3 in Algorithm 2 can be combined to be the $t$th update of $\mathbf{z}^t$

$$\mathbf{z}^t = Proj\left(\mathbf{z}^{t-1} - \frac{1}{l}\left(\mathbf{B}\mathbf{z}^{t-1} - \mathbf{q} - \rho\mathbf{1}_K^{t-1} + \rho\mathbf{1}\right)\right)$$

$$= \left(\mathbf{z}^{t-1} - \frac{1}{l}\left(\mathbf{B}\mathbf{z}^{t-1} - \mathbf{q} - \rho\mathbf{1}_K^{t-1} + \rho\mathbf{1}\right)\right)_+. \tag{39}$$

Thus, we can clearly see that the SlDC-GPSR-Basic is just a gradient projection descent method to solve the following nonconvex problem with the required step size $1/l$ for $l = \lambda_{\max}(\mathbf{\Phi}^T\mathbf{\Phi})$

$$\min_{\mathbf{z}} \quad \frac{1}{2}\mathbf{z}^T\mathbf{B}\mathbf{z} - \mathbf{q}^T\mathbf{z} - \rho\mathbf{1}_K^T\mathbf{z} + \rho\mathbf{1}^T\mathbf{z} := F(\mathbf{z})$$

$$\text{s.t.} \quad \mathbf{z} \succeq \mathbf{0} \tag{40}$$

where $\mathbf{z}$, $\mathbf{B}$ and $\mathbf{q}$ have been defined in (33). The problem (40) is an equivalent form of the sparse reconstruction problem (16), but it uses a nonnegative double-sized variable $\mathbf{z} = [(\mathbf{x})_+^T, (-\mathbf{x})_+^T]^T$ to express the sparse vector $\mathbf{x}$. The gradient projection descent is a mature method for convex optimizations. However, its convergence is unguaranteed for solving a nonconvex problem. Thus, in general, the gradient projection descent method cannot be directly applied to a

nonconvex problem. Interestingly, as shown, the proposed SlDC-GPSR-Basic algorithm can be viewed as the gradient projection descent method applied to nonconvex optimizations. Since this algorithm is derived from DC programming, it carries the same global convergence property as the other DC programming algorithms [31].

Although the fixed step size $1/l$ ensures a global convergence for the SlDC-GPSR-Basic algorithm, this step size can be too small in practice. Thus, we design a variable step size to accelerate the single-loop DC-GPSR algorithm. We denote $\alpha^t$ by a variable step size for the $(t+1)$th update, and extend the update (39) to be a gradient projection descent update for the problem (40) with step size $\alpha^t$, i.e.,

$$
\begin{aligned}
\tilde{\mathbf{z}}^{t+1} &= Proj\left(\mathbf{z}^t - \alpha^t\left(\nabla F(\mathbf{z}^t)\right)\right) \\
&= \left(\mathbf{z}^t - \alpha^t\left(\mathbf{B}\mathbf{z}^t - \mathbf{q} - \rho\mathbf{1}_K^t + \rho\mathbf{1}\right)\right)_+ 
\end{aligned}
\tag{41}
$$

where $\alpha^t$ can be explicitly calculated by the BB step size method by

$$
\alpha^t = \frac{\|\mathbf{z}^t - \mathbf{z}^{t-1}\|^2}{(\mathbf{z}^t - \mathbf{z}^{t-1})^T\left(F(\mathbf{z}^t) - F(\mathbf{z}^{t-1})\right)}.
\tag{42}
$$

To prevent the step size being overly large, we employ a scaler $\beta^{t+1} \in (0, 1]$ to limit the update so that the objective will descend monotonically

$$
\mathbf{z}^{t+1} = \mathbf{z}^t + \beta^t(\tilde{\mathbf{z}}^{t+1} - \mathbf{z}^t).
\tag{43}
$$

We can calculate $\beta^t$ to minimize the objective $F(\mathbf{z}^{t+1})$ by

$$
\beta^t = \frac{(\boldsymbol{\delta}^t)^T \nabla F(\mathbf{z}^t)}{(\boldsymbol{\delta}^t)^T \mathbf{B}\boldsymbol{\delta}^t}
\tag{44}
$$

where $\boldsymbol{\delta}^t = \tilde{\mathbf{z}}^{t+1} - \mathbf{z}^t$. We summarize the updates (41) and (43) in Algorithm 3 and name it SlDC-GPSR-BB, which is an extension of single-loop DC-GPSR algorithm since it extends Algorithm 2 by introducing the BB step size.

---

**Algorithm 3** Single-loop DC-GPSR with monotonic BB step size (SlDC-GPSR-BB)

---

**Input:** measurements $\mathbf{y}$, measurement matrix $\boldsymbol{\Phi}$ and a small number $\epsilon$

**Output:** reconstruction $\hat{\mathbf{x}}$

**Initialization:** $\mathbf{u}^0, \mathbf{v}^0, \mathbf{z}^0 \leftarrow [(\mathbf{u}^0)^T, (\mathbf{v}^0)^T]^T, \alpha^0$

1: **for** $t = 0, 1, 2, \ldots$ **do**
2:     Compute update $\tilde{\mathbf{z}}^{t+1}$ using (41).
3:     Compute step size $\beta^t$ using (44).
4:     Compute update $\mathbf{z}^{t+1}$ using (43).
5:     Compute step size $\alpha^{t+1}$ using (42).
6:     Check terminate condition $\|\mathbf{z}^{t+1} - \mathbf{z}^t\|_2 \leq \epsilon$, return to Step 1 if not satisfied; otherwise, terminate with $\mathbf{z}^{t+1} = [(\mathbf{u}^{t+1})^T, (\mathbf{v}^{t+1})^T]^T$ and return the reconstruction $\hat{\mathbf{x}} = \mathbf{u}^{t+1} - \mathbf{v}^{t+1}$.
7: **end for**

---

## VI. COMPLEXITY ANALYSIS

Table I shows the computational cost of the main computing operations and compares the computational complexities of the proposed algorithms. Common operations for all three algorithms mainly contain matrix-vector multiplications, vector inner products, vector sums, scaler-vector multiplications and the subgradient of the top-$(K, 1)$ norm for a nonnegative vector. Among these operations, the computationally intensive terms have the subgradient computation $\partial\|\mathbf{z}\|_{K,1}$ and the matrix-vector product $\mathbf{B}\mathbf{z}$ whose computational costs are shown in Table I. To analyze the overall time complexities of the proposed algorithms, we consider both the computational costs at each iteration and the convergence rates. The computational costs at each iteration [9] for these algorithms are shown in Table I, while the convergence rates will be illustrated in Section VII.B. By taking into account of the convergence rates, we can conclude that, among the three proposed algorithms, SlDC-GPSR-BB has the lowest time complexity.

## VII. NUMERICAL RESULTS

### A. Experiment Setup and Performance Metrics

We consider a downlink massive MIMO system having half-wavelength spaced ULA at the BS. The number of BS antennas is set as $N_t = 256$. Without loss of generality, the number of UE antennas is set as $N_r = 1$ [12]. We consider the narrowband block-fading channels, and set a channel coherence block as $L_c = 600$ symbols. We randomly generate $1,000$ channel vectors according to the Saleh-Valenzuela channel model described in (1), where the number of paths is set as $N_p = 3$. For each path, the complex channel gain $\alpha_l$ follows a complex Gaussian distribution; the AoA and AoD, i.e., $\theta_{r,l}$ and $\theta_{t,l}$, are uniformly distributed over $[-\pi/2, \pi/2]$. We transform the generated channel vectors into the angular domain using a DFT. To consider power leakage, we set the number of nonzero channel coefficients as $N_s = 16$ by neglecting the small-value beamspace channel coefficients [13]. Unless stated otherwise, we adopt the random Gaussian matrix as the measurement matrix $\mathbf{S}_G \in \mathbb{R}^{L \times N_t}$, and the pilot matrix can be obtained by $\mathbf{P} = \mathbf{S}_G^T \mathbf{U}_t \in \mathbb{C}^{N_t \times L}$. All the simulations were implemented on a desktop computer equipped with a 3.2 GHz Intel Core i7-8700 CPU with 8GB of physical memory.

We adopt the normalized mean squared error (NMSE) to evaluate the channel reconstruction accuracy. The NMSE is defined as $\text{NMSE} = \frac{1}{N}\sum_{i=1}^{N}\frac{\|\mathbf{H}_i - \hat{\mathbf{H}}_i\|_F^2}{\|\mathbf{H}_i\|_F^2}$, where $N$ is the number of channel samples. We use achievable spectral efficiency to evaluate the influences of pilot overhead and SNR penalty caused by imperfect channel estimation on data communication. The achievable spectral efficiency $R/B$ is defined as $R/B = (1 - \alpha)C(\text{SNR}_{\text{eff}})$ [50, eq. (5.195)], where $R$

---

[9]To compare fairly, one iteration is considered from Step 2 to Step 5 for DlDC-GPSR, Step 2 and Step 3 for SlDC-GPSR-Basic and from Step 2 to Step 5 for SlDC-GPSR-BB.

[12]According to the proposed column-wise broadcasting method for vectorizing beamspace channel matrix $\mathbf{H} \in \mathbb{C}^{N_t \times N_r}$, to reconstruct $\mathbf{H}$ we simply reconstruct its $N_r$ columns independently.

[13]Although in practice the number of nonzero channel coefficients can be unknown and uncertain, in the simulation we fix the number of nonzero channel coefficients for fair performance comparisons. We set $N_s = 16$ because our observations revealed that the 16 largest-magnitude channel coefficients can cover most of the energy of a channel vector, i.e., $\|\mathbf{h}_i\|_2^2$ for $1 \leq i \leq N_r$.

TABLE I: Computational costs for operations (left) and computational complexity per iteration for different algorithms (right), $n$ is the dimension of the sparse vector $\mathbf{x}$, and $2n$ is the length of vector $\mathbf{z}$, $m$ is the dimension of the measurements $\mathbf{y}$

| Operation/Algorithm | $\partial\|\mathbf{z}\|_{K,1}$ | $\mathbf{Bz}$ | DlDC-GPSR | SlDC-GPSR-Basic | SlDC-GPSR-BB |
|---|---|---|---|---|---|
| Time complexity | $O(2n\log(2n))$ [10] | $O(mn)$ | $I \times O(mn)$ [11] | $O(mn)$ | $O(mn)$ |



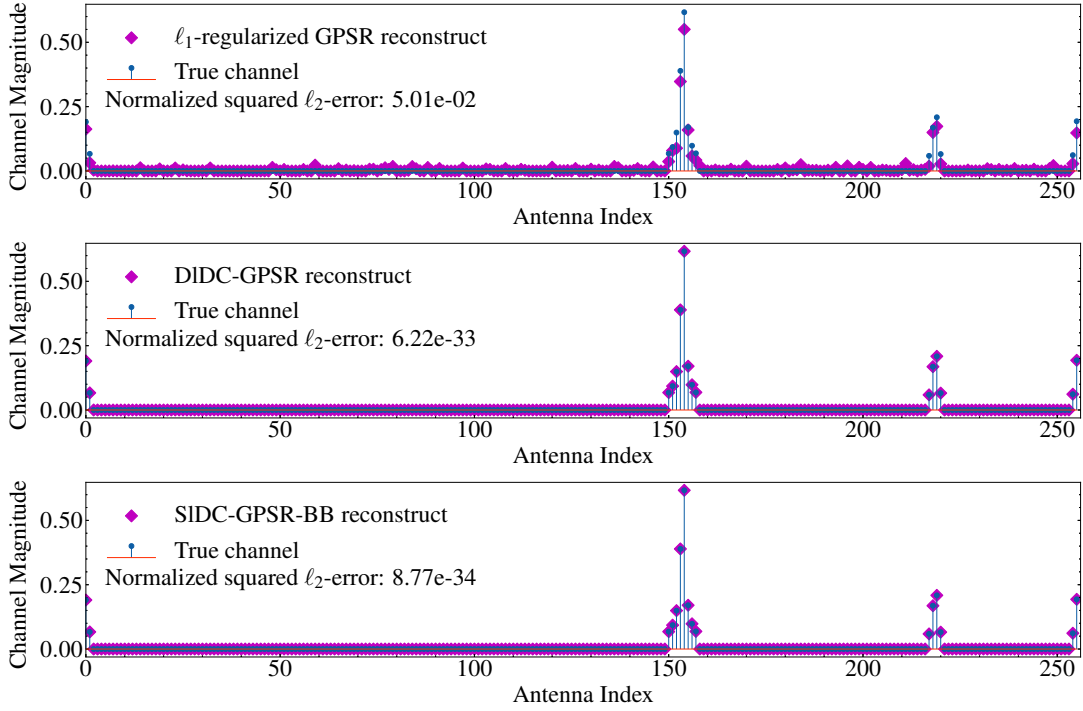Fig. 1: Channel magnitudes of a true channel sample $\mathbf{x}_{\text{opt}}$ and the reconstruction $\hat{\mathbf{x}}$ by the conventional $\ell_1$-regularized GPSR algorithm, the proposed algorithms DlDC-GPSR and SlDC-GPSR-BB. The normalized squared $\ell_2$-error is defined as $\|\mathbf{x}_{\text{opt}} - \hat{\mathbf{x}}\|_2^2/\|\mathbf{x}_{\text{opt}}\|_2^2$

is the achievable rate; $B$ is the bandwidth; $\alpha$ is the ratio of pilot training in a channel coherence block duration, and it can be calculated as $L/L_c$, where $L_c$ is the length of channel coherence block; $C(\text{SNR}_{\text{eff}})$ is the channel capacity at $\text{SNR}_{\text{eff}}$, and $\text{SNR}_{\text{eff}}$ refers to the efficient SNR by considering imperfect channel estimation [50]. According to the linear equation $\mathbf{R} = \mathbf{SH} + \mathbf{W}$ in (6), we define the system SNR as $\text{SNR} = \frac{\|\mathbf{SH}\|_F^2}{\|\mathbf{W}\|_F^2}$, where $\mathbf{S}$ is the measurement matrix, $\mathbf{H}$ is the beamspace channel, and $\mathbf{W}$ is the noise.

*B. Illustrations of Beamspace Channel Reconstructions and Algorithm Convergence Property*

In this subsection, we evaluate the reconstruction accuracies and convergence rates of DlDC-GPSR (Algorithm 1) and SlDC-GPSR-BB (Algorithm 3) [14] by reconstructing an arbitrary channel vector sample $\mathbf{x}$ and compare the performances with conventional $\ell_1$-regularized GPSR algorithm [34]. For fair comparisons, we adopt BB step sizes for the inner loops of DlDC-GPSR. The pilot length is set as $L = 128$. Fig. 1 shows the magnitudes of the true sparse beamspace channel

[14]As we have shown, the SlDC-GPSR-Basic (Algorithm 2) has an interesting theoretical interpretation as a pure gradient projection algorithm, but the required step size $1/l$ is often too small to converge within a reasonable period. Therefore, we do not consider Algorithm 2 in simulations.

and the reconstructions. We can see that the DlDC-GPSR and SlDC-GPSR-BB algorithm can achieve perfect reconstructions with normalized squared $\ell_2$-errors $6.22 \times 10^{-33}$ and $8.77 \times 10^{-34}$, whereas the $\ell_1$-regularized GPSR reconstruction has noticeable errors with the error $5.01 \times 10^{-2}$. By setting the termination condition as $\|\hat{\mathbf{x}}^t - \hat{\mathbf{x}}^{t-1}\|_2 \leq 10^{-30}$, the runtimes are about 0.08, 0.08 and 0.05 seconds for $\ell_1$-regularized GPSR, DlDC-GPSR and SlDC-GPSR-BB, respectively. To show their convergence properties within runtimes, we plot objective values versus iterations in Fig. 2. It should be noted that the objective of the proposed DlDC-GPSR and SlDC-GPSR-BB algorithms is $\frac{1}{2}\|\mathbf{y} - \mathbf{\Phi x}\|_2^2 + \rho(\|\mathbf{x}\|_1 - \|\mathbf{x}\|_{K,1})$, whereas the objective of conventional $\ell_1$-regularized GPSR is $\frac{1}{2}\|\mathbf{y} - \mathbf{\Phi x}\|_2^2 + \rho\|\mathbf{x}\|_1$. For DlDC-GPSR algorithm, the red dots indicate objective values for outer iterations, and the blue solid line shows the objective values along all inner iterations. We can see the proposed DlDC-GPSR algorithm decreases its objective significantly after the sixth outer-step and approaches the optimal value at the eighth step. The SlDC-GPSR-BB algorithm has the fastest convergence and can achieve almost the same objective value as the DlDC-GPSR algorithm. On the contrary, the objective of $\ell_1$-regularized GPSR converges to a relatively large value. Fig. 3 shows the evolution of $\ell_1$-norm penalized least-square objective values, i.e., $\|\mathbf{y} - \mathbf{\Phi x}\|_2^2 + \rho\|\mathbf{x}\|_1$. In Fig. 3, we also plot the optimal
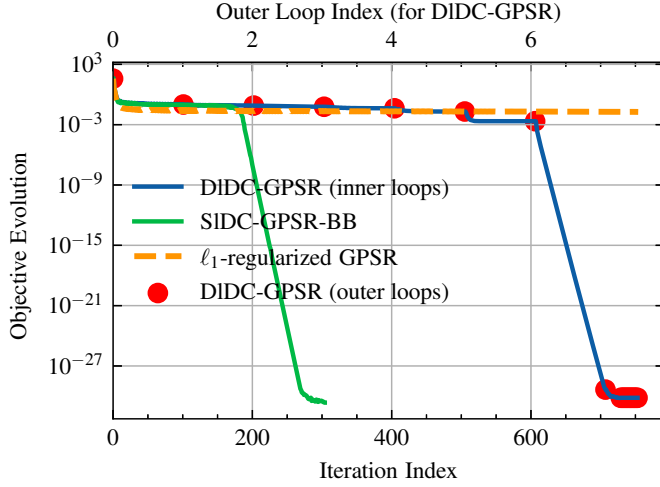
Fig. 2: Objective values versus iterations for reconstructing a sample of beamspace channel vector. The evaluating objectives are $\frac{1}{2}\|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2^2 + \rho(\|\mathbf{x}\|_1 - \|\mathbf{x}\|_{K,1})$ for DlDC-GPSR and SlDC-GPSR-BB, and is $\frac{1}{2}\|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2^2 + \rho\|\mathbf{x}\|_1$ for $\ell_1$-regularized GPSR
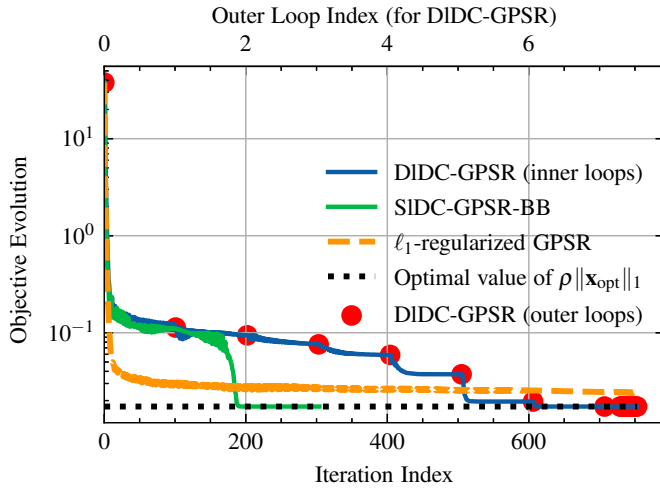


Fig. 3: Objective values ($\frac{1}{2}\|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2^2 + \rho\|\mathbf{x}\|_1$) versus iterations for the reconstruction of a sample of beamspace channel vector

values of the penalty term $\rho\|\mathbf{x}_{\text{opt}}\|_1$, where $\mathbf{x}_{\text{opt}}$ represents the true sample of a beamspace channel vector. We can see that the DlDC-GPSR and SlDC-GPSR-BB algorithms can arrive and stay at the objective value $\rho\|\mathbf{x}_{\text{opt}}\|_1$, which is the optimal value that the objective $\|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2^2 + \rho\|\mathbf{x}\|_1$ can reach when $\|\mathbf{y} - \mathbf{\Phi}\hat{\mathbf{x}}\|_2^2 = 0$ and $\rho\|\hat{\mathbf{x}}\|_1 = \rho\|\mathbf{x}_{\text{opt}}\|_1$. However, the $\ell_1$-regularized GPSR cannot achieve this optimal value and has a noticeable gap from the optimal value. It is meaningful to observe this minimum objective gap between our proposed algorithms and the conventional $\ell_1$-regularized GPSR algorithm, because this gap can provide important insight into the approximation error introduced by relaxing the $\ell_0$-norm as $\ell_1$-norm. Fig. 4 shows the normalized squared-$\ell_2$ errors of reconstruction versus the number of iterations.
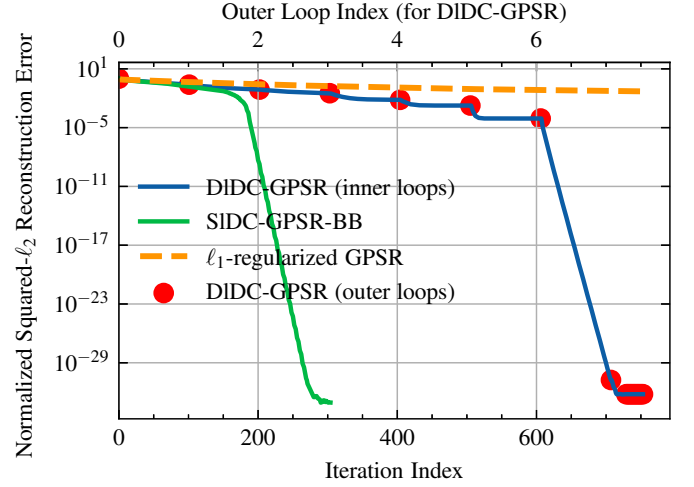


Fig. 4: Reconstruction error ($\|\hat{\mathbf{x}} - \mathbf{x}_{\text{opt}}\|_2^2 / \|\mathbf{x}_{\text{opt}}\|_2^2$) versus iterations for the reconstruction of a sample of beamspace channel vector

We observe that both the DlDC-GPSR and SlDC-GPSR-BB algorithm can achieve accurate reconstructions having errors on the order of $10^{-33}$ and $10^{-34}$, which is far more accurate than the reconstruction by the conventional $\ell_1$-regularized GPSR algorithm having an error on the order of $10^{-2}$.

*C. Performance Comparisons With Other Algorithms*

We compare the reconstruction performances of the proposed algorithms with several existing sparse reconstruction algorithms including the $\ell_1$-regularized GPSR [34], ISTA [36], and OMP [41], as well as the existing state-the-art channel estimation schemes including the SD [15] [25] and the SCAPMI algorithms [26].
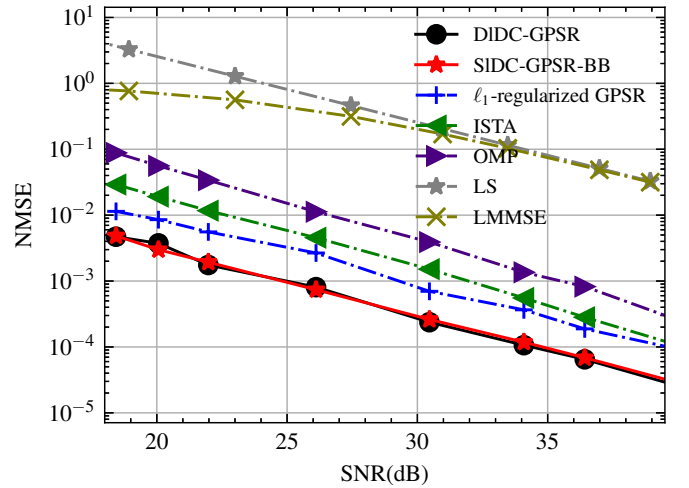


Fig. 5: Channel reconstruction NMSE versus SNR (dB) with training pilot length $L = 128$
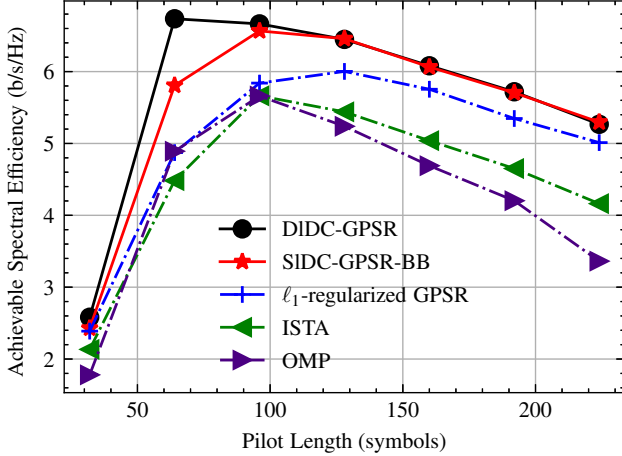
In Fig. 5, we compare the channel reconstruction NMSE with conventional channel estimation methods including the

[15]The SD algorithm is the support detection-based channel estimation algorithm [25].
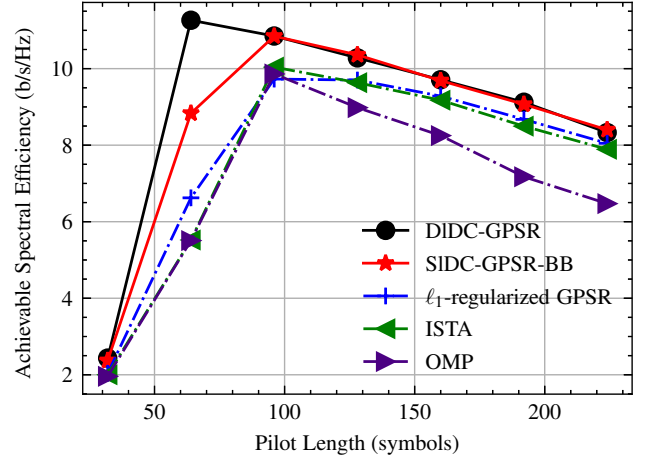
TABLE II: Comparison of the average runtimes per channel sample reconstruction

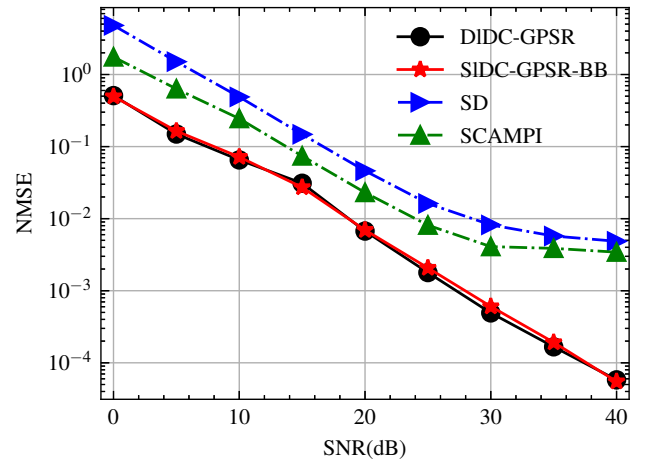| Algorithm | SlDC-GPSR-BB | DlDC-GPSR | $\ell_1$-regularized GPSR | ISTA | OMP |
|---|---|---|---|---|---|
| SNR = 40 dB | 0.03s | 0.67s | 0.12s | 0.61s | 0.16s |
| SNR = 30 dB | 0.07s | 1.61s | 0.19s | 0.86s | 0.21s |
| SNR = 18 dB | 0.16s | 2.39s | 0.26s | 1.89s | 0.31s |

LS and LMMSE, and popular sparse reconstruction algorithms including the OMP, ISTA and $\ell_1$-regularized GPSR. For conventional LS and LMMSE channel estimation methods, we use the optimal pilot matrix that contains orthonormal training pilot sequences having the length of $L = 256$. For the sparse reconstruction algorithms including DlDC-GPSR, SlDC-GPSR-BB, OMP, ISTA and $\ell_1$-regularized GPSR, we set the training pilot length as $L = 128$. Fig. 5 shows that the proposed DlDC-GPSR and SlDC-GPSR-BB algorithms achieve approximately the same accuracy. Even consuming more training pilots, both LS and LMMSE channel estimation methods have higher channel reconstruction errors than the sparse reconstruction methods. Also, the proposed algorithms DlDC-GPSR and SlDC-GPSR-BB have higher accuracies than the other sparse reconstruction algorithms. The average runtimes required for reconstructing a channel sample are summarized in Table II for different sparse reconstruction algorithms.



Fig. 7: Achievable spectral efficiency $R/B$ versus training pilot length $L$ at SNR $= 40$ dB

decreased by increasing the number of training symbols, such that the efficient SNR (SNR$_{\text{eff}}$) increases and dominates the achievable spectral efficiency. However, when $L$ is sufficiently large, a further increase of $L$ will give a diminishing benefit while the data communication ratio $(1 - \alpha)$ decreases linearly, where $\alpha = L/L_c$ is the ratio of pilot training overhead in a coherence block of fading channels. Fig. 6 and Fig. 7 show that the proposed DlDC-GPSR and SlDC-GPSR-BB algorithms can attain higher achievable spectral efficiencies than the $\ell_1$-regularized GPSR, ISTA and OMP sparse reconstruction algorithms.



Fig. 6: Achievable spectral efficiency $R/B$ versus training pilot length $L$ at SNR $= 25$ dB

For sparse channel reconstructions, increasing the pilot length can improve reconstruction accuracy. However, longer pilot length will also increase the training overhead and subsequently degrade the spectral efficiency. In other words, there exists a tradeoff between the training pilot overhead and spectral efficiency. To illustrate this tradeoff and comparison for different sparse reconstruction algorithms, we plot the achievable spectral efficiency versus training pilot length $L$ for SNR $= 25$ dB and SNR $= 40$ dB in Fig. 6 and Fig. 7, where the length of a channel coherence block is set as $L_c = 600$ symbols, and the training pilot length is set as $L \in \{32, 64, ..., 256\}$. We can see that the achievable spectral efficiencies increase when the training pilot length $L$ increases, and then decrease after reaching the peak values. Since when $L$ has a small value, the channel estimation error can be



Fig. 8: NMSE comparison between the SCAMPI algorithm under uniform distribution with a lens antenna array in the size of $16 \times 16$, the SD algorithm and the proposed SlDC-GPSR-BB and DlDC-GPSR algorithms with a ULA in the size of $256$ while using a training pilot length of $L = 96$
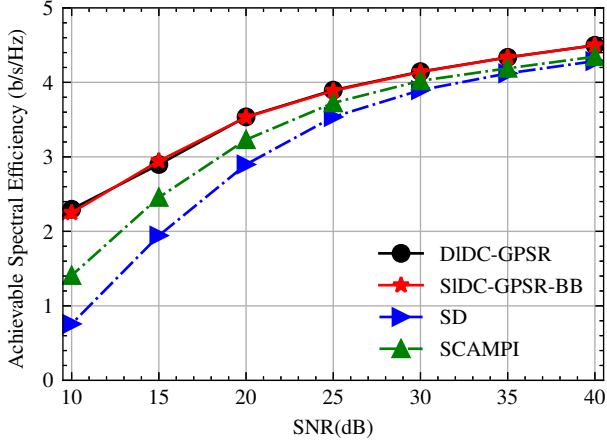
Fig. 9: Comparison of the achievable spectral efficiency $R/B$ between the SCAMPI algorithm under uniform distribution with a lens antenna array in the size of $16 \times 16$, the SD algorithm and the proposed SlDC-GPSR-BB and DlDC-GPSR algorithms with a ULA in the size of 256 while using a training pilot length of $L = 96$

We now compare the performances of the proposed algorithms with the state-of-the-art beamspace channel estimation methods including the SD and the SCAMPI algorithms. For fair comparisons, we selected a $16 \times 16$ lens antenna array size for SCAMPI, and chose $N_t = 256$ for the ULA antenna number for both the SD and the proposed algorithms. The training pilot length is set as $L = 96$. To allow the proposed algorithms to be contrasted fairly with the SD and SCAMPI algorithms, we constructed the same type of measurement matrix used in [25] and [26]. The measurement matrix is a Rademacher matrix, which contains entries drawn from $\{+1, -1\}$ with equal probabilities. Fig. 8 shows the results of channel reconstruction NMSE versus SNR. At high SNR, the SD and SCAMPI algorithms appear to reach an error floor of approximately $10^{-2}$, while the proposed SlDC-GPSR-BB and DlDC-GPSR algorithms achieve lower reconstruction errors. Fig. 9 compares the achievable spectral efficiencies by different channel estimation schemes. The proposed SlDC-GPSR-BB and DlDC-GPSR algorithms achieve higher spectral efficiency than the SD and SCAMPI algorithms.

## VIII. CONCLUSION

We proposed three DC programming gradient projection sparse reconstruction algorithms for massive MIMO beamspace channel estimation, and they are DlDC-GPSR, SlDC-GPSR-Basic and SlDC-GPSR-BB. We derived these algorithms by solving a least squares problem having a nonconvex regularizer. The regularizer is simply the difference between an $\ell_1$-norm and a top-$(K, 1)$ norm, which was introduced to remove the penalties on $K$ largest-magnitude elements of the reconstructing vector. We employed DC programming and a gradient projection method to solve the nonconvex sparse reconstruction, and derived different double-loop and single-loop algorithms by different DC decompositions. The double-loop DlDC-GPSR algorithm is simple, accurate, and

robust. The SlDC-GPSR-Basic algorithm has simple single-loop updates, and it shares the same theoretical convergence property as the other DC programming methods. We observed that the SlDC-GPSR-Basic updates can be perfectly interpreted as a simple gradient projection descent update. Based on this observation, we adopted the BB step-size strategy and proposed an extension algorithm SlDC-GPSR-BB, which has a significantly improved convergence rate and can achieve approximately the same level accuracy as the double-loop algorithm DlDC-GPSR. Several numerical demonstrations were provided to show the advantages of the proposed algorithms, such as high accuracies and fast computations.

## IX. ACKNOWLEDGEMENT

## APPENDIX

### A. Proof of Theorem 1:

Suppose that $\mathbf{x}_{\rho^*}$ is an optimal solution to (16) with given $\rho^*$, then $\mathbf{x}_{\rho^*}$ is also optimal to (15) as long as $\|\mathbf{x}_{\rho^*}\|_0 \le K$ (or $\|\mathbf{x}_{\rho^*}\|_1 - \|\mathbf{x}_{\rho^*}\|_{K,1} = 0$) satisfies. Thus, we only need to prove that $\|\mathbf{x}_{\rho^*}\|_0 > K$ is infeasible. We assume that $\mathbf{x}_{\rho^*}$ is an optimal solution to (16) with $\rho^* > \max_i\{|(\boldsymbol{\Phi}^T\mathbf{y})_i| + q(\|\boldsymbol{\Phi}^T\boldsymbol{\Phi}\mathbf{e}_i\|_2 + |(\boldsymbol{\Phi}^T\boldsymbol{\Phi})_{ii}|/2)\}$. For $\|\mathbf{x}_{\rho^*}\|_0 > K$, we construct a feasible solution to (15) as $\mathbf{x}' = \mathbf{x}_{\rho^*} - x_i\mathbf{e}_i$, where $i$ represents the index of the $(K+1)$th largest-magnitude element of vector $\mathbf{x}_{\rho^*}$; $\mathbf{e}_i$ represents the unit vector in which the $i$th element is one while the other elements are zeros; $x_i$ represents the $i$th element of vector $\mathbf{x}_{\rho^*}$. By writing the objective of (16) as

$$F(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\mathbf{x} - (\boldsymbol{\Phi}^T\mathbf{y})^T\mathbf{x} + \rho^*\|\mathbf{x}\|_1 - \rho^*\|\mathbf{x}\|_{K,1}$$
(45)

we have

$$\begin{aligned}
&F(\mathbf{x}_{\rho^*}) - F(\mathbf{x}') \\
&= F(\mathbf{x}_{\rho^*}) - F(\mathbf{x}_{\rho^*} - x_i\mathbf{e}_i) \\
&= \frac{1}{2}\mathbf{x}_{\rho^*}^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\mathbf{x}_{\rho^*} - (\boldsymbol{\Phi}^T\mathbf{y})^T\mathbf{x}_{\rho^*} \\
&\quad -\frac{1}{2}(\mathbf{x}_{\rho^*} - x_i\mathbf{e}_i)^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}(\mathbf{x}_{\rho^*} - x_i\mathbf{e}_i) \\
&\quad + (\boldsymbol{\Phi}^T\mathbf{y})^T(\mathbf{x}_{\rho^*} - x_i\mathbf{e}_i) + \rho^*|x_i| \\
&= x_i\mathbf{e}_i^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\mathbf{x}_{\rho^*} - \frac{1}{2}x_i^2(\boldsymbol{\Phi}^T\boldsymbol{\Phi})_{i,i} - x_i(\boldsymbol{\Phi}^T\mathbf{y})_i + \rho^*|x_i| \\
&\ge -x_i\mathbf{e}_i^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\mathbf{x}_{\rho^*} - \frac{1}{2}x_i^2(\boldsymbol{\Phi}^T\boldsymbol{\Phi})_{i,i} - x_i(\boldsymbol{\Phi}^T\mathbf{y})_i + \rho^*|x_i| \\
&\ge -|x_i|\|\mathbf{e}_i^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\|_2\|\mathbf{x}_{\rho^*}\|_2 - \frac{1}{2}|x_i|\|\mathbf{x}_{\rho^*}\|_2(\boldsymbol{\Phi}^T\boldsymbol{\Phi})_{i,i} \\
&\quad -|x_i| \cdot |(\boldsymbol{\Phi}^T\mathbf{y})_i| + \rho^*|x_i| \\
&= |x_i|\big(\rho - \|\mathbf{e}_i^T\boldsymbol{\Phi}^T\boldsymbol{\Phi}\|_2\|\mathbf{x}_{\rho^*}\|_2 - \frac{1}{2}\|\mathbf{x}_{\rho^*}\|_2(\boldsymbol{\Phi}^T\boldsymbol{\Phi})_{i,i} \\
&\quad -|(\boldsymbol{\Phi}^T\mathbf{y})_i|\big) \\
&> 0
\end{aligned}$$
(46)

which contradicts the assumption that $\mathbf{x}_{\rho^*}$ is an optimal solution to (16).

## B. Proof of Theorem 2:

We first derive the Lipschitz constant $l = \lambda_{\max}(\mathbf{\Phi}^T\mathbf{\Phi})$, then we prove the convexity of $h(\mathbf{x})$. We denote the least squares objective by the function $l(\mathbf{x})$. Since the least squares objective $w(\mathbf{x}) = \frac{1}{2}\|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2^2$ is smooth if and only if its gradient function is Lipschitz continuous, we assume there exists $l < \infty$, which is named as a Lipschitz constant, such that

$$\|\nabla w(\mathbf{x}) - \nabla w(\mathbf{z})\|_2 \leq l\|\mathbf{x} - \mathbf{z}\|_2. \tag{47}$$

For $w(\mathbf{x}) = \frac{1}{2}\|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2^2$, we have

$$\begin{aligned}
\|\nabla w(\mathbf{x}) - \nabla w(\mathbf{z})\|_2 &= \|\mathbf{\Phi}^T(\mathbf{\Phi}\mathbf{x} - \mathbf{y}) - \mathbf{\Phi}^T(\mathbf{\Phi}\mathbf{z} - \mathbf{y})\|_2 \\
&= \|\mathbf{\Phi}^T\mathbf{\Phi}(\mathbf{x} - \mathbf{z})\|_2 \\
&\leq \|\|\mathbf{\Phi}^T\mathbf{\Phi}\|\|_2 \|\mathbf{x} - \mathbf{z}\|_2 \\
&= \lambda_{max}(\mathbf{\Phi}^T\mathbf{\Phi})\|\mathbf{x} - \mathbf{z}\|_2
\end{aligned} \tag{48}$$

where $\|\|\cdot\|\|_2$ represents the spectral norm of a matrix, and $\lambda_{max}(\cdot)$ represents the largest eigenvalue of a matrix. Thus, we obtain the Lipschitz constant $l = \lambda_{\max}(\mathbf{\Phi}^T\mathbf{\Phi})$.

For $h(\mathbf{x}) = \frac{l}{2}\|\mathbf{x}\|_2^2 - \frac{1}{2}\|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2^2$, we have the Hessian matrix $\nabla^2 h(\mathbf{x})$ as

$$\begin{aligned}
\nabla^2 h(\mathbf{x}) &= \frac{\partial(\nabla h(\mathbf{x}))}{\partial\mathbf{x}} \\
&= \frac{\partial(l\mathbf{x}^T - \mathbf{\Phi}^T(\mathbf{\Phi}\mathbf{x} - \mathbf{y}))}{\partial\mathbf{x}} \\
&= l\mathbf{I} - \mathbf{\Phi}^T\mathbf{\Phi}
\end{aligned} \tag{49}$$

where $\mathbf{I}$ is the identity matrix. For $l = \lambda_{\max}(\mathbf{\Phi}^T\mathbf{\Phi})$, we have $l\mathbf{I} - \mathbf{\Phi}^T\mathbf{\Phi} \succeq 0$, which means that the Hessian matrix $l\mathbf{I} - \mathbf{\Phi}^T\mathbf{\Phi}$ of $h(\mathbf{x})$ is semidefinite positive. Thus, $h(\mathbf{x})$ is a convex function of vector $\mathbf{x}$.

## REFERENCES

[1] W. U. Bajwa, J. Haupt, A. M. Sayeed, and R. Nowak, "Compressed channel sensing: A new approach to estimating sparse multipath channels," *Proc. IEEE*, vol. 98, no. 6, pp. 1058–1076, June 2010.

[2] S. F. Cotter and B. D. Rao, "Sparse channel estimation via matching pursuit with application to equalization," *IEEE Trans. Commun.*, vol. 50, no. 3, pp. 374–377, Mar. 2002.

[3] M. R. Raghavendra and K. Giridhar, "Improving channel estimation in OFDM systems for sparse multipath channels," *IEEE Signal Process. Lett.*, vol. 12, no. 1, pp. 52–55, Jan. 2005.

[4] J. L. Paredes, G. R. Arce, and Z. Wang, "Ultra-wideband compressed sensing: Channel estimation," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 3, pp. 383–395, Otc. 2007.

[5] G. Tauböck, F. Hlawatsch, D. Eiwen, and H. Rauhut, "Compressive estimation of doubly selective channels in multicarrier systems: Leakage effects and sparsity-enhancing processing," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 2, pp. 255–271, Apr. 2020.

[6] C. R. Berger, S. Zhou, J. C. Preisig, and P. Willett, "Sparse channel estimation for multicarrier underwater acoustic communication: From subspace methods to compressed sensing," *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1708–1721, Mar. 2010.

[7] J. Brady, N. Behdad, and A. M. Sayeed, "Beamspace MIMO for millimeter-wave communications: System architecture, modeling, analysis, and measurements," *IEEE Trans. Antennas Propag.*, vol. 61, no. 7, pp. 3814–3827, July 2013.

[8] Z. Gao, L. Dai, Z. Wang, and S. Chen, "Spatially common sparsity based adaptive channel estimation and feedback for FDD massive MIMO," *IEEE Trans. Signal Process.*, vol. 63, no. 23, pp. 6169–6183, Dec. 2015.

[9] Z. Gao, L. Dai, W. Dai, B. Shim, and Z. Wang, "Structured compressive sensing-based spatio-temporal joint channel estimation for FDD massive MIMO," *IEEE Trans. Commun.*, vol. 64, no. 2, pp. 601–617, Feb. 2016.

[10] C. R. Berger, Z. Wang, J. Huang, and S. Zhou, "Application of compressive sensing to sparse channel estimation," *IEEE Commun. Mag.*, vol. 48, no. 11, pp. 164–174, Nov. 2010.

[11] M. E. Eltayeb, T. Y. Al-Naffouri, and H. R. Bahrami, "Compressive sensing for feedback reduction in MIMO broadcast channels," *IEEE Trans. Commun.*, vol. 62, no. 9, pp. 3209–3222, Sept. 2014.

[12] J. W. Choi, B. Shim, and S. Chang, "Downlink pilot reduction for massive MIMO systems via compressed sensing," *IEEE Commun. Lett.*, vol. 19, no. 11, pp. 1889–1892, Nov. 2015.

[13] C.-C. Tseng, J.-Y. Wu, and T.-S. Lee, "Enhanced compressive downlink CSI recovery for FDD massive MIMO systems using weighted block $l_1$-minimization," *IEEE Trans. Commun.*, vol. 64, no. 3, pp. 1055–1067, Mar. 2016.

[14] H. Almosa, S. Mosleh, E. Perrins, and L. Liu, "Downlink channel estimation with limited feedback for FDD multi-user massive MIMO with spatial channel correlation," in *Proc. IEEE ICC*, Kansas City, MO, May 2018, pp. 1–6.

[15] W. Shen, L. Dai, Y. Shi, B. Shim, and Z. Wang, "Joint channel training and feedback for FDD massive MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 65, no. 10, pp. 8762–8767, Otc. 2016.

[16] X. Rao and V. K. N. Lau, "Distributed compressive CSIT estimation and feedback for FDD multi-user massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 62, no. 12, pp. 3261–3271, June 2014.

[17] A. Liu, F. Zhu, and V. K. N. Lau, "Closed-loop autonomous pilot and compressive CSIT feedback resource adaptation in multi-user FDD massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 65, no. 1, pp. 173–183, Jan. 2017.

[18] X. Gao, L. Dai, S. Han, C. I, and X. Wang, "Reliable beamspace channel estimation for millimeter-wave massive MIMO systems with lens antenna array," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 6010–6021, Sept. 2017.

[19] Z. Gao, C. Hu, L. Dai, and Z. Wang, "Channel estimation for millimeter-wave massive MIMO with hybrid precoding over frequency-selective fading channels," *IEEE Commun. Lett.*, vol. 20, no. 6, pp. 1259–1262, June 2016.

[20] K. Venugopal, A. Alkhateeb, N. G. Prelcic, and R. W. Heath Jr., "Channel estimation for hybrid architecture-based wideband millimeter wave systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 1996–2009, Sept. 2017.

[21] J. Lee and G.-T. Gil, "Exploiting spatial sparsity for estimating channels of hybrid MIMO systems in millimeter wave communications," in *IEEE Globecom*, Austin, TX USA, Dec. 2014.

[22] J. Rodríguez-Fernández, N. González-Prelcic, K. Venugopal, and R. W. Heath Jr., "Frequency-domain compressive channel estimation for frequency-selective hybrid millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 5, pp. 2946–2960, May 2018.

[23] J. P. González-Coma, J. Rodríguez-Fernández, N. González-Prelcic, L. Castedo, and R. W. Heath Jr., "Channel estimation and hybrid precoding for frequency selective multiuser mmWave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 2, pp. 353–367, May 2018.

[24] P. S. Jianhua Mo and R. W. Heath Jr., "Channel estimation in broadband millimeter wave MIMO systems with few-bit ADCs," *IEEE Trans. Signal Process.*, vol. 66, no. 5, pp. 1141–1154, Mar. 2018.

[25] L. Dai, X. Gao, S. Han, C.-L. I, and X. Wang, "Beamspace channel estimation for millimeter-wave massive MIMO systems with lens antenna array," in *IEEE Int. Conf. Commun. China (ICCC)*, Chengdu, China, July 2016.

[26] J. Yang, C.-K. Wen, S. Jin, and F. Gao, "Beamspace channel estimation in mmWave systems via cosparse image reconstruction technique," *IEEE Trans. Commun.*, vol. 66, no. 10, pp. 4767–4782, Otc. 2018.

[27] A. Alkhateeb, O. E. Ayach, G. Leus, and R. W. Heath Jr., "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 831–846, Otc. 2014.

[28] Y. Wang, Y. Zhang, Z. Tian, G. Leus, and G. Zhang, "Super-resolution channel estimation for arbitrary arrays in hybrid millimeter-wave massive MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 5, pp. 947–959, Sept. 2019.

[29] W. Zhang, T. Kim, and S.-H. Leung, "A sequential subspace method for millimeter wave MIMO channel estimation," *IEEE Trans. Veh. Technol.*, vol. 69, no. 5, pp. 5355–5368, May 2020.

[30] S. Buzzi and C. D'Andrea, "Subspace tracking algorithms for millimeter wave MIMO channel estimation with hybrid beamforming," in *WSA*, Berlin, Germany, Mar. 2017.

[31] H. A. L. Thi and T. P. Dinh, "DC programming and DCA: Thirty years of developments," *Math. Program.*, vol. 169, pp. 5–68, 2018.

[32] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "An interior-point method for large-scale $\ell_1$-regularized least squares," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 4, pp. 606–617, Dec. 2007.

[33] W. Yin, S. Osher, D. Goldfarb, and J. Darbon, "Bregman iterative algorithms for $\ell_1$-minimization with applications to compressed sensing," *SIAM Journal on Imaging Sciences*, vol. 1, no. 1, pp. 143–168, Mar. 2008.

[34] M. A. T. Figueiredo, R. D. Nowak, and S. J. Wright, "Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 4, pp. 586–597, Dec 2007.

[35] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo, "Sparse reconstruction by separable approximation," *IEEE Trans. Signal Process.*, vol. 57, no. 7, pp. 2479–2493, July 2009.

[36] T. Blumensath and M. E. Davies, "Iterative thresholding for sparse approximations," *Journal of Fourier Analysis and Applications*, vol. 14, pp. 629–654, 2008.

[37] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM J. Imag. Sci.*, vol. 2, no. 1, pp. 183–202, 2009.

[38] S. Foucart and M.-J. Lai, "Sparsest solutions of underdetermined linear systems via $\ell_q$-minimization for $0 \leq q \leq 1$," *Appl. Comput. Harmonic Anal.*, vol. 26, no. 3, pp. 395–407, 2009.

[39] G. Gasso, A. Rakotomamonjy, and S. Canu, "Recovering sparse signals with a certain family of nonconvex penalties and DC programming," *IEEE Trans. Signal Process.*, vol. 57, no. 12, pp. 4686–4698, Dec. 2009.

[40] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," in *Conf. Rec. 27th Asilomar Conf. Signals, Syst. Comput.*, vol. 1, 1993, pp. 40–44.

[41] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inform. Theory*, vol. 53, no. 12, pp. 4655–4666, Dec. 2007.

[42] D. Needell and J. A. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Applied and Computational Harmonic Analysis*, vol. 26, no. 3, pp. 301–321, 2008.

[43] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[44] J. Gotoh, A. Takeda, and K. Tono, "DC formulations and algorithms for sparse optimization problems," *Mathematical Programming*, vol. 169, no. 1, pp. 141–176, 2018.

[45] D. Tse and V. Pramod, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[46] I. Rish and G. Y. Grabarnik, *Sparse Modeling: Theory, Algorithms, and Applications*. Boca Raton, Florida: Chapman & Hall/CRC Press, 2015.

[47] P. D. Tao and L. T. H. An, "Convex analysis approach to DC programming: Theory, algorithms and applications," *Acta Mathematica Vietnamica*, vol. 22, no. 1, pp. 289–355, 1997.

[48] B. K. Sriperumbudur and G. R. Lanckriet, "A proof of convergence of the concave-convex procedure using Zangwill's theory," *Neural Computation*, vol. 24, no. 6, pp. 1391–1407, 2012.

[49] T. P. Dinh and H. A. L. Thi, "Recent advances in DC programming and DCA," *Transactions on Computational Intelligence XIII*, vol. 8432, pp. 1–37, 2014.

[50] R. W. Heath Jr. and A. Lozano, *Foundations of MIMO Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2018.

[51] P. Wu, H. Ma, and J. Cheng, "Sparse channel reconstruction with nonconvex regularizer via DC programming for massive MIMO systems," in *IEEE Globecom*, Taipei, Taiwan, Dec. 2020.