A Communication Efficient Stochastic Multi-Block Alternating Direction Method of Multipliers

Hao Yu
Amazon
eeyuhao@gmail.com

Abstract

The alternating direction method of multipliers (ADMM) has recently received tremendous interests for distributed large scale optimization in machine learning, statistics, multi-agent networks and related applications. In this paper, we propose a new parallel multi-block stochastic ADMM for distributed stochastic optimization, where each node is only required to perform simple stochastic gradient descent updates. The proposed ADMM is fully parallel, can solve problems with arbitrary block structures, and has a convergence rate comparable to or better than existing state-of-the-art ADMM methods for stochastic optimization. Existing stochastic (or deterministic) ADMMs require each node to exchange its updated primal variables across nodes at each iteration and hence cause significant amount of communication overhead. Existing ADMMs require roughly the same number of inter-node communication rounds as the number of in-node computation rounds. In contrast, the number of communication rounds required by our new ADMM is only the square root of the number of computation rounds.

1 Introduction

Fix integer $N \geq 2$. Consider multi-block linearly constrained stochastic convex programs given by:

$$\min_{\mathbf{x}_i \in \mathcal{X}_i, \forall i} \quad f(\mathbf{x}) = \sum_{i=1}^{N} f_i(\mathbf{x}_i) \ \text{ s.t. } \sum_{i=1}^{N} \mathbf{A}_i \mathbf{x}_i = \mathbf{b}, \tag{1}$$

where $\mathbf{x}_i \in \mathbb{R}^{d_i}, \mathbf{A}_i \in \mathbb{R}^{m \times d_i}, \mathbf{b} \in \mathbb{R}^m$, $\mathcal{X}_i \subseteq \mathbb{R}^{d_i}$ are closed convex sets, and $f_i(\mathbf{x}_i) = \mathbb{E}_\xi[f_i(\mathbf{x}_i; \xi)]$ are convex functions. To have a compact representation of (1), we define $\mathbf{x} = [\mathbf{x}_1; \mathbf{x}_2; \ldots; \mathbf{x}_N] \in \mathbb{R}^{\sum_{i=1}^N d_i}$, $\mathcal{X} = \prod_{i=1}^N \mathcal{X}_i$, $f(\mathbf{x}) = \sum_{i=1}^N f_i(\mathbf{x}_i)$ and $\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_N] \in \mathbb{R}^{m \times \sum_{i=1}^N d_i}$. Note that constraint $\sum_{i=1}^N \mathbf{A}_i \mathbf{x}_i = \mathbf{b}$ now can be written as $\mathbf{A}\mathbf{x} = \mathbf{b}$.

The problem (1) captures many important applications in machine learning, network scheduling, statistics and finance. For example, (stochastic) linear programs that are too huge to be solved over a single node can be written as (1). To solve such large scale linear programs in a distributed manner, we can save each $\mathbf{A}_i$ and $f_i(\cdot)$ at a separate node and let each node iteratively solves smaller sub-problems (with necessary inter-node communication). Another important application of formulation (1) is the distributed consensus training of a machine learning model over N nodes [15, 17, 23] described as follows:

- In an online training setup, i.i.d. realizations of $f_i(\cdot; \xi)$ are sampled at each node. In an offline training setup, $f_i(\mathbf{x}_i) = \mathbb{E}_\xi[f_i(\mathbf{x}_i; \xi)]$ are approximated by $\frac{1}{N_i} \sum_{j=1}^{N_i} f_{ij}(\mathbf{x}_i)$ where N_i is the number of training samples at node i and each $f_{ij}(\cdot)$ represents one training sample.
- To enforce all N nodes are training the same model, our constraint $\mathbf{A}\mathbf{x} = \mathbf{b}$ is given by $\mathbf{x}_i = \mathbf{x}_j$ for all $i \neq j \in \{1, 2, \ldots, N\}$. (In fact, we only need such constraints for pairs (i, j) that construct a connected graph for all nodes.)

The Alternating Direction Method of Multipliers (ADMM) is an effective and popular method to solve linearly constrained convex programs, especially distributed consensus optimiation [28, 5],

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

since it often yields distributed implementations with low complexity [4]. Conventional ADMMs are developed for the special case of problem (1) with $N = 2$ and/or deterministic $f_i(\mathbf{x}_i)$. To solve a two-block problem (1) where f_1 is a stochastic function and f_2 is a deterministic function, previous works [21, 25, 31, 1] have developed stochastic (two-block) ADMMs to solve problem (1) with $N = 2$. It is unclear whether these methods can be extended to solve the case $N \geq 3$. In fact, even for problem (1) where all $f_i(\mathbf{x}_i)$ are deterministic, [6] proves that the classical (two-block) ADMM, on which the stochastic versions in [21, 25] are built, converges for $N = 2$ but diverges for $N \geq 3$. To solve stochastic convex program (1) with $N \geq 3$, randomized block coordinate updated ADMMs with $O(1/\epsilon^2)$ convergence are developed in [27, 11]. Due to the challenging stochastic objective functions, the convergence rate of stochastic ADMMs is fundamentally slower than deterministic ADMMs, i.e., $O(1/\epsilon^2)$ v.s. $O(1/\epsilon)$ [13, 7, 11]. The $O(1/\epsilon^2)$ convergence is optimal since it is optimal even for unconstrained stochastic convex optimization without strong convexity [20]. However, in distributive implementations of ADMMs, each node has to pass its most recent $\mathbf{x}_i$ value to its neighbors or a fusion center and then updates the dual variable $\boldsymbol{\lambda}$. Existing stochastic ADMM methods [21, 25, 11] require a communication step immediately after each $\mathbf{x}_i$ computation step. In practice, the inter-node communication over TCP/IP is much slower than in-node memory computations and often requires additional set-up time such that communication overhead is the performance bottleneck of most distributed optimization methods.

As a consequence, communication efficient optimization recently attracted a lot of research interests [29, 14, 24, 15, 17, 18, 23]. Work [17] proposes a primal-dual method that can solve problem (1) with stochastic objective functions using $O(1/\epsilon^2)$ computation iterations and $O(1/\epsilon)$ communication iterations. However, the method in [17] requires each objective function $f_i(\cdot)$ to satisfy the stringent condition that there exists M such that $f_i(\mathbf{u}) \leq f_i(\mathbf{v}) + \langle \mathbf{d}, \mathbf{u} - \mathbf{v} \rangle + M\|\mathbf{u} - \mathbf{v}\|$ for any $\mathbf{u}, \mathbf{v}$ and $\mathbf{d} \in \partial f_i(\mathbf{v})$. Such a condition is more stringent than the smoothness when $\mathbf{u}$ and $\mathbf{v}$ are far apart from each other. For example, the simple scalar smooth function $f(x) = x^2$ does not satisfy this condition over $\mathcal{X} = \mathbb{R}$. Work [18] proposes a communication efficient method to solve **deterministic** convex programs based on the quadratic penalty method and can obtain an ϵ-optimal solution with $O(1/\epsilon^{2+\delta})$ computation rounds (δ is a positive constant) and $O(1/\epsilon)$ communication rounds. For distributed consensus optimization over a network, which can be formulated as a special case of problem (1) where $\mathbf{A}_i$ and $\mathbf{b}$ are chosen to ensure all $\mathbf{x}_i$ are identical, mixing or local averaging based methods with fast convergence (and low communication overhead) are recently developed in [26, 22, 23, 19].

Our Contributions: This paper proposes a new communication efficient stochastic multi-block ADMM which has communication rounds less frequently than computation rounds. For stochastic convex programs with general convex objective functions, our algorithm can achieve an ϵ-solution with $O(1/\epsilon^2)$ computation[1] rounds and $O(1/\epsilon)$ communication rounds. That is, our communication efficient ADMM has the same computation convergence rate as the ADMM in [11] but only requires the square root of communication rounds required by the method in [11]. For stochastic convex programs with strongly convex objective functions, our algorithm can achieve an ϵ-accuracy solution with $\tilde{O}(1/\epsilon)$ computation rounds and $\tilde{O}(1/\sqrt{\epsilon})$ communication rounds[2]. The fast computation convergence (and even faster communication convergence) for strongly convex stochastic programs is not possessed by the ADMM in [11]. When applying our new multi-block ADMM to the special case of two-block problems, our algorithm has the same computation convergence as existing two-block stochastic ADMM methods in [21, 25, 31, 1]. However, the number of communication rounds used by our ADMM is only the squared root of these previous methods.

Notations: This paper uses $\|\mathbf{A}\|$ to denote the spectral norm of matrix $\mathbf{A}$; $\|\mathbf{z}\|$ to denote the Euclidean norm of vector $\mathbf{z}$; and $\langle \mathbf{y}, \mathbf{z} \rangle = \mathbf{y}^\mathsf{T}\mathbf{z}$ to denote the inner product of vectors $\mathbf{y}$ and $\mathbf{z}$. If symmetric matrix $\mathbf{Q} \succeq \mathbf{0}$ is positive semi-definite, then we define $\|\mathbf{z}\|_\mathbf{Q}^2 = \mathbf{z}^\mathsf{T}\mathbf{Q}\mathbf{z}$ for any vector $\mathbf{z}$.

2 Formulation and New Algorithm

Following the convention in [8], a function $h(\mathbf{x})$ is said to be *convex with modulus* μ, or equivalently, μ-convex, if $h(\mathbf{x}) - \frac{\mu}{2}\|\mathbf{x}\|^2$ is convex. The μ-convex definition unifies the conventional definitions of convexity and strong convexity. That is, a general convex function, which is not necessarily strongly convex, is convex with modulus $\mu = 0$; and a strongly convex function is convex with modulus $\mu > 0$. Throughout this paper, convex program (1) is assumed to satisfy the following standard assumption:

[1] A computation round of our algorithm is a just a single iteration of the SGD update.

[2] A logarithm factor $\log(\frac{1}{\epsilon})$ is hidden in the notation $\tilde{O}(\cdot)$.

Assumption 1. *Convex program* (1) *has a saddle point* $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$*. That is,* $\mathbf{x}^*$ *is an optimal solution and* $\boldsymbol{\lambda}^* \in \mathbb{R}^m$ *is a Lagrange multiplier attaining strong duality* $q(\boldsymbol{\lambda}^*) = f(\mathbf{x}^*)$*, where* $q(\boldsymbol{\lambda}^*) \triangleq \inf_{\{\mathbf{x}_i \in \mathcal{X}_i, \forall i\}} \{f(\mathbf{x}) + \langle \boldsymbol{\lambda}^*, \mathbf{A}\mathbf{x} - \mathbf{b}\rangle\}$ *is the Lagrangian dual function.*

Note that strong duality in Assumption 1 is often stated as its equivalent "KKT conditions", e.g., in [7]. A mild sufficient condition for Assumption 1 to hold is (1) has at least one feasible point and the domain of each $f_i(\mathbf{x}_i)$ includes $\mathcal{X}_i$ as an interior [3].

Assume unbiased subgradients $G_i(\mathbf{x}_i;\xi)$ satisfying $\mathbb{E}_\xi[G_i(\mathbf{x}_i;\xi)] = \partial f_i(\mathbf{x}_i), \forall \mathbf{x}_i \in \mathcal{X}_i$ for each function $f_i(\mathbf{x}_i)$ can be sampled. Denote the stacked column vector $\mathbf{G}(\mathbf{x};\xi) \triangleq [G_1(\mathbf{x}_1;\xi)^\mathsf{T}, \ldots, G_N(\mathbf{x}_N;\xi)^\mathsf{T}]^\mathsf{T} \in \mathbb{R}^{\sum_{i=1}^N d_i}$. We have $\mathbb{E}_\xi[\mathbf{G}(\mathbf{x};\xi)] = \partial f(\mathbf{x})$.

Consider the communication efficient stochastic multi-block ADMM described in Algorithm 1. Since $f_i(\mathbf{x}_i)$ are stochastic, $\phi_i(\mathbf{x}_i)$ defined in (2) is fundamentally unknown. However, each $\phi_i(\mathbf{x}_i)$ is $\nu^{(t)}$-convex and its unbiased stochastic subgradient is available as long as we have unbiased stochastic subgradients of $f_i(\mathbf{x}_i)$. The sub-procedure STO-LOCAL involved in Algorithm 1 is a simple stochastic subgradient decent (SGD) procedure (with particular choices of parameters, starting points and averaging schemes) to minimize $\phi_i^{(t)}(\cdot)$ over set $\mathcal{X}_i$ and is described in Algorithm 2.

Algorithm 1 Two-Layer Communication Efficient ADMM

1: **Input:** Algorithm parameters T, $\{\rho^{(t)}\}_{t\geq 1}$, $\{\nu^{(t)}\}_{t\geq 1}$ and $\{K^{(t)}\}_{t\geq 1}$.
2: Initialize arbitrary $\mathbf{y}_i^{(0)} \in \mathcal{X}_i, \forall i$, $\mathbf{r}^{(0)} = \sum_{i=1}^N \mathbf{A}_i \mathbf{y}_i^{(0)} - \mathbf{b}$, $\boldsymbol{\lambda}^{(0)} = \mathbf{0}$, and $t = 1$.
3: **while** $t \leq T$ **do**
4: Each node i defines

$$\phi_i^{(t)}(\mathbf{x}_i) \triangleq f_i(\mathbf{x}_i) + \rho^{(t)}\big\langle \mathbf{r}^{(t-1)} + \frac{1}{\rho^{(t)}}\boldsymbol{\lambda}^{(t-1)}, \mathbf{A}_i\mathbf{x}_i - \frac{\mathbf{b}}{N}\big\rangle + \frac{\nu^{(t)}}{2}\|\mathbf{x}_i - \mathbf{y}_i^{(t-1)}\|^2 \tag{2}$$

and **in parallel** updates $\mathbf{x}_i^{(t)}, \mathbf{y}_i^{(t)}$ using local sub-procedure Algorithm 2 via

$$(\mathbf{x}_i^{(t)}, \mathbf{y}_i^{(t)}) = \text{STO-LOCAL}(\phi_i^{(t)}(\cdot), \mathcal{X}_i, \mathbf{y}_i^{(t-1)}, K^{(t)}) \tag{3}$$

5: Each node i passes $\mathbf{x}_i^{(t)}$ and $\mathbf{y}_i^{(t)}$ between nodes or to a parameter server. Update $\boldsymbol{\lambda}^{(t)}$ and $\mathbf{r}^{(t)}$ via

$$\boldsymbol{\lambda}^{(t)} = \boldsymbol{\lambda}^{(t-1)} + \rho^{(t)}\Big(\sum_{i=1}^N \mathbf{A}_i\mathbf{x}_i^{(t)} - \mathbf{b}\Big) \tag{4}$$

$$\mathbf{r}^{(t)} = \sum_{i=1}^N \mathbf{A}_i\mathbf{y}_i^{(t)} - \mathbf{b}. \tag{5}$$

6: Update $t \leftarrow t + 1$.
7: **end while**
8: **Output:** $\overline{\mathbf{x}}^{(T)} = \frac{1}{\sum_{t=1}^T \rho^{(t)}} \sum_{t=1}^T \rho^{(t)}\mathbf{x}^{(t)}$

Algorithm 2 STO-LOCAL$(\phi(\mathbf{z}), \mathcal{Z}, \mathbf{z}^{\text{init}}, K)$

1: **Input:** μ: strong convexity modulus of $\phi(\mathbf{z})$; Algorithm parameters: $k_0 > 0$; $\gamma^{(k)} = \frac{2}{\mu(k+k_0)}, \forall k \in \{1, 2, \ldots, K\}$.
2: Initialize $\mathbf{z}^{(0)} = \mathbf{z}^{\text{init}}$ and $k = 1$.
3: **while** $k \leq K$ **do**
4: Observe an unbiased gradient $\boldsymbol{\zeta}^{(k)}$ such that $\mathbb{E}[\boldsymbol{\zeta}^{(k)}] = \partial\phi(\mathbf{z}^{(k-1)})$ and update $\mathbf{z}^{(k)}$ via

$$\mathbf{z}^{(k)} = \mathcal{P}_{\mathcal{Z}}\Big[\mathbf{z}^{(k-1)} - \gamma^{(k)}\boldsymbol{\zeta}^{(k)}\Big] \tag{6}$$

where $\mathcal{P}_{\mathcal{Z}}[\cdot]$ is the projection onto $\mathcal{Z}$.
5: **end while**
6: **Output:** $(\widehat{\mathbf{z}}, \mathbf{z}^{(K)})$ where $\widehat{\mathbf{z}}$ is the time average of $\{\mathbf{z}^{(0)}, \ldots, \mathbf{z}^{(K)}\}$ defined in Lemmas 1 or 2.

We now justify why Algorithm 1 is a two-layer ADMM method. (See Supplement 6.1 for a more detailed discussion.)

- The Lagrange multiplier update (4) is identical to that used in existing ADMM methods or other Lagrangian based methods. It is helpful to enforce the linear constraint.
- At the first sight, the primal update in Algorithm (4) is quite different from existing deterministic ADMMs in [10, 4, 7], which require to solve an "*argmin*" problem, or stochastic ADMMs in [21, 25, 11], which perform a single gradient descent step . However, with a simple manipulation, it is not difficult to show that that function $\phi_i^{(t)}(\mathbf{x}_i)$ in (2) is similar to the "*argmin*" target in the proximal Jacobi ADMM method [7] with the distinction that the proximal term $\|\mathbf{x}_i - \mathbf{y}_i^{(t-1)}\|^2$ is regarding a newly introduced variable $\mathbf{y}_i^{(t-1)}$ rather than $\mathbf{x}_i^{(t-1)}$.

Recall that the fastest stochastic ADMMs in [21, 25, 11] can solve general convex problem (1) (with $N = 2$) with $O(1/\sqrt{T})$ convergence. That is, to obtain a solution with ϵ errors for both the objective value and the constraint violation, the ADMMs in [21, 25, 11] require $O(1/\epsilon^2)$ computation steps, each of which uses a single gradient evaluation and variable update. The ADMMs in [21, 25, 11] has a single layer structure and hence are communication inefficient in the sense that each computation step involves a communication steps. Thus, the communication complexity of these stochastic ADMMs is also $O(1/\epsilon^2)$. Compared with existing ADMMs in [21, 25, 11], Algorithm 1 has a two layer structure where each outer layer step involves a single inter-node communication step given by (4)-(5) and calls the sub-procedure, i.e. Algorithm 2, STO-LOCAL$(\phi_i^{(t)}(\cdot), \mathcal{X}_i, \mathbf{y}_i^{(t)}, K^{(t)})$, which is run by each node locally and in parallel and hence does not incur any inter-node communication overhead. Since each call of Algorithm 2 incurs $K^{(t)}$ SGD update, T iterations of Algorithm 1 use $\sum_{t=1}^{T} K^{(t)}$ computation steps. We shall show that to achieve an ϵ solution for general convex problem (1), Algorithm 1 uses $T = O(1/\epsilon)$ communication rounds and $\sum_{t=1}^{T} K^{(t)} = O(1/\epsilon^2)$ computation steps. That is, Algorithm 1 is as fast as existing fastest stochastic ADMMs but uses only a square root of the number of communications rounds in [21, 25, 11].

Note that inter-node communication in Algoirthm 1 can be either centralized or decentralized. To use centralized communication, we can let all nodes pass their $\mathbf{x}_i^{(t)}$ to a parameter server, where (4)-(5) are executed, and then pull the updated $\boldsymbol{\lambda}^{(t)}$ and $\mathbf{r}^{(t)}$ from the server. It is possible to implement (4)-(5) using decentralized communication by exploring the structure of matrix $\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_N]$. For example, consider distributed machine learning in a line network where $\mathbf{Ax} = \mathbf{b}$ is given by $N - 1$ equality constraints $\mathbf{x}_i - \mathbf{x}_{i+1} = 0, i \in \{1, 2, \ldots, N-1\}$. In this case, $\boldsymbol{\lambda}_i^{(t)}$ and $\mathbf{r}_i^{(t)}$ only depend on $\mathbf{x}_i^{(t)}$ and $\mathbf{x}_{i+1}^{(t)}$ and are only used to updates $\mathbf{x}_i^{(t+1)}$ and $\mathbf{x}_{i+1}^{(t+1)}$. Thus, to implement Algorithm 1, each node only needs to send its local $\mathbf{x}_i^{(t)}$ to and pull $\boldsymbol{\lambda}_j^{(t)}$ and $\mathbf{r}_j^{(t)}$ from its neighbors in the line network.

2.1 Basic Facts of Algorithm 2

Since each iteration of Algorithm 1 calles Algorithm 2, which essentially applies SGD with carefully designed step size rules to newly introduced objective functions $\phi_i^{(t)}(\cdot)$. This subsection provides some useful insight of SGD for strongly convex stochastic minimization.

It is known that SGD can have $O(1/\epsilon)$ convergence for strongly convex minimization. The next two lemmas summarize the convergence of SGD Algorithm 2. When characterizing $O(1/\epsilon)$ rate, our lemmas also include a push-back term involving the last iteration solution. This term ensures when the SGD solution from Algorithm 2 is used in the outer-level ADMM dynamics, the accumulated error of our final solution does not explode. It also explains why we use $\mathbf{y}_i^{(t-1)}$, which is the last iteration solution from the SGD sub-procedure, rather than conventional $\mathbf{x}_i^{(t-1)}$ to define $\phi_i^{(t)}(\mathbf{x}_i)$.

Lemma 1 ([16])**.** *Assume $\phi(\mathbf{z})$ is a μ-convex function ($\mu > 0$) over set $\mathcal{Z}$ and there exists a constant B such that the unbiased subgradient $\boldsymbol{\zeta}^{(k)}$ used in Algorithm 2 satisfies $\mathbb{E}[\|\boldsymbol{\zeta}^{(k)}\|^2] \leq B^2, \forall k \in \{1, 2, \ldots, K\}$. If we take $k_0 = 1$ in Algorithm 2, then for all $\mathbf{z} \in \mathcal{Z}$, we have*

$$\mathbb{E}[\phi(\widehat{\mathbf{z}})] \leq \phi(\mathbf{z}) - \underbrace{\frac{\mu}{2}\mathbb{E}[\|\mathbf{z}^{(K)} - \mathbf{z}\|^2]}_{\text{(7)-term (I)}} + \frac{2B^2}{\mu(K+1)}, \tag{7}$$

where $\widehat{\mathbf{z}} = \frac{1}{\sum_{k=0}^{K-1}(k+k_0)} \sum_{k=0}^{K-1}(k + k_0)\mathbf{z}^{(k)}$.

Remark 1. *It is firstly shown in [16] that Algorithm 2 with $k_0 = 1$ (vanilla SGD with a particular averaging scheme) has $O(1/\epsilon)$ convergence for non-smooth strongly convex problems. Note that* (7) *holds for all $\mathbf{z} \in \mathcal{Z}$ (not necessarily the minimizer of $\phi(\cdot)$). The push-back term* (7)*-term (I) is often ignored in convergence rate analysis for SGD but is important for our analysis of Algorithm 1.*

Recall that a function $h(\mathbf{x})$ is said to be L-smooth if its gradient $\nabla h(\mathbf{x})$ is Lipschitz with modulus L. The next lemma is new and extends Lemma 1 to smooth minimization such that the error term depends only on the variance of stochastic gradients (using a different averaging scheme).

Lemma 2. *Assume $\phi(\mathbf{z})$ is a L-smooth and μ-convex function ($\mu > 0$) with conditional number $\kappa = \frac{L}{\mu}$ and there exists $\sigma > 0$ such that unbiased gradient $\boldsymbol{\zeta}^{(k)}$ (at point $\mathbf{z}^{(k-1)}$) in Algorithm 2 satisfies $\mathbb{E}[\|\boldsymbol{\zeta}^{(k)} - \nabla\phi(\mathbf{z}^{(k-1)})\|^2] \leq \sigma^2, \forall k \in \{1, 2, \ldots, K\}$. If we take integer $k_0 > 2\kappa$, then for any $\mathbf{z} \in \mathcal{Z}$, we have*

$$\mathbb{E}[\phi(\widehat{\mathbf{z}})] \leq \phi(\mathbf{z}) + \frac{\mu(k_0^2 - k_0)}{2K(K + 2k_0 - 1)}\left(\mathbb{E}[\|\mathbf{z} - \mathbf{z}^{(0)}\|^2] - \mathbb{E}[\|\mathbf{z} - \mathbf{z}^{(K)}\|^2]\right) - \frac{\mu}{2}\mathbb{E}[\|\mathbf{z} - \mathbf{z}^{(K)}\|^2] + \frac{2k_0\sigma^2}{(K + 2k_0 - 1)\mu} \quad (8)$$

where $\widehat{\mathbf{z}} = \frac{1}{\sum_{k=1}^K (k+k_0-1)} \sum_{k=1}^K (k + k_0 - 1)\mathbf{z}^{(k)}$.

Proof. See Supplement 6.6. □

3 Performance Analysis of Algorithm 1

This section shows that Algorithm 1 can achieve an ϵ-accuracy solution using $O(1/\epsilon^2)$ computation rounds and $O(1/\epsilon)$ communication rounds for general convex stochastic programs; or using $\tilde{O}(1/\epsilon)$ computation rounds and $\tilde{O}(1/\sqrt{\epsilon})$ communication rounds for strongly convex stochastic programs.

3.1 General objective functions (possibly non-smooth non-strongly convex)

Theorem 1. *Consider convex program* (1) *under Assumption 1. Let $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ be any saddle point defined in Assumption 1. Assume that*

- *The constraint set $\mathcal{X}$ is bounded, i.e., there exists constant $R > 0$ such that $\|\mathbf{x}\| \leq R, \forall \mathbf{x} \in \mathcal{X}$.*
- *The function $f(\mathbf{x})$ has unbiased stochastic subgradients with a bounded second order moment, i.e., there exists constant $D > 0$ such that $\mathbb{E}_\xi[\|\mathbf{G}(\mathbf{x}; \xi)\|^2] \leq D^2, \forall \mathbf{x} \in \mathcal{X}$.*

For all $T \geq 1$, if we choose any fixed $\rho^{(t)} = \rho > 0$, $\nu^{(t)} = \nu \geq 8\rho\|\mathbf{A}\|^2$, $K^{(t)} = K \geq T$ in Algorithm 1 and the sub-procedure STO-LOCAL (Algorithm 2) uses $\widehat{\mathbf{z}}$ defined in Lemma 1 as the output, then

$$\mathbb{E}[f(\overline{\mathbf{x}}^{(T)})] \leq f(\mathbf{x}^*) + \frac{\nu}{2T}\|\mathbf{x}^* - \mathbf{y}^{(0)}\|^2 + \frac{C}{2\nu T} \quad (9)$$

$$\mathbb{E}[\|\mathbf{A}\overline{\mathbf{x}}^{(T)} - \mathbf{b}\|] \leq \frac{1}{T}\frac{\sqrt{Q}}{\rho} \quad (10)$$

where $\overline{\mathbf{x}}^{(T)} = \frac{1}{t}\sum_{t=1}^T \mathbf{x}^{(t)}$; $Q = (2\|\boldsymbol{\lambda}^*\| + \sqrt{\rho\nu\|\mathbf{x}^* - \mathbf{y}^{(0)}\|^2 + \frac{24\rho D^2}{\nu} + \frac{24(\rho)^3\|\mathbf{A}\|^2(\|\mathbf{A}\|R + \|\mathbf{b}\|)^2}{\nu} + 96\nu\rho R^2}/(1 - \sqrt{\frac{8\rho\|\mathbf{A}\|^2}{\nu}}))^2$ is an absolute constant (irrelevant to T); and $C \overset{\Delta}{=} 4\|\mathbf{A}\|^2 Q + 12D^2 + 12\rho^2\|\mathbf{A}\|^2(\|\mathbf{A}\|R + \|\mathbf{b}\|)^2 + 48\nu^2R^2$ is also an absolute constant.

Proof. See Supplement 6.7. □

Remark 2. *After T outer-level rounds, Algorithm 1 yields a solution with error $O(1/T)$. Note that the number of communication rounds is equal to the number of outer-level rounds and the number of computation rounds is $\sum_{t=1}^T K^{(t)} = O(T^2)$ when $K^{(t)} = T, \forall t$. Thus, to obtain an ϵ-solution, Algorithm 1 uses $O(1/\epsilon)$ communication rounds and $O(1/\epsilon^2)$ computation rounds.*

Remark 3. *If we choose $\nu^{(t)} = \nu = 8\rho\|\mathbf{A}\|^2$ in Theorem 1 and further analyze the dependence on $\|\mathbf{A}\|$ in* (9)*-*(10)*, we have $\mathbb{E}[f(\overline{\mathbf{x}}^{(T)})] \leq f(\mathbf{x}^*) + O(\frac{1}{T}\rho\|\mathbf{A}\|^2)$ and $\mathbb{E}[\|\mathbf{A}\overline{\mathbf{x}}^{(T)} - \mathbf{b}\|] \leq O(\frac{1}{T}(\frac{1}{\rho} + \|\mathbf{A}\|))$. If $\|\mathbf{A}\|$ is large, to balance the dependence on $\|\mathbf{A}\|$ in* (9)*-*(10)*, we shall choose $\rho = \frac{1}{\|\mathbf{A}\|}$ such that the error terms in both* (9) *and* (10) *are order $O(\frac{1}{T}\|\mathbf{A}\|)$. In general, ρ can be controlled to trade off between objective error and constraint error. For distributed consensus optimization considered in [26, 22, 23, 19] (assuming $d_i = 1$ without loss of generality), we can choose any $\mathbf{A}, \mathbf{b}$ that suffices to ensure the consistence of local solutions, e.g., Null$\{\mathbf{A}\}$ =Span$\{\mathbf{1}\}$ and $\mathbf{b} = \mathbf{0}$. Our method does not necessarily require $\mathbf{A} = \mathbf{I} - \mathbf{W}$ with a stochastic matrix $\mathbf{W}$ encoding the network topology as some methods in [26, 22, 23, 19]. Nevertheless, even when ung $\mathbf{A} = \mathbf{I} - \mathbf{W}$, our communication overhead can possibly have a better dependence on $\mathbf{W}$. Note that a stochastic matrix $\mathbf{W}$ ensures $\|\mathbf{A}\| \leq 2$. The convergence in [26, 22, 23, 19] (using a doubly stochastic or symmetric PSD $\mathbf{W}$ for mixing) further depends on $1/(1 - \max\{|\lambda_2(\mathbf{W})|, |\lambda_N(\mathbf{W})|\})$ or the eigen-gap $\lambda_1(\mathbf{W})/\lambda_{N-1}(\mathbf{W})$, which can be much larger than constant 2 when some eigenvalues are extreme.*

3.2 Smooth objective functions

For unconstrained stochastic smooth minimization, the constant factor in the SGD convergence rate is determined by the variance that can be significantly less than the second order moment for non-smooth stochastic minimization[20]. Such a property enable us to speed up SGD by averaging multiple i.i.d. stochastic gradients, e.g., mini-batch SGD. In this subsection, we show that Algorithm 1 has a similar property when $f(\cdot)$ in problem (1) is smooth.

Theorem 2. *Consider convex program* (1) *with μ-convex (possibly $\mu = 0$) objective function under Assumption 1. Let $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ be any saddle point defined in Assumption 1. Assume that*

- *The function $f(\mathbf{x})$ is L-smooth.*
- *The function $f(\mathbf{x})$ has unbiased stochastic gradients with a bounded variance, i.e., there exists constant $\sigma > 0$ such that $\mathbb{E}_\xi[\|\mathbf{G}(\mathbf{x};\xi) - \nabla f(\mathbf{x})\|^2] \leq \sigma^2, \forall \mathbf{x} \in \mathcal{X}$.*

If the sub-procedure STO-LOCAL (Algorithm 2) uses $\widehat{\mathbf{z}}$ defined in Lemma 2 as the output, then Algorithm 1 ensures:

- **General Convex ($\mu = 0$)***: For all $T \geq 1$, if we choose any fixed $\rho^{(t)} = \rho > 0$, $\nu^{(t)} = \nu \geq \rho\|\mathbf{A}\|^2$, $K^{(t)} = K = T$ and positive integer $k_0 \geq 2\frac{L+\nu}{\nu}$, then we have*

$$\mathbb{E}[f(\overline{\mathbf{x}}^{(T)})] \leq f(\mathbf{x}^*) + \frac{1}{T}\frac{\nu(k_0+1)}{4}\|\mathbf{x}^* - \mathbf{y}^{(0)}\|^2 + \frac{1}{T}\frac{2k_0\sigma^2}{\nu} \tag{11}$$

$$\mathbb{E}[\|\mathbf{A}\overline{\mathbf{x}}^{(T)} - \mathbf{b}\|] \leq \frac{1}{T}\Big(\frac{2}{\rho}\|\boldsymbol{\lambda}^*\| + \sqrt{\frac{\nu(k_0+1)}{2\rho}}\|\mathbf{x}^* - \mathbf{y}^{(0)}\| + 2\sqrt{\frac{k_0\sigma^2}{\rho\nu}}\Big) \tag{12}$$

  where $\overline{\mathbf{x}}^{(T)} = \frac{1}{t}\sum_{t=1}^T \mathbf{x}^{(t)}$.

- **Strongly Convex ($\mu > 0$)***: For all $T \geq 1$, if we choose $\rho \leq \frac{\mu}{3\|\mathbf{A}\|^2}$, $\rho^{(t)} = t\rho$, $\nu^{(t)} = t\rho\|\mathbf{A}\|^2$, positive integer $k_0 \geq 2(1 + \frac{L}{\mu})$ and $K^{(t)} = (2k_0 - 1)t$, then we have*

$$\mathbb{E}[f(\overline{\mathbf{x}}^{(T)})] \leq f(\mathbf{x}^*) + \frac{1}{T(T+1)}\Big(c_1\|\mathbf{x}^* - \mathbf{y}^{(0)}\|^2 + \frac{c_2}{\rho}\log(T+1)\Big) \tag{13}$$

$$\mathbb{E}[\|\mathbf{A}\overline{\mathbf{x}}^{(T)} - \mathbf{b}\|] \leq \frac{2}{T(T+1)}\Big(\frac{4\|\boldsymbol{\lambda}^*\|}{\rho} + \frac{\sqrt{c_1}}{\sqrt{\rho}}\|\mathbf{x}^* - \mathbf{y}^{(0)}\| + \frac{\sqrt{c_2\log(T+1)}}{\rho}\Big) \tag{14}$$

  where $\overline{\mathbf{x}}^{(T)} = \frac{1}{\sum_{t=1}^T \rho^{(t)}}\sum_{t=1}^T \rho^{(t)}\mathbf{x}^{(t)}$; and $c_1 \triangleq \rho\|\mathbf{A}\|^2 + \frac{(\rho\|\mathbf{A}\|^2+\mu)(k_0^2-k_0)}{2(2k_0-1)^2}$ and $c_2 \triangleq \frac{4k_0\sigma^2}{(2k_0-1)\|\mathbf{A}\|^2}$ are two constants.

Proof. See Supplement 6.8. □

Remark 4. *If $f(\mathbf{x})$ in convex program* (1) *is strongly convex, Algorithm 1 can obtain a solution with error $O(\frac{\log(T)}{T^2})$ after T outer-level rounds. Recall the number of communication rounds is equal to the number of outer-level rounds and the number of computation rounds is equal to $\sum_{t=1}^T K^{(t)} = \frac{2k_0-1}{2}T(T+1) = O(T^2)$, Algorithm 1 requires $\tilde{O}(\frac{1}{\epsilon})$ communication rounds and $\tilde{O}(\frac{1}{\epsilon^2})$ computation rounds to obtain an ϵ-solution.*

3.3 Non-smooth strongly convex objective functions

There is a fourth case, where the stochastic objective function $f(\mathbf{x})$ is strongly convex but possibly non-smooth, uncovered in the previous subsections. In this case, we assume the following condition (originally introduced in [17]): There exists constant $M > 0$ such that

$$f(\mathbf{x}) \leq f(\mathbf{y}) + \langle \mathbf{d}, \mathbf{x} - \mathbf{y}\rangle + M\|\mathbf{x} - \mathbf{y}\|, \tag{15}$$

for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ and $\mathbf{d} \in \partial f(\mathbf{y})$. This condition is assumed throughout [17] to develop a different communication efficient primal-dual method. Supplement 6.9 shows this condition is almost as useful as smoothness and under this condition, our communication efficient ADMM can achieve an ϵ-accuracy solution with $\tilde{O}(1/\epsilon)$ computation rounds and $\tilde{O}(1/\sqrt{\epsilon})$ communication rounds for non-smooth strongly convex stochastic optimization.

4 Experiments

4.1 Distributed Stochastic Optimization with Noisy Stochastic Gradient Information

Consider simple stochastic optimization given by

$$\min \quad \sum_{i=1}^{3} \mathbb{E}_{\mathbf{c}_i}[\|\mathbf{x}_i - \mathbf{c}_i\|_2^2] \tag{16}$$

$$\text{s.t.} \quad \mathbf{x}_1 = \mathbf{x}_2, \mathbf{x}_2 = \mathbf{x}_3 \tag{17}$$

$$\mathbf{x}_i \in [-1, 1]^3, \forall i \in \{1, 2, \ldots, 3\} \tag{18}$$

where $\mathbf{c}_i \sim \mathcal{N}(\bar{\mathbf{c}}_i, \sigma_i^2 \mathbf{I})$ satisfy normal distributions with $\bar{\mathbf{c}}_1 = [-2.0871, -0.3702, 0.2302]^\mathsf{T}$, $\sigma_1 = 0.1$, $\bar{\mathbf{c}}_2 = [-0.5556, -0.4413, 0.2869]^\mathsf{T}$, $\sigma_2 = 0.2$, $\bar{\mathbf{c}}_3 = [-1.4991, -1.8286, -2.0477]^\mathsf{T}$ and $\sigma_3 = 0.1$. Solving this problem with Algorithm 1 only requires each node to access samples of local $\mathbf{c}_i$ and does not use the true value $\bar{\mathbf{c}}_i$ and σ_i,which are fundamentally unavailable. However, by assuming the knowledge of $\bar{\mathbf{c}}_i$ and σ_i, we can convert this stochastic optimization to a deterministic problem and use CVXPY [9] to obtain the unique solution $\mathbf{x}_1^* = \mathbf{x}_2^* = \mathbf{x}_3^* = [-1, -0.88003599, -0.51020207]^\mathsf{T}$ such that we can evaluate the performance of Algorithm 1. Since the objective function is smooth and strongly convex, by Theorem 2, using time-varying parameters in Algorithm 1 has faster convergence. We run Algorithm 1 with constant ρ, ν according to[3] Theorem 1 and with time-varying $\rho^{(t)}, \nu^{(t)}$ according to Theorem 2, respectively. Note that if an algorithm has $O(1/\epsilon^\beta)$ convergence, then its error should decay like $O(1/t^{1/\beta})$ where t is the iteration index.

Figures 1 plots the distance to $\mathbf{x}^*$ versus the computation round index or the communication round index in a log-log scale. It also plots baseline curves $1/t^{\frac{1}{\beta}}$ corresponding to $O(1/\epsilon^\beta)$ convergence proven in the theorems. Note that in a log-log scale, curves $1/t^{\frac{1}{\beta}}$ become straight lines with slopes $-\frac{1}{\beta}$. That is, if our algorithm has the proven convergence rate, the error curves should be eventually parallel to corresponding baseline for large t. In Figures 1, we observe the numerical result is consistent with our theoretical rate proven in our theorems. This simple experiment verifies the correctness of our theorems. Our multi-core implementation of Algorithm 1 uses Python 3.7 and MPI4PY. In an experiment over a machine with a multi-core Intel Xeon Processor E5-2682 2.5GHz. Each computation round takes 0.3ms and each communication round takes 43.7ms. Note communication becomes more relatively expensive as more parallel nodes/cores are involved.

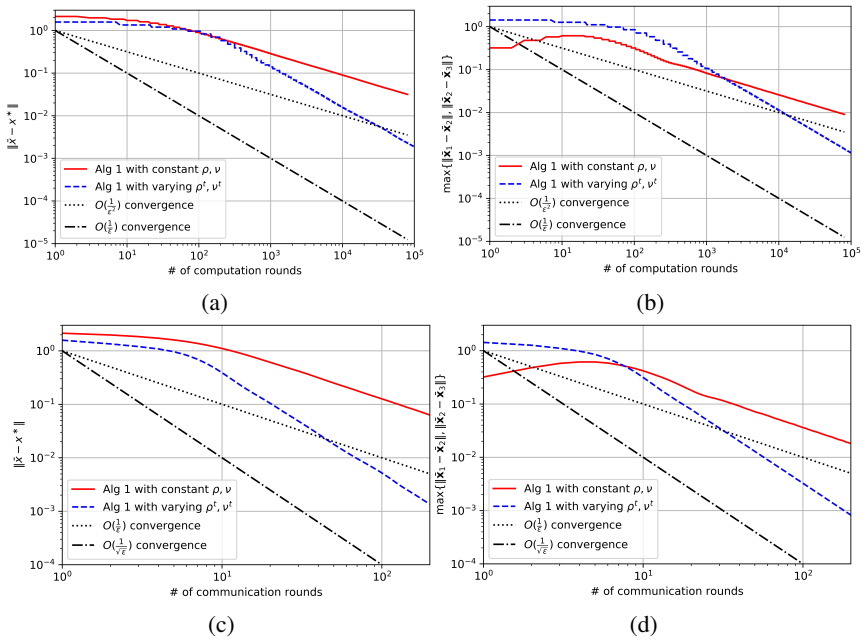

Figure 1: Performance of Algorithm 1 to solve stochastic optimization (16)-(18): (a)& (b) convergence w.r.t. # of computation rounds; (c)&(d) convergence w.r.t. # of communication rounds.

[3]Since $f(\mathbf{x})$ is also smooth, using constant ρ, ν according to Theorem 2 can give a similar (slightly better) performance. Theoretically, by using $K^{(t)} = t$ rather than $K^{(t)} = T$ for a fixed T, the rate is slightly worse, i.e. $O(\log(T)/T)$ v.s. $O(1/T)$. However, we find the performance degradation for large T regions is negligible when using $K^{(t)} = t$. In contrast, using $K^{(t)} = t$ enable the algorithm converge faster for small t. We use $K^{(t)} = t$ when performing the numerical experiments in this paper.

4.2 Distributed l_1 Regularized Logistic Regression

Consider a distributed l_1 regularized logistic regression problem (over 10 nodes) given by:

$$\min \quad \frac{1}{10}\sum_{i=1}^{10}\frac{1}{N_i}\sum_{j=1}^{N_i}\log(1+\exp(b_{ij}(\mathbf{a}_{ij}^{\mathsf{T}}\mathbf{x}_i))+\mu\|\mathbf{x}_i\|_1 \tag{19}$$

with each optimization variable $\mathbf{x}_i \in \mathbb{R}^d$. Each node contains N_i training pairs $(\mathbf{a}_{ij}, b_{ij})$, where $\mathbf{a}_{ij} \in \mathbb{R}^d$ is a feature vector and $b_{ij} \in \{-1, 1\}$ is the corresponding label. To ensure all nodes yield a consistent model, consensus constraints are needed to enforce all $\mathbf{x}_i$ are equal. Note that conventional two-block ADMMs must introduce a dummy block (server node) $\mathbf{z}$ and add constraints $\mathbf{x}_i = \mathbf{z}$. (See e.g., [4, 21, 25].) However, such an ADMM method requires all nodes to pass the updated $\mathbf{x}_i$ value to the (server) node corresponding to the $\mathbf{z}$ block and hence can turn $\mathbf{z}$ node into a communication bottleneck in large networks. In contrast, using a multi-block ADMM method allows arbitrary linear constraints, e.g., constraints $\mathbf{x}_i = \mathbf{x}_{i+1}, \forall i$ that ensure all $\mathbf{x}_i$ are equal, and the corresponding multi-block ADMM only uses communication between adjacent blocks. Alternatively, consider a line network where only one-hop transmission is allowed, then our ADMM naturally yields a protocol that is faithful to the network communication restriction. In general, given an arbitrary network communication topology, our multi-block ADMM can always yield an implementable distributed protocol by adding constraints $\mathbf{x}_i = \mathbf{x}_j$ for links (i, j) existing in the network.

We generate a problem instance in a way similarly to [4]. Our problem instance uses $d = 100$, $N_i = 10^5$ for all i and $\mu = 0.002$. Each feature vector $\mathbf{a}_{ij}$ is generated from a standard normal distribution. We choose a true weight vector $\mathbf{x}^{\text{true}} \in \mathbb{R}^d$ with 10 non-zero entries from a standard normal distribution and then generate the label $b_{ij} = \text{sign}(\mathbf{a}_{ij}^{\mathsf{T}}\mathbf{x}^{\text{true}} + n_i)$ where noise $n_i \sim \mathcal{N}(0, \sigma_i^2)$ with fixed constants σ_i randomly generated from a uniform distribution Unif$[0, 1]$. Figures 2 compares Algorithm 1 with RPDBUS ADMM proposed in [11], where the number of communication rounds is the same that of computation rounds, and DCS in [17], where the number of communication rounds is the square root of that of computation rounds. We observe that Algorithm 1 has fastest convergence with respect to both computation and communication.

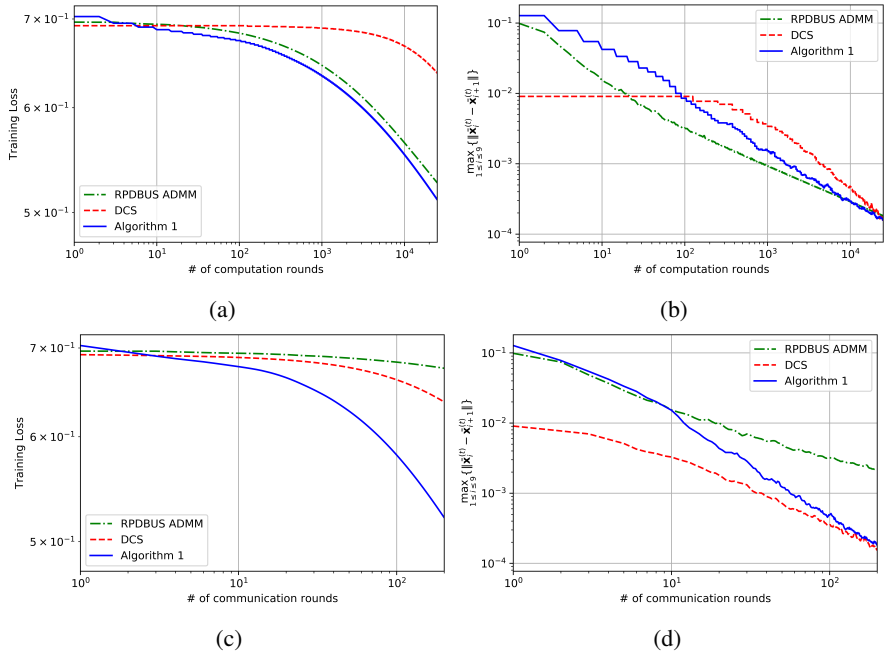

Figure 2: Distributed l_1 regularized logistic regression: (a)& (b) performance w.r.t. # of computation rounds; (c)&(d) performance w.r.t. # of communication rounds

5 Conclusions

This paper proposes a new communication efficient multi-block ADMM for linearly constrained stochastic optimization. This method is as fast as (or faster than) existing stochastic ADMMs but the associated communication overhead is only the square root of that required by existing ADMMs.

References

[1] Samaneh Azadi and Suvrit Sra. Towards an optimal stochastic alternating direction method of multipliers. In *International Conference on Machine Learning (ICML)*, pages 620–628, 2014.

[2] Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, second edition, 1999.

[3] Dimitri P. Bertsekas, Angelia Nedić, and Asuman E. Ozdaglar. *Convex Analysis and Optimization*. Athena Scientific, 2003.

[4] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

[5] Tsung-Hui Chang, Mingyi Hong, Wei-Cheng Liao, and Xiangfeng Wang. Asynchronous distributed admm for large-scale optimization—part i: Algorithm and convergence analysis. *IEEE Transactions on Signal Processing*, 64(12):3118–3130, 2016.

[6] Caihua Chen, Bingsheng He, Yinyu Ye, and Xiaoming Yuan. The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent. *Mathematical Programming*, 155:57–79, 2016.

[7] Wei Deng, Ming-Jun Lai, Zhimin Peng, and Wotao Yin. Parallel multi-block ADMM with $o(1/k)$ convergence. *Journal of Scientific Computing*, 71(2):712–736, 2017.

[8] Wei Deng and Wotao Yin. On the global and linear convergence of the generalized alternating direction method of multipliers. *Journal of Scientific Computing*, 66(3):889–916, 2016.

[9] Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

[10] Jonathan Eckstein and Dimitri P Bertsekas. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1-3):293–318, 1992.

[11] Xiang Gao, Yangyang Xu, and Shuzhong Zhang. Randomized primal-dual proximal block coordinate updates. *arXiv:1605.05969*, 2016.

[12] Bingsheng He, Hong-Kun Xu, and Xiaoming Yuan. On the proximal Jacobian decomposition of ALM for multiple-block separable convex minimization problems and its relationship to ADMM. *Journal of Scientific Computing*, 66(3):1204–1217, 2016.

[13] Bingsheng He and Xiaoming Yuan. On the $O(1/n)$ convergence rate of the Douglas-Rachford alternating direction method. *SIAM Journal on Numerical Analysis*, 50(2):700–709, 2012.

[14] Martin Jaggi, Virginia Smith, Martin Takác, Jonathan Terhorst, Sanjay Krishnan, Thomas Hofmann, and Michael I Jordan. Communication-efficient distributed dual coordinate ascent. *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[15] Jakub Konečnỳ, H Brendan McMahan, Daniel Ramage, and Peter Richtárik. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv:1610.02527*, 2016.

[16] Simon Lacoste-Julien, Mark Schmidt, and Francis Bach. A simpler approach to obtaining an $O(1/t)$ convergence rate for the projected stochastic subgradient method. *arXiv:1212.2002*, 2012.

[17] Guanghui Lan, Soomin Lee, and Yi Zhou. Communication-efficient algorithms for decentralized and stochastic optimization. *arXiv:1701.03961*, 2017.

[18] Huan Li, Cong Fang, and Zhouchen Lin. Convergence rates analysis of the quadratic penalty method and its applications to decentralized distributed optimization. *arXiv:1711.10802*, 2017.

[19] Angelia Nedić, Alex Olshevsky, and Michael G Rabbat. Network topology and communication-computation tradeoffs in decentralized optimization. *Proceedings of the IEEE*, 106(5):953–976, 2018.

[20] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

[21] Hua Ouyang, Niao He, Long Tran, and Alexander Gray. Stochastic alternating direction method of multipliers. In *International Conference on Machine Learning (ICML)*, 2013.

[22] Shi Pu and Angelia Nedić. Distributed stochastic gradient tracking methods. *arXiv:1805.11454*, 2018.

[23] Kevin Scaman, Francis Bach, Sébastien Bubeck, Yin Tat Lee, and Laurent Massoulié. Optimal algorithms for non-smooth distributed optimization in networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[24] Virginia Smith, Simone Forte, Chenxin Ma, Martin Takác, Michael I Jordan, and Martin Jaggi. CoCoA: A general framework for communication-efficient distributed optimization. *arXiv:1611.02189*, 2016.

[25] Taiji Suzuki. Dual averaging and proximal gradient descent for online alternating direction multiplier method. In *International Conference on Machine Learning (ICML)*, 2013.

[26] César A Uribe, Soomin Lee, Alexander Gasnikov, and Angelia Nedić. Optimal algorithms for distributed optimization. *ArXiv:1712.00232*, 2017.

[27] Huahua Wang, Arindam Banerjee, and Zhi-Quan Luo. Parallel direction method of multipliers. *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[28] Ermin Wei and Asuman Ozdaglar. Distributed alternating direction method of multipliers. In *IEEE Conference on Decision and Control (CDC)*, pages 5445–5450, 2012.

[29] Tianbao Yang. Trading computation for communication: Distributed stochastic dual coordinate ascent. *Advances in Neural Information Processing Systems (NIPS)*, 2013.

[30] Hao Yu and Michael J. Neely. A simple parallel algorithm with an $O(1/t)$ convergence rate for general convex programs. *SIAM Journal on Optimization*, 27(2):759–783, 2017.

[31] Wenliang Zhong and James Kwok. Fast stochastic alternating direction method of multipliers. In *International Conference on Machine Learning (ICML)*, pages 46–54, 2014.

6 Supplement

6.1 Connection between Algorithm 1 and Existing ADMMs

Note that Algorithm 1 uses the same Lagrange multiplier update as most existing ADMM methods. The Lagrange multiplier update (4) is helpful to enforce the linear constraint. (See Lemma 3 in Supplement 6.3 for a more technical justification.) However, the per-iteration $\mathbf{x}_i^{(t)}$ updates in Algorithm 1 introduce new stochastic functions $\phi_i^{(t)}(\mathbf{x}_i)$ and let each node call a SGD sub-procedure locally to minimize $\phi_i^{(t)}(\mathbf{x}_i)$. This is quite different from existing deterministic ADMMs [4, 7], which require to solve an "*argmin*" problem exactly, or existing stochastic ADMMs [21, 25, 11], which perform a single gradient descent step. The "*argmin*" update is fundamentally impossible for stochastic minimization since the stochastic objective function is fundamentally unknown and can only be sampled. Our intuition is existing stochastic ADMMs are too conservative in updating $\mathbf{x}_i$ by restricting themselves to a **single** gradient descent update and then communicate immediately for the Lagrange multiplier update. In contrast, our Algorithm 1 introduces the SGD sub-procedure (Algorithm 2) for each node to update $\mathbf{x}_i$ using $K^{(t)}$ gradient descent steps. Such SGD sub-procedures only involve local computations and do not incur any inter-node communication. This is the key reason why our Algorithm 1 requires fewer communication rounds than computation rounds. It is tempting to interpret Algorithm 1 as an ADMM variant where the "*argmin*" primal update is only approximately solved using local SGD sub-procedures. Previous work [10] considers ADMM variants with inexact "*argmin*" primal updates for **deterministic** optimization without analyzing the convergence rate. However, our Algorithm 1 is different from the method in [10] and can solve more challenging stochastic optimization with convergence rate guarantees.

It remains to see how we come up with $\phi_i^{(t)}(\mathbf{x}_i)$ in (2). To see so, we introduce Algorithm 3 that generalizes the deterministic multi-block ADMM in [7] and provide new insight and analysis.

Algorithm 3 Deterministic Multi-Block Proximal Jacobi ADMM (generalized from [7])

1: **Input:** Algorithm parameters: $\{\mathbf{P}_i^{(t)}\}_{t\geq 1, i\in\{1,2,\ldots,N\}}$ with $\mathbf{P}_i^{(t)} \succeq 0, \forall i, \forall t$; $\{\rho^{(t)}\}_{t\geq 1}$.
2: Initialize arbitrary $\mathbf{x}_i^{(0)} \in \mathcal{X}_i, \forall i$, $\boldsymbol{\lambda}^{(0)} = \mathbf{0}$ and $t = 1$.
3: **while** $t \leq T$ **do**
4: Update each $\mathbf{x}_i^{(t)}$ **in parallel** equal to

$$\operatorname*{argmin}_{\mathbf{x}_i\in\mathcal{X}_i} \Big\{ f_i(\mathbf{x}_i) + \frac{\rho^{(t)}}{2}\|\mathbf{A}_i\mathbf{x}_i + \sum_{j\neq i}\mathbf{A}_j\mathbf{x}_j^{(t-1)} - \mathbf{b} + \frac{1}{\rho^{(t)}}\boldsymbol{\lambda}^{(t-1)}\|^2 + \frac{1}{2}\|\mathbf{x}_i - \mathbf{x}_i^{(t-1)}\|^2_{\mathbf{P}_i^{(t)}} \Big\}. \quad (20)$$

5: Update $\boldsymbol{\lambda}^{(t)}$ according to (4).
6: Update $t \leftarrow t + 1$.
7: **end while**
8: **Output:** $\overline{\mathbf{x}}^{(T)} = \frac{1}{\sum_{t=1}^T \rho^{(t)}} \sum_{t=1}^T \rho^{(t)}\mathbf{x}^{(t)}$

Algorithm 3 is almost identical to the original parallel multi-block ADMM proposed in [12, 7] except that it allows $\mathbf{P}_i^{(t)}$ and $\rho^{(t)}$ to be time-varying. We will show later that time-varying $\mathbf{P}_i^{(t)}$ and $\rho^{(t)}$ are useful for Algorithm 3 to achieve faster $O(1/\sqrt{\epsilon})$ convergence for problems with strongly convex objective functions. Note that if we take $\mathbf{P}_i^{(t)} = \nu^{(t)}\mathbf{I} - \rho^{(t)}\mathbf{A}_i^\mathsf{T}\mathbf{A}_i$ with scalar $\nu^{(t)} > 0$, then (20) in Algorithm 3 is equivalent to

$$\mathbf{x}_i^{(t)} = \operatorname*{argmin}_{\mathbf{x}_i\in\mathcal{X}_i} \Big\{ f_i(\mathbf{x}_i) + \rho^{(t)}\langle \mathbf{A}_i^\mathsf{T}\Big(\sum_{i=1}^N \mathbf{A}_i\mathbf{x}_i^{(t-1)} - \mathbf{b} + \frac{1}{\rho^{(t)}}\boldsymbol{\lambda}^{(t-1)}\Big), \mathbf{x}_i\rangle + \frac{\nu^{(t)}}{2}\|\mathbf{x}_i - \mathbf{x}_i^{(t-1)}\|^2 \Big\} \quad (21)$$

$$= \operatorname*{argmin}_{\mathbf{x}_i\in\mathcal{X}_i} \Big\{ f_i(\mathbf{x}_i) + \rho^{(t)}\langle \sum_{i=1}^N \mathbf{A}_i\mathbf{x}_i^{(t-1)} - \mathbf{b} + \frac{1}{\rho^{(t)}}\boldsymbol{\lambda}^{(t-1)}, \mathbf{A}_i\mathbf{x}_i - \frac{\mathbf{b}}{N}\rangle + \frac{\nu^{(t)}}{2}\|\mathbf{x}_i - \mathbf{x}_i^{(t-1)}\|^2 \Big\} \quad (22)$$

Since both $\langle \mathbf{c}, \mathbf{x}_i\rangle$ and $\|\mathbf{x}_i - \mathbf{x}_i^{(t-1)}\|^2$ are separable (with respect to each component of vector $\mathbf{x}_i$), the equivalent minimization step (21) or (22) can be further decomposed into d_i simple scalar

minimization subproblems if $f_i(\mathbf{x}_i)$ is also separable, e.g. linear $f_i(\mathbf{x}_i)$ or $f_i(\mathbf{x}_i) = \|\mathbf{x}_i\|_1$. Thus, suitable choices of $\mathbf{P}_i^{(t)}$ can remarkably reduce the implementation complexity of Algorithm 3 and enable the parallelism of $\mathbf{x}_i^{(t)}$ updates for different i. See [8, 7] for more discussions on the benefit of introducing $\mathbf{P}_i^{(t)}$.

Note $\mathbf{r}^{(t-1)} = \sum_{i=1}^N \mathbf{A}_i \mathbf{y}_i^{(t-1)} - \mathbf{b}$ in Algorithm 1, it now becomes transparent how $\phi_i^{(t)}(\mathbf{x}_i)$ in (2) is developed in Algorithm 1. Each $\phi_i^{(t)}(\mathbf{x}_i)$ is obtained by replacing each $\mathbf{x}_i^{(t-1)}$ in expression (22) with a newly introduced variable $\mathbf{y}_i^{(t-1)}$. Algorithm 1 then further call a SGD sub-procedure (Algorithm 2) to minimize $\phi_i^{(t)}(\mathbf{x}_i)$. The introduction of $\mathbf{y}_i^{(t-1)}$ is to compensate the error accumulated in the SGD sub-procedures and is further justified in Section 2.1.

To further motivate the development of Algorithm 1 from Algorithm 3, the next theorem summarizes the convergence of Algorithm 3 for **deterministic** convex programs:

Theorem 3. *Consider convex programs in the form of* (1) *with μ-convex (possibly non-smooth) deterministic $f(\mathbf{x})$. Let $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ be a saddle point in Assumption 1.*

1. **General Convex ($\mu = 0$)***: If we choose any fixed $\rho^{(t)} = \rho > 0$ and $\mathbf{P}_i^{(t)} = \mathbf{P}_i = \nu\mathbf{I} - \rho\mathbf{A}_i^\mathsf{T}\mathbf{A}_i$ with $\nu \geq \rho\|\mathbf{A}\|^2$ in Algorithm 3, then we have*

$$f(\overline{\mathbf{x}}^{(T)}) \leq f(\mathbf{x}^*) + \frac{1}{2T}\|\mathbf{x}^* - \mathbf{x}^{(0)}\|_{\mathbf{Q}}^2 \tag{23}$$

$$\|\mathbf{A}\overline{\mathbf{x}}^{(T)} - \mathbf{b}\| \leq \frac{1}{T}\frac{2\|\boldsymbol{\lambda}^*\|}{\rho} + \frac{1}{T}\frac{\|\mathbf{x}^* - \mathbf{x}^{(0)}\|_{\mathbf{Q}}}{\sqrt{\rho}} \tag{24}$$

where $\overline{\mathbf{x}}^{(T)} = \frac{1}{\sum_{t=1}^T \rho}\sum_{t=1}^T \rho\mathbf{x}^{(t)} = \frac{1}{T}\sum_{t=1}^T \mathbf{x}^{(t)}$; $\mathbf{Q} = Diag(\mathbf{Q}_1, \ldots, \mathbf{Q}_N) = Diag(\mathbf{P}_1 + \rho\mathbf{A}_1^\mathsf{T}\mathbf{A}_1, \ldots, \mathbf{P}_N + \rho\mathbf{A}_N^\mathsf{T}\mathbf{A}_N)$.

2. **Strongly Convex ($\mu > 0$)***: If we choose $\rho \leq \frac{\mu}{3\|\mathbf{A}\|^2}$, $\rho^{(t)} = t\rho$ and $\mathbf{P}_i^{(t)} = t\rho\|\mathbf{A}\|^2\mathbf{I} - t\rho\mathbf{A}_i^\mathsf{T}\mathbf{A}_i$ in Algorithm 3, then we have*

$$f(\overline{\mathbf{x}}^{(T)}) \leq f(\mathbf{x}^*) + \frac{\rho}{T(T+1)}\|\mathbf{A}\|^2\|\mathbf{x}^* - \mathbf{x}^{(0)}\|^2 \tag{25}$$

$$\|\mathbf{A}\overline{\mathbf{x}}^{(T)} - \mathbf{b}\| \leq \frac{4\|\boldsymbol{\lambda}^*\|}{\rho T(T+1)} + \frac{2\|\mathbf{A}\|\|\mathbf{x}^* - \mathbf{x}^{(0)}\|}{T(T+1)} \tag{26}$$

Proof. See Supplement 6.5. □

Remark 5. *It is sufficient to use any constant ρ to ensure $O(1/T)$ convergence for the $\mu = 0$ case. However, a larger ρ yields larger objective error (note that $\|\cdot\|_{\mathbf{Q}}^2 = O(\rho)$) and smaller constraint error. Thus, ρ can be controlled to trade off between objective error and constraint error. Similar tradeoffs also hold for the $\mu > 0$ case (as long as ρ satisfies the condition ensuring the algorithm convergence) and other algorithms in this paper.*

Remark 6. *For the $\mu = 0$ case, Algorithm 3 with fixed algorithm parameters degrades to the proximal Jacobi ADMM considered in [7]). However, the convergence rate shown in [7] is in the weak form of $\|\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)}\|^2 \leq o(1/t)$ and does not necessarily mean[4] the convergence for the objective value or feasibility shown in Theorem 3. In contrast, our Theorem 3 proves the $O(1/T)$ convergence rate of Algorithm 3 regarding the objective value and feasibility, which is the concern for math optimization.*

A similar $O(1/T)$ convergence rate, or equivalently, $O(1/\epsilon)$ convergence time, for $\mu = 0$ case is independently shown in [11] for an ADMM variant different from Algorithm 3. In Supplement 6.5, we provide a different analysis that unifies both $\mu = 0$ and $\mu > 0$ cases. To our knowledge, the $O(1/T^2)$ convergence rate of Algorithm 3 with time-varying parameters for $\mu > 0$ case (with possibly non-smooth $f(\mathbf{x})$ and arbitrary matrix $\mathbf{A}$) is new. Existing faster convergence of ADMM for strongly convex programs requires additional conditions of $f(\mathbf{x})$ and/or $\mathbf{A}$.

[4] In fact, the $\|\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)}\|^2 \leq o(1/t)$ convergence is so weak that it does not even imply $\mathbf{x}^t$ converges to a fixed $\mathbf{x}^*$. For example, the scalar sequence $x^{(t)} = t^{1/4}$ satisfies $\|x^{(t+1)} - x^{(t)}\|^2 \leq o(1/t)$ but diverges to ∞.

6.2 Analysis Technique in This Paper

Note that our analysis technique for the proximal Jacobi ADMM described in Algorithm 3 and the communication efficient stochastic ADMM in Algorithm 1 is different from the analysis for conventional Jacobi type ADMM as in [7]. The analysis in this paper is extended from [30] where a proximal Lagrangian based method is developed for convex programs with possibly non-linear constraints. By utilizing the simpler linear constraint structure, we obtain finer convergence rate results for Algorithm 3 and further establish the computation and communication complexity for Algorithm 1.

6.3 Basic Facts from Lagrange Multiplier Updates

In this section, we present two lemmas that hold for any algorithm using (4) to update $\boldsymbol{\lambda}$. These two lemmas are frequently used to analyze the feasibility violations in this paper.

Lemma 3. *Let $\boldsymbol{\lambda}^{(0)} = \mathbf{0}$ and $\boldsymbol{\lambda}^{(t)}, t \geq 1$ be updated according to* (4)*.*

1. *For any $T \geq 1$, we have $\sum_{t=1}^{T} \rho^{(t)} \left(\mathbf{A}\mathbf{x}^{(t)} - \mathbf{b}\right) = \boldsymbol{\lambda}^{(T)}$*
2. *For all $t \geq 1$, we have $\left\langle \boldsymbol{\lambda}^{(t-1)}, \mathbf{A}\mathbf{x}^{(t)} - \mathbf{b} \right\rangle = \frac{1}{2\rho^{(t)}} \left(\|\boldsymbol{\lambda}^{(t)}\|^2 - \|\boldsymbol{\lambda}^{(t-1)}\|^2 \right) - \frac{\rho^{(t)}}{2} \|\mathbf{A}\mathbf{x}^{(t)} - \mathbf{b}\|^2$.*

Proof.

1. This follows directly from the update equation (4).
2. Fix $t \geq 1$. Taking the squared vector l_2 norm on both sides of (4) yields

$$\|\boldsymbol{\lambda}^{(t)}\|^2 = \|\boldsymbol{\lambda}^{(t-1)}\|^2 + (\rho^{(t)})^2 \| \sum_{i=1}^{N} \mathbf{A}_i \mathbf{x}_i^{(t)} - \mathbf{b}\|^2 + 2\rho^{(t)} \left\langle \boldsymbol{\lambda}^{(t-1)}, \mathbf{A}\mathbf{x}^{(t)} - \mathbf{b} \right\rangle.$$

   This part follows by dividing by $2\rho^{(t)}$ on both sides and rearranging terms.

□

Note that part (1) of lemma implies that to analyze the accumulated feasibility violations over T iterations, it is sufficient to analyze the boundedness of $\boldsymbol{\lambda}^{(T)}$. The next lemma follows directly from the saddle point assumption (Assumption 1) and relates $\boldsymbol{\lambda}^T$ with the accumulated objective performance.

Lemma 4. *Consider convex program* (1) *under Assumption 1 such that $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ is any saddle point defined in Assumption 1. For any $T \geq 1$, if an algorithm generates $\mathbf{x}^{(t)} \in \mathcal{X}$ and updates $\boldsymbol{\lambda}^{(t)}$ according to* (4)*(with $\boldsymbol{\lambda}^{(0)} = \mathbf{0}$) at each iteration $t \in \{1, 2, \ldots, T\}$, then we have*

$$\sum_{t=1}^{T} \rho^{(t)} f(\mathbf{x}^{(t)}) \geq \sum_{t=1}^{T} \rho^{(t)} f(\mathbf{x}^*) - \|\boldsymbol{\lambda}^*\| \|\boldsymbol{\lambda}^{(T)}\|$$

Proof. Fix $T > 0$. For any $t \in \{1, \ldots, T\}$, by Assumption 1, we have

$$f(\mathbf{x}^*) = q(\boldsymbol{\lambda}^*) \triangleq \inf_{\mathbf{x} \in \mathcal{X}} \{ f(\mathbf{x}) + \langle \boldsymbol{\lambda}^*, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle \} \overset{(a)}{\leq} f(\mathbf{x}^{(t)}) + \left\langle \boldsymbol{\lambda}^*, \mathbf{A}\mathbf{x}^{(t)} - \mathbf{b} \right\rangle$$

where (a) trivially follows because $\mathbf{x}^{(t)} \in \mathcal{X}$. Multiplying $\rho^{(t)}$ on both sides and summing over $t \in \{1, 2, \ldots, T\}$ yields

$$\begin{aligned} \sum_{t=1}^{T} \rho^{(t)} f(\mathbf{x}^*) &\leq \sum_{t=1}^{T} \rho^{(t)} f(\mathbf{x}^{(t)}) + \left\langle \boldsymbol{\lambda}^*, \sum_{t=1}^{T} \rho^{(t)} (\mathbf{A}\mathbf{x}^{(t)} - \mathbf{b}) \right\rangle \\ &\overset{(a)}{=} \sum_{t=1}^{T} \rho^{(t)} f(\mathbf{x}^{(t)}) + \langle \boldsymbol{\lambda}^*, \boldsymbol{\lambda}^{(T)} \rangle \end{aligned}$$

$$\overset{(b)}{\leq} \sum_{t=1}^{T} \rho^{(t)} f(\mathbf{x}^{(t)}) + \|\boldsymbol{\lambda}^*\| \|\boldsymbol{\lambda}^{(T)}\|$$

where (a) follows from part (1) of Lemma 3 and (b) follows from the Cauchy-Schwarz inequality. □

6.4 New Facts on Convex Analysis

Recall the following important fact on the minimizer of strongly convex functions:

Lemma 5 (See e.g. Corollary 1 in [30])**.** *Let $h : \mathcal{X} \to \mathbb{R}$ be a strongly convex function, i.e., μ-convex with $\mu > 0$, and $\mathbf{x}^{min} \in \mathcal{X}$ be a point that minimizes h over set $\mathcal{X}$, then*

$$h(\mathbf{x}^{min}) \leq h(\mathbf{x}) - \frac{\mu}{2}\|\mathbf{x} - \mathbf{x}^{min}\|^2 \quad \forall \mathbf{x} \in \mathcal{X}.$$

Note that this fact holds trivially for convex functions without strong convexity (μ-convex functions with $\mu = 0$). We now extend the above property for a convex function given by $h(\mathbf{x}) = g(\mathbf{x}) + \frac{1}{2}\|\mathbf{x}\|^2_{\mathbf{Q}}$ where $g(\mathbf{x})$ is a μ-convex function and $\mathbf{Q} \succeq \mathbf{0}$ is a symmetric semidefinite positive matrix, in the following lemma:

Lemma 6. *Let $h : \mathcal{X} \to \mathbb{R}$ be defined as $h(\mathbf{x}) = g(\mathbf{x}) + \frac{1}{2}\|\mathbf{x}\|^2_{\mathbf{Q}}$ where $g(\mathbf{x})$ is a μ-convex function and $\mathbf{Q} \succeq \mathbf{0}$ is a symmetric semidefinite positive matrix. If $\mathbf{x}^{min} \in \mathcal{X}$ is a point that minimizes h over set $\mathcal{X}$, then*

$$h(\mathbf{x}^{min}) \leq h(\mathbf{x}) - \frac{1}{2}\|\mathbf{x} - \mathbf{x}^{min}\|^2_{\mathbf{Q}+\mu\mathbf{I}} \quad \forall \mathbf{x} \in \mathcal{X}.$$

Since matrix $\mathbf{Q}$ can be rank deficient, the function $\frac{1}{2}\|\mathbf{x}\|^2_{\mathbf{Q}}$ is not necessarily strongly convex. Thus, $h(\mathbf{x}) = g(\mathbf{x}) + \frac{1}{2}\|\mathbf{x}\|^2_{\mathbf{Q}}$ is in general μ-convex. By Lemma 5, we can only say $h(\mathbf{x}^{min}) \leq h(\mathbf{x}) - \frac{1}{2}\|\mathbf{x} - \mathbf{x}^{min}\|^2_{\mu\mathbf{I}}$ for all $\mathbf{x} \in \mathcal{X}$, which is weaker than the inequality in Lemma 6.

The following lemma will be useful to prove Lemma 6

Lemma 7. *Let $h : \mathcal{X} \to \mathbb{R}$ be defined as $h(\mathbf{x}) = g(\mathbf{x}) + \frac{1}{2}\|\mathbf{x}\|^2_{\mathbf{Q}}$ where $g(\mathbf{x})$ is a μ-convex function and $\mathbf{Q} \succeq \mathbf{0}$ is a symmetric semidefinite positive matrix. Let $\partial h(\mathbf{x})$ be the set of all subgradients of h at point $\mathbf{x}$. Then*

$$h(\mathbf{y}) \geq h(\mathbf{x}) + \langle \mathbf{d}, \mathbf{y} - \mathbf{x} \rangle + \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|^2_{\mathbf{Q}+\mu\mathbf{I}}$$

for all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$ and all $\mathbf{d} \in \partial h(\mathbf{x})$.

Proof. Define $\phi(\mathbf{x}) = h(\mathbf{x}) - \frac{\mu}{2}\|\mathbf{x}\|^2 - \frac{1}{2}\|\mathbf{x}\|^2_{\mathbf{Q}} = h(\mathbf{x}) - \frac{1}{2}\|\mathbf{x}\|^2_{\mathbf{Q}+\mu\mathbf{I}}$. Since $h(\mathbf{x}) = g(\mathbf{x}) + \frac{1}{2}\|\mathbf{x}\|^2_{\mathbf{Q}}$, we known $\phi(\mathbf{x})$ is a convex function. Let $\partial\phi(\mathbf{x})$ denote the set of all subgradients of ϕ at point $\mathbf{x}$, then $\partial\phi(\mathbf{x}) = \partial h(\mathbf{x}) - (\mathbf{Q} + \mu\mathbf{I})\mathbf{x} = \{\mathbf{d} - (\mathbf{Q} + \mu\mathbf{I})\mathbf{x} \mid \mathbf{d} \in \partial h(\mathbf{x})\}$. By convexity of ϕ, for all $\mathbf{d} \in \partial h(\mathbf{x})$ and all $\mathbf{x}, \mathbf{y} \in \mathcal{X}$, we have

$$\begin{aligned}\phi(\mathbf{y}) \geq& \phi(\mathbf{x}) + \langle \mathbf{d} - (\mathbf{Q} + \mu\mathbf{I})\mathbf{x}, \mathbf{y} - \mathbf{x} \rangle \\ =& \phi(\mathbf{x}) + \|\mathbf{x}\|^2_{\mathbf{Q}+\mu\mathbf{I}} + \langle \mathbf{d}, \mathbf{y} - \mathbf{x} \rangle - \langle (\mathbf{Q} + \mu\mathbf{I})\mathbf{x}, \mathbf{y} \rangle\end{aligned}$$

Substituting $\phi(\mathbf{x}) = h(\mathbf{x}) - \frac{1}{2}\|\mathbf{x}\|^2_{\mathbf{Q}+\mu\mathbf{I}}$ and $\phi(\mathbf{y}) = h(\mathbf{y}) - \frac{1}{2}\|\mathbf{y}\|^2_{\mathbf{Q}+\mu\mathbf{I}}$ into it and rearranging terms (noting that $\mathbf{Q} + \mu\mathbf{I}$ is symmetric) yields

$$h(\mathbf{y}) \geq h(\mathbf{x}) + \langle \mathbf{d}, \mathbf{y} - \mathbf{x} \rangle + \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|^2_{\mathbf{Q}+\mu\mathbf{I}}$$

□

Now we are ready to prove Lemma 6:

Proof of Lemma 6: Fix $\mathbf{x} \in \mathcal{X}$. Note that h is also convex. By the first order optimality condition of convex functions, e.g., Proposition B.24 (f) in [2], there exists $\mathbf{d} \in \partial h(\mathbf{x}^{min})$ such that $\langle \mathbf{d}, \mathbf{x} - \mathbf{x}^{min} \rangle \geq 0$. By Lemma 7, we also have

$$\begin{aligned}h(\mathbf{x}) &\geq h(\mathbf{x}^{min}) + \langle \mathbf{d}, \mathbf{x} - \mathbf{x}^{min} \rangle + \frac{1}{2}\|\mathbf{x} - \mathbf{x}^{min}\|^2_{\mathbf{Q}+\mu\mathbf{I}} \\ &\overset{(a)}{\geq} h(\mathbf{x}^{min}) + \frac{1}{2}\|\mathbf{x} - \mathbf{x}^{min}\|^2_{\mathbf{Q}+\mu\mathbf{I}},\end{aligned}$$

where (a) follows from the fact that $\langle \mathbf{d}, \mathbf{x} - \mathbf{x}^{min} \rangle \geq 0$.

Corollary 1. *Let* $\mathbf{c}$ *be a fixed constant vector and* $h(\mathbf{x}) = g(\mathbf{x}) + \frac{1}{2}\|\mathbf{x}-\mathbf{c}\|^2_{\mathbf{Q}}$ *where* $g(\mathbf{x})$ *is a* μ*-convex function and* $\mathbf{Q} \succeq \mathbf{0}$ *is a symmetric semidefinite positive matrix. If* $\mathbf{x}^{min} \in \mathcal{X}$ *be a point that minimizes* h *over set* $\mathcal{X}$*, then*

$$h(\mathbf{x}^{min}) \leq h(\mathbf{x}) - \frac{1}{2}\|\mathbf{x}-\mathbf{x}^{min}\|^2_{\mathbf{Q}+\mu\mathbf{I}} \quad \forall \mathbf{x} \in \mathcal{X}.$$

Proof. Let $\tilde{g}(\mathbf{x}) = g(\mathbf{x}) + \frac{1}{2}\|\mathbf{c}\|^2_{\mathbf{Q}} + \langle \mathbf{c}, \mathbf{x}\rangle$. Note that $\tilde{g}(\mathbf{x})$ is μ-convex as long as $g(\mathbf{x})$ is. We further note that $h(\mathbf{x}) = \tilde{g}(\mathbf{x}) + \frac{1}{2}\|\mathbf{x}\|^2_{\mathbf{Q}}$, which is a summation of μ-convex function and $\frac{1}{2}\|\mathbf{x}\|^2_{\mathbf{Q}}$. Thus, this corollary follows directly from Lemma 6. □

6.5 Proof of Theorem 3

The proof is built upon Corollary 1 from Section 6.4 and a different interpretation of the $\mathbf{x}^{(t)}$ update in Algorithm 3.

Lemma 8. *The update in* (20) *(Algorithm 3) is equivalent to*

$$\mathbf{x}_i^{(t)} = \underset{\mathbf{x}_i \in \mathcal{X}_i}{\textit{argmin}} \left\{ f_i(\mathbf{x}_i) + \rho^{(t)} \left\langle \sum_{i=1}^N \mathbf{A}_i \mathbf{x}_i^{(t-1)} - \mathbf{b} + \frac{1}{\rho^{(t)}}\boldsymbol{\lambda}^{(t-1)}, \mathbf{A}_i\mathbf{x}_i - \frac{\mathbf{b}}{N} \right\rangle + \frac{1}{2}\|\mathbf{x}_i - \mathbf{x}_i^{(t-1)}\|^2_{\mathbf{Q}_i^{(t)}} \right\}, \tag{27}$$

$\forall i \in \{1, 2, \ldots, N\}$ *with* $\mathbf{Q}_i^{(t)} = \mathbf{P}_i^{(t)} + \rho^{(t)}\mathbf{A}_i^\top \mathbf{A}_i \succeq 0$.

Proof. Note that $\boldsymbol{\lambda}^{(t-1)}$ and $\mathbf{x}_i^{(t-1)}$ are given constants in (27). This lemma follows by noting that (27) is equivalent to

$$\begin{aligned}
&\underset{\mathbf{x}_i \in \mathcal{X}_i}{\text{argmin}} \left\{ f_i(\mathbf{x}_i) + \rho^{(t)} \left\langle \sum_{i=1}^N \mathbf{A}_i \mathbf{x}_i^{(t-1)} - \mathbf{b} + \frac{1}{\rho^{(t)}}\boldsymbol{\lambda}^{(t-1)}, \mathbf{A}_i\mathbf{x}_i - \frac{\mathbf{b}}{N} \right\rangle + \frac{\rho^{(t)}}{2}\|\mathbf{A}_i(\mathbf{x}_i - \mathbf{x}_i^{(t-1)})\|^2 \right. \\
&\qquad \left. + \frac{1}{2}\|\mathbf{x}_i - \mathbf{x}_i^{(t-1)}\|^2_{\mathbf{P}_i^{(t)}} \right\} \\
\overset{(a)}{\Leftrightarrow} &\underset{\mathbf{x}_i \in \mathcal{X}_i}{\text{argmin}} \left\{ f_i(\mathbf{x}_i) + \rho^{(t)} \left\langle \sum_{i=1}^N \mathbf{A}_i \mathbf{x}_i^{(t-1)} - \mathbf{b} + \frac{1}{\rho^{(t)}}\boldsymbol{\lambda}^{(t-1)}, \mathbf{A}_i(\mathbf{x}_i - \mathbf{x}_i^{(t-1)}) \right\rangle + \frac{\rho^{(t)}}{2}\|\mathbf{A}_i(\mathbf{x}_i - \mathbf{x}_i^{(t-1)})\|^2 \right. \\
&\qquad \left. + \frac{1}{2}\|\mathbf{x}_i - \mathbf{x}_i^{(t-1)}\|^2_{\mathbf{P}_i^{(t)}} \right\} \\
\overset{(b)}{\Leftrightarrow} &\underset{\mathbf{x}_i \in \mathcal{X}_i}{\text{argmin}} \left\{ f_i(\mathbf{x}_i) + \frac{\rho^{(t)}}{2} \left\| \mathbf{A}_i(\mathbf{x}_i - \mathbf{x}_i^{(t-1)}) + \sum_{i=1}^N \mathbf{A}_i \mathbf{x}_i^{(t-1)} - \mathbf{b} + \frac{1}{\rho^{(t)}}\boldsymbol{\lambda}^{(t-1)} \right\|^2 + \frac{1}{2}\|\mathbf{x}_i - \mathbf{x}_i^{(t-1)}\|^2_{\mathbf{P}_i^{(t)}} \right\} \\
\Leftrightarrow &\underset{\mathbf{x}_i \in \mathcal{X}_i}{\text{argmin}} \left\{ f_i(\mathbf{x}_i) + \frac{\rho^{(t)}}{2} \left\| \mathbf{A}_i\mathbf{x}_i + \sum_{j \neq i} \mathbf{A}_j \mathbf{x}_j^{(t-1)} - \mathbf{b} + \frac{1}{\rho^{(t)}}\boldsymbol{\lambda}^{(t-1)} \right\|^2 + \frac{1}{2}\|\mathbf{x}_i - \mathbf{x}_i^{(t-1)}\|^2_{\mathbf{P}_i^{(t)}} \right\}
\end{aligned}$$

where (a) follows because an argmin solution does not change if we add constant terms to the expression to minimize and (b) follows by completing the square (and adding necessary constant terms for this). □

Corollary 2. *The update in* (20) *(Algorithm 3) is equivalent to*

$$\mathbf{x}^{(t)} = \underset{\mathbf{x} \in \mathcal{X}}{\textit{argmin}} \left\{ f(\mathbf{x}) + \rho^{(t)} \left\langle \mathbf{A}\mathbf{x}^{(t-1)} - \mathbf{b} + \frac{1}{\rho^{(t)}}\boldsymbol{\lambda}^{(t-1)}, \mathbf{A}\mathbf{x} - \mathbf{b} \right\rangle + \frac{1}{2}\|\mathbf{x} - \mathbf{x}^{(t-1)}\|^2_{\mathbf{Q}^{(t)}} \right\}, \tag{28}$$

with

$$\mathbf{Q}^{(t)} \triangleq \textit{Diag}(\mathbf{Q}_1^{(t)}, \ldots, \mathbf{Q}_N^{(t)}) = \textit{Diag}(\mathbf{P}_1^{(t)} + \rho^{(t)}\mathbf{A}_1^\top\mathbf{A}_1, \ldots, \mathbf{P}_N^{(t)} + \rho^{(t)}\mathbf{A}_N^\top\mathbf{A}_N) \tag{29}$$

Proof. Note that the update of each $\mathbf{x}_i^{(t)}$ is fully decoupled in (27). That is, $\mathbf{x}^{(t)}$ chosen by Algorithm 3 is to jointly minimize

$$\begin{aligned}&\sum_{i=1}^N\left[f_i(\mathbf{x}_i)+\rho^{(t)}\left\langle\sum_{i=1}^N\mathbf{A}_i\mathbf{x}_i^{(t-1)}-\mathbf{b}+\frac{1}{\rho^{(t)}}\boldsymbol{\lambda}^{(t-1)},\mathbf{A}_i\mathbf{x}_i-\frac{\mathbf{b}}{N}\right\rangle+\frac{1}{2}\|\mathbf{x}_i-\mathbf{x}_i^{(t-1)}\|_{\mathbf{Q}_i^{(t)}}^2\right]\\=&f(\mathbf{x})+\rho^{(t)}\left\langle\mathbf{A}\mathbf{x}^{(t-1)}-\mathbf{b}+\frac{1}{\rho^{(t)}}\boldsymbol{\lambda}^{(t-1)},\mathbf{A}\mathbf{x}-\mathbf{b}\right\rangle+\frac{1}{2}\|\mathbf{x}-\mathbf{x}^{(t-1)}\|_{\mathbf{Q}^{(t)}}^2\end{aligned}$$

over set $\mathcal{X}=\prod_{i=1}^N\mathbf{x}_i$

□

Lemma 9. *Let $\mathbf{x}^*$ be any optimal solution of problem* (1)*. Let $\mathbf{Q}^{(t)}$ be defined in* (29)*. If $\mathbf{P}_i^{(t)}\succeq 0$ and $\rho^{(t)}>0$ in Algorithm 3 are chosen to satisfy*

$$\mathbf{Q}^{(t)}\succeq\rho^{(t)}\mathbf{A}^\mathsf{T}\mathbf{A} \tag{30}$$

Then, for all $T\geq 1$, Algorithm 3 ensures that

$$\sum_{t=1}^T\rho^{(t)}f(\mathbf{x}^{(t)})\leq\sum_{t=1}^T\rho^{(t)}f(\mathbf{x}^*)+\frac{1}{2}\sum_{t=1}^T\rho^{(t)}\Theta^{(t)}-\frac{1}{2}\|\boldsymbol{\lambda}^{(T)}\|^2$$

where $\Theta^{(t)}\triangleq\|\mathbf{x}^*-\mathbf{x}^{(t-1)}\|_{\mathbf{Q}^{(t)}}^2-\|\mathbf{x}^*-\mathbf{x}^{(t)}\|_{\mu\mathbf{I}+\mathbf{Q}^{(t)}}^2$.

Proof. Fix $T\geq 1$. For any $t\in\{1,2,\ldots,T\}$, by Corollary 2, $\mathbf{x}^{(t)}$ is chosen to minimize $f(\mathbf{x})+\rho^{(t)}\left\langle\mathbf{A}\mathbf{x}^{(t-1)}-\mathbf{b}+\frac{1}{\rho^{(t)}}\boldsymbol{\lambda}^{(t-1)},\mathbf{A}\mathbf{x}-\mathbf{b}\right\rangle+\frac{1}{2}\|\mathbf{x}-\mathbf{x}^{(t-1)}\|_{\mathbf{Q}^{(t)}}^2$ over $\mathbf{x}\in\mathcal{X}$. Note that $f(\mathbf{x})+\rho^{(t)}\left\langle\mathbf{A}\mathbf{x}^{(t-1)}-\mathbf{b}+\frac{1}{\rho^{(t)}}\boldsymbol{\lambda}^{(t-1)},\mathbf{A}\mathbf{x}-\mathbf{b}\right\rangle$ is μ-convex since $f(\mathbf{x})$ is μ-convex. By Corollary 1 (note that $\mathbf{x}^*\in\mathcal{X}$), we have

$$\begin{aligned}&f(\mathbf{x}^{(t)})+\rho^{(t)}\left\langle\mathbf{A}\mathbf{x}^{(t-1)}-\mathbf{b}+\frac{1}{\rho^{(t)}}\boldsymbol{\lambda}^{(t-1)},\mathbf{A}\mathbf{x}^{(t)}-\mathbf{b}\right\rangle+\frac{1}{2}\|\mathbf{x}^{(t)}-\mathbf{x}^{(t-1)}\|_{\mathbf{Q}^{(t)}}^2\\\leq&f(\mathbf{x}^*)+\rho^{(t)}\left\langle\mathbf{A}\mathbf{x}^{(t-1)}-\mathbf{b}+\frac{1}{\rho^{(t)}}\boldsymbol{\lambda}^{(t-1)},\mathbf{A}\mathbf{x}^*-\mathbf{b}\right\rangle+\frac{1}{2}\|\mathbf{x}^*-\mathbf{x}^{(t-1)}\|_{\mathbf{Q}^{(t)}}^2 \qquad (31)\\&-\frac{1}{2}\|\mathbf{x}^*-\mathbf{x}^{(t)}\|_{\mu\mathbf{I}+\mathbf{Q}^{(t)}}^2\\\overset{(a)}{=}&f(\mathbf{x}^*)+\frac{1}{2}\Theta^{(t)} \qquad (32)\end{aligned}$$

where (a) follows because $\mathbf{A}\mathbf{x}^*-\mathbf{b}=\mathbf{0}$ and $\Theta^{(t)}\triangleq\|\mathbf{x}^*-\mathbf{x}^{(t-1)}\|_{\mathbf{Q}^{(t)}}^2-\|\mathbf{x}^*-\mathbf{x}^{(t)}\|_{\mu\mathbf{I}+\mathbf{Q}^{(t)}}^2$.

Recall that by part (2) of Lemma 3, we have

$$\left\langle\boldsymbol{\lambda}^{(t-1)},\mathbf{A}\mathbf{x}^{(t)}-\mathbf{b}\right\rangle=\frac{1}{2\rho^{(t)}}\left(\|\boldsymbol{\lambda}^{(t)}\|^2-\|\boldsymbol{\lambda}^{(t-1)}\|^2\right)-\frac{\rho^{(t)}}{2}\|\mathbf{A}\mathbf{x}^{(t)}-\mathbf{b}\|^2 \tag{33}$$

By the basic identity $\langle\mathbf{u},\mathbf{v}\rangle=\frac{1}{2}\|\mathbf{u}\|_2^2+\frac{1}{2}\|\mathbf{v}\|^2-\frac{1}{2}\|\mathbf{u}-\mathbf{v}\|^2$ for any vector $\mathbf{u},\mathbf{v}$, we have

$$\left\langle\mathbf{A}\mathbf{x}^{(t-1)}-\mathbf{b},\mathbf{A}\mathbf{x}^{(t)}-\mathbf{b}\right\rangle=\frac{1}{2}\|\mathbf{A}\mathbf{x}^{(t-1)}-\mathbf{b}\|^2+\frac{1}{2}\|\mathbf{A}\mathbf{x}^{(t)}-\mathbf{b}\|^2-\frac{1}{2}\|\mathbf{A}(\mathbf{x}^{(t)}-\mathbf{x}^{(t-1)})\|^2 \tag{34}$$

Substituting (33)-(34) into (32) and rearranging terms yields

$$\begin{aligned}f(\mathbf{x}^{(t)})\leq&f(\mathbf{x}^*)+\frac{1}{2}\Theta^{(t)}+\frac{1}{2\rho^{(t)}}\left(\|\boldsymbol{\lambda}^{(t-1)}\|^2-\|\boldsymbol{\lambda}^{(t)}\|^2\right)-\frac{\rho^{(t)}}{2}\|\mathbf{A}\mathbf{x}^{(t-1)}-\mathbf{b}\|^2\\&+\frac{\rho^{(t)}}{2}\|\mathbf{A}(\mathbf{x}^{(t)}-\mathbf{x}^{(t-1)})\|^2-\frac{1}{2}\|\mathbf{x}^{(t)}-\mathbf{x}^{(t-1)}\|_{\mathbf{Q}^{(t)}}^2\\\overset{(a)}{\leq}&f(\mathbf{x}^*)+\frac{1}{2}\Theta^{(t)}+\frac{1}{2\rho^{(t)}}\left(\|\boldsymbol{\lambda}^{(t-1)}\|^2-\|\boldsymbol{\lambda}^{(t)}\|^2\right)-\frac{1}{2}\|\mathbf{x}^{(t)}-\mathbf{x}^{(t-1)}\|_{\mathbf{Q}^{(t)}-\rho^{(t)}\mathbf{A}^\mathsf{T}\mathbf{A}}^2\\\overset{(b)}{\leq}&f(\mathbf{x}^*)+\frac{1}{2}\Theta^{(t)}+\frac{1}{2\rho^{(t)}}\left(\|\boldsymbol{\lambda}^{(t-1)}\|^2-\|\boldsymbol{\lambda}^{(t)}\|^2\right)\end{aligned}$$

where (a) follows by ignoring the negative term $-\frac{\rho^{(t)}}{2}\|\mathbf{A}\mathbf{x}^{(t-1)}-\mathbf{b}\|^2$ and noting that $\frac{\rho^{(t)}}{2}\|\mathbf{A}(\mathbf{x}^{(t)}-\mathbf{x}^{(t-1)})\|^2=\frac{1}{2}\|\mathbf{x}^{(t)}-\mathbf{x}^{(t-1)}\|^2_{\rho^{(t)}\mathbf{A}^\mathsf{T}\mathbf{A}}$; and (b) follows because $\mathbf{Q}^{(t)}\succeq\rho^{(t)}\mathbf{A}^\mathsf{T}\mathbf{A}$.

Multiplying $\rho^{(t)}$ on both sides and summing over $t\in\{1,\dots,T\}$ yields

$$\begin{aligned}\sum_{t=1}^{T}\rho^{(t)}f(\mathbf{x}^{(t)})\leq&\sum_{t=1}^{T}\rho^{(t)}f(\mathbf{x}^*)+\frac{1}{2}\sum_{t=1}^{T}\rho^{(t)}\Theta^{(t)}+\frac{1}{2}\sum_{t=1}^{T}\left(\|\boldsymbol{\lambda}^{(t-1)}\|^2-\|\boldsymbol{\lambda}^{(t)}\|^2\right)\\\overset{(a)}{=}&\sum_{t=1}^{T}\rho^{(t)}f(\mathbf{x}^*)+\frac{1}{2}\sum_{t=1}^{T}\rho^{(t)}\Theta^{(t)}-\frac{1}{2}\|\boldsymbol{\lambda}^{(T)}\|^2\end{aligned}$$

where (a) follows by simplifying the telescoping sums and recalling that $\boldsymbol{\lambda}^{(0)}=\mathbf{0}$. □

The following lemma provides a few practical sufficient conditions that ensure (30)

Lemma 10. *The condition* (30) *holds if* **any** *of the following three conditions holds*

1. $\mathbf{P}_i^{(t)}=\nu^{(t)}\mathbf{I}-\rho^{(t)}\mathbf{A}_i^\mathsf{T}\mathbf{A}_i$ *with* $\nu^{(t)}\geq\rho^{(t)}\|\mathbf{A}\|^2$.
2. $\mathbf{P}_i^{(t)}=\nu^{(t)}\mathbf{I}$ *with* $\nu^{(t)}\geq\rho^{(t)}\|\mathbf{A}\|^2$.
3. $\mathbf{P}_i^{(t)}=\nu_i^{(t)}\mathbf{I}$ *with* $\nu_i^{(t)}\geq\rho^{(t)}(N-1)\|\mathbf{A}_i\|_2^2$

Proof. Note that (30) holds trivially when the first or the second condition holds. To see (30) holds when $\mathbf{P}_i=\nu_i^{(t)}\mathbf{I}$ with $\nu_i^{(t)}\geq\rho^{(t)}(N-1)\|\mathbf{A}_i\|_2^2$, we note that for any $\mathbf{z}=[\mathbf{z}_1;\dots,\mathbf{z}_N]\in\mathbb{R}^{\sum_{i=1}^N d_i}$,

$$\begin{aligned}\|\mathbf{z}\|^2_{\rho^{(t)}\mathbf{A}^\mathsf{T}\mathbf{A}-\mathbf{Q}^{(t)}}=&\rho^{(t)}\|\sum_{i=1}^{N}\mathbf{A}_i\mathbf{z}_i\|^2-\sum_{i=1}^{N}\|\mathbf{z}_i\|^2_{\mathbf{P}_i^{(t)}+\rho^{(t)}\mathbf{A}_i^\mathsf{T}\mathbf{A}_i}\\\overset{(a)}{\leq}&\rho^{(t)}N\sum_{i=1}^{N}\|\mathbf{z}_i\|^2_{\mathbf{A}_i^\mathsf{T}\mathbf{A}_i}-\sum_{i=1}^{N}\|\mathbf{z}_i\|^2_{\mathbf{P}_i^{(t)}+\rho^{(t)}\mathbf{A}_i^\mathsf{T}\mathbf{A}_i}\\=&-\sum_{i=1}^{N}\|\mathbf{z}_i\|^2_{\mathbf{P}_i^{(t)}-\rho^{(t)}(N-1)\mathbf{A}_i^\mathsf{T}\mathbf{A}_i}\end{aligned}$$

where (a) follows from the Cauchy-Schwarz inequality. □

Remark 7. *Note that the sufficient conditions developed in Lemma 10 are similar to the conditions from [7], under which [7] shows Algorithm 3 with constant ρ and $\mathbf{P}_i$ can ensure $\mathbf{x}^{(t)}$ eventually converge to an optimal solution $\mathbf{x}^*$ and has an $o(1/t)$ non-ergodic convergence rate in the sense $\|\mathbf{x}^{t+1}-\mathbf{x}^{(t)}\|^2=o(1/t)$. However, work [7] does not establish the convergence rate of objective violations and feasibility violations shown in our Theorem 3. Furthermore, the fast $O(1/T^2)$ convergence for strongly convex case is not considered in [7].*

Now we are ready to prove both parts of the theorem.

1. **Proof of case $\mu=0$:** Note that $\rho^{(t)}=\rho$ and $\mathbf{P}_i^{(t)}=\mathbf{P}_i$ are chosen to satisfy (30) by Lemma 10. By Lemma 9 (with $\mu=0$), we have

$$\begin{aligned}\rho\sum_{t=1}^{T}f(\mathbf{x}^{(t)})\leq&\rho Tf(\mathbf{x}^*)+\frac{1}{2}\sum_{t=1}^{T}\rho\left(\|\mathbf{x}^*-\mathbf{x}^{(t-1)}\|^2_{\mathbf{Q}}-\|\mathbf{x}^*-\mathbf{x}^{(t)}\|^2_{\mathbf{Q}}\right)-\frac{1}{2}\|\boldsymbol{\lambda}^{(T)}\|^2\\\leq&\rho Tf(\mathbf{x}^*)+\frac{\rho}{2}\|\mathbf{x}^*-\mathbf{x}^{(0)}\|^2_{\mathbf{Q}}-\frac{1}{2}\|\boldsymbol{\lambda}^{(T)}\|^2\end{aligned}\tag{35}$$

Ignoring the (negative term) $-\frac{1}{2}\|\boldsymbol{\lambda}^{(T)}\|^2$, dividing both sides by ρT, applying Jensen's inequality yields

$$f(\overline{\mathbf{x}}^T)\leq f(\mathbf{x}^*)+\frac{1}{2T}\|\mathbf{x}^*-\mathbf{x}^{(0)}\|^2_{\mathbf{Q}},$$

which is (23) of our theorem.
By Lemma 4 (with $\rho^{(t)} = \rho$), we have

$$\rho \sum_{t=1}^{T} f(\mathbf{x}^{(t)}) \geq \rho T f(\mathbf{x}^*) - \|\boldsymbol{\lambda}^*\| \|\boldsymbol{\lambda}^{(T)}\|$$

Combining this with (35) and cancelling the common term yields

$$\|\boldsymbol{\lambda}^{(T)}\|^2 - 2\|\boldsymbol{\lambda}^*\| \|\boldsymbol{\lambda}^{(T)}\| \leq \rho \|\mathbf{x}^* - \mathbf{x}^{(0)}\|_{\mathbf{Q}}^2$$

This quadratic inequality can be further be rewritten as

$$\left(\|\boldsymbol{\lambda}^{(T)}\| - \|\boldsymbol{\lambda}^*\|\right)^2 \leq \|\boldsymbol{\lambda}^*\|^2 + \rho \|\mathbf{x}^* - \mathbf{x}^{(0)}\|_{\mathbf{Q}}^2$$

Thus, we have

$$\begin{aligned} \|\boldsymbol{\lambda}^{(T)}\| \leq & \|\boldsymbol{\lambda}^*\| + \sqrt{\|\boldsymbol{\lambda}^*\|^2 + \rho \|\mathbf{x}^* - \mathbf{x}^{(0)}\|_{\mathbf{Q}}^2} \\ \overset{(a)}{\leq} & 2\|\boldsymbol{\lambda}^*\| + \sqrt{\rho} \|\mathbf{x}^* - \mathbf{x}^{(0)}\|_{\mathbf{Q}} \end{aligned} \tag{36}$$

where (a) follows from the basic inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for all $a, b \geq 0$.
By part (1) of Lemma 3 (with $\rho^{(t)} = \rho$), we have

$$\rho \sum_{t=1}^{T} \left(\mathbf{A}\mathbf{x}^{(t)} - \mathbf{b}\right) = \boldsymbol{\lambda}^{(T)}$$

Dividing both sides by ρT and taking the vector l_2 norm on both sides yields

$$\begin{aligned} \|\mathbf{A}\overline{\mathbf{x}}^{(T)} - \mathbf{b}\| \leq & \frac{1}{\rho T} \|\boldsymbol{\lambda}^{(T)}\| \\ \overset{(a)}{\leq} & \frac{1}{T} \frac{2\|\boldsymbol{\lambda}^*\|}{\rho} + \frac{1}{T} \frac{\|\mathbf{x}^* - \mathbf{x}^{(0)}\|_{\mathbf{Q}}}{\sqrt{\rho}} \end{aligned}$$

where (a) follows from (36). Note this is (24) of our theorem.

2. **Proof of case** $\mu > 0$: Note that $\rho^{(t)} = \rho t$ and $\mathbf{P}_i^{(t)} = t\rho \|\mathbf{A}\|^2 - t\rho \mathbf{A}_i^{\mathsf{T}} \mathbf{A}_i$ are chosen to satisfy (30) by Lemma 10. By Lemma 9, we have

$$\begin{aligned} & \sum_{t=1}^{T} \rho^{(t)} f(\mathbf{x}^{(t)}) \\ \leq & \sum_{t=1}^{T} \rho^{(t)} f(\mathbf{x}^*) + \frac{1}{2} \sum_{t=1}^{T} \rho^{(t)} \left(\|\mathbf{x}^* - \mathbf{x}^{(t-1)}\|_{\mathbf{Q}^{(t)}}^2 - \|\mathbf{x}^* - \mathbf{x}^{(t)}\|_{\mu\mathbf{I}+\mathbf{Q}^{(t)}}^2 \right) - \frac{1}{2} \|\boldsymbol{\lambda}^{(T)}\|^2 \end{aligned} \tag{37}$$

where $\mathbf{Q}^{(t)} = t\rho \|\mathbf{A}\|^2 \mathbf{I}$.
Note that

$$\begin{aligned} & \sum_{t=1}^{T} \rho^{(t)} \left(\|\mathbf{x}^* - \mathbf{x}^{(t-1)}\|_{\mathbf{Q}^{(t)}}^2 - \|\mathbf{x}^* - \mathbf{x}^{(t)}\|_{\mu\mathbf{I}+\mathbf{Q}^{(t)}}^2 \right) \\ = & \rho \|\mathbf{x}^* - \mathbf{x}^{(0)}\|_{\rho\|\mathbf{A}\|^2\mathbf{I}}^2 - \sum_{t=1}^{T-1} \left(\rho^{(t)} \|\mathbf{x}^* - \mathbf{x}^{(t)}\|_{\mu\mathbf{I}+\mathbf{Q}^{(t)}}^2 - \rho^{(t+1)} \|\mathbf{x}^* - \mathbf{x}^{(t)}\|_{\mathbf{Q}^{(t+1)}}^2 \right) \\ & - \rho^T \|\mathbf{x}^* - \mathbf{x}^{(t)}\|_{\mu\mathbf{I}+\mathbf{Q}^{(T)}}^2 \\ = & \rho^2 \|\mathbf{A}\|^2 \|\mathbf{x}^* - \mathbf{x}^{(0)}\|^2 - \sum_{t=1}^{T-1} \|\mathbf{x}^* - \mathbf{x}^{(t)}\|_{\rho^{(t)}(\mu\mathbf{I}+\mathbf{Q}^{(t)}) - \rho^{(t+1)}\mathbf{Q}^{(t+1)}}^2 - \rho^T \|\mathbf{x}^* - \mathbf{x}^{(t)}\|_{\mu\mathbf{I}+\mathbf{Q}^{(T)}}^2 \end{aligned} \tag{38}$$

$$\overset{(a)}{\leq} \rho^2 \|\mathbf{A}\|^2 \|\mathbf{x}^* - \mathbf{x}^{(0)}\|^2 \tag{39}$$

where (a) follows by ignoring the negative term $-\rho^T\|\mathbf{x}^*-\mathbf{x}^{(t)}\|^2_{\mu\mathbf{I}+\mathbf{Q}^{(T)}}$ and noting that $\rho^{(t)}(\mu\mathbf{I}+\mathbf{Q}^{(t)})-\rho^{(t+1)}\mathbf{Q}^{(t+1)}=(\rho t\mu+\rho^2t^2\|\mathbf{A}\|^2-\rho^2(t+1)^2\|\mathbf{A}\|^2)\mathbf{I}=\rho(t\mu-\rho(2t+1)\|\mathbf{A}\|^2)\mathbf{I}\succeq\rho\mu(t-\frac{2t+1}{3})\mathbf{I}\succeq 0$ where the first $\succeq$ follows because $\rho\leq\frac{\mu}{3\|\mathbf{A}\|^2}$ by our algorithm parameter selection and the second $\succeq$ follows because $t\geq 1$.

Substituting (39) into (37) yields

$$\sum_{t=1}^{T}\rho^{(t)}f(\mathbf{x}^{(t)})\leq\sum_{t=1}^{T}\rho^{(t)}f(\mathbf{x}^*)+\frac{\rho^2}{2}\|\mathbf{A}\|^2\|\mathbf{x}^*-\mathbf{x}^{(0)}\|^2-\frac{1}{2}\|\boldsymbol{\lambda}^{(T)}\|^2 \tag{40}$$

Ignoring the (negative term) $-\frac{1}{2}\|\boldsymbol{\lambda}^{(T)}\|_2^2$, dividing both sides by $\sum_{t=1}^{T}\rho^{(t)}$, applying Jensen's inequality yields

$$\begin{aligned}f(\overline{\mathbf{x}}^T)\leq&f(\mathbf{x}^*)+\frac{\rho^2}{2\sum_{t=1}^{T}\rho^{(t)}}\|\mathbf{A}\|^2\|\mathbf{x}^*-\mathbf{x}^{(0)}\|^2\\ \overset{(a)}{=}&f(\mathbf{x}^*)+\frac{\rho}{T(T+1)}\|\mathbf{A}\|^2\|\mathbf{x}^*-\mathbf{x}^{(0)}\|^2\end{aligned}$$

where (a) follows because $\sum_{t=1}^{T}\rho^{(t)}=\rho\sum_{t=1}^{T}t=\rho\frac{T(T+1)}{2}$. Note this is (25) of our theorem.

By Lemma 4, we have

$$\sum_{t=1}^{T}\rho^{(t)}f(\mathbf{x}^{(t)})\geq\sum_{t=1}^{T}\rho^{(t)}f(\mathbf{x}^*)-\|\boldsymbol{\lambda}^*\|\|\boldsymbol{\lambda}^{(T)}\|$$

Combining this with (40) and cancelling the common term yields

$$\|\boldsymbol{\lambda}^{(T)}\|^2-2\|\boldsymbol{\lambda}^*\|\|\boldsymbol{\lambda}^{(T)}\|\leq\rho^2\|\mathbf{A}\|^2\|\mathbf{x}^*-\mathbf{x}^{(0)}\|^2$$

This quadratic inequality can be further be rewritten as

$$\left(\|\boldsymbol{\lambda}^{(T)}\|-\|\boldsymbol{\lambda}^*\|\right)^2\leq\|\boldsymbol{\lambda}^*\|^2+\rho^2\|\mathbf{A}\|^2\|\mathbf{x}^*-\mathbf{x}^{(0)}\|^2$$

Thus, we have

$$\begin{aligned}\|\boldsymbol{\lambda}^{(T)}\|\leq&\|\boldsymbol{\lambda}^*\|+\sqrt{\|\boldsymbol{\lambda}^*\|^2+\rho^2\|\mathbf{A}\|^2\|\mathbf{x}^*-\mathbf{x}^{(0)}\|^2}\\ \overset{(a)}{\leq}&2\|\boldsymbol{\lambda}^*\|+\rho\|\mathbf{A}\|\|\mathbf{x}^*-\mathbf{x}^{(0)}\|\end{aligned} \tag{41}$$

where (a) follows from the basic inequality $\sqrt{a+b}\leq\sqrt{a}+\sqrt{b}$ for all $a,b\geq 0$.

By part (1) of Lemma 3, we have

$$\sum_{t=1}^{T}\rho^{(t)}\left(\mathbf{A}\mathbf{x}^{(t)}-\mathbf{b}\right)=\boldsymbol{\lambda}^{(T)}$$

Dividing both sides by $\sum_{t=1}^{T}\rho^{(t)}$ and taking the vector l_2 norm on both sides yields

$$\begin{aligned}\|\mathbf{A}\overline{\mathbf{x}}^{(T)}-\mathbf{b}\|=&\frac{1}{\sum_{t=1}^{T}\rho^{(t)}}\|\boldsymbol{\lambda}^{(T)}\|\\ \overset{(a)}{\leq}&\frac{4\|\boldsymbol{\lambda}^*\|}{\rho T(T+1)}+\frac{2\|\mathbf{A}\|\|\mathbf{x}^*-\mathbf{x}^{(0)}\|}{T(T+1)}\end{aligned}$$

where (a) follows from (41) and the fact that $\sum_{t=1}^{T}\rho^{(t)}=\rho\sum_{t=1}^{T}t=\rho\frac{T(T+1)}{2}$. Note this is (26) of our theorem.

6.6 Proof of Lemma 2

Fix $\mathbf{z} \in \mathcal{Z}$. At each iteration k, the projected gradient update (6) in Algorithm 2 can be rewritten as

$$\mathbf{z}^{(k)} = \operatorname*{argmin}_{\mathbf{z}\in\mathcal{Z}} \{\langle \boldsymbol{\zeta}^{(k)}, \mathbf{z} - \mathbf{z}^{(k-1)}\rangle + \frac{1}{2\gamma^{(k)}}\|\mathbf{z} - \mathbf{z}^{(k-1)}\|^2\}.$$

Since the objective function is $\frac{1}{\gamma^{(k)}}$-convex, by Lemma 5, we have

$$\begin{aligned}\langle \boldsymbol{\zeta}^{(k)}, \mathbf{z}^{(k)} - \mathbf{z}^{(k-1)}\rangle + \frac{1}{2\gamma^{(k)}}\|\mathbf{z}^{(k)} - \mathbf{z}^{(k-1)}\|^2 \leq &\langle \boldsymbol{\zeta}^{(k)}, \mathbf{z} - \mathbf{z}^{(k-1)}\rangle + \frac{1}{2\gamma^{(k)}}\|\mathbf{z} - \mathbf{z}^{(k-1)}\|^2 \\ &- \frac{1}{2\gamma^{(k)}}\|\mathbf{z} - \mathbf{z}^{(k)}\|^2\end{aligned}$$

Adding $\phi(\mathbf{z}^{(k-1)}) + \langle \nabla\phi(\mathbf{z}^{(k-1)}) - \boldsymbol{\zeta}^{(k)}, \mathbf{z}^{(k)} - \mathbf{z}^{(k-1)}\rangle + \frac{L}{2}\|\mathbf{z}^{(k)} - \mathbf{z}^{(k-1)}\|^2$ on both sides and rearranging terms yields

$$\begin{aligned}&\phi(\mathbf{z}^{(k-1)}) + \langle \nabla\phi(\mathbf{z}^{(k-1)}), \mathbf{z}^{(k)} - \mathbf{z}^{(k-1)}\rangle + \frac{L}{2}\|\mathbf{z}^{(k)} - \mathbf{z}^{(k-1)}\|^2 \\ \leq &\phi(\mathbf{z}^{(k-1)}) + \langle \boldsymbol{\zeta}^{(k)}, \mathbf{z} - \mathbf{z}^{(k-1)}\rangle + \frac{1}{2\gamma^{(k)}}\|\mathbf{z} - \mathbf{z}^{(k-1)}\|^2 - \frac{1}{2\gamma^{(k)}}\|\mathbf{z} - \mathbf{z}^{(k)}\|^2 \\ &- \frac{1}{2}(\frac{1}{\gamma^{(k)}} - L)\|\mathbf{z}^{(k)} - \mathbf{z}^{(k-1)}\|^2 + \langle \nabla\phi(\mathbf{z}^{(k-1)}) - \boldsymbol{\zeta}^{(k)}, \mathbf{z}^{(k)} - \mathbf{z}^{(k-1)}\rangle \end{aligned} \tag{42}$$

Since $\phi(\cdot)$ is L-smooth, by the descent lemma, e.g., Proposition A.24 in [2], we have

$$\phi(\mathbf{z}^{(k)}) \leq \phi(\mathbf{z}^{(k-1)}) + \langle \nabla\phi(\mathbf{z}^{(k-1)}), \mathbf{z}^{(k)} - \mathbf{z}^{(k-1)}\rangle + \frac{L}{2}\|\mathbf{z}^{(k)} - \mathbf{z}^{(k-1)}\|^2 \tag{43}$$

By Young's inequality, for any $\eta^{(k)} > 0$, we have

$$\langle \nabla\phi(\mathbf{z}^{(k-1)}) - \boldsymbol{\zeta}^{(k)}, \mathbf{z}^{(k)} - \mathbf{z}^{(k-1)}\rangle \leq \frac{1}{2\eta^{(k)}}\|\nabla\phi(\mathbf{x}^{k-1}) - \boldsymbol{\zeta}^{(k)}\|^2 + \frac{\eta^{(k)}}{2}\|\mathbf{z}^{(k)} - \mathbf{z}^{(k-1)}\|^2 \tag{44}$$

Substituting (43) and (44) into (42) yields

$$\begin{aligned}\phi(\mathbf{z}^{(k)}) \leq &\phi(\mathbf{z}^{(k-1)}) + \langle \boldsymbol{\zeta}^{(k)}, \mathbf{z} - \mathbf{z}^{(k-1)}\rangle + \frac{1}{2\gamma^{(k)}}\|\mathbf{z} - \mathbf{z}^{(k-1)}\|^2 - \frac{1}{2\gamma^{(k)}}\|\mathbf{z} - \mathbf{z}^{(k)}\|^2 \\ &- \frac{1}{2}(\frac{1}{\gamma^{(k)}} - L - \eta^{(k)})\|\mathbf{z}^{(k)} - \mathbf{z}^{(k-1)}\|^2 + \frac{1}{2\eta^{(k)}}\|\nabla\phi(\mathbf{x}^{k-1}) - \boldsymbol{\zeta}^{(k)}\|^2 \end{aligned} \tag{45}$$

For any fixed $\mathbf{z}$, since $\boldsymbol{\zeta}^{(k)}$ is an unbiased i.i.d. stochastic gradient and $\mathbf{z}^{(t-1)}$ is determined by $\boldsymbol{\zeta}^{(0)}, \ldots, \boldsymbol{\zeta}^{(k-1)}$, we have

$$\mathbb{E}[\langle \boldsymbol{\zeta}^{(k)}, \mathbf{z} - \mathbf{z}^{(k-1)}\rangle] = \langle \nabla\phi(\mathbf{z}^{(k-1)}), \mathbf{z} - \mathbf{z}^{(k-1)}\rangle \tag{46}$$

By the bounded variance assumption, we have

$$\mathbb{E}[\|\nabla\phi(\mathbf{x}^{k-1}) - \boldsymbol{\zeta}^{(k)}\|^2] \leq \sigma^2 \tag{47}$$

By the μ-strong convexity of $\phi(\cdot)$, we have

$$\phi(\mathbf{z}^{(k-1)}) + \langle \nabla\phi(\mathbf{z}^{(k-1)}), \mathbf{z} - \mathbf{z}^{(k-1)}\rangle \leq \phi(\mathbf{z}) - \frac{\mu}{2}\|\mathbf{z} - \mathbf{z}^{(k-1)}\|^2 \tag{48}$$

Taking expectations on both sides of (45) and substituting (46)-(48) into it yields

$$\begin{aligned}\mathbb{E}[\phi(\mathbf{z}^{(k)})] \leq &\phi(\mathbf{z}) + \frac{1}{2}(\frac{1}{\gamma^{(k)}} - \mu)\mathbb{E}[\|\mathbf{z} - \mathbf{z}^{(k-1)}\|^2] - \frac{1}{2}\frac{1}{\gamma^{(k)}}\mathbb{E}[\|\mathbf{z} - \mathbf{z}^{(k)}\|^2] \\ &- \frac{1}{2}(\frac{1}{\gamma^{(k)}} - L - \eta^{(k)})\mathbb{E}[\|\mathbf{z}^{(k)} - \mathbf{z}^{(k-1)}\|^2] + \frac{1}{2\eta^{(k)}}\sigma^2 \end{aligned} \tag{49}$$

Note that if we take $\gamma^{(k)} = \frac{2}{\mu(k+k_0)}$, $\eta^{(k)} = \frac{\mu}{2}k$, then $\frac{1}{\gamma^{(k)}} - L - \eta^{(k)} \geq 0$ since $k_0 \geq 2\kappa = 2\frac{L}{\mu}$. Thus, under the current choice of $\gamma^{(k)}$ and $\eta^{(k)}$, (49) implies that

$$\mathbb{E}[\phi(\mathbf{z}^{(k)})] \leq \phi(\mathbf{z}) + \frac{\mu}{4}(k+k_0-2)\mathbb{E}[\|\mathbf{z}-\mathbf{z}^{(k-1)}\|^2] - \frac{\mu}{4}(k+k_0)\mathbb{E}[\|\mathbf{z}-\mathbf{z}^{(k)}\|^2] + \frac{1}{k\mu}\sigma^2$$

Multiplying both sides by $k+k_0-1$ yields

$$\begin{aligned}(k+k_0-1)\mathbb{E}[\phi(\mathbf{z}^{(k)})] \leq & (k+k_0-1)\phi(\mathbf{z}) + \frac{\mu}{4}(k+k_0-2)(k+k_0-1)\mathbb{E}[\|\mathbf{z}-\mathbf{z}^{(k-1)}\|^2] \\ & - \frac{\mu}{4}(k+k_0-1)(k+k_0)\mathbb{E}[\|\mathbf{z}-\mathbf{z}^{(k)}\|^2] + \frac{k+k_0-1}{k\mu}\sigma^2 \\ \leq & (k+k_0-1)\phi(\mathbf{z}) + \frac{\mu}{4}(k+k_0-2)(k+k_0-1)\mathbb{E}[\|\mathbf{z}-\mathbf{z}^{(k-1)}\|^2] \\ & - \frac{\mu}{4}(k+k_0-1)(k+k_0)\mathbb{E}[\|\mathbf{z}-\mathbf{z}^{(k)}\|^2] + \frac{k_0}{\mu}\sigma^2\end{aligned}$$

Summing over $k \in \{1, 2, \ldots, K\}$ and dividing both sides by $\sum_{k=1}^{K}(k+k_0-1)$ yields

$$\begin{aligned}&\mathbb{E}[\frac{1}{\sum_{k=1}^{K}(k+k_0-1)}\sum_{k=1}^{K}(k+k_0-1)\phi(\mathbf{z}^{(k)})] \\ \leq & \phi(\mathbf{z}) + \frac{\mu(k_0^2-k_0)}{2K(K+2k_0-1)}\mathbb{E}[\|\mathbf{z}-\mathbf{z}^{(0)}\|^2] - \frac{\mu(k_0^2-k_0)}{2K(K+2k_0-1)}\mathbb{E}[\|\mathbf{z}-\mathbf{z}^{(K)}\|^2] - \frac{\mu}{2}\mathbb{E}[\|\mathbf{z}-\mathbf{z}^{(K)}\|^2] \\ & + \frac{2k_0\sigma^2}{(K+2k_0-1)\mu}\end{aligned}$$

Define $\widehat{\mathbf{z}} \triangleq \frac{1}{\sum_{k=1}^{K}(k+k_0-1)}(k+k_0-1)\mathbf{z}^{(k)}$. By Jensen's inequality, we have

$$\begin{aligned}\mathbb{E}[\phi(\widehat{\mathbf{z}})] \leq & \phi(\mathbf{z}) + \frac{\mu(k_0^2-k_0)}{2K(K+2k_0-1)}\mathbb{E}[\|\mathbf{z}-\mathbf{z}^{(0)}\|^2] - \frac{\mu(k_0^2-k_0)}{2K(K+2k_0-1)}\mathbb{E}[\|\mathbf{z}-\mathbf{z}^{(K)}\|^2] \\ & - \frac{\mu}{2}\mathbb{E}[\|\mathbf{z}-\mathbf{z}^{(K)}\|^2] + \frac{2k_0\sigma^2}{(K+2k_0-1)\mu}\end{aligned}$$

6.7 Proof of Theorem 1

For convenience of our presentation, we extract the assumption in Theorem 1 and call it Assumption 2:

Assumption 2. *Convex program 1 satisfies the following:*

1. *The constraint set $\mathcal{X}$ is bounded, i.e., there exists constant $R > 0$ such that $\|\mathbf{x}\| \leq R, \forall \mathbf{x} \in \mathcal{X}$.*
2. *The function $f(\mathbf{x})$ has unbiased stochastic subgradients with a bounded second order moment, i.e., there exists constant $D > 0$ such that $\mathbb{E}_\xi[\|\mathbf{G}(\mathbf{x};\xi)\|^2] \leq D^2, \forall \mathbf{x} \in \mathcal{X}$.*

Lemma 11. *Consider convex program* (1) *under Assumption 2. Let $\mathbf{x}^*$ be any optimal solution. If $\nu^{(t)} > 0$ and $\rho^{(t)} > 0$ in Algorithm 1 are chosen to satisfy*

$$\nu^{(t)} \geq \rho^{(t)}\|\mathbf{A}\|^2, \forall t,$$

and the sub-procedure STO-LOCAL (Algorithm 2) uses $\widehat{\mathbf{z}}$ defined in Lemma 1 as the output then, for all $T \geq 1$, Algorithm 1 ensures

$$\sum_{t=1}^{T}\mathbb{E}[\rho^{(t)}f(\mathbf{x}^{(t)})] \leq \sum_{t=1}^{T}\mathbb{E}[\rho^{(t)}f(\mathbf{x}^*)] + \frac{1}{2}\sum_{t=1}^{T}\mathbb{E}[\rho^{(t)}\Lambda^{(t)}] - \frac{1}{2}\mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|^2] + \sum_{t=1}^{T}\frac{2\rho^{(t)}(B^{(t)})^2}{\nu^{(t)}(K^{(t)}+1)}$$

with

$$\Gamma^{(t)} \triangleq \nu^{(t)}\|\mathbf{x}^* - \mathbf{y}^{(t-1)}\|^2 - \nu^{(t)}\|\mathbf{x}^* - \mathbf{y}^{(t)}\|^2, \tag{50}$$

and

$$(B^{(t)})^2 \triangleq 2\|\mathbf{A}\|^2\mathbb{E}[\|\boldsymbol{\lambda}^{(t-1)}\|^2] + 6D^2 + 6(\rho^{(t)})^2\|\mathbf{A}\|^2(\|\mathbf{A}\|R + \|\mathbf{b}\|)^2 + 24(\nu^{(t)})^2R^2 \tag{51}$$

where D and R are constants defined in Assumption 2.

Proof. Fix $t \in \{1, 2, \ldots, T\}$. Define $\phi^{(t)}(\mathbf{x}) = \sum_{i=1}^{N} \phi_i^{(t)}(\mathbf{x}_i)$. Since $\phi^{(t)}(\mathbf{x})$ is separable with respect to each $\mathbf{x}_i$, the fact that Algorithm 1 updates each $\mathbf{x}_i$ and $\mathbf{y}_i$ locally and in parallel by calling sub-procedure $(\mathbf{x}_i^{(t)}, \mathbf{y}_i^{(t)}) = \text{STO-LOCAL}(\phi_i^{(t)}(\cdot), \mathcal{X}_i, \mathbf{y}_i^{(t-1)}, K^{(t)})$ can be interpreted as all N nodes jointly update $\mathbf{x}$ and $\mathbf{y}$ via calling sub-procedure $(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) = \text{STO-LOCAL}(\phi^{(t)}(\cdot), \mathcal{X}, \mathbf{y}^{(t-1)}, K^{(t)})$. (Note that the synchronization of parallel sub-procedures $\text{STO-LOCAL}(\phi_i^{(t)}(\cdot), \mathcal{X}_i, \mathbf{y}_i^{(t-1)}, K^{(t)})$ is not needed since these sub-procedures are fully decoupled. We just need to aggregate the variables with the same index together and write it into the above compact form.)

For each $i \in \{1, 2, \ldots, N\}$, the unbiased stochastic subgradient used in each iteration $k \in \{1, 2, \ldots, K^{(t)}\}$ of Algorithm 2 is given by

$$\boldsymbol{\zeta}_i^{(k)} = \mathbf{G}_i^{(k)} + \mathbf{A}_i^{\mathsf{T}}(\rho^{(t)}(\mathbf{A}\mathbf{y}^{(t-1)} - \mathbf{b}) + \boldsymbol{\lambda}^{(t-1)}) + \nu^{(t)}(\mathbf{z}_i^{(k)} - \mathbf{y}_i^{(t-1)})$$

where $\mathbf{G}_i^{(k)}$ is an unbiased stochastic subgradient for $f_i(\mathbf{x}_i)$ at point $\mathbf{x}_i = \mathbf{z}_i^{(k)}$.

Define $\boldsymbol{\zeta}^{(k)} = [\boldsymbol{\zeta}_1^{(k)}; \ldots; \boldsymbol{\zeta}_N^{(k)}] = \mathbf{G}^{(k)} + \mathbf{A}^{\mathsf{T}}(\rho^{(t)}(\mathbf{A}\mathbf{y}^{(t-1)} - \mathbf{b}) + \boldsymbol{\lambda}^{(t-1)}) + \nu^{(t)}(\mathbf{z}^{(k)} - \mathbf{y}^{(t-1)}), \forall k \in \{1, 2, \ldots, K^{(t)}\}$. Then, $\boldsymbol{\zeta}^{(k)}$ is the unbiased stochastic subgradient used in each iteration of the joint sub-procedure $(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) = \text{STO-LOCAL}(\phi^{(t)}(\cdot), \mathcal{X}, \mathbf{y}^{(t-1)}, K^{(t)})$.

Note that

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{\zeta}^{(k)}\|^2] &\overset{(a)}{\leq} 2\mathbb{E}[\|\mathbf{A}^{\mathsf{T}}\boldsymbol{\lambda}^{(t-1)}\|^2] + 6(\mathbb{E}[\|\mathbf{G}^{(k)}\|^2] + 6\mathbb{E}[\|\rho^{(t)}\mathbf{A}^{\mathsf{T}}(\mathbf{A}\mathbf{y}^{(t-1)} - \mathbf{b})\|^2] \\
&\quad + 6\mathbb{E}[\|\nu^{(t)}(\mathbf{z}^{(k)} - \mathbf{y}^{(t-1)})\|^2]) \\
&\overset{(b)}{\leq} 2\|\mathbf{A}\|^2\mathbb{E}[\|\boldsymbol{\lambda}^{(t-1)}\|^2] + 6D^2 + 6(\rho^{(t)})^2\|\mathbf{A}\|^2(\|\mathbf{A}\|R + \|\mathbf{b}\|)^2 + 24(\nu^{(t)})^2R^2 \\
&\overset{(c)}{=} (B^{(t)})^2
\end{aligned}$$

where (a) follows from the basic inequality $\|\mathbf{v}_1 + \mathbf{v}_2 + \mathbf{v}_3 + \mathbf{v}_4\|^2 \leq 2\|\mathbf{v}_1\|^2 + 6\|\mathbf{v}_2\|^2 + 6\|\mathbf{v}_3\|^2 + 6\|\mathbf{v}_4\|^2$, which can be easily shown by noting that $\|\mathbf{v}_1 + \mathbf{v}_2 + \mathbf{v}_3 + \mathbf{v}_4\|^2 \leq 2\|\mathbf{v}_1\|^2 + 2\|\mathbf{v}_2 + \mathbf{v}_3 + \mathbf{v}_4\|^2 \leq 2\|\mathbf{v}_1\|^2 + 2 \cdot (3\|\mathbf{v}_2\|^2 + 3\|\mathbf{v}_3\|^2 + 3\|\mathbf{v}_4\|^2)$; (b) follows from Assumption 2 and basic matrix norm inequalities; and (c) follows from the definition of $B^{(t)}$ in (51).

Since $\phi^{(t)}(\cdot)$ is $\nu^{(t)}$-convex (with $\nu^{(t)} > 0$), by Lemma 1, we have

$$\mathbb{E}[\phi^{(t)}(\mathbf{x}^{(t)})] \leq \mathbb{E}[\phi^{(t)}(\mathbf{x}^*)] - \frac{\nu^{(t)}}{2}\mathbb{E}[\|\mathbf{y}^{(t)} - \mathbf{x}^*\|^2] + \frac{2(B^{(t)})^2}{\nu^{(t)}(K^{(t)} + 1)}.$$

Substituting the expression of $\phi_i^{(t)}(\cdot)$ (defined in (2)) into the above equation yields

$$\begin{aligned}
&\mathbb{E}[f(\mathbf{x}^{(t)})] + \rho^{(t)}\mathbb{E}[\langle \mathbf{A}\mathbf{y}^{(t-1)} - \mathbf{b}, \mathbf{A}\mathbf{x}^{(t)} - \mathbf{b}\rangle] + \mathbb{E}[\langle \boldsymbol{\lambda}^{(t-1)}, \mathbf{A}\mathbf{x}^{(t)} - \mathbf{b}\rangle] \\
&+ \frac{\nu^{(t)}}{2}\mathbb{E}[\|\mathbf{x}^{(t)} - \mathbf{y}^{(t-1)}\|^2] \\
\leq &\mathbb{E}[f(\mathbf{x}^*)] + \rho^{(t)}\mathbb{E}[\langle \mathbf{A}\mathbf{y}^{(t-1)} - \mathbf{b}, \mathbf{A}\mathbf{x}^* - \mathbf{b}\rangle] + \mathbb{E}[\langle \boldsymbol{\lambda}^{(t-1)}, \mathbf{A}\mathbf{x}^* - \mathbf{b}\rangle] + \frac{\nu^{(t)}}{2}\mathbb{E}[\|\mathbf{x}^* - \mathbf{y}^{(t-1)}\|^2] \\
&- \frac{\nu^{(t)}}{2}\mathbb{E}[\|\mathbf{x}^* - \mathbf{y}^{(t)}\|^2] + \frac{2(B^{(t)})^2}{\nu^{(t)}(K^{(t)} + 1)} \\
\overset{(a)}{=} &\mathbb{E}[f(\mathbf{x}^*)] + \frac{1}{2}\mathbb{E}[\Gamma^{(t)}] + \frac{2(B^{(t)})^2}{\nu^{(t)}(K^{(t)} + 1)}
\end{aligned} \tag{52}$$

where (a) follows because $\mathbf{A}\mathbf{x}^* - \mathbf{b} = \mathbf{0}$ and $\Gamma^{(t)} = \nu^{(t)}\|\mathbf{x}^* - \mathbf{y}^{(t-1)}\|^2 - \nu^{(t)}\|\mathbf{x}^* - \mathbf{y}^{(t)}\|^2$.

Recall that by part (2) of Lemma 3, we have

$$\left\langle \boldsymbol{\lambda}^{(t-1)}, \mathbf{A}\mathbf{x}^{(t)} - \mathbf{b} \right\rangle = \frac{1}{2\rho^{(t)}}\left(\|\boldsymbol{\lambda}^{(t)}\|^2 - \|\boldsymbol{\lambda}^{(t-1)}\|^2\right) - \frac{\rho^{(t)}}{2}\|\mathbf{A}\mathbf{x}^{(t)} - \mathbf{b}\|^2 \tag{53}$$

By the basic identity $\langle \mathbf{u}, \mathbf{v} \rangle = \frac{1}{2}\|\mathbf{u}\|_2^2 + \frac{1}{2}\|\mathbf{v}\|^2 - \frac{1}{2}\|\mathbf{u}-\mathbf{v}\|^2$ for any vector $\mathbf{u}, \mathbf{v}$, we have

$$\left\langle \mathbf{A}\mathbf{y}^{(t-1)} - \mathbf{b}, \mathbf{A}\mathbf{x}^{(t)} - \mathbf{b} \right\rangle = \frac{1}{2}\|\mathbf{A}\mathbf{y}^{(t-1)} - \mathbf{b}\|^2 + \frac{1}{2}\|\mathbf{A}\mathbf{x}^{(t)} - \mathbf{b}\|^2 - \frac{1}{2}\|\mathbf{A}(\mathbf{x}^{(t)} - \mathbf{y}^{(t-1)})\|^2 \tag{54}$$

Substituting (53) and (54) into (52) and rearranging terms yields

$$\begin{aligned}
\mathbb{E}[f(\mathbf{x}^{(t)})] \leq & \mathbb{E}[f(\mathbf{x}^*)] + \frac{1}{2}\mathbb{E}[\Lambda^{(t)}] + \frac{1}{2\rho^{(t)}}\mathbb{E}[\|\boldsymbol{\lambda}^{(t-1)}\|^2 - \|\boldsymbol{\lambda}^{(t)}\|^2] - \frac{\rho^{(t)}}{2}\mathbb{E}[\|\mathbf{A}\mathbf{y}^{(t-1)} - \mathbf{b}\|^2] \\
& + \frac{\rho^{(t)}}{2}\mathbb{E}[\|\mathbf{A}(\mathbf{x}^{(t)} - \mathbf{y}^{(t-1)})\|^2] - \frac{\nu^{(t)}}{2}\mathbb{E}[\|\mathbf{x}^{(t)} - \mathbf{y}^{(t-1)}\|^2] + \frac{2(B^{(t)})^2}{\nu^{(t)}(K^{(t)}+1)} \\
\overset{(a)}{\leq} & \mathbb{E}[f(\mathbf{x}^*)] + \frac{1}{2}\mathbb{E}[\Gamma^{(t)}] + \frac{1}{2\rho^{(t)}}\mathbb{E}[\|\boldsymbol{\lambda}^{(t-1)}\|^2 - \|\boldsymbol{\lambda}^{(t)}\|^2] + \frac{2(B^{(t)})^2}{\nu^{(t)}(K^{(t)}+1)}
\end{aligned}$$

where (a) follows by ignoring the negative term $-\frac{\rho^{(t)}}{2}\mathbb{E}[\|\mathbf{A}\mathbf{y}^{(t-1)} - \mathbf{b}\|^2]$ and noting that $\rho^{(t)}\|\mathbf{A}(\mathbf{x}^{(t)} - \mathbf{y}^{(t-1)})\|^2 - \nu^{(t)}\|\mathbf{x}^{(t)} - \mathbf{y}^{(t-1)}\|^2 \leq 0$ for any $\mathbf{x}^{(t)}$ and $\mathbf{y}^{(t-1)}$ as long as $\nu^{(t)} \geq \rho^{(t)}\|\mathbf{A}\|^2$.

Multiplying both sides by $\rho^{(t)}$ and summing over $t \in \{1, 2, \ldots, T\}$ yields

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}[\rho^{(t)}f(\mathbf{x}^{(t)})] \leq & \sum_{t=1}^{T}\mathbb{E}[\rho^{(t)}f(\mathbf{x}^*)] + \frac{1}{2}\sum_{t=1}^{T}\mathbb{E}[\rho^{(t)}\Gamma^{(t)}] + \frac{1}{2}\sum_{t=1}^{T}\mathbb{E}[\|\boldsymbol{\lambda}^{(t-1)}\|^2 - \|\boldsymbol{\lambda}^{(t)}\|^2] \\
& + \sum_{t=1}^{T}\frac{2\rho^{(t)}(B^{(t)})^2}{\nu^{(t)}(K^{(t)}+1)} \\
\overset{(a)}{=} & \sum_{t=1}^{T}\mathbb{E}[\rho^{(t)}f(\mathbf{x}^*)] + \frac{1}{2}\sum_{t=1}^{T}\mathbb{E}[\rho^{(t)}\Gamma^{(t)}] - \frac{1}{2}\mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|^2] + \sum_{t=1}^{T}\frac{2\rho^{(t)}(B^{(t)})^2}{\nu^{(t)}(K^{(t)}+1)}
\end{aligned}$$

where (a) follows by simplifying the telescoping sum and recalling that $\boldsymbol{\lambda}^{(0)} = \mathbf{0}$. □

Lemma 12. *Consider convex program* (1) *under Assumption 1-2. Let* $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ *be any saddle point defined in Assumption 1. For all* $T \geq 1$*, if we choose any fixed* $\rho^{(t)} = \rho > 0$*,* $\nu^{(t)} = \nu > 8\rho\|\mathbf{A}\|^2$*,* $K^t = K \geq T$ *in Algorithm 1 and the sub-procedure STO-LOCAL (Algorithm 2) uses* $\widehat{\mathbf{z}}$ *defined in Lemma 1 as the output, then we have*

$$\mathbb{E}[\|\boldsymbol{\lambda}^{(t)}\|^2] \leq Q, \forall t \in \{0, 1, \ldots, T\}$$

where

$$Q \triangleq \left(\frac{2\|\boldsymbol{\lambda}^*\| + \sqrt{\rho\nu\|\mathbf{x}^* - \mathbf{y}^{(0)}\|^2 + \frac{24\rho D^2}{\nu} + \frac{24(\rho)^3\|\mathbf{A}\|^2(\|\mathbf{A}\|R + \|\mathbf{b}\|)^2}{\nu} + 96\nu\rho R^2}}{1 - \sqrt{\frac{8\rho\|\mathbf{A}\|^2}{\nu}}} \right)^2$$

is an absoute constant (independent of T*) with constants* R, D *defined in Assumption 2.*

Proof. We prove this lemma by inductions. Note that $\boldsymbol{\lambda}^{(0)} = \mathbf{0}$ trivially satisfies $\mathbb{E}[\|\boldsymbol{\lambda}^{(0)}\|^2] \leq Q$. Assume $\mathbb{E}[\|\boldsymbol{\lambda}^{(t)}\|^2] \leq Q$ holds for all $t \leq t_0$ with $0 \leq t_0 \leq T-1$ and consider $t = t_0 + 1$.

By Lemma 4, we have

$$\sum_{t=1}^{t_0+1}\rho f(\mathbf{x}^*) - \sum_{t=1}^{t_0+1}\rho f(\mathbf{x}^{(t)}) \leq \|\boldsymbol{\lambda}^*\|\|\boldsymbol{\lambda}^{(t_0+1)}\|$$

Taking expectations on both sides yields

$$\sum_{t=1}^{t_0+1}\mathbb{E}[\rho f(\mathbf{x}^*)] - \sum_{t=1}^{t_0+1}\mathbb{E}[\rho f(\mathbf{x}^{(t)})] \leq \|\boldsymbol{\lambda}^*\|\mathbb{E}[\|\boldsymbol{\lambda}^{(t_0+1)}\|] \overset{(a)}{\leq} \|\boldsymbol{\lambda}^*\|\sqrt{\mathbb{E}[\|\boldsymbol{\lambda}^{(t_0+1)}\|^2]} \tag{55}$$

where (a) follows because $\mathbb{E}[X^2] \geq (\mathbb{E}[X])^2$ for any random variable X.

Note that our selection of $\rho^{(t)} = \rho$ and $\nu^{(t)} = \nu > 8\rho\|\mathbf{A}\|^2$ satisfies the condition in Lemma 11. By Lemma 11, we have

$$\begin{aligned}
&\sum_{t=1}^{t_0+1} \mathbb{E}[\rho f(\mathbf{x}^{(t)})] \\
\leq &\sum_{t=1}^{t_0+1} \mathbb{E}[\rho f(\mathbf{x}^*)] + \frac{\rho\nu}{2} \sum_{t=1}^{t_0+1} \mathbb{E}[\|\mathbf{x}^* - \mathbf{y}^{(t-1)}\|^2 - \|\mathbf{x}^* - \mathbf{y}^{(t)}\|^2] - \frac{1}{2}\mathbb{E}[\|\boldsymbol{\lambda}^{(t_0+1)}\|^2] + \sum_{t=1}^{t_0+1} \frac{2\rho(B^{(t)})^2}{\nu(K+1)} \\
\leq &\sum_{t=1}^{t_0+1} \mathbb{E}[\rho f(\mathbf{x}^*)] + \frac{\rho\nu}{2}\|\mathbf{x}^* - \mathbf{y}^{(0)}\|^2 - \frac{1}{2}\mathbb{E}[\|\boldsymbol{\lambda}^{(t_0+1)}\|^2] + \sum_{t=1}^{t_0+1} \frac{2\rho(B^{(t)})^2}{\nu(K+1)}
\end{aligned}$$

Rearranging terms yields

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{\lambda}^{(t_0+1)}\|^2] \leq &2\left(\sum_{t=1}^{t_0+1} \mathbb{E}[\rho f(\mathbf{x}^*)] - \sum_{t=1}^{t_0+1} \mathbb{E}[\rho f(\mathbf{x}^{(t)})]\right) + \rho\nu\|\mathbf{x}^* - \mathbf{y}^{(0)}\|^2 + \sum_{t=1}^{t_0+1} \frac{4\rho(B^{(t)})^2}{\nu(K+1)} \\
\overset{(a)}{\leq} &2\|\boldsymbol{\lambda}^*\|\sqrt{\mathbb{E}[\|\boldsymbol{\lambda}^{(t_0+1)}\|^2]} + \rho\nu\|\mathbf{x}^* - \mathbf{y}^{(0)}\|^2 + \sum_{t=1}^{t_0+1} \frac{4\rho(B^{(t)})^2}{\nu(K+1)} \qquad (56)
\end{aligned}$$

where (a) follows by using (55).

Recalling the definition of $(B^{(t)})^2$ in (51) (with $\rho^{(t)} = \rho$ and $\nu^{(t)} = \nu$), we have

$$\begin{aligned}
&\sum_{t=1}^{t_0+1} \frac{4\rho(B^{(t)})^2}{\nu(K+1)} \\
= &\frac{4\rho}{\nu(K+1)} \sum_{t=1}^{t_0+1} \left(2\|\mathbf{A}\|^2\mathbb{E}[\|\boldsymbol{\lambda}^{(t-1)}\|^2] + 6D^2 + 6(\rho)^2\|\mathbf{A}\|^2(\|\mathbf{A}\|R + \|\mathbf{b}\|)^2 + 24(\nu)^2R^2\right) \\
\overset{(a)}{\leq} &\frac{24\rho D^2}{\nu} + \frac{24(\rho)^3\|\mathbf{A}\|^2(\|\mathbf{A}\|R + \|\mathbf{b}\|)^2}{\nu} + 96\nu\rho R^2 + \frac{8\rho\|\mathbf{A}\|^2}{\nu}Q \qquad (57)
\end{aligned}$$

where (a) follows because $t_0 + 1 \leq (K+1)$ by our selection of K and $\mathbb{E}[\|\boldsymbol{\lambda}^{(t)}\|^2] \leq Q, \forall 0 \leq t \leq t_0$ by induction hypothesis.

Denote $c \triangleq \rho\nu\|\mathbf{x}^* - \mathbf{y}^{(0)}\|^2 + \frac{24\rho D^2}{\nu} + \frac{24(\rho)^3\|\mathbf{A}\|^2(\|\mathbf{A}\|R+\|\mathbf{b}\|)^2}{\nu} + 96\nu\rho R^2$. Note that $Q = \left(\frac{2\|\boldsymbol{\lambda}^*\|+\sqrt{c}}{1-\sqrt{\frac{8\rho\|\mathbf{A}\|^2}{\nu}}}\right)^2$. Substituting (57) into (56) yields

$$\mathbb{E}[\|\boldsymbol{\lambda}^{(t_0+1)}\|^2]] \leq 2\|\boldsymbol{\lambda}^*\|\sqrt{\mathbb{E}[\|\boldsymbol{\lambda}^{(t_0+1)}\|^2]} + c + \frac{8\rho\|\mathbf{A}\|^2}{\nu}Q.$$

This can be rewritten as

$$\left(\sqrt{\mathbb{E}[\|\boldsymbol{\lambda}^{(t_0+1)}\|^2]} - \|\boldsymbol{\lambda}^*\|\right)^2 \leq \|\boldsymbol{\lambda}^*\|^2 + c + \frac{8\rho\|\mathbf{A}\|^2}{\nu}Q,$$

which further implies that

$$\begin{aligned}
\sqrt{\mathbb{E}[\|\boldsymbol{\lambda}^{(t_0+1)}\|^2]} \leq &\|\boldsymbol{\lambda}^*\| + \sqrt{\|\boldsymbol{\lambda}^*\|^2 + c + \frac{8\rho\|\mathbf{A}\|^2}{\nu}Q} \\
\overset{(a)}{\leq} &2\|\boldsymbol{\lambda}^*\| + \sqrt{c} + \sqrt{\frac{8\rho\|\mathbf{A}\|^2}{\nu}}\sqrt{Q} \\
= &\frac{2\|\boldsymbol{\lambda}^*\| + \sqrt{c}}{1 - \sqrt{\frac{8\rho\|\mathbf{A}\|^2}{\nu}}}
\end{aligned}$$

where (a) follows from the basic inequality $\sqrt{a_1+a_2+a_3} \leq \sqrt{a_1}+\sqrt{a_2}+\sqrt{a_3}$ for all $a_1, a_2, a_3 \geq 0$; and (b) follows by substituting $Q = \left(\frac{2\|\boldsymbol{\lambda}^*\|+\sqrt{c}}{1-\sqrt{\frac{8\rho\|\mathbf{A}\|^2}{\nu}}}\right)^2$.

Squaring both sides yields

$$\mathbb{E}[\|\boldsymbol{\lambda}^{(t_0+1)}\|^2] \leq \left(\frac{2\|\boldsymbol{\lambda}^*\|+\sqrt{c}}{1-\sqrt{\frac{8\rho\|\mathbf{A}\|^2}{\nu}}}\right)^2 = Q$$

By far, we have shown $\mathbb{E}[\|\boldsymbol{\lambda}^{(t)}\|^2] \leq Q$ for $t = t_0+1$. Thus, this lemma follows by inductions. □

Main proof of Theorem 1: Now, we are ready to prove the theorem. Fix $T \geq 1$. By Lemma 11,

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}[\rho f(\mathbf{x}^{(t)})] \\
\leq &\sum_{t=1}^{Ts} \mathbb{E}[\rho f(\mathbf{x}^*)] + \frac{\rho\nu}{2}\sum_{t=1}^{T}\mathbb{E}[\|\mathbf{x}^*-\mathbf{y}^{(t-1)}\|^2 - \|\mathbf{x}^*-\mathbf{y}^{(t)}\|^2] - \frac{1}{2}\mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|^2] + \sum_{t=1}^{T}\frac{2\rho(B^{(t)})^2}{\nu(K+1)} \\
\leq &\sum_{t=1}^{t_0+1} \mathbb{E}[\rho f(\mathbf{x}^*)] + \frac{\rho\nu}{2}\|\mathbf{x}^*-\mathbf{y}^{(0)}\|^2 + \sum_{t=1}^{T}\frac{2\rho(B^{(t)})^2}{\nu(K+1)} \qquad (58)
\end{aligned}$$

Dividing both sides by ρT and applying Jensen's inequality yields

$$\mathbb{E}f(\overline{\mathbf{x}}^T)] \leq f(\mathbf{x}^*)] + \frac{\nu}{2T}\|\mathbf{x}^*-\mathbf{y}^{(0)}\|^2 + \frac{2}{\nu T}\sum_{t=1}^{T}\frac{(B^{(t)})^2}{K+1} \qquad (59)$$

Note that for all $t \in \{1, 2, \ldots, T\}$, we have

$$\begin{aligned}
\frac{(B^{(t)})^2}{K+1} =& \frac{2\|\mathbf{A}\|^2\mathbb{E}[\|\boldsymbol{\lambda}^{(t-1)}\|^2] + 6D^2 + 6\rho^2\|\mathbf{A}\|^2(\|\mathbf{A}\|R+\|\mathbf{b}\|)^2 + 24\nu^2R^2}{K+1} \\
\overset{(a)}{\leq}& \frac{2\|\mathbf{A}\|^2Q + 6D^2 + 6\rho^2\|\mathbf{A}\|^2(\|\mathbf{A}\|R+\|\mathbf{b}\|)^2 + 24\nu^2R^2}{T} \qquad (60)
\end{aligned}$$

where (a) follows because $\mathbb{E}[\|\boldsymbol{\lambda}^{(t)}\|^2] \leq Q, \forall t \in \{0, 1, \ldots, T\}$ by Lemma 12 and $K \geq T$.

Substituting (60) into (59) yields

$$\mathbb{E}f(\overline{\mathbf{x}}^T)] \leq f(\mathbf{x}^*) + \frac{\nu}{2T}\|\mathbf{x}^*-\mathbf{y}^{(0)}\|^2 + \frac{C}{\nu T}$$

with $C \triangleq 4\|\mathbf{A}\|^2Q + 12D^2 + 12\rho^2\|\mathbf{A}\|^2(\|\mathbf{A}\|R+\|\mathbf{b}\|)^2 + 48\nu^2R^2$. This is (9) of our theorem.

By part (1) of Lemma 3 (with $\rho^{(t)} = \rho$), we have $\sum_{t=1}^{T}\rho\left(\mathbf{A}\mathbf{x}^{(t)}-\mathbf{b}\right) = \boldsymbol{\lambda}^{(T)}$. Dividing both sides by ρT, taking the vector l_2 norm and then taking expectations on both sides yields

$$\begin{aligned}
\mathbb{E}[\|\mathbf{A}\overline{\mathbf{x}}^{(T)}-\mathbf{b}\|] =& \frac{1}{\rho T}\mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|] \\
\leq& \frac{1}{\rho T}\sqrt{\mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|^2]} \\
\overset{(a)}{\leq}& \frac{\sqrt{Q}}{\rho T}
\end{aligned}$$

where (a) follows from Lemma 12. This is (10) of our theorem.

6.8 Proof of Theorem 2

For convenience of our presentation, we extract the assumptions in Theorem 2 and call it Assumption 3:

Assumption 3. *Convex program* (1) *satisfies the following:*

- *The function $f(\mathbf{x})$ is L-smooth.*
- *The function $f(\mathbf{x})$ has unbiased stochastic gradients with a bounded variance, i.e., there exists constant $\sigma > 0$ such that $\mathbb{E}_\xi[\|\mathbf{G}(\mathbf{x};\xi) - \nabla f(\mathbf{x})\|^2] \leq \sigma^2, \forall \mathbf{x} \in \mathcal{X}$.*

Lemma 13. *Consider convex program* (1) *with μ-convex stochastic objective functions under Assumption 3. Let $\mathbf{x}^*$ be any optimal solution. If $\nu^{(t)} > 0$ and $\rho^{(t)} > 0$ in Algorithm 1 are chosen to satisfy*

$$\nu^{(t)} \geq \rho^{(t)}\|\mathbf{A}\|^2, \forall t,$$

and the sub-procedure STO-LOCAL (Algorithm 2) uses $k_0 \geq 2\frac{\nu^{(t)}+L}{\nu^{(t)}+\mu}, \forall t$ and $\widehat{\mathbf{z}}$ defined in Lemma 2 as the output, then, for all $T \geq 1$, Algorithm 1 ensures

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}[\rho^{(t)}f(\mathbf{x}^{(t)})]\\
\leq&\sum_{t=1}^{T}\mathbb{E}[\rho^{(t)}f(\mathbf{x}^*)] + \frac{1}{2}\sum_{t=1}^{T}\mathbb{E}[\rho^{(t)}\Lambda^{(t)}] - \frac{1}{2}\mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|^2] + \sum_{t=1}^{T}\frac{2\rho^{(t)}k_0\sigma^2}{(\nu^{(t)}+\mu)(K^{(t)}+2k_0-1)}
\end{aligned}$$

$$\Lambda^{(t)} \triangleq \left(\nu^{(t)} + \frac{(\nu^{(t)}+\mu)(k_0^2-k_0)}{K^t(K^t+2k_0-1)}\right)\|\mathbf{x}^* - \mathbf{y}^{(t-1)}\|^2 - \left(\nu^{(t)} + \mu + \frac{(\nu^{(t)}+\mu)(k_0^2-k_0)}{K^t(K^t+2k_0-1)}\right)\|\mathbf{x}^* - \mathbf{y}^{(t)}\|^2 \tag{61}$$

and σ^2 is the constant defined in Assumption 3.

Proof. Fix $t \in \{1, 2, \ldots, T\}$. Define $\phi^{(t)}(\mathbf{x}) = \sum_{i=1}^{N}\phi_i^{(t)}(\mathbf{x}_i)$. Note that $\phi^{(t)}(\mathbf{x})$ is $(L+\nu^{(t)})$-smooth and $(\mu+\nu^{(t)})$-convex. Similarly to the observation in the proof of Lemma 11, each iteration of Algorithm 1 is to jointly update $\mathbf{x}$ and $\mathbf{y}$ via the sub-procedure $(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) = \text{STO-LOCAL}(\phi^{(t)}(\cdot), \mathcal{X}, \mathbf{y}^{(t-1)}, K^{(t)})$. Note that the stochastic gradient used in each iteration of the sub-procedure is given by

$$\boldsymbol{\zeta}^{(k)} = \mathbf{G}^{(k)} + \mathbf{A}^\mathsf{T}(\rho^{(t)}(\mathbf{A}\mathbf{y}^{(t-1)} - \mathbf{b}) + \nu^{(t)}(\mathbf{z}^{(k)} - \mathbf{y}^{(t-1)}))$$

and has the same variance bound σ^2 as the stochastic gradient of $f(\mathbf{x})$.

By Lemma 2, we have

$$\begin{aligned}
&\mathbb{E}[\phi^{(t)}(\mathbf{x}^{(t)})]\\
\leq&\mathbb{E}[\phi^{(t)}(\mathbf{x}^*)] + \frac{(\nu^{(t)}+\mu)(k_0^2-k_0)}{2K^{(t)}(K^{(t)}+2k_0-1)}\mathbb{E}[\|\mathbf{x}^* - \mathbf{y}^{(t-1)}\|^2] - \frac{(\nu^{(t)}+\mu)(k_0^2-k_0)}{2K^{(t)}(K^{(t)}+2k_0-1)}\mathbb{E}[\|\mathbf{x}^* - \mathbf{y}^{(t)}\|^2]\\
&- \frac{\nu^{(t)}+\mu}{2}\mathbb{E}[\|\mathbf{x}^* - \mathbf{y}^{(t)}\|^2] + \frac{2k_0\sigma^2}{(K^{(t)}+2k_0-1)(\nu^{(t)}+\mu)}
\end{aligned} \tag{62}$$

Substituting the expression of $\phi_i^{(t)}(\cdot)$ (defined in (2)) into the above equation (and recalling $\mathbf{r}^{(t)} = \mathbf{A}\mathbf{y}^{(t-1)} - \mathbf{b}$) yields

$$\begin{aligned}
&\mathbb{E}[f(\mathbf{x}^{(t)})] + \rho^{(t)}\mathbb{E}[\langle \mathbf{A}\mathbf{y}^{(t-1)} - \mathbf{b}, \mathbf{A}\mathbf{x}^{(t)} - \mathbf{b}\rangle] + \mathbb{E}[\langle \boldsymbol{\lambda}^{(t-1)}, \mathbf{A}\mathbf{x}^{(t)} - \mathbf{b}\rangle] \\
&+ \frac{\nu^{(t)}}{2}\mathbb{E}[\|\mathbf{x}^{(t)} - \mathbf{y}^{(t-1)}\|^2] \\
\leq &\mathbb{E}[f(\mathbf{x}^*)] + \rho^{(t)}\mathbb{E}[\langle \mathbf{A}\mathbf{y}^{(t-1)} - \mathbf{b}, \mathbf{A}\mathbf{x}^* - \mathbf{b}\rangle] + \mathbb{E}[\langle \boldsymbol{\lambda}^{(t-1)}, \mathbf{A}\mathbf{x}^* - \mathbf{b}\rangle] + \frac{\nu^{(t)}}{2}\mathbb{E}[\|\mathbf{x}^* - \mathbf{y}^{(t-1)}\|^2] \\
&+ \frac{(\nu^{(t)} + \mu)(k_0^2 - k_0)}{2K^{(t)}(K^{(t)} + 2k_0 - 1)}\mathbb{E}[\|\mathbf{x}^* - \mathbf{y}^{(t-1)}\|^2] - \frac{(\nu^{(t)} + \mu)(k_0^2 - k_0)}{2K^{(t)}(K^{(t)} + 2k_0 - 1)}\mathbb{E}[\|\mathbf{x}^* - \mathbf{y}^{(t)}\|^2] \\
&- \frac{(\nu^{(t)} + \mu)}{2}\mathbb{E}[\|\mathbf{x}^* - \mathbf{y}^{(t)}\|^2] + \frac{2k_0\sigma^2}{(K^{(t)} + 2k_0 - 1)(\nu^{(t)} + \mu)} \\
\overset{(a)}{=} &\mathbb{E}[f(\mathbf{x}^*)] + \frac{1}{2}\mathbb{E}[\Lambda^{(t)}] + \frac{2k_0\sigma^2}{(K^{(t)} + 2k_0 - 1)(\nu^{(t)} + \mu)}
\end{aligned} \tag{63}$$

where (a) follows from $\mathbf{A}\mathbf{x}^* - \mathbf{b} = \mathbf{0}$ and the definition of $\Lambda^{(t)}$ in (61).

The remaining part of our proof is almost identical to the proof of Lemma 11. Recall that by part (2) of Lemma 3, we have

$$\left\langle \boldsymbol{\lambda}^{(t-1)}, \mathbf{A}\mathbf{x}^{(t)} - \mathbf{b} \right\rangle = \frac{1}{2\rho^{(t)}}\left(\|\boldsymbol{\lambda}^{(t)}\|^2 - \|\boldsymbol{\lambda}^{(t-1)}\|^2\right) - \frac{\rho^{(t)}}{2}\|\mathbf{A}\mathbf{x}^{(t)} - \mathbf{b}\|^2 \tag{64}$$

By the basic identity $\langle \mathbf{u}, \mathbf{v}\rangle = \frac{1}{2}\|\mathbf{u}\|^2 + \frac{1}{2}\|\mathbf{v}\|^2 - \frac{1}{2}\|\mathbf{u} - \mathbf{v}\|^2$ for any vector $\mathbf{u}, \mathbf{v}$, we have

$$\left\langle \mathbf{A}\mathbf{y}^{(t-1)} - \mathbf{b}, \mathbf{A}\mathbf{x}^{(t)} - \mathbf{b} \right\rangle = \frac{1}{2}\|\mathbf{A}\mathbf{y}^{(t-1)} - \mathbf{b}\|^2 + \frac{1}{2}\|\mathbf{A}\mathbf{x}^{(t)} - \mathbf{b}\|^2 - \frac{1}{2}\|\mathbf{A}(\mathbf{x}^{(t)} - \mathbf{y}^{(t-1)})\|^2 \tag{65}$$

Substituting (64) and (65) into (63) and rearranging terms yields

$$\begin{aligned}
&\mathbb{E}[f(\mathbf{x}^{(t)})] \\
\leq &\mathbb{E}[f(\mathbf{x}^*)] + \frac{1}{2}\mathbb{E}[\Lambda^{(t)}] + \frac{1}{2\rho^{(t)}}\mathbb{E}[\|\boldsymbol{\lambda}^{(t-1)}\|^2 - \|\boldsymbol{\lambda}^{(t)}\|^2] - \frac{\rho^{(t)}}{2}\|\mathbf{A}\mathbf{y}^{(t-1)} - \mathbf{b}\|^2 \\
&+ \frac{\rho^{(t)}}{2}\mathbb{E}[\|\mathbf{A}(\mathbf{x}^{(t)} - \mathbf{y}^{(t-1)})\|^2] - \frac{\nu^{(t)}}{2}\mathbb{E}[\|\mathbf{x}^{(t)} - \mathbf{y}^{(t-1)}\|^2] + \frac{2k_0\sigma^2}{(K^{(t)} + 2k_0 - 1)(\nu^{(t)} + \mu)} \\
\overset{(a)}{\leq} &\mathbb{E}[f(\mathbf{x}^*)] + \frac{1}{2}\mathbb{E}[\Lambda^{(t)}] + \frac{1}{2\rho^{(t)}}\mathbb{E}[\|\boldsymbol{\lambda}^{(t-1)}\|^2 - \|\boldsymbol{\lambda}^{(t)}\|^2] + \frac{2k_0\sigma^2}{(K^{(t)} + 2k_0 - 1)(\nu^{(t)} + \mu)}
\end{aligned}$$

where (a) follows by ignoring the negative term $-\frac{\rho^{(t)}}{2}\|\mathbf{A}\mathbf{y}^{(t-1)} - \mathbf{b}\|^2$ and noting that $\rho^{(t)}\|\mathbf{A}(\mathbf{x}^{(t)} - \mathbf{y}^{(t-1)})\|^2 - \nu^{(t)}\|\mathbf{x}^{(t)} - \mathbf{y}^{(t-1)}\|^2 \leq 0$ for any $\mathbf{x}^{(t)}$ and $\mathbf{y}^{(t-1)}$ as long as $\nu^{(t)} \geq \rho^{(t)}\|\mathbf{A}\|^2$.

Multiplying both sides by $\rho^{(t)}$ and summing over $t \in \{1, 2, \ldots, T\}$ yields

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}[\rho^{(t)}f(\mathbf{x}^{(t)})] \\
\leq &\sum_{t=1}^{T}\mathbb{E}[\rho^{(t)}f(\mathbf{x}^*)] + \frac{1}{2}\sum_{t=1}^{T}\mathbb{E}[\rho^{(t)}\Lambda^{(t)}] + \frac{1}{2}\sum_{t=1}^{T}\mathbb{E}[\|\boldsymbol{\lambda}^{(t-1)}\|^2 - \|\boldsymbol{\lambda}^{(t)}\|^2] \\
&+ \sum_{t=1}^{T}\frac{2\rho^{(t)}k_0\sigma^2}{(K^{(t)} + 2k_0 - 1)(\nu^{(t)} + \mu)} \\
\overset{(a)}{=} &\sum_{t=1}^{T}\mathbb{E}[\rho^{(t)}f(\mathbf{x}^*)] + \frac{1}{2}\sum_{t=1}^{T}\mathbb{E}[\rho^{(t)}\Lambda^{(t)}] - \frac{1}{2}\mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|^2] + \sum_{t=1}^{T}\frac{2\rho^{(t)}k_0\sigma^2}{(K^{(t)} + 2k_0 - 1)(\nu^{(t)} + \mu)}
\end{aligned}$$

where (a) follows by simplifying the telescoping sums and recalling that $\boldsymbol{\lambda}^{(0)} = \mathbf{0}$. □

Now we are ready to prove both parts of the theorem.

1. **Proof of case $\mu = 0$:** Fix $T \geq 1$. Note that our selection of $\rho^{(t)} = \rho$, $\nu^{(t)} = \nu \geq \rho\|\mathbf{A}\|^2$ and positive integer $k_0 \geq 2\frac{L+\nu}{\nu}$ satisfies the condition in Lemma 13. By Lemma 13 (with $\mu = 0$), we have

$$\sum_{t=1}^{T} \mathbb{E}[\rho f(\mathbf{x}^{(t)})] \leq \sum_{t=1}^{T} \mathbb{E}[\rho f(\mathbf{x}^*)] + \frac{\rho}{2}\sum_{t=1}^{T} \mathbb{E}[\Lambda^{(t)}] - \frac{1}{2}\mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|^2] + \sum_{t=1}^{T} \frac{2\rho k_0 \sigma^2}{(T + 2k_0 - 1)\nu} \tag{66}$$

Note that

$$\begin{aligned}\sum_{t=1}^{T} \Lambda^{(t)} =& \nu\Big(1 + \frac{k_0^2 - k_0}{T(T + 2k_0 - 1)}\Big) \sum_{t=1}^{T} \Big(\|\mathbf{x}^* - \mathbf{y}^{(t-1)}\|^2 - \|\mathbf{x}^* - \mathbf{y}^{(t)}\|^2\Big) \\ =& \nu\Big(1 + \frac{k_0^2 - k_0}{T(T + 2k_0 - 1)}\Big) \Big(\|\mathbf{x}^* - \mathbf{y}^{(0)}\|^2 - \|\mathbf{x}^* - \mathbf{y}^{T}\|^2\Big) \\ \leq& \nu\Big(1 + \frac{k_0^2 - k_0}{T(T + 2k_0 - 1)}\Big)\|\mathbf{x}^* - \mathbf{y}^{(0)}\|^2 \end{aligned} \tag{67}$$

and

$$\sum_{t=1}^{T} \frac{2\rho k_0 \sigma^2}{(T + 2k_0 - 1)\nu} = \frac{2\rho k_0 \sigma^2}{\nu} \frac{T}{T + 2k_0 - 1} \leq \frac{2\rho k_0 \sigma^2}{\nu} \tag{68}$$

Substituting (67)-(68) into (66) yields

$$\begin{aligned}&\sum_{t=1}^{T} \mathbb{E}[\rho f(\mathbf{x}^{(t)})] \\ \leq& \sum_{t=1}^{T} \mathbb{E}[\rho f(\mathbf{x}^*)] + \frac{\rho\nu}{2}\Big(1 + \frac{k_0^2 - k_0}{T(T + 2k_0 - 1)}\Big)\|\mathbf{x}^* - \mathbf{y}^{(0)}\|^2 + \frac{2\rho k_0 \sigma^2}{\nu} - \frac{1}{2}\mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|^2] \\ \overset{(a)}{\leq}& \sum_{t=1}^{T} \mathbb{E}[\rho f(\mathbf{x}^*)] + \frac{\rho\nu(k_0 + 1)}{4}\|\mathbf{x}^* - \mathbf{y}^{(0)}\|^2 + \frac{2\rho k_0 \sigma^2}{\nu} - \frac{1}{2}\mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|^2] \end{aligned} \tag{69}$$

where (a) follows because $\frac{k_0^2 - k_0}{T(T+2k_0-1)} \leq \frac{k_0 - 1}{2}$ when $T \geq 1$.

Ignoring the negative term $-\frac{1}{2}\mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|^2]$, dividing both sides by ρT and applying Jensen's inequality yields

$$\mathbb{E}[f(\overline{\mathbf{x}}^{(T)})] \leq f(\mathbf{x}^*) + \frac{1}{T}\frac{\nu(k_0 + 1)}{4}\|\mathbf{x}^* - \mathbf{y}^{(0)}\|^2 + \frac{1}{T}\frac{2k_0\sigma^2}{\nu}$$

This is (11) of our theorem.

By Lemma 4 (after taking expectations on both sides), we have

$$\mathbb{E}[\sum_{t=1}^{T} \rho f(\mathbf{x}^{(t)})] \geq \sum_{t=1}^{T} \mathbb{E}[\rho f(\mathbf{x}^*)] - \|\boldsymbol{\lambda}^*\|\mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|]$$

Combining this inequality with (69) and cancelling the common terms yields

$$\mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|^2] \leq 2\|\boldsymbol{\lambda}^*\|\mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|] + \frac{\rho\nu(k_0 + 1)}{2}\|\mathbf{x}^* - \mathbf{y}^{(0)}\|^2 + \frac{4\rho k_0 \sigma^2}{\nu}$$

Since $\Big(\mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|]\Big)^2 \leq \mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|^2]$, we further have

$$\Big(\mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|]\Big)^2 \leq 2\|\boldsymbol{\lambda}^*\|\mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|] + \frac{\rho\nu(k_0 + 1)}{2}\|\mathbf{x}^* - \mathbf{y}^{(0)}\|^2 + \frac{4\rho k_0 \sigma^2}{\nu}$$

This quadratic inequality can be rewritten as

$$\left(\mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|] - \|\boldsymbol{\lambda}^*\|\right)^2 \leq \|\boldsymbol{\lambda}^*\|^2 + \frac{\rho\nu(k_0+1)}{2}\|\mathbf{x}^* - \mathbf{y}^{(0)}\|^2 + \frac{4\rho k_0\sigma^2}{\nu}$$

Thus, we have

$$\begin{aligned}
\mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|] \leq & \|\boldsymbol{\lambda}^*\| + \sqrt{\|\boldsymbol{\lambda}^*\|^2 + \frac{\rho\nu(k_0+1)}{2}\|\mathbf{x}^* - \mathbf{y}^{(0)}\|^2 + \frac{4\rho k_0\sigma^2}{\nu}} \\
\leq & 2\|\boldsymbol{\lambda}^*\| + \sqrt{\frac{\rho\nu(k_0+1)}{2}\|\mathbf{x}^* - \mathbf{y}^{(0)}\|^2 + \frac{4\rho k_0\sigma^2}{\nu}} \qquad (70) \\
\leq & 2\|\boldsymbol{\lambda}^*\| + \sqrt{\frac{\rho\nu(k_0+1)}{2}}\|\mathbf{x}^* - \mathbf{y}^{(0)}\| + \sqrt{\frac{4\rho k_0\sigma^2}{\nu}} \qquad (71)
\end{aligned}$$

By part (1) of Lemma 3 (with $\rho^{(t)} = \rho$), we have

$$\sum_{t=1}^{T} \rho\left(\mathbf{A}\mathbf{x}^{(t)} - \mathbf{b}\right) = \boldsymbol{\lambda}^{(T)}$$

Dividing both sides by ρT, taking the vector l_2 norm and then taking expectations on both sides yields

$$\begin{aligned}
\mathbb{E}[\|\mathbf{A}\overline{\mathbf{x}}^{(T)} - \mathbf{b}\|] = & \frac{1}{\rho T}\mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|] \\
\overset{(a)}{\leq} & \frac{1}{T}\left(\frac{2}{\rho}\|\boldsymbol{\lambda}^*\| + \sqrt{\frac{\nu(k_0+1)}{2\rho}}\|\mathbf{x}^* - \mathbf{y}^{(0)}\| + 2\sqrt{\frac{k_0\sigma^2}{\rho\nu}}\right)
\end{aligned}$$

where (a) follows from (71). This is (78) of our theorem.

2. **Proof of Case** $\mu > 0$: Note that our selection of $\rho^{(t)} = t\rho$, $\nu^{(t)} = t\rho\|\mathbf{A}\|^2$ and $k_0 \geq 2(1 + \frac{L}{\mu})$ satisfies the conditions in Lemma 13. By Lemma 13, we have

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}[\rho^{(t)}f(\mathbf{x}^{(t)})] \leq & \sum_{t=1}^{T}\mathbb{E}[\rho^{(t)}f(\mathbf{x}^*)] + \frac{1}{2}\sum_{t=1}^{T}\mathbb{E}[\rho^{(t)}\Lambda^{(t)}] - \frac{1}{2}\mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|^2] \\
& + \sum_{t=1}^{T}\frac{2\rho^{(t)}k_0\sigma^2}{(\nu^{(t)}+\mu)(K^{(t)}+2k_0-1)} \qquad (72)
\end{aligned}$$

Recalling the definition of Λ^t in (61), we have

$$\begin{aligned}
& \sum_{t=1}^{T}\rho^{(t)}\Lambda^{(t)} \\
= & \rho\left(\rho\|\mathbf{A}\|^2 + \frac{(\rho\|\mathbf{A}\|^2+\mu)(k_0^2-k_0)}{2(2k_0-1)^2}\right)\|\mathbf{x}^* - \mathbf{y}^{(0)}\|^2 \\
& - \sum_{t=1}^{T-1}\left(\rho\left(\rho t^2\|\mathbf{A}\|^2 + t\mu - \rho(t+1)^2\|\mathbf{A}\|^2\right) + \rho\left(\frac{\rho t\|\mathbf{A}\|^2+\mu}{t+1} - \frac{\rho(t+1)\|\mathbf{A}\|^2+\mu}{t+2}\right)\frac{k_0^2-k_0}{(2k_0-1)^2}\right)\|\mathbf{x}^* - \mathbf{y}^{(t)}\|^2 \\
& - T\rho\left(T\rho\|\mathbf{A}\|^2 + \mu + \left(\frac{\rho T\|\mathbf{A}\|^2+\mu}{T(T+1)}\right)\frac{k_0^2-k_0}{(2k_0-1)^2}\right)\|\mathbf{x}^* - \mathbf{y}^{(T)}\|^2 \\
\overset{(a)}{\leq} & \rho\left(\rho\|\mathbf{A}\|^2 + \frac{(\rho\|\mathbf{A}\|^2+\mu)(k_0^2-k_0)}{2(2k_0-1)^2}\right)\|\mathbf{x}^* - \mathbf{y}^{(0)}\|^2 \qquad (73)
\end{aligned}$$

where (a) follows by ignoring the last negative term and noting that $\rho t^2\|\mathbf{A}\|^2 + t\mu - \rho(t+1)^2\|\mathbf{A}\|^2 = t\mu - \rho(2t+1)\|\mathbf{A}\|^2 \geq \mu(t - \frac{2t+1}{3}) \geq 0$ for all $t \geq 1$, where the first inequality uses $\rho \leq \frac{\mu}{3\|\mathbf{A}\|^2}$; and $\frac{\rho t\|\mathbf{A}\|^2+\mu}{t+1} - \frac{\rho(t+1)\|\mathbf{A}\|^2+\mu}{t+2} = \frac{\mu-\rho\|\mathbf{A}\|^2}{(t+1)(t+2)} \geq 0$, where the inequality also uses $\rho \leq \frac{\mu}{3\|\mathbf{A}\|^2}$.

We also note that

$$\begin{aligned}\sum_{t=1}^{T}\frac{2\rho^{(t)}k_0\sigma^2}{(\nu^{(t)}+\mu)(K^{(t)}+2k_0-1)} =& \frac{2k_0\sigma^2\rho}{2k_0-1}\sum_{t=1}^{T}\frac{t}{(\rho t\|\mathbf{A}\|^2+\mu)(t+1)}\\ \overset{(a)}{\leq}& \frac{2k_0\sigma^2}{2k_0-1}\sum_{t=1}^{T}\frac{t}{(t+1)(t+3)\|\mathbf{A}\|^2}\\ \leq& \frac{2k_0\sigma^2}{(2k_0-1)\|\mathbf{A}\|^2}\sum_{t=1}^{T}\frac{1}{t+1}\\ \leq& \frac{2k_0\sigma^2}{(2k_0-1)\|\mathbf{A}\|^2}\log(T+1) \end{aligned} \tag{74}$$

where (a) follows because $\mu \geq 3\rho\|\mathbf{A}\|^2$ by $\rho \leq \frac{\mu}{3\|\mathbf{A}\|^2}$.
Substituting (73) and (74) into (72) yields

$$\begin{aligned}&\sum_{t=1}^{T}\mathbb{E}[\rho^{(t)}f(\mathbf{x}^{(t)})]\\ \leq&\sum_{t=1}^{T}\mathbb{E}[\rho^{(t)}f(\mathbf{x}^*)]+\frac{1}{2}\rho\left(\rho\|\mathbf{A}\|^2+\frac{(\rho\|\mathbf{A}\|^2+\mu)(k_0^2-k_0)}{2(2k_0-1)^2}\right)\|\mathbf{x}^*-\mathbf{y}^{(0)}\|^2-\frac{1}{2}\mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|^2]\\ &+\frac{2k_0\sigma^2}{(2k_0-1)\|\mathbf{A}\|^2}\log(T+1)\\ \overset{(a)}{\leq}&\sum_{t=1}^{T}\mathbb{E}[\rho^{(t)}f(\mathbf{x}^*)]+\frac{1}{2}\rho c_1\|\mathbf{x}^*-\mathbf{y}^{(0)}\|^2-\frac{1}{2}\mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|^2]+\frac{1}{2}c_2\log(T+1)\end{aligned} \tag{75}$$

where (a) follows because $c_1=\rho\|\mathbf{A}\|^2+\frac{(\rho\|\mathbf{A}\|^2+\mu)(k_0^2-k_0)}{2(2k_0-1)^2}$ and $c_2=\frac{4k_0\sigma^2}{(2k_0-1)\|\mathbf{A}\|^2}$.
Ignoring the negative term $-\frac{1}{2}\mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|^2]$, dividing both sides by $\sum_{t=1}^{T}\rho^{(t)}$ and applying Jensen's inequality yields

$$\begin{aligned}\mathbb{E}[f(\overline{\mathbf{x}}^{(T)})] \leq& f(\mathbf{x}^*)+\frac{1}{2\sum_{t=1}^{T}\rho^{(t)}}\left(\rho c_1\|\mathbf{x}^*-\mathbf{y}^{(0)}\|^2+c_2\log(T+1)\right)\\ \overset{(a)}{=}& f(\mathbf{x}^*)+\frac{1}{T(T+1)}\left(c_1\|\mathbf{x}^*-\mathbf{y}^{(0)}\|^2+\frac{c_2}{\rho}\log(T+1)\right)\end{aligned}$$

where (a) follows because $\sum_{t=1}^{T}\rho^{(t)}=\rho\sum_{t=1}^{T}t=\frac{\rho T(T+1)}{2}$. This is (77) of our theorem.
By Lemma 4 (after taking expectations on both sides), we have

$$\mathbb{E}[\sum_{t=1}^{T}\rho^{(t)}f(\mathbf{x}^{(t)})]\geq\sum_{t=1}^{T}\mathbb{E}[\rho^{(t)}f(\mathbf{x}^*)]-\|\boldsymbol{\lambda}^*\|\mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|]$$

Combining this inequality with (75) and cancelling the common terms yields

$$\mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|^2]\leq 2\|\boldsymbol{\lambda}^*\|\mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|]+\rho c_1\|\mathbf{x}^*-\mathbf{y}^{(0)}\|^2+c_2\log(T+1)$$

Since $\left(\mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|]\right)^2\leq\mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|^2]$, we further have

$$\left(\mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|]\right)^2\leq 2\|\boldsymbol{\lambda}^*\|\mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|]+\rho c_1\|\mathbf{x}^*-\mathbf{y}^{(0)}\|^2+c_2\log(T+1)$$

This quadratic inequality can be rewritten as

$$\left(\mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|]-\|\boldsymbol{\lambda}^*\|\right)^2\leq\|\boldsymbol{\lambda}^*\|^2+\rho c_1\|\mathbf{x}^*-\mathbf{y}^{(0)}\|^2+c_2\log(T+1)$$

Thus, we have

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|] \leq & \|\boldsymbol{\lambda}^*\| + \sqrt{\|\boldsymbol{\lambda}^*\|^2 + \rho c_1 \|\mathbf{x}^* - \mathbf{y}^{(0)}\|^2 + c_2 \log(T+1)} \\ \leq & 2\|\boldsymbol{\lambda}^*\| + \sqrt{\rho c_1}\|\mathbf{x}^* - \mathbf{y}^{(0)}\| + \sqrt{c_2 \log(T+1)} \end{aligned} \tag{76}$$

By part (1) of Lemma 3 (with $\rho^{(t)} = \rho$), we have

$$\sum_{t=1}^{T} \rho \left(\mathbf{A}\mathbf{x}^{(t)} - \mathbf{b} \right) = \boldsymbol{\lambda}^{(T)}$$

Dividing both sides by $\sum_{t=1}^{T} \rho^{(t)} = \rho \frac{T(T+1)}{2}$, taking the vector l_2 norm and then taking expectations on both sides yields

$$\begin{aligned}\mathbb{E}[\|\mathbf{A}\overline{\mathbf{x}}^{(T)} - \mathbf{b}\|] = & \frac{2}{T(T+1)} \mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|] \\ \overset{(a)}{\leq} & \frac{2}{T(T+1)} \left(\frac{4\|\boldsymbol{\lambda}^*\|}{\rho} + \frac{\sqrt{c_1}}{\sqrt{\rho}} \|\mathbf{x}^* - \mathbf{y}^{(0)}\| + \frac{\sqrt{c_2 \log(T+1)}}{\rho} \right) \end{aligned}$$

where (a) follows from (76). This is (78) of our theorem.

6.9 Performance of Algorithm 1 for strongly convex non-smooth problems

In this subsection, we consider stochastic convex program (1) under the the following assumption.

Assumption 4. *Convex program* (1) *satisfies the following:*

- *The function $f(\mathbf{x})$ satisfies* (15)*.*
- *The function $f(\mathbf{x})$ has unbiased stochastic subgradients with a bounded second order moment, i.e., there exists constant $D > 0$ such that $\mathbb{E}_{\xi}[\|\mathbf{G}(\mathbf{x}; \xi)\|^2] \leq D^2, \forall \mathbf{x} \in \mathcal{X}$.*

Theorem 4. *Consider convex program* (1) *with μ-convex ($\mu > 0$) possibly non-smooth function under Assumption 1 and Assumption 4. Let $(\mathbf{x}^*, \boldsymbol{\lambda}^*)$ be any saddle point defined in Assumption 1.*

If the sub-procedure STO-LOCAL (Algorithm 2) uses $\widehat{\mathbf{z}} \triangleq \frac{1}{\sum_{k=1}^{K}(k+k_0-1)}(k+k_0-1)\mathbf{z}^{(k)}$ as the output and $\rho \leq \frac{\mu}{3\|\mathbf{A}\|^2}$, $\rho^{(t)} = t\rho$, $\nu^{(t)} = t\rho\|\mathbf{A}\|^2$, $k_0 = 2$ and $K^t = 3t$ in Algorithm 1, then for all $T \geq 1$,

$$\mathbb{E}[f(\overline{\mathbf{x}}^{T})] \leq f(\mathbf{x}^*) + \frac{1}{T(T+1)} \left(c_1 \|\mathbf{x}^* - \mathbf{y}^{(0)}\|^2 + \frac{c_2}{\rho} \log(T+1) \right) \tag{77}$$

$$\mathbb{E}[\|\mathbf{A}\overline{\mathbf{x}}^{(T)} - \mathbf{b}\|] \leq \frac{2}{T(T+1)} \left(\frac{4\|\boldsymbol{\lambda}^*\|}{\rho} + \frac{\sqrt{c_1}}{\sqrt{\rho}} \|\mathbf{x}^* - \mathbf{y}^{(0)}\| + \frac{\sqrt{c_2 \log(T+1)}}{\rho} \right) \tag{78}$$

where $\overline{\mathbf{x}}^{(T)} = \frac{1}{\sum_{t=1}^{T} \rho^{(t)}} \sum_{t=1}^{T} \rho^{(t)} \mathbf{x}^{(t)}$; and $c_1 \triangleq \rho\|\mathbf{A}\|^2 + \frac{2(\rho\|\mathbf{A}\|^2 + \mu)}{18}$, $c_2 \triangleq \frac{16(B^2+M^2)}{3\|\mathbf{A}\|^2}$ are two constants.

We first develop a lemma that summarizes that Algorithm 2 behaves well as a sub-procedure under Assumption 4. This lemma essentially says Algorithm 2 has good performance when used to minimize a stochastic function that is the sum of a smooth part and a part that satisfying (15).

Lemma 14. *Assume $\phi(\mathbf{z}) = \dot{\phi}(\mathbf{z}) + \ddot{\phi}(\mathbf{z})$ where $\dot{\phi}(\mathbf{z})$ is μ_1-convex and satisfies that the assumption that there exists a constant $M > 0$ such that*

$$\dot{\phi}(\mathbf{z}_1) \leq \dot{\phi}(\mathbf{z}_2) + \langle \mathbf{d}, \mathbf{z}_1 - \mathbf{z}_2 \rangle + M\|\mathbf{z}_1 - \mathbf{z}_2\|, \tag{79}$$

for all $\mathbf{z}_1, \mathbf{z}_2 \in \mathcal{Z}$ and $\mathbf{d} \in \partial\dot{\phi}(\mathbf{z}_1)$; and $\ddot{\phi}(\mathbf{z})$ is a L-smooth and μ_2-convex deterministic function ($\mu_2 > 0$) over set $\mathcal{Z}$ with conditional number $\kappa = \frac{L}{\mu_2} = 1$. Assume there exists a constant B such that the unbiased subgradient $\boldsymbol{\zeta}^{(k)}$ used in Algorithm 2 satisfy

$$\mathbb{E}[\|\boldsymbol{\zeta}^{(k)}\|^2] \leq B^2, \forall k \in \{1, 2, \ldots, K\}$$

If we take $k_0 = 2$ in Algorithm 2, then we have

$$\mathbb{E}[\phi(\widehat{\mathbf{z}})] \leq \phi(\mathbf{z}) + \frac{\mu}{K(K+3)}\mathbb{E}[\|\mathbf{z}-\mathbf{z}^{(0)}\|^2] - \frac{\mu}{K(K+3)}\mathbb{E}[\|\mathbf{z}-\mathbf{z}^{(K)}\|^2] - \frac{\mu}{2}\mathbb{E}[\|\mathbf{z}-\mathbf{z}^{(K)}\|^2] + \frac{8(B^2+M^2)}{(K+3)\mu} \tag{80}$$

where $\widehat{\mathbf{z}} \triangleq \frac{1}{\sum_{k=1}^{K}(k+k_0-1)}(k+k_0-1)\mathbf{z}^{(k)}$ *and* $\mu \triangleq \mu_1 + \mu_2$

Proof. Fix $\mathbf{z} \in \mathcal{Z}$. At each iteration k, the projected gradient update (6) in Algorithm 2 can be rewritten as

$$\mathbf{z}^{(k)} = \operatorname*{argmin}_{\mathbf{z}\in\mathcal{Z}} \{\langle \boldsymbol{\zeta}^{(k)}, \mathbf{z}-\mathbf{z}^{(k-1)}\rangle + \frac{1}{2\gamma^{(k)}}\|\mathbf{z}-\mathbf{z}^{(k-1)}\|^2\}.$$

Since the objective function is $\frac{1}{\gamma^{(k)}}$-convex, by Lemma 5, we have

$$\begin{aligned}&\langle \boldsymbol{\zeta}^{(k)}, \mathbf{z}^{(k)}-\mathbf{z}^{(k-1)}\rangle + \frac{1}{2\gamma^{(k)}}\|\mathbf{z}^{(k)}-\mathbf{z}^{(k-1)}\|^2\\ \leq&\langle \boldsymbol{\zeta}^{(k)}, \mathbf{z}-\mathbf{z}^{(k-1)}\rangle + \frac{1}{2\gamma^{(k)}}\|\mathbf{z}-\mathbf{z}^{(k-1)}\|^2 - \frac{1}{2\gamma^{(k)}}\|\mathbf{z}-\mathbf{z}^{(k)}\|^2\end{aligned}$$

Since $\boldsymbol{\zeta}^{(k)}$ is an unbiased stochastic subgradient of $\phi(\mathbf{z})$ at $\mathbf{z} = \mathbf{z}^{(k-1)}$ and $\ddot{\phi}(\mathbf{z})$ is a deterministic function, we have $\mathbb{E}[\boldsymbol{\zeta}^{(k)}] = \mathbf{d} + \nabla\ddot{\phi}(\mathbf{z}^{(k-1)})$ for some $\mathbf{d} \in \partial\dot{\phi}(\mathbf{z}^{(k-1)})$.

Adding $\phi(\mathbf{z}^{(k-1)}) + \langle \mathbf{d} + \nabla\ddot{\phi}(\mathbf{z}^{(k-1)}) - \boldsymbol{\zeta}^{(k)}, \mathbf{z}^{(k)} - \mathbf{z}^{(k-1)}\rangle + \frac{L}{2}\|\mathbf{z}^{(k)}-\mathbf{z}^{(k-1)}\|^2 + M\|\mathbf{z}^{(k)}-\mathbf{z}^{(k-1)}\|$ on both sides and rearranging terms yields

$$\begin{aligned}&\phi(\mathbf{z}^{(k-1)}) + \langle \mathbf{d} + \nabla\ddot{\phi}(\mathbf{z}^{(k-1)}), \mathbf{z}^{(k)}-\mathbf{z}^{(k-1)}\rangle + \frac{L}{2}\|\mathbf{z}^{(k)}-\mathbf{z}^{(k-1)}\|^2 + M\|\mathbf{z}^{(k)}-\mathbf{z}^{(k-1)}\|\\ \leq&\phi(\mathbf{z}^{(k-1)}) + \langle \boldsymbol{\zeta}^{(k)}, \mathbf{z}-\mathbf{z}^{(k-1)}\rangle + \frac{1}{2\gamma^{(k)}}\|\mathbf{z}-\mathbf{z}^{(k-1)}\|^2 - \frac{1}{2\gamma^{(k)}}\|\mathbf{z}-\mathbf{z}^{(k)}\|^2\\ &- \frac{1}{2}(\frac{1}{\gamma^{(k)}} - L)\|\mathbf{z}^{(k)}-\mathbf{z}^{(k-1)}\|^2 + M\|\mathbf{z}^{(k)}-\mathbf{z}^{(k-1)}\| + \langle \mathbb{E}[\boldsymbol{\zeta}^{(k)}] - \boldsymbol{\zeta}^{(k)}, \mathbf{z}^{(k)}-\mathbf{z}^{(k-1)}\rangle\end{aligned} \tag{81}$$

Since $\ddot{\phi}(\cdot)$ is L-smooth, by the descent lemma, e.g., Proposition A.24 in [2], we have

$$\ddot{\phi}(\mathbf{z}^{(k)}) \leq \ddot{\phi}(\mathbf{z}^{(k-1)}) + \langle \nabla\ddot{\phi}(\mathbf{z}^{(k-1)}), \mathbf{z}^{(k)}-\mathbf{z}^{(k-1)}\rangle + \frac{L}{2}\|\mathbf{z}^{(k)}-\mathbf{z}^{(k-1)}\|^2 \tag{82}$$

By Young's inequality, for any $\eta^{(k)} > 0$, we have

$$\langle \mathbb{E}[\boldsymbol{\zeta}^{(k)}] - \boldsymbol{\zeta}^{(k)}, \mathbf{z}^{(k)}-\mathbf{z}^{(k-1)}\rangle \leq \frac{1}{2\eta^{(k)}}\|\mathbb{E}[\boldsymbol{\zeta}^{(k)}] - \boldsymbol{\zeta}^{(k)}\|^2 + \frac{\eta^{(k)}}{2}\|\mathbf{z}^{(k)}-\mathbf{z}^{(k-1)}\|^2 \tag{83}$$

Substituting (79), (82) and (83) into (81) and recalling that $\phi(\mathbf{z}^{(k)}) = \dot{\phi}(\mathbf{z}^{(k)}) + \ddot{\phi}(\mathbf{z}^{(k)})$ yields

$$\begin{aligned}\phi(\mathbf{z}^{(k)}) \leq&\phi(\mathbf{z}^{(k-1)}) + \langle \boldsymbol{\zeta}^{(k)}, \mathbf{z}-\mathbf{z}^{(k-1)}\rangle + \frac{1}{2\gamma^{(k)}}\|\mathbf{z}-\mathbf{z}^{(k-1)}\|^2 - \frac{1}{2\gamma^{(k)}}\|\mathbf{z}-\mathbf{z}^{(k)}\|^2\\ &- \frac{1}{2}(\frac{1}{\gamma^{(k)}} - L - \eta^{(k)})\|\mathbf{z}^{(k)}-\mathbf{z}^{(k-1)}\|^2 + M\|\mathbf{z}^{(k)}-\mathbf{z}^{(k-1)}\| + \frac{1}{2\eta^{(k)}}\|\mathbb{E}[\boldsymbol{\zeta}^{(k)}] - \boldsymbol{\zeta}^{(k)}\|^2\end{aligned} \tag{84}$$

Recall that $L = \mu_2$ and $\mu = \mu_1 + \mu_2$. If we take $\gamma^{(k)} = \frac{2}{\mu(k+k_0)}$ with $k_0 = 2$, $\eta^{(k)} = \frac{\mu}{4}k$, then $\frac{1}{\gamma^{(k)}} - L - \eta^{(k)} = \frac{\mu}{4}k + \mu_1 \geq \frac{\mu}{4}k$. Thus, we have

$$\begin{aligned}&- \frac{1}{2}(\frac{1}{\gamma^{(k)}} - L - \eta^{(k)})\|\mathbf{z}^{(k)}-\mathbf{z}^{(k-1)}\|^2 + M\|\mathbf{z}^{(k)}-\mathbf{z}^{(k-1)}\|\\ =&-\frac{\mu}{8}k\|\mathbf{z}^{(k)}-\mathbf{z}^{(k-1)}\|^2 + M\|\mathbf{z}^{(k)}-\mathbf{z}^{(k-1)}\|\\ \overset{(a)}{\leq}&\frac{2M^2}{k\mu}\end{aligned} \tag{85}$$

where (a) follows from the basic inequality $-au^2+bu \le b^2/(4a)$ for any $a<0$ and $u\in\mathbb{R}$.

Substituting (85) into (84) yields

$$\begin{aligned}\phi(\mathbf{z}^{(k)}) \le& \phi(\mathbf{z}^{(k-1)}) + \langle \boldsymbol{\zeta}^{(k)}, \mathbf{z}-\mathbf{z}^{(k-1)}\rangle + \frac{1}{2\gamma^{(k)}}\|\mathbf{z}-\mathbf{z}^{(k-1)}\|^2 - \frac{1}{2\gamma^{(k)}}\|\mathbf{z}-\mathbf{z}^{(k)}\|^2 + \frac{2M^2}{k\mu}\\ &+ \frac{1}{2\eta^{(k)}}\|\mathbb{E}[\boldsymbol{\zeta}^{(k)}]-\boldsymbol{\zeta}^{(k)}\|^2 \end{aligned} \tag{86}$$

For any fixed $\mathbf{z}$, since $\boldsymbol{\zeta}^{(k)}$ is an unbiased i.i.d. stochastic subgradient and $\mathbf{z}^{(k-1)}$ is determined by $\boldsymbol{\zeta}^{(0)},\ldots,\boldsymbol{\zeta}^{(k-1)}$, we have

$$\mathbb{E}[\langle \boldsymbol{\zeta}^{(k)}, \mathbf{z}-\mathbf{z}^{(k-1)}\rangle] = \mathbb{E}[\mathbb{E}[\langle \boldsymbol{\zeta}^{(k)}, \mathbf{z}-\mathbf{z}^{(k-1)}\rangle | \boldsymbol{\zeta}^{[0:k-1]}]] = \mathbb{E}[\langle \mathbf{d}+\nabla\ddot{\phi}(\mathbf{z}^{(k-1)}), \mathbf{z}-\mathbf{z}^{(k-1)}\rangle] \tag{87}$$

By the basic probability fact, we have

$$\mathbb{E}[\|\mathbb{E}[\boldsymbol{\zeta}^{(k)}]-\boldsymbol{\zeta}^{(k)}\|^2] \le \mathbb{E}[\|\boldsymbol{\zeta}^{(k)}\|^2] \le B^2 \tag{88}$$

By the μ-convexity of $\phi(\cdot)$, we have

$$\phi(\mathbf{z}^{(k-1)}) + \langle \mathbf{d}+\nabla\ddot{\phi}(\mathbf{z}^{(k-1)}), \mathbf{z}-\mathbf{z}^{(k-1)}\rangle \le \phi(\mathbf{z}) - \frac{\mu}{2}\|\mathbf{z}-\mathbf{z}^{(k-1)}\|^2 \tag{89}$$

Taking expectations on both sides of (86) and substituting (87)-(89) into it yields

$$\begin{aligned}\mathbb{E}[\phi(\mathbf{z}^{(k)})] \le& \phi(\mathbf{z}) + \frac{1}{2}(\frac{1}{\gamma^{(k)}}-\mu)\mathbb{E}[\|\mathbf{z}-\mathbf{z}^{(k-1)}\|^2] - \frac{1}{2}\frac{1}{\gamma^{(k)}}\mathbb{E}[\|\mathbf{z}-\mathbf{z}^{(k)}\|^2] + \frac{2M^2}{k\mu} + \frac{1}{2\eta^{(k)}}B^2\\ =& \phi(\mathbf{z}) + \frac{\mu}{4}(k+k_0-2)\mathbb{E}[\|\mathbf{z}-\mathbf{z}^{(k-1)}\|^2] - \frac{\mu}{4}(k+k_0)\mathbb{E}[\|\mathbf{z}-\mathbf{z}^{(k)}\|^2] + \frac{2(B^2+M^2)}{k\mu}\end{aligned} \tag{90}$$

Multiplying both sides by $k+k_0-1$ yields

$$\begin{aligned}&(k+k_0-1)\mathbb{E}[\phi(\mathbf{z}^{(k)})]\\ \le& (k+k_0-1)\phi(\mathbf{z}) + \frac{\mu}{4}(k+k_0-2)(k+k_0-1)\mathbb{E}[\|\mathbf{z}-\mathbf{z}^{(k-1)}\|^2]\\ &- \frac{\mu}{4}(k+k_0-1)(k+k_0)\mathbb{E}[\|\mathbf{z}-\mathbf{z}^{(k)}\|^2] + \frac{k+k_0-1}{k\mu}(2(B^2+M^2))\\ \overset{(a)}{\le}& (k+k_0-1)\phi(\mathbf{z}) + \frac{\mu}{4}(k+k_0-2)(k+k_0-1)\mathbb{E}[\|\mathbf{z}-\mathbf{z}^{(k-1)}\|^2]\\ &- \frac{\mu}{4}(k+k_0-1)(k+k_0)\mathbb{E}[\|\mathbf{z}-\mathbf{z}^{(k)}\|^2] + \frac{4(B^2+M^2)}{\mu}\end{aligned}$$

where (a) follows by recalling $k_0=2$. Summing over $k\in\{1,2,\ldots,K\}$ and dividing both sides by $\sum_{k=1}^K(k+k_0-1)$ yields

$$\begin{aligned}&\mathbb{E}[\frac{1}{\sum_{k=1}^K(k+k_0-1)}\sum_{k=1}^K(k+k_0-1)\phi(\mathbf{z}^{(k)})]\\ \le& \phi(\mathbf{z}) + \frac{\mu(k_0^2-k_0)}{2K(K+2k_0-1)}\mathbb{E}[\|\mathbf{z}-\mathbf{z}^{(0)}\|^2] - \frac{\mu(k_0^2-k_0)}{2K(K+2k_0-1)}\mathbb{E}[\|\mathbf{z}-\mathbf{z}^{(K)}\|^2] - \frac{\mu}{2}\mathbb{E}[\|\mathbf{z}-\mathbf{z}^{(K)}\|^2]\\ &+ \frac{8(B^2+M^2)}{(K+2k_0-1)\mu}\end{aligned}$$

Define $\widehat{\mathbf{z}} \triangleq \frac{1}{\sum_{k=1}^K(k+k_0-1)}(k+k_0-1)\mathbf{z}^{(k)}$. By Jensen's inequality and recalling $k_0=2$, we have

$$\begin{aligned}&\mathbb{E}[\phi(\widehat{\mathbf{z}})]\\ \le& \phi(\mathbf{z}) + \frac{\mu}{K(K+3)}\mathbb{E}[\|\mathbf{z}-\mathbf{z}^{(0)}\|^2] - \frac{\mu}{K(K+3)}\mathbb{E}[\|\mathbf{z}-\mathbf{z}^{(K)}\|^2] - \frac{\mu}{2}\mathbb{E}[\|\mathbf{z}-\mathbf{z}^{(K)}\|^2] + \frac{8(B^2+M^2)}{(K+3)\mu}\end{aligned}$$

□

Lemma 15. *Consider convex program* (1) *with μ-convex stochastic objective functions under Assumption 4. Let $\mathbf{x}^*$ be any optimal solution. If $\nu^{(t)} > 0$ and $\rho^{(t)} > 0$ in Algorithm 1 are chosen to satisfy*

$$\nu^{(t)} \geq \rho^{(t)} \|\mathbf{A}\|^2, \forall t,$$

and the sub-procedure STO-LOCAL (Algorithm 2) uses $k_0 = 2$ and $\widehat{\mathbf{z}}$ defined in Lemma 14 as the output, then, for all $T \geq 1$, Algorithm 1 ensures

$$\sum_{t=1}^{T} \mathbb{E}[\rho^{(t)} f(\mathbf{x}^{(t)})] \leq \sum_{t=1}^{T} \mathbb{E}[\rho^{(t)} f(\mathbf{x}^*)] + \frac{1}{2} \sum_{t=1}^{T} \mathbb{E}[\rho^{(t)} \Phi^{(t)}] - \frac{1}{2} \mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|^2] + \sum_{t=1}^{T} \frac{8\rho^{(t)}(B^2 + M^2)}{(\nu^{(t)} + \mu)(K^{(t)} + 3)}$$

where

$$\Phi^{(t)} \triangleq \left(\nu^{(t)} + \frac{2(\nu^{(t)} + \mu)}{K^t(K^t + 3)} \right) \|\mathbf{x}^* - \mathbf{y}^{(t-1)}\|^2 - \left(\nu^{(t)} + \mu + \frac{2(\nu^{(t)} + \mu)}{K^t(K^t + 3)} \right) \|\mathbf{x}^* - \mathbf{y}^{(t)}\|^2 \tag{91}$$

and B and M are constants defined in Assumption 4.

Proof. Fix $t \in \{1, 2, \ldots, T\}$. Define $\phi^{(t)}(\mathbf{x}) = \sum_{i=1}^{N} \phi_i^{(t)}(\mathbf{x}_i)$. Note that if we define $\dot{\phi}^{(t)}(\mathbf{x}) = f(\mathbf{x})$ and $\ddot{\phi}^t(\mathbf{x}) = \rho^{(t)} \langle \mathbf{r}^{t-1} + \frac{1}{\rho^{(t)}} \boldsymbol{\lambda}^{(t-1)}, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle + \frac{\nu^{(t)}}{2} \|\mathbf{x} - \mathbf{y}^{(t-1)}\|^2$, then $\phi^{(t)}(\mathbf{x}) = \dot{\phi}^{(t)}(\mathbf{x}) + \ddot{\phi}^t(\mathbf{x})$ where $\dot{\phi}^{(t)}(\mathbf{x})$ is μ-convex and satisfies (79) by Assumption 4 and $\ddot{\phi}^{(t)}(\mathbf{x})$ is $\nu^{(t)}$-convex and $\nu^{(t)}$-smooth. As in the observation in the proof of Lemma 11 or Lemma 13 , each iteration of Algorithm 1 is to jointly update $\mathbf{x}$ and $\mathbf{y}$ via the sub-procedure $(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}) = \text{STO-LOCAL}(\phi^{(t)}(\cdot), \mathcal{X}, \mathbf{y}^{(t-1)}, K^{(t)})$.

Thus, by Lemma 14, we have

$$\begin{aligned} \mathbb{E}[\phi^{(t)}(\mathbf{x}^{(t)})] \leq & \phi^{(t)}(\mathbf{x}^*) + \frac{(\nu^{(t)} + \mu)}{K(K+3)} \mathbb{E}[\|\mathbf{z} - \mathbf{z}^{(0)}\|^2] - \frac{(\nu^{(t)} + \mu)}{K(K+3)} \mathbb{E}[\|\mathbf{z} - \mathbf{z}^{(K)}\|^2] \\ & - \frac{(\nu^{(t)} + \mu)}{2} \mathbb{E}[\|\mathbf{z} - \mathbf{z}^{(K)}\|^2] + \frac{8(B^2 + M^2)}{(K+3)(\nu^{(t)} + \mu)} \end{aligned}$$

This is almost identical to (62) (with $k_0 = 2$) in Lemma 13 except the constant σ^2 is replaced by $2(B^2 + M^2)$.

Following the same lines (after (62)) in the proof of Lemma 13, we can show

$$\sum_{t=1}^{T} \mathbb{E}[\rho^{(t)} f(\mathbf{x}^{(t)})] \leq \sum_{t=1}^{T} \mathbb{E}[\rho^{(t)} f(\mathbf{x}^*)] + \frac{1}{2} \sum_{t=1}^{T} \mathbb{E}[\rho^{(t)} \Phi^{(t)}] - \frac{1}{2} \mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|^2] + \sum_{t=1}^{T} \frac{8\rho^{(t)}(B^2 + M^2)}{(\nu^{(t)} + \mu)(K^{(t)} + 3)}$$

where $\Phi^{(t)} \triangleq \left(\nu^{(t)} + \frac{2(\nu^{(t)} + \mu)}{K^t(K^t + 3)} \right) \|\mathbf{x}^* - \mathbf{y}^{(t-1)}\|^2 - \left(\nu^{(t)} + \mu + \frac{2(\nu^{(t)} + \mu)}{K^t(K^t + 3)} \right) \|\mathbf{x}^* - \mathbf{y}^{(t)}\|^2$.

Since the conclusion from Lemma 15 is quite similar to that from Lemma 13 (with $k_0 = 2$) with the minor distinction that constant σ^2 is replaced by $2(B^2 + M^2)$, the main proof of Theorem 4 is similar to the proof of $\mu > 0$ case of Theorem 2.

Main Proof of Theorem 4: Note that our selection of $\rho^{(t)} = t\rho$, $\nu^{(t)} = t\rho \|\mathbf{A}\|^2$ and $k_0 = 2$ satisfies the conditions in Lemma 15. By Lemma 15, we have

$$\sum_{t=1}^{T} \mathbb{E}[\rho^{(t)} f(\mathbf{x}^{(t)})] \leq \sum_{t=1}^{T} \mathbb{E}[\rho^{(t)} f(\mathbf{x}^*)] + \frac{1}{2} \sum_{t=1}^{T} \mathbb{E}[\rho^{(t)} \Phi^{(t)}] - \frac{1}{2} \mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|^2] + \sum_{t=1}^{T} \frac{8\rho^{(t)}(B^2 + M^2)}{(\nu^{(t)} + \mu)(K^{(t)} + 3)} \tag{92}$$

Recalling the definition of Φ^t in (91) and $K^{(t)} = 3t$, we have

$$\begin{aligned}
&\sum_{t=1}^{T} \rho^{(t)} \Phi^{(t)} \\
=&\rho \left(\rho \|\mathbf{A}\|^2 + \frac{(\rho\|\mathbf{A}\|^2 + \mu)}{18} \right) \|\mathbf{x}^* - \mathbf{y}^{(0)}\|^2 \\
&- \sum_{t=1}^{T-1} \left(\rho\Big(\rho t^2 \|\mathbf{A}\|^2 + t\mu - \rho(t+1)^2 \|\mathbf{A}\|^2\Big) + \rho\Big(\frac{\rho t \|\mathbf{A}\|^2 + \mu}{t+1} - \frac{\rho(t+1)\|\mathbf{A}\|^2 + \mu}{t+2}\Big)\frac{2}{9} \right) \|\mathbf{x}^* - \mathbf{y}^{(t)}\|^2 \\
&- T\rho \left(T\rho \|\mathbf{A}\|^2 + \mu + \Big(\frac{\rho T \|\mathbf{A}\|^2 + \mu}{T(T+1)}\Big)\frac{2}{9} \right) \|\mathbf{x}^* - \mathbf{y}^{T}\|^2 \\
\overset{(a)}{\leq} &\rho \left(\rho \|\mathbf{A}\|^2 + \frac{2(\rho\|\mathbf{A}\|^2 + \mu)}{18} \right) \|\mathbf{x}^* - \mathbf{y}^{(0)}\|^2
\end{aligned} \tag{93}$$

where (a) follows by ignoring the last negative term and noting that $\rho t^2 \|\mathbf{A}\|^2 + t\mu - \rho(t+1)^2\|\mathbf{A}\|^2 = t\mu - \rho(2t+1)\|\mathbf{A}\|^2 \geq \mu(t - \frac{2t+1}{3}) \geq 0$ for all $t \geq 1$, where the first inequality uses $\rho \leq \frac{\mu}{3\|\mathbf{A}\|^2}$; and $\frac{\rho t\|\mathbf{A}\|^2+\mu}{t+1} - \frac{\rho(t+1)\|\mathbf{A}\|^2+\mu}{t+2} = \frac{\mu-\rho\|\mathbf{A}\|^2}{(t+1)(t+2)} \geq 0$, where the inequality also uses $\rho \leq \frac{\mu}{3\|\mathbf{A}\|^2}$.

We also note that

$$\begin{aligned}
\sum_{t=1}^{T} \frac{8\rho^{(t)}(B^2+M^2)}{(\nu^{(t)}+\mu)(K^{(t)}+3)} =& \frac{8\rho(B^2+M^2)}{3} \sum_{t=1}^{T} \frac{t}{(\rho t\|\mathbf{A}\|^2 + \mu)(t+1)} \\
\overset{(a)}{\leq}& \frac{8\rho(B^2+M^2)}{3} \sum_{t=1}^{T} \frac{t}{(t+1)(t+3)\|\mathbf{A}\|^2} \\
\leq& \frac{8\rho(B^2+M^2)}{3\|\mathbf{A}\|^2} \sum_{t=1}^{T} \frac{1}{t+1} \\
\leq& \frac{8\rho(B^2+M^2)}{3\|\mathbf{A}\|^2} \log(T+1)
\end{aligned} \tag{94}$$

where (a) follows because $\mu \geq 3\rho\|\mathbf{A}\|^2$ by $\rho \leq \frac{\mu}{3\|\mathbf{A}\|^2}$.

Substituting (93) and (94) into (92) yields

$$\begin{aligned}
&\sum_{t=1}^{T} \mathbb{E}[\rho^{(t)} f(\mathbf{x}^{(t)})] \\
\leq &\sum_{t=1}^{T} \mathbb{E}[\rho^{(t)} f(\mathbf{x}^*)] + \frac{1}{2}\rho \left(\rho\|\mathbf{A}\|^2 + \frac{2(\rho\|\mathbf{A}\|^2+\mu)}{18} \right) \|\mathbf{x}^* - \mathbf{y}^{(0)}\|^2 - \frac{1}{2}\mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|^2] \\
&+ \frac{8\rho(B^2+M^2)}{\|\mathbf{A}\|^2} \log(T+1) \\
\overset{(a)}{\leq} &\sum_{t=1}^{T} \mathbb{E}[\rho^{(t)} f(\mathbf{x}^*)] + \frac{1}{2}\rho c_1 \|\mathbf{x}^* - \mathbf{y}^{(0)}\|^2 - \frac{1}{2}\mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|^2] + \frac{1}{2} c_2 \log(T+1)
\end{aligned} \tag{95}$$

where (a) follows because $c_1 = \rho\|\mathbf{A}\|^2 + \frac{2(\rho\|\mathbf{A}\|^2+\mu)}{18}$ and $c_2 = \frac{16\rho(B^2+M^2)}{3\|\mathbf{A}\|^2}$.

Ignoring the negative term $-\frac{1}{2}\mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|^2]$, dividing both sides by $\sum_{t=1}^{T} \rho^{(t)}$ and applying Jensen's inequality yields

$$\begin{aligned}
\mathbb{E}[f(\overline{x}^T)] \leq& f(\mathbf{x}^*) + \frac{1}{2\sum_{t=1}^{T} \rho^{(t)}} \Big(\rho c_1 \|\mathbf{x}^* - \mathbf{y}^{(0)}\|^2 + c_2 \log(T+1) \Big) \\
\overset{(a)}{=}& f(\mathbf{x}^*) + \frac{1}{T(T+1)} \left(c_1 \|\mathbf{x}^* - \mathbf{y}^{(0)}\|^2 + \frac{c_2}{\rho} \log(T+1) \right)
\end{aligned}$$

where (a) follows because $\sum_{t=1}^{T} \rho^{(t)} = \rho \sum_{t=1}^{T} t = \frac{\rho T(T+1)}{2}$. This is (77) of our theorem.

By Lemma 4 (after taking expectations on both sides), we have

$$\mathbb{E}[\sum_{t=1}^{T} \rho^{(t)} f(\mathbf{x}^{(t)})] \geq \sum_{t=1}^{T} \mathbb{E}[\rho^{(t)} f(\mathbf{x}^*)] - \|\boldsymbol{\lambda}^*\| \mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|]$$

Combining this inequality with (75) and cancelling the common terms yields

$$\mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|^2] \leq 2\|\boldsymbol{\lambda}^*\| \mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|] + \rho c_1 \|\mathbf{x}^* - \mathbf{y}^{(0)}\|^2 + c_2 \log(T+1)$$

Since $\left(\mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|]\right)^2 \leq \mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|^2]$, we further have

$$\left(\mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|]\right)^2 \leq 2\|\boldsymbol{\lambda}^*\| \mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|] + \rho c_1 \|\mathbf{x}^* - \mathbf{y}^{(0)}\|^2 + c_2 \log(T+1)$$

This quadratic inequality can be rewritten as

$$\left(\mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|] - \|\boldsymbol{\lambda}^*\|\right)^2 \leq \|\boldsymbol{\lambda}^*\|^2 + \rho c_1 \|\mathbf{x}^* - \mathbf{y}^{(0)}\|^2 + c_2 \log(T+1)$$

Thus, we have

$$\begin{aligned} \mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|] \leq & \|\boldsymbol{\lambda}^*\| + \sqrt{\|\boldsymbol{\lambda}^*\|^2 + \rho c_1 \|\mathbf{x}^* - \mathbf{y}^{(0)}\|^2 + c_2 \log(T+1)} \\ \leq & 2\|\boldsymbol{\lambda}^*\| + \sqrt{\rho c_1} \|\mathbf{x}^* - \mathbf{y}^{(0)}\| + \sqrt{c_2 \log(T+1)} \end{aligned} \tag{96}$$

By part (1) of Lemma 3 (with $\rho^{(t)} = \rho$), we have

$$\sum_{t=1}^{T} \rho \left(\mathbf{A}\mathbf{x}^{(t)} - \mathbf{b}\right) = \boldsymbol{\lambda}^{(T)}$$

Dividing both sides by $\sum_{t=1}^{T} \rho^{(t)} = \rho \frac{T(T+1)}{2}$, taking the vector l_2 norm and then taking expectations on both sides yields

$$\begin{aligned} \mathbb{E}[\|\mathbf{A}\bar{\mathbf{x}}^{(T)} - \mathbf{b}\|] = & \frac{2}{T(T+1)} \mathbb{E}[\|\boldsymbol{\lambda}^{(T)}\|] \\ \overset{(a)}{\leq} & \frac{2}{T(T+1)} \left(\frac{4\|\boldsymbol{\lambda}^*\|}{\rho} + \frac{\sqrt{c_1}}{\sqrt{\rho}} \|\mathbf{x}^* - \mathbf{y}^{(0)}\| + \frac{\sqrt{c_2 \log(T+1)}}{\rho} \right) \end{aligned}$$

where (a) follows from (76). This is (78) of our theorem.

□