# Stochastic Smoothed Primal-Dual Algorithms for Nonconvex Optimization with Linear Inequality Constraints

Ruichuan Huang<sup>\*</sup> Jia

Jiawei Zhang<sup>†¶</sup>

Ahmet Alacaoglu<sup>‡¶</sup>

#### Abstract

We propose smoothed primal-dual algorithms for solving stochastic and smooth nonconvex optimization problems with linear inequality constraints. Our algorithms are single-loop and only require a single stochastic gradient based on one sample at each iteration. A distinguishing feature of our algorithm is that it is based on an inexact gradient descent framework for the Moreau envelope, where the gradient of the Moreau envelope is estimated using one step of a stochastic primal-dual augmented Lagrangian method. To handle inequality constraints and stochasticity, we combine the recently established global error bounds in constrained optimization with a Moreau envelope-based analysis of stochastic proximal algorithms. For obtaining  $\varepsilon$ -stationary points, we establish the optimal  $O(\varepsilon^{-4})$  sample complexity guarantee for our algorithms and provide extensions to stochastic linear constraints. We also show how to improve this complexity to  $O(\varepsilon^{-3})$  by using variance reduction and the expected smoothness assumption. Unlike existing methods, the iterations of our algorithms are free of subproblems, large batch sizes or increasing penalty parameters and use dual variable updates to ensure feasibility.

### 1 Introduction

We focus on the problem template

$$\min_{\mathbf{x}\in X} f(\mathbf{x}) \text{ subject to } A\mathbf{x} = \mathbf{b}, \tag{1.1}$$

where  $f : \mathbb{R}^n \to \mathbb{R}$  is  $L_f$ -smooth, the set  $X \subseteq \mathbb{R}^n$  is polyhedral, and easy to project. In particular, let X be given as  $X = \{\mathbf{x} : H\mathbf{x} \leq h\}$  for some matrix H and vector h. Taking H = I, for example, gives this template the ability to model *linear inequality* constraints. In particular, when we have the problem

$$\min_{\mathbf{x}} f(\mathbf{x}) \text{ subject to } A\mathbf{x} \le \mathbf{b}, \tag{1.2}$$

we introduce a slack variable  $\mathbf{t} = A\mathbf{x} - \mathbf{b}$  so that  $A\mathbf{x} - \mathbf{t} = \mathbf{b}$  and our optimization variable becomes  $\binom{\mathbf{x}}{\mathbf{t}}$  and we can equivalently write the problem in the template (1.1) by using the constraint  $\mathbf{t} \leq 0$ . As such, we focus on (1.1) and our results directly apply to solving (1.2) by using this standard slack variable reformulation.

Throughout, we assume that we have access to an unbiased oracle  $F(\mathbf{x})$  such that

$$\mathbb{E}[F(\mathbf{x})] = \nabla f(\mathbf{x}), \text{ and } \mathbb{E} \|F(\mathbf{x}) - \nabla f(\mathbf{x})\|^2 \le \sigma^2.$$
(1.3)

A common setting is when  $f(\mathbf{x}) = \mathbb{E}_{\xi \sim \Xi}[f(\mathbf{x}, \xi)]$  where  $\Xi$  is an unknown distribution that we can draw i.i.d. samples from. In this case, it is common to set  $F(\mathbf{x}) = \nabla f(\mathbf{x}, \xi)$  and assume that  $\mathbb{E}[\nabla f(\mathbf{x}, \xi)] = \nabla f(\mathbf{x})$ .

Inclusion of X in the template (1.1) increases the modeling power significantly, while causing difficulties in the analysis. Many problems fit this template, including constrained and distributed optimization, nonnegative matrix factorization, sparse subspace estimation and collaborative learning, see for example

<sup>\*</sup>Department of Mathematics, University of British Columbia. hrc22@student.ubc.ca

<sup>&</sup>lt;sup>†</sup>Laboratory of Information and Decision Systems, Massachusetts Institute of Technology. jwzhang@mit.edu

<sup>&</sup>lt;sup>‡</sup>Department of Mathematics, University of British Columbia. alacaoglu@math.ubc.ca

<sup>&</sup>lt;sup>¶</sup>Co-last authors and corresponding authors.

Zhang et al. (2022); Hong (2016) and also Section 6. Moreover, reformulations of nonconvex minimization problems are also common by using linear inequality constraints (Zhang et al., 2022).

Algorithm development for (1.1) and related templates have been active in the last couple of years (Alacaoglu & Wright, 2024; Zhang & Luo, 2020; Zhang et al., 2020; Lu et al., 2024; Li et al., 2021; Lin et al., 2022; Yan & Xu, 2022; Li et al., 2024; Boob et al., 2023; Hong, 2016), mainly due to the applications of functionally constrained nonconvex optimization problems in the context of neural network training (Katz-Samuels et al., 2022; Dener et al., 2020). Stochastic augmented Lagrangian methods (ALM) have found widespread use in practice with problems involving nonconvex functional constraints (Katz-Samuels et al., 2022; Dener et al., 2020), whereas their behavior for even linearly constrained nonconvex optimization of the form (1.1) remain poorly understood. The focus of this work is to improve our understanding of stochastic ALM in the context of nonconvex optimization, by focusing on the fundamental template (1.1).

Compared to the setting of convex f, where the global complexity analysis is mostly settled for ALM and its stochastic version (Yan & Xu, 2022), nonconvexity of f poses significant difficulties in the analysis of ALM. Many works in the literature focus on penalty based algorithms (which will be formally introduced later in this section) that do not perform dual updates (or perform negligible dual updates that we clarify later) (Lu et al., 2024; Li et al., 2021; Lin et al., 2022), rather than primal-dual algorithms such as ALM. However, in practice, dual updates are known to be essential for accelerating convergence. Penalty methods are known to be unstable since increasing penalty parameter causes Lipschitz constant of the subproblems to increase and can lead to numerical issues. These differences in behavior between penalty and augmented Lagrangian methods are well-known, see for example classical books, such as (Bertsekas, 2014, 2016; Nocedal & Wright, 1999).

For problem (1.1) with access to full gradients of f and the full matrix A, the optimal complexity with primal-dual methods are obtained in the work of Zhang & Luo (2022). When one has access to stochastic gradients of f and the matrix A, a recent work by Alacaoglu & Wright (2024) showed optimal complexity guarantees under expected smoothness (see Assumption 5.2), for the special case of (1.1) when  $X = \mathbb{R}^n$ , where this latter restriction significantly reduces the generality of the template. For example, modeling the standard quadratic programming problem requires X to be a half-space, which was not supported in the analysis of Alacaoglu & Wright (2024). Our goal is to go beyond these results by handling both the case when  $X \neq \mathbb{R}^n$  as well as the case when we do not have access to the matrix A but only to an unbiased estimate of A, by keeping optimal complexity guarantees. A more detailed comparison of complexity guarantees will be made in Section 7 and a summary is provided in Table 1.

Lagrangian, penalty and augmented Lagrangian functions The standard approach to tackle (1.1) is to design algorithms operating on the Lagrangian, augmented Lagrangian or penalty functions. In particular, the Lagrangian function is given as

$$L(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}) + \langle A\mathbf{x} - \mathbf{b}, \mathbf{y} \rangle$$

with the dual variable **y**, whereas the penalty function has the form of

$$\operatorname{Pen}_{\rho}(\mathbf{x}) = f(\mathbf{x}) + \frac{\rho}{2} \|A\mathbf{x} - \mathbf{b}\|^2.$$

It is common for algorithms based on the penalty function to require  $\rho \to \infty$  for convergence (Bertsekas, 2014). One major disadvantage of this strategy is that  $\rho$  getting larger makes the subproblem of minimizing the penalty function more and more ill-conditioned.

An influential idea was the introduction of the augmented Lagrangian (AL) function which combined the idea of the Lagrangian and penalty formulations (Hestenes, 1969). In particular, the AL function is defined as

$$L_{\rho}(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}) + \langle A\mathbf{x} - \mathbf{b}, \mathbf{y} \rangle + \frac{\rho}{2} ||A\mathbf{x} - \mathbf{b}||^2.$$

Augmented Lagrangian methods in the classical literature were favored because these methods worked without requiring  $\rho$  to grow arbitrarily large. In fact, many instances of ALM converge to the optimal solution with fixed  $\rho$  since the inclusion of the dual variable y aids in satisfying feasibility (Bertsekas, 2014). **Primal vs primal-dual algorithms.** The algorithms based on the penalty function are generally referred to as *penalty algorithms* and are easier to analyze in different settings since they are primal-only algorithms, meaning that they only perform updates on primal variable  $\mathbf{x}$  where approximate feasibility is ensured by  $\rho \to \infty$ . In particular, a classical penalty method iterates for  $k = 1, 2, \ldots$  as

$$\mathbf{x}_{k+1} \approx \arg\min_{\mathbf{x}\in X} f(\mathbf{x}) + \frac{\rho_k}{2} \|A\mathbf{x} - \mathbf{b}\|^2,$$
  
Select  $\rho_{k+1} > \rho_k.$ 

The algorithms based on the augmented Lagrangian are generally more difficult to analyze due to the additional dynamics coming from the dual updates where the dual updates are critical to ensure that the approximate feasibility is attained with constant  $\rho$ . An ALM iteration proceeds for  $k = 1, 2, \ldots$  by updating

$$\mathbf{x}_{k+1} \approx \arg\min_{\mathbf{x}\in X} f(\mathbf{x}) + \langle \mathbf{y}_k, A\mathbf{x} - \mathbf{b} \rangle + \frac{\rho}{2} \|A\mathbf{x} - \mathbf{b}\|^2,$$
$$\mathbf{y}_{k+1} = \mathbf{y}_k + \sigma(A\mathbf{x}_{k+1} - \mathbf{b}).$$

For both penalty methods and ALM, different strategies exist to generate  $\mathbf{x}_{k+1}$  that approximately minimize the penalty or augmented Lagrangian functions by either iterating multiple steps of gradient descent (GD), known as *inexact* algorithms, or applying one step of GD, known as *linearized* algorithms(Ouyang et al., 2015).

In view of the earlier discussion, when f is nonconvex, most of the literature focuses on either analyzing penalty methods, or analyzing ALM with *negligible* dual updates and increasing penalty parameters  $\rho$ , due to the inherent difficulty in analyzing the dual variable and its effect in convergence. In particular, as also highlighted in Alacaoglu & Wright (2024), many of the recent analysis of ALM is of the form of a *perturbed penalty analysis*, meaning that the feasibility is driven by increasing penalty parameters, and the dual updates are designed so that they do not deteriorate the estimates too much. Because of this, the dual step sizes are selected to be small to ensure boundedness of the dual variable (or controlling the growth of the dual variable). We refer to such updates as *negligible* dual updates since the analyses do not harness the benefit of such updates in ensuring feasibility. Feasibility is driven by large penalty parameters. Some representative examples are Lu et al. (2024), Li et al. (2021), Lin et al. (2022), Li et al. (2024).

This is the case even in the deterministic setting and the only method that we are aware that can handle true ALM with fixed penalty parameters and non-negligible dual updates are due to Zhang & Luo (2022) that uses a linearized *proximal* AL function with a dynamic adjustment on the proximal center, which will be clarified in Section 2 since it will form the basis of our algorithmic development.

#### 1.1 Contributions

In this paper, we propose a stochastic smoothed linearized augmented Lagrangian algorithm for solving (1.1) that only uses a single sample of stochastic gradient at every iteration. This algorithm also works with a constant penalty parameter and incorporates non-negligible dual updates for feasibility where the dual step sizes have the same order as the primal step sizes. We show that this method has its iteration complexity and sample complexity guarantees in the order of  $O(\varepsilon^{-4})$ . Such a sample complexity result is optimal even in the unconstrained nonconvex case under our assumptions (see Assumption 1.1) (Arjevani et al., 2023). In contrast, the prior results with optimal complexity required large penalty parameters, no dual updates and further assumptions (Lu et al., 2024). We then prove that this complexity can be improved to  $O(\varepsilon^{-3})$  with variance reduction when an additional expected smoothness assumption is made (see Assumption 5.2). Under this stronger assumption, this is the optimal complexity even without constraints (Arjevani et al., 2023).

We consider extensions of this framework when we have linear constraints that hold in expectation, that is, when the constraints are given as  $\mathbb{E}_{\xi}[A_{\xi}\mathbf{x} - \mathbf{b}_{\xi}] = 0$ . Our algorithm can also handle this stochastic constrained case with the same complexity guarantees. To our knowledge, this is the first algorithm achieving the optimal  $O(\varepsilon^{-4})$  benchmark sample complexity for nonconvex optimization with stochastic constraints using one sample per iteration, going beyond the best-known  $O(\varepsilon^{-5})$  complexity that is achieved for a more general problem that does not capture the structure of linear constraints (Li et al., 2024; Alacaoglu & Wright, 2024).

Reference	Constraint	Oracle	Complexity	Loops	Method
Alacaoglu & Wright (2024)	$A\mathbf{x} = \mathbf{b}$	Eq. $(1.3)$ and Asmp. 5.2	$\widetilde{O}(\varepsilon^{-3})$	1	ALM
Alacaoglu & Wright (2024)	$\mathbb{E}[c(\mathbf{x}, \zeta)] = 0,$ and $\mathbf{x} \in X$ where X is easy to project	Eq. $(1.3)$ and Asmp. $5.2$	$\widetilde{O}(\varepsilon^{-5})$	1	Penalty
Lu et al. (2024)	$c(\mathbf{x}) = 0,$ and $\mathbf{x} \in X$ where X is easy to project	Eq. $(1.3)$ and Asmp. $5.2$	$O(\varepsilon^{-3})$	1	Penalty
Li et al. (2024)	$\mathbb{E}[c(\mathbf{x}, \zeta)] = 0,$ and $\mathbf{x} \in X$ where X is easy to project	Eq. $(1.3)$ and Asmp. $5.2$	$O(\varepsilon^{-5})$	2	Penalty*
This work	$A\mathbf{x} = \mathbf{b},$ and $\mathbf{x} \in X$ is a polyhedral	Eq. (1.3)	$O(\varepsilon^{-4})$	1	ALM
This work	$\mathbb{E}_{\zeta}[A(\zeta)\mathbf{x} - \mathbf{b}(\zeta)] = 0,$ and $\mathbf{x} \in X$ is a polyhedral	Eq. (1.3)	$O(\varepsilon^{-4})$	1	ALM
This work	$A\mathbf{x} = \mathbf{b},$ and $\mathbf{x} \in X$ is a polyhedral	Eq. $(1.3)$ and Asmp. 5.2	$O(\varepsilon^{-3})$	1	ALM

A more detailed comparison with the related works is given in Section 7. A summary is given in Table 1.

Table 1: Comparison of methods. \*This method is referred to as a penalty method because the penalty parameter is taken to infinity to ensure feasibility and dual updates do not contribute in achieving feasibility.

#### **1.2** Preliminaries

We denote the indicator function of a set X as

$$I_X(\mathbf{x}) = \begin{cases} 0 & \text{if } \mathbf{x} \in X, \\ +\infty & \text{if } \mathbf{x} \notin X. \end{cases}$$

The notation  $\partial f$  for a convex, closed function denotes the subdifferential set and  $\partial I_X(\mathbf{x})$  is the normal cone of X at  $\mathbf{x}$  by definition. For a matrix A, we use ||A|| to denote its operator norm.

Given closed and convex X, we denote the projection onto this set as

$$\operatorname{proj}_X(\mathbf{x}) = \arg\min_{\mathbf{v}\in X} \|\mathbf{x} - \mathbf{v}\|^2$$

Similarly, we define the proximal operator of f as

$$\operatorname{prox}_{f}(\mathbf{x}) = \operatorname{arg\,min}_{\mathbf{v}} f(\mathbf{v}) + \frac{1}{2} \|\mathbf{v} - \mathbf{x}\|^{2}.$$

We say that f is L-smooth when the gradient of f is L-Lipschitz:

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \le L \|\mathbf{x} - \mathbf{y}\|.$$

We say that f is  $\rho$ -weakly convex when  $f + \frac{\rho}{2} \|\cdot\|^2$  is convex. An L-smooth function is automatically L-weakly convex. The Moreau envelope of a weakly convex function f is defined as

$$\varphi_{\lambda}(\mathbf{z}) = \min_{\mathbf{v}} f(\mathbf{v}) + \frac{1}{2\lambda} \|\mathbf{v} - \mathbf{z}\|^2,$$

which can be interpreted as a notion of *smoothing*. Moreau envelope has many useful properties such as being smooth when f is nonsmooth and weakly convex, and when  $\lambda$  is selected accordingly. Moreover, the stationary points of f and the Moreau envelope coincide (Drusvyatskiy & Paquette, 2019, Lemma 4.3).

The gradient of the Moreau envelope is computed as

$$\lambda^{-1}(\mathbf{x} - \operatorname{prox}_{\lambda\varphi}(\mathbf{x})).$$

**Stationary points.** A succinct and standard way of characterizing a stationary point of (1.1) is the following: we call  $\mathbf{x}^*$  to be a stationary point if there exists  $\mathbf{y}^*$  such that the following requirements hold:

$$0 \in \nabla f(\mathbf{x}^{\star}) + A^{\top} \mathbf{y}^{\star} + \partial I_X(\mathbf{x}^{\star}),$$
  
$$0 = A\mathbf{x}^{\star} - \mathbf{b}.$$

One may, for example, refer to Rockafellar (2000).

Accordingly, we say that  $(\mathbf{x}, \mathbf{y})$  is an  $\varepsilon$ -stationary point if we have

$$\|A\mathbf{x} - \mathbf{b}\| \le \varepsilon$$
 and  
 $\|\mathbf{v}\| \le \varepsilon$  where  $\mathbf{v} \in \nabla f(x) + A^{\top}\mathbf{y} + \partial I_X(\mathbf{x})$ ,

which is a common notion used in related works, for example Zhang & Luo (2022).

We also use the following related and weaker notion of *near-stationarity*, as used in Davis & Drusvyatskiy (2019). We say that  $\mathbf{x}$  is  $\varepsilon$ -near stationary, if it satisfies

$$\|\nabla \Psi(\mathbf{x})\| \le \varepsilon,\tag{1.4}$$

where  $\Psi(\mathbf{x})$  is the Moreau envelope of the objective function  $f(\mathbf{x}) + I_X(\mathbf{x}) + I_{\{\mathbf{v}:A\mathbf{v}=\mathbf{b}\}}(\mathbf{x})$  in (1.1). We refer to Davis & Drusvyatskiy (2019) for the precise notion of near stationarity.

#### **1.3** Assumptions

We proceed to state the assumptions that will be used throughout. These assumptions are standard and to our knowledge, the weakest, in the literature for both deterministic and stochastic nonconvex problems with linear constraints (Zhang & Luo, 2022; Alacaoglu & Wright, 2024). A more detailed comparison of assumptions will be made in Section 7.

**Assumption 1.1.** For the problem (1.1), the following holds:

1. The function f is  $L_f$ -smooth and is lower bounded over the feasible set, that is,

$$f(\mathbf{x}) \ge \underline{f} > -\infty,$$

for any  $\mathbf{x} \in X$  and  $A\mathbf{x} = \mathbf{b}$ .

- 2. The set X admits an efficient projection and is polyhedral. That is, it has the form  $X = \{\mathbf{x} : H\mathbf{x} \le h\}$  for some H, h.
- 3. We have access to a stochastic gradient F of f that satisfies (1.3).

## 2 Algorithm

We introduce Algorithm 1 in this section. To gain a deeper understanding of the algorithm, we will go over two different ways of interpreting it. **Interpretation 1: Linearized proximal ALM.** Algorithm 1 incorporates a single-step stochastic gradient descent approximation of the proximal augmented Lagrangian function. This strategy is also known as the linearized proximal ALM. In particular, the first step of the algorithm approximates the proximal AL function, that is,

$$\mathbf{x}_{t+1} \approx \arg\min_{\mathbf{x}\in X} L_{\rho}(\mathbf{x}, \mathbf{y}_{t+1}) + \frac{\lambda}{2} \|\mathbf{x} - \mathbf{z}_t\|^2,$$

by a single step of projected SGD, followed by a dual variable update and updating the proximal center  $\mathbf{z}_t$ , by a combination of  $\mathbf{z}_t$  and  $\mathbf{x}_t$ , resulting in the terminology *smoothed* that we use for the algorithm.

Interpretation 2: Inexact GD on the Moreau envelope. Algorithm 1 can also be interpreted as an inexact gradient descent step on the Moreau envelope of the function in (1.1). In particular, the Moreau envelope of (1.1) is given as

$$\Psi(\mathbf{z}_t) = \min_{\mathbf{x} \in X, A\mathbf{x} = \mathbf{b}} \left\{ f(\mathbf{x}) + \frac{\lambda}{2} \|\mathbf{x} - \mathbf{z}_t\|^2 \right\}.$$
(2.1)

By observing that minimizing the Moreau envelope helps in obtaining a near-stationary point in view of (1.4) (cf. Davis & Drusvyatskiy (2019)), inexact gradient update on this function requires the computation of

$$\underset{\mathbf{x}\in X, A\mathbf{x}=\mathbf{b}}{\operatorname{arg\,min}} \left\{ f(\mathbf{x}) + \frac{\lambda}{2} \|\mathbf{x} - \mathbf{z}_t\|^2 \right\},\$$

which is a nontrivial optimization subproblem. However, it is easier than (1.1) because the regularization (given that  $\lambda$  is larger than  $L_f$ ) provides us a *strongly convex objective* in the subproblem. As a result, we can approximate the solution of this problem by applying one iteration of ALM since this problem is a strongly convex optimization problem over linear constraints. We show that just one step of stochastic ALM is sufficient at every iteration by using a stochastic gradient computed with a single sample and one dual update, followed by the update of the proximal center  $\mathbf{z}_t$ .

On the surface, this algorithm strongly resembles the algorithm of Zhang & Luo (2022) where we draw many ideas. However, in addition to using stochastic gradients, there is another subtle change, on the update of  $\mathbf{z}_{t+1}$ . Unlike Zhang & Luo (2022), we update  $\mathbf{z}_{t+1}$  by using  $\mathbf{x}_t$  to be able to continue the analysis with the bounded variance assumption on G instead of boundedness assumption on G, since the latter would require bounded domains. Thanks to this small change, in this section, we can handle the case where both primal and dual domains are unbounded.

### Algorithm 1 Stochastic smoothed and linearized ALM

**Initialize:**  $\mathbf{x}_0 = \mathbf{z}_0 \in X$ ,  $\mathbf{y}_0 \in \mathbb{R}^m$  and  $\rho \ge 0$ . **for** t = 0 **to** T - 1 **do**   $\mathbf{y}_{t+1} = \mathbf{y}_t + \eta(A\mathbf{x}_t - \mathbf{b})$ Sample  $\xi_t \in \Xi$  i.i.d. and generate  $\mathbb{E}_{\xi_t}[G(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t, \xi_t)] = \nabla_{\mathbf{x}} L_{\rho}(\mathbf{x}_t, \mathbf{y}_{t+1}) + \mu(\mathbf{x}_t - \mathbf{z}_t)$   $\mathbf{x}_{t+1} = \operatorname{proj}_X(\mathbf{x}_t - \tau G(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t, \xi_t))$   $\mathbf{z}_{t+1} = \mathbf{z}_t + \beta(\mathbf{x}_t - \mathbf{z}_t)$ end for

### **3** Convergence Analysis

In this section, we first provide the main complexity results, then introduce the main analysis tools and a proof sketch.

#### 3.1 Main Theorem

In view of the two stationary notions given in Section 1.2, we start with the result showing that Algorithm 1 outputs a point at which the norm of the gradient of Moreau envelope is small, in expectation.

For the result, we state the algorithmic parameters. To avoid clutter, we write the orders of the parameters by highlighting their dependencies on the problem parameters. The explicit forms of the parameters with all the constants are given in (A.1) in Appendix A.

$$\tau \approx \frac{1}{\sqrt{T}}, \quad \eta \approx \frac{1}{\sqrt{T}}, \quad \beta \approx \frac{1}{\sqrt{T}}, \quad \mu \approx L_f, \quad \lambda \approx L_f + \mu(\|A\|^2 + 1). \tag{3.1}$$

We are now ready to state the first main result.

**Theorem 3.1.** Let Assumption 1.1 hold and run Algorithm 1 with the parameters given in (3.1) (see also (A.1)). We have that  $\mathbb{E} \| \nabla \Psi(\mathbf{z}_{t^*}) \| \leq \varepsilon$  where  $t^*$  is selected uniformly at random from  $\{1, \ldots, T\}$  with  $T = \Theta(\varepsilon^{-4})$ . The stochastic oracle complexity is  $O(\varepsilon^{-4})$ .

In particular, the above result gives us an  $\varepsilon$ -near stationary point in view of Davis & Drusvyatskiy (2019).

Next, to get an  $\varepsilon$ -stationary point, we perform a post-processing procedure to obtain the following output from the result of Algorithm 1:

$$\hat{\mathbf{x}} = \operatorname{proj}_X(\mathbf{x}_{t^*} - \tau \widehat{G}(\mathbf{x}_{t^*}, \mathbf{y}_{t^*+1}, \mathbf{z}_{t^*})),$$
(3.2)

with  $\tau \leq \frac{1}{L_K}$  where  $L_K$  is the Lipschitz constant of  $L_{\rho}(\cdot, \mathbf{y}, \mathbf{z}) + \frac{\lambda}{2} \| \cdot -\mathbf{x} \|^2$  (cf. (A.1)) and

$$\hat{G}(\mathbf{x}_{t^*}, \mathbf{y}_{t^*+1}, \mathbf{z}_{t^*}) = \frac{1}{B} \sum_{i=1}^{B} G(\mathbf{x}_{t^*}, \mathbf{y}_{t^*+1}, \mathbf{z}_{t^*}, \xi_i)$$

for  $\xi_i$  i.i.d. and  $B = \Omega(\varepsilon^{-2})$ . We note that this is the only place where we use a large batch size and our algorithm only runs with a single sample at every iteration. This post processing step is only done once and does not affect the overall complexity. The details are given in Appendix A.3.

**Corollary 3.2.** Let Assumption 1 hold. From the output of Algorithm 1, we can obtain an output  $\hat{\mathbf{x}}$  which is an  $\varepsilon$ -stationary point. The complexity of the whole procedure is  $O(\varepsilon^{-4})$ .

#### 3.2 Analysis Tools

In our analysis, Moreau envelopes of two functions are critical. The first was the Moreau envelope of the composite objective in (1.1), defined in (2.1). We next define the Moreau envelope on the proximal AL function which is the main function to analyze projected SGD update (cf. Davis & Drusvyatskiy (2019)):

$$\varphi_{1/\lambda}(\mathbf{x}, \mathbf{y}, \mathbf{z}) = \min_{\mathbf{u} \in X} \left\{ L_{\rho}(\mathbf{u}, \mathbf{y}) + \frac{\mu}{2} \|\mathbf{u} - \mathbf{z}\|^2 + \frac{\lambda}{2} \|\mathbf{u} - \mathbf{x}\|^2 \right\}.$$
(3.3)

Another important quantity that has a significant role in the analysis is the proximal point with respect to the last definition.

$$\mathbf{u}^*(\mathbf{x}, \mathbf{y}, \mathbf{z}) = \operatorname*{arg\,min}_{\mathbf{u} \in X} L_{\rho}(\mathbf{u}, \mathbf{y}) + \frac{\mu}{2} \|\mathbf{u} - \mathbf{z}\|^2 + \frac{\lambda}{2} \|\mathbf{u} - \mathbf{x}\|^2.$$
(3.4)

With this, we have

$$\varphi_{1/\lambda}(\mathbf{x},\mathbf{y},\mathbf{z}) = L_{\rho}(\mathbf{u}^*(\mathbf{x},\mathbf{y},\mathbf{z}),\mathbf{y}) + \frac{\mu}{2} \|\mathbf{u}^*(\mathbf{x},\mathbf{y},\mathbf{z}) - \mathbf{z}\|^2 + \frac{\lambda}{2} \|\mathbf{u}^*(\mathbf{x},\mathbf{y},\mathbf{z}) - \mathbf{x}\|^2.$$

Note that this is the main point of departure from Zhang & Luo (2022) where the proximal AL function is used in the analysis and the potential function. This is because Zhang & Luo (2022) used a projected *full* 

GD step on the proximal AL function for which, a descent inequality follows directly. In our case, because we apply a projected SGD step, to be able to handle updates with single-sample stochastic gradients, we need to use the Moreau envelope of the proximal AL function in our potential. This analysis of projected SGD was pioneered in Davis & Drusvyatskiy (2019).

The first result is a descent-type result on the Moreau envelope.

**Lemma 3.3.** (cf. Lemma A.2) Let Assumption 1.1 hold and set  $\lambda = L_K, \tau \leq \frac{1}{6\lambda}$ . Then for the  $\mathbf{x}_{t+1}$  update given in Algorithm 1, we have

$$\mathbb{E}\left[\varphi_{1/\lambda}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}, \mathbf{z}_{t+1})\right] \leq \mathbb{E}\left[\varphi_{1/\lambda}(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_{t+1})\right] - \frac{\tau\lambda^2}{16}\mathbb{E}\|\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{x}_t\|^2 + \lambda\tau^2\sigma^2 + (\lambda\tau\mu + 2\lambda\tau^2\mu^2 + \tau\lambda^2\mu^2/8\gamma_s^2)\mathbb{E}\|\mathbf{z}_t - \mathbf{z}_{t+1}\|^2,$$

where  $\gamma_s = 2\mu + \rho ||A||, L_K = L_f + \rho ||A|| + \mu$ .

This follows mostly from Davis & Drusvyatskiy (2019) and handles the transition from  $\mathbf{x}_t$  to  $\mathbf{x}_{t+1}$  in our analysis. One additional error term we have here is  $\|\mathbf{z}_{t+1} - \mathbf{z}_t\|^2$ , due to the change in the proximal center  $\mathbf{z}_t$ , a term that was not involved in the analysis of Davis & Drusvyatskiy (2019).

Next, we have to incorporate the dynamics of the updates on the dual variable  $\mathbf{y}_t$  and the proximal center  $\mathbf{z}_t$  on top of the previous result. These results will use some ideas from Zhang & Luo (2022) with some additional insights. The reason is that since Zhang & Luo (2022) uses the function  $L_{\rho}(\mathbf{x}, \mathbf{y}) + \frac{\lambda}{2} ||\mathbf{x} - \mathbf{z}||^2$  in their potential function, their analysis only characterizes the change in  $\mathbf{y}$  and  $\mathbf{z}$  in this function. Our analysis however, needs to characterize this change in the Moreau envelope of this function. This requires further estimations using the properties of the Moreau envelope, as well as the proximal point  $\mathbf{u}^*(\mathbf{x}, \mathbf{y}, \mathbf{z})$  (see, for example, Lemma A.3 and Appendix A).

**Lemma 3.4.** (cf. Lemma A.3) Let Assumption 1.1 hold, then for the iterates generated by Algorithm 1, we have

$$\mathbb{E}\left[\varphi_{1/\lambda}(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_{t+1})\right] \le \mathbb{E}\left[\varphi_{1/\lambda}(\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t)\right] - \mathbb{E}\langle \mathbf{y}_{t+1} - \mathbf{y}_t, A\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t) - \mathbf{b}\rangle - \frac{\mu}{2}\mathbb{E}\langle \mathbf{z}_t - \mathbf{z}_{t+1}, 2\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{z}_{t+1} - \mathbf{z}_t\rangle.$$

It is easy to notice that combining the last two lemmas will give us a bound on the change of  $\varphi_{1/\lambda}$  between timesteps t and t + 1. On the other hand, the inner products appearing on the right-hand side of the last bound will require an intricate analysis after combining with the terms coming from other components in the potential function, introduced next. One aim, is to make sure we get enough slack to be able to cancel error terms coming from  $\|\mathbf{z}_{t+1} - \mathbf{z}_t\|^2$  in the previous lemma and further errors that will arise as we handle the inner products.

#### 3.3 Proof Sketch

#### 3.3.1 One iteration inequality on the potential function

As alluded to earlier, we introduce the potential function we work with, which incorporates the Moreau envelopes defined earlier in (2.1) and (3.3):

$$V_t = \varphi_{1/\lambda}(\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t) - 2d(\mathbf{y}_t, \mathbf{z}_t) + 2\Psi(\mathbf{z}_t),$$

where we use

$$d(\mathbf{y}, \mathbf{z}) = \min_{\mathbf{x} \in X} L_{\rho}(\mathbf{x}, \mathbf{y}) + \frac{\mu}{2} \|\mathbf{x} - \mathbf{z}\|^2.$$
(3.5)

There are two main changes compared to the analysis of Zhang & Luo (2022). The first is that the *primal* descent portion of our analysis investigates the behavior of the Moreau envelope of the proximal AL function

(given in (3.3)) whereas the analysis of Zhang & Luo (2022) analyzes the proximal AL function (given in (5.3)) directly.

The reason for this departure is the well-known difficulty while analyzing SGD for constrained problems with single sample of stochastic gradients. Hence, it is not clear if it is possible to show descent for the proximal AL function in the constrained case without using large minibatch sizes. In particular, until the work of Davis & Drusvyatskiy (2019), convergence analyses of projected SGD required large batches.

In addition to combining the bounds from the previous section on the change of  $\varphi_{1/\lambda}$ , we have to characterize the change in  $d(\mathbf{y}, \mathbf{z})$  and  $\Psi(\mathbf{z})$ , for which we can use the following estimations, which only use the definition of  $\mathbf{y}_{t+1}$  and hence have the same proof as the previous work.

**Lemma 3.5.** (Zhang & Luo, 2020, Lemma 3.2, Lemma 3.3) For the functions  $d(\mathbf{y}, \mathbf{z})$  and  $\Psi(\mathbf{z})$  defined in (2.1) and (3.5), we have

$$d(\mathbf{y}_{t+1}, \mathbf{z}_{t+1}) - d(\mathbf{y}_t, \mathbf{z}_t) \ge \eta \langle A\mathbf{x}_t - \mathbf{b}, A\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{b} \rangle + \frac{\mu}{2} \langle \mathbf{z}_{t+1} - \mathbf{z}_t, \mathbf{z}_{t+1} + \mathbf{z}_t - 2\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_{t+1}) \rangle,$$

and

$$\Psi(\mathbf{z}_{t+1}) - \Psi(\mathbf{z}_t) \le \mu \langle \mathbf{z}_{t+1} - \mathbf{z}_t, \mathbf{z}_t - \bar{\mathbf{x}}^*(\mathbf{z}_t) \rangle + \frac{\mu}{2\sigma_4} \|\mathbf{z}_t - \mathbf{z}_{t+1}\|^2,$$

where  $\sigma_4 = \frac{\mu - L_f}{\mu}$  and

$$\mathbf{x}^{*}(\mathbf{y}, \mathbf{z}) = \operatorname*{arg\,min}_{\mathbf{x} \in X} L_{\rho}(\mathbf{x}, \mathbf{y}) + \frac{\mu}{2} \|\mathbf{x} - \mathbf{z}\|^{2}, \tag{3.6}$$

$$\bar{\mathbf{x}}^*(\mathbf{z}) = \operatorname*{arg\,min}_{\mathbf{x}\in X, A\mathbf{x}=\mathbf{b}} f(\mathbf{x}) + \frac{\mu}{2} \|\mathbf{x} - \mathbf{z}\|^2.$$
(3.7)

We continue with the main *descent*-type inequality on the potential function after one iteration of the algorithm. The proof of this lemma is rather intricate and requires a careful combination of the inner products coming from the previous lemmas, and using the particular update of the proximal center  $\mathbf{z}_{t+1}$  as well as parameter selections. Let us recall that  $\mathbf{u}^*(\mathbf{x}, \mathbf{y}, \mathbf{z})$  and  $\mathbf{x}^*(\mathbf{y}, \mathbf{z})$  that appear in the lemma statement are defined in (3.4) and (3.6).

**Lemma 3.6.** (cf. Lemma A.6) Under Assumption 1.1, with the parameters selected as (3.1) (see also (A.1)), the iterates of Algorithm 1 satisfy the inequality

$$\mathbb{E}V_t - \mathbb{E}V_{t+1} \ge c_\beta \mathbb{E}\|\mathbf{z}_{t+1} - \mathbf{z}_t\|^2 - \lambda \tau^2 \sigma^2 + c_\tau \mathbb{E}\|\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{x}_t\|^2 + c_\eta \mathbb{E}\|A\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{b}\|^2, \quad (3.8)$$

where  $c_{\tau} = \Theta(1/\sqrt{T}), c_{\eta} = \Theta(1/\sqrt{T}), c_{\beta} = \Theta(1/\sqrt{T})$  with their precise definitions given in Lemma A.6.

One novelty in our analysis is to show that this potential function is still lower bounded and decreases, in expectation, up to an error term depends on  $\tau^2$  and the variance. To integrate this change into the framework of Zhang & Luo (2022) under reasonable assumptions on the stochastic oracle as mentioned earlier in Section 2, we also slightly changed the definition of  $\mathbf{z}_{t+1}$  in the algorithm, due to technical reasons. In particular, in our case, we lose the control over  $\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2$  (since we do not assume bounded domains in this section), whereas the deterministic analysis of Zhang & Luo (2022) have a natural control over such terms.

The other change is the error coming from the variance of the stochastic gradients of f. This error causes the complexity to deteriorate compared to the deterministic case, but this is an effect that is common with algorithms based on SGD. In particular, with a correctly selected step size, we still obtain the same-order sample complexity as SGD, which is optimal even for unconstrained smooth nonconvex optimization (Arjevani et al., 2023).

#### 3.3.2 Complexity analysis

After Lemma 3.6, it is straightforward to obtain

$$\begin{split} \mathbb{E} \| \mathbf{z}_{t+1} - \mathbf{z}_t \|^2 &\leq \varepsilon^2, \\ \mathbb{E} \| A \mathbf{x}^* (\mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{b} \|^2 &\leq \varepsilon^2, \\ \mathbb{E} \| \mathbf{u}^* (\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{x}_t \|^2 &\leq \varepsilon^2, \end{split}$$

when  $T = \Theta(\varepsilon^{-4})$ . Then, by tedious but straightforward calculations, we can directly get the bound on the norm of the gradient of the Moreau envelope,  $\nabla \Psi(\mathbf{z}_t)$ , obtaining near-stationarity. The details for these estimations appear in Appendix A.2.

There are a couple more steps to go from this result to obtaining  $\varepsilon$ -stationary points, but the idea is simple. Since we know that small norm of  $\nabla \Psi(\mathbf{z}_t)$  means that we are near a stationary point, we can perform one more iteration of SGD with a batch size depending on  $\varepsilon^{-2}$  to get an  $\varepsilon$  stationary point, without changing the dominant term in the complexity. The details are given in Appendix A.3.

### 4 Extension to Random Linear Constraints

We turn to the case when constraints are sampled, that is, we do not have access to the full matrix A, or vector **b** but only to unbiased samples of them. This is a suitable setting, when, for example, we have a large matrix A. In particular, we have  $A = \mathbb{E}_{\zeta \sim P}[A_{\zeta}]$ ,  $\mathbf{b} = \mathbb{E}_{\zeta \sim P}[\mathbf{b}_{\zeta}]$  and use  $A_{\zeta}$ ,  $\mathbf{b}_{\zeta}$  in the algorithm. We rewrite the template for convenience, as

$$\min_{\mathbf{x}\in X} f(\mathbf{x}) \text{ subject to } \mathbb{E}_{\zeta \sim P}[A_{\zeta}\mathbf{x} - \mathbf{b}_{\zeta}] = 0,$$
(4.1)

where  $f(\mathbf{x}) = \mathbb{E}_{\xi \sim \Xi}[f(\mathbf{x}, \xi)]$ . In this case, to get an unbiased stochastic gradient for the proximal augmented Lagrangian, we need to sample two i.i.d. samples of  $\zeta$  and compute

$$G(\mathbf{x}, \mathbf{y}, \mathbf{z}, \xi) = f(\mathbf{x}, \xi) + A_{\zeta^1}^\top \mathbf{y} + A_{\zeta^1}^\top (A_{\zeta^2} \mathbf{x} - \mathbf{b}_{\zeta^2}).$$
(4.2)

An immediate issue that arises is that the variance of the stochastic gradients of the proximal AL function now scale as  $\mathbf{x}$  and  $\mathbf{y}$ . As such, assuming bounded variance would require assuming bounded dual variables, which is a strong assumption that is not satisfied in practice. To go around this difficulty, we have two adjustments, (i) we need to assume a constraint qualification (CQ) condition and compactness of X and (ii) we include a safeguarding procedure in the algorithm to monitor when the dual variable gets too large. We will show that under these two modifications, we can obtain the same complexity guarantees as our previous setting with deterministic linear constraints.

#### Algorithm 2 Stochastic smoothed and linearized ALM for stochastic constraints with dual safeguarding

Input:  $M_y > \frac{M_V - M_{\Psi} + 2M}{r}$  (check also Remark 4.1) Initialize:  $\mathbf{x}_0 = \mathbf{z}_0 \in X$ ,  $\mathbf{y}_0 \in \mathbb{R}^m$ ,  $\rho \ge 0$ . for t = 0 to T - 1 do  $\mathbf{y}_{t+1} = \mathbf{y}_t + \eta(A_{\zeta_t}\mathbf{x}_t - \mathbf{b}_{\zeta_t})$  where  $\zeta_t \sim P$  is generated i.i.d. if  $\|\mathbf{y}_{t+1}\| \ge M_y$  then  $\mathbf{y}_{t+1} = 0$ end if Sample  $\xi_t \sim \Xi$  i.i.d. and generate  $\mathbb{E}_{\xi_t}[G(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t, \xi_t)] = \nabla_{\mathbf{x}} L_{\rho}(\mathbf{x}_t, \mathbf{y}_{t+1}) + \mu(\mathbf{x}_t - \mathbf{z}_t)$  as in (4.2)  $\mathbf{x}_{t+1} = \operatorname{proj}_X(\mathbf{x}_t - \tau G(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t, \xi_t))$   $\mathbf{z}_{t+1} = \mathbf{z}_t + \beta(\mathbf{x}_t - \mathbf{z}_t)$ end for Remark 4.1. We give the choice of  $M_y$  as follows. Let  $M_V = \max_{\mathbf{x}, \mathbf{z} \in X} \{K(\mathbf{x}, 0, \mathbf{z}) - 2d(0, \mathbf{z}) + 2\Psi(\mathbf{z})\}$ ,  $M = \max_{\mathbf{x}, \mathbf{z} \in X} \{|f(\mathbf{x})| + \frac{\mu}{2} \| \mathbf{x} - \mathbf{z} \|^2 + \frac{\rho}{2} \| A\mathbf{x} - \mathbf{b} \|^2 \}$ , where K is defined in (5.3) and  $M_{\Psi}$  is a uniform lower bound of  $\Psi(\mathbf{z}_t)$ , for example,  $\underline{f}$ . According to Assumption 4.2, there exists a positive r > 0 such that for any direction  $\mathbf{d} \in \operatorname{Range}(A)$ , we can find a  $\mathbf{x} \in X$  satisfying  $\|A\mathbf{x} - \mathbf{b}\| = r$  and  $A\mathbf{x} - \mathbf{b}$  has the same direction as  $\mathbf{d}$ . Then, we choose  $M_y$  as

$$M_y > \frac{M_V - M_\Psi + 2M}{r}.$$

Assumption 4.2. For the problem given in (4.1), the following holds:

- 1. The feasible set  $\{\mathbf{x} : \mathbf{x} \in X, A\mathbf{x} = \mathbf{b}\}$  is bounded.
- 2. The origin is in the relative interior of the set  $\{A\mathbf{x} \mathbf{b} \colon \mathbf{x} \in X\}$
- 3. A has full row-rank.

Here, in addition to the assumptions in the earlier setting, we require a Slater's condition as well as compact domains to ensure boundedness of the dual variable. Slater's condition is a classical CQ, see for example Bertsekas et al. (2003).

In this setting, we only state our theorem for the near-stationarity. The  $\varepsilon$ -stationarity follows in the same way as the previous section by a post-processing step.

**Theorem 4.3.** Let Assumptions 1.1 and 4.2 hold and run Algorithm 2 with the parameters given in (3.1) (also (A.1)). We have that  $\mathbb{E} \| \nabla \Psi(\mathbf{z}_{t^*}) \| \leq \varepsilon$  where  $t^*$  is randomly selected from  $\{1, \ldots, T\}$  with  $T = \Omega(\varepsilon^{-4})$ . The stochastic oracle complexity is  $O(\varepsilon^{-4})$ .

For the proof of this theorem, we refer to Appendix B.

As mentioned earlier, the optimal sample complexity for nonconvex optimization with Lipschitz  $\nabla f$  is  $O(\varepsilon^{-4})$  (Arjevani et al., 2023). Our result achieves this optimal complexity while handling linear constraints with random sampling.

### 5 Extension with Variance Reduction

Algorithm 3 Stochastic smoothed and linearized ALM with STORM

Initialize:  $x_0 = z_0 \in X, y_0 \in \mathbb{R}^m$ ,  $\widehat{\nabla} f_0 = \frac{1}{N} \sum_{i=1}^N \nabla f(x_0, \zeta_i)$ ,  $N = T^{1/6}$  and  $\rho \ge 0$ for t = 0 to T - 1 do  $\mathbf{y}_{t+1} = \mathbf{y}_t + \eta(A\mathbf{x}_t - \mathbf{b})$   $G(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) = \widehat{\nabla} f_t + A^\top \mathbf{y} + A^\top (A\mathbf{x}_t - \mathbf{b}) + \lambda(\mathbf{x}_t - \mathbf{z}_t)$   $\mathbf{x}_{t+1} = \operatorname{proj}_X(\mathbf{x}_t - \tau G(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t))$   $\mathbf{z}_{t+1} = \mathbf{z}_t + \beta(\mathbf{x}_t - \mathbf{z}_t)$ Sample  $\xi_{t+1} \sim \Xi$  i.i.d. and set  $\widehat{\nabla} f_{t+1} = \nabla f(\mathbf{x}_{t+1}, \xi_{t+1}) + (1 - \alpha)(\widehat{\nabla} f_t - \nabla f(\mathbf{x}_t, \xi_{t+1}))$ end for

We now introduce the STORM variance reduction technique from Cutkosky & Orabona (2019) into our algorithm, which improves the iteration and oracle complexity from  $O(\varepsilon^{-4})$  to  $O(\varepsilon^{-3})$  under a stronger assumption on the oracle, compared to our earlier sections. With this variant, we reduce the variance of the stochastic gradients of the objective function, which leads to a faster convergence rate, and, also a simpler analysis that does not rely on the Moreau envelope  $\varphi_{1/\lambda}$ .

The ALM-STORM algorithm is given in Algorithm 3. The main difference between ALM-STORM and the original algorithm is the update of the stochastic gradient estimate  $\hat{\nabla} f_t$ . In STORM, the stochastic gradient estimate is updated using the previous stochastic gradient estimate, to reduce the variance of the stochastic gradients. The update of  $\hat{\nabla} f_t$  is given by

$$\widehat{\nabla}f_{t+1} = \nabla f(\mathbf{x}_{t+1}, \xi_{t+1}) + (1 - \alpha)(\widehat{\nabla}f_t - \nabla f(\mathbf{x}_t, \xi_t)),$$
(5.1)

where  $\alpha \in (0, 1)$  is a parameter to be determined.

The update of  $\nabla f_t$  is a linear combination of the current stochastic gradient estimate, the previous stochastic gradient and a correction term involving  $\nabla f(\mathbf{x}_{t+1}, \xi_t)$  and  $\nabla f(\mathbf{x}_t, \xi_t)$ . It is easy to see that when  $\alpha = 0$ , Algorithm 3 reduces to Algorithm 1, but we will see that a particular choice of  $\alpha$  will help us obtain a better complexity under Assumption 5.2, which is stronger than the oracle access and smoothness required in Assumption 1.1.

Remark 5.1. We only use a minibatch in the initialization, which does not affect the overall complexity. The minibatch size is  $N = T^{1/6}$ , which is small compared to the total number of iterations T. The iterations of our algorithm only require 2 stochastic gradients,  $\nabla f(\mathbf{x}_t, \xi_{t+1})$  and  $\nabla f(\mathbf{x}_{t+1}, \xi_{t+1})$ .

For the analysis of ALM-STORM, we introduce the new assumption mentioned above. This is used, for example, in Arjevani et al. (2023). In particular, Arjevani et al. (2023) showed that the oracle complexity  $O(\varepsilon^{-3})$  is tight under Assumption 5.2 even with no constraints.

Assumption 5.2. For a given  $\xi \sim \Xi$ , we can query  $\nabla f(\mathbf{x}, \xi)$  and  $\nabla f(\mathbf{y}, \xi)$  for different points  $\mathbf{x}, \mathbf{y}$ . There exists a constant  $L_0 > 0$  such that for all  $\mathbf{x}, \mathbf{y} \in X$ , we have

$$\mathbb{E}_{\xi \sim \Xi} \|\nabla f(\mathbf{x}, \xi) - \nabla f(\mathbf{y}, \xi)\|^2 \le L_0^2 \|\mathbf{x} - \mathbf{y}\|^2.$$

We also have access to a stochastic gradient of f that satisfies (1.3).

The proof of the following lemma, taken from Cutkosky & Orabona (2019), is given in Appendix C for completeness.

**Lemma 5.3.** (from Cutkosky & Orabona (2019)) Let Assumption 5.2 hold. We have the estimation of the variance as:

$$\mathbb{E}\|\widehat{\nabla}f_{t+1} - \nabla f(\mathbf{x}_{t+1})\|^2 \le (1-\alpha)^2 \mathbb{E}\|\widehat{\nabla}f_t - \nabla f(\mathbf{x}_t)\|^2 + 3(L_0^2 + L_f^2) \mathbb{E}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + 3\alpha^2 \sigma^2.$$

We introduce a different potential function  $\bar{V}_t$  for the ALM-STORM algorithm compared to Sections 3 and 4. The potential function we use in this section is similar to the one defined in Zhang & Luo (2022), with the exception of the last term that helps us control the error coming from the variance. In particular, we have

$$\bar{V}_t = K(\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t) - 2d(\mathbf{y}_t, \mathbf{z}_t) + 2\Psi(\mathbf{z}_t) + \frac{1}{48(L_0^2 + L_f^2)\tau} \mathbb{E} \|\widehat{\nabla}f_t - \nabla f(\mathbf{x}_t)\|^2,$$
(5.2)

where

$$K(\mathbf{x}, \mathbf{y}, \mathbf{z}) = L_{\rho}(\mathbf{x}, \mathbf{y}) + \frac{\mu}{2} \|\mathbf{x} - \mathbf{z}\|^2$$
(5.3)

and  $\mathbf{x} \mapsto K(\mathbf{x}, \mathbf{y}, \mathbf{z})$  is  $L_K$ -smooth with  $L_K = L_f + \rho ||A|| + \mu$ .

We first establish the descent-type lemma of this potential function, which is the key step in the analysis of the ALM-STORM algorithm. Compared to the deterministic settings as in Zhang & Luo (2022), we have the extra error due to using  $\widehat{\nabla} f_t$  instead of the full gradient  $\nabla f(\mathbf{x}_t)$ .

**Lemma 5.4.** Let Assumption 1.1 hold. For the iterates generated by Algorithm 3, we have

$$K(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}, \mathbf{z}_t) - K(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) \le \frac{\tau}{2} \|\nabla f(\mathbf{x}_t) - \widehat{\nabla} f_t\|^2 - (\frac{1}{2\tau} - \frac{L_K}{2}) \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2$$

The proof of Lemma 5.4 could be found in Appendix C.

**Lemma 5.5.** Let Assumption 1.1 hold. For the iterates generated by Algorithm 3, we have

$$K(\mathbf{x}_{t}, \mathbf{y}_{t}, \mathbf{z}_{t}) - K(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) \geq -\eta \|A\mathbf{x}_{t} - \mathbf{b}\|^{2} + (\frac{\mu}{\beta} - \frac{3\mu}{4})\|\mathbf{z}_{t+1} - \mathbf{z}_{t}\|^{2} - \frac{\tau}{2} \|\nabla f(\mathbf{x}_{t}) - \widehat{\nabla}f_{t}\|^{2} + (\frac{1}{2\tau} - \frac{L_{K}}{2} - \mu)\|\mathbf{x}_{t+1} - \mathbf{x}_{t}\|^{2}$$
(5.4)

The proof of Lemma 5.5 also could be found in Appendix C.

Then we can combine the above lemma analyzing one step change of  $K(\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t)$  with the lemmas analyzing one step changes of  $d(\mathbf{y}_t, \mathbf{z}_t), \Psi(\mathbf{z}_t)$  (Lemma A.5) as well as the variance term (Lemma 5.3), to obtain the final lemma for the change in the potential function  $\bar{V}_t$  from t to t + 1. For the proof, we refer to Appendix C.

**Theorem 5.6.** Under Assumption 1.1 and Assumption 5.2, with the parameters chosen as:

$$\mu = \max\{2, 4L_f\}, \quad \tau \le \min\left\{\frac{1}{4L_K + 8\mu}, \frac{1}{\sqrt{48(L_0^2 + L_f^2)}}\right\}$$
  

$$\eta = \min\left\{\frac{(\mu - L_f)^2 \tau}{4\|A\|^2}, \frac{2\mu + \rho\|A\|}{4\|A\|^4}, \frac{\tau}{200\|A\|^2}, \frac{\tau(2\mu + \rho\|A\|^2)}{20\|A\|^2}\right\}, \quad (5.5)$$
  

$$\beta = \min\left\{\frac{\tau}{100}, \frac{1}{50}, \frac{\eta}{36\mu\bar{\sigma}^2}\right\},$$
  

$$\alpha = 48(L_0^2 + L_f^2)\tau^2,$$

where  $L_K = L_f + \rho ||A|| + \mu$ ,  $\bar{\sigma}$  is defined in Lemma A.9, we have

$$\mathbb{E}\bar{V}_{t} - \mathbb{E}\bar{V}_{t+1} \ge \frac{\mu}{2\beta} \mathbb{E}\|\mathbf{z}_{t} - \mathbf{z}_{t+1}\|^{2} + \frac{1}{8\tau} \mathbb{E}\|\mathbf{x}_{t} - \mathbf{x}_{t+1}\|^{2} + \frac{\eta}{2} \mathbb{E}\|A\mathbf{x}^{*}(\mathbf{y}_{t+1}, \mathbf{z}_{t}) - b\|^{2} + \frac{\tau}{4} \mathbb{E}\|\widehat{\nabla}f_{t} - \nabla f(\mathbf{x}_{t})\|^{2} - 144(L_{0}^{2} + L_{f}^{2})\sigma^{2}\tau^{3}.$$
(5.6)

Note that, on a high level, the main difference between Theorem 5.6 and Lemma 3.6 is that the order of  $\tau$  in the error term is different. In Theorem 5.6, the order of  $\tau$  is  $O(\tau^3)$ , while in Lemma 3.6, the order of  $\tau$  is  $O(\tau^2)$ , which contribute to a faster convergence rate in the ALM-STORM algorithm.

**Theorem 5.7.** Let Assumptions 1.1 and 5.2 hold. We have that  $(x_{t^*}, y_{t^*})$  is an  $\varepsilon$ -stationary point, where  $t^*$  is selected uniformly at random from  $\{1, \ldots, T\}$  with  $T = \Theta(\varepsilon^{-3})$ . The complexity of the whole procedure is  $O(\varepsilon^{-3})$ .

For the proof of this theorem, we refer to Appendix C.

Remark 5.8. Under Assumptions 1.1, 4.2 and 5.2, we can combine this variance reduction technique with our extension to stochastic constraints in Section 4 to obtain the same  $O(\varepsilon^{-3})$  complexity result for the stochastic linear constraints case. We provide a brief justification for this claim in Appendix C.

### 6 Applications

#### 6.1 Distributed Optimization

In this section, we consider the distributed optimization problem with the following form

$$\min_{\mathbf{x}\in X} \Big\{ f(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} f_i(\mathbf{x}) \Big\},\tag{6.1}$$

where  $X \subset \mathbb{R}^n$  is a polyhedral set.

Typically, this problem is addressed using a network with N nodes, represented as an undirected graph G = (V, E), where V is the set of nodes and E is the set of edges. The number of nodes is |V| = N and the number of edges is |E| = M. Each node *i* can only access its own local function  $f_i$  and communicate with with

its neighboring node j, meaning that an edge (i, j) exists in E. Classically, we now model this communication setting by introducing N local variables  $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$  for each node and define the concatenated vector as

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_N \end{pmatrix}$$

With this, the problem (6.1) can be formulated as follows:

$$\min_{\mathbf{x}_1,\dots,\mathbf{x}_N} \frac{1}{N} \sum_{i=1}^N f_i(\mathbf{x}_i)$$
s.t.  $\mathbf{x}_i = \mathbf{x}_j, \quad \forall (i,j) \in E \text{ and } \mathbf{x}_i \in X \quad \forall i = 1,\dots,N.$ 

$$(6.2)$$

Next, we work to reformulate this into a more concise representation. Specifically, we introduce the *edge-agent incidence matrix*  $W \in \mathbb{R}^{\frac{N(N-1)}{2} \times N}$ . Each row of W corresponds to a node in the graph G. In particular, if we take the *k*th row and if this row represents the pair (i, j) in the graph, then if there is an edge between (i, j), we define  $W_{k,i} = 1$ ,  $W_{k,j} = -1$ . Then, we set all other entries in the *k*th row to zero.

Let us recall that  $\mathbf{x}_i \in \mathbb{R}^n$  and  $\mathbf{x} \in \mathbb{R}^{nN}$ . To represent the relationships of nodes by using W, we define  $A = W \otimes I_n$ , where  $\otimes$  denotes the Kronecker product. Then, we can rewrite the constraints (6.2) as  $A\mathbf{x} = 0$ .

Here, we consider a more general case where the network structure graph is random, that is, the connections between the nodes may change from iteration to iteration. We consider a discrete-time random graph model, which is discussed in Chaintreau et al. (2007). In this model, the network is represented as a time-varying graph  $G_t = (V, E_t)$ , where V is the set of nodes and  $E_t$  is the set of edges at time t. The edges  $E_t$  are determined by a probabilistic process, capturing the dynamic nature of the network.  $\mathbb{P}[(i, j) \in E_t] = p_{ij}$  for any pair of nodes i, j and the events  $\{(i, j) \in E_t\}$ , for all pair of nodes i, j are mutually independent.

This model is particularly useful for analyzing communication networks where connections between nodes are not static but change over time due to mobility, interference, or other dynamic factors. The random graph model allows us to study the behavior of algorithms and protocols under realistic, time-varying network conditions.

Hence, we model this situation with  $W = W(\zeta)$ , where  $\zeta$  is a random variable. Then the constraints  $A\mathbf{x} = 0$  changes to  $\mathbb{E}_{\zeta \sim P}[A(\zeta)]\mathbf{x} = A\mathbf{x} = 0$ . In the discrete-time model, a row in  $\mathbb{E}[W]$  represents the likelihood of a connection between (i, j), that is, the i-th entry equals to  $p_{ij}$ , the j-th entry equals to  $-p_{ij}$ .

The problem (6.2) comes to the following form:

$$\min_{\mathbf{x}\in X^N} f(\mathbf{x})$$
s.t.  $\mathbb{E}_{\zeta\sim P}[A(\zeta)]\mathbf{x} = 0,$ 
(6.3)

where  $f(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} f_i(\mathbf{x}_i)$ . Then, we can use Algorithm 2 to solve the problem (6.3).

In the discrete-time model, for example, the first row of  $\mathbb{E}_{\zeta \sim P}[A(\zeta)]$  equals to  $(p_{12}, -p_{12}, 0, ..., 0)$ , hence the first entry of  $\mathbb{E}_{\zeta \sim P}[A(\zeta)\mathbf{x}] = p_{12}x_1 - p_{12}x_2$ . Then by the definition of  $\varepsilon$ -stationary point, particularly the feasibility bound, we will get

$$\sum_{i < j} \mathbb{E} \| p_{ij} x_i - p_{ij} x_j \|^2 \le \varepsilon^2.$$

We assume  $\forall i, j, p_{ij} \ge p > 0$ , then this condition assures that each  $\mathbf{x}_i$  converges to the same point, which is also the stationary solution of the original problem (6.1).

With our method, we do not need the assumption that the graph is connected at any iteration and our developments in Section 4 apply for this problem. We can convert our stochastic primal-dual algorithm to distributed form, where we refer to Chen et al. (2021).

#### 6.2 Discrete Optimization with Smooth Nonconvex Regularizers

In this section, we follow an idea from Zhang et al. (2017) to deal with discrete optimization problems by using continuous nonconvex regularizers to relax the discrete constraints. Then, this brings the need to handle objective functions with nonconvexity.

We consider a communication network represented by a directed graph  $G = (\mathcal{V}, \mathcal{L})$ , where  $\mathcal{V}$  denotes the set of nodes and  $\mathcal{L}$  represents the set of directed links. We define  $\mathcal{V}_f$  as the subset of function nodes capable of providing service function f, where each node has computational capacity  $\mu_i$ . The network serves K data flows, each requiring a service function chain  $\mathcal{F}(k)$  that must be executed in sequence. Then we denote  $r_{ij}(k)$  as the rate of flow k on link (i, j),  $r_{ij}(k, f)$  as the rate of virtual flow (k, f) on link (i, j) and  $x_{i,j}(k)$  as the binary variable indicating whether or not f is used by flow k in node i. The network slicing problem aims to determine optimal routes and flow rates that satisfy both service function chain requirements and capacity constraints of all links and function nodes.

We omit some details of the constraints about the network slicing problem for brevity, and just write down those as linear constraints which are discussed extensively in Zhang et al. (2017). Then abstract form of this problem is

$$\min_{\mathbf{r},\mathbf{x}} \quad g(\mathbf{r}) = \sum_{k,(i,j)} r_{ij}(k)$$
s.t.  $A\begin{bmatrix}\mathbf{r}\\\mathbf{x}\end{bmatrix} = 0,$ 

$$\sum_{i \in V_f} x_{i,f}(k) = 1, \forall f \in \mathcal{F}(k), \forall k \quad (6.4)$$

$$r_{ij}(k) \ge 0, \quad \forall k, \forall (i,j) \in \mathcal{L},$$

$$r_{ij}(k, f) \ge 0, \quad \forall f \in \mathcal{F}(k), \forall k, \forall (i,j) \in \mathcal{L},$$

$$x_{i,f}(k) \in \{0,1\}, \quad \forall i \in \mathcal{V}_f, \forall f \in \mathcal{F}(k), \forall k,$$

where  $\mathbf{r} = \{r_{ij}(k), r_{ij}(k, f)\}$  and  $\mathbf{x} = \{x_{i,f}(k)\}$ . This is a linear programming problem with binary constraints.

Then Zhang et al. (2017) uses the continuous relaxation of the binary constraints and add a nonconvex regularizer to solve the problem. In particular, this work shows that the solution of the binary LP can be approximated by this continuous but *nonconvex* problem

$$\begin{array}{ll}
\min_{\mathbf{r},\mathbf{x}} & g(\mathbf{r}) + \sigma P_{\epsilon}(\mathbf{x}) \\
\text{s.t.} & A\begin{bmatrix}\mathbf{r}\\\mathbf{x}\end{bmatrix} = 0, \\
& r_{ij}(k) \ge 0, \quad \forall k, \forall (i,j) \in \mathcal{L}, \\
& r_{ij}(k,f) \ge 0, \quad \forall f \in \mathcal{F}(k), \forall k, \forall (i,j) \in \mathcal{L}, \\
& x_{i,f}(k) \in [0,1], \quad \forall i \in \mathcal{V}_f, \forall f \in \mathcal{F}(k), \forall k, \\
\end{array}$$
(6.5)

where  $\sigma > 0$  is the penalty parameter and nonconvexity stems from  $P_{\epsilon}$ . In particular, we have  $P_{\epsilon}(\mathbf{x}) = \sum_{k} \sum_{f \in \mathcal{F}(k)} (\|x_f(k) + \epsilon \mathbf{1}\|_p^p - c_{\epsilon,f})$ , where  $x_f(k) = \{x_{i,f}(k)\}_{i \in V_f}$ ,  $c_{\epsilon,f} = (1 + \epsilon)^p + (|V_f| - 1)\epsilon^p$  and  $p \in (0, 1)$ ,  $\epsilon$  is any nonnegative constant. We can then apply Algorithm 1 to solve the problem (6.5).

#### 6.3 Classification with Fairness

We consider the setting of a *binary classification* task, where the goal is to learn a decision rule

$$f_{\theta}: \mathbb{R}^d \to \{-1, +1\},$$

where  $\theta$  is the parameter.

We note that we use a different notation in this section than the rest of our text to be compatible with the application we consider.

Given a training set of labeled examples  $\{(x_i, y_i)\}_{i=1}^N$ , each  $x_i$  is a feature vector in  $\mathbb{R}^d$  and  $y_i \in \{-1, +1\}$ . The task is to find parameters  $\theta$  that define a decision boundary and minimize a chosen loss function  $L(\theta)$  on the training data. Once trained, the classifier predicts +1 if a test point's signed distance to the boundary, denoted as  $d_{\theta^*}(x)$ , is non-negative, and -1 otherwise, where  $\theta^* = \arg \min_{\theta} L(\theta)$ .

In Zafar et al. (2017), the authors define the measure of (un)fairness of a decision boundary as the covariance between the set of sensitive attributes  $\{z_i\}_{i=1}^N$  and the signed distance of each sample's feature vector to the decision boundary  $\{d_{\theta}(x_i)\}_{i=1}^N$ . Formally,

$$\operatorname{Cov}(z, d_{\theta}(x)) = \mathbb{E}[(z - \bar{z}) d_{\theta}(x)] - \mathbb{E}[z - \bar{z}] \bar{d}_{\theta}(x) \approx \frac{1}{N} \sum_{i=1}^{N} (z_i - \bar{z}) d_{\theta}(x_i),$$
(2)

where  $\mathbb{E}[z-\bar{z}] \bar{d}_{\theta}(x) = 0$  since  $\mathbb{E}[z-\bar{z}] = 0$ . Here,  $\bar{z}$  denotes the average of the sensitive attribute over the training set, and  $\bar{d}_{\theta}(x)$  is the mean signed distance.

Considering the setting of linear classifier, that is,  $f_{\theta}(x) = \langle \theta, x \rangle$ , one has the following problem:

$$\min_{\theta} \quad L(\theta) = \frac{1}{N} \sum_{i=1}^{N} V(f_{\theta}(x_i), y_i)$$
s.t. 
$$\frac{1}{N} \sum_{i=1}^{N} (z_i - \bar{z}) \theta^{\top} x_i \le c$$

$$\frac{1}{N} \sum_{i=1}^{N} (z_i - \bar{z}) \theta^{\top} x_i \ge -c,$$
(6.6)

where c is the covariance threshold.

Although L for logistic regression that is considered in Zafar et al. (2017) is indeed convex, there are many nonconvex loss functions for this classification problem. For example Krause & Singer (2004) proposes a smooth nonconvex loss function called Logistic difference loss function for classification problems, which is defined as follows:

$$V(f(x), y) = \log(1 + e^{-yf(x)}) - \log(1 + e^{-yf(x)-\mu}),$$
(6.7)

where the  $\mu$  is a parameter.

In Zhao et al. (2010), the authors propose smoothed 0-1 loss function as follow:

$$V(f(x), y) = \begin{cases} 0, & yf(x) > 1\\ \frac{1}{4}yf(x)^3 - \frac{3}{4}yf(x) + \frac{1}{2}, & -1 \le yf(x) \le 1\\ 1, & yf(x) < -1. \end{cases}$$
(6.8)

We refer to the review about the nonconvex loss functions used in classification problems in Zhao et al. (2010). This work showcases certain advantages of using nonconvex loss functions, such as robustness to outliers, better approximation to 0-1 loss and improved generalization, which are supported by experiment results in Zhao et al. (2010).

When we use a nonconvex smooth loss function in the classification problem, we can apply our Algorithm 1 to solve the problem (6.6).

### 7 Related Works

Since the literature of algorithms solving the problem (1.1) is broad with different focuses, we will survey the related results in three sub-cases, covering different stochastic or deterministic access to objective and constraints. When we mention oracle or sample complexity results in the sequel, we always consider the complexity for obtaining an  $\varepsilon$ -stationary point, in view of the definition in Section 1.2. Deterministic objective and deterministic constraints. The setting when objective f in (1.1) is deterministic is the most well-studied with many results in the classical literature (Bertsekas, 2016). Recent work focused on characterizing the global oracle complexity of Lagrangian or augmented Lagrangian algorithms. With nonlinear and nonconvex constraints, many of the existing algorithms analyzing AL-based algorithms need to rely on strong constraint qualification and boundedness assumptions and use large penalty parameters to ensure feasibility (Li et al., 2021; Lin et al., 2022; Kong et al., 2019; Kong & Monteiro, 2023; Kong et al., 2023). The existing frameworks so far fail to capture the importance of dual variable updates, which are, in fact, the main reason behind the ability to use constant penalty parameters while ensuring convergence, see for example Bertsekas (2014). The recent works mentioned above obtained the complexity bounds  $O(\varepsilon^{-3})$  for general nonlinear constraints with no specialization for linear constraints. When specialized to convex functional constraints, the best-known rate for these methods has been  $O(\varepsilon^{-2.5})$  (Lin et al., 2022).

In the case when the constraints are linear, such as (1.1) with  $X = \mathbb{R}^n$ , the work of Hong (2016) managed to analyze ALM with constant penalty parameters and non-negligible dual updates to get optimal complexity  $O(\varepsilon^{-2})$ . The case of  $X \neq \mathbb{R}^n$  turned out to be significantly more challenging with many works focusing on variants of ALM with large penalty parameters (depending on the inverse of the final accuracy) to ensure near-feasibility and *negligible* dual updates that do not help with feasibility and obtaining the suboptimal complexity  $\tilde{O}(\varepsilon^{-2.5})$  (Kong & Monteiro, 2023; Kong et al., 2023). The exceptions are the works Zhang & Luo (2020, 2022) that showed, for the case X polyhedral, near-optimal complexity  $O(\varepsilon^{-2})$  with a constant penalty parameter and dual steps with constant step sizes, with no constraint qualification. The key step was the global error bound that our work also relied on.

Stochastic objective and deterministic constraints One important step in generalizing the template to tasks arising in machine learning was to consider stochastic objectives where we have access to an unbiased gradient. With general nonlinear constraints and Lipschitzness of  $\nabla f$ , the optimal sample complexity is  $O(\varepsilon^{-4})$  which is obtained with double loop algorithms (Curtis et al., 2024; Boob et al., 2023; Ma et al., 2020). These works also come with strong assumptions on the boundedness of the primal domain as well as constraint qualifications, which are often not necessary with linear constraints.

Another set of results concern stochastic optimization with deterministic nonlinear constraints with penalty-based algorithms and, requiring large penalty parameters to ensure near-feasibility rather than dual updates (Lu et al., 2024; Alacaoglu & Wright, 2024). These works assume expected Lipschitzness of the stochastic gradients, stated in Assumption 5.2, which is stronger than Lipschitzness of  $\nabla f$  (we will unpack this further in the sequel). Since these works focuses on nonlinear functional constraints, the analysis requires strong boundedness assumptions as well as constraint qualifications, unlike our results in Section 3 for deterministic linear constraints.

One of the most related to our setting is Alacaoglu & Wright (2024) that considered an augmented Lagrangian algorithm with a constant penalty parameter and non-negligible dual updates and obtained the complexity  $O(\varepsilon^{-3})$  for linear equality constraints and expected Lipschitzness. In particular, this work only covered the case  $X = \mathbb{R}^n$  and left open the question of handling the case of more general X (see (Alacaoglu & Wright, 2024, Section 5)).

In this work, we address an important special case of this open question when X is polyhedral, allowing our analysis to cover linear inequality constraints. The work of Alacaoglu & Wright (2024) focused on applying variance reduction on estimation of the gradient of f, which means that the assumption on the stochastic gradients was Assumption 5.2, stronger than Assumption 1.1. We show in Section 5 how to obtain the same optimal complexity as this paper while handling the case when X is polyhedral to cover problems with linear inequality constraints, which cannot be solved by Alacaoglu & Wright (2024).

Moreover, we also get the complexity  $O(\varepsilon^{-4})$  under Assumption 1.1. This complexity is optimal under Assumption 1.1 and we refer to Arjevani et al. (2023) for further details on the lower bounds. In contrast, the work in (Alacaoglu & Wright, 2024) does not have guarantees without Assumption 5.2. Stochastic objective and stochastic constraints. This is the most general class, where the existing results come with many assumptions that are not always easy to interpret, similar to the case of stochastic objective and deterministic nonconvex functional constraints described above (Li et al., 2024; Alacaoglu & Wright, 2024). The state-of-the-art complexity result is  $O(\varepsilon^{-5})$  and is obtained by using the expected Lipschitzness assumption above, by an inexact, double-loop, augmented Lagrangian algorithm in Li et al. (2024) and by a single loop penalty algorithm in Alacaoglu & Wright (2024). These results concerning augmented Lagrangian methods all need to use large penalty parameters, which renders them as penalty methods since the dual updates do not contribute to the analysis for ensuring the feasibility. Other approaches for solving this sub-case also require double-loop algorithms and stronger assumptions since they focus on a generic nonconvex constraint (Boob et al., 2023; Ma et al., 2020). These works obtain the complexity  $O(\varepsilon^{-6})$  since they do not assume expected Lipschitzness.

In conclusion, in this sub-case, none of the existing surveyed results used the structure of linear constraints, which we do in Section 4 to achieve improved complexity guarantees.

### Acknowledgements

J. Zhang was supported by the MIT School of Engineering Postdoctoral Fellowship Program for Engineering Excellence.

### References

- Alacaoglu, A. and Wright, S. J. Complexity of single loop algorithms for nonlinear programming with stochastic objective and constraints. In *International Conference on Artificial Intelligence and Statistics*, pp. 4627–4635. PMLR, 2024.
- Arjevani, Y., Carmon, Y., Duchi, J. C., Foster, D. J., Srebro, N., and Woodworth, B. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1):165–214, 2023.
- Bertsekas, D. Constrained optimization and Lagrange multiplier methods. Academic press, 2014.
- Bertsekas, D. Nonlinear Programming, volume 4. Athena Scientific, 2016.
- Bertsekas, D., Nedic, A., and Ozdaglar, A. *Convex analysis and optimization*, volume 1. Athena Scientific, 2003.
- Boob, D., Deng, Q., and Lan, G. Stochastic first-order methods for convex and nonconvex functional constrained optimization. *Mathematical Programming*, 197(1):215–279, 2023.
- Chaintreau, A., Mtibaa, A., Massoulie, L., and Diot, C. The diameter of opportunistic mobile networks. In *Proceedings of the 2007 ACM CoNEXT Conference*, CoNEXT '07, New York, NY, USA, 2007. Association for Computing Machinery. ISBN 9781595937704. doi: 10.1145/1364654.1364670. URL https://doi.org/10.1145/1364654.1364670.
- Chen, C., Zhang, J., Shen, L., Zhao, P., and Luo, Z. Communication efficient primal-dual algorithm for nonconvex nonsmooth distributed optimization. In Banerjee, A. and Fukumizu, K. (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1594–1602. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/chen21c.html.
- Curtis, F. E., O'Neill, M. J., and Robinson, D. P. Worst-case complexity of an sqp method for nonlinear equality constrained stochastic optimization. *Mathematical Programming*, 205(1):431–483, 2024.
- Cutkosky, A. and Orabona, F. Momentum-based variance reduction in non-convex sgd. Advances in neural information processing systems, 32, 2019.

- Davis, D. and Drusvyatskiy, D. Stochastic model-based minimization of weakly convex functions. SIAM Journal on Optimization, 29(1):207–239, 2019.
- Dener, A., Miller, M. A., Churchill, R. M., Munson, T., and Chang, C.-S. Training neural networks under physical constraints using a stochastic augmented lagrangian approach. arXiv preprint arXiv:2009.07330, 2020.
- Drusvyatskiy, D. and Paquette, C. Efficiency of minimizing compositions of convex functions and smooth maps. *Mathematical Programming*, 178:503–558, 2019.
- Hestenes, M. R. Multiplier and gradient methods. *Journal of optimization theory and applications*, 4(5): 303–320, 1969.
- Hiriart-Urruty, J.-B. and Lemarechal, C. Convex Analysis and Minimization Algorithms II: Advanced Theory and Bundle Methods, volume 306. Springer Berlin Heidelberg, Berlin, Heidelberg, 1st 1993.;1; edition, 1993. ISBN 0072-7830.
- Hong, M. Decomposing linearly constrained nonconvex problems by a proximal primal dual approach: Algorithms, convergence, and applications. *arXiv preprint arXiv:1604.00543*, 2016.
- Katz-Samuels, J., Nakhleh, J. B., Nowak, R., and Li, Y. Training OOD detectors in their natural habitats. In *ICML*, 2022.
- Kong, W. and Monteiro, R. D. An accelerated inexact dampened augmented lagrangian method for linearlyconstrained nonconvex composite optimization problems. *Computational Optimization and Applications*, 85(2):509–545, 2023.
- Kong, W., Melo, J. G., and Monteiro, R. D. Complexity of a quadratic penalty accelerated inexact proximal point method for solving linearly constrained nonconvex composite programs. SIAM Journal on Optimization, 29(4):2566–2593, 2019.
- Kong, W., Melo, J. G., and Monteiro, R. D. Iteration complexity of an inner accelerated inexact proximal augmented lagrangian method based on the classical lagrangian function. SIAM Journal on Optimization, 33(1):181–210, 2023.
- Krause, N. and Singer, Y. Leveraging the margin more carefully. In Proceedings of the Twenty-First International Conference on Machine Learning, ICML '04, pp. 63, New York, NY, USA, 2004. Association for Computing Machinery. ISBN 1581138385. doi: 10.1145/1015330.1015344. URL https://doi.org/10.1145/1015330.1015344.
- Lan, G. First-order and stochastic optimization methods for machine learning. Springer, 2020.
- Li, Z., Chen, P.-Y., Liu, S., Lu, S., and Xu, Y. Rate-improved inexact augmented lagrangian method for constrained nonconvex optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 2170–2178. PMLR, 2021.
- Li, Z., Chen, P.-Y., Liu, S., Lu, S., and Xu, Y. Stochastic inexact augmented lagrangian method for nonconvex expectation constrained optimization. *Computational Optimization and Applications*, 87(1): 117–147, 2024.
- Lin, Q., Ma, R., and Xu, Y. Complexity of an inexact proximal-point penalty method for constrained smooth non-convex optimization. Computational optimization and applications, 82(1):175–224, 2022.
- Lu, Z., Mei, S., and Xiao, Y. Variance-reduced first-order methods for deterministically constrained stochastic nonconvex optimization with strong convergence guarantees. arXiv preprint arXiv:2409.09906, 2024.

- Ma, R., Lin, Q., and Yang, T. Quadratically regularized subgradient methods for weakly convex optimization with weakly convex constraints. In *International Conference on Machine Learning*, pp. 6554–6564. PMLR, 2020.
- Nocedal, J. and Wright, S. J. Numerical optimization. Springer, 1999.
- Ouyang, Y., Chen, Y., Lan, G., and Pasiliao Jr, E. An accelerated linearized alternating direction method of multipliers. SIAM Journal on Imaging Sciences, 8(1):644–681, 2015.
- Planiden, C. and Wang, X. Strongly convex functions, moreau envelopes, and the generic nature of convex functions with strong minimizers. SIAM Journal on Optimization, 26(2):1341–1364, 2016.
- Rockafellar, R. T. Extended nonlinear programming. Nonlinear optimization and related topics, pp. 381–399, 2000.
- Yan, Y. and Xu, Y. Adaptive primal-dual stochastic gradient method for expectation-constrained convex stochastic programs. *Mathematical Programming Computation*, 14(2):319–363, 2022.
- Zafar, M. B., Valera, I., Rogriguez, M. G., and Gummadi, K. P. Fairness Constraints: Mechanisms for Fair Classification. In Singh, A. and Zhu, J. (eds.), Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, volume 54 of Proceedings of Machine Learning Research, pp. 962–970. PMLR, 20–22 Apr 2017. URL https://proceedings.mlr.press/v54/zafar17a.html.
- Zhang, J. and Luo, Z.-Q. A proximal alternating direction method of multiplier for linearly constrained nonconvex minimization. SIAM Journal on Optimization, 30(3):2272–2302, 2020.
- Zhang, J. and Luo, Z.-Q. A global dual error bound and its application to the analysis of linearly constrained nonconvex optimization. SIAM Journal on Optimization, 32(3):2319–2346, 2022. doi: 10.1137/20M135474X. URL https://doi.org/10.1137/20M135474X.
- Zhang, J., Xiao, P., Sun, R., and Luo, Z. A single-loop smoothed gradient descent-ascent algorithm for nonconvex-concave min-max problems. Advances in neural information processing systems, 33:7377–7389, 2020.
- Zhang, J., Ge, S., Chang, T.-H., and Luo, Z.-Q. Decentralized non-convex learning with linearly coupled constraints: Algorithm designs and application to vertical learning problem. *IEEE Transactions on Signal Processing*, 70:3312–3327, 2022.
- Zhang, N., Liu, Y.-F., Farmanbar, H., Chang, T.-H., Hong, M., and Luo, Z.-Q. Network slicing for serviceoriented networks under resource constraints. *IEEE Journal on Selected Areas in Communications*, 35 (11):2512–2521, 2017. doi: 10.1109/JSAC.2017.2760147.
- Zhao, L., Mammadov, M., and Yearwood, J. From convex to nonconvex: A loss function analysis for binary classification. In *Proceedings of the 2010 IEEE International Conference on Data Mining Workshops*, ICDMW '10, pp. 1281–1288, USA, 2010. IEEE Computer Society. ISBN 9780769542577. doi: 10.1109/ ICDMW.2010.57. URL https://doi.org/10.1109/ICDMW.2010.57.

## A Proofs for Section 3

Let us recall the following definition from (5.3) which will be used extensively in the proofs

$$K(\mathbf{x}, \mathbf{y}, \mathbf{z}) = L_{\rho}(\mathbf{x}, \mathbf{y}) + \frac{\mu}{2} \|\mathbf{x} - \mathbf{z}\|^2$$

With this notation, we have

$$\mathbf{u}^*(\mathbf{x}, \mathbf{y}, \mathbf{z}) = \min_{\mathbf{u} \in X} K(\mathbf{u}, \mathbf{y}, \mathbf{z}) + \frac{\lambda}{2} \|\mathbf{u} - \mathbf{x}\|^2.$$

We also introduce here some parameters that are used throughout, for convenience.

$$\mu = \max\{2, 4L_f\}, 
L_K = L_f + \rho ||A|| + \mu, 
\lambda = L_K, 
\sigma_4 = \frac{\mu - L_f}{\mu}, 
\tau = \frac{1}{6\lambda^2 \sqrt{T}}, 
\eta = \min\left\{\frac{2\mu + \rho ||A||}{4||A||^4}, \frac{\tau}{200||A||^2}, \frac{\tau(2\mu + \rho ||A||^2)}{20||A||^2}\right\}, 
\beta = \min\left\{\frac{\tau}{100}, \frac{1}{50\lambda}, \frac{\eta}{36\mu\bar{\sigma}^2}\right\}, 
\gamma_s = 2\mu + \rho ||A||, \gamma = \frac{(\mu - L_f)\lambda}{\mu - L_f + \lambda}, \gamma_K = \mu - L_f,$$
(A.1)

where  $\bar{\sigma}$  is defined in A.9.

### A.1 Proofs for Lemma 3.6

In the next lemma, the first part is using the idea of Davis & Drusvyatskiy (2019) to handle bounded variance assumption instead of the restricted bounded stochastic gradient assumption. The second part of the lemma also follows a similar idea as this work, with the exception of the dependence of the changing center point  $\mathbf{z}_t$ . This introduces additional issues, since the stochastic gradient in the update of  $\mathbf{x}_{t+1}$  depends on  $\mathbf{z}_t$  whereas the proximal point  $\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_{t+1})$  depends on  $\mathbf{z}_{t+1}$ . Our analysis below estimates this additional error and shows it to be in the order of  $\|\mathbf{z}_{t+1} - \mathbf{z}_t\|^2$ , which will be handled later.

**Lemma A.1.** Suppose that Assumption 1.1 holds, for the proximal point  $\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t)$ , defined in (3.4) we have the characterization

$$\mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) = \operatorname{proj}_{X}(\tau \lambda \mathbf{x}_{t} + (1 - \tau \lambda)\mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) - \tau \nabla_{\mathbf{x}} K(\mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}), \mathbf{y}_{t+1}, \mathbf{z}_{t+1})).$$
(A.2)

Moreover, for the sequence  $\mathbf{x}_{t+1}$  calculated as Algorithm 1, if  $\lambda = L_K = L_f + \rho \|A\|^2 + \mu$  and  $\tau \leq \frac{1}{6\lambda}$ , we have

$$\mathbb{E}\|\mathbf{u}^{*}(\mathbf{x}_{t},\mathbf{y}_{t+1},\mathbf{z}_{t+1}) - \mathbf{x}_{t+1}\|^{2} \leq (1 - \frac{\tau\lambda}{4})\mathbb{E}\|\mathbf{u}^{*}(\mathbf{x}_{t},\mathbf{y}_{t+1},\mathbf{z}_{t+1}) - \mathbf{x}_{t}\|^{2} + (\tau\mu + 2\tau^{2}\mu^{2})\mathbb{E}\|\mathbf{z}_{t} - \mathbf{z}_{t+1}\|^{2} + \tau^{2}\sigma^{2}$$

*Proof.* From the definition of  $\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_{t+1})$  in (3.4), we have

$$\lambda(\mathbf{x}_t - \mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_{t+1})) \in \nabla_{\mathbf{x}} K(\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}), \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) + \partial I_X(\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_{t+1})).$$

Multiplying both sides by the step size  $\tau$  and rearranging give

$$\tau \lambda \mathbf{x}_{t} - \tau \nabla_{\mathbf{x}} K(\mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}), \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) + (1 - \tau \lambda) \mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) \\ \in \mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) + \tau \partial I_{X}(\mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t+1})).$$

Since  $(I + \tau \partial I_X)^{-1} = \operatorname{prox}_{I_X} = \operatorname{proj}_X$  due to  $\partial I_X$  being the normal cone and proximal operator of a normal cone being the projection to the set, we have the first assertion.

We next establish the second assertion. Using the just established identity (A.2), the update rule of  $\mathbf{x}_{t+1}$  and nonexpansiveness of the projection, we derive

$$\begin{aligned} \|\mathbf{u}^{*}(\mathbf{x}_{t},\mathbf{y}_{t+1},\mathbf{z}_{t+1}) - \mathbf{x}_{t+1}\|^{2} \\ \leq \|\tau\lambda\mathbf{x}_{t} + (1-\tau\lambda)\mathbf{u}^{*}(\mathbf{x}_{t},\mathbf{y}_{t+1},\mathbf{z}_{t+1}) - \tau\nabla_{\mathbf{x}}K(\mathbf{u}^{*}(\mathbf{x}_{t},\mathbf{y}_{t+1},\mathbf{z}_{t+1}),\mathbf{y}_{t+1},\mathbf{z}_{t+1}) - [\mathbf{x}_{t} - \tau G(\mathbf{x}_{t},\mathbf{y}_{t+1},\mathbf{z}_{t},\xi_{t})]\|^{2}. \end{aligned}$$

We add and subtract  $\nabla_{\mathbf{x}} K(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t)$  inside the squared norm, expand and take conditional expectation to obtain

$$\begin{split} & \mathbb{E}_{t} \| \mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) - \mathbf{x}_{t+1} \|^{2} \\ & = \| (1 - \tau \lambda) (\mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) - \mathbf{x}_{t}) - \tau \nabla_{\mathbf{x}} K(\mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}), \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) + \tau \nabla_{\mathbf{x}} K(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t}) \|^{2} \\ & + \tau^{2} \mathbb{E}_{t} \| G(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t}, \xi_{t}) - \nabla_{\mathbf{x}} K(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t}) \|^{2}. \end{split}$$

where the cross term disappeared because

$$\mathbb{E}_t[G(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t, \xi_t)] = \nabla_{\mathbf{x}} K(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t)$$

and  $\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}, \mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_{t+1})$  are deterministic under the conditioning since  $\mathbf{z}_{t+1}$  only depends on  $\mathbf{x}_t$ . The second term on the right-hand side is trivially bounded by the oracle assumptions, that is,

$$\mathbb{E}_t \| G(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}, \xi_t) - \nabla_{\mathbf{x}} K(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) \|^2 \le \sigma^2.$$

For the first term, we further estimate as

$$\begin{aligned} \| (1 - \tau \lambda) (\mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) - \mathbf{x}_{t}) - \tau \nabla_{\mathbf{x}} K(\mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}), \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) + \tau \nabla_{\mathbf{x}} K(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t}) \|^{2} \\ &\leq (1 - \tau \lambda)^{2} \| \mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) - \mathbf{x}_{t} \|^{2} \\ &+ \tau (1 - \tau \lambda) \langle \mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) - \mathbf{x}_{t}, \nabla_{\mathbf{x}} K(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t}) - \nabla_{\mathbf{x}} K(\mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}), \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) \rangle \\ &+ \tau^{2} \| \nabla_{\mathbf{x}} K(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t}) - \nabla_{\mathbf{x}} K(\mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}), \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) \|^{2}. \end{aligned}$$
(A.3)

Next, we turn to estimating

$$\begin{aligned} \|\nabla_{\mathbf{x}}K(\mathbf{x}_{t},\mathbf{y}_{t+1},\mathbf{z}_{t}) - \nabla_{\mathbf{x}}K(\mathbf{u}^{*}(\mathbf{x}_{t},\mathbf{y}_{t+1},\mathbf{z}_{t+1}),\mathbf{y}_{t+1},\mathbf{z}_{t+1})\| \\ &\leq \|\nabla_{\mathbf{x}}K(\mathbf{x}_{t},\mathbf{y}_{t+1},\mathbf{z}_{t}) - \nabla_{\mathbf{x}}K(\mathbf{x}_{t},\mathbf{y}_{t+1},\mathbf{z}_{t+1})\| + \|\nabla_{\mathbf{x}}K(\mathbf{x}_{t},\mathbf{y}_{t+1},\mathbf{z}_{t+1}) - \nabla_{\mathbf{x}}K(\mathbf{u}^{*}(\mathbf{x}_{t},\mathbf{y}_{t+1},\mathbf{z}_{t+1}),\mathbf{y}_{t+1},\mathbf{z}_{t+1})\|. \end{aligned}$$

Note that, by definition, we have

$$\nabla_{\mathbf{x}} K(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) - \nabla_{\mathbf{x}} K(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) = \mu(\mathbf{z}_{t+1} - \mathbf{z}_t).$$

Using this and the Lipschitzness of  $\nabla_{\mathbf{x}} K(\cdot, \mathbf{y}_{t+1}, \mathbf{z}_{t+1})$ , we then obtain

$$\|\nabla_{\mathbf{x}}K(\mathbf{x}_{t},\mathbf{y}_{t+1},\mathbf{z}_{t}) - \nabla_{\mathbf{x}}K(\mathbf{u}^{*}(\mathbf{x}_{t},\mathbf{y}_{t+1},\mathbf{z}_{t+1}),\mathbf{y}_{t+1},\mathbf{z}_{t+1})\| \leq \mu \|\mathbf{z}_{t+1} - \mathbf{z}_{t}\| + L_{K}\|\mathbf{u}^{*}(\mathbf{x}_{t},\mathbf{y}_{t+1},\mathbf{z}_{t+1}) - \mathbf{x}_{t}\|.$$

Plug this bound into the second term in the right-hand side of (A.3) after using Cauchy-Schwarz and Young's inequalities, we get

$$\begin{aligned} &\tau(1-\tau\lambda)\langle \mathbf{u}^{*}(\mathbf{x}_{t},\mathbf{y}_{t+1},\mathbf{z}_{t+1})-\mathbf{x}_{t},\nabla_{\mathbf{x}}K(\mathbf{x}_{t},\mathbf{y}_{t+1},\mathbf{z}_{t})-\nabla_{\mathbf{x}}K(\mathbf{u}^{*}(\mathbf{x}_{t},\mathbf{y}_{t+1},\mathbf{z}_{t+1}),\mathbf{y}_{t+1},\mathbf{z}_{t+1})\rangle \\ &\leq \tau(1-\tau\lambda)\|\mathbf{u}^{*}(\mathbf{x}_{t},\mathbf{y}_{t+1},\mathbf{z}_{t+1})-\mathbf{x}_{t}\|(\mu\|\mathbf{z}_{t+1}-\mathbf{z}_{t}\|+L_{K}\|\mathbf{u}^{*}(\mathbf{x}_{t},\mathbf{y}_{t+1},\mathbf{z}_{t+1})-\mathbf{x}_{t}\|) \\ &\leq \tau(1-\tau\lambda)(L_{K}+\mu/2)\|\mathbf{u}^{*}(\mathbf{x}_{t},\mathbf{y}_{t+1},\mathbf{z}_{t+1})-\mathbf{x}_{t}\|^{2}+\frac{\tau(1-\tau\lambda)\mu}{2}\|\mathbf{z}_{t+1}-\mathbf{z}_{t}\|^{2}.\end{aligned}$$

Using the last two inequalities in (A.3), we obtain

$$\begin{aligned} &\| (1-\tau\lambda) (\mathbf{u}^* (\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) - \mathbf{x}_t) - \tau \nabla_{\mathbf{x}} K (\mathbf{u}^* (\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}), \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) + \tau \nabla_{\mathbf{x}} K (\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) \|^2 \\ &\leq [(1-\tau\lambda)^2 + \tau (1-\tau\lambda) (L_K + \mu) + 2\tau^2 L_K^2] \| \mathbf{u}^* (\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) - \mathbf{x}_t \|^2 + (\tau (1-\tau\lambda) (\mu + 2\tau^2 \mu^2) \| \mathbf{z}_{t+1} - \mathbf{z}_t \|^2. \end{aligned}$$

We estimate the coefficient of the first term. First, note that  $\tau \leq \frac{1}{\lambda}$  and  $\mu \leq L_K = \lambda$ . As a result, we have

$$\begin{aligned} (1-\tau\lambda)^2 + \tau(1-\tau\lambda)(L_K+\mu/2) + 2\tau^2 L_K^2 &\leq 1 - 2\tau\lambda + \tau^2\lambda^2 + \frac{3\tau L_K}{2} - \frac{3\tau^2\lambda L_K}{2} + 2\tau^2 L_K^2 \\ &\leq 1 - \frac{\tau\lambda}{2} + \tau^2\lambda^2 + \frac{\tau^2 L_K^2}{2} \\ &\leq 1 - \frac{\tau\lambda}{4}, \end{aligned}$$

since  $\tau \leq \frac{1}{6\lambda}$ . Finally, since  $\tau(1-\tau\lambda)\mu + 2\tau^2\mu^2 \leq \tau\mu + 2\tau^2\mu^2$ , the proof is completed after taking full expectation of the resulting equality. 

**Lemma A.2.** Let Assumption 1.1 hold, then if  $\lambda = L_K$  and  $\tau \leq \frac{1}{6\lambda}$  we have

$$\mathbb{E}\varphi_{1/\lambda}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) \leq \mathbb{E}\varphi_{1/\lambda}(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) - \frac{\tau\lambda^2}{16} \mathbb{E}\|\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{x}_t\|^2 + (\lambda\tau\mu + 2\lambda\tau^2\mu^2 + \frac{\tau\lambda^2\mu^2}{8\gamma_s^2})\mathbb{E}\|\mathbf{z}_t - \mathbf{z}_{t+1}\|^2 + \lambda\tau^2\sigma^2,$$
(A.4)

where  $\gamma_s = 2\mu + \rho \|A\|$ .

*Proof.* By the definition of  $\varphi_{1/\lambda}$  and  $\mathbf{u}^*(\mathbf{x}, \mathbf{y}_{t+1}, \mathbf{z}_{t+1})$ , we have

$$\mathbb{E}\varphi_{1/\lambda}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) \leq \mathbb{E}K(\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}), \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) + \frac{\lambda}{2}\mathbb{E}\|\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) - \mathbf{x}_{t+1}\|^2 \\
\leq \mathbb{E}K(\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}), \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) + \left(\frac{\lambda}{2} - \frac{\tau\lambda^2}{8}\right)\mathbb{E}\|\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) - \mathbf{x}_t\|^2 \\
+ (\lambda\tau\mu + 2\lambda\tau^2\mu^2)\mathbb{E}\|\mathbf{z}_t - \mathbf{z}_{t+1}\|^2 + \lambda\tau^2\sigma^2 \\
= \mathbb{E}\varphi_{1/\lambda}(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) - \frac{\tau\lambda^2}{8}\mathbb{E}\|\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) - \mathbf{x}_t\|^2 \\
+ (\lambda\tau\mu + 2\lambda\tau^2\mu^2)\mathbb{E}\|\mathbf{z}_t - \mathbf{z}_{t+1}\|^2 + \lambda\tau^2\sigma^2.$$
(A.5)

We next bound the second term on the right-hand side by using

$$\|\mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) - \mathbf{x}_{t}\|^{2} = \|\mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) - \mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t}) + \mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t}) - \mathbf{x}_{t}\|^{2}$$

$$\geq \frac{1}{2} \|\mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t}) - \mathbf{x}_{t}\|^{2} - \|\mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) - \mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t})\|^{2}$$

$$\geq \frac{1}{2} \|\mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t}) - \mathbf{x}_{t}\|^{2} - \frac{p^{2}}{\gamma_{s}^{2}} \|\mathbf{z}_{t} - \mathbf{z}_{t+1}\|^{2}, \qquad (A.6)$$

(A.7)

where the last line used (A.32).

We substitute the last inequality into (A.5) to conclude.

Since the previous result only allowed us to connect  $\varphi_{1/\lambda}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}, \mathbf{z}_{t+1})$  to  $\varphi_{1/\lambda}(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_{t+1})$ , we now need to analyze the effect of changing  $\mathbf{y}_{t+1}$  and  $\mathbf{z}_{t+1}$  in  $\varphi_{1/\lambda}$ . The main idea of this lemma is similar to Zhang & Luo (2022), where the difference lies in the fact that our potential involves the Moreau envelope of  $K(\mathbf{x}, \mathbf{y}, \mathbf{z})$  whereas the potential of Zhang & Luo (2022) involves  $K(\mathbf{x}, \mathbf{y}, \mathbf{z})$  hence this work considers the change of the arguments in the function K instead of  $\varphi_{1/\lambda}$ . Therefore, our proof uses the properties of the Moreau envelope which was not needed in Zhang & Luo (2022).

**Lemma A.3.** Suppose that Assumption 1.1 holds, for  $\varphi_{1/\lambda}$  defined in (3.3), we have that

$$\begin{split} \varphi_{1/\lambda}(\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t) &- \varphi_{1/\lambda}(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) \geq \langle \mathbf{y}_t - \mathbf{y}_{t+1}, A \mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t) - b \rangle + \frac{\gamma_s}{2} \| \mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t) - \mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) \|^2, \\ \varphi_{1/\lambda}(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) - \varphi_{1/\lambda}(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) \geq \frac{\mu}{2} \langle \mathbf{z}_{t+1} - \mathbf{z}_t, 2 \mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{z}_{t+1} - \mathbf{z}_t \rangle \\ &+ \frac{\gamma_s}{2} \| \mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) - \mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) \|^2, \end{split}$$

where  $\gamma_s = 2\mu + \rho \|A\|$ .

*Proof.* We first consider the change in **y** argument of  $\varphi_{1/\lambda}$ . By using the definition of  $\varphi_{1/\lambda}$ , we have

$$\varphi_{1/\lambda}(\mathbf{x}_{t}, \mathbf{y}_{t}, \mathbf{z}_{t}) - \varphi_{1/\lambda}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t}) = K(\mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t}, \mathbf{z}_{t}), \mathbf{y}_{t}, \mathbf{z}_{t}) + \frac{\lambda}{2} \|\mathbf{x}_{t} - \mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t}, \mathbf{z}_{t})\|^{2} - K(\mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t}), \mathbf{y}_{t+1}, \mathbf{z}_{t}) - \frac{\lambda}{2} \|\mathbf{x}_{t} - \mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t})\|^{2} = K(\mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t}, \mathbf{z}_{t}), \mathbf{y}_{t}, \mathbf{z}_{t}) - K(\mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t}, \mathbf{z}_{t}), \mathbf{y}_{t+1}, \mathbf{z}_{t}) + K(\mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t}, \mathbf{z}_{t}), \mathbf{y}_{t+1}, \mathbf{z}_{t}) + \frac{\lambda}{2} \|\mathbf{x}_{t} - \mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t}, \mathbf{z}_{t})\|^{2} - K(\mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t}), \mathbf{y}_{t+1}, \mathbf{z}_{t}) - \frac{\lambda}{2} \|\mathbf{x}_{t} - \mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t})\|^{2}, \quad (A.8)$$

where the second equality adds and subtracts  $K(\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t), \mathbf{y}_{t+1}, \mathbf{z}_t)$ .

From the definition of  $\mathbf{y}_{t+1}$ , it trivially follows that

$$K(\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t), \mathbf{y}_t, \mathbf{z}_t) - K(\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t), \mathbf{y}_{t+1}, \mathbf{z}_t) = \langle \mathbf{y}_t - \mathbf{y}_{t+1}, A\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t) - \mathbf{b} \rangle.$$

Next, we use the property that  $K(\cdot, \mathbf{y}_{t+1}, \mathbf{z}_t) + \frac{\lambda}{2} \| \cdot - \mathbf{x}_t \|^2$  is  $\gamma_s$ -strongly convex with minimizer  $\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t)$  to obtain

$$\begin{split} &K(\mathbf{u}^{*}(\mathbf{x}_{t},\mathbf{y}_{t},\mathbf{z}_{t}),\mathbf{y}_{t+1},\mathbf{z}_{t}) + \frac{\lambda}{2} \|\mathbf{x}_{t} - \mathbf{u}^{*}(\mathbf{x}_{t},\mathbf{y}_{t},\mathbf{z}_{t})\|^{2} - K(\mathbf{u}^{*}(\mathbf{x}_{t},\mathbf{y}_{t+1},\mathbf{z}_{t}),\mathbf{y}_{t+1},\mathbf{z}_{t}) - \frac{\lambda}{2} \|\mathbf{x}_{t} - \mathbf{u}^{*}(\mathbf{x}_{t},\mathbf{y}_{t+1},\mathbf{z}_{t})\|^{2} \\ &\geq \frac{\gamma_{s}}{2} \|\mathbf{u}^{*}(\mathbf{x}_{t},\mathbf{y}_{t},\mathbf{z}_{t}) - \mathbf{u}^{*}(\mathbf{x}_{t},\mathbf{y}_{t+1},\mathbf{z}_{t})\|^{2}. \end{split}$$

Combining the last two estimates in (A.8) gives the first assertion.

Next, we analyze the effect of changing the  $\mathbf{z}$  component in  $\varphi_{1/\lambda}$ . Similar to the proof of the first assertion, we start with the definition of  $\varphi_{1/\lambda}$  and then add and subtract  $K(\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}))$  to obtain

$$\begin{split} \varphi_{1/\lambda}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t}) &- \varphi_{1/\lambda}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) \\ = K(\mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t}), \mathbf{y}_{t+1}, \mathbf{z}_{t}) + \frac{\lambda}{2} \|\mathbf{x}_{t} - \mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t})\|^{2} \\ &- K(\mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}), \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) - \frac{\lambda}{2} \|\mathbf{x}_{t} - \mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t+1})\|^{2} \\ &= K(\mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t}), \mathbf{y}_{t+1}, \mathbf{z}_{t}) - K(\mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t}), \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) \\ &+ K(\mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t}), \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) + \frac{\lambda}{2} \|\mathbf{x}_{t} - \mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t})\|^{2} \\ &- K(\mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}), \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) - \frac{\lambda}{2} \|\mathbf{x}_{t} - \mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t+1})\|^{2}. \end{split}$$
(A.9)

First, by definition, of K, it trivially follows that

$$K(\mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t}), \mathbf{y}_{t+1}, \mathbf{z}_{t}) - K(\mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t}), \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) = \frac{\mu}{2} \|\mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t}) - \mathbf{z}_{t}\|^{2} - \frac{\mu}{2} \|\mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t}) - \mathbf{z}_{t+1}\|^{2}.$$

For the remaining terms in the right-hand side, we again use that  $K(\cdot, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) + \frac{\lambda}{2} \| \cdot - \mathbf{x}_t \|^2$  is  $\gamma_s$ -strongly convex with minimizer  $\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_{t+1})$  to deduce

$$K(\mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t}), \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) + \frac{\lambda}{2} \|\mathbf{x}_{t} - \mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t})\|^{2} - K(\mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}), \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) - \frac{\lambda}{2} \|\mathbf{x}_{t} - \mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t+1})\|^{2} \geq \frac{\gamma_{s}}{2} \|\mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) - \mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t})\|^{2}.$$

Plugging in the last two estimates in (A.9) gives the second assertion.

**Corollary A.4.** Suppose that Assumption 1.1 holds, for  $\varphi_{1/\lambda}$  defined in (3.3), we have that

$$\begin{split} \varphi_{1/\lambda}(\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t) - \varphi_{1/\lambda}(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) &\geq \frac{\tau \lambda^2}{16} \mathbb{E} \| \mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{x}_t \|^2 \\ &- (\lambda \tau \mu + 2\lambda \tau^2 \mu^2 + \frac{\tau \lambda^2 \mu^2}{8\gamma_s^2}) \mathbb{E} \| \mathbf{z}_t - \mathbf{z}_{t+1} \|^2 - \lambda \tau^2 \sigma^2 \\ &- \eta \langle A \mathbf{x}_t - \mathbf{b}, A \mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t) - b \rangle \\ &+ \frac{\mu}{2} \langle \mathbf{z}_{t+1} - \mathbf{z}_t, 2 \mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{z}_{t+1} - \mathbf{z}_t \rangle, \end{split}$$

where  $\gamma_s = 2\mu + \rho \|A\|$ .

*Proof.* We sum up the results in Lemma A.2 and Lemma A.3, plug in the definition of  $\mathbf{y}_{t+1}$  and discard two nonnegative terms on the right-hand side to get the result.

Next, we analyze the rest of the terms appearing in the potential function. This lemma is only using the definition of  $d(\mathbf{y}, \mathbf{z})$  and  $\Psi(\mathbf{z})$  and is equivalent to Zhang & Luo (2022) and hence we omit its proof. Notably, these bounds are agnostic to the algorithm used to generate the sequences. Note that the only difference is that in the result below, we do not use the definition of  $\mathbf{y}_{t+1}$  whereas the proof in Zhang & Luo (2022) uses this definition. The rest of the estimations are precisely the same.

**Lemma A.5.** (*Zhang & Luo, 2020, Lemma 3.2, Lemma 3.3*) For the functions  $d(\mathbf{y}, \mathbf{z})$  and  $\Psi(\mathbf{z})$  defined in (3.5) and (2.1), we have

$$d(\mathbf{y}_{t+1}, \mathbf{z}_{t+1}) - d(\mathbf{y}_t, \mathbf{z}_t) \ge \eta \langle A\mathbf{x}_t - \mathbf{b}, A\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{b} \rangle + \frac{\mu}{2} \langle \mathbf{z}_{t+1} - \mathbf{z}_t, \mathbf{z}_{t+1} + \mathbf{z}_t - 2\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_{t+1}) \rangle,$$
  
$$\Psi(\mathbf{z}_{t+1}) - \Psi(\mathbf{z}_t) \le \mu \langle \mathbf{z}_{t+1} - \mathbf{z}_t, \mathbf{z}_t - \bar{\mathbf{x}}^*(\mathbf{z}_t) \rangle + \frac{\mu}{2\sigma_4} \|\mathbf{z}_t - \mathbf{z}_{t+1}\|^2,$$

where  $\sigma_4$  is defined in (A.1).

In the next lemma, we will join the previous lemmas and characterize the change in the potential function. **Lemma A.6** (cf. Lemma 3.6). Let Assumption 1.1 hold. By using the parameters (A.1) in Algorithm 1, we obtain

$$\mathbb{E}V_{t} - \mathbb{E}V_{t+1} \ge c_{\beta}\mathbb{E}\|\mathbf{z}_{t+1} - \mathbf{z}_{t}\|^{2} + c_{\tau}\mathbb{E}\|\mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t}) - \mathbf{x}_{t}\|^{2} + c_{\eta}\mathbb{E}\|A\mathbf{x}^{*}(\mathbf{y}_{t+1}, \mathbf{z}_{t}) - \mathbf{b}\|^{2} - \lambda\tau^{2}\sigma^{2}, \quad (A.10)$$
where  $c_{\beta} = \frac{\mu}{50\beta}, c_{\tau} = \frac{7\tau\lambda^{2}}{400}, c_{\eta} = \frac{\eta}{4}.$ 

Proof. Combining Corollary A.4 and Lemma A.5, we obtain

$$\begin{split} \mathbb{E}[V_t] - \mathbb{E}[V_{t+1}] &= \mathbb{E}\varphi_{1/\lambda}(x_t, y_t, z_t) - \mathbb{E}\varphi_{1/\lambda}(x_{t+1}, y_{t+1}, z_{t+1}) + 2\mathbb{E}d(y_{t+1}, z_{t+1}) - 2\mathbb{E}d(y_t, z_t) + 2\Psi(z_t) - 2\Psi(z_{t+1}) \\ &\geq \frac{\tau\lambda^2}{16}\mathbb{E}\|\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{x}_t\|^2 - (\lambda\tau\mu + 2\lambda\tau^2\mu^2 + \frac{\tau\lambda^2\mu^2}{8\gamma_s^2})\mathbb{E}\|\mathbf{z}_t - \mathbf{z}_{t+1}\|^2 - \lambda\tau^2\sigma^2 \\ &- \mathbb{E}\eta\langle A\mathbf{x}_t - \mathbf{b}, A\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t) - \mathbf{b}\rangle + \frac{\mu}{2}\mathbb{E}\langle \mathbf{z}_{t+1} - \mathbf{z}_t, 2\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{z}_t - \mathbf{z}_{t+1}\rangle \\ &+ 2\eta\mathbb{E}\langle A\mathbf{x}_t - \mathbf{b}, A\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{b}\rangle + \mu\mathbb{E}\langle \mathbf{z}_{t+1} - \mathbf{z}_t, \mathbf{z}_{t+1} + \mathbf{z}_t - 2\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_{t+1})\rangle \\ &- 2\mu\mathbb{E}\langle \mathbf{z}_{t+1} - \mathbf{z}_t, \mathbf{z}_t - \bar{\mathbf{x}}^*(\mathbf{z}_t)\rangle - \frac{\mu}{\sigma_4}\mathbb{E}\|\mathbf{z}_t - \mathbf{z}_{t+1}\|^2. \end{split}$$

We next manipulate the terms on the right-hand side.

First, by adding and subtracting  $A\mathbf{x}_t$  on the second argument of the inner product, we get

$$-\eta \langle A\mathbf{x}_t - \mathbf{b}, A\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t) - \mathbf{b} \rangle = -\eta \|A\mathbf{x}_t - \mathbf{b}\|^2 - \eta \langle A\mathbf{x}_t - \mathbf{b}, A\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t) - A\mathbf{x}_t \rangle.$$

Consequently, we have

$$-\eta \langle A\mathbf{x}_t - \mathbf{b}, A\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t) - \mathbf{b} \rangle + 2\eta \langle A\mathbf{x}_t - \mathbf{b}, A\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{b} \rangle$$
  
=  $-\eta \|A\mathbf{x}_t - A\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t)\|^2 + \eta \|A\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{b}\|^2 - \eta \langle A\mathbf{x}_t - \mathbf{b}, A\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t) - A\mathbf{x}_t \rangle.$ 

Second, adding and subtracting  $2\mathbf{x}_t$  in the second argument of the inner product gives

$$\frac{\mu}{2} \langle \mathbf{z}_{t+1} - \mathbf{z}_t, 2\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{z}_t - \mathbf{z}_{t+1} \rangle = \frac{\mu}{2} \langle \mathbf{z}_{t+1} - \mathbf{z}_t, 2\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) - 2\mathbf{x}_t \rangle + \frac{\mu}{2} \langle \mathbf{z}_{t+1} - \mathbf{z}_t, 2\mathbf{x}_t - \mathbf{z}_t - \mathbf{z}_{t+1} \rangle.$$
We continue estimating the inner products involving  $\mathbf{z}_{t+1} - \mathbf{z}_t$ . Note that  $\mathbf{z}_{t+1} = \mathbf{z}_t + \beta(\mathbf{x}_t - \mathbf{z}_t) \iff 2\mathbf{x}_t - 2\mathbf{z}_t = \frac{2}{\beta}(\mathbf{z}_{t+1} - \mathbf{z}_t)$ 

$$\begin{aligned} \frac{\mu}{2} \langle \mathbf{z}_{t+1} - \mathbf{z}_t, 2\mathbf{x}_t - \mathbf{z}_t - \mathbf{z}_{t+1} \rangle &= \frac{\mu}{2} \langle \mathbf{z}_{t+1} - \mathbf{z}_t, 2\mathbf{x}_t - 2\mathbf{z}_t \rangle + \frac{\mu}{2} \langle \mathbf{z}_{t+1} - \mathbf{z}_t, \mathbf{z}_t - \mathbf{z}_{t+1} \rangle \\ &= \frac{\mu}{2} \left( \frac{2}{\beta} - 1 \right) \| \mathbf{z}_t - \mathbf{z}_{t+1} \|^2 \ge \frac{\mu}{2\beta} \| \mathbf{z}_t - \mathbf{z}_{t+1} \|^2, \end{aligned}$$

where the last inequality is due to  $\beta \leq 1$ . Next, we have

$$\mu \langle \mathbf{z}_{t+1} - \mathbf{z}_t, \mathbf{z}_{t+1} + \mathbf{z}_t - 2\mathbf{x}^* (\mathbf{y}_{t+1}, \mathbf{z}_{t+1}) \rangle - 2\mu \langle \mathbf{z}_{t+1} - \mathbf{z}_t, \mathbf{z}_t - \bar{\mathbf{x}}^* (\mathbf{z}_t) \rangle$$
  
=  $\mu \| \mathbf{z}_{t+1} - \mathbf{z}_t \|^2 + 2\mu \langle \mathbf{z}_{t+1} - \mathbf{z}_t, \bar{\mathbf{x}}^* (\mathbf{z}_t) - \mathbf{x}^* (\mathbf{y}_{t+1}, \mathbf{z}_{t+1}) \rangle.$ 

We can use Cauchy-Schwarz, triangle and Young's inequalities on the second term to get

$$\begin{aligned} \langle \mathbf{z}_{t+1} - \mathbf{z}_t, \bar{\mathbf{x}}^*(\mathbf{z}_t) - \mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_{t+1}) \rangle &\geq -\|\mathbf{z}_{t+1} - \mathbf{z}_t\| (\|\bar{\mathbf{x}}^*(\mathbf{z}_t) - \mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t)\| + \|\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_{t+1})\|) \\ &\geq -\left(\frac{1}{2\zeta} + \frac{1}{\sigma_4}\right) \|\mathbf{z}_{t+1} - \mathbf{z}_t\|^2 - \frac{\zeta}{2} \|\bar{\mathbf{x}}^*(\mathbf{z}_t) - \mathbf{x}^*(y_{t+1}, \mathbf{z}_t)\|^2, \end{aligned}$$

where the last step also used (A.34). Consequently, we obtain

$$\mu \langle \mathbf{z}_{t+1} - \mathbf{z}_t, \mathbf{z}_{t+1} + \mathbf{z}_t - 2\mathbf{x}^* (\mathbf{y}_{t+1}, \mathbf{z}_{t+1}) \rangle - 2\mu \langle \mathbf{z}_{t+1} - \mathbf{z}_t, \mathbf{z}_t - \bar{\mathbf{x}}^* (\mathbf{z}_t) \rangle$$

$$\geq \left( \mu - \frac{\mu}{\zeta} - \frac{2\mu}{\sigma_4} \right) \|\mathbf{z}_{t+1} - \mathbf{z}_t\|^2 - \mu \zeta \|\bar{\mathbf{x}}^* (\mathbf{z}_t) - \mathbf{x}^* (y_{t+1}, \mathbf{z}_t)\|^2.$$

As a result, we get

$$\mathbb{E}[V_t] - \mathbb{E}[V_{t+1}]$$

$$\geq \frac{\tau\lambda^2}{16} \mathbb{E}\|\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{x}_t\|^2 - (\lambda\tau\mu + 2\lambda\tau^2\mu^2 + \frac{\tau\lambda^2\mu^2}{8\gamma_s^2} + \frac{\mu}{\zeta} + \frac{3\mu}{\sigma_4} - \mu - \frac{\mu}{2\beta})\mathbb{E}\|\mathbf{z}_t - \mathbf{z}_{t+1}\|^2 - \lambda\tau^2\sigma^2$$

$$- \eta\langle A\mathbf{x}_t - \mathbf{b}, A\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t) - A\mathbf{x}_t\rangle - \eta\|A\mathbf{x}_t - A\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t)\|^2 + \eta\|A\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{b}\|^2$$

$$- \mu\zeta\|\bar{\mathbf{x}}^*(\mathbf{z}_t) - \mathbf{x}^*(y_{t+1}, \mathbf{z}_t)\|^2 + \mu\langle \mathbf{z}_{t+1} - \mathbf{z}_t, \mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{x}_t\rangle.$$
(A.11)

We will now operate on some of terms from the right-hand side of (A.11), by using Lemma A.8 and A.9. First, we have by Cauchy-Schwarz and Young's inequalities that

$$\begin{aligned} &-\eta \langle A\mathbf{x}_t - \mathbf{b}, A\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t) - A\mathbf{x}_t \rangle \\ &\geq -\frac{\eta}{4} \|A\mathbf{x}_t - \mathbf{b}\|^2 - \eta \|A\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t) - A\mathbf{x}_t\|^2 \\ &\geq -\frac{\eta}{4} \|A\mathbf{x}_t - \mathbf{b}\|^2 - 2\eta \|A\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t) - A\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t)\|^2 - 2\eta \|A\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) - A\mathbf{x}_t\|^2. \end{aligned}$$

Next, by using the Lipschitzness of  $\mathbf{u}^*(\mathbf{x}_t, \cdot, \mathbf{z}_t)$  from (A.31), we have

$$\begin{split} \|A\mathbf{u}^{*}(\mathbf{x}_{t},\mathbf{y}_{t},\mathbf{z}_{t}) - A\mathbf{u}^{*}(\mathbf{x}_{t},\mathbf{y}_{t+1},\mathbf{z}_{t})\|^{2} &\leq \|A\|^{2} \|\mathbf{u}^{*}(\mathbf{x}_{t},\mathbf{y}_{t},\mathbf{z}_{t}) - \mathbf{u}^{*}(\mathbf{x}_{t},\mathbf{y}_{t+1},\mathbf{z}_{t})\|^{2} \\ &\leq \frac{\|A\|^{4}}{\gamma_{s}^{2}} \|\mathbf{y}_{t} - \mathbf{y}_{t+1}\|^{2} \\ &= \frac{\|A\|^{4}\eta^{2}}{\gamma_{s}^{2}} \|A\mathbf{x}_{t} - \mathbf{b}\|^{2}, \end{split}$$

where the last step also used the definition of  $\mathbf{y}_{t+1}$ . Using this estimation along with (A.38) gives

$$\begin{aligned} &-\eta \langle A\mathbf{x}_{t} - \mathbf{b}, A\mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t}, \mathbf{z}_{t}) - A\mathbf{x}_{t} \rangle \\ &\geq -(\frac{\eta}{4} + \frac{2\|A\|^{4}\eta^{3}}{\gamma_{s}^{2}})\|A\mathbf{x}_{t} - \mathbf{b}\|^{2} - 2\eta\|A\|^{2}\|\mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t}) - \mathbf{x}_{t}\|^{2} \\ &\geq -(\frac{\eta\|A\|^{2}\lambda^{2}}{2\gamma^{2}} + \frac{4\|A\|^{6}\eta^{3}\lambda^{2}}{\gamma^{2}\gamma_{s}^{2}} + 2\eta\|A\|^{2})\|\mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t}) - \mathbf{x}_{t}\|^{2} \\ &- (\frac{\eta}{2} + \frac{4\|A\|^{4}\eta^{3}}{\gamma_{s}^{2}})E\|A\mathbf{x}^{*}(\mathbf{y}_{t+1}, \mathbf{z}_{t}) - \mathbf{b}\|^{2}. \end{aligned}$$

We next have by Young's inequality that for any  $\theta > 0$ :

$$\mu \langle \mathbf{z}_{t+1} - \mathbf{z}_t, \mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{x}_t \rangle \geq -\frac{\mu}{4\theta} \|\mathbf{z}_{t+1} - \mathbf{z}_t\|^2 - \theta \mu \|\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{x}_t\|^2.$$

The inequality derived in (A.37) directly implies

$$-\eta \|A\mathbf{x}_t - A\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t)\|^2 \ge -\frac{\eta \|A\|^2 \lambda^2}{\gamma^2} \|\mathbf{x}_t - \mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t)\|^2.$$

The key global error bound given in Lemma A.9 originally proved in Zhang & Luo (2022) results in

$$-6\mu\beta \|\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t) - \bar{\mathbf{x}}^*(\mathbf{z}_t)\|^2 \ge -6\mu\beta\bar{\sigma}^2 \|A\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{b}\|^2$$

Combining these estimates lead to

$$\mathbb{E}[V_{t}] - \mathbb{E}[V_{t+1}] \geq -(\lambda\tau\mu + 2\lambda\tau^{2}\mu^{2} + \frac{\tau\lambda^{2}\mu^{2}}{8\gamma_{s}^{2}} + \frac{\mu}{\zeta} + \frac{3\mu}{\sigma_{4}} - \mu - \frac{\mu}{2\beta} + \frac{\mu}{4\theta})\mathbb{E}\|\mathbf{z}_{t} - \mathbf{z}_{t+1}\|^{2} - \lambda\tau^{2}\sigma^{2} + \left(\frac{\tau\lambda^{2}}{16} - \frac{3\|A\|^{2}\lambda^{2}\eta}{2\gamma^{2}} - \frac{4\|A\|^{6}\eta^{3}\lambda^{2}}{\gamma_{s}^{2}\gamma^{2}} - 2\eta\|A\|^{2} - \mu\theta\right)\mathbb{E}\|\mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t}) - \mathbf{x}_{t}\|^{2} + \left(\frac{\eta}{2} - \frac{4\|A\|^{4}\eta^{3}}{\gamma_{s}^{2}} - 6\mu\beta\bar{\sigma}^{2}\right)\|A\mathbf{x}^{*}(\mathbf{y}_{t+1}, \mathbf{z}_{t}) - \mathbf{b}\|^{2}.$$
(A.12)

We now estimate the coefficients inside the parantheses, with straightforward but tedious calculations which follow from the parameter settings.

First, we estimate the coefficient of  $\mathbb{E} \| \mathbf{z}_t - \mathbf{z}_{t+1} \|^2$  in (A.12): Let  $\mu \geq 4L_f$ , we have  $\sigma_4 \geq \frac{1}{2}$  because  $\sigma_4 = \frac{\mu - L_f}{\mu}$ . Then letting  $\zeta = 6\beta, \beta < \frac{1}{30}$ , we have

$$\mu - \frac{3\mu}{\sigma_4} \ge -5\mu \ge -\frac{\mu}{6\beta}, \quad \frac{\mu}{\zeta} = \frac{\mu}{6\beta}$$

Therefore,

$$\frac{\mu}{2\beta} + \mu - \frac{3\mu}{\sigma_4} - \frac{\mu}{\zeta} \ge (\frac{1}{2} - \frac{1}{6} - \frac{1}{6})\frac{\mu}{\beta} \ge \frac{\mu}{6\beta}.$$
(A.13)

Hence,

c

coefficient of 
$$\mathbb{E} \| \mathbf{z}_t - \mathbf{z}_{t+1} \| \ge -\lambda \tau \mu - 2\lambda \tau^2 \mu^2 - \frac{\tau \lambda^2 \mu^2}{8\gamma_s^2} + \frac{\mu}{6\beta} - \frac{\mu}{8\beta}$$
.

Let  $\eta = \frac{\eta'}{2\|A\|^2}, \theta = 2\beta, \eta' \leq \frac{1}{40}$ , and  $\mu = \max\{2, 4L_f\}, \lambda = L_K = L_f + \rho \|A\| + \mu, \tau \leq \frac{1}{10\lambda^2}$ , and  $\gamma_s = \mu - L_f + \gamma$  from Fact A.10. We have  $-\lambda \tau \mu \geq -\frac{\mu}{10}$  and  $-2\lambda \tau^2 \mu^2 \geq -\frac{\mu}{50}$ , then

coefficient of 
$$\mathbb{E} \| \mathbf{z}_t - \mathbf{z}_{t+1} \| \ge \frac{\mu}{24\beta} - \frac{\mu}{10} - \frac{\mu}{50} - \tau \lambda^2 \frac{\mu^2}{(\mu - L_f + \lambda)^2}$$

By  $\beta \leq 1/30$ , we have  $\frac{1}{24\beta} - \frac{1}{10} - \frac{1}{50} \geq \frac{1}{30\beta}$ . In addition, using  $\tau \lambda^2 \frac{\mu^2}{(\mu - L_f + \lambda)^2} \leq \tau \lambda^2 \leq \frac{1}{10}$ , we famally obtain:

coefficient of 
$$\mathbb{E} \| \mathbf{z}_t - \mathbf{z}_{t+1} \| \ge \frac{\mu}{30\beta} - \frac{1}{10} \stackrel{\mu \ge 2}{\ge} \frac{\mu}{50\beta}.$$
 (A.14)

Then we estimate the coefficient of  $\mathbb{E} \| \mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{x}_t \|^2$  in (A.12). From above assumptions, we can easily get  $\gamma = \frac{(p-L_f)\lambda}{p-L_f+\lambda} \ge \frac{1}{2}$  because  $\lambda \ge \mu \ge 2$ . Moreover, we assume  $\eta' \leq \frac{\tau}{40}, \frac{\eta'}{\mu - L_f + \lambda} \leq \frac{\tau}{10}, \beta \leq \frac{\tau}{40}$  First, by our new notations, we have

coefficient of 
$$\mathbb{E} \| \mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) - \mathbf{x}_t \|^2 = \frac{\tau \lambda^2}{16} - \frac{3\eta' \lambda^2}{4\gamma^2} - \frac{\eta'^3 \lambda^2}{2\gamma^2 \gamma_s^2} - \eta' - 2\mu\beta$$

By  $\gamma \geq \frac{1}{2}$  and the definition of  $\gamma_s$ , we have  $-\frac{3\eta'\lambda^2}{4\gamma^2} \geq -3\eta'\lambda^2$ ,  $-\frac{\eta'^3\lambda^2}{2\gamma^2\gamma_s^2} \geq \frac{\eta'^2\lambda^2}{(\mu-L_f+\lambda)^2}$ , Then

coefficient of 
$$\mathbb{E} \| \mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) - \mathbf{x}_t \|^2 \ge \frac{\tau \lambda^2}{16} - 3\eta' \lambda^2 - \frac{2\eta'^3 \lambda^2}{(\mu - L_f + \lambda)^2} - \eta' - 2\mu\beta$$

With  $2 \leq \mu \leq \lambda$ ,  $\eta' \leq \frac{\tau}{100}, \frac{\eta'}{\mu - L_f + \lambda} \leq \frac{\tau}{10}, \beta \leq \frac{\tau}{200}$ , we can obtain  $-3\eta'\lambda^2 \geq -\frac{3\tau\lambda^2}{400}, -\frac{2\eta'^3\lambda^2}{(\mu - L_f + \lambda)^2} \geq -\frac{3\tau\lambda^2}{400}$  $-\frac{\lambda^2 \tau^2}{400} \ge -\frac{\lambda^2 \tau}{400}, \ -\eta' \ge -\frac{\tau}{100} \ge -\frac{\tau \lambda^2}{100}, \ -2\mu\beta \ge \frac{\tau \mu}{50} \stackrel{\mu \le \lambda}{\ge} -\frac{\tau \lambda}{50} \ge \frac{\tau \lambda^2}{100}.$  Hence,

coefficient of 
$$\mathbb{E} \| \mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) - \mathbf{x}_t \|^2 \ge \frac{\tau \lambda^2}{16} - \frac{3\tau \lambda^2}{100} - \frac{\tau \lambda^2}{400} - \frac{\tau \lambda^2}{400} - \frac{\tau \lambda^2}{100} = \frac{7\tau \lambda^2}{400}$$
(A.15)

Last, we estimate the coefficient of  $\mathbb{E} ||A\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{b}||^2$  in (A.12):

By  $6\mu\beta\bar{\sigma}^2 \leq \frac{\eta}{6}$  and the definition  $\eta', \gamma_s$ , we have  $-\frac{4\|A\|^2\eta^3}{\gamma_s^2} = -\frac{\eta'^2\eta}{(\mu-L_f+\lambda)^2} \stackrel{\overline{\eta'}=\frac{\tau}{\mu-L_f+\lambda}\leq\frac{\tau}{10}}{\geq} -\frac{\eta\tau^2}{100} \geq -\frac{\eta}{100}$  and  $-6\mu\beta\bar{\sigma}^2 \geq -\frac{\eta}{6}$ . Hence, we have

coefficient of 
$$\mathbb{E} \| A \mathbf{x}^* (\mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{b} \|^2 \ge \frac{\eta}{4}.$$
 (A.16)

Plug A.14, A.15 and A.16 to A.12, we finishe the proof.

### A.2 Proof of Theorem 3.1

*Proof.* We start from the result in Lemma A.6. First, it follows from the definition of  $\mathbf{z}_{t+1}$  that

$$\|\mathbf{z}_t - \mathbf{z}_{t+1}\| = \beta \|\mathbf{x}_t - \mathbf{z}_t\|$$

So, we rewrite (A.10), as:

$$\mathbb{E}V_t - \mathbb{E}V_{t+1} \ge \beta^2 c_\beta \mathbb{E} \|\mathbf{x}_t - \mathbf{z}_t\|^2 + c_\tau \mathbb{E} \|\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{x}_t)\|^2 + c_\eta \mathbb{E} \|A\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{b}\|^2 - \lambda \tau^2 \sigma^2.$$
(A.17)

For t > 0, we have  $V_t \ge \underline{f}$ , which is proven in Lemma A.13. It then follows that

$$\sum_{t=0}^{T-1} (\mathbb{E}V_t - \mathbb{E}V_{t+1}) = \mathbb{E}V_0 - \mathbb{E}V_T \le \mathbb{E}V_0 - \underline{f}.$$
 (A.18)

Then, summing up (A.17), using (A.18), and the fact that  $c_{\tau} = \Theta(\tau)$ ,  $c_{\eta} = \Theta(\tau)$ ,  $\beta^2 c_{\beta} = \Theta(\tau)$  from (A.1), we have

$$V_0 - \underline{f} + T\lambda\tau^2\sigma^2 \ge \sum_{t=1}^T C_0\tau \left[ \mathbb{E} \|\mathbf{x}_t - \mathbf{z}_t\|^2 + \mathbb{E} \|\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{x}_t\|^2 + \mathbb{E} \|A\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{b}\|^2 \right],$$

for some explicit constant  $C_0$ .

Dividing both sides by T, rearranging and using the definition  $\tau = \frac{1}{6\lambda\sqrt{T}}$  gives

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\mathbf{x}_t - \mathbf{z}_t\|^2 + \mathbb{E}\|\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{x}_t\|^2 + \mathbb{E}\|A\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{b}\|^2 \le \frac{1}{C_0\sqrt{T}} \left(6\lambda(V_0 - \underline{f}) + \frac{\sigma^2}{6}\right).$$
(A.19)

Since we have

$$\nabla \Psi(\mathbf{z}_t) = \mu(\mathbf{z}_t - \bar{\mathbf{x}}^*(\mathbf{z}_t)),$$

by Danskin's theorem, we deduce for any t

$$\begin{split} \frac{1}{\mu^2} \| \nabla \Psi(\mathbf{z}_t) \| &= \| \mathbf{z}_t - \bar{\mathbf{x}}^*(\mathbf{z}_t) \| \\ &\leq \| \mathbf{z}_t - \mathbf{x}^*(\mathbf{y}_{t+1}, z_t) \| + \| \mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t) - \bar{\mathbf{x}}^*(\mathbf{z}_t) \| \\ &\leq \| \mathbf{z}_t - \mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t) \| + \bar{\sigma} \| A \mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_s) - \mathbf{b} \| \\ &\leq \| \mathbf{z}_t - \mathbf{x}_t \| + \| \mathbf{x}_t - \mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t) \| + \bar{\sigma} \| A \mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{b} \| \\ &\leq \| \mathbf{z}_t - \mathbf{x}_t \| + \frac{\lambda}{\gamma} \| \mathbf{x}_t - \mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) \| + \bar{\sigma} \| A \mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{b} \| . \end{split}$$

where the first inequality is by triangle inequality, the second by (A.9), the third by triangle inequality and the fourth by (A.29).

Next, we take square of both sides, take expectation, use Young's inequality, sum for all t = 0, ..., T - 1, divide by T and use (A.19) to derive

$$\frac{1}{\mu^2} \frac{1}{T} \sum_{t=1}^T \mathbb{E} \|\nabla \Psi(\mathbf{z}_t)\|^2 \le \frac{1}{T} \sum_{t=1}^T \mathbb{E} \left[ 3\|\mathbf{z}_t - \mathbf{x}_t\|^2 + \frac{3\lambda^2}{\gamma^2} \|\mathbf{x}_t - \mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t)\|^2 + 3\bar{\sigma}^2 \|A\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{b}\|^2 \right] = O\left(\frac{1}{\sqrt{T}}\right).$$

The result then follows since  $t^*$  is selected uniformly at random from  $\{1, 2, \ldots, T\}$ .

#### Proof of Corollary 3.2 A.3

*Proof.* From the definition of  $\hat{x}$ , we have

$$0 \in \hat{G}(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) + \frac{2}{\tau}(\hat{\mathbf{x}} - \mathbf{x}_t) + \partial I_X(\hat{\mathbf{x}}).$$

Let us set

$$\mathbf{v} = \nabla_{\mathbf{x}} K(\hat{\mathbf{x}}, \mathbf{y}_{t+1}, \mathbf{z}_t) - \hat{G}(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) - \frac{2}{\tau} (\hat{\mathbf{x}} - \mathbf{x}_t) - \rho A^T (A\hat{\mathbf{x}} - \mathbf{b}) - \mu (\hat{\mathbf{x}} - \mathbf{z}_t).$$
(A.20)

Combining with the optimality condition, we have

$$\mathbf{v} \in \nabla_{\mathbf{x}} K(\hat{\mathbf{x}}, \mathbf{y}_{t+1}, \mathbf{z}_t) - \rho A^T (A \hat{\mathbf{x}} - \mathbf{b}) - \mu(\hat{\mathbf{x}} - \mathbf{z}_t) + \partial I_X(\hat{\mathbf{x}})$$
  
=  $\nabla f(\hat{\mathbf{x}}) + A^T \mathbf{y}_{t+1} + \partial I_X(\hat{\mathbf{x}}).$ 

Hence, we need to estimate  $\mathbb{E} \|A\hat{\mathbf{x}} - \mathbf{b}\|$  and  $\mathbb{E} \|\mathbf{v}\|$ .

For the mini-batch gradient in the post-processing step, we have

$$\mathbb{E}\|\hat{G}(\mathbf{x},\mathbf{y},\mathbf{z}) - \nabla K(\mathbf{x},\mathbf{y},\mathbf{z})\|^2 \le \frac{\sigma^2}{B}.$$
(A.21)

which is a standard calculation (Lan, 2020, Section 5.2.3). Since  $B = \Omega(\varepsilon^{-2})$ , this gives us

$$\mathbb{E}\|\hat{G}(\mathbf{x},\mathbf{y},\mathbf{z}) - \nabla K(\mathbf{x},\mathbf{y},\mathbf{z})\|^2 \le \varepsilon^2.$$
(A.22)

First, let us note that the purpose of  $\hat{\mathbf{x}}$  is to estimate  $\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t)$ , where

$$\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) = \arg\min_{\mathbf{u} \in X} \{ l(\mathbf{u}) := K(\mathbf{u}, \mathbf{y}_{t+1}, \mathbf{z}_t) + \frac{\lambda}{2} \|\mathbf{x}_t - \mathbf{u}\|^2 \}$$

Note that the gradient of this objective is

$$\nabla l(\mathbf{u}) = \nabla_{\mathbf{x}} K(\mathbf{x}, \mathbf{y}_{t+1}, \mathbf{z}_t) + \lambda(\mathbf{x} - \mathbf{x}_t).$$

As a result, we have  $\nabla l(\mathbf{x}_t) = \nabla_{\mathbf{x}} K(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t)$ . Let us also denote

$$\mathbf{x}_t^* = \operatorname{proj}_X(\mathbf{x}_t - \tau \nabla l(\mathbf{x}_t))$$

That is,  $\mathbf{x}_t^*$  is the output of doing a full-gradient step on  $\mathbf{x}_t$ . Of course, this is not tractable in our setting, but we only use this as a theoretical tool.

Since this is a GD step on the objective l which is  $L_K$ -smooth and convex with optimizer  $\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t)$ , the standard analysis for GD gives

$$\|\mathbf{x}_{t}^{*} - \mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t})\|^{2} \le \|\mathbf{x}_{t} - \mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t})\|^{2},$$
(A.23)

as long as  $\tau \leq \frac{1}{L_K}$ . Next, by the definitions of  $\mathbf{x}_t^*$  and  $\hat{\mathbf{x}}$ , along with nonexpansiveness of the projection, we have

$$\mathbb{E} \|\mathbf{x}_{t}^{*} - \hat{\mathbf{x}}\|^{2} \leq \mathbb{E}\tau^{2} \|\hat{G}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) - \nabla_{\mathbf{x}} K(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t})\|^{2} \\ \leq \tau^{2} \varepsilon^{2},$$
(A.24)

where the second inequality used (A.22).

In view of (A.20), we estimate  $\|\mathbf{v}\|$  as

$$\|\mathbf{v}\| \le \|\nabla_x K(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) - \hat{G}(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t)\| + L_K \|\mathbf{x}_t - \hat{\mathbf{x}}\| + \frac{2}{\tau} \|\hat{\mathbf{x}} - \mathbf{x}_t\| + \rho \|A\| \|A\hat{\mathbf{x}} - \mathbf{b}\| + \mu \|\hat{\mathbf{x}} - \mathbf{z}_t\|.$$

On this, multiple applications of triangle inequality gives

$$\begin{aligned} \|\hat{\mathbf{x}} - \mathbf{x}_t\| &\leq \|\hat{\mathbf{x}} - \mathbf{x}_t^*\| + \|\mathbf{x}_t^* - \mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t)\| + \|\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{x}_t\| \\ &\leq \|\hat{\mathbf{x}} - \mathbf{x}_t^*\| + 2\|\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{x}_t) - \mathbf{x}_t\|, \end{aligned}$$
(A.25)

where the second line is due to (A.23).

For the feasibility, we have by triangle inequality that

$$\|\hat{\mathbf{x}} - \mathbf{z}_t\| \le \|\hat{\mathbf{x}} - \mathbf{x}_t\| + \|\mathbf{x}_t - \mathbf{z}_t\|.$$
(A.26)

As a result, we have that

$$\|\mathbf{v}\| = O(\|\hat{\mathbf{x}} - \mathbf{x}_{t}^{*}\| + \|\mathbf{x}_{t} - \mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t})\| + \|A\hat{\mathbf{x}} - \mathbf{b}\| + \|\mathbf{x}_{t} - \mathbf{z}_{t}\| + \|\nabla_{x}K(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t}) - \hat{G}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t})\|).$$
(A.27)

For the feasibility, we have

$$\begin{aligned} \|A\hat{\mathbf{x}} - \mathbf{b}\| &\leq \|A\hat{\mathbf{x}} - A\mathbf{x}_t\| + \|A\mathbf{x}_t - \mathbf{b}\| \\ &\leq \|A\| \|\hat{\mathbf{x}} - \mathbf{x}_t\| + \|A\mathbf{x}_t - \mathbf{b}\|. \end{aligned}$$

Now, by invoking the above inequality for  $t = t^*$ , taking expectation, using Young's inequality, (A.25), (A.24) and (A.19) along with (A.38), we get that

$$\mathbb{E}\|A\hat{\mathbf{x}} - \mathbf{b}\|^2 \le \varepsilon^2,\tag{A.28}$$

since  $T = \Omega(\varepsilon^{-4})$ .

Finally, using  $t = t^*$ , taking square and then expectation of (A.27), using Young's inequality and then combining (A.28), (A.24), (A.22) and (A.19) gives the result since  $T = \Omega(\varepsilon^{-4})$ .

### A.4 Auxiliary Results

**Lemma A.7.** Under Assumption 1.1, for any  $\mathbf{x}, \mathbf{z}, \mathbf{z}' \in X$ , we have

$$\frac{\lambda}{\gamma} \|\mathbf{x} - \mathbf{u}^*(\mathbf{x}, \mathbf{y}, \mathbf{z})\| \ge \|\mathbf{x} - \mathbf{x}^*(\mathbf{y}, \mathbf{z})\|, \tag{A.29}$$

$$\|\mathbf{u}^*(\mathbf{x}, \mathbf{y}, \mathbf{z}) - \mathbf{x}\| \le \|\mathbf{x} - \mathbf{x}^*(\mathbf{y}, \mathbf{z})\|,$$
(A.30)

$$\|\mathbf{u}^*(\mathbf{x}, \mathbf{y}, \mathbf{z}) - \mathbf{u}^*(\mathbf{x}, \mathbf{y}', \mathbf{z})\| \le \frac{\|A\|}{\gamma_s} \|\mathbf{y} - \mathbf{y}'\|,\tag{A.31}$$

$$\|\mathbf{u}^*(\mathbf{x}, \mathbf{y}, \mathbf{z}') - \mathbf{u}^*(\mathbf{x}, \mathbf{y}, \mathbf{z}')\| \le \frac{\mu}{\gamma_s} \|\mathbf{z} - \mathbf{z}'\|,$$
(A.32)

$$\|\mathbf{z}' - \mathbf{z}\| \ge \frac{\mu - L_f}{\mu} \|\mathbf{x}^*(\mathbf{y}, \mathbf{z}') - \mathbf{x}^*(\mathbf{y}, \mathbf{z})\|,$$
(A.33)

$$\|\mathbf{y}' - \mathbf{y}\| \ge \frac{\gamma_K}{\|A\|} \|\mathbf{x}^*(\mathbf{y}', \mathbf{z}) - \mathbf{x}^*(\mathbf{y}, \mathbf{z})\|,$$
(A.34)

$$\|\bar{\mathbf{x}}^*(\mathbf{z}) - \bar{\mathbf{x}}^*(\mathbf{z}')\| \le \frac{\mu}{\mu - L_f} \|\mathbf{z} - \mathbf{z}'\|$$
(A.35)

(A.36)

where  $\gamma = \frac{(\mu - L_f)\lambda}{\mu - L_f + \lambda}, \gamma_s = \mu - L_f + \lambda, \gamma_K = \mu - L_f.$ 

Proof. The proofs for (A.33), (A.34), and (A.35) appear in Zhang & Luo (2022), so we omit these proofs.

We first prove (A.29). Let us note that  $\mathbf{x}^*(\mathbf{y}, \mathbf{z})$  minimizes  $\varphi_{1/\lambda}$ , see for example (Hiriart-Urruty & Lemarechal, 1993, Theorem XV4.1.7). As a result, we have  $\nabla_{\mathbf{x}} \varphi_{1/\lambda}(\mathbf{x}^*(\mathbf{y}, \mathbf{z}), \mathbf{y}, \mathbf{z}) = 0$ . From Lemma A.12, we have that  $\varphi_{1/\lambda}(\cdot, y, z)$  is  $\gamma = \frac{(\mu - L_f)\lambda}{\mu - L_f + \lambda}$ -strongly convex. Then, by strong convexity, we have

$$\begin{aligned} \langle \nabla_{\mathbf{x}} \varphi_{1/\lambda}(\mathbf{x}^*(\mathbf{y}, \mathbf{z}), \mathbf{y}, \mathbf{z}) - \nabla_{\mathbf{x}} \varphi_{1/\lambda}(\mathbf{x}, \mathbf{y}, \mathbf{z}), \mathbf{x}^*(\mathbf{y}, \mathbf{z}) - \mathbf{x} \rangle &\geq \gamma \|\mathbf{x} - \mathbf{x}^*(\mathbf{y}, \mathbf{z})\|^2 \\ \iff \|\nabla_{\mathbf{x}} \varphi_{1/\lambda}(\mathbf{x}, \mathbf{y}, \mathbf{z})\| \geq \gamma \|\mathbf{x} - \mathbf{x}^*(\mathbf{y}, \mathbf{z})\|, \end{aligned}$$

where the inclusion used  $\nabla_{\mathbf{x}}\varphi_{1/\lambda}(\mathbf{x}^*(\mathbf{y},\mathbf{z}),\mathbf{y},\mathbf{z}) = 0$  established in the previous paragraph as well as Cauchy-Schwarz inequality. Then, using  $\nabla_{\mathbf{x}}\varphi_{1/\lambda}(\mathbf{x},\mathbf{y},\mathbf{z}) = \lambda(\mathbf{x} - \mathbf{u}^*(\mathbf{x},\mathbf{y},\mathbf{z}))$ , we obtain (A.29).

From definition of  $\mathbf{u}^*(\mathbf{x}, \mathbf{y}, \mathbf{z})$  in (3.4), we have,

$$K(\mathbf{u}^*(\mathbf{x}, \mathbf{y}, \mathbf{z}), \mathbf{y}, \mathbf{z}) + \frac{\lambda}{2} \|\mathbf{x} - \mathbf{u}^*(\mathbf{x}, \mathbf{y}, \mathbf{z})\|^2 \le K(\mathbf{x}^*(\mathbf{y}, \mathbf{z}), \mathbf{y}, \mathbf{z}) + \frac{\lambda}{2} \|\mathbf{x} - \mathbf{x}^*(\mathbf{y}, \mathbf{z})\|^2,$$

where we also remark that  $\mathbf{x}^*(\mathbf{y}, \mathbf{z}) \in X$ . Combining with  $K(\mathbf{x}^*(\mathbf{y}, \mathbf{z}), \mathbf{y}, \mathbf{z}) \leq K(\mathbf{u}^*(\mathbf{x}, \mathbf{y}, \mathbf{z}), \mathbf{y}, \mathbf{z})$ , which follows from the definition of  $\mathbf{x}^*(\mathbf{y}, \mathbf{z})$  in (3.6), we have (A.30).

The proofs of the other two assertions will use a similar idea to Zhang & Luo (2022), but there will be differences in the estimations since this previous work did not use the function  $\varphi_{1/\lambda}$ .

For (A.31), we proceed by using the definition of  $\varphi_{1/\lambda}$  and adding and subtracting  $K(\mathbf{u}^*(\mathbf{x}, \mathbf{y}', \mathbf{z}), \mathbf{y}, \mathbf{z})$ to get

$$\begin{split} & K(\mathbf{u}^*(\mathbf{x}, \mathbf{y}, \mathbf{z}), \mathbf{y}, \mathbf{z}) + \frac{\lambda}{2} \|\mathbf{u}^*(\mathbf{x}, \mathbf{y}, \mathbf{z}) - \mathbf{x}\|^2 - K(\mathbf{u}^*(\mathbf{x}, \mathbf{y}', \mathbf{z}), \mathbf{y}', \mathbf{z}) - \frac{\lambda}{2} \|\mathbf{u}^*(\mathbf{x}, \mathbf{y}', \mathbf{z}) - \mathbf{x}\|^2 \\ &= K(\mathbf{u}^*(\mathbf{x}, \mathbf{y}, \mathbf{z}), \mathbf{y}, \mathbf{z}) + \frac{\lambda}{2} \|\mathbf{u}^*(\mathbf{x}, \mathbf{y}, \mathbf{z}) - \mathbf{x}\|^2 \\ &- K(\mathbf{u}^*(\mathbf{x}, \mathbf{y}', \mathbf{z}), \mathbf{y}, \mathbf{z}) - \frac{\lambda}{2} \|\mathbf{u}^*(\mathbf{x}, \mathbf{y}', \mathbf{z}) - \mathbf{x}\|^2 \\ &+ K(\mathbf{u}^*(\mathbf{x}, \mathbf{y}', \mathbf{z}), \mathbf{y}, \mathbf{z}) - K(\mathbf{u}^*(\mathbf{x}, \mathbf{y}', \mathbf{z}), \mathbf{y}', \mathbf{z}) \\ &\leq \frac{-\gamma_s}{2} \|\mathbf{u}^*(\mathbf{x}, \mathbf{y}, \mathbf{z}) - \mathbf{u}^*(\mathbf{x}, \mathbf{y}', \mathbf{z})\|^2 + \langle \mathbf{y} - \mathbf{y}', A\mathbf{u}^*(\mathbf{x}, \mathbf{y}', \mathbf{z}) - \mathbf{b} \rangle. \end{split}$$

The last step uses  $\mathbf{u} \mapsto K(\mathbf{u}, \mathbf{y}, \mathbf{z}) + \frac{\lambda}{2} \|\mathbf{u} - \mathbf{x}\|^2$  being  $\gamma_s$ -strongly convex (cf. Fact A.10) with minimizer  $\mathbf{u}^*(\mathbf{x}, \mathbf{y}, \mathbf{z})$ , as well as the definition of K.

We then argue similarly, this time adding and subtracting  $K(\mathbf{u}^*(\mathbf{x}, \mathbf{y}, \mathbf{z}), \mathbf{y}', \mathbf{z})$ :

$$\begin{split} & K(\mathbf{u}^*(\mathbf{x},\mathbf{y},\mathbf{z}),\mathbf{y},\mathbf{z}) + \frac{\lambda}{2} \|\mathbf{u}^*(\mathbf{x},\mathbf{y},\mathbf{z}) - \mathbf{x}\|^2 - K(\mathbf{u}^*(\mathbf{x},\mathbf{y}',\mathbf{z}),\mathbf{y}',\mathbf{z}) - \frac{\lambda}{2} \|\mathbf{u}^*(\mathbf{x},\mathbf{y}',\mathbf{z}) - \mathbf{x}\|^2 \\ &= K(\mathbf{u}^*(\mathbf{x},\mathbf{y},\mathbf{z}),\mathbf{y}',\mathbf{z}) + \frac{\lambda}{2} \|\mathbf{u}^*(\mathbf{x},\mathbf{y},\mathbf{z}) - \mathbf{x}\|^2 \\ &- K(\mathbf{u}^*(\mathbf{x},\mathbf{y}',\mathbf{z}),\mathbf{y}',\mathbf{z}) - \frac{\lambda}{2} \|\mathbf{u}^*(\mathbf{x},\mathbf{y}',\mathbf{z}) - \mathbf{x}\|^2 \\ &- K(\mathbf{u}^*(\mathbf{x},\mathbf{y},\mathbf{z}),\mathbf{y}',\mathbf{z}) + K(\mathbf{u}^*(\mathbf{x},\mathbf{y},\mathbf{z}),\mathbf{y},\mathbf{z}) \\ &\geq \frac{\gamma_s}{2} \|\mathbf{u}^*(\mathbf{x},\mathbf{y},\mathbf{z}) - \mathbf{u}^*(\mathbf{x},\mathbf{y}',\mathbf{z})\|^2 + \langle \mathbf{y} - \mathbf{y}', A\mathbf{u}^*(\mathbf{x},\mathbf{y},\mathbf{z}) - \mathbf{b} \rangle. \end{split}$$

The last step uses that  $\mathbf{u} \mapsto K(\mathbf{u}, \mathbf{y}', \mathbf{z}) + \frac{\lambda}{2} \|\mathbf{u} - \mathbf{x}\|^2$  is  $\gamma_s$ -strongly convex (cf. Fact A.10) with minimizer  $\mathbf{u}^*(\mathbf{x}, \mathbf{y}', \mathbf{z})$  and the definition of K.

Combining the last two estimates give

$$\langle \mathbf{y} - \mathbf{y}', A\mathbf{u}^*(\mathbf{x}, \mathbf{y}', \mathbf{z}) - A\mathbf{u}^*(\mathbf{x}, \mathbf{y}, \mathbf{z}) \rangle \geq \gamma_s \|\mathbf{u}^*(\mathbf{x}, \mathbf{y}, \mathbf{z}) - \mathbf{u}^*(\mathbf{x}, \mathbf{y}', \mathbf{z})\|^2.$$

Using Cauchy-Schwarz inequality and the definition of operator norm gives (A.31).

The proof of (A.32) is similar to the proof of (A.31), just completed. In particular, by adding and subtracting  $K(\mathbf{u}^*(\mathbf{x}, \mathbf{y}, \mathbf{z}), \mathbf{y}, \mathbf{z}')$ , we have

$$\begin{split} K(\mathbf{u}^*(\mathbf{x},\mathbf{y},\mathbf{z}),\mathbf{y},\mathbf{z}) &+ \frac{\lambda}{2} \|\mathbf{u}^*(\mathbf{x},\mathbf{y},\mathbf{z}) - \mathbf{x}\|^2 - K(\mathbf{u}^*(\mathbf{x},\mathbf{y},\mathbf{z}'),\mathbf{y},\mathbf{z}') + \frac{\lambda}{2} \|\mathbf{u}^*(\mathbf{x},\mathbf{y},\mathbf{z}') - \mathbf{x}\|^2 \\ &= K(\mathbf{u}^*(\mathbf{x},\mathbf{y},\mathbf{z}),\mathbf{y},\mathbf{z}) + \frac{\lambda}{2} \|\mathbf{u}^*(\mathbf{x},\mathbf{y},\mathbf{z}) - \mathbf{x}\|^2 - K(\mathbf{u}^*(\mathbf{x},\mathbf{y},\mathbf{z}'),\mathbf{y},\mathbf{z}) - \frac{\lambda}{2} \|\mathbf{u}^*(\mathbf{x},\mathbf{y},\mathbf{z}') - \mathbf{x}\|^2 \\ &- K(\mathbf{u}^*(\mathbf{x},\mathbf{y},\mathbf{z}'),\mathbf{y},\mathbf{z}') + K(\mathbf{u}^*(\mathbf{x},\mathbf{y},\mathbf{z}'),\mathbf{y},\mathbf{z}) \\ &\leq -\frac{\gamma_s}{2} \|\mathbf{u}^*(\mathbf{x},\mathbf{y},\mathbf{z}) - \mathbf{u}^*(\mathbf{x},\mathbf{y},\mathbf{z}')\|^2 + \frac{\mu}{2} (\|\mathbf{u}^*(\mathbf{x},\mathbf{y},\mathbf{z}') - \mathbf{z}\|^2 - \|\mathbf{u}^*(\mathbf{x},\mathbf{y},\mathbf{z}') - \mathbf{z}'\|^2), \end{split}$$

where we used that  $\mathbf{u} \mapsto K(\mathbf{u}, \mathbf{y}, \mathbf{z}) + \frac{\lambda}{2} \|\mathbf{u} - \mathbf{x}\|^2$  is  $\gamma_s$ -strongly convex with minimizer  $\mathbf{u}^*(\mathbf{x}, \mathbf{y}, \mathbf{z})$  and the definition of K.

Finally, we add and subtract  $K(\mathbf{u}^*(\mathbf{x}, \mathbf{y}, \mathbf{z}'), \mathbf{y}, \mathbf{z})$  to get

$$\begin{split} &K(\mathbf{u}^*(\mathbf{x},\mathbf{y},\mathbf{z}),\mathbf{y},\mathbf{z}) + \frac{\lambda}{2} \|\mathbf{u}^*(\mathbf{x},\mathbf{y},\mathbf{z}) - \mathbf{x}\|^2 - K(\mathbf{u}^*(\mathbf{x},\mathbf{y},\mathbf{z}'),\mathbf{y},\mathbf{z}') - \frac{\lambda}{2} \|\mathbf{u}^*(\mathbf{x},\mathbf{y},\mathbf{z}') - \mathbf{x}\|^2 \\ &= K(\mathbf{u}^*(\mathbf{x},\mathbf{y},\mathbf{z}),\mathbf{y},\mathbf{z}') + \frac{\lambda}{2} \|\mathbf{u}^*(\mathbf{x},\mathbf{y},\mathbf{z}) - \mathbf{x}\|^2 - K(\mathbf{u}^*(\mathbf{x},\mathbf{y},\mathbf{z}'),\mathbf{y},\mathbf{z}) - \frac{\lambda}{2} \|\mathbf{u}^*(\mathbf{x},\mathbf{y},\mathbf{z}') - \mathbf{x}\|^2 \\ &+ K(\mathbf{u}^*(\mathbf{x},\mathbf{y},\mathbf{z}),\mathbf{y},\mathbf{z}) - K(\mathbf{u}^*(\mathbf{x},\mathbf{y},\mathbf{z}),\mathbf{y},\mathbf{z}') \\ &\geq \frac{\gamma_s}{2} \|\mathbf{u}^*(\mathbf{x},\mathbf{y},\mathbf{z}) - \mathbf{u}^*(\mathbf{x},\mathbf{y},\mathbf{z}')\|^2 + \frac{\mu}{2} (\|\mathbf{u}^*(\mathbf{x},\mathbf{y},\mathbf{z}) - \mathbf{z}\|^2 - \|\mathbf{u}^*(\mathbf{x},\mathbf{y},\mathbf{z}) - \mathbf{z}'\|^2), \end{split}$$

where we used that  $\mathbf{u} \mapsto K(\mathbf{u}, \mathbf{y}, \mathbf{z}') + \frac{\lambda}{2} \|\mathbf{u} - \mathbf{x}\|^2$  is  $\gamma_s$ -strongly convex with minimizer  $\mathbf{u}^*(\mathbf{x}, \mathbf{y}, \mathbf{z}')$  and the definition of K.

Combining the last two inequalities give

$$\mu \langle \mathbf{u}^*(\mathbf{x}, \mathbf{y}, \mathbf{z}') - \mathbf{u}^*(\mathbf{x}, \mathbf{y}, \mathbf{z}), \mathbf{z}' - \mathbf{z} \rangle \geq \gamma_s \| \mathbf{u}^*(\mathbf{x}, \mathbf{y}, \mathbf{z}) - \mathbf{u}^*(\mathbf{x}, \mathbf{y}, \mathbf{z}') \|^2$$

Using Cauchy-Schwarz inequality concludes the proof.

Lemma A.8. Under Assumption 1.1, we have that

$$\|A\mathbf{x}_{t} - A\mathbf{x}^{*}(\mathbf{y}_{t+1}, \mathbf{z}_{t})\|^{2} \leq \frac{\|A\|^{2}\lambda^{2}}{\gamma^{2}} \|\mathbf{x}_{t} - \mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t})\|^{2},$$
(A.37)

$$\|A\mathbf{x}_{t} - \mathbf{b}\|^{2} \leq \frac{2\|A\|^{2}\lambda^{2}}{\gamma^{2}} \|\mathbf{x}_{t} - \mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t})\|^{2} + 2\|A\mathbf{x}^{*}(\mathbf{y}_{t+1}, \mathbf{z}_{t}) - \mathbf{b}\|^{2},$$
(A.38)

$$\|A\mathbf{u}^{*}(\mathbf{x}_{t},\mathbf{y}_{t},\mathbf{z}_{t}) - A\mathbf{x}_{t}\|^{2} \leq \frac{2\|A\|^{4}}{\gamma_{s}^{2}} \|\mathbf{y}_{t} - \mathbf{y}_{t+1}\|^{2} + 2\|A\|^{2} \|\mathbf{u}^{*}(\mathbf{x}_{t},\mathbf{y}_{t+1},\mathbf{z}_{t}) - \mathbf{x}_{t}\|^{2},$$
(A.39)

where  $\gamma, \gamma_s$  are defined in (A.1).

*Proof.* The assertion in (A.37) follows directly from (A.29) since

$$\|A\mathbf{x}_{t} - A\mathbf{x}^{*}(\mathbf{y}_{t+1}, \mathbf{z}_{t})\|^{2} \leq \|A\|^{2} \|\mathbf{x}_{t} - \mathbf{x}^{*}(\mathbf{y}_{t+1}, \mathbf{z}_{t})\|^{2} \leq \frac{\|A\|^{2}\lambda^{2}}{\gamma^{2}} \|\mathbf{x}_{t} - \mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t})\|^{2}.$$

Combining the first assertion with Young's inequality gives the second assertion, since

$$\begin{aligned} \|A\mathbf{x}_{t} - \mathbf{b}\|^{2} &\leq 2\|A\mathbf{x}_{t} - A\mathbf{x}^{*}(\mathbf{y}_{t+1}, \mathbf{z}_{t})\|^{2} + 2\|A\mathbf{x}^{*}(\mathbf{y}_{t+1}, \mathbf{z}_{t}) - \mathbf{b}\|^{2} \\ &\leq \frac{2\|A\|^{2}\lambda^{2}}{\gamma^{2}}\|\mathbf{x}_{t} - \mathbf{u}^{*}(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t})\|^{2} + 2\|A\mathbf{x}^{*}(\mathbf{y}_{t+1}, \mathbf{z}_{t}) - \mathbf{b}\|^{2}. \end{aligned}$$

Young's inequality and (A.31) gives the third assertion

$$\begin{aligned} \|A\mathbf{u}^{*}(\mathbf{x}_{t},\mathbf{y}_{t},\mathbf{z}_{t}) - A\mathbf{x}_{t}\|^{2} &\leq 2\|A\mathbf{u}^{*}(\mathbf{x}_{t},\mathbf{y}_{t},\mathbf{z}_{t}) - A\mathbf{u}^{*}(\mathbf{x}_{t},\mathbf{y}_{t+1},\mathbf{z}_{t})\|^{2} + 2\|A\mathbf{u}^{*}(\mathbf{x}_{t},\mathbf{y}_{t+1},\mathbf{z}_{t}) - A\mathbf{x}_{t}\|^{2} \\ &\leq \frac{2\|A\|^{4}}{\gamma_{s}^{2}}\|\mathbf{y}_{t} - \mathbf{y}_{t+1}\|^{2} + 2\|A\|^{2}\|\mathbf{u}^{*}(\mathbf{x}_{t},\mathbf{y}_{t+1},\mathbf{z}_{t}) - \mathbf{x}_{t}\|^{2}. \end{aligned}$$

The completes the proof.

The following important lemma is known as the global error bound in Zhang & Luo (2022). This global result holds in its entirety in our case, so we only state it here and refer to where it appeared originally for the precise definition of the constant  $\bar{\sigma}$  which depends on Hoffman constant of certain linear systems.

**Lemma A.9.** (*Zhang & Luo, 2022, Lemma 3.2*) If  $\mu > L_f$ , then we have

$$\|\mathbf{x}^*(\mathbf{y}, \mathbf{z}) - \bar{\mathbf{x}}^*(\mathbf{z})\| \le \bar{\sigma} \|A\mathbf{x}^*(\mathbf{y}, \mathbf{z}) - \mathbf{b}\| \quad for \ any \quad \mathbf{y}, \mathbf{z}$$

where  $\bar{\sigma} > 0$  depends only on the constants  $C_1 = (L_f + \rho ||A||^2 + \mu)$ ,  $C_2 = -L_f + \mu$ , and the matrices A, H and is always finite.

**Fact A.10.** For  $\mathbf{x} \in X$ , we have that  $\mathbf{x} \mapsto K(\mathbf{x}, \mathbf{y}, \mathbf{z})$  is strongly convex with modulus  $\gamma_K = \mu - L_f$ , and  $\mathbf{x} \mapsto \nabla_{\mathbf{x}} K(\mathbf{x}, \mathbf{y}, \mathbf{z})$  is  $(L_f + \rho ||A||^2 + \mu)$ -Lipschitz continuous.

For  $\mathbf{u} \in X$ ,  $\mathbf{u} \mapsto K(\mathbf{u}, \mathbf{y}, \mathbf{z}) + \frac{\lambda}{2} \|\mathbf{x} - \mathbf{u}\|^2$  is strongly convex with modulus  $\gamma_s = \mu - L_f + \lambda$ , and  $\mathbf{u}^*(\mathbf{x}, \mathbf{y}, \mathbf{z})$  is the minimizer of  $K(\cdot, \mathbf{y}, \mathbf{z}) + I_X(\mathbf{u}) + \frac{\lambda}{2} \|\mathbf{x} - \mathbf{u}\|^2$ .

**Lemma A.11.** (*Planiden & Wang, 2016, Lemma 2.19*) Let r > 0. The function f is r-strongly convex if and only if  $f_1(\mathbf{x}) = \min_{\mathbf{u}} f(\mathbf{u}) + \frac{1}{2} ||\mathbf{x} - \mathbf{u}||^2$  is  $\frac{r}{r+1}$ -strongly convex.

**Lemma A.12.** The function  $\mathbf{x} \mapsto \varphi_{1/\lambda}(\mathbf{x}, \mathbf{y}, \mathbf{z})$  is  $\gamma = \frac{(\mu - L_f)\lambda}{\mu - L_f + \lambda}$ -strongly convex.

*Proof.* By definition, we have

$$\varphi_{1/\lambda}(\mathbf{x}, \mathbf{y}, \mathbf{z}) = \min_{\mathbf{u}} \left\{ K(\mathbf{u}, \mathbf{y}, \mathbf{z}) + I_X(\mathbf{u}) + \frac{\lambda}{2} \|\mathbf{x} - \mathbf{u}\|^2 \right\} = \lambda \min_{\mathbf{u}} \left\{ \frac{K(\mathbf{u}, \mathbf{y}, \mathbf{z}) + I_X(\mathbf{u})}{\lambda} + \frac{1}{2} \|\mathbf{x} - \mathbf{u}\|^2 \right\}$$

Recall that  $\gamma_K = \mu - L_f$ . Then, since  $K(\mathbf{x}, \mathbf{y}, \mathbf{z})/\lambda$  is  $\frac{\gamma_K}{\lambda}$ -strongly convex, we have  $\min_{\mathbf{u}} \frac{K(\mathbf{u}, \mathbf{y}, \mathbf{z}) + I_X(\mathbf{u})}{\lambda} + \frac{1}{2} \|\mathbf{x} - \mathbf{u}\|^2$  is  $\frac{\gamma_K/\lambda}{\gamma_K/\lambda + 1}$ -strongly convex, by Lemma A.11. Hence,  $\varphi_{1/\lambda}(\mathbf{x}, \mathbf{y}, \mathbf{z})$  is strongly convex with modulus  $\frac{\gamma_K}{\gamma_K/\lambda + 1} = \frac{\lambda\gamma_K}{\lambda + \gamma_K} = \frac{(\mu - L_f)\lambda}{\mu - L_f + \lambda}$ .

Lemma A.13. If  $\mathbf{x} \in X$ , we have  $\varphi_{1/\lambda}(\mathbf{x}, \mathbf{y}, \mathbf{z}) - 2d(\mathbf{y}, \mathbf{z}) + 2\Psi(\mathbf{z}) \geq f$ .

*Proof.* Because  $\mathbf{x}^*(\mathbf{y}, \mathbf{z})$  minimizes  $\varphi_{1/\lambda}(\cdot, \mathbf{y}, \mathbf{z})$  (see for example (Hiriart-Urruty & Lemarechal, 1993, Theorem XV4.1.7)), we have

$$\varphi_{1/\lambda}(\mathbf{x},\mathbf{y},\mathbf{z}) \ge \varphi_{1/\lambda}(\mathbf{x}^*(\mathbf{y},\mathbf{z})) = K(\mathbf{x}^*(\mathbf{y},\mathbf{z})).$$

We can then deduce

$$\begin{split} \varphi_{1/\lambda}(x,y,z) - 2d(y,z) + 2\Psi(z) &\geq K(x(y,z),y,z) - 2d(y,z) + 2\Psi(z) \\ &= d(y,z) - 2d(y,z) + 2\Psi(z) \\ &= \Psi(z) + \Psi(z) - d(y,z) \\ &\geq \Psi(z) \\ &\geq f \end{split}$$

The second inequality in the above chain comes from definition, that is, denoting  $\mathbf{x}_{\mu}^{*} = \arg \min_{x \in X, A\mathbf{x}=\mathbf{b}} \{f(\mathbf{x}) + \frac{\mu}{2} \|\mathbf{x} - \mathbf{z}\|^{2} \}$  in view of (2.1), we have

$$d(\mathbf{y}, \mathbf{z}) = \min_{\mathbf{x} \in X} K(\mathbf{x}, \mathbf{y}, \mathbf{z}) \le K(\mathbf{x}_{\mu}^*, \mathbf{y}, \mathbf{z}) = f(\mathbf{x}_{\mu}^*) + \frac{\mu}{2} \|\mathbf{x}_{\mu}^* - \mathbf{z}\|^2 = \Psi(\mathbf{z}),$$

where the first inequality also uses  $\mathbf{x}^*_{\mu} \in X$ , which is by definition.

### **B** Proofs for Section 4

### B.1 Proof of Theorem 4.3

**Lemma B.1.** Let Assumption 4.2 hold. With the update rule of  $\mathbf{y}_{t+1} = \mathbf{y}_t + \eta(A_{\zeta}\mathbf{x}_t - \mathbf{b}_{\zeta})$ , where  $\mathbb{E}_{\zeta}[A_{\zeta}\mathbf{x}_t - \mathbf{b}_{\zeta}] = A\mathbf{x}_t - \mathbf{b}$ , we have

$$\mathbb{E}d(\mathbf{y}_{t+1}, \mathbf{z}_{t+1}) - \mathbb{E}d(\mathbf{y}_t.\mathbf{z}_t) \ge \eta \mathbb{E}\langle (A\mathbf{x}_t - \mathbf{b}), Ax(\mathbf{y}_{t+1}, \mathbf{z}_t) - b \rangle - \frac{\eta^2}{32} \mathbb{E} \|A\mathbf{x}_t - \mathbf{b}\|^2 - (\frac{1}{2} + \frac{17\|A\|^2}{2\gamma_K^2})\eta^2 L^2 + \frac{\mu}{2} \mathbb{E}\langle \mathbf{z}_{t+1} - \mathbf{z}_t, \mathbf{z}_{t+1} + \mathbf{z}_t - 2\mathbf{x}(\mathbf{y}_{t+1}, \mathbf{z}_{t+1})\rangle, \\ \mathbb{E}\Psi(\mathbf{z}_{t+1}) - \mathbb{E}\Psi(\mathbf{z}_t) \le \mu \mathbb{E}\langle \mathbf{z}_{t+1} - \mathbf{z}_t, \mathbf{z}_t - \bar{\mathbf{x}}^*(\mathbf{z}_t)\rangle + \frac{\mu}{2\sigma_4} \mathbb{E}\|\mathbf{z}_t - \mathbf{z}_{t+1}\|^2$$
(B.1)

where  $\gamma_K, \sigma_4$  are introduced in A.7, and we assume  $\mathbb{E} \|A_{\zeta_t} \mathbf{x}_t - \mathbf{b}_{\zeta_t}\|^2 \leq L$ .

Proof. It is easy to derive, for example as (Zhang & Luo, 2020, Lemma 3.2), that

$$d(\mathbf{y}_{t+1}, \mathbf{z}_{t+1}) - d(\mathbf{y}_t, \mathbf{z}_t) \ge \langle \mathbf{y}_{t+1} - \mathbf{y}_t, A\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{b} \rangle + \frac{\mu}{2} \langle \mathbf{z}_{t+1} - \mathbf{z}_t, \mathbf{z}_{t+1} + \mathbf{z}_t - 2\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_{t+1}) \rangle.$$

Hence, by using the update rule of  $\mathbf{y}_{t+1}$ , we get

$$\begin{aligned} d(\mathbf{y}_{t+1}, \mathbf{z}_{t+1}) - d(\mathbf{y}_t, \mathbf{z}_t) &\geq \langle \mathbf{y}_{t+1} - \mathbf{y}_t, A\mathbf{x}^*(\mathbf{y}_t, \mathbf{z}_t) - \mathbf{b} \rangle + \langle \mathbf{y}_{t+1} - \mathbf{y}_t, A\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t) - A\mathbf{x}^*(\mathbf{y}_t, \mathbf{z}_t) \rangle \\ &+ \frac{\mu}{2} \langle \mathbf{z}_{t+1} - \mathbf{z}_t, \mathbf{z}_{t+1} + \mathbf{z}_t - 2\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_{t+1}) \rangle \\ &\geq \langle \mathbf{y}_{t+1} - \mathbf{y}_t, A\mathbf{x}^*(\mathbf{y}_t, \mathbf{z}_t) - \mathbf{b} \rangle - \frac{1}{2} \| \mathbf{y}_{t+1} - \mathbf{y}_t \|^2 - \frac{1}{2} \| A\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t) - A\mathbf{x}^*(\mathbf{y}_t, \mathbf{z}_t) \|^2 \\ &+ \frac{\mu}{2} \langle \mathbf{z}_{t+1} - \mathbf{z}_t, \mathbf{z}_{t+1} + \mathbf{z}_t - 2\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_{t+1}) \rangle \\ &\geq \langle \eta(A(\omega_t)\mathbf{x}_t - \mathbf{b}(\omega_t)), A\mathbf{x}^*(\mathbf{y}_t, \mathbf{z}_t) - \mathbf{b} \rangle - (\frac{1}{2} + \frac{\|A\|^2}{2\gamma_K^2}) \eta^2 L^2 \\ &+ \frac{p}{2} \langle \mathbf{z}_{t+1} - \mathbf{z}_t, \mathbf{z}_{t+1} + \mathbf{z}_t - 2\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_{t+1}) \rangle, \end{aligned}$$

where we introduce a term  $A\mathbf{x}^*(\mathbf{y}_t, \mathbf{z}_t)$  in the first inequality. Then we use Cauchy-Schwarz inequality in the second step, and the last inequality comes from (A.34),  $||A\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t) - A\mathbf{x}^*(\mathbf{y}_t, \mathbf{z}_t)||^2 \leq \frac{||A||^2}{\gamma_K^2} ||\mathbf{y}_{t+1} - \mathbf{y}_t||^2$  and the bound of  $\mathbb{E}||A_{\zeta_t}\mathbf{x}_t - \mathbf{b}_{\zeta_t}||^2$ .

After taking expectation and using tower property along with  $\mathbf{y}_t, \mathbf{z}_t$  being deterministic under the conditioning, we have

$$\mathbb{E}d(\mathbf{y}_{t+1}, \mathbf{z}_{t+1}) - \mathbb{E}d(\mathbf{y}_t, \mathbf{z}_t) \ge \eta \mathbb{E}\langle (A\mathbf{x}_t - \mathbf{b}), A\mathbf{x}^*(\mathbf{y}_t, \mathbf{z}_t) - \mathbf{b} \rangle - \left(\frac{1}{2} + \frac{\|A\|^2}{2\gamma_K^2}\right) \eta^2 L^2 + \frac{\mu}{2} \mathbb{E}\langle \mathbf{z}_{t+1} - \mathbf{z}_t, \mathbf{z}_{t+1} + \mathbf{z}_t - 2\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_{t+1}) \rangle$$
(B.2)

Then we estimate the first term in the above inequality. We have

$$\begin{split} \eta \mathbb{E} \langle A\mathbf{x}_{t} - \mathbf{b}, A\mathbf{x}^{*}(\mathbf{y}_{t}, \mathbf{z}_{t}) - \mathbf{b} \rangle \\ &= \eta \mathbb{E} [\langle (A\mathbf{x}_{t} - \mathbf{b}), A\mathbf{x}^{*}(\mathbf{y}_{t+1}, \mathbf{z}_{t}) - \mathbf{b} \rangle + \eta \langle (A\mathbf{x}_{t} - \mathbf{b}), A\mathbf{x}^{*}(\mathbf{y}_{t+1}, \mathbf{z}_{t}) - A\mathbf{x}^{*}(\mathbf{y}_{t}, \mathbf{z}_{t}) \rangle] \\ &\geq \eta \mathbb{E} [\langle (A\mathbf{x}_{t} - \mathbf{b}), A\mathbf{x}^{*}(\mathbf{y}_{t+1}, \mathbf{z}_{t}) - \mathbf{b} \rangle - 8 \|A\mathbf{x}^{*}(\mathbf{y}_{t+1}, \mathbf{z}_{t}) - A\mathbf{x}^{*}(\mathbf{y}_{t}, \mathbf{z}_{t})\|^{2}] - \frac{\eta^{2}}{32} \|A\mathbf{x}_{t} - \mathbf{b}\|^{2} \\ &\geq \eta \mathbb{E} [\langle (A\mathbf{x}_{t} - \mathbf{b}), A\mathbf{x}^{*}(\mathbf{y}_{t+1}, \mathbf{z}_{t}) - \mathbf{b} \rangle - \frac{8 \|A\|^{2}}{\gamma_{k}^{2}} \eta^{2} L^{2}] - \frac{\eta^{2}}{32} \|A\mathbf{x}_{t} - \mathbf{b}\|^{2}, \end{split}$$

where we introduce a term  $A\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t)$  to get the first equality. The second inequality comes from Young inequality  $\langle a, b \rangle \leq \frac{1}{32} \|a\|^2 + 8\|b\|^2 \forall a, b$ . In last inequality, we use (A.34) and  $\mathbb{E}[\|A_{\zeta_t}\mathbf{x}_t - \mathbf{b}_{\zeta_t}\|^2] \leq L$  again. Finally, plug (B.3) to (B.2), we obtain the desired result.

**Lemma B.2.** Let Assumption 1.1 and 4.2 hold. By using the parameters (A.1) in Algorithm 1 with the dual update changed to  $\mathbf{y}_{t+1} = \mathbf{y}_t + \eta(A_{\zeta}\mathbf{x}_t - \mathbf{b}_{\zeta})$ , we obtain

$$\mathbb{E}V_t - \mathbb{E}V_{t+1} \ge \widetilde{c}_{\beta} \mathbb{E} \|\mathbf{z}_{t+1} - \mathbf{z}_t\|^2 + \widetilde{c}_{\tau} \mathbb{E} \|\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{x}_t\|^2 + \widetilde{c}_{\eta} \mathbb{E} \|A\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{b}\|^2 - \lambda \tau^2 \sigma_2^2 - (1 + \frac{17\|A\|^2}{\gamma_K^2})\eta^2 L^2$$
(B.3)

where  $\widetilde{c}_{\beta} = \frac{\mu}{50\beta}$ ,  $\widetilde{c}_{\tau} = \frac{6\tau\lambda^2}{400}$ ,  $\widetilde{c}_{\eta} = \frac{\eta}{8}$  and  $\mathbb{E}\|\widehat{G}(\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t, \xi_t) - \nabla_x K(\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t)\|^2 \le \sigma_2^2$ 

*Proof.* First, we show  $\mathbb{E} \| \hat{G}(\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t, \xi_t) - \nabla_x K(\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t) \|^2$  is bounded.

$$\begin{split} & \mathbb{E} \| \hat{G}(\mathbf{x}_{t}, \mathbf{y}_{t}, \mathbf{z}_{t}, \xi_{t}) - \nabla_{x} K(\mathbf{x}_{t}, \mathbf{y}_{t}, \mathbf{z}_{t}) \|^{2} \\ & \leq \mathbb{E} 2 \| \hat{G}(\mathbf{x}_{t}, \mathbf{y}_{t}, \mathbf{z}_{t}, \xi_{t}) - \hat{G}(\mathbf{x}_{t}, 0, \mathbf{z}_{t}, \xi_{t}) \|^{2} + \mathbb{E} 2 \| \hat{G}(\mathbf{x}_{t}, 0, \mathbf{z}_{t}, \xi_{t}) - K(\mathbf{x}_{t}, \mathbf{y}_{t}, \mathbf{z}_{t}) \|^{2} \\ & \leq 2 \mathbb{E} L_{G} \| \mathbf{y}_{t} \|^{2} + 2 \mathbb{E} \| \hat{G}(\mathbf{x}_{t}, 0, \mathbf{z}_{t}, \xi_{t}) - \nabla_{x} K(\mathbf{x}_{t}, \mathbf{y}_{t}, \mathbf{z}_{t}) \|^{2} \\ & \leq 2 \mathbb{E} L_{G} \| \mathbf{y}_{t} \|^{2} + 4 \mathbb{E} \| \hat{G}(\mathbf{x}_{t}, 0, \mathbf{z}_{t}, \xi_{t}) - \nabla_{x} K(\mathbf{x}_{t}, 0, \mathbf{z}_{t}) \|^{2} \\ & \leq 2 \mathbb{E} L_{G} \| \mathbf{y}_{t} \|^{2} + 4 \mathbb{E} \| \hat{G}(\mathbf{x}_{t}, 0, \mathbf{z}_{t}, \xi_{t}) - \nabla_{x} K(\mathbf{x}_{t}, 0, \mathbf{z}_{t}) \|^{2} + 4 \mathbb{E} \| \nabla_{x} K(\mathbf{x}_{t}, 0, \mathbf{z}_{t}) - \nabla_{x} K(\mathbf{x}_{t}, \mathbf{y}_{t}, \mathbf{z}_{t}) \|^{2} \\ & \leq 2 L_{G} M_{y}^{2} + 4 \| A \|^{2} \| \mathbf{y}_{t} \|^{2} + 4 \mathbb{E} \| \hat{G}(\mathbf{x}_{t}, 0, \mathbf{z}_{t}, \xi_{t}) - \nabla_{x} K(\mathbf{x}_{t}, 0, \mathbf{z}_{t}) \|^{2} \end{split}$$

Because  $\mathbf{x}, \mathbf{y}, \mathbf{z}$  are all bounded,  $\mathbb{E} \| \hat{G}(\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t, \xi_t) - \nabla_x K(\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t) \|^2$  is bounded, we denote the upper bound as  $\sigma_2^2$ .

Combining with deterministic linear result, we have:

$$\begin{split} \mathbb{E}V_t - \mathbb{E}V_{t+1} &\geq c_{\beta} \mathbb{E} \|\mathbf{z}_{t+1} - \mathbf{z}_t\|^2 + c_{\tau} \mathbb{E} \|\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{x}_t\|^2 + c_{\eta} \mathbb{E} \|A\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{b}\|^2 - \lambda \tau^2 \sigma_2^2 \\ &- \frac{\eta^2}{16} \mathbb{E} \|A\mathbf{x}_t - \mathbf{b}\|^2 - (1 + \frac{17\|A\|^2}{\gamma_K^2}) \eta^2 L^2 \end{split}$$

where  $c_{\beta} = \frac{\mu}{50\beta}$ ,  $c_{\tau} = \frac{7\tau\lambda^2}{400}$ ,  $c_{\eta} = \frac{\eta}{4}$ . Because

$$-\frac{\eta^2}{16}\mathbb{E}\|A\mathbf{x}_t - \mathbf{b}\|^2 \ge -\frac{\|A\|^2\lambda^2\eta^2}{8\gamma^2}\|\mathbf{x}_t - \mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t)\|^2 - \frac{\eta^2}{8}\|A\mathbf{x}(\mathbf{y}_{t+1}, \mathbf{z}_t) - b\|^2$$

By the parameter choices, we have  $\frac{7\tau\lambda^2}{400} - \frac{\|A\|^2\lambda^2\eta^2}{8\gamma^2} \ge \frac{6\tau\lambda^2}{400}$  and  $\frac{\eta}{4} - \frac{\eta^2}{8} \ge \frac{\eta}{8}$ .

**Proposition B.3.** Under Assumption 4.2,  $\|\mathbf{y}_t\| \leq \frac{\Psi(\mathbf{z}_t) - d(\mathbf{y}_t, \mathbf{z}_t) + 2M}{r}$ , where  $M = \max_{\mathbf{x}, \mathbf{z} \in X} \{|f(\mathbf{x})| + \frac{\mu}{2} \|\mathbf{x} - \mathbf{z}\|^2 + \frac{\rho}{2} \|A\mathbf{x} - \mathbf{b}\|^2 \}$  and r > 0 is defined as  $\|A\hat{\mathbf{x}} - \mathbf{b}\| = r$  where  $\hat{\mathbf{x}}$  is in the relative interior of the constraints. The existence of this is guaranteed by our assumption.

*Proof.* Given  $\widetilde{\mathbf{x}} \in X$ , we have

$$\begin{split} \Psi(\mathbf{z}_{t}) - d(\mathbf{y}_{t}, \mathbf{z}_{t}) &\geq f(\bar{\mathbf{x}}^{*}(\mathbf{z}_{t})) + \frac{\mu}{2} \|\bar{\mathbf{x}}^{*}(\mathbf{z}_{t}) - \mathbf{z}_{t}\|^{2} - K(\tilde{\mathbf{x}}, \mathbf{y}_{t}, \mathbf{z}_{t}) \\ &\geq f(\bar{\mathbf{x}}^{*}(\mathbf{z}_{t})) + \frac{\mu}{2} \|\bar{\mathbf{x}}^{*}(\mathbf{z}_{t}) - \mathbf{z}_{t}\|^{2} - [f(\tilde{\mathbf{x}}) + \langle \mathbf{y}_{t}, A\tilde{\mathbf{x}} \rangle + \frac{\rho}{2} \|A\tilde{\mathbf{x}} - \mathbf{b}\|^{2} + \frac{\mu}{2} \|\tilde{\mathbf{x}} - \mathbf{z}_{t}\|^{2}] \\ &= [f(\bar{\mathbf{x}}^{*}(\mathbf{z}_{t})) + \frac{\mu}{2} \|\bar{\mathbf{x}}^{*}(\mathbf{z}_{t}) - \mathbf{z}_{t}\|^{2} - f(\tilde{\mathbf{x}}) - \frac{\mu}{2} \|\tilde{\mathbf{x}} - \mathbf{z}_{t}\|^{2}] - \langle \mathbf{y}_{t}, A\tilde{\mathbf{x}} - \mathbf{b} \rangle - \frac{\rho}{2} \|A\tilde{\mathbf{x}} - \mathbf{b}\|^{2} \\ &= [f(\bar{\mathbf{x}}^{*}(\mathbf{z}_{t})) + \frac{\mu}{2} \|\bar{\mathbf{x}}^{*}(\mathbf{z}_{t}) - \mathbf{z}_{t}\|^{2} - f(\tilde{\mathbf{x}}) - \frac{\mu}{2} \|\tilde{\mathbf{x}} - \mathbf{z}_{t}\|^{2} - \frac{\rho}{2} \|A\tilde{\mathbf{x}} - \mathbf{b}\|^{2}] - \langle \mathbf{y}_{t}, A\tilde{\mathbf{x}} - \mathbf{b} \rangle \\ &\geq -2M - \langle \mathbf{y}_{t}, A\tilde{\mathbf{x}} - \mathbf{b} \rangle. \end{split}$$

Where the first inequality comes from the definition of  $\Psi(\mathbf{z}_t)$  and

$$d(\mathbf{y}_t, \mathbf{z}_t) = \min_{\mathbf{x} \in X} K(\mathbf{x}, \mathbf{y}, \mathbf{z})$$

And in the last inequality, we let

$$M = \max_{\mathbf{x}, \mathbf{z} \in X} \{ |f(\mathbf{x})| + \frac{\mu}{2} \|\mathbf{x} - \mathbf{z}\|^2 + \frac{\rho}{2} \|A\mathbf{x} - \mathbf{b}\|^2 \}$$

So we have the last inequality.

According to Assumption 4.2(2), there exists a positive r > 0 such that for any direction  $\mathbf{d} \in \text{Range}(A)$ , we can find a  $\mathbf{x} \in X$  satisfying  $||A\mathbf{x} - \mathbf{b}|| = r$  and  $A\mathbf{x} - \mathbf{b}$  has the same direction as  $\mathbf{d}$ . Because  $\mathbf{y}_t \in \text{Range}(A)$ (by assumption 4.2(3), Range(A) =  $\mathbb{R}^m$ ) we can choose  $\widetilde{\mathbf{x}}$  such that  $A\widetilde{\mathbf{x}} - \mathbf{b}$  is of the same direction as  $-\mathbf{y}_t$ and  $||A\widetilde{\mathbf{x}} - \mathbf{b}|| = r$ . Then we obtain

$$\Psi(\mathbf{z}_t) - d(\mathbf{y}_t, \mathbf{z}_t) \ge -2M + r \|\mathbf{y}_t\| \Longrightarrow \|\mathbf{y}_t\| \le \frac{\Psi(\mathbf{z}_t) - d(\mathbf{y}_t, \mathbf{z}_t) + 2M}{r}, \forall t \in \{0, 1, ..., T\}.$$
  
udes the proof.

This concludes the proof.

#### Then we start the proof for Theorem 4.3

*Proof.* First, let  $M_V = \max_{\mathbf{x}, \mathbf{z} \in X} \{K(\mathbf{x}, 0, \mathbf{z}) - 2d(0, \mathbf{z}) + 2\Psi(\mathbf{z})\}$  and  $M_y > \frac{M_V - M_\Psi + 2M}{r}$  where  $M_\Psi$  is a uniform lower bound of  $\Psi(\mathbf{z}_t)$ , for example, f.

If  $\|\mathbf{y}_{t+1}\| \leq M_y$ , then

$$\mathbb{E}V_t - \mathbb{E}V_{t+1} \ge \widetilde{c}_{\beta}\mathbb{E}\|\mathbf{z}_{t+1} - \mathbf{z}_t\|^2 + \widetilde{c}_{\tau}\mathbb{E}\|\mathbf{u}^*(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{x}_t\|^2 + \widetilde{c}_{\eta}\mathbb{E}\|A\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{b}\|^2 - \lambda\tau^2\sigma_2^2 - (1 + \frac{17\|A\|^2}{\gamma_K^2})\eta^2 L^2$$
(B.4)

according to the analysis of Lemma B.2.

If  $\|\mathbf{y}_{t+1}\| > M_y$ . For distinction, if we perform the procedure  $\mathbf{y}_{t+1} = 0$ , let us denote the update as  $\mathbf{y}_{t+1}, \mathbf{x}_{t+1}, \mathbf{z}_{t+1}$  as  $\hat{\mathbf{y}}_{t+1}, \hat{\mathbf{x}}_{t+1}, \hat{\mathbf{z}}_{t+1}$  and  $\mathbf{y}_{t+1}, \mathbf{x}_{t+1}, \mathbf{z}_{t+1}$  denote the iteration generated without taking  $y_{t+1} = 0$ . Then

$$\begin{split} K(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) &- 2d(\mathbf{y}_{t+1}, \mathbf{z}_{t+1}) + 2\Psi(\mathbf{z}_{t+1}) \ge \Psi(\mathbf{z}_{t+1}) - d(\mathbf{y}_{t+1}, \mathbf{z}_{t+1}) + \Psi(\mathbf{z}_{t+1}) \\ &\ge r \|\mathbf{y}_{t+1}\| - 2M + M_{\Psi} \\ &\ge rM_y - 2M + M_{\Psi} \\ &\ge M_V \\ &= \max_{\mathbf{x}, \mathbf{z} \in X} \{K(\mathbf{x}, 0, \mathbf{z}) - 2d(0, \mathbf{z}) + 2\Psi(\mathbf{z})\} \\ &\ge K(\hat{\mathbf{x}}_{t+1}, 0, \hat{\mathbf{z}}_{t+1}) - 2d(0, \hat{\mathbf{z}}_{t+1}) + 2\Psi(\hat{\mathbf{z}}_{t+1}), \end{split}$$

where the first step used  $d(\mathbf{y}_{t+1}, \mathbf{z}_{t+1}) \leq K(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}, \mathbf{z}_{t+1})$  and the second line used  $\Psi(\mathbf{z}_{t+1}) \geq M_{\Psi}$ .

Hence,  $\mathbb{E}V_t - \mathbb{E}V_{t+1}$  becomes larger if we run  $\mathbf{y}_{t+1} = 0$ . So (B.4) still holds, then the convergence result follows.

The rest of the proof for the complexity result proceeds the same as Appendix A.2 up to simple changes in the constants, and hence is omitted. 

#### С **Proof for Section 5**

Lemma C.1. (Zhang & Luo, 2020, Lemma 3.10) Under Assumption 1.1, we have

$$\|\mathbf{x} - \operatorname{proj}_{X}(\mathbf{x} - \tau \nabla K(\mathbf{x}, \mathbf{y}, \mathbf{z}))\| \ge \tau (\mu - L_{f}) \|\mathbf{x} - \mathbf{x}^{*}(\mathbf{y}, \mathbf{z})\|,$$

where  $K(\mathbf{x}, \mathbf{y}, \mathbf{z}) = L_{\rho}(\mathbf{x}, \mathbf{y}) + \frac{\mu}{2} \|\mathbf{x} - \mathbf{z}\|^2$ , and  $\mathbf{x}^*(\mathbf{y}, \mathbf{z}) = \min_{\mathbf{x} \in X} K(\mathbf{x}, \mathbf{y}, \mathbf{z})$ .

Lemma C.2. Under Assumption 1.1, for the iterates generated by Algorithm 3 we have

$$\|\mathbf{x}_t - \mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t)\| \le \frac{1}{\tau(\mu - L_f)} \|\mathbf{x}_t - \mathbf{x}_{t+1}\| + \frac{1}{(\mu - L_f)} \|\widehat{\nabla}f_t - \nabla f(\mathbf{x}_t)\|$$

*Proof.* Taking  $\mathbf{x}, \mathbf{y}, \mathbf{z}$  as  $\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t$  in Lemma C.1, we have

$$\begin{aligned} \|\mathbf{x}_{t} - \mathbf{x}^{*}(\mathbf{y}_{t+1}, \mathbf{z}_{t})\| &\leq \frac{1}{\tau(\mu - L_{f})} \|\mathbf{x}_{t} - \operatorname{proj}_{X}(\mathbf{x}_{t} - \tau \nabla K(\mathbf{x}, \mathbf{y}_{t+1}, \mathbf{z}_{t}))\| \\ &\leq \frac{1}{\tau(\mu - L_{f})} \|\mathbf{x}_{t} - \operatorname{proj}_{X}(\mathbf{x} - \tau G(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t}))\| \\ &+ \frac{1}{\tau(\mu - L_{f})} \|\operatorname{proj}_{X}(\mathbf{x}_{t} - \tau \nabla K(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t})) - \operatorname{proj}_{X}(\mathbf{x}_{t} - \tau G(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t}))\| \\ &\leq \frac{1}{\tau(\mu - L_{f})} \|\mathbf{x}_{t} - \mathbf{x}_{t+1}\| + \frac{1}{(\mu - L_{f})} \|\widehat{\nabla}f_{t} - \nabla f(\mathbf{x}_{t})\|, \end{aligned}$$

where the second inequality comes form triangle inequality and the last inequality comes from the fact that  $\operatorname{proj}_X$  is nonexpansive and  $\nabla K(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) - G(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) = \widehat{\nabla} f_t - \nabla f(\mathbf{x}_t)$ 

Proof of Lemma 5.3. By the definition of  $\widehat{\nabla} f_{t+1}$ , we have

$$\begin{split} \widehat{\nabla}f_{t+1} &- \nabla f(\mathbf{x}_{t+1}) \\ &= \nabla f(\mathbf{x}_{t+1}, \xi_{t+1}) + (1-\alpha)(\widehat{\nabla}f_t - \nabla f(\mathbf{x}_t, \xi_{t+1})) - \nabla f(\mathbf{x}_{t+1}) \\ &= \nabla f(\mathbf{x}_{t+1}, \xi_{t+1}) + (1-\alpha)(\widehat{\nabla}f_t - \nabla f(\mathbf{x}_t)) + (1-\alpha)(\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_t, \xi_{t+1})) - \nabla f(\mathbf{x}_{t+1}) \\ &= (1-\alpha)(\widehat{\nabla}f_t - \nabla f(\mathbf{x}_t)) + (1-\alpha)(\nabla f(\mathbf{x}_t) - f(\mathbf{x}_t, \xi_{t+1})) + f(\mathbf{x}_{t+1}, \xi_{t+1}) - \nabla f(\mathbf{x}_{t+1}), \end{split}$$
(C.1)

where in the second equality, we added and subtracted  $(1 - \alpha)\nabla f(\mathbf{x}_t)$ .

Then, we compute the squared norm of (C.1) and expand to get

$$\begin{split} \|\widehat{\nabla}f_{t+1} - \nabla f(\mathbf{x}_{t+1})\|^2 \\ &= (1-\alpha)^2 \|\widehat{\nabla}f_t - \nabla f(\mathbf{x}_t)\|^2 + \|(1-\alpha)(\nabla f(\mathbf{x}_t) - f(\mathbf{x}_t, \xi_{t+1})) + f(\mathbf{x}_{t+1}, \xi_{t+1}) - \nabla f(\mathbf{x}_{t+1})\|^2 \\ &+ 2(1-\alpha)\langle\widehat{\nabla}f_t - \nabla f(\mathbf{x}_t), (1-\alpha)(\nabla f(\mathbf{x}_t) - f(\mathbf{x}_t, \xi_{t+1})) + f(\mathbf{x}_{t+1}, \xi_{t+1}) - \nabla f(\mathbf{x}_{t+1})\rangle. \end{split}$$

Next, we take expectation with respect to  $\xi_{t+1}$  to obtain

$$\mathbb{E}_{\xi_{t+1}} \|\nabla f_{t+1} - \nabla f(\mathbf{x}_{t+1})\|^{2} = (1 - \alpha)^{2} \mathbb{E}_{\xi_{t+1}} \|\nabla f_{t} - \nabla f(\mathbf{x}_{t})\|^{2} + \mathbb{E}_{\xi_{t+1}} \|(1 - \alpha)(\nabla f(\mathbf{x}_{t}) - f(\mathbf{x}_{t}, \xi_{t+1})) + f(\mathbf{x}_{t+1}, \xi_{t+1}) - \nabla f(\mathbf{x}_{t+1})\|^{2}, \tag{C.2}$$

which is due to  $\widehat{\nabla} f_t - \nabla f(\mathbf{x}_t)$  being independent of  $\xi_{t+1}$ , and

$$\mathbb{E}_{\xi_{t+1}}[\nabla f(\mathbf{x}_t) - f(\mathbf{x}_t, \xi_{t+1})] = 0, \mathbb{E}_{\xi_{t+1}}[f(\mathbf{x}_{t+1}, \xi_{t+1}) - \nabla f(\mathbf{x}_{t+1})] = 0.$$

Finally, we estimate the last term in the right-hand side of (C.2):

$$\begin{split} & \mathbb{E}_{\xi_{t+1}} \| (1-\alpha) (\nabla f(\mathbf{x}_{t}) - f(\mathbf{x}_{t}, \xi_{t+1})) + f(\mathbf{x}_{t+1}, \xi_{t+1}) - \nabla f(\mathbf{x}_{t+1}) \|^{2} \\ & = \mathbb{E}_{\xi_{t+1}} \| f(\mathbf{x}_{t+1}, \xi_{t+1}) - f(\mathbf{x}_{t}, \xi_{t+1}) + \nabla f(\mathbf{x}_{t}) - \nabla f(\mathbf{x}_{t+1}) + \alpha (f(\mathbf{x}_{t}, \xi_{t+1}) - \nabla f(\mathbf{x}_{t})) \|^{2} \\ & \leq 3 \mathbb{E}_{\xi_{t+1}} \| f(\mathbf{x}_{t+1}, \xi_{t+1}) - f(\mathbf{x}_{t}, \xi_{t+1}) \|^{2} + 3 \mathbb{E}_{\xi_{t+1}} \| \nabla f(\mathbf{x}_{t}) - \nabla f(\mathbf{x}_{t+1}) \|^{2} + 3 \mathbb{E}_{\xi_{t+1}} \| \alpha (f(\mathbf{x}_{t}, \xi_{t+1}) - \nabla f(\mathbf{x}_{t})) \|^{2} \\ & \leq 3 \mathbb{E}_{\xi_{t+1}} \| f(\mathbf{x}_{t+1} - \mathbf{x}_{t} \|^{2} + 3 L_{f}^{2} \| \mathbf{x}_{t} - \mathbf{x}_{t+1} \|^{2} + 3 \alpha^{2} \sigma^{2}, \end{split}$$

where in the first equality, we rearrange the terms, and in the first inequality, we use Young's inequality. Then in the second inequality, we use the Assumption 5.2,  $L_f$ -smoothness of  $f(\mathbf{x})$  and  $\mathbb{E}_{\xi} \|\nabla f(\mathbf{x}, \xi) - \nabla f(\mathbf{x})\|^2 \leq \sigma^2$ . We use this estimation in (C.2) and take total expectation to get the result. Proof of Lemma 5.4. We have, by smoothness of K:

$$K(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}, \mathbf{z}_t) \le K(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) + \langle \nabla_{\mathbf{x}} K(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle + \frac{L_K}{2} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2$$

We estimate the inner product here as

 $\langle \nabla_{\mathbf{x}} K(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle = \langle G(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t), \mathbf{x}_{t+1} - \mathbf{x}_k \rangle + \langle \nabla_{\mathbf{x}} K(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) - G(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle.$ We first have  $\overline{}$ 7

$$\nabla_{\mathbf{x}} K(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) - G(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) = \nabla f(\mathbf{x}_t) - \nabla f_t,$$

The definition of  $\mathbf{x}_{t+1}$  gives

$$\langle \mathbf{x}_{t+1} - \mathbf{x}_t + \tau G(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t), \mathbf{x}_t - \mathbf{x}_{t+1} \rangle \ge 0 \quad \Longleftrightarrow \quad \langle G(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t), \mathbf{x}_t - \mathbf{x}_{t+1} \rangle \ge \frac{1}{\tau} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2,$$

Using  $\langle \nabla_{\mathbf{x}} K(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) - G(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle \leq \frac{\tau}{2} \|\nabla f(\mathbf{x}_t) - \widehat{\nabla} f_t\| + \frac{1}{2\tau} \|\mathbf{x}_{t+1} - \mathbf{x}_t\|$ , we have

$$\langle \nabla_{\mathbf{x}} K(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t), \mathbf{x}_{t+1} - \mathbf{x}_t \rangle \leq \frac{\tau}{2} \| \nabla f(\mathbf{x}_t) - \widehat{\nabla} f_t \|^2 - \frac{1}{2\tau} \| \mathbf{x}_{t+1} - \mathbf{x}_t \|^2.$$

Then the result follows.

Proof of Lemma 5.5. First, from the definition we have

$$K(\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t) - K(\mathbf{x}_t, \mathbf{y}_{t+1}, \mathbf{z}_t) = -\eta \|A\mathbf{x}_t - \mathbf{b}\|^2$$

and also

$$\begin{split} &K(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}, \mathbf{z}_{t}) - K(\mathbf{x}_{t+1}, \mathbf{y}_{t+1}, \mathbf{z}_{t+1}) \\ &= \frac{\mu}{2} (\|\mathbf{x}_{t+1} - \mathbf{z}_{t}\|^{2} - \|\mathbf{x}_{t+1} - \mathbf{z}_{t+1}\|^{2}) \\ &= \frac{\mu}{2} \langle \mathbf{z}_{t+1} - \mathbf{z}_{t}, 2\mathbf{x}_{t+1} - \mathbf{z}_{t} - \mathbf{z}_{t+1} \rangle \\ &= \frac{\mu}{2} \langle \mathbf{z}_{t+1} - \mathbf{z}_{t}, 2\mathbf{x}_{t+1} - 2\mathbf{x}_{t} + 2\mathbf{x}_{t} - 2\mathbf{z}_{t} + \mathbf{z}_{t} - \mathbf{z}_{t+1} \rangle \\ &= \frac{\mu}{2} \langle \mathbf{z}_{t+1} - \mathbf{z}_{t}, 2\mathbf{x}_{t+1} - 2\mathbf{x}_{t} \rangle + \frac{\mu}{2} \langle \mathbf{z}_{t+1} - \mathbf{z}_{t}, 2\mathbf{x}_{t} - 2\mathbf{z}_{t} \rangle - \frac{\mu}{2} \|\mathbf{z}_{t+1} - \mathbf{z}_{t}\|^{2} \\ &\geq -\frac{\mu}{4} \|\mathbf{z}_{t+1} - \mathbf{z}_{t}\|^{2} - \mu \|\mathbf{x}_{t+1} - \mathbf{x}_{t}\|^{2} + \frac{\mu}{\beta} \|\mathbf{z}_{t} - \mathbf{z}_{t+1}\|^{2} - \frac{\mu}{2} \|\mathbf{z}_{t+1} - \mathbf{z}_{t}\|^{2}, \end{split}$$

where the first equality comes from the definition of K. In the last inequality, we use  $\langle a, b \rangle \ge -\frac{1}{4} \|a\|^2 - \|b\|^2$ and  $\mathbf{x}_t - \mathbf{z}_t = \frac{\mathbf{z}_{t+1} - \mathbf{z}_t}{\beta}$  by the definition of  $\mathbf{z}_{t+1}$  in Algorithm 3. Then combining the above two results with Lemma 5.4 yields the claim. 

Proof of Theorem 5.6. We denote that

$$V_t = K(\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t) - 2d(\mathbf{y}_t, \mathbf{z}_t) + 2\Psi(\mathbf{z}_t).$$
(C.3)

Joining (5.4) with Lemma A.5, we have

$$\mathbb{E}V_{t} - \mathbb{E}V_{t+1} \geq -\eta \mathbb{E} \|A\mathbf{x}_{t} - \mathbf{b}\|^{2} + \left(\frac{\mu}{\beta} - \frac{3\mu}{4}\right) \mathbb{E} \|\mathbf{z}_{t+1} - \mathbf{z}_{t}\|^{2}$$

$$- \frac{\tau}{2} \mathbb{E} \|\nabla f(\mathbf{x}_{t}) - \widehat{\nabla}f_{t}\|^{2} + \left(\frac{1}{2\tau} - \frac{L_{K}}{2} - \mu\right) \mathbb{E} \|\mathbf{x}_{t+1} - \mathbf{x}_{t}\|^{2}$$

$$+ 2\eta \mathbb{E} \langle A\mathbf{x}_{t} - \mathbf{b}, A\mathbf{x}^{*}(\mathbf{y}_{t+1}, \mathbf{z}_{t}) - \mathbf{b} \rangle + \mu \mathbb{E} \langle \mathbf{z}_{t+1} - \mathbf{z}_{t}, \mathbf{z}_{t+1} + \mathbf{z}_{t} - 2\mathbf{x}^{*}(\mathbf{y}_{t+1}, \mathbf{z}_{t+1}) \rangle$$

$$- 2\mu \mathbb{E} \langle \mathbf{z}_{t+1} - \mathbf{z}_{t}, \mathbf{z}_{t} - \bar{\mathbf{x}}^{*}(\mathbf{z}_{t}) \rangle - \frac{\mu}{\sigma_{4}} \mathbb{E} \|\mathbf{z}_{t} - \mathbf{z}_{t+1}\|^{2}. \qquad (C.4)$$

		1

First, let us combine the first and fifth terms on the right-hand side to obtain

$$-\eta \mathbb{E} \|A\mathbf{x}_t - \mathbf{b}\|^2 + 2\eta \langle A\mathbf{x}_t - \mathbf{b}, A\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{b} \rangle = -\eta \|A\mathbf{x}_t - A\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t)\|^2 + \eta \|A\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{b}\|^2.$$
(C.5)

Next, we combine the sixth and seventh terms on the right-hand side of (C.4)

$$\mu \langle \mathbf{z}_{t+1} - \mathbf{z}_t, \mathbf{z}_{t+1} + \mathbf{z}_t - 2\mathbf{x}^* (\mathbf{y}_{t+1}, \mathbf{z}_{t+1}) \rangle - 2\mu \langle \mathbf{z}_{t+1} - \mathbf{z}_t, \mathbf{z}_t - \bar{\mathbf{x}}^* (\mathbf{z}_t) \rangle$$

$$= \mu \langle \mathbf{z}_{t+1} - \mathbf{z}_t, \mathbf{z}_{t+1} - \mathbf{z}_t - 2\mathbf{x}^* (\mathbf{y}_{t+1}, \mathbf{z}_{t+1}) + 2\bar{\mathbf{x}}^* (\mathbf{z}_t) \rangle$$

$$= \mu \| \mathbf{z}_{t+1} - \mathbf{z}_t \|^2 + 2 \langle \mathbf{z}_{t+1} - \mathbf{z}_t, -\mathbf{x}^* (\mathbf{y}_{t+1}, \mathbf{z}_{t+1}) + \bar{\mathbf{x}}^* (\mathbf{z}_t) \rangle$$

$$(C.6)$$

We now single out the inner product in the last equality and estimate it by adding and subtracting  $\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t)$  in the second argument of the inner product:

$$2\langle \mathbf{z}_{t+1} - \mathbf{z}_{t}, -\mathbf{x}^{*}(\mathbf{y}_{t+1}, \mathbf{z}_{t+1}) + \bar{\mathbf{x}}^{*}(\mathbf{z}_{t}) \rangle = 2\mu \langle \mathbf{z}_{t+1} - \mathbf{z}_{t}, -\mathbf{x}^{*}(\mathbf{y}_{t+1}, \mathbf{z}_{t+1}) + \mathbf{x}^{*}(\mathbf{y}_{t+1}, \mathbf{z}_{t}) \rangle + 2\mu \langle \mathbf{z}_{t+1} - \mathbf{z}_{t}, -\mathbf{x}^{*}(\mathbf{y}_{t+1}, \mathbf{z}_{t}) + \bar{\mathbf{x}}^{*}(\mathbf{z}_{t}) \rangle \geq -\mu \|\mathbf{z}_{t+1} - \mathbf{z}_{t}\|^{2} - \mu \|\mathbf{x}^{*}(\mathbf{y}_{t+1}, \mathbf{z}_{t}) - \mathbf{x}^{*}(\mathbf{y}_{t+1}, \mathbf{z}_{t+1})\|^{2} - \frac{\mu}{\zeta} \|\mathbf{z}_{t+1} - \mathbf{z}_{t}\|^{2} - \mu \zeta \|\bar{\mathbf{x}}^{*}(\mathbf{z}_{t}) - \mathbf{x}^{*}(\mathbf{y}_{t+1}, \mathbf{z}_{t})\|,$$
(C.7)

for any  $\zeta$ , where we used Young's inequality twice. Then, we plug this into (C.6) to obtain

$$\mu \langle \mathbf{z}_{t+1} - \mathbf{z}_t, \mathbf{z}_{t+1} + \mathbf{z}_t - 2\mathbf{x}^* (\mathbf{y}_{t+1}, \mathbf{z}_{t+1}) \rangle - 2\mu \langle \mathbf{z}_{t+1} - \mathbf{z}_t, \mathbf{z}_t - \bar{\mathbf{x}}^* (\mathbf{z}_t) \rangle$$

$$\geq -\frac{\mu}{\sigma_4^2} \| \mathbf{z}_{t+1} - \mathbf{z}_t \|^2 - \frac{\mu}{\zeta} \| \mathbf{z}_{t+1} - \mathbf{z}_t \|^2 - \mu \zeta \| \bar{\mathbf{x}}^* (\mathbf{z}_t) - \mathbf{x}^* (\mathbf{y}_{t+1}, \mathbf{z}_t) \|^2,$$
(C.8)

where we use (A.33) to bound the second term on the right-hand side of (C.7), where  $\sigma_4$  is as defined in (A.1).

Then we use (C.5) and (C.8) in (C.4) to obtain

$$\begin{split} \mathbb{E}V_{t} - \mathbb{E}V_{t+1} &\geq (\frac{\mu}{\beta} - \frac{3\mu}{4}) \mathbb{E} \|\mathbf{z}_{t+1} - \mathbf{z}_{t}\|^{2} - \frac{\tau}{2} \mathbb{E} \|\nabla f(\mathbf{x}_{t}) - \widehat{\nabla}f_{t}\|^{2} + (\frac{1}{2\tau} - \frac{L_{K}}{2} - \mu) \mathbb{E} \|\mathbf{x}_{t+1} - \mathbf{x}_{t}\|^{2} \\ &- \eta \mathbb{E} \|A\mathbf{x}_{t} - A\mathbf{x}^{*}(\mathbf{y}_{t+1}, \mathbf{z}_{t})\|^{2} + \eta \mathbb{E} \|A\mathbf{x}^{*}(\mathbf{y}_{t+1}, \mathbf{z}_{t}) - \mathbf{b}\|^{2} \\ &- \frac{\mu}{\sigma_{4}^{2}} \mathbb{E} \|\mathbf{z}_{t+1} - \mathbf{z}_{t}\|^{2} - \frac{\mu}{\zeta} \mathbb{E} \|\mathbf{z}_{t+1} - \mathbf{z}_{t}\|^{2} - \mu \zeta \|\bar{\mathbf{x}}^{*}(\mathbf{z}_{t}) - \mathbf{x}^{*}(\mathbf{y}_{t+1}, \mathbf{z}_{t})\|^{2} - \frac{\mu}{\sigma_{4}} \mathbb{E} \|\mathbf{z}_{t+1} - \mathbf{z}_{t}\|^{2} \\ &\geq (\frac{\mu}{\beta} - \frac{3\mu}{4} - \frac{\mu}{\sigma_{4}^{2}} - \frac{\mu}{\zeta} - \frac{\mu}{\sigma_{4}}) \mathbb{E} \|\mathbf{z}_{t+1} - \mathbf{z}_{t}\|^{2} \\ &- \frac{\tau}{2} \mathbb{E} \|\nabla f(\mathbf{x}_{t}) - \widehat{\nabla}f_{t}\|^{2} - \frac{2\eta \|A\|^{2}}{(\mu - L_{f})^{2}} \mathbb{E} \|\nabla f(\mathbf{x}_{t}) - \widehat{\nabla}f_{t}\|^{2} \\ &+ (\frac{1}{2\tau} - \frac{L_{K}}{2} - \mu - \eta \|A\|^{2} \frac{2}{\tau^{2}(\mu - L_{f})^{2}}) \mathbb{E} \|\mathbf{x}_{t+1} - \mathbf{x}_{t}\|^{2} \\ &+ \eta \mathbb{E} \|A\mathbf{x}^{*}(\mathbf{y}_{t+1}, \mathbf{z}_{t}) - \mathbf{b}\|^{2} - \mu \zeta \overline{\sigma}^{2} \mathbb{E} \|A\mathbf{x}^{*}(\mathbf{y}_{t+1}, \mathbf{z}_{t}) - \mathbf{b}\| \end{split}$$

Where in the last inequality, we use the C.2 and A.9, then combine the same terms together.

Then we need to estimate the coefficients of each terms in the above inequality. From the choosen parameters, we easily have  $\sigma_4 = \frac{\mu - L_f}{\mu} > \frac{1}{2}$  and let  $\zeta = 6\beta$ 

We now estimate the coefficient of  $\mathbb{E} \| \mathbf{z}_t - \mathbf{z}_{t+1} \|^2$  in the last inequality. First, by  $\sigma_4 > \frac{1}{2}$ , we have  $\frac{\mu}{\sigma_4^2} \leq 4\mu$  and  $\frac{\mu}{\sigma_4} \leq 2\mu$ . By also using  $\zeta = 6\beta$ , we have

The coefficient of 
$$\mathbb{E} \|\mathbf{z}_t - \mathbf{z}_{t+1}\|^2 \ge \frac{\mu}{\beta} - \frac{3\mu}{4} - 4\mu - \frac{\mu}{6\beta} - 2\mu$$

Second, using  $\beta \leq 1/50$ , we obtain  $(\frac{3}{4} + 4 + 2)\mu \leq \frac{\mu}{5\beta}$ , then

The coefficient of 
$$\mathbb{E} \|\mathbf{z}_t - \mathbf{z}_{t+1}\|^2 \ge \frac{\mu}{\beta} - \frac{\mu}{5\beta} - \frac{\mu}{6\beta} \ge \frac{\mu}{\beta}.$$

We move on to estimating the coefficient of  $\mathbb{E} \|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2$ . With  $\eta \leq \frac{(\mu - L_f)^2 \tau}{8\|A\|^2}$ , we have  $2\eta \|A\|^2 \frac{1}{\tau^2(\mu - L_f)^2} \leq \frac{1}{4\tau}$ , we have

The coefficient of 
$$\mathbb{E} \|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 \ge \frac{1}{4\tau} - \frac{L_K}{2} - \mu$$

Last, we work on the coefficient of  $\mathbb{E} ||A\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{b}||^2$ . Because  $\zeta = 6\beta$ , it follows that  $\eta - \mu \zeta \bar{\sigma}^2 = \eta - 6\mu\beta\bar{\sigma}^2$ .

With  $\beta \leq \frac{\eta}{36\mu\bar{\sigma}^2}$ , we have  $6\mu\beta\bar{\sigma}^2 \leq \frac{\eta}{6}$ , then

The coefficient of 
$$\mathbb{E} \|A\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{b}\|^2 \ge \eta - \frac{\eta}{6} \ge \frac{\eta}{2}.$$

Next, we estimate the coefficient of  $\mathbb{E} \|\nabla f(\mathbf{x}_t) - \widehat{\nabla} f_t\|^2$ . With  $\eta \leq \frac{(\mu - L_f)^2 \tau}{8\|A\|^2}$ , we have  $-\frac{\tau}{2} - \frac{2\eta \|A\|^2}{(\mu - L_f)^2} \geq -\frac{3}{4}\tau$ . Finally, we have

$$\begin{split} & \mathbb{E}V_{t} - \mathbb{E}V_{t+1} \\ & \geq \frac{\mu}{2\beta} \mathbb{E} \|\mathbf{z}_{t} - \mathbf{z}_{t+1}\|^{2} + (\frac{1}{4\tau} - \frac{L_{K}}{2} - \mu) \mathbb{E} \|\mathbf{x}_{t} - \mathbf{x}_{t+1}\|^{2} + \frac{\eta}{2} \mathbb{E} \|A\mathbf{x}^{*}(\mathbf{y}_{t+1}, \mathbf{z}_{t}) - \mathbf{b}\|^{2} - \frac{3\tau}{4} \mathbb{E} \|\nabla f(\mathbf{x}_{t}) - \widehat{\nabla} f_{t}\|^{2} \\ & = \frac{\mu}{2\beta} \mathbb{E} \|\mathbf{z}_{t} - \mathbf{z}_{t+1}\|^{2} + (\frac{1}{4\tau} - \frac{L_{K}}{2} - \mu) \mathbb{E} \|\mathbf{x}_{t} - \mathbf{x}_{t+1}\|^{2} + \frac{\eta}{2} \mathbb{E} \|A\mathbf{x}^{*}(\mathbf{y}_{t+1}, \mathbf{z}_{t}) - \mathbf{b}\|^{2} \\ & \quad + \frac{\tau}{4} \mathbb{E} \|\nabla f(\mathbf{x}_{t}) - \widehat{\nabla} f_{t}\|^{2} - \tau \mathbb{E} \|\nabla f(\mathbf{x}_{t}) - \widehat{\nabla} f_{t}\|^{2} \end{split}$$

Then recalling Lemma 5.3 and assuming  $0 < \alpha \leq 1$ , we have

$$\mathbb{E}\|\widehat{\nabla}f_{t+1} - \nabla f(\mathbf{x}_{t+1})\|^2 \le (1-\alpha)\mathbb{E}\|\widehat{\nabla}f_t - \nabla f(\mathbf{x}_t)\|^2 + 3(L_0^2 + L_f^2)\mathbb{E}\|\mathbf{x}_{t+1} - \mathbf{x}_t\|^2 + 3\alpha^2\sigma^2 \tag{C.9}$$

We multiply (C.9) by  $\frac{\tau}{\alpha}$  and plug into (C.9), to get

$$\mathbb{E}V_{t} - \mathbb{E}V_{t+1} \geq \frac{\mu}{2\beta} \mathbb{E} \|\mathbf{z}_{t} - \mathbf{z}_{t+1}\|^{2} + (\frac{1}{4\tau} - \frac{L_{K}}{2} - \mu)\mathbb{E} \|\mathbf{x}_{t} - \mathbf{x}_{t+1}\|^{2} + \frac{\eta}{2} \mathbb{E} \|A\mathbf{x}^{*}(\mathbf{y}_{t+1}, \mathbf{z}_{t}) - \mathbf{b}\|^{2} + \frac{\tau}{4} \mathbb{E} \|\nabla f(\mathbf{x}_{t}) - \widehat{\nabla} f_{t}\|^{2} + \frac{\tau}{\alpha} \mathbb{E} \|\widehat{\nabla} f_{t+1} - \nabla f(\mathbf{x}_{t+1})\|^{2} - \frac{\tau}{\alpha} \mathbb{E} \|\widehat{\nabla} f_{t} - \nabla f(\mathbf{x}_{t})\|^{2} - \frac{3(L_{0}^{2} + L_{f}^{2})\tau}{\alpha} \mathbb{E} \|\mathbf{x}_{t} - \mathbf{x}_{t+1}\|^{2} - 3\alpha\sigma^{2}\tau.$$
(C.10)

Because  $\alpha = 48(L_0^2 + L_f^2)\tau^2$  and  $\tau \le \min\{\frac{1}{4L_K + 8\mu}, \frac{1}{\sqrt{48(L_0^2 + L_f^2)}}\}$ , we obtain

$$\mathbb{E}V_{t} - \mathbb{E}V_{t+1} \geq \frac{\mu}{2\beta} \mathbb{E}\|\mathbf{z}_{t} - \mathbf{z}_{t+1}\|^{2} + \frac{1}{8\tau} \mathbb{E}\|\mathbf{x}_{t} - \mathbf{x}_{t+1}\|^{2} + \frac{\eta}{2} \mathbb{E}\|A\mathbf{x}^{*}(\mathbf{y}_{t+1}, \mathbf{z}_{t}) - \mathbf{b}\|^{2} + \frac{\tau}{4} \mathbb{E}\|\nabla f(\mathbf{x}_{t}) - \widehat{\nabla}f_{t}\|^{2} \\ + \frac{1}{48(L_{0}^{2} + L_{f}^{2})\tau} \mathbb{E}\|\widehat{\nabla}f_{t+1} - \nabla f(\mathbf{x}_{t+1})\|^{2} - \frac{1}{48(L_{0}^{2} + L_{f}^{2})\tau} \mathbb{E}\|\widehat{\nabla}f_{t} - \nabla f(\mathbf{x}_{t})\|^{2} - 144(L_{0}^{2} + L_{f}^{2})\sigma^{2}\tau^{3}$$

Finally, we move  $\frac{1}{48(L_0^2+L_f^2)\tau} \mathbb{E} \|\widehat{\nabla}f_{t+1} - \nabla f(\mathbf{x}_{t+1})\|^2 - \frac{1}{48(L_0^2+L_f^2)\tau} \mathbb{E} \|\widehat{\nabla}f_t - \nabla f(\mathbf{x}_t)\|^2$  to the left-hand side of the above inequality and use the definition of  $\overline{V}_t$  in (5.2) to get the desired result.

Proof of Theorem 5.7. Because  $\mathbf{z}_{t+1} - \mathbf{z}_t = \beta(\mathbf{x}_t - \mathbf{z}_t), \frac{\mu\beta}{2} = \Theta(\tau)$  and  $\frac{\eta}{2} = \Theta(\tau)$ , hence there exists a constant C such that

$$\mathbb{E}\bar{V}_{t} - \mathbb{E}\bar{V}_{t+1} \ge C\tau\{\mathbb{E}\|\mathbf{x}_{t} - \mathbf{z}_{t}\|^{2} + \mathbb{E}\|\tau^{-1}(\mathbf{x}_{t} - \mathbf{x}_{t+1})\|^{2} + \mathbb{E}\|A\mathbf{x}^{*}(\mathbf{y}_{t+1}, \mathbf{z}_{t}) - \mathbf{b}\|^{2} + \mathbb{E}\|\nabla f(\mathbf{x}_{t}) - \widehat{\nabla}f_{t}\|^{2}\} - 144(L_{0}^{2} + L_{f}^{2})\sigma^{2}\tau^{3}.$$
(C.11)

Then, summing up (C.11) over  $t = 0, 1, \ldots, T - 1$ , we have

$$\mathbb{E}\bar{V}_{0} - \mathbb{E}\bar{V}_{T} \geq \sum_{t=0}^{T-1} C\tau \{\mathbb{E}\|\mathbf{x}_{t} - \mathbf{z}_{t}\|^{2} + \mathbb{E}\|\tau^{-1}(\mathbf{x}_{t} - \mathbf{x}_{t+1})\|^{2} + \mathbb{E}\|A\mathbf{x}^{*}(\mathbf{y}_{t+1}, \mathbf{z}_{t}) - \mathbf{b}\|^{2} + \mathbb{E}\|\nabla f(\mathbf{x}_{t}) - \widehat{\nabla}f_{t}\|^{2} \} - 144(L_{0}^{2} + L_{f}^{2})\sigma^{2}\tau^{3}T.$$
(C.12)

Form the definition, we have  $K(\mathbf{x}, \mathbf{y}, \mathbf{z}) \ge d(\mathbf{y}, \mathbf{z})$  and  $\Psi(\mathbf{z}) \ge d(\mathbf{y}, \mathbf{z})$  (see also Lemma A.13), then

$$V_t = K(\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t) - 2d(\mathbf{y}_t, \mathbf{z}_t) + 2\Psi(\mathbf{z}_t) \ge \Psi(\mathbf{z}_t) \ge \underline{f}$$

Then, we have

$$\bar{V}_t = K(\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t) - 2d(\mathbf{y}_t, \mathbf{z}_t) + 2\Psi(\mathbf{z}_t) + \frac{1}{48(L_0^2 + L_f^2)\tau} \mathbb{E}\|\widehat{\nabla}f_t - \nabla f(\mathbf{x}_t)\|^2 \ge \underline{f}.$$
(C.13)

Let  $\tau = T^{-1/3}$  and use mini-batch in the initial step where we will have  $\mathbb{E}\|\widehat{\nabla}f_0 - \nabla f(\mathbf{x}_0)\|^2 \leq T^{-1/3}\sigma^2$ , then

$$\bar{V}_{0} = K(\mathbf{x}_{0}, \mathbf{y}_{0}, \mathbf{z}_{0}) - 2d(\mathbf{y}_{0}, \mathbf{z}_{0}) + 2\Psi(\mathbf{z}_{0}) + \frac{1}{48(L_{0}^{2} + L_{f}^{2})\tau} \mathbb{E}\|\widehat{\nabla}f_{0} - \nabla f(\mathbf{x}_{0})\|^{2} \\
\leq K(\mathbf{x}_{0}, \mathbf{y}_{0}, \mathbf{z}_{0}) - 2d(\mathbf{y}_{0}, \mathbf{z}_{0}) + 2\Psi(\mathbf{z}_{0}) + \frac{\sigma^{2}}{48(L_{0}^{2} + L_{f}^{2})}$$
(C.14)

where the right-hand is a constant independent of T, we denote it as  $C_0$ .

Combining (C.12) with (C.13) and (C.14), we have

$$\frac{1}{T} \sum_{t=0}^{T-1} C\{\mathbb{E} \|\mathbf{x}_t - \mathbf{z}_t\|^2 + \mathbb{E} \|\tau^{-1}(\mathbf{x}_t - \mathbf{x}_{t+1})\|^2 + \mathbb{E} \|A\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{b}\|^2 + \mathbb{E} \|\nabla f(\mathbf{x}_t) - \widehat{\nabla} f_t\|^2 \}$$

$$\leq T^{-2/3} (C_0 - \underline{f} + 144(L_0^2 + L_f^2)\sigma^2).$$
(C.15)

Then, for index s selected uniformly at random from  $\{0, 1, ..., T-1\}$ , we have

$$\mathbb{E} \|\mathbf{x}_{s} - \mathbf{z}_{s}\|^{2} = O(T^{-2/3}), \quad \mathbb{E} \|\tau^{-1}(\mathbf{x}_{s} - \mathbf{x}_{s+1})\|^{2} = O(T^{-2/3}), \\ \mathbb{E} \|A\mathbf{x}^{*}(\mathbf{y}_{s+1}, \mathbf{z}_{s}) - \mathbf{b}\|^{2} = O(T^{-2/3}), \quad \mathbb{E} \|\nabla f(\mathbf{x}_{t}) - \widehat{\nabla} f_{t}\|^{2} = O(T^{-2/3}).$$
(C.16)

According to Algorithm 3, we have

$$\mathbf{x}_{s+1} = \arg\min_{\mathbf{x}} \left\{ \langle G(\mathbf{x}_s, \mathbf{y}_{s+1}, \mathbf{z}_s), \mathbf{x} - \mathbf{x}^s \rangle + \frac{1}{\tau} \|\mathbf{x} - \mathbf{x}_s\|^2 + \partial I_X(\mathbf{x}) \right\}$$

By the definition of  $\mathbf{x}_{s+1}$ , we have

$$0 \in G(\mathbf{x}_s, \mathbf{y}_{s+1}, \mathbf{z}_s) + \frac{2}{\tau} (\mathbf{x}_{s+1} - \mathbf{x}_s) + \partial I_X(\mathbf{x}_{s+1}).$$
(C.17)

We now set

$$\mathbf{v} = \nabla_{\mathbf{x}} K(\mathbf{x}_{s+1}, \mathbf{y}_{s+1}, \mathbf{z}_s) - G(\mathbf{x}_s, \mathbf{y}_{s+1}, \mathbf{z}_s) - \frac{2}{\tau} (\mathbf{x}_{s+1} - \mathbf{x}_s) - \rho A^{\top} (A\mathbf{x}_{s+1} - \mathbf{b}) - \mu (\mathbf{x}_{s+1} - \mathbf{z}_s).$$

Now, by using the definition of  $K(\mathbf{x}, \mathbf{y}, \mathbf{z})$  from (5.3) and (C.17), we obtain

$$\mathbf{v} \in \nabla f(\mathbf{x}_{s+1}) + A^{\top} \mathbf{y}_{s+1} + \partial I_X(\mathbf{x}_{s+1})$$

We now derive the guarantees on the feasibility and the norm of  $\mathbf{v}$ .

First, by triangle inequality, we have

$$\begin{aligned} \|A\mathbf{x}_{s+1} - \mathbf{b}\| &\leq \|A\mathbf{x}^{*}(\mathbf{y}_{s+1}, \mathbf{z}_{s}) - \mathbf{b}\| + \|A\mathbf{x}_{s+1} - A\mathbf{x}_{s}\| + \|A(\mathbf{x}_{s} - \mathbf{x}^{*}(\mathbf{y}_{s+1}, \mathbf{z}_{s}))\| \\ &\leq \|A\mathbf{x}^{*}(\mathbf{y}_{s+1}, \mathbf{z}_{s}) - \mathbf{b}\| + \|A\| \|\mathbf{x}_{s+1} - \mathbf{x}_{s}\| + \frac{\|A\|}{\tau(\mu - L_{f})} \|\mathbf{x}_{s} - \mathbf{x}_{s+1}\| + \frac{\|A\|}{\mu - L_{f}} \|\widehat{\nabla}f_{s} - \nabla f(\mathbf{x}_{s})\| \\ &= O(T^{-1/3}), \end{aligned}$$
(C.18)

where in the second inequality, we use Lemma C.2.

Then we have

$$\begin{aligned} \|\mathbf{v}\| &\leq \|\nabla_{\mathbf{x}} K(\mathbf{x}_{s+1}, \mathbf{y}_{s+1}, \mathbf{z}_{s}) - \nabla_{\mathbf{x}} K(\mathbf{x}_{s}, \mathbf{y}_{s+1}, \mathbf{z}_{s})\| + \|\nabla_{\mathbf{x}} K(\mathbf{x}_{s}, \mathbf{y}_{s+1}, \mathbf{z}_{s}) - G(\mathbf{x}_{s}, \mathbf{y}_{s+1}, \mathbf{z}_{s})\| \\ &+ \frac{2}{\tau} \|\mathbf{x}_{s+1} - \mathbf{x}_{s}\| + \rho \|A\| \|A\mathbf{x}_{s+1} - \mathbf{b}\| + \mu \|\mathbf{x}_{s+1} - \mathbf{z}_{s}\| \\ &\leq \left(L_{K} + \frac{2}{\tau}\right) \|\mathbf{x}_{s+1} - \mathbf{x}_{s}\| + \|\nabla f(\mathbf{x}_{s}) - \widehat{\nabla} f_{s}\| + \rho \|A\| \|A\mathbf{x}_{s+1} - \mathbf{b}\| + \mu (\|\mathbf{x}_{s} - \mathbf{z}_{s}\| + \|\mathbf{x}_{s+1} - \mathbf{x}_{s}\|) \\ &= O(T^{-1/3}), \end{aligned}$$

where in first inequality, we introduce a term  $\nabla_{\mathbf{x}} K(\mathbf{x}_s, \mathbf{y}_{s+1}, \mathbf{z}_s)$  and then use triangle inequality. The second inequality used Lipschitzness of K, the definition of G, and the triangle inequality. The last step uses (C.16) and (C.18).

Proof of Remark 5.8. This is because for the proof for Section 4, we only need to obtain a result similar to Lemma 5.6 when  $\|\mathbf{y}\| < M_y$  (where the latter result will be shown in the same way as Proposition B.3).

First, there is a difference in Lemma 5.4 that analyzes the progress of  $K(\mathbf{x}_t, \mathbf{y}_t, \mathbf{z}_t)$ . In particular, we now have

$$\begin{aligned} \langle \nabla_{\mathbf{x}} K(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t}) - G(\mathbf{x}_{t}, \mathbf{y}_{t+1}, \mathbf{z}_{t}), \mathbf{x}_{t+1} - \mathbf{x}_{t} \rangle \\ &= \langle \nabla f(\mathbf{x}_{t}) + A^{\top} \mathbf{y}_{t+1} + A^{\top} (A \mathbf{x}_{t} - \mathbf{b}) - \widehat{\nabla} f_{t} - A^{\top}_{\zeta_{t}^{\top}} \mathbf{y}_{t+1} + A^{\top}_{\zeta_{t}^{\top}} (A_{\zeta_{t}^{2}} \mathbf{x}_{t} - \mathbf{b}_{\zeta_{t}^{2}}), \mathbf{x}_{t+1} - \mathbf{x}_{t} \rangle \\ &\leq \frac{\tau}{2} \| (\nabla f(\mathbf{x}_{t}) - \widehat{\nabla} f_{t} + (A^{\top} \mathbf{y}_{t+1} - A^{\top}_{\zeta_{t}^{\top}} \mathbf{y}_{t+1}) + (A^{\top} (A \mathbf{x}_{t} - \mathbf{b}) - A^{\top}_{\zeta_{t}^{\top}} (A_{\zeta_{t}^{2}} \mathbf{x}_{t} - \mathbf{b}_{\zeta_{t}^{2}})) \|^{2} + \frac{1}{2\tau} \| \mathbf{x}_{t+1} - \mathbf{x}_{t} \|^{2} \\ &\leq \frac{3\tau}{2} (\| \nabla f(\mathbf{x}_{t}) - \widehat{\nabla} f_{t} \|^{2} + \| A^{\top} \mathbf{y}_{t+1} - A^{\top}_{\zeta_{t}^{1}} \mathbf{y}_{t+1} \|^{2} + \| A^{\top} (A \mathbf{x}_{t} - \mathbf{b}) - A^{\top}_{\zeta_{t}^{1}} (A_{\zeta_{t}^{2}} \mathbf{x}_{t} - \mathbf{b}_{\zeta_{t}^{2}}) \|^{2}) + \frac{1}{2\tau} \| \mathbf{x}_{t+1} - \mathbf{x}_{t} \|^{2} \end{aligned}$$

The only difference is that we have more variance-type error terms. We use the same STORM technique to update the stochastic gradient  $A_{\zeta_t}^{\top} \mathbf{y}_{t+1}$  and  $A_{\zeta_t}^{\top} (A_{\zeta_t}^2 \mathbf{x}_t - \mathbf{b}_{\zeta_t}^2)$ . Then, under Assumption 4.2, we will have a similar result that the first three terms could be bounded by similar recurrence relations to C.9 and, as in (C.10), the order of  $\tau$  will be 3 since  $\alpha \approx \tau^2$ .

Next, there is the second difference in Lemma A.5 for  $d(\mathbf{y}_t, \mathbf{z}_t)$ . In particular, we have

$$\begin{aligned} d(\mathbf{y}_{t+1}, \mathbf{z}_{t+1}) &- d(\mathbf{y}_t, \mathbf{z}_t) \\ &\geq \langle \mathbf{y}_{t+1} - \mathbf{y}_t, A\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{b} \rangle + \frac{\mu}{2} \langle \mathbf{z}_{t+1} - \mathbf{z}_t, \mathbf{z}_{t+1} + \mathbf{z}_t - 2\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_{t+1}) \rangle \\ &= \langle \eta(A_{\zeta_t}\mathbf{x}_t - b_{\zeta_t}) - \eta(A\mathbf{x}_t - b), A\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{b} \rangle + \langle \eta(A\mathbf{x}_t - b), A\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{b} \rangle \\ &+ \frac{\mu}{2} \langle \mathbf{z}_{t+1} - \mathbf{z}_t, \mathbf{z}_{t+1} + \mathbf{z}_t - 2\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_{t+1}) \rangle \\ &\geq -\eta \| (A_{\zeta_t}\mathbf{x}_t - b_{\zeta_t}) - (A\mathbf{x}_t - b) \|^2 - \frac{\eta}{4} \| A\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{b} \|^2 \\ &+ \langle \eta(A\mathbf{x}_t - b), A\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_t) - \mathbf{b} \rangle + \frac{\mu}{2} \langle \mathbf{z}_{t+1} - \mathbf{z}_t, \mathbf{z}_{t+1} + \mathbf{z}_t - 2\mathbf{x}^*(\mathbf{y}_{t+1}, \mathbf{z}_{t+1}) \rangle \end{aligned}$$

Note that the only difference between the above estimate and Lemma A.5 are the first two (nonpositive) error terms on the right-hand side. We also use the STORM technique to update the stochastic gradient  $A_{\zeta_t} \mathbf{x}_t - b_{\zeta_t}$ , then the first error term will be bounded with a similar recurrence relation to (C.9), and the error, as before, will be a constant term in the order of  $\tau^3$ . In addition the second error term will be canceled by the third term on the right-hand side of (5.6).

Hence we will have an inequality for the change of  $\mathbb{E}\overline{V}_t$  to  $\mathbb{E}\overline{V}_{t+1}$ , similar to (5.6). In particular, the main error term is of the order  $\tau^3$ . Then we will obtain the same  $O(\varepsilon^{-3})$  complexity result by arguing the same way as in Theorem 5.7.