Journal of Machine Learning Research 21 (2020) 1-45 Submitted 6/18; Revised 8/19; Published 3/20

# On the Complexity Analysis of the Primal Solutions for the Accelerated Randomized Dual Coordinate Ascent

**Huan Li** [*] LIHUANSS@NUAA.EDU.CN
*College of Computer Science and Technology*
*Nanjing University of Aeronautics and Astronautics*
*Nanjing, China*
*Key Lab. of Machine Perception (MOE), School of EECS*
*Peking University*
*Beijing, China*

**Zhouchen Lin** [†] ZLIN@PKU.EDU.CN
*Key Lab. of Machine Perception (MOE), School of EECS*
*Peking University*
*Beijing, China*

**Editor:** Miguel Carreira-Perpinan

## Abstract

Dual first-order methods are essential techniques for large-scale constrained convex optimization. However, when recovering the primal solutions, we need $T(\epsilon^{-2})$ iterations to achieve an $\epsilon$-optimal primal solution when we apply an algorithm to the non-strongly convex dual problem with $T(\epsilon^{-1})$ iterations to achieve an $\epsilon$-optimal dual solution, where $T(x)$ can be $x$ or $\sqrt{x}$. In this paper, we prove that the iteration complexity of the primal solutions and dual solutions have the same $O\left(\frac{1}{\sqrt{\epsilon}}\right)$ order of magnitude for the accelerated randomized dual coordinate ascent. When the dual function further satisfies the quadratic functional growth condition, by restarting the algorithm at any period, we establish the linear iteration complexity for both the primal solutions and dual solutions even if the condition number is unknown. When applied to the regularized empirical risk minimization problem, we prove the iteration complexity of $O\left(n\log n+\sqrt{\frac{n}{\epsilon}}\right)$ in both primal space and dual space, where $n$ is the number of samples. Our result takes out the $\left(\log\frac{1}{\epsilon}\right)$ factor compared with the methods based on smoothing/regularization or Catalyst reduction. As far as we know, this is the first time that the optimal $O\left(\sqrt{\frac{n}{\epsilon}}\right)$ iteration complexity in the primal space is established for the dual coordinate ascent based stochastic algorithms. We also establish the accelerated linear complexity for some problems with nonsmooth loss, e.g., the least absolute deviation and SVM.

**Keywords:** accelerated randomized dual coordinate ascent, restart at any period, iteration complexity, primal solutions, dual first-order methods

[*] Part of this work was done when Huan Li was a Ph.D student at Peking University.

[†] Zhouchen Lin is the corresponding author.

## 1. Introduction

In this paper, we study the following structured constrained convex optimization problem:

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^t} \quad & F(\mathbf{x}) \equiv f(\mathbf{x}) + \frac{1}{n}\sum_{i=1}^n \phi_i(\boldsymbol{A}_i^T\mathbf{x}), \\ s.t. \quad & \boldsymbol{B}\mathbf{x} + \mathbf{b} = \mathbf{0}, \\ & g_i(\mathbf{x}) \le 0, \quad i = 1, \cdots, m, \end{aligned} \tag{1}$$

where $\boldsymbol{A} \in \mathbb{R}^{t\times n}$, $\boldsymbol{B} \in \mathbb{R}^{p\times t}$, each $\phi_i$ and $g_i$ is convex and $f$ is $\mu$-strongly convex. Both $f$ and $\phi_i$ can be non-differentiable. In machine learning, each column of $\boldsymbol{A}$ often represents a data point. $\phi_i$ is often the loss function, e.g., $\phi_i(y) = |y|$ for the absolute deviation and $\phi_i(y) = \max\{0, 1 - l_i y\}$ for SVM, where $l_i \in \{\pm 1\}$ is the label for the $i$-th data. $f$ is often the regularizer, e.g., the $L_2$ regularization $f(\mathbf{x}) = \|\mathbf{x}\|_2^2$ and $L_1$-$L_2$ regularization $f(\mathbf{x}) = \|\mathbf{x}\|_2^2 + \sigma\|\mathbf{x}\|_1$. Problem (1) is actually very general to incorporate many existing problems in machine learning. When dropping the constraints, problem (1) becomes the regularized empirical risk minimization (ERM) problem associated with linear predictors:

$$\min_{\mathbf{x}\in\mathbb{R}^t} \quad F(\mathbf{x}) \equiv f(\mathbf{x}) + \frac{1}{n}\sum_{i=1}^n \phi_i(\boldsymbol{A}_i^T\mathbf{x}). \tag{2}$$

The ERM problem is widely used in machine learning. Please see (Shalev-Shwartz and Zhang, 2013, 2016) for examples.

Due to the complicated constraints, people often do not solve problem (1) directly. Instead, they solve its dual problem by introducing the Lagrangian function. Many first-order methods can be used to solve the dual problem, e.g., the dual full gradient ascent (DFGA) (Tseng, 1990), the accelerated DFGA (ADFGA) (Beck and Teboulle, 2014; Huang et al., 2013), the randomized dual coordinate ascent (RDCA) (Nesterov, 2012a; Lu and Xiao, 2015; Richtárik and Takáč, 2014; Shalev-Shwartz and Zhang, 2013) and the accelerated RDCA (ARDCA) (Nesterov, 2012a; Fercoq and Richtárik, 2015; Lin et al., 2015b; Shalev-Shwartz and Zhang, 2016)[1]. They need $O\left(\frac{1}{\epsilon}\right)$, $O\left(\frac{1}{\sqrt{\epsilon}}\right)$, $O\left(\frac{\hat{n}}{\epsilon}\right)$ and $O\left(\frac{\hat{n}}{\sqrt{\epsilon}}\right)$ iterations to achieve an $\epsilon$-optimal dual solution, respectively, where $\hat{n}$ is the dimension in the dual space. At each iteration, RDCA and ARDCA choose one coordinate to sufficiently increase the dual objective value while keeping the others fixed. The cost at each iteration of RDCA and ARDCA may be much lower than that of DFGA and ADFGA. Since both $f$ and $\phi_i$ can be non-differentiable, the dual function is non-strongly convex. So only the sublinear complexity can be obtained.

It is not satisfactory to establish the iteration complexity only in the dual space. We should recover the primal solutions from the dual iterates and need to estimate how quickly the primal solutions converge. Unfortunately, Lu and Johansson (2016) established the algorithm independent result that the iteration complexity in the primal space is worse than

[1] Although the algorithm studied in this paper is a special case of APCG in (Lin et al., 2015b) and APPROX in (Fercoq and Richtárik, 2015), we name it ARDCA to emphasize the application to the dual problem.

that in the dual space. Specifically, Lu and Johansson (2016) studied the following problem, which is a special case of problem (1),

$$\begin{aligned} \min_{\mathbf{x}\in\mathbb{R}^t} \quad & f(\mathbf{x}), \\ s.t. \quad & \boldsymbol{B}\mathbf{x}+\mathbf{b}=\mathbf{0}, \\ & g_i(\mathbf{x})\leq 0, \quad i=1,\cdots,m. \end{aligned} \tag{3}$$

For a pair of approximate primal-dual solution $\{\mathbf{x}^*(\mathbf{u}^K),\mathbf{u}^K\}$[2], the precision between $\mathbf{x}^*(\mathbf{u}^K)$ and $\mathbf{u}^K$ satisfies

$$\begin{aligned} &|f(\mathbf{x}^*(\mathbf{u}^K))-f(\mathbf{x}^*)|\leq O\left(\sqrt{D(\mathbf{u}^K)-D(\mathbf{u}^*)}+D(\mathbf{u}^K)-D(\mathbf{u}^*)\right), \\ &\left\|\begin{bmatrix} \boldsymbol{B}\mathbf{x}^*(\mathbf{u}^K)+\mathbf{b} \\ \max\left\{0,g(\mathbf{x}^*(\mathbf{u}^K))\right\} \end{bmatrix}\right\|\leq O\left(\sqrt{D(\mathbf{u}^K)-D(\mathbf{u}^*)}\right), \end{aligned}$$

where $D(\mathbf{u})$ is the negative of the dual function and $(\mathbf{x}^*,\mathbf{u}^*)$ is a pair of optimal primal-dual solution. Thus if some algorithm achieves an $\epsilon$-optimal dual solution[3] of $\mathbf{u}^K$ after $T(\epsilon^{-1})$ iterations, it only achieves an $\sqrt{\epsilon}$-optimal primal solution[4] of $\mathbf{x}^*(\mathbf{u}^K)$ after the same time. Lu and Johansson (2016) studied DFGA and ADFGA and established the $O\left(\frac{1}{\epsilon^2}\right)$ and $O\left(\frac{1}{\epsilon}\right)$ iteration complexity in the primal space to find an $\epsilon$-optimal primal solution. Dünner et al. (2016) proved the similar algorithm independent results for problem (2). Kim and Fessler (2016) proved the $O\left(\frac{1}{\epsilon^{2/3}}\right)$ iteration complexity to achieve an $\epsilon$-optimal primal solution for the deterministic accelerated full gradient methods for problem (2).

Some researchers used regularization/smoothing to improve the iteration complexity of the primal solutions. They added a small regularization term $\epsilon\|\mathbf{u}\|^2$ to the dual function to smooth the primal objective and solved a regularized dual problem by some algorithm with linear convergence rate. Devolder et al. (2012) applied ADFGA to a smoothed problem of (3) and Necoara and Patrascu (2016) used ADFGA to solve a regularized dual problem of conic convex programming. However, they established the suboptimal iteration complexity of $O\left(\frac{1}{\sqrt{\epsilon}}\log\frac{1}{\epsilon}\right)$ to achieve an $\epsilon$-optimal primal solution recovered from the last dual iterate, which has an additional $\left(\log\frac{1}{\epsilon}\right)$ factor. The drawback of this strategy in practice is that it needs to choose the parameter $\epsilon$ in advance, which is related to the target accuracy. It is desirable to develop direct support for problems with non-smooth primal objective or non-strongly convex dual objective.

Other researchers improved the iteration complexities of DFGA and ADFGA in the primal space via averaging the primal solutions appropriately. Tseng (2008) studied the problem of $\min_{\mathbf{x}}\max_{\mathbf{v}}\psi(\mathbf{x},\mathbf{v})+P(\mathbf{x})$ and established the $O\left(\frac{1}{\sqrt{\epsilon}}\right)$ iteration complexity measured by the duality gap for the accelerated full gradient method. Necoara and Nedelcu (2014) and Patrinos and Bemporad (2013) used Tseng (2008)'s result for ADFGA to solve

[2] $\mathbf{x}^*(\mathbf{u})$ is recovered form $\mathbf{u}$ and will be defined in (7) later.

[3] We define an $\epsilon$-optimal dual solution as $D(\mathbf{u})-D(\mathbf{u}^*)\leq\epsilon$.

[4] We define an $\epsilon$-optimal primal solution as $|F(\mathbf{x})-F(\mathbf{x}^*)|\leq\epsilon$ and $\left\|\begin{bmatrix}\boldsymbol{B}\mathbf{x}+\mathbf{b}\\ \max\{0,g(\mathbf{x})\}\end{bmatrix}\right\|\leq\epsilon$. $\|\cdot\|$ can be a general norm.

the embedded linear model predictive control problem, which is a special case of problem (3). Necoara and Patrascu (2016) proved the $O\left(\frac{1}{\epsilon}\right)$ iteration complexity for DFGA and $O\left(\frac{1}{\sqrt{\epsilon}}\right)$ iteration complexity for ADFGA to achieve an $\epsilon$-optimal averaged primal solution for conic convex programming. None of them studied the general problem (1) and none of them studied the methods based on randomized dual coordinate ascent.

The randomized coordinate descent and its accelerated version have received extensive attention recently for solving large-scale optimization problems since it can break down the problem into smaller pieces. Shalev-Shwartz and Zhang (2013) showed that the Stochastic Dual Coordinate Ascent (SDCA) needs $O\left(n\log n+\frac{1}{\epsilon}\right)$ iterations to reach an $\epsilon$-optimal solution in both the primal space and dual space for problem (2). Shalev-Shwartz and Zhang (2016) then developed an accelerated SDCA (ASDCA) and attained the suboptimal $O\left(\left(n+\sqrt{\frac{n}{\epsilon}}\right)\log\frac{1}{\epsilon}\right)$ iteration complexity to achieve an $\epsilon$-optimal primal solution via solving a regularized dual problem, which has the additional $\left(\log\frac{1}{\epsilon}\right)$ factor due to the smoothing/regularization technique. Catalyst (Lin et al., 2015a), a general scheme for accelerating first-order optimization methods, also yields the additional $\left(\log\frac{1}{\epsilon}\right)$ factor. The Accelerated randomized Proximal Coordinate Gradient (APCG) method (Lin et al., 2015b) is another famous method for problem (2), which needs $O\left(\frac{n}{\sqrt{\epsilon}}\right)$ iterations to find a dual solution in $\epsilon$ accuracy. However, the sublinear complexity in the primal space is not established in (Lin et al., 2015b). Zhang and Xiao (2017) proposed a Stochastic Primal-Dual Coordinate method (SPDC) and Lan and Zhou (2018) proposed a Randomized Primal-Dual Gradient method (RPDG) for problem (2). They smoothed $\phi_i$ and achieved the $O\left(\left(n+\sqrt{\frac{n}{\epsilon}}\right)\log\frac{1}{\epsilon}\right)$ iteration complexity. When $\phi_i$ has $\frac{1}{\gamma}$-Lipschitz continuous gradient, ASDCA, APCG, SPDC and RPDG all have the accelerated linear complexity of $O\left(\left(n+\sqrt{\frac{n}{\gamma\mu}}\right)\log\frac{1}{\epsilon}\right)$.

### 1.1. Contributions

In this paper, we study the iteration complexity of the primal solutions when using ARDCA to solve the non-strongly convex dual problem. Specifically, we aim to prove that the complexity of the primal solutions has the same order of magnitude as that of the dual solutions.

For the general problem (1), when applying ARDCA to solve its dual problem, we prove the $O\left(\frac{\hat{n}}{\sqrt{\epsilon}}\right)$ iteration complexity of the primal solutions simply by averaging the last few primal iterates appropriately. This complexity has the same order of magnitude as that of the dual solutions and thus improves the theoretical results in (Lu and Johansson, 2016; Dünner et al., 2016). As a comparison, literature (Tseng, 2008; Necoara and Nedelcu, 2014; Patrinos and Bemporad, 2013; Necoara and Patrascu, 2016) only studied ADFGA, which is much simpler than the analysis of ARDCA. Since we use ARDCA to solve the dual problem directly, rather than a regularized dual problem or a smoothed primal problem, our result takes out the $\left(\log\frac{1}{\epsilon}\right)$ factor compared with the smoothing/regularization based methods.

When the dual function satisfies the quadratic functional growth condition, by restarting ARDCA at any period, we prove the linear iteration complexity for both the primal solutions and dual solutions. Moreover, our analysis does not require the parameters of the algorithm depend on the condition number $\kappa$, which will be defined in Assumption 2 later and it is often difficult to estimate in practice. We show that ARDCA with restart outperforms

RCDA for a wide range of inner iteration numbers and the optimal $O\left(\left(\hat{n}+\frac{\hat{n}}{\sqrt{\kappa}}\right)\log\frac{1}{\epsilon}\right)$ complexity can be attained when the inner iteration number is equal to $O\left(\hat{n}+\frac{\hat{n}}{\sqrt{\kappa}}\right)$. The difference with respect to (Fercoq and Qu, 2020) is that our analysis does not require the uniqueness of the optimal dual solution.

When applied to problem (2), our work extends the theoretical results of (Lin et al., 2015b) and improves those of (Shalev-Shwartz and Zhang, 2016). We prove that ARDCA needs $O\left(n\log n+\sqrt{\frac{n}{\epsilon}}\right)$ iterations to find an $\epsilon$-optimal solution in both the primal space and dual space, while Lin et al. (2015b) only proved the iteration complexity in the dual space. This complexity matches the theoretical lower bound (Woodworth and Srebro, 2016) and state-of-the-art upper bound (Allen-Zhu, 2018). Our theory outperforms ASDCA (Shalev-Shwartz and Zhang, 2016) and Catalyst (Lin et al., 2015a) by the factor of $\left(\log\frac{1}{\epsilon}\right)$. As far as we know, we are the first to establish the optimal $O\left(\sqrt{\frac{n}{\epsilon}}\right)$ complexity in the primal space for the dual coordinate ascent based stochastic algorithms. When $\phi_i$ has $\frac{1}{\gamma}$-Lipschitz continuous gradient, ARDCA with restart has the optimal $O\left(\left(n+\sqrt{\frac{n}{\gamma\mu}}\right)\log\frac{1}{\epsilon}\right)$ complexity. Moreover, we establish the accelerated linear complexity of ARDCA with restart for some special problems with nonsmooth $\phi_i$, e.g., the least absolute deviation problem and support vector machine (SVM).

### 1.2. Assumption, Notation and Problem Formulation

We describe the assumptions, notations and problem formulation in this section.

#### 1.2.1. Assumption

We study problem (1) under the following assumptions:

**Assumption 1**

1. *$f$ is $\mu$-strongly convex over $\mathbb{R}^t$, i.e., $f(\mathbf{y})\geq f(\mathbf{x})+\langle\mathbf{s},\mathbf{y}-\mathbf{x}\rangle+\frac{\mu}{2}\|\mathbf{y}-\mathbf{x}\|^2,\forall\mathbf{x},\mathbf{y}$, for every subgradient $\mathbf{s}\in\partial f(\mathbf{x})$.*
2. *$\phi_i$ is convex and $M$-Lipschitz continuous over $\mathbb{R}$, i.e., $|\phi_i(x)-\phi_i(y)|\leq M|x-y|,\forall x,y$.*
3. *$g_i$ is convex and has bounded subgradients over $\mathbb{R}^t$, i.e., $\|\mathbf{s}\|\leq L_{g_i}$, $\forall\mathbf{s}\in\partial g_i(\mathbf{x})$.*
4. *There exists $\overline{\mathbf{x}}$ such that $g_i(\overline{\mathbf{x}})<0$ and $\boldsymbol{B}\overline{\mathbf{x}}+\mathbf{b}=\mathbf{0}$.*
5. *The optimal objective value of problem (1) is finite.*

Assumption 1.4 is the Slater's condition. Assumption 1.4 and 1.5 ensure that the strong duality holds, i.e., the dual optimal value is equal to the primal optimal value (Bertsekas, 1999). Assumptions 1.1 and 1.3 will be used to establish the Lipschitz smoothness of part of the dual function in Lemma 1. Assumption 1.2 will be used to get the complexity for problem (1) from that of the reformulated problem (5).

To make each iteration of the randomized dual coordinate ascent computationally efficient, we only consider the case that $g(\mathbf{x})$ is a linear function for simplicity, i.e.,

$$g(\mathbf{x})=\boldsymbol{J}\mathbf{x}+\mathbf{q},\tag{4}$$

where $\boldsymbol{J} \in \mathbb{R}^{m\times t}$ and $\mathbf{q} \in \mathbb{R}^m$. However, the analysis in this paper suits for the general function $g(\mathbf{x})$ satisfying Assumption 1.3.

In Section 4, we will prove the linear complexity of ARDCA with restart under the quadratic functional growth condition (Necoara et al., 2019; Ma et al., 2016). This condition is equivalent to the error bound condition in (Luo and Tseng, 1992; Drusvyatskiy and Lewis, 2018) and is satisfied for broad applications in machine learning, e.g., the least absolute deviation and SVM (Wang and Lin, 2014).

**Assumption 2** *$D(\mathbf{u})$ satisfies the quadratic functional growth condition with respect to the norm $\|\cdot\|_L$, i.e., $\kappa\|\mathbf{u}-Proj_{\mathbb{D}^*}(\mathbf{u})\|_L^2 \leq D(\mathbf{u})-D(\mathbf{u}^*), \forall \mathbf{u}$, where $\kappa>0$ is the condition number, $\mathbf{u}^*$ is the optimal dual solution, $Proj_{\mathbb{D}^*}(\mathbf{u})$ is the projection of $\mathbf{u}$ onto the optimal dual solution set $\mathbb{D}^*$ and $D(\mathbf{u})$ is the negative of the dual function.*

### 1.2.2. Notation

Lowercase bold letters, e.g., $\mathbf{u}, \mathbf{v}, \mathbf{z}, \mathbf{x}$ and $\mathbf{y}$, represent vectors, uppercase bold letters, e.g., $\boldsymbol{A}$ and $\boldsymbol{B}$, represent matrices and non-bold letters, e.g., $\theta$ and $\alpha$, represent scalars. Let $\mathbb{R}_+^m$ be the set of nonnegative vectors in $\mathbb{R}^m$ and $\hat{n}=n+p+m$ be the dimension of the dual variable. Denote $\mathbf{u}_i$ and $\nabla_i d(\mathbf{u})$ as the $i$-th element of $\mathbf{u}$ and $\nabla d(\mathbf{u})$, respectively. Let $\mathbf{u}_{i:j}$ and $g_{1:m}(\mathbf{x})$ be the vectors consisting of $\mathbf{u}_i, \cdots, \mathbf{u}_j$ and $g_1(\mathbf{x}), \cdots, g_m(\mathbf{x})$, respectively. The value of $\mathbf{u}$ at iteration $k$ is denoted by $\mathbf{u}^k$. For scalars, e.g., $\theta$, $\theta_k$ represents its value at iteration $k$ and $\theta^2$ denotes its squares. $\boldsymbol{A}_i \in \mathbb{R}^t$ and $\boldsymbol{A}_{j,:} \in \mathbb{R}^n$ are the $i$-th column and $j$-th row of $\boldsymbol{A}$, respectively. We use $\|\cdot\|$ as the $l_2$ Euclidean norm and $\|\cdot\|_\infty$ as the infinite norm for a vector. Define the weighted norm $\|\mathbf{x}\|_L=\sqrt{\sum_i L_i\|\mathbf{x}_i\|^2}$ and its dual norm $\|\mathbf{x}\|_L^*=\sqrt{\sum_i \frac{1}{L_i}\|\mathbf{x}_i\|^2}$. For any matrix $\boldsymbol{A}$, $\|\boldsymbol{A}\|_2=\sigma_{\max}(\boldsymbol{A})$ is the largest singular value of $\boldsymbol{A}$. $\lfloor x\rfloor$ ($\lceil x\rceil$) means the largest (smallest) integer less (larger) than or equal to $x$. For a function $\phi_i$, we use $\phi_i^*(u)=\sup_v \langle u,v\rangle-\phi_i(v)$ to denote its conjugate and $\text{Prox}_{\phi_i}(v)=\text{argmin}_u \phi_i(u)+\frac{1}{2}||u-v\|^2$ to denote its proximal mapping. Define $\phi(\mathbf{y})=\sum_{i=1}^n \phi_i(\mathbf{y}_i)$.

### 1.2.3. Problem Formulation

Reformulate problem (1) as

$$\begin{aligned}
\min_{\mathbf{x}\in\mathbb{R}^t,\mathbf{y}\in\mathbb{R}^n} \quad & f(\mathbf{x})+\frac{1}{n}\sum_{i=1}^n \phi_i(\mathbf{y}_i),\\
s.t. \quad & \frac{1}{n}(\boldsymbol{A}^T\mathbf{x}-\mathbf{y})=\mathbf{0},\\
& \boldsymbol{B}\mathbf{x}+\mathbf{b}=\mathbf{0},\\
& g_i(\mathbf{x})\leq 0, \quad i=1,\cdots,m,
\end{aligned} \tag{5}$$

and introduce the Lagrangian function as

$$L_F(\mathbf{x},\mathbf{y},\mathbf{u})=f(\mathbf{x})+\frac{1}{n}\sum_{i=1}^n\phi_i(\mathbf{y}_i)+\frac{1}{n}\langle\mathbf{u}_{1:n},\boldsymbol{A}^T\mathbf{x}-\mathbf{y}\rangle+\langle\mathbf{u}_{n+1:n+p},\boldsymbol{B}\mathbf{x}+\mathbf{b}\rangle+\sum_{i=1}^m\mathbf{u}_{n+p+i}g_i(\mathbf{x}), \tag{6}$$

where $\mathbf{u}\in\mathbb{R}^{\hat{n}}$ is the vector of the Lagrange multipliers. Define the dual feasible set as $\mathbb{D}=\{\mathbf{u}\in\mathbb{R}^{\hat{n}}:\mathbf{u}_{n+p+1:n+p+m}\in\mathbb{R}_+^m\}$ and denote $\mathbb{D}^*$ to be the optimal dual solution set.

Then the Lagrange dual problem of (5) can be expressed as

$$
\begin{aligned}
&\max_{\mathbf{u}\in\mathbb{D}} \min_{\mathbf{x}\in\mathbb{R}^t,\mathbf{y}\in\mathbb{R}^n} L_F(\mathbf{x},\mathbf{y},\mathbf{u})\\
=&\max_{\mathbf{u}\in\mathbb{D}}\left(\min_{\mathbf{x}\in\mathbb{R}^t}\left(f(\mathbf{x})+\left\langle\mathbf{u}_{1:n},\boldsymbol{A}^T\mathbf{x}/n\right\rangle+\langle\mathbf{u}_{n+1:n+p},\boldsymbol{B}\mathbf{x}+\mathbf{b}\rangle+\sum_{i=1}^{m}\mathbf{u}_{n+p+i}g_i(\mathbf{x})\right)\right.\\
&\qquad\left.+\min_{\mathbf{y}\in\mathbb{R}^n}\frac{1}{n}\left(\sum_{i=1}^{n}\phi_i(\mathbf{y}_i)-\langle\mathbf{u}_{1:n},\mathbf{y}\rangle\right)\right)\\
=&\max_{\mathbf{u}\in\mathbb{D}}\left(L_f(\mathbf{x}^*(\mathbf{u}),\mathbf{u})-\frac{1}{n}\sum_{i=1}^{n}\phi_i^*(\mathbf{u}_i)\right),
\end{aligned}
$$

where

$$
\begin{aligned}
&L_f(\mathbf{x},\mathbf{u})=f(\mathbf{x})+\left\langle\mathbf{u}_{1:n},\boldsymbol{A}^T\mathbf{x}/n\right\rangle+\langle\mathbf{u}_{n+1:n+p},\boldsymbol{B}\mathbf{x}+\mathbf{b}\rangle+\sum_{i=1}^{m}\mathbf{u}_{n+p+i}g_i(\mathbf{x}),\\
&\mathbf{x}^*(\mathbf{u})=\operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^t}L_f(\mathbf{x},\mathbf{u}).
\end{aligned}
\tag{7}
$$

Define

$$
d(\mathbf{u})=-L_f(\mathbf{x}^*(\mathbf{u}),\mathbf{u})\quad\text{and}\quad h_i(\mathbf{u}_i)=\begin{cases}\frac{1}{n}\phi_i^*(\mathbf{u}_i), & i=1,\cdots,n,\\ 0, & n+1\le i\le n+p,\\ I_{u\ge 0}(\mathbf{u}_i), & i>n+p,\end{cases}\tag{8}
$$

where $I_{u\ge0}(u)=\begin{cases}0, & \text{if } u\ge 0,\\ \infty, & \text{otherwise.}\end{cases}$ Then we can rewrite the Lagrange dual problem as

$$
\min_{\mathbf{u}\in\mathbb{R}^{\hat{n}}} D(\mathbf{u})=d(\mathbf{u})+h(\mathbf{u}),\tag{9}
$$

where we define $h(\mathbf{u})=\sum_{i=1}^{\hat{n}}h_i(\mathbf{u}_i)$ and $D(\mathbf{u})$ means the negative of the dual function. We study the negative of the dual function, rather than the dual function directly, since $D(\mathbf{u})$ is convex. Let $(\mathbf{x}^*,\mathbf{y}^*)$ and $\mathbf{u}^*$ be the optimal primal solution and dual solution of problem (5), respectively. Then they satisfy the KKT condition (Bertsekas, 1999). Since the strong duality holds, we have $f(\mathbf{x}^*)+\frac{1}{n}\phi(\mathbf{y}^*)=-D(\mathbf{u}^*)$. Since $L_f(\mathbf{x},\mathbf{u})$ is strongly convex over $\mathbf{x}$ for every $\mathbf{u}\in\mathbb{D}$, then $\mathbf{x}^*(\mathbf{u})$ is unique. Due to Danskin's theorem (Bertsekas, 1999) we know that $d(\mathbf{u})$ is convex, differentiable and

$$
\nabla d(\mathbf{u})=-\left[\left(\boldsymbol{A}^T\mathbf{x}^*(\mathbf{u})/n\right)^T,(\boldsymbol{B}\mathbf{x}^*(\mathbf{u})+\mathbf{b})^T,g_1(\mathbf{x}^*(\mathbf{u})),\cdots,g_m(\mathbf{x}^*(\mathbf{u}))\right]^T.\tag{10}
$$

From Proposition 3.3 in (Lu and Johansson, 2016), we have a Lipschitz smooth condition[5] of $\|\nabla d(\mathbf{u})-\nabla d(\mathbf{v})\|\le L\|\mathbf{u}-\mathbf{v}\|,\forall\mathbf{u},\mathbf{v}\in\mathbb{D}$, where

$$
L=\frac{\sqrt{m+1}\max\{\|[\boldsymbol{A}^T/n,\boldsymbol{B}]\|_2,\max_i L_{g_i}\}}{\mu}\sqrt{\|[\boldsymbol{A}^T/n,\boldsymbol{B}]\|_2^2+\sum_{i=1}^{m}L_{g_i}^2}.\tag{11}
$$

[5] Lu and Johansson (2016) studied the projected gradient method under the local Lipschitz smooth condition and the fast gradient method under the global Lipschitz smooth condition. The former condition is ensured by replacing Assumption 1.3 with $|g_i(\mathbf{x})-g_i(\mathbf{y})|\le L_{g_i}\|\mathbf{x}-\mathbf{y}\|$ and a further assumption that $g_i$ is differentiable. In this paper, we use the global Lipschitz smooth condition over the dual feasible set $\mathbb{D}$ for simplicity.

Similarly, we can also prove a coordinatewise Lipschitz smooth condition in the following lemma, whose proof is given in Appendix B.

**Lemma 1** *For any* $\mathbf{u}, \mathbf{v} \in \mathbb{D}$ *and any* $j$*, assume that* $\mathbf{u}_i = \mathbf{v}_i, \forall i \neq j$*. Then* $\|\nabla_j d(\mathbf{u}) - \nabla_j d(\mathbf{v})\| \leq L_j \|\mathbf{u} - \mathbf{v}\|$*, where*

$$L_j = \begin{cases} \frac{\|\boldsymbol{A}_j\|^2}{n^2\mu}, & j \leq n, \\ \frac{\|\boldsymbol{B}_{j-n,:}\|^2}{\mu}, & n < j \leq n+p, \\ \frac{L_{g_{j-n-p}}^2}{\mu}, & j > n+p. \end{cases} \tag{12}$$

An immediate consequence of Lemma 1 is (Nesterov, 2004, Lemma 1.2.3)

$$|d(\mathbf{u}) - d(\mathbf{v}) - \langle \nabla_j d(\mathbf{v}), \mathbf{u}_j - \mathbf{v}_j \rangle| \leq \frac{L_j}{2} \|\mathbf{u}_j - \mathbf{v}_j\|^2 \tag{13}$$

for all $\mathbf{u}, \mathbf{v} \in \mathbb{D}$ satisfying $\mathbf{u}_i = \mathbf{v}_i, \forall i \neq j$.

## 2. Accelerated Randomized Dual Coordinate Ascent

In this section, we use the standard accelerated randomized coordinate descent (Fercoq and Richtárik, 2015; Lin et al., 2015b) to solve the dual problem (9), which consists of the following steps at each iteration:

$$\begin{aligned}
&\mathbf{v}^k = \theta_k \mathbf{z}^k + (1-\theta_k)\mathbf{u}^k, \\
&\text{select } i_k \text{ randomly with probability of } 1/\hat{n}, \\
&\mathbf{z}_{i_k}^{k+1} = \operatorname*{argmin}_u \frac{\hat{n}\theta_k L_{i_k}}{2} \|u - \mathbf{z}_{i_k}^k\|^2 + \left\langle \nabla_{i_k} d(\mathbf{v}^k), u - \mathbf{z}_{i_k}^k \right\rangle + h_{i_k}(u), \\
&\mathbf{z}_j^{k+1} = \mathbf{z}_j^k, \forall j \neq i_k, \\
&\mathbf{u}^{k+1} = \mathbf{v}^k + \hat{n}\theta_k(\mathbf{z}^{k+1} - \mathbf{z}^k), \\
&\theta_{k+1} = \frac{\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k^2}{2}.
\end{aligned}$$

At each iteration, accelerated randomized coordinate descent picks a random coordinate $i_k \in \{1, 2, \cdots, \hat{n}\}$ and generates $\mathbf{z}^{k+1}, \mathbf{u}^{k+1}$ and $\mathbf{v}^{k+1}$. Only the $i_k$-th coordinate of $\mathbf{z}^{k+1}$ is updated and the other coordinates remain unchanged. However, the above algorithm performs full-dimensional vector operations of $\mathbf{v}^k$ and $\mathbf{u}^k$, which makes the per-iteration cost higher than the simple non-accelerated coordinate descent. To avoid such operations, we can use a change of variables scheme proposed in (Lee and Sidford, 2013; Fercoq and Richtárik, 2015). Specifically, introduce $\hat{\mathbf{u}}^k$ initialized at $\hat{\mathbf{u}}^0 = 0$ and the new algorithm consists of the following steps at each iteration:

$$\begin{aligned}
&\text{select } i_k \text{ randomly with probability of } 1/\hat{n},\\
&\hat{\mathbf{z}}_{i_k}^{k+1} = \operatorname*{argmin}_u \frac{\hat{n}\theta_k L_{i_k}}{2}\|u - \hat{\mathbf{z}}_{i_k}^k\|^2 + \left\langle \nabla_{i_k} d\left(\theta_k^2 \hat{\mathbf{u}}^k + \hat{\mathbf{z}}^k\right), u - \hat{\mathbf{z}}_{i_k}^k \right\rangle + h_{i_k}(u),\\
&\hat{\mathbf{z}}_j^{k+1} = \hat{\mathbf{z}}_j^k, \forall j \neq i_k,\\
&\hat{\mathbf{u}}^{k+1} = \hat{\mathbf{u}}^k - \frac{1 - \hat{n}\theta_k}{\theta_k^2}(\hat{\mathbf{z}}^{k+1} - \hat{\mathbf{z}}^k),\\
&\theta_{k+1} = \frac{\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k^2}{2}.
\end{aligned}$$

From Proposition 1 in (Fercoq and Richtárik, 2015), we know that the above two algorithms are equivalent in the sense of $\mathbf{z}^k = \hat{\mathbf{z}}^k$, $\mathbf{u}^{k+1} = \theta_k^2 \hat{\mathbf{u}}^{k+1} + \mathbf{z}^{k+1}$ and $\mathbf{v}^k = \theta_k^2 \hat{\mathbf{u}}^k + \mathbf{z}^k$ given $\mathbf{u}^0 = \mathbf{z}^0 = \hat{\mathbf{z}}^0$.

Next, we discuss the computation of $\nabla_{i_k} d(\mathbf{v}^k)$, and we expect that it is best $\hat{n}$ times faster than the computation of $\nabla d(\mathbf{v}^k)$. Consider the simple case of (4). Define

$$\boldsymbol{S} = \left[\boldsymbol{A}/n, \boldsymbol{B}^T, \boldsymbol{J}^T\right] \in \mathbb{R}^{t \times \hat{n}} \qquad \text{and} \qquad \mathbf{p} = \left[\mathbf{0}^T, \mathbf{b}^T, \mathbf{q}^T\right]^T \in \mathbb{R}^{\hat{n}}. \tag{14}$$

Then from the definitions in (8) and (7), we have $d(\mathbf{u}) = -\min_{\mathbf{x}}(f(\mathbf{x}) + \langle \boldsymbol{S}\mathbf{u}, \mathbf{x}\rangle + \langle \mathbf{u}, \mathbf{p}\rangle)$. So we can prove $\mathbf{x}^*(\mathbf{v}^k) = \nabla f^*(-\boldsymbol{S}\mathbf{v}^k)$, $\nabla d(\mathbf{v}^k) = -\boldsymbol{S}^T\mathbf{x}^*(\mathbf{v}^k) - \mathbf{p}$ and $\nabla_i d(\mathbf{v}^k) = -\boldsymbol{S}_i^T\mathbf{x}^*(\mathbf{v}^k) - \mathbf{p}_i$. Thus if we keep a variable $\mathbf{s}_{\mathbf{v}}^k \equiv \boldsymbol{S}\mathbf{v}^k \in \mathbb{R}^t$ and update it without the full matrix-vector multiplication, $\mathbf{x}^*(\mathbf{v}^k)$ and $\nabla_i d(\mathbf{v}^k)$ can be efficiently computed. We describe the explicit update of $\mathbf{s}_{\mathbf{v}}^k$ in Algorithm 1. At each iteration, only the $i_k$-th column of $\boldsymbol{S}$ is used, rather than the full matrix $\boldsymbol{S}$. So Algorithm 1 only needs to deal with the $i_k$-th constraint, rather than all the constraints at each iteration.

The only thing left to do is to compute the gradient of $f^*$ and the proximal mapping of $\phi_i^*$. In machine learning, $f$ is often the regularizer and $\phi_i$ is the loss function. The most commonly used strongly convex regularizer is the $L_2$ regularization $f(\mathbf{x}) = \frac{\mu}{2}\|\mathbf{x}\|^2$. In this case, $f^*(\mathbf{x}) = \frac{1}{2\mu}\|\mathbf{x}\|^2$ and thus the computation time of $\nabla f^*(-\mathbf{s}_{\mathbf{v}}^k)$ is $O(t)$. We can also use some non-smooth strongly convex regularizers in the form of $f(\mathbf{x}) = \frac{\mu}{2}\|\mathbf{x}\|^2 + \sigma(\mathbf{x})$. In this case, $\nabla f^*(\mathbf{x}) = \text{Prox}_{\sigma/\mu}(\mathbf{x}/\mu)$, which can be efficiently computed when the proximal mapping of $\sigma(\mathbf{x})$ has closed form solution. On the other hand, when computing the proximal mapping of $\phi_i^*$ (or the proximal mapping of $\phi_i$ since $u = \text{Prox}_{\phi_i}(u) + \text{Prox}_{\phi_i^*}(u)$), we only need to solve an optimization problem with only one dimension, which can be efficiently done, e.g., by the nonsmooth Newton or quasi Newton method (Lewis and Overton, 2013). For many machine learning problems, the proximal mapping of $\phi_i^*$ can be computed efficiently (see, e.g., (Shalev-Shwartz and Zhang, 2013, 2016)). Thus, Algorithm 1 needs about $O(t)$ time at each iteration while DFGA and ADFGA need $O(t\hat{n})$ time.

**Remark 2** *A main difference between Algorithm 1 and the original accelerated randomized coordinate descent is that Algorithm 1 uses the step-size of $\frac{1}{\hat{n}\theta_k L_{i_k}}$ when computing $\mathbf{z}_{i_k}^{k+1}$ while the original algorithm uses a larger one of $\frac{2}{\hat{n}\theta_k L_{i_k}}$, which makes the original algorithm faster than Algorithm 1 in practice. The reason of the smaller step-size is to fit the proof.*

**Algorithm 1** Accelerated Randomized Dual Coordinate Ascent (ARDCA)

```
Input $\mathbf{u}^0 \in \mathbb{D}$, $K_0$, $K$
Initialize $\mathbf{z}^0 = \mathbf{u}^0$, $\hat{\mathbf{u}}^0 = \mathbf{0}$, $\mathbf{s}_{\mathbf{z}}^0 = \boldsymbol{S}\mathbf{z}^0$, $\mathbf{s}_{\hat{\mathbf{u}}}^0 = 0$, $\theta_0 = \frac{1}{\hat{n}}$.
for $k = 0, 1, 2, 3, \cdots, K$ do
  $\mathbf{s}_{\mathbf{v}}^k = \theta_k^2 \mathbf{s}_{\hat{\mathbf{u}}}^k + \mathbf{s}_{\mathbf{z}}^k$,
  $\mathbf{x}^*(\mathbf{v}^k) = \nabla f^*(-\mathbf{s}_{\mathbf{v}}^k)$,
  select $i_k$ randomly with probability of $1/\hat{n}$,
  $\nabla_{i_k} d(\mathbf{v}^k) = -\boldsymbol{S}_{i_k}^T \mathbf{x}^*(\mathbf{v}^k) - \mathbf{p}_{i_k}$,
  $\mathbf{z}_{i_k}^{k+1} = \operatorname{argmin}_u \hat{n}\theta_k L_{i_k} \|u - \mathbf{z}_{i_k}^k\|^2 + \left\langle \nabla_{i_k} d(\mathbf{v}^k), u - \mathbf{z}_{i_k}^k \right\rangle + h_{i_k}(u)$,
  $\hat{\mathbf{u}}_{i_k}^{k+1} = \hat{\mathbf{u}}_{i_k}^k - \frac{1-\hat{n}\theta_k}{\theta_k^2}(\mathbf{z}_{i_k}^{k+1} - \mathbf{z}_{i_k}^k)$,
  $\mathbf{z}_j^{k+1} = \mathbf{z}_j^k$ and $\hat{\mathbf{u}}_j^{k+1} = \hat{\mathbf{u}}_j^k$ for $j \neq i_k$,
  $\mathbf{s}_{\mathbf{z}}^{k+1} = \mathbf{s}_{\mathbf{z}}^k + \boldsymbol{S}_{i_k}(\mathbf{z}_{i_k}^{k+1} - \mathbf{z}_{i_k}^k)$,
  $\mathbf{s}_{\hat{\mathbf{u}}}^{k+1} = \mathbf{s}_{\hat{\mathbf{u}}}^k + \boldsymbol{S}_{i_k}(\hat{\mathbf{u}}_{i_k}^{k+1} - \hat{\mathbf{u}}_{i_k}^k)$,
  $\theta_{k+1} = \frac{\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k^2}{2}$,
end for
Output $\hat{\mathbf{x}}^K = \frac{\sum_{k=K_0}^K \frac{\mathbf{x}^*(\mathbf{v}^k)}{\theta_k}}{\sum_{k=K_0}^K \frac{1}{\theta_k}}$ and $\mathbf{u}^{K+1} = \theta_K^2 \hat{\mathbf{u}}^{K+1} + \mathbf{z}^{K+1}$.
```

*Specifically, it allows us to keep an additional term* $\frac{\hat{n}^2\theta_k^2}{2}\|\mathbf{z}^{k+1} - \mathbf{z}^k\|_L^2$ *in Lemma 10, which is crucial in the proof of Lemma 12. Otherwise, we may only bound* $\|\mathbb{E}_{\xi_K}[\cdot]\|_L^*$ *for the constraint functions, rather than* $\mathbb{E}_{\xi_K}[\|\cdot\|_L^*]$ *in Lemma 12. The former is less interesting since the expectation is inside the norm.*

In Algorithm 1, we output the average of $\mathbf{x}^*(\mathbf{v}^k)$ from some $K_0$ to $K$[6], rather than $\mathbf{x}^*(\mathbf{u}^{K+1})$. This little change allows us to give a faster convergence rate in the primal space. We average from $K_0$, rather than from the first iteration, since the first few iterations often produce poor solutions. We now state our main result on the convergence rate of the primal solutions for ARDCA. Let

$$\xi_k = \{i_0, i_1, \cdots, i_k\}$$

denote the random sequence, $\mathbb{E}_{\xi_k}$ be the expectation with respect to $\xi_k$ and $\mathbb{E}_{i_k|\xi_{k-1}}$ be the conditional expectation with respect to $i_k$ conditioned on $\xi_{k-1}$, then we have the following theorem.

**Theorem 3** *Suppose Assumption 1 holds. Let* $K_0 \leq \left\lfloor \frac{K}{\upsilon(1+1/\hat{n})} + 1 \right\rfloor$ *with any* $\upsilon > 1$. *Then for Algorithm 1, we have*

$$\left|\mathbb{E}_{\xi_K}[F(\hat{\mathbf{x}}^K)] - F(\mathbf{x}^*)\right| \leq \frac{9\hat{n}^2\left((1-\theta_0)\left(D(\mathbf{u}^0) - D(\mathbf{u}^*)\right) + \|\mathbf{u}^0 - \mathbf{u}^*\|_L^2 + \|\mathbf{u}^*\|_L^2 + M^2\sum_{i=1}^n L_i\right)}{(K^2/4 + \hat{n}K)\,(1 - 1/\upsilon)},$$

$$\mathbb{E}_{\xi_K}\left[\left\|\begin{bmatrix} \boldsymbol{B}\hat{\mathbf{x}}^K + \mathbf{b} \\ \max\left\{0, g(\hat{\mathbf{x}}^K)\right\} \end{bmatrix}\right\|_L^*\right] \leq \frac{7\hat{n}^2}{(K^2/4 + \hat{n}K)(1 - 1/\upsilon)}\sqrt{(1-\theta_0)(D(\mathbf{u}^0) - D(\mathbf{u}^*)) + \|\mathbf{u}^0 - \mathbf{u}^*\|_L^2}.$$

[6] We leave the efficient computation of the average in Appendix A.

Now we compare the convergence rate of the primal solutions with that of the dual solutions. For the dual problem (9), Lin et al. (2015b) proved the $O\left(\frac{\hat{n}^2}{K^2}\right)$ convergence rate, which is described in the following proposition.

**Proposition 4** *(Lin et al., 2015b) Suppose Assumption 1 holds. Then for Algorithm 1, we have*

$$\mathbb{E}_{\xi_K}[D(\mathbf{u}^{K+1})]-D(\mathbf{u}^*)\leq\left(\frac{2\hat{n}}{2\hat{n}+K\hat{n}/\sqrt{\hat{n}^2-1}}\right)^2\left(D(\mathbf{u}^0)-D(\mathbf{u}^*)+\frac{\hat{n}^2}{2(\hat{n}^2-1)}\|\mathbf{u}^0-\mathbf{u}^*\|_L^2\right).\tag{15}$$

Thus we can see that Algorithm 1 needs $O\left(\frac{\hat{n}}{\sqrt{\epsilon}}\right)$ iterations to achieve an $\epsilon$-optimal primal solution and dual solution, i.e., the iteration complexity of the primal solutions has the same order of magnitude as that of the dual solutions for ARDCA.

To make a better comparison to the existing result, we describe the relation of the primal objective and constraint functions with the dual objective without averaging the primal solutions in the following proposition, which only applies to problem (3). Combing with (15), we can immediately get the $O\left(\frac{\hat{n}}{K+\hat{n}}\right)$ convergence rate in the primal space, which verifies that averaging the primal solutions helps to improve the convergence rate in the primal space.

**Proposition 5** *(Lu and Johansson, 2016) Suppose Assumption 1 holds for problem (3). Then for Algorithm 1, we have*

$$\mathbb{E}_{\xi_K}[f(\mathbf{x}^*(\mathbf{u}^{K+1}))]-f(\mathbf{x}^*)\leq\left(\|\mathbf{u}^{K+1}\|_\infty\sqrt{2L\hat{n}}+\sqrt{\mathbb{E}_{\xi_K}[D(\mathbf{u}^{K+1})]-D(\mathbf{u}^*)}\right)\sqrt{\mathbb{E}_{\xi_K}[D(\mathbf{u}^{K+1})]-D(\mathbf{u}^*)},$$

$$\mathbb{E}_{\xi_K}[f(\mathbf{x}^*(\mathbf{u}^{K+1}))]-f(\mathbf{x}^*)\geq-\|\mathbf{u}^*\|\sqrt{2L\left(\mathbb{E}_{\xi_K}[D(\mathbf{u}^{K+1})]-D(\mathbf{u}^*)\right)},$$

$$\mathbb{E}_{\xi_K}\left[\left\|\begin{bmatrix}\boldsymbol{B}\mathbf{x}^*(\mathbf{u}^{K+1})+\mathbf{b}\\ \max\left\{0,g(\mathbf{x}^*(\mathbf{u}^{K+1}))\right\}\end{bmatrix}\right\|_L^*\right]\leq\sqrt{2L\left(\mathbb{E}_{\xi_K}[D(\mathbf{u}^{K+1})]-D(\mathbf{u}^*)\right)}.$$

## 3. Convergence Rate Analysis of the Primal Solutions

In this section, we prove Theorem 3. First we study a simple case in Section 3.1 to intuitively show how to relate the primal objective with the dual objective and why average helps to improve the convergence rate. Then we give the detailed analysis for the general case in Section 3.2.

### 3.1. Intuition: A Case Study

In this section, we study a simple case of problem (1):

$$\min_{\mathbf{x}\in\mathbb{R}^t} f(\mathbf{x}),\quad s.t.\quad \boldsymbol{B}\mathbf{x}=\mathbf{b}.$$

We only consider the gradient descent to solve the dual problem for simplicity, which has the recursion of

$$\mathbf{u}^{k+1}=\mathbf{u}^k-\frac{1}{L}\nabla d(\mathbf{u}^k).\tag{16}$$

Then $d(\mathbf{u})$ and $\nabla d(\mathbf{u})$ reduce to

$$\begin{aligned} d(\mathbf{u}) &= -f(\mathbf{x}^*(\mathbf{u})) - \langle \mathbf{u}, \boldsymbol{B}\mathbf{x}^*(\mathbf{u}) - \mathbf{b} \rangle , \\ \nabla d(\mathbf{u}) &= -(\boldsymbol{B}\mathbf{x}^*(\mathbf{u}) - \mathbf{b}), \end{aligned} \tag{17}$$

which further leads to

$$d(\mathbf{u}) - \langle \mathbf{u}, \nabla d(\mathbf{u}) \rangle = -f(\mathbf{x}^*(\mathbf{u})). \tag{18}$$

(18) is a crucial property to relate the primal objective and dual objective. From the $L$-smoothness of $d(\mathbf{u})$, we have

$$d(\mathbf{u}^{k+1}) \leq d(\mathbf{u}^k) + \left\langle \nabla d(\mathbf{u}^k), \mathbf{u}^{k+1} - \mathbf{u}^k \right\rangle + \frac{L}{2}\|\mathbf{u}^{k+1} - \mathbf{u}^k\|^2.$$

From $-f(\mathbf{x}^*) = d(\mathbf{u}^*) \leq d(\mathbf{u}^{k+1})$, (18), (17) and (16), we have

$$\begin{aligned} -f(\mathbf{x}^*) \leq & - f(\mathbf{x}^*(\mathbf{u}^k)) + \left\langle \nabla d(\mathbf{u}^k), \mathbf{u}^{k+1} \right\rangle + \frac{L}{2}\|\mathbf{u}^{k+1} - \mathbf{u}^k\|^2 \\ = & - f(\mathbf{x}^*(\mathbf{u}^k)) - \left\langle \boldsymbol{B}\mathbf{x}^*(\mathbf{u}^k) - \mathbf{b}, \mathbf{u} \right\rangle + \left\langle \nabla d(\mathbf{u}^k), \mathbf{u}^{k+1} - \mathbf{u} \right\rangle + \frac{L}{2}\|\mathbf{u}^{k+1} - \mathbf{u}^k\|^2 \\ = & - f(\mathbf{x}^*(\mathbf{u}^k)) - \left\langle \boldsymbol{B}\mathbf{x}^*(\mathbf{u}^k) - \mathbf{b}, \mathbf{u} \right\rangle - L\left\langle \mathbf{u}^{k+1} - \mathbf{u}^k, \mathbf{u}^{k+1} - \mathbf{u} \right\rangle + \frac{L}{2}\|\mathbf{u}^{k+1} - \mathbf{u}^k\|^2 \\ = & - f(\mathbf{x}^*(\mathbf{u}^k)) - \left\langle \boldsymbol{B}\mathbf{x}^*(\mathbf{u}^k) - \mathbf{b}, \mathbf{u} \right\rangle + \frac{L}{2}\|\mathbf{u}^k - \mathbf{u}\|^2 - \frac{L}{2}\|\mathbf{u}^{k+1} - \mathbf{u}\|^2. \end{aligned}$$

Define $\hat{\mathbf{x}}^K = \frac{\sum_{k=0}^K \mathbf{x}^*(\mathbf{u}^k)}{K+1}$, letting $\mathbf{u} = \mathbf{u}^*$ and summing over $k = 0, 1, \cdots, K$, we have

$$\begin{aligned} f(\hat{\mathbf{x}}^K) + \left\langle \boldsymbol{B}\hat{\mathbf{x}}^K - \mathbf{b}, \mathbf{u}^* \right\rangle - f(\mathbf{x}^*) \overset{a}{\leq} & \frac{\sum_{k=0}^K \left( f(\mathbf{x}^*(\mathbf{u}^k)) + \left\langle \boldsymbol{B}\mathbf{x}^*(\mathbf{u}^k) - \mathbf{b}, \mathbf{u}^* \right\rangle \right)}{K+1} - f(\mathbf{x}^*) \\ \leq & \frac{L}{2(K+1)}\|\mathbf{u}^0 - \mathbf{u}^*\|^2, \end{aligned}$$

where we use Jensen's inequality for $f(\mathbf{x})$ in $\overset{a}{\leq}$. On the other hand, from (17) and (16), we have

$$\begin{aligned} &\|\boldsymbol{B}\hat{\mathbf{x}}^K - \mathbf{b}\| = \frac{1}{K+1}\left\|\sum_{k=0}^K (\boldsymbol{B}\mathbf{x}^*(\mathbf{u}^k) - \mathbf{b})\right\| = \frac{1}{K+1}\left\|\sum_{k=0}^K \nabla d(\mathbf{u}^k)\right\| = \frac{L}{K+1}\left\|\sum_{k=0}^K (\mathbf{u}^{k+1} - \mathbf{u}^k)\right\| \\ &= \frac{L}{K+1}\|\mathbf{u}^{K+1} - \mathbf{u}^0\| \leq \frac{L}{K+1}\left(\|\mathbf{u}^0 - \mathbf{u}^*\| + \|\mathbf{u}^{K+1} - \mathbf{u}^*\|\right) \overset{b}{\leq} \frac{2L}{K+1}\|\mathbf{u}^0 - \mathbf{u}^*\|, \end{aligned} \tag{19}$$

where we use the non-increasing of $\|\mathbf{u}^k - \mathbf{u}^*\|$ for gradient descent in $\overset{b}{\leq}$. Thus, from the above two inequalities, we have

$$f(\hat{\mathbf{x}}^K) - f(\mathbf{x}^*) \leq O\left(\frac{1}{K}\right) \quad \text{and} \quad \|\boldsymbol{B}\hat{\mathbf{x}}^K - \mathbf{b}\| \leq O\left(\frac{1}{K}\right),$$

which gives the $O\left(\frac{1}{K}\right)$ convergence rate of the primal solutions. When replacing gradient descent with accelerated gradient descent, the convergence rate can be improved to $O\left(\frac{1}{K^2}\right)$. However, more efforts are required for the analysis in the dual space and the averaging weights should be designed carefully. When solving the general problem (1), we should consider the separable part $\phi_i(\boldsymbol{A}_i^T\mathbf{x})$ and the inequality constraints more carefully. When using the accelerated randomized dual coordinate ascent, we should pay more efforts to deal with the expectation, especially that the deduction in (19) is not enough to deal with the constraint functions and it requires more skillful analysis. We give the detailed analysis in the following section.

From the above analysis, we can see that (18) is a critical trick in our analysis. To show the importance of averaging the primal solutions, we give a simple convergence rate analysis measured at the non-averaged primal solution, which is adapted from (Lu and Johansson, 2016). From (18) and $-f(\mathbf{x}^*) = d(\mathbf{u}^*)$, we have

$$f(\mathbf{x}^*(\mathbf{u})) - f(\mathbf{x}^*) = d(\mathbf{u}^*) - d(\mathbf{u}) + \langle \mathbf{u}, \nabla d(\mathbf{u}) \rangle \leq \langle \mathbf{u}, \nabla d(\mathbf{u}) \rangle \leq \sqrt{\hat{n}} \|\mathbf{u}\|_\infty \|\nabla d(\mathbf{u})\|.$$

So we only need to bound $\|\nabla d(\mathbf{u})\|$. From the smoothness of $d(\mathbf{u})$, we have

$$\begin{aligned} & d(\mathbf{u}) - d(\mathbf{u}^*) \geq d(\mathbf{u}) - d(\mathbf{u} + \nabla d(\mathbf{u})/L) \\ & \geq \langle \nabla d(\mathbf{u}), \mathbf{u} + \nabla d(\mathbf{u})/L - \mathbf{u} \rangle - \frac{L}{2}\|\mathbf{u} + \nabla d(\mathbf{u})/L - \mathbf{u}\|^2 = \frac{1}{2L}\|\nabla d(\mathbf{u})\|^2. \end{aligned}$$

Thus, we have $\|\nabla d(\mathbf{u})\| \leq \sqrt{2L(d(\mathbf{u}) - d(\mathbf{u}^*))}$. Since $d(\mathbf{u}^K) - d(\mathbf{u}^*) \leq O\left(\frac{1}{K}\right)$ for gradient descent and (17), we have

$$f(\mathbf{x}^*(\mathbf{u}^K)) - f(\mathbf{x}^*) \leq O\left(\frac{1}{\sqrt{K}}\right) \quad \text{and} \quad \|\boldsymbol{B}\mathbf{x}^*(\mathbf{u}^K) - \mathbf{b}\| \leq O\left(\frac{1}{\sqrt{K}}\right).$$

Generally speaking, average helps to improve the convergence rate for many algorithms. Typical examples include the Douglas-Rachford splitting and ADMM (Davis and Yin, 2017). Davis and Yin (2017) proved that for the two algorithms, the $O\left(\frac{1}{\sqrt{K}}\right)$ and $O\left(\frac{1}{K}\right)$ rates are tight for the non-averaged and averaged solutions, respectively.

From the above analysis for the non-averaged solutions, we can see that $\|\nabla d(\mathbf{u})\|$ serves as a measure of the optimality and feasibility of the primal solutions. Nesterov (2012b) studied how to make gradient small. Specifically, to find a point $\mathbf{u}$ with $\|\nabla d(\mathbf{u})\| \leq \epsilon$, accelerated dual ascent needs $O\left(\frac{1}{\epsilon^{2/3}}\right)$ iterations. When using some regularization technique, the $O\left(\frac{1}{\sqrt{\epsilon}}\log\frac{1}{\epsilon}\right)$ complexity can be attained. We can see that none of them reaches the optimal $O\left(\frac{1}{\sqrt{\epsilon}}\right)$ complexity. On the other hand, when averaging the primal solutions, we do not need to make the gradient small. Instead, we only need to make the average of gradients small, i.e., $\frac{1}{K+1}\left\|\sum_{k=0}^K \nabla d(\mathbf{u}^k)\right\|$ in (19). This is the reason why average helps to improve the convergence rate.

### 3.2. ARPCA for the General Problem

To better analyze the method, we give an equivalent algorithm of ARDCA and describe it in Algorithm 2. In Algorithm 2, variables $\widetilde{\mathbf{z}}_j^k, \forall j \neq i_k$, are only used for analysis. In practice, we do not need to compute $\widetilde{\mathbf{z}}_j^k, \forall j \neq i_k$.

**Algorithm 2** Equivalent ARDCA only for analysis

```
Initialize $\mathbf{z}^0 = \mathbf{u}^0 \in \mathbb{D}$, $\theta_0 = \frac{1}{\hat{n}}$.
for $k = 0, 1, 2, 3, \cdots$ do
    $\mathbf{v}^k = \theta_k \mathbf{z}^k + (1-\theta_k)\mathbf{u}^k$,
    for $i = 1, 2, \cdots, \hat{n}$ do
        $\widetilde{\mathbf{z}}_i^k = \operatorname{argmin}_u \hat{n}\theta_k L_i \|u - \mathbf{z}_i^k\|^2 + \left\langle \nabla_i d(\mathbf{v}^k), u - \mathbf{z}_i^k \right\rangle + h_i(u)$,
    end for
    select $i_k$ randomly with probability $\frac{1}{\hat{n}}$,
    $\mathbf{z}_{i_k}^{k+1} = \widetilde{\mathbf{z}}_{i_k}^k$,
    $\mathbf{z}_j^{k+1} = \mathbf{z}_j^k$ for $j \neq i_k$,
    $\mathbf{u}^{k+1} = \mathbf{v}^k + \hat{n}\theta_k(\mathbf{z}^{k+1} - \mathbf{z}^k)$,
    $\theta_{k+1} = \frac{\sqrt{\theta_k^4 + 4\theta_k^2} - \theta_k^2}{2}$.
end for
Output $\hat{\mathbf{x}}^K = \frac{\sum_{k=K_0}^K \frac{\mathbf{x}^*(\mathbf{v}^k)}{\theta_k}}{\sum_{k=K_0}^K \frac{1}{\theta_k}}$ and $\mathbf{u}^{K+1}$.
```

The definition of $\{\theta_0, \theta_1, \cdots, \theta_K\}$ satisfies $\frac{1-\theta_k}{\theta_k^2} = \frac{1}{\theta_{k-1}^2}$. Define $\theta_{-1} = 1/\sqrt{\hat{n}^2 - \hat{n}}$, which also satisfies $\frac{1-\theta_k}{\theta_k^2} = \frac{1}{\theta_{k-1}^2}$ for $k = 0$. For the sequence $\{\theta_0, \theta_1, \cdots, \theta_K\}$, we can simply prove the following properties.

**Lemma 6** *For the sequence $\{\theta_0, \theta_1, \cdots, \theta_K\}$ satisfying $\theta_0 = \frac{1}{\hat{n}}$ and $\frac{1-\theta_k}{\theta_k^2} = \frac{1}{\theta_{k-1}^2}, \forall k \geq 0$, we have*

1. $0 \leq \theta_k \leq \theta_{k-1} \leq \cdots \leq \theta_1 \leq \theta_0 = \frac{1}{\hat{n}}$.

2. $\sum_{k=K_0}^K \frac{1}{\theta_k} = \frac{1}{\theta_K^2} - \frac{1}{\theta_{K_0-1}^2}$.

3. $\frac{k}{2} + \frac{k}{2\hat{n}} + \hat{n} \geq \frac{1}{\theta_k} \geq \frac{k}{2} + \hat{n}$.

4. *Letting $K_0 \leq \left\lfloor \frac{K}{\upsilon(1+1/\hat{n})} + 1 \right\rfloor$, we have $\frac{1}{\theta_K^2} - \frac{1}{\theta_{K_0-1}^2} \geq \left(\frac{K^2}{4} + \hat{n}K\right)\left(1 - \frac{1}{\upsilon}\right)$ for any $\upsilon > 1$.*

We follow (Fercoq and Richtárik, 2015) to define the sequence $\{\alpha_{k,t} : 0 \leq t \leq k, k = 0, 1, \cdots\}$ satisfying

$$\alpha_{0,0} = 1, \alpha_{1,t} = \begin{cases} 1 - \hat{n}\theta_0, & t = 0, \\ \hat{n}\theta_0, & t = 1, \end{cases} \quad \alpha_{k+1,t} = \begin{cases} (1-\theta_k)\alpha_{k,t}, & t \leq k-1, \\ (1-\theta_k)\alpha_{k,k} - (\hat{n}-1)\theta_k, & t = k, \\ \hat{n}\theta_k, & t = k+1. \end{cases}$$

From Lemma 2 in (Fercoq and Richtárik, 2015), we have $0 \leq \alpha_{k,t} \leq 1, \forall t = 0, \cdots, k$, $\sum_{t=0}^k \alpha_{k,t} = 1$ and $\mathbf{u}^{k+1} = \sum_{t=0}^{k+1} \alpha_{k+1,t}\mathbf{z}^t$. Define $H^{k+1} = \sum_{t=0}^{k+1} \alpha_{k+1,t} h(\mathbf{z}^t)$, then $h(\mathbf{u}^{k+1}) \leq H^{k+1}$ due to Jensen's inequality for $h(\mathbf{x})$. We can easily verify that variables $\mathbf{z}^k$, $\mathbf{u}^k$ and $\mathbf{v}^k$ remain in $\mathbb{D}$ at all times by induction, described in the following lemma. Then we can use the Lipschitz smooth condition described in Lemma 1.

**Lemma 7** *For Algorithm 2, we have* $\mathbf{z}^k \in \mathbb{D}$*,* $\mathbf{u}^k \in \mathbb{D}$ *and* $\mathbf{v}^k \in \mathbb{D}, \forall k \geq 0$*.*

For Algorithm 2, let

$$\mathbf{y}_i^k = -2\hat{n}\theta_k L_i(\widetilde{\mathbf{z}}_i^k - \mathbf{z}_i^k) - \nabla_i d(\mathbf{v}^k), \quad 1 \leq i \leq n. \tag{20}$$

From the optimality condition of $\widetilde{\mathbf{z}}_i^k$ in Algorithm 2, we have $\mathbf{y}_i^k \in \partial h_i(\widetilde{\mathbf{z}}_i^k), 1 \leq i \leq n$. Define

$$\sigma_1(\mathbf{u}_i, \widetilde{\mathbf{z}}_i^k) = \begin{cases} h_i(\widetilde{\mathbf{z}}_i^k) + \left\langle \mathbf{y}_i^k, \mathbf{u}_i - \widetilde{\mathbf{z}}_i^k \right\rangle - h_i(\mathbf{u}_i), & \text{if } i \leq n, \\ 0, & \text{if } n < i \leq n+p+m, \end{cases} \tag{21}$$

and

$$\sigma_2(\mathbf{u}, \mathbf{v}^k) = d(\mathbf{v}^k) + \left\langle \nabla d(\mathbf{v}^k), \mathbf{u} - \mathbf{v}^k \right\rangle - d(\mathbf{u}). \tag{22}$$

From the convexity of $h_i$ and $d$, we have $\sigma_1(\mathbf{u}_i, \widetilde{\mathbf{z}}_i^k) \leq 0$ and $\sigma_2(\mathbf{u}, \mathbf{v}^k) \leq 0$. We use $\sigma_1(\mathbf{u}_i, \widetilde{\mathbf{z}}_i^k)$ and $\sigma_2(\mathbf{u}, \mathbf{v}^k)$ to relate the primal objective function, primal constraint functions and dual objective function in the following lemma.

**Lemma 8** *Suppose Assumption 1 holds. For any* $\mathbf{u} \in \mathbb{D}$*, we have*

$$-\sum_{i=1}^n \sigma_1(\mathbf{u}_i, \widetilde{\mathbf{z}}_i^k) - \sigma_2(\mathbf{u}, \mathbf{v}^k) = \left\langle \triangle(\mathbf{x}^*(\mathbf{v}^k), n\mathbf{y}^k), \mathbf{u} \right\rangle + D(\mathbf{u}) + f(\mathbf{x}^*(\mathbf{v}^k)) + \frac{1}{n}\phi(n\mathbf{y}^k), \tag{23}$$

*where*

$$\triangle(\mathbf{x}, \mathbf{y}) = \left[(\boldsymbol{A}^T\mathbf{x} - \mathbf{y})^T/n, (\boldsymbol{B}\mathbf{x} + \mathbf{b})^T, g_1(\mathbf{x}), \cdots, g_m(\mathbf{x})\right]^T.$$

**Proof** From (10), (7) and the definition of $d(\mathbf{u})$ in (8), we have

$$f(\mathbf{x}^*(\mathbf{u})) = -d(\mathbf{u}) + \nabla d(\mathbf{u})^T \mathbf{u}.$$

Thus, by the definition of $\sigma_2(\mathbf{u}, \mathbf{v}^k)$ in (22), we have

$$\begin{aligned} \sigma_2(\mathbf{u}, \mathbf{v}^k) =& \left\langle \nabla d(\mathbf{v}^k), \mathbf{u} \right\rangle - d(\mathbf{u}) - \left[ \left\langle \nabla d(\mathbf{v}^k), \mathbf{v}^k \right\rangle - d(\mathbf{v}^k) \right] \\ =& \left\langle \nabla d(\mathbf{v}^k), \mathbf{u} \right\rangle - d(\mathbf{u}) - f(\mathbf{x}^*(\mathbf{v}^k)). \end{aligned}$$

From the definition of $\sigma_1(\mathbf{u}_i, \widetilde{\mathbf{z}}_i^k)$ in (21) and the fact that $\mathbf{y}_i^k \in \partial h_i(\widetilde{\mathbf{z}}_i^k), 1 \leq i \leq n$, we can also have

$$\begin{aligned} \sum_{i=1}^n \sigma_1(\mathbf{u}_i, \widetilde{\mathbf{z}}_i^k) =& \sum_{i=1}^n \left( \left\langle \mathbf{y}_i^k, \mathbf{u}_i \right\rangle - h_i(\mathbf{u}_i) - \left[ \left\langle \mathbf{y}_i^k, \widetilde{\mathbf{z}}_i^k \right\rangle - h_i(\widetilde{\mathbf{z}}_i^k) \right] \right) \\ \overset{a}{=}& \sum_{i=1}^n \left( \left\langle \mathbf{y}_i^k, \mathbf{u}_i \right\rangle - h_i(\mathbf{u}_i) - h_i^*(\mathbf{y}_i^k) \right) \\ \overset{b}{=}& \sum_{i=1}^n \left( \left\langle \mathbf{y}_i^k, \mathbf{u}_i \right\rangle - h_i(\mathbf{u}_i) - \frac{1}{n}\phi_i(n\mathbf{y}_i^k) \right), \end{aligned}$$

where we use the definition of conjugate in $\overset{a}{=}$ and the fact that $\varphi(x) = \alpha\psi(x) \Rightarrow \varphi^*(y) = \alpha\psi^*(y/\alpha)$ and $h_i(\mathbf{u}_i) = \frac{1}{n}\phi_i^*(\mathbf{u}_i), \forall i \leq n$ in $\overset{b}{=}$. So the result of (23) immediately follows by adding the above two equations and using (10). ■

### 3.2.1. Primal Objective

In the following lemma, we use the relation (23) to bound the Lagrangian function $L_F(\hat{\mathbf{x}}^K, \hat{\mathbf{y}}^K, \mathbf{u})$. We also bound $\sum_{k=K_0}^K \mathbb{E}_{\xi_K}[\|\mathbf{z}^{k+1} - \mathbf{z}^k\|_L^2]$ and $\mathbb{E}_{\xi_K}[\|\mathbf{z}^{K+1} - \mathbf{u}^*\|_L^2]$, which will be used to bound the constraint functions later. As a by-product, we also give the convergence rate of the dual solutions.

**Lemma 9** *Suppose Assumption 1 holds. Define* $\hat{\mathbf{x}}^K = \frac{\sum_{k=K_0}^K \frac{\mathbf{x}^*(\mathbf{v}^k)}{\theta_k}}{\sum_{k=K_0}^K \frac{1}{\theta_k}}$ *and* $\hat{\mathbf{y}}^K = \frac{\sum_{k=K_0}^K \frac{n\mathbf{y}^k}{\theta_k}}{\sum_{k=K_0}^K \frac{1}{\theta_k}}$. *Then we have*

$$\begin{aligned}&\left(\frac{1}{\theta_K^2} - \frac{1}{\theta_{K_0-1}^2}\right)\mathbb{E}_{\xi_K}\left[\left\langle \triangle(\hat{\mathbf{x}}^K, \hat{\mathbf{y}}^K), \mathbf{u}\right\rangle + D(\mathbf{u}^*) + f(\hat{\mathbf{x}}^K) + \frac{1}{n}\phi(\hat{\mathbf{y}}^K)\right]\\ \leq& 2(\hat{n}^2 - \hat{n})\left(D(\mathbf{u}^0) - D(\mathbf{u}^*)\right) + 2\hat{n}^2\|\mathbf{u}^0 - \mathbf{u}^*\|_L^2 + 2\hat{n}^2\|\mathbf{u} - \mathbf{u}^*\|_L^2\end{aligned} \tag{24}$$

*for any* $\mathbf{u} \in \mathbb{D}$ *independent on* $\xi_K$. *We also have*

$$\frac{1}{2}\sum_{k=K_0}^K \mathbb{E}_{\xi_K}[\|\mathbf{z}^{k+1} - \mathbf{z}^k\|_L^2] \leq (1-\theta_0)\left(D(\mathbf{u}^0) - D(\mathbf{u}^*)\right) + \|\mathbf{u}^0 - \mathbf{u}^*\|_L^2, \tag{25}$$

$$\mathbb{E}_{\xi_K}[\|\mathbf{z}^{K+1} - \mathbf{u}^*\|_L^2] \leq (1-\theta_0)\left(D(\mathbf{u}^0) - D(\mathbf{u}^*)\right) + \|\mathbf{u}^0 - \mathbf{u}^*\|_L^2, \tag{26}$$

$$\frac{\mathbb{E}_{\xi_K}[D(\mathbf{u}^{K+1})] - D(\mathbf{u}^*)}{\theta_K^2} + \hat{n}^2\|\mathbf{z}^{K+1} - \mathbf{u}^*\|_L^2 \leq \hat{n}^2\left((1-\theta_0)\left(D(\mathbf{u}^0) - D(\mathbf{u}^*)\right) + \|\mathbf{u}^0 - \mathbf{u}^*\|_L^2\right). \tag{27}$$

The proof of Lemma 9 is based on the following lemma, which gives the progress in one iteration in the dual space. Lemma 10 can be proved by the techniques in the proof of Theorem 3 in (Fercoq and Richtárik, 2015), except that we keep the additional terms $\sum_{i=1}^n \sigma_1(\mathbf{u}_i, \widetilde{\mathbf{z}}_i^k) + \sigma_2(\mathbf{u}, \mathbf{v}^k)$ and $\|\mathbf{z}^{k+1} - \mathbf{z}^k\|_L^2$. We leave the proof of Lemma 10 in Appendix D.

**Lemma 10** *Suppose Assumption 1 holds. Then we have*

$$\begin{aligned}&\mathbb{E}_{\xi_K}\left[d(\mathbf{u}^{k+1}) + H^{k+1} - D(\mathbf{u}) + \hat{n}^2\theta_k^2\|\mathbf{z}^{k+1} - \mathbf{u}\|_L^2 + \frac{\hat{n}^2\theta_k^2}{2}\|\mathbf{z}^{k+1} - \mathbf{z}^k\|_L^2\right]\\ \leq& (1-\theta_k)\mathbb{E}_{\xi_K}[d(\mathbf{u}^k) + H^k - D(\mathbf{u})] + \hat{n}^2\theta_k^2\mathbb{E}_{\xi_K}[\|\mathbf{z}^k - \mathbf{u}\|_L^2] + \theta_k\mathbb{E}_{\xi_K}\left[\sum_{i=1}^n \sigma_1(\mathbf{u}_i, \widetilde{\mathbf{z}}_i^k) + \sigma_2(\mathbf{u}, \mathbf{v}^k)\right]\end{aligned} \tag{28}$$

*for any* $\mathbf{u} \in \mathbb{D}$ *independent on* $\xi_K$.

Based on Lemma 10, we are ready to prove Lemma 9.

**Proof** Dividing both sides of (28) by $\theta_k^2$ and using $\frac{1-\theta_k}{\theta_k^2} = \frac{1}{\theta_{k-1}^2}$, we have

$$\begin{aligned}&\frac{\mathbb{E}_{\xi_K}[d(\mathbf{u}^{k+1}) + H^{k+1} - D(\mathbf{u})]}{\theta_k^2} + \hat{n}^2\mathbb{E}_{\xi_K}[\|\mathbf{z}^{k+1} - \mathbf{u}\|_L^2] + \frac{\hat{n}^2}{2}\mathbb{E}_{\xi_K}[\|\mathbf{z}^{k+1} - \mathbf{z}^k\|_L^2]\\ \leq& \frac{\mathbb{E}_{\xi_K}[d(\mathbf{u}^k) + H^k - D(\mathbf{u})]}{\theta_{k-1}^2} + \hat{n}^2\mathbb{E}_{\xi_K}[\|\mathbf{z}^k - \mathbf{u}\|_L^2] + \frac{\mathbb{E}_{\xi_K}\left[\sum_{i=1}^n \sigma_1(\mathbf{u}_i, \widetilde{\mathbf{z}}_i^k) + \sigma_2(\mathbf{u}, \mathbf{v}^k)\right]}{\theta_k}.\end{aligned} \tag{29}$$

Letting $\mathbf{u} = \mathbf{u}^*$, we know $\frac{\mathbb{E}_{\xi_K}[d(\mathbf{u}^k)+H^k]-D(\mathbf{u}^*)}{\theta_{k-1}^2} + \hat{n}^2\mathbb{E}_{\xi_K}[\|\mathbf{z}^k - \mathbf{u}^*\|_L^2]$ is decreasing since $\sigma_1(\mathbf{u}_i, \mathbf{z}_i^{k+1}) \leq 0$ and $\sigma_2(\mathbf{u}, \mathbf{v}_i^k) \leq 0, \forall \mathbf{u}$. Thus, we have

$$\mathbb{E}_{\xi_K}\left[\frac{d(\mathbf{u}^{K_0})+H^{K_0}-D(\mathbf{u}^*)}{\theta_{K_0-1}^2}+\hat{n}^2\|\mathbf{z}^{K_0}-\mathbf{u}^*\|_L^2\right]\leq(\hat{n}^2-\hat{n})\left(D(\mathbf{u}^0)-D(\mathbf{u}^*)\right)+\hat{n}^2\|\mathbf{z}^0-\mathbf{u}^*\|_L^2, \tag{30}$$

where we use $H^0 = h(\mathbf{u}^0)$ and $\frac{1}{\theta_{-1}^2} = \hat{n}^2 - \hat{n}$. Summing (29) over $k = K_0, K_0+1, \cdots, K$ and using $h(\mathbf{u}^{k+1}) \leq H^{K+1}$, we have

$$\begin{aligned}
&\frac{\mathbb{E}_{\xi_K}[D(\mathbf{u}^{K+1}) - D(\mathbf{u})]}{\theta_K^2} + \hat{n}^2\mathbb{E}_{\xi_K}[\|\mathbf{z}^{K+1} - \mathbf{u}\|_L^2] \\
\leq&\frac{\mathbb{E}_{\xi_K}[d(\mathbf{u}^{K_0}) + H^{K_0} - D(\mathbf{u})]}{\theta_{K_0-1}^2} + \hat{n}^2\mathbb{E}_{\xi_K}[\|\mathbf{z}^{K_0} - \mathbf{u}\|_L^2] \\
&+ \sum_{k=K_0}^{K}\frac{\mathbb{E}_{\xi_K}\left[\sum_i \sigma_1(\mathbf{u}_i, \widetilde{\mathbf{z}}_i^k) + \sigma_2(\mathbf{u}, \mathbf{v}^k)\right]}{\theta_k} - \frac{\hat{n}^2}{2}\sum_{k=K_0}^{K}\mathbb{E}_{\xi_K}[\|\mathbf{z}^{k+1} - \mathbf{z}^k\|_L^2].
\end{aligned}$$

Letting $\mathbf{u} = \mathbf{u}^*$, from (30), $\sigma_1(\mathbf{u}_i, \widetilde{\mathbf{z}}_i^k) \leq 0$ and $\sigma_2(\mathbf{u}, \mathbf{v}_i^k) \leq 0$, we can immediately have (25), (26) and (27). On the other hand, we also have

$$\begin{aligned}
&-\sum_{k=K_0}^{K}\frac{\mathbb{E}_{\xi_K}\left[\sum_i \sigma_1(\mathbf{u}_i, \widetilde{\mathbf{z}}_i^k) + \sigma_2(\mathbf{u}, \mathbf{v}^k)\right]}{\theta_k} \\
\leq&\frac{\mathbb{E}_{\xi_K}[d(\mathbf{u}^{K_0}) + H^{K_0} - D(\mathbf{u})]}{\theta_{K_0-1}^2} + \hat{n}^2\mathbb{E}_{\xi_K}[\|\mathbf{z}^{K_0} - \mathbf{u}\|_L^2] - \frac{\mathbb{E}_{\xi_K}[D(\mathbf{u}^{K+1})] - D(\mathbf{u})}{\theta_K^2}.
\end{aligned} \tag{31}$$

From (23), for any $\mathbf{u} \in \mathbb{D}$ we have

$$\begin{aligned}
&-\sum_{k=K_0}^{K}\frac{\mathbb{E}_{\xi_K}\left[\sum_{i=1}^n \sigma_1(\mathbf{u}_i, \widetilde{\mathbf{z}}_i^k) + \sigma_2(\mathbf{u}, \mathbf{v}^k)\right]}{\theta_k} \\
=&\left\langle\sum_{k=K_0}^{K}\frac{\mathbb{E}_{\xi_K}[\boldsymbol{A}^T\mathbf{x}^*(\mathbf{v}^k) - n\mathbf{y}^k]}{n\theta_k}, \mathbf{u}_{1:n}\right\rangle + \left\langle\sum_{k=K_0}^{K}\frac{\mathbb{E}_{\xi_K}[\boldsymbol{B}\mathbf{x}^*(\mathbf{v}^k) + \mathbf{b}]}{\theta_k}, \mathbf{u}_{n+1:n+p}\right\rangle \\
&+\left\langle\left[\sum_{k=K_0}^{K}\frac{\mathbb{E}_{\xi_K}[g_1(\mathbf{x}^*(\mathbf{v}^k))]}{\theta_k}, \cdots, \sum_{k=K_0}^{K}\frac{\mathbb{E}_{\xi_K}[g_m(\mathbf{x}^*(\mathbf{v}^k))]}{\theta_k}\right]^T, \mathbf{u}_{n+p+1:n+p+m}\right\rangle \\
&+\sum_{k=K_0}^{K}\frac{D(\mathbf{u})}{\theta_k} + \sum_{k=K_0}^{K}\frac{\mathbb{E}_{\xi_K}[f(\mathbf{x}^*(\mathbf{v}^k))]}{\theta_k} + \frac{1}{n}\sum_{k=K_0}^{K}\frac{\mathbb{E}_{\xi_K}[\phi(n\mathbf{y}^k)]}{\theta_k}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{k=K_0}^{K} \frac{1}{\theta_k}\left[\left\langle \frac{\sum_{k=K_0}^{K}\frac{\mathbb{E}_{\xi_K}[\boldsymbol{A}^T\mathbf{x}^*(\mathbf{v}^k)-n\mathbf{y}^k]}{n\theta_k}}{\sum_{k=K_0}^{K}\frac{1}{\theta_k}}, \mathbf{u}_{1:n}\right\rangle + \left\langle \frac{\sum_{k=K_0}^{K}\frac{\mathbb{E}_{\xi_K}[\boldsymbol{B}\mathbf{x}^*(\mathbf{v}^k)+\mathbf{b}]}{\theta_k}}{\sum_{k=K_0}^{K}\frac{1}{\theta_k}}, \mathbf{u}_{n+1:n+p}\right\rangle\right]\\
&\quad + \sum_{k=K_0}^{K} \frac{1}{\theta_k}\left\langle \left[\frac{\sum_{k=K_0}^{K}\frac{\mathbb{E}_{\xi_K}[g_1(\mathbf{x}^*(\mathbf{v}^k))]}{\theta_k}}{\sum_{k=K_0}^{K}\frac{1}{\theta_k}}, \cdots, \frac{\sum_{k=K_0}^{K}\frac{\mathbb{E}_{\xi_K}[g_m(\mathbf{x}^*(\mathbf{v}^k))]}{\theta_k}}{\sum_{k=K_0}^{K}\frac{1}{\theta_k}}\right]^T, \mathbf{u}_{n+p+1:n+p+m}\right\rangle\\
&\quad + \sum_{k=K_0}^{K}\frac{D(\mathbf{u})}{\theta_k} + \sum_{k=K_0}^{K}\frac{1}{\theta_k}\frac{\sum_{k=K_0}^{K}\frac{\mathbb{E}_{\xi_K}[f(\mathbf{x}^*(\mathbf{v}^k))]}{\theta_k}}{\sum_{k=K_0}^{K}\frac{1}{\theta_k}} + \frac{1}{n}\sum_{k=K_0}^{K}\frac{1}{\theta_k}\frac{\sum_{k=K_0}^{K}\frac{\mathbb{E}_{\xi_K}[\phi(n\mathbf{y}^k)]}{\theta_k}}{\sum_{k=K_0}^{K}\frac{1}{\theta_k}}\\
&\overset{a}{\geq} \sum_{k=K_0}^{K}\frac{1}{\theta_k}\left\langle \frac{1}{n}\mathbb{E}_{\xi_K}[\boldsymbol{A}^T\hat{\mathbf{x}}^K-\hat{\mathbf{y}}^K], \mathbf{u}_{1:n}\right\rangle + \sum_{k=K_0}^{K}\frac{1}{\theta_k}\left\langle \mathbb{E}_{\xi_K}[\boldsymbol{B}\hat{\mathbf{x}}^K+\mathbf{b}], \mathbf{u}_{n+1:n+p}\right\rangle\\
&\quad + \sum_{k=K_0}^{K}\frac{1}{\theta_k}\left\langle \left[\mathbb{E}_{\xi_K}[g_1(\hat{\mathbf{x}}^K)], \cdots, \mathbb{E}_{\xi_K}[g_m(\hat{\mathbf{x}}^K)]\right]^T, \mathbf{u}_{n+p+1:n+p+m}\right\rangle\\
&\quad + \sum_{k=K_0}^{K}\frac{D(\mathbf{u})}{\theta_k} + \sum_{k=K_0}^{K}\frac{\mathbb{E}_{\xi_K}[f(\hat{\mathbf{x}}^K)]}{\theta_k} + \frac{1}{n}\sum_{k=K_0}^{K}\frac{\mathbb{E}_{\xi_K}[\phi(\hat{\mathbf{y}}^K)]}{\theta_k}\\
&= \sum_{k=K_0}^{K}\frac{1}{\theta_k}\left[\left\langle \mathbb{E}_{\xi_K}[\triangle(\hat{\mathbf{x}}^K,\hat{\mathbf{y}}^K)], \mathbf{u}\right\rangle + D(\mathbf{u}) + \mathbb{E}_{\xi_K}[f(\hat{\mathbf{x}}^K)] + \frac{1}{n}\mathbb{E}_{\xi_K}[\phi(\hat{\mathbf{y}}^K)]\right],
\end{aligned}
$$

where we use the definition of $\hat{\mathbf{x}}^K$ and $\hat{\mathbf{y}}^K$, $\mathbf{u}_{n+p+1:n+p+m} \geq 0$ and Jensen's inequality for $g_i$, $f$ and $\phi_i$ in $\overset{a}{\geq}$. Thus, from (31) and the second property in Lemma 6, we have

$$
\begin{aligned}
&\left(\frac{1}{\theta_K^2}-\frac{1}{\theta_{K_0-1}^2}\right)\mathbb{E}_{\xi_K}\left[\left\langle \triangle(\hat{\mathbf{x}}^K,\hat{\mathbf{y}}^K),\mathbf{u}\right\rangle + f(\hat{\mathbf{x}}^K) + \frac{1}{n}\phi(\hat{\mathbf{y}}^K)\right]\\
\leq& \frac{\mathbb{E}_{\xi_K}[d(\mathbf{u}^{K_0})+H^{K_0}]}{\theta_{K_0-1}^2} + \hat{n}^2\mathbb{E}_{\xi_K}[\|\mathbf{z}^{K_0}-\mathbf{u}\|_L^2] - \frac{\mathbb{E}_{\xi_K}[D(\mathbf{u}^{K+1})]}{\theta_K^2}.
\end{aligned}
$$

where we eliminate $D(\mathbf{u})$ from both sides. Adding $\left(\frac{1}{\theta_K^2}-\frac{1}{\theta_{K_0-1}^2}\right)D(\mathbf{u}^*)$ to both sides and using $\|\mathbf{z}^{K_0}-\mathbf{u}\|_L^2 \leq 2\|\mathbf{z}^{K_0}-\mathbf{u}^*\|_L^2 + 2\|\mathbf{u}-\mathbf{u}^*\|_L^2$, we have

$$
\begin{aligned}
&\left(\frac{1}{\theta_K^2}-\frac{1}{\theta_{K_0-1}^2}\right)\mathbb{E}_{\xi_K}\left[\left\langle \triangle(\hat{\mathbf{x}}^K,\hat{\mathbf{y}}^K),\mathbf{u}\right\rangle + D(\mathbf{u}^*) + f(\hat{\mathbf{x}}^K) + \frac{1}{n}\phi(\hat{\mathbf{y}}^K)\right]\\
\leq& \mathbb{E}_{\xi_K}\left[\frac{d(\mathbf{u}^{K_0})+H^{K_0}-D(\mathbf{u}^*)}{\theta_{K_0-1}^2} + 2\hat{n}^2\|\mathbf{z}^{K_0}-\mathbf{u}^*\|_L^2 + 2\hat{n}^2\|\mathbf{u}-\mathbf{u}^*\|_L^2\right] - \frac{\mathbb{E}_{\xi_K}[D(\mathbf{u}^{K+1})]-D(\mathbf{u}^*)}{\theta_K^2}\\
\overset{b}{\leq}& 2(\hat{n}^2-\hat{n})\left(D(\mathbf{u}^0)-D(\mathbf{u}^*)\right) + 2\hat{n}^2\|\mathbf{z}^0-\mathbf{u}^*\|_L^2 + 2\hat{n}^2\|\mathbf{u}-\mathbf{u}^*\|_L^2,
\end{aligned}
$$

where we use $d(\mathbf{u}^{K_0}) + H^{K_0} \geq D(\mathbf{u}^{K_0}) \geq D(\mathbf{u}^*)$ and (30) in $\overset{b}{\leq}$. ■

Now we are ready to prove the convergence rate of the primal solution. We first consider the primal objective function of problem (5) in the following lemma.

**Lemma 11** *Suppose Assumption 1 holds. Define* $\hat{\mathbf{x}}^K = \frac{\sum_{k=K_0}^K \frac{\mathbf{x}^*(\mathbf{v}^k)}{\theta_k}}{\sum_{k=K_0}^K \frac{1}{\theta_k}}$ *and* $\hat{\mathbf{y}}^K = \frac{\sum_{k=K_0}^K \frac{n\mathbf{y}^k}{\theta_k}}{\sum_{k=K_0}^K \frac{1}{\theta_k}}$. *Then we have*

$$\begin{aligned}
&\left|\mathbb{E}_{\xi_K}[f(\hat{\mathbf{x}}^K)] + \frac{1}{n}\mathbb{E}_{\xi_K}[\phi(\hat{\mathbf{y}}^K)] - f(\mathbf{x}^*) - \frac{1}{n}\phi(\mathbf{y}^*)\right| \\
\leq & \frac{2\hat{n}^2\left((1-\theta_0)\left(D(\mathbf{u}^0) - D(\mathbf{u}^*)\right) + \|\mathbf{u}^0 - \mathbf{u}^*\|_L^2 + \|\mathbf{u}^*\|_L^2\right)}{\frac{1}{\theta_K^2} - \frac{1}{\theta_{K_0-1}^2}} + \sqrt{\mathbb{E}_{\xi_K}\left[\left(\|\widetilde{\triangle}(\hat{\mathbf{x}}^K, \hat{\mathbf{y}}^K)\|_L^*\right)^2\right]}\|\mathbf{u}^*\|_L,
\end{aligned}$$

*where*

$$\widetilde{\triangle}(\mathbf{x},\mathbf{y}) = \left[(\boldsymbol{A}^T\mathbf{x} - \mathbf{y})^T/n, (\boldsymbol{B}\mathbf{x} + \mathbf{b})^T, \max\{0, g_1(\mathbf{x})\}, \cdots, \max\{0, g_m(\mathbf{x})\}\right]^T.$$

**Proof** Define $\hat{\mathbf{u}}_i^* = \begin{cases} \mathbf{u}_i^*, & \text{if } i \leq n+p, \\ \mathbf{u}_i^*, & \text{if } i > n+p \text{ and } \mathbb{E}_{\xi_K}[g_i(\hat{\mathbf{x}}^K)] \geq 0, \\ 0, & \text{if } i > n+p \text{ and } \mathbb{E}_{\xi_K}[g_i(\hat{\mathbf{x}}^K)] < 0. \end{cases}$ Since $\mathbf{u}_{n+p+1:n+p+m}^* \geq \mathbf{0}$, then $\hat{\mathbf{u}}_{n+p+1:n+p+m}^* \geq \mathbf{0}$. We also have $\|\hat{\mathbf{u}}^*\|_L \leq \|\mathbf{u}^*\|_L$ and $\|\hat{\mathbf{u}}^* - \mathbf{u}^*\|_L \leq \|\mathbf{u}^*\|_L$. Moreover, $\hat{\mathbf{u}}^*$ is independent on $\xi_K$ since we use $\mathbb{E}_{\xi_K}[g_i(\hat{\mathbf{x}}^K)]$ in the definition, rather than $g_i(\hat{\mathbf{x}}^K)$. So we can let $\mathbf{u} = \hat{\mathbf{u}}^*$ in (24). Define $\triangle_E(\mathbf{x},\mathbf{y}) = \begin{bmatrix} \mathbb{E}_{\xi_K}[(\boldsymbol{A}^T\mathbf{x} - \mathbf{y})/n] \\ \mathbb{E}_{\xi_K}[\boldsymbol{B}\mathbf{x} + \mathbf{b}] \\ \max\{\mathbf{0}, \mathbb{E}_{\xi_K}[g_{1:m}(\mathbf{x})]\} \end{bmatrix}$, then we have

$$\begin{aligned}
&\left\langle \mathbb{E}_{\xi_K}[\triangle(\hat{\mathbf{x}}^K, \hat{\mathbf{y}}^K)], \hat{\mathbf{u}}^*\right\rangle \overset{a}{=} \left\langle \triangle_E(\hat{\mathbf{x}}^K, \hat{\mathbf{y}}^K), \hat{\mathbf{u}}^*\right\rangle \geq -\|\triangle_E(\hat{\mathbf{x}}^K, \hat{\mathbf{y}}^K)\|_L^*\|\hat{\mathbf{u}}^*\|_L \\
&\overset{b}{\geq} -\|\mathbb{E}_{\xi_K}[\widetilde{\triangle}(\hat{\mathbf{x}}^K, \hat{\mathbf{y}}^K)]\|_L^*\|\hat{\mathbf{u}}^*\|_L \overset{c}{\geq} -\sqrt{\mathbb{E}_{\xi_K}\left[\left(\|\widetilde{\triangle}(\hat{\mathbf{x}}^K, \hat{\mathbf{y}}^K)\|_L^*\right)^2\right]}\|\hat{\mathbf{u}}^*\|_L,
\end{aligned} \tag{32}$$

where we use $\hat{\mathbf{u}}_{n+p+i}^* = 0$ if $\mathbb{E}_{\xi_K}[g_i(\hat{\mathbf{x}}^K)] < 0$ in $\overset{a}{=}$, $\max\{0, \mathbb{E}_{\xi_K}[a]\} \leq \mathbb{E}_{\xi_K}[\max\{0, a\}]$ in $\overset{b}{\geq}$ and $(\mathbb{E}[a])^2 \leq \mathbb{E}[a^2]$ in $\overset{c}{\geq}$. Thus, letting $\mathbf{u} = \hat{\mathbf{u}}^*$ in (24), we have

$$\begin{aligned}
&\mathbb{E}_{\xi_K}[f(\hat{\mathbf{x}}^K)] + \frac{1}{n}\mathbb{E}_{\xi_K}[\phi(\hat{\mathbf{y}}^K)] - f(\mathbf{x}^*) - \frac{1}{n}\phi(\mathbf{y}^*) \\
=&\mathbb{E}_{\xi_K}[f(\hat{\mathbf{x}}^K)] + \frac{1}{n}\mathbb{E}_{\xi_K}[\phi(\hat{\mathbf{y}}^K)] + D(\mathbf{u}^*) \\
\leq& \frac{2\hat{n}^2\left((1-\theta_0)\left(D(\mathbf{u}^0) - D(\mathbf{u}^*)\right) + \|\mathbf{u}^0 - \mathbf{u}^*\|_L^2 + \|\mathbf{u}^*\|_L^2\right)}{\frac{1}{\theta_K^2} - \frac{1}{\theta_{K_0-1}^2}} + \sqrt{\mathbb{E}_{\xi_K}\left[\left(\|\widetilde{\triangle}(\hat{\mathbf{x}}^K, \hat{\mathbf{y}}^K)\|_L^*\right)^2\right]}\|\mathbf{u}^*\|_L.
\end{aligned}$$

On the other hand, since $(\mathbf{x}^*, \mathbf{y}^*, \mathbf{u}^*)$ is a KKT point, we have

$$L_F(\hat{\mathbf{x}}^K, \hat{\mathbf{y}}^K, \mathbf{u}^*) \geq L_F(\mathbf{x}^*, \mathbf{y}^*, \mathbf{u}^*) = f(\mathbf{x}^*) + \frac{1}{n}\phi(\mathbf{y}^*).$$

From the definition in (6), we have

$$\begin{aligned}
&f(\hat{\mathbf{x}}^K)+\frac{1}{n}\phi(\hat{\mathbf{y}}^K)-f(\mathbf{x}^*)-\frac{1}{n}\phi(\mathbf{y}^*)\\
\geq&-\frac{1}{n}\left\langle\mathbf{u}_{1:n}^*,\boldsymbol{A}^T\hat{\mathbf{x}}^K-\hat{\mathbf{y}}^K\right\rangle-\left\langle\mathbf{u}_{n+1:n+p}^*,\boldsymbol{B}\hat{\mathbf{x}}^K+\mathbf{b}\right\rangle-\sum_{i=1}^m\mathbf{u}_{n+p+i}^*g_i(\hat{\mathbf{x}}^K)\\
\overset{d}{\geq}&-\frac{1}{n}\left\langle\mathbf{u}_{1:n}^*,\boldsymbol{A}^T\hat{\mathbf{x}}^K-\hat{\mathbf{y}}^K\right\rangle-\left\langle\mathbf{u}_{n+1:n+p}^*,\boldsymbol{B}\hat{\mathbf{x}}^K+\mathbf{b}\right\rangle-\sum_{i=1}^m\mathbf{u}_{n+p+i}^*\max\{0,g_i(\hat{\mathbf{x}}^K)\}\\
\geq&-\|\widetilde{\triangle}(\hat{\mathbf{x}}^K,\hat{\mathbf{y}}^K)\|_L^*\|\mathbf{u}^*\|_L,
\end{aligned}$$

where we use $\mathbf{u}_{n+p+1:n+p+m}^*\geq 0$ in $\overset{d}{\geq}$. So we have

$$\begin{aligned}
&\mathbb{E}_{\xi_K}[f(\hat{\mathbf{x}}^K)]+\frac{1}{n}\mathbb{E}_{\xi_K}[\phi(\hat{\mathbf{y}}^K)]-f(\mathbf{x}^*)-\frac{1}{n}\phi(\mathbf{y}^*)\\
\geq&-\mathbb{E}_{\xi_K}\left[\|\widetilde{\triangle}(\hat{\mathbf{x}}^K,\hat{\mathbf{y}}^K)\|_L^*\right]\|\mathbf{u}^*\|_L\geq-\sqrt{\mathbb{E}_{\xi_K}\left[\left(\|\widetilde{\triangle}(\hat{\mathbf{x}}^K,\hat{\mathbf{y}}^K)\|_L^*\right)^2\right]}\|\mathbf{u}^*\|_L,
\end{aligned}\tag{33}$$

which completes the proof. ■

### 3.2.2. Constraint Functions

Lemma 12 establishes the convergence rate for the constraint functions of problem (5). A straightforward extension of (19) only leads to $\|\mathbb{E}_{\xi_K}[\widetilde{\triangle}(\hat{\mathbf{x}}^K,\hat{\mathbf{y}}^K)]\|_L^*\leq O\left(\frac{1}{K^2}\right)$, which is less interesting since the expectation is inside the norm. We should take the expectation outside the norm, i.e., $\mathbb{E}_{\xi_K}[\|\widetilde{\triangle}(\hat{\mathbf{x}}^K,\hat{\mathbf{y}}^K)\|_L^*]\leq O\left(\frac{1}{K^2}\right)$. The later one cannot be attained by the simple techniques in (19) and requires more skillful tricks. The proof sketch of Lemma 12 is that we first establish a recursion (38) and then using (25) and the definition of $\mathbf{s}^k$ in (37) to bound the constraint functions.

**Lemma 12** *Suppose Assumption 1 holds. Define* $\hat{\mathbf{x}}^K=\frac{\sum_{k=K_0}^K\frac{\mathbf{x}^*(\mathbf{v}^k)}{\theta_k}}{\sum_{k=K_0}^K\frac{1}{\theta_k}}$ *and* $\hat{\mathbf{y}}^K=\frac{\sum_{k=K_0}^K\frac{n\mathbf{y}^k}{\theta_k}}{\sum_{k=K_0}^K\frac{1}{\theta_k}}$. *Then we have*

$$\sqrt{\mathbb{E}_{\xi_K}\left[\left(\|\widetilde{\triangle}(\hat{\mathbf{x}}^K,\hat{\mathbf{y}}^K)\|_L^*\right)^2\right]}\leq\frac{\sqrt{48}\hat{n}^2\sqrt{(1-\theta_0)\left(D(\mathbf{u}^0)-D(\mathbf{u}^*)\right)+\|\mathbf{u}^0-\mathbf{u}^*\|_L^2}}{\frac{1}{\theta_K^2}-\frac{1}{\theta_{K_0-1}^2}}.\tag{34}$$

**Proof** From the update of $\widetilde{\mathbf{z}}^k$ and the definitions of $\mathbf{y}^k$ and $\nabla d(\mathbf{v}^k)$ in (20) and (10), we have

$$\begin{aligned}
\widetilde{\mathbf{z}}_i^k&=\mathbf{z}_i^k-\frac{\nabla_i d(\mathbf{v}^k)+\mathbf{y}_i^k}{2\hat{n}\theta_k L_i}=\mathbf{z}_i^k-\frac{(-\boldsymbol{A}_i^T\mathbf{x}^*(\mathbf{v}^k)+n\mathbf{y}_i^k)/n}{2\hat{n}\theta_k L_i},i\leq n,\\
\widetilde{\mathbf{z}}_i^k&=\mathbf{z}_i^k-\frac{\nabla d_i(\mathbf{v}^k)}{2\hat{n}\theta_k L_i}=\mathbf{z}_i^k+\frac{\boldsymbol{B}_{i,:}^T\mathbf{x}^*(\mathbf{v}^k)+\mathbf{b}_i}{2\hat{n}\theta_k L_i},n<i\leq n+p,\\
\widetilde{\mathbf{z}}_i^k&=\left[\mathbf{z}_i^k-\frac{\nabla d_i(\mathbf{v}^k)}{2\hat{n}\theta_k L_i}\right]_+=\mathbf{z}_i^k+\frac{\max\left\{g_i(\mathbf{x}^*(\mathbf{v}^k)),-2\hat{n}\theta_k L_i\mathbf{z}_i^k\right\}}{2\hat{n}\theta_k L_i},i>n+p.
\end{aligned}\tag{35}$$

Define $\pi^k \in \mathbb{R}^{\hat{n}}$ such that $\pi_i^k = \begin{cases} \frac{\boldsymbol{A}_i^T \mathbf{x}^*(\mathbf{v}^k) - n\mathbf{y}_i^k}{n}, & i \le n, \\ \boldsymbol{B}_{i,:}^T \mathbf{x}^*(\mathbf{v}^k) + \mathbf{b}_i, & n < i \le n+p, \\ \max\left\{g_i(\mathbf{x}^*(\mathbf{v}^k)), -2\hat{n}\theta_k L_i \mathbf{z}_i^k\right\}, & i > n+p, \end{cases}$ then we have

$$\widetilde{\mathbf{z}}_i^k - \mathbf{z}_i^k = \tfrac{\pi_i^k}{2\hat{n}\theta_k L_i} \tag{36}$$

and

$$\mathbb{E}_{i_k|\xi_{k-1}}[\mathbf{z}_i^{k+1}] = \tfrac{1}{\hat{n}}\widetilde{\mathbf{z}}_i^k + (1-\tfrac{1}{\hat{n}})\mathbf{z}_i^k = \tfrac{1}{\hat{n}}\left(\mathbf{z}_i^k + \tfrac{\pi_i^k}{2\hat{n}\theta_k L_i}\right) + (1-\tfrac{1}{\hat{n}})\mathbf{z}_i^k = \mathbf{z}_i^k + \tfrac{\pi_i^k}{2\hat{n}^2\theta_k L_i}.$$

Define $\mathbf{g}^k \in \mathbb{R}^{\hat{n}}$ and $\mathbf{s}^k \in \mathbb{R}^{\hat{n}}$ such that

$$\mathbf{g}_i^k = \frac{\pi_i^k}{2\hat{n}^2\theta_k L_i} + \mathbf{z}_i^k - \mathbf{z}_i^{k+1} \quad \text{and} \quad \mathbf{s}_i^k = \sum_{t=K_0}^{k} \mathbf{g}_i^t \text{ (specially, } \mathbf{s}_i^k = 0, k < K_0), \tag{37}$$

then we get $\mathbb{E}_{i_k|\xi_{k-1}}[\mathbf{g}_i^k] = 0$ and $\mathbb{E}_{i_k|\xi_{k-1}}[\mathbf{s}_i^k] = \mathbf{s}_i^{k-1}$. Moreover, for $k \ge K_0$, we have

$$\begin{aligned}
\mathbb{E}_{i_k|\xi_{k-1}}[\|\mathbf{g}^k\|_L^2] \overset{a}{=}& \mathbb{E}_{i_k|\xi_{k-1}}\left[\left\|\frac{1}{\hat{n}}(\widetilde{\mathbf{z}}^k - \mathbf{z}^k) + \mathbf{z}^k - \mathbf{z}^{k+1}\right\|_L^2\right] \\
=& \frac{1}{\hat{n}}\sum_{i=1}^{\hat{n}}\left[L_i\left\|\frac{1}{\hat{n}}(\widetilde{\mathbf{z}}_i^k - \mathbf{z}_i^k) + \mathbf{z}_i^k - \widetilde{\mathbf{z}}_i^k\right\|^2 + \sum_{j\neq i} L_j \left\|\frac{1}{\hat{n}}(\widetilde{\mathbf{z}}_j^k - \mathbf{z}_j^k) + \mathbf{z}_j^k - \mathbf{z}_j^k\right\|^2\right] \\
=& \left(\frac{1}{\hat{n}}\left(1 - \frac{1}{\hat{n}}\right)^2 + \frac{\hat{n}-1}{\hat{n}^3}\right)\sum_{i=1}^{\hat{n}} L_i \|\widetilde{\mathbf{z}}_i^k - \mathbf{z}_i^k\|^2 \\
\le& \frac{1}{\hat{n}}\sum_{i=1}^{\hat{n}} L_i\|\widetilde{\mathbf{z}}_i^k - \mathbf{z}_i^k\|^2 \overset{b}{=} \mathbb{E}_{i_k|\xi_{k-1}}[\|\mathbf{z}^{k+1} - \mathbf{z}^k\|_L^2],
\end{aligned}$$

where we use (36) and (37) in $\overset{a}{=}$ and (67) in $\overset{b}{=}$. Then for any $k \ge K_0$, we have

$$\begin{aligned}
\mathbb{E}_{\xi_k}[\|\mathbf{s}^k\|_L^2] =& \mathbb{E}_{\xi_{k-1}}\left[\mathbb{E}_{i_k|\xi_{k-1}}[\|\mathbf{s}^k\|_L^2]\right] \\
=& \mathbb{E}_{\xi_{k-1}}\left[\mathbb{E}_{i_k|\xi_{k-1}}\left[\|\mathbf{s}^k - \mathbb{E}_{i_k|\xi_{k-1}}[\mathbf{s}^k] + \mathbb{E}_{i_k|\xi_{k-1}}[\mathbf{s}^k]\|_L^2\right]\right] \\
=& \mathbb{E}_{\xi_{k-1}}\left[\mathbb{E}_{i_k|\xi_{k-1}}\left[\|\mathbf{s}^k - \mathbb{E}_{i_k|\xi_{k-1}}[\mathbf{s}^k]\|_L^2\right] + \|\mathbb{E}_{i_k|\xi_{k-1}}[\mathbf{s}^k]\|_L^2\right] \\
=& \mathbb{E}_{\xi_{k-1}}\left[\mathbb{E}_{i_k|\xi_{k-1}}[\|\mathbf{g}^k\|_L^2] + \|\mathbf{s}^{k-1}\|_L^2\right] \\
\le& \mathbb{E}_{\xi_k}[\|\mathbf{z}^{k+1} - \mathbf{z}^k\|_L^2] + \mathbb{E}_{\xi_{k-1}}[\|\mathbf{s}^{k-1}\|_L^2].
\end{aligned} \tag{38}$$

Summing over $k = K_0, K_0+1, \cdots, K$ and using $\mathbf{s}^{K_0-1} = \mathbf{0}$, we have

$$\begin{aligned}
\mathbb{E}_{\xi_K}\left[\|\mathbf{s}^K\|_L^2\right] \le \sum_{k=K_0}^{K}\mathbb{E}_{\xi_k}[\|\mathbf{z}^{k+1} - \mathbf{z}^k\|_L^2] =& \sum_{k=K_0}^{K}\mathbb{E}_{\xi_K}[\|\mathbf{z}^{k+1} - \mathbf{z}^k\|_L^2] \\
\overset{a}{\le}& 2(1-\theta_0)\left(D(\mathbf{u}^0) - D(\mathbf{u}^*)\right) + 2\|\mathbf{u}^0 - \mathbf{u}^*\|_L^2,
\end{aligned} \tag{39}$$

where we use (25) in $\overset{a}{\leq}$. On the other hand, from the definition of $\mathbf{s}^K$ and $\mathbf{g}^k$, we have

$$\begin{aligned}&\|\mathbf{s}^K\|_L^2 = \left\|\sum_{k=K_0}^{K}\mathbf{g}^k\right\|_L^2 = \sum_i L_i \left\|\sum_{k=K_0}^{K}\left(\frac{\pi_i^k}{2\hat{n}^2\theta_k L_i}+\mathbf{z}_i^k-\mathbf{z}_i^{k+1}\right)\right\|^2\\ &=\sum_i L_i\left\|\sum_{k=K_0}^{K}\frac{\pi_i^k}{2\hat{n}^2\theta_k L_i}+\mathbf{z}_i^{K_0}-\mathbf{z}_i^{K+1}\right\|^2 \geq \frac{1}{3}\left(\left\|\sum_{k=K_0}^{K}\frac{\pi^k}{2\hat{n}^2\theta_k}\right\|_L^*\right)^2 - \|\mathbf{z}^{K_0}-\mathbf{u}^*\|_L^2 - \|\mathbf{z}^{K+1}-\mathbf{u}^*\|_L^2.\end{aligned}$$

So from (39), (26) and (30), we have

$$\begin{aligned}&\mathbb{E}_{\xi_K}\left[\left(\left\|\sum_{k=K_0}^{K}\frac{\pi^k}{2\hat{n}^2\theta_k}\right\|_L^*\right)^2\right] \leq 3\mathbb{E}_{\xi_K}\left[\|\mathbf{s}^K\|_L^2\right]+3\mathbb{E}_{\xi_K}\left[\|\mathbf{z}^{K+1}-\mathbf{u}^*\|_L^2\right]+3\mathbb{E}_{\xi_K}\left[\|\mathbf{z}^{K_0}-\mathbf{u}^*\|_L^2\right]\\ &\quad\leq 12(1-\theta_0)\left(D(\mathbf{u}^0)-D(\mathbf{u}^*)\right)+12\|\mathbf{u}^0-\mathbf{u}^*\|_L^2.\end{aligned} \tag{40}$$

For $i \leq n$, we have

$$\begin{aligned}&\sum_{k=K_0}^{K}\frac{\pi_i^k}{2\hat{n}^2\theta_k} = \frac{1}{2\hat{n}^2}\sum_{k=K_0}^{K}\frac{(\boldsymbol{A}_i^T\mathbf{x}^*(\mathbf{v}^k)-n\mathbf{y}_i^k)/n}{\theta_k}\\ &=\frac{\sum_{k=K_0}^{K}\frac{1}{\theta_k}}{2\hat{n}^2}\frac{\sum_{k=K_0}^{K}\frac{\boldsymbol{A}_i^T\mathbf{x}^*(\mathbf{v}^k)-n\mathbf{y}_i^k}{n\theta_k}}{\sum_{k=K_0}^{K}\frac{1}{\theta_k}} = \frac{\frac{1}{\theta_K^2}-\frac{1}{\theta_{K_0-1}^2}}{2\hat{n}^2}(\boldsymbol{A}_i^T\hat{\mathbf{x}}^K-\hat{\mathbf{y}}_i^K)/n.\end{aligned} \tag{41}$$

Similarly, for $n < i \leq n+p$, we have

$$\sum_{k=K_0}^{K}\frac{\pi_i^k}{2\hat{n}^2\theta_k} = \frac{\frac{1}{\theta_K^2}-\frac{1}{\theta_{K_0-1}^2}}{2\hat{n}^2}(\boldsymbol{B}_{i,:}^T\hat{\mathbf{x}}^K+\mathbf{b}_i),$$

and for $n+p < i \leq n+p+m$, we have

$$\begin{aligned}&\sum_{k=K_0}^{K}\frac{\pi_i^k}{2\hat{n}^2\theta_k} \geq \frac{1}{2\hat{n}^2}\sum_{k=K_0}^{K}\frac{g_i(\mathbf{x}^*(\mathbf{v}^k))}{\theta_k} \geq \frac{\sum_{k=K_0}^{K}\frac{1}{\theta_k}}{2\hat{n}^2}g_i(\hat{\mathbf{x}}^K) = \frac{\frac{1}{\theta_K^2}-\frac{1}{\theta_{K_0-1}^2}}{2\hat{n}^2}g_i(\hat{\mathbf{x}}^K)\\ &\Rightarrow \left(\sum_{k=K_0}^{K}\frac{\pi_i^k}{2\hat{n}^2\theta_k}\right)^2 \geq \left(\frac{1}{\theta_K^2}-\frac{1}{\theta_{K_0-1}^2}\right)^2\frac{\left(\max\{0,g_i(\hat{\mathbf{x}}^K)\}\right)^2}{4\hat{n}^4}.\end{aligned}$$

Then we have

$$\frac{1}{4\hat{n}^4}\left(\frac{1}{\theta_K^2}-\frac{1}{\theta_{K_0-1}^2}\right)^2\left(\|\widetilde{\triangle}(\hat{\mathbf{x}}^K,\hat{\mathbf{y}}^K)\|_L^*\right)^2 \leq \left(\left\|\sum_{k=K_0}^{K}\frac{\pi^k}{2\hat{n}^2\theta_k}\right\|_L^*\right)^2.$$

Taking expectation with respect to $\xi_K$ and from (40), we can immediately have the conclusion. ■

### 3.2.3. Proof of Theorem 3

From Lemma 11, Lemma 12, $\frac{1}{\theta_K^2} - \frac{1}{\theta_{K_0-1}^2} \geq \left(\frac{K^2}{4} + \hat{n}K\right)\left(1-\frac{1}{\upsilon}\right)$, $(\mathbb{E}[a])^2 \leq \mathbb{E}[a^2]$ and

$$\begin{aligned}&\left|\frac{1}{n}\phi(\boldsymbol{A}^T\hat{\mathbf{x}}^K) - \frac{1}{n}\phi(\hat{\mathbf{y}}^K)\right| \leq \frac{1}{n}\sum_{i=1}^n M|\boldsymbol{A}_i^T\hat{\mathbf{x}}^K - \hat{\mathbf{y}}_i^K| \\ &\leq \sqrt{\sum_{i=1}^n M^2 L_i}\left\|\frac{1}{n}(\boldsymbol{A}^T\hat{\mathbf{x}}^K - \hat{\mathbf{y}}^K)\right\|_L^* \leq \sqrt{\sum_{i=1}^n M^2 L_i \|\widetilde{\triangle}(\hat{\mathbf{x}}^K, \hat{\mathbf{y}}^K)\|_L^*},\end{aligned} \tag{42}$$

we have

$$\begin{aligned}\left|\mathbb{E}_{\xi_K}\left[F(\hat{\mathbf{x}}^K)\right] - F(\mathbf{x}^*)\right| \leq & \frac{2\hat{n}^2\left((1-\theta_0)\left(D(\mathbf{u}^0) - D(\mathbf{u}^*)\right) + \|\mathbf{u}^0 - \mathbf{u}^*\|_L^2 + \|\mathbf{u}^*\|_L^2\right)}{\left(\frac{K^2}{4} + \hat{n}K\right)\left(1-\frac{1}{\upsilon}\right)} \\ & + \frac{6\sqrt{2}\hat{n}^2(\|\mathbf{u}^*\|_L + M\sqrt{\sum_{i=1}^n L_i})}{\left(\frac{K^2}{4} + \hat{n}K\right)\left(1-\frac{1}{\upsilon}\right)}\sqrt{(1-\theta_0)\left(D(\mathbf{u}^0) - D(\mathbf{u}^*)\right) + \|\mathbf{u}^0 - \mathbf{u}^*\|_L^2}.\end{aligned}$$

From Cauchy-Schwartz inequality, we have the desired result.

**Remark 13** *In Assumption 1, we assume that $\phi_i$ is $M$-Lipschitz continuous. From the above analysis, we can see that it is only used in (42). Lemmas 11 and 12 do not need this assumption. Moreover, it only affects the convergence rate in the primal space and the analysis in the dual space does not require it.*

## 4. Extension under the Quadratic Growth Condition

In this section, we give the linear complexity under stronger assumptions. Specifically, we use both Assumptions 1 and 2 in this section. The quadratic functional growth condition in Assumption 2 is equivalent to the global error bound condition (Drusvyatskiy and Lewis, 2018) and is satisfied for broad applications. We give a simple example satisfying Assumption 2 and refer the reader to (Bolte et al., 2017; Li, 2013; Yang, 2009; Liu and Yang, 2017) for more examples.

***Example***: Consider problem (1) with strongly convex and smooth $f$ and the simple form (4) of $g(\mathbf{x})$. Furthermore, we require that $\sum_{i=1}^n \phi_i^*(\mathbf{u}_i)$ has the form of $\langle \mathbf{c}, \mathbf{u}\rangle + P(\mathbf{u})$, where $P(\mathbf{u})$ is a polyhedral function or an indicator function of a polyhedral set. In this case, $d(\mathbf{u}) = f^*(-\boldsymbol{S}\mathbf{u}) - \langle \mathbf{p}, \mathbf{u}\rangle$, where $\boldsymbol{S}$ and $\mathbf{p}$ are defined in (14). It may not be strongly convex since $\boldsymbol{S}$ may not be full column rank. However, $D(\mathbf{u})$ satisfies the error bound condition (Luo and Tseng, 1992; Wang and Lin, 2014) and thus satisfies Assumption 2. The least absolute deviation, SVM and multiclass SVM (Shalev-Shwartz and Zhang, 2016) have the required form.

To exploit Assumption 2, we use Algorithm 1 with restart (O'Donoghue and Candès, 2015; Fercoq and Qu, 2020) and establish the faster convergence rate. Namely, at each iteration, Algorithm 1 is called with fixed and finite iterations with the output of the previous iteration being the initializer of current iteration. We describe the method in Algorithm 3. We use the inner-outer iteration procedure, rather than the one loop accelerated algorithms with direct support to strongly convex dual functions, e.g., APCG (Lin et al., 2015b), since the quadratic functional growth condition is generally not enough to prove the accelerated linear convergence rate for the one loop accelerated algorithms (Necoara et al., 2019).

**Algorithm 3** ARDCA with restart

```
Input $\mathbf{u}^{-1,K+1} = \mathbf{u}^{0,0} \in \mathbb{D}$.
for $t = 0, 1, 2, \cdots, N$ do
  Run ARDCA($\mathbf{u}^{t-1,K+1}$,$K_0$,$K$) and output $\mathbf{u}^{t,K+1}$ and $\hat{\mathbf{x}}^{t,K}$.
end for
Output $\hat{\mathbf{x}}^{N,K}$.
```

Define $\mathbf{u}^{t,0,*} = \mathrm{Proj}_{\mathbb{D}^*}(\mathbf{u}^{t,0}) = \mathrm{argmin}_{\mathbf{u} \in \mathbb{D}^*} \|\mathbf{u}^{t,0} - \mathbf{u}\|_L$ to be the nearest optimal solution to $\mathbf{u}^{t,0}$. Denote $\xi_{t,K} = \{i_{t,0}, i_{t,1}, \cdots, i_{t,K}\}$ and $\zeta_t = \cup_{r=0}^{t} \xi_{r,K}$ to be the random sequence, where $i_{t,s}$ is the random index chosen at the $t$-th outer iteration and $s$-th inner iteration of Algorithm 3. We give the linear convergence of Algorithm 3 for both primal solutions and dual solutions in Theorem 14.

**Theorem 14** *Suppose Assumptions 1 and 2 hold. For Algorithm 3, we have*

$$(1-\theta_0)\left(\mathbb{E}_{\zeta_N}[D(\mathbf{u}^{N,0})] - D(\mathbf{u}^*)\right) + \mathbb{E}_{\zeta_N}[\|\mathbf{u}^{N,0} - \mathbf{u}^{N,0,*}\|_L^2] \le \left(\frac{1+(1-\theta_0)\kappa}{1+\frac{\kappa}{2}\left(\frac{K}{2\hat{n}}+1\right)^2}\right)^N T_{0.0}, \tag{43}$$

*where* $T_{0,0} = (1-\theta_0)\left(D(\mathbf{u}^{0,0}) - D(\mathbf{u}^*)\right) + \|\mathbf{u}^{0,0} - \mathbf{u}^{0,0,*}\|_L^2$.

*Suppose Assumptions 1 and 2 hold. Assume that* $\mathbb{D}^*$ *is bounded, i.e.,* $\|\mathbf{u}^*\|_L \le C_{\mathbb{D}^*}, \forall \mathbf{u}^* \in \mathbb{D}^*$. *Let* $K_0 \le \left\lfloor \frac{K}{\upsilon(1+1/\hat{n})} + 1 \right\rfloor$ *with any* $\upsilon > 1$ *and* $K \ge \hat{n}$. *Then for Algorithm 3, we have*

$$\begin{aligned} \left|\mathbb{E}_{\zeta_N}\left[F(\hat{\mathbf{x}}^{N,K})\right] - F(\mathbf{x}^*)\right| &\le C_1 \left(\frac{1+(1-\theta_0)\kappa}{1+\frac{\kappa}{2}\left(\frac{K}{2\hat{n}}+1\right)^2}\right)^{N/2} + C_2 \left(\frac{1+(1-\theta_0)\kappa}{1+\frac{\kappa}{2}\left(\frac{K}{2\hat{n}}+1\right)^2}\right)^N, \\ \mathbb{E}_{\zeta_N}\left[\left\|\left[\begin{array}{c} \boldsymbol{B}\hat{\mathbf{x}}^{N,K} + \mathbf{b} \\ \max\left\{0, g(\hat{\mathbf{x}}^{N,K})\right\} \end{array}\right]\right\|_L^*\right] &\le C_3 \left(\frac{1+(1-\theta_0)\kappa}{1+\frac{\kappa}{2}\left(\frac{K}{2\hat{n}}+1\right)^2}\right)^{N/2}, \end{aligned} \tag{44}$$

*where* $C_1 = \frac{6C_{\mathbb{D}^*}\sqrt{T_{0,0}} + 6M\sqrt{\sum_{i=1}^n L_i}\sqrt{T_{0,0}}}{1-1/\nu} + \frac{2\sqrt{m}C_{\mathbb{D}^*}\sqrt{T_{0,0}}}{\sqrt{1-1/\nu}}$, $C_2 = \frac{2T_{0,0}}{1-1/\nu}$ *and* $C_3 = \frac{6\sqrt{T_{0,0}}}{1-1/\nu}$.

In the traditional analysis for the restart scheme, the inner iteration number heavily depends on the condition number $\kappa$ (Fercoq and Qu, 2019, 2020). Specifically, letting $\mathbf{u}^* = \mathbf{u}^{t,0,*}$ in (27) and from Assumption 2, we have

$$\mathbb{E}_{\zeta_N}[D(\mathbf{u}^{t+1,0})] - D(\mathbf{u}^*) \le \hat{n}^2 \theta_K^2 \left(1 - \theta_0 + \frac{1}{\kappa}\right)\left(\mathbb{E}_{\zeta_N}[D(\mathbf{u}^{t,0})] - D(\mathbf{u}^*)\right).$$

To make $\hat{n}^2\theta_K^2\left(1-\theta_0+\frac{1}{\kappa}\right) < 1$, we should require $K = O\left(\hat{n} + \frac{\hat{n}}{\sqrt{\kappa}}\right)$, otherwise, the traditional analysis cannot prove the decrement of the objective. However, in practice, $\kappa$ is often difficult to estimate. Different from the traditional analysis, in Theorem 14, we show the linear convergence when the algorithm restarts at any period. In other words, $\frac{1+(1-\theta_0)\kappa}{1+\frac{\kappa}{2}\left(\frac{K}{2\hat{n}}+1\right)^2} < 1$ for any $K \ge \hat{n}$. Our analysis applies to the problems where $\kappa$ is unknown.

**Remark 15** *Our result (43) is motivated by (Fercoq and Qu, 2020). However, when applied to the dual problem (9), Fercoq and Qu (2020) requires that $D(\mathbf{u})$ has the unique optimal dual solution $\mathbf{u}^*$ and needs a stronger quadratic functional growth condition of*

$$\kappa\|\mathbf{u}-\mathbf{u}^*\|_L^2 \leq D(\mathbf{u})-D(\mathbf{u}^*). \tag{45}$$

*As a comparison, in Assumption 2, we do not need the uniqueness of the optimal dual solution and only assume*

$$\kappa\|\mathbf{u}-Proj_{\mathbb{D}^*}(\mathbf{u})\|_L^2 \leq D(\mathbf{u})-D(\mathbf{u}^*). \tag{46}$$

*Comparing (46) with (45), we can see that (45) can deduce (46) and not vice versa. The analysis in (Fercoq and Qu, 2020) cannot be applied under our assumptions since a critical property in (Fercoq and Qu, 2020, Equ. (31)) does not hold any more. To deal with the more general assumption (46), we develop a totally different proof framework and it is much simpler and more general.*

Let us compare the complexity of Algorithm 3 with that of randomized dual coordinate ascent, i.e., $O\left(\left(\hat{n}+\frac{\hat{n}}{\kappa}\right)\log\frac{1}{\epsilon}\right)$ (Lu and Xiao, 2015). If $\kappa < 1$, Algorithm 3 attains the optimal complexity of $O\left(\left(\hat{n}+\frac{\hat{n}}{\sqrt{\kappa}}\right)\log\frac{1}{\epsilon}\right)$ for both the primal solutions and dual solutions when $K=O\left(\hat{n}+\frac{\hat{n}}{\sqrt{\kappa}}\right)$. When $\hat{n}\leq K\leq \hat{n}+\frac{\hat{n}}{\kappa}$, Algorithm 3 outperforms the randomized dual coordinate ascent[7]. When $K$ is larger than $\hat{n}+\frac{\hat{n}}{\kappa}$, Algorithm 3 performs worse than randomized dual coordinate ascent. On the other hand, if $\kappa>1$, the complexity of Algorithm 3 has the same order of magnitude as that of randomized dual coordinate ascent when $\hat{n}\leq K\leq\hat{n}+\frac{\hat{n}}{\kappa}$. When $K>\hat{n}+\frac{\hat{n}}{\kappa}$, Algorithm 3 performs worse. For most practical problems, we have $\kappa<1$ and thus Algorithm 3 is a better and safe choice for a wide range of $K$.

### 4.1. Convergence Rate Analysis of the Dual Solutions

In this section, we prove (43). We first consider one outer iteration of Algorithm 3 and use the symbols in Algorithm 2 for simplicity. In the following lemma, we show that $\mathbf{u}^{k+1}$ is the convex combination of $\mathbf{u}^1,\cdots,\mathbf{u}^k$ and $\mathbf{z}^{k+1}$.

**Lemma 16** *For Algorithm 2, we have*

$$\mathbf{u}^{k+1}=\frac{\theta_k}{\theta_0}\mathbf{z}^{k+1}+\theta_k\sum_{t=1}^{k}\left(\frac{\theta_t(1-\theta_0)^{k-t}}{\theta_{t-1}^2}-\frac{(1-\theta_0)^{k+1-t}}{\theta_{t-1}}\right)\mathbf{u}^t, \tag{47}$$

*where* $\frac{\theta_t(1-\theta_0)^{k-t}}{\theta_{t-1}^2}-\frac{(1-\theta_0)^{k+1-t}}{\theta_{t-1}}>0$ *and*

$$\frac{\theta_k}{\theta_0}+\theta_k\sum_{t=1}^{k}\left(\frac{\theta_t(1-\theta_0)^{k-t}}{\theta_{t-1}^2}-\frac{(1-\theta_0)^{k+1-t}}{\theta_{t-1}}\right)=1. \tag{48}$$

[7] Please see the details in Appendix F.

**Proof** For Algorithm 2, we have

$$\mathbf{u}^{k+1} = (1-\theta_k)\mathbf{u}^k + \frac{\theta_k}{\theta_0}\mathbf{z}^{k+1} - \frac{\theta_k(1-\theta_0)}{\theta_0}\mathbf{z}^k. \tag{49}$$

Decomposing term $(1-\theta_k)\mathbf{u}^k$ into $(1-\theta_k)\left(1-\frac{\theta_{k-1}(1-\theta_0)}{\theta_k}\right)\mathbf{u}^k$ and $(1-\theta_k)\frac{\theta_{k-1}(1-\theta_0)}{\theta_k}\mathbf{u}^k$ and using (49) for the later one, we have

$$\begin{aligned}\mathbf{u}^{k+1} =&(1-\theta_k)\left(1-\frac{\theta_{k-1}(1-\theta_0)}{\theta_k}\right)\mathbf{u}^k + \frac{\theta_k}{\theta_0}\mathbf{z}^{k+1} - \frac{\theta_k(1-\theta_0)}{\theta_0}\mathbf{z}^k\\ &+(1-\theta_k)\frac{\theta_{k-1}(1-\theta_0)}{\theta_k}\left[(1-\theta_{k-1})\mathbf{u}^{k-1} + \frac{\theta_{k-1}}{\theta_0}\mathbf{z}^k - \frac{\theta_{k-1}(1-\theta_0)}{\theta_0}\mathbf{z}^{k-1}\right]\\ =&(1-\theta_k)\left(1-\frac{\theta_{k-1}(1-\theta_0)}{\theta_k}\right)\mathbf{u}^k + \frac{\theta_k}{\theta_0}\mathbf{z}^{k+1} + \frac{\theta_k\theta_{k-1}(1-\theta_0)}{\theta_{k-2}^2}\mathbf{u}^{k-1} - \frac{\theta_k(1-\theta_0)^2}{\theta_0}\mathbf{z}^{k-1}.\end{aligned}$$

Decomposing $\frac{\theta_k\theta_{k-1}(1-\theta_0)}{\theta_{k-2}^2}\mathbf{u}^{k-1}$ into $\frac{\theta_k\theta_{k-1}(1-\theta_0)}{\theta_{k-2}^2}\left(1-\frac{\theta_{k-2}(1-\theta_0)}{\theta_{k-1}}\right)\mathbf{u}$ and $\frac{\theta_k\theta_{k-1}(1-\theta_0)}{\theta_{k-2}^2}\frac{\theta_{k-2}(1-\theta_0)}{\theta_{k-1}}\mathbf{u}$ and using (49) for the later one again, we have

$$\begin{aligned}\mathbf{u}^{k+1} =&(1-\theta_k)\left(1-\frac{\theta_{k-1}(1-\theta_0)}{\theta_k}\right)\mathbf{u}^k + \frac{\theta_k}{\theta_0}\mathbf{z}^{k+1} + \frac{\theta_k\theta_{k-1}(1-\theta_0)}{\theta_{k-2}^2}\left(1-\frac{\theta_{k-2}(1-\theta_0)}{\theta_{k-1}}\right)\mathbf{u}^{k-1}\\ &-\frac{\theta_k(1-\theta_0)^2}{\theta_0}\mathbf{z}^{k-1} + \frac{\theta_k\theta_{k-1}(1-\theta_0)}{\theta_{k-2}^2}\frac{\theta_{k-2}(1-\theta_0)}{\theta_{k-1}}\left((1-\theta_{k-2})\mathbf{u}^{k-2} + \frac{\theta_{k-2}}{\theta_0}\mathbf{z}^{k-1} - \frac{\theta_{k-2}(1-\theta_0)}{\theta_0}\mathbf{z}^{k-2}\right)\\ =&(1-\theta_k)\left(1-\frac{\theta_{k-1}(1-\theta_0)}{\theta_k}\right)\mathbf{u}^k + \frac{\theta_k}{\theta_0}\mathbf{z}^{k+1} + \frac{\theta_k\theta_{k-1}(1-\theta_0)}{\theta_{k-2}^2}\left(1-\frac{\theta_{k-2}(1-\theta_0)}{\theta_{k-1}}\right)\mathbf{u}^{k-1}\\ &+\frac{\theta_k\theta_{k-2}(1-\theta_0)^2}{\theta_{k-3}^2}\mathbf{u}^{k-2} - \frac{\theta_k(1-\theta_0)^3}{\theta_0}\mathbf{z}^{k-2}.\end{aligned}$$

Do the above operations recursively, we have

$$\begin{aligned}\mathbf{u}^{k+1} =&\frac{\theta_k}{\theta_0}\mathbf{z}^{k+1} + \sum_{t=1}^k \frac{\theta_k\theta_t(1-\theta_0)^{k-t}}{\theta_{t-1}^2}\left(1-\frac{\theta_{t-1}(1-\theta_0)}{\theta_t}\right)\mathbf{u}^t + \frac{\theta_k\theta_0(1-\theta_0)^k}{\theta_{-1}^2}\mathbf{u}^0 - \frac{\theta_k(1-\theta_0)^{k+1}}{\theta_0}\mathbf{z}^0\\ \overset{a}{=}&\frac{\theta_k}{\theta_0}\mathbf{z}^{k+1} + \theta_k\sum_{t=1}^k\left(\frac{\theta_t(1-\theta_0)^{k-t}}{\theta_{t-1}^2} - \frac{(1-\theta_0)^{k+1-t}}{\theta_{t-1}}\right)\mathbf{u}^t,\end{aligned}$$

where we use $\frac{1}{\theta_{-1}^2} = \frac{1-\theta_0}{\theta_0^2}$ and $\mathbf{u}^0 = \mathbf{z}^0$ in $\overset{a}{=}$. On the other hand, we can easily check that

$$\begin{aligned}&\sum_{t=1}^k\left(\frac{\theta_t(1-\theta_0)^{k-t}}{\theta_{t-1}^2} - \frac{(1-\theta_0)^{k+1-t}}{\theta_{t-1}}\right) = \sum_{t=1}^k\left(\frac{(1-\theta_t)(1-\theta_0)^{k-t}}{\theta_t} - \frac{(1-\theta_0)^{k-(t-1)}}{\theta_{t-1}}\right)\\ &=\frac{1}{\theta_k} - \sum_{t=1}^k(1-\theta_0)^{k-t} - \frac{(1-\theta_0)^k}{\theta_0} = \frac{1}{\theta_k} - \frac{1-(1-\theta_0)^k}{\theta_0} - \frac{(1-\theta_0)^k}{\theta_0} = \frac{1}{\theta_k} - \frac{1}{\theta_0},\end{aligned}$$

which leads to (48). Next, we prove $\frac{\theta_t(1-\theta_0)^{k-t}}{\theta_{t-1}^2} - \frac{(1-\theta_0)^{k+1-t}}{\theta_{t-1}} > 0$, which is equivalent to $\frac{\theta_t}{\theta_{t-1}} > 1-\theta_0$, which is true since $\frac{\theta_t}{\theta_{t-1}} = \sqrt{1-\theta_t} \geq 1-\theta_t > 1-\theta_0$. ■

Based on Lemma 16, we can establish the decreasing of $(1-\theta_0)\left(\mathbb{E}_{\xi_K}[D(\mathbf{u}^{k+1})]-D(\mathbf{u}^*)\right)+\mathbb{E}_{\xi_K}[\|\mathbf{u}^{k+1}-\mathbf{u}^{k+1,*}\|_L^2]$ for any $\theta_k$ in the following lemma. The proof sketch of Lemma 17 is that we first establish a recursion (52) based on Lemma 16 and (27) and then prove the result in Lemma 17 by induction.

**Lemma 17** *Suppose Assumptions 1 and 2 hold. For Algorithm 2, we have*

$$\begin{aligned}&(1-\theta_0)\left(\mathbb{E}_{\xi_K}[D(\mathbf{u}^{k+1})]-D(\mathbf{u}^*)\right)+\mathbb{E}_{\xi_K}[\|\mathbf{u}^{k+1}-\mathbf{u}^{k+1,*}\|_L^2]\\ \leq&\frac{1+(1-\theta_0)\kappa}{1+\frac{\kappa\theta_0^2}{2\theta_k^2}}\left((1-\theta_0)\left(D(\mathbf{u}^0)-D(\mathbf{u}^*)\right)+\|\mathbf{u}^0-\mathbf{u}^{0,*}\|_L^2\right),\end{aligned}$$

*where we define* $\mathbf{u}^{k,*}=\mathit{Proj}_{\mathbb{D}^*}(\mathbf{u}^k)=\operatorname{argmin}_{\mathbf{u}\in\mathbb{D}^*}\|\mathbf{u}^k-\mathbf{u}\|_L$.

**Proof** We consider $(1-\theta_0)\left(D(\mathbf{u}^{k+1})-D(\mathbf{u}^*)\right)+\|\mathbf{u}^{k+1}-\mathbf{u}^{k+1,*}\|_F^2$. Decomposing the second term into $\sigma_k\|\mathbf{u}^{k+1}-\mathbf{u}^{k+1,*}\|_F^2$ and $(1-\sigma_k)\|\mathbf{u}^{k+1}-\mathbf{u}^{k+1,*}\|_F^2$ and using Assumption 2 for the first term, we have

$$\begin{aligned}&(1-\theta_0)\left(\mathbb{E}_{\xi_K}[D(\mathbf{u}^{k+1})]-D(\mathbf{u}^*)\right)+\mathbb{E}_{\xi_K}[\|\mathbf{u}^{k+1}-\mathbf{u}^{k+1,*}\|_L^2]\\ \leq&\left(1-\theta_0+\frac{\sigma_k}{\kappa}\right)\left(\mathbb{E}_{\xi_K}[D(\mathbf{u}^{k+1})]-D(\mathbf{u}^*)\right)+(1-\sigma_k)\mathbb{E}_{\xi_K}[\|\mathbf{u}^{k+1}-\mathbf{u}^{k+1,*}\|_L^2].\end{aligned}\tag{50}$$

From the definition of $\mathbf{u}^{k+1,*}$, (48) and (47), we have

$$\begin{aligned}&\|\mathbf{u}^{k+1}-\mathbf{u}^{k+1,*}\|_L^2\\ \leq&\left\|\mathbf{u}^{k+1}-\left(\frac{\theta_k}{\theta_0}\mathbf{u}^{0,*}+\theta_k\sum_{t=1}^k\left(\frac{\theta_t(1-\theta_0)^{k-t}}{\theta_{t-1}^2}-\frac{(1-\theta_0)^{k+1-t}}{\theta_{t-1}}\right)\mathbf{u}^{t,*}\right)\right\|_L^2\\ \leq&\frac{\theta_k}{\theta_0}\|\mathbf{z}^{k+1}-\mathbf{u}^{0,*}\|_L^2+\theta_k\sum_{t=1}^k\left(\frac{\theta_t(1-\theta_0)^{k-t}}{\theta_{t-1}^2}-\frac{(1-\theta_0)^{k+1-t}}{\theta_{t-1}}\right)\|\mathbf{u}^t-\mathbf{u}^{t,*}\|_L^2.\end{aligned}$$

Plugging it into (50) and letting $\sigma_k=\frac{\frac{\kappa\theta_0}{\theta_k}-(1-\theta_0)\kappa}{1+\frac{\kappa\theta_0}{\theta_k}}$, we have

$$\begin{aligned}&(1-\theta_0)\left(\mathbb{E}_{\xi_K}[D(\mathbf{u}^{k+1})]-D(\mathbf{u}^*)\right)+\mathbb{E}_{\xi_K}[\|\mathbf{u}^{k+1}-\mathbf{u}^{k+1,*}\|_L^2]\\ \leq&\frac{1+(1-\theta_0)\kappa}{1+\frac{\kappa\theta_0}{\theta_k}}\left(\frac{\theta_0}{\theta_k}\left(\mathbb{E}_{\xi_K}[D(\mathbf{u}^{k+1})]-D(\mathbf{u}^*)\right)+\frac{\theta_k}{\theta_0}\mathbb{E}_{\xi_K}[\|\mathbf{z}^{k+1}-\mathbf{u}^{0,*}\|_L^2]\right.\\ &\qquad\left.+\theta_k\sum_{t=1}^k\left(\frac{\theta_t(1-\theta_0)^{k-t}}{\theta_{t-1}^2}-\frac{(1-\theta_0)^{k+1-t}}{\theta_{t-1}}\right)\mathbb{E}_{\xi_K}[\|\mathbf{u}^t-\mathbf{u}^{t,*}\|_L^2]\right)\\ \overset{a}{\leq}&\frac{1+(1-\theta_0)\kappa}{1+\frac{\kappa\theta_0}{\theta_k}}\left(\frac{\theta_k}{\theta_0}T_0+\theta_k\sum_{t=1}^k\left(\frac{\theta_t(1-\theta_0)^{k-t}}{\theta_{t-1}^2}-\frac{(1-\theta_0)^{k+1-t}}{\theta_{t-1}}\right)\mathbb{E}_{\xi_K}[\|\mathbf{u}^t-\mathbf{u}^{t,*}\|_L^2]\right),\end{aligned}\tag{51}$$

where we denote $T_k=(1-\theta_0)\left(\mathbb{E}_{\xi_K}[D(\mathbf{u}^k)]-D(\mathbf{u}^*)\right)+\mathbb{E}_{\xi_K}[\|\mathbf{u}^k-\mathbf{u}^{k,*}\|_L^2]$ for simplicity and use (27) with $\mathbf{u}^*=\mathbf{u}^{0,*}$ in $\overset{a}{\leq}$. On the other hand, decomposing $\|\mathbf{u}^t-\mathbf{u}^{t,*}\|_L^2$ into

$\sigma\|\mathbf{u}^t-\mathbf{u}^{t,*}\|_L^2$ and $(1-\sigma)\|\mathbf{u}^t-\mathbf{u}^{t,*}\|_L^2$, using Assumption 2 for the first term and letting $\sigma=\frac{(1-\theta_0)\kappa}{1+(1-\theta_0)\kappa}$, we have

$$\mathbb{E}_{\xi_K}[\|\mathbf{u}^t-\mathbf{u}^{t,*}\|_L^2]\leq\frac{\sigma}{\kappa}\left(\mathbb{E}_{\xi_K}[D(\mathbf{u}^t)]-D(\mathbf{u}^*)\right)+(1-\sigma)\mathbb{E}_{\xi_K}[\|\mathbf{u}^t-\mathbf{u}^{t,*}\|_L^2]=\frac{T_t}{1+(1-\theta_0)\kappa}.$$

Plugging it into (51), we have

$$T_{k+1}\leq\frac{1+(1-\theta_0)\kappa}{1+\frac{\kappa\theta_0}{\theta_k}}\left(\frac{\theta_k}{\theta_0}T_0+\frac{\theta_k}{1+(1-\theta_0)\kappa}\sum_{t=1}^{k}\left(\frac{\theta_t(1-\theta_0)^{k-t}}{\theta_{t-1}^2}-\frac{(1-\theta_0)^{k+1-t}}{\theta_{t-1}}\right)T_t\right). \tag{52}$$

Next, we prove $T_{k+1}\leq\frac{1+(1-\theta_0)\kappa}{1+\kappa\theta_0^2/(2\theta_k^2)}T_0$ by induction. From (52), we have $T_1\leq\frac{1+(1-\theta_0)\kappa}{1+\kappa}T_0\leq\frac{1+(1-\theta_0)\kappa}{1+\kappa\theta_0^2/(2\theta_0^2)}T_0$. Assume that $T_t\leq\frac{1+(1-\theta_0)\kappa}{1+\kappa\theta_0^2/(2\theta_{t-1}^2)}T_0$ holds for $t\leq k$. Now, we consider $t=k+1$. From (52), we have

$$T_{k+1}\leq\frac{1+(1-\theta_0)\kappa}{1+\frac{\kappa\theta_0}{\theta_k}}\left(\frac{\theta_k}{\theta_0}+\theta_k\sum_{t=1}^{k}\left(\frac{\theta_t(1-\theta_0)^{k-t}}{\theta_{t-1}^2}-\frac{(1-\theta_0)^{k-(t-1)}}{\theta_{t-1}}\right)\frac{1}{1+\frac{\kappa\theta_0^2}{2\theta_{t-1}^2}}\right)T_0. \tag{53}$$

We can easily check

$$\begin{aligned}
&\sum_{t=1}^{k}\left(\frac{\theta_t(1-\theta_0)^{k-t}}{\theta_{t-1}^2}-\frac{(1-\theta_0)^{k-(t-1)}}{\theta_{t-1}}\right)\frac{1}{1+\frac{\kappa\theta_0^2}{2\theta_{t-1}^2}}\\
=&\sum_{t=1}^{k}\left(\frac{\theta_t}{\theta_{t-1}^2+\frac{\kappa\theta_0^2}{2}}(1-\theta_0)^{k-t}-\frac{\theta_{t-1}}{\theta_{t-1}^2+\frac{\kappa\theta_0^2}{2}}(1-\theta_0)^{k-(t-1)}\right)\\
=&\sum_{t=1}^{k}\left(\frac{\theta_t}{\theta_t^2+\frac{\kappa\theta_0^2}{2}}(1-\theta_0)^{k-t}-\frac{\theta_{t-1}}{\theta_{t-1}^2+\frac{\kappa\theta_0^2}{2}}(1-\theta_0)^{k-(t-1)}\right)+\sum_{t=1}^{k}\left(\frac{\theta_t}{\theta_{t-1}^2+\frac{\kappa\theta_0^2}{2}}-\frac{\theta_t}{\theta_t^2+\frac{\kappa\theta_0^2}{2}}\right)(1-\theta_0)^{k-t}\\
=&\frac{\theta_k}{\theta_k^2+\frac{\kappa\theta_0^2}{2}}-\frac{(1-\theta_0)^k}{\theta_0+\frac{\kappa\theta_0}{2}}+\sum_{t=1}^{k}\left(\frac{\theta_t}{\theta_{t-1}^2+\frac{\kappa\theta_0^2}{2}}-\frac{\theta_t}{\theta_t^2+\frac{\kappa\theta_0^2}{2}}\right)(1-\theta_0)^{k-t}
\end{aligned}$$

and

$$\frac{\theta_t}{\theta_{t-1}^2+\frac{\kappa\theta_0^2}{2}}-\frac{\theta_t}{\theta_t^2+\frac{\kappa\theta_0^2}{2}}\overset{a}{=}-\frac{\theta_t^2\theta_{t-1}^2}{\left(\theta_{t-1}^2+\frac{\kappa\theta_0^2}{2}\right)\left(\theta_t^2+\frac{\kappa\theta_0^2}{2}\right)}=-\frac{1}{\left(1+\frac{\kappa\theta_0^2}{2\theta_{t-1}^2}\right)\left(1+\frac{\kappa\theta_0^2}{2\theta_t^2}\right)}\leq-\frac{1}{\left(1+\frac{\kappa\theta_0^2}{2\theta_k^2}\right)^2},$$

where we use $\theta_{t-1}^2-\theta_t^2=\theta_{t-1}^2\theta_t^2\left(\frac{1}{\theta_t^2}-\frac{1}{\theta_{t-1}^2}\right)=\theta_{t-1}^2\theta_t$ in $\overset{a}{=}$. Plugging them into (53), we only need to prove

$$\left(1+\frac{\kappa\theta_0^2}{2\theta_k^2}\right)\left(\frac{\theta_k}{\theta_0}+\frac{\theta_k^2}{\theta_k^2+\frac{\kappa\theta_0^2}{2}}-\frac{\theta_k(1-\theta_0)^k}{\theta_0+\frac{\kappa\theta_0}{2}}-\frac{\theta_k\sum_{t=1}^{k}(1-\theta_0)^{k-t}}{\left(1+\frac{\kappa\theta_0^2}{2\theta_k^2}\right)^2}\right)\leq1+\frac{\kappa\theta_0}{\theta_k}. \tag{54}$$

After some simple calculations, (54) is equivalent to

$$1 \le \frac{\theta_0 \sum_{t=1}^{k}(1-\theta_0)^{k-t}}{\left(1+\frac{\kappa\theta_0^2}{2\theta_k^2}\right)^2} + \frac{(1-\theta_0)^k}{1+\frac{\kappa}{2}} + \frac{\frac{\kappa\theta_0^2}{\theta_k^2}}{1+\frac{\kappa\theta_0^2}{2\theta_k^2}}.$$

Since $\theta_0 \sum_{t=1}^{k}(1-\theta_0)^{k-t} = 1-(1-\theta_0)^k$, we only need to ensure $\frac{(1-\theta_0)^k}{1+\frac{\kappa}{2}} - \frac{(1-\theta_0)^k}{\left(1+\kappa\theta_0^2/(2\theta_k^2)\right)^2} \ge 0$ and $\frac{1}{\left(1+\kappa\theta_0^2/(2\theta_k^2)\right)^2} + \frac{\kappa\theta_0^2/\theta_k^2}{1+\kappa\theta_0^2/(2\theta_k^2)} \ge 1$. Both inequalities hold for any $\theta_k$ and any $\kappa$. ■

Now we consider the outer iterations of Algorithm 3. Replace $\mathbf{u}^{k+1}$, $\mathbf{u}^0$, $\mathbf{u}^{k+1,*}$ and $\mathbf{u}^{0,*}$ in Lemma 17 by $\mathbf{u}^{t,K+1}$, $\mathbf{u}^{t,0}$, $\mathbf{u}^{t,K+1,*}$ and $\mathbf{u}^{t,0,*}$, respectively. From $\frac{1}{\theta_K^2} \ge \left(\frac{K}{2}+\hat{n}\right)^2$ and $\mathbf{u}^{t,K+1} = \mathbf{u}^{t+1,0}$, we can immediately have (43).

### 4.2. Convergence Rate Analysis of the Primal Solutions

In this section, we prove (44). We first establish the relation between the primal objective and dual objective in the following lemma and then (44) can be attained by Lemmas 17 and 18 immediately.

**Lemma 18** *Suppose Assumptions 1 and 2 hold. Assume that $\mathbb{D}^*$ is bounded, i.e., $\|\mathbf{u}^*\|_L \le C_{\mathbb{D}^*}, \forall \mathbf{u}^* \in \mathbb{D}^*$. Let $K_0 \le \left\lfloor \frac{K}{\upsilon(1+1/\hat{n})} + 1 \right\rfloor$ with any $\upsilon > 1$ and $K \ge \hat{n}$. Then for Algorithm 3 we have*

$$\left|\mathbb{E}_{\zeta_t}\left[F(\hat{\mathbf{x}}^{t,K})\right] - F(\mathbf{x}^*)\right| \le \left(\frac{2T_{t,0} + 6C_{\mathbb{D}^*}\sqrt{T_{t,0}} + 6M\sqrt{\sum_{i=1}^n L_i}\sqrt{T_{t,0}}}{1-1/\nu} + \frac{2\sqrt{m}C_{\mathbb{D}^*}\sqrt{T_{t,0}}}{\sqrt{1-1/\nu}}\right),$$

$$\mathbb{E}_{\zeta_t}\left[\left\|\begin{bmatrix} \boldsymbol{B}\hat{\mathbf{x}}^{t,K} + \mathbf{b} \\ \max\left\{0, g(\hat{\mathbf{x}}^{t,K})\right\}\end{bmatrix}\right\|_L^*\right] \le \frac{6\sqrt{T_{t,0}}}{1-1/\nu},$$

*where $T_{t,0} = (1-\theta_0)\left(\mathbb{E}_{\zeta_{t-1}}[D(\mathbf{u}^{t,0})] - D(\mathbf{u}^*)\right) + \mathbb{E}_{\zeta_{t-1}}[\|\mathbf{u}^{t,0} - \mathbf{u}^{t,0,*}\|_L^2]$.*

**Proof** We denote $\mathbf{u}^{t,k}, \mathbf{z}^{t,k}, \mathbf{v}^{t,k}, i_{t,k}, \xi_{t,k}, \hat{\mathbf{x}}^{t,K}, \hat{\mathbf{y}}^{t,K}$ to be the variables at the $t$-th outer iteration of Algorithm 3, which are the counterparts of $\mathbf{u}^k, \mathbf{z}^k, \mathbf{v}^k, i_k, \xi_k, \hat{\mathbf{x}}^K$ and $\hat{\mathbf{y}}^K$ in Algorithm 1. Choose $\mathbf{u}^* = \mathbf{u}^{t,0,*}$ and let $\mathbf{u} = \mathbf{u}^{t,0,*}$ in (24), which is independent on $\xi_{t,K}$ conditioned on $\zeta_{t-1}$. For the $t$-th outer iteration of Algorithm 3, we have

$$\begin{aligned}&\left(\frac{1}{\theta_K^2} - \frac{1}{\theta_{K_0-1}^2}\right)\mathbb{E}_{\xi_{t,K}|\zeta_{t-1}}\left[\left\langle \triangle(\hat{\mathbf{x}}^{t,K}, \hat{\mathbf{y}}^{t,K}), \mathbf{u}^{t,0,*}\right\rangle + D(\mathbf{u}^*) + f(\hat{\mathbf{x}}^{t,K}) + \frac{1}{n}\phi(\hat{\mathbf{y}}^{t,K})\right]\\ &\le 2\hat{n}^2\left((1-\theta_0)\left(D(\mathbf{u}^{t,0}) - D(\mathbf{u}^*)\right) + \|\mathbf{u}^{t,0} - \mathbf{u}^{t,0,*}\|_L^2\right).\end{aligned}$$

Taking expectation with respect to $\zeta_{t-1}$ and using the forth property of Lemma 6, we have

$$\mathbb{E}_{\zeta_t}\left[\left\langle \triangle(\hat{\mathbf{x}}^{t,K}, \hat{\mathbf{y}}^{t,K}), \mathbf{u}^{t,0,*}\right\rangle + D(\mathbf{u}^*) + f(\hat{\mathbf{x}}^{t,K}) + \frac{1}{n}\phi(\hat{\mathbf{y}}^{t,K})\right] \le \frac{2\hat{n}^2}{(K^2/4 + \hat{n}K)(1-1/\nu)}T_{t,0}. \quad (55)$$

Choosing $\mathbf{u}^* = \mathbf{u}^{t,0,*}$ in (34) and using a similar induction, we have

$$\mathbb{E}_{\zeta_t}[\|\widetilde{\triangle}(\hat{\mathbf{x}}^{t,K},\hat{\mathbf{y}}^{t,K})\|_L^*]\leq\mathbb{E}_{\zeta_{t-1}}\sqrt{\mathbb{E}_{\xi_{t,K}|\zeta_{t-1}}\left[\left(\|\widetilde{\triangle}(\hat{\mathbf{x}}^K,\hat{\mathbf{y}}^K)\|_L^*\right)^2\right]}\leq\frac{7\hat{n}^2}{(K^2/4+\hat{n}K)(1-1/\nu)}\sqrt{T_{t,0}}. \tag{56}$$

Now, we consider the objective function. From (55) and the definition in (6), we have

$$\mathbb{E}_{\zeta_t}\left[L_F(\hat{\mathbf{x}}^{t,K},\hat{\mathbf{y}}^{t,K},\mathbf{u}^{t,0,*}) - L_F(\mathbf{x}^*,\mathbf{y}^*,\mathbf{u}^{t,0,*})\right] \leq \frac{2\hat{n}^2}{(K^2/4+\hat{n}K)(1-1/\nu)}T_{t,0}.$$

Since $(\mathbf{x}^*,\mathbf{y}^*,\mathbf{u}^{t,0,*})$ satisfies the KKT condition, we have $\mathbf{0}\in\partial_{\mathbf{x},\mathbf{y}}L_F(\mathbf{x}^*,\mathbf{y}^*,\mathbf{u}^{t,0,*})$. On the other hand, $L_F(\mathbf{x},\mathbf{y},\mathbf{u}^{t,0,*})$ is convex with respect to $(\mathbf{x},\mathbf{y})$ and $\mu$-strongly convex with respect to $\mathbf{x}$, so we have

$$L_F(\hat{\mathbf{x}}^{t,K},\hat{\mathbf{y}}^{t,K},\mathbf{u}^{t,0,*}) - L_F(\mathbf{x}^*,\mathbf{y}^*,\mathbf{u}^{t,0,*}) \geq \frac{\mu}{2}\|\hat{\mathbf{x}}^{t,K}-\mathbf{x}^*\|^2,$$

which leads to $\mathbb{E}_{\zeta_t}\left[\|\hat{\mathbf{x}}^{t,K}-\mathbf{x}^*\|^2\right] \leq \frac{4\hat{n}^2}{\mu(K^2/4+\hat{n}K)(1-1/\nu)}T_{t,0}$ and

$$\mathbb{E}_{\zeta_t}[\|\hat{\mathbf{x}}^{t,K}-\mathbf{x}^*\|] \leq \frac{2\hat{n}}{\sqrt{\mu(K^2/4+\hat{n}K)(1-1/\nu)}}\sqrt{T_{t,0}}. \tag{57}$$

Let $\mathbb{I}$ be the index set such that for any $i\in\mathbb{I}$, we have $\mathbf{u}_{n+p+i}^{t,0,*}>0$ and $g_i(\hat{\mathbf{x}}^{t,K})<0$. So we have

$$\begin{aligned}
&\mathbb{E}_{\zeta_t}\left[\left\langle\triangle(\hat{\mathbf{x}}^{t,K},\hat{\mathbf{y}}^{t,K}),\mathbf{u}^{t,0,*}\right\rangle\right]\\
=&\mathbb{E}_{\zeta_t}\left[\left\langle\widetilde{\triangle}(\hat{\mathbf{x}}^{t,K},\hat{\mathbf{y}}^{t,K}),\mathbf{u}^{t,0,*}\right\rangle\right] + \sum_{i\in\mathbb{I}}\mathbb{E}_{\zeta_t}\left[\mathbf{u}_{n+p+i}^{t,0,*}g_i(\hat{\mathbf{x}}^{t,K})\right]\\
\overset{a}{=}&\mathbb{E}_{\zeta_t}\left[\left\langle\widetilde{\triangle}(\hat{\mathbf{x}}^{t,K},\hat{\mathbf{y}}^{t,K}),\mathbf{u}^{t,0,*}\right\rangle\right] + \sum_{i\in\mathbb{I}}\mathbb{E}_{\zeta_t}\left[\mathbf{u}_{n+p+i}^{t,0,*}(g_i(\hat{\mathbf{x}}^{t,K})-g_i(\mathbf{x}^*))\right]\\
\geq&-\mathbb{E}_{\zeta_t}\left[\|\mathbf{u}^{t,0,*}\|_L\|\widetilde{\triangle}(\hat{\mathbf{x}}^{t,K},\hat{\mathbf{y}}^{t,K})\|_L^*\right] - \mathbb{E}_{\zeta_t}\left[\|\mathbf{u}^{t,0,*}\|_L\sqrt{\sum_{i\in\mathbb{I}}\frac{1}{L_{n+p+i}}\left(g_i(\hat{\mathbf{x}}^{t,K})-g_i(\mathbf{x}^*)\right)^2}\right]\\
\overset{b}{\geq}&-\mathbb{E}_{\zeta_t}\left[\|\mathbf{u}^{t,0,*}\|_L\|\widetilde{\triangle}(\hat{\mathbf{x}}^{t,K},\hat{\mathbf{y}}^{t,K})\|_L^*\right] - \sqrt{m\mu}\mathbb{E}_{\zeta_t}\left[\|\mathbf{u}^{t,0,*}\|_L\|\hat{\mathbf{x}}^{t,K}-\mathbf{x}^*\|\right]\\
\overset{c}{\geq}&-\left(\frac{7\hat{n}^2}{(K^2/4+\hat{n}K)(1-1/\nu)}+\frac{2\hat{n}\sqrt{m}}{\sqrt{(K^2/4+\hat{n}K)(1-1/\nu)}}\right)C_{\mathbb{D}^*}\sqrt{T_{t,0}},
\end{aligned}$$

where in $\overset{a}{=}$ we use $\mathbf{u}_{n+p+i}^{t,0,*}g_i(\mathbf{x}^*)=0$ from the complementary slackness in the KKT condition. From Assumption 1.3 and (12), we have $\sum_{i\in\mathbb{I}}\frac{1}{L_{n+p+i}}\left(g_i(\hat{\mathbf{x}}^{t,K})-g_i(\mathbf{x}^*)\right)^2 \leq \sum_{i\in\mathbb{I}}\frac{L_{g_i}^2}{L_{n+p+i}}\|\hat{\mathbf{x}}^{t,K}-\mathbf{x}^*\|^2\leq m\mu\|\hat{\mathbf{x}}^{t,K}-\mathbf{x}^*\|^2$, which leads to $\overset{b}{\geq}$. In $\overset{c}{\geq}$, we use $\|\mathbf{u}^{t,0,*}\|_L\leq C_{\mathbb{D}^*}$, (56) and (57). From (55), we have

$$\mathbb{E}_{\zeta_t}\left[D(\mathbf{u}^*)+f(\hat{\mathbf{x}}^{t,K})+\frac{1}{n}\phi(\hat{\mathbf{y}}^{t,K})\right]\leq\frac{2\hat{n}^2T_{t,0}+7\hat{n}^2C_{\mathbb{D}^*}\sqrt{T_{t,0}}}{(K^2/4+\hat{n}K)(1-1/\nu)}+\frac{2\hat{n}\sqrt{m}C_{\mathbb{D}^*}\sqrt{T_{t,0}}}{\sqrt{(K^2/4+\hat{n}K)(1-1/\nu)}}.$$

From $D(\mathbf{u}^*)=-F(\mathbf{x}^*)$, (42), (56), (33) and $K\geq\hat{n}$, we have the conclusion. ■

## 5. Application to the Regularized ERM

The regularized empirical risk minimization problem (2) has broad applications in machine learning. For the special problem (2), its dual problem (9) becomes

$$\min_{\mathbf{u}\in\mathbb{R}^n} D(\mathbf{u}) = d(\mathbf{u}) + h(\mathbf{u}) \equiv f^*\left(-\frac{\boldsymbol{A}\mathbf{u}}{n}\right) + \frac{1}{n}\sum_{i=1}^{n}\phi_i^*(\mathbf{u}_i). \tag{58}$$

We follow (Shalev-Shwartz and Zhang, 2013, 2016) to assume $\|\boldsymbol{A}_i\| \leq 1, \forall i$, which can be guaranteed by normalizing the data. Then we have $L = \frac{\|\boldsymbol{A}\|_2^2}{n^2\mu}$ and $L_j = \frac{\|\boldsymbol{A}_j\|^2}{n^2\mu} \leq \frac{1}{n^2\mu}$ from (11) and (12). From Lemmas 21 and 22 in (Shalev-Shwartz and Zhang, 2013), we have $|\mathbf{z}_i^k| \leq M, |\widetilde{\mathbf{z}}_i^k| \leq M, k = 0, 1, \cdots, K$, and $|\mathbf{u}_i^*| \leq M$, which leads to $\|\mathbf{u}^0 - \mathbf{u}^*\|_L^2 \leq \frac{4M^2}{n\mu}$ and $L\|\mathbf{u}^0 - \mathbf{u}^*\|^2 \leq \frac{4M^2}{n\mu}\|\boldsymbol{A}\|_2^2$. We will discuss the iteration complexity of ARDCA in three scenarios.

### 5.1. Strongly Convex and Nonsmooth $f$, Convex and Nonsmooth $\phi_i$

From Theorem 3, we know that the convergence rate of ARDCA for problem (2) is:

$$\mathbb{E}_{\xi_K}[F(\hat{\mathbf{x}}^K)] - F(\mathbf{x}^*) \leq \frac{9nM^2\left(6 + \frac{n\mu}{M^2}\left(D(\mathbf{u}^0) - D(\mathbf{u}^*)\right)\right)}{\mu(K^2/4 + nK)(1 - 1/\upsilon)}.$$

In order to have the $O\left(\frac{nM^2}{\mu K^2}\right)$ convergence rate for ARDCA, we should find an initializer good enough such that $D(\mathbf{u}^0) - D(\mathbf{u}^*) \leq O\left(\frac{M^2}{n\mu}\right)$. We use ARDCA with fixed $\theta_k = \frac{1}{n}$, i.e., non-accelerated RDCA, to find such initializer. Specifically, we describe the method in Algorithm 4. Lemma 19 establishes the convergence rates of both the primal solutions and dual solutions for the first step of Algorithm 4, whose proof is given in Appendix E.

**Algorithm 4** ARDCA for ERM

```
  Input $\mathbf{u}^0 \in \mathbb{D}$, $K'$, $K_0$, $K$.
  Run ARDCA($\mathbf{u}^0$,0,$K'$) with fixed $\theta_k = \frac{1}{n}$ and output $\mathbf{u}^{K'+1}$ and $\mathbf{x}^*(\mathbf{v}^{K'})$.
  Run ARDCA($\mathbf{u}^{K'+1}$,$K_0$,$K$) with decreasing $\theta_k$ and output $\mathbf{u}^{K+1}$ and $\hat{\mathbf{x}}^K$.
```

**Lemma 19** *Let $K' = \left\lceil n\log\left(\min\left\{\frac{1}{\epsilon}, \frac{n\mu}{M^2}\right\}\left(D(\mathbf{u}^0) + F(\mathbf{x}^*(\mathbf{u}^0))\right)\right) - 1\right\rceil$. Suppose Assumptions 1.1, 1.2 and 1.5 hold. Then for step 1 of Algorithm 4, we have.*

$$\mathbb{E}_{\xi_{K'}}[D(\mathbf{u}^{K'+1})] - D(\mathbf{u}^*) \leq 9\max\left\{\epsilon, \frac{M^2}{n\mu}\right\}, \tag{59}$$

$$\mathbb{E}_{\xi_{K'}}[F(\mathbf{x}^*(\mathbf{v}^{K'})] - F(\mathbf{x}^*) \leq 17\max\left\{\epsilon, \frac{M^2}{n\mu}\right\}. \tag{60}$$

An immediate consequence of Lemma 19 is that if $\epsilon \geq O\left(\frac{M^2}{n\mu}\right)$, we only need to run step 1 of Algorithm 4 with linear complexity to achieve an $\epsilon$-optimal solution. We describe the results in Corollary 20. However, in statistical learning, $\mu$ is usually on the order of $\frac{1}{\sqrt{n}}$

or $\frac{1}{n}$ (Bousquet and Elisseeff, 2002; Zhang and Xiao, 2017), thus $\frac{M^2}{n\mu}$ is often not too small. In the following discussions, we only consider the case of $\epsilon < O\left(\frac{M^2}{n\mu}\right)$.

**Corollary 20** *If $\epsilon \geq O\left(\frac{M^2}{n\mu}\right)$, we only need to run ARDCA($\mathbf{u}^0$,0,$K'$) with fixed $\theta_k = \frac{1}{n}$, i.e., non-accelerated RDCA, for $K' = \left\lceil n \log\left(\frac{D(\mathbf{u}^0)+F(\mathbf{x}^*(\mathbf{u}^0))}{\epsilon}\right)\right\rceil$ iterations to find an $\epsilon$-optimal solution such that*

$$\mathbb{E}_{\xi_{K'}}[F(\mathbf{x}^*(\mathbf{v}^{K'}))] - F(\mathbf{x}^*) \leq \epsilon, \quad \mathbb{E}_{\xi_{K'}}[D(\mathbf{u}^{K'+1})] - D(\mathbf{u}^*) \leq \epsilon.$$

*If $\epsilon < O\left(\frac{M^2}{n\mu}\right)$, Algorithm 4 needs $K'+K = O\left(n \log\left(\frac{n\mu}{M^2}\left(D(\mathbf{u}^0) + F(\mathbf{x}^*(\mathbf{u}^0))\right)\right) + M\sqrt{\frac{n}{\mu\epsilon}}\right)$ iterations to find an $\epsilon$-optimal solution such that*

$$\mathbb{E}_{\xi_K \cup \xi_{K'}}[F(\hat{\mathbf{x}}^K)] - F(\mathbf{x}^*) \leq \epsilon, \quad \mathbb{E}_{\xi_K \cup \xi_{K'}}[D(\mathbf{u}^{K+1})] - D(\mathbf{u}^*) \leq \epsilon.$$

Algorithm 1 is a special case of APCG (Lin et al., 2015b). Lin et al. (2015b) only established the $O\left(n\sqrt{\frac{C}{\epsilon}}\right)$ iteration complexity in the dual space to achieve an $\epsilon$-optimal dual solution[8], where $C = D(\mathbf{u}^0) - D(\mathbf{u}^*) + \frac{1}{2}\|\mathbf{z}^0 - \mathbf{u}^*\|_L^2 \leq D(\mathbf{u}^0) - D(\mathbf{u}^*) + \frac{2M^2}{n\mu}$. Shalev-Shwartz and Zhang (2016) developed an accelerated SDCA with an inner-outer iteration procedure, where the outer loop is a full-dimensional accelerated proximal point method. At each iteration of the outer loop, SDCA is called to solve a subproblem inexactly. ASDCA is mainly used for the problems with smooth $\phi_i$. When $\phi_i$ is nonsmooth, Shalev-Shwartz and Zhang (2016) used ASDCA to solve a smoothed problem of (2), i.e., a regularized problem of (58), and achieved a slightly worse iteration complexity of $O\left(\left(n + M\sqrt{\frac{n}{\mu\epsilon}}\right)\log\frac{1}{\epsilon}\right)$ to find an $\epsilon$-optimal primal solution. We can also use Catalyst (Lin et al., 2015a) to solve the problems with nonsmooth $\phi_i$ without using smoothing. However, Catalyst also yields the additional $\left(\log\frac{1}{\epsilon}\right)$ factor. To make ASDCA faster than SDCA, which has the $O\left(n\log\frac{n\mu}{M^2} + \frac{M^2}{\mu\epsilon}\right)$ complexity (Shalev-Shwartz and Zhang, 2013), Shalev-Shwartz and Zhang (2016) required $\epsilon \leq \frac{M^2}{n\mu}$. Katyusha (Allen-Zhu, 2018), a primal-only algorithm, obtains the state-of-the-art iteration complexity of $O\left(n\log\frac{F(\mathbf{x}^0)-F(\mathbf{x}^*)}{\epsilon} + M\sqrt{\frac{n}{\mu\epsilon}}\right)$, which is worse than our result when $\epsilon \leq \frac{M^2}{n\mu}$. Our result matches the theoretical lower bound of $\left(n + M\sqrt{\frac{n}{\mu\epsilon}}\right)$ (Woodworth and Srebro, 2016) when ignoring the constant term of $n\log n$. All the compared methods need $O(t)$ runtime at each iteration.

### 5.2. Strongly Convex and Nonsmooth $f$, Convex and Smooth $\phi_i$

When each $\phi_i$ is $1/\gamma$-smooth, which is defined as $\phi_i(u) \leq \phi_i(v) + \langle\nabla\phi_i(v), u - v\rangle + \frac{1}{2\gamma}\|u-v\|^2$, then $\phi_i^*$ is $\gamma$-strongly convex and $D(\mathbf{u})$ is $\frac{\gamma}{n}$-strongly convex. In this case, Assumption 2 is satisfied with $\kappa = n\gamma\mu$. From the discussion at the end of Section 4, we know that

[8] When $\phi_i$ has Lipchitz continuous gradient, Lin et al. (2015b) proved the linear convergence rate in the primal space. However, when $\phi_i$ is only Lipchitz continuous, the convergence rate in the primal space is not established in (Lin et al., 2015b).

Algorithm 3 needs $O\left(\left(n+\sqrt{\frac{n}{\gamma\mu}}\right)\log\frac{1}{\epsilon}\right)$ iterations in total to achieve an $\epsilon$-optimal primal solution and dual solution in the best case scenario, which matches or outperforms the complexities of the dual based stochastic algorithms established in (Lin et al., 2015b; Shalev-Shwartz and Zhang, 2016). Specifically, their complexities are $O\left(\left(n+\sqrt{\frac{n}{\gamma\mu}}\right)\log\frac{1}{\epsilon}\right)$ and $O\left(\left(n+\sqrt{\frac{n}{\gamma\mu}}\right)\log\frac{1}{\epsilon}\log^2\frac{1}{n\gamma\mu}\right)$, respectively. To make the accelerated algorithms faster than the non-accelerated counterparts, Shalev-Shwartz and Zhang (2016) required $\frac{1}{n\gamma\mu}\gg 1$ (i.e., $\kappa\ll 1$). Thus, our complexity has a better dependence on $\left(\log\frac{1}{n\gamma\mu}\right)$ than (Shalev-Shwartz and Zhang, 2016). Note that APCG (Lin et al., 2015b) needs an extra proximal full gradient step to establish the linear convergence rate in the primal space. Our analysis does not need such an additional operation. On the other hand, Lin et al. (2015b) and Shalev-Shwartz and Zhang (2016) did not study the case when $\kappa$ is unknown.

### 5.3. Strongly Convex and Smooth $f$, Convex and Nonsmooth $\phi_i$

As discussed in Section 4, Assumption 2 is weaker than the strong convexity of $D(\mathbf{u})$ and some special cases of problem (1) with nonsmooth $\phi_i$ also satisfy Assumption 2. We take SVM and the least absolute deviation as examples. The primal problem and dual problem of SVM are

$$\begin{aligned}&\min_{\mathbf{x}\in\mathbb{R}^t} F(\mathbf{x})=\frac{\mu}{2}\|\mathbf{x}\|^2+\frac{1}{n}\sum_{i=1}^n\max\{0,1-l_i\boldsymbol{A}_i^T\mathbf{x}\},\\ &\min_{\mathbf{u}\in\mathbb{R}^n} D(\mathbf{u})=\frac{1}{2\mu}\left\|\frac{\widetilde{\boldsymbol{A}}\mathbf{u}}{n}\right\|^2-\frac{1}{n}\sum_{i=1}^n\mathbf{u}_i+I_{[0,1]}(\mathbf{u}),\end{aligned}\tag{61}$$

where $\widetilde{\boldsymbol{A}}_i=l_i\boldsymbol{A}_i$ and $l_i$ is the label for the $i$-th data $\boldsymbol{A}_i$, $I_{[0,1]}(\mathbf{u})=\begin{cases}0 & \text{if } 0\leq\mathbf{u}\leq 1,\\ \infty & \text{otherwise.}\end{cases}$ Wang and Lin (2014) proved that $D(\mathbf{u})$ in (61) satisfies the global error bound condition. From (Drusvyatskiy and Lewis, 2018), we know that Assumption 2 holds. From the discussion at the end of Section 4, we can see that Algorithm 3 needs $O\left(\frac{n}{\sqrt{\kappa}}\log\frac{1}{\epsilon}\right)$ iterations in total to achieve an $\epsilon$-optimal primal solution and dual solution in the best case scenario. As a comparison, Ma et al. (2016) studied the randomized coordinate descent and established the $O\left(\frac{n}{\kappa}\log\frac{1}{\epsilon}\right)$ iteration complexity. The better dependence on $\kappa$ in our iteration complexity is significant when $\kappa$ is small and this is often the case in practice. We refer the reader to Section 5 of (Ma et al., 2016) for the discussion on the size of $\kappa$. When $\kappa$ is unknown, Algorithm 3 is still a better choice for a wide range of inner iteration number than the randomized coordinate descent.

For the least absolute deviation, its primal problem and dual problem are

$$\begin{aligned}&\min_{\mathbf{x}\in\mathbb{R}^t} F(\mathbf{x})=\frac{\mu}{2}\|\mathbf{x}\|^2+\|\boldsymbol{A}^T\mathbf{x}-\mathbf{b}\|_1,\\ &\min_{\mathbf{u}\in\mathbb{R}^n} D(\mathbf{u})=\frac{1}{2\mu}\|\boldsymbol{A}\mathbf{u}\|^2+\langle\mathbf{u},\mathbf{b}\rangle+I_{[-1,1]}(\mathbf{u}).\end{aligned}\tag{62}$$

Similar to SVM, $D(\mathbf{u})$ in (62) also satisfies Assumption 2 and Algorithm 3 needs $O\left(\frac{n}{\sqrt{\kappa}}\log\frac{1}{\epsilon}\right)$ iterations to achieve an $\epsilon$-optimal primal solution and dual solution in the best case scenario.

## 6. Numerical Experiments

In this section, we test the performance of Algorithms 1, 3 and 4 on the sparse recovery problem. Consider the sparse linear regression problem of $\mathbf{b} = \boldsymbol{A}^T\mathbf{x} + \mathbf{w}$, where $\mathbf{x} \in \mathbb{R}^t$ is the unknown sparse vector to estimate, $\mathbf{b} \in \mathbb{R}^n$ is the observation and $\mathbf{w}$ is some additive noise. A particular instance of this problem is compressed sensing (Candes et al., 2006). In order to recovery $\mathbf{x}$, a popular regularization is the $l_1$-norm, in which case people often solve the following problems:

$$\min_{\mathbf{x}\in\mathbb{R}^t} f(\mathbf{x}), \quad s.t. \quad \|\boldsymbol{A}^T\mathbf{x} - \mathbf{b}\|_\alpha \le \tau \qquad \text{or} \qquad \min_{\mathbf{x}\in\mathbb{R}^t} \lambda f(\mathbf{x}) + \|\boldsymbol{A}^T\mathbf{x} - \mathbf{b}\|_\alpha,$$

where $f(\mathbf{x}) = \|\mathbf{x}\|_1 + \frac{\mu}{2}\|\mathbf{x}\|^2$. We add the term $\frac{\mu}{2}\|\mathbf{x}\|^2$ to make the objective function strongly convex and thus we can use some fast convergent algorithms. When the noise is generated from the Gaussian distribution, people often use the $l_2$ loss function, i.e., $\alpha = 2$. When the noise is sparse and the data contains some outliers, the $l_1$ loss is often used, i.e., $\alpha = 1$. When the noise is generated from a uniform distribution, we often use the $l_\infty$ loss instead. In this section, we solve the following three problems

$$\min_{\mathbf{x}\in\mathbb{R}^t} F(\mathbf{x}) \equiv \lambda f(\mathbf{x}) + \frac{1}{2n}\|\boldsymbol{A}^T\mathbf{x} - \mathbf{b}\|_2^2, \tag{63}$$

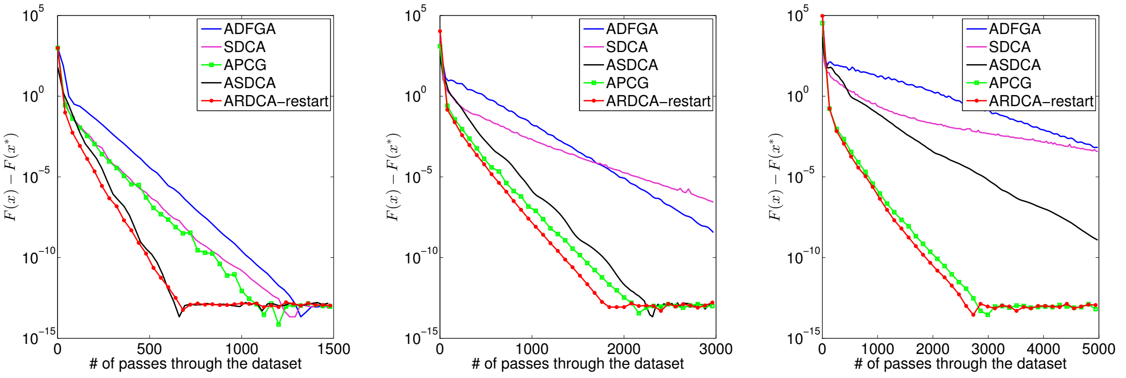
$$\min_{\mathbf{x}\in\mathbb{R}^t} F(\mathbf{x}) \equiv \lambda f(\mathbf{x}) + \frac{1}{n}\|\boldsymbol{A}^T\mathbf{x} - \mathbf{b}\|_1, \tag{64}$$

$$\min_{\mathbf{x}\in\mathbb{R}^t} f(\mathbf{x}), \quad s.t. \quad -\tau\mathbf{1} \le \boldsymbol{A}^T\mathbf{x} - \mathbf{b} \le \tau\mathbf{1}. \tag{65}$$

Problems (63) and (64) are special cases of problem (2) satisfying the assumptions in Sections 5.2 and 5.1 and problem (65) is a special case of problem (3), respectively. In our numerical experiment, we set $t = 1000$, $n = 200$ and $\mu = 0.1$. We generate the entries of $\boldsymbol{A}$ from the uniform distribution in $[0, 1]$ and normalize each column of $\boldsymbol{A}$ such that $\|\boldsymbol{A}_i\| = 1$. We set $t/10$ entries of $\mathbf{x}$ to be nonzeros. $\mathbf{b}$ is generated by $\boldsymbol{A}^T\mathbf{x} + \mathbf{w}$, where we generate each entry of noise $\mathbf{w}$ from the Gaussian distribution $N(0, \tau)$ for problem (63), generate $n/10$ entries of $\mathbf{w}$ from $N(0, \tau)$ and set the others to be 0 for problem (64), and generate each entry of $\mathbf{w}$ from the uniform distribution in $[-\tau, \tau]$ for problem (65). We vary $\lambda$ in the range $\{10^{-3}, 10^{-4}, 10^{-5}\}$ in problems (63) and (64) and $\tau$ in the range $\{10^{-3}, 10^{-4}, 10^{-5}\}$ in problem (65).

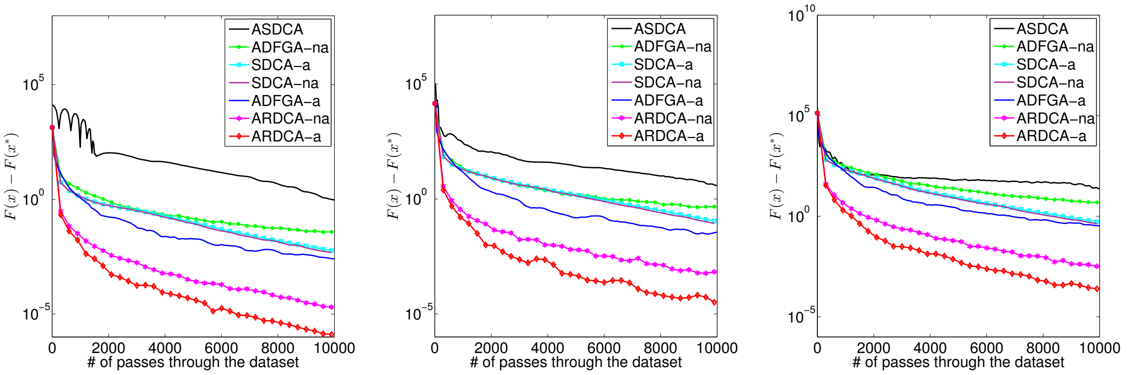
For problem (63), we compare ARDCA-restart (Algorithm 3) with ASDCA (Shalev-Shwartz and Zhang, 2016), APCG (Lin et al., 2015b), SDCA (Shalev-Shwartz and Zhang, 2013) and ADFGA (Beck and Teboulle, 2014). Figure 1 plots the primal gap as functions of the number of passes over the data, where each $n$ (inner) iterations are equivalent to a single pass over the data for APCG and SDCA (ARDCA and ASDCA). We use the maximal dual objective value produced by the compared methods to approximate the optimal primal objective value $F(\mathbf{x}^*)$. We can see that ARDCA-restart outperforms the non-accelerated

SDCA and non-randomized ADFGA for a wide range of $\lambda$ and ARDCA-restart is superior to APCG and ASDCA for some values of $\lambda$.



$\lambda = 10^{-3}$ $\lambda = 10^{-4}$ $\lambda = 10^{-5}$

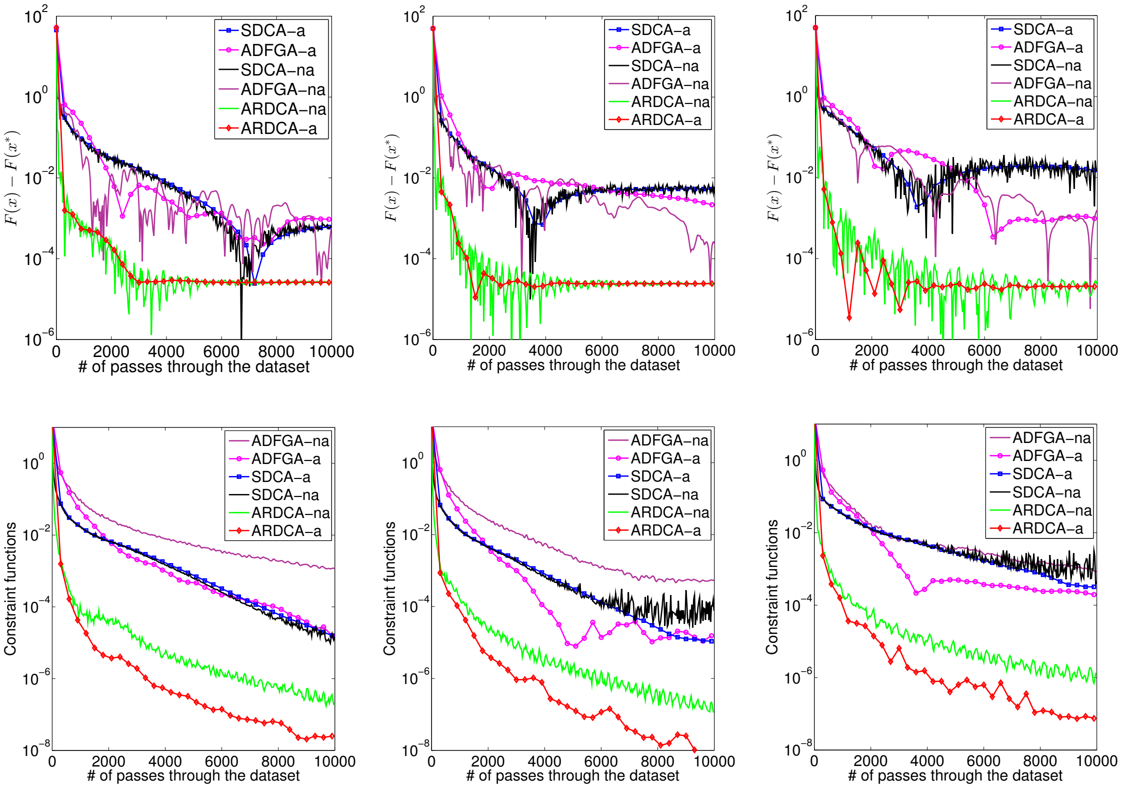
Figure 1: Comparing ARDCA-restart with SDCA, APCG, ASDCA and ADFGA on problem (63).



$\lambda = 10^{-3}$ $\lambda = 10^{-4}$ $\lambda = 10^{-5}$

Figure 2: Comparing ARDCA with SDCA, ASDCA and ADFGA on problem (64).

For problem (64), we compare ARDCA (Algorithm 4) with ASDCA (Shalev-Shwartz and Zhang, 2016), SDCA (Shalev-Shwartz and Zhang, 2013) and ADFGA (Beck and Teboulle, 2014), where ASDCA solves a regularized dual problem of (64) by adding term $\frac{\epsilon}{2}\|\mathbf{u}\|^2$ to the dual objective with $\epsilon = 10^{-6}$. We set $\upsilon = 1.1$ in Algorithm 1. Figure 2 plots the results, where ARDCA-a means that we test the averaged primal solution and ARDCA-na means the non-averaged primal solution. We can see that ARDCA-a yields the best result by orders of magnitude. Specially, ASDCA with regularization does not perform well although it converges linearly when $\phi_i$ is smooth. Thus, although the regularization/smoothing based ASDCA has the near optimal theoretical result (the sub-optimality comes from the $\left(\log \frac{1}{\epsilon}\right)$ factor), its practical performance is not satisfactory.



$\tau = 10^{-3}$ $\tau = 10^{-4}$ $\tau = 10^{-5}$

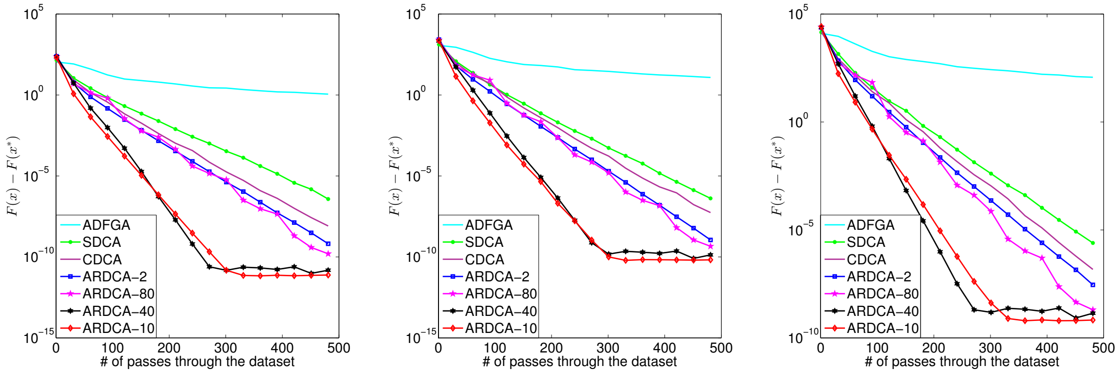
Figure 3: Comparing ARDCA with SDCA and ADFGA on problem (65). Top: objective function. Bottom: constraint functions.

For problem (65), we compare ARDCA with SDCA and ADFGA. As demonstrated in Figure 3, we can see that ARDCA-a performs the best in both reducing the primal gap and constraint function value.

At last, we consider problem (64) with $f(\mathbf{x}) = \frac{\mu}{2}\|\mathbf{x}\|^2$ to verify the conclusions in Section 5.3. In this scenario, we generate $\mathbf{x}$ to be a dense vector. We compare ARDCA with restart (Algorithm 3) with SDCA (Shalev-Shwartz and Zhang, 2013), the Cyclic Dual Coordinate Ascent (CDCA) (Wang and Lin, 2014) and ADFGA (Beck and Teboulle, 2014). Since the quadratic functional growth parameter $\kappa$ is unknown, we test Algorithm 3 with different inner iteration number $K_t \in \{2n, 10n, 40n, 80n\}$. From Figure 4 we can see that ARDCA, SDCA and CDCA all converge linearly and ARDCA with suitable $K$ performs the best. This verifies our theories in Section 5.3.

## 7. Conclusion

In this paper, we prove that the iteration complexities of the primal solutions and dual solutions have the same order of magnitude for the accelerated randomized dual coordinate ascent. Specifically, when $f(\mathbf{x})$ is $\mu$-strongly convex and the objectives are nonsmooth, we



$\lambda = 10^{-3}$ $\lambda = 10^{-4}$ $\lambda = 10^{-5}$

Figure 4: Comparing ARDCA-restart with SDCA, CDCA and ADFGA on problem (64) with smooth $f(\mathbf{x})$.

establish the $O\left(\frac{1}{\sqrt{\epsilon}}\right)$ iteration complexity. When the dual function further satisfies the quadratic functional growth condition, we prove the linear iteration complexity even if the condition number is unknown. When applied to the regularized empirical risk minimization problem, we prove the iteration complexity of $O\left(n\log n + \sqrt{\frac{n}{\epsilon}}\right)$, which outperforms the existing results by a $\left(\log\frac{1}{\epsilon}\right)$ factor. We also prove the accelerated linear convergence rate for some special problems with nonsmooth loss, e.g., the least absolute deviation and SVM. All the above results are established for both the primal solutions and dual solutions. The topic on the complexity analysis of the primal solutions is significant not only in stochastic optimization but also in distributed optimization. We hope that the analysis in this paper could facilitate more studies on this topic.

## Acknowledgments

The authors would like to thank the reviewers of JMLR for their suggestions to improve this paper. Li is sponsored by Zhejiang Lab (grant no. 2019KB0AB02). Lin is supported by NSF China (grant no.s 61625301 and 61731018), Major Research Project of Zhejiang Lab (grant no.s 2019KB0AC01 and 2019KB0AB02) and Beijing Academy of Artificial Intelligence.

## Appendix A. Efficient Computation of the Average

We discuss the efficient computation of $\hat{\mathbf{x}}^K = \frac{\sum_{k=K_0}^K \frac{\mathbf{x}^*(\mathbf{v}^k)}{\theta_k}}{\sum_{k=K_0}^K \frac{1}{\theta_k}}$. We define two variables $\text{sum}(\mathbf{x},k)$ and $\text{sum}(\theta,k)$, update $\text{sum}(\mathbf{x},k) = \text{sum}(\mathbf{x},k-1) + \frac{\mathbf{x}^*(\mathbf{v}^k)}{\theta_k}$ and $\text{sum}(\theta,k) = \text{sum}(\theta,k-1) + \frac{1}{\theta_k}$ at each iteration of Algorithm 1. We only store $\text{sum}(\mathbf{x},k)$ and $\text{sum}(\theta,k)$ when $k = 1, \lceil\nu(1+1/\hat{n})\rceil, \lceil\nu(1+1/\hat{n})\rceil^2, \lceil\nu(1+1/\hat{n})\rceil^3, \cdots$. When the algorithm terminate at the $K$-th iteration, we let $K_0 = \lceil\nu(1+1/\hat{n})\rceil^p$ such that $\lceil\nu(1+1/\hat{n})\rceil^{p+1} \le K < \lceil\nu(1+1/\hat{n})\rceil^{p+2}$ and compute

$\hat{\mathbf{x}}^K = \frac{\mathrm{sum}(\mathbf{x},K)-\mathrm{sum}(\mathbf{x},K_0)}{\mathrm{sum}(\theta,K)-\mathrm{sum}(\theta,K_0)}$. Thus, we only need $O(t)$ computation time at each iteration and $O(t\log K)$ storage space in total, where $t$ is the dimension of $\mathbf{x}$.

## Appendix B. Proof of Lemma 1

**Proof** Let $\hat{\boldsymbol{A}} = [\boldsymbol{A}/n, \boldsymbol{B}^T]$. From the proof of Theorem 3.1 in (Lu and Johansson, 2016), we have

$$\|\mathbf{x}^*(\mathbf{u}) - \mathbf{x}^*(\mathbf{v})\| \leq \frac{1}{\mu}\|\hat{\boldsymbol{A}}\mathbf{u}_{1:n+p} - \hat{\boldsymbol{A}}\mathbf{v}_{1:n+p}\| + \frac{1}{\mu}\sum_{i=1}^{m} L_{g_i}|\mathbf{u}_{n+p+i} - \mathbf{v}_{n+p+i}|.$$

If $j \leq n+p$, then we have

$$\|\mathbf{x}^*(\mathbf{u}) - \mathbf{x}^*(\mathbf{v})\| \leq \frac{1}{\mu}\|\hat{\boldsymbol{A}}_j\mathbf{u}_j - \hat{\boldsymbol{A}}_j\mathbf{v}_j\| \leq \frac{\|\hat{\boldsymbol{A}}_j\|}{\mu}|\mathbf{u}_j - \mathbf{v}_j| = \frac{\|\hat{\boldsymbol{A}}_j\|}{\mu}\|\mathbf{u} - \mathbf{v}\|,$$

$$\|\nabla_j d(\mathbf{u}) - \nabla_j d(\mathbf{v})\|^2 = \|\hat{\boldsymbol{A}}_j^T\mathbf{x}^*(\mathbf{u}) - \hat{\boldsymbol{A}}_j^T\mathbf{x}^*(\mathbf{v})\|^2 \leq \|\hat{\boldsymbol{A}}_j\|^2\|\mathbf{x}^*(\mathbf{u}) - \mathbf{x}^*(\mathbf{v})\|^2 \leq \frac{\|\hat{\boldsymbol{A}}_j\|^4}{\mu^2}\|\mathbf{u} - \mathbf{v}\|^2.$$

If $j > n+p$, then we have

$$\|\mathbf{x}^*(\mathbf{u}) - \mathbf{x}^*(\mathbf{v})\| \leq \frac{L_{g_{j-n-p}}}{\mu}|\mathbf{u}_j - \mathbf{v}_j| = \frac{L_{g_{j-n-p}}}{\mu}\|\mathbf{u} - \mathbf{v}\|,$$

$$\|\nabla d_j(\mathbf{u}) - \nabla d_j(\mathbf{v})\|^2 = |g_{j-n-p}(\mathbf{x}^*(\mathbf{u})) - g_{j-n-p}(\mathbf{x}^*(\mathbf{v}))|^2 \leq L_{g_{j-n-p}}^2\|\mathbf{x}^*(\mathbf{u}) - \mathbf{x}^*(\mathbf{v})\|^2 \leq \frac{L_{g_{j-n-p}}^4}{\mu^2}\|\mathbf{u} - \mathbf{v}\|^2,$$

which completes the proof. ■

## Appendix C. Proof of Lemma 6

**Proof** From $\frac{1-\theta_k}{\theta_k^2} = \frac{1}{\theta_{k-1}^2}$, we can immediately prove the first two properties. We also have $\left(\frac{1}{\theta_k} - \frac{1}{2} - \frac{1}{2\hat{n}}\right)^2 \leq \frac{1}{\theta_{k-1}^2} \leq \left(\frac{1}{\theta_k} - \frac{1}{2}\right)^2$, which leads to $\frac{k}{2} + \frac{k}{2\hat{n}} + \hat{n} \geq \frac{1}{\theta_k} \geq \frac{k}{2} + \hat{n}$. So we get $\frac{1}{\theta_K^2} - \frac{1}{\theta_{K_0-1}^2} \geq \left(\frac{K^2}{4} + \hat{n}K\right)\left(1 - \frac{1}{\upsilon}\right)$ by letting $K_0 \leq \left\lfloor \frac{K}{\upsilon(1+1/\hat{n})} + 1 \right\rfloor$ for any $\upsilon > 1$. ■

## Appendix D. Proof of Lemma 10

Lemma 10 can be proved by the techniques in (Fercoq and Richtárik, 2015) with only a little changes. We give the proof for the sake of completeness.

**Proof** From the optimality condition of $\widetilde{\mathbf{z}}_i^k$, we have

$$0 \in 2\hat{n}\theta_k L_i(\widetilde{\mathbf{z}}_i^k - \mathbf{z}_i^k) + \nabla_i d(\mathbf{v}^k) + \partial h_i(\widetilde{\mathbf{z}}_i^k), \quad \forall i = 1, 2, \cdots, \hat{n}.$$

Thus, for any $\mathbf{u} \in \mathbb{D}$ and any $i = 1, \cdots, n+p+m$, we have

$$h_i(\mathbf{u}_i) - h_i(\widetilde{\mathbf{z}}_i^k) \geq 2\hat{n}\theta_k L_i\left\langle \widetilde{\mathbf{z}}_i^k - \mathbf{z}_i^k, \widetilde{\mathbf{z}}_i^k - \mathbf{u}_i \right\rangle + \left\langle \nabla_i d(\mathbf{v}^k), \widetilde{\mathbf{z}}_i^k - \mathbf{u}_i \right\rangle - \sigma_1(\mathbf{u}_i, \widetilde{\mathbf{z}}_i^k), \qquad (66)$$

where we use the convexity of $h_i(u)$ for $i > n$ and the definition of $\sigma_1(\mathbf{u}_i, \widetilde{\mathbf{z}}_i^k)$ for $i \le n$. Since $\mathbf{u}^{k+1} \in \mathbb{D}$, $\mathbf{v}^k \in \mathbb{D}$ and $\mathbf{u}_j^{k+1} = \mathbf{v}_j^k, \forall j \neq i_k$, then from (13), we have

$$d(\mathbf{u}^{k+1}) \le d(\mathbf{v}^k) + \left\langle \nabla_{i_k} d(\mathbf{v}^k), \mathbf{u}_{i_k}^{k+1} - \mathbf{v}_{i_k}^k \right\rangle + \frac{L_{i_k}}{2}\|\mathbf{u}_{i_k}^{k+1} - \mathbf{v}_{i_k}^k\|^2.$$

Using the relations of $\mathbf{u}^{k+1} - \mathbf{v}^k = \hat{n}\theta_k(\mathbf{z}^{k+1} - \mathbf{z}^k)$ and $\theta_k \mathbf{z}^k = \mathbf{v}^k - (1-\theta_k)\mathbf{u}^k$ in Algorithm 2, we have

$$\begin{aligned}
d(\mathbf{u}^{k+1}) \le& d(\mathbf{v}^k) + \left\langle \nabla_{i_k} d(\mathbf{v}^k), \hat{n}\theta_k(\widetilde{\mathbf{z}}_{i_k}^k - \mathbf{z}_{i_k}^k) \right\rangle + \frac{\hat{n}^2\theta_k^2 L_{i_k}}{2}\|\widetilde{\mathbf{z}}_{i_k}^k - \mathbf{z}_{i_k}^k\|^2 \\
=& d(\mathbf{v}^k) + \left\langle \nabla_{i_k} d(\mathbf{v}^k), \hat{n}\left[\theta_k \widetilde{\mathbf{z}}_{i_k}^k + (1-\theta_k)\mathbf{u}_{i_k}^k - \mathbf{v}_{i_k}^k\right] \right\rangle + \frac{\hat{n}^2\theta_k^2 L_{i_k}}{2}\|\widetilde{\mathbf{z}}_{i_k}^k - \mathbf{z}_{i_k}^k\|^2 \\
=& d(\mathbf{v}^k) + \left\langle \nabla_{i_k} d(\mathbf{v}^k), \hat{n}\left[\theta_k(\widetilde{\mathbf{z}}_{i_k}^k - \mathbf{v}_{i_k}^k) + (1-\theta_k)(\mathbf{u}_{i_k}^k - \mathbf{v}_{i_k}^k)\right] \right\rangle + \frac{\hat{n}^2\theta_k^2 L_{i_k}}{2}\|\widetilde{\mathbf{z}}_{i_k}^k - \mathbf{z}_{i_k}^k\|^2.
\end{aligned}$$

Taking expectation with respect to $i_k$ conditioned on $\xi_{k-1}$, we have

$$\begin{aligned}
&\mathbb{E}_{i_k|\xi_{k-1}}\left[d(\mathbf{u}^{k+1})\right] \\
\le& \frac{1}{\hat{n}}\sum_{i=1}^{\hat{n}}\left[d(\mathbf{v}^k) + \hat{n}\left\langle \nabla_i d(\mathbf{v}^k), \theta_k(\widetilde{\mathbf{z}}_i^k - \mathbf{v}_i^k) + (1-\theta_k)(\mathbf{u}_i^k - \mathbf{v}_i^k)\right\rangle + \frac{\hat{n}^2\theta_k^2 L_i}{2}\|\widetilde{\mathbf{z}}_i^k - \mathbf{z}_i^k\|^2\right] \\
=& d(\mathbf{v}^k) + (1-\theta_k)\left\langle \nabla d(\mathbf{v}^k), \mathbf{u}^k - \mathbf{v}^k\right\rangle + \sum_{i=1}^{\hat{n}}\left[\theta_k\left\langle \nabla_i d(\mathbf{v}^k), \widetilde{\mathbf{z}}_i^k - \mathbf{v}_i^k\right\rangle + \frac{\hat{n}\theta_k^2 L_i}{2}\|\widetilde{\mathbf{z}}_i^k - \mathbf{z}_i^k\|^2\right] \\
\overset{a}{\le}& (1-\theta_k)d(\mathbf{u}^k) + \theta_k d(\mathbf{v}^k) \\
&+ \sum_{i=1}^{\hat{n}}\left[\theta_k\left\langle \nabla_i d(\mathbf{v}^k), \mathbf{u}_i - \mathbf{v}_i^k\right\rangle + \theta_k\left\langle \nabla_i d(\mathbf{v}^k), \widetilde{\mathbf{z}}_i^k - \mathbf{u}_i\right\rangle + \frac{\hat{n}\theta_k^2 L_i}{2}\|\widetilde{\mathbf{z}}_i^k - \mathbf{z}_i^k\|^2\right] \\
\overset{b}{=}& (1-\theta_k)d(\mathbf{u}^k) + \theta_k d(\mathbf{u}) + \theta_k\sigma_2(\mathbf{u}, \mathbf{v}^k) + \sum_{i=1}^{\hat{n}}\left[\theta_k\left\langle \nabla_i d(\mathbf{v}^k), \widetilde{\mathbf{z}}_i^k - \mathbf{u}_i\right\rangle + \frac{\hat{n}\theta_k^2 L_i}{2}\|\widetilde{\mathbf{z}}_i^k - \mathbf{z}_i^k\|^2\right] \\
\overset{c}{\le}& (1-\theta_k)d(\mathbf{u}^k) + \theta_k d(\mathbf{u}) + \theta_k\sigma_2(\mathbf{u}, \mathbf{v}^k) + \sum_{i=1}^{\hat{n}}\left[\theta_k\left(h_i(\mathbf{u}_i) - h_i(\widetilde{\mathbf{z}}_i^k)\right) + \theta_k\sigma_1(\mathbf{u}_i, \widetilde{\mathbf{z}}_i^k)\right. \\
&\left. -2\hat{n}\theta_k^2 L_i\left\langle \widetilde{\mathbf{z}}_i^k - \mathbf{z}_i^k, \widetilde{\mathbf{z}}_i^k - \mathbf{u}_i\right\rangle + \frac{\hat{n}\theta_k^2 L_i}{2}\|\widetilde{\mathbf{z}}_i^k - \mathbf{z}_i^k\|^2\right] \\
\overset{d}{=}& (1-\theta_k)d(\mathbf{u}^k) + \theta_k d(\mathbf{u}) + \theta_k\sigma_2(\mathbf{u}, \mathbf{v}^k) + \sum_{i=1}^{\hat{n}}\left[\theta_k\left(h_i(\mathbf{u}_i) - h_i(\widetilde{\mathbf{z}}_i^k)\right) + \theta_k\sigma_1(\mathbf{u}_i, \widetilde{\mathbf{z}}_i^k)\right. \\
&\left. +\hat{n}\theta_k^2 L_i\left[\|\mathbf{z}_i^k - \mathbf{u}_i\|^2 - \|\widetilde{\mathbf{z}}_i^k - \mathbf{u}_i\|^2\right] - \frac{\hat{n}\theta_k^2 L_i}{2}\|\widetilde{\mathbf{z}}_i^k - \mathbf{z}_i^k\|^2\right] \\
\overset{e}{=}& (1-\theta_k)d(\mathbf{u}^k) + \theta_k D(\mathbf{u}) + \theta_k\sigma_2(\mathbf{u}, \mathbf{v}^k) + \theta_k\sum_{i=1}^{n}\sigma_1(\mathbf{u}_i, \widetilde{\mathbf{z}}_i^k) + (1-\theta_k)H^k \\
&-\mathbb{E}_{i_k|\xi_{k-1}}[H^{k+1}] + \hat{n}^2\theta_k^2\Big[\|\mathbf{z}^k - \mathbf{u}\|_L^2 - \mathbb{E}_{i_k|\xi_{k-1}}[\|\mathbf{z}^{k+1} - \mathbf{u}\|_L^2]\Big] - \frac{\hat{n}^2\theta_k^2}{2}\mathbb{E}_{i_k|\xi_{k-1}}[\|\mathbf{z}^{k+1} - \mathbf{z}^k\|_L^2],
\end{aligned}$$

where we use the convexity of $d$ in $\overset{a}{\leq}$, the definition of $\sigma_2(\mathbf{u},\mathbf{v}^k)$ in $\overset{b}{=}$, (66) in $\overset{c}{\leq}$, $2\langle a-b, a-c\rangle = \|a-b\|^2+\|a-c\|^2-\|b-c\|^2$ in $\overset{d}{=}$ and the following three equations in $\overset{e}{=}$, which can be obtained from Lemma 4 and Equation (45) in (Fercoq and Richtárik, 2015),

$$\begin{aligned}
\mathbb{E}_{i_k|\xi_{k-1}}[\|\mathbf{z}^{k+1}-\mathbf{u}\|_L^2] &= \frac{1}{\hat{n}}\sum_{i=1}^{\hat{n}}\left[L_i\|\widetilde{\mathbf{z}}_i^k-\mathbf{u}_i\|^2+\sum_{j\neq i}L_j\|\mathbf{z}_j^k-\mathbf{u}_j\|^2\right]\\
&= \frac{1}{\hat{n}}\sum_{i=1}^{\hat{n}}L_i\|\widetilde{\mathbf{z}}_i^k-\mathbf{u}_i\|^2+\frac{\hat{n}-1}{\hat{n}}\sum_{j=1}^{\hat{n}}L_j\|\mathbf{z}_j^k-\mathbf{u}_j\|^2,\\
\mathbb{E}_{i_k|\xi_{k-1}}[H^{k+1}] &= \sum_{t=0}^{k}\alpha_{k+1,t}h(\mathbf{z}^t)+\mathbb{E}_{i_k|\xi_{k-1}}[\hat{n}\theta_k h(\mathbf{z}^{k+1})]\\
&= \sum_{t=0}^{k}\alpha_{k+1,t}h(\mathbf{z}^t)+\frac{1}{\hat{n}}\sum_{i=1}^{\hat{n}}\hat{n}\theta_k\left(h_i(\widetilde{\mathbf{z}}_i^k)+\sum_{j\neq i}h_j(\mathbf{z}_j^k)\right)\\
&= \sum_{t=0}^{k}\alpha_{k+1,t}h(\mathbf{z}^t)+\theta_k\sum_{i=1}^{\hat{n}}h_i(\widetilde{\mathbf{z}}_i^k)+(\hat{n}-1)\theta_k\sum_{i=1}^{\hat{n}}h_i(\mathbf{z}_i^k)\\
&= \sum_{t=0}^{k-1}\alpha_{k+1,t}h(\mathbf{z}^t)+\alpha_{k+1,k}h(\mathbf{z}^k)+(\hat{n}-1)\theta_k h(\mathbf{z}^k)+\theta_k\sum_{i=1}^{\hat{n}}h_i(\widetilde{\mathbf{z}}_i^k)\\
&= (1-\theta_k)\sum_{t=0}^{k-1}\alpha_{k,t}h(\mathbf{z}^t)+(1-\theta_k)\alpha_{k,k}h(\mathbf{z}^k)+\theta_k\sum_{i=1}^{\hat{n}}h_i(\widetilde{\mathbf{z}}_i^k)\\
&= (1-\theta_k)H^k+\theta_k\sum_{i=1}^{\hat{n}}h_i(\widetilde{\mathbf{z}}_i^k),\\
\mathbb{E}_{i_k|\xi_{k-1}}[\|\mathbf{z}^{k+1}-\mathbf{z}^k\|_L^2] &= \frac{1}{\hat{n}}\sum_{i=1}^{\hat{n}}\left(L_i\|\widetilde{\mathbf{z}}_i^k-\mathbf{z}_i^k\|^2+\sum_{j\neq i}L_j\|\mathbf{z}_j^k-\mathbf{z}_j^k\|^2\right)\\
&= \frac{1}{\hat{n}}\sum_{i=1}^{\hat{n}}L_i\|\widetilde{\mathbf{z}}_i^k-\mathbf{z}_i^k\|^2. \qquad (67)
\end{aligned}$$

Rearranging the terms, we have

$$\begin{aligned}
&\mathbb{E}_{i_k|\xi_{k-1}}\left[d(\mathbf{u}^{k+1})+H^{k+1}-D(\mathbf{u})+\hat{n}^2\theta_k^2\|\mathbf{z}^{k+1}-\mathbf{u}\|_L^2+\frac{\hat{n}^2\theta_k^2}{2}\|\mathbf{z}^{k+1}-\mathbf{z}^k\|_L^2\right]\\
\leq&(1-\theta_k)\left[d(\mathbf{u}^k)+H^k-D(\mathbf{u})\right]+\hat{n}^2\theta_k^2\|\mathbf{z}^k-\mathbf{u}\|_L^2+\theta_k\left(\sum_{i=1}^{n}\sigma_1(\mathbf{u}_i,\widetilde{\mathbf{z}}_i^k)+\sigma_2(\mathbf{u},\mathbf{v}^k)\right).
\end{aligned}$$

Taking expectation with respect to $\xi_{k-1}$ on both sides, we have

$$\begin{aligned}&\mathbb{E}_{\xi_k}\left[d(\mathbf{u}^{k+1})+H^{k+1}-D(\mathbf{u})+\hat{n}^2\theta_k^2\|\mathbf{z}^{k+1}-\mathbf{u}\|_L^2+\frac{\hat{n}^2\theta_k^2}{2}\|\mathbf{z}^{k+1}-\mathbf{z}^k\|_L^2\right]\\ \leq&(1-\theta_k)\mathbb{E}_{\xi_{k-1}}\Big[d(\mathbf{u}^k)+H^k-D(\mathbf{u})\Big]+\hat{n}^2\theta_k^2\mathbb{E}_{\xi_{k-1}}[\|\mathbf{z}^k-\mathbf{u}\|_L^2]+\theta_k\mathbb{E}_{\xi_{k-1}}\left[\sum_{i=1}^n\sigma_1(\mathbf{u}_i,\widetilde{\mathbf{z}}_i^k)+\sigma_2(\mathbf{u},\mathbf{v}^k)\right].\end{aligned}$$

Since $\{\mathbf{u}^{k+1}, \mathbf{z}^{k+1}, H^{k+1}\}$ are independent on $\{i_{k+1},\cdots,i_K\}$ and $\{\mathbf{u}^k, \mathbf{z}^k, \widetilde{\mathbf{z}}^k, H^k, \sigma_1(\mathbf{u},\widetilde{\mathbf{z}}^k), \sigma_2(\mathbf{u},\mathbf{v}^k)\}$ are independent on $\{i_k,\cdots,i_K\}$, we have (28). ■

## Appendix E. Proof of Lemma 19

.

**Proof** We first prove (59). From (28) and $\|\mathbf{z}^k-\mathbf{u}^*\|_L^2 \leq \frac{4M^2}{n\mu}$, we have

$$\begin{aligned}&\mathbb{E}_{\xi_{K'}}[D(\mathbf{u}^{K'+1})]-D(\mathbf{u}^*)\leq\mathbb{E}_{\xi_{K'}}[d(\mathbf{u}^{K'+1})+H^{K'+1}]-D(\mathbf{u}^*)\\ \leq&\left(1-\frac{1}{n}\right)\left(\mathbb{E}_{\xi_{K'}}[d(\mathbf{u}^{K'})+H^{K'}]-D(\mathbf{u}^*)\right)+\frac{1}{n}\mathbb{E}_{\xi_{K'}}\left[\sum_{i=1}^n\sigma_1(\mathbf{u}_i^*,\widetilde{\mathbf{z}}_i^{K'})+\sigma_2(\mathbf{u}^*,\mathbf{v}^{K'})\right]\\ &+\mathbb{E}_{\xi_{K'}}[\|\mathbf{z}^{K'}-\mathbf{u}^*\|_L^2]-\mathbb{E}_{\xi_{K'}}[\|\mathbf{z}^{K'+1}-\mathbf{u}^*\|_L^2]\\ \leq&\left(1-\frac{1}{n}\right)^{K'+1}\left(D(\mathbf{u}^0)-D(\mathbf{u}^*)\right)+\frac{1}{n}\mathbb{E}_{\xi_{K'}}\left[\sum_{i=1}^n\sigma_1(\mathbf{u}_i^*,\widetilde{\mathbf{z}}_i^{K'})+\sigma_2(\mathbf{u}^*,\mathbf{v}^{K'})\right]\\ &+\sum_{k=1}^{K'}\left(1-\frac{1}{n}\right)^{K'-k}\frac{1}{n}\mathbb{E}_{\xi_{K'}}[\|\mathbf{z}^k-\mathbf{u}^*\|_L^2]+\left(1-\frac{1}{n}\right)^{K'}\|\mathbf{z}^0-\mathbf{u}^*\|_L^2\\ \leq&\exp\left(-\frac{K'+1}{n}\right)\left(D(\mathbf{u}^0)-D(\mathbf{u}^*)\right)+\frac{1}{n}\mathbb{E}_{\xi_{K'}}\left[\sum_{i=1}^n\sigma_1(\mathbf{u}_i^*,\widetilde{\mathbf{z}}_i^{K'})+\sigma_2(\mathbf{u}^*,\mathbf{v}^{K'})\right]+\frac{8M^2}{n\mu}\\ \leq&9\max\left\{\epsilon,\frac{M^2}{n\mu}\right\}+\frac{1}{n}\mathbb{E}_{\xi_{K'}}\left[\sum_{i=1}^n\sigma_1(\mathbf{u}_i^*,\widetilde{\mathbf{z}}_i^{K'})+\sigma_2(\mathbf{u}^*,\mathbf{v}^{K'})\right]\leq9\max\left\{\epsilon,\frac{M^2}{n\mu}\right\},\end{aligned}$$

which leads to (59) and

$$-\frac{1}{n}\mathbb{E}_{\xi_{K'}}\left[\sum_{i=1}^n\sigma_1(\mathbf{u}_i^*,\widetilde{\mathbf{z}}_i^{K'})+\sigma_2(\mathbf{u}^*,\mathbf{v}^{K'})\right]\leq9\max\left\{\epsilon,\frac{M^2}{n\mu}\right\}. \tag{68}$$

Then we prove (60). From (23) and (68), we have

$$\begin{aligned}&\mathbb{E}_{\xi_{K'}}[f(\mathbf{x}^*(\mathbf{v}^{K'})]+\frac{1}{n}\mathbb{E}_{\xi_{K'}}[\phi(n\mathbf{y}^{K'})]+D(\mathbf{u}^*)-\|\mathbf{u}^*\|_L\mathbb{E}_{\xi_{K'}}[\|\boldsymbol{A}^T\mathbf{x}^*(\mathbf{v}^{K'})/n-\mathbf{y}^{K'}\|_L^*]\\ \leq&-\frac{1}{n}\mathbb{E}_{\xi_{K'}}\left[\sum_{i=1}^n\sigma_1(\mathbf{u}_i^*,\widetilde{\mathbf{z}}_i^{K'})+\sigma_2(\mathbf{u}^*,\mathbf{v}^{K'})\right]\leq9\max\left\{\epsilon,\frac{M^2}{n\mu}\right\}.\end{aligned}$$

From (35) with $\theta_k = \frac{1}{\hat{n}}$ and $\|\widetilde{\mathbf{z}}^{K'} - \mathbf{z}^{K'}\|_L^2 \leq \frac{4M^2}{n\mu}$, we have

$$(\|\boldsymbol{A}^T\mathbf{x}^*(\mathbf{v}^{K'})/n - \mathbf{y}^{K'}\|_L^*)^2 = 4\|\widetilde{\mathbf{z}}^{K'} - \mathbf{z}^{K'}\|_L^2 \leq 16\max\left\{\epsilon, \frac{M^2}{n\mu}\right\}.$$

So from a similar induction to (42), we have

$$\begin{aligned}
&\mathbb{E}_{\xi_{K'}}[f(\mathbf{x}^*(\mathbf{v}^{K'})] + \frac{1}{n}\mathbb{E}_{\xi_{K'}}[\phi(\boldsymbol{A}^T\mathbf{x}^*(\mathbf{v}^{K'}))] - f(\mathbf{x}^*) - \frac{1}{n}\phi(\boldsymbol{A}^T\mathbf{x}^*)\\
\leq& 9\max\left\{\epsilon, \frac{M^2}{n\mu}\right\} + \left(\|\mathbf{u}^*\|_L + M\sqrt{\sum_{i=1}^n L_i}\right)\mathbb{E}_{\xi_{K'}}[\|\boldsymbol{A}^T\mathbf{x}^*(\mathbf{v}^{K'})/n - \mathbf{y}^{K'}\|_L^*]\\
\leq& 17\max\left\{\epsilon, \frac{M^2}{n\mu}\right\},
\end{aligned}$$

where we use $\|\mathbf{u}^*\|_L^2 \leq \frac{M^2}{n\mu}$ and $M\sqrt{\sum_{i=1}^n L_i} \leq \frac{M}{\sqrt{n\mu}}$. ■

## Appendix F. Analysis for the Complexity Comparisons in Section 4

The complexity of Algorithm 3 is

$$O\left(K\frac{\log\frac{1}{\epsilon}}{\log\frac{1+\frac{\kappa}{2}\left(\frac{K}{2\hat{n}}+1\right)^2}{1+(1-\theta_0)\kappa}}\right). \tag{69}$$

Case 1: $\kappa < 1$.

Letting $\frac{1+\frac{\kappa}{2}\left(\frac{K}{2\hat{n}}+1\right)^2}{1+(1-\theta_0)\kappa}$ be a constant, e.g., 2, we have $K = 2\hat{n}\left(\sqrt{\frac{2}{\kappa}+4(1-\theta_0)}-1\right)$ and $2\hat{n}\left(\sqrt{\frac{1}{\kappa}}+\sqrt{2(1-\theta_0)}-1\right) \leq K \leq 2\hat{n}\left(\sqrt{\frac{2}{\kappa}}+2\sqrt{1-\theta_0}-1\right)$. So $K = O\left(\hat{n}+\frac{\hat{n}}{\sqrt{\kappa}}\right)$ and (69) has the same order of magnitude as $\left(\hat{n}+\frac{\hat{n}}{\sqrt{\kappa}}\right)\log\frac{1}{\epsilon}$.

When $\hat{n} < K < \hat{n}+\frac{\hat{n}}{\sqrt{\kappa}}$, (69) has the same order of magnitude as $K\frac{\log\frac{1}{\epsilon}}{\log\left(1+\frac{\kappa K^2}{\hat{n}^2}\right)} = O\left(\frac{\hat{n}^2}{\kappa K}\log\frac{1}{\epsilon}\right)$ and it is smaller that $\frac{\hat{n}}{\kappa}\log\frac{1}{\epsilon}$.

When $\hat{n}+\frac{\hat{n}}{\sqrt{\kappa}} < K < \hat{n}+\frac{\hat{n}}{\kappa}$, (69) has the same order of magnitude as $K\frac{\log\frac{1}{\epsilon}}{\log\left(1+\frac{\kappa K^2}{\hat{n}^2}\right)}$ and it is also smaller than $\left(\hat{n}+\frac{\hat{n}}{\kappa}\right)\log\frac{1}{\epsilon}$.

Case 2: $\kappa > 1$.

(69) has the same order of magnitude as $K\log\frac{1}{\epsilon} = O\left(\left(\hat{n}+\frac{\hat{n}}{\kappa}\right)\log\frac{1}{\epsilon}\right)$ when $\hat{n} < K < \hat{n}+\frac{\hat{n}}{\kappa}$.

## References

Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *SIAM Journal on Optimization*, 18(1):8194–8244, 2018.

Amir Beck and Marc Teboulle. A fast dual proximal gradient algorithm for convex minimization and applications. *Operations Research Letters*, 42(1):1–6, 2014.

Dimitri Bertsekas. *Nonlinear Programming.* Athena Scientific, Belmont, Ma, 1999.

Jérôme Bolte, Trong Phong Nguyen, Juan Peypouquet, and Bruce W. Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 165(2):471–507, 2017.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2(May):499–526, 2002.

Emmanuel Candes, Justin Romberg, and Terence Tau. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8): 1207–1223, 2006.

Damek Davis and Wotao Yin. Convergence rate analysis of several splitting schemes. *Splitting Methods in Communication, Imaging, Science, and Engineering*, pages 115–163, 2017.

Olivier Devolder, Francois Glineur, and Yurii Nesterov. Double smoothing technique for large-scale linearly constrained convex optimization. *SIAM Journal on Optimization*, 22 (2):702–727, 2012.

Dmitriy Drusvyatskiy and Adrian S. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *Mathematics of Operations Research*, 43(3):919–948, 2018.

Celestine Dünner, Simone Forte, Martin Takác, and Martin Jaggi. Primal-dual rates and certificates. In *International Conference on Machine Learning (ICML)*, 2016.

Olivier Fercoq and Zheng Qu. Adaptive restart of accelerated gradient methods under local quadratic growth condition. *IMA Journal of Numerical Analysis*, 39(4):2069–2095, 2019.

Olivier Fercoq and Zheng Qu. Restarting the accelerated coordinate descent method with a rough strong convexity estimate. *Computational Optimization and Applications*, 75(1): 63–91, 2020.

Olivier Fercoq and Peter Richtárik. Accelerated, parallel, and proximal coordinate descent. *SIAM Journal on Optimization*, 25(4):1997–2023, 2015.

Bohuang Huang, Shiqian Ma, and Donald Goldfarb. Accelerated linearized Bregman method. *Journal of Scientific Computing*, 54(2-3):428–453, 2013.

Donghwan Kim and Jeffrey A. Fessler. Fast dual proximal gradient algorithms with rate $o(1/k^{1.5})$ for convex minimization. In *arxiv:1609.09441*, 2016.

Guanghui Lan and Yi Zhou. An optimal randomized incremental gradient method. *Mathematical Programming*, 171(1-2):167–215, 2018.

Yin Tat Lee and Aaron Sidford. Efficient accelerated coordinate descent methods and faster algorithms for solving linear systems. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, 2013.

Adrian S. Lewis and Michael L. Overton. Nonsmooth optimization via quasi-newton methods. *Mathematical Programming*, 141(1-2):135–163, 2013.

Guoyin Li. Global error bounds for piecewise convex polynomials. *Mathematical Programming*, 137(1-2):37–64, 2013.

Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. A universal Catalyst for first-order optimization. In *Conference on Neural Information Processing Systems (NIPS)*, 2015a.

Qihang Lin, Zhaosong Lu, and Lin Xiao. An accelerated randomized proximal coordinate gradient method and its application to regularized empirical risk minimization. *SIAM Journal on Optimization*, 25(4):2244–2273, 2015b.

Mingrui Liu and Tianbao Yang. Adaptive accelerated gradient converging method under hölderian error bound condition. In *Conference on Neural Information Processing Systems (NIPS)*, 2017.

Jie Lu and Mikael Johansson. Convergence analysis of aproximate primal solutions in dual first order methods. *SIAM Journal on Optimization*, 26(4):2430–2467, 2016.

Zhaosong Lu and Lin Xiao. On the complexity analysis of randomized block-coordinate descent methods. *Mathematical Programming*, 152(1-2):615–642, 2015.

Zhiquan Luo and Paul Tseng. On the linear convergence of descent methods for convex essentially smooth minimization. *SIAM Journal on Control and Optimization*, 30(2): 408–425, 1992.

Chenxin Ma, Rachael Tappenden, and Martin Takáč. Linear convergence of randomized feasible descent methods under the weak strong convexity assumption. *Journal of Machine Learning Research*, 230(17):1–24, 2016.

I. Necoara, Yu. Nesterov, and F. Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, 175(1-2):69–107, 2019.

Ion Necoara and Valentin Nedelcu. Rate analysis of inexact dual first-order methods application to dual decomposition. *IEEE Transactions on Automatic Control*, 59(5): 1232–1243, 2014.

Ion Necoara and Andrei Patrascu. Iteration complexity analysis of dual first-order methods for conic convex programming. *Optimization Methods and Software*, 31(3):645–678, 2016.

Yurii Nesterov. *Introductory Lectures on Convex Optimization.* Springer Science & Business Media, 2004.

Yurii Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012a.

Yurii Nesterov. How to make the gradients small. *Optima*, 88:10–11, 2012b.

B. O'Donoghue and E. Candès. Adaptive restart for accelerated gradient schemes. *Foundations of Computational Mathematics*, 15(3):715–732, 2015.

Panagiotis Patrinos and Alberto Bemporad. An accelerated dual gradient projection algorithm for embedded linear model predictive control. *IEEE Transactions on Automatic Control*, 59(1):18–33, 2013.

Peter Richtárik and Martin Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144 (1-2):1–38, 2014.

Shai Shalev-Shwartz and Tong Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *Journal of Machine Learning Research*, 14(1):567–599, 2013.

Shai Shalev-Shwartz and Tong Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. *Mathematical Programming*, 155(1):105–145, 2016.

Paul Tseng. Dual ascent methods for problems with strictly convex costs and linear constraints: A unified approach. *SIAM Journal on Optimization*, 28(1):214–242, 1990.

Paul Tseng. On accelerated proximal gradient methods for convex-concave optimization. Technical report, University of Washington, Seattle, 2008.

Powei Wang and Chih-Jen Lin. Iteration complexity of feasible descent methods for convex optimization. *Journal of Machine Learning Research*, 15(1):1523–1548, 2014.

Blake Woodworth and Nathan Srebro. Tight complexity bounds for optimizing composite objectives. In *Conference on Neural Information Processing Systems (NIPS)*, 2016.

Weihong Yang. Error bounds for convex polynomials. *SIAM Journal on Optimization*, 19 (4):1633–1647, 2009.

Yuchen Zhang and Lin Xiao. Stochastic primal-dual coordinate method for regularized empirical risk minimization. *Journal of Machine Learning Research*, 18(1):2939–2980, 2017.