Penalty Dual Decomposition Method for Nonsmooth Nonconvex Optimization—Part I: Algorithms and Convergence Analysis

Qingjiang Shi and Mingyi Hong

***Abstract*—Many contemporary signal processing, machine learning and wireless communication applications can be formulated as nonconvex nonsmooth optimization problems. Often there is a lack of efficient algorithms for these problems, especially when the optimization variables are nonlinearly coupled in some nonconvex constraints. In this work, we propose an algorithm named penalty dual decomposition (PDD) for these difficult problems and discuss its various applications. The PDD is a double-loop iterative algorithm. Its inner iteration is used to inexactly solve a nonconvex nonsmooth augmented Lagrangian problem via block-coordinate-descent-type methods, while its outer iteration updates the dual variables and/or a penalty parameter. In Part I of this work, we describe the PDD algorithm and establish its convergence to KKT solutions. In Part II we evaluate the performance of PDD by customizing it to three applications arising from signal processing and wireless communications.**

***Index Terms*—Penalty method, dual decomposition, BSUM, KKT, augmented Lagrangian, nonconvex optimization.**

I. Introduction

MANY important engineering problems arising from signal processing, wireless communications and machine learning can be modeled as nonconvex nonsmooth optimization problems. These problems are generally difficult to solve, especially when the optimization variables are nonlinearly coupled in some (possibly nonconvex) constraints. This two-part paper provides an algorithmic framework that can fully exploit the problem structure, for optimizing a nonconvex nonsmooth function subject to nonconvex but continuously differentiable *coupling* constraints.

Manuscript received September 26, 2019; revised May 1, 2020 and June 8, 2020; accepted June 8, 2020. Date of publication June 18, 2020; date of current version July 24, 2020. The associate editor coordinating the review of this manuscript and approving it for publication was Nicolas Gillis. The work of Qingjiang Shi was supported in part by the National Key Research and Development Project under Grant 2017YFE0119300, and in part by the NSFC under Grants 61671411, 61731018, and U1709219. The work of Mingyi Hong was supported in part by the National Science Foundation under Grants CIF-1910385 and CMMI-172775 and in part by Army Research Office under Grant W911NF-19-1-0247. Part of this paper has been presented in IEEE International Conference on Acoustics, Speech and Signal Processing, New Orleans, LA, USA, 5–9 Mar. 2017 [1]. *(Corresponding author: Mingyi Hong.)*

Qingjiang Shi is with the School of Software Engineering, Tongji University, Shanghai 201804, China, and also with the Shenzhen Research Institute of Big Data, Shenzhen 518172, China (e-mail: shiqj@tongji.edu.cn).

Mingyi Hong is with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55454, USA (e-mail: mhong@umn.edu).

Digital Object Identifier 10.1109/TSP.2020.3001906

Nonconvex problems with constraints that couple a few design variables often arise in contemporary applications. For example, in the joint source-relay design of many multiple-input-multiple-output (MIMO) relay systems [2]–[4], the relay power constraints often couple the source or relay precoders in a *bi-quadratic* manner, meaning that fixing one variable (i.e., the source precoders), then the constraint function becomes quadratic with respect to the other variable (i.e., the relay precoders). Another popular example arises in the family of quality-of-service (QoS)-constrained power minimization problems, in which the signal-to-interference-plus-noise ratio (SINR) functions or the (weighted) mean-square-error (MSE) functions are also quadratic in the beamformers [5]–[9]. In problems such as dictionary learning [10], [11], nonnegative matrix factorization [12]–[14], and geometry-based blind source separation [15], the variables are coupled in a *bi-linear* manner by certain equality constraints. Other problems with nonlinear and nonconvex constraints coupling can be found in [16]–[19]. Such constraint coupling makes developing efficient low-complexity, and parallel algorithms a very challenging task.

Generally speaking, when designing algorithms for an engineering problem, it is important to exploit, as much as possible, its fundamental structures in order to improve solution quality and/or speed. For problems with multi-blocks and coupling constraint, it is the *block structure* that often gets exploited. One such popular method is the alternating optimization (AO) method, which replaces difficult joint optimization over all variables with a sequence of easier optimization over individual (block) variable. For instance, for two-hop relay broadcast channel, the authors of [9] considered joint source-relay design for achieving power minimization subject to SINR constraints, where the source precoder and relay precoder are coupled with each other. Observing that the power minimization problem is convex with respect to the source precoder or the relay precoder, the work [9] used the AO method to address the power minimization problem. Similar to [9], the work [2] also used the AO method to address the joint source-relay design to achieve sum rate maximization in a MIMO relay interference channel. However, the AO method can only provide feasible solutions in the coupling constraint case and cannot guarantee convergence to stationary solutions (or KKT points) unless the objective has some special structure; see for example [8]. In particular, the AO method easily gets trapped in some unexpected points in the equality coupling constraint case; see [20] for illustrative examples. To deal with a special class of equality coupling constraint $\mathbf{Z} = \mathbf{XY}$ (where $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}$ are all matrix variables) that arises from relay network design, the work [21] first transformed the equality coupling constraint into two matrix inequalities and then used

concave-convex procedure to solve the resulting problem. However, this method is not only computationally expensive, but also lacks convergence guarantee to stationary solutions.

Another popular approach that can deal with the coupled constraint, especially the equality coupling constraints, is the penalty method [22]. The basic idea of penalty method is to move the difficult constraints to the objective function as a penalty term, so that infeasible points can get relatively high cost compared with the feasible ones. For example, in [17], Kuang *et al*. used penalty method to approximate the solution of the symmetric nonnegative matrix factorization problem. In [3], Shi *et al*. used penalty method to solve the joint source-relay design problem for full-duplex MIMO relay systems. The work [23] showed that penalty method can be applied to solve the rank minimization problem, an important class of problems that often arises from signal processing. However, penalty methods could be very inefficient, because it usually requires that certain penalty parameter goes to infinity, resulting in ill-conditioning for its subproblems. Augmented Lagrangian (AL) methods [24], [25] were proposed to overcome the limitations of penalty methods by introducing an additional dual-related term. In the AL methods, a sequence of *AL subproblems* (i.e., the problems of minimization of the augmented Lagrangian) need to be exactly or approximately solved [22]. When the AL subproblems are easily solvable, the AL methods are attractive as they can be often easily implemented (often in a matrix-free manner) [26] and have at least local convergence guarantees under relatively mild assumptions [27], [28]. However, the AL subproblems are generally hard to solve especially when they have complicated constraints. Further, the AL method generally cannot deal with nonsmooth penalty function in the objective.

As an important variant of augmented Lagrangian method, alternating direction method of multipliers (ADMM) has recently regained popularity due to its applicability in many large-scale problems [29]. Differently from the standard AL method, a single iteration of *block coordinate descent* (BCD) or AO is used to *approximately* minimize the augmented Lagrangian at each iteration of ADMM. That is, the *AL subproblem* is minimized only *approximately*, by solving a sequence of smaller, and potentially easier, subproblems generated by the block coordinate decomposition. Indeed, it is the idea of combining block decomposition and approximate AL subproblem minimization that enables the ADMM to fully exploit the block structure of the problem. Although the ADMM has been widely used in the areas of signal processing [15], [30], [31], wireless communication [5], [7], [32], [33], and machine learning [14], [29], [34], [35], they are primarily developed for convex problems with linearly coupling constraints. Generally speaking, ADMM does not converge for nonconvex problems, except for a few special cases; see recent developments in [36]–[39] and the references therein.

Other relevant works in the literatures include [40]–[45]. In [40], the authors proposed to use sequential quadratic programming (SQP) based method to optimize a nonsmooth problem with both equality and inequality constraints. In particular, the proposed algorithm is based on solving certain smooth version of the problem, by using adaptive smoothing parameters, and by utilizing the state-of-the-art SQP solvers. This is a very general scheme that can deal with a fairly wide class of problems. However, the proposed algorithm requires the computation of Hessian matrices, or some approximation of them, which could be expensive to obtain in practice. Further, it is not clear how to deal with block structures. The work [41] dealt with nonconvex nonsmooth optimization variables assuming that, at each step certain proximity operator can be evaluated exactly. In [42], the authors dealt with difference of convex (dc) problems, where both the objective and constraints can take the nonconvex dc structure. The authors developed feasible and infeasible algorithms for these problems, and discussed various extensions such as distributed schemes for problems with finite-sum structures. The authors of [45] proposed distributed and/or parallel algorithms to deal with nonsmooth objective function and difficult constraints. The algorithm generates feasible iterates by solving certain strongly convex subproblem with inner convex approximation of the original feasible set. The authors of [44] considered problems where both the objective function and the constraints can be represented by certain LC^1 functions. The authors proposed methods based on the idea of combining SQP methods and successive convex approximation with appropriately diminishing stepsizes. We note that these algorithms and the associated analysis can be extended to the multiple-block setting, but it is not trivial to include constraints that couple all the variables, while still being able to fully utilize the block structure of the problem.

In this work, we propose an optimization framework named penalty dual decomposition (PDD), which integrates the penalty method, the AL method and the ADMM method. Specifically, our framework is a double-loop algorithm where the inner loop *approximately* solves the AL subproblem, while the outer loop updates the dual variable and/or a certain penalty parameter. To exploit the problem structure as fully as possible, a block-coordinate-descent (BCD) based method is used to approximately solve the AL subproblem. In Part I of the paper, we first introduce the notion of generalized gradient [47], [48] and provide conditions under which a KKT point exists. We then rigorously prove the convergence of the PDD to KKT points under some constraint qualification (CQ) condition. Furthermore, to address AL problems with nonconvex constraints using BCD-type algorithms, we propose stochastic BSUM algorithm and prove its convergence. Our proof is critically dependent on the randomization introduced to the original BSUM algorithm, which provides the algorithm with good convergence behavior even in the presence of *nonconvex constraints*. In the second part of this paper, we customize the PDD to several engineering problems arising from signal processing and wireless communications. Our numerical results show that PDD outperforms a number of state-of-the-art algorithms, therefore validating the effectiveness of the PDD method in solving nonconvex nonsmooth problem with coupling constraints.

Notations: Throughout this paper, we use uppercase bold letters for matrices, lowercase bold letters for column vectors, and regular letters for scalars (unless otherwise specified). The notations $\mathbb{R}^n$, $\mathbb{R}^n_+$ and $\mathbb{R}^n_-$ denote the n-dimensional space of real number, nonnegative real number, nonpositive real number, respectively. For a vector $\boldsymbol{x}$, $\|\boldsymbol{x}\|$ and $\|\boldsymbol{x}\|_\infty$ denote Euclidean norm and element-wise infinity norm, respectively. $B_\delta(\boldsymbol{x}_0)$ denotes a Euclidean ball centered at $\boldsymbol{x}_0$ with radius δ. For a scalar function $f(\cdot)$, $f'(\cdot)$ and $\nabla f(\cdot)$ respectively denote its derivative and gradient with respect to its argument. For a multivariate function $f(\boldsymbol{x},\boldsymbol{y})$, $\nabla_{\boldsymbol{x}} f(\boldsymbol{x},\boldsymbol{y})$ denotes its gradient with respect to $\boldsymbol{x}$. For vector functions $\boldsymbol{g}(\boldsymbol{x})$ and $\boldsymbol{h}(\boldsymbol{x},\boldsymbol{y})$, $\nabla \boldsymbol{g}(\boldsymbol{x})$ denotes the Jacobian matrix of $\boldsymbol{g}(\boldsymbol{x})$ and $\nabla_{\boldsymbol{x}} \boldsymbol{h}(\boldsymbol{x},\boldsymbol{y})$ denotes the Jacobian matrix of $\boldsymbol{h}(\boldsymbol{x},\boldsymbol{y})$ with respect to $\boldsymbol{x}$. For a convex function $\hbar(\boldsymbol{x})$, $\partial \hbar(\boldsymbol{x})$ denotes its subdifferential. $T_{\mathcal{Z}}(\boldsymbol{z})$ and $N_{\mathcal{Z}}(\boldsymbol{z})$ denotes the tangent cone and normal cone [49] of the set $\mathcal{Z}$ at point $\boldsymbol{z}$,

respectively, and these definitions are formally given in Appendix A. The notation $int\mathcal{Z}$ denotes the interior of the set $\mathcal{Z}$ while the notation $(\boldsymbol{x}_i)_i$ denotes a vector stacked by all subvectors $\boldsymbol{x}_i$'s.

II. Nonconvex Nonsmooth Optimization and KKT Characterization

Consider the following problem

$$\begin{array}{ll} \min_{\boldsymbol{x}\in\mathcal{X},\boldsymbol{y}} & F(\boldsymbol{x},\boldsymbol{y}) \triangleq f(\boldsymbol{x},\boldsymbol{y}) + \sum_{j=1}^{n_y} \tilde{\phi}(\boldsymbol{y}_j) \\ \text{s.t.} & \boldsymbol{h}(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{0}, \\ & \boldsymbol{g}_i(\boldsymbol{x}_i) \leq 0,\ i = 1,2,\ldots,n \end{array} \tag{P}$$

where

- the feasible set $\mathcal{X}$ is the Cartesian product of n *closed convex* sets: $\mathcal{X} \triangleq \Pi_{i=1}^n \mathcal{X}_i$ with $\mathcal{X}_i \subseteq \mathbb{R}^{n_i}$ and $\sum_{i=1}^n n_i = N$;
- the optimization variable $\boldsymbol{x} \in \mathbb{R}^N$ is decomposed as $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$ with $\boldsymbol{x}_i \in \mathcal{X}_i$ $i = 1,2,\ldots,n$, and $\boldsymbol{y} \in \mathbb{R}^M$ is decomposed as $\boldsymbol{y}_j \in \mathbb{R}^{m_j}$, $j = 1,2,\ldots,m$, with $\sum_{j=1}^m m_j = M$;
- $f(\boldsymbol{x},\boldsymbol{y})$ is a scalar *continuously differentiable* function; $\tilde{\phi}(\boldsymbol{y}_j)$ is a composite function in the form of $\phi_j(s_j(\boldsymbol{y}_j))$, with each $s_j(\boldsymbol{y}_j)$ being a *convex* but possibly nondifferentiable function while $\phi_j(x)$ being a *nondecreasing* and *continuously differentiable* function;
- for each i, $\boldsymbol{g}_i(\boldsymbol{x}_i) \in \mathbb{R}^{q_i}$ is a vector of q_i *continuously differentiable* functions, and we define $q \triangleq \sum_{i=1}^n q_i$;
- $\boldsymbol{h}(\boldsymbol{x},\boldsymbol{y}) \in \mathbb{R}^p$ is a vector of p *continuously differentiable* functions.
- The feasible set of problem (P), given below, is nonempty

$$\begin{aligned} \mathcal{Z} \triangleq \{(\boldsymbol{x},\boldsymbol{y}) \in \mathbb{R}^N \times \mathbb{R}^M \mid \boldsymbol{x} \in \mathcal{X}, \\ \boldsymbol{h}(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{0},\ \boldsymbol{g}_i(\boldsymbol{x}_i) \leq 0,\ \forall\, i\}. \end{aligned} \tag{1}$$

In the above problem, the constraint coupling is mainly represented by the equality constraint $\boldsymbol{h}(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{0}$, while for each i, the inequality constraint $g_i(\boldsymbol{x}_i) \leq 0$ represents the possibly nonconvex constraints for $\boldsymbol{x}_i$. Note that we do not explicitly write down the constraint set for the block $\boldsymbol{y}$ for ease of exposition. However, the constraint on $\boldsymbol{y}$ can be similarly treated as that for $\boldsymbol{x}$.

Further, we remark that the term $\sum_{j=1}^{n_y} \tilde{\phi}(\boldsymbol{y}_j)$ represents the nonsmooth part of the objective function. Typically, the composite function $\tilde{\phi}(\boldsymbol{y}_j) = \phi_j(s_j(\boldsymbol{y}_j))$ can take the form of sparsity promoting functions. For instance, in the case of log-based sparsity promotion function, we have $\phi_j(z) = \lambda \log(1 + \frac{z}{\epsilon})$ and $s_j(\boldsymbol{y}_j) = \|\boldsymbol{y}_j\|$, and thus $\tilde{\phi}(\boldsymbol{y}_j) = \lambda \log(1 + \frac{\|\boldsymbol{y}_j\|}{\epsilon})$. Here λ and ϵ are two positive sparsity-related control parameters. We refer readers to [50, Table I] for more examples of sparsity promotion functions, e.g, lasso penalty function, SCAD penalty function, etc. Since the term $\sum_{j=1}^{n_y} \tilde{\phi}(\boldsymbol{y}_j)$ could be neither convex nor differentiable, we need to use generalized gradient [48] to characterize the first-order optimality condition, which is the main topic of the following two subsections.

A. Preliminaries

First, we introduce the definition of the local Lipschitz continuity and the locally Lipschitz function.

Definition 2.1 (Local Lipschitz continuity [47], [48]): A function $\hbar(\boldsymbol{x})$ is Lipschitz near a point $\boldsymbol{x}_0 \in int\ dom\hbar$ if there exists $K \geq 0$ such that $|\hbar(\boldsymbol{x}) - \hbar(\boldsymbol{x}')| \leq K\|\boldsymbol{x} - \boldsymbol{x}'\|, \forall \boldsymbol{x}, \boldsymbol{x}' \in B_\delta(\boldsymbol{x}_0)$ where $\delta > 0$ is sufficiently small so as to have $B_\delta(\boldsymbol{x}_0) \subset$ dom$\hbar$. A locally Lipschitz function is a function that is Lipschitz near every point in $int\ dom\hbar$.

Two important special cases of locally Lipschitz functions are continuously differentiable functions and convex functions [47], [48]. Combining this with the boundedness of continuous functions over a compact set, it can be shown that each $\phi_j(s_j(\boldsymbol{y}_j))$ is locally Lipschitz. As a result, the objective function of problem (P) is locally Lipschtiz as well. This fact will be used in establishing the optimality condition.

Next, we introduce the concept of generalized gradient which is defined for nonconvex nondifferentiable functions.

Definition 2.2 (Generalized gradient [46], [47], [48]): Clarke's generalized directional derivative of $\hbar(\boldsymbol{x})$ at $\boldsymbol{x}_0$ in the direction $\boldsymbol{d}$, denoted as $\hbar^o(\boldsymbol{x}_0;\boldsymbol{d})$, is defined by

$$\begin{aligned} \hbar^o(\boldsymbol{x}_0;\boldsymbol{d}) &= \limsup_{\substack{\boldsymbol{u}\to\boldsymbol{0} \\ \lambda\downarrow 0}} \frac{\hbar(\boldsymbol{x}_0 + \boldsymbol{u} + \lambda\boldsymbol{d}) - \hbar(\boldsymbol{x}_0 + \boldsymbol{u})}{\lambda} \\ &= \lim_{\delta\downarrow 0} \sup_{\boldsymbol{u}\in\mathcal{B}_\delta(\boldsymbol{0}),\lambda\in(0,\delta)} \frac{\hbar(\boldsymbol{x}_0 + \boldsymbol{u} + \lambda\boldsymbol{d}) - \hbar(\boldsymbol{x}_0 + \boldsymbol{u})}{\lambda} \end{aligned} \tag{2}$$

Also, Clarke's generalized subdifferential of $\hbar$ at $\boldsymbol{x}_0$ is defined by

$$\bar{\partial}\hbar(\boldsymbol{x}_0) = \{\boldsymbol{\xi} : \hbar^o(\boldsymbol{x}_0;\boldsymbol{d}) \geq \boldsymbol{\xi}^T\boldsymbol{d}, \forall \boldsymbol{d}\}.$$

For any $\boldsymbol{\xi} \in \bar{\partial}\hbar(\boldsymbol{x}_0)$, we refer to it as generalized gradient of $\hbar$ at $\boldsymbol{x}_0$.

As compared to the conventional directional derivative [51], the generalized directional derivative in (2) is defined with a new "base point", i.e., $\boldsymbol{x}_0 + \boldsymbol{u}$, for taking the difference. Moreover, due to the *supremum* taken before the limit, it is shown in [47, Lemma 2.6] [52] that the generalized directional derivative $\hbar^o(\boldsymbol{x}_0;\boldsymbol{d})$ is convex with respect to $\boldsymbol{d}$ even when $\hbar$ itself is nonconvex. Hence, by convex analysis, we have Theorem 2.1, whose proof is relegated to Appendix B.

Theorem 2.1: Let $\hbar(\boldsymbol{x})$ be Lipschitz near $\boldsymbol{x}_0$ with local Lipschitz constant K. Then the following holds:

1) $\hbar^o(\boldsymbol{x}_0;\boldsymbol{0}) = 0$;
2) $\bar{\partial}\hbar(\boldsymbol{x}_0)$ is not empty and is a compact set;
3) $\|\boldsymbol{\xi}\| \leq K, \forall\, \boldsymbol{\xi} \in \bar{\partial}\hbar(\boldsymbol{x}_0)$;
4) $\hbar^o(\boldsymbol{x}_0;\boldsymbol{d}) = \max_{\boldsymbol{\xi}\in\bar{\partial}\hbar(\boldsymbol{x}_0)} \boldsymbol{\xi}^T\boldsymbol{d}, \forall \boldsymbol{d}$.

Furthermore, the following theorem establishes the connections between the generalized gradient and two classical concepts: the ordinary gradient and the subdifferential of convex analysis. The proof can be found in [52], [47, Prop. 2.7 & 2.8].

Theorem 2.2: The following holds

1) If $\hbar(\boldsymbol{x})$ is continuously differentiable at $\boldsymbol{x}_0$, then $\bar{\partial}\hbar(\boldsymbol{x}_0) = \{\nabla\hbar(\boldsymbol{x}_0)\}$.
2) If $\hbar(\boldsymbol{x})$ is a convex function, then the Clarke's generalized gradient coincides with the subdifferential of $\hbar$, i.e., $\bar{\partial}\hbar(\boldsymbol{x}) = \partial\hbar(\boldsymbol{x})$.

Theorem 2.2 implies $\bar{\partial}\tilde{\phi}(\boldsymbol{y}_j) = \nabla\phi(s_j(\boldsymbol{y}_j))\partial s_j(\boldsymbol{y}_j), \forall j$. Moreover, considering that both convex functions and continuously differentiable functions are locally Lipschtz, according to the result of the above two theorems, we can deduce that $\nabla\hbar(\boldsymbol{x}_0)$

is bounded if $\hbar(\boldsymbol{x})$ is continuously differentiable at $\boldsymbol{x}_0$, and that any subgradient of $\hbar(\boldsymbol{x})$ is also bounded if $\hbar(\boldsymbol{x})$ is convex.

B. KKT Characterization Under Robinson's Condition

To describe optimality condition for nonlinear optimization, it is often required to assume that the problem satisfies some regularity conditions [22], [49]. In this paper, we use *Robinson's condition*, whose precise definition is given below. Note that we have provided in Appendix A some basics for understanding Robinson's condition.

Definition 2.3 (Robinson's condition [22], [49]): Robinson's condition is satisfied at $\hat{\boldsymbol{z}} \triangleq (\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})$ for problem (P), if the following holds [49, Chap. 3]

$$\left\{ \begin{pmatrix} \nabla \boldsymbol{h}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) \boldsymbol{d}_z \\ \nabla \boldsymbol{g}_1(\hat{\boldsymbol{x}}_i) \boldsymbol{d}_{x_1} - \boldsymbol{v}_1 \\ \vdots \\ \nabla \boldsymbol{g}_n(\hat{\boldsymbol{x}}_n) \boldsymbol{d}_{x_n} - \boldsymbol{v}_n \end{pmatrix} \middle| \begin{array}{l} \boldsymbol{d}_x \in T_{\mathcal{X}}(\hat{\boldsymbol{x}}), \boldsymbol{d}_y \in \mathbb{R}^M, \\ \boldsymbol{v} \in \mathbb{R}^q, \boldsymbol{v}_{i,\ell} \leq 0, \\ \forall \ell \in I_i(\hat{\boldsymbol{x}}_i),\ \forall\, i \end{array} \right\} = \mathbb{R}^p \times \mathbb{R}^q \tag{3}$$

where $\boldsymbol{d}_z \triangleq (\boldsymbol{d}_x, \boldsymbol{d}_y)$, $\boldsymbol{v} \triangleq (\boldsymbol{v}_i)_i$, $\boldsymbol{v}_{i,\ell}$ denotes the ℓ-th element of $\boldsymbol{v}_i$, $I_i(\hat{\boldsymbol{x}}_i)$ is the i-th index set of active inequality constraints at $\hat{\boldsymbol{x}}$, i.e.,

$$I_i(\hat{\boldsymbol{x}}_i) \triangleq \{\ell | \boldsymbol{g}_{i,\ell}(\hat{\boldsymbol{x}}_i) = 0,\ 0 \leq \ell \leq q_i\},$$

where $\boldsymbol{g}_{i,\ell}(\hat{\boldsymbol{x}}_i)$ denotes the ℓ-th component function of $\boldsymbol{g}_i(\hat{\boldsymbol{x}}_i)$.

According to Theorem A.2 in Appendix A, when the system of constraints of problem (P) satisfies Robinson's condition at point $\hat{\boldsymbol{z}} \triangleq (\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})$, the tangent cone to the feasible set $\mathcal{Z}$ of problem (P) exists and takes the following form [49, Chap. 3]

$$T_{\mathcal{Z}}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) = \Big\{ \boldsymbol{d}_z \triangleq (\boldsymbol{d}_x, \boldsymbol{d}_y) | \boldsymbol{d}_x \in T_{\mathcal{X}}(\hat{\boldsymbol{x}}), \boldsymbol{d}_y \in \mathbb{R}^M, \nabla \boldsymbol{h}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) \boldsymbol{d}_z = 0, \nabla \boldsymbol{g}_{i,\ell}(\hat{\boldsymbol{x}}_i)^T \boldsymbol{d}_{x_i} \leq 0, \ell \in I_i(\hat{\boldsymbol{x}}_i),\ \forall\, i \Big\} \tag{4}$$

where $\boldsymbol{d}_{x_i} \in \mathbb{R}^{n_i}$ is the i-th subvector of $\boldsymbol{d}_x$ with $\boldsymbol{d}_x = (\boldsymbol{d}_{x_i})_i$.

Now we are ready to establish the KKT condition for problem (P) in the following theorem. As shown in Appendix B, our proof for this theorem is extended from Theorem 3.25 in [49] which deals with the case where the objective function is differentiable. Here we deal with the possibly nonconvex and nondifferentiable objective function of problem (P) by using the notion of generalized directional derivative/gradient.

Theorem 2.3: Let $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})$ be a local minimum of problem (P). Assume that Robinson's condition holds for problem (P) at $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})$. Then there exist multipliers $\hat{\boldsymbol{\mu}} \in \mathbb{R}^p$ and $\hat{\boldsymbol{v}}_i \in \mathbb{R}^{q_i}$, $i = 1, 2, \ldots, n$, such that the following generalized KKT system is satisfied

$$\big(\nabla_{\boldsymbol{x}_i} f(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) + \nabla_{\boldsymbol{x}_i} \boldsymbol{h}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})^T \hat{\boldsymbol{\mu}} + \nabla_{\boldsymbol{x}_i} \boldsymbol{g}_i(\hat{\boldsymbol{x}}_i)^T \hat{\boldsymbol{v}}_i\big)^T \times (\boldsymbol{x}_i - \hat{\boldsymbol{x}}_i) \geq 0,\ \forall\, \boldsymbol{x}_i \in \mathcal{X}_i, \tag{5a}$$

$$0 \in \bar{\partial} \tilde{\phi}(\boldsymbol{y}_j) + \nabla_{\boldsymbol{y}_j} f(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) + \nabla_{\boldsymbol{y}_j} \boldsymbol{h}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})^T \hat{\boldsymbol{\mu}},\ \forall\, j, \tag{5b}$$

$$(\hat{\boldsymbol{v}}_i)^T \boldsymbol{g}_i(\hat{\boldsymbol{x}}_i) = 0,\ \forall i, \tag{5c}$$

$$\boldsymbol{g}_i(\hat{\boldsymbol{x}}_i) \leq 0,\ \forall\, i, \tag{5d}$$

$$\hat{\boldsymbol{v}}_i \geq 0,\ \forall\, i, \tag{5e}$$

$$\boldsymbol{h}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}) = 0. \tag{5f}$$

TABLE I
ALGORITHM 1: PDD METHOD FOR PROBLEM (P)

0. initialize $\boldsymbol{z}^0 = (\boldsymbol{x}^0, \boldsymbol{y}^0)$, $\varrho_0 > 0$, $\boldsymbol{\lambda}_0$, and set $k = 1$
   pick two sequences $\{\eta_k > 0\}$, $\{\epsilon_k > 0\}$
1. **repeat**
2. $\quad \{\boldsymbol{z}^k, \boldsymbol{v}^k\} = optimize(P_{\varrho_k, \boldsymbol{\lambda}_k}, \boldsymbol{z}^{k-1}, \epsilon_k)$
3. $\quad$ if $\|\boldsymbol{h}(\boldsymbol{z}^k)\|_\infty \leq \eta_k$ // *case 1*—AL method
4. $\quad\quad \boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k + \frac{1}{\varrho_k} \boldsymbol{h}(\boldsymbol{z}^k)$
5. $\quad\quad \varrho_{k+1} = \varrho_k$
6. $\quad$ **else** // *case 2*—penalty method
7. $\quad\quad \boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k$
8. $\quad\quad$ update ϱ_{k+1} by decreasing ϱ_k
9. $\quad$ **end**
10. $\quad k = k + 1$
11. **until** some termination criterion is met

Robinson's condition is more general than the well-known Mangasarian-Fromovitz constraint qualification (MFCQ) condition. Their relation will be discussed in Section V.A. Here, it is worth mentioning that, if Robinson's condition is replaced with MFCQ condition in Theorem 2.3, the above result is readily implied by the standard KKT conditions [[46] Theorem 6.1.1].

III. PDD METHOD AND ITS CONVERGENCE

Besides the nonconvexity and nondifferentiability, the variable coupling introduced by the equality constraint $\boldsymbol{h}(\boldsymbol{x}, \boldsymbol{y}) = 0$ further complicates problem (P). Without such a coupling constraint, efficient block decomposition algorithms such as BCD, BSUM or FLEXA [53] can be applied to decompose problem (P) into a sequence of small-scale problems. Unfortunately, these block decomposition methods can fail to reach any interesting solution in the presence of coupling constraint [20]. In this section we propose the PDD algorithm that relaxes the difficult coupling constraints (by using Lagrangian relaxation), performs block decomposition over the resulting augmented Lagrangian function, and utilizes appropriate penalty parameters to eventually enforce the relaxed equality constraint.

A. The Basic PDD Method

To introduce the algorithm, denote by $\mathcal{L}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\lambda})$ the augmented Lagrange function with penalty parameter ϱ and dual variable $\boldsymbol{\lambda}$ corresponding to the coupling constraint $\boldsymbol{h}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{0}$. Further, let us define an *AL problem* $(P_{\varrho, \boldsymbol{\lambda}})$ as follows

$$(P_{\varrho, \boldsymbol{\lambda}}) \min_{\boldsymbol{x}_i \in \widetilde{\mathcal{X}}_i, \boldsymbol{y}} \Big\{ \mathcal{L}(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\lambda}) \triangleq f(\boldsymbol{x}, \boldsymbol{y}) + \sum_{j=1}^{n_y} \phi_j(s_j(\boldsymbol{y}_j)) + \boldsymbol{\lambda}^T \boldsymbol{h}(\boldsymbol{x}, \boldsymbol{y}) + \frac{1}{2\varrho} \|\boldsymbol{h}(\boldsymbol{x}, \boldsymbol{y})\|^2 \Big\} \tag{6}$$

where $\widetilde{\mathcal{X}}_i \triangleq \{\boldsymbol{x}_i \mid \boldsymbol{g}_i(\boldsymbol{x}_i) \leq 0,\ \boldsymbol{x}_i \in \mathcal{X}_i\}$.

The basic PDD method, presented in Table I, is a double-loop iterative algorithm, where the inner loop approximately solves the AL subproblem (6) while the outer loop updates the dual variable or the penalty parameter if necessary. In Table I, the notation 'optimize$(P_{\varrho_k, \boldsymbol{\lambda}_k}, \boldsymbol{z}^{k-1}, \epsilon_k)$' represents some optimization oracle used to iteratively solve problem $(P_{\varrho_k, \boldsymbol{\lambda}_k})$. It returns the tuple $(\boldsymbol{z}^k, \boldsymbol{v}^k)$, where $\boldsymbol{v}^k$ is the dual variable associated with the constraint $\boldsymbol{g}(\boldsymbol{x}) \leq 0$, and such a tuple *approximatley* solves $(P_{\varrho_k, \boldsymbol{\lambda}_k})$ to some accuracy ϵ_k. In particular, we require that the

output $(\boldsymbol{z}^k, \boldsymbol{v}^k)$ should satisfy the following condition (for some properly chosen sequence $\{\epsilon_k\}$)

$$\max\left(\|\boldsymbol{e}^k\|_\infty, \|\boldsymbol{\Delta}^k\|_\infty\right) \le \epsilon_k, \forall k \tag{7}$$

where $\boldsymbol{e}^k$ and $\boldsymbol{\Delta}^k$ are defined in (8) and (9), respectively. It is easy to show that, when the problem $(P_{\varrho_k,\boldsymbol{\lambda}_k})$ is solved to a KKT solution, then condition (7) should be satisfied with $\epsilon_k = 0$. Typically, to fully exploit the problem structure, one could instantiate the optimization oracle using some BCD-type algorithms, such as the classical BCD algorithm [22], or some inexact variants of BCD, such as the BSUM [54] algorithm.

Furthermore, we update the dual variable $\boldsymbol{\lambda}_k$ when the constraint violation $\|\boldsymbol{h}(\boldsymbol{z}^k)\|_\infty$ is relatively small (i.e., Step 4); otherwise we decrease the penalty parameter ϱ_k (i.e., Step 8). Therefore, the PDD method adaptively switches between the AL and the penalty method. This adaptive strategy is expected to find an appropriate penalty parameter ϱ, with which the AL method could eventually converge. In the PDD method, the parameter $\eta_k > 0$ measures the constraint violation and the parameter $\epsilon_k > 0$ controls the accuracy of the optimization oracle, with both parameters going to zero as the number of outer iterations k increases.

B. Convergence Analysis for PDD

In the following, we address the convergence issue of the PDD method. To do so, we define $\boldsymbol{e}^k$ and $\boldsymbol{\Delta}_j^k$ in (8) and (9) (see the bottom of the page), where $\boldsymbol{g}(\boldsymbol{x}) \triangleq (\boldsymbol{g}_i(\boldsymbol{x}_i))_i$. We will show that, when these two terms go to zero, the first order optimality condition of the AL problem with respect to $\boldsymbol{x}$ and $\boldsymbol{y}$ holds true. The main convergence result is presented in Theorem 3.1.

Theorem 3.1: Let $\{\boldsymbol{x}^k, \boldsymbol{y}^k, \boldsymbol{v}^k\}$ be the sequence generated by Algorithm 1 for problem (P), where $\boldsymbol{v}^k = (\boldsymbol{v}_i^k)_i$ denotes the Lagrange multipliers associated with the constraints $\boldsymbol{g}_i(\boldsymbol{x}_i) \le 0, \forall i$. The termination condition for the optimization oracle involved in Algorithm 1 is

$$\max\left(\|\boldsymbol{e}^k\|_\infty, \|\boldsymbol{\Delta}^k\|_\infty\right) \le \epsilon_k, \forall k \tag{10}$$

with $\epsilon_k, \eta_k, \varrho_k \to 0$ as $k \to \infty$. Suppose that $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ is a limit point of the sequence $\{\boldsymbol{x}^k, \boldsymbol{y}^k\}$ and at the limit point $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ the Robinson's condition holds for problem (P). Then $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ satisfies $\boldsymbol{h}(\boldsymbol{x}^*, \boldsymbol{y}^*) = \boldsymbol{0}$, and it is a KKT point of problem (P) that satisfies (5).

Proof: Our proof consists of two steps, in the first step we will utilize Robinson's condition to argue that $\{\mu^k\}$ (cf. (13)) is a bounded sequence. Then based on this result we will argue that the sequence converges to KKT points.

Step 1: First, we show that a key inequality [see (17)] holds for $\{(\boldsymbol{x}^k, \boldsymbol{y}^k)\}$. Without loss of generality, we assume that the sequence $\{(\boldsymbol{x}^k, \boldsymbol{y}^k)\}$ converges to $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ (otherwise we can restrict to a convergent subsequence of $\{(\boldsymbol{x}^k, \boldsymbol{y}^k)\}$). By noting that $\mathcal{X}$ is a closed convex set, we have $\boldsymbol{x}^* \in \mathcal{X}$. By the definition of $\boldsymbol{e}^k$ and using *projection theorem* [22, Prop. 2.1.3 (b)], we have

$$\begin{aligned}\left(\boldsymbol{x}-(\boldsymbol{x}^k+\boldsymbol{e}^k)\right)^T \big(&(\boldsymbol{x}^k - \nabla_{\boldsymbol{x}}\mathcal{L}_k(\boldsymbol{x}^k, \boldsymbol{y}^k)\\ &-\nabla \boldsymbol{g}(\boldsymbol{x}^k)^T\boldsymbol{v}^k\big) - (\boldsymbol{x}^k+\boldsymbol{e}^k)\big) \le 0, \forall \boldsymbol{x}\in\mathcal{X}, \forall k.\end{aligned} \tag{11}$$

It follows that

$$\begin{aligned}-\left(\boldsymbol{x} - (\boldsymbol{x}^k+\boldsymbol{e}^k)\right)^T \big(&\nabla_{\boldsymbol{x}}\mathcal{L}_k(\boldsymbol{x}^k, \boldsymbol{y}^k)\\ &+\nabla \boldsymbol{g}(\boldsymbol{x}^k)^T\boldsymbol{v}^k + \boldsymbol{e}^k\big) \le 0, \forall \boldsymbol{x} \in \mathcal{X}, \forall k.\end{aligned} \tag{12}$$

Let us define a "virtual" multiplier vector as

$$\boldsymbol{\mu}^k \triangleq \frac{1}{\varrho_k}\boldsymbol{h}(\boldsymbol{x}^k, \boldsymbol{y}^k) + \boldsymbol{\lambda}_k. \tag{13}$$

Then we have

$$\nabla_{\boldsymbol{x}}\mathcal{L}_k(\boldsymbol{x}^k, \boldsymbol{y}^k) = \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^k, \boldsymbol{y}^k) + \nabla_{\boldsymbol{x}}\boldsymbol{h}(\boldsymbol{x}^k, \boldsymbol{y}^k)^T\boldsymbol{\mu}^k.$$

Plugging the above equality into (12) , we obtain

$$\begin{aligned}-\left(\boldsymbol{x} - (\boldsymbol{x}^k+\boldsymbol{e}^k)\right)^T \big(&\nabla_{\boldsymbol{x}} f(\boldsymbol{x}^k, \boldsymbol{y}^k) + \nabla_{\boldsymbol{x}}\boldsymbol{h}(\boldsymbol{x}^k, \boldsymbol{y}^k)^T\boldsymbol{\mu}^k\\ &+\nabla \boldsymbol{g}(\boldsymbol{x}^k)^T\boldsymbol{v}^k + \boldsymbol{e}^k\big) \le 0, \forall \boldsymbol{x} \in \mathcal{X}, \forall k.\end{aligned} \tag{14}$$

On the other hand, by the definition of $\boldsymbol{\Delta}_j^k$ and (13), we have that for all j the following identity holds

$$\begin{aligned}\boldsymbol{y}_j^k - \boldsymbol{\Delta}_j^k = \arg\min_{\boldsymbol{y}_j} \Big\{&\phi_j'(s_j(\boldsymbol{y}_j^k))s_j(\boldsymbol{y}_j) + \frac{1}{2}||\boldsymbol{y}_j - \boldsymbol{y}_j^k||^2\\ &+\left(\nabla_{\boldsymbol{y}_j} f(\boldsymbol{x}^k, \boldsymbol{y}^k)+\nabla_{\boldsymbol{y}_j}\boldsymbol{h}(\boldsymbol{x}^k, \boldsymbol{y}^k)^T\boldsymbol{\mu}^k\right)^T(\boldsymbol{y}_j-\boldsymbol{y}_j^k)\Big\}\end{aligned} \tag{15}$$

By the optimality condition of the above problem, we have, $\exists \boldsymbol{\xi}_j^k \in \phi_j'(s_j(\boldsymbol{y}_j^k))\partial s_j(\boldsymbol{y}_j^k - \boldsymbol{\Delta}_j^k), \forall j$ such that

$$\begin{aligned}\sum_{j=1}^{n_y}\big(\boldsymbol{\xi}_j^k - \boldsymbol{\Delta}_j^k + \nabla_{\boldsymbol{y}_j} f(\boldsymbol{x}^k, \boldsymbol{y}^k) + \nabla_{\boldsymbol{y}_j}\boldsymbol{h}(\boldsymbol{x}^k, \boldsymbol{y}^k)^T\boldsymbol{\mu}^k\big)^T\\ \times(\boldsymbol{y}_j - \boldsymbol{y}_j^k + \boldsymbol{\Delta}_j^k)) = 0.\end{aligned} \tag{16}$$

Combining (16) with (14), we have

$$\begin{aligned}&(\nabla f(\boldsymbol{x}^k, \boldsymbol{y}^k) + \boldsymbol{\chi}^k + \nabla\boldsymbol{h}(\boldsymbol{x}^k, \boldsymbol{y}^k)^T\boldsymbol{\mu}^k)^T\\ &\quad\times(\boldsymbol{x} - \boldsymbol{x}^k - \boldsymbol{e}^k, \boldsymbol{y} - \boldsymbol{y}^k + \boldsymbol{\Delta}^k) \ge 0, \forall \boldsymbol{x} \in \mathcal{X}, \boldsymbol{y} \in \mathbb{R}^M.\end{aligned} \tag{17}$$

where

$$\begin{aligned}\boldsymbol{\chi}^k &\triangleq \begin{Bmatrix}\nabla \boldsymbol{g}(\boldsymbol{x}^k)^T\boldsymbol{v}^k + \boldsymbol{e}^k\\ \boldsymbol{\xi}^k + \boldsymbol{\Delta}^k\end{Bmatrix}\\ \boldsymbol{\Delta}^k &\triangleq (\boldsymbol{\Delta}_j^k)_j, \boldsymbol{\xi} = (\boldsymbol{\xi}_j^k)_j.\end{aligned} \tag{18}$$

Next, we prove that $\boldsymbol{\mu}^k$ is bounded by contradiction and using Robinson condition. Assume, to the contrary, that $\boldsymbol{\mu}^k$ is

$$\boldsymbol{e}^k = \mathcal{P}_{\mathcal{X}}\{\boldsymbol{x}^k - \nabla_{\boldsymbol{x}}\mathcal{L}_k(\boldsymbol{x}^k, \boldsymbol{y}^k) - \nabla \boldsymbol{g}(\boldsymbol{x}^k)^T\boldsymbol{v}^k\} - \boldsymbol{x}^k, \tag{8}$$

$$\boldsymbol{\Delta}_j^k = \boldsymbol{y}_j^k - \arg\min_{\boldsymbol{y}_j}\left\{\begin{array}{c}\phi_j'(s_j(\boldsymbol{y}_j^k))s_j(\boldsymbol{y}_j) + \frac{1}{2}||\boldsymbol{y}_j - \boldsymbol{y}_j^k||^2\\ +\left(\nabla_{\boldsymbol{y}_j} f(\boldsymbol{x}^k, \boldsymbol{y}^k) + \nabla_{\boldsymbol{y}_j}\boldsymbol{h}(\boldsymbol{x}^k, \boldsymbol{y}^k)^T\left(\frac{1}{\varrho_k}\boldsymbol{h}(\boldsymbol{x}^k, \boldsymbol{y}^k) + \boldsymbol{\lambda}^k\right)\right)^T(\boldsymbol{y}_j - \boldsymbol{y}_j^k)\end{array}\right\} \tag{9}$$

unbounded. Define $\bar{\boldsymbol{\mu}}^k \triangleq \frac{\boldsymbol{\mu}^k}{\|\boldsymbol{\mu}^k\|}$. Since $\{\bar{\boldsymbol{\mu}}^k\}$ is bounded, there must exist a convergent subsequence $\{\bar{\boldsymbol{\mu}}^{k_r}\}$. Let $\boldsymbol{\mu}^{k_r} \to \bar{\boldsymbol{\mu}}$ as $r \to \infty$. On the other hand, since $f(\boldsymbol{x}, \boldsymbol{y})$ and $\boldsymbol{g}(\boldsymbol{x})$ are continuously differentiable, $\nabla f(\boldsymbol{x}^k, \boldsymbol{y}^k)$ and $\nabla \boldsymbol{g}(\boldsymbol{x}^k)$ are bounded. Moreover, by Theorem 2.1, we know that $\boldsymbol{\xi}^k$ is bounded. Also, by Robinson's condition and Lemma 3.26 in [49], we conclude that $\boldsymbol{v}^k$ is bounded. As a result, $\boldsymbol{\chi}^k$ is bounded.[1] By dividing both sides of (17) by $\|\boldsymbol{\mu}^k\|$ and using the boundedness of $\nabla f(\boldsymbol{x}^k, \boldsymbol{y}^k)$ and $\boldsymbol{\chi}^k$, we have for sufficiently large r

$$-\left(\boldsymbol{x} - (\boldsymbol{x}^{k_r} + \boldsymbol{e}^{k_r}), \boldsymbol{y} - (\boldsymbol{y}^{k_r} - \boldsymbol{\Delta}^{k_r})\right)^T \times \left(\nabla \boldsymbol{h}(\boldsymbol{x}^{k_r}, \boldsymbol{y}^{k_r})^T \bar{\boldsymbol{\mu}}^{k_r}\right) \le 0, \forall \boldsymbol{x} \in \mathcal{X}. \quad (19)$$

Note that $\nabla \boldsymbol{h}(\boldsymbol{x}, \boldsymbol{y})$ is continuous in $(\boldsymbol{x}, \boldsymbol{y})$. Moreover, by assumption

$$\max\left(\|\boldsymbol{e}^k\|_\infty, \|\boldsymbol{\Delta}^k\|_\infty\right) \le \epsilon_k, \forall k, \quad (20)$$

we have $\boldsymbol{e}^k \to 0$ and $\boldsymbol{\Delta}^k \to 0$ due to $\epsilon_k \to 0$ as $k \to 0$. In addition, it holds that $(\boldsymbol{x}^{k_r}, \boldsymbol{y}^{k_r}) \to (\boldsymbol{x}^*, \boldsymbol{y}^*)$ and $\boldsymbol{\mu}^{k_r} \to \bar{\boldsymbol{\mu}}$ as $r \to \infty$. Hence, taking limits on both sides of (17), we have

$$-(\boldsymbol{x} - \boldsymbol{x}^*, \boldsymbol{y} - \boldsymbol{y}^*)^T \nabla \boldsymbol{h}(\boldsymbol{x}^*, \boldsymbol{y}^*)^T \bar{\boldsymbol{\mu}} \le 0, \forall \boldsymbol{x} \in \mathcal{X}, \boldsymbol{y} \in \mathbb{R}^M. \quad (21)$$

Utilizing the first part of the Robinson's condition, that is

$$\left\{\nabla \boldsymbol{h}(\boldsymbol{x}^*, \boldsymbol{y}^*)(\boldsymbol{d}_{\boldsymbol{x}}, \boldsymbol{d}_{\boldsymbol{y}}) : \boldsymbol{d}_{\boldsymbol{x}} \in T_{\mathcal{X}}(\boldsymbol{x}^*), \boldsymbol{d}_y \in \mathbb{R}^M\right\} = \mathbb{R}^p, \quad (22)$$

it follows that there exists some $\boldsymbol{x} \in \mathcal{X}$, $\boldsymbol{y} \in \mathbb{R}^M$ and $c > 0$ such that $-\bar{\boldsymbol{\mu}} = c\nabla \boldsymbol{h}(\boldsymbol{x}^*, \boldsymbol{y}^*)(\boldsymbol{x} - \boldsymbol{x}^*, \boldsymbol{y} - \boldsymbol{y}^*)$. This together with (21) implies $\bar{\boldsymbol{\mu}} = \boldsymbol{0}$, contradicting the identity $\|\bar{\boldsymbol{\mu}}\| = 1$. Hence, $\{\boldsymbol{\mu}^k\}$ is bounded.

Step 2: Next we show that the algorithm indeed reaches the KKT points. From Steps 3-9, we observe that, either both $\{\boldsymbol{\mu}^k\}$ and $\{\boldsymbol{\lambda}_k\}$ are bounded with $\varrho_k \to 0$ (i.e., case 2 in Algorithm 1) , or $\boldsymbol{\mu}^k - \boldsymbol{\lambda}_k \to 0$ with ϱ_k bounded (i.e., case 1 in Algorithm 1). Hence, from the definition (13) we must have

$$\boldsymbol{h}(\boldsymbol{x}^k, \boldsymbol{y}^k) = \varrho_k(\boldsymbol{\mu}^k - \boldsymbol{\lambda}_k) \to 0.$$

which implies that $\boldsymbol{h}(\boldsymbol{x}^*, \boldsymbol{y}^*) = 0$. That is, the equality constraint will be satisfied in the limit. In addition, due to the boundedness of $\{\boldsymbol{\mu}^k\}$, there exists a convergent subsequence $\{\boldsymbol{\mu}^{k_r}\}$ that we assume converge to $\boldsymbol{\mu}^*$. By restricting to the subsequence $\{\boldsymbol{\mu}^{k_r}\}$ and taking limits on both sides of (14), we have

$$(\boldsymbol{x} - \boldsymbol{x}^*)^T \left(\nabla_{\boldsymbol{x}} f(\boldsymbol{x}^*, \boldsymbol{y}^*) + \nabla_{\boldsymbol{x}} \boldsymbol{h}(\boldsymbol{x}^*, \boldsymbol{y}^*)^T \boldsymbol{\mu}^* + \nabla \boldsymbol{g}(\boldsymbol{x}^*)^T \boldsymbol{v}^*\right) \ge 0, \forall \boldsymbol{x} \in \mathcal{X}, \quad (23)$$

On the other hand, since problem (15) has a unique solution, by restricting to a convergent subsequence, we can take limit on both sides of (15), leading to

$$\boldsymbol{y}_j^* = \arg\min_{\boldsymbol{y}_j} \phi_j'(s_j(\boldsymbol{y}_j^*)) s_j(\boldsymbol{y}_j) + \frac{1}{2}||\boldsymbol{y}_j - \boldsymbol{y}_j^*||^2 + \left(\nabla_{\boldsymbol{y}_j} f(\boldsymbol{x}^*, \boldsymbol{y}^*) + \nabla_{\boldsymbol{y}_j} \boldsymbol{h}(\boldsymbol{x}^*, \boldsymbol{y}^*)^T \boldsymbol{\mu}^*\right)^T (\boldsymbol{y}_j - \boldsymbol{y}_j^*), \forall j. \quad (24)$$

[1]Note that the objective function of problem $(P_{\varrho,\boldsymbol{\lambda}})$ is continuously differentiable in $\boldsymbol{x}$. Thus we can apply here Lemma 3.26 in [49].

TABLE II
ALGORITHM 2: PDD ALGORITHM FOR PROBLEM (P)

0. initialize $\boldsymbol{z}^0 = (\boldsymbol{x}^0, \boldsymbol{y}^0)$, $\varrho_0 > 0$, $\boldsymbol{\lambda}_0$, and $k = 1$
   pick two sequences $\{\eta_k > 0\}$, $\{\epsilon_k > 0\}$
1. **repeat**
2. $(\boldsymbol{z}^k, \boldsymbol{v}^k) = \text{optimize}(P_{\varrho_k, \boldsymbol{\lambda}_k}, \boldsymbol{z}^{k-1}, \epsilon_k)$
4. $\boldsymbol{\lambda}_{k+1} = \boldsymbol{\lambda}_k + \frac{1}{\varrho_k} \boldsymbol{h}(\boldsymbol{z}^k)$
5. update ϱ_{k+1} by decreasing ϱ_k
6. $k = k + 1$
7. **until** some termination criterion is met

It follows that

$$\boldsymbol{0} \in \phi_j'(s_j(\boldsymbol{y}_j^*)) \partial s_j(\boldsymbol{y}_j^*) + \nabla_{\boldsymbol{y}_j} f(\boldsymbol{x}^*, \boldsymbol{y}^*) + \nabla_{\boldsymbol{y}_j} \boldsymbol{h}(\boldsymbol{x}^*, \boldsymbol{y}^*)^T \boldsymbol{\mu}^*, \forall j \quad (25)$$

In addition, $\boldsymbol{g}(\boldsymbol{x}^k) \le 0$ implies $\boldsymbol{g}(\boldsymbol{x}^*) \le 0$. Moreover, since $\boldsymbol{v}^k$ are the Lagrange multiplier associated with the constraints $\boldsymbol{g}(\boldsymbol{x}) \le 0$, we have $\boldsymbol{g}(\boldsymbol{x}^k)^T \boldsymbol{v}^k = 0$ and $\boldsymbol{v}^k \ge 0$. It follows that

$$\boldsymbol{g}(\boldsymbol{x}^*)^T \boldsymbol{v}^* = 0 \text{ and } \boldsymbol{v}^* \ge 0. \quad (26)$$

Combining Eqs. (23), (25), (26) and the fact $\boldsymbol{h}(\boldsymbol{x}^*, \boldsymbol{y}^*) = 0$, $\boldsymbol{g}(\boldsymbol{x}^*) \le 0$, and $\boldsymbol{x}^* \in \mathcal{X}$, we conclude that $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ satisfies the KKT condition of problem (P). This completes the proof. ■

Remark 3.1: We note that in the above proof, the Robinson's condition has been used in a slightly different way than in the proof of Theorem 2.3. In particular, in Theorem 2.3, the condition is assumed on a local minimizer $(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})$, which is obviously a *feasible* solution for problem (P). On the other hand, in Theorem 3.1, the Robinson's condition is assumed on a limit point $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ generated by the PDD algorithm, and such a point may not be feasible for the constraints $\boldsymbol{h}(\boldsymbol{x}^*, \boldsymbol{y}^*) = \boldsymbol{0}$ to begin with. Therefore, in practical applications, in order to use Theorem 3.1, one has to check whether Robinson's condition holds for all $(\boldsymbol{x}, \boldsymbol{y})$ satisfying the constraints that $\boldsymbol{x} \in \mathcal{X}$, $\boldsymbol{g}_i(\boldsymbol{y}) \le 0$, $\forall i$ (but not necessarily satisfying $\boldsymbol{h}(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{0}$). This will be done for each application that we will study in Part II of this paper.

C. PDD Method With Increasing Penalty

We expect that in practice, the basic PDD method can achieve convergence with *finite* penalty in many applications. However, it requires frequent evaluation of constraint violation, an operation that can be costly for certain applications. To overcome this weakness, we propose a simple variant of the basic PDD method; see Table II for the detailed description. The main difference lies in that we always keep *increasing* the penalty and updating the dual variable. Hence this variant is referred to as *increasing penalty dual decomposition* (IPDD) method. The following theorem shows that every limit point of the iterates generated by the IPDD is a KKT point of problem (P) under Robinson's condition.

Theorem 3.2: Let $\{\boldsymbol{x}^k, \boldsymbol{y}^k, \boldsymbol{v}^k\}$ be the sequence generated by Algorithm 2 for problem (P), where $\boldsymbol{v}^k = (\boldsymbol{v}_i^k)_i$ denote the Lagrange multipliers associated with the constraints $\boldsymbol{g}_i(\boldsymbol{x}_i) \le 0, \forall i$. The termination condition for the optimization oracle involved in Algorithm 2 is given in (10) with $\epsilon_k, \eta_k, \varrho_k \to 0$ as $k \to \infty$. Suppose that $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ is a limit point of the sequence $\{\boldsymbol{x}^k, \boldsymbol{y}^k\}$ and the condition (22) holds at $(\boldsymbol{x}^*, \boldsymbol{y}^*)$, then the point $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ satisfies $\boldsymbol{h}(\boldsymbol{x}^*, \boldsymbol{y}^*) = \boldsymbol{0}$. Furthermore, suppose

that Robinson's condition holds for problem (P) at $(\boldsymbol{x}^*, \boldsymbol{y}^*)$. Then $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ is a KKT point of problem (P), i.e., it satisfies the KKT system (5) of problem (P).

Proof: Following the same argument as that of the proof of Theorem 3.1, we can show 1) all the KKT equations except $\boldsymbol{h}(\boldsymbol{x}^*, \boldsymbol{y}^*) = 0$ and 2) that the sequence $\{\boldsymbol{\mu}^k\}$ is bounded. By checking the definition of $\boldsymbol{\mu}^k$ and the dual update in Step 4 of Algorithm 2, we have $\boldsymbol{\lambda}_{k+1} = \boldsymbol{\mu}^k$. It follows that the sequence $\{\boldsymbol{\lambda}^k\}$ is bounded, implying $\|\boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}_k\|$ is bounded. Since it holds that $\varrho_k \to 0$ as $k \to \infty$, we have from the dual update that $\|\boldsymbol{h}(\boldsymbol{x}^k, \boldsymbol{y}^k)\| = \varrho_k \|\boldsymbol{\lambda}_{k+1} - \boldsymbol{\lambda}_k\| \to 0$ as $k \to \infty$, implying $\boldsymbol{h}(\boldsymbol{x}^*, \boldsymbol{y}^*) = 0$. This completes the proof. ■

IV. Randomized BSUM for Problem $(P_{\varrho_k, \boldsymbol{\lambda}_k})$

In the PDD/IPDD method, BCD-type algorithms are typically used as optimization oracles in Step 2 to solve problem $(P_{\varrho_k, \boldsymbol{\lambda}_k})$, and it is assumed to be able to guarantee Eq. (10). However, the convergence theory of the basic BSUM algorithm [54] (which includes the exact BCD method [22] as a special case) is established only for convex constraint cases. By considering a random block update rule, we here provide an extension of the basic BSUM algorithm, termed rBSUM, which is applicable for problems with nonconvex constraints. In the following, we present the rBSUM algorithm with a convergence analysis. In particular, we show that the proposed rBSUM can reach KKT solutions of problem $(P_{\varrho_k, \boldsymbol{\lambda}_k})$, therefore ensuring Eq. (10).

To proceed, we define $\boldsymbol{z} = (\boldsymbol{z}_i)_i$ with $\boldsymbol{z}_i = \boldsymbol{x}_i$ for $i = 1, 2, \ldots, n$ and $\boldsymbol{z}_{n+j} = \boldsymbol{y}_j$ for $j = 1, 2, \ldots, n_y$, i.e., $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y})$. Let $n_z = n + n_y$ denote the total number of block variables, and define the set $[n_1 : n_2] \triangleq \{n_1, n_1 + 1 \cdots, n_2\}$. Hence, the sets $[1 : n]$ and $[n + 1 : n_z]$ contain the indices of the $\boldsymbol{x}_i$ variables and $\boldsymbol{y}_j$ variables in $\boldsymbol{z}$, respectively. Furthermore, for notational simplicity, we omit k for problem $(P_{\varrho_k, \boldsymbol{\lambda}_k})$ and denote its objective function simply as $\mathcal{L}(\boldsymbol{z})$. Thus, let us consider the rBSUM algorithm for solving

$$\begin{aligned} \min_{\boldsymbol{z}} \quad & \mathcal{L}(\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_{n_z}) \\ \text{s.t.} \quad & \boldsymbol{z}_i \in \mathcal{X}_i, i \in [1 : n], \\ & \boldsymbol{g}_i(\boldsymbol{z}_i) \le 0, i \in [1 : n]. \end{aligned} \tag{27}$$

At each iteration, the rBSUM updates one block variable by minimizing a locally tight upper bound $u_i(\cdot; \cdot)$ of the objective function, while fixing the rest of the blocks. Let $\tilde{\mathcal{X}}_{i+n} = \mathbb{R}^{m_i}$, $i \in [1 : n_y]$ and define $\tilde{\mathcal{X}} \triangleq \tilde{\mathcal{X}}_1 \times \tilde{\mathcal{X}}_2 \times \cdots \times \tilde{\mathcal{X}}_{n_z}$. The rBSUM algorithm is summarized in Table III, where Steps 3 and 4 generate a random index set $\mathcal{I}$ specifying the update order of block variables. In what follows, we study the convergence of the rBSUM algorithm.

First, we make the following assumption on $u_i(\cdot; \cdot)$.

Assumption 4.1:

$$u_i(\boldsymbol{z}_i; \boldsymbol{z}) = \mathcal{L}(\boldsymbol{z}), \forall \boldsymbol{z} \in \tilde{\mathcal{X}}, \forall i; \tag{28a}$$

$$u_i(\boldsymbol{v}_i; \boldsymbol{z}) \ge \mathcal{L}(\boldsymbol{z}_{<i}, \boldsymbol{v}_i, \boldsymbol{z}_{>i}), \forall \boldsymbol{v}_i \in \tilde{\mathcal{X}}_i, \forall \boldsymbol{z} \in \tilde{\mathcal{X}}, \forall i; \tag{28b}$$

$$u_i^o(\boldsymbol{v}_i; \boldsymbol{z}, \boldsymbol{d}_i)|_{\boldsymbol{v}_i = \boldsymbol{z}_i} = \mathcal{L}^o(\boldsymbol{z}; \boldsymbol{d}), \forall \boldsymbol{d} = (\boldsymbol{0}, \ldots, \boldsymbol{0}, \boldsymbol{d}_i, \boldsymbol{0}, \ldots, \boldsymbol{0})$$
$$\text{s.t. } \boldsymbol{x}_i + \boldsymbol{d}_i \in \tilde{\mathcal{X}}_i, \forall i; \tag{28c}$$

$$u_i(\boldsymbol{v}_i; \boldsymbol{z}) \text{is continuous in} (\boldsymbol{v}_i, \boldsymbol{z}), \forall i. \tag{28d}$$

TABLE III
Algorithm 3: rBSUM Algorithm

0. initialize $\boldsymbol{z}^0 \in \tilde{\mathcal{X}}$ and set $k = 0$
1. **repeat**
2. $\quad \boldsymbol{w} = \boldsymbol{z}^k$
3. $\quad$ uniformly randomly pick $i_k \in \{1, \ldots, n_z\}$
4. $\quad \mathcal{I}_k = \{i_k, 1, 2, \ldots, i_k - 1, i_k + 1, \ldots, n_z\}$
5. $\quad$ **for** each $i \in \mathcal{I}_k$
6. $\quad\quad \mathcal{A}_i^k = \arg\min_{\boldsymbol{z}_i \in \tilde{\mathcal{X}}_i} u_i(\boldsymbol{z}_i; \boldsymbol{w})$
7. $\quad\quad$ set $\boldsymbol{w}_i$ to be an arbitrary element in $\mathcal{A}_i^k$
8. $\quad$ **end**
9. $\quad \boldsymbol{z}^{k+1} = \boldsymbol{w}$
10. $\quad k = k + 1$
11. **until** some termination criterion is met

In the above assumption, $\boldsymbol{v}_i$ is the i-th block component of $\boldsymbol{v}$, having the same size as $\boldsymbol{z}_i$; the notations $\boldsymbol{z}_{<i}$ and $\boldsymbol{z}_{>i}$ represent the block components of $\boldsymbol{z}$ with their indices less than i or larger than i, respectively; $u_i^o(\boldsymbol{v}_i; \boldsymbol{z}, \boldsymbol{d}_i)$ denotes the generalized directional derivative of $u_i(\cdot; \boldsymbol{z})$ with respect to $\boldsymbol{v}_i$ along the direction $\boldsymbol{d}_i$; and $\mathcal{L}^o(\boldsymbol{z}; \boldsymbol{d})$ denotes the generalized directional derivative of $\mathcal{L}(\cdot)$ with respect to $\boldsymbol{z}$ along the direction $\boldsymbol{d}$. The assumption (28c) guarantees that the first order behavior of $u_i(\cdot, \boldsymbol{z})$ is the same as $\mathcal{L}(\cdot)$ locally [54], hence it is referred to as the gradient consistency assumption.

Second, we give the definition of regular functions which will be used later.

Definition 4.1 (Regularity of a function): A function $\hbar(\cdot)$ is *regular* at $\boldsymbol{x} = (\boldsymbol{x}_i)_i$ if the following implication holds

$$\hbar^o(\boldsymbol{x}; \boldsymbol{d}) \ge 0, \forall \boldsymbol{d} = (\boldsymbol{d}_i)_i \Longleftarrow \hbar^o(\boldsymbol{x}; \boldsymbol{d}_i^0) \ge 0, \quad \forall \boldsymbol{d}_i^0 \triangleq (\boldsymbol{0}, \ldots, \boldsymbol{0}, \boldsymbol{d}_i, \boldsymbol{0}, \ldots, \boldsymbol{0}), \forall i.$$

Based on the above assumption and the definition of regular functions, we next prove that, with probability one (w.p.1.) the sequence generated by the rBSUM algorithm converges to the set of stationary/KKT solutions of problem (27).

Theorem 4.1: Let Assumption 4.1 hold. Furthermore, assume that $\mathcal{L}(\cdot)$ is bounded below in $\tilde{\mathcal{X}}$ and it is regular at every point in $\tilde{\mathcal{X}}$. Then with probability one, every limit point of the iterates generated by the rBSUM algorithm, denoted as $\boldsymbol{z}^\infty$, is a stationary point of problem (27), which satisifes the following condition

$$\mathcal{L}^o(\boldsymbol{z}^\infty; \boldsymbol{d}) \ge 0, \ \ \forall \boldsymbol{d} = (\boldsymbol{d}_1, \boldsymbol{d}_2, \ldots, \boldsymbol{d}_{n_z}) \text{ with } \boldsymbol{d}_i \in T_{\tilde{\mathcal{X}}_i}(\boldsymbol{z}_i^\infty), \ \forall i. \tag{29}$$

Moreover, if Robinson's condition holds for problem (27) at the limit point, then the limit point is also a KKT point of problem (27).

Proof: It is easily seen that Steps 3 and 4 can generate n_z permutations of the index set in total. Let π denote the index of permutation and $\pi(1)$ denote the first number of the π-th permutation. Moreover, let $q_\pi > 0$ denote the probability of permutation π, with $\sum_{\pi=1}^{n_z} q_\pi = 1$. Then, we have

$$\mathbb{E}[\mathcal{L}(\boldsymbol{z}^{k+1}) \mid \boldsymbol{z}^k] = \sum_{\pi=1}^{n_z} q_\pi \mathcal{L}(\boldsymbol{z}^{\pi, k+1}) \tag{30}$$

where $\boldsymbol{z}^{\pi,k+1}$ denotes the update obtained by running one iteration of rBSUM (given $\boldsymbol{z}^k$) according to the block selection rule specified by the π-th permutation. Due to the upper bound assumption (28b) and the update rule, it must hold that

$$\mathcal{L}(\boldsymbol{z}^{\pi,k+1}) \leq \min_{\boldsymbol{z}_{\pi(1)}\in\tilde{\mathcal{X}}_{\pi(1)}} u_{\pi(1)}(\boldsymbol{z}_{\pi(1)};\boldsymbol{z}^k),\ \forall\pi. \tag{31}$$

Combining (30) and (31), we have

$$\mathbb{E}[\mathcal{L}(\boldsymbol{z}^{k+1}) \mid \boldsymbol{z}^k] \leq \mathcal{L}(\boldsymbol{z}^k) - \sum_{\pi=1}^{n_z} q_\pi \Bigg(\mathcal{L}(\boldsymbol{z}^k) - \min_{\boldsymbol{z}_{\pi(1)}\in\tilde{\mathcal{X}}_{\pi(1)}} u_{\pi(1)}(\boldsymbol{z}_{\pi(1)};\boldsymbol{z}^k) \Bigg) \tag{32}$$

which implies that $\mathcal{L}(\boldsymbol{z}^k)$ is a supermartingale and thus converges [55], and moreover the following holds w.p.1.,

$$\sum_{k=1}^{\infty}\sum_{\pi=1}^{n_z} q_\pi \left(\mathcal{L}(\boldsymbol{z}^k) - \min_{\boldsymbol{z}_{\pi(1)}\in\tilde{\mathcal{X}}_{\pi(1)}} u_{\pi(1)}(\boldsymbol{z}_{\pi(1)};\boldsymbol{z}^k) \right) < \infty \tag{33}$$

as $\mathcal{L}(\cdot)$ is bounded from below. Thus, by noting $\mathcal{L}(\boldsymbol{z}^k) \geq \min_{\boldsymbol{z}_{\pi(1)}\in\tilde{\mathcal{X}}_{\pi(1)}} u_{\pi(1)}(\boldsymbol{z}_{\pi(1)};\boldsymbol{z}^k), \forall\pi$, we must have, w.p.1.,

$$\lim_{k\to\infty} \left(\mathcal{L}(\boldsymbol{z}^k) - \min_{\boldsymbol{z}_{\pi(1)}\in\tilde{\mathcal{X}}_{\pi(1)}} u_{\pi(1)}(\boldsymbol{z}_{\pi(1)};\boldsymbol{z}^k) \right) = 0, \forall\pi. \tag{34}$$

Now let us restrict our analysis to a convergent subsequence $\{\boldsymbol{z}^{k_j}\}$ with $\lim_{j\to\infty}\boldsymbol{z}^{k_j} = \boldsymbol{z}^\infty$. We have from (34) and the continuity of $\mathcal{L}(\cdot)$ that

$$\lim_{j\to\infty} \min_{\boldsymbol{z}_{\pi(1)}\in\tilde{\mathcal{X}}_{\pi(1)}} u_{\pi(1)}(\boldsymbol{z}_{\pi(1)};\boldsymbol{z}^{k_j}) = \mathcal{L}(\boldsymbol{z}^\infty), \forall\pi, \text{w.p.1.} \tag{35}$$

On the other hand, according to the update rule, we have

$$\min_{\boldsymbol{z}_{\pi(1)}\in\tilde{\mathcal{X}}_{\pi(1)}} u_{\pi(1)}(\boldsymbol{z}_{\pi(1)};\boldsymbol{z}^{k_j}) \leq u_{\pi(1)}(\boldsymbol{z}_{\pi(1)};\boldsymbol{z}^{k_j}), \quad \forall \boldsymbol{z}_{\pi(1)} \in \tilde{\mathcal{X}}_{\pi(1)}, \forall\pi, \text{w.p.1.} \tag{36}$$

By taking limit as $j\to\infty$ on both sides of (36), and using (35) and the continuity of $u_i(\cdot;\cdot)$, we obtain

$$\mathcal{L}(\boldsymbol{z}^\infty) \leq u_{\pi(1)}(\boldsymbol{z}_{\pi(1)};\boldsymbol{z}^\infty), \forall \boldsymbol{z}_{\pi(1)} \in \tilde{\mathcal{X}}_{\pi(1)}, \forall\pi, \text{w.p.1.} \tag{37}$$

Due to the function value consistency assumption (28a), we have $\mathcal{L}(\boldsymbol{z}^\infty) = u(\boldsymbol{z}_i^\infty;\boldsymbol{z}^\infty), \forall i$, and thus

$$u(\boldsymbol{z}_{\pi(1)}^\infty;\boldsymbol{z}^\infty) \leq u_{\pi(1)}(\boldsymbol{z}_{\pi(1)};\boldsymbol{z}^\infty), \forall \boldsymbol{z}_{\pi(1)} \in \tilde{\mathcal{X}}_{\pi(1)}, \forall\pi, \text{w.p.1.} \tag{38}$$

Note that the above inequality holds for all permutations. Therefore, we have that w.p.1.,

$$u_i(\boldsymbol{z}_i^\infty;\boldsymbol{z}^\infty) \leq u_i(\boldsymbol{z}_i;\boldsymbol{z}^\infty), \forall \boldsymbol{z}_i \in \tilde{\mathcal{X}}_i, \forall i. \tag{39}$$

It follows that[2]

$$u_i^o(\boldsymbol{z}_i^\infty;\boldsymbol{z}^\infty,\boldsymbol{d}_i) \geq 0, \forall \boldsymbol{d}_i \in T_{\tilde{\mathcal{X}}_i}(\boldsymbol{z}_i^\infty), \forall i. \tag{40}$$

[2]This can be proven following a similar argument as that for the first part of proof of Theorem 2.3.

where

$$T_{\tilde{\mathcal{X}}_i}(\boldsymbol{z}_i^\infty) = \begin{cases} \{\boldsymbol{d}_i | \boldsymbol{d}_i \in T_{\mathcal{X}_i}(\boldsymbol{z}_i^\infty), \\ \nabla \boldsymbol{g}_{i,\ell}(\boldsymbol{z}_i^\infty)^T \boldsymbol{d}_i \leq 0, \ell \in I_i(\boldsymbol{z}_i^\infty)\}, \ \forall i \in [n] \\ \mathbb{R}^{m_{i-n}}, i = n+1,\ldots,n_z. \end{cases} \tag{41}$$

Thus, by the gradient consistency assumption (28c), we have from (40) that w.p.1.,

$$\mathcal{L}^o(\boldsymbol{z}^\infty;\boldsymbol{d}_i^0) \geq 0, \forall \boldsymbol{d}_i^0 \triangleq (\boldsymbol{0},\ldots,\boldsymbol{0},\boldsymbol{d}_i,\boldsymbol{0},\ldots,\boldsymbol{0}), \quad \boldsymbol{d}_i \in T_{\tilde{\mathcal{X}}_i}(\boldsymbol{z}_i^\infty), \forall i. \tag{42}$$

Since $\mathcal{L}(\boldsymbol{z})$ is regular at $\boldsymbol{z}^\infty$, it follows that w.p.1.,

$$\mathcal{L}^o(\boldsymbol{z}^\infty;\boldsymbol{d}) \geq 0, \forall \boldsymbol{d} = (\boldsymbol{d}_1,\boldsymbol{d}_2,\ldots,\boldsymbol{d}_{n_z}) \text{ with } \boldsymbol{d}_i \in T_{\tilde{\mathcal{X}}_i}(\boldsymbol{z}_i^\infty),\ \forall i. \tag{43}$$

By applying to Eq. (43) a similar argument as that for the second part of proof of Theorem 2.3, we can show under Robinson's condition that, there exists multipliers $(\hat{\boldsymbol{\nu}}_j)_j$ associated with the inequality constraints such that, the KKT condition of problem (27), i.e., Eqs. (5a–5e) with $(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}}) = \boldsymbol{z}^\infty$, holds true w.p.1. This completes the proof. ■

Next, we present a result regarding the convergence rates for the rBSUM algorithm. We start with imposing the following additional assumptions. Let us define:

$$\mathcal{L}(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\lambda}) = \tilde{\mathcal{L}}(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\lambda}) + \sum_{j=1}^{n_y} \phi_j(s_j(\boldsymbol{y}_j))$$

$$\text{where } \tilde{\mathcal{L}}(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\lambda}) \triangleq f(\boldsymbol{x},\boldsymbol{y}) + \boldsymbol{\lambda}^T\boldsymbol{h}(\boldsymbol{x},\boldsymbol{y}) + \frac{1}{2\varrho}\|\boldsymbol{h}(\boldsymbol{x},\boldsymbol{y})\|^2.$$

Further we assume that the upper bound functions only approximate the smooth part $\tilde{\mathcal{L}}$, that is

$$u_i(\boldsymbol{z}_i;\boldsymbol{z}) = q_i(\boldsymbol{z}_i;\boldsymbol{z}), \forall i \in [1:n]$$
$$u_i(\boldsymbol{z}_i;\boldsymbol{z}) = q_i(\boldsymbol{z}_i;\boldsymbol{z}) + \phi_{i-n}(s_{i-n}(\boldsymbol{z}_i)), \forall i \in [n+1:n_z]$$

where $q_i(\cdot)$'s are the new *differentiable* upper bound functions that satisfy a set of conditions similarly as in Assumption 4.1, given below:

Assumption 4.2:

$$q_i(\boldsymbol{z}_i;\boldsymbol{z}) = \tilde{\mathcal{L}}(\boldsymbol{z}), \forall \boldsymbol{z} \in \tilde{\mathcal{X}}, \forall i; \tag{44a}$$

$$q_i(\boldsymbol{v}_i;\boldsymbol{z}) \geq \tilde{\mathcal{L}}(\boldsymbol{z}_{<i},\boldsymbol{v}_i,\boldsymbol{z}_{>i}), \forall \boldsymbol{v}_i \in \tilde{\mathcal{X}}_i, \forall \boldsymbol{z} \in \tilde{\mathcal{X}}, \forall i; \tag{44b}$$

$$\nabla q_i(\boldsymbol{z}_i;\boldsymbol{z}) = \nabla_{\boldsymbol{z}_i}\tilde{\mathcal{L}}(\boldsymbol{z}), \forall \boldsymbol{z} \in \tilde{\mathcal{X}}, \forall i; \tag{44c}$$

$$q_i(\boldsymbol{v}_i;\boldsymbol{z}) \text{is continuous in} (\boldsymbol{v}_i,\boldsymbol{z}), \forall i. \tag{44d}$$

Please note that here $q_i(\cdot)$'s are differentiable, so to satisfy Assumption 4.1, the nonsmooth terms $\phi_{i-n}(s_{i-n}(\boldsymbol{z}_i))$'s can be chosen as standard convex non-smooth functions such as ℓ_1 and ℓ_2 norms.

Further, we make the following *additional assumptions* on q_i's and the Lagrangian functions:

Assumption 4.3:

$$q_i(\boldsymbol{x}_i;\boldsymbol{z}) - q_i(\boldsymbol{y}_i;\boldsymbol{z}) \geq \langle \nabla q_i(\boldsymbol{y}_i;\boldsymbol{z}), \boldsymbol{x}_i - \boldsymbol{y}_i \rangle + \frac{\theta}{2}\|\boldsymbol{x}_i - \boldsymbol{y}_i\|^2,\ \forall \boldsymbol{x}_i,\boldsymbol{y}_i \in \tilde{\mathcal{X}}_i,\ \forall i. \tag{45a}$$

$$\|\nabla q_i(\boldsymbol{x}_i;\boldsymbol{z}) - \nabla q_i(\boldsymbol{x}_i;\boldsymbol{w})\| \le H_i \|\boldsymbol{z}-\boldsymbol{w}\|, \quad \forall \boldsymbol{z},\boldsymbol{w} \in \tilde{\mathcal{X}}, \boldsymbol{x}_i \in \tilde{\mathcal{X}}_i \ \forall i \tag{45b}$$

$$\left\|\nabla \tilde{\mathcal{L}}(\boldsymbol{z}) - \nabla \tilde{\mathcal{L}}(\boldsymbol{w})\right\| \le L\|\boldsymbol{z}-\boldsymbol{w}\|, \ \forall \boldsymbol{z},\boldsymbol{w} \in \tilde{\mathcal{X}}. \tag{45c}$$

$$\boldsymbol{g}_i(\boldsymbol{x}_i), s_j(\boldsymbol{y}_j) \text{ are convex functions, } \forall i,j \tag{45d}$$

$$\phi_j(s_j(\boldsymbol{y}_j)) = s_j(\boldsymbol{y}_j), \ \forall j. \tag{45e}$$

In Assumption 4.3, the first item says $q_i(\cdot)$ is strongly convex, and the second to the fourth items are related to the Lipschitzness of the gradients of the upper bound functions and the Lagrangian function. We note that the Lipschitz assumption (45b) typically holds when the original Lagrangian function has Lipschitz gradient, i.e., when (45c) holds. The last two conditions restrict the constraint $\boldsymbol{g}_i(\boldsymbol{x}_i) \le 0$ to be a convex constraint, and the nonsmooth regularizer to be a simple convex one. The proof of the result below can be found in Appendix C.

Theorem 4.2: Let Assumptions 4.1–4.3 hold. Further assume that $\mathcal{L}(\cdot)$ is bounded from below. Then if rBSUM is run for T iterations, we have the following convergence rate estimate:

$$\min_{1\le k\le T}\{\max\{\|\boldsymbol{\Delta}^{k+1}\|_\infty^2, \|\boldsymbol{e}^{k+1}\|_\infty^2\}\} \le \frac{d}{T}, \tag{46}$$

where d is some constant independent of T.

Remark 4.1: In the derivation of the above result, we showed that under Assumption 4.1–4.3, *for each* $i \in [1:n], j \in [1:n_y]$, the following holds:

$$\|\boldsymbol{\Delta}_j^{k+1}\| \le H_{n+j}\|\boldsymbol{z}^{k+1}-\boldsymbol{z}^k\|, \ \|\boldsymbol{e}_i^{k+1}\| \le H_i\|\boldsymbol{z}^{k+1}-\boldsymbol{z}^k\|.$$

Therefore, a simple way to verify the stopping criteria (7) is to check if the following holds:

$$\sqrt{\sum_{i=1}^{n_z} H_i^2} \times \|\boldsymbol{z}^{k+1}-\boldsymbol{z}^k\| \le \epsilon. \tag{47}$$

Since the rBSUM algorithm achieves convergence to KKT points *w.p.1* under Robinson's condition, we modify the claims of Theorem 3.1 & 3.2 for the case where rBSUM is used as the optimization oracle.

Corollary 4.1: Suppose that the parameter settings and the termination conditions of the optimization oracle in Theorem 3.1 & 3.2 are used, and that rBSUM is used as the optimization oracle. Then every limit point of the sequence generated by the PDD/IPDD method is a KKT point w.p.1 when rBSUM is used as the optimization oracle, provided that the Robinson's condition is satisfied at the limit point.

V. Discussion

A. The Robinson's Condition

It is well-known that constraint qualification (CQ) conditions (or regularity conditions) are often needed to precisely describe the first-order optimality condition for nonlinear optimization. In our KKT and convergence analysis, Robinson's condition is assumed as a type of CQ. Similarly to many other CQs, such condition is generally difficult to check, but it is a standard one and has been used in many existing works on constrained optimization, e.g., [23], [49], [56], [57]. For ease of understanding Robinson's condition, [49, Lemma 3.16] has provided a simple sufficient condition. That is, *if the rows of* $\nabla \boldsymbol{h}(\boldsymbol{z}^*)$ *are linearly independent and moreover there exists* $\boldsymbol{z}^{int} = (\boldsymbol{x}^{int}, \boldsymbol{y}^{int}) \in int(\mathcal{X} \times \mathbb{R}^M)$ *such that* $\nabla \boldsymbol{h}(\boldsymbol{z}^*)(\boldsymbol{z}^{int} - \boldsymbol{z}^*) = \boldsymbol{0}$ *and* $\nabla g_{i\ell}(\boldsymbol{x}_i^*)(\boldsymbol{x}_i^{int} - \boldsymbol{x}_i^*) < 0, \ \forall \ell \in I_i(\boldsymbol{x}_i^*), \forall i$, *then Robinson's condition (3) holds true*.

Below, we summarize the relationship between the Robinson's condition and a few commonly used CQs.

1) MFCQ: When $\mathcal{X} = \mathbb{R}^N$, the above sufficient condition for Ronbinson's condition reduces to the well-known *Mangasarian-Fromovitz constraint qualification* (MFCQ). Moreover, it is shown in [49, Lemma 3.17] that Robinson's condition is equivalent to the MFCQ when $\mathcal{X} = \mathbb{R}^N$.

2) LICQ: When $\mathcal{X} = \mathbb{R}^N$ and the rows of $\nabla \boldsymbol{h}(\boldsymbol{z}^*)$ as well as the gradients of the *active* inequality constraint functions $\nabla g_{i\ell}(\boldsymbol{x}_i^*)$'s are linearly independent, we can easily find $\boldsymbol{z}^{int}$ such that $\nabla \boldsymbol{h}(\boldsymbol{z}^*)(\boldsymbol{z}^{int} - \boldsymbol{z}^*) = \boldsymbol{0}$ and $\nabla g_{i\ell}(\boldsymbol{x}_i^*)(\boldsymbol{x}_i^{int} - \boldsymbol{x}_i^*) < 0, \ \forall \ell \in I_i(\boldsymbol{x}_i^*), \forall i$. This means that Robinson's condition is implied by the *linear independence constraint qualification* (LICQ).

3) Slater's condition: When the constraint set of problem (P) is convex (i.e., $\boldsymbol{h}(\cdot)$ is affine and $\boldsymbol{g}_i(\cdot)$'s are convex) and the Slater's condition holds, i.e., there exists a point $\boldsymbol{z}^s = (\boldsymbol{x}^s, \boldsymbol{y}^s) \in int(\mathcal{X} \times \mathbb{R}^M)$ such that $\boldsymbol{h}(\boldsymbol{z}^s) = 0$ and $\boldsymbol{g}_i(\boldsymbol{x}_i^s) < \boldsymbol{0}$, $\forall i$, it can be easily shown that the following relations hold

$$\nabla g_{i\ell}(\boldsymbol{x}_i^*)(\boldsymbol{x}_i^s - \boldsymbol{x}_i^*) \le g_{i\ell}(\boldsymbol{x}_i^s) - g_{i\ell}(\boldsymbol{x}_i^*) < 0, \ \forall \ell \in I_i(\boldsymbol{x}_i^*), \forall i,$$

$$\nabla \boldsymbol{h}(\boldsymbol{z}^*)(\boldsymbol{z}^s - \boldsymbol{z}^*) = \boldsymbol{h}(\boldsymbol{z}^s) - \boldsymbol{h}(\boldsymbol{z}^*) = \boldsymbol{0}.$$

Hence, the Slater's condition is sufficient for Robinson's condition for problems with convex constraints.

4) Linearly Constraint Qualification: When problem (P) has linear constraints (i.e., $\mathcal{X} = \mathbb{R}^N$ and $\boldsymbol{h}(\cdot)$ and $\boldsymbol{g}_i(\cdot)$'s are affine), as in the case of Slater's condition, it can be readily verified that Robinson's condition holds true.

B. Practical Considerations on Parameter Selection and Termination Conditions

In the PDD method, the control parameter η_k determines how often the AL method and the penalty method are carried out. If η_k is decreased too fast, then the penalty method will often take place, resulting in a large penalty and slow convergence. On the other hand, when η_k is decrease very slowly, then the AL method will be more often performed. However, if the AL method does not converge in this case, such a choice will also slow down the convergence of the PDD. A more adaptive way to set η_k is to make it explicitly related to the constraint violation. For example, we can set $\eta_k = \tau \min(\eta_{k-1}, \|\boldsymbol{h}(\boldsymbol{z}^{k-1})\|_\infty)$ where $0 < \tau < 1$. Similarly, the penalty parameter ϱ_k can impact the convergence the PDD method. Specifically, when ϱ_k decreases too fast, the AL problem will become ill-conditioned which impact the convergence of the optimization oracle. A simple way to set ϱ_k is to let $\varrho_{k+1} = c\varrho_k$ where the parameter c is a fraction which should be appropriately chosen to control the decreasing speed of the penalty parameter. Various choices of the parameter settings will be examined extensively in the second part of this paper.

Besides the parameter choice, the termination condition of the optimization oracle also affects the convergence of the PDD/IPDD. To guarantee theoretical convergence, we have used Eq. (10) to terminate the optimization oracle. However, it is sometime difficult to evaluate $\boldsymbol{e}^k$ and $\boldsymbol{\Delta}^k$ when the set $\mathcal{X}$ is complicated and the function $s_j(\boldsymbol{y}_j)$ is not simple. In practice, it is reasonable to terminate the optimization oracle based on the

progress of the objective value $\mathcal{L}_k(\boldsymbol{z}^k)$, called relative objective progress (RBP) condition, i.e.,

$$\frac{|\mathcal{L}_k(\boldsymbol{z}^k) - \mathcal{L}_{k-1}(\boldsymbol{z}^{k-1})|}{|\mathcal{L}_{k-1}(\boldsymbol{z}^{k-1})|} \leq \epsilon_k. \tag{48}$$

Another practical choice of the termination condition for the optimization oracle is simple by setting the maximum number of iterations. Such termination condition is very suitable for in-network distributed implementation as it does not require coordination among network agents. Although the latter condition lacks theoretical guarantee, it is actually perform well in our numerical experience.

C. Optimization Oracle

To make use of the problem structure, we advocate using BCD-type algorithms as optimization oracle to address the AL problem $(P_{\varrho_k,\boldsymbol{\lambda}_k})$. Certainly, any other reasonable optimization method can be used, as long as they can guarantee the theoretical termination condition (10). For example, when some inequality constraint $\boldsymbol{g}_j(\boldsymbol{x}_j) \leq 0$ is complicated, we can use concave-convex procedure [58], [59] to address the AL problem, i.e., we can replace $\boldsymbol{g}_j(\boldsymbol{x}_j)$ with its simple upper bound function [60] and solve the resulting problem (which is often easier) instead of the AL problem. In addition, when the AL problem can be globally solved by certain solver, the PDD/IPDD method with such global solver (instead of the BCD-type algorithms) could provide globally optimal solution to problem (P).

D. Sharper Solution Concepts

We note that there is a vast literature on developing generalized subdifferentials/subgradients of different types, beyond the classical Clarke subdifferential/subgradient. For example, in [61] Mordukhovich introduced a notion of KKT conditions based on the generalized gradient of Mordukhovich (MGG) (under suitable nonsmooth extensions of the Mangasarian-Fromovitz Constraint Qualifications), which is sharper than the condition bsed on the Clark's subdifferential. Also there are many extensions of the above results, see for example, those defined using linear generalized gradient [62]–[64]. The current work uses the relative weaker notion of KKT conditions, mainly because of the following reasons. First, such a notion is still a valid necessary condition for the original problem (P); Second, it is relatively easy to design efficient algorithms (i.e., the proposed PDD algorithm) such that, combined with relatively easily checkable conditions (i.e., the Robinson's condition), the KKT condition can be computed. Indeed, in the applications to be presented in Part II of this work. we can check that the Robinson's CQ are all satisfied. Third, the solutions obtained by the PDD algorithm (which are only guaranteed to achieve those weaker KKT conditions), achieve very good practical performance. This may suggest that for signal processing applications that are of interest to this work, such weaker notions of KKT are reasonable solution concepts. We note that, by using MGG or related notions, the problems to be covered could certainly be more general, and solutions with better qualities can be obtained (at least in theory). Therefore it will be an interesting future work to analyze if one can extend the PDD algorithm to compute those sharper KKT solutions.

VI. Conclusion

In this paper, we design an optimization algorithm for a class of nonsmooth and nonconvex problems. The proposed algorithm, named PDD, can deal with difficult nonconvex coupling constraints, and it is further able to fully explore the problem structure for efficient numerical implementation. The PDD can be used to address a wide range of difficult engineering problems arising from areas such as signal processing, wireless communication and machine learning. In the second part of this paper we will demonstrate the strength of our algorithm by customizing it to a number of applications.

Appendix A
Some Basics

To improve the readability, we here list a few definitions and facts which are from [49, Chap 2&3] and used throughout the paper.

A. Tangent Cone, Polar Cone and Normal Cone

Tangent cone is the set of tangent directions whose definition is given as follows.

Definition A.1: [49, Def. 3.11] A direction $\boldsymbol{d}$ is called *tangent* to the set $X \subset \mathbb{R}^n$ at the point $\boldsymbol{x} \in X$ if there exist sequences of points $\boldsymbol{x}^k \in X$ and scalars $\tau_k > 0$, $k = 1, 2, \ldots$, such that $\tau_k \downarrow 0$ and

$$\boldsymbol{d} = \lim_{k\to\infty} \frac{\boldsymbol{x}^k - \boldsymbol{x}}{\tau_k}$$

Further, define the cone of feasible directions at $\boldsymbol{x} \in X$:

$$K_X(\boldsymbol{x}) = \{\boldsymbol{d} \in \mathbb{R}^n \mid \boldsymbol{d} = \beta(\boldsymbol{y} - \boldsymbol{x}), \boldsymbol{y} \in X, \beta \geq 0\}.$$

Then we have

Lemma A.1: [49, Lemma 3.13] Let $X \subset \mathbb{R}^n$ be a convex set and let $\boldsymbol{x} \in X$. Then the ***tangent cone***, i.e., the set of tangent direction, of the set X at $\boldsymbol{x}$ is

$$T_X(\boldsymbol{x}) = col\ K_X(\boldsymbol{x})$$

where $col\ X$ means the closure of the set X.

The polar cone is defined as follows.

Definition A.2: [49, Def. 2.23] Let K be a cone in $\mathbb{R}^n$. The set

$$K^\circ \triangleq \{\boldsymbol{y} \in \mathbb{R}^n \mid \boldsymbol{y}^T\boldsymbol{x} \leq 0 \quad \forall \boldsymbol{x} \in K\}$$

is called the ***polar cone*** of K.

Define

$$K \triangleq \{\boldsymbol{x} \in K_1 \mid \mathbf{A}\boldsymbol{x} \in K_2\}. \tag{49}$$

Given the definition of polar cone, the following fact holds true.

Theorem A.1: [49, Theorem 2.36] Assume that K_1 and K_2 are closed convex cones, and K is defined by (49). If

$$0 \in \text{int}\{\mathbf{A}\boldsymbol{x} - \boldsymbol{y} : \boldsymbol{x} \in K_1, \boldsymbol{y} \in K_2\}, \tag{50}$$

then

$$K^\circ = K_1^\circ + \{\mathbf{A}^T\boldsymbol{\lambda} : \boldsymbol{\lambda} \in K_2^\circ\}. \tag{51}$$

The definition of normal cone is given as follows.

Definition A.3: [49, Def. 2.37] Let X be a closed convex set and let $\boldsymbol{x} \in X$. Then

$$N_X(\boldsymbol{x}) = \{\boldsymbol{v} \in \mathbb{R}^n \mid \boldsymbol{v}^T(\boldsymbol{y} - \boldsymbol{x}) \leq 0, \forall \boldsymbol{y} \in X\}$$

is called ***normal cone*** to X at $\boldsymbol{x}$.

Following the above definitions, we have for a closed convex set X

$$[T_X(x)]^o = [K_X(x)]^o = N_X(x). \tag{52}$$

B. Robinson's Condition

Consider the problem

$$\begin{aligned} &\min f(\boldsymbol{x}) \\ &\text{s.t. } g_i(\boldsymbol{x}) \leq 0, i = 1, \ldots, m, \\ &\quad h_i(\boldsymbol{x}) = 0, i = 1, \ldots, p, \\ &\quad \boldsymbol{x} \in X \end{aligned} \tag{53}$$

with continuously differentiable functions $f : \mathbb{R}^n \to \mathbb{R}$, $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^m$, $\boldsymbol{h} : \mathbb{R}^n \to \mathbb{R}^p$ and with a closed convex set X. We consider a feasible point $\boldsymbol{x}_0$ of problem (53) and define the set of active inquality constraints:

$$I(\boldsymbol{x}_0) = \{1 \leq i \leq m : g_i(\boldsymbol{x}_0) = 0\}.$$

Robinson's condition with respect to the constraint set of problem (53) takes on the form

$$\left\{ \begin{pmatrix} \nabla \boldsymbol{h}(\boldsymbol{x}_0)\boldsymbol{d} \\ \nabla \boldsymbol{g}(\boldsymbol{x}_0)\boldsymbol{d} - \boldsymbol{v} \end{pmatrix} \middle| \boldsymbol{d} \in T_X(\boldsymbol{x}_0), \boldsymbol{v} \in \mathbb{R}^m, \right. \\ \left. v_i \leq 0, i \in I(\boldsymbol{x}_0) \right\} = \mathbb{R}^p \times \mathbb{R}^m. \tag{54}$$

Let Z denote the feasible set of problem (53). Then we have

Theorem A.2: [49, Theorem 3.15] If Robinson's condition holds for problem (53) at $\boldsymbol{x}_0$, then $T_Z(\boldsymbol{x}_0)$ takes the form

$$T_Z(\boldsymbol{x}_0) = \{\boldsymbol{d} \in \mathbb{R}^n \mid \boldsymbol{d} \in T_X(\boldsymbol{x}_0), \nabla \boldsymbol{h}(\boldsymbol{x}_0)\boldsymbol{d} = \boldsymbol{0}, \\ \nabla g_i(\boldsymbol{x}_0)^T \boldsymbol{d} \leq 0, i \in I(\boldsymbol{x}_0)\}. \tag{55}$$

C. The Boundedness of Lagrange Multipliers

The following theorem gives a necessary optimality condition (i.e., KKT conation) for problem (53) with continuously differentiable function $f(\cdot)$.

Theorem A.3: [49, Theorem 3.25] Let $\hat{\boldsymbol{x}}$ be a local minimum of problem (53). Assume that at $\hat{\boldsymbol{x}}$ the constraint qualification condition[3] is satisfied for problem (53). Then there exist multipliers $\hat{\lambda}_i \geq 0, i = 1, \ldots, m$, and $\hat{\mu}_i \in \mathbb{R}, i = 1, \ldots, p$, such that

$$0 \in \nabla f(\hat{\boldsymbol{x}}) + \sum_{i=1}^{m} \hat{\lambda}_i \nabla g_i(\hat{\boldsymbol{x}}) + \sum_{i=1}^{p} \hat{\mu}_i \nabla h_i(\hat{\boldsymbol{x}}) + N_X(\hat{\boldsymbol{x}}), \tag{56}$$

and

$$\hat{\lambda}_i g_i(\hat{\boldsymbol{x}}) = 0, i = 1, \ldots, m. \tag{57}$$

[3]If any of the sufficient conditions for (55) is satisfied, we say that problem (53) satisfies the constraint qualification condition.

Furthermore, it is shown in the following lemma that, once (56) and (57) are satisfied, the corresponding Lagrange multipliers are all bounded under Robinson's condition.

Lemma A.2: [49, Lemma 3.26] Let $\hat{\boldsymbol{x}}$ be a local minimum of problem (53) and let $\hat{\Lambda}(\hat{\boldsymbol{x}})$ be the set of Lagrange multipliers $\hat{\lambda} \in \mathbb{R}^m_+$ and $\hat{\mu} \in \mathbb{R}^p$ satisfying (56), (57).

1) The set $\hat{\Lambda}(\hat{\boldsymbol{x}})$ is convex and closed.
2) If problem (53) satisfies Robinson's condition at $\hat{\boldsymbol{x}}$, then the set $\hat{\Lambda}(\hat{\boldsymbol{x}})$ is also bounded.

APPENDIX B
SOME PROOFS

A. The Proof of Theorem 2.1

Proof: Since $\hbar(\boldsymbol{x})$ is Lipschitz near $\boldsymbol{x}_0$, we have $\hbar^o(\boldsymbol{x}_0; \boldsymbol{0}) = 0$ and $\|\boldsymbol{\xi}\| \leq K$ for all $\boldsymbol{\xi} \in \bar{\partial}\hbar(\boldsymbol{x}_0)$ by [47, Lemma 2.6]. Moreover, it is known from [47, Lemma 2.6] that $\hbar^o(\boldsymbol{x}_0; \boldsymbol{d})$ is a convex function with respect to $\boldsymbol{d}$. It follows that[4] $\bar{\partial}\hbar(\boldsymbol{x}_0) = \partial_{\boldsymbol{d}}\hbar^o(\boldsymbol{x}_0; \boldsymbol{0})$ is not empty and compact [51, Lemma 2.16 & Theorem 2.15], and moreover $\hbar^o(\boldsymbol{x}_0; \boldsymbol{d}) = \sup_{\boldsymbol{\xi} \in \bar{\partial}\hbar(\boldsymbol{x}_0)} \boldsymbol{\xi}^T \boldsymbol{d}$, $\forall \boldsymbol{d}$ [47, Theorem 2.5], implying part 4). This completes the proof. ■

B. The Proof of Theorem 2.3

Proof: The proof is divided into two steps. We first establish a necessary optimality condition, which is then shown to be equivalent to the KKT system.

Step 1: Recall that $F(\boldsymbol{x}, \boldsymbol{y}) \triangleq f(\boldsymbol{x}, \boldsymbol{y}) + \sum_{j=1}^{n_y} \phi_j(s_j(\boldsymbol{y}_j))$ is the objective function of problem (P) (5). In the first step, we show that a local optimal solution point $\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}$ must satisfy the following condition

$$F^o(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}; \boldsymbol{d}) \geq 0, \quad \forall \boldsymbol{d} \in T_{\mathcal{Z}}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}).$$

Assume on the contrary that there exists a direction $\boldsymbol{d} \in T_{\mathcal{Z}}(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})$ such that $F^o(\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}; \boldsymbol{d}) < 0$. Because $\boldsymbol{d}$ is a tangent direction, there exists a sequence $\boldsymbol{z}^k \triangleq (\boldsymbol{x}^k, \boldsymbol{y}^k) \in \mathcal{Z}$ converging to $\hat{\boldsymbol{z}} \triangleq (\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}})$ and a sequence of nonnegative scalars $\tau^k \to 0$ as $k \to \infty$, such that [49, Def. 3.11]

$$\lim_{k \to \infty} \frac{\boldsymbol{z}^k - \hat{\boldsymbol{z}}}{\tau^k} = \boldsymbol{d}.$$

It follows that

$$\lim_{k \to \infty} \frac{\boldsymbol{w}^k}{\|\boldsymbol{w}^k\|} = \frac{\boldsymbol{d}}{\|\boldsymbol{d}\|}, \text{ where } \boldsymbol{w}^k \triangleq \boldsymbol{z}^k - \hat{\boldsymbol{z}}. \tag{58}$$

Define a sequence $\{\delta_k\}$ such that the following conditions are satisfied

$$\delta_k > \|\boldsymbol{w}^k\|, \quad \forall k \quad \text{and} \quad \lim_{k \to \infty} \delta_k \to 0. \tag{59}$$

Then we have

$$\begin{aligned} &\limsup_{k \to \infty} \frac{F(\boldsymbol{z}^k) - F(\hat{\boldsymbol{z}})}{\|\boldsymbol{z}^k - \hat{\boldsymbol{z}}\|} \\ &= \limsup_{k \to \infty} \frac{F(\hat{\boldsymbol{z}} + \boldsymbol{w}^k) - F(\hat{\boldsymbol{z}})}{\|\boldsymbol{w}^k\|} \end{aligned}$$

[4]Considering that $\hbar^o(\boldsymbol{x}_0; \boldsymbol{d})$ is a convex function with respect to $\boldsymbol{d}$, we use $\partial_{\boldsymbol{d}}\hbar^o(\boldsymbol{x}_0; \boldsymbol{0})$ to denote its subdifferential evaluated at $\boldsymbol{d} = \boldsymbol{0}$.

$$\overset{\text{(i)}}{\leq} \limsup_{k\to\infty} \sup_{\substack{\boldsymbol{u}\in B_{\delta_k}(\mathbf{0}),\\ \lambda\in(0,\delta_k)}} \frac{F(\hat{\boldsymbol{z}}+\boldsymbol{u}+\lambda\frac{\boldsymbol{w}^k}{\|\boldsymbol{w}^k\|})-F(\hat{\boldsymbol{z}}+\boldsymbol{u})}{\lambda}$$

$$\overset{\text{(ii)}}{=} \lim_{k\to\infty} \sup_{\substack{\boldsymbol{u}\in B_{\delta_k}(\mathbf{0}),\\ \lambda\in(0,\delta_k)}} \frac{F(\hat{\boldsymbol{z}}+\boldsymbol{u}+\lambda\frac{\boldsymbol{w}^k}{\|\boldsymbol{w}^k\|})-F(\hat{\boldsymbol{z}}+\boldsymbol{u})}{\lambda}$$

$$= F^o\left(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}};\frac{\boldsymbol{d}}{\|\boldsymbol{d}\|}\right) \overset{\text{(iii)}}{<} 0,$$

where (i) is due to the fact that $\|\boldsymbol{w}_k\|<\delta_k$ and $\mathbf{0}\in B_{\delta_k}(\mathbf{0})$; (ii) follows from the existence of the limit, and (iii) is due to the the assumption $F^o(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}};\boldsymbol{d})<0$ as well as the positive homogeneity of generalized gradient [47, Lemma 2.6], i.e., $F^o(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}};\alpha\boldsymbol{d})=\alpha F^o(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}};\boldsymbol{d})$ for all $\alpha\geq 0$. The above result contradicts to the fact that $(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}})$ is a local optimum. Hence, for any local optimum $(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}})$, we must have $F^o(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}};\boldsymbol{d})\geq 0,\ \forall \boldsymbol{d}\in T_{\mathcal{Z}}(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}})$.

Step 2: Based on the necessary optimality condition established in the first step, we then show that the KKT system holds for a locally optimal solution $(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}})$. First, by noting that $F(\boldsymbol{x},\boldsymbol{y})$ is locally Lipschitz near $(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}})$ (see the arguments under Definition 2.1) and using the result of Part 4) of Theorem 2.1, we have, $\exists \boldsymbol{\xi}\in\bar{\partial}F(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}})$ such that

$$\boldsymbol{\xi}^T\boldsymbol{d}=F^o(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}};\boldsymbol{d})\geq 0,\ \forall \boldsymbol{d}\in T_{\mathcal{Z}}(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}}). \tag{60}$$

Recall the definition of polar cone (see Appendix A). Eq. (60) can be equivalently expressed as: $-\boldsymbol{\xi}\in(T_{\mathcal{Z}}(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}}))^o$. Define

$$\mathbf{A}\triangleq\begin{pmatrix}\nabla \boldsymbol{h}(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}})\\ \left[\text{blkdiag}\{\nabla \boldsymbol{g}_i(\boldsymbol{x}_i)^T\}_i\mathbf{0}_{q\times M}\right]\end{pmatrix},$$
$$\mathcal{K}_1\triangleq T_{\mathcal{X}}(\hat{\boldsymbol{x}})\times\mathbb{R}^M,\mathcal{K}_2\triangleq\{0\}^p\times\mathbb{R}^q_-$$

where the notation $\text{blkdiag}\{\nabla \boldsymbol{g}_i(\boldsymbol{x}_i)^T\}_i$ denotes a q by N matrix which is block diagonal concatenation of matrices $\nabla \boldsymbol{g}_i(\boldsymbol{x}_i)^T$, $i=1,2,\ldots,n$, that is,

$$\begin{aligned}&\text{blkdiag}\{\nabla \boldsymbol{g}_i(\boldsymbol{x}_i)^T\}_i\\ &\triangleq\begin{pmatrix}\nabla \boldsymbol{g}_1(\boldsymbol{x}_1)^T & & & \\ & \nabla \boldsymbol{g}_2(\boldsymbol{x}_2)^T & & \\ & & \ddots & \\ & & & \nabla \boldsymbol{g}_n(\boldsymbol{x}_n)^T\end{pmatrix}\end{aligned} \tag{61}$$

Assume for simplicity $I_i(\hat{\boldsymbol{x}}_i)=\{1,2,\ldots,q_i\}$, then $T_{\mathcal{Z}}(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}})$ defined by (4) can be compactly expressed as

$$T_{\mathcal{Z}}(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}})=\{\boldsymbol{d}\in\mathcal{K}_1|\mathbf{A}\boldsymbol{d}\in\mathcal{K}_2\}.$$

Moreover, Robinson's condition (3) is equivalent to [49, pp. 102]

$$\mathbf{0}\in int\left(\{\mathbf{A}\boldsymbol{\theta}-\boldsymbol{\eta}:\boldsymbol{\theta}\in\mathcal{K}_1,\boldsymbol{\eta}\in\mathcal{K}_2\}\right).$$

It follows that (see [49, Theorem 2.36], or Theorem A.1)

$$-\boldsymbol{\xi}\in(T_{\mathcal{Z}}(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}}))^o=\mathcal{K}_1^o+\{\mathbf{A}^T\hat{\boldsymbol{\lambda}}:\hat{\boldsymbol{\lambda}}\in\mathcal{K}_2^o\} \tag{62}$$

where $\mathcal{K}_1^o=N_{\mathcal{X}}(\hat{\boldsymbol{x}})\times\{0\}^M$ and $\mathcal{K}_2^o=\mathbb{R}^p\times\mathbb{R}^q_+$ are obtained by the definition of polar cone; see Appendix A. Eq. (62) is further equivalent to, $\exists\hat{\boldsymbol{\mu}}\in\mathbb{R}^p$, $\hat{\boldsymbol{v}}_i\in\mathbb{R}^{q_i}_+$, and $\hat{\boldsymbol{\lambda}}\triangleq(\hat{\boldsymbol{\mu}},(\hat{\boldsymbol{v}}_i)_i)$ such that

$$-\mathbf{A}^T\hat{\boldsymbol{\lambda}}\in\mathcal{K}_1^o+\bar{\partial}F(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}}). \tag{63}$$

By using the following facts

$$\bar{\partial}F(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}})=\prod_{i=1}^n\{(\nabla_{\boldsymbol{x}_i}f(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}}))\}\times\prod_{j=1}^{n_y}\bar{\partial}\tilde{\phi}(\boldsymbol{y}_j) \tag{64}$$

$$\begin{aligned}&\mathbf{A}^T\hat{\boldsymbol{\lambda}}=\\ &\left((\nabla_{\boldsymbol{x}_i}\boldsymbol{h}(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}})^T\hat{\boldsymbol{\mu}}+\nabla_{\boldsymbol{x}_i}\boldsymbol{g}_i(\hat{\boldsymbol{x}})^T\hat{\boldsymbol{v}}_i)_i,(\nabla_{\boldsymbol{y}_j}\boldsymbol{h}(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}})^T\hat{\boldsymbol{\mu}})_j\right)\end{aligned} \tag{65}$$

$$\mathcal{K}_1^o=\prod_i^n N_{\mathcal{X}_i}(\hat{\boldsymbol{x}})\times\prod_j^{n_y}\{0\}^{m_j} \tag{66}$$

we can recast Eq. (63) into Eqs. (5a) and (5b). Now let us take $I_i(\hat{\boldsymbol{x}}_i)$ into consideration. By the definition of polar cone, we have $\boldsymbol{v}_{i,\ell}=0$ for $i\notin I_i(\hat{\boldsymbol{x}}_i)$. Thus Eq. (5c) follows. The rest of equations in KKT system (5) are trivial. This completes the proof. ■

APPENDIX C
PROOF OF THEOREM 4.2

Proof: First, let us fix some notations. For simplicity, let us assume that at a given iteration k that we are going to focus on, the order of update is fixed as $1,2,\cdots,n_z$. Let us define

$$\begin{aligned}\boldsymbol{w}^{k,1}&=\boldsymbol{z}^k,\\ \boldsymbol{w}^{k,i+1}&=[\boldsymbol{w}^{k,i}_{<i},\boldsymbol{z}^{k+1}_i,\boldsymbol{w}^{k,i}_{>i}],\ \forall i,\\ \boldsymbol{w}^{k,n_z+1}&=\boldsymbol{z}^{k+1}.\end{aligned}$$

That is, $\boldsymbol{w}^{k,i}$ is the point where block i constructs its upper bound. Further define:

$$\begin{aligned}\tilde{\boldsymbol{y}}^k_j\triangleq\arg\min_{\boldsymbol{y}_j}\Big\{&s_j(\boldsymbol{y}_j)+\nabla_{\boldsymbol{y}_j}\tilde{\mathcal{L}}(\boldsymbol{x}^k,\boldsymbol{y}^k)(\boldsymbol{y}_j-\boldsymbol{y}^k_j)\\ &+\frac{1}{2}\|\boldsymbol{y}_j-\boldsymbol{y}^k_j\|^2\Big\},\forall j\in[1:n_y].\end{aligned}$$

Note that the computation of $\tilde{\boldsymbol{y}}^k_j$ is called the *proximity operator*, and for typical nonsmooth regularizers (such as ℓ_1 norm), the above problem has closed-form solutions.

Now we can simplify $\boldsymbol{\Delta}^k_j$ in (9) as the following:

$$\boldsymbol{\Delta}^k_j=\boldsymbol{y}^k_j-\tilde{\boldsymbol{y}}^k_j. \tag{67}$$

Second, for a given $\boldsymbol{x}_i$, when it is optimized at iteration k, it must satisfy

$$\begin{aligned}\boldsymbol{x}^{k+1}_i&=\mathcal{P}_{\tilde{\mathcal{X}}_i}(\boldsymbol{x}^{k+1}_i-\nabla u_i(\boldsymbol{x}^{k+1}_i;\boldsymbol{w}^{k,i}))\\ &=\mathcal{P}_{\mathcal{X}_i}(\boldsymbol{x}^{k+1}_i-\nabla u_i(\boldsymbol{x}^{k+1}_i;\boldsymbol{w}^{k,i})-\nabla g_i(\boldsymbol{x}^{k+1}_i)^T\boldsymbol{v}^{k+1}_i)\end{aligned} \tag{68}$$

where $\boldsymbol{v}^{k+1}_i$ is the dual variable corresponding to the constraint $g_i(\boldsymbol{x}_i)\leq 0$. Therefore, when block $i\in[1:n]$ has been selected to update at iteration $k+1$, we have

$$\begin{aligned}\|\boldsymbol{e}^{k+1}_i\|\overset{(8)}{=}\|&\boldsymbol{x}^{k+1}_i-\mathcal{P}_{\mathcal{X}_i}(\boldsymbol{x}^{k+1}_i-\nabla_{\boldsymbol{x}_i}\mathcal{L}(\boldsymbol{x}^{k+1},\boldsymbol{y}^{k+1})\\ &-\nabla g_i(\boldsymbol{x}^{k+1}_i)^T\boldsymbol{v}^{k+1}_i))\|\end{aligned}$$

$$\begin{aligned}
&\leq \|\nabla_{\boldsymbol{x}_i}\mathcal{L}(\boldsymbol{x}^{k+1},\boldsymbol{y}^{k+1}) - \nabla u_i(\boldsymbol{x}_i^{k+1};\boldsymbol{w}^{k,i})\| \\
&= \|\nabla u_i(\boldsymbol{x}_i^{k+1};\boldsymbol{z}^{k+1}) - \nabla u_i(\boldsymbol{x}_i^{k+1};\boldsymbol{w}^{k,i})\| \\
&\overset{(45b)}{\leq} H_i\|\boldsymbol{z}^{k+1} - \boldsymbol{w}^{k,i}\| \\
&\leq H_i\|\boldsymbol{z}^{k+1} - \boldsymbol{z}^{k}\|,\ \forall i \in [1:n] \qquad (69)
\end{aligned}$$

where the first inequality uses (68), as well as the non-expansiveness of the projection operator; the equality comes from (28c); the last inequality comes from the definition of $\boldsymbol{w}^{k,i}$.

Next, by applying the same line of argument as above on $\boldsymbol{\Delta}_j^{k+1}$, and by using the expression (67) and the non-expansiveness of the proximity operator, we can also obtain that, if $\boldsymbol{y}_j$ is updated, then

$$\|\boldsymbol{\Delta}_j^{k+1}\| \leq H_{n+j}\|\boldsymbol{z}^{k+1} - \boldsymbol{z}^{k}\|, \forall j \in [1:n_y]. \qquad (70)$$

Since all the blocks are updated once at iteration $k+1$, we have

$$(\|\boldsymbol{e}^{k+1}\| + \|\boldsymbol{\Delta}^{k+1}\|) \leq Q\|\boldsymbol{z}^{k+1} - \boldsymbol{z}^{k}\|, \qquad (71)$$

where $Q \triangleq \sqrt{\sum_{i=1}^{n_z} H_i^2}$. Based on the above properties, let us analyze the convergence rate of rBSUM.

Let us first investigate the optimality conditions for $\boldsymbol{z}_i^{k+1}$, $i = 1, \cdots, n_z$. If $i \in [1:n]$, then $\boldsymbol{z}_i$'s are optimized based on the following problem:

$$\min\ q_i(\boldsymbol{z}_i;\boldsymbol{w}^{k,i})\ \ \text{s.t.}\ \ \boldsymbol{z}_i \in \tilde{\mathcal{X}}_i. \qquad (72)$$

Then we have the following optimality condition

$$\langle \nabla q_i(\boldsymbol{z}_i^{k+1};\boldsymbol{w}^{k,i}), \boldsymbol{z}_i - \boldsymbol{z}_i^{k+1}\rangle \geq 0,\ \forall \boldsymbol{z}_i \in \tilde{\mathcal{X}}_i \qquad (73)$$

where we have used the fact that $\boldsymbol{g}_i(\boldsymbol{z}_i)$ is a convex function, so $\tilde{\mathcal{X}}_i$ is a convex set for all i.

If $i \in [n{+}1 : n_z]$, then $\boldsymbol{z}_i^{k+1}$'s are optimized based on the following problem:

$$\min\ q_i(\boldsymbol{z}_i;\boldsymbol{w}^{k,i}) + s_{i-n}(\boldsymbol{z}_i) \qquad (74)$$

and the optimality condition is:

$$\nabla q_i(\boldsymbol{z}_i^{k+1};\boldsymbol{w}^{k,i}) + \xi_i^{k+1} = 0 \qquad (75)$$

where ξ_i^{k+1} is a vector belonging to the subdifferential set of $s_{i-n}(\boldsymbol{z}_i^{k+1})$. Using the fact that for each $j \in [1:n_y]$, $s_j(\cdot)$ is a convex function, we further have for $i \in [n{+}1 : n_z]$

$$s_{i-n}(\boldsymbol{z}_i) - s_{i-n}(\boldsymbol{z}_i^{k+1}) + \langle \nabla q_i(\boldsymbol{z}_i^{k+1};\boldsymbol{w}^{k,i}), \boldsymbol{z}_i - \boldsymbol{z}_i^{k+1}\rangle \geq 0. \qquad (76)$$

Using (73) as well as the strong convexity assumption of $q_i(\cdot)$ in (45a), we have

$$\begin{aligned}
&q_i(\boldsymbol{z}_i^{k};\boldsymbol{w}^{k,i}) - q_i(\boldsymbol{z}_i^{k+1};\boldsymbol{w}^{k,i}) \\
&\geq \langle \nabla q_i(\boldsymbol{z}_i^{k+1};\boldsymbol{w}^{k,i}), \boldsymbol{z}_i^{k} - \boldsymbol{z}_i^{k+1}\rangle + \frac{\theta}{2}\|\boldsymbol{z}_i^{k+1} - \boldsymbol{z}_i^{k}\|^2 \\
&\geq \frac{\theta}{2}\|\boldsymbol{z}_i^{k+1} - \boldsymbol{z}_i^{k}\|^2,\ \forall i \in [1:n]. \qquad (77)
\end{aligned}$$

Similarly (i.e., using (76) and (45a))

$$\begin{aligned}
&(s_{i-n}(\boldsymbol{z}_i^{k}) + q_i(\boldsymbol{z}_i^{k};\boldsymbol{w}^{k,i})) - (s_{i-n}(\boldsymbol{z}_i^{k+1}) + q_i(\boldsymbol{z}_i^{k+1};\boldsymbol{w}^{k,i})) \\
&\geq \frac{\theta}{2}\|\boldsymbol{z}_i^{k+1} - \boldsymbol{z}_i^{k}\|^2,\ \forall i \in [n{+}1 : n_z]. \qquad (78)
\end{aligned}$$

Therefore, after each update of $i \in [1:n]$, we can estimate the descent of the objective:

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{w}^{k,i+1}) - \mathcal{L}(\boldsymbol{w}^{k,i}) \\
&\leq -\left(\tilde{\mathcal{L}}(\boldsymbol{w}^{k,i}) - \min_{\boldsymbol{z}_{i\in\tilde{\mathcal{X}}_i}} q_i(\boldsymbol{z}_i;\boldsymbol{w}^{k,i})\right) \\
&= -\left(q_i(\boldsymbol{z}_i^{k};\boldsymbol{w}^{k,i}) - q_i(\boldsymbol{z}_i^{k+1};\boldsymbol{w}^{k,i})\right) \\
&\leq -\frac{\theta}{2}\|\boldsymbol{z}_i^{k} - \boldsymbol{z}_i^{k+1}\|^2,\ \forall i \in [1:n],
\end{aligned}$$

where the first inequality is due to (44b), and the equality is due to (44a) and the definition of $\boldsymbol{z}_i^{k+1}$.

Similarly, after each update of $i \in [n{+}1 : n_z]$, we can estimate the descent of the objective:

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{w}^{k,i+1}) - \mathcal{L}(\boldsymbol{w}^{k,i}) \\
&\leq -\left(\mathcal{L}(\boldsymbol{w}^{k,i}) - \min_{\boldsymbol{z}_i}(s_{i-n}(\boldsymbol{z}_i) + q_i(\boldsymbol{z}_i;\boldsymbol{w}^{k,i}))\right) \\
&= -\left(q_i(\boldsymbol{z}_i^{k};\boldsymbol{w}^{k,i}) + s_{i-n}(\boldsymbol{z}_i^{k}) - \left(q_i(\boldsymbol{z}_i^{k+1};\boldsymbol{w}^{k,i}) + s_{i-n}(\boldsymbol{z}_i^{k+1})\right)\right) \\
&\leq -\frac{\theta}{2}\|\boldsymbol{z}_i^{k} - \boldsymbol{z}_i^{k+1}\|^2,\ \forall i \in [n{+}1 : n_z].
\end{aligned}$$

Overall, by summing up all the indices from $i = 1$ to n_z, and use the definition of $\boldsymbol{z}^k = \boldsymbol{w}^{k,1}$ and $\boldsymbol{z}^{k+1} = \boldsymbol{w}^{k,n_z+1}$ we obtain

$$\mathcal{L}(\boldsymbol{z}^{k+1}) - \mathcal{L}(\boldsymbol{z}^{k}) \leq -\frac{\theta}{2}\|\boldsymbol{z}^{k} - \boldsymbol{z}^{k+1}\|^2.$$

Clearly, the above result is not dependent on how the set $\mathcal{I}_k$ is chosen, so it holds true for all k. Using the telescope sum from $k = 1$ to T, we obtain the following

$$\sum_{k=1}^{T}\|\boldsymbol{z}^{k} - \boldsymbol{z}^{k+1}\|^2 \leq -\frac{2}{\theta}(\mathcal{L}(\boldsymbol{z}^{T}) - \mathcal{L}(\boldsymbol{z}^{0})). \qquad (79)$$

Since $\mathcal{L}(\cdot)$ is bounded from below, we use $\underline{\mathcal{L}}$ to denote its lower bound. Dividing both sides by T, we have

$$\frac{1}{T}\sum_{k=1}^{T}\|\boldsymbol{z}^{k} - \boldsymbol{z}^{k+1}\|^2 \leq \frac{2}{\theta T}(\mathcal{L}(\boldsymbol{z}^{0}) - \underline{\mathcal{L}}). \qquad (80)$$

Utilizing (71), we obtain

$$\begin{aligned}
&\min_{k=1:T} \max\{\|\boldsymbol{\Delta}^{k+1}\|_\infty^2, \|\boldsymbol{e}^{k+1}\|_\infty^2\} \\
&\leq \frac{1}{T}\sum_{k=1}^{T}\max\{\|\boldsymbol{\Delta}^{k+1}\|_\infty^2, \|\boldsymbol{e}^{k+1}\|_\infty^2\}
\end{aligned}$$

$$\leq \frac{1}{T}\sum_{k=1}^{T}\|\boldsymbol{\Delta}^{k+1}\|_\infty^2 + \|\boldsymbol{e}^{k+1}\|_\infty^2$$

$$\leq \frac{c}{T}\sum_{k=1}^{T}\|\boldsymbol{\Delta}^{k+1}\|^2 + \|\boldsymbol{e}^{k+1}\|^2$$

$$\leq cQ^2\frac{1}{T}\sum_{k=1}^{T}\|\boldsymbol{z}^k - \boldsymbol{z}^{k+1}\|^2$$

$$\leq \frac{2cQ^2}{\theta T}(\mathcal{L}(\boldsymbol{z}^0) - \underline{\mathcal{L}}). \tag{81}$$

Here c is a constant depending on the problem dimension, which relates ℓ_2 norm and ℓ_∞ norm. Therefore, we complete the proof by letting $d = \frac{2cQ^2}{\theta}(\mathcal{L}(\boldsymbol{z}^0) - \underline{\mathcal{L}})$. ■

REFERENCES

[1] Q. Shi and M. Hong, "Penalty dual decomposition with application in signal processing," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2017, pp. 1–5.

[2] K. T. Truong, P. Sartori, and R. W. Heath, "Cooperative algorithms for MIMO amplify-and-forward relay networks," *IEEE Trans. Signal Process.*, vol. 61, no. 5, pp. 1272–1287, Mar. 2013.

[3] Q. Shi, M. Hong, X. Gao, E. Song, Y. Cai, and W. Xu, "Joint source-relay design for full–duplex MIMO AF relay systems," *IEEE Trans. Signal Process.*, vol. 62, no. 23, pp. 6118–6131, Dec. 2016.

[4] Y. Rong, "Joint source and relay optimization for two-way linear non-regenerative MIMO relay communications," *IEEE Trans. Signal Process.*, vol. 60, no. 12, pp. 6533–6546, Dec. 2012.

[5] W. C. Liao, M. Hong, H. Farmanbar, X. Li, Z. Q. Luo, and H. Zhang, "Min flow rate maximization for software defined radio access networks," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1282–1294, Jun. 2014.

[6] N. Vucic and H. Boche, "Robust QoS-constrained optimization of downlink multiuser MISO systems," *IEEE Trans. Signal Process.*, vol. 57, no. 2, pp. 714–725, Feb. 2009.

[7] W. C. Liao, M. Hong, Y. F. Liu, and Z. Q. Luo, "Base station activation and linear transceiver design for optimal resource management in heterogeneous networks," *IEEE Trans. Signal Process.*, vol. 62, no. 15, pp. 3939–3952, Aug. 2014.

[8] Q. Shi, M. Razaviyayn, M. Hong, and Z. Q. Luo, "SINR constrained beamforming for a MIMO multi-user downlink system: Algorithms and convergence analysis," *IEEE Trans. Signal Process.*, vol. 64, no. 11, pp. 2920–2933, Jun. 2016.

[9] R. Zhang, C. C. Chai, and Y. C. Liang, "Joint beamforming and power control for multiantenna relay broadcast channel with QoS constraints," *IEEE Trans. Signal Process.*, vol. 57, no. 2, pp. 726–737, Feb. 2009.

[10] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[11] B. Friedlander and T. Strohmer, "Bilinear compressed sensing for array self-calibration," in *Proc. 48th Asilomar Conf. Signals, Syst. Comput.*, Nov. 2014, pp. 363–367.

[12] D. L. Sun and C. Fevotte, "Alternating direction method of multipliers for non-negative matrix factorization with the beta-divergence," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2014, pp. 6201–6205.

[13] Y. Xu, W. Yin, Z. Wen, and Y. Zhang, "An alternating direction algorithm for matrix completion with nonnegative factors," *Frontiers Math. China*, vol. 7, no. 2, pp. 365–384, 2012.

[14] D. Hajinezhad, T. H. Chang, X. Wang, Q. Shi, and M. Hong, "Nonnegative matrix factorization using ADMM: Algorithm and convergence analysis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Mar. 2016, pp. 4742–4746.

[15] X. Fu, W. K. Ma, K. Huang, and N. D. Sidiropoulos, "Blind separation of quasi-stationary sources: Exploiting convex geometry in covariance domain," *IEEE Trans. Signal Process.*, vol. 63, no. 9, pp. 2306–2320, May 2015.

[16] B. Dai and W. Yu, "Sparse beamforming and user-centric clustering for downlink cloud radio access network," *IEEE Access*, vol. 2, pp. 1326–1339, Nov. 2014.

[17] D. Kuang, S. Yun, and H. Park, "SymNMF: Nonnegative low-rank approximation of a similarity matrix for graph clustering," *J. Global Optim.*, vol. 62, no. 3, pp. 545–574, 2015.

[18] Q. T. Dinh, S. Gumussoy, W. Michiels, and M. Diehl, "Combining convex-concave decompositions and linearization approaches for solving BMIs, with application to static output feedback," *IEEE Trans. Autom. Control*, vol. 57, no. 6, pp. 1377–1390, Jun. 2012.

[19] J. Wang and J. Q. Zhang, "A globally optimal bilinear programming approach to the design of approximate Hilbert pairs of orthonormal wavelet bases," *IEEE Trans. Signal Process.*, vol. 58, no. 1, pp. 233–241, Jan. 2010.

[20] M. Hong, M. Razaviyayn, Z.-Q. Luo, and J.-S. Pang, "A unified algorithmic framework for block-structured optimization involving big data: With applications in machine learning and signal processing," *IEEE Signal Process. Mag.*, vol. 33, no. 1, pp. 57–77, Jan. 2016.

[21] U. Rashid, H. D. Tuan, H. H. Kha, and H. H. Nguyen, "Joint optimization of source precoding and relay beamforming in wireless MIMO relay networks," *IEEE Trans. Commun.*, vol. 62, no. 2, pp. 488–499, Feb. 2014.

[22] D. Bertsekas, *Nonlinear Programming*, 2nd ed., Belmont, MA, USA: Athena Scientific, 1999.

[23] Z. Lu, Y. Zhang, and X. Li, "Penalty decomposition methods for rank minimization," *Optim. Methods Softw.*, vol. 30, no. 3, pp. 531–558, 2015.

[24] M. R. Hestenes, "Multiplier and gradient methods," *J. Optim. Theory Appl.*, vol. 4, no. 5, pp. 303–320, 1969.

[25] M. J. D. Powell, "A method for nonlinear constraints in minimization problems," in *Optimization*, R. Fletcher, Ed. New York, NY, USA: Academic Press, 1969, pp. 283–298.

[26] M. Kočvara and M. Stingl, *PENNON: A Generalized Augmented Lagrangian Method for Semidefinite Programming*. Boston, MA, USA: Springer, 2003, pp. 303–321.

[27] A. F. Izmailov and M. V. Solodov, "On attraction of linearly constrained lagrangian methods and of stabilized and quasi-newton SQP methods to critical multipliers," *Math. Programm.*, vol. 126, no. 2, pp. 231–257, 2011.

[28] D. Fernandez and M. V. Solodov, "Local convergence of exact and inexact augmented Lagrangian methods under the second-order sufficient optimality condition," *SIAM J. Optim.*, vol. 22, no. 2, pp. 384–407, 2012.

[29] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan. 2011.

[30] T. H. Chang, M. Hong, and X. Wang, "Multi-agent distributed optimization via inexact consensus ADMM," *IEEE Trans. Signal Process.*, vol. 63, no. 2, pp. 482–497, Jan. 2015.

[31] P. A. Thouvenin, N. Dobigeon, and J. Y. Tourneret, "Hyperspectral unmixing with spectral variability using a perturbed linear mixing model," *IEEE Trans. Signal Process.*, vol. 64, no. 2, pp. 525–538, Jan. 2016.

[32] C. Shen, T. H. Chang, K. Y. Wang, Z. Qiu, and C. Y. Chi, "Distributed robust multicell coordinated beamforming with imperfect CSI: An ADMM approach," *IEEE Trans. Signal Process.*, vol. 60, no. 6, pp. 2988–3003, Jun. 2012.

[33] M. Leinonen, M. Codreanu, and M. Juntti, "Distributed joint resource and routing optimization in wireless sensor networks via alternating direction method of multipliers," *IEEE Trans. Wireless Commun.*, vol. 12, no. 11, pp. 5454–5467, Nov. 2013.

[34] K. Huang, N. D. Sidiropoulos, and A. P. Liavas, "A flexible and efficient algorithmic framework for constrained matrix and tensor factorization," *IEEE Trans. Signal Process.*, vol. 64, no. 19, pp. 5052–5065, Oct. 2016.

[35] C. Zhang, H. Lee, and K. Shin, "Efficient distributed linear classification algorithms via the alternating direction method of multipliers," *J. Mach. Learn. Res.*, vol. 22, pp. 1398–1406, 2012.

[36] M. Hong, Z.-Q. Luo, and M. Razaviyayn, "Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems," *SIAM J. Optim.*, vol. 26, no. 1, pp. 337–364, 2016.

[37] B. Ames and M. Hong, "Alternating directions method of multipliers for l1-penalized zero variance discriminant analysis and principal component analysis," *Comput. Optim. Appl.*, vol. 64, no. 3, pp. 725–754, 2016.

[38] Y. Wang and J. Z. W. Yin, "Global convergence of ADMM in nonconvex nonsmooth optimization," *J. Scientific Comput.*, vol. 78, pp. 29–63, 2019.

[39] G. Li and T.-K. Pong, "Splitting methods for nonconvex composite optimization," *SIAM J. Optim.*, vol. 25, no. 4, pp. 2434–2460, Dec. 2015.

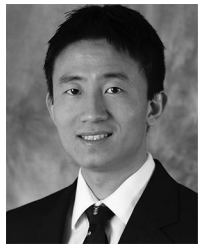
[40] M. Xu, J. Ye, and L. Zhang, "Smoothing SQP methods for solving degenerate nonsmooth constrained optimization problems with applications to bilevel programs," *SIAM J. Optim.*, vol. 25, no. 3, pp. 1388–1410, 2015.

[41] J. Bolte, S. Sabach, and M. Teboulle, "Proximal alternating linearized minimization for nonconvex and nonsmooth problems," *Math. Program.*, vol. 146, pp. 459–494, 2014.

[42] J.-S. Pang, M. Razaviyayn, and A. Alvarado, "Computing B-stationary points of nonsmooth DC programs," *Math. Oper. Res.*, vol. 42, no. 1, pp. 95–118, Jan. 2017.
[43] C. Cartis, N. I. M. Gould, and P. L. Toint, "On the complexity of steepest descent, Newton's and regularized newton's methods for nonconvex unconstrained optimization," *SIAM J. Optim.*, vol. 20, no. 6, pp. 2833–2852, 2010.
[44] F. Facchinei, V. Kungurtsev, L. Lampariello, and G. Scutari, "Ghost penalties in nonconvex constrained optimization: Diminishing stepsizes and iteration complexity," 2017, *arXiv:1709.03384*.
[45] F. Facchinei, L. Lampariello, and G. Scutari, "Feasible methods for nonconvex nonsmooth problems with applications in green communications," *Math. Program.*, vol. 164, no. 1, pp. 55–90, Jul. 2017.
[46] F. H. Clarke, *Optimization and Nonsmooth Analysis.* New York, NY, USA: Wiley, 2008.
[47] E. J. Balder, "On generalized gradients and optimization," Sep. 2008. [Online]. Available: http://www.staff.science.uu.nl/balde101/cao10/cursus08_3.pdf
[48] F. H. Clarke, "Generalized gradients and applications," *Trans. Amer. Math. Soc.*, vol. 205, pp. 247–262, Apr. 1975.
[49] A. Ruszczynski, *Nonlinear Optimization*. Princeton, NJ, USA: Princeton Univ. Press, 2011.
[50] 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.
[51] E. J. Balder, "On subdifferential calculus," Sep. 2008. [Online]. Available: http://www.staff.science.uu.nl/balde101/cao10/cursus10_1.pdf
[52] J. V. Tiel, *Convex Analysis*. Hoboken, NJ, USA: Wiley, 1984.
[53] F. Facchinei, G. Scutari, and S. Sagratella, "Parallel selective algorithms for nonconvex big data optimization," *IEEE Trans. Signal Process.*, vol. 63, no. 7, pp. 1874–1889, Apr. 2015.
[54] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM J. Optim.*, vol. 23, no. 2, pp. 1126–1153, 2013.
[55] D. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*. Belmont, MA, USA: Athena Scientific, 1996.
[56] Z. Lu and Y. Zhang, "Sparse approximation via penalty decomposition methods," *SIAM J. Optim.*, vol. 23, no. 4, pp. 2448–2478, 2013.
[57] A. F. Izmailov and M. V. Solodov, "Optimality conditions for irregular inequality-constrained problems," *SIAM J. Control Optim.*, vol. 40, no. 4, pp. 1280–1295, 2002.
[58] A. L. Yuille and A. Rangarajan, "The concave-convex procedure," *Neural Comput.*, vol. 15, no. 4, pp. 915–936, Apr. 2003.
[59] G. R. Lanckriet and B. K. Sriperumbudur, "On the convergence of the concave-convex procedure," in *Proc. Advances Neural Inf. Process. Syst.*, 2009, pp. 1759–1767.
[60] G. Scutari, F. Facchinei, and L. Lampariello, "Parallel and distributed methods for constrained nonconvex optimization–Part I: Theory," *IEEE Trans. Signal Process.*, vol. 65, no. 8, pp. 1929–1944, Apr. 2017.
[61] B. Mordukhovich, *Approximation Methods in Problems of Optimization and Control*. Moscow, Russia: Nauka, 1988.
[62] J. Treiman, "Lagrange multipliers for nonconvex generalized gradients with equality, inequality, and set constraints," *SIAM J. Control Optim.*, vol. 37, no. 5, pp. 1313–1329, 1999.
[63] J. Treiman, "The linear nonconvex generalized gradient and lagrange multipliers," *SIAM J. Optim.*, vol. 5, no. 3, pp. 670–680, 1995.
[64] B. Mordukhovich, *Variational Analysis and Generalized Differentiation I: Basic Theory*. Berlin, Germany: Springer, 2006.

Qingjiang Shi received his Ph.D. degree in electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 2011. From September 2009 to September 2010, he visited Prof. Z.-Q. (Tom) Luo's research group at the University of Minnesota, Twin Cities. In 2011, he worked as a Research Scientist at Bell Labs China. From 2012, he was with the School of Information and Science Technology at Zhejiang Sci-Tech University. From Feb. 2016 to Mar. 2017, he worked as a Research Fellow at Iowa State University, USA. From Mar. 2018, he is currently a Full Professor with the School of Software Engineering at Tongji University. He is also with the Shenzhen Research Institute of Big Data. His interests lie in algorithm design and analysis with applications in machine learning, signal processing and wireless networks. So far he has published more than 50 IEEE journals (six papers were ESI highly cited papers) and filed about 30 national patents.

Dr. Shi was an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING. He was awarded Golden Medal at the 46th International Exhibition of Inventions of Geneva in 2018, and also was the recipient of the First Prize of Science and Technology Award from China Institute of Communications in 2017, the National Excellent Doctoral Dissertation Nomination Award in 2013, the Shanghai Excellent Doctorial Dissertation Award in 2012, and the Best Paper Award from the IEEE PIMRC'09 conference.



Mingyi Hong received his Ph.D. degree from the University of Virginia, Charlottesville, in 2011. He is an Assistant Professor in the Department of Electrical and Computer Engineering at the University of Minnesota, Minneapolis. His research interests include optimization theory and applications in signal processing and machine learning. He is a Member of the IEEE. He serves on the IEEE Signal Processing for Communications and Networking and Machine Learning for Signal Processing Technical Committees.