Mathematical Programming (2019) 176:207–245
https://doi.org/10.1007/s10107-019-01365-4

FULL LENGTH PAPER

Series B

Perturbed proximal primal–dual algorithm for nonconvex nonsmooth optimization

Davood Hajinezhad[1] · Mingyi Hong[2]

Received: 2 November 2017 / Accepted: 17 January 2019 / Published online: 20 February 2019

Abstract

In this paper, we propose a perturbed proximal primal–dual algorithm (PProx-PDA) for an important class of linearly constrained optimization problems, whose objective is the sum of smooth (possibly nonconvex) and convex (possibly nonsmooth) functions. This family of problems can be used to model many statistical and engineering applications, such as high-dimensional subspace estimation and the distributed machine learning. The proposed method is of the Uzawa type, in which a primal gradient descent step is performed followed by an (approximate) dual gradient ascent step. One distinctive feature of the proposed algorithm is that the primal and dual steps are both perturbed appropriately using past iterates so that a number of asymptotic convergence and rate of convergence results (to first-order stationary solutions) can be obtained. Finally, we conduct extensive numerical experiments to validate the effectiveness of the proposed algorithm.

Keywords Nonconvex optimization · Nonsmooth optimization · Primal–dual algorithms

Mathematics Subject Classification 49 · 90

This work was completed when Davood Hajinezhad was a Ph.D. student at Iowa State University. Mingyi Hong is supported by NSF Grants CMMI-1727757, and an AFOSR Grant 15RT0767.

✉ Mingyi Hong
mhong@umn.edu

Davood Hajinezhad
davood.hajinezhad@sas.com

[1] SAS Institute, Cary, NC, USA

[2] Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, USA

1 Introduction

1.1 The problem and the proposed algorithm

Consider the following optimization problem

$$\min_{x\in X} \; f(x)+h(x), \quad \text{s.t.} \quad Ax=b, \tag{1}$$

where $f(x):\mathbb{R}^N\to\mathbb{R}$ is a continuous smooth function (possibly nonconvex); $A\in\mathbb{R}^{M\times N}$ is a rank deficient matrix; $b\in\mathbb{R}^M$ is a given vector; $X\subset\mathbb{R}^N$ is a convex compact set; $h(x):\mathbb{R}^N\to\mathbb{R}$ is a lower semi-continuous nonsmooth convex function. Problem (1) is an interesting class that can be specialized to a number of statistical and engineering applications. We provide a few of these applications in Sect. 1.2.

To develop an efficient algorithm for problem (1), let us first construct its augmented Lagrangian as below

$$L_\rho(x,\lambda)=f(x)+h(x)+\langle\lambda, Ax-b\rangle+\frac{\rho}{2}\|Ax-b\|^2, \tag{2}$$

where $\lambda\in\mathbb{R}^M$ is the dual variable associated with the equality constraint $Ax=b$, and $\rho>0$ is the penalty parameter for the augmented term $\|Ax-b\|^2$.

Define $B\in\mathbb{R}^{M\times N}$ as a scaling matrix, and introduce two new parameters $\gamma\in(0,1)$ and $\beta>0$, where γ is a small positive parameter related to the size of the tolerable equality constraint violation; β is the proximal parameter that regularizes the primal update. Let us choose $\gamma>0$ and $\rho>0$ such that $\rho\gamma<1$. The steps of the proposed perturbed proximal primal-dual algorithm (PProx-PDA) are given below (Algorithm 1).

Algorithm 1: The perturbed proximal primal–dual algorithm (PProx-PDA)

```
Initialize: $\lambda^0$ and $x^0$
Repeat: update variables by
    $x^{r+1} = \arg\min_{x\in X} \Big\{ \langle\nabla f(x^r), x-x^r\rangle + h(x) + \langle(1-\rho\gamma)\lambda^r, Ax-b\rangle$
        $+ \frac{\rho}{2}\|Ax-b\|^2 + \frac{\beta}{2}\|x-x^r\|^2_{B^TB} \Big\}$    (3a)
    $\lambda^{r+1} = (1-\rho\gamma)\lambda^r + \rho\left(Ax^{r+1}-b\right)$
        $= \lambda^r + \rho\left(Ax^{r+1}-b-\gamma\lambda^r\right).$    (3b)
Until Convergence.
```

In contrast to the classical Augmented Lagrangian (AL) method [37,59], in which the primal variable is updated by minimizing the augmented Lagrangian given in (2), in

PProx-PDA the primal step minimizes an approximated augmented Lagrangian, where the approximation comes from: (1) replacing function $f(x)$ with the surrogate function $\langle \nabla f(x^r), x - x^r \rangle$; (2) perturbing dual variable λ by a positive factor $1 - \rho\gamma > 0$; (3) adding proximal term $\frac{\beta}{2}\|x - x^r\|^2_{B^TB}$. We make a few remarks about these algorithmic choices.

First, the use of the linear surrogate function $\langle \nabla f(x^r), x - x^r \rangle$ ensures that only first-order information is used for the primal update. Also it is worth mentioning that one can replace the function $\langle \nabla f(x^r), x - x^r \rangle$ with a wider class of "surrogate" functions satisfying certain gradient consistency conditions [60,64], and our subsequent analysis will still hold true. However, in order to stay focused, we choose not to present those variations.

Second, the primal and dual perturbations are added to facilitate convergence analysis. In particular, the analysis for the PProx-PDA algorithm differs from the recent analysis on nonconvex primal/dual type algorithms, which is first presented in Ames and Hong [2] and later generalized by [31,33,35,40,45,53,69]. Those analyses have been critically dependent on bounding the size of the successive dual variables with that of the successive primal variables. Unfortunately, this can only be done when the primal step immediately preceding the dual step is *smooth* and *unconstrained*. Therefore the analysis presented in these works cannot be applied to our general formulation with nonsmooth terms and constraints.

Our perturbation scheme is strongly motivated by the dual perturbation scheme developed for the *convex* problems, for example in [43]. Conceptually, the perturbed dual step can be viewed as performing a dual ascent step on certain *regularized Lagrangian* in the dual space; see [43, Sec.3.1]. As pointed out in this reference, and in many related works, the main benefit for introducing the dual perturbation/regularization, is to ensure that the dual update is well-behaved and easy to analyze. One of the main contributions of this work is to develop a similar but slightly more refined perturbation technique, so that first-order primal–dual methods can be applied to a much wider class of problems, as compared to those that can be handled in existing works reviewed above [2,31,33,35,40,45,53,69].

Third, the proximal term $\frac{\beta}{2}\|x - x^r\|^2_{B^TB}$ is used for two purposes: (1) to make the primal subproblem strongly convex; (2) for certain applications to ensure that the primal subproblem is decomposable over the variables. We will discuss how this can be done in the subsequent sections.

1.2 Motivating applications

Sparse subspace estimation Suppose that $\Sigma \in \mathbb{R}^{p\times p}$ is an unknown covariance matrix, $\lambda_1 \geq \lambda_2 \geq \cdots \lambda_p$ and $u_1, u_2, \ldots, u_p$ are its eigenvalues and eigenvectors, respectively, and they satisfy $\Sigma = \sum_{i=1}^p \lambda_i u_i u_i^\top$. Principal Component Analysis (PCA) aims to recover $u_1, u_2, \ldots, u_k$, where $k \leq p$, from a sample covariance matrix $\hat{\Sigma}$ obtained from i.i.d samples $\{x_i\}_{i=1}^n$. The subspace spanned by $\{u_i\}_{i=1}^k$ is called k-dimensional principal subspace, whose projection matrix is given by $\Pi^* = \sum_{i=1}^k u_i u_i^\top$. Therefore, PCA reduces to finding an estimate of Π^*, denoted by $\hat{\Pi}$, from the sample covariance matrix $\hat{\Sigma}$. In high dimensional setting where the number of data points is significantly smaller than the dimension i.e. ($n \ll p$), it is desirable to find a *sparse* $\hat{\Pi}$, using the following formulation [29]

$$\min_{\Pi} \left\langle \hat{\Sigma}, \Pi \right\rangle + \mathcal{P}_\nu(\Pi), \qquad \text{s.t.} \quad \Pi \in \mathcal{F}^k, \tag{4}$$

where, $\mathcal{F}^k$ denotes the Fantope set [68], given by $\mathcal{F}^k = \{X : 0 \preceq X \preceq I, \text{trace}(X) = k\}$, which promotes low rankness in X. The function $\mathcal{P}_\nu(\Pi)$ is a nonconvex regularizer that enforces sparsity on Π. Typical forms of this regularization are smoothly clipped absolute deviation (SCAD) [22], and minimax concave penalty (MCP) [73]. For example, MCP with parameters b and ν for some scalar ϕ is given below

$$\mathcal{P}_\nu(\phi) = \iota_{|\phi| \le b\nu} \left(\nu|\phi| - \frac{\phi^2}{2b} \right) + \iota_{|\phi| > b\nu} \left(\frac{b\nu^2}{2} \right), \tag{5}$$

where, ι_X denotes the indicator function for a convex set X, which is defined as

$$\iota_X(y) = 0, \text{ when } y \in X, \quad \iota_X(y) = \infty, \text{ otherwise.} \tag{6}$$

Notice that $\mathcal{P}_\nu(\Pi)$ in problem (4) is an element-wise operator over all entries of matrix Π. One particular characterization for these nonconvex penalties is that they can be decomposed as a sum of an ℓ_1-norm function (i.e. for $x \in \mathbb{R}^N$, $\|x\|_1 = \sum_{i=1}^N |x_i|$) and a concave function $q_\nu(x)$ as $\mathcal{P}_\nu(\phi) = \nu|\phi| + q_\nu(\phi)$ for some $\nu \ge 0$. In a recent work [29], it is shown that with high probability, every first-order stationary solution of problem (4) (denoted as $\hat{\Pi}$) is of high-quality. See [29, Theorem 3] for detailed description. In order to deal with the Fantope and the nonconvex regularizer separately, one can introduce a new variable Φ and reformulate problem (4) in the following manner [68]

$$\min_{\Pi, \Phi} \left\langle \hat{\Sigma}, \Pi \right\rangle + \mathcal{P}_\nu(\Phi), \text{ s.t. } \Pi \in \mathcal{F}^k, \; \Pi - \Phi = 0. \tag{7}$$

Clearly this is a special case of problem (1), with $x = [\Pi, \Phi]$, $f(x) = \left\langle \hat{\Sigma}, \Pi \right\rangle + q_\nu(\Phi)$, $h(x) = \nu\|\Phi\|_1$, $X = \mathcal{F}^k$, $A = [I, -I]$, $b = 0$.

The exact consensus problem over networks Consider a network which consists of N agents who collectively optimize the following problem

$$\min_{y \in \mathbb{R}} f(y) + h(y) := \sum_{i=1}^{N} \left(f_i(y) + h_i(y) \right), \tag{8}$$

where $f_i(y) : \mathbb{R} \to \mathbb{R}$ is a smooth function, and $h_i(y) : \mathbb{R} \to \mathbb{R}$ is a convex, possibly nonsmooth regularizer (here y is assumed to be scalar for ease of presentation). Note that both f_i and h_i are only accessible by agent i. In particular, each local loss function f_i can represent: (1) a mini-batch of (possibly nonconvex) loss functions modeling data fidelity [4]; (2) nonconvex activation functions of neural networks [1]; (3) nonconvex utility functions used in applications such as resource allocation [13]. The regularization function h_i usually takes the following forms: (1) convex regularizers such as nonsmooth ℓ_1 or smooth ℓ_2 functions; (2) the indicator function for a closed convex set X, i.e. the ι_X function defined in (6). This problem has found applications in various domains such as distributed statistical learning [52], distributed consensus [67], distributed communication networking [46,75], distributed and parallel machine learning [25,38] and distributed signal processing [63,75]; for more applications we refer the readers to a recent survey [27].

To integrate the structure of the network into problem (8), we assume that the agents are connected through a network defined by an undirected, connected graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, with $|\mathcal{V}| = N$ vertices and $|\mathcal{E}| = E$ edges. For agent $i \in \mathcal{V}$ the neighborhood set is defined as $\mathcal{N}_i := \{j \in \mathcal{V} \text{ s.t. } (i,j) \in \mathcal{E}\}$. Each agent can only communicate with its neighbors, and it is responsible for optimizing one component function f_i regularized by h_i. Define the incidence matrix $A \in \mathbb{R}^{E\times N}$ as following: if $e \in \mathcal{E}$ and it connects vertex i and j with $i > j$, then $A_{ev} = 1$ if $v = i$, $A_{ev} = -1$ if $v = j$ and $A_{ev} = 0$ otherwise. Using this definition, the *signed graph Laplacian matrix* L_- is given by $L_- := A^T A \in \mathbb{R}^{N\times N}$. Introducing N new variables x_i as the local copy of the global variable y, and define $x := [x_1; \cdots; x_N] \in \mathbb{R}^N$, problem (8) can be equivalently expressed as

$$\min_{x\in\mathbb{R}^N} \; f(x) + h(x) := \sum_{i=1}^{N} \left(f_i(x_i) + h_i(x_i)\right), \text{ s.t. } Ax = 0. \tag{9}$$

This problem is precisely the original problem (1) with the correspondence: $X = \mathbb{R}^N$, $b = 0$, $f(x) := \sum_{i=1}^N f_i(x_i)$, and $h(x) := \sum_{i=1}^N h_i(x_i)$.

For this problem, let us see how the proposed PProx-PDA can be applied. The first observation is that choosing the scaling matrix B is critical because the appropriate choice of B ensures that problem (3a) is decomposable over different variables (or variable blocks), thus the PProx-PDA algorithm can be performed fully distributedly. Let us define the *signless incidence matrix* $B := |A|$, where A is the signed incidence matrix defined above, and the absolute value is taken for each component of A. Using this choice of B, we have $B^T B = L_+ \in \mathbb{R}^{N\times N}$, which is the signless graph Laplacian whose (i,i)th diagonal entry is the degree of node i, and its (i,j)th entry is 1 if $e = (i,j) \in \mathcal{E}$, and 0 otherwise. Further, let us set $\rho = \beta$. Then x-update step (3a) becomes

$$x^{r+1} = \arg\min_x \left\{ \sum_{i=1}^{N} \langle \nabla f_i(x_i^r), x_i - x_i^r\rangle + \langle (1-\rho\gamma)\lambda^r, Ax\rangle + \rho x^T D x - \rho x^T L_+ x^r \right\},$$

where $D := \mathrm{diag}[d_1, \ldots, d_N] \in \mathbb{R}^{N\times N}$ is the diagonal degree matrix, with d_i denoting the degree of node i. Clearly this problem is separable over the variable x_i for all $i = 1, 2, \ldots, N$. To perform this update, each agent i only requires local information as well as information from its neighbors $\mathcal{N}_i$. This is because D is a diagonal matrix and the structure of matrix L^+ ensures that the ith block vector of $L^+ x^r$ is only related to x_j^r and x_i^r, where $j \in \mathcal{N}_i$.

The partial consensus problem In the previous application, the agents are required to reach *exact* consensus, and such constraint is imposed through $Ax = 0$ in (9). In practice, however, consensus is rarely required exactly, for example due to potential disturbances in network communication; see detailed discussion in [42]. Further, in applications ranging from distributed estimation to rare event detection, the data obtained by the agents, such as harmful algal blooms, network activities, and local temperature, often exhibit distinctive spatial structure [17]. The distributed problem in these settings can be best formulated by using certain partial consensus model in which the local variables of an agent are only required to be close to those of its neighbors. To model such a *partial* consensus constraint, we denote ξ as the permissible tolerance for $e = (i, j) \in \mathcal{E}$, and define the link variable $z_e = x_i - x_j$. Then we replace the strict consensus constraint $z_e = 0$ with $-\xi \leq [z_e]_k \leq \xi$, where $[z_e]_k$ denotes the kth entry of vector z_e (for the sake of simplicity we assume that the permissible tolerance ξ is identical for all $e \in \mathcal{E}$). Setting

$$z := \{z_e\}_{e\in\mathcal{E}} \text{ and } Z := \{z;\ \ |[z_e]_k| \leq \xi,\quad \forall\, e \in \mathcal{E}, \forall\, k\},$$

the partial consensus problem can be formulated as

$$\min_{x,z}\ \sum_{i=1}^{N} \left(f_i(x_i) + h_i(x_i)\right),\ \text{ s.t. } Ax - z = 0,\ z \in Z, \tag{10}$$

which is again a special case of problem (1).

1.3 Literature review and contribution

1.3.1 Literature on related algorithms

The Augmented Lagrangian (AL) method, also known as the methods of multipliers, is pioneered by Hestenes [37] and Powell [59]. It is a classical algorithm for solving nonconvex smooth constrained problems and its convergence is guaranteed under rather weak assumptions [8,23,58]. A modified version of AL has been developed by Rockafellar in [61], in which a proximal term has been added to the objective function in order to make it strongly convex in each iteration. Later Wright [70] specialized this algorithm to the linear programming problem. Many existing packages such as LANCELOT are implemented based on AL method. Recently, due to the need to solve very large-scale nonlinear optimization problems, the AL and its variants regain their popularity. For example, in [18] a line search AL method has been proposed for solving problem (1) with $h \equiv 0$ and $X = \{x;\ l \leq x \leq u\}$. Also reference [15] has developed an AL-based algorithm for nonconvex nonsmooth optimization, where sub-gradients of the augmented Lagrangian are used in the primal update. When the problem is convex, smooth and the constraints are linear, Lan and Monterio [44] have analyzed the iteration complexity for the AL method. More specifically, the authors analyzed the total number of Nesterov's optimal iterations [57] that are required to reach high-quality primal–dual solutions. Subsequently, Liu et al. [48] proposed an inexact AL (IAL) algorithm which only requires an $\epsilon-$approximated solution for the primal subproblem at each iteration. Hong et al. [38] proposed a proximal primal–dual algorithm (Prox-PDA), an AL-based method mainly used to solve smooth and unconstrained distributed nonconvex problem [by unconstrained we refer to the problem (1) with $h \equiv 0$ and $X \in \mathbb{R}^N$; however, the linear constraint $Ax = b$ is always imposed]. Another AL-based algorithm, which is called ALADIN [41], is designed for nonconvex smooth optimization problem with coupled affine constraints in the distributed setting. In ALADIN the objective function is separable over different nodes and the loss function is assumed to be twice differentiable. To implement ALADIN a fusion center is needed in the network to propagate global variable to the agents. A comprehensive survey about AL-based methods in both convex and nonconvex setting can be found in [36]. See more practical AL algorithms in [12]. Overall, the AL-based methods often require sophisticated stepsize selection and an accurate oracle for solving the primal problem. Further, they cannot deal with problems that have both nonsmooth regularizer $h(x)$ and a general convex constraint. Therefore, it is not straightforward to apply these methods to problems such as distributed learning and high-dimensional sparse subspace estimation mentioned in the previous subsection.

Recently, the alternating direction method of multipliers (ADMM), a variant of the AL, has gained popularity for decomposing large-scale nonsmooth optimization problems [14]. The method originates in early 1970s [26,28], and has since been studied extensively [10,20,39]. The main strength of this algorithm is that it is capable of decomposing a large problem into a series of small and simple subproblems, therefore making the overall algorithm scalable and easy to implement. However, unlike the AL method, the ADMM is designed for convex problems, despite its good numerical performance in nonconvex problems such as the nonnegative matrix factorization [66], phase retrieval [71], distributed clustering [25], tensor decomposition [47] and so on. Only very recently, researchers have begun to rigorously investigate the convergence of ADMM (to first-order stationary solutions) for nonconvex problems. Zhang [74] have analyzed a class of splitting algorithms (which includes the ADMM as a special case) for a very special class of nonconvex quadratic problems. Ames and Hong in [2] have developed an analysis for ADMM for certain ℓ_1 penalized problem arising in the high-dimensional discriminant analysis. Other works along this line include [34,40,45,53] and [69]; See Table 1 in [69] for a comparison of the conditions required for these works. Despite the recent progress, it appears that the aforementioned works still pose very restrictive assumptions on the problem types in order to achieve convergence. For example, it is not clear whether the ADMM can be used for the distributed nonconvex optimization problem (9) over an arbitrary connected graph with regularizers and constraints, despite the fact that for a convex problem such application is popular, and the resulting algorithms are efficient.

1.3.2 Literature on applications

The sparse subspace estimation problem formulations (4) and (7) have been first considered in [19,68] and subsequently considered in [29]. The work [68] proposes a semidefinite convex optimization problem to estimate principal subspace of a population matrix Σ based on a sample covariance matrix. The authors of [29] further show that by utilizing nonconvex regularizers it is possible to significantly improve the estimation accuracy for a given number of data points. However, the algorithm considered in [29] is not guaranteed to reach any stationary solutions.

The consensus problem (8) and (9) have been studied extensively in the literature when the objective functions are all convex; see for example [7,49,54,55,65]. Without assuming convexity of f_i's, the literature has been very scant; see recent developments in [11,32,40,50]. However, all of these recent results require that the nonsmooth terms h_i's, if present, have to be identical for all agents in the network. This assumption is unnecessarily strong and it defeats the purpose of *distributed* consensus since *global* information about the objective function has to be shared among the agents. Further, in the nonconvex setting, we are not aware of any existing distributed algorithm with convergence guarantee that can deal with the more practical problem (10) with partial consensus.

1.3.3 Contributions of this work

In this paper, we develop an AL-based algorithm, named the perturbed proximal primal–dual algorithm (PProx-PDA), for the challenging linearly constrained nonconvex nonsmooth problem (1). The proposed method, listed in Algorithm 1, is of the Uzawa type [5] and it has a very simple update rule. It is a *single-loop* algorithm that alternates between a primal (scaled) proximal gradient descent step, and an (approximate) dual gradient ascent step. Further, by appropriately selecting the scaling matrix in the primal step, the variables can be easily updated in parallel. These features make the algorithm attractive for applications such as the high-dimensional subspace estimation and the distributed learning problems discussed in Sect. 1.2.

One distinctive feature of the PProx-PDA is that it incorporates a novel primal–dual perturbation scheme, which is designed to ensure a number of asymptotic convergence and rate of convergence properties (to approximate first-order stationary solutions). Specifically, we show that when certain perturbation parameter remains *constant* across the iterations, the algorithm converges globally sub-linearly to the set of approximate first-order stationary solutions. Further, when the perturbation parameter reduces to zero with an appropriate rate, the algorithm converges to the set of exact first-order stationary solutions. To the best of our knowledge, the proposed algorithm represents one of the first first-order methods with convergence and rate of convergence guarantees (to certain approximate stationary solutions) for problems in the form of (1).

Notation *We use $\|\cdot\|$, $\|\cdot\|_1$, and $\|\cdot\|_F$ to denote the Euclidean norm, ℓ_1-norm, and Frobenius norm respectively. For given vector x, and matrix H, we denote $\|x\|_H^2 := x^T Hx$. For two vectors a, b we use $\langle a, b\rangle$ to denote their inner product. We use $\sigma_{\max}(A)$ to denote the maximum eigenvalue for a matrix A. We use I_N to denote an $N \times N$ identity matrix. For a nonsmooth convex function $h(x)$, $\partial h(x)$ denotes the sub-differential set defined by*

$$\partial h(x) = \{v \in \mathbb{R}^N : h(y) \geq h(x) + \langle v, y - x\rangle \; \forall y \in \mathbb{R}^N\}. \tag{11}$$

For a convex function $h(x)$ and a constant $\alpha > 0$ the proximity operator is defined as below

$$prox_h^{1/\alpha}(x) := \operatorname*{argmin}_z \left\{ \frac{1}{2\alpha}\|x - z\|^2 + h(z) \right\}. \tag{12}$$

2 Convergence analysis of PProx-PDA

In this section, we provide the convergence analysis for PProx-PDA presented in Algorithm 1. We will frequently use the following identity

$$\langle b, b - a\rangle = \frac{1}{2}\left(\|b - a\|^2 + \|b\|^2 - \|a\|^2\right). \tag{13}$$

Also, for the notation simplicity we define

$$w^r := (x^{r+1} - x^r) - (x^r - x^{r-1}). \tag{14}$$

To proceed, let us make the following blanket assumptions on problem (1).

Assumption A

A1. The gradient of function $f(x)$ is Lipschitz-continuous on X i.e., there exists $L > 0$ such that

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|, \quad \forall\, x, y \in X. \tag{15}$$

Further, without loss of generality, assume that $f(x) \ge 0$ for all $x \in X$.

A2. The function $h(x)$ is nonsmooth lower semi-continuous convex function, lower bounded (for simplicity we assume $h(x) \ge 0,\ \forall\, x \in X$), and its sub-gradient is bounded.

A3. The problem (1) is feasible.

A4. The feasible set X is a convex and compact set.

A5. The scaling matrix B is chosen such that $A^T A + B^T B \succeq I$.

Our first lemma characterizes the relationship between the primal and dual variables for two consecutive iterations.

Lemma 1 *Under Assumptions A, the following holds true for PProx-PDA for every $r \ge 1$*

$$\begin{aligned}
&\frac{1 - \rho\gamma}{2\rho}\|\lambda^{r+1} - \lambda^r\|^2 + \frac{\beta}{2}\|x^{r+1} - x^r\|^2_{B^T B} + \frac{L}{2}\|x^{r+1} - x^r\|^2 \\
&\le \frac{1 - \rho\gamma}{2\rho}\|\lambda^r - \lambda^{r-1}\|^2 + \frac{\beta}{2}\|x^r - x^{r-1}\|^2_{B^T B} \\
&+ \frac{L}{2}\|x^r - x^{r-1}\|^2 + L\|x^{r+1} - x^r\|^2 - \gamma\|\lambda^{r+1} - \lambda^r\|^2.
\end{aligned} \tag{16}$$

Proof From the optimality condition of the x-update in (3a) we obtain

$$\begin{aligned}&\langle \nabla f(x^r) + A^T\lambda^r(1-\rho\gamma) + \rho A^T(Ax^{r+1}-b)\\&\quad + \beta B^TB(x^{r+1}-x^r) + \xi^{r+1}, x^{r+1}-x\rangle \le 0,\ \forall\, x\in X,\end{aligned} \tag{17}$$

for some $\xi^{r+1}\in\partial h(x^{r+1})$. Using the dual update rule (3b) we obtain

$$\langle \nabla f(x^r) + A^T\lambda^{r+1} + \beta B^TB(x^{r+1}-x^r) + \xi^{r+1}, x^{r+1}-x\rangle \le 0,\ \forall\, x\in X. \tag{18}$$

Using this equation for $r-1$, we have, for all $r\ge 1$

$$\langle \nabla f(x^{r-1}) + A^T\lambda^{r} + \beta B^TB(x^{r}-x^{r-1}) + \xi^{r}, x^{r}-x\rangle \le 0,\ \forall\, x\in X, \tag{19}$$

for some $\xi^r\in\partial h(x^r)$. Let $x=x^r$ in the first inequality and $x=x^{r+1}$ in the second, we can then add the resulting inequalities to obtain the following for all $r\ge 1$

$$\begin{aligned}&\langle \nabla f(x^r)-\nabla f(x^{r-1}), x^{r+1}-x^r\rangle + \langle A^T(\lambda^{r+1}-\lambda^r), x^{r+1}-x^r\rangle\\&+\beta\langle B^TBw^r, x^{r+1}-x^r\rangle \le \langle \xi^r-\xi^{r+1}, x^{r+1}-x^r\rangle \le 0,\end{aligned} \tag{20}$$

where in the last inequality we have utilized the monotonicity of the sub-differential.

Now let us analyze each terms on the left hand side (LHS) of (20). For the first term we have the following

$$\langle \nabla f(x^{r-1})-\nabla f(x^r), x^{r+1}-x^r\rangle \le \frac{L}{2}\|x^r-x^{r-1}\|^2 + \frac{L}{2}\|x^{r+1}-x^r\|^2, \tag{21}$$

where we applied Young's inequality and the Lipschitz continuity of the gradient of function f. Then we can express the second term in the LHS of (20)

$$\begin{aligned}&\langle A^T(\lambda^{r+1}-\lambda^r), x^{r+1}-x^r\rangle = \langle A(x^{r+1}-x^r), \lambda^{r+1}-\lambda^r\rangle\\&\quad = \langle (Ax^{r+1}-b-\gamma\lambda^r)-(Ax^r-b-\gamma\lambda^{r-1}), \lambda^{r+1}-\lambda^r\rangle + \gamma\langle \lambda^r-\lambda^{r-1}, \lambda^{r+1}-\lambda^r\rangle\\&\overset{(3b),(13)}{=} \frac{1}{2}\left(\frac{1}{\rho}-\gamma\right)\Big(\|\lambda^{r+1}-\lambda^r\|^2 - \|\lambda^r-\lambda^{r-1}\|^2\\&\qquad + \|(\lambda^{r+1}-\lambda^r)-(\lambda^r-\lambda^{r-1})\|^2\Big) + \gamma\|\lambda^{r+1}-\lambda^r\|^2.\end{aligned} \tag{22}$$

For the term $\beta\langle B^TBw^r, x^{r+1}-x^r\rangle$, we have

$$\begin{aligned}&\beta\langle B^TBw^r, x^{r+1}-x^r\rangle \overset{(13)}{=} \frac{\beta}{2}\left(\|x^{r+1}-x^r\|^2_{B^TB} - \|x^r-x^{r-1}\|^2_{B^TB} + \|w^r\|^2_{B^TB}\right)\\&\ge \frac{\beta}{2}\left(\|x^{r+1}-x^r\|^2_{B^TB} - \|x^r-x^{r-1}\|^2_{B^TB}\right).\end{aligned} \tag{23}$$

Therefore, combining (21)–(23), we obtain the desired result in (16). □

Next we analyze the behavior of the primal iterations. Towards this end, let us define the following new quantity

$$T(x,\lambda) := f(x) + h(x) + \langle (1-\rho\gamma)\lambda, Ax - b - \gamma\lambda\rangle + \frac{\rho}{2}\|Ax-b\|^2. \tag{24}$$

Note that this quantity is identical to the augmented Lagrangian when $\gamma = 0$. It is constructed to track the behavior of the algorithm. Even though the function f is not convex, it is easy to show that $T(x,\lambda) + \frac{\beta}{2}\|x - x^r\|^2_{B^T B}$ is strongly convex with respect to the variable x, and with modulus $\beta - L$ when $\rho \geq \beta$, and $\beta > L$. First let us define $g(x,\lambda;x^r) = T(x,\lambda) - h(x) + \frac{\beta}{2}\|x - x^r\|^2_{B^T B}$, which is a smooth function. For this function we have

$$\begin{aligned}
&\langle \nabla_x g(x,\lambda;x^r) - \nabla_y g(y,\lambda;x^r), x - y\rangle\\
&= \langle \nabla f(x) - \nabla f(y) + \rho A^T A(x-y) + \beta B^T B(x-y), x - y\rangle\\
&\overset{\text{(i)}}{\geq} \langle \nabla f(x) - \nabla f(y), x - y\rangle + \beta(A^T A + B^T B)\|x-y\|^2\\
&\overset{\text{(ii)}}{\geq} -L\|x-y\|^2 + \beta(A^T A + B^T B)\|x-y\|^2\\
&\overset{\text{(iii)}}{\geq} (\beta - L)\|x-y\|^2,
\end{aligned} \tag{25}$$

where (i) is true because $\rho \geq \beta$; (ii) is from the Lipschitz continuity of ∇f; (iii) is true because we assumed that [see Assumption A.5] $A^T A + B^T B \succeq I$. This proves that $g(x,\lambda;x^r)$ is strongly convex with modulus $\beta - L$ when $\beta > L$. Since $h(x)$ is assumed to be convex, we conclude that $T(x,\lambda) + \frac{\beta}{2}\|x - x^r\|^2_{B^T B}$ is also strongly convex with modulus $\beta - L$. The next lemma analyzes the change of T in two successive iterations of the algorithm.

Lemma 2 *Suppose that $\beta > 3L$ and $\rho \geq \beta$. Then we have the following*

$$\begin{aligned}
&T(x^{r+1},\lambda^{r+1}) + \frac{(1-\rho\gamma)\gamma}{2}\|\lambda^{r+1}\|^2\\
&\leq T(x^r,\lambda^r) + \frac{(1-\rho\gamma)\gamma}{2}\|\lambda^r\|^2 + \left(\frac{(1-\rho\gamma)(2-\rho\gamma)}{2\rho}\right)\|\lambda^{r+1} - \lambda^r\|^2\\
&\quad - \left(\frac{\beta - 3L}{2}\right)\|x^{r+1} - x^r\|^2, \quad \forall\, r \geq 0.
\end{aligned} \tag{26}$$

Proof It is easy to see that the change of x results in the following reduction of T

$$\begin{aligned}
&T(x^{r+1},\lambda^r) - T(x^r,\lambda^r)\\
&\overset{\text{(i)}}{\leq} \langle \nabla f(x^{r+1}) + \xi^{r+1} + (1-\rho\gamma)A^T\lambda^r + \rho A^T(Ax^{r+1} - b) + \beta B^T B(x^{r+1} - x^r),\\
&\qquad x^{r+1} - x^r\rangle - \frac{\beta - L}{2}\|x^{r+1} - x^r\|^2\\
&\overset{\text{(ii)}}{\leq} -\left(\frac{\beta - 3L}{2}\right)\|x^{r+1} - x^r\|^2,
\end{aligned} \tag{27}$$

where (i) is true because from (25) we know that when $\beta > 3L$, $\rho \ge \beta$ and $A^TA + B^TB \succeq I$, the function $T(x,\lambda) + \frac{\beta}{2}\|x - x^r\|_{B^TB}$ is strongly convex with modulus $\beta - L$; (ii) is true due to the optimality condition (17) for x-subproblem, and the assumption that the gradient of $f(x)$ is Lipschitz continuous. Second, let us analyze $T(x^{r+1},\lambda^{r+1}) - T(x^{r+1},\lambda^r)$ as the following

$$\begin{aligned}
&T(x^{r+1},\lambda^{r+1}) - T(x^{r+1},\lambda^r)\\
&= (1-\rho\gamma)\left(\langle \lambda^{r+1} - \lambda^r, Ax^{r+1} - b - \gamma\lambda^r\rangle\right) - (1-\rho\gamma)\langle \gamma\lambda^{r+1} - \gamma\lambda^r, \lambda^{r+1}\rangle\\
&\overset{(3b)}{=} \frac{1}{\rho}\left\langle (\lambda^{r+1} - \lambda^r) - (\lambda^r - \lambda^{r-1}), \lambda^{r+1} - \lambda^r\right\rangle + \gamma\langle \lambda^r - \lambda^{r+1}, \lambda^{r+1} - \lambda^r\rangle\\
&\overset{(13)}{=} (1-\rho\gamma)\left(\frac{1}{\rho}\|\lambda^{r+1} - \lambda^r\|^2 + \frac{\gamma}{2}(\|\lambda^r\|^2 - \|\lambda^{r+1}\|^2 - \|\lambda^{r+1} - \lambda^r\|^2)\right). \quad (28)
\end{aligned}$$

Combining the previous two steps, we obtain the desired inequality in (26). □

Comparing the results of Lemmas 1 and 2, from (16) we can observe that the term $\frac{1}{2}(\frac{1}{\rho} - \gamma)\|\lambda^{r+1} - \lambda^r\|^2 + \frac{\beta}{2}\|x^{r+1} - x^r\|^2_{B^TB}$ is descending in $\|\lambda^{r+1} - \lambda^r\|^2$ and ascending in $\|x^{r+1} - x^r\|^2$, while from (26) we can see that $T(x^{r+1},\lambda^{r+1}) + \frac{(1-\rho\gamma)\gamma}{2}\|\lambda^{r+1}\|^2$ behaves in an opposite manner. Therefore, let us define the following potential function P_c as a conic combination of these two terms such that it is descending in each iteration. For some $c > 0$ let us define

$$\begin{aligned}
&P_c(x^{r+1},\lambda^{r+1}; x^r,\lambda^r) := T(x^{r+1},\lambda^{r+1}) + \frac{(1-\rho\gamma)\gamma}{2}\|\lambda^{r+1}\|^2\\
&+ c\left(\frac{1-\rho\gamma}{\rho}\|\lambda^{r+1} - \lambda^r\|^2 + \beta\|x^{r+1} - x^r\|^2_{B^TB} + L\|x^{r+1} - x^r\|^2\right). \quad (29)
\end{aligned}$$

Then according to the previous two lemmas, one can conclude that the following holds

$$\begin{aligned}
&P_c(x^{r+1},\lambda^{r+1}; x^r,\lambda^r) - P_c(x^r,\lambda^r; x^{r-1},\lambda^{r-1})\\
&\le -a_1\|\lambda^{r+1} - \lambda^r\|^2 - a_2\|x^{r+1} - x^r\|^2, \quad (30)
\end{aligned}$$

where a_1 and a_2 are given below

$$a_1 = \left((1-\rho\gamma)\frac{\gamma}{2} + 2c\gamma - \frac{1-\rho\gamma}{\rho}\right), \quad \text{and} \quad a_2 = \left(\frac{\beta - 3L}{2} - 2cL\right). \quad (31)$$

Therefore, in order to make the function P_c decrease at each iteration, it suffices to ensure that

$$(1-\rho\gamma)\frac{\gamma}{2} + 2c\gamma - \frac{1-\rho\gamma}{\rho} > 0, \text{ and } \beta > (3+4c)L. \quad (32)$$

Therefore a sufficient condition is that

$$\tau := \rho\gamma \in (0,1), \quad c > \frac{1}{\tau} - 1 > 0, \quad \beta > (3+4c)L, \quad \rho > \beta. \tag{33}$$

From the discussion here we can see the necessity for having perturbation parameter $\gamma > 0$. In particular, if $\gamma = 0$ the constant in front of the $\|\lambda^{r+1} - \lambda^r\|^2$ would be $\frac{1}{\rho}$, which is always positive. Therefore, it is difficult to construct a potential function that has descent on the dual variable.

Next, let us show that the potential function P_c is lower bounded, when choosing particular parameters given in Lemma 2.

Lemma 3 *Suppose Assumptions A are satisfied, and the algorithm parameters are chosen according to (33). Then the following statement holds true*

$$\exists\, \underline{P} \text{ s.t. } P_c(x^{r+1}, \lambda^{r+1}; x^r, \lambda^r) \geq \underline{P} > -\infty, \quad \forall\, r \geq 0. \tag{34}$$

Proof First, we analyze terms related to $T(x^{r+1}, \lambda^{r+1})$. The inner product term in (24) can be bounded as

$$\begin{aligned}
&\langle \lambda^{r+1} - \rho\gamma\lambda^{r+1}, Ax^{r+1} - b - \gamma\lambda^{r+1}\rangle \\
&\overset{(13)}{=} \frac{1}{2}\left(\frac{1-\rho\gamma}{\rho} - (1-\rho\gamma)\gamma\right)(\|\lambda^{r+1}\|^2 - \|\lambda^r\|^2 + \|\lambda^{r+1} - \lambda^r\|^2) \\
&= \frac{(1-\rho\gamma)^2}{2\rho}(\|\lambda^{r+1}\|^2 - \|\lambda^r\|^2 + \|\lambda^{r+1} - \lambda^r\|^2).
\end{aligned} \tag{35}$$

Clearly, the constant in front of the above equality is positive. Summing over R iterations of $T(x^{r+1}, \lambda^{r+1})$, we obtain

$$\begin{aligned}
&\sum_{r=1}^{R} T(x^{r+1}, \lambda^{r+1}) = \sum_{r=1}^{R}\left(f(x^{r+1}) + h(x^{r+1}) + \frac{\rho}{2}\|Ax^{r+1} - b\|^2\right) \\
&+ \frac{(1-\rho\gamma)^2}{2\rho}(\|\lambda^{R+1}\|^2 - \|\lambda^1\|^2 + \sum_{r=1}^{R}\|\lambda^{r+1} - \lambda^r\|^2) \\
&\geq \sum_{r=1}^{R}\left(f(x^{r+1}) + h(x^{r+1}) + \frac{\rho}{2}\|Ax^{r+1} - b\|^2\right) + \frac{(1-\rho\gamma)^2}{2\rho}(\|\lambda^{R+1}\|^2 - \|\lambda^1\|^2) \\
&\geq -\frac{(1-\rho\gamma)^2}{2\rho}\|\lambda^1\|^2,
\end{aligned} \tag{36}$$

where the last inequality comes from the fact that f and h are both assumed to be lower bounded by 0. Since $\|\lambda^1\|^2 \leq \infty$, it follows that the sum of the $T(\cdot,\cdot)$ function is lower bounded. From (36) we conclude that $\sum_{r=1}^{R} P_c(x^{r+1}, \lambda^{r+1}; x^r, \lambda^r)$ is also lower bounded by $-\frac{(1-\rho\gamma)^2}{2\rho}\|\lambda^1\|^2$ for any R, because besides the term $\sum_{r=1}^{R} T(x^{r+1}, \lambda^{r+1})$, the rest of the terms are all positive. Combined with the fact that P_c is non-increasing we conclude that the potential function is lower bounded. This proves the claim. □

To present the main result on the convergence of the PProx-PDA, we need the following notion of approximate stationary solutions for the problem (1).

Definition 1 *Approximate stationary solution* Consider problem (1). Then for given $\epsilon > 0$, we say the tuple (x^*, λ^*) is an ϵ-stationary solution for the problem (1) if the following holds

$$\|Ax^* - b\|^2 \le \epsilon, \quad \langle \nabla f(x^*) + A^T\lambda^* + \xi^*, x^* - x\rangle \le 0, \quad \forall\, x \in X, \tag{37}$$

where $x^* \in X$ and ξ^* is some vector that satisfies $\xi^* \in \partial h(x^*)$.

It is important to note that, suppose that x^* is a local optimal solution and it satisfies appropriate constraint qualification (CQ) condition, then it must also satisfy (37) with $\epsilon = 0$. We note that the ϵ-stationary solution slightly violates the constraint $\|Ax - b\| = 0$. This definition is closely related to the approximate KKT (AKKT) condition in the existing literature [3,21,30]. It can be verified that when $X = \mathbb{R}^N$, and $h \equiv 0$, then the condition in (37) satisfies the stopping criteria for reaching AKKT condition Eqs. (9)–(11) in [3]. We refer the readers to [3, Section 3.1] for detailed discussion of the relationship between AKKT and KKT conditions.

We show below that by appropriately choosing the algorithm parameters, the PProx-PDA converges to the set of approximate stationary solutions.

Theorem 1 *Suppose Assumptions A hold. Further, assume that the parameters γ, ρ, β, c satisfy (33). Then the following is true for the sequence $\{(x^r, \lambda^r)\}$ generated by the PProx-PDA*

- *The sequences $\{x^r\}$ and $\{\lambda^r\}$ are bounded, and that*

$$\lambda^{r+1} - \lambda^r \to 0, \quad x^{r+1} - x^r \to 0.$$

- *Let (x^*, λ^*) denote any accumulation point of the sequence $\{(x^r, \lambda^r)\}$. Then (x^*, λ^*) is a $(\gamma^2\|\lambda^*\|^2)$-stationary solution of problem (1).*

Proof Using the assumption that X is a compact set we have that the sequence $\{x^r\}$ is bounded. Further, combining the bound given in (30) with the fact that the potential function P_c is decreasing and lower bounded, we have

$$\lambda^{r+1} - \lambda^r \to 0, \quad x^{r+1} - x^r \to 0. \tag{38}$$

Also, from the dual update Eq. (3b) we have $\lambda^{r+1} - \lambda^r = \rho\left(Ax^{r+1} - b - \gamma\lambda^r\right)$. Combining with $\lambda^{r+1} - \lambda^r \to 0$ we can see that $\{\lambda^r\}$ is also bounded. This proves the first part.

In order to prove the second part let (x^*, λ^*) be any accumulation point of the sequence $\{(x^r, \lambda^r)\}$. From (3b) we have $\lambda^{r+1} - \lambda^r = \rho(Ax^{r+1} - b - \gamma\lambda^r)$. Then combining this with (38) we obtain

$$Ax^* - b - \gamma\lambda^* = 0. \tag{39}$$

Thus, we have $\|Ax^* - b\|^2 \le \gamma^2\|\lambda^*\|^2$; which proves the first inequality in (37).

Further, from the optimality condition of (18) we have

$$\begin{aligned}&\langle \nabla f(x^r) + A^T\lambda^r(1-\rho\gamma) + \rho A^T(Ax^{r+1}-b) + \beta B^T B(x^{r+1}-x^r), x^{r+1}-x\rangle\\&\quad\le \langle \xi^{r+1}, x - x^{r+1}\rangle,\ \forall\, x \in X.\end{aligned}\tag{40}$$

Recall that $\xi^{r+1} \in \partial h(x^{r+1})$ and h is convex. Then it follows that for all $x \in X$ we have $\langle \xi^{r+1}, x - x^{r+1}\rangle \le h(x) - h(x^{r+1})$. Plugging this inequality into (40), using the update Eq. (3b) and rearranging the terms we obtain for all $x \in X$

$$h(x^{r+1}) + \langle \nabla f(x^r) + A^T\lambda^{r+1} + \beta B^T B(x^{r+1}-x^r), x^{r+1}-x\rangle \le h(x).\tag{41}$$

Rearranging the terms, we have for all $x \in X$

$$\begin{aligned}&h(x^{r+1}) + \langle \nabla f(x^r), x^{r+1}-x\rangle + \langle \lambda^{r+1}, Ax^{r+1}\rangle + \langle \beta B(x^{r+1}-x^r), B(x^{r+1}-x)\rangle\\&\le\ h(x) + \langle \lambda^{r+1}, Ax\rangle.\end{aligned}\tag{42}$$

Note the following relation holds

$$\begin{aligned}&\langle B(x^{r+1}-x^r), B(x^{r+1}-x)\rangle = \|x^{r+1}-x^r\|^2_{B^TB} + \langle B(x^{r+1}-x^r), B(x^r-x)\rangle\\&= \frac{1}{2}\|x^{r+1}-x^r\|^2_{B^TB} - \frac{1}{2}\|x^r - x\|^2_{B^TB} + \frac{1}{2}\|x^{r+1}-x^r+(x^r-x)\|^2_{B^TB}\\&\ge\ \frac{1}{2}\|x^{r+1}-x^r\|^2_{B^TB} - \frac{1}{2}\|x^r-x\|^2_{B^TB}.\end{aligned}\tag{43}$$

Plugging the above into (42) and add and subtract x^r in the term $\langle \nabla f(x^r), x^{r+1}-x\rangle$ we obtain

$$\begin{aligned}&h(x^{r+1}) + \langle \nabla f(x^r), x^{r+1}-x^r\rangle + \langle \lambda^{r+1}, Ax^{r+1}\rangle + \frac{\beta}{2}\|B(x^{r+1}-x^r)\|^2\\&\le\ h(x) + \langle \nabla f(x^r), x-x^r\rangle + \langle \lambda^{r+1}, Ax\rangle + \frac{\beta}{2}\|B(x-x^r)\|^2,\ \forall\, x \in X.\end{aligned}$$

Let (x^*, λ^*) be an accumulation point for the sequence $\{x^{r+1}, \lambda^{r+1}\}$. Taking the limit, and using the fact that $x^{r+1} - x^r \to 0$, we have

$$h(x^*) + \langle \lambda^*, Ax^*\rangle \le h(x) + \langle \nabla f(x^*), x - x^*\rangle + \langle \lambda^*, Ax\rangle + \frac{\beta}{2}\|B(x-x^*)\|^2,\ \forall\, x \in X.$$

The above inequality suggests that $x = x^*$ achieves the optimality for the right hand side. In particular, we have

$$x^* = \arg\min_{x\in X}\ h(x) + \langle \nabla f(x^*), x - x^*\rangle + \langle \lambda^*, Ax\rangle + \frac{\beta}{2}\|B(x-x^*)\|^2.\tag{44}$$

The optimality of the above problem becomes

$$\langle \nabla f(x^*) + A^T\lambda^* + \xi^*, x^* - x\rangle \le 0, \quad \forall\, x \in X, \tag{45}$$

for some $\xi^* \in \partial h(x^*)$. □

2.1 The choice of perturbation parameter

In this section, we discuss how to obtain ϵ-stationary solution. First, note that in Theorem 1 we proved that the sequence $\{\lambda^r\}$ is bounded. Therefore, if the bound is independent of the choice of parameters γ, ρ, β, c, then one can choose $\gamma = \mathcal{O}(\sqrt{\epsilon})$ to reach an ϵ-optimal solution. In the rest of this section, we take an alternative approach to argue ϵ-stationary solution. Our general strategy is to take $\frac{1}{\rho}$ and γ proportional to the accuracy parameter ϵ, while keeping $\tau = \rho\gamma \in (0, 1)$ and c fixed to some ϵ-independent constants. Let us define the following constants for problem (1)

$$\begin{aligned} &d_1 = \max\{\|Ax - b\|^2 \mid x \in X\}, \quad d_2 = \max\{\|x - y\|^2 \mid x, y \in X\},\\ &d_3 = \max\{\|x - y\|^2_{B^TB} \mid x, y \in X\}, \quad d_4 = \max\{f(x) + h(x) \mid x \in X\}. \end{aligned} \tag{46}$$

The lemma below provides a parameter independent bound for $\frac{\rho}{2}\|Ax^1 - b\|^2$.

Lemma 4 *Suppose $\lambda^0 = 0$, $Ax^0 = b$, $\rho \ge \beta$, and $\beta - 3L > 0$. Then we have*

$$\frac{\rho}{2}\|Ax^1 - b\|^2 \le d_4, \quad \frac{\beta}{2}\|x^1 - x^0\|^2 \le d_4 + \frac{3L}{2}d_2. \tag{47}$$

Proof From Lemma 2 and using x^0 and λ^0 in the statement of the lemma, we obtain

$$\begin{aligned} &T(x^1, \lambda^1) + \frac{(1-\rho\gamma)\gamma}{2}\|\lambda^1\|^2 + \frac{\beta - 3L}{2}\|x^1 - x^0\|^2\\ &\le\ T(x^0, \lambda^0) + \left(\frac{1-\rho\gamma}{\rho} - \frac{\gamma}{2}(1-\rho\gamma)\right)\|\lambda^1\|^2. \end{aligned}$$

Utilizing the definition of $T(x, \lambda)$ and (35), we obtain

$$\begin{aligned} T(x^1, \lambda^1) &= f(x^1) + h(x^1) + \frac{(1-\rho\gamma)^2}{\rho}\|\lambda^1\|^2\\ &\quad + \frac{\rho}{2}\|Ax^1 - b\|^2,\\ T(x^0, \lambda^0) &= f(x^0) + h(x^0). \end{aligned}$$

Combining the above, we obtain

$$\left((1-\rho\gamma)\gamma - \frac{1-\rho\gamma}{\rho} + \frac{(1-\rho\gamma)^2}{\rho}\right)\|\lambda^1\|^2 + \frac{\rho}{2}\|Ax^1 - b\|^2 + \frac{\beta - 3L}{2}\|x^1 - x^0\|^2$$

$$\leq\ T(x^0, \lambda^0) - f(x^1) - h(x^1).$$

By some simple calculation we can show that $\left((1-\rho\gamma)\gamma - \frac{1-\rho\gamma}{\rho} + \frac{(1-\rho\gamma)^2}{\rho}\right) = 0.$ By using the assumption $f(x^1) \geq 0$, $h(x^1) \geq 0$, it follows that

$$\frac{\beta - 3L}{2}\|x^1 - x^0\|^2 \leq d_4, \quad \frac{\rho}{2}\|Ax^1 - b\|^2 \leq d_4. \tag{48}$$

This leads to the desired claim. □

Combining Lemma 4 with the dual update (3b), we can conclude that

$$\frac{1}{2\rho}\|\lambda^1\|^2 = \frac{\rho}{2}\|Ax^1 - b\|^2 \leq d_4. \tag{49}$$

Next, we derive an upper bound for the initial potential function $P_c(x^1, \lambda^1; x^0, \lambda^0)$. Assuming that $Ax^0 = b$, $\lambda^0 = 0$, we have

$$\begin{aligned}
&P_c(x^1, \lambda^1; x^0, \lambda^0) \overset{(29)}{=} T(x^1, \lambda^1) + \frac{(1-\rho\gamma)(\gamma + 2c/\rho)}{2}\|\lambda^1\|^2 \\
&\quad + c\left(\beta\|x^1 - x^0\|^2_{B^TB} + L\|x^1 - x^0\|^2\right) \\
&\overset{(24),(35)}{\leq} \quad f(x^1) + h(x^1) + \frac{(1-\rho\gamma)^2}{\rho}\|\lambda^1\|^2 + \frac{\rho}{2}\|Ax^1 - b\|^2 \\
&\quad + \ \frac{(1-\rho\gamma)(\gamma + 2c/\rho)}{2}\|\lambda^1\|^2 + c\left(\beta\|x^1 - x^0\|^2_{B^TB} + L\|x^1 - x^0\|^2\right) \\
&\overset{(47)}{\leq} \left[2 + 2(1-\rho\gamma)^2 + (1-\rho\gamma)(2c + \rho\gamma)\right] d_4 \\
&+ c\left(2\sigma_{\max}(B^TB)(d_4 + \frac{3L}{2}d_2) + Ld_2\right) \\
&= \left[2 + 2(1-\tau)^2 + (1-\tau)(2c + \tau)\right] d_4 + c\left(2\sigma_{\max}(B^TB)(d_4 + \frac{3L}{2}d_2) + Ld_2\right) \\
&:= P_c^0
\end{aligned} \tag{50}$$

It is important to note that P_c^0 does not depend on ρ, γ, β individually, but only on $\rho\gamma$ and c, both of which can be chosen as absolute constants. The next lemma bounds the size of $\|\lambda^r\|^2$.

Lemma 5 *Suppose that (ρ, γ, β) are chosen according to (33), and the assumptions in Lemma 4 hold true. Then the following holds true for all $r \geq 0$*

$$\frac{\gamma(1-\rho\gamma)}{2}\|\lambda^r\|^2 \leq P_c^0. \tag{51}$$

Proof We use induction to prove the lemma. The initial step $r = 0$ is clearly true. In the inductive step we assume that

$$\frac{\gamma(1-\rho\gamma)}{2}\|\lambda^r\|^2 \le P_c^0, \quad \text{for some } r \ge 0. \tag{52}$$

Using the fact that the potential function is decreasing (cf. (30)), we have

$$P_c(x^{r+1}, \lambda^{r+1}; x^r, \lambda^r) \le P_c(x^1, \lambda^1; x^0, \lambda^0) \le P_c^0. \tag{53}$$

Combining (53) with (35), and using the definition of P_c in (29), we obtain

$$\frac{(1-\rho\gamma)^2}{2\rho}(\|\lambda^{r+1}\|^2 - \|\lambda^r\|^2) + \frac{\gamma(1-\rho\gamma)}{2}\|\lambda^{r+1}\|^2 \le P_c^0. \tag{54}$$

If $\|\lambda^{r+1}\| - \|\lambda^r\| \ge 0$, then we have

$$\frac{\gamma(1-\rho\gamma)}{2}\|\lambda^{r+1}\|^2 \le \frac{\gamma(1-\rho\gamma)}{2}\|\lambda^{r+1}\|^2 + \frac{(1-\rho\gamma)^2}{2\rho}(\|\lambda^{r+1}\|^2 - \|\lambda^r\|^2) \overset{(54)}{\le} P_c^0.$$

If $\|\lambda^{r+1}\| - \|\lambda^r\| < 0$, then we have

$$\frac{\gamma(1-\rho\gamma)}{2}\|\lambda^{r+1}\|^2 < \frac{\gamma(1-\rho\gamma)}{2}\|\lambda^r\|^2 \le P_c^0,$$

where the second inequality comes from the induction assumption (52). This concludes the proof. □

From Lemma 5, and the fact that $\rho\gamma = \tau \in (0, 1)$, we have

$$\gamma\|\lambda^{r+1}\|^2 \le \frac{2P_c^0}{1-\tau}, \quad \forall\, r \ge 0. \tag{55}$$

Therefore, we get

$$\gamma^2\|\lambda^{r+1}\|^2 \le \frac{2P_c^0\gamma}{1-\tau}, \quad \forall\, r \ge 0. \tag{56}$$

Also note that ρ and β should satisfy (33), restated below

$$\tau := \rho\gamma \in (0, 1), \quad c > \frac{1}{\tau} - 1 > 0, \quad \beta > (3+4c)L, \quad \rho \ge \beta. \tag{57}$$

Combining the above results, we have the following corollary about the choice of parameters to achieve ϵ-stationary solution.

Corollary 1 *Consider the following choices of the algorithm parameters.*

$$\gamma = \min\left\{\epsilon, \frac{1}{\beta}\right\}, \quad \rho = \frac{1}{2}\max\left\{\beta, \frac{1}{\epsilon}\right\}, \quad \beta > 11L, \quad c = 2. \tag{58}$$

Further suppose Assumptions A are satisfied, and that $Ax^0 = b$, $\lambda^0 = 0$. Then the sequence of dual variables $\{\lambda^r\}$ lies in a bounded set, and $\lambda^{r+1} - \lambda^r \to 0$, $x^{r+1} - x^r \to 0$. Further, every accumulation point generated by the PProx-PDA algorithm is an ϵ-stationary solution.

Proof Using the parameters in (58), we have the following relation

$$\tau = \rho\gamma = \frac{1}{2}, \quad \frac{\gamma}{1-\rho\gamma} \leq 2\epsilon.$$

Then we can bound P_c^0 by the following

$$\begin{aligned} P_c^0 &= \Big[2 + 2(1-\rho\gamma)^2 + (1-\rho\gamma)(c+\rho\gamma)\Big] d_4 \\ &\quad + \frac{c}{2}(2\sigma_{\max}(B^T B)(d_4 + \frac{3L}{2}d_2) + Ld_2) \\ &\leq (6 + 2\sigma_{\max}(B^T B))d_4 + (3\sigma_{\max}(B^T B)L + L)d_2. \end{aligned}$$

Therefore using (56) we conclude

$$\gamma^2\|\lambda^{r+1}\|^2 \leq \frac{2P_c^0\gamma}{1-\rho\gamma} \leq 4((6 + 2\sigma_{\max}(B^T B))d_4 + (3\sigma_{\max}(B^T B)L + L)d_2)\epsilon.$$

Note that the constant in front of ϵ is not dependent on algorithm parameters. This implies that $\gamma^2\|\lambda^{r+1}\|^2 = \mathcal{O}(\epsilon)$. □

Remark 1 First, in the above result, the ϵ-stationary solution is obtained by imposing the additional assumption that the initial solution is feasible for the linear constraint (i.e., $Ax^0 = b$), and that $\lambda^0 = 0$. Admittedly, obtaining a feasible initial solution could be challenging, but for problems such as distributed optimization (9) and subspace estimation (4), finding feasible x^0 is relatively easy. For the former case either the agents can agree on a trivial solution (such as $x_i = x_j = 0$), or they can run an average consensus-based algorithm such as [67] to reach consensus. For the latter case, one can just set $\Pi = \Phi = 0$. Second, the penalty parameter could be large because it is inversely proportional to the accuracy. Having a large penalty parameter at the beginning can make the algorithm progress slowly. In practice, one can start with a smaller ρ and gradually increase it until reaching the predefined threshold. Following this idea, in the next section, we will design an algorithm that allows ρ to increase unboundedly, such that the exact first-order stationary solution can be obtained in the limit.

2.2 Convergence rate analysis

In this subsection, we briefly discuss the convergence rate of the algorithm. To begin with, assume that parameters are chosen according to (33), and $Ax^0 = b$, $\lambda^0 = 0$. Also we will choose $1/\rho$ and γ proportional to certain accuracy parameter, while keeping $\tau = \rho\gamma \in (0, 1)$ and c fixed to some absolute constants. To proceed, let us define the following quantities

$$H(x^r, \lambda^r) := f(x^r) + h(x^r) + \langle \lambda^r, Ax^r - b\rangle, \tag{59a}$$

$$G(x^r, \lambda^r) := \|\tilde{\nabla} H(x^r, \lambda^r)\|^2 + \frac{1}{\rho^2}\|\lambda^{r+1} - \lambda^r\|^2, \tag{59b}$$

$$Q(x^r, \lambda^r) := \|\tilde{\nabla} H(x^r, \lambda^r)\|^2 + \|Ax^r - b\|^2, \tag{59c}$$

where $\tilde{\nabla} H(x^r, \lambda^r)$ is the proximal gradient defined as

$$\tilde{\nabla} H(x^r, \lambda^r) = x^r - \operatorname{prox}^{\beta}_{h+\iota(X)}\left[x^r - \frac{1}{\beta}\nabla(H(x^r, \lambda^r) - h(x^r))\right]. \tag{60}$$

It can be checked that $Q(x^r, \lambda^r) \to 0$ if and only if a stationary solution for problem (1) is obtained. Therefore we say that a θ-stationary solution is obtained if $Q(x^r, \lambda^r) \le \theta$. Note that the θ-stationary solution has been used in [40] for characterizing the rate for ADMM method. Compared with the ϵ-stationary solution defined in Definition 1, its progress is easier to quantify. Using the definition of proximity operator, the optimality condition of the x-subproblem (3a) can be equivalently written as

$$x^{r+1} = \operatorname{prox}^{\beta}_{h+\iota(X)}\left[x^{r+1} - \frac{1}{\beta}\big[\nabla f(x^r) + A^T\lambda^{r+1} + \beta B^T B(x^{r+1} - x^r)\big]\right].$$

By using the non-expansiveness of the proximity operator, we obtain the following

$$\begin{aligned}
\|\tilde{\nabla} H(x^r, \lambda^r)\|^2 &= \left\|x^r - \operatorname{prox}^{\beta}_{h+\iota(X)}\left[x^r - \frac{1}{\beta}\nabla\big[H(x^r, \lambda^r) - h(x^r)\big]\right]\right\|^2 \\
&= \Bigg\|x^{r+1} - \operatorname{prox}^{\beta}_{h+\iota(X)}\left[x^{r+1} - \frac{1}{\beta}\big[\nabla f(x^r) + A^T\lambda^{r+1} + \beta B^T B(x^{r+1} - x^r)\big]\right] \\
&\quad - x^r + \operatorname{prox}^{\beta}_{h+\iota(X)}\left[x^r - \frac{1}{\beta}\nabla\big[H(x^r, \lambda^r) - h(x^r)\big]\right]\Bigg\|^2 \\
&\le 2\|x^{r+1} - x^r\|^2 + \frac{4}{\beta^2}\|A^T(\lambda^{r+1} - \lambda^r)\|^2 + 4\|(I - B^T B)(x^{r+1} - x^r)\|^2 \\
&\le (2 + 4\sigma^2_{\max}(\hat{B}))\|x^{r+1} - x^r\|^2 + \frac{4\sigma_{\max}(A^T A)}{\beta^2}\|\lambda^{r+1} - \lambda^r\|^2,
\end{aligned}$$

where in the last inequality we define $\hat{B} := I - B^T B$. Therefore,

$$G(x^r, \lambda^r) \le b_1\|\lambda^{r+1} - \lambda^r\|^2 + b_2\|x^{r+1} - x^r\|^2, \tag{61}$$

where we have defined

$$b_1 = \frac{4\sigma_{\max}(A^T A)}{\beta^2} + \frac{1}{\rho^2}, \quad b_2 = 2 + 4\sigma^2_{\max}(\hat{B}). \tag{62}$$

Combining (61) with the descent estimate for the potential function P_c given in (30), we obtain

$$G(x^r, \lambda^r) \le V\left[P_c(x^r, \lambda^r; x^{r-1}, \lambda^{r-1}) - P_c(x^{r+1}, \lambda^{r+1}; x^r, \lambda^r)\right], \tag{63}$$

where we have defined

$$V := \frac{\max(b_1, b_2)}{\min(a_1, a_2)}.$$

It is easy to verify that V is in the order of $\mathcal{O}(1/\gamma)$ because a_1 is in the order of γ; cf. (31). From part 1 of Theorem 1 and (61) we conclude that $G(x^r, \lambda^r) \to 0$. Let R denote the first time that $G(x^{r+1}, \lambda^{r+1})$ reaches below a given number $\theta > 0$. Summing both sides of (63) over R iterations, and utilizing the fact that P_c is lower bounded by $\underline{P}$, it follows that

$$\theta \le \frac{V\left(P_c^0 - \underline{P}\right)}{R} \le \frac{V\left(P_c^0 + \frac{(1-\rho\gamma)^2}{2\rho}\|\lambda^1\|^2\right)}{R} \overset{(49)}{\le} \frac{V\left(P_c^0 + (1-\tau)^2 d_4\right)}{R}, \tag{64}$$

where d_4 is given in (46), and P_c^0 is given in (50). Note that $G^{r+1} \le \theta$ implies that $1/\rho^2\|\lambda^{r+1} - \lambda^r\|^2 = \|Ax^{r+1} - b - \gamma\lambda^r\|^2 \le \theta$. From (51) we have that

$$\|\gamma\lambda^{r+1}\| \le \sqrt{\frac{2P_c^0\gamma}{1-\rho\gamma}}, \quad \forall\, r \ge 0.$$

It follows that

$$\|Ax^{r+1} - b\| \le \frac{1}{\rho}\|\lambda^{r+1} - \lambda^r\| + \|\gamma\lambda^r\| \le \sqrt{\theta} + \sqrt{\frac{2P_c^0\gamma}{1-\rho\gamma}}.$$

It follows that whenever $G(x^r, \lambda^r) \le \theta$ we have

$$Q(x^r, \lambda^r) := \|\tilde{\nabla} H(x^r, \lambda^r)\|^2 + \|Ax^r - b\|^2 \le \theta + \left(\sqrt{\theta} + \sqrt{\frac{2P_c^0\gamma}{1-\rho\gamma}}\right)^2. \tag{65}$$

Let us pick the parameters such that they satisfy (33) and the following [note that this is always possible by fixing τ and making γ in the order of $\mathcal{O}(\theta)$]

$$\frac{2P_c^0\gamma}{1-\rho\gamma} = \frac{2P_c^0\gamma}{1-\tau} = \theta.$$

Then whenever $G(x^r, \lambda^r) \leq \theta$, we have $Q(x^r, \lambda^r) \leq 5\theta$. Using (64), it follows that the total number of iterations it takes for $Q(x^r, \lambda^r)$ to reach below 5θ is given by

$$R \leq \frac{V\big(P_c^0 + (1-\tau)^2 d_4\big)}{\theta} = \mathcal{O}\left(\frac{1}{\theta^2}\right), \tag{66}$$

where the last relation holds because V is in the order of $\mathcal{O}(\frac{1}{\gamma})$, γ is chosen in the order of $\mathcal{O}(\theta)$, and P_c^0, d_4 and τ are not dependent on the problem accuracy. The result below summarizes our discussion above.

Corollary 2 *Suppose that $Ax^0 = b$ and $\lambda^0 = 0$. Additionally, for a given $\theta > 0$, and $\tau \in (0, 1)$, choose γ, ρ, c, β as follows*

$$\gamma = \frac{\theta(1-\tau)}{2P_c^0}, \quad \rho = \frac{\tau}{\gamma}, \quad c > \frac{1}{\tau} - 1, \quad \rho \geq \beta \quad \textit{and}\ \beta > (3+4c)L.$$

Let R denote the first time that $Q(x^r, \lambda^r)$ reaches below 5θ. Then we have $R = \mathcal{O}\left(\frac{1}{\theta^2}\right)$.

3 An algorithm with increasing accuracy

So far we have shown that PProx-PDA converges to the set of *approximate* stationary solutions by properly choosing the algorithm parameters. The inaccuracy of the algorithm can be attributed to the use of perturbation parameter γ. Is it possible to gradually reduce the perturbation so that asymptotically the algorithm reaches the *exact* stationary solutions? Is it possible to avoid using very large penalty parameter ρ at the beginning of the algorithm? This section designs an algorithm that addresses the above questions. We consider a modified algorithm in which the parameters (ρ, β, γ) are *iteration-dependent*. In particular, we choose ρ^{r+1}, β^{r+1} and $1/\gamma^{r+1}$ to be increasing sequences. The new algorithm, named PProx-PDA with increasing accuracy (PProx-PDA-IA), is listed in Algorithm 2.

Algorithm 2: PProx-PDA with increasing accuracy (PProx-PDA-IA)

```
Initialize: $\lambda^0$ and $x^0$
Repeat: update variables by
    $x^{r+1} = \arg\min_{x \in X} \Big\{ \langle \nabla f(x^r), x - x^r \rangle + h(x) + \langle (1 - \rho^{r+1}\gamma^{r+1})\lambda^r, Ax - b \rangle$
        $+ \frac{\rho^{r+1}}{2}\|Ax - b\|^2 + \frac{\beta^{r+1}}{2}\|x - x^r\|^2_{B^TB} \Big\}.$    (67a)
    $\lambda^{r+1} = (1 - \rho^{r+1}\gamma^{r+1})\lambda^r + \rho^{r+1}\left(Ax^{r+1} - b\right)$
        $= \lambda^r + \rho^{r+1}\left(Ax^{r+1} - b - \gamma^{r+1}\lambda^r\right).$    (67b)
Until Convergence.
```

Below we analyze the convergence of the new algorithm. Besides assuming that the optimization problem under consideration satisfies Assumptions A, we make the following additional assumptions:

Assumption B

B1. Assume that

$$\rho^{r+1}\gamma^{r+1} = \tau \in (0, 1), \quad \text{for some fixed constant } \tau.$$

B2. The sequence $\{\rho^r\}$ satisfies

$$\rho^{r+1} \to \infty, \; \sum_{r=1}^{\infty} \frac{1}{\rho^{r+1}} = \infty, \; \sum_{r=1}^{\infty} \frac{1}{(\rho^{r+1})^2} < \infty, \; \rho^{r+1} - \rho^r = D > 0,$$

for some $D > 0$. A simple choice of ρ^{r+1} is $\rho^{r+1} = r+1$. Similarly, the sequence $\{\gamma^{r+1}\}$ satisfies

$$\gamma^{r+1} - \gamma^r \le 0, \; \gamma^{r+1} \to 0, \; \sum_{r=1}^{\infty} \gamma^{r+1} = \infty, \; \sum_{r=1}^{\infty} (\gamma^{r+1})^2 < \infty. \tag{68}$$

B3. Assume that

$$\exists \, c_0 > 1 \text{ s.t. } \beta^{r+1} = c_0 \rho^{r+1}, \quad \text{for } r \text{ large enough.} \tag{69}$$

B4. There exists $\Lambda > 0$ such that for every $r > 0$ we have $\|\lambda^r\| \le \Lambda$.

We note that Assumption [B4] is somewhat restrictive because it is an assumption on the iterates, therefore, it is difficult to verify *a priori*. In the "Appendix", we will show that such an assumption can be satisfied under some additional regularity condition about problem (1). We choose to state [B4] here to avoid lengthy discussion on those regularity conditions before the main convergence analysis. The main idea of the proof is similar to that of Theorem 1. We first construct certain potential function and show that with appropriate choices of algorithm parameters, it will decrease *eventually*. A classical result in [9] is then used to argue the convergence. The main challenge is to carefully analyze how different algorithm parameters affect the dynamics of the iterates. Due to space limitation, below we provide an outline of the proof. The full proof can be found on the authors' website.[1]

Theorem 2 *Suppose that Assumptions A–B hold true, and that τ, c and D are picked such that the following relations are satisfied*

$$0 < c < \frac{c_0}{2(L + c_0\|B^T B\|)}, \quad 0 < D < \frac{1-\tau}{c}. \tag{70}$$

Then every accumulation point of the sequence generated by PProx-PDA-IA is a stationary solution of problem (1).

[1] http://people.ece.umn.edu/~mhong/PProx_PDA.pdf.

Proof Sketch Similar to Lemma 1, our first step utilizes the optimality condition of two consecutive iterates to analyze the change of the primal and dual differences.
Step 1 Suppose that the Assumptions A and [B1]–[B3] hold true, and that τ, D, are constants defined in Assumptions B. Then for large enough r, there exists constant C_1 such that

$$\begin{aligned}
&\frac{(1-\tau)}{2}\|\lambda^{r+1}-\lambda^r\|^2+\frac{\tau}{2}(\frac{\rho^{r+1}}{\rho^{r+2}}-1)\|\lambda^{r+1}\|^2\\
&+\frac{\beta^{r+1}\rho^{r+1}}{2}\|x^{r+1}-x^r\|^2_{B^TB}+\frac{\rho^{r+1}L}{2}\|x^{r+1}-x^r\|^2_{B^TB}\\
&\le\frac{(1-\tau)}{2}\|\lambda^r-\lambda^{r-1}\|^2+\frac{\tau}{2}(\frac{\rho^r}{\rho^{r+1}}-1)\|\lambda^r\|^2+\frac{\beta^r\rho^r}{2}\|x^r-x^{r-1}\|^2_{B^TB}\\
&\quad+\frac{\rho^rL}{2}\|x^r-x^{r-1}\|^2_{B^TB}\\
&\quad-\frac{\tau}{2}\|\lambda^{r+1}-\lambda^r\|^2+\frac{C_1(\gamma^{r+1})^2}{2}\|\lambda^{r+1}\|^2\\
&\quad+\frac{L\rho^r+D(L+\beta^{r+1}\|B^TB\|)}{2}\|x^{r+1}-x^r\|^2_{B^TB}-\frac{\beta^r\rho^r}{2}\|w^r\|^2_{B^TB}.
\end{aligned}\tag{71}$$

Step 2 The second step analyzes the behavior of $T(x,\lambda)$ which is originally defined in (24) in order to bound the descent of the primal variable. In this case, because T is also a function of time varying parameters ρ and γ, we denote it as $T(x,\lambda;\rho,\gamma)$. Suppose that the Assumptions A and [B1]–[B3] hold true, τ and D are constants defined in Assumptions B. Then we have

$$\begin{aligned}
&T(x^{r+1},\lambda^{r+1};\rho^{r+2},\gamma^{r+2})+\left((1-\tau)\frac{\gamma^{r+2}}{2}+D\frac{2\tau-1}{2\tau}(\gamma^{r+2})^2\right)\|\lambda^{r+1}\|^2\\
&\le T(x^r,\lambda^r;\rho^{r+1},\gamma^{r+1})+\left((1-\tau)\frac{\gamma^{r+1}}{2}+D\frac{2\tau-1}{2\tau}(\gamma^{r+1})^2\right)\|\lambda^r\|^2\\
&-\left(\frac{\beta^{r+1}-3L}{2}\right)\|x^{r+1}-x^r\|^2+(1-\tau)\left(\frac{2-\tau}{2\rho^{r+1}}+\frac{D(\gamma^{r+1})^2}{2\tau^2(1-\tau)}\right)\|\lambda^{r+1}-\lambda^r\|^2\\
&+\left[\frac{(1-\tau)(\gamma^{r+1}-\gamma^{r+2})}{2}+\frac{D(\gamma^{r+2})^2}{2}+D\frac{(\gamma^{r+1})^2-(\gamma^{r+2})^2}{2\tau}\right]\|\lambda^{r+1}\|^2.
\end{aligned}\tag{72}$$

[Step 3] In the next step we construct and estimate the descent of the potential function. For some given $c>0$, satisfying (70) we construct the following potential function

$$\begin{aligned}
P_c^{r+1}:=&T(x^{r+1},\lambda^{r+1};\rho^{r+2},\gamma^{r+2})\\
&+\left((1-\tau)\frac{\gamma^{r+2}}{2}+D\frac{2\tau-1}{2\tau}(\gamma^{r+2})^2\right)\|\lambda^{r+1}\|^2
\end{aligned}$$

$$+ c\bigg(\frac{(1-\tau)}{2}\|\lambda^{r+1}-\lambda^r\|^2 + \frac{\tau}{2}(\frac{\rho^{r+1}}{\rho^{r+2}}-1)\|\lambda^{r+1}\|^2 + \frac{\beta^{r+1}\rho^{r+1}}{2}\|x^{r+1}-x^r\|^2_{B^TB} + \frac{\rho^{r+1}L}{2}\|x^{r+1}-x^r\|^2_{B^TB}\bigg). \tag{73}$$

Suppose that the Assumptions A and [B1]–[B3] hold true, and let τ and D be the constants defined in Assumptions B. Then for large enough r we have the following for the potential function P_c

$$P_c^{r+1} - P_c^r \le -\left(\frac{\beta^{r+1}-3L}{2} - cL\rho^r - cDL - c\beta^{r+1}\|B^TB\|\right)\|x^{r+1}-x^r\|^2 - c\frac{\tau}{4}\|\lambda^{r+1}-\lambda^r\|^2 + D_0\|\lambda^{r+1}\|^2(\gamma^{r+1})^2 - c\frac{\beta^r\rho^r}{2}\|w^r\|^2_{B^TB}, \tag{74}$$

where D_0 is a positive constant. Further, the potential function is lower bounded. *Sketch Proof of Theorem* 2 First, fix a small enough $c > 0$ to make the constant in front of $\|x^{r+1}-x^r\|^2$ in (74) proportional to $-\beta^{r+1}$ for large enough r. Then using Steps 2–3 and the boundedness assumption of λ^{r+1}, we have

$$\sum_{r=1}^{\infty}\beta^{r+1}\|x^{r+1}-x^r\|^2 < \infty, \quad \sum_{r=1}^{\infty}\|\lambda^{r+1}-\lambda^r\|^2 < \infty, \tag{75}$$

$$\sum_{r=1}^{\infty}(\beta^{r+1})^2\|w^{r+1}\|^2_{B^TB} < \infty. \tag{76}$$

From (75) we have $\lambda^{r+1}-\lambda^r \to 0$, which implies that $(\rho^{r+1})(Ax^{r+1}-b)-\tau\lambda^r \to 0$. Combined with the assumption that $\{\lambda^r\}$ is a bounded sequence, and $\rho^{r+1} \to \infty$, we conclude $Ax^{r+1}-b \to 0$. Let (x^*, λ^*) be an accumulation point of (x^{r+1}, λ^{r+1}). Comparing the optimality condition of the problem (1) and the optimality condition of x-subproblem (67a), in order to argue convergence to stationary solutions, we need to show $\beta^{r+1}\|x^{r+1}-x^r\| \to 0$. To proceed, let us define $v^{r+1} := \beta^{r+1}(x^{r+1}-x^r)$. From the first inequality in (75) (and after shifting the indices)

$$\sum_{r=1}^{\infty}\frac{1}{\beta^r}\|v^r\|^2 < \infty. \tag{77}$$

This relation combined with Assumption [B3] implies: $\liminf_{r\to\infty}\|v^r\| = 0$. The remaining part of the proof is to show $\limsup_{r\to\infty}\|v^r\| = 0$. The proof is a modification of the classical result for using diminishing stepsize for the unconstrained problem in [9, Proposition 3.5], and recent extension to the constrained problems in [64, Theorem 4]. The difference is that none of these works involves relaxing the equality constraints. Due to space limitations, we omit the rest of the proof.

4 Numerical results

In this section, we customize the proposed algorithms to a number of applications in Sect. 1.2, and compare with the state-of-the-art algorithms.

4.1 Distributed nonconvex quadratic problem

In this subsection we consider the nonconvex regularized, nonnegative, sparse principal component analysis (SPCA) problem [6]. Distributed version of this problem [which is a special case of problem (1)] can be modeled as below

$$\begin{aligned} \min_{x} \quad & \sum_{i=1}^{N} \left\{ -x_i^\top \Sigma_i x_i + h_i(x_i) \right\} \\ \text{s.t.} \quad & \|x_i\|^2 \le 1, \quad x_i \ge 0, \quad i = 1, \ldots, N \\ & Ax = 0, \quad \texttt{[Consensus Constraint]} \end{aligned} \tag{78}$$

where $x_i \in \mathbb{R}^d$ for each i; $x := \{x_i\}_{i=1}^N$ stacks all x_i's, $\Sigma_i \in \mathbb{R}^{d\times d}$ is the covariance matrix for the mini-batch data in node i; $h_i(\cdot)$ is some regularizer to promote structure in the solution (e.g., sparsity). For simplicity define $h(x) := \sum_{i=1}^N h_i(x_i)$

Define $\bar{x} := \frac{1}{N}\sum_{i=1}^N x_i$, $f(\bar{x}) := \sum_{i=1}^N \bar{x}^T \Sigma_i \bar{x}$, and $X := \{x_i \mid \|\bar{x}\|^2 \le 1, \bar{x} \ge 0\}$. The stationarity-gap (stat-gap) and the constraint violation (const-vio) are defined below

$$\text{stat-gap} = \left\| \bar{x} - \text{prox}_{h+\iota_X}\left[\bar{x} - \nabla f(\bar{x})\right] \right\|^2, \quad \text{const-vio} = \|Ax\|^2. \tag{79}$$

At this point, one can certainly use Algorithm 1 or Algorithm 2 to solve problem (78) directly. However, the resulting x- subproblems for both algorithms are difficult to solve due to the fact that computing the proximity operator for nonsmooth function $h(x) + \iota_{\|x\|^2\le 1}(x) + \iota_{x\ge 0}(x)$ does not have a closed form. On the contrary, in most of the problems encountered in practice, the proximity operators for the individual component functions all have closed-form. To utilize such a problem structure, we divide the agents into three subsets, each with a distinctive regularizer or constraint set. Let us denote $r = \lfloor N/3 \rfloor$. The new reformulation is given below

$$\begin{aligned} \min \quad & \sum_{i=1}^{r} \left\{ -x_i^\top \Sigma_i x_i + \frac{N}{r} h_i(x_i) \right\} - \sum_{i=r+1}^{2r} x_i^\top \Sigma_i x_i - \sum_{i=2r+1}^{N} x_i^\top \Sigma_i x_i \\ \text{s.t.} \quad & \|x_i\|^2 \le 1, \quad i = r+1, \ldots, 2r \\ & x_i \ge 0, \qquad i = 2r+1, \ldots, N \\ & Ax = 0, \quad \texttt{[Consensus Constraint]}. \end{aligned} \tag{80}$$

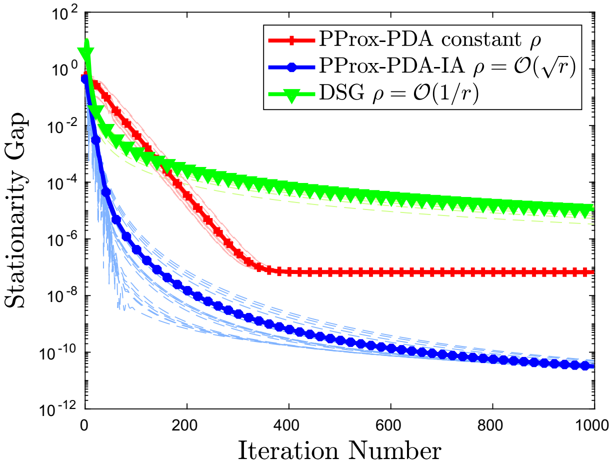

Fig. 1 Comparison of proposed algorithms with DSG [56] in terms of stationarity-gap for problem (80) with parameters $N = 20$, $R = 0.7$, $d = 10$, $\alpha = 0.01$

To the best of our knowledge, no existing methods for nonconvex distributed optimization can effectively deal with the above problem (at least not with theoretical convergence guarantee to the stationary solution). The major difficulty is to deal with the *agent-specific* nonsmooth terms. In our numerical result, the graph $\mathcal{G}$ is generated based on the scheme proposed in [72]. In this scheme, a random graph with N nodes and radius R is generated with nodes uniformly distributed over a unit square, and two nodes connect to each other if their distance is less than R. The test problems are generated in the following manner. We specialize the regularizer to be $h_i(x_i) := \alpha\|x_i\|_1, \ \forall\, i$. The number of agents, the network radius, the problem dimension, and the sparsity parameter are chosen to be: $N = 20, R = 0.7, d = 10, \alpha = 0.01$, respectively. For PProx-PDA we set perturbation parameter $\gamma = 10^{-4}$, and ρ and β are picked such that they satisfy the theoretical bounds given in (57). For PProx-PDA-IA we set the increasing penalty and the proximal coefficients to be $\rho^r = \beta^r = 30r, \ \forall\, r$, and decreasing perturbation parameter to be $\gamma^r = 10^{-3}/r, \ \forall\, r$. The proposed methods are compared with the DSG algorithm [56], whose parameters are given below. For the DSG algorithm the stepsize is set to $0.1/r$ (this choice is made so that DSG has the best performance). Each algorithm is run for 20 independents trials, with random initialization and randomly generated data. All algorithms stop after 1000 iterations. The results are plotted in Figs. 1 and 2. In the figures, dashed lines with light colors are used to show the performance for each individual trial, while the solid dark lines are the average performance over all 20 trials. From the plots it can be observed that the proposed algorithms, especially the increasing penalty version, outperform the DSG algorithm.

To see more numerical results we compare different algorithms with different problem setups. The algorithms are run for 20 independent trials with randomly generated data and random initial solutions in each individual trials. All algorithm parameters are

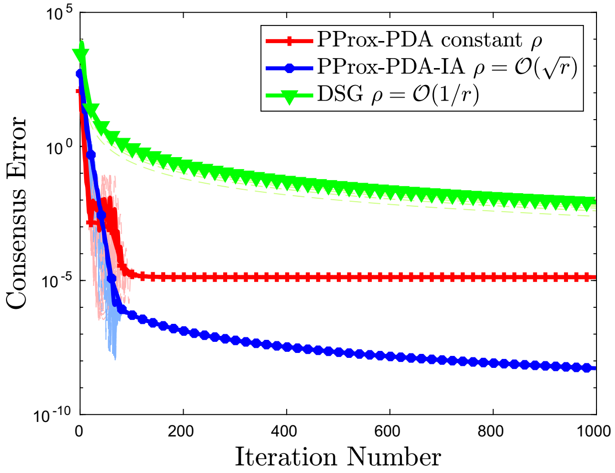

Fig. 2 Comparison of proposed algorithms with DSG [56] in terms of constraint violation for problem (80) with parameters $N = 20$, $R = 0.7$, $d = 10$, $\alpha = 0.01$

Table 1 Comparison of proposed algorithms with DSG algorithm

Parameters	Stationarity-gap			Cons-Vio		
	Alg1	Alg2	DSG	Alg1	Alg2	DSG
$N = 5, n = 80, R = 0.7$	1.9E−4	6.0E−5	9.0E−4	6.0E−6	9.5E−7	4.3E−5
$N = 20, n = 15, R = 0.7$	1.3E−4	5.0E−8	9.4E−5	1.7E−3	6.8E−6	0.013
$N = 30, n = 20, R = 0.5$	6.3E−5	2.1E−8	2.6E−4	7.0E−3	6.4E−7	0.06
$N = 40, n = 30, R = 0.5$	2.0E−4	4.9E−8	1.5E−3	8.1E−3	1.5E−6	0.05

Alg1 and Alg2 denote PProx-PDA and PProx-PDA-IA algorithms respectively

set to be the same as in the previous experiment. The comparison results are displayed in Table 1. The first column describes the problem parameters including the number of agents N, the number of variables n, and the network radius R, while ‘Alg1’ and ‘Alg2’ stand for PProx-PDA and PProx-PDA-IA, respectively. It can be observed that in all scenarios the proposed algorithms outperform DSG.

4.2 Nonconvex subspace estimation

In this subsection we study the problem of *sparse subspace estimation* (4). We compare the proposed PProx-PDA and PProx-PDA-IA with the ADMM algorithm studied in [29, Algorithm 1]. Note that the latter is a heuristic algorithm that does not have convergence guarantee. We first consider a problem with the number of samples, problem dimension, and MCP parameters chosen as $n = 80$, $p = 128$, $\nu = 3$, $b = 3$, respectively. For PProx-PDA we set perturbation parameter $\gamma = 10^{-4}$, and ρ and β are chosen to satisfy the theoretical bounds given in (57). For PProx-PDA-IA we set increasing penalty $\rho = \beta = 5r$, and decreasing perturbation $\gamma = 10^{-4}/r$. The data set is generated following the same procedure as in [29]. In particular, we set $s = 5$

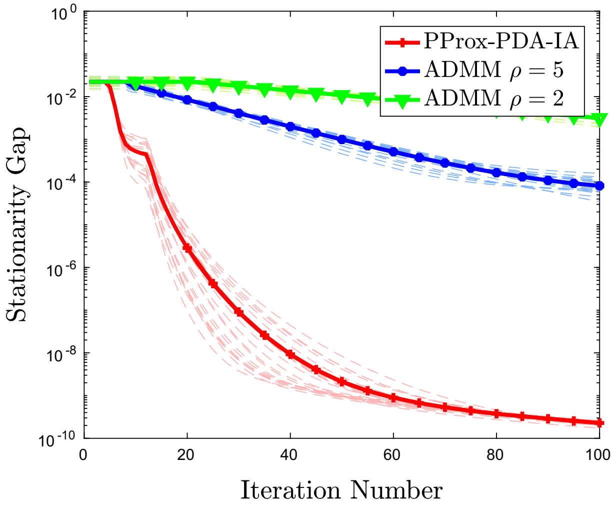

Fig. 3 Comparison of proposed algorithms with ADMM in terms of stationarity-gap for nonconvex subspace estimation problem with MCP regularization. Each dotted line represents the performance of one realization, and each solid line represents an average of 20 independent trials

and $k = 1$, the leading eigenvalue of its covariance matrix Σ is set as $\nu_1 = 100$, and its corresponding eigenvector is sparse such that only the first $s = 5$ entries are nonzero, and they take the value $1/\sqrt{5}$. The rest of the eigenvalues are set to be 1, and their eigenvectors are chosen arbitrarily. For all three algorithms, we measure the stationarity-gap, the constraint violation, the objective value, and the distance to the global optimal solution (i.e. $\|\hat{\Pi} - \Pi^*\|$). The results, which are from 20 independent trials with random initial solutions, are plotted in Figs. 3, 4, 5, 6. As shown in these figures, compared to the ADMM algorithm, the PProx-PDA-IA algorithm converges faster, and to better solutions.

Our next experiment is designed to understand the effect that the problem parameters (i.e. n, p, k, and s) have on the solution quality. Here, we compare the PProx-PDA-IA [with $\rho = \mathcal{O}(r), \gamma = \mathcal{O}(1/r)$] with ADMM algorithm with step-size $\rho = 5$. Both algorithms will be run for 200 iterations. In this experiment we generate data sets with $s = 10$, $k = 5$, and vary other problem parameter. For this dataset the top five eigenvalues are set as $\lambda_1 = \cdots = \lambda_4 = 100$ and $\lambda_5 = 10$. To generate their corresponding eigenvectors we sample its nonzero entries from a standard Gaussian distribution, and then orthnormalize them while retaining the first $s = 10$ rows to be nonzero [29]. The rest of the eigenvalues are set as $\lambda_6 = \cdots = \lambda_p = 1$, and the associated eigenvectors are chosen arbitrarily. The results in terms of the error $\|\hat{\Pi} - \Pi^*\|$ are shown in Table 2. In all scenarios the proposed algorithm PProx-PDA-IA outperforms ADMM.

Further, the True Positive Rate (TPR) and False Positive Rate (FPR) [24] are measured and the results are displayed in Table 3 to see the recovery results. For this problem the event of being zero in vector $v = |\text{supp}(\text{diag}(\hat{\Pi}))|$ (here $\hat{\Pi}$ denotes the output of the algorithm) is considered as a positive event. Let P denotes the number of

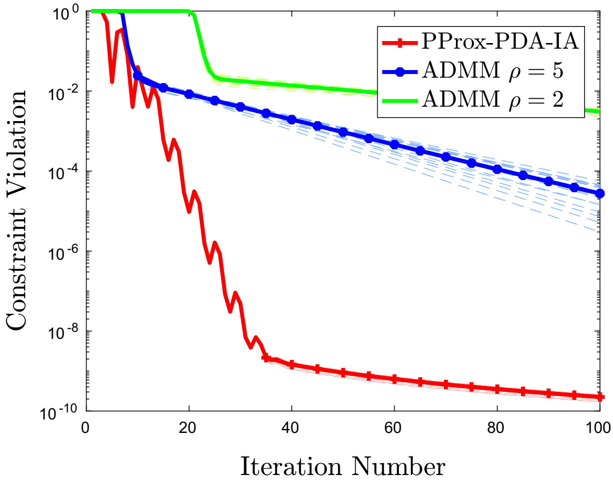

Fig. 4 Comparison of proposed algorithms with ADMM in terms of constraint violation $\|Ax\|^2$ for nonconvex subspace estimation problem with MCP regularization. Each dotted line represents the performance of one realization, and each solid line represents an average of 20 independent trials

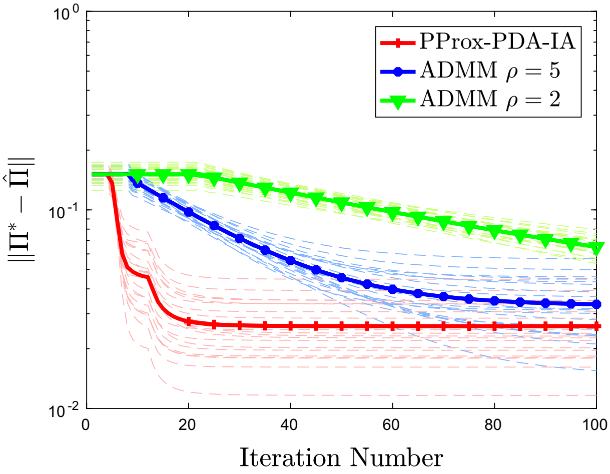

Fig. 5 Comparison of proposed algorithms with ADMM in terms of global error for nonconvex subspace estimation problem with MCP regularization. The problem parameters are $n = 80$, $p = 128$, $\nu = 3$, $b = 3$. Each dotted line represents the performance of one realization, and each solid line represents an average of 20 independent trials

positives, and S denotes the number of non-zeros in the ground truth vector denoted by Π^*. Further, let us use FP and TP to denote *false positive* and *true positive* respectively. In particular, FP counts the number of positive events (i.e. zeros in our case) in vector $\hat{\Pi}$ which are nonzero in ground truth vector Π^*. In contrast, TP counts the number of zeros in $\hat{\Pi}$ which are true zeros in Π^*. Given these notations, the FPR and TPR are defined as follows

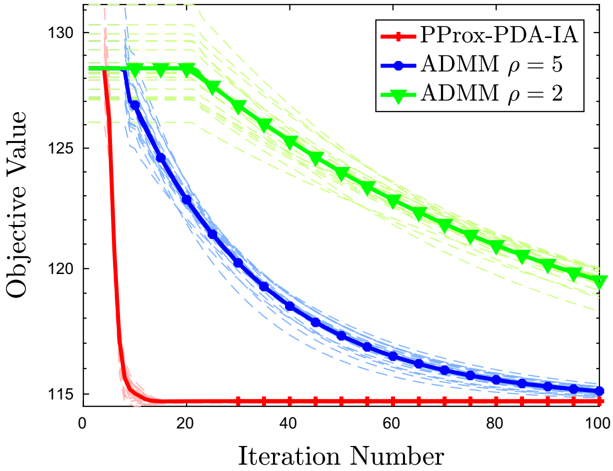

Fig. 6 Comparison of proposed algorithms with ADMM in terms of objective value for nonconvex subspace estimation problem with MCP regularization. The solid lines and dotted lines represent the single performance and the average performance, respectively

Table 2 Comparison of PPox-PDA-IA with ADMM in terms of global error $\|\hat{\Pi} - \Pi^*\|$ for nonconvex subspace estimation problem with MCP regularization

Parameters	$\|\hat{\Pi} - \Pi^*\|$	
	PProx-PDA-IA	ADMM
$n = 30, p = 128, k = 1, s = 5$	0.045 ± 0.02	0.052 ± 0.02
$n = 80, p = 128, k = 1, s = 5$	0.024 ± 0.01	0.028 ± 0.08
$n = 120, p = 128, k = 1, s = 5$	0.020 ± 0.07	0.021 ± 0.06
$n = 150, p = 200, k = 1, s = 5$	0.022 ± 0.07	0.022 ± 0.07
$n = 80, p = 128, k = 1, s = 10$	0.048 ± 0.01	0.062 ± 0.01
$n = 80, p = 128, k = 5, s = 10$	0.21 ± 0.05	0.29 ± 0.02
$n = 128, p = 128, k = 5, s = 10$	0.18 ± 0.02	0.25 ± 0.02
$n = 70, p = 128, k = 5, s = 10$	0.26 ± 0.03	0.33 ± 0.03

$$TPR = \frac{TP}{P}, \quad FPR = \frac{FP}{S}. \tag{81}$$

In terms of TPR both algorithms work perfectly well. However, PProx-PDA-IA gets lower FPR compare to the ADMM algorithm.

4.3 Partial consensus

The partial consensus optimization problem has been introduced in (10). As stated in the introduction, we are not aware of any existing algorithm that is able to perform nonconvex partial consensus optimization with guaranteed performance. Let us consider *regularized logistic regression* problem [4] in a network with N nodes, in

Table 3 Recovery results for PPox-PDA-IA and ADMM in terms of TPR and FPR

Parameters	TPR		FPR	
	PProx-PDA-IA	ADMM	PProx-PDA-IA	ADMM
$n = 30,\ p = 128,\ k = 1,\ s = 5$	1.0 ± 0.0	1.0 ± 0.0	0.00 ± 0.00	0.00 ± 0.00
$n = 80,\ p = 128,\ k = 1,\ s = 5$	1.0 ± 0.0	1.0 ± 0.0	0.00 ± 0.00	0.00 ± 0.00
$n = 120,\ p = 128,\ k = 1,\ s = 5$	1.0 ± 0.0	1.0 ± 0.0	0.00 ± 0.00	0.00 ± 0.00
$n = 150,\ p = 200,\ k = 1,\ s = 5$	1.0 ± 0.0	1.0 ± 0.0	0.00 ± 0.00	0.00 ± 0.00
$n = 80,\ p = 128,\ k = 1,\ s = 10$	1.0 ± 0.0	1.0 ± 0.0	0.00 ± 0.00	0.00 ± 0.00
$n = 80,\ p = 128,\ k = 5,\ s = 10$	1.0 ± 0.0	1.0 ± 0.0	0.53 ± 0.03	0.56 ± 0.04
$n = 128,\ p = 128,\ k = 5,\ s = 10$	1.0 ± 0.0	1.0 ± 0.0	0.57 ± 0.01	0.59 ± 0.02
$n = 70,\ p = 128,\ k = 5,\ s = 10$	1.0 ± 0.0	1.0 ± 0.0	0.53 ± 0.05	0.54 ± 0.01

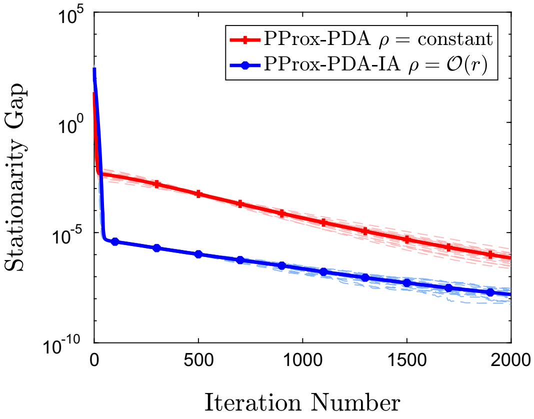

Fig. 7 The stationarity-gap achieved by the proposed methods for the partial consensus problem. The solid lines and dotted lines represent the single performance and the average performance, respectively

mini-batch setup i.e. each node stores b (batch size) data points, and each component function is given by

$$f_i(x_i) = \frac{1}{Nb}\left[\sum_{j=1}^{b} \log(1 + \exp(-y_{ij} x_i^T v_{ij})) + \sum_{k=1}^{M} \frac{\hat{\beta}\hat{\alpha} x_{i,k}^2}{1 + \hat{\alpha} x_{i,k}^2}\right],$$

where $v_{ij} \in \mathbb{R}^M$ and $y_{ij} \in \{1, -1\}$ are the feature vector and the label for the jth date point in i-th agent, $\hat{\alpha}$ and $\hat{\beta}$ are the regularization parameters [4].

We set $N = 20$, $M = 10$, $b = 100$, $\hat{\beta} = 0.01$, $\hat{\alpha} = 1$, and $\xi = 0.001$. The graph $\mathcal{G}$ is generated similar to the problem in Sect. 4.1. The PProx-PDA and PProx-PDA-IA algorithms are implemented for the above problem. Both algorithms stop after 1000 iterations, and we measure the averaged performance over 20 trials, where in each trial the data matrix and the initial solutions are generated randomly independent. In Fig. 7 the stationarity-gap for the problem has been plotted. It can be observed that the gap is vanishing as the algorithm proceeds, and it appears that PProx-PDA-IA is faster than PProx-PDA. Figure 8 displays the constraint violation for the

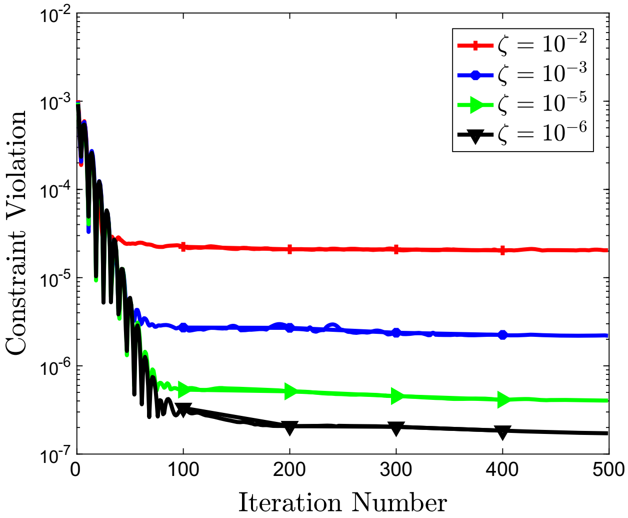

Fig. 8 Constraint violation $\|Ax\|$ achieved by the proposed method for the partial consensus problem with different permissible tolerance ξ

PProx-PDA algorithm with different tolerance ξ. It is also interesting to observe that when reducing the constraint violation error (represented by $\xi > 0$), the resulting solution indeed achieves higher degrees of consensus.

5 Conclusion

In this paper, we proposed a class of perturbed primal–dual based algorithms for optimizing nonconvex and linearly constrained problems. The proposed methods are of Uzawa type, in which a primal gradient descent step is performed followed by an (approximate) dual gradient ascent step. We performed theoretical convergence analysis and tested their performance on a number of statistical and engineering applications. In the future, we plan to investigate, both in theory and in practice, whether the perturbation is necessary for primal–dual type algorithms to reach stationary solutions. Further, we plan to extend the proposed algorithms to problems with stochastic objective functions.

Acknowledgements The authors would like to thank Dr. Quanquan Gu who provided us with the codes of [29]. The authors would also like to thanks Dr. Gesualdo Scutari for helpful discussions about the numerical results.

Appendix A

In this section, we justify Assumption [B4], which imposes the boundedness of the sequence of dual variables. Throughout this section we will assume that Assumptions A and [B1]–[B3] hold. First, we prove that when $\|\lambda^{t+1}\| \to \infty$, we have

$\liminf_{r\to\infty} \frac{\beta^{r+1}\|x^{r+1}-x^r\|}{\|\lambda^{r+1}\|} = 0$. Using Assumption [B3] we have the following identity

$$\frac{\beta^{r+1}\rho^{r+1}}{2}\|x^{r+1}-x^r\|^2 = \frac{(\beta^{r+1})^2}{2c_0}\|x^{r+1}-x^r\|^2. \tag{82}$$

Assume the contrary, that there exists $c_1 > 0$ such that

$$\lim_{r\to\infty} \beta^{r+1}\|x^{r+1}-x^r\|^2 \ge \frac{c_1}{\beta^{r+1}}\|\lambda^{r+1}\|^2. \tag{83}$$

Then from (74), it is easy to show that when r is large enough, P is decreasing. Similarly as in Lemma 3, it is relatively easy to show that the potential function is lower and upper bounded (the proof is included in Lemma 10–11 in the online version). The lower boundedness of the potential function and the fact that it is descending, implies that (75) holds true, which further implies that $\frac{1}{\beta^{r+1}}\|\lambda^{r+1}\|^2 \to 0$, according to (83). Examine the definition of the potential function in (73) and use the choice of c in (70) we conclude that $\frac{\beta^{r+1}\rho^{r+1}}{2}\|x^{r+1}-x^r\|^2$ in the potential function is bounded. Therefore, there exists D_1 such that

$$\beta^{r+1}\|x^{r+1}-x^r\| \le D_1. \tag{84}$$

It follows that $c_1\|\lambda^{r+1}\|^2$ is also upper bounded. This contradicts our assumption that $\|\lambda^r\| \to \infty$.

Next, we make use of some constraint qualification to argue the boundedness of the dual variables. The technique used in the proof is relatively standard, see recent works [21,51]. Assume that the so-called Robinson's condition is satisfied for problem (1) at $\hat{x}$ [62, Chap. 3]. This means $\{Ad_x \mid d_x \in \mathcal{T}_X(\hat{x})\} = \mathbb{R}^M$, where d_x is the tangent direction for convex set X, and $\mathcal{T}_X(\hat{x})$ is the tangent cone to the feasible set X at the point $\hat{x}$. Utilizing this assumption we will prove that the dual variable is bounded.

Lemma 6 *Suppose the Robinson's condition holds true for problem (1). Then the sequence of dual variable $\{\lambda^r\}$ generated by (67b) is bounded.*

Proof We argue by contradiction. Suppose that the dual variable sequence is not bounded, i.e.,

$$\|\lambda^r\| \to \infty. \tag{85}$$

From the optimality condition of x^{r+1} we have for all $x \in X$

$$\langle \nabla f(x^r) + \xi^{r+1} + A^T\lambda^{r+1} + \beta^{r+1}B^TB(x^{r+1}-x^r), x - x^{r+1}\rangle \ge 0.$$

Note that $\liminf_{r\to\infty} \frac{\beta^{r+1}\|x^{r+1}-x^r\|}{\|\lambda^{r+1}\|} = 0$, so the following holds:

$$\liminf_{r\to\infty} \frac{\beta^{r+1}\|B^TB(x^{r+1}-x^r)\|}{\|\lambda^{r+1}\|} = 0.$$

Let us define a new *bounded* sequence as $\mu^r = \lambda^r/\|\lambda^r\|, r = 1, 2, \ldots$. Let (x^*, μ^*) be an accumulation point of $\{x^{r+1}, \mu^{r+1}\}$. Assume that the Robinson's condition holds at x^*. Dividing both sides of the above inequality by $\|\lambda^{r+1}\|$ we obtain for all $x \in X$

$$\begin{aligned}&\langle \nabla f(x^r)/\|\lambda^{r+1}\| + \xi^{r+1}/\|\lambda^{r+1}\| + A^T\mu^{r+1}\\ &\quad + \beta^{r+1}B^TB(x^{r+1} - x^r)/\|\lambda^{r+1}\|, x - x^{r+1}\rangle \geq 0.\end{aligned}$$

Taking the limit, passing a subsequence if necessary and utilizing the assumption that $\|\lambda^{r+1}\| \to \infty$, and that X is a compact set, we obtain

$$\langle A^T\mu^*, x - x^*\rangle \geq 0, \ \forall\, x \in X.$$

Utilizing the Robinson's condition, we know that there exists $x \in X$ and a scaling constant $c > 0$ that such $cA(x - x^*) = -\mu^*$, which combined with the above relation yields: $-c\|\mu^*\|^2 \leq 0$. Therefore we must have $\mu^* = 0$. However, this contradicts to the fact that $\|\mu^*\| = 1$. Therefore, we conclude that $\{\lambda^r\}$ is a bounded sequence. □

Appendix B

We show how the sufficient conditions developed in "Appendix A" can be applied to the problems discussed in Sect. 1.2. We will focus on the partial consensus problem (10).

To proceed, we note that the Robinson's condition reduces to the well-known Mangasarian–Fromovitz constraint qualification (MFCQ) if we set $X = \mathbb{R}^N$, and write out explicitly the inequality constraints as $g(x) \leq 0$ [62, Lemma 3.16]. To state the MFCQ, consider the following system

$$\begin{aligned}&p_i(y) = 0, \ i = 1, \ldots, M\\ &g_j(y) \leq 0, \ j = 1, \ldots, P\end{aligned} \tag{86}$$

where $p_i : \mathbb{R}^N \to \mathbb{R}$ and $g_j : \mathbb{R}^N \to \mathbb{R}$ are all continuously differentiable functions. For a given feasible solution $\hat{y}$ let us use $\mathcal{A}(\hat{y})$ to denote the indices for active inequality constraints, that is

$$\mathcal{A}(\hat{y}) := \{1 \leq j \leq P \ \mid \ g_j(\hat{y}) = 0\}. \tag{87}$$

Let us define

$$p(y) := [p_1(y); p_2(y); \cdots; p_M(y)], \quad g(y) := [g_1(y); g_2(y); \cdots; g_P(y)].$$

Then the MFCQ holds for system (86) at point $\hat{y}$ if we have: 1) The rows of Jacobian matrix of $p(y)$ denoted by $\nabla p(\hat{y})$ are linearly independent. 2) There exists a vector $d_y \in \mathbb{R}^N$ such that

$$\nabla p(\hat{y})d_y = 0, \quad \nabla g_j(\hat{y})^T d_y < 0, \ \forall\, j \in \mathcal{A}(\hat{y}). \tag{88}$$

See [62, Lemma 3.17] for more details. In the following, we show that MFCQ holds true for problem (10) at any point (x, z) that satisfies $z \in Z$. Comparing the constraint set of this problem with system (86) we have the following specifications. The optimization variable $y = [x; z]$, where $x \in \mathbb{R}^N$ stacks all $x_i \in \mathbb{R}$ from N nodes (here we assume $x_i \in \mathbb{R}$ only for the ease of presentation). Also, $z \in \mathbb{R}^E$ stacks all $z_e \in \mathbb{R}$ for $e \in \mathcal{E}$. The equality constraint is written as $p(y) = [A, -I]y = 0$, where $A \in \mathbb{R}^{E\times N}$ and I is an $E \times E$ identity matrix. Finally, for the inequality constraint we have $g_e(y) = |z_e| - \xi$, and the active set is given by $\mathcal{A}(y) := \mathcal{A}^+(y) \cup \mathcal{A}^-(y)$, where

$$\mathcal{A}^+(y) = \{e \in \mathcal{E} \mid z_e = \xi\}, \quad \mathcal{A}^-(y) = \{e \in \mathcal{E} \mid z_e = -\xi\}.$$

Without loss of generality we assume $\xi = 1$. To show that MFCQ holds, consider a solution $\hat{y} := (\hat{x}, \hat{z})$. First observe that the Jacobian of equality constraint is $\nabla p(\hat{y}) = [A, -I]$ which has full row rank. In order to verify the second condition we need to find a vector $d_y := [d_x; d_z] \in \mathbb{R}^{N+E}$ such that

$$Ad_x = d_z, \tag{89a}$$

$$[d_z]_e < 0 \quad \text{for } e \in \mathcal{A}^+(\hat{y}) \tag{89b}$$

$$[d_z]_e > 0 \quad \text{for } e \in \mathcal{A}^-(\hat{y}) \tag{89c}$$

where $[d_z]_e$ denotes the eth component of vector d_z. Let us denote an all-one vector and all-zero vector by $\mathbf{1}$ and $\mathbf{0}$ respectively. To proceed, let us consider two different cases:

Case 1 For the vector $\hat{z} \in \mathbb{R}^E$ we have $\hat{z} \neq \mathbf{1}$ and $\hat{z} \neq -\mathbf{1}$. Let us take

$$d_z = \frac{1}{E}(\hat{z}^T\mathbf{1})\mathbf{1} - \hat{z}.$$

First we can show that $d_z \in \text{col}(A)$. Note that for our problem when the graph is *connected*, the only null space of A (which is the incidence of the graph) is spanned by the vector $\mathbf{1}$ [16]. Using this fact, we have $\mathbf{1}^T d_z = \hat{z}^T\mathbf{1} - \mathbf{1}^T\hat{z} = 0$, therefore, $Ad_x = d_z$ holds true. Second, for $e \in \mathcal{A}^+(\hat{y})$ we have that $\hat{z}_e = 1$. Therefore, we can check that $[d_z]_e = \left[\frac{1}{E}(\hat{z}^T\mathbf{1})\mathbf{1} - \hat{z}\right]_e < 0$, because $\frac{1}{E}(\hat{z}^T\mathbf{1})\mathbf{1} < 1$ from the fact that $\hat{z} \neq \mathbf{1}$. Condition (89b) is verified. Using similar argument we can verify condition (89c).

Case 2 Suppose we have $\hat{z} = \mathbf{1}$ (resp. $\hat{z} = -\mathbf{1}$). Since $\hat{z} \in \text{null}(A)$ let us set $d_x = \mathbf{0}$ and $d_z = -\hat{z}$ (resp. $d_z = \hat{z}$). First we have $Ad_x = d_z$. Second, for $e \in \mathcal{A}^+(\hat{y})$ we have that $[d_z]_e < 0$. Similarly, we have $[d_z]_e > 0$ for $e \in \mathcal{A}^-(\hat{y})$. All conditions (89a)–(89c) are verified. The above proof shows that MFCQ holds true for the sequence $\{(x^r, z^r)\}$ generated by the PProx-PDA algorithm, since in the algorithm it is always guaranteed that $z^r \in Z$.

References

1. Allen-Zhu, Z., Hazan, E.: Variance reduction for faster non-convex optimization. In: Proceedings of the 33rd International Conference on Machine Learning, ICML, pp. 699–707 (2016)
2. Ames, B., Hong, M.: Alternating directions method of multipliers for l1-penalized zero variance discriminant analysis and principal component analysis. Comput. Optim. Appl. **64**(3), 725–754 (2016)
3. Andreani, R., Haeser, G., Martnez, J.M.: On sequential optimality conditions for smooth constrained optimization. Optimization **60**(5), 627–641 (2011)
4. Antoniadis, A., Gijbels, I., Nikolova, M.: Penalized likelihood regression for generalized linear models with non-quadratic penalties. Ann. Inst. Stat. Math. **63**(3), 585–615 (2009)
5. Arrow, K.J., Hurwicz, L., Uzawa, H.: Studies in Linear and Non-linear Programming. Stanford University Press, Palo Alto (1958)
6. Asteris, M., Papailiopoulos, D., Dimakis, A.: Nonnegative sparse PCA with provable guarantees. In: Proceedings of the 31st International Conference on Machine Learning (ICML), vol. 32, pp. 1728–1736 (2014)
7. Aybat, N.S., Hamedani, E.Y.: A primal–dual method for conic constrained distributed optimization problems. Adv. Neural Inf. Process. Syst. (NIPS) 5049–5057 (2016)
8. Bertsekas, D.P.: Constrained Optimization and Lagrange Multiplier Method. Academic Press, Cambridge (1982)
9. Bertsekas, D.P., Tsitsiklis, J.N.: Neuro-Dynamic Programming. Athena Scientific, Belmont (1996)
10. Bertsekas, D.P., Tsitsiklis, J.N.: Parallel and Distributed Computation: Numerical Methods, 2nd edn. Athena Scientific, Belmont (1997)
11. Bianchi, P., Jakubowicz, J.: Convergence of a multi-agent projected stochastic gradient algorithm for non-convex optimization. IEEE Trans. Autom. Control **58**(2), 391–405 (2013)
12. Birgin, E., Martínez, J.: Practical Augmented Lagrangian Methods for Constrained Optimization. Society for Industrial and Applied Mathematics, Philadelphia (2014)
13. Björnson, E., Jorswieck, E.: Optimal resource allocation in coordinated multi-cell systems. Found. Trends Commun. Inf. Theory **9**, 113–381 (2013)
14. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Found. Trends Mach. Learn. **3**(1), 1–122 (2011)
15. Burachik, R.S., Kaya, C.Y., Mammadov, M.: An inexact modified subgradient algorithm for nonconvex optimization. Comput. Optim. Appl. **45**(1), 1–24 (2008)
16. Chung, F.R.K.: Spectral Graph Theory. The American Mathematical Society, Providence (1997)
17. Cressie, N.: Statistics for Spatial Data. Wiley, Hoboken (2015)
18. Curtis, F.E., Gould, N.I.M., Jiang, H., Robinson, D.P.: Adaptive augmented Lagrangian methods: algorithms and practical numerical experience. Optim. Methods Softw. **31**(1), 157–186 (2016)
19. D'Aspremont, A., Ghaoui, L.E., Jordan, M.I., Lanckriet, G.R.G.: A direct formulation for sparse PCA using semidefinite programming. SIAM Rev. **49**(3), 434–448 (2007)
20. Deng, W., Yin, W.: On the global and linear convergence of the generalized alternating direction method of multipliers. J. Sci. Comput. **66**(3), 889–916 (2016)
21. Dutta, J., Deb, K., Tulshyan, R., Arora, R.: Approximate KKT points and a proximity measure for termination. J. Glob. Optim. **56**(4), 1463–1499 (2013)
22. Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. **96**(456), 1348–1360 (2001)
23. Fernández, D., Solodov, M.V.: Local convergence of exact and inexact augmented Lagrangian methods under the second-order sufficient optimality condition. SIAM J. Optim. **22**(2), 384–407 (2012)
24. Fleiss, J.L., Levin, B., Paik, M.C.: Statistical Methods for Rates and Proportions. Wiley, Hoboken (2003)
25. Forero, P.A., Cano, A., Giannakis, G.B.: Distributed clustering using wireless sensor networks. IEEE J. Sel. Top. Signal Proces. **5**(4), 707–724 (2011)
26. Gabay, D., Mercier, B.: A dual algorithm for the solution of nonlinear variational problems via finite element approximation. Comput. Math. Appl. **2**, 17–40 (1976)
27. Giannakis, G.B., Ling, Q., Mateos, G., Schizas, I.D., Zhu, H.: Decentralized learning for wireless communications and networking. In: Splitting Methods in Communication and Imaging. Springer, New York (2015)

28. Glowinski, R., Marroco, A.: Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problémes de dirichlet non linéares. Revue Franqaise d'Automatique, Informatique et Recherche Opirationelle **9**, 41–76 (1975)
29. Gu, Q., Z. Wang, Z., Liu, H.: Sparse PCA with oracle property. In: Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS), pp. 1529–1537 (2014)
30. Haeser, G., Melo, V.: On sequential optimality conditions for smooth constrained optimization. Preprint (2013)
31. Hajinezhad, D., Chang, T.H., Wang, X., Shi, Q., Hong, M.: Nonnegative matrix factorization using ADMM: algorithm and convergence analysis. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4742–4746 (2016)
32. Hajinezhad, D., Hong, M.: Nonconvex alternating direction method of multipliers for distributed sparse principal component analysis. In: IEEE Global Conference on Signal and Information Processing (GlobalSIP). IEEE (2015)
33. Hajinezhad, D., Hong, M., Garcia, A.: Zeroth order nonconvex multi-agent optimization over networks. arXiv preprint arXiv:1710.09997 (2017)
34. Hajinezhad, D., Hong, M., Zhao, T., Wang, Z.: NESTT: A nonconvex primal–dual splitting method for distributed and stochastic optimization. In: Advances in Neural Information Processing Systems (NIPS), pp. 3215–3223 (2016)
35. Hajinezhad, D., Shi, Q.: Alternating direction method of multipliers for a class of nonconvex bilinear optimization: convergence analysis and applications. J. Glob. Optim. **70**, 1–28 (2018)
36. Hamdi, A., Mishra, S.K.: Decomposition Methods Based on Augmented Lagrangians: A Survey, pp. 175–203. Springer, New York (2011)
37. Hestenes, M.R.: Multiplier and gradient methods. J. Optim. Appl. **4**, 303–320 (1969)
38. Hong, M., Hajinezhad, D., Zhao, M.M.: Prox-PDA: the proximal primal-dual algorithm for fast distributed nonconvex optimization and learning over networks. In: Proceedings of the 34th International Conference on Machine Learning (ICML), (70), pp. 1529–1538 (2017)
39. Hong, M., Luo, Z.Q.: On the linear convergence of the alternating direction method of multipliers. Math. Program. **162**(1), 165–199 (2017)
40. Hong, M., Luo, Z.Q., Razaviyayn, M.: Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems. SIAM J. Optim. **26**(1), 337–364 (2016)
41. Houska, B., Frasch, J., Diehl, M.: An augmented Lagrangian based algorithm for distributed nonconvex optimization. SIAM J. Optim. **26**(2), 1101–1127 (2016)
42. Koppel, A., Sadler, B.M., Ribeiro, A.: Proximity without consensus in online multi-agent optimization. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3726–3730 (2016)
43. Koshal, J., Nedić, A., Shanbhag, Y.V.: Multiuser optimization: distributed algorithms and error analysis. SIAM J. Optim. **21**(3), 1046–1081 (2011)
44. Lan, G., Monteiro, R.D.C.: Iteration-complexity of first-order augmented Lagrangian methods for convex programming. Math. Program. **155**(1), 511–547 (2015)
45. Li, G., Pong, T.K.: Global convergence of splitting methods for nonconvex composite optimization. SIAM J. Optim. **25**(4), 2434–2460 (2015)
46. Liao, W., Hong, M., Farmanbar, H., Luo, Z.: Semi-asynchronous routing for large scale hierarchical networks. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2894–2898 (2015)
47. Liavas, A.P., Sidiropoulos, N.D.: Parallel algorithms for constrained tensor factorization via alternating direction method of multipliers. IEEE Trans. Signal Process. **63**(20), 5450–5463 (2015)
48. Liu, Y.F., Liu, X., Ma, S.: On the non-ergodic convergence rate of an inexact augmented Lagrangian framework for composite convex programming. arXiv preprint arXiv:1603.05738 (2016)
49. Lobel, I., Ozdaglar, A.: Distributed subgradient methods for convex optimization over random networks. IEEE Trans. Autom. Control **56**(6), 1291–1306 (2011)
50. Lorenzo, P.D., Scutari, G.: NEXT: in-network nonconvex optimization. IEEE Trans. Signal Inf. Process Over Netw. **2**(2), 120–136 (2016)
51. Lu, Z., Zhang, Y.: Sparse approximation via penalty decomposition methods. SIAM J. Optim. **23**(4), 2448–2478 (2013)
52. Mateos, G., Bazerque, J.A., Giannakis, G.B.: Distributed sparse linear regression. IEEE Trans. Signal Process. **58**(10), 5262–5276 (2010)

53. Max L.N. Goncalves, J.G.M., Monteiro, R.D.: Convergence rate bounds for a proximal ADMM with over-relaxation stepsize parameter for solving nonconvex linearly constrained problems (2017). Preprint arXiv:1702.01850
54. Nedić, A., Olshevsky, A.: Distributed optimization over time-varying directed graphs. IEEE Trans. Autom. Control **60**(3), 601–615 (2015)
55. Nedić, A., Ozdaglar, A.: Distributed subgradient methods for multi-agent optimization. IEEE Trans. Autom. Control **54**(1), 48–61 (2009)
56. Nedić, A., Ozdaglar, A., Parrilo, P.A.: Constrained consensus and optimization in multi-agent networks. IEEE Trans. Autom. Control **55**(4), 922–938 (2010)
57. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course. Springer, Berlin (2004)
58. Nocedal, J., Wright, S.J.: Numerical Optimization. Springer, Berlin (1999)
59. Powell, M.M.D.: An efficient method for nonlinear constraints in minimization problems. In: Optimization. Academic Press, pp. 283–298 (1969)
60. Razaviyayn, M., Hong, M., Luo, Z.Q.: A unified convergence analysis of block successive minimization methods for nonsmooth optimization. SIAM J. Optim. **23**(2), 1126–1153 (2013)
61. Rockafellar, R.T.: Augmented Lagrangians and applications of the proximal point algorithm in convex programming. Math. Oper. Res. **1**(2), 97–116 (1976)
62. Ruszczyński, A.: Nonlinear Optimization. Princeton University, Princeton (2011)
63. Schizas, I., Ribeiro, A., Giannakis, G.: Consensus in ad hoc WSNs with noisy links—part I: distributed estimation of deterministic signals. IEEE Trans. Signal Process. **56**(1), 350–364 (2008)
64. Scutari, G., Facchinei, F., Song, P., Palomar, D.P., Pang, J.S.: Decomposition by partial linearization: parallel optimization of multi-agent systems. IEEE Trans. Signal Process. **63**(3), 641–656 (2014)
65. Shi, W., Ling, Q., Wu, G., Yin, W.: EXTRA: an exact first-order algorithm for decentralized consensus optimization. SIAM J. Optim. **25**(2), 944–966 (2014)
66. Sun, Y., Scutari, G., Palomar, D.: Distributed nonconvex multiagent optimization over time-varying networks. In: 50th Asilomar Conference on Signals, Systems and Computers, pp. 788–794 (2016)
67. Tsitsiklis, J.: Problems in decentralized decision making and computation. Ph.D. thesis, Massachusetts Institute of Technology (1984)
68. Vu, V.Q., Cho, J., Lei, J., Rohe, K.: Fantope projection and selection: a near-optimal convex relaxation of sparse PCA. In: Advances in Neural Information Processing Systems (NIPS), pp. 2670–2678 (2013)
69. Wang, Y., Yin, W., Zeng, J.: Global convergence of ADMM in nonconvex nonsmooth optimization. J. Sci. Comput. **78**(1), 29–63 (2019)
70. Wright, S.J.: Implementing proximal point methods for linear programming. J. Optim. Theory Appl. **65**(3), 531–554 (1990)
71. Wen, Z., Yang, C., Liu, X., Marchesini, S.: Alternating direction methods for classical and ptychographic phase retrieval. Inverse Probl. **28**(11), 1–18 (2012)
72. Yildiz, M.E., Scaglione, A.: Coding with side information for rate-constrained consensus. IEEE Trans. Signal Process. **56**(8), 3753–3764 (2008)
73. Zhang, C.H.: Nearly unbiased variable selection under minimax concave penalty. Ann. Stat. **38**(2), 894–942 (2010)
74. Zhang, Y.: Convergence of a class of stationary iterative methods for saddle point problems. Preprint (2010)
75. Zhu, H., Cano, A., Giannakis, G.: Distributed consensus-based demodulation: algorithms and error analysis. IEEE Trans. Wirel. Commun. **9**(6), 2044–2054 (2010)