

Online Alternating Direction Method

Huahua Wang

Dept of Computer Science & Engg
University of Minnesota, Twin Cities
huwang@cs.umn.edu

Arindam Banerjee

Dept of Computer Science & Engg
University of Minnesota, Twin Cities
banerjee@cs.umn.edu

Abstract

Online optimization has emerged as powerful tool in large scale optimization. In this paper, we introduce efficient online optimization algorithms based on the alternating direction method (ADM), which can solve online convex optimization under linear constraints where the objective could be non-smooth. We introduce new proof techniques for ADM in the batch setting, which yields a $O(1/T)$ convergence rate for ADM and forms the basis for regret analysis in the online setting. We consider two scenarios in the online setting, based on whether an additional Bregman divergence is needed or not. In both settings, we establish regret bounds for both the objective function as well as constraints violation for general and strongly convex functions. We also consider inexact ADM updates where certain terms are linearized to yield efficient updates and show the stochastic convergence rates. In addition, we briefly discuss that online ADM can be used as projection-free online learning algorithm in some scenarios. Preliminary results are presented to illustrate the performance of the proposed algorithms.

1 Introduction

In recent years, online optimization [16, 74, 40] and its batch counterpart stochastic gradient descent [61, 44] has contributed substantially to advances in large scale optimization techniques for machine learning. Online convex optimization has been generalized to handle time-varying and non-smooth convex functions [26, 27, 69]. Distributed optimization, where the problem is divided into parts on which progress can be made in parallel, has also contributed to advances in large scale optimization [11, 9, 14].

Important advances have been made based on the above ideas in the recent literature. Composite objective mirror descent (COMID) [26] generalizes mirror descent [6] to the online setting. COMID also includes certain other proximal splitting methods such as FOBOS [27] as special cases. Regularized dual averaging (RDA) [69] generalizes dual averaging [57] to online and composite optimization, and can be used for distributed optimization [24]. The three methods consider the following composite objective optimization [56]:

$$\min_{\mathbf{x} \in \mathcal{X}} \sum_{t=1}^T (f_t(\mathbf{x}) + g(\mathbf{x})) , \quad (1)$$

where the functions f_t, g are convex functions and \mathcal{X} is a convex set. Solving (1) usually involves the projection onto \mathcal{X} . In some cases, e.g., when g is the ℓ_1 norm or \mathcal{X} is the unit simplex, the projection can be done efficiently. In general, the full projection requires an inner loop algorithm, leading to a double loop algorithm for solving (1) [41].

In this paper, we propose single loop online optimization algorithms for composite objective optimization subject to linear constraints. In particular, we consider optimization problems of the following form:

$$\min_{\mathbf{x} \in \mathcal{X}, \mathbf{z} \in \mathcal{Z}} \sum_{t=1}^T (f_t(\mathbf{x}) + g(\mathbf{z})) \quad \text{s.t.} \quad \mathbf{Ax} + \mathbf{Bz} = \mathbf{c}, \quad (2)$$

where $\mathbf{A} \in \mathbb{R}^{m \times n_1}$, $\mathbf{B} \in \mathbb{R}^{m \times n_2}$, $\mathbf{c} \in \mathbb{R}^m$, $\mathbf{x} \in \mathcal{X} \in \mathbb{R}^{n_1 \times 1}$, $\mathbf{z} \in \mathcal{Z} \in \mathbb{R}^{n_2 \times 1}$ and \mathcal{X} and \mathcal{Z} are convex sets. The linear equality constraint introduces splitting variables and thus splits functions and feasible sets into simpler constraint sets $\mathbf{x} \in \mathcal{X}$ and $\mathbf{z} \in \mathcal{Z}$. (2) can easily accommodate linear inequality constraints by introducing a slack variable, which will be discussed in Section 6.4. In the sequel, we drop the convex sets \mathcal{X} and \mathcal{Z} for ease of exposition, noting that one can consider g and other additive functions to be the indicators of suitable convex feasible sets. f_t and g can be non-smooth, including piecewise linear and indicator functions. In the context of machine learning, f_t is usually a loss function such as ℓ_1 , ℓ_2 , hinge and logistic loss, while g is a regularizer, e.g., ℓ_1 , ℓ_2 , nuclear norm, mixed-norm and total variation.

In the batch setting, where $f_t = f$, (2) can be solved by the well known alternating direction method of multipliers (ADMM or ADM) [11]. First introduced in [35], ADM has since been extensively explored in recent years due to its ease of applicability and empirical performance in a wide variety of problems, including composite objectives [11, 29, 45]. It has been shown as a special case of Douglas-Rachford splitting method [18, 23, 29], which in turn is a special case of the proximal point method [63]. Recent literature has illustrated the empirical efficiency of ADM in a broad spectrum of applications ranging from image processing [58, 31, 1, 17] to applied statistics and machine learning [65, 1, 72, 73, 70, 45, 5, 51, 49]. ADM has been shown to outperform state-of-the-art methods for sparse problems, including LASSO [67, 39, 1, 11], total variation [36], sparse inverse covariance selection [21, 4, 32, 50, 65, 72], and sparse and low rank approximations [73, 45, 13]. ADM have also been used to solve linear programs (LPs) [28], LP decoding [5] and MAP inference problems in graphical models [49, 51, 33]. In addition, an advantage of ADM is that it can handle linear equality constraint of the form $\{\mathbf{x}, \mathbf{z} | \mathbf{Ax} + \mathbf{Bz} = \mathbf{c}\}$, which makes distributed optimization by variable splitting in a batch setting straightforward [9, 52, 11, 60, 66]. For further understanding of ADM, we refer the readers to the comprehensive review by [11] and references therein.

Although the proof of global convergence of ADM can be found in [34, 29, 11], the literature does not have the convergence rate for ADM¹ or even the convergence rate for the objective, which is fundamentally important to regret analysis in the online setting. We introduce new proof techniques for the rate of convergence of ADM in the batch setting, which establish a $O(1/T)$ convergence rate for the objective, the optimality conditions (constraints) and ADM based on variational inequalities [30]. The $O(1/T)$ convergence rate for ADM is in line with gradient methods for composite objective [55, 56, 27]². Our proof requires rather weak assumptions compared to the Lipschitz continuous gradient required in general in gradient methods [55, 56, 27]. Further, the convergence analysis for the batch setting forms the basis of regret analysis in the online setting.

In an online or stochastic gradient descent setting, where f_t is a time-varying function, (2) amounts to solving a sequence of equality-constrained subproblems, which in general leads to a double-loop algorithm where the inner loop ADM iterations have to be run till convergence after every new data point or function is revealed. As a result, ADM has not yet been generalized to the online setting.

¹During/after the publication of our preliminary version [68], the convergence rate for ADM was shown in [43, 42, 46, 22, 10, 37], but our proof is different and self-contained. In particular, the other approaches do not prove the convergence rate for the objective, which is the key for the regret analysis in the online setting or stochastic setting.

²The gradient methods can be accelerated to achieve the $O(1/T^2)$ convergence rate [55, 56].

Problem	$\min_{\mathbf{Ax}+\mathbf{Bz}=\mathbf{c}} \sum_t f_t(\mathbf{x}) + g(\mathbf{z})$		$\min_{\mathbf{x}} \sum_t f_t(\mathbf{x}) + g(\mathbf{x})$
Methods	OADM		OGD, FOBOS, COMID, RDA
Regret Bounds	Objective	constraint	Objective
General Convex	$O(\sqrt{T})$	$O(\sqrt{T})$	$O(\sqrt{T})$
Strongly Convex	$O(\log(T))$	$O(\log(T))$	$O(\log(T))$

Table 1: Main results for regret bounds of OADM in solving linearly-constrained composite objective optimization, in comparison with OGD, FOBOS, COMID and RDA in solving composite objective optimization. In both general and strongly convex cases, OADM achieves the optimal regret bounds for the objective, matching the results of the state-of-the-art methods. In addition, OADM also achieves the optimal regret bounds for constraint violation, showing the equality constraint will be satisfied on average in the long.

We consider two scenarios in the online setting, based on whether an additional Bregman divergence is needed or not for a proximal function in each step. We propose efficient online ADM (OADM) algorithms for both scenarios which make a single pass through the update equations and avoid a double loop algorithm. In the online setting, while a single pass through the ADM update equations is not guaranteed to satisfy the linear constraint $\mathbf{Ax} + \mathbf{Bz} = \mathbf{c}$ in each iteration, we consider two types of regret: regret in the *objective* as well as regret in *constraint violation*. We establish both types of regret bounds for general and strongly convex functions. In Table 1, we summarize the main results of OADM and also compare with OGD [74], FOBOS [27], COMID [26] and RDA [69]. While OADM aims to solve linearly-constrained composite objective optimization problems, OGD, FOBOS and RDA are for such problems without explicit constraints. In both general and strongly convex cases, our methods achieve the optimal regret bounds for the objective as well as the constraint violation, while start-of-the-art methods achieve the optimal regret bounds for the objective. We also present preliminary experimental results illustrating the performance of the proposed OADM algorithms in comparison with FOBOS and RDA [27, 69].

The key advantage of the OADM algorithms can be summarized as follows: Like COMID and RDA, OADM can solve online composite optimization problems, matching the regret bounds for existing methods. The ability to additionally handle linear equality constraint of the form $\mathbf{Ax} + \mathbf{Bz} = \mathbf{c}$ makes non-trivial variable splitting possible yielding efficient distributed online optimization algorithms [20] and projection-free online learning [41] based on OADM. Further, the notion of regret in both the objective as well as constraint may contribute towards development of suitable analysis tools for online constrained optimization problems [48, 47].

We summarize our main contributions as follows:

- We establish a $O(1/T)$ convergence rate for the objective, optimality conditions (constraints) and variational inequality for ADM.
- We propose online ADM (OADM), which is the first single loop online algorithm to explicitly solve the linearly-constrained problem (2) by just doing a single pass over examples.
- In OADM, we establish the optimal regret bounds for both objective and constraint violation for general as well as strongly convex functions. The introduction of regret for constraint violation which allows constraints to be violated at each round but guarantees constraints to be satisfied on average in the long run.
- We show some inexact updates in the OADM through the use of an additional Bregman divergence,

including OGD and COMID as special cases. For OADM with inexact updates, we also show the stochastic convergence rates.

- For an intersection of simple constraints, e.g., linear constraint (simplex), OADM is a projection-free online learning algorithm achieving the optimal regret bounds for both general and strongly convex functions.

The rest of the paper is organized as follows. In Section 2, we analyze batch ADM and establish its convergence rate. In Section 3, we propose OADM to solve the online optimization problem with linear constraints. In Sections 4 and 5, we present the regret analysis in two different scenarios based on whether an additional Bregman divergence is added or not. In Section 6, we discuss inexact ADM updates and show the stochastic convergence rates, show the connection to related works and projection-free online learning based on OADM. We present preliminary experimental results in Section 7, and conclude in Section 8.

2 Analysis for Batch Alternating Direction Method

We consider the batch ADM problem where f_t is fixed in (2), i.e.,

$$\min_{\mathbf{x}, \mathbf{z}} f(\mathbf{x}) + g(\mathbf{z}) \quad \text{s.t.} \quad \mathbf{Ax} + \mathbf{Bz} = \mathbf{c}. \quad (3)$$

The Lagrangian [12, 8] for the equality-constrained optimization problem (3) is

$$L(\mathbf{x}, \mathbf{y}, \mathbf{z}) = f(\mathbf{x}) + g(\mathbf{z}) + \langle \mathbf{y}, \mathbf{Ax} + \mathbf{Bz} - \mathbf{c} \rangle, \quad (4)$$

where \mathbf{x}, \mathbf{z} are the primal variables and \mathbf{y} is the dual variable. To penalize the violation of equality constraint, augmented Lagrangian methods use an additional quadratic penalty term. In particular, the augmented Lagrangian [8] for (2) is

$$L_\rho(\mathbf{x}, \mathbf{y}, \mathbf{z}) = f(\mathbf{x}) + g(\mathbf{z}) + \langle \mathbf{y}, \mathbf{Ax} + \mathbf{Bz} - \mathbf{c} \rangle + \frac{\rho}{2} \|\mathbf{Ax} + \mathbf{Bz} - \mathbf{c}\|^2, \quad (5)$$

where $\rho > 0$ is a penalty parameter. Batch ADM updates the three variables $(\mathbf{x}, \mathbf{z}, \mathbf{y})$ by alternately minimizing the augmented Lagrangian. It executes the following three steps iteratively till convergence [11]:

$$\mathbf{x}_{t+1} = \underset{\mathbf{x}}{\operatorname{argmin}} \left\{ f(\mathbf{x}) + \langle \mathbf{y}_t, \mathbf{Ax} + \mathbf{Bz}_t - \mathbf{c} \rangle + \frac{\rho}{2} \|\mathbf{Ax} + \mathbf{Bz}_t - \mathbf{c}\|^2 \right\}, \quad (6)$$

$$\mathbf{z}_{t+1} = \underset{\mathbf{z}}{\operatorname{argmin}} \left\{ g(\mathbf{z}) + \langle \mathbf{y}_t, \mathbf{Ax}_{t+1} + \mathbf{Bz} - \mathbf{c} \rangle + \frac{\rho}{2} \|\mathbf{Ax}_{t+1} + \mathbf{Bz} - \mathbf{c}\|^2 \right\}, \quad (7)$$

$$\mathbf{y}_{t+1} = \mathbf{y}_t + \rho(\mathbf{Ax}_{t+1} + \mathbf{Bz}_{t+1} - \mathbf{c}). \quad (8)$$

At step $(t + 1)$, the equality constraint in (3) is not necessarily satisfied in ADM. However, one can show that the equality constraint is satisfied in the long run so that $\lim_{t \rightarrow \infty} \mathbf{Ax}_t + \mathbf{Bz}_t - \mathbf{c} \rightarrow \mathbf{0}$.

While global convergence of ADMM has been established under appropriate conditions, we are interested in the rate of convergence of ADM in terms of iteration complexity, i.e., the number of iterations needed to obtain an ϵ -optimal solution. Most first-order methods require functions to be smooth or having Lipschitz continuous gradient to establish the convergence rate [55, 56, 27]. The assumptions in establishing convergence rate of ADM are relatively simple [11], and are stated below for the sake of completeness:

Assumption 1

- (a) $f : \mathbb{R}^{n_1} \rightarrow \mathbb{R} \cup \{+\infty\}$ and $g : \mathbb{R}^{n_2} \rightarrow \mathbb{R} \cup \{+\infty\}$ are closed, proper and convex.
- (b) An optimal solution to (3) exists. Let $\{\mathbf{x}^*, \mathbf{z}^*, \mathbf{y}^*\}$ be an optimal solution. Denote $\|\mathbf{y}^*\|_2 = D_{\mathbf{y}}, \|\mathbf{z}^*\|_2 = D_{\mathbf{z}}$.
- (c) Without loss of generality, $\mathbf{z}_0 = 0, \mathbf{y}_0 = 0$. Let $\lambda_{\max}^{\mathbf{B}}$ be the largest eigenvalue of $\mathbf{B}^T \mathbf{B}$.

We first analyze the convergence rate for the objective and optimality conditions (constraints) separately using new proof techniques, which play an important role for the regret analysis in the online setting. Then, a joint analysis of the objective and constraints using a variational inequality [30] establishes the $O(1/T)$ convergence rate for ADM.

2.1 Convergence Rate for the Objective

The updates of \mathbf{x}, \mathbf{z} implicitly generate the (sub)gradients of $f(\mathbf{x}_{t+1})$ and $g(\mathbf{z}_{t+1})$, as given in the following lemma.

Lemma 1 *Let $\partial f(\mathbf{x}_{t+1})$ be the subgradient of $f(\mathbf{x})$ at \mathbf{x}_{t+1} , we have*

$$-\mathbf{A}^T(\mathbf{y}_t + \rho(\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_t - \mathbf{c})) \in \partial f(\mathbf{x}_{t+1}), \quad (9)$$

$$-\mathbf{A}^T(\mathbf{y}_{t+1} + \rho(\mathbf{B}\mathbf{z}_t - \mathbf{B}\mathbf{z}_{t+1})) \in \partial f(\mathbf{x}_{t+1}) \quad (10)$$

Let $\partial g(\mathbf{z}_{t+1})$ be the subgradient of $g(\mathbf{z})$ at \mathbf{z}_{t+1} , we have

$$-\mathbf{B}^T \mathbf{y}_{t+1} \in \partial g(\mathbf{z}_{t+1}). \quad (11)$$

Proof: Since \mathbf{x}_{t+1} minimizes (6), we have

$$0 \in \partial f(\mathbf{x}_{t+1}) + \mathbf{A}^T \mathbf{y}_t + \rho \mathbf{A}^T (\mathbf{A}\mathbf{x}_{t+1} - \mathbf{B}\mathbf{z}_t - \mathbf{c}).$$

Rearranging the terms gives (9). using (8) yield (10).

Similarly, \mathbf{z}_{t+1} minimizes (7), then

$$\partial g(\mathbf{z}_{t+1}) + \mathbf{B}^T \mathbf{y}_t + \rho \mathbf{B}^T (\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_{t+1} - \mathbf{c}) \in 0.$$

Rearranging the terms and using (8) yield (11). ■

The following lemma shows the inaccuracy of the objective with respect to the optimum at $(t+1)$ is bounded by step differences of \mathbf{y} and \mathbf{z} .

Lemma 2 *Let the sequences $\{\mathbf{x}_t, \mathbf{z}_t, \mathbf{y}_t\}$ be generated by ADM. Then for any $\mathbf{x}^*, \mathbf{z}^*$ satisfying $\mathbf{A}\mathbf{x}^* + \mathbf{B}\mathbf{z}^* = \mathbf{c}$, we have*

$$\begin{aligned} & f(\mathbf{x}_{t+1}) + g(\mathbf{z}_{t+1}) - (f(\mathbf{x}^*) + g(\mathbf{z}^*)) \\ & \leq \frac{1}{2\rho} (\|\mathbf{y}_t\|_2^2 - \|\mathbf{y}_{t+1}\|_2^2) - \frac{\rho}{2} \|\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_t - \mathbf{c}\|_2^2 + \frac{\rho}{2} (\|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_t\|_2^2 - \|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_{t+1}\|_2^2). \end{aligned} \quad (12)$$

Proof: Since $f(\mathbf{x})$ is a convex function and its subgradient is given in (10),

$$\begin{aligned}
f(\mathbf{x}_{t+1}) - f(\mathbf{x}^*) &\leq -\langle \mathbf{A}^T(\mathbf{y}_{t+1} + \rho(\mathbf{B}\mathbf{z}_t - \mathbf{B}\mathbf{z}_{t+1})), \mathbf{x}_{t+1} - \mathbf{x}^* \rangle \\
&= -\langle \mathbf{y}_{t+1} + \rho(\mathbf{B}\mathbf{z}_t - \mathbf{B}\mathbf{z}_{t+1}), \mathbf{A}\mathbf{x}_{t+1} - \mathbf{A}\mathbf{x}^* \rangle \\
&= -\langle \mathbf{y}_{t+1}, \mathbf{A}\mathbf{x}_{t+1} - \mathbf{c} + \mathbf{B}\mathbf{z}^* \rangle + \rho \langle \mathbf{B}\mathbf{z}_{t+1} - \mathbf{B}\mathbf{z}_t, \mathbf{A}\mathbf{x}_{t+1} - \mathbf{c} + \mathbf{B}\mathbf{z}^* \rangle \\
&= -\langle \mathbf{y}_{t+1}, \mathbf{A}\mathbf{x}_{t+1} - \mathbf{c} + \mathbf{B}\mathbf{z}^* \rangle + \frac{\rho}{2} (\|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_t\|_2^2 - \|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_{t+1}\|_2^2 \\
&\quad + \|\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_{t+1} - \mathbf{c}\|_2^2 - \|\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_t - \mathbf{c}\|_2^2) .
\end{aligned} \tag{13}$$

where the last equality uses

$$\langle \mathbf{u}_1 - \mathbf{u}_2, \mathbf{u}_3 + \mathbf{u}_4 \rangle = \frac{1}{2} (\|\mathbf{u}_4 - \mathbf{u}_2\|_2^2 - \|\mathbf{u}_4 - \mathbf{u}_1\|_2^2 + \|\mathbf{u}_3 + \mathbf{u}_1\|_2^2 - \|\mathbf{u}_3 + \mathbf{u}_2\|_2^2). \tag{14}$$

Similarly, for convex function $g(\mathbf{z})$, using its subgradient in (11), we have

$$g(\mathbf{z}_{t+1}) - g(\mathbf{z}^*) \leq -\langle \mathbf{B}^T \mathbf{y}_{t+1}, \mathbf{z}_{t+1} - \mathbf{z}^* \rangle = -\langle \mathbf{y}_{t+1}, \mathbf{B}\mathbf{z}_{t+1} - \mathbf{B}\mathbf{z}^* \rangle. \tag{15}$$

Adding (13) and (15) together yields

$$\begin{aligned}
&f(\mathbf{x}_{t+1}) + g(\mathbf{z}_{t+1}) - (f(\mathbf{x}^*) + g(\mathbf{z}^*)) \\
&\leq -\langle \mathbf{y}_{t+1}, \mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_{t+1} - \mathbf{c} \rangle + \frac{\rho}{2} \|\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_{t+1} - \mathbf{c}\|_2^2 \\
&\quad - \frac{\rho}{2} \|\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_t - \mathbf{c}\|_2^2 + \frac{\rho}{2} (\|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_t\|_2^2 - \|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_{t+1}\|_2^2) .
\end{aligned} \tag{16}$$

Recalling (8), the first two terms in (16) can be rewritten as

$$\begin{aligned}
&-\langle \mathbf{y}_{t+1}, \mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_{t+1} - \mathbf{c} \rangle + \frac{\rho}{2} \|\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_{t+1} - \mathbf{c}\|_2^2 \\
&= \frac{1}{2\rho} (2\langle \mathbf{y}_{t+1}, \mathbf{y}_t - \mathbf{y}_{t+1} \rangle + \|\mathbf{y}_t - \mathbf{y}_{t+1}\|_2^2) \\
&= \frac{1}{2\rho} (\|\mathbf{y}_t\|_2^2 - \|\mathbf{y}_{t+1}\|_2^2) .
\end{aligned} \tag{17}$$

Plugging back into (16) yields the result. \blacksquare

As observed in several experiments [11], the objective is not monotonically non-increasing. The following theorem shows the objective of ADM has the $O(1/T)$ convergence rate in an ergodic sense.

Theorem 1 *Let the sequences $\{\mathbf{x}_t, \mathbf{z}_t, \mathbf{y}_t\}$ be generated by ADM and $\bar{\mathbf{x}}_T = \frac{1}{T} \sum_{t=1}^T \mathbf{x}_t$, $\bar{\mathbf{z}}_T = \frac{1}{T} \sum_{t=1}^T \mathbf{z}_t$. For any $\mathbf{x}^*, \mathbf{z}^*$ satisfying $\mathbf{A}\mathbf{x}^* + \mathbf{B}\mathbf{z}^* = \mathbf{c}$, for any T , we have*

$$f(\bar{\mathbf{x}}_T) + g(\bar{\mathbf{z}}_T) - (f(\mathbf{x}^*) + g(\mathbf{z}^*)) \leq \frac{\lambda_{\max}^{\mathbf{B}} D_{\mathbf{z}}^2 \rho}{2T} . \tag{18}$$

Proof: In (12), ignoring $-\frac{\rho}{2} \|\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_t - \mathbf{c}\|_2^2$ and summing over t from 0 to $T-1$, we have the following telescoping sum

$$\begin{aligned}
&\sum_{t=0}^{T-1} [f(\mathbf{x}_{t+1}) + g(\mathbf{z}_{t+1}) - (f(\mathbf{x}^*) + g(\mathbf{z}^*))] \\
&\leq \frac{1}{2\rho} (\|\mathbf{y}_0\|_2^2 - \|\mathbf{y}_T\|_2^2) + \frac{\rho}{2} (\|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_0\|_2^2 - \|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_T\|_2^2) .
\end{aligned}$$

Since both f and g are convex, dividing by T , applying Jensen's inequality and letting the assumptions hold complete the proof. ■

Although (18) shows that the objective value converges to the optimal value, $\{\mathbf{x}_{t+1}, \mathbf{z}_{t+1}\}$ may not be feasible and the equality constraint may not necessarily be satisfied.

2.2 Convergence Rate for the Optimality Conditions (Constraints)

Assume that $\{\mathbf{x}^*, \mathbf{z}^*, \mathbf{y}^*\}$ satisfies the KKT conditions of the Lagrangian (4), i.e.,

$$-\mathbf{A}^T \mathbf{y}^* \in \partial f(\mathbf{x}^*), \quad (19)$$

$$-\mathbf{B}^T \mathbf{y}^* \in \partial g(\mathbf{z}^*), \quad (20)$$

$$\mathbf{A}\mathbf{x}^* + \mathbf{B}\mathbf{z}^* - \mathbf{c} = 0. \quad (21)$$

According to (10), condition (19) holds if $\mathbf{B}\mathbf{z}_{t+1} - \mathbf{B}\mathbf{z}_t = 0$. According to (11), condition (20) holds for every iterate. Therefore, the KKT conditions (19)-(21) hold if the following optimality conditions are satisfied:

$$\mathbf{B}\mathbf{z}_{t+1} - \mathbf{B}\mathbf{z}_t = 0, \quad (22)$$

$$\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_{t+1} - \mathbf{c} = 0, \quad (23)$$

The LHS of (22) is called *primal residual* and the LHS of (23) is called *equality constraint violation* or *dual residual* [11] when considering (8).

Define a *residual function* of optimality conditions as

$$R(s, t) = \|\mathbf{A}\mathbf{x}_s + \mathbf{B}\mathbf{z}_t - \mathbf{c}\|_2^2 + \|\mathbf{B}\mathbf{z}_t - \mathbf{B}\mathbf{z}_{s-1}\|_2^2, \quad (24)$$

where $s \in \{t, t+1\}$. In particular, the residual after the \mathbf{z} update (7) at iteration $(t+1)$ is

$$R(t+1, t+1) = \|\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_{t+1} - \mathbf{c}\|_2^2 + \|\mathbf{B}\mathbf{z}_{t+1} - \mathbf{B}\mathbf{z}_t\|_2^2. \quad (25)$$

and the residual after the \mathbf{x} -update (6) at $(t+1)$ is

$$R(t+1, t) = \|\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_t - \mathbf{c}\|_2^2. \quad (26)$$

Therefore, the convergence of $R(t+1, t+1)$ implies the convergence of the optimality conditions.

The following two lemmas show the residuals of optimality conditions (constraints) are monotonically non-increasing.

Lemma 3 *Let the sequences $\{\mathbf{x}_t, \mathbf{z}_t, \mathbf{y}_t\}$ be generated by ADM. Then*

$$R(t+1, t) \leq R(t, t) \quad (27)$$

Proof: Since $f(\mathbf{x})$ is a convex function and its subgradient is given in (9), for any \mathbf{x} , we have

$$\begin{aligned} f(\mathbf{x}_{t+1}) - f(\mathbf{x}) &\leq -\langle \mathbf{A}^T(\mathbf{y}_t + \rho(\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_t - \mathbf{c})), \mathbf{x}_{t+1} - \mathbf{x} \rangle \\ &= \langle \mathbf{y}_t, \mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}_{t+1} \rangle + \rho \langle \mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_t - \mathbf{c}, \mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}_{t+1} \rangle. \end{aligned} \quad (28)$$

Letting $\mathbf{x} = \mathbf{x}_t$, we have

$$\begin{aligned} f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) &\leq \langle \mathbf{y}_t, \mathbf{Ax}_t - \mathbf{Ax}_{t+1} \rangle + \rho \langle \mathbf{Ax}_{t+1} + \mathbf{Bz}_t - \mathbf{c}, \mathbf{Ax}_t - \mathbf{Ax}_{t+1} \rangle \\ &= \langle \mathbf{y}_t, \mathbf{Ax}_t - \mathbf{Ax}_{t+1} \rangle + \frac{\rho}{2} (\|\mathbf{Ax}_t + \mathbf{Bz}_t - \mathbf{c}\|_2^2 - \|\mathbf{Ax}_{t+1} + \mathbf{Bz}_t - \mathbf{c}\|_2^2 - \|\mathbf{Ax}_t - \mathbf{Ax}_{t+1}\|_2^2) . \end{aligned} \quad (29)$$

where the last equality uses

$$\langle \mathbf{u}_1 - \mathbf{u}_2, \mathbf{u}_3 - \mathbf{u}_1 \rangle = \frac{1}{2} (\|\mathbf{u}_2 - \mathbf{u}_3\|_2^2 - \|\mathbf{u}_1 - \mathbf{u}_2\|_2^2 - \|\mathbf{u}_1 - \mathbf{u}_3\|_2^2) . \quad (30)$$

Using the subgradient of f given in (10) at \mathbf{x}_t , for any \mathbf{x} ,

$$f(\mathbf{x}_t) - f(\mathbf{x}) \leq -\langle \mathbf{A}^T(\mathbf{y}_t + \rho(\mathbf{Bz}_{t-1} - \mathbf{Bz}_t)), \mathbf{x}_t - \mathbf{x} \rangle . \quad (31)$$

Letting $\mathbf{x} = \mathbf{x}_{t+1}$, we have

$$\begin{aligned} f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}) &\leq -\langle \mathbf{y}_t, \mathbf{Ax}_t - \mathbf{Ax}_{t+1} \rangle + \rho \langle \mathbf{Bz}_{t-1} - \mathbf{Bz}_t, \mathbf{Ax}_{t+1} - \mathbf{Ax}_t \rangle \\ &\leq \langle \mathbf{Ax}_{t+1} - \mathbf{Ax}_t, \mathbf{y}_t \rangle + \frac{\rho}{2} (\|\mathbf{Ax}_{t+1} - \mathbf{Ax}_t\|_2^2 + \|\mathbf{Bz}_{t-1} - \mathbf{Bz}_t\|_2^2) . \end{aligned} \quad (32)$$

Adding (29) and (32) together and rearranging the terms complete the proof. \blacksquare

Lemma 4 *Let the sequences $\{\mathbf{x}_t, \mathbf{z}_t, \mathbf{y}_t\}$ be generated by ADM. Then*

$$R(t+1, t+1) \leq R(t+1, t) \quad (33)$$

Proof: Recalling the subgradient of convex function $g(\mathbf{z})$ given in (11), we have

$$g(\mathbf{z}_{t+1}) - g(\mathbf{z}_t) \leq \langle -\mathbf{B}^T \mathbf{y}_{t+1}, \mathbf{z}_{t+1} - \mathbf{z}_t \rangle , \quad (34)$$

$$g(\mathbf{z}_t) - g(\mathbf{z}_{t+1}) \leq \langle -\mathbf{B}^T \mathbf{y}_t, \mathbf{z}_t - \mathbf{z}_{t+1} \rangle . \quad (35)$$

Adding (34) and (35) together yields

$$0 \leq \langle \mathbf{B}^T(\mathbf{y}_{t+1} - \mathbf{y}_t), \mathbf{z}_t - \mathbf{z}_{t+1} \rangle = \langle \mathbf{y}_{t+1} - \mathbf{y}_t, \mathbf{Bz}_t - \mathbf{Bz}_{t+1} \rangle . \quad (36)$$

According to (8), the right-hand side can be rewritten as

$$\begin{aligned} &\langle \mathbf{y}_{t+1} - \mathbf{y}_t, \mathbf{Bz}_t - \mathbf{Bz}_{t+1} \rangle \\ &= \rho \langle \mathbf{Ax}_{t+1} + \mathbf{Bz}_{t+1} - \mathbf{c}, (\mathbf{Bz}_t - \mathbf{c}) - (\mathbf{Bz}_{t+1} - \mathbf{c}) \rangle \\ &= \frac{\rho}{2} (\|\mathbf{Ax}_{t+1} + \mathbf{Bz}_t - \mathbf{c}\|_2^2 - \|\mathbf{Bz}_{t+1} - \mathbf{Bz}_t\|_2^2 - \|\mathbf{Ax}_{t+1} + \mathbf{Bz}_{t+1} - \mathbf{c}\|_2^2) . \end{aligned} \quad (37)$$

Plugging into (36) and rearranging the terms complete the proof. \blacksquare

The above two lemmas together shows that

$$R(t+1, t+1) \leq R(t+1, t) \leq R(t, t) \leq R(t, t-1) , \quad (38)$$

meaning $R(s, t)$ is monotonically non-increasing. The following lemma shows $R(t+1, t)$ is bounded by step differences of a telescoping series of \mathbf{y} and \mathbf{z} .

Lemma 5 *Let the sequences $\{\mathbf{x}_t, \mathbf{z}_t, \mathbf{y}_t\}$ be generated by ADM and $\{\mathbf{x}^*, \mathbf{z}^*, \mathbf{y}^*\}$ satisfy the KKT conditions (19)-(21), then*

$$R(t+1, t) \leq \|\mathbf{Bz}^* - \mathbf{Bz}_t\|_2^2 - \|\mathbf{Bz}^* - \mathbf{Bz}_{t+1}\|_2^2 + \frac{1}{\rho^2} (\|\mathbf{y}^* - \mathbf{y}_t\|_2^2 - \|\mathbf{y}^* - \mathbf{y}_{t+1}\|_2^2). \quad (39)$$

Proof: Assume $\{\mathbf{x}^*, \mathbf{y}^*\}$ satisfies (19). Since f is convex, then

$$f(\mathbf{x}^*) - f(\mathbf{x}_{t+1}) \leq -\langle \mathbf{A}^T \mathbf{y}^*, \mathbf{x}^* - \mathbf{x}_{t+1} \rangle = -\langle \mathbf{y}^*, \mathbf{Ax}^* - \mathbf{Ax}_{t+1} \rangle. \quad (40)$$

Similarly, for convex function g and $\{\mathbf{z}^*, \mathbf{y}^*\}$ satisfies (20), we have

$$g(\mathbf{z}^*) - g(\mathbf{z}_{t+1}) \leq -\langle \mathbf{B}^T \mathbf{y}^*, \mathbf{z}^* - \mathbf{z}_{t+1} \rangle = -\langle \mathbf{y}^*, \mathbf{Bz}^* - \mathbf{Bz}_{t+1} \rangle. \quad (41)$$

Adding them together and using the fact that $\mathbf{Ax}^* + \mathbf{Bz}^* = \mathbf{c}$, we have

$$f(\mathbf{x}^*) + g(\mathbf{z}^*) - (f(\mathbf{x}_{t+1}) + g(\mathbf{z}_{t+1})) \leq \langle \mathbf{y}^*, \mathbf{Ax}_{t+1} + \mathbf{Bz}_{t+1} - \mathbf{c} \rangle. \quad (42)$$

Adding (16) and (42) together yields

$$\begin{aligned} 0 &\leq \frac{\rho}{2} (\|\mathbf{Bz}^* - \mathbf{Bz}_t\|_2^2 - \|\mathbf{Bz}^* - \mathbf{Bz}_{t+1}\|_2^2 - \|\mathbf{Ax}_{t+1} + \mathbf{Bz}_t - \mathbf{c}\|_2^2 + \|\mathbf{Ax}_{t+1} + \mathbf{Bz}_{t+1} - \mathbf{c}\|_2^2) \\ &\quad + \langle \mathbf{y}^* - \mathbf{y}_{t+1}, \mathbf{Ax}_{t+1} + \mathbf{Bz}_{t+1} - \mathbf{c} \rangle. \end{aligned} \quad (43)$$

The last term can be rewritten as

$$\begin{aligned} \langle \mathbf{y}^* - \mathbf{y}_{t+1}, \mathbf{Ax}_{t+1} + \mathbf{Bz}_{t+1} - \mathbf{c} \rangle &= \frac{1}{\rho} \langle \mathbf{y}^* - \mathbf{y}_{t+1}, \mathbf{y}_{t+1} - \mathbf{y}_t \rangle \\ &= -\frac{1}{2\rho} (\|\mathbf{y}^* - \mathbf{y}_t\|_2^2 - \|\mathbf{y}^* - \mathbf{y}_{t+1}\|_2^2 - \|\mathbf{y}_{t+1} - \mathbf{y}_t\|_2^2). \end{aligned} \quad (44)$$

Substituting it into (43) and rearranging the terms gives

$$\begin{aligned} &\|\mathbf{Bz}^* - \mathbf{Bz}_t\|_2^2 - \|\mathbf{Bz}^* - \mathbf{Bz}_{t+1}\|_2^2 + \frac{1}{\rho^2} (\|\mathbf{y}^* - \mathbf{y}_t\|_2^2 - \|\mathbf{y}^* - \mathbf{y}_{t+1}\|_2^2) \\ &\geq \|\mathbf{Ax}_{t+1} + \mathbf{Bz}_t - \mathbf{c}\|_2^2 + \|\mathbf{Ax}_{t+1} + \mathbf{Bz}_{t+1} - \mathbf{c}\|_2^2 - \frac{1}{\rho^2} \|\mathbf{y}_{t+1} - \mathbf{y}_t\|_2^2 \\ &= \|\mathbf{Ax}_{t+1} + \mathbf{Bz}_t - \mathbf{c}\|_2^2, \end{aligned} \quad (45)$$

which completes the proof. ■

Now, we are ready to show that the optimality conditions have a $O(1/T)$ convergence rate.

Theorem 2 *Let the sequences $\{\mathbf{x}_t, \mathbf{z}_t, \mathbf{y}_t\}$ be generated by ADM. For any $\mathbf{x}^*, \mathbf{z}^*$ satisfying $\mathbf{Ax}^* + \mathbf{Bz}^* = \mathbf{c}$, for any T , we have*

$$R(T, T) \leq R(T, T-1) \leq \frac{\lambda_{\max}^{\mathbf{B}} D_{\mathbf{z}}^2 + D_{\mathbf{y}}^2 / \rho^2}{T}, \quad (46)$$

where $R(T, T) = \|\mathbf{Ax}_T + \mathbf{Bz}_T - \mathbf{c}\|_2^2 + \|\mathbf{Bz}_T - \mathbf{Bz}_{T-1}\|_2^2$.

Proof: Since $\|\mathbf{Ax}_{t+1} + \mathbf{Bz}_t - \mathbf{c}\|_2^2$ is monotonically non-increasing, we have

$$\begin{aligned}
TR(T, T-1) &= T\|\mathbf{Ax}_T + \mathbf{Bz}_{T-1} - \mathbf{c}\|_2^2 \leq \sum_{t=0}^{T-1} \|\mathbf{Ax}_{t+1} + \mathbf{Bz}_t - \mathbf{c}\|_2^2 \\
&\leq \|\mathbf{Bz}^* - \mathbf{Bz}_0\|_2^2 - \|\mathbf{Bz}^* - \mathbf{Bz}_T\|_2^2 + \frac{1}{\rho^2}(\|\mathbf{y}^* - \mathbf{y}_0\|_2^2 - \|\mathbf{y}^* - \mathbf{y}_T\|_2^2) \\
&\leq \|\mathbf{Bz}^* - \mathbf{Bz}_0\|_2^2 + \frac{1}{\rho^2}\|\mathbf{y}^* - \mathbf{y}_0\|_2^2.
\end{aligned} \tag{47}$$

Divide both sides by T . Letting Assumption 1 hold and using Lemma 4 yield (46). \blacksquare

Results similar to Lemma 4 and 5 have appeared in [11], but Lemma 3 is new. The monotonicity and $O(1/T)$ convergence rate for optimality conditions have also been shown in [42], but our proof is different and self-contained.

2.3 Rate of Convergence of ADM based on Variational Inequality

We now prove the $O(1/T)$ convergence rate for ADM using a variational inequality (VI) based on the Lagrangian given in (4). In this section, we need the following assumption [8, 7]:

Assumption 2 \mathbf{y} is bounded in \mathbb{R}^m and $\|\mathbf{y}\|_2 \leq D$, i.e., $\mathbf{y} \in \mathcal{Y} \subset \mathbb{R}^m$ and \mathcal{Y} is a bounded set.

Let $\Omega = \mathcal{X} \times \mathcal{Z} \times \mathcal{Y}$, where \mathcal{X} and \mathcal{Z} are defined in (2). Any $\mathbf{w}^* = (\mathbf{x}^*, \mathbf{z}^*, \mathbf{y}^*) \in \Omega$ solves the original problem in (3) optimally if it satisfies the following variational inequality [30, 54, 43]:

$$\forall \mathbf{w} \in \Omega, \quad h(\mathbf{w}) - h(\mathbf{w}^*) + \langle \mathbf{w} - \mathbf{w}^*, F(\mathbf{w}^*) \rangle \geq 0, \tag{48}$$

where $h(\mathbf{w}) = f(\mathbf{x}) + g(\mathbf{z})$ and

$$F(\mathbf{w}) = \begin{bmatrix} \mathbf{A}^T \mathbf{y} \\ \mathbf{B}^T \mathbf{y} \\ -(\mathbf{Ax} + \mathbf{Bz} - \mathbf{c}) \end{bmatrix} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{A}^T \\ \mathbf{0} & \mathbf{0} & \mathbf{B}^T \\ -\mathbf{A} & -\mathbf{B} & \mathbf{0} \end{bmatrix} \mathbf{w} + \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{c} \end{bmatrix} = \mathbf{M}\mathbf{w} + \mathbf{q}$$

is the gradient of the last term of the Lagrangian. \mathbf{M} is an anti-symmetric matrix and $\mathbf{w}^T \mathbf{M} \mathbf{w} = 0$. Then, $\tilde{\mathbf{w}} = (\tilde{\mathbf{x}}, \tilde{\mathbf{z}}, \tilde{\mathbf{y}})$ approximately solves the problem with accuracy ϵ if it satisfies

$$\forall \mathbf{w} \in \Omega, \quad h(\tilde{\mathbf{w}}) - h(\mathbf{w}) + \langle \tilde{\mathbf{w}} - \mathbf{w}, F(\tilde{\mathbf{w}}) \rangle \leq \epsilon. \tag{49}$$

We show that after T iterations, the average $\bar{\mathbf{w}}_T = \frac{1}{T} \sum_{t=1}^T \mathbf{w}_t$, where $\mathbf{w}_t = (\mathbf{x}_t, \mathbf{z}_t, \mathbf{y}_t)$ are from (6)-(8), satisfies the above inequality with $\epsilon = O(1/T)$.

Theorem 3 Let $\bar{\mathbf{w}}_T = \frac{1}{T} \sum_{t=1}^T \mathbf{w}_t$, where $\mathbf{w}_t = (\mathbf{x}_t, \mathbf{z}_t, \mathbf{y}_t)$ from (6)-(8). Let Assumption 1 and 2 hold, then

$$\forall \mathbf{w} \in \Omega, \quad h(\bar{\mathbf{w}}_T) - h(\mathbf{w}) + \langle \bar{\mathbf{w}}_T - \mathbf{w}, F(\bar{\mathbf{w}}_T) \rangle \leq \frac{L}{T}.$$

where $L = \frac{\rho}{2}\|\mathbf{Ax} - \mathbf{c}\|_2^2 + \frac{1}{2\rho}\|\mathbf{y}\|^2$.

Proof: Considering $f(\mathbf{x})$ is a convex function and its subgradient is given in (10), $\forall \mathbf{x} \in \mathcal{X}$,

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}) \leq -\langle \mathbf{A}^T(\mathbf{y}_{t+1} + \rho(\mathbf{B}\mathbf{z}_t - \mathbf{B}\mathbf{z}_{t+1})), \mathbf{x}_{t+1} - \mathbf{x} \rangle .$$

Rearranging the terms gives

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}) + \langle \mathbf{x}_{t+1} - \mathbf{x}, \mathbf{A}^T \mathbf{y}_{t+1} \rangle \leq \rho \langle \mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}_{t+1}, \mathbf{B}\mathbf{z}_t - \mathbf{B}\mathbf{z}_{t+1} \rangle . \quad (50)$$

Using the subgradient of g given in (11), we have $\forall \mathbf{z} \in \mathcal{Z}$

$$g(\mathbf{z}_{t+1}) - g(\mathbf{z}) + \langle \mathbf{z}_{t+1} - \mathbf{z}, \mathbf{B}^T \mathbf{y}_{t+1} \rangle \leq 0 . \quad (51)$$

Adding (50) and (51) and denoting $h(\mathbf{w}) = f(\mathbf{x}) + g(\mathbf{z})$, we have $\forall \mathbf{w} \in \Omega$

$$\begin{aligned} h(\mathbf{w}_{t+1}) - h(\mathbf{w}) + \langle \mathbf{w}_{t+1} - \mathbf{w}, F(\mathbf{w}_{t+1}) \rangle \\ \leq \rho \langle \mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}_{t+1}, \mathbf{B}\mathbf{z}_t - \mathbf{B}\mathbf{z}_{t+1} \rangle + \frac{1}{\rho} \langle \mathbf{y}_{t+1} - \mathbf{y}, -(\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_{t+1} - \mathbf{c}) \rangle \\ = \rho \langle \mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}_{t+1}, \mathbf{B}\mathbf{z}_t - \mathbf{B}\mathbf{z}_{t+1} \rangle + \frac{1}{\rho} \langle \mathbf{y} - \mathbf{y}_{t+1}, \mathbf{y}_{t+1} - \mathbf{y}_t \rangle . \end{aligned} \quad (52)$$

The first term can be rewritten as

$$\begin{aligned} 2\langle \mathbf{A}\mathbf{x} - \mathbf{A}\mathbf{x}_{t+1}, \mathbf{B}\mathbf{z}_t - \mathbf{B}\mathbf{z}_{t+1} \rangle \\ = 2\langle \mathbf{A}\mathbf{x} - \mathbf{c} - (\mathbf{A}\mathbf{x}_{t+1} - \mathbf{c}), \mathbf{B}\mathbf{z}_t - \mathbf{B}\mathbf{z}_{t+1} \rangle \\ = \|\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z}_t - \mathbf{c}\|^2 - \|\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z}_{t+1} - \mathbf{c}\|^2 + \|\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_{t+1} - \mathbf{c}\|^2 - \|\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_t - \mathbf{c}\|^2 , \end{aligned} \quad (53)$$

where the last equality uses (14). The second term in (52) is equivalent to

$$2\langle \mathbf{y} - \mathbf{y}_{t+1}, \mathbf{y}_{t+1} - \mathbf{y}_t \rangle = \|\mathbf{y} - \mathbf{y}_t\|^2 - \|\mathbf{y} - \mathbf{y}_{t+1}\|^2 - \|\mathbf{y}_t - \mathbf{y}_{t+1}\|^2 , \quad (54)$$

which uses (30). Substituting (53) and (54) into (52) and using (8), we have

$$\begin{aligned} h(\mathbf{w}_{t+1}) - h(\mathbf{w}) + \langle \mathbf{w}_{t+1} - \mathbf{w}, F(\mathbf{w}_{t+1}) \rangle \\ \leq \frac{\rho}{2} (\|\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z}_t - \mathbf{c}\|^2 - \|\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z}_{t+1} - \mathbf{c}\|^2) + \frac{1}{2\rho} (\|\mathbf{y} - \mathbf{y}_t\|^2 - \|\mathbf{y} - \mathbf{y}_{t+1}\|^2) . \end{aligned} \quad (55)$$

Summing over t from 0 to $T - 1$, we have the following telescoping sum

$$\sum_{t=1}^T [h(\mathbf{w}_t) - h(\mathbf{w}) + \langle \mathbf{w}_t - \mathbf{w}, F(\mathbf{w}_t) \rangle] \leq L , \quad (56)$$

where the constant $L = \frac{\rho}{2} \|\mathbf{A}\mathbf{x} - \mathbf{c}\|_2^2 + \frac{1}{2\rho} \|\mathbf{y}\|^2$. Recall that $h(\tilde{\mathbf{w}})$ is a convex function of $\tilde{\mathbf{w}}$. Further, from the definition of $F(\tilde{\mathbf{w}})$, we have

$$\langle \tilde{\mathbf{w}} - \mathbf{w}, F(\tilde{\mathbf{w}}) \rangle = \langle \tilde{\mathbf{w}} - \mathbf{w}, \mathbf{M}\tilde{\mathbf{w}} + \mathbf{q} \rangle = -\langle \mathbf{w}, \mathbf{M}\tilde{\mathbf{w}} \rangle + \langle \tilde{\mathbf{w}} - \mathbf{w}, \mathbf{q} \rangle , \quad (57)$$

which is a linear function of $\tilde{\mathbf{w}}$. Dividing both sides of (56) by T , recalling that $\bar{\mathbf{w}}_T = \frac{1}{T} \sum_{t=1}^T \mathbf{w}_t$, and using Jensen's inequality, we have

$$\begin{aligned} & h(\bar{\mathbf{w}}_T) - h(\mathbf{w}) + \langle \bar{\mathbf{w}}_T - \mathbf{w}, F(\bar{\mathbf{w}}_T) \rangle \\ & \leq \frac{1}{T} \sum_{t=1}^T h(\mathbf{w}_t) - h(\mathbf{w}) + \frac{1}{T} \sum_{t=1}^T \langle \mathbf{w}_t - \mathbf{w}, F(\mathbf{w}_t) \rangle \\ & \leq \frac{L}{T} = O\left(\frac{1}{T}\right), \end{aligned}$$

which establishes convergence rate for ADM. \blacksquare

The bound requires \mathbf{x} and \mathbf{y} to be bounded. In general, L is larger compared to the results in Theorem 1 and 2. According to (6),

$$\rho \sum_{t=0}^{T-1} (\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_{t+1} - \mathbf{c}) = \sum_{t=0}^{T-1} (\mathbf{y}_{t+1} - \mathbf{y}_t) = \mathbf{y}_T - \mathbf{y}_0 = \mathbf{y}_T, \quad (58)$$

meaning \mathbf{y}_T is the sum of all past residuls of constraint violation and thus $\|\mathbf{y}\|_2$ is large. [43] also shows a similar result based on an auxiliary sequence $\{\mathbf{x}_{t+1}, \mathbf{z}_{t+1}, \tilde{\mathbf{y}}_{t+1} = \mathbf{y}_t + \rho(\mathbf{A}\mathbf{x}_{t+1} + \mathbf{A}\mathbf{z}_t - \mathbf{c})\}$ instead of the sequence $\{\mathbf{x}_{t+1}, \mathbf{z}_{t+1}, \mathbf{y}_{t+1}\}$ generated by ADM. Compared to their proof, our proof is arguably simple and easier to understand. In fact, their proof is based on weak VI [54, 19, 30], while our proof is based on strong VI [54, 19, 30]. According to Minty's lemma [19, 30], they are equivalent if the solution set Ω is closed bounded and VI operator F is continuous and monotone.

3 Online Alternating Direction Method

In this section, we extend ADM to the online learning setting. Specifically, we focus on using online ADM (OADM) to solve the problem (2). For our analysis, \mathbf{A} and \mathbf{B} are assumed to be fixed. At round t , we consider solving the following regularized optimization problem:

$$\mathbf{x}_{t+1} = \underset{\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} = \mathbf{c}}{\operatorname{argmin}} f_t(\mathbf{x}) + g(\mathbf{z}) + \eta B_\phi(\mathbf{x}, \mathbf{x}_t), \quad (59)$$

where $\eta \geq 0$ is a learning rate and $B_\phi(\mathbf{x}, \mathbf{x}_t)$ is a Bregman divergence [3, 14].

Let $\phi : \Omega \rightarrow \mathbb{R}$ be a continuously differentiable and strictly convex function. Denote $\nabla\phi(\mathbf{y})$ as the gradient of ϕ at \mathbf{y} . The Bregman divergence $B_\phi : \Omega \times \operatorname{ri}(\Omega) \rightarrow \mathbb{R}_+$ is defined as

$$B_\phi(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x}) - \phi(\mathbf{y}) - \langle \nabla\phi(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle.$$

Two widely used examples are squared Euclidian distance $B_\phi(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$ and KL divergence $B_\phi(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^n x_i \log \frac{x_i}{y_i}$.

If the problem (59) is solved exactly in every step, standard analysis techniques [40] can be suitably adopted to obtain sublinear regret bounds. While (59) can be solved by batch ADM, we essentially obtain a double loop algorithm where the function f_t changes in the outer loop and the inner loop runs ADM iteratively till convergence so that the constraint are satisfied. Note that existing online methods, such as projected gradient descent and variants [40, 26] do assume a black-box approach for projecting onto the feasible set, which for linear constraint may require iterative cyclic projections [14].

For our analysis, instead of requiring the equality constraint to be satisfied at each time t , we only require the equality constraint to be satisfied in the long run, with a notion of regret associated with constraint. In particular, we consider the following constrained cumulative regret for the online learning problem:

$$\begin{aligned} & \sum_{t=1}^T f_t(\mathbf{x}_t) + g(\mathbf{z}_t) - \min_{\mathbf{Ax} + \mathbf{Bz} = \mathbf{c}} \sum_{t=1}^T f_t(\mathbf{x}) + g(\mathbf{z}) \\ \text{s.t. } & \sum_{t=1}^T \|\mathbf{Ax}_t + \mathbf{Bz}_t - \mathbf{c}\|_2^2 = o(T), \end{aligned} \quad (60)$$

where the cumulative constraint violation is sublinear in T . The goal is to design a single-loop algorithm for (60), which has sublinear regret in both the objective and the constraint violation.

The augmented Lagrangian of (59) at time t is

$$L_\rho^t(\mathbf{x}, \mathbf{y}, \mathbf{z}) = f_t(\mathbf{x}) + g(\mathbf{z}) + \langle \mathbf{y}, \mathbf{Ax} + \mathbf{Bz} - \mathbf{c} \rangle + \eta B_\phi(\mathbf{x}, \mathbf{x}_t) + \frac{\rho}{2} \|\mathbf{Ax} + \mathbf{Bz} - \mathbf{c}\|^2. \quad (61)$$

At time t , OADM (**Algorithm 1**) consists of just one pass through the following three update steps:

$$\mathbf{x}_{t+1} = \underset{\mathbf{x}}{\operatorname{argmin}} \{f_t(\mathbf{x}) + \langle \mathbf{y}_t, \mathbf{Ax} + \mathbf{Bz}_t - \mathbf{c} \rangle + \frac{\rho}{2} \|\mathbf{Ax} + \mathbf{Bz}_t - \mathbf{c}\|^2 + \eta B_\phi(\mathbf{x}, \mathbf{x}_t)\}, \quad (62)$$

$$\mathbf{z}_{t+1} = \underset{\mathbf{z}}{\operatorname{argmin}} \{g(\mathbf{z}) + \langle \mathbf{y}_t, \mathbf{Ax}_{t+1} + \mathbf{Bz} - \mathbf{c} \rangle + \frac{\rho}{2} \|\mathbf{Ax}_{t+1} + \mathbf{Bz} - \mathbf{c}\|^2\}, \quad (63)$$

$$\mathbf{y}_{t+1} = \mathbf{y}_t + \rho(\mathbf{Ax}_{t+1} + \mathbf{Bz}_{t+1} - \mathbf{c}). \quad (64)$$

Operationally, in round t , the algorithm presents a solution $\{\mathbf{x}_t, \mathbf{z}_t\}$ as well as \mathbf{y}_t . Then, nature reveals function f_t and we encounter two types of losses. The first type is the traditional loss measured by $f_t(\mathbf{x}_t) + g(\mathbf{z}_t)$, with corresponding cumulative regret

$$R_1(T) = \sum_{t=1}^T f_t(\mathbf{x}_t) + g(\mathbf{z}_t) - \min_{\mathbf{Ax} + \mathbf{Bz} = \mathbf{c}} \sum_{t=1}^T f_t(\mathbf{x}) + g(\mathbf{z}). \quad (65)$$

The second type is the residual of constraint violation, i.e., $\|\mathbf{Ax}_t + \mathbf{Bz}_t - \mathbf{c}\|^2$. As the updates include the primal and dual variables, in line with batch ADM, we use the following cumulative regret for constraint violation:

$$R^c(T) = \sum_{t=1}^T \|\mathbf{Ax}_{t+1} + \mathbf{Bz}_{t+1} - \mathbf{c}\|_2^2 + \|\mathbf{Bz}_{t+1} - \mathbf{Bz}_t\|_2^2. \quad (66)$$

The goal is to establish sublinear regret bounds for both the objective and constraint violation.

The OADM updates (62)-(63) are similar as ADM updates (6)-(7) except the \mathbf{x} update in OADM uses a time varying function f_t and an additional Bregman divergence, which is the first scenario where the regret bounds of R_1 (65) and R^c (66) will be presented in Section 4. We also consider another scenario, where $\eta = 0$ in (62) and thus the Bregman divergence is eliminated and only the quadratic penalty term is involved in the \mathbf{x} -update. \mathbf{x}_{t+1} is kept close to \mathbf{x}_t indirectly through the quadratic penalty term at \mathbf{z}_t . Instead of using $\{\mathbf{x}_t, \mathbf{z}_t\}$ as the solution at round t , we use a solution $\{\hat{\mathbf{x}}_t, \mathbf{z}_t\}$ based on \mathbf{z}_t such that $\mathbf{A}\hat{\mathbf{x}}_t + \mathbf{Bz}_t = \mathbf{c}$.

Algorithm 1 Online Alternating Direction Method (OADM)

- 1: **Input:** $f_t(\mathbf{x}) + g(\mathbf{z}), \mathbf{A}, \mathbf{B}, \mathbf{c}, \rho, \eta, \phi(\mathbf{x})$
 - 2: **Initialization:** $\mathbf{x}_1, \mathbf{z}_1, \mathbf{u}_1 = \mathbf{0}$
 - 3: **for** $t = 1$ to T **do**
 - 4: $\mathbf{x}_{t+1} = \operatorname{argmin}_{\mathbf{x}} f_t(\mathbf{x}) + \langle \mathbf{y}_t, \mathbf{Ax} + \mathbf{Bz}_t - \mathbf{c} \rangle + \frac{\rho}{2} \|\mathbf{Ax} + \mathbf{Bz}_t - \mathbf{c}\|^2 + \eta B_\phi(\mathbf{x}, \mathbf{x}_t)$,
 - 5: $\mathbf{z}_{t+1} = \operatorname{argmin}_{\mathbf{z}} g(\mathbf{z}) + \langle \mathbf{y}_t, \mathbf{Ax}_{t+1} + \mathbf{Bz} - \mathbf{c} \rangle + \frac{\rho}{2} \|\mathbf{Ax}_{t+1} + \mathbf{Bz} - \mathbf{c}\|^2$,
 - 6: $\mathbf{y}_{t+1} = \mathbf{y}_t + \rho(\mathbf{Ax}_{t+1} + \mathbf{Bz}_{t+1} - \mathbf{c})$.
 - 7: Receive a cost function f_{t+1} and incur loss $f_{t+1}(\mathbf{x}_{t+1}) + g(\mathbf{z}_{t+1})$ and constraint violation $\|\mathbf{Ax}_{t+1} + \mathbf{Bz}_{t+1} - \mathbf{c}\|_2^2$;
 - 8: **end for**
-

Regret bounds	$\eta > 0$		$\eta = 0$	
	R_1	R^c	R_2	R^c
general convex	$O(\sqrt{T})$	$O(\sqrt{T})$	$O(\sqrt{T})$	$O(\sqrt{T})$
strongly convex	$O(\log T)$	$O(\log T)$	$O(\log T)$	$O(\log T)$

Table 2: Regret Bounds for Online Alternating Direction Method

While $\{\hat{\mathbf{x}}_t, \mathbf{z}_t\}$ satisfies the constraint by design, the goal is to establish sublinear regret of the objective $f_t(\hat{\mathbf{x}}_t) + g(\mathbf{z}_t)$, i.e.,

$$R_2(T) = \sum_{t=1}^T f_t(\hat{\mathbf{x}}_t) + g(\mathbf{z}_t) - \min_{\mathbf{Ax} + \mathbf{Bz} = \mathbf{c}} \sum_{t=1}^T f_t(\mathbf{x}) + g(\mathbf{z}) . \quad (67)$$

The sublinear regret of constraint violation for the true $\{\mathbf{x}_t, \mathbf{z}_t\}$ defined in (66) should still be achieved. The regret bounds for OADM in the two scenarios are summarized in Table 2.

Before getting into the regret analysis, we discuss some example problems which can be solved using OADM. Like FOBOS and RDA, OADM can deal with machine learning problems where f_t is a loss function and g is a regularizer, e.g., generalized lasso and group lasso [11, 67, 69] using ℓ_1 or mixed norm, or an indicator function of a convex set. OADM can also be used to solve the batch optimization problems mentioned in Section 1, including linear programs, e.g., MAP LP relaxation [51] and LP decoding [5], and non-smooth optimization, e.g. robust PCA [13, 45]. Another promising scenario for OADM is consensus optimization [11] where distributed local variables are updated separately and reach a global consensus in the long run. More examples can be found in [11] and references therein.

In the sequel, we need the following assumptions:

Assumption 3

(a) For a p -norm $\|\cdot\|_p$, the dual norm of subgradient of $f_t(\mathbf{x})$ is bounded by G_f , i.e., $\|\nabla f'_t(\mathbf{x})\|_q \leq G_f$, where $f'_t(\mathbf{x}) \in \partial f_t(\mathbf{x}), \forall \mathbf{x} \in \mathcal{X}$ and $\frac{1}{p} + \frac{1}{q} = 1$.

(b) The Bregman divergence B_ϕ is defined on an α -strongly convex function ϕ with respect to a p -norm $\|\cdot\|_p$, i.e., $B_\phi(\mathbf{u}, \mathbf{v}) \geq \frac{\alpha}{2} \|\mathbf{x} - \mathbf{y}\|_p^2$ where $\alpha > 0$.

(c) $\mathbf{x}_1 = \mathbf{0}, \mathbf{y}_1 = \mathbf{0}, \mathbf{z}_1 = \mathbf{0}$. For any $\mathbf{x}^*, \mathbf{z}^*$ satisfying $\mathbf{Ax}^* + \mathbf{Bz}^* = \mathbf{c}$, $B_\phi(\mathbf{x}^*, \mathbf{x}_1) \leq D_{\mathbf{x}}^2, \|\mathbf{z}^* - \mathbf{z}_1\|_2 \leq D_{\mathbf{z}}$.

(d) $g(\mathbf{z}_1) = 0$ and $g(\mathbf{z}) \geq 0$.

(e) For any t , $f_t(\mathbf{x}_{t+1}) + g(\mathbf{z}_{t+1}) - (f_t(\mathbf{z}^*) + g(\mathbf{z}^*)) \geq -F$, where F is a positive constant.

In Assumption 3, (a) and (b) are in general required in the online learning setting [74, 27, 69]. (c) and (d) are simply for the ease of exposition of regret bounds and is commonly assumed for composite objective [27, 69], e.g., g is a regularizer in machine learning. We may assume the convex sets of \mathbf{x} and \mathbf{z} are bounded [74, 40] in (c). To obtain a sublinear regret bound for constraint violation, we need (e), which is true if functions are bounded from below or Lipschitz continuous in the convex set [47].

4 Regret Analysis for OADM

We consider two types of regret in OADM. The first type is the regret of the objective based on splitting variables, i.e., R_1 defined in (65). Aside from using splitting variables, R_1 is the standard regret in the online learning setting. The second is the regret of the constraint violation R^c defined in (66). We establish sublinear regret bounds for several cases whether f_t and g are strongly convex or not.

4.1 General Convex Functions

The following establishes the regret bounds for OADM for general convex functions.

Theorem 4 *Let the sequences $\{\mathbf{x}_t, \mathbf{z}_t, \mathbf{y}_t\}$ be generated by OADM (62)-(64) and let Assumption 3 hold. For any $\mathbf{x}^*, \mathbf{z}^*$ satisfying $\mathbf{A}\mathbf{x}^* + \mathbf{B}\mathbf{z}^* = \mathbf{c}$, setting $\eta = \frac{G_f\sqrt{T}}{D_{\mathbf{x}}\sqrt{2\alpha}}$ and $\rho = \sqrt{T}$, we have*

$$R_1(T) \leq \frac{\lambda_{\max}^{\mathbf{B}} D_{\mathbf{z}}^2 \sqrt{T}}{2} + \frac{\sqrt{2} G_f D_{\mathbf{x}} \sqrt{T}}{\sqrt{\alpha}}, \quad (68)$$

$$R^c(T) \leq \lambda_{\max}^{\mathbf{B}} D_{\mathbf{z}}^2 + \frac{2\sqrt{2} D_{\mathbf{x}} G_f}{\sqrt{\alpha}} + 2F\sqrt{T}. \quad (69)$$

Proof: Since \mathbf{x}_{t+1} minimizes (62), we have

$$0 \in \partial f_t(\mathbf{x}_{t+1}) + \mathbf{A}^T \mathbf{y}_t + \rho \mathbf{A}^T (\mathbf{A}\mathbf{x}_{t+1} - \mathbf{B}\mathbf{z}_t - \mathbf{c}) + \eta (\nabla \phi(\mathbf{x}_{t+1}) - \nabla \phi(\mathbf{x}_t)). \quad (70)$$

Rearranging the terms and using (64) give the subgradient of $f_t(\mathbf{x}_{t+1})$,

$$-\mathbf{A}^T (\mathbf{y}_{t+1} + \rho(\mathbf{B}\mathbf{z}_t - \mathbf{B}\mathbf{z}_{t+1})) - \eta (\nabla \phi(\mathbf{x}_{t+1}) - \nabla \phi(\mathbf{x}_t)) \in \partial f_t(\mathbf{x}_{t+1}) \quad (71)$$

Compared to (10) in Lemma 1, the additional terms introduced by Bregman divergence are included in the subgradient. Therefore, replacing f by f_t in Lemma 2 and adding the terms $-\eta (\nabla \phi(\mathbf{x}_{t+1}) - \nabla \phi(\mathbf{x}_t))$, we have

$$\begin{aligned} & f_t(\mathbf{x}_{t+1}) + g(\mathbf{z}_{t+1}) - (f_t(\mathbf{x}^*) + g(\mathbf{z}^*)) \\ & \leq \frac{1}{2\rho} (\|\mathbf{y}_t\|_2^2 - \|\mathbf{y}_{t+1}\|_2^2) - \frac{\rho}{2} \|\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_t - \mathbf{c}\|_2^2 + \frac{\rho}{2} (\|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_t\|_2^2 - \|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_{t+1}\|_2^2) \\ & \quad - \eta \langle \nabla \phi(\mathbf{x}_{t+1}) - \nabla \phi(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}^* \rangle. \end{aligned} \quad (72)$$

Using the three point property of Bregman divergence, the last term can be written as

$$-\langle \nabla \phi(\mathbf{x}_{t+1}) - \nabla \phi(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}^* \rangle = B_{\phi}(\mathbf{x}^*, \mathbf{x}_t) - B_{\phi}(\mathbf{x}^*, \mathbf{x}_{t+1}) - B_{\phi}(\mathbf{x}_{t+1}, \mathbf{x}_t). \quad (73)$$

Let $f'_t(\mathbf{x}_t) \in \partial f_t(\mathbf{x}_t)$. According to the Fenchel-Young's inequality [62], i.e., $2|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\|_q^2 + \|\mathbf{y}\|_p^2$, we have

$$\begin{aligned} f_t(\mathbf{x}_t) - f_t(\mathbf{x}_{t+1}) &\leq \langle f'_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t+1} \rangle = \langle \frac{1}{\sqrt{\alpha\eta}} f'_t(\mathbf{x}_t), \sqrt{\alpha\eta}(\mathbf{x}_t - \mathbf{x}_{t+1}) \rangle \\ &\leq \frac{1}{2\alpha\eta} \|f'_t(\mathbf{x}_t)\|_q^2 + \frac{\alpha\eta}{2} \|\mathbf{x}_t - \mathbf{x}_{t+1}\|_p^2. \end{aligned} \quad (74)$$

Recalling $B_\phi(\mathbf{x}_{t+1}, \mathbf{x}_t) \geq \frac{\alpha}{2} \|\mathbf{x}_t - \mathbf{x}_{t+1}\|_p^2$ and combining (72)-(74), we have

$$\begin{aligned} &f_t(\mathbf{x}_t) + g(\mathbf{z}_{t+1}) - (f_t(\mathbf{x}^*) + g(\mathbf{z}^*)) \\ &\leq \frac{1}{2\rho} (\|\mathbf{y}_t\|_2^2 - \|\mathbf{y}_{t+1}\|_2^2) - \frac{\rho}{2} \|\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_t - \mathbf{c}\|_2^2 + \frac{\rho}{2} (\|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_t\|_2^2 - \|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_{t+1}\|_2^2) \\ &\quad + \frac{1}{2\alpha\eta} \|f'_t(\mathbf{x}_t)\|_q^2 + \eta(B_\phi(\mathbf{x}^*, \mathbf{x}_t) - B_\phi(\mathbf{x}^*, \mathbf{x}_{t+1})). \end{aligned} \quad (75)$$

From Assumption 3, $g(\mathbf{z}) \geq 0$ and $g(\mathbf{z}_1) = 0$ for $\mathbf{z}_1 = \mathbf{0}$, $R_1(T)$ is bounded as follows :

$$\begin{aligned} R_1(T) &= \sum_{t=1}^T f_t(\mathbf{x}_t) + g(\mathbf{z}_{t+1}) - (f_t(\mathbf{x}^*) + g(\mathbf{z}^*)) + g(\mathbf{z}_1) - g(\mathbf{z}_{T+1}) \\ &\leq \frac{1}{2\rho} (\|\mathbf{y}_1\|_2^2 - \|\mathbf{y}_{T+1}\|_2^2) + \frac{\rho}{2} (\|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_1\|_2^2 - \|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_{T+1}\|_2^2) \\ &\quad + \eta(B_\phi(\mathbf{x}^*, \mathbf{x}_1) - B_\phi(\mathbf{x}^*, \mathbf{x}_{T+1})) + \frac{1}{2\alpha\eta} \sum_{t=1}^T \|f'_t(\mathbf{x}_t)\|_q^2 \\ &\leq \frac{\lambda_{\max}^{\mathbf{B}} D_{\mathbf{z}}^2 \rho}{2} + \eta D_{\mathbf{x}}^2 + \frac{G_f^2 T}{2\alpha\eta}. \end{aligned} \quad (76)$$

Setting $\eta = \frac{G_f \sqrt{T}}{D_{\mathbf{x}} \sqrt{2\alpha}}$ and $\rho = \sqrt{T}$ yields (68).

Now we prove (69). Rearranging the terms in (72), we have

$$\begin{aligned} \|\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_t - \mathbf{c}\|_2^2 &\leq \frac{2F}{\rho} + \frac{1}{\rho^2} (\|\mathbf{y}_t\|_2^2 - \|\mathbf{y}_{t+1}\|_2^2) + (\|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_t\|_2^2 - \|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_{t+1}\|_2^2) \\ &\quad + \frac{2\eta}{\rho} (B_\phi(\mathbf{x}^*, \mathbf{x}_t) - B_\phi(\mathbf{x}^*, \mathbf{x}_{t+1}) - B_\phi(\mathbf{x}_{t+1}, \mathbf{x}_t)). \end{aligned} \quad (77)$$

Letting Assumption 3 hold and summing over t from 1 to T , we have

$$\begin{aligned} &\sum_{t=1}^T \|\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_t - \mathbf{c}\|_2^2 \\ &\leq \frac{2FT}{\rho} + \frac{1}{\rho^2} (\|\mathbf{y}_1\|_2^2 - \|\mathbf{y}_{T+1}\|_2^2) + (\|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_1\|_2^2 - \|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_{T+1}\|_2^2) \\ &\quad + \frac{2\eta}{\rho} (B_\phi(\mathbf{x}^*, \mathbf{x}_1) - B_\phi(\mathbf{x}^*, \mathbf{x}_{T+1})) \\ &\leq \frac{2FT}{\rho} + \lambda_{\max}^{\mathbf{B}} D_{\mathbf{z}}^2 + \frac{2\eta}{\rho} D_{\mathbf{x}}^2. \end{aligned} \quad (78)$$

Setting $\eta = \frac{G_f \sqrt{T}}{D_{\mathbf{x}} \sqrt{2\alpha}}$ and $\rho = \sqrt{T}$, we have (69) by using Lemma 4. \blacksquare

Note the bounds are achieved without any explicit assumptions on $\mathbf{A}, \mathbf{B}, \mathbf{c}$.³ The subgradient of f_t is required to be bounded, but the subgradient of g is not necessarily bounded. Thus, the bounds hold for the case where g is an indicator function of a convex set. Compared to regret bound for COMID which is $\frac{G_f D_{\mathbf{x}} \sqrt{T}}{\sqrt{\alpha}}$ [26], the regret bound for the objective of ADMM has an additional term $\frac{\lambda_{\max}^{\mathbf{B}} D_{\mathbf{z}}^2 \sqrt{T}}{2}$ which is for the splitting variable \mathbf{z} . In addition to the $O(\sqrt{T})$ regret bound, OADM achieves the $O(\sqrt{T})$ bound for the constraint violation, which is not considered in the start-of-the-art online learning algorithms [26, 27, 69], since they do not explicitly handle linear constraint of the form $\mathbf{Ax} + \mathbf{Bz} = \mathbf{c}$. In fact, the bound for constraint violation could be reduced to a constant if \mathbf{y}_t is assumed to be bounded (see Assumption 2), which is shown in the following theorem.

Theorem 5 *Let the sequences $\{\mathbf{x}_t, \mathbf{z}_t, \mathbf{y}_t\}$ be generated by OADM. Assume that $\|\mathbf{y}_t\|_2 \leq D$. Setting $\rho = \sqrt{T}$, then*

$$\sum_{t=1}^T \|\mathbf{Ax}_{t+1} + \mathbf{Bz}_{t+1} - \mathbf{c}\|_2^2 \leq 4D^2. \quad (79)$$

Proof: According to (64), we have

$$\|\mathbf{Ax}_{t+1} + \mathbf{Bz}_{t+1} - \mathbf{c}\|_2^2 = \left\| \frac{1}{\rho} (\mathbf{y}_{t+1} - \mathbf{y}_t) \right\|_2^2 \leq \frac{2}{\rho^2} (\|\mathbf{y}_{t+1}\|_2^2 + \|\mathbf{y}_t\|_2^2) \leq \frac{4D^2}{\rho^2}. \quad (80)$$

Summing over t from 1 to T and setting $\rho = \sqrt{T}$ yield (79). \blacksquare

4.2 Strongly Convex Functions

We assume both $f_t(\mathbf{x})$ and g are strongly convex. Specifically, we assume $f_t(\mathbf{x})$ is β_1 -strongly convex with respect to a differentiable convex function ϕ , i.e.,

$$f_t(\mathbf{x}^*) \geq f_t(\mathbf{x}) + \langle f'_t(\mathbf{x}), \mathbf{x}^* - \mathbf{x} \rangle + \beta_1 B_{\phi}(\mathbf{x}^*, \mathbf{x}), \quad (81)$$

where $f'_t(\mathbf{x})$ denotes the subgradient of f_t at \mathbf{x} and $\beta_1 > 0$. Assume g is a β_2 -strongly convex function, i.e.,

$$g(\mathbf{z}^*) \geq g(\mathbf{z}) + \langle g'(\mathbf{z}), \mathbf{z}^* - \mathbf{z} \rangle + \frac{\beta_2}{2} \|\mathbf{z}^* - \mathbf{z}\|_2^2, \quad (82)$$

where $g'(\mathbf{z})$ denotes the subgradient of g at \mathbf{z} and $\beta_2 > 0$.

Instead of using fixed ρ and η , we allow them to change over time, i.e., ρ_t and η_t , which is fairly standard in the proof of logarithmic regret bounds [40, 27, 69] where the curvature of a sequence of strongly convex functions f_t is considered. The following theorem establishes logarithmic regret bounds for R_1 as well as R^c .

³We do assume that $\mathbf{Ax} + \mathbf{Bz} = \mathbf{c}$ is feasible.

Theorem 6 Let Assumption 3 hold. Assume $f_t(\mathbf{x})$ and g are strongly convex given in (81) and (82). Setting $\eta_t = \beta_1 t, \rho_t = \beta_2 t / \lambda_{\max}^{\mathbf{B}}$, we have

$$R_1(T) \leq \frac{G_f^2}{2\alpha\beta_1} \log(T+1) + \frac{\beta_2 D_{\mathbf{z}}^2}{2} + \beta_1 D_{\mathbf{x}}^2, \quad (83)$$

$$R^c(T) \leq \frac{2F\lambda_{\max}^{\mathbf{B}}}{\beta_2} \log(T+1) + \lambda_{\max}^{\mathbf{B}} D_{\mathbf{z}}^2 + \frac{2\beta_1 \lambda_{\max}^{\mathbf{B}} D_{\mathbf{x}}^2}{\beta_2}. \quad (84)$$

Proof: Assume $f_t(\mathbf{x})$ and g are strongly convex (81)-(82). Let \mathbf{x} be \mathbf{x}_{t+1} and \mathbf{z} be \mathbf{z}_{t+1} in (81)-(82) respectively. Adding them together and rearranging the terms give

$$\begin{aligned} & f_t(\mathbf{x}_{t+1}) + g(\mathbf{z}_{t+1}) - (f_t(\mathbf{x}^*) + g(\mathbf{z}^*)) \\ & \leq \langle f'_t(\mathbf{x}_{t+1}), \mathbf{x}_{t+1} - \mathbf{x}^* \rangle - \beta_1 B_\phi(\mathbf{x}^*, \mathbf{x}_{t+1}) + \langle g'(\mathbf{z}_{t+1}), \mathbf{z}_{t+1} - \mathbf{z}^* \rangle - \frac{\beta_2}{2} \|\mathbf{z}^* - \mathbf{z}_{t+1}\|_2^2. \end{aligned} \quad (85)$$

Compared to the general convex case in Theorem 4, the right hand side has two additional strongly convex terms. (85) can be obtained by letting ρ, η be ρ_{t+1}, η_{t+1} respectively in (72) and adding the two strongly convex term as follows:

$$\begin{aligned} & f_t(\mathbf{x}_{t+1}) + g(\mathbf{z}_{t+1}) - (f_t(\mathbf{x}^*) + g(\mathbf{z}^*)) \\ & \leq \frac{1}{2\rho_{t+1}} (\|\mathbf{y}_t\|_2^2 - \|\mathbf{y}_{t+1}\|_2^2) - \frac{\rho_{t+1}}{2} \|\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_t - \mathbf{c}\|_2^2 + \frac{\rho_{t+1}}{2} (\|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_t\|_2^2 - \|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_{t+1}\|_2^2) \\ & \quad + \eta_{t+1} (B_\phi(\mathbf{x}^*, \mathbf{x}_t) - B_\phi(\mathbf{x}^*, \mathbf{x}_{t+1}) - B_\phi(\mathbf{x}_{t+1}, \mathbf{x}_t)) - \beta_1 B_\phi(\mathbf{x}^*, \mathbf{x}_{t+1}) - \frac{\beta_2}{2} \|\mathbf{z}^* - \mathbf{z}_{t+1}\|_2^2. \end{aligned} \quad (86)$$

Let η be η_{t+1} in (74). Adding to (86) and ignoring the negative term $-\frac{\rho_{t+1}}{2} \|\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_t - \mathbf{c}\|_2^2$, we have

$$\begin{aligned} & f_t(\mathbf{x}_t) + g(\mathbf{z}_{t+1}) - (f_t(\mathbf{x}^*) + g(\mathbf{z}^*)) \\ & \leq \frac{1}{\eta_{t+1}} \|f'_t(\mathbf{x}_t)\|_*^2 + \frac{1}{2\rho_{t+1}} (\|\mathbf{y}_t\|_2^2 - \|\mathbf{y}_{t+1}\|_2^2) + \frac{\rho_{t+1}}{2} (\|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_t\|_2^2 - \|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_{t+1}\|_2^2) \\ & \quad - \frac{\beta_2}{2} \|\mathbf{z}^* - \mathbf{z}_{t+1}\|_2^2 + (\eta_{t+1} (B_\phi(\mathbf{x}^*, \mathbf{x}_t) - B_\phi(\mathbf{x}^*, \mathbf{x}_{t+1})) - \beta_1 B_\phi(\mathbf{x}^*, \mathbf{x}_{t+1})) - \eta_{t+1} B_\phi(\mathbf{x}_{t+1}, \mathbf{x}_t). \end{aligned} \quad (87)$$

Summing over t from 1 to T , we have

$$\begin{aligned} R_1(T) & \leq \frac{1}{2\alpha} \sum_{t=1}^T \frac{1}{\eta_{t+1}} \|f'_t(\mathbf{x}_t)\|_*^2 + \sum_{t=1}^T \frac{1}{2\rho_{t+1}} (\|\mathbf{y}_t\|_2^2 - \|\mathbf{y}_{t+1}\|_2^2) \\ & \quad + \sum_{t=1}^T \left(\frac{\rho_{t+1}}{2} (\|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_t\|_2^2 - \|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_{t+1}\|_2^2) - \frac{\beta_2}{2} \|\mathbf{z}^* - \mathbf{z}_{t+1}\|_2^2 \right) \\ & \quad + \sum_{t=1}^T (\eta_{t+1} (B_\phi(\mathbf{x}^*, \mathbf{x}_t) - B_\phi(\mathbf{x}^*, \mathbf{x}_{t+1})) - \beta_1 B_\phi(\mathbf{x}^*, \mathbf{x}_{t+1})). \end{aligned} \quad (88)$$

Assuming ρ_t is non-decreasing, we have

$$\sum_{t=1}^T \frac{1}{2\rho_{t+1}} (\|\mathbf{y}_t\|_2^2 - \|\mathbf{y}_{t+1}\|_2^2) \leq \frac{1}{2\rho_2} \|\mathbf{y}_1\|_2^2 = 0. \quad (89)$$

Using $\|\mathbf{Bz}^* - \mathbf{Bz}_{t+1}\|_2^2 \leq \lambda_{\max}^{\mathbf{B}} \|\mathbf{z}^* - \mathbf{z}_{t+1}\|_2^2$ and setting $\rho_t = \beta_2 t / \lambda_{\max}^{\mathbf{B}}$, we have

$$\begin{aligned}
& \sum_{t=1}^T [\rho_{t+1}(\|\mathbf{Bz}^* - \mathbf{Bz}_t\|_2^2 - \|\mathbf{Bz}^* - \mathbf{Bz}_{t+1}\|_2^2) - \beta_2 \|\mathbf{z}^* - \mathbf{z}_{t+1}\|_2^2] \\
& \leq \sum_{t=1}^T \left[\rho_{t+1}(\|\mathbf{Bz}^* - \mathbf{Bz}_t\|_2^2 - \|\mathbf{Bz}^* - \mathbf{Bz}_{t+1}\|_2^2) - \frac{\beta_2}{\lambda_{\max}^{\mathbf{B}}} \|\mathbf{Bz}^* - \mathbf{Bz}_{t+1}\|_2^2 \right] \\
& \leq \rho_2 \|\mathbf{Bz}^* - \mathbf{Bz}_1\|_2^2 + \sum_{t=2}^T \|\mathbf{Bz}^* - \mathbf{Bz}_t\|_2^2 (\rho_{t+1} - \rho_t - \frac{\beta_2}{\lambda_{\max}^{\mathbf{B}}}) \\
& = 2\beta_2 D_{\mathbf{z}}^2,
\end{aligned} \tag{90}$$

where the last equality uses the Assumption 3. Similarly, setting $\eta_t = \beta_1 t$, the last term in (88) can be rewritten as

$$\begin{aligned}
& \sum_{t=1}^T [\eta_{t+1}(B_{\phi}(\mathbf{x}^*, \mathbf{x}_t) - B_{\phi}(\mathbf{x}^*, \mathbf{x}_{t+1})) - \beta_1 B_{\phi}(\mathbf{x}^*, \mathbf{x}_{t+1})] \\
& = \eta_2 B_{\phi}(\mathbf{x}^*, \mathbf{x}_1) + \sum_{t=2}^T B_{\phi}(\mathbf{x}^*, \mathbf{x}_t)(\eta_{t+1} - \eta_t - \beta_1) - \eta_{T+1} B_{\phi}(\mathbf{x}^*, \mathbf{x}_{T+1}) - \beta_1 B_{\phi}(\mathbf{x}^*, \mathbf{x}_{T+1}) \\
& \leq \eta_2 B_{\phi}(\mathbf{x}^*, \mathbf{x}_1) + \sum_{t=2}^T B_{\phi}(\mathbf{x}^*, \mathbf{x}_t)(\eta_{t+1} - \eta_t - \beta_1) \\
& = 2\beta_1 D_{\mathbf{x}}^2.
\end{aligned} \tag{91}$$

Setting $\rho_t = \beta_2 t / \lambda_{\max}^{\mathbf{B}}$, $\eta_t = \beta_1 t$ and combining (88), (89), (90) and (91), we have

$$R_1(T) \leq \frac{G_f^2}{2\alpha} \sum_{t=1}^T \frac{1}{\beta_1(t+1)} + \beta_2 D_{\mathbf{z}}^2 + 2\beta_1 D_{\mathbf{x}}^2. \tag{92}$$

Applying $\sum_{t=1}^T \frac{1}{t+1} \leq \int_{t=0}^T \frac{1}{t+1} dt = \log(T+1)$ gives (83).

Now we prove (84). Rearranging terms in (86), we have

$$\begin{aligned}
\|\mathbf{Ax}_{t+1} + \mathbf{Bz}_t - \mathbf{c}\|_2^2 & \leq \frac{2F}{\rho_{t+1}} + \frac{1}{\rho_{t+1}^2} (\|\mathbf{y}_t\|_2^2 - \|\mathbf{y}_{t+1}\|_2^2) + (\|\mathbf{Bz}^* - \mathbf{Bz}_t\|_2^2 - \|\mathbf{Bz}^* - \mathbf{Bz}_{t+1}\|_2^2) \\
& \quad + \frac{2\eta_{t+1}}{\rho_{t+1}} (B_{\phi}(\mathbf{x}^*, \mathbf{x}_t) - B_{\phi}(\mathbf{x}^*, \mathbf{x}_{t+1}) - B_{\phi}(\mathbf{x}_{t+1}, \mathbf{x}_t)).
\end{aligned} \tag{93}$$

Letting $\rho_t = \beta_2 t / \lambda_{\max}^{\mathbf{B}}$ and $\eta_t = \beta_1 t$ and summing over t from 0 to T , we have

$$\begin{aligned}
& \sum_{t=1}^T \|\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_t - \mathbf{c}\|_2^2 \\
& \leq 2F \sum_{t=1}^T \frac{1}{\rho_{t+1}} + \sum_{t=1}^T \frac{1}{\rho_{t+1}^2} (\|\mathbf{y}_t\|_2^2 - \|\mathbf{y}_{t+1}\|_2^2) + (\|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_0\|_2^2 - \|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_{T+1}\|_2^2) \\
& \quad + \sum_{t=1}^T \frac{2\eta_{t+1}}{\rho_{t+1}} (B_\phi(\mathbf{x}^*, \mathbf{x}_t) - B_\phi(\mathbf{x}^*, \mathbf{x}_{t+1})) \\
& \leq \frac{2F\lambda_{\max}^{\mathbf{B}} \log(T+1)}{\beta_2} + \lambda_{\max}^{\mathbf{B}} D_{\mathbf{z}}^2 + \frac{2\beta_1 \lambda_{\max}^{\mathbf{B}} D_{\mathbf{x}}^2}{\beta_2}. \tag{94}
\end{aligned}$$

We use (89) in the last inequality. According to Lemma 4, we have (84). \blacksquare

To guarantee logarithmic regret bounds for both objective and constraints violation, OADM requires both f_t and g to be strongly convex. FOBOS, COMID, and RDA only require g to be strongly convex although they do not consider linear constraints explicitly. Further, the logarithmic regret bounds for the constraints violation could reduce to constant bound if assuming \mathbf{y}_t is bounded.

Theorem 7 *Let the sequences $\{\mathbf{x}_t, \mathbf{z}_t, \mathbf{y}_t\}$ be generated by OADM and $\|\mathbf{y}_t\|_2 \leq D$. Setting $\rho_t = \beta_2 t / \lambda_{\max}^{\mathbf{B}}$, then*

$$\sum_{t=1}^T \|\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_{t+1} - \mathbf{c}\|_2^2 \leq \frac{2\pi D^2 \lambda_{\max}^{\mathbf{B}}}{3\beta_2^2}. \tag{95}$$

Proof: Replacing ρ by ρ_{t+1} in (80) and summing over t from 1 to T , we have

$$\sum_{t=1}^T \|\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_{t+1} - \mathbf{c}\|_2^2 \leq \sum_{t=1}^T \frac{4D^2}{\rho_{t+1}^2}. \tag{96}$$

Setting $\rho_t = \beta_2 t / \lambda_{\max}^{\mathbf{B}}$ and using $\sum_{t=1}^T \frac{1}{t^2} \leq \frac{\pi}{6}$ complete the proof. \blacksquare

5 Regret Analysis for OADM with $\eta = 0$

We analyze the regret bound when $\eta = 0$. In this case, OADM has the same updates as ADM except f_t is changing over time. The \mathbf{x} -update only including the quadratic penalty term is easier to solve than the one with an additional Bregman divergence, particularly when the Bregman divergence is not a quadratic function. Without a Bregman divergence to keep two consecutive iterates of \mathbf{x} close, the quadratic penalty term is qualified for this task through variable \mathbf{z} . We consider \mathbf{z}_t to be the key primal variable, and compute $\hat{\mathbf{x}}_t$ using \mathbf{z}_t so that $\mathbf{A}\hat{\mathbf{x}}_t + \mathbf{B}\mathbf{z}_t = \mathbf{c}$. Therefore, we use the regret bound R_2 defined in (67). While $\{\hat{\mathbf{x}}_t, \mathbf{z}_t\}$ satisfies the equality constraint, $\{\mathbf{x}_t, \mathbf{z}_t\}$ need not satisfy $\mathbf{A}\mathbf{x}_t + \mathbf{B}\mathbf{z}_t - \mathbf{c} = \mathbf{0}$. Therefore, we also consider bounds for R^c as defined in (66). A common case we often encounter is when $\mathbf{A} = \mathbf{I}, \mathbf{B} = -\mathbf{I}, \mathbf{c} = \mathbf{0}$, thus $\hat{\mathbf{x}}_t = \mathbf{z}_t$. Consensus optimization is a typical example of this form [11, 9, 53]. In machine learning, many examples like (group) lasso [11, 71] can be reformulated in this way.

In this section, we need additional assumptions. In Assumption 3 (a), we specify the dual norm $\|\cdot\|_q$ to be ℓ_2 , i.e., $\|f_t(\mathbf{x})\|_2 \leq G_f$. To guarantee that $\mathbf{A}\hat{\mathbf{x}}_t + \mathbf{B}\mathbf{z}_t = \mathbf{c}, \mathbf{A} \in \mathbb{R}^{m \times n_1}$ is feasible, the equality

constraint, in particular, implicitly requires the assumption $m \leq n_1$. On the other hand, to establish a bound for R_2 , \mathbf{A} should be full-column rank, i.e., $\text{rank}(\mathbf{A}) = n_1$. Therefore, we need the following assumption in this scenario:

Assumption 4 \mathbf{A} is a square and full rank matrix, i.e., \mathbf{A} is invertible. Let $\lambda_{\min}^{\mathbf{A}}$ be the smallest eigenvalue of $\mathbf{A}\mathbf{A}^T$, then $\lambda_{\min}^{\mathbf{A}} > 0$.

Assumption 4 is satisfied in most examples like lasso and consensus optimization. Considering the subgradient of f_t given in (9), if there always exists a vector \mathbf{v}_t such that $-\mathbf{A}^T \mathbf{v}_t \in \partial f_t(\mathbf{x}_t)$, Assumption 4 can be safely removed under the implicit assumption that $\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} = \mathbf{c}$ is feasible.

5.1 General Convex Functions

The following theorem shows the regret bounds for R_2 as well as R^c .

Theorem 8 Let $\eta = 0$ in OADM. Let Assumption 3 and 4 hold. For any $\mathbf{x}^*, \mathbf{z}^*$ satisfying $\mathbf{A}\mathbf{x}^* + \mathbf{B}\mathbf{z}^* = \mathbf{c}$, setting $\rho = \frac{G_f \sqrt{T}}{D_{\mathbf{z}} \sqrt{\lambda_{\min}^{\mathbf{A}} \lambda_{\max}^{\mathbf{B}}}}$, we have

$$R_2(T) \leq \frac{G_f D_{\mathbf{z}} \sqrt{\lambda_{\max}^{\mathbf{B}}}}{\sqrt{\lambda_{\min}^{\mathbf{A}}}} \sqrt{T}, \quad (97)$$

$$R^c(T) \leq \lambda_{\max}^{\mathbf{B}} D_{\mathbf{z}}^2 + \frac{2F D_{\mathbf{z}} \sqrt{\lambda_{\min}^{\mathbf{A}} \lambda_{\max}^{\mathbf{B}} T}}{G_f}. \quad (98)$$

Proof: Replacing f by f_t in Lemma 2, we have

$$\begin{aligned} & f_t(\mathbf{x}_{t+1}) + g(\mathbf{z}_{t+1}) - (f(\mathbf{x}^*) + g(\mathbf{z}^*)) \\ & \leq \frac{1}{2\rho} (\|\mathbf{y}_t\|_2^2 - \|\mathbf{y}_{t+1}\|_2^2) - \frac{\rho}{2} \|\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_t - \mathbf{c}\|_2^2 + \frac{\rho}{2} (\|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_t\|_2^2 - \|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_{t+1}\|_2^2). \end{aligned} \quad (99)$$

Let $f'_t(\hat{\mathbf{x}}_t) \in \partial f_t(\hat{\mathbf{x}}_t)$. Recalling $\mathbf{A}\hat{\mathbf{x}}_t + \mathbf{B}\mathbf{z}_t = \mathbf{c}$, then

$$\begin{aligned} & f_t(\hat{\mathbf{x}}_t) - f_t(\mathbf{x}_{t+1}) \leq \langle f'_t(\hat{\mathbf{x}}_t), \hat{\mathbf{x}}_t - \mathbf{x}_{t+1} \rangle = \langle (\mathbf{A}^{-1})^T f'_t(\hat{\mathbf{x}}_t), \mathbf{A}\hat{\mathbf{x}}_t - \mathbf{A}\mathbf{x}_{t+1} \rangle \\ & = -\langle (\mathbf{A}^{-1})^T f'_t(\hat{\mathbf{x}}_t), \mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_t - \mathbf{c} \rangle \leq \frac{1}{2\lambda_{\min}^{\mathbf{A}} \rho} \|f'_t(\hat{\mathbf{x}}_t)\|_2^2 + \frac{\rho}{2} \|\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_t - \mathbf{c}\|_2^2. \end{aligned} \quad (100)$$

Adding to (99) gives

$$\begin{aligned} & f_t(\hat{\mathbf{x}}_t) + g(\mathbf{z}_{t+1}) - (f_t(\mathbf{x}^*) + g(\mathbf{z}^*)) \\ & \leq \frac{1}{2\lambda_{\min}^{\mathbf{A}} \rho} \|f'_t(\hat{\mathbf{x}}_t)\|_2^2 + \frac{1}{2\rho} (\|\mathbf{y}_t\|_2^2 - \|\mathbf{y}_{t+1}\|_2^2) + \frac{\rho}{2} (\|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_t\|_2^2 - \|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_{t+1}\|_2^2). \end{aligned} \quad (101)$$

Letting the assumptions hold, $R_2(T)$ is bounded as:

$$\begin{aligned}
R_2(T) &\leq \sum_{t=1}^T [f_t(\hat{\mathbf{x}}_t) + g(\mathbf{z}_{t+1}) - (f_t(\mathbf{x}^*) + g(\mathbf{z}^*))] \\
&\leq \frac{1}{2\lambda_{\min}^{\mathbf{A}}\rho} \sum_{t=1}^T \|f'_t(\hat{\mathbf{x}}_t)\|_2^2 + \frac{1}{2\rho} (\|\mathbf{y}_1\|_2^2 - \|\mathbf{y}_{T+1}\|_2^2) + \frac{\rho}{2} (\|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_1\|_2^2 - \|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_{T+1}\|_2^2) \\
&\leq \frac{G_f^2 T}{2\lambda_{\min}^{\mathbf{A}}\rho} + \frac{\lambda_{\max}^{\mathbf{B}} D_{\mathbf{z}}^2 \rho}{2}.
\end{aligned} \tag{102}$$

Setting $\rho = \frac{G_f \sqrt{T}}{D_{\mathbf{z}} \sqrt{\lambda_{\min}^{\mathbf{A}} \lambda_{\max}^{\mathbf{B}}}}$ yields (97).

Now we prove (98). Rearranging the terms in (99), we have

$$\|\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_t - \mathbf{c}\|_2^2 \leq \frac{2F}{\rho} + \frac{1}{\rho^2} (\|\mathbf{y}_t\|_2^2 - \|\mathbf{y}_{t+1}\|_2^2) + (\|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_t\|_2^2 - \|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_{t+1}\|_2^2). \tag{103}$$

Letting the assumptions hold and summing over t from 1 to T , we have

$$\begin{aligned}
&\sum_{t=1}^T \|\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_t - \mathbf{c}\|_2^2 \\
&\leq \frac{2FT}{\rho} + \frac{1}{\rho^2} (\|\mathbf{y}_1\|_2^2 - \|\mathbf{y}_{T+1}\|_2^2) + (\|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_1\|_2^2 - \|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_{T+1}\|_2^2) \\
&\leq \frac{2FT}{\rho} + \lambda_{\max}^{\mathbf{B}} D_{\mathbf{z}}^2.
\end{aligned} \tag{104}$$

Setting $\rho = \frac{G_f \sqrt{T}}{D_{\mathbf{z}} \sqrt{\lambda_{\min}^{\mathbf{A}} \lambda_{\max}^{\mathbf{B}}}}$ and using Lemma 4 give (98). ■

The following theorem shows that R^c has a constant bound when assuming $\|\mathbf{y}\|_2 \leq D^2$.

Theorem 9 *Let the sequences $\{\mathbf{x}_t, \mathbf{z}_t, \mathbf{y}_t\}$ be generated by OADM with $\eta = 0$. Let Assumption 4 hold. Assuming $\|\mathbf{y}_t\|_2 \leq D^2$ and setting $\rho = \frac{G_f \sqrt{T}}{D_{\mathbf{z}} \sqrt{\lambda_{\min}^{\mathbf{A}} \lambda_{\max}^{\mathbf{B}}}}$, we have*

$$R^c(T) \leq \frac{2D_{\mathbf{z}}^2 \lambda_{\min}^{\mathbf{A}} \lambda_{\max}^{\mathbf{B}}}{G_f^2} (D^2 + \frac{G_f^2}{\lambda_{\min}^{\mathbf{A}}}). \tag{105}$$

Proof: Let f be f_t in (9). Define

$$f'_t(\mathbf{x}_{t+1}) = -(\mathbf{A}^T \mathbf{y}_t + \rho \mathbf{A}^T (\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_t - \mathbf{c})). \tag{106}$$

Multiplying both sides by $(\mathbf{A}^T)^{-1}$ gives

$$(\mathbf{A}^T)^{-1} f'_t(\mathbf{x}_{t+1}) = -(\mathbf{y}_t + \rho (\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_t - \mathbf{c})). \tag{107}$$

Rearranging the terms, we have

$$\begin{aligned}
\|\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_t - \mathbf{c}\|_2^2 &= \frac{1}{\rho^2} \|\mathbf{y}_t + (\mathbf{A}^T)^{-1} f'_t(\mathbf{x}_{t+1})\|_2^2 \\
&\leq \frac{2}{\rho^2} (\|\mathbf{y}_t\|_2^2 + \|(\mathbf{A}^T)^{-1} f'_t(\mathbf{x}_{t+1})\|_2^2) \\
&\leq \frac{2}{\rho^2} (D^2 + \frac{G_f^2}{\lambda_{\min}^{\mathbf{A}}}) .
\end{aligned} \tag{108}$$

Summing over t from 1 to T and setting $\rho = \frac{G_f \sqrt{T}}{D_{\mathbf{z}} \sqrt{\lambda_{\min}^{\mathbf{A}} \lambda_{\max}^{\mathbf{B}}}}$, we have (105) according to Lemma 2. \blacksquare

Without requiring an additional Bregman divergence, R_2 achieves the same \sqrt{T} bound as R_1 . While R_1 depends on \mathbf{x}_t which may not stay in the feasible set, R_2 is defined on $\hat{\mathbf{x}}_t$ which always satisfies the equality constraint. The corresponding algorithm requires finding $\hat{\mathbf{x}}_t$ in each iteration such that $\mathbf{A}\hat{\mathbf{x}}_t = \mathbf{c} - \mathbf{B}\mathbf{z}_t$, which involves solving a linear system. The algorithm will be efficient in some settings, e.g., consensus optimization where $\mathbf{A} = \mathbf{I}$.

5.2 Strongly Convex Functions

If $g(\mathbf{z})$ is a β_2 -strongly convex function given in (82), we show that R_2 and R^c have logarithmic bounds.

Theorem 10 *Let $\eta = 0$ in OADM. Assume that $g(\mathbf{z})$ is β_2 -strongly convex and Assumption 3 and 4 hold. Setting $\rho_t = \beta_2 t / \lambda_{\max}^{\mathbf{B}}$, we have*

$$R_2(T) \leq \frac{G_f^2 \lambda_{\max}^{\mathbf{B}}}{2 \lambda_{\min}^{\mathbf{A}} \beta_2} (\log(T+1)) + \beta_2 D_{\mathbf{z}}^2 , \tag{109}$$

$$R^c(T) \leq \lambda_{\max}^{\mathbf{B}} D_{\mathbf{z}}^2 + \frac{2F \lambda_{\max}^{\mathbf{B}}}{\beta_2} \log(T+1) . \tag{110}$$

Proof: Assuming $g(\mathbf{z})$ is strongly convex (82), we can show the regret bound by replacing ρ by ρ_{t+1} and subtracting the strongly convex term $\frac{\beta_2}{2} \|\mathbf{z}^* - \mathbf{z}_{t+1}\|_2^2$ in (101), i.e.,

$$\begin{aligned}
f_t(\hat{\mathbf{x}}_t) + g(\mathbf{z}_{t+1}) - (f_t(\mathbf{x}^*) + g(\mathbf{z}^*)) &\leq \frac{1}{2 \lambda_{\min}^{\mathbf{A}} \rho_{t+1}} \|f'_t(\hat{\mathbf{x}}_t)\|_2^2 + \frac{1}{2 \rho_{t+1}} (\|\mathbf{y}_t\|_2^2 - \|\mathbf{y}_{t+1}\|_2^2) \\
&+ \frac{\rho_{t+1}}{2} (\|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_t\|_2^2 - \|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_{t+1}\|_2^2) - \frac{\beta_2}{2} \|\mathbf{z}^* - \mathbf{z}_{t+1}\|_2^2 .
\end{aligned} \tag{111}$$

Summing over t from 1 to T , we have

$$\begin{aligned}
R_2(T) &\leq \frac{G_f^2}{2 \lambda_{\min}^{\mathbf{A}}} \sum_{t=1}^T \frac{1}{\rho_{t+1}} + \sum_{t=1}^T \frac{1}{2 \rho_{t+1}} (\|\mathbf{y}_t\|_2^2 - \|\mathbf{y}_{t+1}\|_2^2) \\
&+ \sum_{t=1}^T \left[\frac{\rho_{t+1}}{2} (\|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_t\|_2^2 - \|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_{t+1}\|_2^2) - \frac{\beta_2}{2} \|\mathbf{z}^* - \mathbf{z}_{t+1}\|_2^2 \right] .
\end{aligned} \tag{112}$$

Using (89), (90) and setting $\rho_t = \beta_2 t / \lambda_{\max}^{\mathbf{B}}$, we get (109) by applying $\sum_{t=1}^T \frac{1}{t+1} \leq \log(T+1)$.

Now we prove (110). Replacing ρ by ρ_{t+1} in (103), we have

$$\|\mathbf{Ax}_{t+1} + \mathbf{Bz}_t - \mathbf{c}\|_2^2 \leq \frac{2F}{\rho_{t+1}} + \frac{1}{\rho_{t+1}^2} (\|\mathbf{y}_t\|_2^2 - \|\mathbf{y}_{t+1}\|_2^2) + (\|\mathbf{Bz}^* - \mathbf{Bz}_t\|_2^2 - \|\mathbf{Bz}^* - \mathbf{Bz}_{t+1}\|_2^2). \quad (113)$$

Letting the assumptions hold and summing over t from 0 to T , we have

$$\begin{aligned} & \sum_{t=1}^T \|\mathbf{Ax}_{t+1} + \mathbf{Bz}_t - \mathbf{c}\|_2^2 \\ & \leq 2F \sum_{t=1}^T \frac{1}{\rho_{t+1}} + \sum_{t=1}^T \frac{1}{\rho_{t+1}^2} (\|\mathbf{y}_t\|_2^2 - \|\mathbf{y}_{t+1}\|_2^2) + (\|\mathbf{Bz}^* - \mathbf{Bz}_1\|_2^2 - \|\mathbf{Bz}^* - \mathbf{Bz}_{T+1}\|_2^2) \\ & \leq 2F \sum_{t=1}^T \frac{1}{\rho_{t+1}} + \lambda_{\max}^{\mathbf{B}} D_{\mathbf{z}}^2. \end{aligned} \quad (114)$$

We use (89) in the last inequality. Setting $\rho_t = \beta_2 t / \lambda_{\max}^{\mathbf{B}}$ and using Lemma 4 give (110). \blacksquare

Similar as the case of general convex functions, the logarithmic regret bound for constraint violation can also be reduced to a constant bound, as shown in the following theorem.

Theorem 11 *Let $\eta = 0$ in OADM. Assume that $g(\mathbf{z})$ is β_2 -strongly convex and Assumption 4 hold. Assuming $\|\mathbf{y}_t\|_2 \leq D$ and setting $\rho_t = \beta_2 t / \lambda_{\max}^{\mathbf{B}}$, we have*

$$R^c(T) \leq \frac{\pi \lambda_{\max}^{\mathbf{B}}{}^2}{3\beta_2^2} \left(D^2 + \frac{G_f^2}{\lambda_{\min}^{\mathbf{A}}} \right) \quad (115)$$

Proof: Setting $\rho_t = \beta_2 t / \lambda_{\max}^{\mathbf{B}}$ in (108), summing over t from 1 to T and using $\sum_{t=1}^T \frac{1}{t^2} \leq \frac{\pi}{6}$ complete the proof. \blacksquare

Theorem 11 shows that OADM can achieve the logarithmic regret bound without requiring f_t to be strongly convex, which is in line with other online learning algorithms for composite objectives.

6 Further Discussions

In this section, we discuss several variants of the \mathbf{x} update in OADM which can lead to efficient updates and show the stochastic convergence rates. The connection to the related work is presented. We also show that OADM can serve as projection-free online learning.

6.1 Inexact ADMM Updates ($\eta > 0$)

In OADM ($\eta > 0$), since the \mathbf{x} update (62) involves the function f_t , the quadratic penalty term and a Bregman divergence, it may be computationally expensive to solve it exactly. We consider several variants which solve the \mathbf{x} update inexactly through the linearization of some terms. The inexact updates can be efficient, and include mirror descent algorithm (MDA) and composite objective mirror descent (COMID) as special cases.

Case 1: Linearization of the quadratic penalty term The linearization of the quadratic penalty term in (62) can be done by removing $\|\mathbf{Ax}\|_2^2$ as follows:

$$\|\mathbf{Ax} + \mathbf{Bz}_t - \mathbf{c}\|_2^2 - \|\mathbf{A}(\mathbf{x} - \mathbf{x}_t)\|_2^2 = 2\langle \mathbf{Ax}_t + \mathbf{Bz}_t - \mathbf{c}, \mathbf{Ax} \rangle + \|\mathbf{Bz}_t - \mathbf{c}\|_2^2 - \|\mathbf{Ax}_t\|_2^2.$$

Let $B_\phi(\mathbf{x}, \mathbf{x}_t) = B_\varphi(\mathbf{x}, \mathbf{x}_t) - \frac{\rho}{2\eta}\|\mathbf{A}(\mathbf{x} - \mathbf{x}_t)\|_2^2$ in (62), where B_φ is a Bregman divergence and the quadratic term is used to linearize the quadratic penalty term. Removing constant terms, (62) becomes

$$\mathbf{x}_{t+1} = \operatorname{argmin}_{\mathbf{x}} f_t(\mathbf{x}) + \langle \mathbf{y}_t + \rho(\mathbf{Ax}_t + \mathbf{Bz}_t - \mathbf{c}), \mathbf{Ax} \rangle + \eta B_\varphi(\mathbf{x}, \mathbf{x}_t). \quad (116)$$

This case mainly solves the problem caused by \mathbf{A} , e.g., \mathbf{Ax} makes \mathbf{x} nonseparable. Several problems have been benefited from the linearization of quadratic term [22], e.g., f is ℓ_1 loss function [39] and projection onto the unit simplex or ℓ_1 ball [25].

Since $B_\phi(\mathbf{x}, \mathbf{x}_t) \geq \frac{\alpha}{2}\|\mathbf{x} - \mathbf{x}_t\|_2^2$ is required for the analysis in Section 4, B_φ should be chosen to satisfy that condition. Note

$$B_\phi(\mathbf{x}, \mathbf{x}_t) = B_\varphi(\mathbf{x}, \mathbf{x}_t) - \frac{\rho}{2\eta}\|\mathbf{A}(\mathbf{x} - \mathbf{x}_t)\|_2^2 \geq B_\varphi(\mathbf{x}, \mathbf{x}_t) - \frac{\rho\lambda_{\max}^{\mathbf{A}}}{2\eta}\|\mathbf{x} - \mathbf{x}_t\|_2^2. \quad (117)$$

Therefore, as long as $B_\varphi(\mathbf{x}, \mathbf{x}_t) \geq \frac{\rho\lambda_{\max}^{\mathbf{A}}/\eta + \alpha}{2}\|\mathbf{x} - \mathbf{x}_t\|_2^2$, the assumption 3(b) holds, meaning Theorem 4 and 6 hold for Case 1.

Case 2: Linearization of function f_t This case is particularly useful when the difficulty of solving (62) is caused by $f_t(\mathbf{x})$, e.g., when f_t is a logistic loss function. Linearizing the function f_t at \mathbf{x}_t in (62), we have

$$\mathbf{x}_{t+1} = \operatorname{argmin}_{\mathbf{x}} \langle f'_t(\mathbf{x}_t), \mathbf{x} - \mathbf{x}_t \rangle + \frac{\rho}{2}\|\mathbf{Ax} + \mathbf{Bz}_t - \mathbf{c}\|_2^2 + \eta B_\phi(\mathbf{x}, \mathbf{x}_t). \quad (118)$$

The updated is called inexact ADMM update if ϕ is a quadratic function [11]. In the Appendix A, we show Theorem 4 and 6 continue to hold in this case.

Case 3: Mirror Descent In this case, we linearize both the function and the quadratic term, which can be done by choosing $B_\phi(\mathbf{x}, \mathbf{x}_t) = B_\varphi(\mathbf{x}, \mathbf{x}_t) - \frac{\rho}{2\eta}\|\mathbf{A}(\mathbf{x} - \mathbf{x}_t)\|_2^2$ in Case 2. Combining the results in Case 1 and 2, (62) becomes the following MDA-type update:

$$\mathbf{x}_{t+1} = \operatorname{argmin}_{\mathbf{x}} \langle F_t(\mathbf{x}_t), \mathbf{x} \rangle + \eta B_\varphi(\mathbf{x}, \mathbf{x}_t), \quad (119)$$

where $F_t(\mathbf{x}_t) = f'_t(\mathbf{x}_t) + \mathbf{A}^T\{\mathbf{y}_t + \rho(\mathbf{Ax}_t + \mathbf{Bz}_t - \mathbf{c})\}$, which is the gradient of the objective in (62). Assuming $B_\varphi(\mathbf{x}, \mathbf{x}_t) \geq \frac{\rho\lambda_{\max}^{\mathbf{A}}/\eta + \alpha}{2}\|\mathbf{x} - \mathbf{x}_t\|_2^2$ in Case 2, the regret bounds in Theorem 4 and 6 still holds in Case 3.

Case 4: COMID Assume f_t is a composite objective consisting of smooth and nonsmooth part, i.e., $f_t(\mathbf{x}) = f_t^S(\mathbf{x}) + f_t^N(\mathbf{x})$, where f_t^S is the smooth part and f_t^N is the nonsmooth part. Let $B_\phi(\mathbf{x}, \mathbf{x}_t) = B_\varphi(\mathbf{x}, \mathbf{x}_t) - \frac{\rho}{2\eta}\|\mathbf{A}(\mathbf{x} - \mathbf{x}_t)\|_2^2$, which is used to linearize the quadratic penalty term. Linearizing the smooth function f_t^S , (62) becomes the following COMID-type update:

$$\mathbf{x}_{t+1} = \operatorname{argmin}_{\mathbf{x}} f_t^N(\mathbf{x}) + \langle F_t^S(\mathbf{x}_t), \mathbf{x} \rangle + \eta B_\varphi(\mathbf{x}, \mathbf{x}_t), \quad (120)$$

where $F_t^S(\mathbf{x}_t) = \nabla f_t^S(\mathbf{x}_t) + \mathbf{A}^T\{\mathbf{y}_t + \rho(\mathbf{Ax}_t + \mathbf{Bz}_t - \mathbf{c})\}$. Applying the analysis in Case 2 on the smooth part, we can get the regret bounds in Theorem 4 and 6.

6.2 Stochastic Convergence Rates

In this section, we present the convergence rates for ADMM in the Case 2-4 in Section 6.1 in the stochastic setting, which solves the following **stochastic learning** problem:

$$\min_{\mathbf{x} \in \mathcal{X}, \mathbf{z} \in \mathcal{Z}} \mathbf{E}_\xi [f(\mathbf{x}, \xi)] + g(\mathbf{z}) \quad \text{s.t.} \quad \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} = \mathbf{c} \quad (121)$$

$f'(\mathbf{x}_t, \xi_t)$ is an unbiased estimate of $f'(\mathbf{x}_t)$ and $f(\mathbf{x}) = \mathbf{E}f(\mathbf{x}, \xi)$. Correspondingly, the \mathbf{x} -update in (118)-(119) uses $f'(\mathbf{x}_t, \xi_t)$ to substitute $f'_t(\mathbf{x}_t)$ and $\nabla f^N(\mathbf{x}_t, \xi_t)$ to substitute $\nabla f_t^N(\mathbf{x}_t)$ in (120). The regret bounds for Case 2-4 in Section 6.1 can be converted to convergence rates in the stochastic setting based on known online-stochastic conversion [15, 27, 69]. More specifically, the stochastic convergence rates in expectation can be obtained by simply dividing regret bounds by T . Using martingale concentration results [15, 27, 69], the high probability bounds can also be obtained by applying the Azuma-Hoeffding inequality [2].

Corollary 1 *Let the sequences $\{\mathbf{x}_t, \mathbf{z}_t, \mathbf{y}_t\}$ be generated by stochastic ADM and Assumption 3 hold. Let $\bar{\mathbf{x}}_T = \frac{1}{T} \sum_{t=1}^T \mathbf{x}_t$ and $\bar{\mathbf{z}}_T = \frac{1}{T} \sum_{t=1}^T \mathbf{z}_t$. For any $\mathbf{x}^*, \mathbf{z}^*$ satisfying $\mathbf{A}\mathbf{x}^* + \mathbf{B}\mathbf{z}^* = \mathbf{c}$, setting $\eta = \frac{G_f \sqrt{T}}{D_{\mathbf{x}} \sqrt{2\alpha}}$ and $\rho = \sqrt{T}$, we have*

(a) *Stochastic convergence rates in expectation*

$$\mathbf{E} [f(\bar{\mathbf{x}}_T) + g(\bar{\mathbf{z}}_T)] - (f(\mathbf{x}^*) + g(\mathbf{z}^*)) \leq \frac{\lambda_{\max}^{\mathbf{B}} D_{\mathbf{z}}^2}{2\sqrt{T}} + \frac{\sqrt{2} G_f D_{\mathbf{x}}}{\sqrt{\alpha} \sqrt{T}}, \quad (122)$$

$$\mathbf{E} [\|\mathbf{A}\bar{\mathbf{x}}_T + \mathbf{B}\bar{\mathbf{z}}_T + \mathbf{c}\|_2^2] \leq \frac{\lambda_{\max}^{\mathbf{B}} D_{\mathbf{z}}^2}{T} + \frac{2\sqrt{2} D_{\mathbf{x}} G_f}{\sqrt{\alpha} T} + \frac{2F}{\sqrt{T}}. \quad (123)$$

(b) *High probability bounds for stochastic convergence rates*

$$P(f(\bar{\mathbf{x}}_T) + g(\bar{\mathbf{z}}_T) - (f(\mathbf{x}^*) + g(\mathbf{z}^*)) \geq \frac{\lambda_{\max}^{\mathbf{B}} D_{\mathbf{z}}^2}{2\sqrt{T}} + \frac{\sqrt{2} G_f D_{\mathbf{x}}}{\sqrt{\alpha} \sqrt{T}} + \varepsilon) \leq \exp \left(-\frac{T \alpha \varepsilon^2}{16 D_{\mathbf{x}}^2 G_f^2} \right), \quad (124)$$

$$P(\|\mathbf{A}\bar{\mathbf{x}}_T + \mathbf{B}\bar{\mathbf{z}}_T - \mathbf{c}\|_2^2 \geq \frac{2F}{\sqrt{T}} + \frac{\lambda_{\max}^{\mathbf{B}} D_{\mathbf{z}}^2}{T} + \frac{2\sqrt{2} D_{\mathbf{x}} G_f}{\sqrt{\alpha} T} + \varepsilon) \leq \exp \left(-\frac{T \alpha \varepsilon^2}{16 D_{\mathbf{x}}^2 G_f^2} \right). \quad (125)$$

The proof is presented in Appendix B. Compared to the stochastic convergence rates for COMID [26], the stochastic convergence rates for the objective of ADM has an extra term $\frac{\lambda_{\max}^{\mathbf{B}} D_{\mathbf{z}}^2}{2\sqrt{T}}$ which bounds the splitting variable \mathbf{z} . For strongly convex functions, we have $O(\frac{\log T}{T})$ stochastic convergence rates by applying the online-stochastic conversion [15, 27, 69] on Theorem 6.

Remark 1 We note that [59] has recently established the stochastic convergence rates for stochastic ADM based on our VI analysis (see Section 2.3), which has the following form in our notation:

$$\mathbf{E} [f(\bar{\mathbf{x}}_T) + g(\bar{\mathbf{z}}_T) - (f(\mathbf{x}^*) + g(\mathbf{z}^*)) + D \|\mathbf{A}\bar{\mathbf{x}}_T + \mathbf{B}\bar{\mathbf{z}}_T + \mathbf{c}\|_2^2] \leq \frac{\lambda_{\max}^{\mathbf{B}} D_{\mathbf{z}}^2 \rho}{2T} + \frac{\sqrt{2} G_f D_{\mathbf{x}}}{\sqrt{T}} + \frac{D^2}{2\rho T}, \quad (126)$$

where $\|\mathbf{y}_t\|_2 \leq D$ (see Assumption 2). The bound in (126) depends on D^2 , which usually is large (see Eq. (58)) and thus worse than our results which do not rely on D^2 . As a matter of fact, we can show the term D^2 can be safely removed (setting $\alpha = 1$ in (162) in Appendix B), i.e.,

$$\mathbf{E} [f(\bar{\mathbf{x}}_T) + g(\bar{\mathbf{z}}_T) - (f(\mathbf{x}^*) + g(\mathbf{z}^*)) + \frac{\rho}{2} \|\mathbf{A}\bar{\mathbf{x}}_T + \mathbf{B}\bar{\mathbf{z}}_T + \mathbf{c}\|_2^2] \leq \frac{\lambda_{\max}^{\mathbf{B}} D_{\mathbf{z}}^2 \rho}{2T} + \frac{\sqrt{2} G_f D_{\mathbf{x}}}{\sqrt{T}}. \quad (127)$$

However, since $\mathbf{x}_t, \mathbf{z}_t$ are not feasible, $f(\bar{\mathbf{x}}_T) + g(\bar{\mathbf{z}}_T) - (f(\mathbf{x}^*) + g(\mathbf{z}^*))$ may be negative. As a result, (126) or (127) may not imply an $O(1/T)$ convergence rate for the equality constraint, in contrast to (123) in Corollary 1. Furthermore, if assuming $\|\mathbf{y}_t\|_2 \leq D$, the residual of equality constraint has an $O(1/T)$ convergence rate by dividing by T on both sides of (79) in Theorem 5 and using the Jensen's inequality.

6.3 Connections to Related Work ($\eta = 0$)

Assume $\eta = 0$, $\mathbf{A} = \mathbf{I}$, $\mathbf{B} = -\mathbf{I}$, $\mathbf{c} = \mathbf{0}$, thus $\mathbf{x} = \mathbf{z}$. Hence, the online optimization problem has the form which is the same as the ones considered in the development of **FOBOS** [27] and RDA [69]. The three steps of **OADM** ($\eta = 0$) reduce to

$$\mathbf{x}_{t+1} = \operatorname{argmin}_{\mathbf{x}} \{f_t(\mathbf{x}) + \langle \mathbf{y}_t, \mathbf{x} - \mathbf{z}_t \rangle + \frac{\rho}{2} \|\mathbf{x} - \mathbf{z}_t\|_2^2\}, \quad (128)$$

$$\mathbf{z}_{t+1} = \operatorname{argmin}_{\mathbf{z}} \{g(\mathbf{z}) + \langle \mathbf{y}_t, \mathbf{x}_{t+1} - \mathbf{z} \rangle + \frac{\rho}{2} \|\mathbf{x}_{t+1} - \mathbf{z}\|_2^2\}, \quad (129)$$

$$\mathbf{y}_{t+1} = \mathbf{y}_t + \rho(\mathbf{x}_{t+1} - \mathbf{z}_{t+1}). \quad (130)$$

Let $f'_t(\mathbf{x}_{t+1}) \in \partial f_t(\mathbf{x})$, $g'(\mathbf{z}_{t+1}) \in \partial g(\mathbf{z})$. The first order optimality conditions for (128) and (129) give

$$\begin{aligned} f'_t(\mathbf{x}_{t+1}) + \mathbf{y}_t + \rho(\mathbf{x}_{t+1} - \mathbf{z}_t) &= \mathbf{0}, \\ g'(\mathbf{z}_{t+1}) - \mathbf{y}_t - \rho(\mathbf{x}_{t+1} - \mathbf{z}_{t+1}) &= \mathbf{0}. \end{aligned}$$

Adding them together yields

$$\mathbf{z}_{t+1} = \mathbf{z}_t - \frac{1}{\rho}(f'_t(\mathbf{x}_{t+1}) + g'(\mathbf{z}_{t+1})). \quad (131)$$

OADM can be considered as taking the implicit subgradient of f_t and g at the yet to be determined \mathbf{x}_{t+1} and \mathbf{z}_{t+1} . FOBOS has the following update [27]:

$$\mathbf{z}_{t+1} = \mathbf{z}_t - \frac{1}{\rho}(f'_t(\mathbf{z}_t) + g'(\mathbf{z}_{t+1})).$$

FOBOS takes the explicit subgradient of f_t at current \mathbf{z}_t . In fact, FOBOS can be considered as a variant of OADM, which linearizes the objective of (128) at \mathbf{z}_t :

$$\mathbf{x}_{t+1} = \operatorname{argmin}_{\mathbf{x}} \langle f'_t(\mathbf{z}_t) + \mathbf{y}_t, \mathbf{x} - \mathbf{z}_t \rangle + \frac{\rho}{2} \|\mathbf{x} - \mathbf{z}_t\|_2^2.$$

It has a closed-form solution, i.e., $\mathbf{x}_{t+1} = \mathbf{z}_t - \frac{1}{\rho}(f'_t(\mathbf{z}_t) + \mathbf{y}_t)$. Denote $\mathbf{z}_{t+\frac{1}{2}} = \mathbf{x}_{t+1} + \frac{1}{\rho}\mathbf{y}_t$, then

$$\mathbf{z}_{t+\frac{1}{2}} = \mathbf{z}_t - \frac{1}{\rho}f'_t(\mathbf{z}_t). \quad (132)$$

(129) is equivalent to the following form:

$$\mathbf{z}_{t+1} = \operatorname{argmin}_{\mathbf{z}} g(\mathbf{z}) + \frac{\rho}{2} \|\mathbf{z} - \mathbf{z}_{t+\frac{1}{2}}\|_2^2. \quad (133)$$

(132) and (133) form the updates of FOBOS [27]. Furthermore, if $g(\mathbf{z})$ is an indicator function of a convex set Ω , substituting (132) into (133), we have

$$\mathbf{z}_{t+1} = \operatorname{argmin}_{\mathbf{z} \in \Omega} \frac{\rho}{2} \|\mathbf{z}_t - \frac{1}{\tau}f'_t(\mathbf{z}_t) - \mathbf{z}\|_2^2 = \mathcal{P}_{\mathbf{z} \in \Omega} \left[\mathbf{z}_t - \frac{1}{\tau}f'_t(\mathbf{z}_t) \right],$$

and we recover projected gradient descent [40].

6.4 Projection-free Online Learning

For an online constrained optimization problem, the state-of-the-art methods like OGD, FOBOS and RDA require a full projection onto the constraint set at each round. In many cases, e.g., an intersection of simple constraints, the full projection can be done by alternating projecting onto simple constraints cyclically [14]. In OADM, we can decompose functions and constraints into simpler subproblems by introducing appropriate splitting variables. If the subproblem for each splitting variable is simple enough to yield efficient projection, the full projection onto the whole constraint set can be done by projections onto simple constraints at each round along with the long term equality constraints. Therefore, OADM and its variants can avoid the full projection at each round. Consider the full projection onto $\mathcal{X} \times \mathcal{Z}$, which in general requires alternating projection onto \mathcal{X} and \mathcal{Z} at each round in OGD, FOBOS and RDA. In OADM, by introducing equality constraint $\mathbf{x} = \mathbf{z}$, the constraint set is split into two parts and $\mathbf{x} \in \mathcal{X}$ and $\mathbf{z} \in \mathcal{Z}$. At each round, the primal updates in OADM and its variants project \mathbf{x}, \mathbf{z} onto \mathcal{X}, \mathcal{Z} separately. In the long run, the equality constraint will be satisfied in expectation, thus \mathbf{x} is a feasible solution. Hence, OADM can be considered as a projection-free online learning algorithm.

In [41], the Frank-Wolfe algorithm is used as a projection-free online learning algorithm, which solves a linear optimization at each round and has $O(T^{3/4})$ regret bound. It assumes linear optimization can be done efficiently in the constraint set. Realizing that solving a linear optimization still requires an inner loop algorithm, the authors pose an open problem whether the optimal regret bound can be achieved by performing one iteration of linear-optimization.

We now show how OADM does projection-free online learning with linear constraints, which includes linear programming and quadratic programming as special cases. Formally, we consider the problem

$$\min_{\mathbf{x}} \sum_{t=1}^T f_t(\mathbf{x}) \quad \text{s.t.} \quad \mathbf{Ax} = \mathbf{a}, \mathbf{Bx} \leq \mathbf{b}. \quad (134)$$

In the setting of OADM, we first introduce an auxiliary variable $\mathbf{z} = \mathbf{Bx}$ to separate inequality constraint from equality constraint. Then (134) can be rewritten as:

$$\min_{\mathbf{x}, \mathbf{z}} \sum_{t=1}^T f_t(\mathbf{x}) + g(\mathbf{z}) \quad \text{s.t.} \quad \mathbf{Ax} = \mathbf{a}, \mathbf{Bx} = \mathbf{z}, \quad (135)$$

where $g(\mathbf{z})$ is the indicator function of box constraint $\mathbf{z} \leq \mathbf{b}$. The augmented Lagrangian for (135) is as follows:

$$\begin{aligned} L_\rho(\mathbf{x}, \mathbf{z}, \mathbf{u}, \mathbf{v}) &= f_t(\mathbf{x}) + g(\mathbf{z}) + \langle \mathbf{u}, \mathbf{Ax} - \mathbf{a} \rangle + \langle \mathbf{v}, \mathbf{Bx} - \mathbf{z} \rangle \\ &\quad + \frac{\rho_{\mathbf{u}}}{2} \|\mathbf{Ax} - \mathbf{a}\|_2^2 + \frac{\rho_{\mathbf{v}}}{2} \|\mathbf{Bx} - \mathbf{z}\|_2^2, \end{aligned} \quad (136)$$

where \mathbf{u}, \mathbf{v} are dual variables and the penalty parameters $\rho_{\mathbf{u}}, \rho_{\mathbf{v}} > 0$. Let the Bregman divergence in the \mathbf{x} update in (62) be the quadratic function. We have the following OADM updates for (135):

$$\mathbf{x}_{t+1} = \underset{\mathbf{x}}{\operatorname{argmin}} \left\{ f_t(\mathbf{x}) + \langle \mathbf{u}_t, \mathbf{A}\mathbf{x} - \mathbf{a} \rangle + \langle \mathbf{v}_t, \mathbf{B}\mathbf{x} - \mathbf{z}_t \rangle + \frac{\rho_u}{2} \|\mathbf{A}\mathbf{x} - \mathbf{a}\|_2^2 + \frac{\rho_v}{2} \|\mathbf{B}\mathbf{x} - \mathbf{z}_t\|_2^2 + \frac{\eta}{2} \|\mathbf{x} - \mathbf{x}_t\|_2^2 \right\}, \quad (137)$$

$$\mathbf{z}_{t+1} = \underset{\mathbf{z} \leq \mathbf{b}}{\operatorname{argmin}} \left\{ \langle \mathbf{v}_t, \mathbf{B}\mathbf{x}_{t+1} - \mathbf{z} \rangle + \frac{\rho_v}{2} \|\mathbf{B}\mathbf{x}_{t+1} - \mathbf{z}\|_2^2 \right\}, \quad (138)$$

$$\mathbf{u}_{t+1} = \mathbf{u}_t + \rho_u (\mathbf{A}\mathbf{x}_{t+1} - \mathbf{a}), \quad (139)$$

$$\mathbf{v}_{t+1} = \mathbf{v}_t + \rho_v (\mathbf{B}\mathbf{x}_{t+1} - \mathbf{z}_{t+1}), \quad (140)$$

where $\eta \geq 0$. The \mathbf{x} -update has a closed-form solution when f_t is a linear or quadratic functions, or the ℓ_1 norm. If the \mathbf{x} -update does not have a closed-form solution, we can linearize f_t at \mathbf{x}_t as in Section 6.1, which leads to a closed-form solution. Further, the \mathbf{z} -update has a closed-form solution of the following form:

$$\mathbf{z}_{t+1} = \min\{\mathbf{B}\mathbf{x}_{t+1} + \mathbf{y}_t/\rho, \mathbf{b}\}. \quad (141)$$

Thus, OADM gives a projection-free online algorithm for optimization problems under linear constraints, e.g., linear and quadratic programming. In contrast, state-of-the-art online learning algorithms require the projection onto the constraints at each round, which amounts to solving a linear or quadratic program [41].

7 Experimental Results

In this section, we use OADM to solve generalized lasso problems [11], including lasso [67] and total variation (TV) problem [64]. We present simulation results to show the convergence of the objective as well as constraints in OADM. We also compare it with batch ADM and two other online learning algorithms: FO-BOS [27] and regularized dual averaging (RDA) [69] in selecting sparse dimension in lasso and recovering data in total variation.

7.1 Generalized Lasso

The generalized lasso problem is formulated as follows:

$$\min_{\mathbf{x}} \frac{1}{N} \sum_{t=1}^N \|\mathbf{a}_t \mathbf{x} - b_t\|_2^2 + \lambda \|\mathbf{D}\mathbf{x}\|_1, \quad (142)$$

where $\mathbf{a}_t \in \mathbb{R}^{1 \times n}$, $\mathbf{x} \in \mathbb{R}^{n \times 1}$, $\mathbf{D} \in \mathbb{R}^{m \times n}$ and b_t is a scalar. If $\mathbf{D} = \mathbf{I}$, (142) yields the lasso. If \mathbf{D} is an upper bidiagonal matrix with diagonal 1 and off-diagonal -1 , (142) becomes the problem of total variation. The ADM form of (142) is:

$$\min_{\mathbf{D}\mathbf{x}=\mathbf{z}} \frac{1}{N} \sum_{t=1}^N \|\mathbf{a}_t \mathbf{x} - b_t\|_2^2 + \lambda \|\mathbf{z}\|_1, \quad (143)$$

where $\mathbf{z} \in \mathbb{R}^{m \times 1}$. The augmented Lagrangian at round t is

$$L_\rho = \|\mathbf{a}_t \mathbf{x} - b_t\|_2^2 + \lambda \|\mathbf{z}\|_1 + \langle \mathbf{y}, \mathbf{D}\mathbf{x} - \mathbf{z} \rangle + \frac{\rho}{2} \|\mathbf{D}\mathbf{x} - \mathbf{z}\|_2^2.$$

The three updates of OADM yield the following closed-form updates:

$$\mathbf{x}_{t+1} = (\mathbf{a}_t^T \mathbf{a}_t + \rho \mathbf{D}^T \mathbf{D} + \eta)^{-1} \mathbf{v}_t, \quad (144)$$

$$\mathbf{z}_{t+1} = S_{\lambda/\rho}(\mathbf{D} \mathbf{x}_{t+1} + \mathbf{u}_t), \quad (145)$$

$$\mathbf{u}_{t+1} = \mathbf{u}_t + \mathbf{D} \mathbf{x}_{t+1} - \mathbf{z}_{t+1}, \quad (146)$$

where $\mathbf{u} = \mathbf{y}/\rho$, $\mathbf{v}_t = \mathbf{a}_t^T \mathbf{b}_t + \rho \mathbf{D}^T (\mathbf{z}_t - \mathbf{u}_t) + \eta \mathbf{x}_t$, and $S_{\lambda/\rho}$ denotes the soft thresholding operator or a shrinkage operator defined as

$$S_{\lambda/\rho}(k) = \begin{cases} k - \lambda/\rho, & k > \lambda/\rho \\ 0, & |k| \leq \lambda/\rho \\ k + \lambda/\rho, & k < -\lambda/\rho \end{cases}, \quad (147)$$

which is a simple element-wise operation.

For lasso, the x-update is

$$\mathbf{x}_{t+1} = (\mathbf{v}_t - (\eta + \rho + \mathbf{a}_t^T \mathbf{a}_t)^{-1} \mathbf{a}_t^T (\mathbf{a}_t \mathbf{v}_t)) / (\eta + \rho),$$

where the inverse term is a scalar. The multiplication terms take $O(n)$ flops [38]. Thus, the x-update can be done in $O(n)$ flops.

For total variation, we set $\eta = 0$ so that

$$\mathbf{x}_{t+1} = (\mathbf{Q} \mathbf{v}_t - (\rho + \mathbf{a}_t^T \mathbf{Q} \mathbf{a}_t)^{-1} \mathbf{Q} \mathbf{a}_t^T (\mathbf{a}_t \mathbf{Q} \mathbf{v}_t)) / \rho,$$

where $\mathbf{Q} = (\mathbf{D}^T \mathbf{D})^{-1}$. Since \mathbf{D} is a bidiagonal matrix, $\mathbf{Q} \mathbf{v}_t$ and $\mathbf{Q} \mathbf{a}_t$ can be done in $O(n)$ flops [38, 11]. The inverse term is scalar and other multiplication terms cost $O(n)$ flops. Overall, the x-update can be carried out in $O(n)$ flops.

In both cases, the three updates (144)-(146) can be done in $O(n)$ flops. In contrast, in batch ADM, the complexity of x-update could be as high as $O(n^3)$ or $O(n^2)$ by caching factorizations [11].

FOBOS and RDA cannot directly solve the TV term. We first reformulate the total variation in the lasso form such that

$$\min_{\mathbf{y}} \frac{1}{N} \sum_{t=1}^N \|\mathbf{a}_t \mathbf{D}^{-1} \mathbf{y} - \mathbf{b}\|_2^2 + \lambda \|\mathbf{y}\|_1, \quad (148)$$

where $\mathbf{y} = \mathbf{D} \mathbf{x}$. FOBOS and RDA can solve the above lasso problem and get \mathbf{y} . \mathbf{x} can be recovered by using $\mathbf{x} = \mathbf{D}^{-1} \mathbf{y}$.

7.2 Simulation

Our experiments mainly follow the lasso and total variation examples in [11],⁴ although we modified the code to accommodate our setup. We first randomly generated \mathbf{A} with N examples of dimensionality n . \mathbf{A} is then normalized along the columns. Then, a true \mathbf{x}_0 is randomly generated with certain sparsity pattern for lasso and TV. For lasso, we set the number of nonzeros (NNZs) k in \mathbf{x}_0 as 100, i.e., $k = 100$. For TV, we first set \mathbf{x}_0 to be a vector of ones, then randomly select some blocks of random size in \mathbf{x}_0 and reset their value to a random value from $[1, 10]$. \mathbf{b} is calculated by adding Gaussian noise to $\mathbf{A} \mathbf{x}_0 / N$. In all

⁴<http://www.stanford.edu/~boyd/papers/admm/>

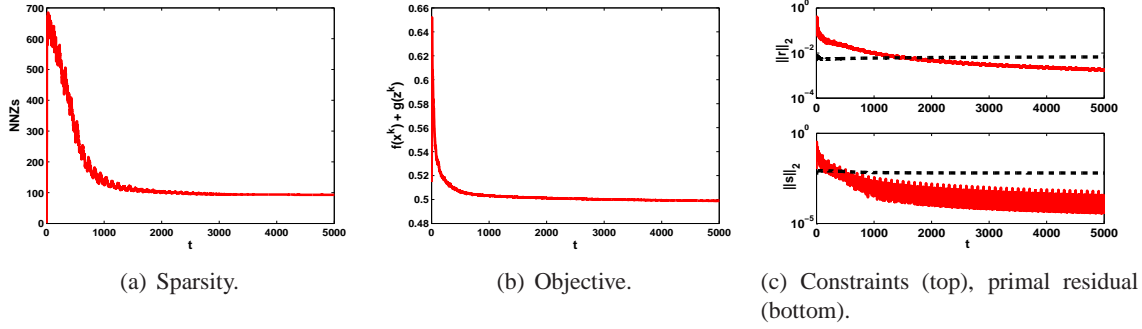


Figure 1: The convergence of sparsity, objective value and constraints for lasso in OADM with $q = 0.5, \rho = 1, \eta = t$.

experiments, $N = 100$, which facilitates the matrix inverse in ADM. For lasso, we try different combination of parameters from $n = [1000, 5000]$, $\rho = [0.1, 1, 10]$ and $q = [0.1, 0.5]$ for $\lambda = q \times |\mathbf{A}^T \mathbf{b} / N|_\infty$. All experiments are implemented in Matlab.

Convergence: We go through the examples 100 times using OADM. Figure 1(a) shows that NNZs converge to a value close to the actual $k = 100$ before $t = 2000$. Figure 1(b) shows the convergence of objective value. In Figure 1(c), the dashed lines are the standard stopping criteria used in ADM [11]. Figure 1(c) shows that the equality constraint (top) and primal residual (bottom) are satisfied in the online setting. While the objective converges fast, the equality constraints take relatively more time to be satisfied.

Sparsity: We compare NNZs found by batch ADM and three online learning algorithms, including OADM, FOBOS, and RDA. We set $\eta = 1000$ for OADM and $\gamma = 1$ for RDA. For FOBOS, we use a time varying parameter $\rho_t = \rho / \sqrt{t}$. For online learning algorithms, we go through the examples 100 times. We run the experiment 20 times and the average results are plotted. We show the results for $q = 0.5$ in Figure 2, where n is 1000 for the first three figures (a)-(c) and 5000 for the last three. While ADM and RDA tend to give the sparsest results, OADM seems more conservative and converges to reasonably sparse solutions. Figure 2 shows OADM is closest to the actual NNZs 100. The NNZs in FOBOS is large and oscillates in a big range, which has also been observed in [69].

Total Variation: We compare the patterns found by the four algorithms. For all algorithms, $N = 100, n = 1000, \lambda = 0.001$ and ρ is chosen through cross validation. In RDA, $\gamma = 100$. Recall that $\eta = 0$ in OADM. While we use a fixed ρ for OADM and RDA, FOBOS uses $\rho_t = \rho / \sqrt{t}$. Figure 3 shows the three different patterns and results found by the algorithms. ADM seems to follow the pattern with oscillation. OADM is smoother and generally follows the trend of the patterns. For the first two examples, FOBOS works well and the patterns found by RDA tend to be flat. In the last example, both FOBOS and RDA oscillate.

8 Conclusions

In this paper, we first developed new proof techniques to analyze the convergence rate for ADM, which establishes a $O(1/T)$ convergence rate for the objective, the optimality conditions (constraints) and the variational inequality form of ADM. The new proof techniques may facilitate the improvement and modifications of ADM which is needed in some scenarios. For example, the quadratic penalty term in the \mathbf{x} and \mathbf{z} updates may not lead to efficient algorithm, while other Bregman divergences like KL divergence may

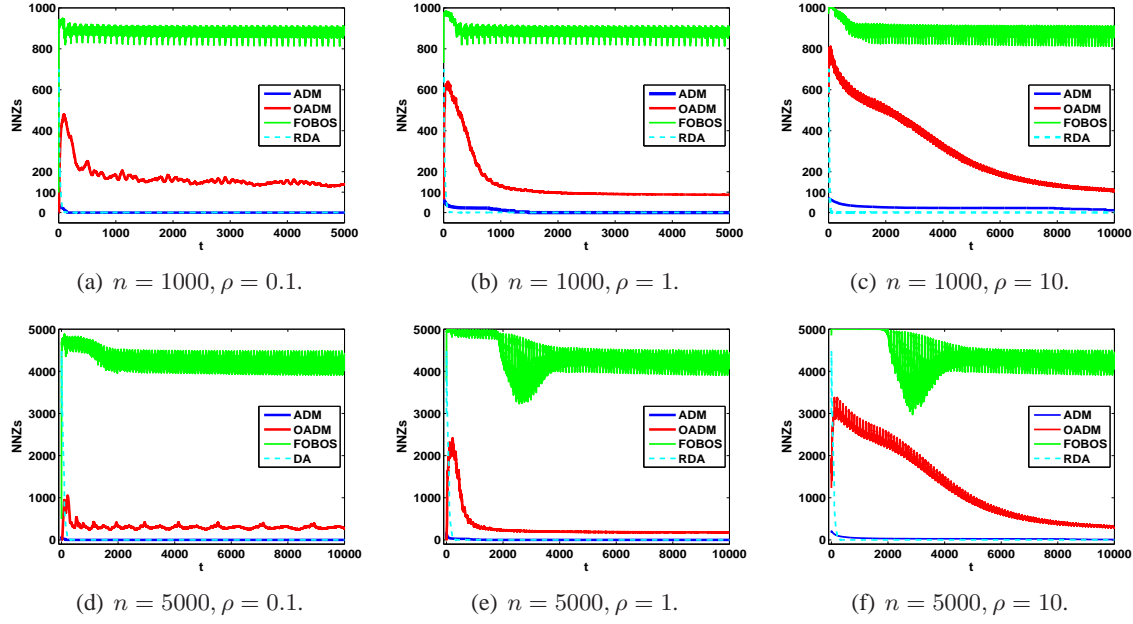


Figure 2: The NNZs found by OADM, ADM, FOBOS and RDA with $q = 0.5$ for lasso. OADM is closest to the actual NNZs.

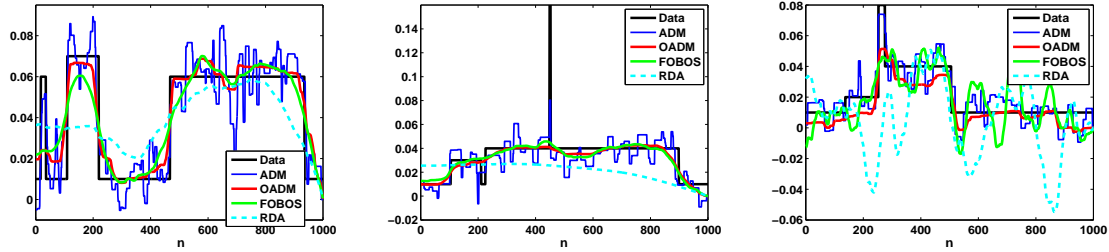


Figure 3: The TV patterns found by OADM, ADM, FOBOS and RDA. OADM is the best in recovering the patterns.

induce efficient updates.

We propose an efficient online learning algorithm named online ADM (OADM). Using the proof technique developed for batch ADM, we establish regret bounds for the objective and constraint violation for general and strongly convex functions in OADM. We also discuss inexact update to yield efficient \mathbf{x} update, including mirror descent and composite objective mirror descent. Finally, we illustrate the efficacy of OADM in solving lasso and total variation problems. Through splitting variables, we show OADM can do projection-free online learning with linear constraints. It would be interesting to explore whether OADM can do projection-free learning with other constraints. Through variables splitting, ADM has been successfully used in distributed optimization. If distributed ADM is extended to the online learning setting, distributed OADM will allow the data to be distributed along the time dimension, which can be particularly useful for spatio-temporal data.

Acknowledgment

The research was supported by NSF CAREER award IIS-0953274, and NSF grants IIS-0916750, IIS-0812183, and IIS-1029711. The authors thank the detailed and insightful comments from reviewers and extend out thanks to Daniel Boley and Stephen Wright for helpful discussion. The authors also thank Bing-sheng He and Xiaoming Yuan for pointing out the relationship with their proof techniques.

References

- [1] M. Afonso, J. Bioucas-Dias, and M. Figueiredo. Fast image recovery using variable splitting and constrained optimization. *IEEE Transactions on Image Processing*, 19(9):2345 – 2356, 2010.
- [2] K. Azuma. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal*, 19:357–367, 1967.
- [3] A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research (JMLR)*, 6:1705–1749, 2005.
- [4] O. Banerjee, L. E. Ghaoui, and A. d’Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research (JMLR)*, 9:485–516, 2008.
- [5] S. Barman, X. Liu, S. Draper, and B. Recht. Decomposition methods for large scale LP decoding. In *Arxiv*, 2012.
- [6] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31:167–175, 2003.
- [7] D. P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods*. Athena Scientific, 1996.
- [8] D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999.
- [9] D. P. Bertsekas and J. N. Tsitsiklis. *Parallel and Distributed Computation: Numerical Methods*. Prentice Hall, 1989.
- [10] D. Boley. Local linear convergence of the alternating direction method of multipliers on quadratic or linear programs. *TR 12-009, Department of Computer Science and Engineering, University of Minnesota*, 2012.
- [11] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundation Trends Machine Learning*, 3(1):1–122, 2011.
- [12] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
- [13] E. Candes, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of ACM*, 58(1):1–37, 2009.
- [14] Y. Censor and S. Zenios. *Parallel Optimization: Theory, Algorithms, and Applications*. Oxford University Press, 1998.

- [15] N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50:2050–2057, 2004.
- [16] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.
- [17] R. H. Chan, J. F. Yang, and X. M. Yuan. Alternating direction method for image inpainting in wavelet domain. *SIAM Journal on Imaging Science*, 4:807–826, 2011.
- [18] P. Combettes and J. Pesquet. Proximal splitting methods in signal processing. *ArXiv*, 2009.
- [19] K. David and S. Guido. *An Introduction to Variational Inequalities and Their Applications*. Society for Industrial and Applied Mathematics, 2000.
- [20] O. Dekel, R. Gilad-Bachrach, O. Shamir, and L. Xiao. Optimal distributed online prediction using mini-batches. *Journal of Machine Learning Research (JMLR)*, 13:165–202, 2012.
- [21] A. P. Dempster. Covariance selection. *Biometrics*, 28(1):157–175, 1972.
- [22] W. Deng and W. Yin. On the global and linear convergence of the generalized alternating direction method of multipliers. *ArXiv*, 2012.
- [23] J. Douglas and H. H. Rachford. On the numerical solution of heat conduction problems in two and three space variables. *Transactions of the American Mathematical Society*, 82:421–439, 1956.
- [24] J. Duchi, A. Agarwal, and M. Wainwright. Dual averaging for distributed optimization: Convergence analysis and network scaling. *IEEE Transactions on Automatic Control*, 57, 2012.
- [25] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the l_1 -ball for learning in high dimensions. In *International Conference on Machine Learning (ICML)*, pages 272–279, 2008.
- [26] J. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. In *Annual Conference on Learning Theory (COLT)*, 2010.
- [27] J. Duchi and Y. Singer. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research (JMLR)*, 10:2873–2898, 2009.
- [28] J. Eckstein and D. P. Bertsekas. An alternating direction method for linear programming. Technical report, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, 1990.
- [29] J. Eckstein and D.P. Bertsekas. On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55:293–318, 1992.
- [30] F. Facchinei and J.-S. Pang. *Finite-Dimensional Variational Inequalities and Complementarity Problems*, volume I. Springer, 2003.
- [31] M. A. T. Figueiredo and J. M. Bioucas-Dias. Restoration of Poissonian images using alternating direction optimization. *IEEE Transactions on Image Processing*, 19:3133–3145, 2010.
- [32] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432, 2008.

- [33] Q. Fu, H. Wang, and A. Banerjee. Bethe-ADMM for tree decomposition based parallel MAP inference. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2013.
- [34] D. Gabay. Applications of the method of multipliers to variational inequalities. In *Augmented Lagrangian Methods: Applications to the Solution of Boundary-Value Problems*. M. Fortin and R. Glowinski, eds., North-Holland: Amsterdam, 1983.
- [35] D. Gabay and B. Mercier. A dual algorithm for the solution of nonlinear variational problems via finite-element approximations. *Computers and Mathematics with Applications*, 2:17–40, 1976.
- [36] T. Goldstein, X. Bresson, and S. Osher. Geometric applications of the split Bregman method: Segmentation and surface reconstruction. *Journal of Scientific Computing*, 45:272–293, 2010.
- [37] T. Goldstein, B. Donoghue, and S. Setzer. Fast alternating direction optimization methods. *CAM report 12-35, UCLA*, 2012.
- [38] G. H. Golub and C. V. Loan. *Matrix Computations*. 3rd ed. Johns Hopkins University Press, 1996.
- [39] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer, 2009.
- [40] E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.
- [41] E. Hazan and S. Kale. Projection-free online learning. In *International Conference on Machine Learning (ICML)*, 2012.
- [42] B. He and X. Yuan. On non-ergodic convergence rate of Douglas-Rachford alternating direction method of multipliers. *Preprint, Available at http://www.optimization-online.org/DB_FILE/2012/01/3318.pdf*, 2012.
- [43] B. He and X. Yuan. On the $O(1/n)$ convergence rate of the Douglas-Rachford alternating direction method. *SIAM Journal on Numerical Analysis*, 50:700–709, 2012.
- [44] A. Juditsky, A. Nemirovski, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.
- [45] Z. Lin, M. Chen, L. Wu, and Y. Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *UIUC Technical Report UILU-ENG-09-2215*, 2009.
- [46] Z. Q. Luo. On the linear convergence of the alternating direction method of multipliers. *ArXiv*, 2012.
- [47] M. Mahdavi, R. Jin, and T. Yang. Trading regret for efficiency: Online convex optimization with long term constraints. *Journal of Machine Learning Research (JMLR)*, 13:2503–2528, 2012.
- [48] S. Mannor and J. N. Tsitsiklis. Online learning with constraints. In *Annual Conference on Learning Theory (COLT)*, 2006.
- [49] A. F. Martins, P. M. Aguiar, M. A. Figueiredo, N. A. Smith, and E. P. Xing. An augmented Lagrangian approach to constrained MAP inference. In *International Conference on Machine Learning (ICML)*, 2011.

- [50] N. Meinshausen and P. Buhlmann. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.
- [51] O. Meshi and A. Globerson. An alternating direction method for dual MAP LP relaxation. In *European Conference on Machine Learning (ECML)*, 2011.
- [52] A. Nedic and A. Ozdaglar. Cooperative distributed multi-agent optimization. In *Convex Optimization in Signal Processing and Communications*. D. P. Palomar and Y. C. Eldar, eds. Cambridge University Press, 2010.
- [53] A. Nedic and A. Ozdaglar. Cooperative distributed multi-agent optimization. in *Convex Optimization in Signal Processing and Communications*, 2010.
- [54] A. Nemirovski. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15:229–251, 2004.
- [55] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic Publishers, 2004.
- [56] Y. Nesterov. Gradient methods for minimizing composite objective function. *Technical Report 76, Center for Operation Research and Economics (CORE), Catholic University of Louvain (UCL)*, 2007.
- [57] Y. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120:221–259, 2009.
- [58] M. Ng, P. Weiss, and X. Yuan. Solving constrained total-variation problems via alternating direction methods. *SIAM Journal on Scientific Computing*, 32(5):2710–2736, 2010.
- [59] H. Ouyang, N. He, L. Q. Tran, and A. Gray. Stochastic alternating direction method of multipliers. In *International Conference on Machine Learning (ICML)*, 2013.
- [60] N. Parikh and S. Boyd. Graph projection block splitting for distributed optimization. *Preprint, Available at http://www.stanford.edu/~boyd/papers/pdf/block_splitting.pdf*, 2012.
- [61] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.
- [62] R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.
- [63] R. T. Rockafellar. Augmented Lagrangians and applications of the proximal point algorithm in convex programming. *Mathematics of Operations Research*, 1:97–116, 1976.
- [64] L. Rudin, S. J. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D*, 60:259–268, 1992.
- [65] K. Scheinberg, S. Ma, and D. Goldfarb. Sparse inverse covariance selection via alternating linearization methods. In *Annual Conference on Neural Information Processing Systems (NIPS)*, 2010.
- [66] I. Schizas, G. Mateos, and G. Giannakis. Distributed LMS for consensus-based in-network adaptive processing. *IEEE Transactions on Signal Processing*, 57(6):2365–2381, 2009.

- [67] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.
- [68] H. Wang and A. Banerjee. Online alternating direction method. In *International Conference on Machine Learning (ICML)*, 2012.
- [69] L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research (JMLR)*, 11:2543–2596, 2010.
- [70] J. Yang and Y. Zhang. Alternating direction algorithms for L1-problems in compressive sensing. *ArXiv*, 2009.
- [71] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68:49–67, 2006.
- [72] X. M. Yuan. Alternating direction methods for sparse covariance selection. *Preprint, Available at http://www.optimization-online.org/DB_FILE/2009/09/2390.pdf*, 2009.
- [73] X. M. Yuan and J. F. Yang. Sparse and low-rank matrix decomposition via alternating direction methods. *Preprint, Available at www.optimization-online.org/DB_FILE/2009/11/2447.pdf*, 2009.
- [74] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *International Conference on Machine Learning (ICML)*, pages 928–936, 2003.

A Proof of Theorem 4 and 6 in Case 2 in Section 6.1

Proof of Theorem 4 The first order derivative is 0, i.e.,

$$f'_t(\mathbf{x}_t) + \mathbf{A}^T \{ \mathbf{y}_t + \rho \mathbf{A}^T (\mathbf{A} \mathbf{x}_t - \mathbf{B} \mathbf{z}_t - \mathbf{c}) \} + \eta (\nabla \phi(\mathbf{x}_{t+1}) - \nabla \phi(\mathbf{x}_t)) = 0, \quad (149)$$

Rearranging the terms yields

$$-\mathbf{A}^T (\mathbf{y}_t + \rho \mathbf{A}^T (\mathbf{A} \mathbf{x}_{t+1} - \mathbf{B} \mathbf{z}_t - \mathbf{c})) - \eta (\nabla \phi(\mathbf{x}_{t+1}) - \nabla \phi(\mathbf{x}_t)) = f'_t(\mathbf{x}_t), \quad (150)$$

where the left hand side is same as (71). Therefore, $\langle f'_t(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}^* \rangle + g(\mathbf{z}_{t+1}) - g(\mathbf{z}^*)$ can be written as the right hand side of (72). Using the convexity of f_t , we have

$$\begin{aligned} f_t(\mathbf{x}_t) + g(\mathbf{z}_{t+1}) - (f_t(\mathbf{x}^*) + g(\mathbf{z}^*)) &\leq \langle f'_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^* \rangle + g(\mathbf{z}_{t+1}) - g(\mathbf{z}^*) \\ &= \langle f'_t(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}^* \rangle + g(\mathbf{z}_{t+1}) - g(\mathbf{z}^*) + \langle f'_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t+1} \rangle. \end{aligned} \quad (151)$$

Applying (74) for the last term, we have (75). Therefore, Theorem 4 holds for Case 2.

Proof of Theorem 6 Using the strong convexity of f_t and g defined in (81) and (82) respectively, we have

$$\begin{aligned} &f_t(\mathbf{x}_t) + g(\mathbf{z}_{t+1}) - (f_t(\mathbf{x}^*) + g(\mathbf{z}^*)) \\ &\leq \langle f'_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^* \rangle - \beta_1 B_\phi(\mathbf{x}^*, \mathbf{x}_t) + \langle g'(\mathbf{z}_{t+1}), \mathbf{z}_{t+1} - \mathbf{z}^* \rangle - \frac{\beta_2}{2} \|\mathbf{z}^* - \mathbf{z}_{t+1}\|_2^2 \\ &= \langle f'_t(\mathbf{x}_t), \mathbf{x}_{t+1} - \mathbf{x}^* \rangle + \langle f'_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}_{t+1} \rangle + \langle g'(\mathbf{z}_{t+1}), \mathbf{z}_{t+1} - \mathbf{z}^* \rangle - \beta_1 B_\phi(\mathbf{x}^*, \mathbf{x}_t) - \frac{\beta_2}{2} \|\mathbf{z}^* - \mathbf{z}_{t+1}\|_2^2. \end{aligned} \quad (152)$$

The first four terms are the same as in (151), which can be reduced to (75). Therefore, adding the last two terms to (75), we have

$$\begin{aligned}
& f_t(\mathbf{x}_t) + g(\mathbf{z}_{t+1}) - (f_t(\mathbf{x}^*) + g(\mathbf{z}^*)) \\
& \leq \frac{1}{2\rho}(\|\mathbf{y}_t\|_2^2 - \|\mathbf{y}_{t+1}\|_2^2) - \frac{\rho}{2}\|\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_t - \mathbf{c}\|_2^2 + \frac{\rho}{2}(\|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_t\|_2^2 - \|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_{t+1}\|_2^2) \\
& \quad + \frac{1}{2\alpha\eta}\|f'_t(\mathbf{x}_t)\|_q^2 + \eta(B_\phi(\mathbf{x}^*, \mathbf{x}_t) - B_\phi(\mathbf{x}^*, \mathbf{x}_{t+1})) - \beta_1 B_\phi(\mathbf{x}^*, \mathbf{x}_t) - \frac{\beta_2}{2}\|\mathbf{z}^* - \mathbf{z}_{t+1}\|_2^2. \tag{153}
\end{aligned}$$

Summing over t from 1 to T , we have

$$\begin{aligned}
R_1(T) & \leq \sum_{t=1}^T \frac{1}{2\rho_{t+1}}(\|\mathbf{y}_t\|_2^2 - \|\mathbf{y}_{t+1}\|_2^2) + \frac{1}{2\beta} \sum_{t=0}^T \frac{1}{\eta_{t+1}}\|f'_t(\mathbf{x}_t)\|_2^2 \\
& \quad + \sum_{t=1}^T \left(\frac{\rho_{t+1}}{2}(\|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_t\|_2^2 - \|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_{t+1}\|_2^2) - \frac{\beta_2}{2}\|\mathbf{z}^* - \mathbf{z}_{t+1}\|_2^2 \right) \\
& \quad + \sum_{t=1}^T (\eta_{t+1}(B_\phi(\mathbf{x}^*, \mathbf{x}_t) - B_\phi(\mathbf{x}^*, \mathbf{x}_{t+1})) - \beta_1 B_\phi(\mathbf{x}^*, \mathbf{x}_t)). \tag{154}
\end{aligned}$$

The difference between (154) and (88) lies in the last term. Setting $\eta_t = \beta_1 t$, we have the following telescoping sum for the last term :

$$\begin{aligned}
& \sum_{t=1}^T (\eta_{t+1}(B_\phi(\mathbf{x}^*, \mathbf{x}_t) - B_\phi(\mathbf{x}^*, \mathbf{x}_{t+1})) - \beta_1 B_\phi(\mathbf{x}^*, \mathbf{x}_t)) \\
& \leq \eta_2 B_\phi(\mathbf{x}^*, \mathbf{x}_1) + \sum_{t=2}^T B_\phi(\mathbf{x}^*, \mathbf{x}_t)(\eta_{t+1} - \eta_t - \beta_1) \\
& = 2\beta_1 D_{\mathbf{x}}^2, \tag{155}
\end{aligned}$$

which is the same as (91). Therefore, Theorem 6 holds for the Case 2.

B Proof of Stochastic Convergence Rates

Although the proof is based on Case 2 in Section 6.1, Case 3 and 4 will follow automatically. In the stochastic setting, replacing $f'_t(\mathbf{x}_t)$ by $f'(\mathbf{x}_t, \xi_t)$ in (150) gives

$$-\mathbf{A}^T(\mathbf{y}_t + \rho\mathbf{A}^T(\mathbf{A}\mathbf{x}_{t+1} - \mathbf{B}\mathbf{z}_t - \mathbf{c})) - \eta(\nabla\phi(\mathbf{x}_{t+1}) - \nabla\phi(\mathbf{x}_t)) = f'(\mathbf{x}_t, \xi_t), \tag{156}$$

(a) Replacing $f_t(\mathbf{x}_t)$, $f'_t(\mathbf{x}_t)$ by $f(\mathbf{x}_t)$, $f'(\mathbf{x}_t, \xi_t)$ respectively in (151) gives

$$\begin{aligned}
& f(\mathbf{x}_t) + g(\mathbf{z}_{t+1}) - (f(\mathbf{x}^*) + g(\mathbf{z}^*)) \leq \langle f'(\mathbf{x}_t, \xi_t), \mathbf{x}_t - \mathbf{x}^* \rangle + g(\mathbf{z}_{t+1}) - g(\mathbf{z}^*) \\
& = \langle f'(\mathbf{x}_t, \xi_t), \mathbf{x}_{t+1} - \mathbf{x}^* \rangle + g(\mathbf{z}_{t+1}) - g(\mathbf{z}^*) + \langle f'(\mathbf{x}_t, \xi_t), \mathbf{x}_t - \mathbf{x}_{t+1} \rangle. \tag{157}
\end{aligned}$$

As a result, we have the following result by replacing $f_t(\mathbf{x}_t)$, $f'_t(\mathbf{x}_t)$ by $f(\mathbf{x}_t)$, $f'(\mathbf{x}_t, \xi_t)$ in (75)

$$\begin{aligned}
& f(\mathbf{x}_t) + g(\mathbf{z}_{t+1}) - (f(\mathbf{x}^*) + g(\mathbf{z}^*)) \\
& \leq \frac{1}{2\rho}(\|\mathbf{y}_t\|_2^2 - \|\mathbf{y}_{t+1}\|_2^2) + \frac{\rho}{2}(\|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_t\|_2^2 - \|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_{t+1}\|_2^2) - \frac{\rho}{2}\|\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_t - \mathbf{c}\|_2^2 \\
& \quad + \frac{1}{2\alpha\eta}\|f'(\mathbf{x}_t, \xi_t)\|_q^2 + \eta(B_\phi(\mathbf{x}^*, \mathbf{x}_t) - B_\phi(\mathbf{x}^*, \mathbf{x}_{t+1})). \tag{158}
\end{aligned}$$

Moving the term $\frac{\rho}{2} \|\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_t - \mathbf{c}\|_2^2$ to the left hand side and using Lemma 4, we have

$$\begin{aligned} & f(\mathbf{x}_t) + g(\mathbf{z}_{t+1}) - (f(\mathbf{x}^*) + g(\mathbf{z}^*)) + \frac{\rho}{2} \|\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_{t+1} - \mathbf{c}\|_2^2 \\ & \leq \frac{1}{2\rho} (\|\mathbf{y}_t\|_2^2 - \|\mathbf{y}_{t+1}\|_2^2) + \frac{\rho}{2} (\|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_t\|_2^2 - \|\mathbf{B}\mathbf{z}^* - \mathbf{B}\mathbf{z}_{t+1}\|_2^2) \\ & \quad + \frac{1}{2\alpha\eta} \|f'(\mathbf{x}_t, \xi_t)\|_q^2 + \eta (B_\phi(\mathbf{x}^*, \mathbf{x}_t) - B_\phi(\mathbf{x}^*, \mathbf{x}_{t+1})) . \end{aligned} \quad (159)$$

Summing over t from 0 to $T - 1$ and following the derivation in (76), we have

$$\sum_{t=1}^T \left[f(\mathbf{x}_t) + g(\mathbf{z}_t) - (f(\mathbf{x}^*) + g(\mathbf{z}^*)) + \frac{\rho}{2} \|\mathbf{A}\mathbf{x}_{t+1} + \mathbf{B}\mathbf{z}_{t+1} - \mathbf{c}\|_2^2 \right] \leq \frac{\lambda_{\max}^{\mathbf{B}} D_{\mathbf{z}}^2 \rho}{2} + \eta D_{\mathbf{x}}^2 + \frac{\|f'(\mathbf{x}_t, \xi_t)\|_q^2 T}{2\alpha\eta} . \quad (160)$$

Dividing both sides by T , applying the Jensen's inequality, we have

$$f(\bar{\mathbf{x}}_T) + g(\bar{\mathbf{z}}_T) - (f(\mathbf{x}^*) + g(\mathbf{z}^*)) + \frac{\rho}{2} \|\mathbf{A}\bar{\mathbf{x}}_T + \mathbf{B}\bar{\mathbf{z}}_T + \mathbf{c}\|_2^2 \leq \frac{\lambda_{\max}^{\mathbf{B}} D_{\mathbf{z}}^2 \rho}{2T} + \frac{\eta D_{\mathbf{x}}^2}{T} + \frac{\|f'(\mathbf{x}_t, \xi_t)\|_q^2}{2\alpha\eta} . \quad (161)$$

Assume $\mathbf{E}[\|f'(\mathbf{x}_t, \xi_t)\|_q^2] \leq G_f^2$. Setting $\eta = \frac{G_f \sqrt{T}}{D_{\mathbf{x}} \sqrt{2\alpha}}$ and taking expectation, we have

$$\mathbf{E} \left[f(\bar{\mathbf{x}}_T) + g(\bar{\mathbf{z}}_T) - (f(\mathbf{x}^*) + g(\mathbf{z}^*)) + \frac{\rho}{2} \|\mathbf{A}\bar{\mathbf{x}}_T + \mathbf{B}\bar{\mathbf{z}}_T + \mathbf{c}\|_2^2 \right] \leq \frac{\lambda_{\max}^{\mathbf{B}} D_{\mathbf{z}}^2 \rho}{2T} + \frac{\sqrt{2} G_f D_{\mathbf{x}}}{\sqrt{\alpha} \sqrt{T}} . \quad (162)$$

(122) follows by setting $\rho = \sqrt{T}$.

Assume $f(\bar{\mathbf{x}}_T) + g(\bar{\mathbf{z}}_T) - (f(\mathbf{x}^*) + g(\mathbf{z}^*)) \geq -F$. Dividing both sides by $\frac{\rho}{2}$ and rearranging the terms yield

$$\mathbf{E} [\|\mathbf{A}\bar{\mathbf{x}}_T + \mathbf{B}\bar{\mathbf{z}}_T + \mathbf{c}\|_2^2] \leq \frac{2F}{\rho} + \frac{\lambda_{\max}^{\mathbf{B}} D_{\mathbf{z}}^2}{T} + \frac{2\sqrt{2} G_f D_{\mathbf{x}}}{\rho \sqrt{\alpha} \sqrt{T}} . \quad (163)$$

Setting $\rho = \sqrt{T}$ gives (123).

(b) Using the convexity of f , we have

$$f(\mathbf{x}_t) - f(\mathbf{x}^*) \leq \langle f'(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^* \rangle = \langle f'(\mathbf{x}_t, \xi_t), \mathbf{x}_{t+1} - \mathbf{x}^* \rangle + \langle f'(\mathbf{x}_t, \xi_t), \mathbf{x}_t - \mathbf{x}_{t+1} \rangle + \epsilon_t . \quad (164)$$

where

$$\epsilon_t = \langle f'(\mathbf{x}_t) - f'(\mathbf{x}_t, \xi_t), \mathbf{x}_t - \mathbf{x}^* \rangle . \quad (165)$$

Let \mathcal{F} be a filtration with $\xi_t \in \mathcal{F}_t$ for $t \leq T$. Since $\mathbf{x}_t \in \mathcal{F}_{t-1}$,

$$\mathbf{E}[\epsilon_t | \mathcal{F}_{t-1}] = \langle f'(\mathbf{x}_t) - \mathbf{E}[f'(\mathbf{x}_t, \xi_t) | \mathcal{F}_{t-1}], \mathbf{x}_t - \mathbf{x}^* \rangle = 0 . \quad (166)$$

Therefore, $\sum_{t=1}^T \epsilon_t$ is a martingale difference sequence. Assuming $B_\phi(\mathbf{x}^*, \mathbf{x}^t) \leq D_{\mathbf{x}}^2$, $\|\mathbf{x}_t - \mathbf{x}^*\|_p \leq \sqrt{\frac{2}{\alpha}} D_{\mathbf{x}}$. We have

$$|\epsilon_t| \leq \|f'(\mathbf{x}_t) - f'(\mathbf{x}_t, \xi_t)\|_q \|\mathbf{x}_t - \mathbf{x}^*\|_p \leq 2\sqrt{\frac{2}{\alpha}} D_{\mathbf{x}} G_f . \quad (167)$$

Applying Azuma-Hoeffding inequality [2] on $\sum_{t=1}^T \epsilon_t$ yields

$$P\left(\sum_{t=1}^T \epsilon_t \geq \varepsilon\right) \leq \exp\left(-\frac{\alpha \varepsilon^2}{16TD_{\mathbf{x}}^2 G_f^2}\right). \quad (168)$$

Combing (157) and (164), we have

$$\begin{aligned} f(\mathbf{x}_t) + g(\mathbf{z}_{t+1}) - (f(\mathbf{x}^*) + g(\mathbf{z}^*)) &\leq \langle f'_t(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^* \rangle + g(\mathbf{z}_{t+1}) - g(\mathbf{z}^*) \\ &= \langle f'(\mathbf{x}_t, \xi_t), \mathbf{x}_{t+1} - \mathbf{x}^* \rangle + g(\mathbf{z}_{t+1}) - g(\mathbf{z}^*) + \langle f'(\mathbf{x}_t, \xi_t), \mathbf{x}_t - \mathbf{x}_{t+1} \rangle + \epsilon_t. \end{aligned} \quad (169)$$

As a result, (161) becomes

$$\begin{aligned} f(\bar{\mathbf{x}}_T) + g(\bar{\mathbf{z}}_T) - (f(\mathbf{x}^*) + g(\mathbf{z}^*)) + \frac{\rho}{2} \|\mathbf{A}\bar{\mathbf{x}}_T + \mathbf{B}\bar{\mathbf{z}}_T + \mathbf{c}\|_2^2 \\ \leq \frac{\lambda_{\max}^{\mathbf{B}} D_{\mathbf{z}}^2 \rho}{2T} + \frac{\eta D_{\mathbf{x}}^2}{T} + \frac{\|f'(\mathbf{x}_t, \xi_t)\|_q^2}{2\alpha\eta} + \frac{1}{T} \sum_{t=1}^T \epsilon_t. \end{aligned} \quad (170)$$

Assuming $\|f'(\mathbf{x}_t, \xi_t)\|_q \leq G_f$ and setting $\eta = \frac{G_f \sqrt{T}}{D_{\mathbf{x}} \sqrt{2\alpha}}$, $\rho = \sqrt{T}$, we have

$$f(\bar{\mathbf{x}}_T) + g(\bar{\mathbf{z}}_T) - (f(\mathbf{x}^*) + g(\mathbf{z}^*)) + \frac{\rho}{2} \|\mathbf{A}\bar{\mathbf{x}}_T + \mathbf{B}\bar{\mathbf{z}}_T + \mathbf{c}\|_2^2 \leq \frac{\lambda_{\max}^{\mathbf{B}} D_{\mathbf{z}}^2}{2\sqrt{T}} + \frac{\sqrt{2} G_f D_{\mathbf{x}}}{\sqrt{\alpha} \sqrt{T}} + \frac{1}{T} \sum_{t=1}^T \epsilon_t. \quad (171)$$

Applying (168) gives (124).

Assume $f(\bar{\mathbf{x}}_T) + g(\bar{\mathbf{z}}_T) - (f(\mathbf{x}^*) + g(\mathbf{z}^*)) \geq -F$. In (171), dividing both sides by $\frac{\rho}{2} = \frac{\sqrt{T}}{2}$ and rearranging the terms yield

$$\|\mathbf{A}\bar{\mathbf{x}}_T + \mathbf{B}\bar{\mathbf{z}}_T + \mathbf{c}\|_2^2 \leq \frac{2F}{\rho} + \frac{\lambda_{\max}^{\mathbf{B}} D_{\mathbf{z}}^2}{T} + \frac{2\sqrt{2} G_f D_{\mathbf{x}}}{\sqrt{\alpha} T} + \frac{1}{T} \sum_{t=1}^T \epsilon_t. \quad (172)$$

Applying (168) yields (125). ■