Noname manuscript No.
(will be inserted by the editor)

Perturbed Proximal Primal Dual Algorithm for Nonconvex Nonsmooth Optimization

Davood Hajinezhad and Mingyi Hong

Received: date / Accepted: date

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 has applications in a number of statistical and engineering problems, for example in high-dimensional subspace estimation, and distributed signal processing and learning over networks. The proposed method is of 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.
AMS(MOS) Subject Classifications: 49, 90.

1 Introduction

1.1 The Problem

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 subsection 1.3.

1.2 The Algorithm

In this section, we present the proposed algorithm. The augmented Lagrangian for problem (1) is given below

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

D. Hajinezhad, Department of Mechannical Engineering and Matrials Science, Duke University, Durham, NC, USA. E-mail: davood.hajinezhad@duke.edu
M. Hong, Department of Electrical and Computer Engineering, University of Minnesota, USA. E-mail: mhong@umn.edu

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 algorithm parameter that is related to the size of the equality constraint violation that is allowed by the algorithm, and 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: λ^0 and x^0

Repeat: update variables by

$$x^{r+1} = \arg\min_{x \in X} \Big\{ \langle \nabla f(x^r), x - x^r \rangle + h(x) + \langle (1 - \rho\gamma)\lambda^r, Ax - b \rangle + \frac{\rho}{2}\|Ax - b\|^2 + \frac{\beta}{2}\|x - x^r\|^2_{B^T B} \Big\} \tag{3a}$$

$$\lambda^{r+1} = (1 - \rho\gamma)\lambda^r + \rho\left(Ax^{r+1} - b\right) \tag{3b}$$

Until Convergence.

In contrast to the classic Augmented Lagrangian (AL) method [34, 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 $\frac{\beta}{2}\|x - x^r\|^2_{B^T B}$. 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 consistent 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 [28,30,32,37,45,53,69]. 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 from the dual perturbation scheme developed for the convex case, for example in [43]. Conceptually, the perturbed dual step can be viewed as performing a dual ascent on certain *regularized Lagrangian* in the dual space; see [42, Sec. 3.1]. The main purpose for introducing the dual perturbation/regularization in this reference, and in many related works, is to ensure that the dual update is well-behaved and easy to analyze. Intuitively, when adopting and modifying such a perturbation strategy in the non-convex setting of interest to this work, we no longer need to bound the size of the successive dual variables with that of the successive primal variables, since the change of the dual variables is now well controlled.

Third, the proximal term $\frac{\beta}{2}\|x - x^r\|^2_{B^T B}$ 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.

1.3 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 $u_1, u_2, \cdots, u_p$ 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 $u_1, u_2, \cdots, u_k$, 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 [26]

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

In the above formulation, $\mathcal{F}^k$ denotes the Fantope set [68], given by $\mathcal{F}^k = \{X : \ 0 \preceq X \preceq I, \mathrm{trace}(X) = k\}$, 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) [20], and minimax concave penalty (MCP) [72]. For example, MCP with parameters b and ν for some scaler ϕ is given below

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

where, ι_X denotes the indicator function for 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 $\nu \geq 0$. In a recent work [26], it is shown that with high probability, every first-order stationary solution of problem (4) (denoted as $\hat{\Pi}$) is of high-quality. See [26, 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]

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

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 [11]. 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 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, 74],

distributed and parallel machine learning [22, 35] and distributed signal processing [63, 74]; for more applications we refer the readers to a recent survey [24].

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 original problem (1) with 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

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

where $D := \text{diag}[d_1, \cdots, d_N] \in \mathbb{R}^{N\times N}$ 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 $i = 1, 2, \cdots, N$. 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 the matrix L^+ ensures that the ith block vector of $L^+ x^r$ is only related to x_j^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 [15]. 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 $-\xi \le [z_e]_k \le \xi$, 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

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

the partial consensus problem can be formulated as

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

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

1.4 **Literature Review and Contribution.**

1.4.1 Literature on Related Algorithms.

The Augmented Lagrangian (AL) method, also known as the methods of multipliers, is pioneered by Hestenes [34] and Powell [59]. It is a classical algorithm for solving nonconvex smooth constrained problems and its convergence is guaranteed under rather week assumptions [7, 21, 58]. A modified version of AL has been develped 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 [40] 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 [16] 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 [13] has developed an AL based algorithm for nonconvex nonsmooth optimization, where subgradients 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 [35] 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 (9) with $h_i \equiv 0$ and $X \in \mathbb{R}^N$; however, the consensus constraint $Ax = 0$ is always imposed]. Another AL based algorithm, which is called ALADIN [38], is designed for nonconvex smooth optimization problem with coupled affine constraints in 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 [33] 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 [12]. The method originates in early 1970s [23, 25], and has since been studied extensively [9, 18, 36]. 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 [70], distributed clustering [22], 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 nonocnvex problems. Zhang [73] 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 high-dimensional discriminant analysis. Other works along this line include [31, 37, 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 convex problem such application is popular, and the resulting algorithms are efficient.

1.4.2 Literature on Applications.

The sparse subspace estimation problem formulations (4) and (7) have been first considered in [17, 68] and subsequently considered in [26]. 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 [26] 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 [26] 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 [6, 49, 54, 55, 65]. Without assuming convexity of f_i's, the literature has been very scant; see recent developments in [10, 29, 37, 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.

1.4.3 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 Uzawa type [41] and it has 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 Section 1.3,

One distinctive feature of the PProx-PDA is the use of a novel perturbation scheme for both the primal and dual steps, 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 sublinearly to the set of approximate first-order stationary solutions. Further, when the perturbation parameter diminishes to zero with 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 $\|x\|_H^2 := x^T H x$. 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 subdifferential set defined by

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

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

$$\text{prox}_h^{1/\alpha}(x) := \underset{z}{\text{argmin}} \left\{ \frac{1}{2\alpha}\|x - z\|^2 + h(z) \right\}. \tag{12}$$

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).
Assumptions 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 subgradient 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 $A^TA + B^TB \succeq I$.

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*

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

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

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

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

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

where in the first inequality we used equation (3b) to obtain

$$\lambda^{r+1} - \lambda^r = A^T\lambda^r(1-\rho\gamma) + \rho A^T(Ax^{r+1} - b) + A^T\lambda^{r-1}(1-\rho\gamma) + \rho A^T(Ax^r - b);$$

In the last inequality we have utilized the convexity of h. Now let us analyze each term in (20). For the first term we have the following:

$$\begin{aligned}\langle \nabla f(x^{r-1}) - \nabla f(x^r), x^{r+1} - x^r\rangle &\leq \frac{1}{2\alpha}\|\nabla f(x^{r-1}) - \nabla f(x^r)\|^2 + \frac{\alpha}{2}\|x^{r+1} - x^r\|^2 \\ &\leq \frac{L^2}{2\alpha}\|x^r - x^{r-1}\|^2 + \frac{\alpha}{2}\|x^{r+1} - x^r\|^2 \\ &= \frac{L}{2}\|x^r - x^{r-1}\|^2 + \frac{L}{2}\|x^{r+1} - x^r\|^2, \end{aligned} \tag{21}$$

where in the first inequality we applied Young's inequality for $\alpha > 0$, the second inequality is due to the Lipschitz continuity of the gradient of function f, and in the last equality we set $\alpha = L$.

For the second term in (20) which is $\langle A^T(\lambda^{r+1} - \lambda^r), x^{r+1} - x^r\rangle$, we have the following series of equalities

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

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

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

Therefore, combining (21) – (23), we obtain the desired result in (16). **Q.E.D.**

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 function f is not convex, below we prove that $T(x, \lambda) + \frac{\beta}{2}\|x - x^r\|^2_{B^TB}$ is strongly convex with modulus $\beta - L$ when $\rho \geq \beta$, and $\beta \geq L$. First let us define $g(x, \lambda; x^r) = T(x, \lambda) - h(x) + \frac{\beta}{2}\|x - x^r\|^2_{B^TB}$, which is a smooth function. For this function we have

$$\begin{aligned}&\nabla g(x, \lambda; x^r) - \nabla g(y, \lambda; x^r) \\ &= \langle \nabla f(x) - \nabla f(y) + \rho A^TA(x - y) + \beta B^TB(x - y), x - y\rangle \\ &\geq \langle \nabla f(x) - \nabla f(y), x - y\rangle + \beta(A^TA + B^TB)\|x - y\|^2 \\ &\geq -L\|x - y\|^2 + \beta(A^TA + B^TB)\|x - y\|^2 \\ &\geq (\beta - L)\|x - y\|^2, \end{aligned} \tag{25}$$

where the first inequality is true because $\rho \geq \beta$, in the second inequality we used the Lipschitz continuity of ∇f, and the last inequality is true because we assumed that $A^TA + B^TB \succeq I$. This proves that function $g(x, \lambda; x^r)$ is strongly convex with modulus $\beta - L$ when $\beta \geq L$. Since $h(x)$ is assumed to be convex we immediately conclude that $T(x, \lambda) + \frac{\beta}{2}\|x - x^r\|^2_{B^TB}$ is 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 if $\beta \geq 3L$, then the change of x results in the reduction of T:

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

where (i) is true because from (25) we know that when $\beta > 3L$, $\rho \geq \beta$ 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 $f(x)$ is gradient Lipschitz continuous . Second, let us analyze $T(x^{r+1},\lambda^{r+1}) - T(x^{r+1},\lambda^r)$ as the following

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

Combining the previous two steps, we obtain the desired inequality in (26). **Q.E.D.**

Comparing the results of Lemmas 1 and Lemma 2, from (16) we can observe that 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 $T(x^{r+1},\lambda^{r+1}) + \frac{(1-\rho\gamma)\gamma}{2}\|\lambda^{r+1}\|^2$ is behaving 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 $c > 0$

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

Then according to the previous two lemmas, one can conclude that there are constants a_1, a_2, such that

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

where $a_1 = \left((1-\rho\gamma)\frac{\gamma}{2} + c\gamma - \frac{1-\rho\gamma}{\rho}\right)$, and $a_2 = (\frac{\beta-3L}{2} - cL)$. Therefore, we can verify that in order to make the function P_c decrease, it is sufficient to ensure that

$$(1-\rho\gamma)\frac{\gamma}{2} + c\gamma - \frac{1-\rho\gamma}{\rho} > 0, \text{ and } \beta > (3+2c)L. \tag{31}$$

Therefore a sufficient condition is that

$$\tau := \rho\gamma \in (0,1), \quad c > \frac{1}{\tau} - 1 > 0, \quad \beta > (3+2c)L, \quad \rho \geq \beta. \tag{32}$$

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* (32)*. Then the following statement holds true*

$$\exists\ \underline{P} \quad \text{s.t. } P_c(x^{r+1}, \lambda^{r+1}; x^r, \lambda^r) \geq \underline{P} > -\infty, \quad \forall\ r \geq 0. \tag{33}$$

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

$$\begin{aligned}
&\langle \lambda^{r+1} - \rho\gamma\lambda^{r+1}, Ax^{r+1} - b - \gamma\lambda^{r+1}\rangle \\
&\overset{(13)}{=} \frac{1}{2}\left(\frac{1-\rho\gamma}{\rho} - (1-\rho\gamma)\gamma\right)(\|\lambda^{r+1}\|^2 - \|\lambda^r\|^2 + \|\lambda^{r+1} - \lambda^r\|^2) \\
&= \frac{(1-\rho\gamma)^2}{2\rho}(\|\lambda^{r+1}\|^2 - \|\lambda^r\|^2 + \|\lambda^{r+1} - \lambda^r\|^2).
\end{aligned} \tag{34}$$

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

$$\begin{aligned}
&\sum_{r=1}^R T(x^{r+1}, \lambda^{r+1}) = \sum_{r=1}^R \left(f(x^{r+1}) + h(x^{r+1}) + \frac{\rho}{2}\|Ax^{r+1} - b\|^2\right) \\
&+ \frac{(1-\rho\gamma)^2}{2\rho}(\|\lambda^{R+1}\|^2 - \|\lambda^1\|^2 + \sum_{r=1}^R \|\lambda^{r+1} - \lambda^r\|^2) \\
&\geq \sum_{r=1}^R \left(f(x^{r+1}) + h(x^{r+1}) + \frac{\rho}{2}\|Ax^{r+1} - b\|^2\right) + \frac{(1-\rho\gamma)^2}{2\rho}(\|\lambda^{R+1}\|^2 - \|\lambda^1\|^2) \\
&\geq -\frac{(1-\rho\gamma)^2}{2\rho}\|\lambda^1\|^2,
\end{aligned} \tag{35}$$

where the last inequality comes from the fact that f and h are both assumed to be lower bounded by 0. Since λ^1 is bounded, if follows that the sum of the $T(\cdot,\cdot)$ function is lower bounded. From (35) 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 nonincreasing we conclude that the potential function is lower bounded, that is we have

$$\underline{P} \geq -\frac{(1-\rho\gamma)^2}{2\rho}\|\lambda^1\|^2. \tag{36}$$

This proves the claim. **Q.E.D.**

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

Definition 1 Stationary solution. Consider problem (1). Given $\epsilon > 0$, the tuple (x^*, λ^*) is an ϵ-stationary solution if the following holds

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

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

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, 19, 27]. 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 Eq. (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* (32). *For any given $\epsilon > 0$, the following is true for the sequence (x^r, λ^r) generated by the PProx-PDA*

- *We have that $\{x^r\}$ 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 limit point of the sequence (x^r, λ^r). Then (x^*, λ^*) is a $(\gamma^2\|\lambda^*\|^2)$-stationary solution of problem* (1).

Proof. Using the fact that set X is a compact set we conclude 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 equation (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 limit point of the sequence (x^r, λ^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 $\|Ax^* - b\|^2 \leq \gamma^2\|\lambda^*\|^2$; which proves the first inequality in (37).

To show the boundedness of the sequences, first note that X has been assumed to be a convex and compact set, it follows that $\{x^r\}$ is bounded. To show the boundedness of $\{\lambda^r\}$, note that (3b) also suggests that

$$(1 - \rho\gamma)(\lambda^{r+1} - \lambda^r) = \rho(Ax^{r+1} - b) - \rho\gamma\lambda^{r+1}. \tag{40}$$

From the boundedness of x^{r+1} and $\lambda^{r+1} \to \lambda^r$, we conclude that λ^{r+1} is bounded.

To show the second part, from the optimality condition of (18) we have

$$\begin{aligned}&\langle \nabla f(x^r) + A^T\lambda^r(1 - \rho\gamma) + \rho A^T(Ax^{r+1} - b) + \beta B^T B(x^{r+1} - x^r), x^{r+1} - x\rangle \\ &\quad \leq \langle \xi^{r+1}, x - x^{r+1}\rangle, \ \forall\ x \in X.\end{aligned} \tag{41}$$

From the convexity of function h we have that for all $x \in X$ it holds that $\langle \xi^{r+1}, x - x^{r+1}\rangle \leq h(x) - h(x^{r+1})$. Plugging this inequality into (41), using the update equation (3b) and rearranging the terms we obtain

$$h(x^{r+1}) + \langle \nabla f(x^r) + A^T\lambda^{r+1} + \beta B^T B(x^{r+1} - x^r), x^{r+1} - x\rangle \leq h(x), \ \forall\ x \in X. \tag{42}$$

Further, adding and subtracting x^r in $x^{r+1}-x$, and bny utilizing relation (13), we can get

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

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

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

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

$$x^*=\arg\min_{x\in X}\ h(x)+\langle\nabla f(x^*),x-x^*\rangle+\langle\lambda^*,Ax\rangle+\frac{\beta}{2}\|B(x-x^*)\|^2. \tag{43}$$

The optimality of the above problem becomes

$$\langle\nabla f(x^*)+A^T\lambda^*+\xi^*,x^*-x\rangle\leq 0,\ \ \forall\ x\in X, \tag{44}$$

for some $\xi^*\in\partial h(x^*)$. **Q.E.D.**

2.1 The Choice of Perturbation Parameter

In this section, we discuss how to obtain ϵ-stationary solution. First, note that Theorem 1 indicates that if the sequence $\{\lambda^r\}$ is bounded, and 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. Such boundedness of λ^* can be ensured by assuming certain constraint qualification (CQ) at (x^*,λ^*); see a related discussion in the Appendix. In the rest of this section, we take an alternative approach to argue ϵ-stationary solution. Our general strategy is to let $\frac{1}{\rho}$ 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)

$$\begin{aligned}&d_1=\max\{\|Ax-b\|^2\mid x\in X\},\quad d_2=\max\{\|x-y\|^2\mid x,y\in X\},\\&d_3=\max\{\|x-y\|^2_{B^TB}\mid x,y\in X\},\quad d_4=\max\{f(x)+h(x)\mid x\in X\}.\end{aligned} \tag{45}$$

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

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

$$\frac{\rho}{2}\|Ax^1-b\|^2\leq d_4,\quad \frac{\beta}{2}\|x^1-x^0\|^2\leq d_4+\frac{3L}{2}d_2 \tag{46}$$

Proof. From Lemma 2, and use the choice of x^0 and λ^0, we obtain

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

Utilizing the definition of $T(x,\lambda)$ and (34), we obtain

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

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 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

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

This leads to the desired claim. **Q.E.D.**

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

$$\frac{1}{2\rho}\|\lambda^1\|^2 = \frac{\rho}{2}\|Ax^1 - b\|^2 \leq d_4. \tag{48}$$

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

$$\begin{aligned} P_c(x^1,\lambda^1;x^0,\lambda^0) &\overset{(29)}{=} T(x^1,\lambda^1) + \frac{(1-\rho\gamma)(\gamma + c/\rho)}{2}\|\lambda^1\|^2 \\ &\quad + \frac{c}{2}\left(\beta\|x^1 - x^0\|^2_{B^TB} + L\|x^1 - x^0\|^2\right) \\ &\overset{(24),(34)}{\leq} f(x^1) + h(x^1) + \frac{(1-\rho\gamma)^2}{\rho}\|\lambda^1\|^2 + \frac{\rho}{2}\|Ax^1 - b\|^2 \\ &\quad + \frac{(1-\rho\gamma)(\gamma + c/\rho)}{2}\|\lambda^1\|^2 + \frac{c}{2}\left(\beta\|x^1 - x^0\|^2_{B^TB} + L\|x^1 - x^0\|^2\right) \\ &\overset{(46)}{\leq} \left[2 + 2(1-\rho\gamma)^2 + (1-\rho\gamma)(c+\rho\gamma)\right]d_4 + \frac{c}{2}\left(2\sigma_{\max}(B^TB)(d_4 + \frac{3L}{2}d_2) + Ld_2\right) \\ &:= P_c^0 \end{aligned} \tag{49}$$

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 $\|\lambda^{r+1}\|^2$.

Lemma 5 *Suppose that (ρ,γ,β) are chosen according to (32), and the assumptions in Lemma 4 hold true. Then the following holds true for all $r \geq 0$*

$$\frac{\gamma(1-\rho\gamma)}{2}\|\lambda^r\|^2 \leq P_c^0. \tag{50}$$

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

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

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

$$P_c(x^{r+1},\lambda^{r+1};x^r,\lambda^r) \leq P_c(x^1,\lambda^1;x^0,\lambda^0) \leq P_c^0. \tag{52}$$

Combining (52) with (34), and use the definition of P_c function in (29), we obtain

$$\frac{(1-\rho\gamma)^2}{2\rho}(\|\lambda^{r+1}\|^2 - \|\lambda^r\|^2) + \frac{\gamma(1-\rho\gamma)}{2}\|\lambda^{r+1}\|^2 \leq P_c^0. \tag{53}$$

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

$$\frac{\gamma(1-\rho\gamma)}{2}\|\lambda^{r+1}\|^2 \leq \frac{\gamma(1-\rho\gamma)}{2}\|\lambda^{r+1}\|^2 + \frac{(1-\rho\gamma)^2}{2\rho}(\|\lambda^{r+1}\|^2 - \|\lambda^r\|^2) \overset{(53)}{\leq} P_c^0.$$

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

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

where the second inequality comes from the induction assumption (51). This concludes the proof of (50). **Q.E.D.**

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

$$\gamma\|\lambda^{r+1}\|^2 \leq \frac{2}{1-\tau}P_c^0, \quad \forall\ r \geq 0. \tag{54}$$

Therefore, we get

$$\gamma^2\|\lambda^{r+1}\|^2 \leq 2P_c^0\frac{\gamma}{1-\tau}, \quad \text{for all } r \geq 0\ . \tag{55}$$

Also note that ρ and β should satisfy (32), restated below

$$\tau := \rho\gamma \in (0,1), \quad c > \frac{1}{\tau} - 1 > 0, \quad \beta > (3+2c)L, \quad \rho \geq \beta. \tag{56}$$

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 algorithm parameters.*

$$\gamma = \min\left\{\epsilon, \frac{1}{\beta}\right\}, \quad \rho = \frac{1}{2}\max\left\{\beta, \frac{1}{\epsilon}\right\}, \quad \beta > 7L, \quad c = 2. \tag{57}$$

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 $\lambda^{r+1} \to \lambda^r$, $x^{r+1} \to x^r$. Further, every limit point generated by the PProx-PDA algorithm is an ϵ-stationary solution.

Proof. Using the parameters in (57), we have the following relation

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

where the first equation is true regardless of whether or not $\epsilon \geq 1/\beta$. Then we can bound P_c^0 by the following

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

Therefore using (55) we conclude

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

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

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 $\mathit{\Pi} = \mathit{\Phi} = 0$. 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 in the limit the exact first-order stationary solution can be obtained.

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 (32), 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

$$H(x^r, \lambda^r) := f(x^r) + h(x^r) + \langle \lambda^r, Ax^r - b\rangle, \tag{58a}$$

$$G(x^r, \lambda^r) := \|\tilde{\nabla} H(x^r, \lambda^r)\|^2 + \frac{1}{\rho^2}\|\lambda^{r+1} - \lambda^r\|^2, \tag{58b}$$

$$Q(x^r, \lambda^r) := \|\tilde{\nabla} H(x^r, \lambda^r)\|^2 + \|Ax^r - b\|^2, \tag{58c}$$

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

$$\tilde{\nabla} H(x^r, \lambda^r) = x^r - \mathrm{prox}^{\beta}_{h+\iota(X)}\left[x^r - \frac{1}{\beta}\nabla(H(x^r, \lambda^r) - h(x^r))\right]. \tag{59}$$

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 an θ-stationary solution is obtained if $Q(x^r, \lambda^r) \le \theta$. Note that the θ-stationary solution has been used in [37] 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

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

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

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

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

$$G(x^r, \lambda^r) \leq b_1 \|\lambda^{r+1} - \lambda^r\|^2 + b_2 \|x^{r+1} - x^r\|^2, \tag{60}$$

where $b_1 = \frac{4\sigma_{\max}(A^T A)}{\beta^2} + \frac{1}{\rho^2}$, and $b_2 = 2 + 4\sigma^2_{\max}(\hat{B}^T \hat{B})$ are positive constants. Combining (60) with the descent estimate for the potential function P_c given in (30), we obtain

$$G(x^r, \lambda^r) \leq V \left[P_c(x^r, \lambda^r; x^{r-1}, \lambda^{r-1}) - P_c(x^{r+1}, \lambda^{r+1}; x^r, \lambda^r) \right], \tag{61}$$

where we have defined

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

and one can check that V is in the order of $\mathcal{O}(1/\gamma)$ because a_1 is in the order of γ; cf. (30). From part 1 of Theorem 1 and equation (60) 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 (61) over R iterations, and utilizing the fact that P_c is lower bounded by $\underline{P}$, it follows that

$$\theta \leq \frac{V\left(P_c^0 - \underline{P}\right)}{R} \overset{(36)}{\leq} \frac{V\left(P_c^0 + \frac{(1-\rho\gamma)^2}{2\rho}\|\lambda^1\|^2\right)}{R} \overset{(48)}{\leq} \frac{V\left(P_c^0 + (1-\tau)^2 d_4\right)}{R}$$

where d_4 is given in (45), and P_c^0 is given in (49). Note that $G^{r+1} \leq \theta$ implies that $1/\rho^2 \|\lambda^{r+1} - \lambda^r\|^2 = \|Ax^{r+1} - b - \gamma\lambda^r\|^2 \leq \theta$. From (50) we have that

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

It follows that

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

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

$$Q(x^r, \lambda^r) := \|\tilde{\nabla} H(x^r, \lambda^r)\|^2 + \|Ax^r - b\|^2 \leq \theta + \left(\sqrt{\theta} + \sqrt{\frac{2P_c^0\gamma}{1-\rho\gamma}}\right)^2. \tag{62}$$

Let us pick the parameters such that they satisfy (32), (57), and the following

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

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

$$R \leq \frac{V\left(P_c^0 + (1-\tau)^2 d_4\right)}{\theta} = \mathcal{O}\left(\frac{1}{\theta^2}\right), \tag{63}$$

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.

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*

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

Let R denote the first time that Q^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 problem 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 these 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. Below we analyze the convergence of the new algorithm. Besides assuming that

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

Initialize: λ^0 and x^0

Repeat: update variables by

$$x^{r+1} = \arg\min_{x\in X} \Big\{ \langle \nabla f(x^r), x - x^r\rangle + h(x) + \langle (1-\rho^{r+1}\gamma^{r+1})\lambda^r, Ax - b\rangle + \frac{\rho^{r+1}}{2}\|Ax-b\|^2 + \frac{\beta^{r+1}}{2}\|x-x^r\|^2_{B^TB} \Big\}. \tag{64a}$$

$$\lambda^{r+1} = (1-\rho^{r+1}\gamma^{r+1})\lambda^r + \rho^{r+1}\left(Ax^{r+1} - b\right). \tag{64b}$$

Until Convergence.

the optimization problem under consideration satisfies Assumptions A, we make the following additional assumptions:

Assumptions 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

$$\gamma^{r+1} - \gamma^r \le 0,\ \gamma^{r+1} \to 0,\ \sum_{r=1}^{\infty} \gamma^{r+1} = \infty,\ \sum_{r=1}^{\infty} \left(\gamma^{r+1}\right)^2 < \infty. \tag{65}$$

B3. Assume that

$$\exists\ c_0 > 1 \text{ s.t. } \beta^{r+1} = c_0\rho^{r+1}, \ \text{ for } r \text{ large enough.} \tag{66}$$

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 dependent on the iterates. In the Appendix we will show that such an assumption can be satisfied under some additional regularity conditions 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*.

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.

Lemma 6 *Suppose that the Assumptions A and [B1]-[B3] hold true, and that τ, D, are constants defined in assumption B. Then for large enough r, there exists constant C_1 such that*

$$\begin{aligned}
&\frac{(1-\tau)}{2}\|\lambda^{r+1}-\lambda^r\|^2+\frac{\tau}{2}(\frac{\rho^{r+1}}{\rho^{r+2}}-1)\|\lambda^{r+1}\|^2+\frac{\beta^{r+1}\rho^{r+1}}{2}\|x^{r+1}-x^r\|^2_{B^TB}\\
&+\frac{\rho^{r+1}L}{2}\|x^{r+1}-x^r\|^2_{B^TB}\\
&\leq\frac{(1-\tau)}{2}\|\lambda^{r}-\lambda^{r-1}\|^2+\frac{\tau}{2}(\frac{\rho^{r}}{\rho^{r+1}}-1)\|\lambda^{r}\|^2+\frac{\beta^{r}\rho^{r}}{2}\|x^{r}-x^{r-1}\|^2_{B^TB}\\
&+\frac{\rho^{r}L}{2}\|x^{r}-x^{r-1}\|^2_{B^TB}-\frac{\tau}{2}\|\lambda^{r+1}-\lambda^r\|^2+\frac{C_1(\gamma^{r+1})^2}{2}\|\lambda^{r+1}\|^2\\
&+\frac{L\rho^r+D(L+\beta^{r+1}\|B^TB\|)}{2}\|x^{r+1}-x^r\|^2_{B^TB}-\frac{\beta^r\rho^r}{2}\|w^r\|^2_{B^TB}. \qquad (67)
\end{aligned}$$

Proof. Suppose that $\xi^{r+1}\in\partial h(x^{r+1})$. From the optimality condition for x-subproblem (64a) we have for all $x\in\mathrm{dom}(h)$

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

Performing the above inequality for the $(r-1)$th iteration, we have

$$\langle\nabla f(x^{r-1})+A^T\lambda^{r}+\beta^{r}B^TB(x^{r}-x^{r-1})+\xi^{r},x^{r}-x\rangle\leq 0,\ \forall\ x\in\mathrm{dom}(h).$$

Plugging in $x=x^r$ in the first inequality, $x=x^{r+1}$ in the second inequality and add them together, and use the fact that h is convex, we obtain

$$\begin{aligned}
&\langle\nabla f(x^r)-\nabla f(x^{r-1})+A^T(\lambda^{r+1}-\lambda^r)\\
&+\beta^{r+1}B^TB(x^{r+1}-x^r)-\beta^rB^TB(x^r-x^{r-1}),x^{r+1}-x^r\rangle\leq 0. \qquad (68)
\end{aligned}$$

Let us analyze the above inequality term by term. First, using Young's inequality and the assumption that f is L-smooth i.e. (15) we have

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

Second, note that we have

$$\begin{aligned}
&\langle A^T(\lambda^{r+1}-\lambda^r),x^{r+1}-x^r\rangle=\langle\lambda^{r+1}-\lambda^r,A(x^{r+1}-x^r)\rangle\\
&=\langle\lambda^{r+1}-\lambda^r,Ax^{r+1}-b-\gamma^{r+1}\lambda^r+\gamma^{r+1}\lambda^r+\gamma^r\lambda^{r-1}-\gamma^r\lambda^{r-1}-Ax^r+b\rangle\\
&\overset{(64b)}{=}\langle\lambda^{r+1}-\lambda^r,\frac{\lambda^{r+1}-\lambda^r}{\rho^{r+1}}+\gamma^{r+1}\lambda^r-\gamma^r\lambda^{r-1}-\frac{\lambda^r-\lambda^{r-1}}{\rho^r}\rangle\\
&=\frac{1}{\rho^r}\langle\lambda^{r+1}-\lambda^r,\lambda^{r+1}-\lambda^r-(\lambda^r-\lambda^{r-1})\rangle+\left(\frac{1}{\rho^{r+1}}-\frac{1}{\rho^r}\right)\|\lambda^{r+1}-\lambda^r\|^2\\
&\quad+\langle\lambda^{r+1}-\lambda^r,\lambda^r-\lambda^{r-1}\rangle\gamma^r+\langle\lambda^{r+1}-\lambda^r,\lambda^r\rangle(\gamma^{r+1}-\gamma^r)\\
&\overset{(13)}{=}\frac{1}{2}\left(\frac{1}{\rho^r}-\gamma^r\right)\left(\|\lambda^{r+1}-\lambda^r\|^2-\|\lambda^r-\lambda^{r-1}\|^2+\|(\lambda^{r+1}-\lambda^r)-(\lambda^r-\lambda^{r-1})\|^2\right)\\
&\quad+\gamma^r\|\lambda^{r+1}-\lambda^r\|^2+\left(\frac{1}{\rho^{r+1}}-\frac{1}{\rho^r}\right)\|\lambda^{r+1}-\lambda^r\|^2\\
&\quad+\frac{1}{2}(\gamma^{r+1}-\gamma^r)(\|\lambda^{r+1}\|^2-\|\lambda^r\|^2-\|\lambda^{r+1}-\lambda^r\|^2).
\end{aligned}$$

Summarizing, we have

$$\begin{aligned}\langle A^T(\lambda^{r+1}-\lambda^r), x^{r+1}-x^r\rangle &\geq \frac{1}{2}\left(\frac{1}{\rho^r}-\gamma^r\right)\left(\|\lambda^{r+1}-\lambda^r\|^2-\|\lambda^r-\lambda^{r-1}\|^2\right)\\ &+\frac{1}{2}(\gamma^{r+1}-\gamma^r)(\|\lambda^{r+1}\|^2-\|\lambda^r\|^2)\\ &+\left(\frac{1}{\rho^{r+1}}-\frac{1}{\rho^r}+\gamma^r-\frac{1}{2}(\gamma^{r+1}-\gamma^r)\right)\|\lambda^{r+1}-\lambda^r\|^2\\ &\overset{(B1)}{=}\frac{1}{2}\left(\frac{1}{\rho^r}-\gamma^r\right)\left(\|\lambda^{r+1}-\lambda^r\|^2-\|\lambda^r-\lambda^{r-1}\|^2\right)\\ &+\frac{1}{2}(\gamma^{r+1}-\gamma^r)(\|\lambda^{r+1}\|^2-\|\lambda^r\|^2)\\ &+\left(\gamma^r-\left(\frac{1}{\tau}-\frac{1}{2}\right)(\gamma^r-\gamma^{r+1})\right)\|\lambda^{r+1}-\lambda^r\|^2.\end{aligned}$$

Third, notice that

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

Therefore, from the above three steps, we can bound (68) by

$$\begin{aligned}&\frac{(1-\tau)}{2\rho^r}\|\lambda^{r+1}-\lambda^r\|^2+\frac{1}{2}(\gamma^{r+2}-\gamma^{r+1})\|\lambda^{r+1}\|^2+\frac{\beta^{r+1}}{2}\|x^{r+1}-x^r\|^2_{B^TB}+\frac{L}{2}\|x^{r+1}-x^r\|^2\\ &\leq\frac{(1-\tau)}{2\rho^r}\|\lambda^r-\lambda^{r-1}\|^2+\frac{1}{2}(\gamma^{r+1}-\gamma^r)\|\lambda^r\|^2+\frac{\beta^r}{2}\|x^r-x^{r-1}\|^2_{B^TB}\\ &+\frac{L}{2}\|x^r-x^{r-1}\|^2-\left(\gamma^r-(\frac{1}{\tau}-\frac{1}{2})(\gamma^r-\gamma^{r+1})\right)\|\lambda^{r+1}-\lambda^r\|^2+L\|x^{r+1}-x^r\|^2\\ &+\frac{1}{2}(\gamma^{r+2}-\gamma^{r+1}-(\gamma^{r+1}-\gamma^r))\|\lambda^{r+1}\|^2-\frac{\beta^r}{2}\|w^r\|^2_{B^TB}.\end{aligned}\tag{69}$$

Multiplying ρ^r on both sides, we obtain

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

where we have used the following fact

$$0\geq\left(\frac{\rho^r}{\rho^{r+2}}-\frac{\rho^r}{\rho^{r+1}}\right)=\frac{\rho^r}{\rho^{r+1}}\left(\frac{\rho^{r+1}}{\rho^{r+2}}-1\right)\geq\left(\frac{\rho^{r+1}}{\rho^{r+2}}-1\right).$$

Further, by Assumption B we have $\rho^{r+1}-\rho^r = D$, also we have $\|x^{r+1}-x^r\|^2_{B^TB} \leq \|B^TB\|\|x^{r+1}-x^r\|^2$. Therefore, we reach

$$\begin{aligned}
&\frac{(1-\tau)}{2}\|\lambda^{r+1}-\lambda^r\|^2+\frac{\tau}{2}(\frac{\rho^{r+1}}{\rho^{r+2}}-1)\|\lambda^{r+1}\|^2+\frac{\beta^{r+1}\rho^{r+1}}{2}\|x^{r+1}-x^r\|^2_{B^TB}\\
&\quad+\frac{\rho^{r+1}L}{2}\|x^{r+1}-x^r\|^2\\
&\leq \frac{(1-\tau)}{2}\|\lambda^{r}-\lambda^{r-1}\|^2+\frac{\tau}{2}(\frac{\rho^{r}}{\rho^{r+1}}-1)\|\lambda^{r}\|^2+\frac{\beta^{r}\rho^{r}}{2}\|x^{r}-x^{r-1}\|^2_{B^TB}\\
&\quad+\frac{\rho^{r}L}{2}\|x^{r}-x^{r-1}\|^2-\frac{\tau}{2}\|\lambda^{r+1}-\lambda^r\|^2+\frac{C_1(\gamma^{r+1})^2}{2}\|\lambda^{r+1}\|^2\\
&\quad+\frac{L\rho^r+D(L+\beta^{r+1}\|B^TB\|)}{2}\|x^{r+1}-x^r\|^2-\frac{\beta^r\rho^r}{2}\|w^r\|^2_{B^TB},
\end{aligned}\tag{70}$$

where the last inequality is true using the following relations:

– To bound the term $\gamma^{r+2}-\gamma^{r+1}-(\gamma^{r+1}-\gamma^r)$ we have

$$\begin{aligned}
\gamma^{r+2}-\gamma^{r+1}-(\gamma^{r+1}-\gamma^r) &= \left(\frac{\tau}{\rho^{r+2}}-\frac{\tau}{\rho^{r+1}}-\frac{\tau}{\rho^{r+1}}+\frac{\tau}{\rho^r}\right)\\
&= \tau D\frac{\rho^{r+2}-\rho^r}{\rho^r\rho^{r+1}\rho^{r+2}}=\frac{2\tau D^2}{\rho^r\rho^{r+1}\rho^{r+2}}.
\end{aligned}$$

Thus there exists a constant C_1 such that

$$\frac{\rho^r}{2}\left(\gamma^{r+2}-\gamma^{r+1}-(\gamma^{r+1}-\gamma^r)\right)\leq\frac{C_1(\gamma^{r+1})^2}{2}.$$

– For large enough r

$$\tau\geq\frac{D(2-\tau)}{\rho^{r+1}}.$$

The proof of the lemma is complete. **Q.E.D.**

Now let us analyze 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 ρ and γ (which is also time varying), we denote it as $T(x,\lambda;\rho,\gamma)$.

Lemma 7 *Suppose that the Assumptions Assumptions A and [B1]-[B3] hold true, τ and D are constants defined in Assumption B. Then we have*

$$\begin{aligned}
&T(x^{r+1},\lambda^{r+1};\rho^{r+2},\gamma^{r+2})+\left((1-\tau)\frac{\gamma^{r+2}}{2}-\frac{D(\gamma^{r+2})^2}{2\tau}+D(\gamma^{r+2})^2\right)\|\lambda^{r+1}\|^2\\
&\leq T(x^r,\lambda^r;\rho^{r+1},\gamma^{r+1})+\left((1-\tau)\frac{\gamma^{r+1}}{2}-\frac{D(\gamma^{r+1})^2}{2\tau}+D(\gamma^{r+1})^2\right)\|\lambda^{r}\|^2\\
&\quad-\left(\frac{\beta^{r+1}-3L}{2}\right)\|x^{r+1}-x^r\|^2\\
&\quad+(1-\tau)\left(\frac{1}{\rho^{r+1}}-\frac{\gamma^{r+1}}{2}+\frac{D(\gamma^{r+1})^2}{2\tau^2(1-\tau)}\right)\|\lambda^{r+1}-\lambda^r\|^2\\
&\quad+\frac{(1-\tau)(\gamma^{r+1}-\gamma^{r+2})}{2}\|\lambda^{r+1}\|^2+\frac{D(\gamma^{r+2})^2}{2}\|\lambda^{r+1}\|^2\\
&\quad+D\frac{(\gamma^{r+1})^2-(\gamma^{r+2})^2}{2\tau}\|\lambda^{r+1}\|^2.
\end{aligned}\tag{71}$$

Proof. Following the same analysis as in](27), we have that the T function has the following descent when only changing the primal variable

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

Second, following (28), it is easy to verify that

$$\begin{aligned} &T(x^{r+1},\lambda^{r+1};\rho^{r+1},\gamma^{r+1}) - T(x^{r+1},\lambda^r;\rho^{r+1},\gamma^{r+1}) \\ &\leq (1-\tau)\left(\frac{\|\lambda^{r+1}-\lambda^r\|^2}{\rho^{r+1}} + \frac{\gamma^{r+1}}{2}\left(\|\lambda^r\|^2 - \|\lambda^{r+1}\|^2 - \|\lambda^{r+1}-\lambda^r\|^2\right)\right) \\ &\leq (1-\tau)\Bigg(\frac{\|\lambda^{r+1}-\lambda^r\|^2}{\rho^{r+1}} + \frac{\gamma^{r+1}}{2}\|\lambda^r\|^2 - \frac{\gamma^{r+2}}{2}\|\lambda^{r+1}\|^2 \\ &\quad - (\frac{\gamma^{r+1}}{2} - \frac{\gamma^{r+2}}{2})\|\lambda^{r+1}\|^2 - \frac{\gamma^{r+1}}{2}\|\lambda^{r+1}-\lambda^r\|^2\Bigg). \end{aligned} \tag{73}$$

The most involving step is the analysis of the change of T when the parameters ρ and γ are changed. We first have the following bound

$$\begin{aligned} &T(x^{r+1},\lambda^{r+1};\rho^{r+2},\gamma^{r+2}) - T(x^{r+1},\lambda^{r+1};\rho^{r+1},\gamma^{r+1}) \\ &= (1-\tau)(\gamma^{r+1}-\gamma^{r+2})\|\lambda^{r+1}\|^2 + \frac{\rho^{r+1}-\rho^r}{2}\|Ax^{r+1}-b\|^2 \\ &= (1-\tau)(\gamma^{r+1}-\gamma^{r+2})\|\lambda^{r+1}\|^2 \\ &+ \underbrace{\frac{D}{2}\|(Ax^{r+1}-b)-\gamma^{r+1}\lambda^r\|^2}_{(a)} - \underbrace{\frac{D}{2}\|\gamma^{r+1}\lambda^r\|^2}_{(b)} + \underbrace{D\langle\gamma^{r+1}\lambda^r, Ax^{r+1}-b\rangle}_{(c)}. \end{aligned} \tag{74}$$

The term (a) in (74) is given by

$$\frac{D}{2}\|(Ax^{r+1}-b)-\gamma^{r+1}\lambda^r\|^2 = \frac{D}{(\rho^{r+1})^2}\|\lambda^{r+1}-\lambda^r\|^2. \tag{75}$$

The term (b) in (74) is given by

$$-\frac{D}{2}\|\gamma^{r+1}\lambda^r\|^2 = -(\gamma^{r+1})^2\frac{D}{2}\|\lambda^r\|^2. \tag{76}$$

The term (c) in (74) is given by

$$\begin{aligned} &D\langle\gamma^{r+1}\lambda^r, Ax^{r+1}-b\rangle = D\langle\gamma^{r+1}\lambda^r, \frac{\lambda^{r+1}-\lambda^r}{\rho^{r+1}} + \gamma^{r+1}\lambda^r\rangle \\ &= D(\gamma^{r+1})^2\|\lambda^r\|^2 + D\frac{(\gamma^{r+1})^2}{2\tau}(\|\lambda^{r+1}\|^2 - \|\lambda^r\|^2 - \|\lambda^{r+1}-\lambda^r\|^2). \end{aligned} \tag{77}$$

So collecting terms, we have

$$
\begin{aligned}
&T(x^{r+1},\lambda^{r+1};\rho^{r+2},\gamma^{r+2})+\left((1-\tau)\frac{\gamma^{r+2}}{2}-\frac{D(\gamma^{r+2})^2}{2\tau}+\frac{D}{2}(\gamma^{r+2})^2\right)\|\lambda^{r+1}\|^2\\
\leq& T(x^{r},\lambda^{r};\rho^{r+1},\gamma^{r+1})+\left((1-\tau)\frac{\gamma^{r+1}}{2}-\frac{D(\gamma^{r+1})^2}{2\tau}+\frac{D}{2}(\gamma^{r+1})^2\right)\|\lambda^{r}\|^2\\
&-(\frac{\beta^{r+1}-3L}{2})\|x^{r+1}-x^r\|^2\\
&+(1-\tau)\left(\frac{1}{\rho^{r+1}}-\frac{\gamma^{r+1}}{2}+\frac{D(\gamma^{r+1})^2}{\tau^2(1-\tau)}\right)\|\lambda^{r+1}-\lambda^r\|^2\\
&+\frac{1-\tau}{2}(\gamma^{r+1}-\gamma^{r+2})\|\lambda^{r+1}\|^2+(\gamma^{r+2})^2\frac{D}{2}\|\lambda^{r+1}\|^2\\
&+D\frac{(\gamma^{r+1})^2-(\gamma^{r+2})^2}{2\tau}\|\lambda^{r+1}\|^2. \qquad (78)
\end{aligned}
$$

The lemma is proved. **Q.E.D.**

In the next step we construct and estimate the descent of the potential function. For some given $c>0$, we construct the following potential function

$$
\begin{aligned}
P_c^{r+1}:=&T(x^{r+1},\lambda^{r+1};\rho^{r+2},\gamma^{r+2}) \qquad (79)\\
&+\left((1-\tau)\frac{\gamma^{r+2}}{2}-\frac{D(\gamma^{r+2})^2}{2\tau}+\frac{D}{2}(\gamma^{r+2})^2\right)\|\lambda^{r+1}\|^2\\
&+c\bigg(\frac{(1-\tau)}{2}\|\lambda^{r+1}-\lambda^r\|^2+\frac{\tau}{2}(\frac{\rho^{r+1}}{\rho^{r+2}}-1)\|\lambda^{r+1}\|^2\\
&+\frac{\beta^{r+1}\rho^{r+1}}{2}\|x^{r+1}-x^r\|^2_{B^TB}+\frac{\rho^{r+1}L}{2}\|x^{r+1}-x^r\|^2_{B^TB}\bigg).
\end{aligned}
$$

Lemma 8 *Suppose that the Assumptions A and [B1]-[B3] hold true, and let τ and D be the constants defined in Assumption B. Then for large enough r we have the following for the potential function P_c*

$$
\begin{aligned}
P_c^{r+1}-P_c^r\leq&-\left(\frac{\beta^{r+1}-3L}{2}-cL\rho^r-cDL-c\beta^{r+1}\|B^TB\|\right)\|x^{r+1}-x^r\|^2\\
&-c\frac{\tau}{4}\|\lambda^{r+1}-\lambda^r\|^2+D_0(\gamma^{r+1})^2-c\frac{\beta^r\rho^r}{2}\|w^r\|^2_{B^TB}, \qquad (80)
\end{aligned}
$$

where D_0 is a positive constant.

Proof. According to Lemma 6 and Lemma 7, for large enough r we have

$$
\begin{aligned}
P_c^{r+1}-P_c^r\leq&-\left(\frac{\beta^{r+1}-3L}{2}-cL\rho^r-cDL-c\beta^{r+1}\|B^TB\|\right)\|x^{r+1}-x^r\|^2\\
&-\left(c\frac{\tau}{2}-(1-\tau)\left(\frac{1}{\rho^{r+1}}-\frac{\gamma^{r+1}}{2}+\frac{D(\gamma^{r+1})^2}{\tau^2(1-\tau)}\right)\right)\|\lambda^{r+1}-\lambda^r\|^2\\
&+\frac{(1-\tau)(\gamma^{r+1}-\gamma^{r+2})}{2}\|\lambda^{r+1}\|^2+\frac{D(\gamma^{r+2})^2}{2}\|\lambda^{r+1}\|^2-c\frac{\beta^r\rho^r}{2}\|w^r\|^2_{B^TB}\\
&+D\frac{(\gamma^{r+1})^2-(\gamma^{r+2})^2}{2\tau}\|\lambda^{r+1}\|^2+c\frac{C_1(\gamma^{r+1})^2}{2}\|\lambda^{r+1}\|^2. \qquad (81)
\end{aligned}
$$

From the properties of perturbation parameter γ^r given in (65) we can observe that

$$\gamma^{r+1}-\gamma^{r+2}\leq\frac{D}{\tau}\gamma^{r+1}\gamma^{r+2}\leq\frac{D}{\tau}(\gamma^{r+1})^2.$$

Utilizing this result together with the Assumption [B6] related to dual variable λ, we obtain the following relations for large enough r

$$\frac{(1-\tau)(\gamma^{r+1}-\gamma^{r+2})}{2}\|\lambda^{r+1}\|^2 \leq D\frac{(1-\tau)(\gamma^{r+1})^2\Lambda}{2\tau}. \tag{82}$$

Similarly we also have

$$c\frac{C_1(\gamma^{r+1})^2}{2}\|\lambda^{r+1}\|^2 \leq \frac{cC_1\Lambda(\gamma^{r+1})^2}{2}.$$

Moreover, since $(\gamma^{r+1})^2-(\gamma^{r+2})^2 \leq (\gamma^{r+1})^2$, and $\gamma^{r+2}\leq\gamma^{r+1}$, we have

$$\begin{aligned} D\frac{(\gamma^{r+1})^2-(\gamma^{r+2})^2}{2\tau}\|\lambda^{r+1}\|^2 &\leq \frac{D\Lambda}{2\tau}(\gamma^{r+1})^2,\\ (\gamma^{r+2})^2\frac{D}{2}\|\lambda^{r+1}\|^2 &\leq \frac{D\Lambda}{2}(\gamma^{r+1})^2. \end{aligned} \tag{83}$$

Let us set

$$D_0 := \frac{D(1-\tau)\Lambda}{2\tau}+\frac{cC_1\Lambda}{2}+\frac{D\Lambda}{2\tau}+\frac{D\Lambda}{2},$$

which adds up the constants in front of $(\gamma^{r+1})^2$ in the above terms. We can therefore bound the difference of the potential function by

$$\begin{aligned} P_c^{r+1}-P_c^r \leq &-\left(\frac{\beta^{r+1}-3L}{2}-cL\rho^r-cDL-c\beta^{r+1}\|B^TB\|\right)\|x^{r+1}-x^r\|^2\\ &-\left(c\frac{\tau}{2}-(1-\tau)\left(\frac{1}{\rho^{r+1}}-\frac{\gamma^{r+1}}{2}+\frac{D(\gamma^{r+1})^2}{\tau^2(1-\tau)}\right)\right)\|\lambda^{r+1}-\lambda^r\|^2\\ &+D_0(\gamma^{r+1})^2-c\frac{\beta^r\rho^r}{2}\|w^r\|^2_{B^TB}. \end{aligned} \tag{84}$$

Since $(1-\tau)\left(\frac{1}{\rho^{r+1}}-\frac{\gamma^{r+1}}{2}+\frac{D(\gamma^{r+1})^2}{\tau^2(1-\tau)}\right)\to 0$, we can find r_0 large enough such that for $r\geq r_0$

$$(1-\tau)\left(\frac{1}{\rho^{r+1}}-\frac{\gamma^{r+1}}{2}+\frac{D(\gamma^{r+1})^2}{2\tau^2(1-\tau)}\right)\leq\frac{c\tau}{4}. \tag{85}$$

Thus, for $r\geq r_0$ we have

$$\begin{aligned} P_c^{r+1}-P_c^r \leq &-\left(\frac{\beta^{r+1}-3L}{2}-cL\rho^r-cDL-c\beta^{r+1}\|B^TB\|\right)\|x^{r+1}-x^r\|^2\\ &-c\frac{\tau}{4}\|\lambda^{r+1}-\lambda^r\|^2+D_0(\gamma^{r+1})^2-c\frac{\beta^r\rho^r}{2}\|w^r\|^2_{B^TB}. \end{aligned} \tag{86}$$

The claim is proved. **Q.E.D.**

Note that by Assumption B we have that

$$\sum_{r=1}^{\infty}(\gamma^{r+1})^2<\infty. \tag{87}$$

Therefore to ensure the potential function decrease eventually, we need to pick the constants in the following way [note that by (66), $c_0\rho^{r+1}=\beta^{r+1}$]

$$\frac{c_0\rho^{r+1}-3L}{2}-cL\rho^r-cDL-cc_0\rho^{r+1}\|B^TB\|\geq 0. \tag{88}$$

It is clear that if constant c is picked such that

$$0<c\leq\frac{c_0}{2(L+c_0\|B^TB\|)}. \tag{89}$$

Then the above inequality is satisfied for large enough r.

In this step we show that the potential function is lower bounded.

Lemma 9 *Suppose that the Assumptions A and [B1]-[B3] hold true, and that the constant c is chosen such that*

$$0 < c \leq \frac{1-\tau}{D}. \tag{90}$$

Then the potential function P_c^r defined in (79) *is lower bounded.*

Proof. Let us rearrange the terms of the potential function

$$\begin{aligned} P_c^{r+1} &= T(x^{r+1}, \lambda^{r+1}; \rho^{r+2}, \gamma^{r+2}) \\ &+ \left(\frac{(1-\tau)D(\gamma^{r+2})^2}{2\tau} + \frac{(1-\tau-cD)\gamma^{r+2}}{2} \right) \|\lambda^{r+1}\|^2 \\ &+ c\bigg(\frac{(1-\tau)}{2}\|\lambda^{r+1}-\lambda^r\|^2 + \frac{\beta^{r+1}\rho^{r+1}}{2}\|x^{r+1}-x^r\|^2_{B^TB} + \frac{L\rho^{r+1}}{2}\|x^{r+1}-x^r\|^2 \bigg). \end{aligned} \tag{91}$$

First of all, we note that if we set $0 < c \leq \frac{1-\tau}{D}$ then the coefficient in front of $\|\lambda^{r+1}\|^2$ is positive. Let us analyze $T(x^{r+1}, \lambda^{r+1}; \rho^{r+2}, \gamma^{r+2})$. We have the following

$$\begin{aligned} &\langle \lambda^{r+1} - \rho^{r+2}\gamma^{r+2}\lambda^{r+1}, Ax^{r+1} - b - \gamma^{r+2}\lambda^{r+1} \rangle \\ &= \frac{1-\tau}{\rho^{r+1}}\langle \lambda^{r+1}, \lambda^{r+1} - \lambda^r \rangle + (1-\tau)\langle \lambda^{r+1}, \gamma^{r+1}\lambda^r - \gamma^{r+2}\lambda^{r+1} \rangle \\ &= \frac{1-\tau}{\rho^{r+1}}\langle \lambda^{r+1}, \lambda^{r+1} - \lambda^r \rangle + (1-\tau)\gamma^{r+1}\langle \lambda^{r+1}, \lambda^r - \lambda^{r+1} \rangle \\ &+ (1-\tau)(\gamma^{r+1} - \gamma^{r+2})\|\lambda^{r+1}\|^2 \\ &\geq \left(\frac{1-\tau}{\rho^{r+1}} - (1-\tau)\gamma^{r+1} \right) \langle \lambda^{r+1}, \lambda^{r+1} - \lambda^r \rangle \\ &= \frac{1}{2\rho^{r+1}}(1-\tau)^2(\|\lambda^{r+1}\|^2 - \|\lambda^r\|^2 + \|\lambda^{r+1} - \lambda^r\|^2) \\ &\geq \frac{(1-\tau)^2}{2}\left(\frac{1}{\rho^{r+1}}\|\lambda^{r+1}\|^2 - \frac{1}{\rho^r}\|\lambda^r\|^2\right). \end{aligned} \tag{92}$$

It follows that the sum $\sum_{r=1}^{\infty} T(x^{r+1}, \lambda^{r+1}; \rho^{r+2}, \gamma^{r+2})$ is lower bounded. The claim can then be proved by using a similar argument as in Lemma 3. **Q.E.D.**

Finally we put all the previous lemmas together to present the main convergence results for the PProx-PDA-IA.

Theorem 2 *Suppose that Assumptions A–B hold true, and that τ, c and D are picked such that* (89) *and* (90) *are satisfied. Then every limit point of the sequence generated by PProx-PDA-IA is a stationary solution of problem* (1).

Proof. In this proof we pick a special case of B satisfying $B^TB = I$, in order to avoid unnecessarily complicated notation. The proof is a modification of the classical result in [8, Proposition 3.5].

Combining Lemma 6 and Lemma 9, we have

$$\sum_{r=1}^{\infty} \beta^{r+1}\|x^{r+1} - x^r\|^2 < \infty, \quad \sum_{r=1}^{\infty} \|\lambda^{r+1} - \lambda^r\|^2 < \infty, \tag{93}$$

$$\sum_{r=1}^{\infty} (\beta^{r+1})^2 \|(x^{r+1} - x^r) - (x^r - x^{r-1})\|^2 < \infty. \tag{94}$$

From (93) we have $\lambda^{r+1} - \lambda^r \to 0$, which implies that From (94), we have

$$(\rho^{r+1})(Ax^{r+1} - b) - \tau\lambda^r \to 0. \tag{95}$$

Combined with the fact that λ^r is bounded, and $\rho^{r+1} \to \infty$, we conclude

$$Ax^{r+1} - b \to 0. \tag{96}$$

Let (x^*, λ^*) be a limit point of (x^{r+1}, λ^{r+1}). Comparing the optimality condition of the problem (1) and the optimality condition of x-subproblem (64a), in order to argue convergence to stationary solutions, we need to show

$$\beta^{r+1}\|x^{r+1} - x^r\| \to 0. \tag{97}$$

Next we show such a claim. To proceed, let us define

$$v^{r+1} := \beta^{r+1}(x^{r+1} - x^r). \tag{98}$$

From (94), it is easy to show that

$$\|v^{r+1} - v^r\| = \|\beta^{r+1}(x^{r+1} - x^r) - \beta^r(x^r - x^{r-1})\| \to 0. \tag{99}$$

From the first inequality in (93), we have

$$\sum_{r=1}^{\infty} \frac{1}{\beta^{r+1}}\|v^{r+1}\|^2 \to 0. \tag{100}$$

This relation combined with Assumption [B3] implies: $\liminf \|v^{r+1}\| = 0$.

Let us pass a subsequence $\mathcal{K}$ to (x^r, λ^r) and denote (x^*, λ^*) as its limit point. For notational simplicity, in the following the index set $\{r\}$ all belongs to the set $\mathcal{K}$. We already know from the previous argument that $\liminf_{r\to\infty} \|v^{r+1}\| = 0$. Then it is clear that $\lim_{r\to\infty} \|v^{r+1}\| = 0$ if and only the following condition is true

$$\lim_{r\to\infty} \|v^{r+1} - v^{r+t}\| = 0, \quad \forall\ t > 0. \tag{101}$$

Let us construct a new sequence

$$z^{r+1} = A^T\lambda^{r+1} + v^{r+1}. \tag{102}$$

Clearly $\liminf_{r\to\infty} z^{r+1} = A^T\lambda^*$, because along the subsequence λ^r converges to λ^*. It is also easy to show that (101) is true if and only if the following is true

$$\lim_{r\to\infty} \|z^{r+1} - z^{r+t}\| = 0, \quad \forall\ t > 0. \tag{103}$$

Suppose that (103) is not true. Hence there exists an $\epsilon > 0$ such that $\|z^r\| < \|A^T\lambda^*\| + \epsilon/2$ for infinitely many r, and $\|z^{r+1}\| > \|A^T\lambda^*\| + \epsilon/2$ for infinitely many r. Then there exists an infinite subset of iteration indices $\mathcal{R}$ such that for each $r \in \mathcal{R}$, there exits a $t(r)$ such that

$$\begin{aligned} &\|z^r\| < \|A^T\lambda^*\| + \epsilon/2, \quad \|z^{t(r)}\| > \|A^T\lambda^*\| + \epsilon, \\ &\|A^T\lambda^*\| + \epsilon/2 < \|z^t\| \le \|A^T\lambda^*\| + \epsilon, \quad \forall\ r < t < t(r). \end{aligned} \tag{104}$$

Also from the fact that $\|v^{r+1} - v^r\| \to 0$ and $\|\lambda^{r+1} - \lambda^r\| \to 0$, we can conclude that $\|z^{r+1} - z^r\| \to 0$. Therefore, we must have

$$\|z^r\| \ge \frac{3\epsilon}{8} + \|A^T\lambda^*\|. \tag{105}$$

Let r be large enough such that

$$\big|\|A^T\lambda^*\| - \|A^T\lambda^r\|\big| \le \|A^T(\lambda^* - \lambda^r)\| \le \frac{\epsilon}{4}. \tag{106}$$

Then we have

$$\|v^t\| \leq \|A^T\lambda^t\| + \|A^T\lambda^*\| + \epsilon \leq 2(\|A^T\lambda^*\| + \epsilon),\ \forall\ r < t < t(r), \tag{107a}$$

$$\|v^t\| \geq \|z^t\| - \|A^T\lambda^t\| \overset{(106)}{\geq} \|z^t\| - \|A^T\lambda^*\| - \frac{\epsilon}{4} \overset{(104)}{\geq} \frac{\epsilon}{4},\ \forall\ r < t < t(r). \tag{107b}$$

$$\|v^r\| \geq \|z^r\| - \|A^T\lambda^r\| \geq \|z^r\| - \|A^T\lambda^*\| - \frac{\epsilon}{4} \overset{(105)}{\geq} \frac{\epsilon}{8}. \tag{107c}$$

From the definition of $t(r)$ we have that for all $r \in \mathcal{R}$ the following is true

$$\frac{\epsilon}{2} \leq \|z^{t(r)}\| - \|z^r\| \leq \sum_{t=r}^{t(r)-1} \|z^{t+1} - z^t\|. \tag{108}$$

Next, we make the following simplification that $X \equiv \mathbb{R}$ and $h \equiv 0$ to avoid lengthy discussion. The subsequent proof holds true for the general case as well, using the same techniques presented in [64, Theorem 4]. From the optimality condition (68), and with the above simplification, we obtain

$$z^{t+1} - z^t = \nabla f(x^t) - \nabla f(x^{t-1}), \tag{109}$$

which implies that

$$\|z^{t+1}\| - \|z^t\| \leq L\|x^t - x^{t-1}\| = \frac{L}{\beta^t}\|v^t\|. \tag{110}$$

Combining this result with (108), we obtain

$$\frac{\epsilon}{2} < L \sum_{t=r}^{t(r)-1} \frac{1}{\beta^t}\|v^t\| \overset{(107a)}{\leq} 2L(\|A^T\lambda^*\| + \epsilon) \sum_{t=r}^{t(r)-1} \frac{1}{\beta^t}. \tag{111}$$

Which implies that

$$\frac{\epsilon}{4L(\|A^T\lambda^*\| + \epsilon)} \leq \sum_{t=r}^{t(r)-1} \frac{1}{\beta^t}. \tag{112}$$

Using the descent of the potential function (84) we have, for $r \in \mathcal{R}$ and r large enough

$$\begin{aligned} P_c^{t(r)} - P_c^r &\leq -\sum_{t=r}^{t(r)-1} \frac{C_5}{\beta^{t+1}}\|v^{t+1}\|^2 + \sum_{t=r}^{t(r)-1} C_3(\gamma^{t+1})^2\|\lambda^{t+1}\|^2 \\ &\leq -\frac{C_5}{L(\|A^T\lambda^*\| + \epsilon)}\frac{\epsilon^2}{64} \end{aligned} \tag{113}$$

where the last inequality we have used the fact that

$$\lim_{R_0 \to \infty} \sum_{r=R^0}^{\infty} C_3(\gamma^{t+1})^2\|\lambda^{t+1}\|^2 \to 0,$$

and equations (107b) and (112). This means that the potential function goes to $-\infty$, a contradiction. Therefore we conclude that

$$\lim_{r\to\infty} \|z^{r+1} - z^{r+t}\| = 0, \quad \forall\ t > 0. \tag{114}$$

which further implies that

$$\lim_{r\to\infty} \|v^{r+1} - v^{r+t}\| = 0, \quad \forall\ t > 0. \tag{115}$$

Combined with the fact that $\liminf \|v^{r+1}\| = 0$, we conclude that

$$\lim_{r\to\infty} \|v^{r+1}\| = 0. \tag{116}$$

We conclude that every limit point of the sequence is a KKT point. **Q.E.D.**

4 Numerical Results

In this section, we customize the proposed algorithms to a number of applications in Section 1.3, and compare with the state-of-the-art algorithms.

4.1 Distributed Nonconvex Quadratic Problem

In this subsection we consider the nonconvex ℓ_1 penalized, nonnegative, sparse principal component analysis (SPCA) problem [5]. Distributed version of this problem [which is a special case of problem (1)] can be modeled as below

$$\begin{aligned} \min_{x} \quad & \sum_{i=1}^{N} \left\{ -x_i^\top \Sigma_i x_i + \alpha \|x_i\|_1 \right\} \\ \text{s.t.} \quad & \|x_i\|^2 \le 1, \quad x_i \ge 0, \quad i = 1, \cdots, N \\ & Ax = 0; \quad \texttt{Consensus Constraint} \end{aligned} \tag{117}$$

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; $\alpha > 0$ is a constant that controls the sparsity. Let us define $\bar{x} := \frac{1}{N}\sum_{i=1}^N x_i$, $h(\bar{x}) := \alpha\|\bar{x}\|_1$, $f(\bar{x}) := \sum_{i=1}^N \bar{x}^T \Sigma_i \bar{x}$, and $X := \{x_i \mid \|\bar{x}\|^2 \le 1, \bar{x} \ge 0\}$. The stationarity gap and the constraint violation for this problem is defined as below

$$\text{staionary-gap} = \left\| \bar{x} - \text{prox}_{h+\iota_X}\left[\bar{x} - \nabla f(\bar{x})\right] \right\|^2, \quad \text{con-vio} = \|Ax\|^2. \tag{118}$$

At this point, one can certainly use Algorithm 1 or Algorithm 2 to solve problem (117). However, the resulting x- subproblems for both algorithms are difficult to solve due to the fact that computing the proximity operator for nonsmooth function $\alpha\|x\|_1 + \iota_{\|x\|^2\le 1}(x) + \iota_{x\ge 0}(x)$ does not have a closed form (where $\iota_X(x)$ represents the indicator function for convex set X). On the contrary, 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. Let us denote $r = \lfloor N/3 \rfloor$. The new reformulation is given below

$$\begin{aligned} \min \quad & \sum_{i=1}^{r} \left\{ -x_i^\top \Sigma_i x_i + \frac{N\alpha}{r}\|x_i\|_1 \right\} - \sum_{i=r+1}^{2r} x_i^\top \Sigma_i x_i - \sum_{i=2r+1}^{N} x_i^\top \Sigma_i x_i \\ \text{s.t.} \quad & \|x_i\|^2 \le 1, \quad i = r+1, \cdots, 2r \\ & x_i \ge 0, \qquad i = 2r+1, \cdots, N \\ & Ax = 0 \quad \texttt{Consensus Constraint}. \end{aligned} \tag{119}$$

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 stationary solution). The major difficulty is to deal with the *agent-specific* nonsmooth terms. For comparison purpose, we consider the DSG algorithm [56], and the NEXT algorithm [50]. In our numerical result, the graph $\mathcal{G}$ is generated based on the scheme proposed in [71]. 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. The number of agents, the network radius, the problem dimension, and the sparsity parameter to be $N = 20, R = 0.7, d = 10, \alpha = 0.01$, respectively. For PProx-PDA algorithm we set perturbation parameter $\gamma = 10^{-4}$, and ρ and β are picked such that they satisfy the theoretical bounds given in (56). For PProx-PDA-IA we set the increasing penalty $\rho = \beta = 40r$, and decreasing perturbation $\gamma = 10^{-3}/r$. For the DSG algorithm the stepsize is set $0.1/r$ (this choice is made so that DSG has the best performance). The parameters for NEXT are tuned according to the description in [50, Theorem 3]. Each algorithm is run for 20 independents

trials, with random initialization and randomly generated data. The results are plotted in Fig. 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 stepsize version, outperform both DSG and NEXT. To see more numerical results we compare different algorithms

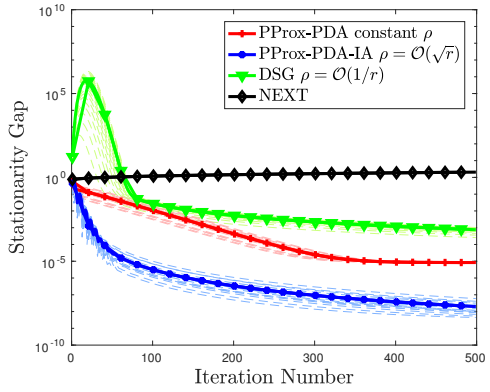

Fig. 1: Comparison of proposed algorithms with DSG [56] and NEXT [50] in terms of stationarity gap for problem 119 with parameters $N = 20, R = 0.7, d = 10, \alpha = 0.01$.

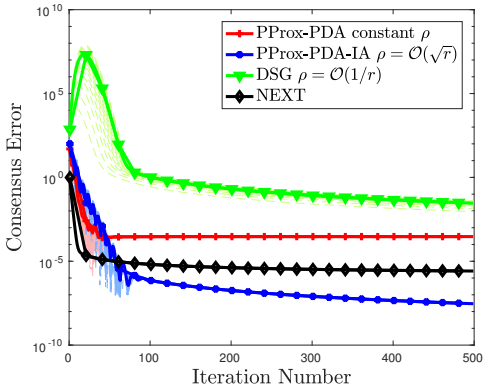

Fig. 2: Comparison of proposed algorithms with DSG [56] and NEXT [50] in terms of constraint violation for problem 119 with parameters $N = 20, R = 0.7, d = 10, \alpha = 0.01$.

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 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 number of agents N, 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.

Table 1: Comparison of proposed algorithms with DSG algorithm. Alg1 and Alg2 denote PProx-PDA and PProx-PDA-IA algorithms respectively.

	Stationarity-Gap			Cons-Vio		
Parameters	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

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 proposed in [26, 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 (56). 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 [26]. In particular, we set $s = 5$ 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 Fig. 3– 6. As shown in these figures, compared to the ADMM algorithm, the PProx-PDA-IA algorithm converges faster, and to better solutions.

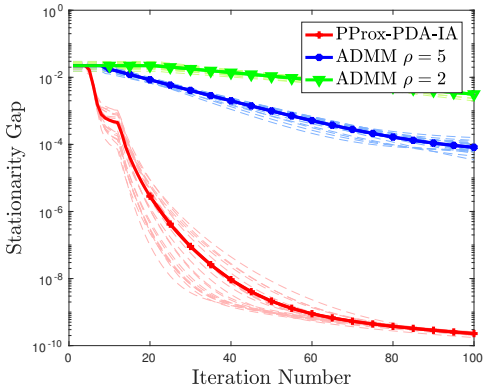

Fig. 3: Comparison of proposed algorithms with ADMM in terms of stationarity gap for nonconvex subspace estimation problem with MCP Regularization. The solid lines and dotted lines represent the single performance and the average performance, respectively.

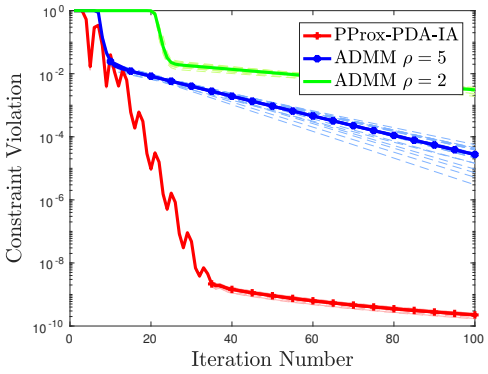

Fig. 4: Comparison of proposed algorithms with ADMM in terms of constraint violation $\|Ax\|^2$ for nonconvex subspace estimation problem with MCP Regularization. The solid lines and dotted lines represent the single performance and the average performance, respectively.

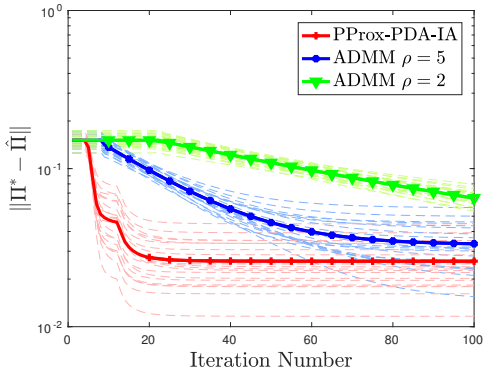

Fig. 5: Comparison of proposed algorithms with ADMM in terms of Global Error for non-convex subspace estimation problem with MCP Regularization. The problem parameters are $n = 80$, $p = 128$, $\nu = 3$, $b = 3$. The solid lines and dotted lines represent the single performance and the average performance, respectively.

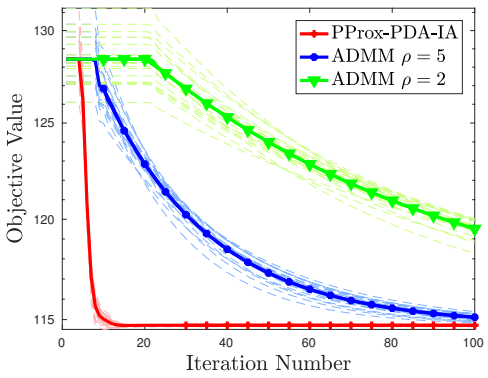

Fig. 6: Comparison of proposed algorithms with ADMM in terms of objective value for nonconvex subspace estimation problem with MCP Regularization. The solid lines and dotted lines represent the single performance and the average performance, respectively.

Our next experiment is designed to see 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 stepsize $\rho = 5$. 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 [26]. 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.

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.

	$\|\hat{\Pi} - \Pi^*\|$	
Parameters	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

Further, the True Positive Rate (TPR) and False Positive Rate (FPR) [39] 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 . Let P denotes the number of 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

$$FPR = \frac{FP}{S}, \quad TPR = \frac{TP}{P}. \tag{120}$$

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

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

	TPR		FPR	
Parameters	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

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

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

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 subsection 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. Fig. 8 displays the constraint violation for the 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.

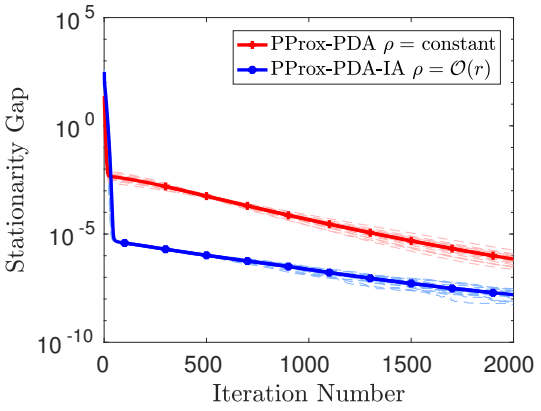

Fig. 7: The stationarity gap achieved by the proposed methods for the partial consensus problem. The solid lines and dotted lines represent the single performance and the average performance, respectively.

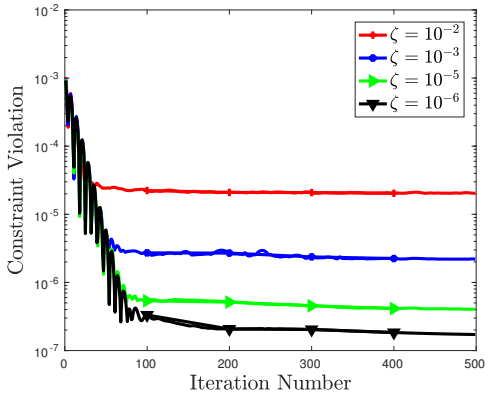

Fig. 8: Constraint Violation $\|Ax\|$ achieved by the proposed method for the partial consensus problem with different permissible tolerance ξ.

5 Conclusion

In this paper, we proposed a 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.

Acknowledgment. The authors would like to thanks Dr. Quanquan Gu who provided us with the codes to perform the numerical results in [26].

6 Appendix

In this section, we justify Assumption [B4], which imposes the boundedness of the dual variable. Throughout this section we will assume that Assumptions A and [B1]–[B3] hold. We present two situations in which the dual variables are guaranteed to be bounded. First we prove that the sequence $\beta^{r+1}\|x^{r+1}-x^r\|$ is bounded (for large enough r). Using Assumption [B3] we have the following identity

$$\frac{\beta^{r+1}\rho^{r+1}}{2}\|x^{r+1}-x^r\|^2=\frac{(\beta^{r+1})^2}{2c_0}\|x^{r+1}-x^r\|^2. \tag{121}$$

From Lemma 8, we have that [also cf. (93)]

$$\sum_{r=1}^{\infty}\|\lambda^{r+1}-\lambda^r\|^2<\infty, \tag{122}$$

which implies that

$$\frac{1}{\rho^{r+1}}\|\lambda^{r+1}\|^2-\frac{1}{\rho^r}\|\lambda^r\|^2\to 0. \tag{123}$$

Plugging this result into (92), we conclude that the following inner product is lower bounded

$$\langle\lambda^{r+1}-\rho^{r+2}\gamma^{r+2}\lambda^{r+1},Ax^{r+1}-b-\gamma^{r+2}\lambda^{r+1}\rangle,$$

and this further implies that $T(x^{r+1},\lambda^{r+1};\rho^{r+2},\gamma^{r+2})$ is lower bounded [by using the definition of T function in (24)]. By Lemma 9 (resp. Lemma 8), we conclude that the potential function is lower (resp. upper) bounded. Examine the definition of the potential function in (91) and use the choice of c in (90) we conclude that except $T(x^{r+1},\lambda^{r+1};\rho^{r+2},\gamma^{r+2})$, all the rest of the terms are all nonnegative. Using the lower boundedness of T, we conclude that the term $\frac{\beta^{r+1}\rho^{r+1}}{2}\|x^{r+1}-x^r\|^2$ in the potential function is bounded. Therefore, there exists D_1 such that

$$\beta^{r+1}\|x^{r+1}-x^r\|\le D_1. \tag{124}$$

Case 1). In this case, 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 [19,51].

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 prove that the dual variable is bounded. Now we prove that the dual variable is bounded.

Lemma 10 *Suppose the Robinson's condition holds true for problem* (1)*. Then the sequence of dual variable λ^r generated by* (64b) *is bounded.*

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

$$\|\lambda^r\|\to\infty. \tag{125}$$

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 $\beta^{r+1}\|x^{r+1}-x^r\|$ is a bounded sequence, so does $\beta^{r+1}B^TB(x^{r+1}-x^r)$. Suppose that $\{\lambda^r\}$ is not bounded, let us define a new *bounded* sequence as $\mu^r=\lambda^r/\|\lambda^r\|$. Let (x^*,μ^*) be a limit 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 $x\in X$

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

Passing limit, 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 $c\langle A, x - x^*\rangle = -\mu^*$. 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. **Q.E.D.**

Case 2). In this section, we verify Assumption [B4] by further imposing conditions on the constraint set and the nonsmooth terms. Specifically we consider the following problem

$$\min_{\{x_k\}} \ f(x) + h(x) := f(x) + \sum_{k=1}^{K} h_k(x_k) \quad \text{s.t.} \quad \sum_{k=1}^{K} A_k x_k = b, \tag{126}$$

where h_k is a convex nonsmooth term that can include both regularizer and indicator functions for convex set X. Setting $K = 1$, the above problem is equivalent to the original problem (1).
Assumption C. Assume that for one of the block, say $k \in 1, \cdots K$, the following is satisfied:

$$X_k = \mathbb{R}^{n_k}. \tag{127}$$

Note that the second of the above condition is possible for example when $h_k(x_k) = \|x_k\|_q$ for some constant $q \geq 1$. Further, we assume that the partial gradient of f with respect to x_k, denoted by $\nabla_k f(x)$, is bounded for all $x_k \in \text{dom}\,(h_k)$, and that A_k has full row rank. ■

Given the above assumption, the following lemma characterizes the bound for the dual variable.

Lemma 11 *Suppose that the Assumption C holds true. Then the sequence of dual variable λ^r generated by* (64b) *is bounded.*

Proof. First, from the optimality condition of x-update (64a) we have that for all k, and for all $x_k \in \text{dom}(h_k)$

$$\langle \nabla_k f(x^r) + A_k^T\lambda^{r+1} + \beta^{r+1}B^TB(x_k^{r+1} - x_k^r) + \xi_k^{r+1}, x_k^{r+1} - x_k\rangle \leq 0, \tag{128}$$

where $\nabla_k f(x^r)$ denotes the partial derivative of $f(x)$ with respect to the block variable x_k at $x = x^r$; and $\xi_k^{r+1} \in \partial h_k(x^{r+1})$. In particular for the block K because it is unconstrained, we have

$$0 = \nabla_K f(x^r) + A_K^T\lambda^{r+1} + \xi_K^{r+1} + \beta^{r+1}B^TB(x_K^{r+1} - x_K^r). \tag{129}$$

Rearranging terms, we obtain

$$-A_K^T\lambda^{r+1} = \nabla_K f(x^r) + \xi_K^{r+1} + \beta^{r+1}B^TB(x_K^{r+1} - x_K^r). \tag{130}$$

From Assumption C we know that theres exists M_0 such that $\|\nabla_K f(x^r) + \xi_K^{r+1}\| \leq M_0$. Together with the previous identity, we get

$$\|A_K^T\lambda^{r+1}\|^2 \leq 2M_0^2 + 2(\beta^{r+1})^2\|B^TB(x_K^{r+1} - x_K^r)\|^2 \quad \forall \ r. \tag{131}$$

Utilizing the fact that $\sigma_K^2\|\lambda^{r+1}\|^2 \leq \|A_K^T\lambda^{r+1}\|^2$, where σ_K^2 denoted the smallest nonzero eigenvalue of $A_K^TA_K$, we further have

$$\|\lambda^{r+1}\|^2 \leq \frac{2}{\sigma_K^2}\left[(\beta^{r+1})^2\|B^TB(x_K^{r+1} - x_K^r)\|^2 + D_0^2\right] \quad \forall \ r. \tag{132}$$

Here $\sigma_K > 0$ because we have assumed that A_K^T is full column rank in Assumption C. Combining this with equation (124) one can find constant Λ such that $\|\lambda^{r+1}\|^2 \leq \Lambda$. The proof is complete. **Q.E.D.**

Appendix B

We show how the sufficient conditions developed in Appendix A can be applied to problems discussed in Section 1.3. We will focus on the sparse subspace estimation problem (7) and the partial consensus problem (10).

We first show that Assumption C is satisfied for sparse subspace estimation problem (7). Recall that for this problem we have two block variables (Π, Φ), and $h(\Phi) := \|\Phi\|_1 = \sum_{i=1}^p \sum_{j=1}^p |\phi_{ij}|$. It is easy to see that the subdifferential of the ℓ_1 function is bounded in $[-1, 1]$. Then we show that $\nabla_\Phi f(\Pi, \Phi)$ is bounded where $f(\Pi, \Phi) = \left\langle \hat{\Sigma}, \Pi \right\rangle + q_\nu(\Phi)$, and $\nabla_\Phi f(\Pi, \Phi) = \nabla_\Phi q_\nu(\Phi)$. For the MCP regularization with parameter b, we have $q_\nu(\Phi) = \sum_{i=1}^p \sum_{j=1}^p q_\nu(\phi_{ij})$, and we can check that where

$$q_\nu(\phi_{ij}) = \begin{cases} \frac{-\phi_{ij}^2}{2b} & \text{if } |\phi_{ij}| \le b\nu; \\ -\nu|\phi_{ij}| + \frac{b\nu^2}{2} & \text{if o.w.} \end{cases}, \quad \frac{\partial q_\nu(\phi_{ij})}{\partial \phi_{ij}} = \begin{cases} \frac{-\phi_{ij}}{b} & \text{if } |\phi_{ij}| < b\nu; \\ -\nu \text{ sign}(\phi_{ij}) & \text{o.w.} \end{cases}$$

This is obviously a bounded function. Finally the matrix $A_\Phi = -I$ has full row rank. In summary, we have validated all the conditions in Assumption C.

Next we consider the partial consensus problem given in (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 $g(x) \le 0$ [62, Lemma 3.16]. To state the MFCQ, consider the following system

$$\begin{aligned} p_i(y) = 0, \ i = 1, \cdots, M \\ g_j(y) \le 0, \ j = 1, \cdots, P \end{aligned} \tag{133}$$

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

$$\mathcal{A}(\hat{y}) := \{1 \le j \le P \ \mid \ g_j(\hat{y}) = 0\}. \tag{134}$$

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 (133) 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

$$\nabla p(\hat{y}) d_y = 0, \quad \nabla g_j(\hat{y})^T d_y < 0, \ \forall \ j \in \mathcal{A}(\hat{y}). \tag{135}$$

See [62, Lemma 3.17] for more details. Below 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 (133) 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

$$Ad_x = d_z, \tag{136a}$$

$$[d_z]_e < 0 \quad \text{for } e \in \mathcal{A}^+(\hat{y}), \tag{136b}$$

$$[d_z]_e > 0 \quad \text{for } e \in \mathcal{A}^-(\hat{y}), \tag{136c}$$

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 $\mathbf{1}$ [14]. 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 (136b) is verified. Using similar argument we can verify condition (136c).
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 (136a)–(136c) 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 (2016)
2. Ames, B., Hong, M.: Alternating directions method of multipliers for l1-penalized zero variance discriminant analysis and principal component analysis. Computational Optimization and Applications **64**(3), 725–754 (2016)
3. Andreani, R., Haeser, G., Martinez, 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. Annals of the Institute of Statistical Mathematics **63**(3), 585–615 (2009)
5. Asteris, M., Papailiopoulos, D., Dimakis, A.: Nonnegative sparse pca with provable guarantees. In: Proceedings of the 31st International Conference on International Conference on Machine Learning (ICML), pp. 1728–1736 (2014)
6. Aybat, N.S., Hamedani, E.Y.: A primal-dual method for conic constrained distributed optimization problems. Advances in Neural Information Processing Systems (2016)
7. Bertsekas, D.P.: Constrained Optimization and Lagrange Multiplier Method. Academic Press (1982)
8. Bertsekas, D.P., Tsitsiklis, J.N.: Neuro-Dynamic Programming. Athena Scientific, Belmont, MA (1996)
9. Bertsekas, D.P., Tsitsiklis, J.N.: Parallel and Distributed Computation: Numerical Methods, 2nd ed. Athena Scientific, Belmont, MA (1997)
10. Bianchi, P., Jakubowicz, J.: Convergence of a multi-agent projected stochastic gradient algorithm for non-convex optimization. IEEE Transactions on Automatic Control **58**(2), 391–405 (2013)
11. Bjornson, E., Jorswieck, E.: Optimal resource allocation in coordinated multi-cell systems. Foundations and Trends in Communications and Information Theory **9** (2013)
12. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Foundations and Trends in Machine Learning **3**(1), 1–122 (2011)
13. Burachik, R.S., Kaya, C.Y., Mammadov, M.: An inexact modified subgradient algorithm for nonconvex optimization. Computational Optimization and Applications **45**(1), 1–24 (2008)
14. Chung, F.R.K.: Spectral Graph Theory. The American Mathematical Society (1997)
15. Cressie, N.: Statistics for spatial data. ohn Wiley & Sons (2015)
16. Curtis, F.E., Gould, N.I.M., Jiang, H., Robinson, D.P.: Adaptive augmented lagrangian methods: algorithms and practical numerical experience. Optimization Methods and Software **31**(1), 157–186 (2016)
17. d'Aspremont, A., Ghaoui, L.E., Jordan, M.I., Lanckriet, G.R.G.: A direct formulation for sparse pca using semidefinite programming. SIAM Review **49**(3), 434–448 (2007)
18. Deng, W., Yin, W.: On the global and linear convergence of the generalized alternating direction method of multipliers. Journal of Scientific Computing pp. 1–28 (2015)
19. Dutta, J., Deb, K., Tulshyan, R., Arora, R.: Approximate kkt points and a proximity measure for termination. Journal of Global Optimization **56**(4), 1463–1499 (2013)
20. Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association **96**(456), 1348–1360 (2001)
21. Fernndez, D., Solodov, M.V.: Local convergence of exact and inexact augmented lagrangian methods under the second-order sufficient optimality condition. SIAM Journal on Optimization **22**(2), 384–407 (2012)
22. Forero, P.A., Cano, A., Giannakis, G.B.: Distributed clustering using wireless sensor networks. IEEE Journal of Selected Topics in Signal Processing **5**(4), 707–724 (2011)

23. Gabay, D., Mercier, B.: A dual algorithm for the solution of nonlinear variational problems via finite element approximation. Computers & Mathematics with Applications **2**, 17–40 (1976)
24. 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)
25. Glowinski, R., Marroco, A.: Sur l'approximation, par elements finis d'ordre un,et la resolution, par penalisation-dualite, d'une classe de problemes de dirichlet non lineares. Revue Franqaise d'Automatique, Informatique et Recherche Opirationelle **9**, 41–76 (1975)
26. 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)
27. Haeser, G., Melo, V.: On sequential optimality conditions for smooth constrained optimization (2013). Preprint
28. 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)
29. 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)
30. Hajinezhad, D., Hong, M., Garcia, A.: Zeroth order nonconvex multi-agent optimization over networks. arXiv preprint arXiv:1710.09997 (2017)
31. 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 29, pp. 3215–3223 (2016)
32. Hajinezhad, D., Shi, Q.: Alternating direction method of multipliers for a class of nonconvex bilinear optimization: convergence analysis and applications. Journal of Global Optimization pp. 1–28 (2018)
33. Hamdi, A., Mishra, S.K.: Decomposition Methods Based on Augmented Lagrangians: A Survey, pp. 175–203. Springer New York, New York, NY (2011)
34. Hestenes, M.R.: Multiplier and gradient methods. Journal of Optimization and Application (4), 303 – 320 (1969)
35. 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), vol. 70, pp. 1529–1538 (2017)
36. Hong, M., Luo, Z.Q.: On the linear convergence of the alternating direction method of multipliers. Mathematical Programming Series A (2016). To appear, available at arXiv:1208.3922
37. Hong, M., Luo, Z.Q., Razaviyayn, M.: Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems (2014). Technical report, University of Minnesota
38. Houska, B., Frasch, J., Diehl, M.: An augmented lagrangian based algorithm for distributed nonconvex optimization. SIAM Journal on Optimization **26**(2), 1101–1127 (2016)
39. J.L. Fleiss B. Levin, M.C.P.J.F.: Statistical Methods for Rates & Proportions. Wiley (2003)
40. J.Wright, S.: Implementing proximal point methods for linear programming. Journal of Optimization Theory and Applications **65**(3), 531–554 (1990)
41. K. J. Arrow, L.H., Uzawa, H.: Studies in Linear and Non-linear Programming. Stanford University Press (1958)
42. Koppel, A., Sadler, B.M., Ribeiro, A.: Proximity without consensus in online multi-agent optimization. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3726–3730 (2016)
43. Koshal, J., Nedic, A., Shanbhag, Y.V.: Multiuser optimization: Distributed algorithms and error analysis. SIAM Journal on Optimization **21**(3), 1046–1081 (2011). DOI 10.1137/090770102
44. Lan, G., Monteiro, R.D.C.: Iteration-complexity of first-order augmented lagrangian methods for convex programming. Mathematical Programming **155**(1), 511–547 (2015)
45. Li, G., Pong, T.K.: Global convergence of splitting methods for nonconvex composite optimization. SIAM Journal on Optimization **25**(4), 2434–2460 (2015)
46. Liao, W.C., Hong, M., Farmanbar, H., Luo, Z.Q.: Semi-asynchronous routing for large-scale hierarchical networks. In: The Proceedings of IEEE ICASSP (2015)
47. Liavas, A.P., Sidiropoulos, N.D.: Parallel algorithms for constrained tensor factorization via alternating direction method of multipliers. IEEE Transactions on Signal Processing **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 Transactions on Automatic Control **56**(6), 1291–1306 (2011)
50. Lorenzo, P.D., Scutari, G.: Next: In-network nonconvex optimization. IEEE Transactions on Signal and Information Processing over Networks **2**(2), 120–136 (2016)
51. Lu, Z., Zhang, Y.: Sparse approximation via penalty decomposition methods. SIAM Journal on Optimization **23**(4), 2448–2478 (2013). DOI 10.1137/100808071
52. Mateos, G., Bazerque, J.A., Giannakis, G.B.: Distributed sparse linear regression. IEEE Transactions on Signal Processing **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, available at: arXiv:1702.01850
54. Nedic, A., Olshevsky, A.: Distributed optimization over time-varying directed graphs. IEEE Transactions on Automatic Control **60**(3), 601–615 (2015)

55. Nedic, A., Ozdaglar, A.: Distributed subgradient methods for multi-agent optimization. IEEE Transactions on Automatic Control **54**(1), 48–61 (2009)
56. Nedic, A., Ozdaglar, A., Parrilo, P.A.: Constrained consensus and optimization in multi-agent networks. IEEE Transactions on Automatic Control **55**(4), 922–938 (2010)
57. Nesterov, Y.: Introductory lectures on convex optimization: A basic course. Springer (2004)
58. Nocedal, J., Wright, S.J.: Numerical Optimization. Springer (1999)
59. Powell, M.M.D.: An efficient method for nonlinear constraints in minimization problems. In: Optimization. Academic Press (1969)
60. Razaviyayn, M., Hong, M., Luo, Z.Q.: A unified convergence analysis of block successive minimization methods for nonsmooth optimization. SIAM Journal on Optimization **23**(2), 1126–1153 (2013)
61. Rockafellar, R.T.: Augmented lagrangians and applications of the proximal point algorithm in convex programming. Mathematics of operations research **1**(2), 97–116 (1976)
62. Ruszczynski, A.: Nonlinear optimization. Princeton University (2011)
63. Schizas, I., Ribeiro, A., Giannakis, G.: Consensus in ad hoc wsns with noisy links - part i: Distributed estimation of deterministic signals. IEEE Transactions on Signal Processing **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 Transactions on Signal Processing **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 Journal on Optimization **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 (1984). Ph.D. thesis, Massachusetts Institute of Technology
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., W. Yin, J.Z.: Global convergence of admm in nonconvex nonsmooth optimization (2015). ArXiv Preprint, arXiv:1511.06324
70. Wen, Z., Yang, C., Liu, X., Marchesini, S.: Alternating direction methods for classical and ptychographic phase retrieval. Inverse Problems **28**(11), 1–18 (2012)
71. Yildiz, M.E., Scaglione, A.: Coding with side information for rate-constrained consensus. IEEE Transactions on Signal Processing. **56**(8), 3753–3764 (2008)
72. Zhang, C.H.: Nearly unbiased variable selection under minimax concave penalty. Ann. Statist. **38**(2), 894–942 (2010)
73. Zhang, Y.: Convergence of a class of stationary iterative methods for saddle point problems (2010). Preprint
74. Zhu, H., Cano, A., Giannakis, G.: Distributed consensus-based demodulation: algorithms and error analysis. IEEE Transactions on Wireless Communications **9**(6), 2044–2054 (2010)