Block-Sparse Recovery with Optimal Block Partition

Hiroki Kuroda, *Member, IEEE*, and Daichi Kitahara, *Member, IEEE*

Abstract—This paper presents a convex recovery method for block-sparse signals whose block partitions are unknown a priori. We first introduce a nonconvex penalty function, where the block partition is adapted for the signal of interest by minimizing the mixed ℓ_2/ℓ_1 norm over all possible block partitions. Then, by exploiting a variational representation of the ℓ_2 norm, we derive the proposed penalty function as a suitable convex relaxation of the nonconvex one. For a block-sparse recovery model designed with the proposed penalty, we develop an iterative algorithm which is guaranteed to converge to a globally optimal solution. Numerical experiments demonstrate the effectiveness of the proposed method.

Index Terms—Block-sparsity, unknown partition, penalty function, convex optimization, proximal splitting algorithm.

I. INTRODUCTION

BLOCK-SPARSITY is a special type of sparsity where nonzero components are clustered in blocks. For recovery of a block-sparse signal whose block partition is known a priori, extensive researches, e.g., [1]–[10], show the effectiveness of the mixed ℓ_2/ℓ_1 norm using the known block partition. However, in fact, the information on the block partition is not available in many applications, e.g., acoustic signal recovery [11]–[13], image restoration [14]–[16], change detection [17]–[19], and radar signal processing [20]–[23]. For instance, the target signal of the phased array weather radar [21]–[25] is block-sparse in the Fourier domain due to the narrow bandwidth of the power spectrum, but the block partition is unknown because it depends on the unknown mean and standard deviation of the Doppler frequency. In such situations, the estimation accuracy of the mixed ℓ_2/ℓ_1 norm often degrades due to a pre-fixed block partition used instead of the ideal one.

Motivated by these observations, we consider the estimation problem of $\mathbf{x}^* \in \mathbb{C}^N$ which is block-sparse across unknown non-overlapping blocks $\mathcal{B}_1^*, \dots, \mathcal{B}_{K^*}^*$ (see Fig. 1 for an illustration). More precisely, we suppose that the subvector $\mathbf{x}_{\mathcal{B}_k^*}^* := (x_n^*)_{n \in \mathcal{B}_k^*}$ contains only (nearly) zero components for many k , i.e.,

$$\mathbf{x}_{\mathcal{B}_k^*}^* \approx \mathbf{0} \text{ for many } k \in \{1, \dots, K^*\}.$$

We use the term *block-sparse* in a strict sense, i.e., $\mathcal{B}_k^*$ consists of only consecutive indices for each $k = 1, \dots, K^*$:

$$\mathcal{B}_k^* = \{n \in \{1, \dots, N\} \mid n_k^* \leq n \leq m_k^*\},$$

This work was supported in part by the Japan Society for the Promotion of Science under Grants-in-Aid 21K17827, 21H01592, and 19K20361. (Corresponding author: Hiroki Kuroda.)

The authors are with the College of Information Science and Engineering, Ritsumeikan University, Shiga 525-8577, Japan (e-mail: kuroda@media.ritsumeik.ac.jp; d-kita@media.ritsumeik.ac.jp).

with some $n_k^*, m_k^* \in \{1, \dots, N\}$. We also suppose that $\mathcal{B}_1^*, \dots, \mathcal{B}_{K^*}^*$ form a partition of $\{1, \dots, N\}$: $\bigcup_{k=1}^{K^*} \mathcal{B}_k^* = \{1, \dots, N\}$, $\mathcal{B}_k^* \neq \emptyset$ ($k = 1, \dots, K^*$), and $\mathcal{B}_k^* \cap \mathcal{B}_{k'}^* = \emptyset$ ($k \neq k'$). Note that the non-overlapping condition $\mathcal{B}_k^* \cap \mathcal{B}_{k'}^* = \emptyset$ is not restrictive since the sizes of $\mathcal{B}_1^*, \dots, \mathcal{B}_{K^*}^*$ can be different.

Several attempts have been made to deal with the unknown blocks $\mathcal{B}_1^*, \dots, \mathcal{B}_{K^*}^*$. A commonly used modification is to use overlapping blocks in the mixed ℓ_2/ℓ_1 norm, e.g., [26]–[33]. Among them, the latent group lasso (LGL) method [32], [33] presents a sophisticated approach to select relevant blocks from pre-defined overlapping blocks. The LGL penalty is defined as the minimum of a convex function, and thus the corresponding regularization model can be solved as a convex optimization problem (see Appendix A for detail). However, since the problem size grows with the number of candidate blocks, the LGL method has a restriction on the number of candidate blocks, which leads to the degradation of the estimation accuracy. Meanwhile, the greedy methods [34], [35] have similar restrictions on the number of candidate blocks due to the computational complexity. Bayesian approaches [36], [37] have also been presented, but they have to solve challenging nonconvex optimization problems. Thus, it is highly desired to realize a convex method which can cope with the unknown blocks $\mathcal{B}_1^*, \dots, \mathcal{B}_{K^*}^*$ whose sizes are different.

In this paper, we propose a convex recovery method for $\mathbf{x}^*$ block-sparse across the unknown blocks $\mathcal{B}_1^*, \dots, \mathcal{B}_{K^*}^*$. Our main contribution is to introduce a novel convex penalty function, named *latent optimally partitioned (LOP)- ℓ_2/ℓ_1 penalty*, where the block partition is automatically adapted for the signal of interest. More precisely, we first design a nonconvex penalty function as the minimum of the mixed ℓ_2/ℓ_1 norm over all possible block partitions. Then, to derive a tight convex relaxation of the nonconvex penalty function, using a variational representation of the ℓ_2 norm (see Lemma 1), we represent the nonconvex penalty function by the minimization of a convex function under the ℓ_0 pseudo-norm constraint on latent variables. Finally, the LOP- ℓ_2/ℓ_1 penalty is derived by replacing the ℓ_0 pseudo-norm constraint with its convex envelope, i.e., the ℓ_1 norm constraint. To compute an optimal solution of a block-sparse recovery model designed with the LOP- ℓ_2/ℓ_1 penalty, we reformulate it into a convex optimization problem involving the latent variables of the LOP- ℓ_2/ℓ_1 penalty. By applying proximal splitting techniques, e.g., [38]–[46], to the reformulated problem with the aid of the computation of the proximity operator shown in [47], [48], we obtain an iterative algorithm with guaranteed convergence to the optimal solution. Numerical experiments on both synthetic examples and real-

Fig. 1. Illustration of a block-sparse signal, where we consider that the block partition is unknown a priori, i.e., the sizes and the positions of $B_1^*, \dots, B_{K^*}^*$ are unknown.

world data show that the proposed method achieves superior estimation accuracy to existing convex methods including the LGL method.

The rest of this paper is organized as follows. In Section II, we introduce the LOP- ℓ_2/ℓ_1 penalty, and also show its extension to higher dimensions. In Section III, we design the block-sparse recovery model with the LOP- ℓ_2/ℓ_1 penalty, and develop the iterative algorithm which converges to the optimal solution of the model. Section IV provides numerical examples, followed by conclusion in Section V.

A preliminary short version of this paper is to be presented at a conference [49].

Notations: $\mathbb{R}$, $\mathbb{R}_+$, and $\mathbb{C}$ respectively denote the sets of all real numbers, all nonnegative real numbers, and all complex numbers. For matrices or vectors, we denote the simple transpose and the Hermitian transpose respectively by $(\cdot)^\top$ and $(\cdot)^H$. For $\mathbf{x} = (x_1, \dots, x_N)^\top \in \mathbb{C}^N$ and an index set $\mathcal{I} \subset \{1, \dots, N\}$, $\mathbf{x}_{\mathcal{I}} := (x_n)_{n \in \mathcal{I}}$ denotes the subvector of $\mathbf{x}$ indexed by $\mathcal{I}$. We define the support of $\mathbf{x} \in \mathbb{C}^N$ by $\text{supp}(\mathbf{x}) := \{n \in \{1, \dots, N\} | x_n \neq 0\}$. For a set $\mathcal{S}$, $|\mathcal{S}|$ denotes the cardinality of $\mathcal{S}$. The ℓ_2 norm, the ℓ_1 norm, and the ℓ_0 pseudo-norm of $\mathbf{x} \in \mathbb{C}^N$ are respectively defined by $\|\mathbf{x}\|_2 := \sqrt{\mathbf{x}^H \mathbf{x}}$, $\|\mathbf{x}\|_1 := \sum_{n=1}^N |x_n|$, and $\|\mathbf{x}\|_0 := |\text{supp}(\mathbf{x})|$. We define the operator norm of $\mathbf{L} \in \mathbb{C}^{M \times N}$ by $\|\mathbf{L}\|_{\text{op}} := \max_{\mathbf{x} \neq 0} [\|\mathbf{L}\mathbf{x}\|_2 / \|\mathbf{x}\|_2]$.

Let $(\mathcal{U}, \|\cdot\|)$ be a finite dimensional Hilbert space. A function $f: \mathcal{U} \rightarrow \mathbb{R} \cup \{\infty\}$ is called proper if its effective domain $\text{dom}(f) := \{\mathbf{u} \in \mathcal{U} | f(\mathbf{u}) < \infty\}$ is nonempty. A function $f: \mathcal{U} \rightarrow \mathbb{R} \cup \{\infty\}$ is lower semicontinuous if its lower level set $\{\mathbf{u} \in \mathcal{U} | f(\mathbf{u}) \leq a\}$ is closed for every $a \in \mathbb{R}$. A function $f: \mathcal{U} \rightarrow \mathbb{R} \cup \{\infty\}$ is convex if $f(\beta \mathbf{u} + (1-\beta)\mathbf{v}) \leq \beta f(\mathbf{u}) + (1-\beta)f(\mathbf{v})$ for every $\mathbf{u}, \mathbf{v} \in \mathcal{U}$ and $\beta \in (0, 1)$. The set of all proper lower semicontinuous convex functions from $\mathcal{U}$ to $\mathbb{R} \cup \{\infty\}$ is denoted by $\Gamma_0(\mathcal{U})$. A function $f: \mathcal{U} \rightarrow \mathbb{R} \cup \{\infty\}$ is called coercive if $\lim_{\|\mathbf{u}\| \rightarrow \infty} f(\mathbf{u}) = \infty$. The proximity operator of $f \in \Gamma_0(\mathcal{U})$ is defined by $\text{prox}_f(\mathbf{u}) := \arg \min_{\mathbf{v} \in \mathcal{U}} [f(\mathbf{v}) + \frac{1}{2}\|\mathbf{u} - \mathbf{v}\|^2]$.

II. DESIGN OF PROPOSED PENALTY FUNCTION

To evaluate the block-sparsity of $\mathbf{x} \in \mathbb{C}^N$ across the fixed non-overlapping blocks $B_1, \dots, B_j \subset \{1, \dots, N\}$, the mixed ℓ_2/ℓ_1 norm is widely used as a convex penalty function. The

Fig. 2. Outline of design of the latent optimally partitioned (LOP)- ℓ_2/ℓ_1 penalty $\Psi_\alpha(\mathbf{x})$.

mixed ℓ_2/ℓ_1 norm is defined as the sum (i.e. the ℓ_1 norm) of the block-wise ℓ_2 norms:

$$\|\mathbf{x}\|_{2,1}^{(B_k)_{k=1}^j} := \sum_{k=1}^j \sqrt{|B_k|} \|\mathbf{x}_{B_k}\|_2, \quad (1)$$

where we employ the weight $\sqrt{|B_k|}$ by following the suggestions in, e.g., [5]–[8]. The mixed ℓ_2/ℓ_1 norm promotes the block-sparsity by pushing components in B_k toward zeros together. However, since the mixed ℓ_2/ℓ_1 norm excessively penalizes blocks composed of both zero and nonzero components, its performance degrades when $B_1, \dots, B_j$ do not match with the ground-truth $B_1^*, \dots, B_{K^*}^*$.

To avoid the block mismatch, the block partition is adapted for $\mathbf{x}$ in the proposed penalty function named *latent optimally partitioned (LOP)- ℓ_2/ℓ_1 penalty*. As shown in Fig. 2, we first introduce a nonconvex penalty function $\psi_K(\mathbf{x})$ as the minimum of the mixed ℓ_2/ℓ_1 norm over the partition of at most K blocks. Then, as a suitable convex relaxation of $\psi_K(\mathbf{x})$, we derive the LOP- ℓ_2/ℓ_1 penalty $\Psi_\alpha(\mathbf{x})$. Concretely, we define $\psi_K(\mathbf{x})$ by

$$\begin{aligned} \psi_K(\mathbf{x}) &:= \min_{j \in \{1, \dots, K\}} \left[\min_{(B_k)_{k=1}^j \in \mathcal{P}_j} \|\mathbf{x}\|_{2,1}^{(B_k)_{k=1}^j} \right] \\ &= \min_{j \in \{1, \dots, K\}} \left[\min_{(B_k)_{k=1}^j \in \mathcal{P}_j} \sum_{k=1}^j \sqrt{|B_k|} \|\mathbf{x}_{B_k}\|_2 \right], \end{aligned} \quad (2)$$

where $\mathcal{P}_j$ consists of all j block partitions of $\{1, \dots, N\}$, i.e.,

$$\begin{aligned} & (B_k)_{k=1}^j \in \mathcal{P}_j \\ & \Leftrightarrow \begin{cases} \bigcup_{k=1}^j B_k = \{1, \dots, N\}, \\ B_k \neq \emptyset \quad (k = 1, \dots, j), \\ B_k \cap B_{k'} = \emptyset \quad (k \neq k'), \\ B_k = \{n \in \{1, \dots, N\} \mid n_k \leq n \leq m_k\} \\ \text{for some } n_k, m_k \in \{1, \dots, N\} \quad (k = 1, \dots, j). \end{cases} \end{aligned} \quad (3)$$

Note that K can be set to an upper bound of the ground-truth K^* . Note also that, since the sizes of $B_1, \dots, B_j$ are adapted,

the non-overlapping condition $\mathcal{B}_k \cap \mathcal{B}_{k'} = \emptyset$ is not restrictive.

To derive the convex relaxation of $\psi_K(\mathbf{x})$, we exploit the following lemma which shows a variational representation of the ℓ_2 norm.

Lemma 1. Define $\phi: \mathbb{C} \times \mathbb{R} \rightarrow \mathbb{R}_+ \cup \{\infty\}$ by

$$\phi(x, \tau) := \begin{cases} \frac{|x|^2}{2\tau} + \frac{\tau}{2}, & \text{if } \tau > 0; \\ 0, & \text{if } x = 0 \text{ and } \tau = 0; \\ \infty, & \text{otherwise,} \end{cases} \quad (4)$$

which is a coercive lower semicontinuous convex function. Then, the block-wise ℓ_2 norm is variationally represented as

$$\sqrt{|\mathcal{B}|} \|\mathbf{x}_{\mathcal{B}}\|_2 = \min_{\tau \in \mathbb{R}} \sum_{n \in \mathcal{B}} \phi(x_n, \tau), \quad (5)$$

for any $\mathcal{B} \subset \{1, \dots, N\}$.

Proof. See Appendix C.

Remark 1. Lemma 1 is a slight modification of the existing result

$$\sqrt{|\mathcal{B}|} \|\mathbf{x}_{\mathcal{B}}\|_2 = \inf_{\tau > 0} \sum_{n \in \mathcal{B}} \left[\frac{|x_n|^2}{2\tau} + \frac{\tau}{2} \right]$$

used in, e.g., [33] for the analysis of the LGL penalty and [50] for the design of a graph-sparse penalty. Since the domain of $\sum_{n \in \mathcal{B}} \left[\frac{|x_n|^2}{2\tau} + \frac{\tau}{2} \right]$ is not closed, this representation is inappropriate for the use in concrete optimization algorithms, as the applicability of the algorithm of [50] is limited [51]. Lemma 1 resolves this issue by using the lower semicontinuous convex function ϕ .

By applying Lemma 1 for each $\sqrt{|\mathcal{B}_k|} \|\mathbf{x}_{\mathcal{B}_k}\|_2$ in (2), we can rewrite $\psi_K(\mathbf{x})$ as

$$\psi_K(\mathbf{x}) = \min_{j \in \{1, \dots, K\}} \left[\min_{(\mathcal{B}_k)_{k=1}^j \in \mathcal{P}_j, \tau \in \mathbb{R}^j} \sum_{k=1}^j \sum_{n \in \mathcal{B}_k} \phi(x_n, \tau_k) \right].$$

Introducing a latent vector $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_N)^\top \in \mathbb{R}^N$ as

$$\sigma_n = \tau_k \quad (n \in \mathcal{B}_k) \quad \text{for } k = 1, \dots, j,$$

as illustrated in Fig. 3, we see that $\boldsymbol{\sigma}$ is characterized by the condition that

$$\|\mathbf{D}\boldsymbol{\sigma}\|_0 \leq j-1,$$

where $\mathbf{D}$ is the first difference operator defined by

$$\mathbf{D} := \begin{bmatrix} -1 & 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \cdots & -1 & 1 \end{bmatrix} \in \mathbb{R}^{(N-1) \times N},$$

because $\mathcal{B}_k$ consists of only consecutive indices for each $k = 1, \dots, j$. Moreover, from $\bigcup_{k=1}^j \mathcal{B}_k = \{1, \dots, N\}$ and $\mathcal{B}_k \cap \mathcal{B}_{k'} = \emptyset$ ($k \neq k'$) in (3), we have

$$\sum_{k=1}^j \sum_{n \in \mathcal{B}_k} \phi(x_n, \tau_k) = \sum_{n=1}^N \phi(x_n, \sigma_n).$$

Fig. 3. Illustration of the latent vector $\boldsymbol{\sigma}$ and its difference $\mathbf{D}\boldsymbol{\sigma}$ for $j = 7$. It can be seen that $\mathbf{D}\boldsymbol{\sigma}$ has only $j-1 = 6$ nonzero components.

Thus, we can represent $\psi_K(\mathbf{x})$ as

$$\begin{aligned} \psi_K(\mathbf{x}) &= \min_{j \in \{1, \dots, K\}} \left[\min_{\boldsymbol{\sigma} \in \mathbb{R}^N} \sum_{n=1}^N \phi(x_n, \sigma_n) \right] \\ &= \min_{\substack{\boldsymbol{\sigma} \in \mathbb{R}^N \\ \|\mathbf{D}\boldsymbol{\sigma}\|_0 \leq K-1}} \sum_{n=1}^N \phi(x_n, \sigma_n). \end{aligned}$$

Finally, based on the fact that the ℓ_1 norm is the tightest convex relaxation of the ℓ_0 pseudo-norm in the constraint, we derive the LOP- ℓ_2/ℓ_1 penalty as

$$\Psi_\alpha(\mathbf{x}) := \min_{\substack{\boldsymbol{\sigma} \in \mathbb{R}^N \\ \|\mathbf{D}\boldsymbol{\sigma}\|_1 \leq \alpha}} \sum_{n=1}^N \phi(x_n, \sigma_n), \quad (6)$$

where $\alpha \in \mathbb{R}_+$ is a tuning parameter related to the number of blocks.

Theorem 1. For any $\alpha \in \mathbb{R}_+$, Ψ_α takes finite values, i.e., $\text{dom}(\Psi_\alpha) = \mathbb{C}^N$. Moreover, Ψ_α is coercive, continuous, nonnegative, and convex. Note that this implies $\Psi_\alpha \in \Gamma_0(\mathbb{C}^N)$.

Proof. See Appendix D.

As its special instances, the LOP- ℓ_2/ℓ_1 penalty reproduces the mixed ℓ_2/ℓ_1 norm with the coarsest partition and the finest partition. This is more precisely shown in the next theorem.

Theorem 2. The LOP- ℓ_2/ℓ_1 penalty reproduces the mixed ℓ_2/ℓ_1 norm with the largest block $\{1, \dots, N\}$ and the smallest blocks $(\{k\})_{k=1}^N$ respectively for $\alpha = 0$ and $\alpha \rightarrow \infty$, i.e.,

$$\Psi_0(\mathbf{x}) = \|\mathbf{x}\|_{2,1}^{\{1, \dots, N\}} = \sqrt{N} \|\mathbf{x}\|_2, \quad (7)$$

$$\lim_{\alpha \rightarrow \infty} \Psi_\alpha(\mathbf{x}) = \|\mathbf{x}\|_{2,1}^{\{k\}} = \|\mathbf{x}\|_1. \quad (8)$$

Proof. See Appendix E.

This theorem also suggests that the tuning parameter α indeed controls the number of blocks in the LOP- ℓ_2/ℓ_1 penalty, though blocks do not explicitly appear in (6). Note that, from our experiments shown in Section IV, the performance of the LOP- ℓ_2/ℓ_1 penalty seems fairly robust against the choice of α .

Remark 2 (Generalization to higher dimensions). For a matrix $\mathbf{X} \in \mathbb{C}^{N \times M}$, the LOP- ℓ_2/ℓ_1 penalty can be naturally extended by replacing the 1D difference operator with a 2D one:

$$\Psi_\alpha^{(2d)}(\mathbf{X}) := \min_{\substack{\boldsymbol{\Sigma} \in \mathbb{R}^{N \times M} \\ \|D_{2d}(\boldsymbol{\Sigma})\|_1 \leq \alpha}} \sum_{n=1}^N \sum_{m=1}^M \phi(X_{n,m}, \Sigma_{n,m}), \quad (9)$$

where $D_{2d}: \mathbb{R}^{N \times M} \rightarrow \mathbb{R}^{(N-1)M+N(M-1)}$ computes differences in vertical and horizontal directions by

$$\begin{aligned} D_{2d}(\boldsymbol{\Sigma}) &= \begin{bmatrix} D_v(\boldsymbol{\Sigma}) \\ D_h(\boldsymbol{\Sigma}) \end{bmatrix}, \\ D_v(\boldsymbol{\Sigma}) &= (\Sigma_{n+1,m} - \Sigma_{n,m})_{(n,m) \in \{1, \dots, N-1\} \times \{1, \dots, M\}}, \\ D_h(\boldsymbol{\Sigma}) &= (\Sigma_{n,m+1} - \Sigma_{n,m})_{(n,m) \in \{1, \dots, N\} \times \{1, \dots, M-1\}}. \end{aligned}$$

Note that we do not only focus on rectangular blocks, and $\Psi_\alpha^{(2d)}$ can manage blocks of various shapes because most components of $D_{2d}(\boldsymbol{\Sigma})$ are zeros when $\boldsymbol{\Sigma}$ is constant on each block, irrespective of block shapes. The LOP- ℓ_2/ℓ_1 penalty can be naturally extended to signals in higher dimensions, e.g., p th-order tensors, by modifying the difference operator in similar ways.

III. PROPOSED BLOCK-SPARSE RECOVERY MODEL AND ITS OPTIMIZATION ALGORITHM

We present a block-sparse recovery model using the LOP- ℓ_2/ℓ_1 penalty (6), and develop an iterative algorithm which converges to an optimal solution of the model. Specifically, we consider the following regularization model:

$$\underset{\mathbf{x} \in \mathbb{C}^N}{\text{minimize}} \quad f(\mathbf{L}\mathbf{x}) + \lambda \Psi_\alpha(\mathbf{x}), \quad (10)$$

where $f(\mathbf{L}\mathbf{x})$ is some convex data-fidelity function with $f: \mathbb{C}^J \rightarrow \mathbb{R}_+$ and $\mathbf{L} \in \mathbb{C}^{J \times N}$, and $\lambda > 0$ is the regularization parameter. We suppose that $f \in \Gamma_0(\mathbb{C}^J)$, and its proximity operator can be computed efficiently. Such examples include the square error¹ $f(\mathbf{u}) = \frac{1}{2} \|\mathbf{y} - \mathbf{A}\mathbf{u}\|_2^2$ and $\mathbf{L} = \mathbf{I}$, and the absolute error $f(\mathbf{u}) = \|\mathbf{y} - \mathbf{u}\|_1$ and $\mathbf{L} = \mathbf{A}$, where $\mathbf{y}$ is the known observation vector, $\mathbf{A}$ is the known measurement matrix, and $\mathbf{I}$ denotes the identity matrix. The existence of the solution of (10) is guaranteed by the coercivity of $\Psi_\alpha \in \Gamma_0(\mathbb{C}^N)$, as shown in the next theorem.

Theorem 3. *For any $\lambda > 0$, the proposed model (10) has an optimal solution.*

Proof. It follows from Theorem 1 by [52, Corollary 11.16].

We show that an optimal solution of the proposed model (10) can be obtained by utilizing proximal splitting techniques [38]–[46] as follows. By plugging the definition of $\Psi_\alpha(\mathbf{x})$ in (6) into (10), the optimization problem (10) is translated into

$$\left\{ \begin{array}{l} \underset{(\mathbf{x}, \boldsymbol{\sigma}) \in \mathbb{C}^N \times \mathbb{R}^N}{\text{minimize}} \quad f(\mathbf{L}\mathbf{x}) + \lambda \varphi(\mathbf{x}, \boldsymbol{\sigma}) \\ \text{subject to} \quad \|\mathbf{D}\boldsymbol{\sigma}\|_1 \leq \alpha \end{array} \right\}, \quad (11)$$

¹For the square error, we can also set as $f(\mathbf{u}) = \frac{1}{2} \|\mathbf{y} - \mathbf{u}\|_2^2$ and $\mathbf{L} = \mathbf{A}$, which would be advantageous for large-scale problems since matrix inversion is avoided.

where we let

$$\varphi(\mathbf{x}, \boldsymbol{\sigma}) := \sum_{n=1}^N \phi(x_n, \sigma_n).$$

To apply the primal-dual algorithm [38]–[43] shown in Appendix B, by introducing auxiliary variables $\mathbf{u} = \mathbf{L}\mathbf{x}$ and $\boldsymbol{\eta} = \mathbf{D}\boldsymbol{\sigma}$, we further rewrite the optimization problem (11) as

$$\left\{ \begin{array}{l} \underset{(\mathbf{x}, \boldsymbol{\sigma}, \mathbf{u}, \boldsymbol{\eta}) \in \mathbb{C}^N \times \mathbb{R}^N \times \mathbb{C}^J \times \mathbb{R}^{N-1}}{\text{minimize}} \quad \underbrace{f(\mathbf{u}) + \lambda \varphi(\mathbf{x}, \boldsymbol{\sigma}) + \iota_{B_1^\alpha}(\boldsymbol{\eta})}_{=: G(\mathbf{x}, \boldsymbol{\sigma}, \mathbf{u}, \boldsymbol{\eta})} \\ \text{subject to} \quad \underbrace{\begin{bmatrix} \mu_1 \mathbf{L} & \mathbf{O} & -\mu_1 \mathbf{I} & \mathbf{O} \\ \mathbf{O} & \mu_2 \mathbf{D} & \mathbf{O} & -\mu_2 \mathbf{I} \end{bmatrix}}_{=: \mathbf{H}} \begin{bmatrix} \mathbf{x} \\ \boldsymbol{\sigma} \\ \mathbf{u} \\ \boldsymbol{\eta} \end{bmatrix} = \mathbf{0} \end{array} \right\}, \quad (12)$$

where $\iota_{B_1^\alpha}(\boldsymbol{\eta})$ is the indicator function of the ℓ_1 ball and defined by

$$\iota_{B_1^\alpha}(\boldsymbol{\eta}) := \begin{cases} 0, & \text{if } \|\boldsymbol{\eta}\|_1 \leq \alpha; \\ \infty, & \text{otherwise,} \end{cases}$$

$\mathbf{I}$ and $\mathbf{O}$ respectively denote the identity and zero matrices of appropriate sizes, and μ_1 and μ_2 satisfying

$$\left\{ \begin{array}{l} 0 < \mu_1 \leq \frac{1}{\sqrt{\|\mathbf{L}\|_{\text{op}}^2 + 1}} \\ 0 < \mu_2 \leq \frac{1}{\sqrt{5}} \left(< \frac{1}{\sqrt{\|\mathbf{D}\|_{\text{op}}^2 + 1}} \right) \end{array} \right\} \quad (13)$$

are introduced so that $\|\mathbf{H}\|_{\text{op}} \leq 1$.

Applying the primal-dual algorithm (26) in Appendix B to the problem (12), we obtain the proposed algorithm shown in Algorithm 1, where the updates are written in terms of $(\mathbf{x}, \boldsymbol{\sigma}, \mathbf{u}, \boldsymbol{\eta})$. The proximity operator of γG in (26) reduces to the proximity operators of $\gamma \lambda \varphi$, γf , and $\gamma \iota_{B_1^\alpha}$, since γG is the separable sum of these functions. Based on [48, Example 2.4], we can compute the proximity operator of $\gamma \lambda \varphi$ as

$$\text{prox}_{\gamma \lambda \varphi}(\mathbf{x}, \boldsymbol{\sigma}) = (\text{prox}_{\gamma \lambda \phi}(x_n, \sigma_n))_{n=1}^N, \quad (14)$$

with

$$\begin{aligned} &\text{prox}_{\gamma \lambda \phi}(x, \sigma) \\ &= \begin{cases} (0, 0), & \text{if } 2\gamma\lambda\sigma + |x|^2 \leq \gamma^2\lambda^2; \\ (0, \sigma - \frac{\gamma\lambda}{2}), & \text{if } x = 0 \text{ and } 2\sigma > \gamma\lambda; \\ \left(x - \gamma\lambda s \frac{x}{|x|}, \sigma + \gamma\lambda \frac{s^2 - 1}{2}\right), & \text{otherwise,} \end{cases} \end{aligned} \quad (15)$$

where $s > 0$ is the unique positive root of

$$s^3 + \left(\frac{2}{\gamma\lambda}\sigma + 1\right)s - \frac{2}{\gamma\lambda}|x| = 0,$$

and can be explicitly given via Cardano's formula as follows. Let $p = \frac{2}{\gamma\lambda}\sigma + 1$, $q = -\frac{2}{\gamma\lambda}|x|$, and $D = -\frac{q^2}{4} - \frac{p^3}{27}$. Then,

$$s = \begin{cases} \sqrt[3]{-\frac{q}{2} + \sqrt{-D}} + \sqrt[3]{-\frac{q}{2} - \sqrt{-D}}, & \text{if } D < 0; \\ 2\sqrt[3]{-\frac{q}{2}}, & \text{if } D = 0; \\ 2\sqrt[3]{\sqrt{\frac{q^2}{4} + D} \cos\left(\frac{\arctan(-2\sqrt{D}/q)}{3}\right)}, & \text{if } D > 0, \end{cases}$$

Algorithm 1: Solver for the proposed model (10)**Input:** $\gamma > 0$, μ_1, μ_2 satisfying (13), and

$$\begin{aligned} \mathbf{x}^{(0)} &\in \mathbb{C}^N, \boldsymbol{\sigma}^{(0)} \in \mathbb{R}^N, \mathbf{u}^{(0)} \in \mathbb{C}^J, \\ \boldsymbol{\eta}^{(0)} &\in \mathbb{R}^{N-1}, \mathbf{r}_1^{(0)} \in \mathbb{C}^J, \mathbf{r}_2^{(0)} \in \mathbb{R}^{N-1}. \end{aligned}$$

for $i = 0, 1, 2, \dots$ **do**

$$\begin{aligned} \tilde{\mathbf{x}}^{(i+1)} &= \mathbf{x}^{(i)} + \mu_1 \mathbf{L}^H (\mathbf{r}_1^{(i)} - \mu_1 (\mathbf{L} \mathbf{x}^{(i)} - \mathbf{u}^{(i)})) \\ \tilde{\boldsymbol{\sigma}}^{(i+1)} &= \boldsymbol{\sigma}^{(i)} + \mu_2 \mathbf{D}^T (\mathbf{r}_2^{(i)} - \mu_2 (\mathbf{D} \boldsymbol{\sigma}^{(i)} - \boldsymbol{\eta}^{(i)})) \\ \tilde{\mathbf{u}}^{(i+1)} &= \mathbf{u}^{(i)} - \mu_1 (\mathbf{r}_1^{(i)} - \mu_1 (\mathbf{L} \mathbf{x}^{(i)} - \mathbf{u}^{(i)})) \\ \tilde{\boldsymbol{\eta}}^{(i+1)} &= \boldsymbol{\eta}^{(i)} - \mu_2 (\mathbf{r}_2^{(i)} - \mu_2 (\mathbf{D} \boldsymbol{\sigma}^{(i)} - \boldsymbol{\eta}^{(i)})) \\ (\mathbf{x}^{(i+1)}, \boldsymbol{\sigma}^{(i+1)}) &= \text{prox}_{\gamma \lambda \varphi}(\tilde{\mathbf{x}}^{(i+1)}, \tilde{\boldsymbol{\sigma}}^{(i+1)}) \\ &\quad \text{// see (14) and (15)} \\ \mathbf{u}^{(i+1)} &= \text{prox}_{\gamma f}(\tilde{\mathbf{u}}^{(i+1)}) \quad \text{// use (16) or (17)} \\ \boldsymbol{\eta}^{(i+1)} &= P_{B_1^\alpha}(\tilde{\boldsymbol{\eta}}^{(i+1)}) \quad \text{// see (18) and (19)} \\ \mathbf{r}_1^{(i+1)} &= \mathbf{r}_1^{(i)} - \mu_1 (\mathbf{L} \mathbf{x}^{(i+1)} - \mathbf{u}^{(i+1)}) \\ \mathbf{r}_2^{(i+1)} &= \mathbf{r}_2^{(i)} - \mu_2 (\mathbf{D} \boldsymbol{\sigma}^{(i+1)} - \boldsymbol{\eta}^{(i+1)}) \end{aligned}$$

where $\sqrt[3]{\cdot}$ denotes the real cubic root. Note that the unique existence of the positive root follows from the uniqueness of the proximity operator [47, Theorem 3.1]. The proximity operator of γf depends on the employed data-fidelity function. For the square error $f(\mathbf{u}) = \frac{1}{2} \|\mathbf{y} - \mathbf{A} \mathbf{u}\|_2^2$, we have

$$\text{prox}_{\frac{\gamma}{2} \|\mathbf{y} - \mathbf{A} \cdot\|_2^2}(\mathbf{u}) = (\gamma \mathbf{A}^H \mathbf{A} + \mathbf{I})^{-1} (\gamma \mathbf{A}^H \mathbf{y} + \mathbf{u}), \quad (16)$$

and for the absolute error $f(\mathbf{u}) = \|\mathbf{y} - \mathbf{u}\|_1$,

$$\text{prox}_{\gamma \|\mathbf{y} - \cdot\|_1}(\mathbf{u}) = \left(u_j + \frac{\gamma}{\max\{|y_j - u_j|, \gamma\}} (y_j - u_j) \right)_{j=1}^J. \quad (17)$$

The proximity operator of $\gamma \iota_{B_1^\alpha}$ reduces to the ℓ_1 ball projection $P_{B_1^\alpha}$, which can be computed as

$$P_{B_1^\alpha}(\boldsymbol{\eta}) = \begin{cases} \boldsymbol{\eta}, & \text{if } \|\boldsymbol{\eta}\|_1 \leq \alpha; \\ (a_n \text{sign}(\eta_n))_{n=1}^{N-1}, & \text{otherwise,} \end{cases} \quad (18)$$

with

$$a_n := \max \left\{ |\eta_n| - \sum_{t=1}^T \frac{\rho_t - \alpha}{T}, 0 \right\}, \quad (19)$$

where $\rho_1, \dots, \rho_{N-1}$ are obtained by sorting $|\eta_1|, \dots, |\eta_{N-1}|$ in descending order, and

$$T := \max \left\{ t \in \{1, \dots, N-1\} \mid \sum_{n=1}^t \frac{\rho_n - \alpha}{t} < \rho_t \right\}.$$

The convergence of Algorithm 1 to an optimal solution of the proposed model (10) is confirmed as follows.

Theorem 4. Choose $\gamma > 0$, μ_1 and μ_2 satisfying (13), and initial values $\mathbf{x}^{(0)} \in \mathbb{C}^N$, $\boldsymbol{\sigma}^{(0)} \in \mathbb{R}^N$, $\mathbf{u}^{(0)} \in \mathbb{C}^J$, $\boldsymbol{\eta}^{(0)} \in \mathbb{R}^{N-1}$, $\mathbf{r}_1^{(0)} \in \mathbb{C}^J$, and $\mathbf{r}_2^{(0)} \in \mathbb{R}^{N-1}$. Then, $(\mathbf{x}^{(i)})_{i=1}^\infty$ generated by Algorithm 1 converges to an optimal solution of (10).

Proof. See Appendix F.

Remark 3 (Computational complexity of proposed method). We summarize the computational cost of Algorithm 1 per an iteration. We can compute $\tilde{\mathbf{x}}^{(i+1)}$, $\tilde{\mathbf{u}}^{(i+1)}$, and $\mathbf{r}_1^{(i+1)}$ with $\mathcal{O}(|\text{supp}(\mathbf{L})|)$ operations, and $\tilde{\boldsymbol{\sigma}}^{(i+1)}$, $\tilde{\boldsymbol{\eta}}^{(i+1)}$, and $\mathbf{r}_2^{(i+1)}$ with $\mathcal{O}(N)$ operations since $\mathbf{D}$ has only $2(N-1)$ nonzero entries. The proximity operator of $\gamma \lambda \varphi$ can be computed by (14) and (15) with $\mathcal{O}(N)$ operations. By setting $\mathbf{L} = \mathbf{A}$, $\text{prox}_{\gamma f}$ can be computed with $\mathcal{O}(N)$ operations for both the absolute error and the square error (see footnote 1). Note that, for the square error, setting $\mathbf{L} = \mathbf{I}$ would be beneficial when, e.g., $\mathbf{A}^H \mathbf{A}$ is diagonal. The computation of the ℓ_1 ball projection $P_{B_1^\alpha}$ by (18) and (19) requires $\mathcal{O}(N \log N)$ operations for sorting. We can also use sophisticated ℓ_1 ball projection algorithms having $\mathcal{O}(N)$ expected complexity, e.g., [53]. In particular, this discussion implies that computations regarding the LOP- ℓ_2/ℓ_1 penalty can be implemented with $\mathcal{O}(N)$ operations. Thus, per an iteration, the proposed method has the same computational complexity with the mixed ℓ_2/ℓ_1 regularization method using the non-overlapping blocks when the primal-dual algorithm (26) in Appendix B is applied.

IV. NUMERICAL EXPERIMENTS

To show the effectiveness of the proposed block-sparse recovery model (10) using the LOP- ℓ_2/ℓ_1 penalty (6), we conduct numerical experiments on both synthetic and real-world data. We compare the LOP- ℓ_2/ℓ_1 model with regularization models using existing convex penalties: the ℓ_1 norm, the mixed ℓ_2/ℓ_1 norm (1), and the LGL penalty (24) in Appendix A.

A. Synthetic Examples

We consider the estimation of a block-sparse signal $\mathbf{x}^* \in \mathbb{R}^N$ from noisy compressive measurements. More precisely, we define the measurements by $\mathbf{y} := \mathbf{A} \mathbf{x}^* + \boldsymbol{\varepsilon}$, where the entries of $\mathbf{A} \in \mathbb{R}^{d \times N}$ ($d < N$) are drawn from i.i.d. Gaussian distribution $\mathcal{N}(0, 1)$, and $\boldsymbol{\varepsilon} \in \mathbb{R}^d$ is the white Gaussian noise. The block-sparse signal $\mathbf{x}^*$ is randomly generated by the following scheme. We set $N = 250$, and $\mathbf{x}^*$ has 80 nonzero components, which are randomly divided into 4 blocks. The blocks are randomly located under the condition $|\text{supp}(\mathbf{x}^*)| = 80$. Amplitudes of nonzero components are drawn from i.i.d. $\mathcal{N}(0, 1)$.

In the proposed regularization model (10), we use the square error by setting $f(\mathbf{u}) = \frac{1}{2} \|\mathbf{y} - \mathbf{A} \mathbf{u}\|_2^2$ and $\mathbf{L} = \mathbf{I}$. The existing penalties are also combined with the square error in similar ways. Note that the regularization parameter is tuned independently for each model to obtain the best accuracy. The proposed model (10) is solved by Algorithm 1, where we set $\gamma = 10^{-1}$, $\mu_1 = 1/\sqrt{2}$, and $\mu_2 = 1/\sqrt{5}$ which satisfy the condition (13) for the convergence to an optimal solution. The existing models are also solved by applying the primal-dual algorithm shown in Appendix B. We terminate the iteration when the norm of the differences between the variables of successive iterates is below the threshold 10^{-4} .

We compare the proposed model (10) and the existing models in terms of the normalized mean square error (NMSE)

$$\frac{\|\mathbf{x}^* - \hat{\mathbf{x}}\|_2^2}{\|\mathbf{x}^*\|_2^2},$$

(a) NMSE against the number of measurements where SNR is fixed to 40dB.

(b) NMSE against SNR where the number of measurements is fixed to 200.

Fig. 4. Comparison of the regularization models for the recovery of block-sparse signals from noisy compressive measurements, where the results are averaged over 100 independent trials.

where $\hat{\mathbf{x}}$ is the solution of each model. In Figs. 4(a) and (b), we show the NMSE respectively against the number d of measurements and the SNR $E[\|\mathbf{A}\mathbf{x}^*\|_2^2]/E[\|\boldsymbol{\varepsilon}\|_2^2]$, where the results are averaged over 100 independent trials. The mixed ℓ_2/ℓ_1 model is tested for block sizes 2 and 5, and the LGL model uses overlapping blocks of size 2 in (24). From Figs. 4(a) and (b), we see that the LOP- ℓ_2/ℓ_1 model outperforms the existing models. It can also be seen that the performance of the LOP- ℓ_2/ℓ_1 model is fairly robust against the choice of the tuning parameter α which indirectly controls the number of blocks.

B. Application to Phased Array Weather Radar

A major goal of the phased array weather radar (PAWR) [24], [25] is to estimate the backscattered signal $\mathbf{X}^* \in \mathbb{C}^{M \times N}$ from noisy observations $\mathbf{Y} := \mathbf{A}\mathbf{X}^* + \boldsymbol{\varepsilon} \in \mathbb{C}^{d \times N}$ obtained by transmitting fan beams, where $\mathbf{A} \in \mathbb{C}^{d \times M}$ is a known array manifold matrix, $\boldsymbol{\varepsilon} \in \mathbb{C}^{d \times N}$ is the white Gaussian noise, and M , N , and d respectively are the numbers of elevation angles, pulses, and array elements. This estimation problem is called *beamforming* in the PAWR literature [21]–[23]. As shown in [22], [23], the backscattered signal $\mathbf{X}^*$ exhibits block-sparsity in the Fourier domain for each elevation angle. Since the block partition depends on the unknown mean and standard deviation of the Doppler frequency at each elevation angle, it is also unknown and different for each elevation angle. Thus, the proposed approach is suitable for the PAWR beamforming. Specifically, we design a LOP- ℓ_2/ℓ_1 model for the PAWR beamforming as

$$\underset{\mathbf{X} \in \mathbb{C}^{M \times N}}{\text{minimize}} \frac{1}{2} \|\mathbf{Y} - \mathbf{A}\mathbf{X}\|_{\text{fro}}^2 + \lambda \tilde{\Psi}_{\alpha}^{(2d)}(\mathbf{F}\mathbf{X}^{\top}), \quad (20)$$

where $\|\cdot\|_{\text{fro}}$ denotes the Frobenius norm, and $\mathbf{F} \in \mathbb{C}^{N \times N}$ is the normalized discrete Fourier transform matrix. To exploit the column-wise block-sparsity of $\mathbf{F}\mathbf{X}^{\top}$, we use $\tilde{\Psi}_{\alpha}^{(2d)}$

slightly modified from (9) by replacing D_{2d} with the vertical difference operator D_v . To apply Algorithm 1, we rewrite (20) in the form of (11) as follows. Since $\mathbf{F}$ is a unitary matrix, we have

$$\|\mathbf{Y} - \mathbf{A}\mathbf{X}\|_{\text{fro}}^2 = \|\mathbf{Y}^{\top} - \mathbf{X}^{\top}\mathbf{A}^{\top}\|_{\text{fro}}^2 = \|\mathbf{F}\mathbf{Y}^{\top} - \mathbf{F}\mathbf{X}^{\top}\mathbf{A}^{\top}\|_{\text{fro}}^2.$$

Thus, by letting $\mathbf{Z} = \mathbf{F}\mathbf{X}^{\top}$ and substituting the definition of $\tilde{\Psi}_{\alpha}^{(2d)}$, the optimization problem (20) can be translated into

$$\left. \begin{aligned} & \underset{(\mathbf{Z}, \boldsymbol{\Sigma}) \in \mathbb{C}^{N \times M} \times \mathbb{R}^{N \times M}}{\text{minimize}} \quad \frac{1}{2} \|\mathbf{F}\mathbf{Y}^{\top} - \mathbf{Z}\mathbf{A}^{\top}\|_{\text{fro}}^2 + \lambda \tilde{\varphi}(\mathbf{Z}, \boldsymbol{\Sigma}) \\ & \text{subject to } \|D_v(\boldsymbol{\Sigma})\|_1 \leq \alpha \end{aligned} \right\}, \quad (21)$$

where

$$\tilde{\varphi}(\mathbf{Z}, \boldsymbol{\Sigma}) := \sum_{n=1}^N \sum_{m=1}^M \phi(Z_{n,m}, \Sigma_{n,m}).$$

Further, by defining $\bar{\mathbf{A}} \in \mathbb{C}^{dN \times NM}$ and $\bar{D}_v \in \mathbb{R}^{(N-1)M \times NM}$ satisfying²

$$\begin{aligned} (\forall \mathbf{Z} \in \mathbb{C}^{N \times M}) \quad & \text{vec}(\mathbf{Z}\mathbf{A}^{\top}) = \bar{\mathbf{A}}\text{vec}(\mathbf{Z}), \\ (\forall \boldsymbol{\Sigma} \in \mathbb{R}^{N \times M}) \quad & D_v(\boldsymbol{\Sigma}) = \bar{D}_v\text{vec}(\boldsymbol{\Sigma}), \end{aligned}$$

and $\bar{\mathbf{y}} := \text{vec}(\mathbf{F}\mathbf{Y}^{\top}) \in \mathbb{C}^{dN}$, we can rewrite (21) as

$$\left. \begin{aligned} & \underset{(\bar{\mathbf{z}}, \bar{\boldsymbol{\sigma}}) \in \mathbb{C}^{dN} \times \mathbb{R}^{NM}}{\text{minimize}} \quad \frac{1}{2} \|\bar{\mathbf{y}} - \bar{\mathbf{A}}\bar{\mathbf{z}}\|_2^2 + \lambda \varphi(\bar{\mathbf{z}}, \bar{\boldsymbol{\sigma}}) \\ & \text{subject to } \|\bar{D}_v\bar{\boldsymbol{\sigma}}\|_1 \leq \alpha \end{aligned} \right\}. \quad (22)$$

Since (22) is in the form of (11), we can solve (22) by Algorithm 1 with setting $f(\bar{\mathbf{z}}) = \|\bar{\mathbf{y}} - \bar{\mathbf{A}}\bar{\mathbf{z}}\|_2^2$, $\mathbf{L} = \mathbf{I}$, and $\mathbf{D} = \bar{D}_v$. From $\hat{\mathbf{z}}$ the solution of (22), we obtain the solution of (20) by $\hat{\mathbf{X}} = (\mathbf{F}^H \text{vec}^{-1}(\hat{\mathbf{z}}))^{\top}$, where vec^{-1} is the inverse of the vectorization operator. Existing penalties are also combined with the square error and solved in similar ways,

²We define the vectorization operator by $\text{vec}(\mathbf{V}) := (\mathbf{v}_1^{\top}, \dots, \mathbf{v}_q^{\top})^{\top} \in \mathbb{C}^{pq}$ for $\mathbf{V} := (\mathbf{v}_1, \dots, \mathbf{v}_q) \in \mathbb{C}^{p \times q}$.

TABLE I
COMPARISON OF THE REGULARIZATION MODELS FOR SIMULATION OF PHASED ARRAY WEATHER RADAR IN TERMS OF NMSE IN dB AVERAGED OVER 50 INDEPENDENT TRIALS.

$\text{LOP-}\ell_2/\ell_1 (\alpha = 120M)$	$\text{LOP-}\ell_2/\ell_1 (\alpha = 130M)$	$\text{LOP-}\ell_2/\ell_1 (\alpha = 140M)$	ℓ_2/ℓ_1 (a)	ℓ_2/ℓ_1 (b)	ℓ_1	LGL
-15.62	-15.72	-15.71	-14.78	-14.03	-13.39	-14.83

Fig. 5. Examples of power spectrums of ground-truth and estimates of the backscattered signal, shown in dB, for phased array weather radar simulation.

where the regularization parameters are tuned independently to obtain the best results.

We conduct a numerical simulation in a setting similar to [21]–[23]. The backscattered signal $\mathbf{X}^* \in \mathbb{C}^{M \times N}$ is generated based on the real reflection intensity measured by the PAWR at Osaka University,³ where the elevation angles are chosen uniformly from -15° to 30° degrees with $M = 110$, and the mean and the standard deviation of the Doppler frequency are respectively drawn from uniform distributions of $[-1.25, 1.25]$ and $[0.0314, 0.1257]$ in kHz for each elevation angle. We set $N = 50$ and $d = 128$, and $\mathcal{E}$ as the white Gaussian noise of the standard deviation $\sqrt{5}$. Table I shows the NMSE

$$\frac{\|\mathbf{X}^* - \hat{\mathbf{X}}\|_{\text{fro}}^2}{\|\mathbf{X}^*\|_{\text{fro}}^2},$$

averaged over 50 independent trails, where the LGL model uses overlapping blocks of size 2, and ℓ_2/ℓ_1 (a) and (b) respectively refer to the mixed ℓ_2/ℓ_1 models using block sizes 2 and 4. The result shows that the $\text{LOP-}\ell_2/\ell_1$ model achieves better estimation accuracy than the existing models, and is quite robust against the choice of the tuning parameter α . In Fig. 5, we show examples of the power spectrums of the ground-truth and the estimates, i.e., squared magnitudes of entries of $\mathbf{F}\mathbf{X}^{*\top}$ and $\mathbf{F}\hat{\mathbf{X}}^\top$. It can be seen that the $\text{LOP-}\ell_2/\ell_1$ model reduces the artifacts caused in the LGL model, e.g., in the zoomed areas of Fig. 5.

C. Application of Proposed 2D Penalty to Speech Denoising

To show the effectiveness of the 2D extension (9) of the $\text{LOP-}\ell_2/\ell_1$ penalty, we conduct experiments on the speech denoising exploiting block-sparsity of the spectrogram. We generate noisy speech as $\mathbf{y} := \mathbf{s}^* + \boldsymbol{\varepsilon} \in \mathbb{R}^d$, where $\mathbf{s}^*$ is a

2-second clip of speech taken from [54] with 16kHz sampling rate (i.e. $d = 32000$), and $\boldsymbol{\varepsilon}$ is the white Gaussian noise. The spectrogram $\mathbf{X}^* := \mathcal{F}(\mathbf{s}^*)$ of the clean speech, where $\mathcal{F}$ denotes the short-time Fourier transform (STFT), is expected to be block-sparse because entries of $\mathbf{X}^*$ in adjacent frequency bins and frames tend to be zeros or nonzeros together. The block partition depends on the characteristics of the speech to be recovered, and thus is unknown a priori. Hence, the proposed approach is suitable for exploiting the block-sparsity of the spectrogram. Specifically, we apply the 2D $\text{LOP-}\ell_2/\ell_1$ penalty in (9) to the speech denoising as

$$\hat{\mathbf{X}} \in \arg \min_{\mathbf{X} \in \mathbb{C}^{N \times M}} \frac{1}{2} \|\mathbf{y} - \mathcal{F}^{-1}(\mathbf{X})\|_2^2 + \lambda \Psi_\alpha^{(2d)}(\mathbf{X}), \quad (23)$$

where N and M are respectively the numbers of frequency bins and frames, and $\mathcal{F}^{-1}$ denotes the inverse STFT. We implement the STFT as a *semi-orthogonal transformation* by using the hann window of 32ms (i.e. 512 samples) with 75% overlap and appropriate zero padding. Note that this setting implies $N = 512$ and $M = 253$. Since D_{2d} and $\mathcal{F}^{-1}$ are linear operators, we can solve (23) by Algorithm 1 through reformulation with the vectorization of $\mathbf{X}$, similarly to the reformulation of (21) into (22), where we should choose $\mu_2 \leq 1/3$ according to the operator norm of D_{2d} . Note that $\mathcal{F}^{-1}$ can be computed as $\mathbf{F}^*$, the adjoint of $\mathcal{F}$, thanks to the semi-orthogonality of $\mathcal{F}$. The existing penalties are used in manners similar to (23), where the regularization parameters are tuned independently to obtain the best NMSE.

Experiments are conducted for male and female speech with input SNRs of 5dB and 10dB. In Table II, we show the NMSE

$$\frac{\|\mathbf{s}^* - \mathcal{F}^{-1}(\hat{\mathbf{X}})\|_2^2}{\|\mathbf{s}^*\|_2^2},$$

averaged over 20 independent trials, where ℓ_2/ℓ_1 (a) and (b) respectively denote the mixed ℓ_2/ℓ_1 models using block sizes 2×4 and 2×2 , which yield best results among

³The PAWR data is courtesy of Prof. Tomoo Ushio, Dr. Hiroshi Kikuchi, and Dr. Eiichi Yoshikawa.

TABLE II
COMPARISON OF THE REGULARIZATION MODELS FOR SPEECH DENOISING IN TERMS OF NMSE IN dB AVERAGED OVER 20 INDEPENDENT TRIALS.

	LOP- ℓ_2/ℓ_1 ($\alpha = 0.006NM$)	LOP- ℓ_2/ℓ_1 ($\alpha = 0.007NM$)	LOP- ℓ_2/ℓ_1 ($\alpha = 0.008NM$)	ℓ_2/ℓ_1 (a)	ℓ_2/ℓ_1 (b)	ℓ_1
Male, 5dB	-12.36	-12.28	-12.23	-12.16	-11.98	-11.85
Male, 10dB	-15.78	-15.73	-15.68	-15.56	-15.41	-15.28
Female, 5dB	-14.05	-14.07	-14.06	-13.69	-13.49	-13.47
Female, 10dB	-17.37	-17.43	-17.41	-17.04	-16.90	-16.85

Fig. 6. Examples of magnitude spectrograms of clean speech and estimates, shown in dB, for the denoising of male speech with the input SNR of 10dB.

block sizes $\{1, 2, 4, 8\} \times \{1, 2, 4, 8\}$. We here do not include the LGL model since the use of 2D overlapping blocks is computationally expensive. From Table II, we see that the 2D LOP- ℓ_2/ℓ_1 model achieves better estimation accuracy than the existing models for every experimental condition. In addition, the performance of the 2D LOP- ℓ_2/ℓ_1 model is fairly robust against the choice of α . Fig. 6 illustrates the magnitude spectrograms of the clean speech and the estimates, i.e., magnitudes of entries of $\mathbf{X}^*$ and $\hat{\mathbf{X}}$, for the denoising of male speech with 10dB input SNR. Rectangular artifacts are observed in Fig. 6(b) which shows the magnitude spectrogram obtained by the mixed ℓ_2/ℓ_1 model with the block size 2×4 . These artifacts are considered to be caused by the rectangular blocks used in the mixed ℓ_2/ℓ_1 norm. Note that, since the blocks are used explicitly, it is difficult to employ blocks of various shapes in the mixed ℓ_2/ℓ_1 norm due to the computational complexity. On the other hand, the 2D LOP- ℓ_2/ℓ_1 penalty can handle blocks of various shapes because the blocks are implicitly controlled with the 2D difference operator D_{2d} (see Remark 2). Indeed, from Fig. 6(c), we see that the 2D LOP- ℓ_2/ℓ_1 model considerably reduces such artifacts, and more accurately recovers detailed shapes of the spectrogram.

V. CONCLUSION

We presented a convex recovery method for block-sparse signals whose block partitions are not available a priori. We first introduce a nonconvex penalty function ψ_K in (2), where the block partition is optimized for the signal of interest by minimizing the mixed ℓ_2/ℓ_1 norm. Then, as the tight convex relaxation of ψ_K , we derive the proposed latent optimally partitioned (LOP)- ℓ_2/ℓ_1 penalty Ψ_α in (6). For the block-sparse recovery model (10) using the LOP- ℓ_2/ℓ_1 penalty, by applying proximal splitting techniques, we develop the iterative solver

as Algorithm 1 which is guaranteed to converge to an optimal solution. Numerical experiments on both synthetic and real-world data illustrate the effectiveness of the proposed method.

APPENDIX A

BRIEF REVIEW OF CONVEX BLOCK-SPARSE PENALTIES USING OVERLAPPING BLOCKS

A simple modification of the mixed ℓ_2/ℓ_1 norm is to use, in (1), overlapping blocks $\bar{\mathcal{B}}_1, \dots, \bar{\mathcal{B}}_{\bar{K}} \subset \{1, \dots, N\}$ such that

$$\begin{aligned} \bigcup_{k=1}^{\bar{K}} \bar{\mathcal{B}}_k &= \{1, \dots, N\}, \\ \bar{\mathcal{B}}_k &\neq \emptyset \quad (k = 1, \dots, \bar{K}), \\ \bar{\mathcal{B}}_k &= \{n \in \{1, \dots, N\} \mid \bar{n}_k \leq n \leq \bar{m}_k\} \\ &\quad \text{for some } \bar{n}_k, \bar{m}_k \in \{1, \dots, N\} \quad (k = 1, \dots, \bar{K}), \end{aligned}$$

where the non-overlapping condition $\mathcal{B}_k \cap \mathcal{B}_{k'} = \emptyset$ ($k \neq k'$) is omitted, e.g., as in [26]–[31]. However, this modification cannot remove the blocks containing both zero and nonzero components. To alleviate this drawback, the latent group lasso (LGL) method [32], [33] is presented to select relevant blocks from the pre-defined overlapping blocks $\bar{\mathcal{B}}_1, \dots, \bar{\mathcal{B}}_{\bar{K}}$. Specifically, the LGL penalty is defined by⁴

$$\text{LGL}(\mathbf{x}) := \min_{\substack{(\mathbf{w}_1, \dots, \mathbf{w}_{\bar{K}}) \in \mathbb{C}^{N \times \bar{K}} \\ \mathbf{x} = \sum_{k=1}^{\bar{K}} \mathbf{w}_k, \\ \text{supp}(\mathbf{w}_k) \subset \bar{\mathcal{B}}_k \quad (k=1, \dots, \bar{K})}} \sum_{k=1}^{\bar{K}} \sqrt{|\bar{\mathcal{B}}_k|} \|\mathbf{w}_k\|_2. \quad (24)$$

⁴By setting $\bar{\mathcal{B}}_k$ to other than blocks, the LGL method is applicable to other structural sparsity, while this paper focuses on the block-sparsity.

The LGL penalty can also be written as

$$\text{LGL}(\mathbf{x}) = \min_{\substack{\tilde{\mathbf{w}} \in \mathbb{C}^{\tilde{N}} \\ \mathbf{x} = \mathbf{G}\tilde{\mathbf{w}}}} \sum_{k=1}^{\tilde{K}} \sqrt{|\tilde{\mathcal{B}}_k|} \|\tilde{\mathbf{w}}_{\mathcal{I}_k}\|_2,$$

where $\tilde{N} = \sum_{k=1}^{\tilde{K}} |\tilde{\mathcal{B}}_k|$,

$$\mathcal{I}_k := \left\{ n \in \{1, \dots, \tilde{N}\} \mid 1 + \sum_{m=1}^{k-1} |\tilde{\mathcal{B}}_m| \leq n \leq \sum_{m=1}^k |\tilde{\mathcal{B}}_m| \right\},$$

$$\mathbf{G} := [\mathbf{G}_1 \ \cdots \ \mathbf{G}_{\tilde{K}}] \in \mathbb{R}^{N \times \tilde{N}},$$

$$\mathbf{G}_k := \begin{bmatrix} \mathbf{O}_{\tilde{n}_k-1, |\tilde{\mathcal{B}}_k|} \\ \mathbf{I}_{|\tilde{\mathcal{B}}_k|} \\ \mathbf{O}_{N-\tilde{n}_k, |\tilde{\mathcal{B}}_k|} \end{bmatrix} \in \mathbb{R}^{N \times |\tilde{\mathcal{B}}_k|},$$

with $\mathbf{I}_p \in \mathbb{R}^{p \times p}$ and $\mathbf{O}_{p,q} \in \mathbb{R}^{p \times q}$ respectively denoting the identity matrix of order p and the zero matrix of size $p \times q$. Thus, when the LGL penalty is combined with some convex data-fidelity function, the resulting regularization model can be solved as a convex optimization problem of $\tilde{\mathbf{w}} \in \mathbb{C}^{\tilde{N}}$. However, it is intractable to use all possible blocks as $\tilde{\mathcal{B}}_1, \dots, \tilde{\mathcal{B}}_{\tilde{K}}$ because the problem size becomes too large, i.e., $\tilde{N} = N(N+1)(N+2)/6$, in this case. Due to this issue, $\tilde{\mathcal{B}}_1, \dots, \tilde{\mathcal{B}}_{\tilde{K}}$ are typically restricted to blocks of a fixed size. On the other hand, the proposed approach is free from such issue because the block partition is implicitly optimized in the proposed penalty (6) with the latent vector $\boldsymbol{\sigma}$ of size N .

APPENDIX B

PRIMAL-DUAL ALGORITHM FOR NONSMOOTH CONVEX OPTIMIZATION PROBLEM

The Chambolle-Pock primal-dual algorithm [38], [39] and the Loris-Verhoeven algorithm [40], [41] are widely used for solving nonsmooth convex optimization problems. As explained in [46], these algorithms are essentially the same for the following problem

$$\underset{\mathbf{v} \in \mathcal{V}}{\text{minimize}} G(\mathbf{v}) \text{ subject to } \mathbf{H}\mathbf{v} = \mathbf{0}, \quad (25)$$

and the iteration can be written as

$$\begin{cases} \tilde{\mathbf{v}}^{(i+1)} = \mathbf{v}^{(i)} + \mathbf{H}^*(\mathbf{r}^{(i)} - \mathbf{H}\mathbf{v}^{(i)}) \\ \mathbf{v}^{(i+1)} = \text{prox}_{\gamma G}(\tilde{\mathbf{v}}^{(i+1)}) \\ \mathbf{r}^{(i+1)} = \mathbf{r}^{(i)} - \mathbf{H}\mathbf{v}^{(i+1)}, \end{cases} \quad (26)$$

where we suppose that $G \in \Gamma_0(\mathcal{V})$, $\mathbf{H}: \mathcal{V} \rightarrow \mathcal{W}$ is a linear operator, $\mathbf{H}^*: \mathcal{W} \rightarrow \mathcal{V}$ is the adjoint of $\mathbf{H}$, $\mathcal{V}$ and $\mathcal{W}$ are finite-dimensional Hilbert spaces, and $\gamma > 0$. This algorithm is also known as the linearized augmented Lagrangian algorithm [42], [43]. The convergence of (26) to an optimal solution of (25) is shown in the following lemma from, e.g., [46, Theorem 4.3] and [42, Corollary 1].

Lemma 2. *Suppose that an optimal solution of (25) exists, the so-called qualification condition⁵ $\mathbf{0} \in \text{ri}(\mathbf{H}(\text{dom}(G)))$ holds, $\|\mathbf{H}\|_{\text{op}} \leq 1$, and $\gamma > 0$. Let $\mathbf{v}^{(0)} \in \mathcal{V}$ and $\mathbf{r}^{(0)} \in \mathcal{W}$. Then,*

⁵For a set $S \subset \mathcal{V}$, $\mathbf{H}(S) := \{\mathbf{H}\mathbf{v} \in \mathcal{W} \mid \mathbf{v} \in S\}$. For a set $C \subset \mathcal{W}$, $\text{ri}(C)$ denotes the relative interior of C (see, e.g., [52, Definition 6.9]).

$(\mathbf{v}^{(i)})_{i=1}^{\infty}$ generated by (26) converges to an optimal solution of (25).

Note that, when $\|\mathbf{H}\|_{\text{op}} > 1$, by replacing $\mathbf{H}$ in (25) with $\mathbf{H}/C$ where $C \geq \|\mathbf{H}\|_{\text{op}}$, the condition $\|\mathbf{H}\|_{\text{op}} \leq 1$ can be satisfied without changing the solution set of (25). Thus, the iteration (26) can compute an optimal solution of (25) for general $\mathbf{H}$.

APPENDIX C

PROOF OF LEMMA 1

First, we prove the properties of ϕ defined by (4). Notice that ϕ can be written as

$$\phi(x, \tau) = \frac{h(x, \tau)}{2} + \frac{\tau}{2},$$

where

$$h(x, \tau) := \begin{cases} \frac{|x|^2}{\tau}, & \text{if } \tau > 0; \\ 0, & \text{if } x = 0 \text{ and } \tau = 0; \\ \infty, & \text{otherwise.} \end{cases}$$

Since h is the *perspective function* of $|x|^2$, it is lower semicontinuous convex [55, Proposition 2.3 and Example 2.6], and thus ϕ is lower semicontinuous convex. The coercivity of ϕ is proven by showing $\phi(x, \tau) \rightarrow \infty$ ($|x| + |\tau| \rightarrow \infty$) as follows.

- a) (Case of $|\tau| \rightarrow \infty$) Clear from $\phi(x, \tau) \geq \frac{\tau}{2}$.
- b) (Case of $|\tau|$ is bounded) In this case, we have $|x| \rightarrow \infty$ and $\exists R > 0$ such that $|\tau| \leq R$. Moreover, since $\phi(x, \tau) = \infty$ when $\tau \leq 0$ and $x \neq 0$, we have

$$(\forall x \neq 0) \quad \phi(x, \tau) \geq \frac{|x|^2}{R},$$

which implies $\phi(x, \tau) \rightarrow \infty$ due to $|x| \rightarrow \infty$.

Next, we prove the equation (5) as follows.

- a) (Case of $\mathbf{x}_{\mathcal{B}} \neq \mathbf{0}$) For $\tau > 0$, we have

$$\sum_{n \in \mathcal{B}} \phi(x_n, \tau) = \sum_{n \in \mathcal{B}} \left(\frac{|x_n|^2}{2\tau} + \frac{\tau}{2} \right) = \frac{\|\mathbf{x}_{\mathcal{B}}\|_2^2}{2\tau} + \frac{|\mathcal{B}|}{2}\tau.$$

Since $\mathbf{x}_{\mathcal{B}} \neq \mathbf{0}$ in this case, $\frac{\|\mathbf{x}_{\mathcal{B}}\|_2^2}{2\tau} + \frac{|\mathcal{B}|}{2}\tau$ is a strictly convex function of $\tau > 0$, and the minimum is attained when $\tau = \|\mathbf{x}_{\mathcal{B}}\|_2 / \sqrt{|\mathcal{B}|}$. Thus, we have

$$\min_{\tau > 0} \sum_{n \in \mathcal{B}} \phi(x_n, \tau) = \sqrt{|\mathcal{B}|} \|\mathbf{x}_{\mathcal{B}}\|_2.$$

For $\tau = 0$, since there exists $n \in \mathcal{B}$ such that $x_n \neq 0$ in this case, we have $\sum_{n \in \mathcal{B}} \phi(x_n, 0) = \infty$. For $\tau < 0$, the definition of ϕ in (4) readily implies $\sum_{n \in \mathcal{B}} \phi(x_n, \tau) = \infty$. Summarizing, we have the equation (5).

- b) (Case of $\mathbf{x}_{\mathcal{B}} = \mathbf{0}$) For $\tau > 0$, since $x_n = 0$ for every $n \in \mathcal{B}$ in this case, we have

$$\sum_{n \in \mathcal{B}} \phi(x_n, \tau) = \sum_{n \in \mathcal{B}} \left(\frac{|x_n|^2}{2\tau} + \frac{\tau}{2} \right) = \frac{\tau}{2} > 0.$$

For $\tau = 0$, we have

$$\sum_{n \in \mathcal{B}} \underbrace{\phi(x_n, 0)}_{=0} = 0.$$

For $\tau < 0$, we have $\sum_{n \in \mathcal{B}} \phi(x_n, \tau) = \infty$. Thus, the minimum is attained when $\tau = 0$, and

$$\min_{\tau \in \mathbb{R}} \sum_{n \in \mathcal{B}} \phi(x_n, \tau) = 0,$$

which implies (5) since $\mathbf{x}_{\mathcal{B}} = \mathbf{0}$ in this case.

APPENDIX D PROOF OF THEOREM 1

Since Ψ_α is defined by the partial minimization of the coercive lower semicontinuous convex function $\sum_{n=1}^N \phi(x_n, \sigma_n)$ under the closed convex constraint $\|\mathbf{D}\boldsymbol{\sigma}\|_1 \leq \alpha$, and

$$\left\{ \boldsymbol{\sigma} \in \mathbb{R}^N \left| \sum_{n=1}^N \phi(x_n, \sigma_n) < \infty \text{ and } \|\mathbf{D}\boldsymbol{\sigma}\|_1 \leq \alpha \right. \right\} \neq \emptyset \quad (27)$$

for any $\alpha \in \mathbb{R}_+$, Ψ_α takes finite values for every $\mathbf{x} \in \mathbb{C}^N$, i.e., $\text{dom}(\Psi_\alpha) = \mathbb{C}^N$, by [52, Proposition 11.15]. Note that (27) holds for $\alpha \in \mathbb{R}_+$ since $\mathbf{D}\boldsymbol{\sigma} = \mathbf{0}$ for any $\boldsymbol{\sigma}$ satisfying $\sigma_1 = \sigma_2 = \dots = \sigma_N$. The convexity of Ψ_α follows from the convexity of $\sum_{n=1}^N \phi(x_n, \sigma_n)$ and $\|\mathbf{D}\boldsymbol{\sigma}\|_1 \leq \alpha$ by [52, Proposition 8.35]. Since Ψ_α is convex and $\text{dom}(\Psi_\alpha) = \mathbb{C}^N$, Ψ_α is continuous by [52, Corollary 8.40]. From $\sum_{n=1}^N \phi(x_n, \sigma_n) \geq 0$, Ψ_α is nonnegative. Lastly, the coercivity of Ψ_α follows from

$$\begin{aligned} \Psi_\alpha(\mathbf{x}) &= \min_{\substack{\boldsymbol{\sigma} \in \mathbb{R}^N \\ \|\mathbf{D}\boldsymbol{\sigma}\|_1 \leq \alpha}} \sum_{n=1}^N \phi(x_n, \sigma_n) \\ &\geq \min_{\boldsymbol{\sigma} \in \mathbb{R}^N} \sum_{n=1}^N \phi(x_n, \sigma_n) \\ &= \sum_{n=1}^N \left[\min_{\sigma_n \in \mathbb{R}} \phi(x_n, \sigma_n) \right] = \|\mathbf{x}\|_1, \end{aligned}$$

where the last equality is obtained by applying Lemma 1 for each of $\min_{\sigma_n \in \mathbb{R}} \phi(x_n, \sigma_n)$ ($n = 1, \dots, N$).

APPENDIX E PROOF OF THEOREM 2

The equation (7) is proven as

$$\begin{aligned} \Psi_0(\mathbf{x}) &= \min_{\substack{\boldsymbol{\sigma} \in \mathbb{R}^N \\ \|\mathbf{D}\boldsymbol{\sigma}\|_1 \leq 0}} \sum_{n=1}^N \phi(x_n, \sigma_n) \\ &= \min_{\boldsymbol{\sigma} \in \mathbb{R}^N} \sum_{n=1}^N \phi(x_n, \sigma) = \sqrt{N} \|\mathbf{x}\|_2, \end{aligned}$$

where the second equality holds due to

$$\|\mathbf{D}\boldsymbol{\sigma}\|_1 \leq 0 \Leftrightarrow \mathbf{D}\boldsymbol{\sigma} = \mathbf{0} \Leftrightarrow \sigma_1 = \sigma_2 = \dots = \sigma_N,$$

and the last equality follows from Lemma 1. Next, (8) is proven as

$$\lim_{\alpha \rightarrow \infty} \Psi_\alpha(\mathbf{x}) = \min_{\boldsymbol{\sigma} \in \mathbb{R}^N} \sum_{n=1}^N \phi(x_n, \sigma_n) = \|\mathbf{x}\|_1,$$

where the last equality holds by Lemma 1.

APPENDIX F PROOF OF THEOREM 4

First, the existence of the solution of (12) is guaranteed from Theorem 3 and the equivalence between (10), (11), and (12). The qualification condition for (12) is confirmed as follows. The effective domain of G is given by

$$\begin{aligned} \text{dom}(G) &= \text{dom}(\varphi) \times \text{dom}(f) \times \text{dom}(\iota_{B_1^\alpha}) \\ &= ((\mathbb{C}^N \times \mathbb{R}_{++}^N) \cup \{\mathbf{0}\}) \times \mathbb{C}^J \times B_1^\alpha, \end{aligned}$$

where $\mathbb{R}_{++}$ denotes the set of all positive real numbers, and $B_1^\alpha = \{\boldsymbol{\eta} \in \mathbb{R}^{N-1} \mid \|\boldsymbol{\eta}\|_1 \leq \alpha\}$. Thus, we have

$$\begin{aligned} \mathbf{H}(\text{dom}(G)) &= \left\{ \begin{bmatrix} \mu_1 \mathbf{L}\mathbf{x} - \mu_1 \mathbf{u} \\ \mu_2 \mathbf{D}\boldsymbol{\sigma} - \mu_2 \boldsymbol{\eta} \end{bmatrix} \mid (\mathbf{x}, \boldsymbol{\sigma}, \mathbf{u}, \boldsymbol{\eta}) \in \text{dom}(G) \right\} \\ &= (\mathbf{L}(\mathbb{C}^N) - \mathbb{C}^J) \times (\mathbf{D}(\mathbb{R}^N) - \mu_2 B_1^\alpha), \quad (28) \end{aligned}$$

where we use the relation $\mathbf{D}(\mathbb{R}_{++}^N) = \mathbf{D}(\mathbb{R}^N)$. This relation holds because, for any $\boldsymbol{\sigma} \in \mathbb{R}^N$, $\mathbf{D}(\boldsymbol{\sigma} + c\mathbf{1}) = \mathbf{D}\boldsymbol{\sigma}$ and $\boldsymbol{\sigma} + c\mathbf{1} \in \mathbb{R}_{++}^N$ can be satisfied with some $c \in \mathbb{R}_{++}$, where $\mathbf{1} := (1, 1, \dots, 1)^\top \in \mathbb{R}_{++}^N$. Since the expression (28) implies that $\mathbf{H}(\text{dom}(G))$ is a nonempty convex set satisfying the symmetry $\mathbf{H}(\text{dom}(G)) = -\mathbf{H}(\text{dom}(G))$, the qualification condition $\mathbf{0} \in \text{ri}(\mathbf{H}(\text{dom}(G)))$ holds by [52, Example 6.10]. The condition $\|\mathbf{H}\|_{\text{op}} \leq 1$ is checked as follows. Since

$$\mathbf{H}\mathbf{H}^\text{H} = \begin{bmatrix} \mu_1^2(\mathbf{L}\mathbf{L}^\text{H} + \mathbf{I}) & \mathbf{O} \\ \mathbf{O} & \mu_2^2(\mathbf{D}\mathbf{D}^\text{T} + \mathbf{I}) \end{bmatrix}$$

is block-diagonal, with $\rho(\cdot)$ denoting the largest eigenvalue of the argument, we have

$$\begin{aligned} \|\mathbf{H}\|_{\text{op}} &= \sqrt{\rho(\mathbf{H}\mathbf{H}^\text{H})} \\ &= \max \left\{ \sqrt{\rho(\mu_1^2(\mathbf{L}\mathbf{L}^\text{H} + \mathbf{I}))}, \sqrt{\rho(\mu_2^2(\mathbf{D}\mathbf{D}^\text{T} + \mathbf{I}))} \right\} \\ &= \max \left\{ \mu_1 \sqrt{\rho(\mathbf{L}\mathbf{L}^\text{H}) + 1}, \mu_2 \sqrt{\rho(\mathbf{D}\mathbf{D}^\text{T}) + 1} \right\} \\ &= \max \left\{ \mu_1 \sqrt{\|\mathbf{L}\|_{\text{op}}^2 + 1}, \mu_2 \sqrt{\|\mathbf{D}\|_{\text{op}}^2 + 1} \right\} \leq 1, \end{aligned}$$

where the last inequality holds since μ_1 and μ_2 are chosen to satisfy (13). Thus, by Lemma 2, $(\mathbf{x}^{(i)}, \boldsymbol{\sigma}^{(i)}, \mathbf{u}^{(i)}, \boldsymbol{\eta}^{(i)})_{i=1}^\infty$ converges to an optimal solution of (12). The convergence of $(\mathbf{x}^{(i)})_{i=1}^\infty$ to the solution of (10) follows from the relation between (10), (11), and (12).

ACKNOWLEDGMENT

We would like to acknowledge helpful discussions with Dr. Akira Hirabayashi. We would also like to thank Prof. Tomoo Ushio, Dr. Hiroshi Kikuchi, and Dr. Eiichi Yoshikawa for providing us with the PAWR data.

REFERENCES

- [1] M. Stojnic, F. Parvaresh, and B. Hassibi, "On the reconstruction of block-sparse signals with an optimal number of measurements," *IEEE Trans. Signal Process.*, vol. 57, no. 8, pp. 3075–3085, Aug. 2009.
- [2] Y. C. Eldar, P. Kuppinger, and H. Bolcskei, "Block-sparse signals: Uncertainty relations and efficient recovery," *IEEE Trans. Signal Process.*, vol. 58, no. 6, pp. 3042–3054, Jun. 2010.
- [3] X. Lv, G. Bi, and C. Wan, "The group lasso for stable recovery of block-sparse signal representations," *IEEE Trans. Signal Process.*, vol. 59, no. 4, pp. 1371–1382, Apr. 2011.

- [4] E. Elhamifar and R. Vidal, "Block-sparse recovery via convex optimization," *IEEE Trans. Signal Process.*, vol. 60, no. 8, pp. 4094–4107, Aug. 2012.
- [5] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Roy. Statist. Soc. Ser. B*, vol. 68, no. 1, pp. 49–67, 2006.
- [6] Y. Nardi and A. Rinaldo, "On the asymptotic properties of the group lasso estimator for linear models," *Electron. J. Statist.*, vol. 2, pp. 605–633, 2008.
- [7] F. Wei and J. Huang, "Consistent group selection in high-dimensional linear regression," *Bernoulli*, vol. 16, no. 4, pp. 1369–1384, 2010.
- [8] J. Huang, P. Breheny, and S. Ma, "A selective review of group selection in high-dimensional models," *Statist. Sci.*, vol. 27, no. 4, pp. 481–499, Nov. 2012.
- [9] F. R. Bach, "Consistency of the group lasso and multiple kernel learning," *J. Mach. Learn. Res.*, vol. 9, pp. 1179–1225, Jun. 2008.
- [10] L. Meier, S. Van De Geer, and P. Bühlmann, "The group lasso for logistic regression," *J. Roy. Statist. Soc. Ser. B*, vol. 70, no. 1, pp. 53–71, 2008.
- [11] G. Yu, S. Mallat, and E. Bacry, "Audio denoising by time-frequency block thresholding," *IEEE Trans. Signal Process.*, vol. 56, no. 5, pp. 1830–1839, May 2008.
- [12] R. Gribonval and E. Bacry, "Harmonic decomposition of audio signals with matching pursuit," *IEEE Trans. Signal Process.*, vol. 51, no. 1, pp. 101–111, Jan. 2003.
- [13] L. Yu, H. Sun, J. Barbot, and G. Zheng, "Bayesian compressive sensing for cluster structured sparse signals," *Signal Process.*, vol. 92, no. 1, pp. 259–269, 2012.
- [14] V. Cevher, A. Sankaranarayanan, M. F. Duarte, D. Reddy, R. G. Baraniuk, and R. Chellappa, "Compressive sensing for background subtraction," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2008, pp. 155–168.
- [15] Z. Gao, L. Cheong, and Y. Wang, "Block-sparse RPCA for salient motion detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 10, pp. 1975–1987, Oct. 2014.
- [16] J. Fang, L. Zhang, and H. Li, "Two-dimensional pattern-coupled sparse Bayesian learning via generalized approximate message passing," *IEEE Trans. Image Process.*, vol. 25, no. 6, pp. 2920–2930, Jun. 2016.
- [17] F. Gustafsson, *Adaptive filtering and change detection*. Wiley, 2000.
- [18] H. Kuroda, M. Yamagishi, and I. Yamada, "Exploiting sparsity in tight-dimensional spaces for piecewise continuous signal recovery," *IEEE Trans. Signal Process.*, vol. 66, no. 24, pp. 6363–6376, Dec. 2018.
- [19] —, "Segmentation of piecewise ARX processes by exploiting sparsity in tight-dimensional spaces," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Jul. 2019, 5 pages.
- [20] L. Wang, L. Zhao, G. Bi, C. Wan, and L. Yang, "Enhanced ISAR imaging by exploiting the continuity of the target scene," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 9, pp. 5736–5750, Sep. 2014.
- [21] E. Yoshikawa, T. Ushio, Z. Kawasaki, S. Yoshida, T. Morimoto, F. Mizutani, and M. Wada, "MMSE beam forming on fast-scanning phased array weather radar," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 5, pp. 3077–3088, May 2013.
- [22] D. Kitahara, M. Nakahara, A. Hirabayashi, E. Yoshikawa, H. Kikuchi, and T. Ushio, "Nonlinear beamforming via convex optimization for phased array weather radar," in *Proc. Asia-Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Nov. 2018, pp. 1831–1835.
- [23] D. Kitahara, H. Kuroda, A. Hirabayashi, E. Yoshikawa, H. Kikuchi, and T. Ushio, "Nonlinear beamforming based on group-sparsities of periodograms for phased array weather radar," preprint TechRxiv:14138537, Mar. 2021.
- [24] F. Mizutani, T. Ushio, E. Yoshikawa, S. Shimamura, H. Kikuchi, M. Wada, S. Satoh, and T. Iguchi, "Fast-scanning phased-array weather radar with angular imaging technique," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 5, pp. 2664–2673, May 2018.
- [25] B. Isom, R. Palmer, R. Kelley, J. Meier, D. Bodine, M. Yeary, B.-L. Cheong, Y. Zhang, T.-Y. Yu, and M. I. Biggerstaff, "The atmospheric imaging radar: Simultaneous volumetric observations using a phased array weather radar," *J. Atmos. Ocean. Technol.*, vol. 30, no. 4, pp. 655–675, Apr. 2013.
- [26] R. Jenatton, J.-Y. Audibert, and F. Bach, "Structured variable selection with sparsity-inducing norms," *J. Mach. Learn. Res.*, vol. 12, pp. 2777–2824, 2011.
- [27] X. Chen, Q. Lin, S. Kim, J. G. Carbonell, and E. P. Xing, "Smoothing proximal gradient method for general structured sparse regression," *Ann. Appl. Stat.*, vol. 6, no. 2, pp. 719–752, 2012.
- [28] İ. Bayram and Ö. D. Akyıldız, "Primal-dual algorithms for audio decomposition using mixed norms," *Signal, Image Video Process.*, vol. 8, no. 1, pp. 95–110, 2014.
- [29] P.-Y. Chen and I. W. Selesnick, "Translation-invariant shrinkage/thresholding of group sparse signals," *Signal Process.*, vol. 94, pp. 476–489, 2014.
- [30] G. Peyré and J. Fadili, "Group sparsity with overlapping partition functions," in *Proc. Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2011, pp. 303–307.
- [31] L. Yuan, J. Liu, and J. Ye, "Efficient methods for overlapping group lasso," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2011, pp. 352–360.
- [32] L. Jacob, G. Obozinski, and J.-P. Vert, "Group lasso with overlap and graph lasso," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jun. 2009, pp. 433–440.
- [33] G. Obozinski, L. Jacob, and J.-P. Vert, "Group lasso with overlaps: The latent group lasso approach," preprint arXiv: 1110.0413, 2011.
- [34] J. Huang, T. Zhang, and D. Metaxas, "Learning with structured sparsity," *J. Mach. Learn. Res.*, vol. 12, no. 11, 2011.
- [35] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Trans. Inf. Theory*, vol. 56, no. 4, pp. 1982–2001, Apr. 2010.
- [36] J. Fang, Y. Shen, H. Li, and P. Wang, "Pattern-coupled sparse Bayesian learning for recovery of block-sparse signals," *IEEE Trans. Signal Process.*, vol. 63, no. 2, pp. 360–372, Jan. 2015.
- [37] L. Wang, L. Zhao, S. Rahardja, and G. Bi, "Alternative to extended block sparse Bayesian learning and its relation to pattern-coupled sparse Bayesian learning," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2759–2771, May 2018.
- [38] A. Chambolle and T. Pock, "A first-order primal-dual algorithm for convex problems with applications to imaging," *J. Math. Imaging Vis.*, vol. 40, no. 1, pp. 120–145, 2011.
- [39] E. Esser, X. Zhang, and T. F. Chan, "A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science," *SIAM J. Imaging Sci.*, vol. 3, no. 4, pp. 1015–1046, 2010.
- [40] I. Loris and C. Verhoeven, "On a generalization of the iterative soft-thresholding algorithm for the case of non-separable penalty," *Inverse Problems*, vol. 27, no. 12, Nov. 2011, 15 pages.
- [41] P. Chen, J. Huang, and X. Zhang, "A primal-dual fixed point algorithm for convex separable minimization with applications to image restoration," *Inverse Problems*, vol. 29, no. 2, Jan. 2013, 33 pages.
- [42] M. Yamagishi and I. Yamada, "Nonexpansiveness of a linearized augmented Lagrangian operator for hierarchical convex optimization," *Inverse Problems*, vol. 33, no. 4, Mar. 2017, 35 pages.
- [43] J. Yang and X. Yuan, "Linearized augmented Lagrangian and alternating direction methods for nuclear norm minimization," *Math. Comput.*, vol. 82, no. 281, pp. 301–329, 2013.
- [44] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite element approximation," *Comput. Math. Appl.*, vol. 2, no. 1, pp. 17–40, 1976.
- [45] J. Eckstein and D. P. Bertsekas, "On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Math. Program.*, vol. 55, no. 1, pp. 293–318, Apr. 1992.
- [46] L. Condat, D. Kitahara, A. Contreras, and A. Hirabayashi, "Proximal splitting algorithms: A tour of recent advances, with new twists," preprint arXiv:1912.00137, 2020.
- [47] P. L. Combettes and C. L. Müller, "Perspective functions: Proximal calculus and applications in high-dimensional statistics," *J. Math. Anal. Appl.*, vol. 457, no. 2, pp. 1283–1306, 2018.
- [48] —, "Perspective maximum likelihood-type estimation via proximal decomposition," *Electron. J. Statist.*, vol. 14, no. 1, pp. 207–238, 2020.
- [49] H. Kuroda, D. Kitahara, and A. Hirabayashi, "A convex penalty for block-sparse signals with unknown structures," to appear in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Jun. 2021, 5 pages.
- [50] C. A. Micchelli, J. M. Morales, and M. Pontil, "Regularizers for structured sparsity," *Adv. Comput. Math.*, vol. 38, no. 3, pp. 455–489, 2013.
- [51] F. Lara, "A note on "Regularizers for structured sparsity,"" *Adv. Comput. Math.*, vol. 44, no. 4, pp. 1321–1323, 2018.
- [52] H. H. Bauschke and P. L. Combettes, *Convex analysis and monotone operator theory in Hilbert spaces*, 2nd ed. Springer, 2011.
- [53] L. Condat, "Fast projection onto the simplex and the ℓ_1 ball," *Math. Program.*, vol. 158, no. 1, pp. 575–585, 2016.
- [54] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Proc. ISCA Speech Synthesis Workshop*, 2004, pp. 223–224.
- [55] P. L. Combettes, "Perspective functions: Properties, constructions, and examples," *Set-Valued and Variational Analysis*, vol. 26, no. 2, pp. 247–264, 2018.

Hiroki Kuroda received the B.E. degree in computer science, and the M.E. and the Ph.D. degrees in information and communications engineering from the Tokyo Institute of Technology, Tokyo, Japan, in 2013, 2015 and 2019, respectively. From April 2019 to March 2020, he was a postdoctoral researcher with the National Institute of Advanced Industrial Science and Technology, Tsukuba, Japan. Since April 2020, he has been an Assistant Professor with the College of Information Science and Engineering, Ritsumeikan University, Shiga, Japan. His current research interests are in signal processing and its applications, sparse modeling, and optimization theory. He received the Seiichi Tejima Doctoral Dissertation Award from the Tokyo Institute of Technology in 2020.

Daichi Kitahara received the B.E. degree in computer science in 2012, and the M.E. and the Ph.D. degrees in communications and computer engineering in 2014 and 2017 from the Tokyo Institute of Technology, Tokyo, Japan, respectively. From April 2014 to March 2017, he was a recipient of the Research Fellowship of the Japan Society for the Promotion of Science. Since April 2017, he has been an Assistant Professor with the College of Information Science and Engineering, Ritsumeikan University, Shiga, Japan. His current research interests include

signal processing and its applications, inverse problem, optimization theory, and multivariate spline theory.