Achieving Convergence Rates for Distributed Constrained and Unconstrained Optimization in Directed Networks

Farzad Yousefian  
School of Industrial Engineering & Management  
Oklahoma State University  
Stillwater, OK 74078, USA  
farzad.yousefian@okstate.edu

Abstract

We consider the problem of distributed optimization over directed networks where the agents communicate their information locally to their neighbors to cooperatively minimize a global cost. The existing theory and algorithms for this class of multi-agent problems are founded on two strong assumptions: (1) The problem is often assumed to be unconstrained, e.g., this is the case in Push-DIGing and Push-Pull algorithms that have been recently developed. (2) The objective function is assumed to be strongly convex, an assumption that appears in many of the existing models and algorithms in the area of multi-agent optimization. In this work, we aim at addressing these shortcomings. We first introduce a new unifying distributed constrained optimization model that is characterized as a bilevel optimization problem. The proposed model captures a wide range of existing problems including the following two classes: (i) Distributed linearly constrained optimization over directed networks. (ii) Distributed unconstrained non-strongly convex optimization over directed networks. To address the proposed optimization model, utilizing a regularization-based relaxation approach, we develop a new push-pull gradient method where at each iteration, the information of iteratively regularized gradients is pushed to the neighbors, while the information about the decision variable is pulled from the neighbors. The analysis of the proposed algorithm leads to two main contributions: (1) For model (i), we derive new convergence rates for suboptimality, infeasibility, and consensus violations. (2) For model (ii), we derive a new convergence rate statement despite the absence of the strong convexity property. The numerical performance of the proposed algorithm is presented.

1 Introduction

We consider a new class of distributed optimization problems in directed networks given as follows:

$$\min_{x \in \mathbb{R}^n} \sum_{i=1}^{m} f_i(x) \quad \text{subject to:} \quad x \in \arg\min_{x \in \mathbb{R}^n} \left\{ \sum_{i=1}^{m} g_i(x) \right\}, \quad (1)$$

under the following assumptions (see the Notation for details):

Assumption 1. (a) Functions $f_i : \mathbb{R}^n \rightarrow \mathbb{R}$ are $\mu_f$-strongly convex and $L_f$-smooth for all $i \in [m]$. (b) Functions $g_i : \mathbb{R}^n \rightarrow \mathbb{R}$ are convex and $L_g$-smooth for all $i \in [m]$. (c) The set $\arg\min_{x \in \mathbb{R}^n} \left\{ \sum_{i=1}^{m} g_i(x) \right\}$ is nonempty.

Here, $m$ agents cooperatively seek to find among the optimal solutions to the problem $\min_{x \in \mathbb{R}^n} \sum_{i=1}^{m} g_i(x)$, one that minimizes a secondary metric, i.e., $\sum_{i=1}^{m} f_i(x)$. Here, functions...
where

\[ f_i \text{ and } g_i \text{ are known locally only by agent } i \text{ and the cooperation among the agents occurs over a directed network. Given a set of nodes } \mathcal{N}, \text{ a directed graph (digraph) is denoted by } \mathcal{G} = (\mathcal{N}, \mathcal{E}) \text{ where } \mathcal{E} \subseteq \mathcal{N} \times \mathcal{N} \text{ is the set of ordered pairs of vertices. For any edge } (i, j) \in \mathcal{E}, \text{ i and j are called parent node and child node, respectively. Graph } \mathcal{G} \text{ is called strongly connected if there is a path between the pair of any two different vertices. The digraph induced by a given nonnegative matrix } \mathbf{B} \in \mathbb{R}^{m \times m} \text{ is denoted by } \mathcal{G}_\mathbf{B} = (\mathcal{N}_\mathbf{B}, \mathcal{E}_\mathbf{B}), \text{ where } \mathcal{N}_\mathbf{B} \triangleq [m] \text{ and } (j, i) \in \mathcal{E}_\mathbf{B} \text{ if and only if } \mathbf{B}_{ij} > 0. \text{ We let } \mathcal{N}_\mathbf{B}^{\text{in}}(i) \text{ and } \mathcal{N}_\mathbf{B}^{\text{out}}(i) \text{ denote the set of parents (in-neighbors) and the set of children (out-neighbors) of vertex } i, \text{ respectively. Also, } \mathcal{R}_\mathcal{G} \text{ denotes the set of roots of all possible spanning trees in } \mathcal{G}. \]

**Significance of the problem formulation:** Problem (1) provides a unifying mathematical framework capturing different variants of existing problems in the distributed optimization literature. From these, we present two important cases below:

(i) Distributed linearly constrained optimization in directed networks: Consider the model given as:

\[
\min_{x \in \mathbb{R}^n} \sum_{i=1}^m f_i(x) \quad \text{subject to: } A_i x = b_i \quad \text{for all } i \in [m],
\]

where \( A_i \in \mathbb{R}^{m_i \times n} \) and \( b_i \in \mathbb{R}^{m_i} \) are known parameters. Let problem (2) be feasible. Then, by choosing \( g_i(x) := \frac{1}{2} \| A_i x - b_i \|_2^2 \) for \( i \in [m] \), problem (2) is equivalent to (1).

(ii) Distributed unconstrained optimization in the absence of strong convexity: Let us define \( f_i(x) := \| x \|_2^2 / m \). Then, problem (1) is equivalent to finding the least \( \ell_2 \)-norm solution of the following canonical distributed unconstrained optimization problem:

\[
\min_{x \in \mathbb{R}^n} \sum_{i=1}^m g_i(x),
\]

where the component functions \( g_i \) are all merely convex and smooth.

**Existing distributed optimization models and algorithms:** The classical mathematical models, tools, and algorithms for consensus-based optimization were introduced and studied as early as the ‘70s [14] and ‘80s [39, 40, 6]. Of these, in the seminal work of Tsitsiklis [39], it was assumed the agents share a global (smooth) objective while their decision component vectors are distributed locally over the network. In the past two decades, in light of the unprecedented growth in data and its imperative role in several broad fields such as social networks, biology, and medicine, the theory of distributed and parallel optimization over networks has much evolved. The distributed optimization problems with local objective functions were first studied in [22, 28]. In this framework, the agents communicate their local information with their neighbors in the network at discrete times to cooperatively minimize the global cost function. Without characterizing the communication rules explicitly, this model can be formulated as \( \sum_{i=1}^m f_i(x) \) subject to \( x \in \mathcal{X} \). Here, the local function \( f_i \) is known only to the agent \( i \) and \( \mathcal{X} \) denotes the system constraint set. This modeling framework captures a wide spectrum of decentralized algorithms in the areas of statistical learning, signal processing, sensor networks, control, and robotics [15, 32, 43, 18, 33, 12, 10]. Because of this, in the past decade, there has been a flurry of research focused on the design and analysis of fast and scalable computational methods to address applications in networks. Among these, average-based consensus methods are one of the most studied approaches. Here, the network is characterized with a stochastic matrix that is possibly time-varying. The underlying idea is that at a given time, each agent uses this matrix and obtains a weighted-average of its neighbors’ local variables. Then, the update is completed by performing a standard subgradient step for the agent. In the case where the model \( \sum_{i=1}^m f_i(x) \) is unconstrained, the seminal work [28] proposed the following scheme:

\[
x_i(k + 1) := \left( \sum_{j=1}^m W_{ij}(k)x_j(k) \right) - \alpha_k \nabla f_i \left( x_i(k) \right),
\]

where \( x_i(k) \in \mathbb{R}^n, W_{ij}(k) \in [0, 1], \text{ and } \nabla f_i \) denote the local decision variable of agent \( i \) at time \( k \), the weight agent \( i \) assigns to the estimate from agent \( j \), and the subgradient of local function \( f_i \), respectively. Random projection variants of this scheme were developed for both synchronous and asynchronous cases, assuming \( \mathcal{X} \triangleq \bigcap_{i=1}^m \mathcal{X}_i \) [19, 56]. For the constrained model where \( \mathcal{X} = \mathbb{R}^n \), the algorithm EXTRA [34] and its proximal variant were developed addressing \( \mathcal{X} = \mathbb{R}^n \). EXTRA is a synchronous and time-invariant scheme and achieves a sublinear and a linear rate of convergence for smooth merely convex and strongly convex problems, respectively. Among many other works such as [21, 16], is the DiGing algorithm [27] which was the first work achieving a linear convergence rate for unconstrained
optimizaiton over a time-varying network. When the graph is directed, a key shortcoming in the weighted-based schemes lies in that the double stochasticity requirement of the weight matrix is impractical. Push-sum protocols were first leveraged in [25][26][29] to weaken this requirement. Recently, the Push-Pull algorithm equipped with a linear convergence rate was developed in [30] for unconstrained strongly convex problems. Extensions of push-sum algorithms to nonconvex regimes have been developed more recently [37][25][38]. Other popular distributed optimization schemes are the dual-based methods, such as ADMM-type methods studied in [7][24][41][20][35][3]. Most of these algorithms can address only static and undirected graphs. Moreover, there are only a few works in the literature that can cope with constraints employing primal-dual methods [9][24][8][2].

Research gap and contributions: Despite the aforementioned advances, the existing theory and algorithms for in-network optimization are founded on two strong assumptions: (1) The problem is often assumed to be unconstrained, e.g., this is the case in Push-DIGing [27] and Push-Pull [30] algorithms that have been recently developed. (2) The objective function is assumed to be strongly convex, an assumption that appears in many of the existing models and algorithms in the area of multi-agent optimization. In this work, we aim at addressing these shortcomings through considering the bilevel framework [1]. To address this model, utilizing a novel regularization-based relaxation approach, we develop a new push-pull gradient algorithm where at each iteration, the information of iteratively regularized (IR) rules that can tackle the inner-level constraints. In this paper, we consider a regularization-based Lagrangian duality does not seem applicable. Overcoming this challenge calls for new relaxation rules that can tackle the inner-level constraints. In this paper, we consider a regularization-based relaxation rule. To this end, motivated by the recent success of so-called iteratively regularized (IR) algorithms in centralized regimes [45][44][17][1], we develop Algorithm [1] that is the philosophy that the regularization parameter \( \lambda_k \) is updated after every step within the algorithm. Here, each agent holds a local copy of the global variable \( x \), denoted by \( x_{i,k} \), and an auxiliary variable...
Assumption 2. (a) The matrix $R = [R_{ij}] \in \mathbb{R}^{m \times m}$ and $C = [C_{ij}] \in \mathbb{R}^{n \times m}$ to update vectors $x_{i,k}$ and $y_{i,k}$, respectively. Below, we state the main assumptions on these two weight mixing matrices.

Algorithm 1 Iteratively Regularized Push-Pull

1: **Input:** For all $i \in [m]$, agent $i$ sets step-size $\gamma_{i,0} \geq 0$, pulling weights $R_{ij} \geq 0$ for all $j \in \mathcal{N}_R(i)$, pushing weights $C_{ij} \geq 0$ for all $j \in \mathcal{N}_C(i)$, an arbitrary initial point $x_{i,0}, y_{i,0} \in \mathbb{R}^n$, an initial

2: for $k = 0, 1, \ldots, \#$ do

3: For all $i \in [m]$, agent $i$ receives (pulls) the vector $x_{j,k} - \gamma_{j,k} y_{j,k}$ from each agent $j \in \mathcal{N}_R(i)$, sends (pushes) $C_{i,j} y_{i,k}$ to each agent $\ell \in \mathcal{N}_C(i)$, and does the following updates:

4: $x_{i,k+1} := \sum_{j=1}^m R_{ij} (x_{j,k} - \gamma_{j,k} y_{j,k})$;

5: $y_{i,k+1} := \sum_{j=1}^m C_{i,j} y_{i,k} + \lambda_{k+1} \nabla f_i(x_{i,k+1}) - \nabla g_i(x_{i,k+1}) - \lambda_k \nabla f_i(x_{i,k})$;

6: end for

Importantly, Assumption 2 does not require the strong condition of a doubly stochastic matrix for communication in a directed network. In turn, utilizing a push-pull protocol and in a similar fashion to the recent work \cite{zhang2019}, it only entails a row stochastic $R$ and a column stochastic matrix $C$. An example is as follows where agent $i$ chooses scalars $r_i, c_i > 0$ and:

$$R_{i,j} = \begin{cases} 1/(|\mathcal{N}_R(i)| + r_i), & \text{if } j \in \mathcal{N}_R(i) \\ r_i/(|\mathcal{N}_R(i)| + r_i), & \text{if } j = i \\ 0, & \text{otherwise} \end{cases}, \quad C_{i,j} = \begin{cases} 1/(|\mathcal{N}_C(i)| + c_i), & \text{if } \ell \in \mathcal{N}_C(i) \\ c_i/(|\mathcal{N}_C(i)| + c_i), & \text{if } \ell = i \\ 0, & \text{otherwise} \end{cases}. \quad (5)$$

Note that Assumption 2(b) is weaker than imposing strong connectivity on $\mathcal{G}_R$ and $\mathcal{G}_C$. The update rules in Algorithm 1 can be compactly represented as the following:

$$x_{k+1} := R(x_k - \gamma_k y_k), \quad (6)$$

$$y_{k+1} := C y_{k+1} + \lambda_{k+1} \nabla f(x_{k+1}) - 2 \nabla g(x_{k+1}) - \lambda_k \nabla f(x_k), \quad (7)$$

where $\gamma_k$ is a nonnegative diagonal matrix defined as $\gamma_k \triangleq \text{diag} (\gamma_{1,k}, \ldots, \gamma_{m,k})$.

3 Preliminaries of convergence analysis

In this section, we provide the mathematical toolbox for the convergence and rate analysis of Algorithm 1. Under Assumption 2, there exit a unique nonnegative left eigenvector $u \in \mathbb{R}^m$ such that $u^T R = u^T$ and $u^T 1 = m$. Similarly, there exits a unique nonnegative right eigenvector $v \in \mathbb{R}^m$ such that $C v = v$ and $1^T v = m$ (cf. Lemma 1 in \cite{zhang2019}). Throughout the analysis, we use the following definitions for any integer $k \geq 0$ and the regularization parameter $\lambda_k > 0$:

$$x^* \triangleq \arg\min_{x \in \mathbb{R}^n} \{ f(x) \} \in \mathbb{R}^{1 \times n}, \quad x_{\lambda_k}^* \triangleq \arg\min_{x \in \mathbb{R}^n} \{ g(x) + \lambda_k f(x) \} \in \mathbb{R}^{1 \times n}, \quad (8)$$

$$G_k(x) \triangleq \nabla g(x) + \lambda_k \nabla f(x) \in \mathbb{R}^{m \times n}, \quad G_k(x) \triangleq \frac{1}{m} 1^T G_k(x) \in \mathbb{R}^{1 \times n}, \quad (9)$$

$$G_k(x) \triangleq G_k(1x^T) \in \mathbb{R}^{1 \times n}, \quad \bar{y}_k \triangleq \bar{y}_k(x_k) \in \mathbb{R}^{1 \times n}, \quad L_k \triangleq L_0 + \lambda_k L_0, \quad (10)$$

$$\bar{x}_k \triangleq \frac{1}{m} u^T x_k \in \mathbb{R}^{1 \times n}, \quad \bar{y}_k \triangleq \frac{1}{m} u^T y_k \in \mathbb{R}^{1 \times n}, \quad \Lambda_k \triangleq \left| 1 - \frac{\lambda_{k+1}}{\lambda_k} \right|. \quad (11)$$

Here, $x^*$ denotes the optimal solution of problem (1) and $x_{\lambda_k}^*$ is defined as the optimal solution to a regularized problem. Note that the strong convexity of $g(x) + \lambda_k f(x)$ implies that $x_{\lambda_k}^*$ exits and is a unique vector (cf. Proposition 1.1.2 in \cite{davis2017}). Also, under Assumption 2, the set $\arg\min g(x)$ is closed and convex. As such, from the strong convexity of $f$ and invoking Proposition 1.1.2 in \cite{davis2017} again, we
We state the following result from [30] introducing two matrix norms induced by matrices $G_k(x)$ denotes the regularized gradient matrix. The vector $\tilde{x}_k$ holds a weighted average of the local copies of the agent’s iterates. Next, we consider a family of update rules for the sequences of the step-size and the regularization parameter under which the convergence and rate analysis can be performed.

**Assumption 3 (Update rules).** Assume the step-size $\gamma_k$ and the regularization parameter $\lambda_k$ are updated satisfying: $\gamma_k := \gamma_{00}/(k+1)^2$ and $\lambda_k := \lambda_0/2(1+1/k)$ where $\gamma_{00} := \max_{j \in [m]} \gamma_{j,k}$ for $k \geq 0$, and $a$ and $b$ satisfy the following conditions: $0 < b < a<1$ and $a + b < 1$. Also, let $\alpha_k \geq \theta \gamma_k$ for $k \geq 0$ for some $\theta > 0$, where $\alpha_k \equiv \frac{1}{m} u^T \gamma_k \nu$.

The constant $\theta$ is Assumption 3 measures the size of the rage within which the agents in $\mathcal{R}_R \cap \mathcal{R}_C$ select their stepsizes. The condition $\alpha_k \geq \theta \gamma_k$ is satisfied in many cases including the case where all the agents choose strictly positive stepsizes (see Remark 4 in [30] for more details). In the following lemma, we list some of the main properties of the update rules in Assumption 3. We will make use of these properties in the convergence and rate analysis.

**Lemma 1 (Properties of the update rules).** Let Assumption 3 hold. Then, the following properties hold: $\{\lambda_k\}_{k=0}^\infty$ is a decreasing strictly positive sequence satisfying $\lambda_k \to 0$, $\frac{\lambda_{k}}{\lambda_{k-1}} \leq \frac{1}{1+k}$ for $k \geq 1$, where $\lambda_k$ is given by (11). Also, $\{\gamma_k\}_{k=0}^\infty$ is a decreasing strictly positive sequence such that $\gamma_k \to 0$ and $\frac{\gamma_k}{\lambda_k} \to 0$. Moreover, for any scalar $\tau > 0$, there exits an integer $K_\tau$ such that $(k+1)^{\frac{\gamma_k \lambda_k}{K_\tau - 1} \lambda_{k-1}} \leq 1 + \tau \gamma_k \lambda_k$ for all $k \geq K_\tau$.

**Proof.** See Appendix A.1

Next, we present a key property of the sequence $\{x_{\lambda_k}^*\}$ and quantify the error between two consecutive elements of the trajectory.

**Lemma 2 (Properties of Tikhonov trajectory).** Let Assumptions 1 and 2 hold. Then, we have: (i) The sequence $\{x_{\lambda_k}^*\}$ converges to the unique solution of problem (1), i.e., $x^*$. (ii) There exits a scalar $M > 0$ such that for any $k \geq 1$, we have $\|x_{\lambda_k}^* - x_{\lambda_{k-1}}^*\|_2 \leq \frac{M}{\mu_f} \lambda_{k-1}$.

**Proof.** See Appendix A.2

In the following, we state the properties of the regularized maps. These results will be used in finding suitable error bounds in the next section.

**Lemma 3.** Consider Algorithm 1. Let Assumptions 1 and 2 hold. For any $k \geq 0$, mappings $G_k$, $\tilde{g}_k$, and $\tilde{y}_k$ given by (9) and (10) satisfy the following relations: (i) We have that $\tilde{y}_k = G_k(x_k)$. (ii) We have $\tilde{g}_k(x_{\lambda_k}^*) = 0$. (iii) The mapping $\tilde{g}_k(x)$ is $(\lambda_k, \mu_f)$-strongly monotone and Lipschitz continuous with parameter $L_k$. (iv) We have $\|\tilde{y}_k - \tilde{g}_k\| \leq \frac{1}{\sqrt{m}} \|x_k - \overline{x}_k\|_2$ and $\|\tilde{g}_k\| \leq L_k \|\tilde{x}_k - x_{\lambda_k}^*\|_2$.

**Proof.** See Appendix A.3

We state the following result from [30] introducing two matrix norms induced by matrices $R$ and $C$.

**Lemma 4 (cf. Lemma 4 and Lemma 6 in [30]).** Let Assumption 2 hold. Then: (i) There exist matrix norms $\|\cdot\|_R$ and $\|\cdot\|_C$ such that for $\sigma_R \equiv \frac{\|\cdot\|_{R}}{\|\cdot\|_{C}}$, and $\sigma_C \equiv \frac{\|\cdot\|_{C}}{\|\cdot\|_{R}}$, we have that $\sigma_R < 1$ and $\sigma_C < 1$. (ii) There exit scalars $\delta_{R,2}, \delta_{C,2}, \delta_{R,C}, \delta_{C,R} > 0$ such that for any $W \in \mathbb{R}^{n \times n}$, we have $\|W\|_R \leq \delta_{R,2} \|W\|_2$, $\|W\|_C \leq \delta_{C,2} \|W\|_2$, $\|W\|_R \leq \delta_{R,C} \|W\|_C$, $\|W\|_C \leq \delta_{C,R} \|W\|_R$, $\|W\|_2 \leq \|W\|_R$, and $\|W\|_2 \leq \|W\|_C$.

### 4 Convergence and rate analysis

In this section, we analyze the convergence of Algorithm 1 by introducing three errors metrics $\|\tilde{x}_{k+1} - x_{\lambda_k}^*\|_2$, $\|x_{k+1} - \overline{x}_{k+1}\|_R$, $\|y_{k+1} - \nu \tilde{y}_{k+1}\|_C$. Of these, the first term relates the averaged iterate with the Tikhonov trajectory, the second term measures the consensus violation for the decision.
matrix, and the third term measures the consensus violation for the matrix of the regularized gradients.

We begin by deriving a system of recursive relations for the three error terms provided below.

**Proposition 1.** Consider Algorithm 7 under Assumptions 7 and 8. Let $\alpha_k$ and $\hat{\gamma}_k$ be given by Assumption 3 and $c_0 = \delta_{C,2} \left\| \left(1 - \frac{1}{m} \nu^T \right)C \right\|_F$. Then, there exist scalars $M > 0$, $B_R > 0$, and an integer $K$ such that for any $k \geq K$, the following recursive relations hold:

\[
\left\| \bar{x}_{k+1} - x_{\lambda_k}^* \right\|_2 \leq \left(1 - \mu_f \alpha_k \lambda_k \right) \left\| \bar{x}_{k} - x_{\lambda_{k-1}}^* \right\|_2 + \frac{MA_{k-1}}{\mu_f} + \frac{\alpha_k L_k}{\sqrt{m}} \left\| x_k - 1 \bar{x}_k \right\|_R \\
+ \frac{\hat{\gamma}_k}{n} \left\| y_k - \nu \bar{y}_k \right\|_C,
\]

\[
\left\| x_{k+1} - 1 \bar{x}_{k+1} \right\|_R \leq \sigma_R \left(1 + \hat{\gamma}_k \sqrt{\frac{L_k}{\sqrt{m}}} \right) \left\| x_k - 1 \bar{x}_k \right\|_R + \sigma_R \hat{\gamma}_k \bar{\delta}_R \left\| y_k - \nu \bar{y}_k \right\|_C \\
+ \sigma_R \hat{\gamma}_k L_k \left\| \bar{y}_k \right\|_R \left\| \bar{x}_k - x_{\lambda_{k-1}}^* \right\|_2 + \frac{M \sigma_R \hat{\gamma}_k \left\| \nu \right\|_R \Lambda_{k-1}}{\mu_f},
\]

\[
\left\| y_{k+1} - \nu \bar{y}_{k+1} \right\|_C \leq \left(\sigma_C + c_0 L_k \hat{\gamma}_k \right) \left\| y_k - \nu \bar{y}_k \right\|_C \\
+ c_0 \left( \frac{\left\| R - I \right\|_2 + \hat{\gamma}_k \left\| \nu \right\|_2 \frac{L_k}{\sqrt{m}} + 2 \Lambda_k \right) \left\| x_k - 1 \bar{x}_k \right\|_2 \\
+ c_0 \left( \hat{\gamma}_k \left\| \nu \right\|_2 L_k + 2 \sqrt{m} \Lambda_k \right) \left\| x_k - x_{\lambda_{k-1}}^* \right\|_2 \\
+ c_0 \left( \hat{\gamma}_k \left\| \nu \right\|_2 \sqrt{m} \Lambda_k + \frac{\mu_f c_0 B_R}{M} \right) \frac{M}{\mu_f} \Lambda_{k-1}.
\]

**Proof.** Proof of (12): From (6) and (11), $\bar{x}_{k+1} = \frac{1}{m} u^T R (x_k - \gamma_k y_k) = \bar{x}_k - \frac{1}{m} u^T \gamma_k y_k$. Thus:

\[
\bar{x}_{k+1} = \bar{x}_k - \frac{1}{m} u^T \gamma_k (y_k - \nu \bar{y}_k) = \bar{x}_k - \frac{1}{m} u^T \gamma_k \nu \bar{y}_k = \frac{1}{m} u^T \gamma_k \nu \bar{y}_k
\]

From Assumption 3, there exits a $K$ such that $\alpha_k < \frac{\sqrt{2}}{L_0} \leq \frac{\sqrt{2}}{L_k}$ for $k \geq K$. From Lemma 3(iii), $g_k(x)$ is $(\mu_f \lambda_k)$-strongly convex and $L_k$-smooth. Invoking Lemma 10 in [31], we have for $\alpha_k \leq \frac{\sqrt{2}}{L_k}$ that $\left\| \bar{x}_k - \alpha_k \bar{g}_k - x_{\lambda_k}^* \right\|_2 \leq \max \left\{ (1 - \mu_f \lambda_k \alpha_k), (1 - L_k \alpha_k) \right\} \left\| \bar{x}_k - x_{\lambda_k}^* \right\|_2$. Thus, since $\mu_f \lambda_k \leq L_k$, we obtain for $\alpha_k \leq \frac{1}{L_k}$ that $\left\| \bar{x}_k - \alpha_k \bar{g}_k - x_{\lambda_k}^* \right\| \leq \left(1 - \mu_f \lambda_k \alpha_k \right) \left\| \bar{x}_k - x_{\lambda_k}^* \right\|$. Using the preceding two relations, we obtain:

\[
\left\| \bar{x}_{k+1} - x_{\lambda_k}^* \right\|_2 = \left\| \bar{x}_k - x_{\lambda_k}^* - \alpha_k \bar{g}_k - \alpha_k (\bar{y}_k - \bar{g}_k) - \frac{1}{m} u^T \gamma_k \nu \bar{y}_k \right\|_2
\]

\[
\leq \left\| \bar{x}_k - x_{\lambda_k}^* - \alpha_k \bar{g}_k \right\|_2 + \alpha_k \left\| \bar{y}_k - \bar{g}_k \right\|_2 + \frac{1}{m} \left\| u^T \gamma_k \nu \bar{y}_k \right\|_2
\]

\[
\leq \left(1 - \mu_f \alpha_k \lambda_k \right) \left\| \bar{x}_k - x_{\lambda_k}^* \right\|_2 + \alpha_k \left\| \bar{y}_k - \bar{g}_k \right\|_2 + \frac{1}{m} \left\| u^T \gamma_k \nu \bar{y}_k \right\|_2.
\]

By adding and subtracting $x_{\lambda_{k-1}}^*$ and using Lemmas 2 and 3(iv), we obtain:

\[
\left\| \bar{x}_{k+1} - x_{\lambda_k}^* \right\|_2 \\
\leq \left(1 - \mu_f \alpha_k \lambda_k \right) \left\| \bar{x}_k - x_{\lambda_{k-1}}^* \right\|_2 + \frac{MA_{k-1}}{\mu_f} + \alpha_k \left\| \bar{y}_k - \bar{g}_k \right\|_2 + \frac{1}{m} \left\| u^T \gamma_k \nu \bar{y}_k \right\|_2
\]

\[
\leq \left(1 - \mu_f \alpha_k \lambda_k \right) \left\| \bar{x}_k - x_{\lambda_{k-1}}^* \right\|_2 + \frac{MA_{k-1}}{\mu_f} + \frac{\alpha_k L_k}{\sqrt{m}} \left\| x_k - 1 \bar{x}_k \right\|_2 + \frac{\left\| \nu \right\|_2 \left\| \gamma_k \right\|_2 \left\| y_k - \nu \bar{y}_k \right\|_2}{m}.
\]

Then, inequality (12) is obtained by invoking Lemma 4, Remark 1 and definition of $\hat{\gamma}_k$.

**Proof of (13):** From (6) and (11) and that $R1 = 1$, we have:

\[
x_{k+1} - 1 \bar{x}_{k+1} = R (x_k - \gamma_k y_k) - 1 \bar{x}_k + \frac{1}{m} u^T \gamma_k y_k = \left( R - \frac{1}{m} u^T \gamma_k \right) ((x_k - 1 \bar{x}_k) - \gamma_k y_k).
\]
Applying Lemma 4, Remark 1 and Lemma 3, from the preceding relation we obtain:

$$\|x_{k+1} - \bar{x}_{k+1}\|_R \leq \sigma \|x_k - x_{k+1}\|_R + \sigma \|y_k - y_{k+1}\|_R + \sigma \|\gamma_k\|_2 \|y_k - y_{k+1}\|_R$$

Next, we estimate the two terms

$$\leq \sigma \|x_k - x_{k+1}\|_R + \sigma \|y_k - y_{k+1}\|_R + \sigma \|\gamma_k\|_2 \|y_k - y_{k+1}\|_R \left( \frac{L_k}{\sqrt{m}} \|x_k - x_{k+1}\|_2 + L_k \|\bar{x}_k - x^*_{k}\|_2 \right)$$

Adding and subtracting $x^*_{k+1}$ and using Lemma 2 we obtain the inequality (13).

**Proof of (14):** From (7) and the definition of $G_k(x)$ in (9), we obtain $y_{k+1} = Cy_k + G_{k+1}(x_{k+1}) - G_k(x_k)$. Multiplying both sides of the preceding relation by $\frac{1}{m} I$ and using the definition of $\bar{y}_k$ in (11), we obtain that $\bar{y}_{k+1} = \bar{y}_k + \frac{1}{m} I^2 G_{k+1}(x_{k+1}) - \frac{1}{m} I^2 G_k(x_k)$. From the last two relations, we have:

$$y_{k+1} - y_{k+1} = \left( C - \frac{1}{m} I^T \right) (y_k - y_{k+1}) + \left( I - \frac{1}{m} I^T \right) (G_{k+1}(x_{k+1}) - G_k(x_k)).$$

Invoking Lemma 4 the definition of $G_k(x)$ in (9) and $c_0$, and we obtain:

\begin{align*}
\|y_{k+1} - y_{k+1}\|_C & \leq \sigma \|y_k - y_{k+1}\|_C + c_0 \|G_{k+1}(x_{k+1}) - G_k(x_k)\|_2 \\
& \leq \sigma \|y_k - y_{k+1}\|_C + c_0 \|\lambda_{k+1} \nabla f(x_k) - \lambda_k \nabla f(x_k)\|_2 \\
& \leq \sigma \|y_k - y_{k+1}\|_C + c_0 \left( \|\nabla g(x_{k+1}) + \lambda_{k+1} \nabla f(x_k) - \lambda_k \nabla f(x_k)\|_2 + L_k \|x_k - x_{k+1}\|_2 \right) \\
& \leq \sigma \|y_k - y_{k+1}\|_C + c_0 \left( \|\nabla g(x_{k+1}) + \lambda_{k+1} \nabla f(x_k) - \lambda_k \nabla f(x_k)\|_2 + L_k \|x_k - x_{k+1}\|_2 \right) \\
& \leq \sigma \|y_k - y_{k+1}\|_C + c_0 \left( \|\nabla g(x_{k+1}) + \lambda_{k+1} \nabla f(x_k) - \lambda_k \nabla f(x_k)\|_2 + L_k \|x_k - x_{k+1}\|_2 \right) \\
& \leq 2L_k \|x_k - x_{k+1}\|_2 + \sqrt{m} L_k \|\bar{x}_k - x^*_{k}\|_2 + 2B_g.
\end{align*}

Next, we estimate the two terms $\|\lambda_k \nabla f(x_k)\|_2$ and $\|x_{k+1} - x_k\|_2$. From Lemma 2 there exists a scalar $B_g < \infty$ such that $L_g \|x^*_k - x^*\|_2 \leq B_g$. Since $\nabla g(x^*) = 0$, we have:

\begin{align*}
\|\lambda_k \nabla f(x_k)\|_2 & \leq \|\nabla g(x_k) + \lambda_k \nabla f(x_k)\|_2 + \|\nabla g(x_k) - \nabla g(1x^*)\|_2 \\
& \leq \|\nabla g(x_k) + \lambda_k \nabla f(x_k)\|_2 + \|\nabla g(x_k) - \nabla g(1x^*)\|_2 + L_g \|x_k - x_{k+1}\|_2 \\
& \leq (L_g + L_g) \|x_k - x_{k+1}\|_2 + 2L_k \|x_k - x_{k+1}\|_2 + 2B_g.
\end{align*}

From row-stochasticity of $R$, we have $(R - I) \bar{x}_k = 0$. Thus, from Lemma 3 we have:

\begin{align*}
\|x_{k+1} - x_k\|_2 &= \|R (x_k - \gamma_k y_k) - x_k\|_2 = \| (R - I) (x_k - \bar{x}_k) - R \gamma_k y_k\|_2 \\
& \leq \|R - I\|_2 \|x_k - x_{k}\|_2 + \|R\|_2 \|\gamma_k\|_2 \|y_k - y_{k}\|_2 + \|\nu\|_2 \|\bar{y}_k - y_{k}\|_2 \\
& \leq \|R - I\|_2 \|x_k - x_{k}\|_2 + \gamma_k \|R\|_2 \left( \|y_k - y_{k}\|_2 + L_k \|\nu\|_2 \right) \left( \|x_k - \bar{x}_k\|_2 \right) \\
& \leq \frac{M}{\mu_f} \Lambda_k - 1.
\end{align*}

From (13), the preceding three relations, we can obtain the desired relation (14).

Next, we derive a unifying recursive bound for the three error bounds in Proposition 1. This result will play a key role in deriving the rate statements.

**Proposition 2.** Consider Algorithm 7. Let Assumptions 2 and 3 hold. Let us define the error metric $\Delta_k \triangleq \left[ \left( \|\bar{x}_k - x^*_{k}\|_2 \right), \|x_k - 1x_k\|_R, \|y_k - y_{k}\|_C \right]^T$ for $k \geq 1$. Then, there exists an integer $K \geq 1$ such that for any $k \geq K$, the following holds:

(a) Let $\Theta \triangleq \max \left\{ c_0 L_o \left( \|\hat{y}_k\|_2 \|\nu\|_2 \|\bar{y}_k\|_2 \right), 1 \right\}$ and $M_k \triangleq \frac{\Theta \Lambda_k}{\mu_f}$. Then:

\begin{align*}
\|\Delta_{k+1}\|_2 & \leq \left( 1 - 0.5 \mu_f \alpha \lambda_k \right) \|\Delta_k\|_2 + \Theta \Lambda_k - 1.
\end{align*}

(b) There exists a scalar $B > 0$ such that $\|\Delta_k\|_2 \leq \frac{B}{\lambda_k^{1.5}}$.
Proof. See Appendix A.4

We next provide a family of convergence results for both suboptimality in both levels of the problem formulation (1).

Theorem 1 (Rate statements for the bilevel model). Consider problem (1) and Algorithm 2. Let Assumptions 1 and 3 hold. Then, we have the following results:

(a) We have \( \lim_{k \to \infty} \| x_k - x^* \| = 0 \). Also, the consensus violation of \( x_k \) and \( y_k \) characterized by \( \| x_{k+1} - 1 \bar{x}_{k+1} \|_{R} \) and \( \| y_{k+1} - \nu \bar{y}_{k+1} \|_{C} \), respectively, are both bounded by \( O \left( 1/k^{1-a-b} \right) \) for any sufficiently large \( k \).

(b) We have \( f(\bar{x}_k) - f(x^*) \leq \frac{Q_0(k_{max} + L_0_{max})}{2} \frac{1}{k^{2-2a-2b}} \) for some \( Q_1 > 0 \) and any sufficiently large \( k \).

(c) \( g(\bar{x}_k) - g(x^*) \leq \frac{Q_2(k_{max} + L_0_{max})}{2} \frac{1}{k^{2-2a-2b}} + \frac{aQ_3}{k^b} \) for \( Q_2, Q_3 > 0 \) and any sufficiently large \( k \).

Proof. See Appendix A.5

Corollary 1 (Rate statements for the linearly constrained model). Consider problem (2) and Algorithm 2. where we set \( g_i(x) := \frac{1}{2} \| A_i x - b_i \|_2^2 \). Let the feasible set be nonempty and Assumption 1(a) and Assumption 2 hold. Suppose Assumption 3 holds with \( a := 0.2 \) and \( b := 0.2 - \epsilon/3 \) where \( \epsilon > 0 \) is a sufficiently small scalar. Then, we have \( \lim_{k \to \infty} \bar{x}_k = x^* \) and for any sufficiently large \( k \):

(a) We have \( \| x_{k+1} - 1 \bar{x}_{k+1} \|_{R} = O \left( 1/k^{0.6+\epsilon/3} \right) \), and \( \| y_{k+1} - \nu \bar{y}_{k+1} \|_{C} = O \left( 1/k^{0.6+\epsilon/3} \right) \).

(b) We have \( f(\bar{x}_k) - f(x^*) = O \left( 1/k^{1-\epsilon} \right) \).

(c) \( \| A \bar{x}_k - b \|_2 = O \left( 1/k^{0.2-\epsilon/3} \right) \) where \( A \triangleq [A_1^T, \ldots, A_m^T]^T \) and \( b \triangleq [b_1^T, \ldots, b_m^T] \).

Proof. See Appendix A.6

Corollary 2 (Rate statements for the unconstrained non-strongly convex model). Consider problem (3) and Algorithm 2. where we set \( f_i(x) := \| x \|_2^2/m \). Let Assumption 1(b), 1(c) and Assumption 2 hold. Suppose Assumption 3 holds with \( a := 0.4 \) and \( b := 0.4 - \epsilon \) where \( \epsilon > 0 \) is a sufficiently small scalar. Let \( x_{r_2}^* \) denote the least \( \ell_2 \)-norm optimal solution of problem (3). Then, for any sufficiently large \( k \):

(a) We have \( \| x_{k+1} - 1 \bar{x}_{k+1} \|_{R} = O \left( 1/k^{0.4+\epsilon} \right) \) and \( \| y_{k+1} - \nu \bar{y}_{k+1} \|_{C} = O \left( 1/k^{0.4+\epsilon} \right) \).

(b) We have \( g(\bar{x}_k) - g(x_{r_2}^*) = O \left( 1/k^{0.4-\epsilon} \right) \) and that \( \| \bar{x}_k - x_{r_2}^* \|_2 = O \left( 1/k^{3\epsilon} \right) \).

Proof. See Appendix A.7

5 Numerical results

We present a numerical comparison of the performance of Algorithm 1 with that of Push-Pull algorithm [30]. Motivated by sensor network applications, we consider the unconstrained ill-posed problem \( \min_{x \in \mathbb{R}^n} \sum_{i=1}^{m} \| z_i - H_i x \|_2^2 \), where \( H_i \in \mathbb{R}^{d \times m} \) and \( z_i \in \mathbb{R}^d \) denote the measurement matrix and the noisy observation of the \( i \)-th sensor. Due to the challenges raised by ill-conditioning and also the lack of convergence and rate guarantees, Push-Pull algorithm needs to be applied to a regularized variant of the problem. To this end, in the implementation of the Push-Pull scheme, we use an \( \ell_2 \) regularizer with a parameter 0.1. Importantly, our comparison would lead to similar results with smaller choices of this parameter as well. Accordingly, in Algorithm 1 we set \( \lambda_0 := 0.1 \). We employ the tuning rules according to Corollary 2 while a constant step-size is used for the Push-Pull method. We generate \( H_i \) and \( z_i \) randomly and choose \( m = 10, n = 20, \) and \( d = 1 \). We generate matrices \( R \) and \( C \) from the same underlying graph with three different directed graphs (see Figure 1). For the star and ring graphs, we use the rule \( R = I - \frac{1}{n} \mathbf{L}_R \) where \( \mathbf{L}_R \) denotes the Laplacian matrix and \( e_{\max} \) denotes the maximum in-degree. We use the same formula for \( C \) using maximum out-degree. For the Lollipop graph, we generate matrices \( R \) and \( C \) using rules (5) with \( r_1 = c_i = 3 \) for all \( i \).

Insights: Figure 1 shows the comparison of the two schemes. We compare objective function values and consensus violations. For the latter, we use the term \( \| x_k - 11/10 \bar{x}_k \|_2 \). In terms of the objective
function value, Algorithm 1 performs significantly better in the star and ring cases. In the Lollipop case, while Algorithm 1 performs almost the same as Push-Pull in terms of the objective value, it achieves a lower consensus violation.

Acknowledgments and Disclosure of Funding

Farzad Yousefian gratefully acknowledges the support of the National Science Foundation through CAREER grant ECCS-1944500.

References

[1] M. Amini and F. Yousefian. An iterative regularized incremental projected subgradient method for a class of bilevel optimization problems. *Proceedings of the American Control Conference (ACC)*, pages 4069–4074, 2019.

[2] N. S. Aybat and E. Y. Hamedani. A primal-dual method for conic constrained distributed optimization problems. *Advances in Neural Information Processing Systems*, pages 5050–5058, 2016.

[3] N. S. Aybat, Z. Wang, T. Lin, and S. Ma. Distributed linearized alternating direction method of multipliers for composite convex consensus optimization. *IEEE Transactions on Automatic Control*, 63(1):5–20, 2017.

[4] D. P. Bertsekas. *Nonlinear Programming: 3rd Edition*. Athena Scientific, Bellmont, MA, 2016.

[5] D. P. Bertsekas, A. Nedich, and A. Ozdaglar. *Convex analysis and optimization*. Athena Scientific, Belmont, MA, 2003.

[6] D. P. Bertsekas and J. N. Tsitsiklis. *Parallel and Distributed Computation: Numerical Methods*. Athena Scientific, Belmont, MA, 1989.

[7] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2010.

[8] T. H. Chang. A proximal dual consensus ADMM method for multi-agent constrained optimization. *IEEE Transactions on Signal Processing*, 64(14):3719–3734, 2016.

[9] T. H. Chang, A. Nedich, and A. Scaglione. Distributed constrained optimization by consensus-based primal-dual perturbation method. *IEEE Transactions on Automatic Control*, 59(6):1524–1538, 2014.

[10] M. F. Duarte and Y. H. Hu. Vehicle classification in distributed sensor networks. *Journal of Parallel and Distributed Computing*, 64(7):826–838, 2014.

[11] J. C. Duchi, P. L. Bartlett, and Martin J. Wainwright. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization (SIOPT)*, 22(2):674–701, 2012.

[12] J. W. Durham, A. Franchi, and F. Bullo. Distributed pursuit-evasion without mapping or global localization via local frontiers. *Autonomous Robots*, 32(1):81—95, 2012.
[13] F. Facchinei and J.-S. Pang. *Finite-dimensional Variational Inequalities and Complementarity Problems. Vols. I, II.* Springer Series in Operations Research. Springer-Verlag, New York, 2003.

[14] M. H. De Groot. Reaching a consensus. *Journal of the American Statistical Association*, 69(345):118–121, 1974.

[15] A. Jadbabaie, J. Lin, and A. S. Morse. Coordination of groups of mobile autonomous agents using nearest neighbor rules. *IEEE Transactions on Automatic Control*, 48(6):988–1001, 2003.

[16] D. Jakovetić, J. Xavier, and J. M. Moura. Fast distributed gradient methods. *IEEE Transactions on Automatic Control*, 59(5):1131–1146, 2014.

[17] H. Kaushik and F. Yousefian. A randomized block coordinate iterative regularized subgradient method for high-dimensional ill-posed convex optimization. *Proceedings of the American Control Conference (ACC)*, pages 3420–3425, 2019.

[18] H. Lakshmanan and D. Farias. Decentralized recourse allocation in dynamic networks of agents. *SIAM Journal on Optimization*, 19(2):911–940, 2008.

[19] S. Lee and A. Nedich. Asynchronous gossip-based random projection algorithms over networks. *IEEE Transactions on Automatic Control*, 61(4):953–958, 2016.

[20] Q. Ling and A. Ribeiro. Decentralized dynamic optimization through the alternating direction method of multiplier. *IEEE Transactions on Signal Processing*, 62(5):1185–1197, 2014.

[21] I. Lobel, A. Ozdaglar, and D. Feijer. Distributed multi-agent optimization with state-dependent communication. *Mathematical Programming*, 129(2):255—284, 2011.

[22] C. G. Lopes and A. H. Sayed. Distributed processing over adaptive networks. In *2007 9th International Symposium on Signal Processing and Its Applications*, pages 1–3, 2007.

[23] P. Di Lorenzo and G. Scutari. Distributed nonconvex optimization over time-varying networks. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

[24] D. Mateos-Nunez and J. Cortes. Distributed subgradient methods for saddle-point problems. *IEEE Conference on Decision and Control (CDC)*, 2015.

[25] A. Nedich and A. Olshevsky. Distributed optimization over time-varying directed graphs. *IEEE Transactions on Automatic Control*, 60(3):601–615, 2015.

[26] A. Nedich and A. Olshevsky. Stochastic gradient-push for strongly convex functions on time-varying directed graphs. *IEEE Transactions on Automatic Control*, 61(12):3936–3947, 2016.

[27] A. Nedich, A. Olshevsky, and W. Shi. Achieving geometric convergence for distributed optimization over time-varying graphs. *SIAM Journal on Optimization*, 27(4):2597–2633, 2017.

[28] A. Nedich and A. Ozdaglar. Subgradient methods for saddle-point problems. *Journal of Optimization Theory and Applications*, 142(1):205–228, 2009.

[29] S. Pu and A. Nedich. Distributed stochastic gradient tracking methods. *Mathematical Programming*, 2020.

[30] S. Pu, W. Shi, J. Xu, and A. Nedich. Push-pull gradient methods for distributed optimization in networks. *IEEE Transactions on Automatic Control*, 2020.

[31] G. Qu and N. Li. Harnessing smoothness to accelerate distributed optimization. *IEEE Transactions on Control of Network Systems*, 5(3):1245–1260, 2018.

[32] M. Rabbat and R. D. Nowak. Distributed optimization in sensor networks. *The International Conference on Information Processing in Sensor Networks*, pages 20–27, 2004.

[33] S. S. Ram, V. V. Veeravalli, and A. Nedich. Distributed non-autonomous power control through distributed convex optimization. *IEEE International Conference on Computer Communications*, page 3001–3005, 2009.

[34] W. Shi, Q. Ling, G. Wu, and W. Yin. EXTRA: an exact first-order algorithm for decentralized consensus optimization. *SIAM Journal on Optimization*, 25(2):5944–5966, 2015.

[35] W. Shi, Q. Ling, K. Yuan, G. Wu, and W. Yin. On the linear convergence of the ADMM in decentralized consensus optimization. *IEEE Transactions on Signal Processing*, 62(7):1750–1761, 2014.

[36] K. Srivastava and A. Nedich. Distributed asynchronous constrained stochastic optimization. *IEEE Journal of Selected Topics in Signal Processing*, 5(4):772—90, 2011.
[37] Y. Sun, G. Scutari, and D. Palomar. Distributed nonconvex multiagent optimization over time-varying networks. In 2016 50th Asilomar Conference on Signals, Systems and Computers, pages 788–794, 2016.

[38] T. Tatarenki and B. Touri. Non-convex distributed optimization. IEEE Transactions on Automatic Control, 62(8):3744–3757, 2017.

[39] J. N. Tsitsiklis. Problems in Decentralized Decision Making and Computation. PhD thesis, Dept. of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, 1984.

[40] J. N. Tsitsiklis, D. P. Bertsekas, and M. Athans. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. IEEE Transactions on Automatic Control, 31(9):803—-812, 1986.

[41] E. Wei and A. Ozdaglar. Distributed alternating direction method of multipliers. IEEE Conference on Decision and Control (CDC), pages 5445–5450, 2012.

[42] E. Wei and A. Ozdaglar. On the O(1/k) convergence of asynchronous distributed alternating direction method of multipliers. In 2013 IEEE Global Conference on Signal and Information Processing, pages 551–554, 2013.

[43] L. Xiao, S. Boyd, and S.-J. Kim. Distributed average consensus with least-mean-square deviation. Journal of Parallel and Distributed Computing, 67(1):33–46, 2007.

[44] F. Yousefiyan, A. Nedich, and U. V. Shanbhag. On smoothing, regularization, and averaging in stochastic approximation methods for stochastic variational inequality problems. Mathematical Programming, 165(1):391–431, 2017.

[45] F. Yousefiyan, A. Nedich, and U. V. Shanbhag. On stochastic and deterministic quasi-Newton methods for nonstrongly convex optimization: Asymptotic convergence and rate analysis. SIAM Journal on Optimization, 30(2):1144–1172, 2020.
Appendices

Appendix A  Supplementary proofs

A.1  Proof of Lemma 1
Recall that \( \hat{\gamma}_k = \frac{\gamma_k}{(k+1)^2} \) and \( \lambda_k = \frac{\lambda_0}{(k+1)^a} \) where \( 0 < b < a < 1 \) and \( a + b < 1 \). Consequently, \( \{\hat{\gamma}_k\}_{k=0}^\infty \) and \( \{\lambda_k\}_{k=0}^\infty \) are strictly positive decreasing sequences and \( \hat{\gamma}_k \to 0, \lambda_k \to 0, \text{ and } \frac{\lambda_k}{\lambda_k} \to 0. \)
Next, we show that \( \Lambda_{k-1} \leq \frac{1}{k+1} \) for \( k \geq 1 \). From (11) and that \( \lambda_k \leq \lambda_{k-1} \), for any \( k \geq 1 \) we have:
\[
\Lambda_{k-1} = 1 - \frac{\lambda_k}{\lambda_{k-1}} = 1 - \frac{\lambda_0(k+1)^{-b}}{\lambda_0k^{-b}} = 1 - \left( \frac{k}{k+1} \right)^b = 1 - \left( 1 - \frac{1}{k+1} \right)^b. \tag{16}
\]
From 0 < \( b < a \) and \( a + b < 1 \), we have \( b < 0.5 \). This implies that \( \left( 1 - \frac{1}{k+1} \right)^b \geq \left( 1 - \frac{1}{k+1} \right)^{0.5} \).
Combining this relation with (16), we have:
\[
\Lambda_{k-1} \leq 1 - \left( 1 - \frac{1}{k+1} \right)^{0.5} = 1 - \left( \frac{1}{1 + \sqrt{1 - \frac{1}{k+1}}} \right) \frac{1}{k+1} \leq \frac{1}{k+1},
\]
where the last inequality is implied from \( k \geq 1 \). Next, we show \( \Lambda_{k+1} \leq \Lambda_k \) for all \( k \geq 0 \). From (16), we have:
\[
\Lambda_{k+1} = 1 - \left( 1 - \frac{1}{k+3} \right)^b \leq 1 - \left( 1 - \frac{1}{k+2} \right)^b = \Lambda_k.
\]
Next, we show that for any scalar \( \tau > 0 \), there exists an integer \( K_\tau \) such that \( \frac{(k+1)^2 \gamma_k}{\gamma_{k-1} \lambda_k} \leq 1 + \tau \hat{\gamma}_k \lambda_k \) for all \( k \geq K_\tau \). It suffices to show \( \lim_{k \to \infty} \frac{(k+1)^2 \gamma_k}{\gamma_{k-1} \lambda_{k-1}} - 1 \frac{1}{\gamma_k \lambda_k} = 0 \). From update formulas of \( \hat{\gamma}_k \) and \( \lambda_k \), we have:
\[
\frac{(k+1)^2 \gamma_k}{\gamma_{k-1} \lambda_{k-1}} - 1 \frac{1}{\gamma_k \lambda_k} = \frac{(k+1)^2 \gamma_k}{\gamma_{k-1} \lambda_{k-1}} \left( \frac{1}{\gamma_k \lambda_k} - 1 \right) = \left( \frac{1}{\gamma_k \lambda_k} - 1 \right) \frac{(k+1)^a + b}{\gamma_0 \lambda_0} \leq \left( \frac{1}{k} + 1 - a \right) \frac{2k^{a+b}}{\gamma_0 \lambda_0} \leq \left( 1 + \frac{1}{k} - a \right) \frac{2k^{a+b}}{\gamma_0 \lambda_0} = \frac{2}{\gamma_0 \lambda_0 k^{1-a-b}}.
\]
Therefore, \( \lim_{k \to \infty} \frac{(k+1)^2 \gamma_k}{\gamma_{k-1} \lambda_{k-1}} - 1 \frac{1}{\gamma_k \lambda_k} = 0 \), implying the existence of the specified integer \( K_\tau \).

A.2  Proof of Lemma 2
(a) From Assumption 1, the function \( g(x) + \lambda_k f(x) \) is \( (\lambda_k \mu_f) \)-strongly convex. Since \( x^*_k \) is the minimizer of this regularized function, we have:
\[
g(x^*) + \lambda_k f(x^*) \geq g(x^*_k) + \lambda_k f(x^*_k) + \frac{\lambda_k \mu_f}{2} \| x^* - x^*_k \|_2^2. \tag{17}
\]
Also, the definition \( x^* \) in (8) implies that \( x^* \in \text{argmin } g(x) \) and so \( g(x^*_k) \geq g(x^*) \). From the preceding two inequalities, we obtain:
\[
f(x^*) \geq f(x^*_k) + \frac{\mu_f}{2} \| x^* - x^*_k \|_2^2 \tag{18}
\]
for all \( k \geq 0 \).
Recall that under Assumption 1, \( x^* \) and \( x^*_k \) both exist and are unique vectors. Thus, we have that \( f(x^*) = f(x^*_k) = \infty \). Therefore, from relation (18), the sequence \( \{x^*_k\} \) is bounded implying that it must have at least one limit point. Let \( \{x^*_{k_n}\}_{k \in K} \) be an arbitrary subsequence such that \( \lim_{k \to \infty} k \in K x^*_{k_n} = \hat{x} \). Next, we show that \( \hat{x} \) is a feasible solution to problem (1). Consider relation (17). Taking the limit from both sides of (17), using the continuity of \( g \) and \( f \) induced from their
Writing the optimality conditions for (7), multiplying both sides of (7) by \( \left( \frac{\mu_f}{2} \right) \), we obtain:

\[
g(x^*) \geq g \left( \lim_{k \to \infty, \ k \in \mathbb{K}} x^*_{\lambda_k} \right) = g(\hat{x}),
\]

implying that \( \hat{x} \) is a minimizer of \( g(x) \) and so it is a feasible solution to problem \( \{1\} \). Next, we show that \( \hat{x} \) is indeed the optimal solution of problem \( \{1\} \). Taking the limit from both sides of (18), using the continuity of \( f \), and neglecting the nonnegative term \( \frac{\mu_f}{2} \left\| x^* - x^*_{\lambda_k} \right\|_2^2 \), we obtain:

\[
f(x^*) \geq f \left( \lim_{k \to \infty, \ k \in \mathbb{K}} x^*_{\lambda_k} \right) = f(\hat{x}).
\]

Hence, from the uniqueness of \( x^* \) we conclude that all limit points of the Tikhonov trajectory are identical to \( x^* \). Therefore, \( \lim_{k \to \infty} x^*_{\lambda_k} = x^* \).

**(b)** If \( x^*_{\lambda_k} = x^*_{\lambda_k-1} \), the desired relation holds. Consider the case where \( x^*_{\lambda_k} \neq x^*_{\lambda_k-1} \) for \( k \geq 1 \). Writing the optimality conditions for \( x^*_{\lambda_k} \) and \( x^*_{\lambda_k-1} \), for any \( x,y \in \mathbb{R}^n \), we have:

\[
g(x) + \lambda_{k-1} f(x) \geq g \left( x^*_{\lambda_{k-1}} \right) + \lambda_{k-1} f \left( x^*_{\lambda_{k-1}} \right) + \frac{\lambda_{k-1} \mu_f}{2} \left\| x - x^*_{\lambda_{k-1}} \right\|_2^2,
\]

\[
g(y) + \lambda_k f(y) \geq g \left( x^*_{\lambda_k} \right) + \lambda_k f \left( x^*_{\lambda_k} \right) + \frac{\lambda_k \mu_f}{2} \left\| y - x^*_{\lambda_k} \right\|_2^2.
\]

By substituting \( x := x^*_{\lambda_k} \) and \( y := x^*_{\lambda_k-1} \) and adding the resulting two relations together, we have:

\[
(\lambda_{k-1} - \lambda_k) \left( f \left( x^*_{\lambda_k} \right) - f \left( x^*_{\lambda_{k-1}} \right) \right) \geq \frac{\mu_f (\lambda_k + \lambda_{k-1})}{2} \left\| x^*_{\lambda_{k-1}} - x^*_{\lambda_k} \right\|_2^2.
\]

From the convexity property of \( f \), we have that:

\[
f \left( x^*_{\lambda_k} \right) - f \left( x^*_{\lambda_{k-1}} \right) \leq \left( x^*_{\lambda_k} - x^*_{\lambda_{k-1}} \right)^T \nabla f \left( x^*_{\lambda_k} \right).
\]

From the preceding two inequalities and that \( \lambda_{k-1} > \lambda_k \), we obtain:

\[
\frac{\mu_f \lambda_{k-1}}{2} \left\| x^*_{\lambda_{k-1}} - x^*_{\lambda_k} \right\|_2^2 \leq (\lambda_{k-1} - \lambda_k) \left( x^*_{\lambda_k} - x^*_{\lambda_{k-1}} \right)^T \nabla f \left( x^*_{\lambda_k} \right).
\]

Dividing both sides of the by \( \lambda_{k-1} \) and using Cauchy-Schwarz inequality, we have:

\[
\left\| x^*_{\lambda_{k-1}} - x^*_{\lambda_k} \right\|_2^2 \leq \frac{2}{\mu_f} \left( 1 - \frac{\lambda_k}{\lambda_{k-1}} \right) \left\| x^*_{\lambda_k} - x^*_{\lambda_{k-1}} \right\|_2 \left\| \nabla f \left( x^*_{\lambda_k} \right) \right\|_2.
\]

Note that since \( x^*_{\lambda_k} \neq x^*_{\lambda_{k-1}} \), we have \( \left\| x^*_{\lambda_k} - x^*_{\lambda_{k-1}} \right\|_2 \neq 0 \). Thus, we obtain:

\[
\left\| x^*_{\lambda_{k-1}} - x^*_{\lambda_k} \right\|_2 \leq \frac{2}{\mu_f} \left( 1 - \frac{\lambda_k}{\lambda_{k-1}} \right) \left\| \nabla f \left( x^*_{\lambda_k} \right) \right\|_2 = \frac{2 \lambda_{k-1}}{\mu_f} \left\| \nabla f \left( x^*_{\lambda_k} \right) \right\|_2.
\]

From part (a), \( \{ x^*_{\lambda_k} \}_{k=0}^\infty \) is a bounded sequence. Thus, there exits a compact ball \( \mathcal{X} \subset \mathbb{R}^n \) such that \( \{ x^*_{\lambda_k} \}_{k=0}^\infty \subseteq \mathcal{X} \). From continuity of the mapping \( \nabla f \), there exists a constant \( M > 0 \) such that:

\[
2 \left\| \nabla f \left( x^*_{\lambda_k} \right) \right\|_2 \leq M \quad \text{for all } k \geq 0.
\]

Combining the preceding two relations, we obtain the desired inequality.

### A.3 Proof of Lemma 3

**(i)** Multiplying both sides of (7) by \( \frac{1}{m} 1^T \) and from the definitions of \( G_k \) and \( G_{k+1} \) in (9), we obtain:

\[
\bar{y}_{k+1} = \frac{1}{m} 1^T y_k + \frac{1}{m} 1^T G_{k+1} (x_{k+1}) - \frac{1}{m} 1^T G_k (x_k) = \bar{y}_k + G_{k+1} (x_{k+1}) - G_k (x_k),
\]

where we used \( 1^T C = 1^T \). From Algorithm 1, we have \( y_0 := \nabla g(x_0) + \lambda_0 \nabla f(x_0) = G_0 (x_0) \), implying that \( y_0 = G_0 (x_0) \). From the two preceding relations, we obtain that \( \bar{y}_k = G_k (x_k) \).
From [9], we have that \( G_k(x) = \frac{1}{m} \sum_{i=1}^{m} (\nabla g_i(x) + \lambda_k \nabla f_i(x)) \). Thus, from the definition of \( \mathbb{G}_k \) we obtain that \( \mathbb{G}_k(x) = G_k(1_{x}^T) = \frac{1}{m} \sum_{i=1}^{m} (\nabla g_i(x) + \lambda_k \nabla f_i(x)) = \frac{1}{m} (\nabla g(x) + \lambda_k \nabla f(x)) \). Thus, from the definition of \( e_{\lambda_k}^* \) in [8], we obtain \( \mathbb{G}_k(e_{\lambda_k}^*) = 0 \). Also, from Assumption [1], we conclude that \( \mathbb{G}_k(x) \) is a \((\lambda_k \mu_f)-\)strongly monotone mapping and Lipschitz continuous with parameter \( L_k \triangleq L_g + \lambda_k L_f \) for \( k \geq 0 \).

For any \( u, v \in \mathbb{R}^{m \times n} \), with \( u_i, v_i \in \mathbb{R}^n \) denoting the \( i \)-th row of \( u, v \), respectively, we have:

\[
\| G_k(u) - G_k(v) \|_2 = \left\| \frac{1}{m} u^T \left( \nabla g(u) + \lambda_k \nabla f(u) \right) - \frac{1}{m} v^T \left( \nabla g(v) + \lambda_k \nabla f(v) \right) \right\|_2 \\
\leq \frac{1}{m} \left\| \sum_{i=1}^{m} \nabla g_i(u_i) - \sum_{i=1}^{m} \nabla g_i(v_i) \right\|_2 + \lambda_k \left\| \sum_{i=1}^{m} \nabla f_i(u_i) - \sum_{i=1}^{m} \nabla f_i(v_i) \right\|_2 \\
\leq \frac{1}{m} \sum_{i=1}^{m} \left( \| \nabla g_i(u_i) - \nabla g_i(v_i) \|_2 + \lambda_k \| \nabla f_i(u_i) - \nabla f_i(v_i) \|_2 \right) \\
\leq \frac{1}{m} \sum_{i=1}^{m} \left( L_g \| u_i - v_i \|_2 + \lambda_k L_f \| u_i - v_i \|_2 \right) \leq \frac{L_k}{\sqrt{m}} \sum_{i=1}^{m} \| u_i - v_i \|_2 \leq \frac{L_k}{\sqrt{m}} \| u - v \|_2.
\]

Consequently, we obtain \( \| \tilde{g}_k - \tilde{g}_k \|_2 = \| G_k(x_k) - G_k(1_{x_k}^T) \|_2 \leq \frac{L_k}{\sqrt{m}} \| x_k - 1_{x_k} \|_2 \). Also, using the Lipschitzian property of \( \mathbb{G}_k \) in part (ii) and \( \mathbb{G}_k(e_{\lambda_k}^*) = 0 \), we obtain:

\[
\| \tilde{g}_k \|_2 = \| G_k(\bar{x}_k) \|_2 = \| G_k(\bar{x}_k) - G_k(e_{\lambda_k}^*) \|_2 \leq L_k \| \bar{x}_k - e_{\lambda_k}^* \|_2.
\]

### A.4 Proof of Proposition [2]

**Proof.**

In the first step, from Proposition [1], we have for all \( k \geq K \) that \( \Delta_{k+1} \leq H_k \Delta_k + h_k \) where \( H_k = [H_{ij,k}]_{3 \times 3} \) and \( h_k = [h_{ij,k}]_{3 \times 3} \) are given as follows:

\[
H_{11,k} := 1 - \mu_f \alpha_k \lambda_k, \quad H_{12,k} := \frac{\alpha_k L_k}{\sqrt{m}}, \quad H_{13,k} := \frac{\gamma_k \| u \|_2^2}{m}, \\
H_{21,k} := \sigma_R \gamma_k L_k \| R \| R, \quad H_{22,k} := \sigma_R \left( 1 + \gamma_k \| R \| R \frac{L_k}{\sqrt{m}} \right), \quad H_{23,k} := \sigma_R \gamma_k \delta R, C, \\
H_{31,k} := c_0 L_k \left( \gamma_k \| R \| R \| u \|_2 L_k + 2 \sqrt{m} \Lambda_k \right), \\
H_{32,k} := c_0 L_k \left( \| R - I \| R \| R \| u \|_2 \frac{L_k}{\sqrt{m}} + 2 \Lambda_k \right), \quad H_{33,k} := \sigma_C + c_0 L_k \gamma_k \| R \| R, \\
h_{1,k} := \frac{M \Lambda_{k-1}}{\mu_f}, \quad h_{2,k} := \frac{M \sigma_R \gamma_k L_k \| R \| R \Lambda_{k-1}}{\mu_f}, \\
h_{3,k} := c_0 L_k \left( \gamma_k \| R \| R \| u \|_2 L_k + \sqrt{m} \Lambda_k + \frac{\mu_f c_0 B_k}{M} \right) \frac{M}{\mu_f} \Lambda_{k-1}.
\]

Let us define the sequence \( \{ \rho_k \} \) as \( \rho_k \triangleq 1 - 0.5 \mu_f \alpha_k \lambda_k \) for \( k \geq 0 \). Next, we the utilize our assumptions to find suitable upper bounds for some of the above terms. We define \( \dot{H}_k = [\dot{H}_{ij,k}]_{3 \times 3} \) and \( \dot{h}_k = [\dot{h}_{ij,k}]_{3 \times 3} \) as follows:

\[
\dot{H}_{11,k} := 1 - \mu_f \alpha_k \lambda_k, \quad \dot{H}_{12,k} := \frac{\alpha_k L_0}{\sqrt{m}}, \quad \dot{H}_{13,k} := \frac{\gamma_k \| u \|_2^2}{m}, \\
\dot{H}_{21,k} := \sigma_R \gamma_k L_0 \| R \| R, \quad \dot{H}_{22,k} := \rho_k - \frac{1}{2} \frac{\sigma R}{R}, \quad \dot{H}_{23,k} := \sigma_R \gamma_k \delta R, C, \\
\dot{H}_{31,k} := c_0 L_0 \left( \gamma_k \| R \| R \| u \|_2 L_0 + 2 \sqrt{m} \Lambda_k \right), \\
\dot{H}_{32,k} := c_0 L_0 \left( \| R - I \| R \| R \| u \|_2 \frac{L_0}{\sqrt{m}} + 2 \Lambda_0 \right), \quad \dot{H}_{33,k} := \rho_k - \frac{1}{2} \frac{\sigma_C}{C}, \\
\dot{h}_{1,k} := \frac{\Theta}{\sqrt{3}} \Lambda_{k-1}, \quad \dot{h}_{2,k} := \frac{\Theta}{\sqrt{3}} \Lambda_{k-1}, \quad \dot{h}_{3,k} := \frac{\Theta}{\sqrt{3}} \Lambda_{k-1}.
\]
Note that we have:

\[
\hat{H}_{22,k} - H_{22,k} = 1 - 0.5\mu_f \alpha_k \lambda_k - \frac{1 - \sigma R}{2} - \sigma R \left( 1 + \hat{\gamma}_k \|\nu\|_R \frac{L_k}{\sqrt{m}} \right) = \frac{1 - \sigma R}{2} - 0.5\mu_f \alpha_k \lambda_k - \hat{\gamma}_k \|\nu\|_R \frac{L_k}{\sqrt{m}}.
\]

From Assumption 3 and the definition of \( \alpha_k \), we have \( \hat{\gamma}_k \to 0 \), \( \alpha_k \to 0 \), and \( \lambda_k \to 0 \). Thus, there exists an integer \( K_R \geq 1 \) such that for all \( k \geq K_R \) we have \( H_{22,k} \leq \hat{H}_{22,k} \). Similarly, there exists an integer \( K_C \geq 1 \) such that for all \( k \geq K_C \) we have \( H_{33,k} \leq \hat{H}_{33,k} \). Thus, by taking to account that \( \lambda_k \) and \( \Lambda_k \) are nonincreasing sequences and invoking the definition of \( \Theta \), we have \( H_k \leq \hat{H}_k \) and \( h_k \leq \hat{h}_k \). This implies that for all \( k \geq \max \{ K, K_R, K_C \} \), we have \( \Delta_{k+1} \leq \hat{H}_k \Delta_k + \hat{h}_k \). Consequently, we obtain:

\[
\|\Delta_{k+1}\|_2 \leq \rho \left( \hat{H}_k \right) \|\Delta_k\|_2 + \Theta \Lambda_{k-1}, \tag{19}
\]

where \( \rho \left( \hat{H}_k \right) \) denotes the spectral norm of \( \hat{H}_k \). Next, we show that for a sufficiently large \( k \), we have that \( \rho \left( \hat{H}_k \right) \leq \rho_k \). To show this relation, employing Lemma 5 in [29], it suffices to show that \( 0 \leq \hat{H}_{i,k} < \rho_k \) for \( i \in \{ 1, 2, 3 \} \) and det \( (\rho_k I - \hat{H}_k) > 0 \). Among these, it can be easily seen that \( H_{i,k} < \rho_k \) holds for all \( i \in \{ 1, 2, 3 \} \). Since \( \alpha_k \to 0 \) and \( \lambda_k \to 0 \), there exists an integer \( K_1 \) such that \( \hat{H}_{11,k} = 1 - \mu_f \alpha_k \lambda_k > 0 \). Similarly, from \( \sigma_R < 1 \) and \( \sigma_C < 1 \), there exists integers \( K_2 \) and \( K_3 \) such that \( \hat{H}_{22,k} > 0 \) and \( \hat{H}_{33,k} > 0 \), respectively. Next, we show det \( (\rho_k I - \hat{H}_k) > 0 \). We have:

\[
det \left( \rho_k I - \hat{H}_k \right) = \prod_{i=1}^{3} \left( \rho_k - \hat{H}_{i,k} \right) - \left( \rho_k - \hat{H}_{11,k} \right) \hat{H}_{23,k} \hat{H}_{32,k} - \left( \rho_k - \hat{H}_{33,k} \right) \hat{H}_{12,k} \hat{H}_{21,k} = (0.5\mu_f \alpha_k \lambda_k) \left( \frac{1 - \sigma R}{2} \right) \left( \frac{1 - \sigma C}{2} \right) - (0.5\mu_f \alpha_k \lambda_k) (\sigma_R \hat{\gamma}_k \delta_{R,C}) c_0 L_0 \left( \|R - I\|_2 + \hat{\gamma}_k \|R\|_2 \|\nu\|_2 \frac{L_0}{\sqrt{m}} + 2\Lambda_0 \right) - \left( \frac{1 - \sigma C}{2} \right) \left( \frac{\alpha_k L_0}{\sqrt{m}} \right) (\sigma_R \hat{\gamma}_k L_0 \|\nu\|_R) - \left( \frac{\alpha_k L_0}{\sqrt{m}} \right) (\sigma_R \hat{\gamma}_k \delta_{R,C}) \left( c_0 L_0 \left( \|R - I\|_2 + \hat{\gamma}_k \|R\|_2 \|\nu\|_2 \right) \frac{L_0}{\sqrt{m}} + 2\Lambda_0 \right) - \left( \frac{1 - \sigma R}{2} \right) \left( \frac{c_0 L_0 \left( \hat{\gamma}_k \|R\|_2 \|\nu\|_2 \right) \frac{L_0}{\sqrt{m}} + 2\Lambda_0} \right) \right).
\]

Next, we find lower and upper bounds for \( \alpha_k \) in terms of \( \hat{\gamma}_k \). Note that Assumption 3 provides \( \theta \hat{\gamma}_k \) as a lower bound for \( \alpha_k \). To find an upper bound, from Lemma 1 in [30], we have that the eigenvector \( u \) is nonzero only on the entries \( i \in R_R \). Similarly, the eigenvector \( \nu \) is nonzero only on the entries \( i \in R_C^T \). Also, we have \( u^T \nu > 0 \). Thus, from the definition of \( \alpha_k \), we can write:

\[
\frac{\alpha_k}{\hat{\gamma}_k} = \frac{1}{m} u^T \frac{\gamma_k}{\gamma_k} \nu = \frac{1}{m} \sum_{i \in R_{R-C^T}} u_i v_i \frac{\gamma_i,k}{\gamma_k} \leq \frac{1}{m} \sum_{i \in R_{R-C^T}} u_i v_i \frac{\gamma_i,k}{\hat{\gamma}_k} = \frac{1}{m} u^T \nu > 0.
\]

Let us define \( \bar{\theta} = \frac{1}{m} u^T \nu \). Thus, we have \( \theta \hat{\gamma}_k \leq \alpha_k \leq \bar{\theta} \hat{\gamma}_k \) for all \( k \geq 0 \). Using these bounds and rearranging the terms, we can obtain:

\[
det \left( \rho_k I - \hat{H}_k \right) \geq -c_1 \hat{\gamma}_k^3 - c_2 \hat{\gamma}_k^2 + c_3 \hat{\gamma}_k \lambda_k - c_4 \hat{\gamma}_k \Lambda_k,
\]
where the scalars $c_1$, $c_2$, $c_3$ are defined as below:

\[
\begin{align*}
c_1 & \triangleq (0.5\mu_f \forall_0) (\sigma_{\mathbf{R}} \delta_{\mathbf{R}, \mathbf{C}}) c_0 L_0 \left( \| \mathbf{R} \|_2 L_0 \sqrt{\frac{L_0}{m}} \right) + \left( \frac{\| u \|_2}{m} \right) (\sigma_{\mathbf{R}} L_0 \| \nu \|_\mathbf{R}) \left( c_0 L_0 (\| \mathbf{R} \|_2 \| \nu \|_2 L_0) \right) \\
c_2 & \triangleq (0.5\mu_f \forall_0) (\sigma_{\mathbf{R}} \delta_{\mathbf{R}, \mathbf{C}}) c_0 L_0 (\| \mathbf{R} - \mathbf{I} \|_2 + 2 \Lambda_0) + \left( \frac{1 - \sigma_{\mathbf{C}}}{2} \right) \left( \frac{\theta L_0}{\sqrt{m}} \right) (\sigma_{\mathbf{R}} L_0 \| \nu \|_\mathbf{R}) \\
c_3 & \triangleq (0.5)^3 \mu_f \forall_0 (1 - \sigma_{\mathbf{R}}) (1 - \sigma_{\mathbf{C}}) \\
c_4 & \triangleq \left( \frac{1 - \sigma_{\mathbf{R}}}{2} \right) \left( c_0 L_0 \sqrt{\frac{m}{\| u \|_2}} \right).
\end{align*}
\]

It suffices to show that $-c_1 \gamma_k^3 - c_2 \gamma_k^2 + c_3 \gamma_k \lambda_k - c_4 \gamma_k \Lambda_k > 0$ for any sufficiently large $k$. From Lemma 1, we have $\hat{\Lambda}_k \rightarrow 0$. Thus, there exists an integer $K_4 \geq 0$ such that for any $k \geq K_4$ we have $c_4 \Lambda_k \leq 0.5c_3 \lambda_k$. As such, it suffices to show that $c_1 \gamma_k^3 + c_2 \gamma_k < 0.5c_3 \lambda_k$. From Lemma 1 since $\gamma_k \rightarrow 0$ and $\frac{\alpha}{\gamma_k} \rightarrow 0$, there exists an integer $K_5 \geq 0$ such that $c_1 \gamma_k^3 + c_2 \gamma_k < 0.5c_3 \lambda_k$ for any $k \geq K_5$. We conclude that for $\mathcal{K} \triangleq \max \{ K, K_1, K_2, K_3, K_4, K_5, k, K_{\hat{\Gamma}, \mathbf{R}, \mathbf{C}} \}$, we have $\det (\rho \mathbf{I} - \hat{H}_k) > 0$ for any $k \geq \mathcal{K}$. Therefore, we have $\rho (\hat{H}_k) \leq 1 - 0.5\mu_f \alpha_k \lambda_k$ for all $k \geq \mathcal{K}$. The desired inequality is obtained from this inequality and the relation (19).

(b) From Lemma 1, we have that $\Lambda_{k-1} \leq \frac{1}{k+1}$. From part (a) and Assumption 3, we obtain:

\[
\| \Delta_{k+1} \|_2 \leq (1 - 0.5\mu_f \alpha_k \gamma_k \theta) \| \Delta_k \|_2 + \frac{\Theta}{k+1} \quad \text{for all } k \geq \mathcal{K}.
\]

We use induction to show that the desired inequality holds for:

\[
\mathcal{B} \triangleq \max \left\{ (\mathcal{K} + 1)^{1-a-b} \| \Delta_k \|_2, \frac{4\Theta}{\mu_f \lambda_0 \gamma_0 \theta} \right\}.
\]

First, we observe that the inequality holds for $k := \mathcal{K}$. This is because:

\[
\| \Delta_\mathcal{K} \|_2 = \left( (\mathcal{K} + 1)^{1-a-b} \right) \frac{1}{(\mathcal{K} + 1)^{1-a-b}} \leq \frac{\mathcal{B}}{(\mathcal{K} + 1)^{1-a-b}}.
\]

Let us assume that $\| \Delta_k \|_2 \leq \frac{\mathcal{B}}{k^{1-a-b}}$ for some $k \geq \mathcal{K}$. We show that this relation also holds for $k + 1$. Consider Lemma 1. Let us choose $\tau := \frac{\mu_f \theta}{4}$. Thus, from Lemma 1 there exits a $K_r$ such that $\frac{(k+1)\gamma_k \lambda_k}{\kappa \alpha_k \lambda_{k-1} \Lambda_{k-1}} \leq 1 + \tau \gamma_k \lambda_k$ for all $k \geq K_r$. This implies that:

\[
\frac{k^{a+b}}{k} \leq \frac{(k+1)^{a+b}}{k+1} (1 + 0.25\mu_f \lambda_0 \gamma_0 \theta).
\]

Let $K_6$ be an integer such that $0.5\mu_f \alpha_k \gamma_k \theta < 1$. Without loss of generality, let us assume $\mathcal{K} \geq \max \{ K_r, K_6 \}$. From (20) and the induction hypothesis, we obtain:

\[
\| \Delta_{k+1} \|_2 \leq (1 - 0.5\mu_f \alpha_k \gamma_k \theta) \frac{\mathcal{B}}{k^{1-a-b}} + \frac{\Theta}{k+1}.
\]

From the preceding relation and (21), we obtain:

\[
\| \Delta_{k+1} \|_2 \leq (1 - 0.5\mu_f \alpha_k \gamma_k \theta)(1 + 0.25\mu_f \alpha_k \gamma_k \theta) \frac{\mathcal{B}}{(k+1)^{1-a-b}} + \frac{\Theta}{k+1}.
\]

From the definition of $\mathcal{B}$ we have $\Theta \leq 0.25\mu_f \lambda_0 \gamma_0 \theta \mathcal{B}$. From this relation and rearranging the terms in the preceding inequality, we obtain:

\[
\| \Delta_{k+1} \|_2 \leq \frac{\mathcal{B}}{(k+1)^{1-a-b}} (1 - 0.25\mu_f \lambda_0 \gamma_0 \theta - 0.125(\mu_f \lambda_0 \gamma_0 \theta)^2) + 0.25\mu_f \lambda_0 \gamma_0 \theta \mathcal{B}.
\]

This implies that $\| \Delta_{k+1} \|_2 \leq \frac{\mathcal{B}}{(k+1)^{1-a-b}}$. Thus, the induction statement holds for $k + 1$ and hence, the proof is completed.
A.5 Proof of Theorem 1

(a) From Lemma 2(a), we have that \(\{x^*_k\}\) converges to \(x^*\). Moreover, from Proposition 2(b), we have that \(\|\bar{x}_k - x^*\|_2\) converges to zero. Therefore, we have \(\lim_{k\to\infty} \bar{x}_k = x^*\). To derive the bounds for \(\|x_k - 1 \bar{x}_k\|_R\) and \(\|y_k - \nu \bar{y}_k\|_C\), from the definition of \(\Delta_k\) in Proposition 2, we can write:

\[
\|x_k - 1 \bar{x}_k\|_R \leq \|\Delta_k\|_2 = O \left( k^{1-a-b} \right).
\]

Similarly, we obtain \(\|y_k - \nu \bar{y}_k\|_C = O \left( k^{1-a-b} \right)\).

(b) Consider the regularized function \(g(x) + \lambda_k f(x)\). Note that it is \(L_k\)-smooth, where \(L_k = L_g + \lambda_k L_f\). Since \(x^*_k\) is the minimizer of \(g(x) + \lambda_k f(x)\), we have:

\[
g(x) + \lambda_k f(x) - g(x^*_k) - \lambda_k f(x^*_k) \leq \frac{L_k}{2} \|x - x^*_k\|_2^2 \quad \text{for all } x \in \mathbb{R}^n.
\]

Also, we can write that \(g(x^*_k) + \lambda_k f(x^*_k) \leq g(x^*) + \lambda_k f(x^*)\). Combining the preceding two relations and substituting \(x\) by \(\bar{x}_{k+1}\), we obtain:

\[
g(\bar{x}_{k+1}) - g(x^*) + \lambda_k (f(\bar{x}_{k+1}) - f(x^*)) \leq \frac{L_k}{2} \|\bar{x}_{k+1} - x^*_k\|_2^2.
\]

Applying the bound from Proposition 2(b), we obtain:

\[
g(\bar{x}_{k+1}) - g(x^*) + \lambda_k (f(\bar{x}_{k+1}) - f(x^*)) \leq \left( \frac{L_k B^2}{2\lambda_k} \right) \frac{1}{(k+1)^{2-2a-2b}} \quad \text{for all } k \geq K.
\]

(22)

Note that from the definition of \(x^*\) in 8, we have \(g(\bar{x}_{k+1}) - g(x^*) \geq 0\). This implies that:

\[
f(\bar{x}_{k+1}) - f(x^*) \leq \left( \frac{L_k B^2}{2\lambda_k} \right) \frac{1}{(k+1)^{2-2a-2b}} \leq \left( \frac{L_0 B^2}{2\lambda_0} \right) \frac{1}{(k+1)^{2-2a-3b}} \quad \text{for all } k \geq K.
\]

Therefore, the desired relation holds for \(\Omega_1 \triangleq \frac{3a^2}{\lambda_0}\).

(c) From part (a), we know that \(\{\overline{x}_k\}\) converges to \(x^*\). This result and that \(f\) is a continuous function imply that there exits a scalar \(\Omega_3 > 0\) such that \(|f(\bar{x}_{k+1}) - f(x^*)| \leq \Omega_3\). Thus, from the inequality (22) and the update rule for \(\lambda_k\), we obtain:

\[
g(\bar{x}_{k+1}) - g(x^*) \leq \left( \frac{L_0 B^2}{2} \right) \frac{1}{(k+1)^{2-2a-2b}} + \frac{\Omega_3 \lambda_0}{(k+1)^b} \quad \text{for all } k \geq K.
\]

Therefore, the desired relation holds for \(\Omega_2 \triangleq \frac{3a^2}{\lambda_0}\).

A.6 Proof of Corollary 1

First, we show that problem 2 is equivalent to problem 1 where \(g_i(x) := \frac{1}{2} \|A_i x - b_i\|^2_2\). Let \(X_1\) and \(X_2\) denote the feasible set of problem 1 and 2, respectively. Suppose \(\hat{x} \in X_1\) is an arbitrary vector. Thus, we have \(\hat{x} \in \text{argmin}_{x \in [\mathbb{R}^n]} \frac{1}{2} \sum_{i=1}^m \|A_i x - b_i\|^2_2\). From the assumption \(X_2 \neq \emptyset\), there exits a point \(\bar{x}\) satisfying \(A \bar{x} = b\). This implies that the minimum of the function \(\frac{1}{2} \sum_{i=1}^m \|A_i x - b_i\|^2_2\) is zero. Therefore, \(\bar{x}\) must satisfy \(A \bar{x} = b\) implying that \(\hat{x} \in X_2\). Next, suppose \(\hat{x} \in X_2\) is an arbitrary vector. Thus, we have \(\frac{1}{2} \sum_{i=1}^m \|A_i \hat{x} - b_i\|^2_2 = 0\) implying that \(\hat{x}\) is a minimizer of the \(\frac{1}{2} \sum_{i=1}^m \|A_i \hat{x} - b_i\|^2_2\). Therefore, we have \(\hat{x} \in X_1\). We conclude that \(X_1 = X_2\) and thus problems 1 and 2 are equivalent. Next, we show that Assumption 1(b) is satisfied. From the definition of function \(g_i\), we have that \(\nabla g_i(x) = A_i^T (A_i x - b_i)\). We can write for all \(x, y \in \mathbb{R}^n\):

\[
\|\nabla g_i(x) - \nabla g_i(y)\| = \|A_i^T (A_i x - b_i) - A_i^T (A_i y - b_i)\| \leq \rho (A_i^T A_i) \|x - y\|,
\]

where \(\rho (A_i^T A_i)\) denotes the spectral norm of \(A_i^T A_i\). Thus, we conclude that Assumption 1(b) is met for \(L_g \triangleq \max_i \rho (A_i^T A_i)\). Therefore, all conditions of Theorem 1 hold. To obtain the rate results in part (a), (b), (c), it suffices to substitute \(a\) by 0.2 and \(b\) by 0.2 - \(\frac{3}{5}\) in the corresponding parts in Theorem 1.
Assumption 1(a) holds with $\mu_\text{graph}$, independent from the choice of the regularization parameter and the topology of the communication graph. It is observed that the proposed algorithm outperforms the standard Push-Pull algorithm.

In this section, we present additional support for the numerical experiments where we choose the regularization parameter employed in the Push-Pull algorithm to take different values. Figure 2 and Figure 3 show the results for the case where the regularization parameter is 0.01 and 0.005, respectively. It is observed that the proposed algorithm outperforms the standard Push-Pull algorithm independent from the choice of the regularization parameter and the topology of the communication graph.

A.7 Proof of Corollary 2

Note that problem (3) is equivalent to problem (1) where $f_i(x) := \|x\|_2^2/m$. This implies that Assumption 1(a) holds with $\mu_f = L_f = 2/m$. Therefore, all conditions of Theorem 1 hold. To obtain the rate results in part (a) and (b), it suffices to substitute $a$ by 0.4 and $b$ by $0.4 - \epsilon$ in the rate results in Theorem 1.

Appendix B Supplementary numerical results

In this section, we present additional support for the numerical experiments where we choose the regularization parameter employed in the Push-Pull algorithm to take different values. Figure 2 and Figure 3 show the results for the case where the regularization parameter is 0.01 and 0.005, respectively. It is observed that the proposed algorithm outperforms the standard Push-Pull algorithm independent from the choice of the regularization parameter and the topology of the communication graph.