On the Convergence of Alternating Direction Lagrangian Methods for Nonconvex Structured Optimization Problems

Sindri Magnússon, Pradeep Chathuranga Weeraddana, *Member, IEEE*, Michael G. Rabbat, *Senior Member, IEEE*, and Carlo Fischione, *Member, IEEE*

***Abstract*—Nonconvex and structured optimization problems arise in many engineering applications that demand scalable and distributed solution methods. The study of the convergence properties of these methods is, in general, difficult due to the nonconvexity of the problem. In this paper, two distributed solution methods that combine the fast convergence properties of augmented Lagrangian-based methods with the separability properties of alternating optimization are investigated. The first method is adapted from the classic quadratic penalty function method and is called the *alternating direction penalty method* (ADPM). Unlike the original quadratic penalty function method, where single-step optimizations are adopted, ADPM uses an alternating optimization which, in turn, makes it scalable. The second method is the well-known *alternating direction method of multipliers* (ADMM). It is shown that ADPM for nonconvex problems asymptotically converges to a primal feasible point under mild conditions and an additional condition ensuring that it asymptotically reaches the standard first-order necessary conditions for local optimality is introduced. In the case of the ADMM, novel sufficient conditions under which the algorithm asymptotically reaches the standard first-order necessary conditions are established. Based on this, complete convergence of the ADMM for a class of low-dimensional problems is characterized. Finally, the results are illustrated by applying ADPM and ADMM to a nonconvex localization problem in wireless-sensor networks.**

***Index Terms*—Alternating direction method of multipliers (ADMM), distributed optimization, localization, nonconvex optimization.**

I. Introduction

In THE last few decades, increasingly rapid technological developments have resulted in vast amounts of dispersed data. Optimization techniques have played a central role in transforming the vast data sets into usable information. However, due to the increasing size of the related optimization problems, it is essential that these optimization techniques scale with data size. Fortunately, many large-scale optimization problems in real-world applications possess appealing structural properties due to the networked nature of the problems. Thus, increasing research efforts have been devoted to the investigation of how these structural properties can be exploited in the algorithm design to achieve scalability. The focal point of these efforts has been on "well-behaved" convex problems, rather than more challenging nonconvex problems. Nevertheless, large-scale nonconvex problems arise in many real-world network applications. Examples of such nonconvex applications include matrix factorization techniques for recommender systems (the Netflix challenge) [1], localization in wireless-sensor networks [2], optimal power flow in smart grids [3], [4], and LDPC decoding [5]. Interestingly, these large-scale nonconvex applications tend to have the structural advantages that are commonly exploited to design-scalable algorithms for their convex counterparts. This suggests that the algorithms used for large-scale convex problems can potentially be applied to nonconvex problems as well. However, theoretical guarantees for these algorithms in the nonconvex regime have not yet been established. This paper investigates convergence properties of a class of scalable and distributed algorithms for *nonconvex* structured optimization problems. Here, 1) by *distributed algorithms*, we mean any algorithm that can be executed by at least two entities where no single entity has access to the full problem data and 2) by *structured optimization problems*, we mean any problem with structures in the problem data that can be exploited to achieve 1).

A. Related Literature

Many recent studies on large-scale optimization have focused on distributed subgradient methods in the context of multiagent networks [6]–[13]. There, multiple agents, each with a private objective function, cooperatively minimize the aggregate objective function by communicating over the network. In contrast to [6]–[11], papers [12] and [13] consider nonconvex multiagent problems. Specifically, [12] applies distributed subgradient methods to the (convex) dual problem and investigates sufficient conditions under which the approach converges to a pair of optimal primal/dual variables. On the other hand, [13] studies the convergence of stochastic subgradient methods to a point satisfying the first-order necessary conditions for local

Manuscript received April 14, 2015; accepted July 22, 2015. Date of publication September 3, 2015; date of current version September 15, 2016. This work was supported in part by the VR Chromos Project and in part by a grant from the KTH School of Electrical Engineering. Recommended by Associate Editor M. di Bernardo.

S. Magnússon and C. Fischione are with the Electrical Engineering School, Access Linnaeus Center, KTH Royal Institute of Technology, 100 44 Stockholm, Sweden (e-mail: sindrim@kth.se; carlofi@kth.se).

P. C. Weeraddana is with Sri Lankan Institute of Information Technology, 10115 Malabe, Sri Lanka (e-mail: chathuranga.we@sliit.lk).

M. G. Rabbat is with the Department of Electrical and Computer Engineering, McGill University, Montréal, QC H3A OE9, Canada (e-mail: michael.rabbat@mcgill.ca).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCNS.2015.2476198

optimality with a probability one. A main drawback of these gradient-based approaches is that they can only converge to an exact optimal (or local optimal) solution when a diminishing step size is used, which results in a poor convergence rate. The diminishing step-size assumption is relaxed in the promising recent work [11] while keeping the exact convergence by introducing a correction term, which significantly improves the convergence rate.

Another widely used approach for structured convex optimization is the alternating direction method of multipliers (ADMM) [14]–[16]. ADMM is a variant of the classical method of multipliers (MM) [17, Ch. 2] [18, Ch. 4.2], where the primal variable update of the MM is split into subproblems, whenever the objective is separable. This structure is common in large-scale optimization problems that arise in practice [16]. Even problems that do not possess such a structure can often be posed equivalently in a form appropriate for ADMM by introducing auxiliary variables and linear constraints. These techniques have been employed in many recent works when designing distributed algorithms for convex as well as nonconvex problems [16], [19]–[25]. A key property of ADMM compared with other existing scalable approaches, such as subgradient and dual descent methods (mentioned above), is its superior convergence behavior, see [16], [20], and [26] for empirical results. Characterizing the exact convergence rate of ADMM is still an ongoing research topic [23], [25]–[27]. Many recent papers have also numerically demonstrated the fast and appealing convergence behavior of ADMM even on nonconvex problems [24], [28]–[31]. Despite these encouraging observations, there are still no theoretical guarantees for ADMM's convergence in the nonconvex regime. Therefore, investigating convergence properties of the ADMM and related algorithms in nonconvex settings is of great importance in theory as well as in practice, and is motivated by the many emerging large-scale nonconvex applications.

B. Notation and Definitions

Vectors and matrices are represented by boldface lowercase and uppercase letters, respectively. The set of real and natural numbers is denoted by $\mathbb{R}$ and $\mathbb{N}$, respectively. The set of real n vectors and $n\times m$ matrices is denoted by $\mathbb{R}^n$ and $\mathbb{R}^{n\times m}$, respectively. The ith component of the vector $\mathbf{x}$ is denoted by $\mathbf{x}_i$. The superscript $(\cdot)^{\mathrm{T}}$ stands for transpose. We use parentheses to construct vectors and matrices from comma separated lists as $(\mathbf{x}_1,\ldots,\mathbf{x}_n)=[\mathbf{x}_1^{\mathrm{T}},\ldots,\mathbf{x}_n^{\mathrm{T}}]^{\mathrm{T}}$ and $(\mathbf{A}_1,\ldots,\mathbf{A}_n)=[\mathbf{A}_1^{\mathrm{T}},\ldots,\mathbf{A}_n^{\mathrm{T}}]^{\mathrm{T}}$, respectively. $\mathrm{diag}(\mathbf{A}_1,\ldots,\mathbf{A}_n)$ denotes the diagonal block matrix with $\mathbf{A}_1,\ldots,\mathbf{A}_n$ on the diagonal. $\mathbf{A}\succ 0$ ($\mathbf{A}\succeq 0$) indicates that the square matrix $\mathbf{A}$ positive (semi)definite. $\|\cdot\|$ denotes the 2-norm. We use the following definition.

Definition 1 (FON): Consider the optimization problem

$$\begin{aligned}\underset{\mathbf{x}\in\mathbb{R}^p}{\text{minimize}} \quad & f(\mathbf{x})\\ \text{subject to} \quad & \boldsymbol{\phi}(\mathbf{x})=\mathbf{0},\quad \boldsymbol{\psi}(\mathbf{x})\leq\mathbf{0}\end{aligned}\tag{1}$$

where $\boldsymbol{\phi}:\mathbb{R}^p\to\mathbb{R}^{q_1}$ and $\boldsymbol{\psi}:\mathbb{R}^p\to\mathbb{R}^{q_2}$ are continuously differentiable functions. We say that $\mathbf{x}^\star\in\mathbb{R}^p$ and $(\boldsymbol{\lambda}^\star,\boldsymbol{\mu}^\star)\in\mathbb{R}^{q_1+q_2}$ satisfy the first-order necessary (FON) conditions for problem (1), if the following hold: 1) primal feasibility $\boldsymbol{\phi}(\mathbf{x}^\star)=\mathbf{0}$ and $\boldsymbol{\psi}(\mathbf{x}^\star)\leq\mathbf{0}$; 2) dual feasibility $\boldsymbol{\mu}^\star\geq\mathbf{0}$; 3) complementary slackness $(\boldsymbol{\mu}^\star)_i\boldsymbol{\psi}_i(\mathbf{x}^\star)=0,\ i=1,\ldots q_2$; 4) Lagrangian vanishes: $\nabla f(\mathbf{x}^\star)=\nabla\boldsymbol{\phi}(\mathbf{x}^\star)\boldsymbol{\lambda}^\star+\nabla\boldsymbol{\psi}(\mathbf{x}^\star)\boldsymbol{\mu}^\star$. We refer to $\mathbf{x}^\star$ and $(\boldsymbol{\lambda}^\star,\boldsymbol{\mu}^\star)$ as the primal and dual variables, respectively.

II. Problem Statement, Related Background, and Contribution of this Paper

This section is organized as follows. Section II-A introduces the class of nonconvex structured problems we study. We give the necessary background on centralized algorithms in Section II-B, before introducing distributed algorithms which exploit the special structures of the related problems in Section II-C. Then we state the contribution and organization of this paper in Section II-D.

A. Problem Statement

We consider the following optimization problem:

$$\begin{aligned}\underset{\mathbf{x}\in\mathbb{R}^{p_1},\mathbf{z}\in\mathbb{R}^{p_2}}{\text{minimize}} \quad & f(\mathbf{x})+g(\mathbf{z})\\ \text{subject to} \quad & \mathbf{x}\in\mathcal{X},\mathbf{z}\in\mathcal{Z}\\ & \mathbf{A}\mathbf{x}+\mathbf{B}\mathbf{z}=\mathbf{c}\end{aligned}\tag{2}$$

where $\mathbf{A}\in\mathbb{R}^{q\times p_1}$, $\mathbf{B}\in\mathbb{R}^{q\times p_2}$, and $\mathbf{c}\in\mathbb{R}^q$. The use of the variable notation $\mathbf{x}$ and $\mathbf{z}$ is consistent with the literature [16]. Functions $f:\mathcal{X}\to\mathbb{R}$ and $g:\mathcal{Z}\to\mathbb{R}$ are continuously differentiable on $\mathbb{R}^{p_1}$ and $\mathbb{R}^{p_2}$, respectively, and may be *nonconvex*. We refer to the affine constraint $\mathbf{A}\mathbf{x}+\mathbf{B}\mathbf{z}=\mathbf{c}$ as *the coupling constraint*. We assume that Problem (2) is feasible. Problem (2) is general in the sense that many interesting large-scale problems, including consensus, and sharing [16, Sec. 7], among others can be equivalently posed in its form. Moreover, as noted in Section I-A, problem (2) commonly appears in multiagent networks, where $\mathbf{x}$ usually represents the private variable of each node/agent, $\mathbf{z}$ represents the coupling between the nodes, and the coupling constraint enforces the network consensus. Therefore, our analytical results in subsequent sections apply to a broad class of problems of practical importance.

Next, we discus centralized solution methods for Problem (2) which are the basis for the distributed methods we study.

B. Penalty and Augmented Lagrangian Methods

Nonconvex problems of the form (2) can be gracefully handled by penalty and augmented Lagrangian methods, such as the quadratic penalty function method and method of multipliers [17, Ch. 2] [18, Ch. 4.2]. The main ingredient of these methods is the augmented Lagrangian, given by

$$\begin{aligned}L_\rho(\mathbf{x},\mathbf{z},\mathbf{y}) = f(\mathbf{x})+g(\mathbf{z})+\mathbf{y}^{\mathrm{T}}(\mathbf{A}\mathbf{x}+\mathbf{B}\mathbf{z}-\mathbf{c})\\ +\left(\frac{\rho}{2}\right)\|\mathbf{A}\mathbf{x}+\mathbf{B}\mathbf{z}-\mathbf{c}\|^2.\end{aligned}$$

Here, $\mathbf{x}$ and $\mathbf{z}$ are the primal variables of Problem (2), and $\mathbf{y}\in\mathbb{R}^q$ and $\rho\in\mathbb{R}$ refer to the multiplier vector and the penalty parameter, respectively.

The penalty and augmented Lagrangian methods consist of iteratively updating the variables $\mathbf{x}$, $\mathbf{z}$, $\mathbf{y}$, and ρ. An update common to all of the methods is the primal variable update, i.e.,

$$(\mathbf{x}(t+1), \mathbf{z}(t+1)) = \underset{(\mathbf{x},\mathbf{z})\in\mathcal{X}\times\mathcal{Z}}{\arg\min}\, L_{\rho(t)}\left(\mathbf{x}, \mathbf{z}, \mathbf{y}(t)\right) \quad (3)$$

where $t \in \mathbb{N}$ is the iteration index. The main difference between the two methods lies in the $\mathbf{y}$ and ρ updates. For example, in the case of the quadratic penalty method, the penalty parameter $\rho(t)$ is chosen such that $\lim_{t\to\infty} \rho(t) = \infty$ with the intention of enforcing the limit points of $\{(\mathbf{x}(t), \mathbf{z}(t))\}_{t\in\mathbb{N}}$ to satisfy the coupling constraint. It turns out that if the Lagrange multipliers are bounded, that is, there exists $M \in \mathbb{R}$ such that $\|\mathbf{y}(t)\| < M$ for all $t \in \mathbb{N}$, then every limit point of the sequence $\{(\mathbf{x}(t), \mathbf{z}(t))\}_{t\in\mathbb{N}}$ is a global minimum of Problem (2) [17, Prop. 2.1].

The motive of the method of multipliers is to choose the sequence of multipliers $\{\mathbf{y}(t)\}_{t\in\mathbb{N}}$ intelligently to enable convergence to local or global optima of (2) without needing $\lim_{t\to\infty} \rho(t) = \infty$. The well-known choice of $\{\mathbf{y}(t)\}_{t\in\mathbb{N}}$ in the method of multipliers follows the recursion:

$$\mathbf{y}(t+1) = \mathbf{y}(t) + \rho(t)\left(\mathbf{A}\mathbf{x}(t+1) + \mathbf{B}\mathbf{z}(t+1) - \mathbf{c}\right). \quad (4)$$

The motivation for (4) is that when $(\mathbf{x}(t+1), \mathbf{z}(t+1))$ is locally optimal for Problem (3) and satisfies the FON conditions (Definition 1)[1] then $(\mathbf{x}(t+1), \mathbf{z}(t+1))$ and $\mathbf{y}(t+1)$ satisfy conditions 2), 3), and 4) of the FON conditions for the original Problem (2), all except 1) primal feasibility. Furthermore, under mild conditions, the method of multipliers converges to a local optimal point $(\mathbf{x}^\star, \mathbf{z}^\star)$ and to a corresponding optimal Lagrangian multiplier $\mathbf{y}^\star$ [17, Prop. 2.4]. In addition to the local convergence, when $(\mathbf{x}(t), \mathbf{z}(t))$ is a global optima of (3), then (4) is a gradient ascent step for the dual problem. However, due to nonzero duality gap in most nonconvex problems, the solution to (2) cannot be recovered from the dual problem. Hence, the method of multipliers can generally only be considered a local method.

In general, the penalty and augmented Lagrangian methods mentioned before are very reliable and effective for handling problems of the form (2). However, these methods entail centralized solvers, especially in the $(\mathbf{x}, \mathbf{z})$-update (3), even if the objective function of problem (2) has a desirable separable structure in $\mathbf{x}$ and $\mathbf{z}$. More specifically, these methods do not allow the possibility of performing the $(\mathbf{x}, \mathbf{z})$-update in two steps: first $\mathbf{x}$-update and then $\mathbf{z}$-update. Otherwise, the assertions on the convergence of the algorithms do not hold anymore. Therefore, the penalty and augmented Lagrangian methods are not applicable in distributed settings, whenever the problems possess decomposition structures. Such restrictions have motivated an adaptation of the classical penalty and augmented Lagrangian methods that have excellent potential for a parallel/distributed implementation which we discussed now.

[1]We do not include the multipliers related to the constraint $\mathcal{X} \times \mathcal{Z}$ to simplify the presentation, but it is easily checked that the claim holds when they are included.

C. Alternating Direction Lagrangian Methods

Recall that problem (2) has a linear coupling constraint and an objective function that is separable in $\mathbf{x}$ and $\mathbf{z}$. This motivates potential solution approaches to Problem (2), where the optimization in (3) is performed in two steps, first in the $\mathbf{x}$ coordinate and then in the $\mathbf{z}$ coordinate, i.e.,

$$\mathbf{x}(t+1) = \underset{\mathbf{x}\in\mathcal{X}}{\arg\min}\, L_{\rho(t)}\left(\mathbf{x}, \mathbf{z}(t), \mathbf{y}(t)\right) \quad (5)$$

$$\mathbf{z}(t+1) = \underset{\mathbf{z}\in\mathcal{Z}}{\arg\min}\, L_{\rho(t)}\left(\mathbf{x}(t+1), \mathbf{z}, \mathbf{y}(t)\right). \quad (6)$$

Let us refer to these approaches as *alternating direction lagrangian methods* (ADLM). We consider two ADLM variants. The first variant is analogous to the quadratic penalty approach, where the sequence of penalty parameters $\{\rho(t)\}_{t\in\mathbb{N}}$ and the multiplier vectors $\{\mathbf{y}(t)\}_{t\in\mathbb{N}}$ are taken to be nondecreasing/divergent and bounded, respectively. We refer to this novel approach as the *alternating direction penalty method* (ADPM). The second variant is the classic ADMM itself, the analog of the method of multipliers. We now pose the question: *can the convergence of the considered ADLM variants, ADPM and ADMM, still be guaranteed when Problem* (2) *is nonconvex*?

D. Contribution and Structure of this Paper

We start by investigating the convergence behavior of the ADPM in Section III when Problem (2) is nonconvex. We consider a) an *unconstrained* case in Section III-B, that is, where $\mathcal{X} = \mathbb{R}^{p_1}$ and $\mathcal{Z} = \mathbb{R}^{p_2}$, and b) a *constrained* case in Section III-C where $\mathcal{X}$ and $\mathcal{Z}$ are compact sets. The analysis in case a) is based on assumptions on (2) which highlight the situation when the $\mathbf{x}$- and $\mathbf{z}$-updates of ADLM are used to achieve distributed algorithms over networks and the coupling constraint expresses the network consensus. Under these assumptions, we show that if $\mathbf{y}(t) = \mathbf{0}$ and $\lim_{t\to} \rho(t) = \infty$, then the primal feasibility of (2) is asymptotically achieved as ADPM proceeds. In addition, if the sequence $1/\rho(t)$ is also nonsummable and $(\mathbf{x}(t), \mathbf{z}(t))$ converges to $(\mathbf{x}^\star, \mathbf{z}^\star)$, then $(\mathbf{x}^\star, \mathbf{z}^\star)$ satisfies the FON conditions (Definition 1) of (2). In case b), we consider more general assumptions on (2) and allow $\mathbf{y}(t)$ to be any bounded sequence. Under these assumptions, we show that if $\mathcal{X}$ and $\mathcal{Z}$ are convex and the sequence $1/\rho(t)$ is summable, then the primal feasibility of (2) is asymptotically achieved as ADPM proceeds. Moreover, we give an intuitive example showing why we need the sets $\mathcal{X}$ and $\mathcal{Z}$ to be convex in general.

Next, we investigate the convergence behavior of the ADMM when (2) is nonconvex in Section IV. We assume that the penalty parameter is fixed, that is, $\rho(t) = \rho$. We consider general assumptions on Problem (2) where the sets $\mathcal{X}$ and $\mathcal{Z}$ can even be nonconvex. We show that when $\mathbf{y}(t)$ converges, then any limit point of $\mathbf{x}(t), \mathbf{z}(t))$ satisfies the FON conditions of Problem (2). We note that the condition can be checked a posteriori or at runtime by inspecting some algorithm parameters as the algorithm proceeds (online). Moreover, we show how our results can be used to completely characterize the convergence

of ADMM for a class of problems, that is, to determine to which point ADMM converges given an initialization. In comparison to [12], we consider ADMM, whereas therein the standard Lagrangian dual function is maximized.

Finally, we illustrate how the considered methods can be applied to design distributed algorithms for cooperative localization in wireless-sensor networks.

III. Alternating Direction Penalty Method

In this section, we study convergence properties of the ADPM for addressing Problem (2). In Section III-A, we give an explicit algorithm description and in Section III-B and C, we investigate properties of the ADPM when $\mathcal{X} \times \mathcal{Z} = \mathbb{R}^{p_1} \times \mathbb{R}^{p_2}$ and when $\mathcal{X} \times \mathcal{Z} \subsetneqq \mathbb{R}^{p_1} \times \mathbb{R}^{p_2}$, respectively.

A. Algorithm Description

The steps of ADPM are shown in Algorithm 1

Algorithm 1: Alternating Direction Penalty Method (ADPM)

---

1) **Initialization**: Set $t = 0$ and initialize $\mathbf{z}(0)$, $\mathbf{y}(0)$, and $\rho(0)$.
2) $\mathbf{x}$-**update**: $\mathbf{x}(t+1) = \arg\min_{\mathbf{x}\in\mathcal{X}} L_{\rho(t)}(\mathbf{x}, \mathbf{z}(t), \mathbf{y}(t))$.
3) $\mathbf{z}$-**update**: $\mathbf{z}(t+1) = \arg\min_{\mathbf{z}\in\mathcal{Z}} L_{\rho(t)}(\mathbf{x}(t+1), \mathbf{z}, \mathbf{y}(t))$.
4) $\rho/\mathbf{y}$-**update**: Update $\rho(t+1)$ and $\mathbf{y}(t+1)$.
5) **Stopping criterion**: If stopping criterion is met terminate, otherwise set $t = t+1$ and go to step 2.

The algorithm parameters $\rho(t)$ and $\mathbf{y}(t)$ are chosen such that $\lim_{t\to\infty} \rho(t) = \infty$ and the sequence $\{\mathbf{y}(t)\}_{t\in\mathbb{N}}$ is taken to be bounded. The $\mathbf{x}$- and $\mathbf{z}$- updates (steps 2 and 3) are the main steps of the algorithm where the augmented Lagrangian is minimized in two steps.

Nonconvexities of f and g suggest potential difficulties in the implementation of the $\mathbf{x}$- and $\mathbf{z}$- updates (see steps 2 and 3). However, it is worth noting that problems encountered in practice often contain structures that can be exploited to successfully implement the $\mathbf{x}$- and $\mathbf{z}$-updates. Several examples are given next.

Example 1: Let $\mathcal{X}$ (or $\mathcal{Z}$) be convex, let f (or g) be twice continuously differentiable, and suppose there exists $\alpha \in \mathbb{R}$ such that $\nabla^2 f(\mathbf{x}) > \alpha$ for all $\mathbf{x} \in \mathcal{X}$, where $\alpha < 0$ if f is nonconvex on $\mathcal{X}$. Moreover, suppose $\mathbf{A}$ (or $\mathbf{B}$) has full-column rank. Then, the optimization problem in the $\mathbf{x}$-update (or $\mathbf{z}$-update) is strongly convex for sufficiently large $\rho(t) > -\alpha/\lambda_{\min}(\mathbf{A}^{\mathrm{T}}\mathbf{A})$. This can be seen by looking at the Hessian $\nabla^2_{\mathbf{x}} L_{\rho(t)}(\mathbf{x}, \mathbf{z}(t), \mathbf{y}(t))$ and using that $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ is positive definite.

Example 2: Let $f(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} + \mathbf{q}^{\mathrm{T}}\mathbf{x}$ where $\mathbf{Q} \in \mathbb{R}^{p_1\times p_1}$ is a symmetric indefinite matrix. Then, if $\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x} > 0$ for all $\mathbf{x} \in \mathbb{R}^{p_1} \setminus \{\mathbf{0}\}$ in the null space of $\mathbf{A}$, then there exists $\bar{\rho} \in \mathbb{R}$ such that $L_{\rho(t)}(\cdot, \mathbf{z}(t), \mathbf{y}(t))$ is convex in $\mathbf{x}$ for all $\rho(t) \geq \bar{\rho}$, see [18, Lemma 3.2.1 and Fig. 3.2.1].

Example 3: A potential feature of the multiagent setting is that the $\mathbf{x}$- update is separable into low-dimensional problems. More specifically, suppose the variable $\mathbf{x}$ is partitioned into low-dimensional subvectors as $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$, where there is no coupling between $\mathbf{x}_i$ and $\mathbf{x}_j$ in the constraints, for all $i, j = 1, \ldots, N$ such that $i \neq j$. Suppose also that the objective function is separable with respect to the partition, that is, $f(\mathbf{x}) = \sum_{i=1}^{N} f_i(\mathbf{x}_i)$. Then, the objective function in the $\mathbf{x}$-update is also separable with respect to the partition. Thus, provided that each subvector $\mathbf{x}_i$ is of low dimension, global methods, such as branch and bound, can be efficiently used to optimally solve the optimization problem in the $\mathbf{x}$-update.

B. Algorithm Properties: Unconstrained Case

In this section, we derive the convergence properties of the ADPM algorithm when $\mathcal{X} = \mathbb{R}^n$ and $\mathcal{Z} = \mathbb{R}^m$. Our convergence results assert that 1) primal feasibility of problem (2) is satisfied and 2) if the sequence $1/\rho(t)$ is nonsummable and $(\mathbf{x}(t), \mathbf{z}(t))$ converges to a point $(\mathbf{x}^\star, \mathbf{z}^\star)$, then $(\mathbf{x}^\star, \mathbf{z}^\star)$ satisfies the FON conditions (Definition 1) of Problem (2). To establish this result precisely, let us first make the following assumptions.

Assumption 1: $g(\mathbf{x}) = 0$, $\mathbf{A} = \mathbf{I}$, $\mathbf{c} = \mathbf{0}$, $\mathbf{B}$ has full-column rank.

Assumption 2: At least one of the following conditions holds true:

a) f is continuously differentiable with bounded gradient, that is, there exists $\kappa \in \mathbb{R}$ such that $\|\nabla f(\mathbf{x})\| \leq \kappa$ for all $\mathbf{x} \in \mathbb{R}^n$.
b) $\|\mathbf{B}\|_\infty = 1$ and $\|(\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}\|_\infty = 1$. Moreover, there exist a scalar $c > 0$ such that: (b.i) $[\nabla f(\mathbf{x})]_i < 0$ if $\mathbf{x}_i < -c$, for component $i \in \{1, \ldots, p_1\}$ and (b.ii) $[\nabla f(\mathbf{x})]_i > 0$ if $\mathbf{x}_i > c$, for $i \in \{1, \ldots, p_1\}$.

Assumption 1 naturally arises when designing distributed algorithm over networks, where $\mathbf{x}$ represents private variables of each node/agent and $\mathbf{z}$ represents the coupling between the nodes. Assumption 2.a is standard in the literature, for example, in relation to (sub)gradient methods [6], [10], [18]. In addition, Assumption 2.b ensures that our results hold for more general classes of practical problems than covered by Assumption 2.a, for example, when f is a polynomial of even degree with positive leading coefficient. (See Problem (58) in Section V.) We note that $\|\mathbf{B}\|_\infty = 1$ and $\|(\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}\|_\infty = 1$ naturally hold when $\mathbf{x}$ and $\mathbf{z}$ represent private and coupling variables of each node/agent in a connected network, for example, see Section V. The main implication of Assumption 2.b is that it ensures that the sequence $(\mathbf{x}(t), \mathbf{z}(t))$ is bounded as we show in the following lemma.

Lemma 1: Suppose Assumption 2.b holds true and $\|\mathbf{z}(t)\|_\infty \leq c$, then $\|\mathbf{x}(t+1)\|_\infty \leq c$ and $\|\mathbf{z}(t+1)\|_\infty \leq c$.

Proof: Let us start by showing that $\|\mathbf{x}(t+1)\|_\infty \leq c$ by using contradiction. Without loss of generality, we assume that $\mathbf{x}_i(t+1) < -c$ for some $i = 1, \ldots, p_1$ (the other cases follow symmetrical arguments). Then, $[\nabla f(\mathbf{x}(t))]_i < 0$, from Assumption 2.b, which, in turn, implies that

$$\left\| \left(\frac{1}{\rho}\right) \nabla f\left(\mathbf{x}(t+1)\right) + \mathbf{x}(t+1) \right\|_\infty > c. \tag{7}$$

However, using the FON conditions of the $\mathbf{x}$-update and that $\|\mathbf{B}\|_\infty = 1$ and $\|\mathbf{z}(t)\|_\infty \leq c$, we also have

$$\left\| \left(\frac{1}{\rho}\right) \nabla f\left(\mathbf{x}(t+1)\right) + \mathbf{x}(t+1) \right\|_\infty = \|\mathbf{B}\mathbf{z}(t)\|_\infty \leq c. \quad (8)$$

Clearly, (7) and (8) contradict each other. Hence, $\|\mathbf{x}(t+1)\|_\infty \leq c$.

Let us next show that $\|\mathbf{z}(t+1)\| \leq c$. From the FON conditions of the $\mathbf{z}$-update, we get that $\mathbf{z}(t+1) = (\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{x}(t+1)$ which, together with $\|(\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}\|_\infty = 1$, ensures that $\|\mathbf{z}(t+1)\| \leq \|\mathbf{x}(t+1)\|_\infty \leq c$. ■

We are now ready to derive the main result of this subsection.

Proposition 1: Suppose assumptions 1 and 2 hold. Let $r(t)$ be the residual at iteration t of the ADPM defined as $r(t) = \|\mathbf{x}(t) + \mathbf{B}\mathbf{z}(t)\|$. Then

1) If $\mathbf{y}(t) = \mathbf{0}$ for all $t \in \mathbb{N}$, then $\lim_{t\to\infty} r(t) = 0$.
2) If in addition $\sum_{t=0} 1/\rho(t) = \infty$ and $\lim_{t\to\infty}(\mathbf{x}(t), \mathbf{z}(t)) = (\mathbf{x}^\star, \mathbf{z}^\star)$, then $(\mathbf{x}^\star, \mathbf{z}^\star)$ satisfies the FON conditions of Problem (2).

Proof: i) Note that Assumption 2 implies that the sequence $\nabla f(\mathbf{x}(t))$ is bounded, when 2.a holds, then the result is obvious and when 2.b holds, the result follows from Lemma 1. In particular, there exists $M \in \mathbb{R}$ such that $\|\nabla f(\mathbf{x}(t))\| < M$ for all $t \in \mathbb{N}$.

Using the FON conditions of the $\mathbf{x}$- and $\mathbf{y}$- updates, we obtain

$$\mathbf{0} = \nabla f\left(\mathbf{x}(t+1)\right) + \rho(t)\left(\mathbf{x}(t+1) + \mathbf{B}\mathbf{z}(t)\right) \quad (9)$$
$$\mathbf{0} = \mathbf{B}^{\mathrm{T}}\left(\mathbf{x}(t+1) - \mathbf{B}\mathbf{z}(t+1)\right) \quad (10)$$

and rearranging (9) and (10), we obtain

$$\mathbf{z}(t) = (\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}\left(\mathbf{x}(t+1) + \frac{1}{\rho(t)}\nabla f(\mathbf{x}(t+1))\right) \quad (11)$$
$$\mathbf{z}(t+1) = (\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{x}(t+1). \quad (12)$$

Using (11), (12), and that $\nabla f(\mathbf{x}(t))$ is bounded, we obtain

$$\|\mathbf{z}(t+1) - \mathbf{z}(t)\| = \frac{1}{\rho(t)}\left\|(\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}\nabla f(\mathbf{x}(t+1))\right\| \quad (13)$$
$$\leq \frac{M}{\rho(t)}\left\|(\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}\right\|. \quad (14)$$

Similarly, using (9) and that $\nabla f(\mathbf{x}(t))$ is bounded, we obtain

$$\|\mathbf{x}(t+1) + \mathbf{B}\mathbf{z}(t)\| = \frac{1}{\rho(t)}\|\nabla f\left(\mathbf{x}(t+1)\right)\| \leq \frac{M}{\rho(t)}. \quad (15)$$

Finally, by using (14), (15) and the triangle inequality give

$$\|\mathbf{x}(t+1) + \mathbf{B}\mathbf{z}(t+1)\| \leq \|\mathbf{x}(t+1) + \mathbf{B}\mathbf{z}(t)\| + \|\mathbf{B}\left(\mathbf{z}(t+1) - \mathbf{z}(t)\right)\|$$
$$\leq \frac{M}{\rho(t)}\left(1 + \left\|\mathbf{B}(\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}\right\|\right). \quad (16)$$

Since $\rho(t)$ diverges to ∞, (16) converges to zero, which concludes the proof.

ii) We need to show that $(\mathbf{x}^\star, \mathbf{z}^\star)$ satisfies the FON conditions (Definition 1) for Problem (2) together with some Lagrangian multiplier. Note that condition 1) of the FON conditions (Primal feasibility) holds because of part i) of this proposition and conditions 2) and 3) of the FON conditions (dual feasibility and complementary slackness) trivially hold since there are no inequality constraints, since $\mathcal{X} = \mathbb{R}^{p_1}$ and $\mathcal{Z} = \mathbb{R}^{p_2}$. Hence, we only need to show condition 4) that the Lagrangian vanishes. We note that the gradient of the Lagrangian is

$$\nabla f(\mathbf{x}^\star) + \boldsymbol{\lambda} = \mathbf{0} \text{ and } \mathbf{B}^{\mathrm{T}}\boldsymbol{\lambda} = \mathbf{0} \quad (17)$$

where $\boldsymbol{\lambda} \in \mathbb{R}^n$ is the dual variable. If $\nabla f(\mathbf{x}^\star)$ is in the null space of $\mathbf{B}^{\mathrm{T}}$, then (17) is satisfied by setting $\boldsymbol{\lambda} = -\nabla f(\mathbf{x}^\star)$, which would conclude the proof. Therefore, in the sequel, we show that $\mathbf{B}^{\mathrm{T}}\nabla f(\mathbf{x}^\star) = \mathbf{0}$.

Using (11) and (12) gives

$$\sum_{t=0}^{\infty}(\mathbf{B}^{\mathrm{T}}\mathbf{B})(\mathbf{z}(t+1) - \mathbf{z}(t)) = \sum_{t=0}^{\infty}\frac{1}{\rho(t)}\mathbf{B}^{\mathrm{T}}\nabla f(\mathbf{x}(t+1)). \quad (18)$$

The left-hand side of (18) is a telescopic series, hence

$$\sum_{t=0}^{\infty}\frac{1}{\rho(t)}\mathbf{B}^{\mathrm{T}}\nabla f\left(\mathbf{x}(t+1)\right) = \left(\mathbf{B}^{\mathrm{T}}\mathbf{B}\right)\left(\mathbf{z}^\star - \mathbf{z}(0)\right) \quad (19)$$

which, in turn, ensures the convergence of (19) and

$$\sum_{t=0}^{\infty}\frac{1}{\rho(t)}\left\|\mathbf{B}^{\mathrm{T}}\nabla f\left(\mathbf{x}(t+1)\right)\right\|. \quad (20)$$

Set $L = \lim_{t\to\infty}\|B^{\mathrm{T}}\nabla f(\mathbf{x}(t))\| = \|B^{\mathrm{T}}\nabla f(\mathbf{x}^\star)\|$. Let us next use the contraction to show that $L = 0$ which, in turn, shows that $B^{\mathrm{T}}\nabla f(\mathbf{x}^\star) = \mathbf{0}$. Without of loss of generality, suppose $L > 0$. Choose $\epsilon > 0$ and $T \in \mathbb{N}$ such that $\|B^{\mathrm{T}}\nabla f(\mathbf{x}(t))\| > L - \epsilon > 0$ for all $t \geq T$. Then

$$\sum_{t=0}^{\infty}\frac{1}{\rho(t)}\left\|\mathbf{B}^{\mathrm{T}}\nabla f\left(\mathbf{x}(t+1)\right)\right\| \geq \sum_{t=0}^{T-1}\frac{1}{\rho(t)}\left\|\mathbf{B}^{\mathrm{T}}\nabla f\left(\mathbf{x}(t+1)\right)\right\| + (L-\epsilon)\sum_{t=T}^{\infty}\frac{1}{\rho(t)}$$

where the right-hand side diverges to ∞, since $\sum_{t=0}^{\infty} 1/\rho(t) = \infty$, which implies that the left-hand side also diverges to ∞. This contradicts that the series (20) converges and, therefore, we can conclude that $L = 0$. ■

Remark 1: In Proposition 1, we considered the case where $\mathbf{y}(t) = \mathbf{0}$, which allowed us to derive the theoretical results. Still, our numerical results in Section V show that it can be beneficial to update $\mathbf{y}$ according to the recursion $\mathbf{y}(t+1) = \mathbf{y}(t) + \rho(\mathbf{A}\mathbf{x}(t+1) + \mathbf{B}\mathbf{z}(t+1) - \mathbf{c})$ [cf. (4)].

C. Algorithm Properties: Constrained Case

In this section, we derive the convergence properties of the ADPM when $\mathcal{X}$ and $\mathcal{Z}$ are proper subsets of $\mathbb{R}^{p_1}$ and $\mathbb{R}^{p_2}$, respectively. Our convergence results assert that the primal feasibility of problem (2), which is a necessary optimality condition, is achieved as ADPM proceeds. More specifically, we show that regardless of whether f, g are convex or nonconvex, whenever $\mathcal{X}$ and $\mathcal{Z}$ are convex, the primal residual at iteration t of the ADPM (i.e., $\mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{z}(t) - \mathbf{c}$) converges to zero as

ADPM proceeds. To establish this result precisely, let us first make the following assumptions:

Assumption 3: The functions f and g of problem (2) are continuously differentiable.

Assumption 4: The sets $\mathcal{X}$ and $\mathcal{Z}$ of problem (2) are convex and compact.

Assumption 5: Slater's condition [18] holds *individually* for $\mathcal{X}$ and $\mathcal{Z}$. In particular, there exists a $\mathbf{x} \in \mathcal{X}$ (respectively, $\mathbf{z} \in \mathcal{Z}$) such that all of the inequality constraints characterizing $\mathcal{X}$ (respectively, $\mathcal{Z}$) are inactive at $\mathbf{x}$ (respectively, $\mathbf{z}$).

Assumption 6: The matrices $\mathbf{A}$ and $\mathbf{B}$ of problem (2) have full-column rank.

Note that we make no convexity assumptions on f and g. However, the convexity assumption on $\mathcal{X}$ and $\mathcal{Z}$ is *essential*. Otherwise, primal feasibility is not guaranteed in general–see Example 4 later in this section. Assumption 5 is an additional technical condition, similar to the constraint qualifications usually used in convex analysis. The last assumption is technically necessary to ensure that $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ and $\mathbf{B}^{\mathrm{T}}\mathbf{B}$ are positive definite. It is quite common in practice that this assumption holds, as desired, see Section V. The following proposition establishes the convergence of ADPM.

Proposition 2: Suppose Assumptions 3–6 hold. Let $\{\rho(t)\}_{t\in\mathbb{N}}$ be a sequence of penalty parameters used in the ADPM algorithm, where $\rho(t+1) \geq \rho(t)$ for all t and suppose there exists an integer $\kappa > 0$ and a scalar $\Delta > 1$ such that $\rho(t+\kappa) \geq \Delta\rho(t)$ for all t. Let $r(t)$ be the residual at iteration t of the ADPM defined as $r(t) = \|\mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{z}(t) - \mathbf{c}\|^2$. Then, $\lim_{t\to\infty} r(t) = 0$.

Proof: Recall that $\{\mathbf{y}(t)\}_{t\in\mathbb{N}}$ is a bounded sequence. Thus, there exists $M_0 > 0$ such that $\|\mathbf{y}(t)\| \leq M_0$, for all $t \in \mathbb{N}$. We denote by $\mathcal{Y}$ the closed ball with radius M_0 centered at the origin $\mathbf{0}$, that is, $\mathcal{Y} = \{\mathbf{y} \in \mathbb{R}^q | \|\mathbf{y}\| \leq M_0\}$.

Since f and g are continuous and the sets $\mathcal{X}$ and $\mathcal{Z}$ are compact, there exists a scalar $M_1 > 0$ such that

$$M_1 = \max_{(\mathbf{x},\mathbf{z},\mathbf{y})\in\mathcal{X}\times\mathcal{Z}\times\mathcal{Y}} \left| f(\mathbf{x}) + g(\mathbf{z}) + \mathbf{y}^{\mathrm{T}}(\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{c}) \right|. \quad (21)$$

In addition, $\hat{\mathbf{x}} : \mathbb{R}^{p_2} \to \mathbb{R}$ and $\hat{\mathbf{z}} : \mathbb{R}^{p_1} \to \mathbb{R}$, defined as

$$\hat{\mathbf{x}}(\mathbf{z}) = \arg\min_{\mathbf{x}\in\mathcal{X}} \|\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{c}\|^2 \quad (22)$$

$$\hat{\mathbf{z}}(\mathbf{x}) = \arg\min_{\mathbf{z}\in\mathcal{Z}} \|\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{c}\|^2 \quad (23)$$

are well-defined *continuous functions* (compared with Assumption 6). By definition, $\mathbf{x}(t+1)$ is a solution of the optimization problem in the $\mathbf{x}$-update of the ADPM. This, together with (21), yields

$$L_{\rho(t)}(\mathbf{x}(t+1), \mathbf{z}(t), \mathbf{y}(t)) \leq M_1 + \left(\frac{\rho(t)}{2}\right) \|\mathbf{A}\hat{\mathbf{x}}(\mathbf{z}(t)) + \mathbf{B}\mathbf{z}(t) - \mathbf{c}\|^2. \quad (24)$$

Similarly, we obtain

$$L_{\rho(t)}(\mathbf{x}(t+1), \mathbf{z}(t+1), \mathbf{y}(t)) \leq M_1 + \left(\frac{\rho(t)}{2}\right) \|\mathbf{A}\mathbf{x}(t+1) + \mathbf{B}\hat{\mathbf{z}}(\mathbf{x}(t+1)) - \mathbf{c}\|^2. \quad (25)$$

Let us first use (24) and (25) to derive a recursive relation for $r(t)$. By rearranging the terms of (24) and by using that $|M_1 - (f(\mathbf{x}) + g(\mathbf{z}) + \mathbf{y}^{\mathrm{T}}(\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{c}))| \leq 2M_1$ for all $(\mathbf{x}, \mathbf{z}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Z} \times \mathcal{Y}$, we have for all $t \in \mathbb{N}$

$$\|\mathbf{A}\mathbf{x}(t+1) + \mathbf{B}\mathbf{z}(t) - \mathbf{c}\|^2 \leq \frac{4M_1}{\rho(t)} + \|\mathbf{A}\hat{\mathbf{x}}(\mathbf{z}(t)) + \mathbf{B}\mathbf{z}(t) - \mathbf{c}\|^2. \quad (26)$$

Moreover, we have for all $t \in \mathbb{N}$

$$r(t+1) \leq \frac{4M_1}{\rho(t)} + \|\mathbf{A}\mathbf{x}(t+1) + \mathbf{B}\hat{\mathbf{z}}(\mathbf{x}(t+1)) - \mathbf{c}\|^2 \quad (27)$$

$$\leq \frac{8M_1}{\rho(t)} + \|\mathbf{A}\hat{\mathbf{x}}(\mathbf{z}(t)) + \mathbf{B}\mathbf{z}(t) - \mathbf{c}\|^2 \quad (28)$$

$$\leq \frac{8M_1}{\rho(t)} + r(t) \quad (29)$$

where (27) follows similarly by rearranging the terms of (25) and by using that $|M_1 - (f(\mathbf{x}) + g(\mathbf{z}) + \mathbf{y}^{\mathrm{T}}(\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{c}))| \leq 2M_1$ for all $(\mathbf{x}, \mathbf{z}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Z} \times \mathcal{Y}$, (28) follows from combining the inequalities (26) and (27), together with the definition of $\hat{\mathbf{x}}$ and $\hat{\mathbf{z}}$, and (29) follows by the definition of $\hat{\mathbf{x}}$.

Let us next use the recursive inequality (29) above to show that $\{r(t)\}_{t\in\mathbb{N}}$ *converges* to a *finite* value. The inequality (29) implies for all $t, n \geq 0$

$$r(t+n) \leq r(t) + 8M_1 \sum_{i=0}^{n-1} \frac{1}{\rho(t+i)}. \quad (30)$$

From the definition of $\{\rho(t)\}_{t\in N}$, we obtain

$$\sum_{i=0}^{n} \frac{1}{\rho(t+i)} \leq \sum_{i=1}^{\lceil \frac{n}{\kappa} \rceil} \sum_{j=0}^{\kappa-1} \frac{1}{\rho(t+i\kappa+j)} \quad (31)$$

$$\leq \sum_{i=0}^{\lceil \frac{n}{\kappa} \rceil} \frac{\kappa}{\Delta^i \rho(t)} \quad (32)$$

$$\leq \frac{\kappa}{\rho(t)} \sum_{i=0}^{\infty} \frac{1}{\Delta^i} \quad (33)$$

where (31) follows because the sum on the right contains all of the terms of the sum on the left (and possibly more) and all of the terms are positive, (32) follows because $1/\rho(t + i\kappa + j) \leq 1/(\Delta^i\rho(t))$ for all $0 \leq j \leq \kappa - 1$, and (33) trivially follows from the non-negativity of summands. Since $\Delta > 1$, $\sum_{i=0}^{\infty} 1/\Delta^i$ is a convergent geometric series and, thus, let $\sum_{i=0}^{\infty} \kappa/\Delta^i = M_2$. This, together with (30) and (31)–(33), implies that for all integers $t, n \geq 0$

$$r(t+n) \leq r(t) + \frac{8M_1M_2}{\rho(t)}. \quad (34)$$

Now note that $\{r(t)\}_{t\in\mathbb{N}}$ is bounded. Moreover, because $\{\rho(t)\}_{t\in\mathbb{N}}$ is an increasing sequence, it follows that for all $\epsilon > 0$, there exists a T such that $(8M_1M_2/\rho(t)) \leq \epsilon$, for all $t \geq T$. These, taken together with (34) and Lemma 2 (see p. 7), ensure that the sequence $\{r(t)\}_{t\in\mathbb{N}}$ *converges* to a *finite* value, denoted by R, that is, $R = \lim_{t\to\infty} r(t)$.

Let us finally show that $R = 0$. Since the set $\mathcal{X} \times \mathcal{Z}$ is compact, the sequence $\{\mathbf{x}(t), \mathbf{z}(t)\}_{t\in\mathbb{N}}$ has a limit point, say $(\bar{\mathbf{x}}, \bar{\mathbf{z}}) \in \mathcal{X} \times \mathcal{Z}$. Moreover, note that the function $\|\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{c}\|^2$ is continuous on $\mathcal{X} \times \mathcal{Z}$. Therefore, taking limits as $t \to \infty$ in $r(t) = \|\mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{z}(t) - \mathbf{c}\|^2$, we have

$$R = \lim_{t\to\infty} \|\mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{z}(t) - \mathbf{c}\|^2 = \|\mathbf{A}\bar{\mathbf{x}} + \mathbf{B}\bar{\mathbf{z}} - \mathbf{c}\|^2. \quad (35)$$

Let us now consider the limits in the inequality (29) as $t \to \infty$. Since $\lim_{t\to\infty} r(t+1) = \lim_{t\to\infty}((8M_1)/\rho(t) + r(t)) = R$, from (27), (28), and the squeezing lemma, together with the continuity of functions $\hat{\mathbf{x}}$ and $\hat{\mathbf{z}}$ it follows that:

$$R = \|\mathbf{A}\hat{\mathbf{x}}(\bar{\mathbf{z}}) + \mathbf{B}\bar{\mathbf{z}} - \mathbf{c}\|^2 = \|\mathbf{A}\bar{\mathbf{x}} + \mathbf{B}\hat{\mathbf{z}}(\bar{\mathbf{x}}) - \mathbf{c}\|^2. \quad (36)$$

By combining (35) and (36), together with the definitions (22) and (23), we obtain

$$\bar{\mathbf{x}} = \arg\min_{\mathbf{x}\in\mathcal{X}} \|\mathbf{A}\mathbf{x} + \mathbf{B}\bar{\mathbf{z}} - \mathbf{c}\|^2 \quad (37)$$

$$\bar{\mathbf{z}} = \arg\min_{\mathbf{z}\in\mathcal{Z}} \|\mathbf{A}\bar{\mathbf{x}} + \mathbf{B}\mathbf{z} - \mathbf{c}\|^2. \quad (38)$$

Since Slater's constraint qualifications condition is satisfied for both sets $\mathcal{X}$ and $\mathcal{Z}$ (Assumption 5), $\bar{\mathbf{x}}$ and $\bar{\mathbf{z}}$ satisfy the first-order necessary conditions for problems (37) and (38), respectively. By combining these first-order necessary conditions and (38), it follows that $(\bar{\mathbf{x}}, \bar{\mathbf{z}})$ satisfies the first-order necessary conditions for the problem:

$$\begin{aligned} \underset{\mathbf{x},\mathbf{z}}{\text{minimize}} \quad & \|\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{c}\|^2 \\ \text{subject to} \quad & (\mathbf{x}, \mathbf{z}) \in \mathcal{X} \times \mathcal{Z}. \end{aligned} \quad (39)$$

Since problem (39) is convex and the constraint sets satisfy Slater's constraint qualifications condition, we conclude that $(\bar{\mathbf{x}}, \bar{\mathbf{z}})$ is the solution to problem (39). Given that problem (2) is feasible, we must have $\|\mathbf{A}\bar{\mathbf{x}} + \mathbf{B}\bar{\mathbf{z}} - \mathbf{c}\|^2 = 0$ and, therefore, $\lim_{t\to\infty} \|\mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{z}(t) - \mathbf{c}\|^2 = 0$ [compared with (35)]. ■

Lemma 2: Let us suppose that $\{a_t\}_{t\in\mathbb{N}}$ is a bounded sequence and for each $\epsilon > 0$, there exists $T \in \mathbb{N}$ such that $a_{t+n} \leq \epsilon + a_t$ for all $n \geq 0$ and $t \geq T$, then $\lim_{t\to\infty} a_t$ exists.

Proof: Let us denote by R the limit inferior of $\{a_t\}_{t\in\mathbb{N}}$, that is, $R = \liminf_{t\to\infty} a_t$, which is finite since $\{a_t\}_{t\in\mathbb{N}}$ is bounded. It follows from elementary properties of the limit inferior that $\{a_t\}_{t\in\mathbb{N}}$ has a subsequence $\{a_{t_j}\}_{j\in\mathbb{N}}$ which converges to R , that is, $\lim_{j\to\infty} a_{t_j} = R$. Subsequently, for a given $\epsilon > 0$, we can find $J_1 \in \mathbb{N}$ such that $|R - a_{t_j}| < \epsilon/2$ for all $j \geq J_1$. Moreover, by using the assumptions of the lemma, there exists $J_2 \in \mathbb{N}$ such that $a_{t_j+n} \leq \epsilon/2 + a_{t_j}$ for all $n \geq 0$ and $j \geq J_2$. If we choose $J = \max\{J_1, J_2\}$, we get that $a_t \leq \epsilon/2 + a_{t_J} < R + \epsilon$ for all $t \geq t_J$. Since this can be done for all $\epsilon > 0$, we get that $\limsup_{t\to\infty} a_t \leq R$, implying that $\limsup_{t\to\infty} a_t = \liminf_{t\to\infty} a_t$. So we can conclude that $\lim_{t\to\infty} a_t = R$. ■

One natural question that arises immediately with Assumption 4 is what if $\mathcal{X}$ and/or $\mathcal{Z}$ are *nonconvex*. The following example shows that the results of Proposition 2 do not generally hold when either $\mathcal{X}$ or $\mathcal{Z}$ are nonconvex.

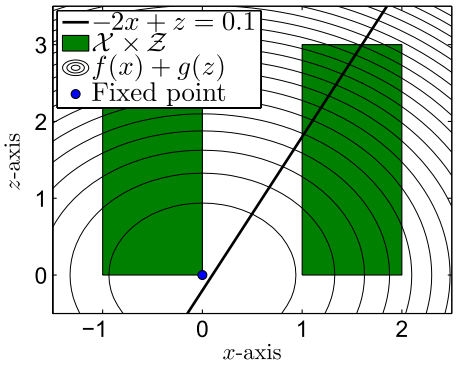

Fig. 1. Example where ADPM fails to converge to a feasible point when sets $\mathcal{X}$ and $\mathcal{Z}$ are nonconvex.

Example 4: Consider the problem

$$\begin{aligned} \underset{x,z}{\text{minimize}} \quad & x^2 + z^2 \\ \text{subject to} \quad & -2x + z = 0.1, \\ & x \in [-1, 0] \cup [1, 2], \quad z \in [0, 3]. \end{aligned} \quad (40)$$

The feasibility set and contours of the objective function are given in Fig. 1. It can be observed that if $z(0) = 0$ and $\mathbf{y}(t) = \mathbf{0}$ for all $t \in \mathbb{N}$, then the optimal solution of the $\mathbf{x}$- and $\mathbf{z}$-updates is 0 for all $t \in \mathbb{N}$, that is, $\lim_{t\to\infty} x(t) = 0$ and $\lim_{t\to\infty} z(t) = 0$. This means that the algorithm converges to (0,0), which is an infeasible point.

Note that our Assumption 3 is a weaker condition than assuming that f and/or g are convex. As a result, generally characterizing the proprieties of the objective value of ADPM after the convergence is technically challenging. Nevertheless, ADPM appears to resemble a sequential optimization approach, which provides degrees of freedom to hover over the true objective function for locating a *good* objective value. In [32], we provide some experiments to numerically show these appealing aspects of the ADPM, besides those ensured by Proposition 2.

IV. Alternating Direction Method of Multipliers

In this section, we investigate some new general properties of the ADMM in a nonconvex setting. We state the algorithm in Section IV-A and study convergence properties in Section IV-B.

A. Algorithm Description

The ADMM can explicitly be stated as follows.

Algorithm 2: The Alternating Direction Method of Multipliers (ADMM)

1) **Initialization**: Set $t = 0$ and put initial values to $\mathbf{z}(t)$, $\mathbf{y}(t)$, and ρ.
2) $\mathbf{x}$**-update**: $\mathbf{x}(t+1) = \arg\min_{\mathbf{x}\in\mathcal{X}} L_\rho(\mathbf{x}, \mathbf{z}(t), \mathbf{y}(t))$.
3) $\mathbf{z}$**-update**: $\mathbf{z}(t+1) = \arg\min_{\mathbf{z}\in\mathcal{Z}} L_\rho(\mathbf{x}(t+1), \mathbf{z}, \mathbf{y}(t))$.
4) $\mathbf{y}$**-update**: $\mathbf{y}(t+1) = \mathbf{y}(t) + \rho(\mathbf{A}\mathbf{x}(t+1) + \mathbf{B}\mathbf{z}(t+1) - \mathbf{c})$.
5) **Stopping criterion**: If stopping criterion is met terminate, otherwise set $t = t + 1$ and go to step 2.

Unlike in Algorithm 1 (ADPM), in Algorithm 2 (ADMM), the penalty parameter is fixed. The first step is the initialization (step 1). As presented before, the $\mathbf{x}$- and $\mathbf{z}$-updates require a *solution* of an optimization problem. This is not as restrictive as it may seem, since under mild conditions such requirements are accomplished, see Examples 1–3. However, we note that no such global optimality requirement of $\mathbf{x}(t+1)$ and $\mathbf{z}(t+1)$ is necessary in our convergence assertions, as we will show in subsequent sections. More specifically, our convergence results apply as long as $\mathbf{x}(t+1)$ [respectively, $\mathbf{z}(t+1)$] is a local minimum.

B. Algorithm Properties

In this section, we show that, under mild assumptions, if the sequence $\{\mathbf{y}(t)\}_{t\in\mathbb{N}}$ converges to some $\bar{\mathbf{y}}$, then any limit point of the sequence $\{\mathbf{x}(t),\mathbf{z}(t)\}_{t\in\mathbb{N}}$, together with $\bar{\mathbf{y}}$, satisfies FON conditions of Problem (2) (compared with Definition 1). It is worth noting that these results hold regardless of whether f, g, $\mathcal{X}$, and $\mathcal{Z}$ are convex or nonconvex.

Let us now scrutinize the above assertion precisely. The analysis is based on the following assumption which can be expected to hold for many problems of practical interest:

Assumption 7: The sets $\mathcal{X}$ and $\mathcal{Z}$ of problem (2) are closed and can be expressed in terms of a finite number of equality and inequality constraints. In particular

$$\mathcal{X}=\{\mathbf{x}\in\mathbb{R}^{p_1}\mid \boldsymbol{\psi}(\mathbf{x})=\mathbf{0},\quad \boldsymbol{\phi}(\mathbf{x})\leq\mathbf{0}\}$$
$$\mathcal{Z}=\{\mathbf{z}\in\mathbb{R}^{p_2}\mid \boldsymbol{\theta}(\mathbf{z})=\mathbf{0},\quad \boldsymbol{\sigma}(\mathbf{z})\leq\mathbf{0}\}$$

where $\boldsymbol{\psi}:\mathbb{R}^{p_1}\to\mathbb{R}^{q_1}$, $\boldsymbol{\phi}:\mathbb{R}^{p_1}\to\mathbb{R}^{q_2}$, $\boldsymbol{\theta}:\mathbb{R}^{p_2}\to\mathbb{R}^{q_3}$, and $\boldsymbol{\sigma}:\mathbb{R}^{p_2}\to\mathbb{R}^{q_4}$ are continuously differentiable functions.

Assumption 8: For every $t\in\mathbb{N}$, $\mathbf{x}(t)$ [respectively, $\mathbf{z}(t)$] computed at step 2 (respectively, step 3) of the ADMM algorithm is locally *or* globally optimal.

Assumption 9: Let $\mathcal{L}$ denote the set of limit points of the sequence $\{(\mathbf{x}(t),\mathbf{z}(t))\}_{t\in\mathbb{N}}$ and let $(\bar{\mathbf{x}},\bar{\mathbf{z}})\in\mathcal{L}$. The set of constraint gradient vectors at $\bar{\mathbf{x}}$

$$\mathcal{C}_{\mathcal{X}}(\bar{\mathbf{x}})=\{\nabla\boldsymbol{\psi}_i(\bar{\mathbf{x}})|i=1,\ldots,q_1\}\cup\{\nabla\boldsymbol{\phi}_i(\bar{\mathbf{x}})|i\in\mathcal{A}_{\mathcal{X}}(\bar{\mathbf{x}})\}\quad(41)$$

associated with the set $\mathcal{X}$ is linearly independent, where $\mathcal{A}_{\mathcal{X}}(\bar{\mathbf{x}})=\{i\mid\boldsymbol{\phi}_i(\bar{\mathbf{x}})=0\}$. Similarly, the corresponding set of constraint gradient vectors $\mathcal{C}_{\mathcal{Z}}$ associated with the set $\mathcal{Z}$ is linearly independent.

Assumption 7 is self-explanatory. Note that steps 2 and 3 of the algorithm involve nonconvex optimization problems, where the computational cost of finding the solutions $\mathbf{x}(t+1)$ and $\mathbf{z}(t+1)$, in general, can be entirely prohibitive. However, Assumption 8 indicates that the solution $\mathbf{x}(t+1)$ [respectively, $\mathbf{z}(t+1)$] of the optimization problem associated with steps 2 (respectively, 3) of the ADMM should *only* be a *local minimum* and not necessarily a global minimum. Thus, Assumption 8 can usually be accomplished by employing efficient local optimization methods (see [33, Sec. 1.4.1]). In the literature, Assumption 9 is called the "regularity assumption" and is usually satisfied in practice. Moreover, any point that complies with the assumption is called regular, see [18, p. 269]. Let us next document two results that will be important later.

Lemma 3: Suppose Assumptions 7 and 9 hold. Let $\{(\mathbf{x}(t_k),\mathbf{z}(t_k))\}_{k\in\mathbb{N}}$ be a subsequence of $\{(\mathbf{x}(t),\mathbf{z}(t))\}_{t\in\mathbb{N}}$ with $\lim_{k\to\infty}(\mathbf{x}(t_k),\mathbf{z}(t_k))=(\bar{\mathbf{x}},\bar{\mathbf{z}})$. Then there exists K such that the sets of vectors $\mathcal{C}_{\mathcal{X}}(\mathbf{x}(t_k))$ and $\mathcal{C}_{\mathcal{Z}}(\mathbf{z}(t_k))$ [cf (41)] are each linearly independent for all $k\geq K$.

Proof: First note that if $i\notin\mathcal{A}(\bar{\mathbf{x}})$, then $\boldsymbol{\phi}_i(\mathbf{x}(t_k))<0$ [or $i\notin\mathcal{A}(\mathbf{x}(t_k))$] for all sufficiently large k, since $\boldsymbol{\phi}_i$ is continuous and the set $\{x\in\mathbb{R}|x\neq0\}$ is open. Therefore, it suffices to show that the columns of the matrix $\mathbf{D}(\mathbf{x}(t_k))\in\mathbb{R}^{p_1\times(q_1+|\mathcal{A}(\bar{\mathbf{x}})|)}$ are linearly independent for all sufficiently large k, where

$$\mathbf{D}(\mathbf{x})=\left[(\nabla\boldsymbol{\psi}_i(\mathbf{x}))_{i=1,\ldots,q_1},(\nabla\boldsymbol{\phi}_i(\mathbf{x}))_{i\in\mathcal{A}_{\mathcal{X}}(\bar{\mathbf{x}})}\right].\quad(42)$$

Since $\mathrm{Det}(\mathbf{D}(\mathbf{x})^{\mathrm{T}}\mathbf{D}(\mathbf{x}))$ is continuous (see Assumption 7), $\mathrm{Det}(\mathbf{D}(\mathbf{x}(t_k))^{\mathrm{T}}\mathbf{D}(\mathbf{x}(t_k)))$ can be made arbitrarily close to $\mathrm{Det}(\mathbf{D}(\bar{\mathbf{x}})^{\mathrm{T}}\mathbf{D}(\bar{\mathbf{x}}))$, which is *nonzero*, see Assumption 9. Equivalently, there exists $K\in\mathbb{N}$ such that $\mathrm{Det}(\mathbf{D}(\mathbf{x}(t_k))^{\mathrm{T}}\mathbf{D}(\mathbf{x}(t_k)))$ is nonzero for all $k\geq K$, which, in turn, ensures that $\mathcal{C}_{\mathcal{X}}(\mathbf{x}(t_k))$ is a linearly independent set for $k\geq K$. The linear independence of $\mathcal{C}_{\mathcal{Z}}(\mathbf{z}(t_k))$ for all sufficiently large k can be proved similarly. ■

Lemma 4: Suppose Assumptions 3, 7, 8, and 9 hold. Let $\{(\mathbf{x}(t_k),\mathbf{z}(t_k))\}_{k\in\mathbb{N}}$ be a subsequence of $\{(\mathbf{x}(t),\mathbf{z}(t))\}_{t\in\mathbb{N}}$ with $\lim_{k\to\infty}(\mathbf{x}(t_k),\mathbf{z}(t_k))=(\bar{\mathbf{x}},\bar{\mathbf{z}})$. Then, for sufficiently large k, there exist Lagrange multipliers $(\boldsymbol{\lambda}(t_k),\boldsymbol{\gamma}(t_k))\in\mathbb{R}^{q_1}\times\mathbb{R}^{q_2}$ [*respectively*, $(\boldsymbol{\mu}(t_k),\boldsymbol{\omega}(t_k))\in\mathbb{R}^{q_3}\times\mathbb{R}^{q_4}$] such that the pair $\mathbf{x}(t_k)$, $(\boldsymbol{\lambda}(t_k),\boldsymbol{\gamma}(t_k))$ [*respectively*, $\mathbf{z}(t_k)$, $(\boldsymbol{\mu}(t_k),\boldsymbol{\omega}(t_k))$] satisfies the FON conditions of the optimization problem in the $\mathbf{x}$-(*respectively*, $\mathbf{z}$-) update of the ADMM algorithm (compare with Definition 1).

Proof: From Lemma 3, we have that $\mathbf{x}(t_k)$ and $\mathbf{z}(t_k)$ are regular for sufficiently large k. This combined with the assumptions yields the result, which is an immediate consequence of [18, Prop. 3.3.1] ■

Lemmas 3 and 4 play a central role when deriving our convergence results, as we will show in the sequel. The following proposition establishes the convergence results of the ADMM algorithm:

Proposition 3: Suppose Assumptions 3, 7, 8, and 9 hold and the sequence $\mathbf{y}(t)$ converges to a point, that is, $\lim_{t\to\infty}\mathbf{y}(t)=\bar{\mathbf{y}}$ for some $\bar{\mathbf{y}}$. Then every limit point of the sequence $\{\mathbf{x}(t),\mathbf{z}(t)\}_{t\in\mathbb{N}}$, together with $\bar{\mathbf{y}}$ and some $\boldsymbol{\lambda}\in\mathbb{R}^{q_1}$, $\boldsymbol{\gamma}\in\mathbb{R}^{q_2}$, $\boldsymbol{\mu}\in\mathbb{R}^{q_3}$, and $\boldsymbol{\omega}\in\mathbb{R}^{q_4}$ satisfy the FON conditions of Problem (2).

Proof: Let $(\bar{\mathbf{x}},\bar{\mathbf{z}})$ be a limit point of $\{(\mathbf{x}(t),\mathbf{z}(t))\}_{t\in\mathbb{N}}$ and $\{(\mathbf{x}(t_k),\mathbf{z}(t_k))\}_{k\in\mathbb{N}}$ be a subsequence such that $\lim_{k\to\infty}(\mathbf{x}(t_k),\mathbf{z}(t_k))=(\bar{\mathbf{x}},\bar{\mathbf{z}})$. We show that the primal variables $\bar{\mathbf{x}},\bar{\mathbf{z}}$ and the Lagrange multipliers $\bar{\mathbf{y}}$, $\boldsymbol{\lambda}$, $\boldsymbol{\gamma}$, $\boldsymbol{\mu}$, and $\boldsymbol{\omega}$ satisfy the first-order necessary conditions, where $\boldsymbol{\lambda}$, $\boldsymbol{\gamma}$, $\boldsymbol{\mu}$, and $\boldsymbol{\omega}$ are chosen as in Lemma 5.

In the sequel, we show that the four conditions of Definition 1 (first-order necessary condition) are all satisfied.

1) Primal feasibility: Since $(\mathbf{x}(t_k),\mathbf{z}(t_k))\in\mathcal{X}\times\mathcal{Z}$ and the set $\mathcal{X}\times\mathcal{Z}$ is closed, it follows that $(\bar{\mathbf{x}},\bar{\mathbf{z}})\in\mathcal{X}\times\mathcal{Z}$. Since $\bar{\mathbf{y}}=\mathbf{y}(0)+\sum_{t=1}^{\infty}\rho(\mathbf{A}\mathbf{x}(t)+\mathbf{B}\mathbf{z}(t)-\mathbf{c})$, we must have $\lim_{t\to\infty}\|\mathbf{A}\mathbf{x}(t)+\mathbf{B}\mathbf{z}(t)-\mathbf{c}\|^2=0$, or $\mathbf{A}\bar{\mathbf{x}}+\mathbf{B}\bar{\mathbf{z}}=\mathbf{c}$.

2) Dual feasibility: It holds for $\boldsymbol{\gamma}(t_k)$ and $\boldsymbol{\omega}(t_k)$ from Lemma 4 that $\boldsymbol{\gamma}(t_k)\geq\mathbf{0}$ and $\boldsymbol{\omega}(t_k)\geq\mathbf{0}$ (compare with Definition 1). Hence, since the closed right half-plane is a closed set, it follows that $\boldsymbol{\gamma}\geq\mathbf{0}$ and $\boldsymbol{\omega}\geq\mathbf{0}$.

3) Complementary slackness: If $\boldsymbol{\phi}_i(\bar{\mathbf{x}})=0$, then $\boldsymbol{\gamma}_i\boldsymbol{\phi}_i(\bar{\mathbf{x}})=0$ trivially holds. On the other hand, if $\boldsymbol{\phi}_i(\bar{\mathbf{x}})<0$, then we showed in the proof of Lemma 5 that $\boldsymbol{\gamma}_i=0$. Hence, it follows that $\boldsymbol{\gamma}_i\boldsymbol{\phi}_i(\bar{\mathbf{x}})=0$.

4) Lagrangian vanishes: We need to show that

$$\nabla_{\mathbf{x}} f(\bar{\mathbf{x}})+\mathbf{A}^{\mathrm{T}}\bar{\mathbf{y}}+\nabla_{\mathbf{x}}\boldsymbol{\psi}(\bar{\mathbf{x}})\boldsymbol{\lambda}+\nabla_{\mathbf{x}}\boldsymbol{\phi}(\bar{\mathbf{x}})\boldsymbol{\gamma}=\mathbf{0} \quad (43)$$

$$\nabla_{\mathbf{z}} g(\bar{\mathbf{z}})+\mathbf{B}^{\mathrm{T}}\bar{\mathbf{y}}+\nabla_{\mathbf{z}}\boldsymbol{\theta}(\bar{\mathbf{z}})\boldsymbol{\mu}+\nabla_{\mathbf{z}}\boldsymbol{\sigma}(\bar{\mathbf{z}})\boldsymbol{\omega}=\mathbf{0}. \quad (44)$$

Let us start by showing (44). From Lemma 4, we obtain for all sufficiently large k that (compare with Definition 1)

$$\nabla_{\mathbf{z}} L\left(\mathbf{x}(t_k),\mathbf{z},\mathbf{y}(t_k-1)\right)+\nabla_{\mathbf{z}}\boldsymbol{\theta}\left(\mathbf{z}(t_k)\right)\boldsymbol{\mu}(t_k)+\nabla_{\mathbf{x}}\boldsymbol{\sigma}\left(\mathbf{x}(t_k)\right)\boldsymbol{\omega}(t_k)=\mathbf{0}. \quad (45)$$

By using $\mathbf{y}(t_k-1)=\mathbf{y}(t_k)-\rho(\mathbf{A}\mathbf{x}(t_k)+\mathbf{B}\mathbf{z}(t_k)-\mathbf{c})$ in (45) and rearranging the terms, we obtain

$$\nabla_{\mathbf{z}} g\left(\mathbf{z}(t_k)\right)+\mathbf{B}^{\mathrm{T}}\mathbf{y}(t_k)+\nabla_{\mathbf{z}}\boldsymbol{\theta}\left(\mathbf{z}(t_k)\right)\boldsymbol{\mu}(t_k)+\nabla_{\mathbf{x}}\boldsymbol{\sigma}\left(\mathbf{x}(t_k)\right)\boldsymbol{\omega}(t_k)=\mathbf{0}. \quad (46)$$

By using that $\lim_{k\to\infty}(\mathbf{x}(t_k),\mathbf{z}(t_k),\mathbf{y}(t_k))=(\bar{\mathbf{x}},\bar{\mathbf{z}},\bar{\mathbf{y}})$ and $\lim_{k\to\infty}(\boldsymbol{\lambda}(t_k),\boldsymbol{\gamma}(t_k),\boldsymbol{\mu}(t_k),\boldsymbol{\omega}(t_k))=(\boldsymbol{\lambda},\boldsymbol{\gamma},\boldsymbol{\mu},\boldsymbol{\omega})$, we conclude that (44) holds. By using the same arguments as before, we get for all sufficiently large k that

$$\nabla_{\mathbf{x}} f\left(\mathbf{x}(t_k)\right)+\mathbf{A}^{\mathrm{T}}\mathbf{y}(t_k)+\nabla_{\mathbf{x}}\boldsymbol{\psi}\left(\mathbf{x}(t_k)\right)\boldsymbol{\lambda}(t_k)+\nabla_{\mathbf{x}}\boldsymbol{\phi}\left(\mathbf{x}(t_k)\right)\boldsymbol{\gamma}(t_k)=\rho\mathbf{A}^{\mathrm{T}}\mathbf{B}\left(\mathbf{z}(t_k)-\mathbf{z}(t_k-1)\right). \quad (47)$$

Therefore, by the arguments above, if we can show that $\lim_{t\to\infty}\rho\mathbf{A}^{\mathrm{T}}\mathbf{B}(\mathbf{z}(t+1)-\mathbf{z}(t))=\mathbf{0}$, then (43) holds. The assumption $\bar{\mathbf{y}}=\lim_{t\to\infty}\mathbf{y}(t)$, together with the relation $\mathbf{y}(t+1)=\mathbf{y}(0)+\rho\sum_{l=1}^{t+1}\mathbf{A}\mathbf{x}(l)+\mathbf{B}\mathbf{z}(l)-\mathbf{c}$, can be used to show that the series

$$\sum_{t=1}^{\infty}\left(\mathbf{A}\mathbf{x}(t)+\mathbf{B}\mathbf{z}(t+1)-\mathbf{c}\right),\quad \sum_{t=1}^{\infty}\left(\mathbf{A}\mathbf{x}(t)+\mathbf{B}\mathbf{z}(t)-\mathbf{c}\right)$$

are convergent. By taking the difference of the two series and using that the sum of convergent series is a convergent series, we get that $\sum_{t=1}^{\infty}\mathbf{B}(\mathbf{z}(t+1)-\mathbf{z}(t))$ is a convergent series, thus implying that $\lim_{t\to\infty}\mathbf{B}(\mathbf{z}(t+1)-\mathbf{z}(t))=0$. By multiplying $\rho\mathbf{A}^{\mathrm{T}}$ from the left side, we get that $\lim_{t\to\infty}\rho\mathbf{A}^{\mathrm{T}}\mathbf{B}(\mathbf{z}(t+1)-\mathbf{z}(t))=0$. ■

Lemma 5: Let $\{t_k\}_{k\in\mathbb{N}}$ be a sequence such that $\lim_{k\to\infty}(\mathbf{x}(t_k),\mathbf{z}(t_k))=(\bar{\mathbf{x}},\bar{\mathbf{z}})$. Then the limits $\lim_{k\to\infty}\boldsymbol{\lambda}(t_k)$, $\lim_{k\to\infty}\boldsymbol{\gamma}(t_k)$, $\lim_{k\to\infty}\boldsymbol{\mu}(t_k)$, and $\lim_{k\to\infty}\boldsymbol{\omega}(t_k)$ exist, where $\boldsymbol{\lambda}(t_k)$, $\boldsymbol{\gamma}(t_k)$, $\boldsymbol{\mu}(t_k)$, and $\boldsymbol{\omega}(t_k)$ are chosen as in Lemma 4.

Proof: We prove the existence of the first two limits. The proof of the existence of the latter two limits follows similarly.

Since ∇f, $\nabla\boldsymbol{\psi}$, and $\nabla\boldsymbol{\phi}$ are continuous functions (see Assumption 3), we have

$$\lim_{k\to\infty}\nabla f\left(\mathbf{x}(t_k)\right)=\nabla f(\bar{\mathbf{x}}),\quad \lim_{k\to\infty}\nabla\boldsymbol{\psi}\left(\mathbf{x}(t_k)\right)=\nabla\boldsymbol{\psi}(\bar{\mathbf{x}})$$
$$\text{and}\qquad \lim_{k\to\infty}\nabla\boldsymbol{\phi}\left(\mathbf{x}(t_k)\right)=\nabla\boldsymbol{\phi}(\bar{\mathbf{x}}).$$

This, together with Lemma 3, implies that there exists K such that $\mathbf{D}(\mathbf{x}(t_k))^{\mathrm{T}}\mathbf{D}(\mathbf{x}(t_k))$ [see (42)] is invertible for all $k\geq K$. Hence, it follows that for all $k\geq K$, we have:

$$\left(\boldsymbol{\lambda}(t_k),(\boldsymbol{\gamma}_i(t_k))_{i\in\mathcal{A}_{\mathcal{X}}(\bar{\mathbf{x}})}\right)=\mathbf{D}(\mathbf{x}(t_k))^{\mathrm{T}}\mathbf{D}(\mathbf{x}(t_k))^{-1}\mathbf{D}(\mathbf{x}(t_k))^{\mathrm{T}}\left(\nabla f(\mathbf{x}(t_k))+\mathbf{A}^{\mathrm{T}}\mathbf{x}(t_k)\right).$$

Since $\mathbf{D}(t_k)$ and $\nabla f(\mathbf{x}(t_k))$ converge when $k\to\infty$, it follows that $\lim_{k\to\infty}(\boldsymbol{\lambda}(t_k),(\boldsymbol{\gamma}_i(t_k))_{i\in\mathcal{A}_{\mathcal{X}}(\bar{\mathbf{x}})})$ exists.

Next, we show that $\lim_{k\to\infty}\boldsymbol{\gamma}_i(t_k)=0$ if $i\notin\mathcal{A}_{\mathcal{X}}(\bar{\mathbf{x}})$. Since $\boldsymbol{\phi}_i(\bar{\mathbf{x}})<0$, there exists an open set $\mathcal{U}\subseteq\mathbb{R}^{p_2}$ containing $\bar{\mathbf{x}}$ such that $\boldsymbol{\phi}_i(\mathbf{x})<0$ for all $\mathbf{x}\in\mathcal{U}$. In particular, there exists $K\in\mathbb{N}$ such that $\boldsymbol{\phi}_i(\mathbf{x}(t_k))<0$ for $k\geq K$. Therefore, there must exist $K\in\mathbb{N}$ such that $\boldsymbol{\gamma}_i(t_k)=0$ for all $k\geq K$, since complementary slackness $[\boldsymbol{\gamma}_i(t_k)\boldsymbol{\phi}_i(\mathbf{x}(t_k))=0]$ holds for all sufficiently large k (compare with Lemma 4). ■

A stronger version of Proposition 3 is shown in the following corollary:

Corollary 1: If $\lim_{t\to}(\mathbf{x}(t),\mathbf{z}(t),\mathbf{y}(t))=(\bar{\mathbf{x}},\bar{\mathbf{z}},\bar{\mathbf{y}})$, then $\bar{\mathbf{x}}$ and $\bar{\mathbf{z}}$ satisfy the FON conditions of Problem (2).

The corollary follows immediately because the hypothesis implies that the set $\mathcal{L}$ defined in Assumption 9 is a Singleton.

Technically, Proposition 3 characterizes the solution of the ADMM algorithm applied on the possibly nonconvex problem (2). More specifically, the proposition claims that under mild assumptions, the solutions computed by ADMM satisfy the FON conditions for problem (2), if at every iteration, the subproblems are locally (or globally) solved and if the dual variables of ADMM converge.

Let us now show how Proposition 3 can be used to completely characterize the convergence of the ADMM for a class of problems identified by the following assumption.

Assumption 10: $f,g:\mathbb{R}\to\mathbb{R}$, $\mathcal{X}=\mathcal{Y}=\mathbb{R}$, and the coupling constraint is $x=z$, that is, $\mathbf{A}=1$ and $\mathbf{B}=-1$. In addition, the derivatives f' and g' are L-Lipschitz continuous.

The following corollary of Proposition 3 shows that under Assumption 10, the ADMM always either converges or diverges to $\pm\infty$ and characterizes the convergence in terms of $z(0)$.

Corollary 2: Suppose Assumption 10 holds, $\rho>L$ and $y(0)=g'(z(0))$. Then

$$\lim_{k\to\infty}(x(k),z(k),y(k))=(z^\star,z^\star,g'(z^\star))$$

where $z^\star$ is determined as follows:

a) If $f'(z(0))+g'(z(0))=0$, then $z^\star=z(0)$.

b) If $f'(z(0))+g'(z(0))<0$, then

$$z^\star=\inf\left\{z\geq z(0)\middle|f'(z)+g'(z)=0\right\}.$$

c) If $f'(z(0))+g'(z(0))>0$, then

$$z^\star=\sup\left\{z\leq z(0)\middle|f'(z)+g'(z)=0\right\}.$$

Proof: We start by writing the steps of the ADMM in a more convenient form. Note that $g'(z(t+1))+y(t)+\rho(x(t+1)-z(t+1))=0$, from the optimality conditions of $z(t+1)$ at the $\mathbf{z}$-update. This, combined with the $\mathbf{y}$-update, yields 1) $y(t)=g'(z(t))$. Moreover, because f' and g' are L-Lipschitz continuous, we have that 2) the functions $L_\rho(\cdot,z(t),y(t))$ and $L_\rho(x(t),\cdot,y(t))$, associated with the $\mathbf{x}$- and $\mathbf{z}$-updates are strongly convex for all $\rho>L$.

From 1) and 2), we get that $x(t+1)$ is the unique solution to

$$0=f'(x)+g'\left(z(t)\right)+\rho\left(x-z(t)\right) \quad (48)$$

and $z(t+1)$ is the unique solution to

$$0=g'(z)-g'\left(z(t)\right)-\rho\left(x(t+1)-z\right). \quad (49)$$

In the sequel, we show each case a), b), and c) separately.

a) If $f'(z(0))+g(z(0))=0$, then $x(t)=z(0)$ and $z(t)=z(0)$ are clearly the unique solutions to (48) and (49), respectively, for all $t\geq 1$. The result follows.

b) In the sequel, we show that $z(t+1)>z(t)$ and $z(t)<z^\star$ for all $t\in\mathbb{N}$, implying that $\bar{z}=\lim_{t\to\infty}z(t)$ exists (it is possible that $\bar{z}=\infty$ when $z^\star=\infty$). Since the interval $\mathcal{U}=[z(0),z^\star]$ (or $\mathcal{U}=[z(0),z^\star[$ when $z^\star=\infty$) is a closed set, $\bar{z}\in\mathcal{U}$. Moreover, by Proposition 3, $(x(t),z(t),y(t))$ can only converge to a point satisfying the first-order necessary conditions, that is, to a point $(z,z,g'(z))$ with $f'(z)+g'(z)=0$. When $z^\star<\infty$, the only $z\in\mathcal{U}$ satisfying the necessary conditions is $z^\star$ and when $z^\star=\infty$, no $z\in\mathcal{U}$ satisfies the necessary conditions. Hence, we can conclude that $\bar{z}=z^\star$.

We show that $z(t+1)>z(t)$ and $z(t+1)<z^\star$ for all $z(t)\in[z(0),z^\star[$, but as an intermediary step, we first show that $x(t+1)>z(k)$ and $x(t+1)<z^\star$ for all $z(t)\in[z(0),z^\star[$. To see that $x(t+1)>z(k)$, we note that $x(t+1)\leq z(k)$ contradicts the L-Lipschitz continuity of f'. In particular, $x(t+1)\leq z(t)$ implies that $\rho|x(t+1)-z(t)|<|f'(x(t+1))-f'(z(t))|$, which is seen by the following inequality:

$$\rho\left(z(t)-x(t+1)\right)<f'\left(x(t+1)\right)-f'\left(z(t)\right) \tag{50}$$

which is obtained by combining (48) and $-f'(z(t))>g'(x(t))$ and rearranging. To see that $x(t+1)<z^\star$, we note that

$$f'(x)<-(g'\left(z(t)\right)+\rho\left(x-z(t)\right)),\ \forall x\in[z(t),x(t+1)[\,, \tag{51}$$
$$g'(x)<g'(z(t))+\rho\left(x-z(t)\right),\ \forall x\in]z(t),x(t+1)] \tag{52}$$

where (51) comes from that $x(t+1)$ is the unique solution of (48) and $f'(z(t))<-g'(z(t))-\rho(z(t)-z(t))$ and (52) come from that $\rho>L$ and g' is L-Lipschitz continuous. Summing (51) and (52) and using the continuity of f' and g' shows that $f'(x)+g'(x)<0$ for all $x\in[z(t),x(t+1)]$ and, hence, $x(t+1)<z^\star$.

We now show that $z(t+1)>z(t)$ and $z(t)<z^\star$. To see that $z(t+1)>z(t)$, we note that $z(t+1)\leq z(t)$ contradicts the L-Lipschitz continuity of g'. In particular, $z(t+1)\leq z(t)$ implies that $\rho|z(t+1)-z(t)|<|g'(x(t+1))-g'(z(t))|$, which is seen by that if $z(t+1)\leq z(t)$, then

$$g'\left(z(t+1)\right)-g'\left(z(t)\right)=\rho\left(x(t+1)-z(t+1)\right) \tag{53}$$
$$>\rho\left(z(t)-z(t+1)\right)>0 \tag{54}$$

where (53) comes by rearranging (49), and (54) comes by assuming that $z(t+1)\leq z(t)$ and using that $x(t+1)>z(t)$. Hence, we can conclude that $z(t+1)>z(t)$. To see that $z(t+1)<z^\star$, we note that if $z(t+1)\leq x(t+1)$, then we are done since $x(t+1)<z^\star$; otherwise, we have that

$$f'(z)<-(g'(z(t))+\rho(x(t+1)-z))\,,\ \forall z\in]x(t+1),z(t+1)], \tag{55}$$
$$g'(z)<g'(z(t))+\rho(x(t+1)-z)\,,\ \forall z\in[x(t+1),z(t+1)[\tag{56}$$

where (55) comes from using that $\rho>L$, and f' is L-Lipschitz continuous together with the inequalities $z\geq x(t+1)$ and $f'(x(t+1))<-g'(z(t))$ which follows from (48), and (56) comes by that $z(t+1)$ is the unique solution of (49) together with $g'(z(t))<g'(z(t))+\rho(x(t+1)-z(t)$ and $z(t)<z(t+1)$. Summing (55) and (56) and using the continuity of f' and g' shows that $f'(z)+g'(z)<0$ for all $z\in[x(t+1),z(t+1)]$, implying that $z(t+1)<z^\star$.

c) Follows from symmetric arguments as those used for showing b) and is thus omitted. ■

Informally, Corollary 2 shows that under Assumption 10 and $\rho>L$, the ADMM converges to the closest stationary point of $z(0)$ in the direction where $f+g$ is decreasing. For example, when $f(x)=\cos(x)$, $g(z)=\sin(z)$, and $\rho>1$, then $\lim_{t\to\infty}(x(t),z(t),y(t))=(z^\star,z^\star,\cos(z^\star))$, where $z^\star=z(0)$ if $z(0)\in\{2\pi n+\pi/4|n\in\mathbb{Z}\}$ and $z^*=2n\pi+5\pi/4$ if $z(0)\in]2n\pi+\pi/4,2(n+1)\pi+\pi/4[$ for $n\in\mathbb{Z}$. If there is no stationary point in the direction where $f+g$ is decreasing, then the ADMM diverges to $\pm\infty$, for example, when $f(x)=g(x)=-x^2$ and $\rho>2$, then $z^\star=0,-\infty,\infty$ for $z(0)=0$, $z(0)<0$, and $z(0)>0$, respectively.

The challenge in multidimensional cases is that we need to know the direction toward the stationary point. Such a direction is easily obtained in the monodimensional, as suppose to the multidimensional case.

The next section demonstrates the potential of the proposed ADLM approaches (see Sections III and IV) in a problem of great practical relevance.

V. Application: Cooperative Localization in Wireless-Sensor Networks

In this section, we use the ADLM methods to design distributed algorithms for cooperative localization (CL) [2] in wireless-sensor networks.

Consider a undirected graph $(\mathcal{N},\mathcal{E})$, where $\mathcal{N}=\{1,\ldots,N\}$ is a set of nodes embedded in $\mathbb{R}^2$, and $\mathcal{E}\subseteq\mathcal{N}\times\mathcal{N}$ is a set of edges. Let $\mathcal{N}=\mathcal{S}\cup\mathcal{A}$, where $\mathcal{S}=\{1,\ldots,S\}$ is the set of sensors with unknown locations and $\mathcal{A}=\{S+1,\ldots,N\}$ is the set of anchors with known locations. We denote the location of node $n\in\mathcal{A}$ by $\mathbf{a}_n$ and an estimate of the location of node $n\in\mathcal{S}$ by $\mathbf{z}_n$.

Suppose the measurements of the squared[2] distance between two nodes $n,m\in\mathcal{N}$, denoted by $d^2_{n,m}$, are available if and only if $(n,m)\in\mathcal{E}$. The additive measurement errors are assumed to be independent and Gaussian distributed with zeros mean and variance σ^2. Then, the CL problem consists of finding the maximum-likelihood estimate of $(\mathbf{z}_n)_{n\in\mathcal{S}}$ by solving the following problem:

$$\underset{\mathbf{z}_1,\ldots,\mathbf{z}_S\in\mathbb{R}^2}{\text{minimize}}\quad\sum_{n\in\mathcal{S}}\left(\sum_{m\in\mathcal{S}_n}\left|d^2_{n,m}-\|\mathbf{z}_n-\mathbf{z}_m\|^2\right|^2+2\sum_{m\in\mathcal{A}_n}\left|d^2_{n,m}-\|\mathbf{z}_n-\mathbf{a}_m\|^2\right|^2\right) \tag{57}$$

where $\mathcal{S}_n=\{m\in\mathcal{S}|(n,m)\in\mathcal{E}\}$, $\mathcal{A}_n=\{m\in\mathcal{A}|(n,m)\in\mathcal{E}\}$ and the coefficient 2 in front of the second term of the sum comes from that $n\in S$ appears twice in the sum. Note that Problem (57) is NP-hard [34].

[2]Using the square ensures that the objective function of (57) is a continuously differentiable (compare with Assumption 3).

To enable distributed implementation (among the nodes) of the proposed ADLM approaches, let us first equivalently reformulate problem (57) into a general consensus form [16, Sec. 7.2]. We start by introducing at each node $n \in \mathcal{N}$, a local copy $\mathbf{x}_n$ of $(\mathbf{z}_m)_{m\in\bar{\mathcal{S}}_n}$, where $\bar{\mathcal{S}}_n = \mathcal{S}_n \cup \{n\}$. More specifically, we let $\mathbf{x}_n = (\mathbf{x}_{n,m})_{m\in\bar{\mathcal{S}}_n}$, where $\mathbf{x}_{n,m} \in \mathbb{R}^2$ denotes the local copy of $\mathbf{z}_m$ at node n. To formally express the consistency between $\mathbf{x}_n$ and $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_S)$, we introduce the matrix $\mathbf{E}_n \in \mathbb{R}^{2|\bar{\mathcal{S}}_n|\times 2S}$, which is a $|\bar{\mathcal{S}}_n| \times S$ block matrix of 2×2 blocks. In particular, the ith, jth block of $\mathbf{E}_n$ is given by $(\mathbf{E}_n)_{i,j} = \mathbf{I}_2$, if $\mathbf{x}_{n,j}$ is the ith block of the vector $\mathbf{x}_n$ and $(\mathbf{E}_n)_{i,j} = \mathbf{0}$ otherwise. Then, Problem (57) is equivalently given by

$$\begin{aligned} \underset{\mathbf{x},\mathbf{z}}{\text{minimize}} \quad & \sum_{n\in\mathcal{N}} f_n(\mathbf{x}_n), \\ \text{subject to} \quad & \mathbf{x}_n = \mathbf{E}_n\mathbf{z}, \text{ for all } n \in \mathcal{N} \end{aligned} \tag{58}$$

where $\mathbf{x} = (\mathbf{x}_1 \cdots, \mathbf{x}_N) \in \mathbb{R}^{\sum_{n\in\mathcal{N}} 2|\bar{\mathcal{S}}_n|}$, $\mathbf{z} \in \mathbb{R}^{2S}$, and

$$f_n(\mathbf{x}_n) = \begin{cases} \sum\limits_{m\in\mathcal{S}_n} \left|d_{n,m}^2 - \|\mathbf{x}_{n,n} - \mathbf{x}_{n,m}\|^2\right|^2 \\ \quad + \sum\limits_{m\in\mathcal{A}_n} \left|d_{n,m}^2 - \|\mathbf{x}_{n,m} - \mathbf{a}_m\|^2\right|^2, & \text{if } n \in \mathcal{S} \\ \sum\limits_{m\in\mathcal{S}_n} \left|d_{n,m}^2 - \|\mathbf{x}_{n,m} - \mathbf{a}_n\|^2\right|^2, & \text{if } n \in \mathcal{A}. \end{cases}$$

Problem (58) fits the form of Problem (2) and the proposed ADLM approaches can readily be applied. The augmented Lagrangian of problem (58) can be written as

$$L_\rho(\mathbf{x}, \mathbf{z}, \mathbf{y}) = \sum_{n\in\mathcal{N}} f_n(\mathbf{x}_n) + \mathbf{y}_n^{\mathrm{T}}(\mathbf{x}_n - \mathbf{E}_n\mathbf{z}) + \frac{\rho}{2}\|\mathbf{x}_n - \mathbf{E}_n\mathbf{z}\|^2$$

where $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ is the Lagrangian multiplier. Note that the variables $\mathbf{x}$ and $\mathbf{y}$ are separable among $n \in \mathcal{N}$. The resulting distributed-ADLM is as follows.

Algorithm 3: DISTRIBUTED ALTERNATING DIRECTION LAGRANGIAN METHOD (D-ADLM)

1) **Initialization**: Set $t = 0$ and put initial values to $\mathbf{z}(t)$, $\mathbf{y}(t)$, and $\rho(t)$.
2) **Subproblem**: Each node $n \in \mathcal{N}$ solves

$$\mathbf{x}_n(t+1) = \underset{\mathbf{x}_n\in\mathbb{R}^{|\mathcal{S}_n|}}{\arg\min}\, L_{\rho(t)}\left(\mathbf{x}_n, \mathbf{z}(t), \mathbf{y}(t)\right). \tag{59}$$

3) **Communication/Averaging**: $\mathbf{z}(t+1)$ is obtained by solving $\arg\min_{\mathbf{z}} L_{\rho(t)}(\mathbf{x}_n(t+1), \mathbf{z}, \mathbf{y}(t))$, that is

$$\mathbf{z}_n(t+1) = \frac{1}{|\mathcal{S}_n|}\sum_{i\in\mathcal{S}_n} \mathbf{E}_{i,n}^{\mathrm{T}}\left(\mathbf{x}_i(t+1) + \frac{\mathbf{y}_i(t)}{\rho(t)}\right) \tag{60}$$

for $n \in \mathcal{S}$, where $\mathbf{E}_{i,n}$ is the column n of the block matrix $\mathbf{E}_i$.
4) **Local parameter update**: Each node $n \in \mathcal{N}$ updates its local parameters $\rho(t)$ and $\mathbf{y}(t)$ accordingly.
5) **Stopping criterion**: If stopping criterion is met terminate, otherwise set $t = t + 1$ and go to step 2.

Note that the D-ADLM can be carried out either as an ADPM or as an ADMM by performing $\rho(t)$ and $\mathbf{y}(t)$ updates at step 4 accordingly (compare with step 4 of the ADPM and ADMM algorithms). In particular, in ADPM, all of the nodes know the value of $\rho(t)$ for each t and the nodes can update $\mathbf{y}_n(t)$, for all $n \in \mathcal{N}$, as they wish, as long as the sequence $\mathbf{y}_n(t)$ is bounded. In ADMM, all of the nodes n know the value of ρ and update y_n according to

$$\mathbf{y}_n(t+1) = \mathbf{y}_n(t) + \rho(t)\left(\mathbf{x}_n(t+1) - \mathbf{E}_n\mathbf{z}(t+1)\right). \tag{61}$$

As indicated in the first step, the initial setting of the algorithm should be agreed on among the nodes. Other steps can be carried out in a distributed manner with local message exchanges. Note that (60) is simply the average of the local copies of $\mathbf{z}_n$ and the corresponding dual variables [scaled by $\rho(t)$], which can be performed by employing standard gossiping algorithms, for example, [35]. Moreover, the last step requires a mechanism to terminate the algorithm. A natural stopping criterion is to fix the number of iterations, which requires no coordination among the nodes except at the beginning. In order to control the accuracy level ϵ of the coupling constraints, one can, for example, terminate the algorithm when $\max_{n\in\mathcal{N}} \|\mathbf{x}_n(t) - \mathbf{E}_n\mathbf{z}(t)\| < \epsilon$. This can be accomplished with an additional coordination among the nodes.

We compare D-ADLM with the following distributed gradient descent algorithm.

Algorithm 4: DISTRIBUTED GRADIENT DESCENT (D-GD)

1) **Initialization**: Set $t = 0$ and initialize $\rho(t)$, $\mathbf{z}(t)$, and $\bar{x}_n(t) = \mathbf{E}_n\mathbf{z}(t)$ for all $n \in \mathcal{N}$.
2) **Subproblem**: Each node $n \in \mathcal{N}$ solves

$$\mathbf{x}_n(t+1) = \bar{\mathbf{x}}_n(t) - \frac{1}{\rho(t)}\nabla f_n\left(\bar{\mathbf{x}}_n(t)\right). \tag{62}$$

3) **Communication/Averaging**: Each sensor $n \in \mathcal{S}$ finds the average estimation of its localization by communicating with neighbors

$$\mathbf{z}_n(t+1) = \frac{1}{|\mathcal{S}_n|}\sum_{i\in\mathcal{S}_n} \mathbf{E}_{i,n}^{\mathrm{T}}\mathbf{x}_i(t+1). \tag{63}$$

Here, $\mathbf{E}_{i,n}$ is the column n of the block matrix $\mathbf{E}_i$. Set $\bar{\mathbf{x}}_n(t+1) = \mathbf{E}_n\mathbf{z}(t+1)$, that is, the average of the components pertaining to $n \in \mathcal{S}$.
4) **Local parameter update**: Each node $n \in \mathcal{N}$ updates $\rho(t)$.
5) **Stopping criterion**: If stopping criterion is met terminate, otherwise set $t = t + 1$ and go to step 2.

Note that D-GD performs almost the same steps as D-ADLM. The main difference is in step 2): (59) in ADLM is a solution to an optimization problem while (62) in D-GD is a gradient descent step. In particular, the required communication is the same for both algorithms. Therefore, D-GD provides a fair comparison to the D-ADLM.

Let us next test the D-ADLM on a CL problem.

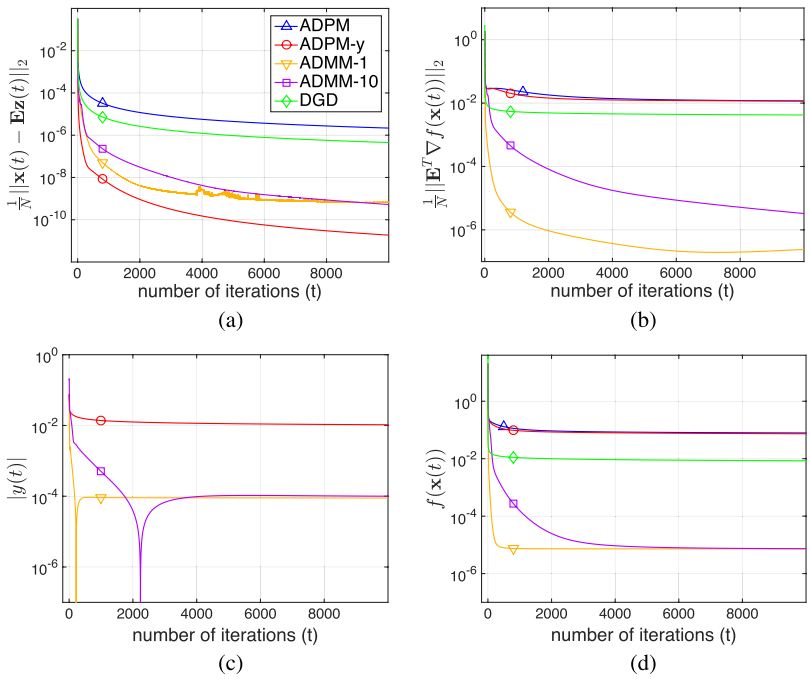

Fig. 2. Results of running all 5 algorithms on the test network. (a) Residuals. (b) Gradient of the objective function. (c) Dual variables. (d) Objective function.

A. *Numerical Results*

We consider a network with $S = 10$, $A = 4$. The 4 anchors are located at (0, 0), (0, 1), (1, 0), and (1, 1). The sensors' are positioned at uniform random in [0, 1] $\times$ [0, 1]. There is an edge between two nodes $n, m \in \mathcal{N}$ if and only if the Euclidean distance between those is less than 0.5. We have $\sigma^2 = 0.05D$, where D is the average squared distance between distinct nodes $(n, m) \in \mathcal{E}$. We consider the following algorithm settings:

name	type	dual update	ρ
ADPM	ADPM	None	$\rho(t) = t$
ADPM-y	ADPM	(61)	$\rho(t) = t$
ADMM-1	ADMM	(61)	$\rho = 1$
ADMM-10	ADMM	(61)	$\rho = 10$
DGD	D-GD	None	$\rho(t) = t$

where the first column identifies each setting, the second column indicates the algorithm used, the third column indicates whether the dual variable update is used or if no dual variable update is used, that is, $\mathbf{y}(t) = \mathbf{0}$, and the fourth column indicates the penalty/steps size used. We initialize the algorithms as $\mathbf{z}_n(0) = (0.5, 0.5)$ for all $n \in \mathcal{S}$. When the dual variable is updated, we initialize it as $\mathbf{y}(0) = \mathbf{0}$.

Fig. 2 depicts the results, where we have compactly written $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ and $\mathbf{E} = (\mathbf{E}_1, \ldots, \mathbf{E}_n)$. Fig. 2(a) and (b) depicts scaled versions of $\|\mathbf{x}(t) - \mathbf{E}\mathbf{z}(t)\|$, the network consensus, and $\|\mathbf{E}^{\mathrm{T}}\nabla f(\mathbf{x}(t))\|$, the gradient of the objective function, respectively, as a function of iterations t. Together, $\|\mathbf{x}(t) - \mathbf{E}\mathbf{z}(t)\|$ and $\|\mathbf{E}^{\mathrm{T}}\nabla f(\mathbf{x}(t))\|$ comprise the FON conditions of Problem (58), that is, when both quantities converge to zero, the FON conditions is asymptotically reached. Both Fig. 2(a) and (b) demonstrates a decreasing trend for all algorithms. In Fig. 2(b), DGD and ADPM have a noticeably slower decay rate than ADPM-y, ADMM-1, and ADMM-10. In Fig. 2(b), DGD, ADPM, and ADPM-y have a noticeably slower decay rate than ADMM-1 and ADMM-10. Therefore, the results suggest that it can be beneficial to use the update (61). The results in Fig. 2(a) and (b) suggest faster convergence of ADMM than DGD and ADPM to the first-order necessary conditions.

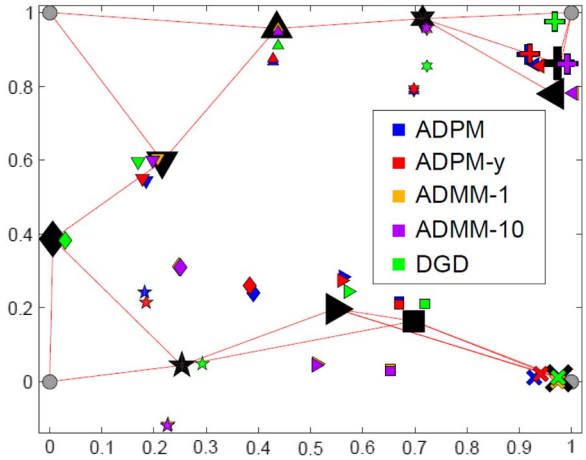

Fig. 3. Position estimate that each algorithm converges to.

Fig. 2(c) depicts an example of a dual variable for each algorithm where the update (61) is used. Similar results were observed for the other dual variables. The figure shows that the dual variables converge, implying that ADMM-1 and ADMM-10 converge based on Proposition 3.

Fig. 2(a) depicts the objective value at each iteration. The algorithms achieve different objective values, which is not surprising since the objective function is nonconvex with multiple local minima. Fig. 3 depicts the resulting location estimations for each algorithm, that is, the estimation at the final iteration. Note that the orange diamond and five-pointed star lay under their purple counterparts and are therefore not visible in the figure. Despite the nonconvexities, all of the algorithms converge to good estimations close to the true locations of the nodes. The DGD achieves a visually better estimation of the diamond

and the five-pointed star in Fig. 3 than the other algorithms. Nevertheless, ADMM-1 and ADMM-10 achieve much better objective function values.

Remark 2: The gradients of f_n for $n \in \mathcal{N}$ are unbounded, but still Assumption (2).b holds, which ensures that the sequence $(\mathbf{x}(t), \mathbf{z}(t))$ of ADPM is bounded, see Lemma 1. Similar results can be derived for ADPM-y, ADMM-1, and ADMM-10 as long as the dual variables $\mathbf{y}(t)$ are bounded. On the other hand, from our numerical experiences, DGD turned out to be unstable for many initializations where it reached floating-point infinity in only a few iterations.

VI. Conclusion

We investigated the convergence behavior of scalable variants of two standard nonconvex optimization methods: a novel method we call ADPM and the well-known ADMMs, variants of the Quadratic Penalty Method and the Method of Multipliers, respectively. Our theoretical results showed that the DPM asymptotically reaches primal feasibility under assumptions that hold widely in practice and provided sufficient conditions for when ADPM asymptotically reaches the first-order necessary conditions for optimality. Furthermore, we provided sufficient conditions for the asymptotic convergence of ADMM to the first-order necessary condition for local optimality and provided a class of problems where these conditions hold. Finally, we demonstrated how the methods can be used to design distributed algorithms for *nonconvex* cooperative localization in wireless-sensor networks.

References

[1] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, Aug. 2009.

[2] N. Patwari, J. Ash, S. Kyperountas, A. Hero, R. Moses, and N. Correal, "Locating the nodes: Cooperative localization in wireless sensor networks," *IEEE Signal Process. Mag.*, vol. 22, no. 4, pp. 54–69, Jul. 2005.

[3] S. Frank, I. Steponavice, and S. Rebennack, "Optimal power flow: A bibliographic survey I," *Energy Syst.*, vol. 3, no. 3, pp. 221–258, 2012.

[4] S. Frank, I. Steponavice, and S. Rebennack, "Optimal power flow: A bibliographic survey II," *Energy Syst.*, vol. 3, no. 3, pp. 259–289, 2012.

[5] L. Xishuo, S. Draper, and B. Recht, "Suppressing pseudocodewords by penalizing the objective of lp decoding," in *Proc. IEEE Inf. Theory Workshop*, Sep. 2012, pp. 367–371.

[6] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Trans. Autom. Control*, vol. 54, no. 1, pp. 48–61, Jan. 2009.

[7] A. Nedic, A. Ozdaglar, and P. Parrilo, "Constrained consensus and optimization in multi-agent networks," *IEEE Trans. Autom. Control*, vol. 55, no. 4, pp. 922–938, Apr. 2010.

[8] M. Zhu and S. Martinez, "On distributed convex optimization under inequality and equality constraints," *IEEE Trans. Autom. Control*, vol. 57, no. 1, pp. 151–164, Jan. 2012.

[9] J. Duchi, A. Agarwal, and M. Wainwright, "Dual averaging for distributed optimization: Convergence analysis and network scaling," *IEEE Trans. Autom. Control*, vol. 57, no. 3, pp. 592–606, Mar. 2012.

[10] D. Jakovetic, J. Xavier, and J. Moura, "Fast distributed gradient methods," *IEEE Trans. Autom. Control*, vol. 59, no. 5, pp. 1131–1146, May 2014.

[11] W. Shi, Q. Ling, G. Wu, and W. Yin, "EXTRA: An exact first-order algorithm for decentralized consensus optimization," *ArXiv e-prints*, Apr. 2014.

[12] M. Zhu and S. Martinez, "An approximate dual subgradient algorithm for multi-agent non-convex optimization," *IEEE Trans. Autom. Control*, vol. 58, no. 6, pp. 1534–1539, Jun. 2013.

[13] P. Bianchi and J. Jakubowicz, "Convergence of a multi-agent projected stochastic gradient algorithm for non-convex optimization," *IEEE Trans. Autom. Control*, vol. 58, no. 2, pp. 391–405, Feb. 2013.

[14] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite element approximation," *Comput. Math. Appl.*, vol. 2, no. 1, pp. 17–40, 1976.

[15] J. Eckstein and D. Bertsekas, "On the Douglas Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Math. Program.*, vol. 55, no. 1-3, pp. 293–318, 1992.

[16] 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. [Online]. Available: http://dx.doi.org/10.1561/2200000016

[17] D. P. Bertsekas, *Constrained Optimization and Lagrange Multipler Methods*. Belmont, MA, USA: Athena Scientific, 1982.

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

[19] Q. Ling and A. Ribeiro, "Decentralized dynamic optimization through the alternating direction method of multipliers," *IEEE Trans. Signal Process.*, vol. 62, no. 5, pp. 1185–1197, Mar. 2014.

[20] G. Mateos, J. Bazerque, and G. Giannakis, "Distributed sparse linear regression," *IEEE Trans. Signal Process.*, vol. 58, no. 10, pp. 5262–5276, Oct. 2010.

[21] I. Schizas, A. Ribeiro, and G. Giannakis, "Consensus in ad hoc WSNs with noisy links-part I: Distributed estimation of deterministic signals," *IEEE Trans. Signal Process.*, vol. 56, no. 1, pp. 350–364, Jan. 2008.

[22] E. Wei and A. Ozdaglar, "On the O(1/k) convergence of asynchronous distributed alternating direction method of multipliers," *ArXiv e-prints*, Jul. 2013.

[23] W. Shi, Q. Ling, K. Yuan, G. Wu, and W. Yin, "On the linear convergence of the ADMM in decentralized consensus optimization," *IEEE Trans. Signal Process.*, vol. 62, no. 7, pp. 1750–1761, Apr. 2014.

[24] S. Magnússon, P. C. Weeraddana, and C. Fischione, "A distributed approach for the optimal power flow problem based on ADMM and sequential convex approximations," Cornell Univ. Library, Ithaca, NY, USA, 2014. [Online]. Available: http://arxiv.org/abs/1401.4621

[25] F. Iutzeler, P. Bianchi, P. Ciblat, and W. Hachem, "Explicit convergence rate of a distributed alternating direction method of multipliers," *ArXiv e-prints*, Dec. 2013.

[26] E. Ghadimi, A. Teixeira, I. Shames, and M. Johansson, "Optimal parameter selection for the alternating direction method of multipliers (admm): Quadratic problems," *IEEE Trans. Autom. Control*, vol. 60, no. 3, pp. 644–658, Mar. 2015.

[27] M. Hong and Z.-Q. Luo, "On the linear convergence of the alternating direction method of multipliers,"*ArXiv e-prints*, Aug. 2012.

[28] T. Kanamori and A. Takeda, "Numerical study of learning algorithms on Stiefel manifold," *Comput. Manage. Sci.*, pp. 1–22, 2013.

[29] D. Liu, T. Zhou, H. Qian, C. Xu, and Z. Zhang, "A nearly unbiased matrix completion approach," in *Machine Learning and Knowledge Discovery in Databases*, vol. 8189. New York, NY, USA: Springer-Verlag, 2013, ser. Lecture Notes in Computer Science, pp. 210–225.

[30] X. Liu, S. Draper, and B. Recht, "The l1 penalized decoder and its reweighted lp," in *Proc. 50th Annu. Allerton Conf. Commun., Control, Comput.*, Oct. 2012, pp. 1572–1579.

[31] R. Chartrand, "Nonconvex splitting for regularized low-rank + sparse decomposition," *IEEE Trans. Signal Process.*, vol. 60, no. 11, pp. 5810–5819, Nov. 2012.

[32] S. Magnússon, P. C. Weeraddana, M. G. Rabbat, and C. Fischione, "On the convergence of an alternating direction penalty method for nonconvex problems," in *Proc. IEEE Asilomar Conf. Signals, Syst., Comput.*, Nov. 2014, pp. 793–797.

[33] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[34] J. Aspnes *et al.*, "A theory of network localization," *IEEE Trans. Mobile Comput.*, vol. 5, no. 12, pp. 1663–1678, Dec. 2006.

[35] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Randomized gossip algorithms," *IEEE Trans. Inf. Theory*, vol. 52, no. 6, pp. 2508–2530, Jun. 2006.

Sindri Magnússon received the B.Sc. degree in mathematics from the University of Iceland, Reykjavik, Iceland, in 2011, the M.Sc. degree in mathematics from KTH Royal Institute of Technology, Stockholm Sweden, in 2013, and is currently pursuing the Ph.D. degree in aomatic control at the School of Electrical Engineering and ACCESS Linnaeus Center, KTH Royal Institute of Technology.

He has held a six-month visiting research position at Harvard University, Cambridge, MA, USA, in 2015, sponsored by the Engblom Foundation. His research interests include distributed optimization, both theory and applications.

Pradeep Chathuranga Weeraddana (S'08–M'11) received the M.Eng degree in telecommunications from the School of Engineering and Technology, Asian Institute of Technology, Khlong Luang, Thailand, in 2007 and the Ph.D. degree telecommunications engineering from the University of Oulu, Oulu, Finland, in 2011.

He was a Postdoctoral Researcher in the Department of Automatic Control, School of Electrical Engineering and ACCESS Linnaeus Center, KTH Royal Institute of Technology, Stockholm, Sweden. Currently, he is a Senior Lecturer at the Sri Lankan Institute of Information Technology, Malabe, Sri Lanka. His research interests include the application of optimization techniques in various application domains, such as signal processing, wireless communications, smart grids, privacy, and security.

Michael G. Rabbat (S'02–M'07–SM'15) received the B.Sc. degree in electrical engineering from the University of Illinois, Urbana-Champaign, IL, USA, in 2001, the M.Sc. degree in electrical engineering from Rice University, Houston, TX, USA, in 2003, and the Ph.D. degree in electrical engineering from the University of Wisconsin, Madison, WI, USA, in 2006.

He joined McGill University, Montréal, QC, Canada, in 2007, where he is currently an Associate Professor. During the 2013–2014 academic year, he held visiting positions at Télécom Bretegne, Brest, France; the Inria Bretagne-Atlantique Reserch Centre, Rennes, France; and KTH Royal Institute of Technology, Stockholm, Sweden. He was a Visiting Researcher at Applied Signal Technology, Inc., Sunnyvale, CA, USA, during the summer of 2003. Currently, he is a Senior Area Editor for the IEEE SIGNAL PROCESSING LETTERS and an Associate Editor for IEEE TRANSACTIONS ON SIGNAL AND INFORMATION PROCESSING OVER NETWORKS and IEEE TRANSACTIONS ON CONTROL OF NETWORK SYSTEMS. His research interests include distributed algorithms for optimization and inference, consensus algorithms, and network modelling and analysis, with applications in distributed sensor systems, large-scale machine learning, statistical signal processing, and social networks.

Dr. Rabbat co-authored the paper which received the Best Paper Award (Signal Processing and Information Theory Track) at the 2010 IEEE International Conference on Distributed Computing in Sensor Systems (DCOSS). He received an Honorable Mention for Outstanding Student Paper Award at the 2006 Conference on Neural Information Processing Systems (NIPS) and a Best Student Paper Award at the 2004 ACM/IEEE International Symposium on Information Processing in Sensor Networks (IPSN). He received the 2014 McGill University Principal's Prize for Excellence in Teaching at the level of assistant professor.

Carlo Fischione (M'05) received the Ph.D. degree in electrical and information engineering and the Dr.Eng. degree in electronic engineering (Hons.) from the University of L'Aquila, L'Aquila, Italy, in 2001 and 2005, respectively.

Currently, he is a tenured Associate Professor at KTH Royal Institute of Technology, Electrical Engineering and ACCESS Linnaeus Center, Automatic Control Lab, Stockholm, Sweden. He has held research positions at the Massachusetts Institute of Technology, Cambridge, MA, USA (2015, Visiting Professor); University of California at Berkeley, Berkeley, CA, USA (2004–2005, Visiting Scholar, and 2007–2008, Research Associate); and the Royal Institute of Technology, Stockholm, Sweden (2005–2007, Research Associate). He has co-authored more than 100 publications, including book, book chapters, international journals and conferences, and international patents. His current research interests include optimization and parallel computation with applications to wireless-sensor networks, cyberphysical systems, and wireless networks.

Prof. Fischione received a number of awards, including the best paper award from the IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS of 2007; the best paper awards at the IEEE International Conference on Mobile Ad-hoc and Sensor System 05 and 09 (IEEE MASS 2005 and IEEE MASS 2009); the Best Paper Award of the IEEE Sweden VT-COM-IT Chapter of 2014; the Best Business Idea award from VentureCup East Sweden in 2010; the Ferdinando Filauro award from University of LAquila, Italy, in 2003; the Higher Education award from Abruzzo Region Government, Italy, in 2004; the Junior Research award from Swedish Research Council in 2007; and the Silver Ear of Wheat award in history from the Municipality of Tornimparte, Italy, in 2012. He has chaired or served as a technical member of program committees of several international conferences and serves as a referee for technical journals. Meanwhile, he also has offered his advice as a Consultant to numerous technology companies, such as Berkeley Wireless Sensor Network Lab, Ericsson Research, Synopsys, and United Technology Research Center. He is co-founder and CTO of the sensor networks start-up companies Lokkupp (indoor navigation) and Modern Ancient Instruments Networked (MIND). He is an Ordinary Member of the academy of history Deputazione Abruzzese di Storia Patria (DASP).