Linear convergence of random dual coordinate incremental aggregated gradient methods

Hui Zhang∗ Yu-Hong Dai† Lei Guo‡

September 3, 2020

Abstract
In this paper, we consider the dual formulation of minimizing \( \sum_{i \in I} f_i(x_i) + \sum_{j \in J} g_j(A_j x) \) with the index sets \( I \) and \( J \) being large. To address the difficulties from the high dimension of the variable \( x \) (i.e., \( I \) is large) and the large number of component functions \( g_j \) (i.e., \( J \) is large), we propose a hybrid method called the random dual coordinate incremental aggregated gradient method by blending the random dual block coordinate descent method and the proximal incremental aggregated gradient method. To the best of our knowledge, no research is done to address the two difficulties simultaneously in this way. Based on a newly established descent-type lemma, we show that linear convergence of the classical proximal gradient method under error bound conditions could be kept even one uses delayed gradient information and randomly updates coordinate blocks. Three application examples are presented to demonstrate the prospect of the proposed method.

Keywords. composition convex optimization, random dual block coordinate descent, proximal incremental aggregated gradient method, error bound, linear convergence

AMS subject classifications. 90C25, 65K05.

1 Introduction
The following structured composition convex optimization has been well studied in the literature
\[
\min_{x \in \mathcal{E}_1} \quad F(x) := f(x) + g(A x),
\]
(1.1)
where \( f, g \) are proper closed convex functions, \( \mathcal{E}_1 \) a Euclidean space, \( A \) a given linear mapping. This class of problems frequently appears in many fields such as mathematical optimization, signal/imaging processing, machine learning and big data. Many efficient numerical algorithms for solving it are proposed in the literature. Among first-order methods, the proximal gradient (PG) method may be the most well-known. A standard assumption, required by the PG method, is

∗Department of Mathematics, National University of Defense Technology, Changsha, Hunan 410073, China. Email: h.zhang1984@163.com
†100190 & School of Mathematical Sciences, Chinese Academy of Sciences, Beijing 100049, China, Email: dyh@lsec.cc.ac.cn
‡Corresponding author. School of Business, East China University of Science and Technology, Shanghai 200237, China. Email: lguo@ecust.edu.cn
that \( g \) is gradient-Lipschitz-continuous and the proximal operator of \( f \) can be easily computed. However, in many cases this assumption fails to hold for the primal problem (1.1) but fortunately it holds for its Fenchel-Rockafellar dual problem

\[
\min_{y \in \mathcal{E}_2} D(y) := f^*(-A^*y) + g^*(y),
\]

where \( \mathcal{E}_2 \) is the dual space to \( \mathcal{E}_1 \), and \( f^* \) and \( g^* \) are conjugate functions of \( f \) and \( g \) respectively. That is to say, \( f^* \) may be gradient-Lipschitz-continuous and the proximal operator of \( g^* \) can be easily computed although \( g \) is not gradient-Lipschitz-continuous or the proximal operator of \( f \) cannot be easily computed. The total variation denoising problem, formulated as minimize \( \frac{1}{2} \| x - v \|^2 + \lambda \| Dx \|_1 \) with given signal \( v \), regularization parameter \( \lambda > 0 \), and matrix \( D \), is such an example. Therefore, the PG method could be directly applied to the dual problem (1.2), and then the dual-based PG methods follows; see e.g. Beck’s recent book [4] for sublinearly convergent dual PG methods and paper [21] for linearly convergent random dual block coordinate descent methods. However, these existing dual-based methods only exploit the separability of \( g \) and then may be not suitable for solving problems where \( f \) has a huge number of component functions. This motivates us to exploit the separability of \( f \) and \( g \) to design efficient methods.

In this paper, we consider problem (1.1) where both \( f \) and \( g \) have separable structure as follows

\[
f(x) := \sum_{i \in I} f_i(x_i), \quad g(Ax) := \sum_{j \in J} g_j(A_jx),
\]

where \( f_i, g_j \) are all proper closed convex functions, \( A_j \) given linear mappings, and \( I, J \) index sets. Specifically, this paper focuses on designing efficient methods for solving

\[
\min_{x \in \mathcal{E}_1} F(x) = \sum_{i \in I} f_i(x_i) + \sum_{j \in J} g_j(A_jx), \quad (1.3)
\]

with \( |I| \) and \( |J| \) being large, whose Fenchel-Rockafellar dual problem can be written as

\[
\min_{y \in \mathcal{E}_2} D(y) := \sum_{i \in I} f_i^*(-\sum_{j \in J} A_{ji}^*y_j) + \sum_{j \in J} g_j^*(y_j), \quad (1.4)
\]

where the linear mappings \( A_{ji} \) will be defined in Section 2. The cardinality \( |I| \) is the dimension of the primal variable \( x \) or the number of component functions \( f_i^* \) in the dual problem, and the cardinality \( |J| \) is the dimension of the dual variable \( y \) or the number of component functions \( g_j \) in the primal problem (1.3). When the PG method is applied to the dual problem, large \( |I| \) implies a huge gradient computation complexity and large \( |J| \) implies a large proximal point computation complexity at each step. Therefore, the case of \( |J| \) being large leads to the development of block-type PG variants such as the random dual block coordinate descent (DBCD) method [21,23]; while the case of \( |I| \) being large leads to the development of incremental-type methods such as the proximal incremental aggregated gradient (PIAG) method [15,30]. To the best of our knowledge, in the literature there is no research that deals with the case where both \( |I| \) and \( |J| \) are large. As a first try, we propose a hybrid method, called the random dual coordinate incremental aggregated gradient (abbreviated by RDCIAG) method, by blending the random DBCD method and the PIAG method. At the algorithmic level, our proposed method could be viewed as a further research for the random DBCD method and the PIAG method.
In analyzing the linear convergence rate of iterate methods, error bound conditions have been shown to be extremely useful [11, 20]. Global error bound conditions may be too stringent in practice, which will substantially restrict the applicability, and local error bound conditions are not sufficient to ensure linear convergence for non-monotone iterative methods. In this paper, we will use the bounded error bound condition to analyze the iteration complexity of the RDCIAG method, which is non-monotone. The bounded error bound condition is actually the bounded metric subregularity of the subdifferential of the dual objective function in problem (1.4). Many sufficient conditions for ensuring bounded error bound condition to hold are given in [13] and [31] (see Section 2.1). Based on the bounded error bound condition, we show that the RDCIAG method converges linearly. The proof depends on two pillars: One is the tail-vanishing lemma introduced in [1], and the other is a newly developed descent-type lemma, which delicately combines the random block coordinate descent and the PIAG descent.

The remainder of the paper is organized as follows. In Section 2, we present the basic notation and some elementary preliminaries, and existing sufficient conditions for ensuring bounded error bound condition to hold. In Section 3, we propose the RDCIAG method by blending the random DBCD method and the PIAG method. In Section 4, we study the linear convergence of the proposed method. In Section 5, we present three application examples to demonstrate the prospect of the proposed method. Finally, section 6 gives some concluding remarks.

2 Preliminaries and preliminary results

In this paper, we restrict our analysis in finite dimensional Euclidean spaces. Let $\mathcal{E}$ be a Euclidean space and $\| \cdot \|$ the associated Euclidean norm. For a closed subset $Q \subseteq \mathcal{E}$ and a point $x \in \mathcal{E}$, we define by $d(x,Q) := \inf_{y \in Q} \| x - y \|$ the distance function from $x$ to $Q$ and by $\mathcal{P}_Q(x) := \{ y \in Q : \| y - x \| = d(x,Q) \}$ the set of projection from $x$ to $Q$. The closed ball around $x \in \mathcal{E}$ with radius $r > 0$ is denoted by $\mathbb{B}_r(x,r) := \{ y \in \mathcal{E} : \| x - y \| \leq r \}$. If the central point is zero and the around space $\mathcal{E}$ is known, we abbreviate the closed ball with radius $r$ as $\mathbb{B}_r$. We let “int” and “ri” denote the interior and relative interior of a given set respectively.

Given $m$ Euclidean spaces $\{ \mathcal{E}_i : i = 1, \cdots, m \}$ with inner products $\langle \cdot, \cdot \rangle_{\mathcal{E}_i}$, their Cartesian product, defined by

$$\mathcal{E} := \bigoplus_{i=1}^{m} \mathcal{E}_i = \{ (x_1, x_2, \cdots, x_m) : x_i \in \mathcal{E}_i, i = 1, \cdots, m \},$$

is a Euclidean space equipped with the component-wise addition and the scalar-vector multiplication. The inner product in $\bigoplus_{i=1}^{m} \mathcal{E}_i$ is defined as

$$\langle (x_i)_{i=1}^{m}, (z_i)_{i=1}^{m} \rangle_{\mathcal{E}} := \sum_{i=1}^{m} \langle x_i, z_i \rangle_{\mathcal{E}_i}.$$

This paper focuses on two Euclidean spaces $\mathcal{E}_1$ and $\mathcal{E}_2$ which are defined as Cartesian products of a group of Euclidean spaces:

$$\mathcal{E}_1 := \bigoplus_{i \in I} \mathcal{E}_{1,i}, \quad \mathcal{E}_2 := \bigoplus_{j \in J} \mathcal{E}_{2,j},$$

3
where $I$ and $J$ are two finite index sets. When no confusion arises, we will omit the subscript. A linear transform $A$ is defined from $E_1$ to $E_2$ as follows:

$$A x = (A_j x)_{j \in J} = \left( \sum_{i \in I} A_{ji} x_i \right)_{j \in J},$$

where $A_{ji}$ are linear transforms from $E_{1,i} \rightarrow E_{2,j}$. The associated adjoint transform $A^* : E_2 \rightarrow E_1$ is defined by

$$A^* y = \sum_{j \in J} A^*_j y_j = \left( \sum_{j \in J} A^*_j y_j \right)_{i \in I}.$$

The norm of the linear transform $A$ is defined by

$$\|A\| := \max \{ \|Ax\|_{E_2} : \|x\|_{E_1} \leq 1 \}.$$

Let $\Gamma_0(E)$ denote the class of proper and lower semicontinuous convex functions from $E$ to $(-\infty, +\infty]$. Let $\phi_i \in \Gamma_0(E_{1,i}), i \in I$ and $\phi \in \Gamma_0(E_1)$. We say that $\phi$ is separable if it has the form $\phi(x) = \sum_{i \in I} \phi_i(x_i)$.

Let $\phi : E \rightarrow (-\infty, +\infty]$ be a proper convex function. The effective domain of $\phi$ is defined by $\text{dom } \phi := \{ x \in E : \phi(x) < +\infty \}$. The proximal mapping of $\phi$ is defined by

$$\text{prox}_{\lambda \phi}(x) := \arg \min_{y \in E} \{ \phi(y) + \frac{1}{2\lambda} \|y - x\|^2 \}.$$

The conjugate (also called Fenchel conjugate, or Legendre tansform, or Legendre-Fenchel tansform) of $\phi$ is

$$\phi^*(y) = \sup_{x \in E} \{ \langle x, y \rangle_E - \phi(x) \}.$$

The subdifferential of $\phi$ at $x$ is defined by

$$\partial \phi(x) := \{ y \in E : \phi(u) \geq \phi(x) + \langle y, u - x \rangle_E, \quad \forall u \in E \}.$$

We say that $\phi$ is subdifferentiable at $x \in E$ if $\partial \phi(x) \neq \emptyset$. The elements of $\partial \phi(x)$ are called the subgradients of $\phi$ at $x$.

A closed proper convex function $\phi$ is called essentially smooth if $\partial \phi$ is a single-valued mapping. In this case, $\partial \phi(x) = \nabla \phi(x)$ when $x \in \text{int dom } \phi$ and $\partial \phi(x) = \emptyset$ otherwise [24 Theorem 26.1]. $\phi$ is called essentially strictly convex if $\phi$ is strictly convex on every convex subset of $\{ x : \partial \phi(x) \neq \emptyset \}$. A closed proper convex function is essentially strictly convex if and only if its conjugate is essentially smooth [24 Theorem 26.3].

We say that $\phi : E \rightarrow (-\infty, +\infty)$ is gradient-Lipschitz-continuous with modulus $L > 0$ if

$$\|\nabla \phi(x) - \nabla \phi(y)\| \leq L \|x - y\|, \quad \forall x, y \in E.$$ 

We say that $\phi : E \rightarrow (-\infty, +\infty]$ is strongly convex with modulus $\mu > 0$ if for any $\alpha \in [0, 1]$,

$$\phi(\alpha x + (1 - \alpha)y) \leq \alpha \phi(x) + (1 - \alpha)\phi(y) - \frac{1}{2} \mu \alpha(1 - \alpha)\|x - y\|^2, \quad \forall x, y \in \text{dom } \phi,$$

or if (when it is differentiable)

$$\langle \nabla \phi(x) - \nabla \phi(y), x - y \rangle \geq \mu \|x - y\|^2, \quad \forall x, y \in \text{dom } \phi.$$
A multi-function $\Psi : \mathcal{E}_1 \rightrightarrows \mathcal{E}_2$ is a mapping assigning each point in $\mathcal{E}_1$ to a subset of $\mathcal{E}_2$. The graph of $\Psi$ is defined by
\[
\text{gph}(\Psi) := \{(u,v) \in \mathcal{E}_1 \times \mathcal{E}_2 : v \in \Psi(u)\}.
\]
The inverse map $\Psi^{-1} : \mathcal{E}_2 \rightrightarrows \mathcal{E}_1$ is defined by
\[
\Psi^{-1}(v) := \{u \in \mathcal{E}_1 : v \in \Psi(u)\}.
\]

The following condition plays an important role in deducing error bound conditions for structured convex optimization problems.

**Definition 2.1** ([2], Bounded linear regularity). We say that the pair sets $\{A,B\}$ have the bounded linear regularity (BLR) if for every bounded set $C$, there exists a constant $\kappa > 0$ such that
\[
d(x, \mathcal{A} \cap \mathcal{B}) \leq \kappa (d(x,A) + d(x,B)), \quad \forall x \in C.
\]

We now introduce two weakened conditions for the strong convexity.

**Definition 2.2** ([31], Bounded metric subregularity). We say that a multi-function $\Psi : \mathcal{E}_1 \rightrightarrows \mathcal{E}_2$ is bounded metrically subregular (BMS) at $(\bar{u},\bar{v}) \in \text{gph}(\Psi)$ if for any compact set $U$ with $\bar{u} \in U$, there exists $\kappa > 0$ such that
\[
d(u, \Psi^{-1}(\bar{v})) \leq \kappa d(\bar{v}, \Psi(u)), \quad \forall u \in U.
\]

It should be noted that the requirement $\bar{u} \in U$ for $\bar{u}$ in Definition 2.2 is not necessary by noting the arbitrariness of compact set $U$. As noted in [31], a polyhedral multi-function must be BMS at every point in the graph of the multi-function. Many functions are polyhedral such as the polyhedral convex function and the convex piecewise linear quadratic function. The norm function $\|\cdot\|$ also satisfies the BMS property; see [31] for more details and examples.

It is well-known that a convex differentiable function $\psi$ is strongly convex if and only if its conjugate is gradient-Lipschitz-continuous [17]. The following result shows that $\partial \psi$ is BMS at $\bar{u}$ when its inverse (or the subdifferential of the conjugate of $\psi$) is upper Lipschitz continuous at $\bar{u}$. This indicates that the BMS property is a weak version for the strong convexity.

**Proposition 2.1.** Let $\psi \in \Gamma^0(\mathcal{E})$. The multi-function $\partial \psi$ is BMS at $\bar{x}$ if $\partial \psi^{-1}$ (or $\partial \psi^*$) is upper Lipschitz continuous at $\bar{x}$ in the sense that there exist $\delta > 0$ and $\kappa > 0$ such that
\[
\partial \psi^{-1}(x) \subseteq \partial \psi^{-1}(\bar{x}) + \kappa \|x - \bar{x}\|_1, \quad \forall x \in B(\bar{x}, \delta).
\] (2.1)

For a function $\phi \in \Gamma^0(\mathcal{E})$, the BMS of $\partial \phi$ is equivalent to the firm convexity of $\phi$ as follows [13].

**Definition 2.3** ([13], Firm convexity). A closed convex function $\phi$ is firmly convex relative to a vector $v$ if the tilted function $\phi_v(x) := \phi(x) - \langle v,x \rangle$ satisfies the quadratic growth condition: for any compact set $V$ there is a constant $\sigma$ satisfying
\[
\phi_v(x) \geq \inf \phi_v + \frac{\sigma}{2} d^2(x,(\partial \phi_v)^{-1}(0)) \quad \forall x \in V.
\]

To ensure the existence of optimal solutions of dual problems and zero duality gap, we make the following standard assumptions throughout this paper.
Assumption 2.1. Let \( f \in \Gamma_0(\mathcal{E}_1) \), \( g \in \Gamma_0(\mathcal{E}_2) \). We assume that an optimal solution \( \bar{x} \) of problem (1.1) exists. Moreover, assume one of the following conditions holds:

(i) The following non-degenerate condition holds, i.e.,
\[ 0 \in \text{ri}(\text{dom } g - \text{Adom } f). \] (2.2)

(ii) If \( g \) is a polyhedral function, \( \text{dom } g \cap \text{Adom } f \neq \emptyset \) holds true.

(iii) If both \( f \) and \( g \) are polyhedral functions, \( \text{dom } g \cap \text{Adom } f \neq \emptyset \) holds true.

A sufficient and necessary condition for ensuring (2.2) to hold is \( \text{ri}(\text{dom } g) \cap \text{Adom } f \neq \emptyset \). [3, Proposition 15.24]. When \( f \) and \( g \) are separable, it is not hard to verify that condition (2.2) can be equivalently written as
\[ \left( \text{ri}(\text{dom } g) \right) \cap \left( \sum_{i \in I} A_{j_i}(\text{dom } f_i) \right) \neq \emptyset, \quad j \in J. \]

The following well-known result is fundamental to study the primal-dual gap and the relationship between primal and dual solutions. Let \( P^* \) and \( D^* \) denote the primal and dual optimal function value respectively, i.e.,
\[ P^* = \inf_{x \in \mathcal{E}_1} f(x) + g(Ax), \quad D^* = \inf_{y \in \mathcal{E}_2} f^*(-A^*y) + g^*(y). \]

Lemma 2.1 (Fenchel-Rockafellar Duality, Theorem 15.23, Fact 15.25, and Theorem 19.1 in [3]). Let \( f \in \Gamma_0(\mathcal{E}_1) \), \( g \in \Gamma_0(\mathcal{E}_2) \), and \( A \) be a linear transform from \( \mathcal{E}_1 \) to \( \mathcal{E}_2 \).

(i) If Assumption 2.1 holds, then the duality gap is zero, i.e., \( P = -D^* \), and the dual problem possesses an optimal solution.

(ii) \( P^* = f(\bar{x}) + g(A\bar{x}) \), \( D^* = f^*(-A^*\bar{y}) + g^*(\bar{y}) \), and \( P^* = -D^* \), if and only if the KKT conditions hold
\[ -A^*\bar{y} \in \partial f(\bar{x}), \quad \bar{y} \in \partial g(A\bar{x}). \] (2.3)

Under Assumption 2.1 since \( \bar{x} \) is an optimal solution of problem (1.1), by Fermat’s rule, we have \( 0 \in \partial f + g \circ A)(\bar{x}) \). Then by Assumption 2.1 and [3, Theorem 16.47], it follows that \( 0 \in \partial f(\bar{x}) + A^*\partial g(A\bar{x}) \). Let \( \bar{y} \in \partial g(A\bar{x}) \) such that \( -A^*\bar{y} \in \partial f(\bar{x}) \). The KKT conditions (2.3) follows immediately. Thus by Lemma 2.1(ii), the duality gap is zero and \( \bar{y} \) is the optimal solution of the dual problem (1.2). We denote \( \mathcal{Y} \) as the optimal solution set of problem (1.2). By Lemma 2.1(ii), we obtain
\[ \mathcal{Y} = \{ \bar{y} \in \mathcal{E}_2 : \bar{y} \in \partial g(A\bar{x}), A^*\bar{y} \in -\partial f(\bar{x}) \} = \partial g(A\bar{x}) \cap (A^*)^{-1}(-\partial f(\bar{x})). \] (2.4)

2.1 Bounded error bound conditions

Error bound condition for the optimal solution set has been shown to be extremely useful in analyzing the linear convergence of iterate methods. Deducing global or local error bounds for mathematical programming as in (1.2) has a long history and been extensively investigated; see e.g. [13,31,35] for general settings and [3,11,18,21,25,28,29] for some special cases.
This subsection focuses on the \textit{bounded} error bound conditions, which lie between global and local error bound conditions, for the optimal solution set \( \mathcal{Y} = \{ y \in \mathcal{E}_2 : 0 \in \partial D(y) \} \) of the dual problem (1.2). That is, for any compact set \( V \subseteq \mathcal{E}_2 \), there exists a constant \( \kappa > 0 \) such that
\[
d(y, \mathcal{Y}) \leq \kappa d(0, \partial D(y)), \quad \forall y \in V.
\] (2.5)

It is easy to see that \( \mathcal{Y} = \partial D^{-1}(0) \). Let \( 0 \in \partial D(\bar{y}) \) and \( \bar{y} \in V \). Then (2.5) can be written as
\[
d(\bar{y}, \partial D^{-1}(0)) \leq \kappa d(0, \partial D(y)), \quad \forall y \in V.
\]

This is exactly the BMS for the multi-function \( \partial D \) at \( (\bar{y}, 0) \).

Since the global error bound result is stronger than the bounded error bound result, we first collect the known sufficient results for ensuring the global error bound to hold in the following. The first condition (a) follows from [14, Lemma 1] (see also Lemma 2.5 in [5]) and the second condition (b) follows from the strong convexity of \( f^* \) immediately:

(a) Assume that the function \( g^* \) is the indicator function of a polyhedral set \( W \), and \( f^* \) is a strongly convex differentiable function with \( \nabla f^* \) Lipschitz continuous on \( W \).

(b) Assume that \( A \) is an identity matrix, and the function \( f^* \) is a strongly convex differentiable function with \( \nabla f^* \) Lipschitz continuous on \( \text{dom}(g^*) \).

In a recent paper [31], for ensuring the linear convergence of the randomized block coordinate proximal gradient method, Ye et al. investigated the BMS property of \( \partial D \) (using our notation of this paper) and gave some sufficient conditions for the BMS to hold by making use of the bounded metric subregular intersection theorem (see [31, Proposition 9]). The following result collects the sufficient conditions for the BMS of \( \partial D \) to hold.

\textbf{Proposition 2.2.} [31, Theorem 2, Theorem 3] Assume that \( f^* \) is strongly convex on any convex compact subset of \( \text{dom} f^* \), that \( f^* \) is continuously differentiable on \( \text{dom} f^* \) which is assumed to be open and \( \nabla f^* \) is Lipschitz continuous on any compact subset of \( \text{dom} f^* \). Then \( \partial D \) is BMS at \( (\bar{y}, 0) \) if one of the following conditions holds

(i) \( \partial g^* \) is a polyhedral multi-function,

(ii) \( \partial g^* \) is BMS at \( (\bar{y}, A\nabla f^*(-A^*\bar{y})) \) and the set \( \{ y : 0 \in -A\nabla f^*(-A^*\bar{y}) + \partial g^*(y) \} \) is a convex polyhedral set.

It is well-known that there is a close relationship between error bound condition and quadratic growth condition. Following the proof technique in [12, Theorem 4.3], it is not hard to verify that the bounded error bound condition (2.5) can imply the bounded quadratic growth condition with \( (r, \sigma) \), i.e., for any \( r > 0 \) there exists \( \sigma > 0 \) such that
\[
D(y) \geq D^* + \frac{\sigma}{2} d^2(y, \mathcal{Y}), \quad \forall y \in B_r.
\] (2.6)

By using the convexity of \( D \), it is also easy to verify that the bounded quadratic growth condition implies the bounded error bound condition (2.5).

Drusvyatskiy and Lewis showed the bounded quadratic growth conditions without assuming the commonly used strong convexity of \( f^* \) in [13, Theorem 4.3]. The main techniques they used are a
newly introduced concept called firm convexity and the bounded linear regularity for the following expression that
\[ Y = \partial g(A\bar{x}) \cap (A^*)^{-1}(-\partial f(\bar{x})). \] (2.7)

The following results essentially follow from [13, Theorem 4.3] and [13, Corollary 4.4]. For the readers’ convenience, we give a brief proof in the appendix.

**Proposition 2.3.** Let \( f \) be an essentially strictly convex function. Then the bounded quadratic growth condition with \((r,\sigma)\) holds if one of the following conditions hold

(i) \( f_i \) is upper Lipschitz continuous at \( \bar{x}_i \) for all \( i \in I \), \( g_j \) is upper Lipschitz continuous at \( A_j\bar{x} \) for all \( j \in J \), and
\[ 0 \in \text{ri}(\partial f(\bar{x}) + A^*\partial g(A\bar{x})) \], \( (2.8) \)

(ii) \( \partial f_i \) and \( \partial g_j \) are polyhedral for all \( i \in I \) and \( j \in J \).

3 The proposed algorithm

In this section, we will propose the RDCIAG method based on the certain structure of the dual objective function. Throughout this section, we always assume that \( f \) is strongly convex and hence its conjugate \( f^* \) in the dual objective function is gradient-Lipschitz-continuous, which is not necessarily strongly convex.

3.1 Warm-up: the random dual block coordinate algorithm

Recently, the authors of [4, 21] introduced (random) dual coordinate descent methods for solving the primal-dual problems \((1.3)-(1.4)\) with \( |I| = 1 \). Here, we first recall their algorithmic idea. Recall that the dual objective function is \( f^*(-A^*y) + \sum_{j\in J} g^*_j(y_j) \) with smooth function \( f^* \). Applying the proximal gradient method to the dual problem, we obtain that
\[ y_{k+1} = \arg\min_{y\in E_2} \left\{ \sum_{j\in J} g^*_j(y_j) + \langle -A\nabla f^*(-A^*y^k), y - y^k \rangle + \frac{1}{2\alpha_k}||y - y^k||^2 \right\}. \] (3.1)

In the following we reformulate \((3.1)\) as a primal-dual scheme by introducing primal variables based on the KKT conditions \((2.3)\). Let
\[ x^k := \nabla f^*(-A^*y^k) = \nabla f^*(-\sum_{j\in J} A^*_j y^k); \] (3.2)
then
\[ -A\nabla f^*(-A^*y^k) = -Ax^k = (-A_jx^k)_{j\in J}. \]

Note that the objective function in \((3.1)\) is separable with respect to \( y \). In terms of the primal variable \( x^k \), the inner product in \((3.1)\) can be simplified and then the update of dual variables can be rewritten as the following form
\[ y_{j,k+1} = \arg\min_{y_j\in E_2,j} \left\{ g^*_j(y_j) + \langle -Ax^k, y_j - y^k_j \rangle + \frac{1}{2\alpha_k}||y_j - y^k_j||^2 \right\}, \quad j \in J, \] (3.3)
or equivalently,

\[ y_{j+1}^k = \text{prox}_{\alpha_k g_j^*}(y_j^k + \alpha_k A_j x^k), \quad j \in J. \]  

(3.4)

Let \( y^0 \in \mathcal{E} \) be the initial point. By collecting the primal-dual updates (3.2)-(3.4) and introducing the idea of random coordinate updates, we present the following random primal-dual algorithm: choose uniformly random index \( j_k \in J \) and update

\[
\begin{align*}
    x^k &= \nabla f^*(- \sum_{j \in J} A_j^* y_j^k), \\
    y_{j_k+1}^k &= \text{prox}_{\alpha_k g_{j_k}^*}(y_{j_k}^k + \alpha_k A_{j_k} x^k), \\
    y_{j+1}^k &= y_j^k, \quad \forall j \neq j_k.
\end{align*}
\]

(3.5)

The convergence of the dual proximal gradient algorithm (3.3) and its randomized variant (3.5) has been extensionally studied; see e.g. \([4, 21, 23, 32]\).

### 3.2 The RDCIAG method

The algorithm (3.5) is very general to include the algorithms in \([4, 21]\) as special cases. But it only exploits the separability of \( g \). In this subsection, we consider how to further exploit the separability of \( f \) in designing algorithms. Recall that the objective function of the dual problem (1.4) is

\[
D(y) = \sum_{i \in I} f_i^*(- \sum_{j \in J} A_{ji}^* y_j) + \sum_{j \in J} g_j^*(y_j).
\]

To simplify the notation, we let \( h_i(y) := f_i^*(- \sum_{j \in J} A_{ji}^* y_j) \) and solve the following problem

\[
\text{minimize}_{y \in \mathcal{E}_2} \quad \sum_{i \in I} h_i(y) + \sum_{j \in J} g_j^*(y_j).
\]

First, let us turn our attention to the first term in the above problem. If the cardinality of the index set \( I \) is very large, which happens in big data models and huge-dimensional problems, evaluating the full gradient of \( \sum_{i \in I} h_i(y) \) at some point as done in the algorithm (3.5) is costly and even prohibitive. On the other hand, in many practical problems such as distributed optimization and network optimization, delay of gradient information update is very common. To overcome these issues, some practice-driven algorithms are developed such as stochastic gradient-type methods and incremental aggregated gradient-type methods. Compared with stochastic gradient-type algorithms, incremental aggregated gradient-type methods are much easier to implement in large-scale setting since the latter does not require independent and random sampling in each iteration. Meanwhile, incremental aggregated gradient-type methods usually outperform their stochastic counterparts because they visit each component function at each iterate and update it in a period while in stochastic settings the selection of component functions is at random and some of them may be not visited in each epoch. Besides, in practice incremental methods are broadly employed for a long history in several advanced fields such as neural networks, reinforcement learning, and optimal control \([8, 9]\). Even though, much less attention is paid to incremental aggregated gradient-type methods possibly due to the difficulty of convergence analysis. During the past few years, some remarkable progresses of determined incremental methods have been made; see e.g. \([7, 15, 16, 22, 30, 33]\). The study may be viewed as a continuum of these progresses.
By borrowing the ideas of random block coordinate descent algorithms and incremental aggregated gradient methods, we propose the following RDCIAG method: Choose uniformly a random index $j_k \in J$ and update

$$y_{jk}^{k+1} = \arg \min_{y_{jk} \in \mathbb{E}_2} \left\{ \sum_{i \in I} \nabla_{jk} h_i(y_{jk}^{k} - y_{jk}^k) + \frac{1}{2\alpha_k} \|y_{jk} - y_{jk}^k\|^2 + g_j^*(y_{jk}) \right\}, \quad (3.6)$$

and let $y_j^{k+1} = y_j^k, \forall j \neq j_k$, or equivalently

$$\begin{cases}
    x_i^k &= \nabla f_i^* \left( -\sum_{j \in J} A_{ji}^* y_j^{k-\tau_i^j} \right), \quad i \in I \\
y_{jk}^{k+1} &= \text{prox}_{\alpha_k g_j^*} \left( y_j^{k} + \alpha_k \sum_{i \in I} A_{jk} x_i^k \right) \\
y_j^{k+1} &= y_j^k, \quad \forall j \neq j_k,
\end{cases} \quad (3.7)$$

where $\tau_i^j \in [0,\tau]$ for all $k$ and $i$ are delayed indexes and $\tau \geq 0$ is the largest delayed factor. The idea of designing dual incremental aggregated methods is not new; see e.g. [7, 10]. Our novelty lies in considering the proximal incremental aggregated gradient descent method and the random dual coordinate descent method in a unified way. This unified scheme sufficiently utilizes the separability of both $f$ and $g$, and should be suitable for large-scale problems and distributed problems due to its low computational load in each iteration and its allowance for delayed gradient computation.

In order to analyze the convergence of the RDCIAG method, we introduce an equivalent expression. Define the embedding operator $U_q : \mathbb{E}_2 \to \mathbb{E}_2$ as $U_q y_q = (z_j)_{j \in J}$ with

$$z_j = \begin{cases}
    y_q, & j = q, \\
    0, & \text{otherwise},
\end{cases}$$

and a determined updated variable $\tilde{y}^{k+1}$ via

$$\tilde{y}_j^{k+1} = \arg \min_{y_j \in \mathbb{E}_2} \left\{ \sum_{i \in I} \nabla_{jk} h_i(y_{jk}^{k-\tau_i^j} - y_{jk}^k) + \frac{1}{2\alpha_k} \|y_{jk} - y_{jk}^k\|^2 + g_j^*(y_{jk}) \right\}, \quad j \in J. \quad (3.8)$$

Then the RDCIAG method can be written as

$$y^{k+1} = U_{jk} \tilde{y}_{jk}^{k+1} + \sum_{q \neq j_k} U_q y_q^k, \quad (3.9)$$

where the random index $j_k$ is chosen uniformly from the index set $J$.

### 4 Convergence analysis

The convergence analysis of the RDCIAG method is built on two pillars. The first is a tail vanishing lemma, which was proposed in [1] for analyzing the linear convergence of IAG and recently developed to study PIAG and its variants. This result states that if the constant $c$ before the sum of tails $w_j$ with $j$ from $k - k_0$ to $k$ can be controlled by (less than) the constant $b$ before $w_k$, then the linear convergence rate $a$ can be conserved.
Lemma 4.1 (Pillar one: tail vanishing lemma). Assume that the nonnegative sequences \( \{V_k\} \) and \( \{w_k\} \) satisfy
\[
V_{k+1} \leq aV_k - bw_k + c \sum_{j=k-k_0}^{k} w_j, \quad \forall k \geq 0,
\]
where \( a \in (0, 1) \), \( b \geq 0 \), \( c \geq 0 \), and \( k_0 \geq 0 \). Assume also that \( w_k = 0 \) for all \( k < 0 \), and the following condition holds:
\[
\frac{c}{1-a} \frac{1-a^{k+1}}{a^{k_0}} \leq b. \tag{4.1}
\]
Then \( V_k \leq a^k V_0 \) for all \( k \geq 0 \).

The second pillar is a generalized descent lemma. Almost all the convergence analysis of first-order methods is built on descent-type lemmas \([4][6][27][32]\). In general, descent lemmas can be easily established by using the gradient-Lipschitz-continuous property and optimality conditions. However, in our case, more issues have to be taken into account to deal with the delayed terms and the random block coordinate updates. There are two ways to seek the required descent lemma. The first way is to take the delayed terms as an error term, as done by Bertesekas in \([7][15]\), and then follow the line of analysis for random block coordinate methods in \([21][23]\). This way will leave the error term hard to copy with. The second way is more delicate. It first deduces a descent result based on the determined function values \( D(y^k) \) and \( D(\tilde{y}^{k+1}) \) conditioned on \((y^1, \ldots, y^k)\) and then establishes a descent lemma for \( D(\tilde{y}^{k+1}) \) following the techniques in \([33]\). As presented below, we succeed in the second way. All the missing proofs can be found in the appendix.

First, we show the gradient-Lipschitz-continuous property of \( h_i(y) = f_i^*(\sum_{j \in J} A_{ij}^* y_j) \).

Proposition 4.1. Assume that for all \( i \in I \), \( f_i \) is strongly convex with modulus \( \mu_i \). Then for all \( i \in I \), \( \nabla h_i \) is Lipschitz continuous with constant \( \ell_i \), where
\[
\ell_i := \sqrt{\sum_{j \in J} \|A_{ij}\|^2 \mu_i^2} \|J\| \max_{j \in J} \{\|A_{ij}^*\|^2\}.
\]

The following is a descent result in terms of the determined function values \( D(y^k) \) and \( D(\tilde{y}^{k+1}) \), conditioned on \((y^1, \ldots, y^k)\).

Proposition 4.2. Let \( \eta_1 := \frac{(|J|-1) \sum_{i \in J} \ell_i}{|J|} \) and \( \xi_k := (y^1, \ldots, y^k) \) with \( \ell_i \) defined as in Proposition 4.1. Then we have
\[
\mathbb{E}_{jk} [D(y^{k+1}) | \xi_k] \leq \frac{|J|-1}{|J|} D(y^k) + \frac{1}{|J|} D(\tilde{y}^{k+1}) + \frac{\eta_1}{|J|} \|\tilde{y}^{k+1} - y^k\|^2. \tag{4.2}
\]

We are ready to present the required descent-type lemma for the convergence analysis.

Lemma 4.2 (Pillar two: descent-type lemma). Let \( \eta_1 \) and \( \xi_k \) be defined as in Proposition 4.2 and let \( \eta_2 := \frac{\ell_{\max}[I](\tau + 1)}{2} \) with \( \ell_{\max} := \max_{i \in I} \{\ell_i\} \). Then it follows that for all \( y \in \mathcal{E}_2 \),
\[
\mathbb{E}_{jk} [D(y^{k+1}) | \xi_k] \leq \frac{|J|-1}{|J|} D(y^k) + \frac{1}{|J|} D(y) + \left( \frac{\eta_1 + \eta_2}{|J|} - \frac{1}{2\alpha |J|} \right) \|\tilde{y}^{k+1} - y^k\|^2
+ \frac{1}{2\alpha |J|} \|y - y^k\|^2 - \frac{1}{2\alpha |J|} \|y - \tilde{y}^{k+1}\|^2 + \frac{\eta_2}{|J|} \sum_{s=k-\tau}^{k-1} \|y^{s+1} - y^s\|^2. \tag{4.3}
\]
4.1 The convergence result

In this subsection, we present the main convergence result for the RDCIAG method. The following result shows that the linear convergence of the PG method under error bound conditions could be kept even one uses delayed gradient information and randomly updates coordinate blocks.

**Theorem 4.1.** Define the Lyapunov function

$$\Gamma_\alpha(y) := D(y) - D^* + \frac{1}{2\alpha} d^2(y, \mathcal{Y}).$$

Assume that the bounded quadratic growth condition with \((r, \sigma)\) in (2.6) holds such that the iterate sequence \(\{y^k\} \subseteq \mathbb{B}_r\). If the stepsize \(\alpha_k \equiv \alpha\) satisfies

$$\alpha \leq \min\left\{\frac{z_0}{\sigma}, \frac{\eta_2}{8|J|}, \frac{1}{4(\eta_1 + \eta_2)}\right\}, \quad (4.4)$$

where \(z_0\) is the solution to the equation (4.18) and \(\eta_1, \eta_2\) are defined as in Lemma 4.2, then the following linear convergence results hold

$$\mathbb{E}[\Gamma_\alpha(y^k)] \leq \left(1 - \frac{\alpha\sigma}{|J|(1 + \alpha\sigma)}\right)^k \Gamma_\alpha(y^0), \quad (4.5)$$

and

$$\mathbb{E}\left[\|x^k - \bar{x}\|^2\right] \leq 2\alpha\Gamma_\alpha(y^0) \sum_{j \in J} \sum_{q \neq j} \frac{\|A_{ji}\|^2}{\mu_i^2} \left(1 - \frac{\alpha\sigma}{|J|(1 + \alpha\sigma)}\right)^{k-\tau}. \quad (4.6)$$

**Proof.** Let \(\bar{y}^k := \mathcal{P}_\mathcal{Y}(y^k)\) be the projection of \(y^k\) onto \(\mathcal{Y}\). First, noting that \(\|\bar{y}^k - y^{k+1}\|^2 \geq d^2(y^{k+1}, \mathcal{Y})\), we have

$$\mathbb{E}_{j_k}[d^2(y^{k+1}, \mathcal{Y})|\xi_k] \leq \mathbb{E}_{j_k}[\|\bar{y}^k - y^{k+1}\|^2|\xi_k]$$

$$= \frac{1}{|J|} \sum_{j \in J} \|U_j \bar{y}^k_j + \sum_{q \neq j} U_q y^k_q - \bar{y}^k\|^2$$

$$= \frac{1}{|J|} \sum_{j \in J} \left(\|\bar{y}^k_j - \bar{y}^k\|^2 + \sum_{q \neq j} \|y^k_q - \bar{y}^k_q\|^2\right)$$

$$\leq \frac{1}{|J|} \|\bar{y}^k - \bar{y}\|^2 + \frac{|J| - 1}{|J|} d^2(y^k, \mathcal{Y}). \quad (4.7)$$

Using (4.3) with \(y = \bar{y}^k\) in Lemma 4.2 and (4.7), and noting that \(D(\bar{y}^k) = D^*\), we obtain

$$\mathbb{E}_{j_k}[\Gamma_\alpha(y^{k+1})|\xi_k] \leq \frac{|J| - 1}{|J|} \Gamma_\alpha(y^k) + \frac{1}{2\alpha|J|} d^2(y^k, \mathcal{Y})$$

$$- \left(\frac{1}{2\alpha} - \eta_1 - \eta_2\right) \frac{1}{|J|} \|\bar{y}^k + y^k\|^2 + \frac{\eta_2}{|J|} \sum_{s = k-\tau}^{k-1} \|y^{s+1} - y^s\|^2. \quad (4.8)$$

By the bounded quadratic growth condition, it follows that for all \(k\),

$$\frac{\sigma}{2} d^2(y^k, \mathcal{Y}) \leq D(y^k) - D^*. \quad (4.9)$$
Then one can verify that the following inequality holds
\[ \frac{1}{2\alpha|J|}d^2(y^k, \mathcal{N}) \leq \frac{1}{(1 + \alpha\sigma)|J|} \Gamma_\alpha(y^k). \] (4.10)

On the other hand, it is not hard to see that
\[ \mathbb{E}_{j_k}[\|y^{k+1} - y^k\|^2|\xi_k] = \frac{1}{|J|}\|y^{k+1} - y^k\|^2, \] (4.11)

which implies that
\[ \mathbb{E}\|y^{k+1} - y^k\|^2 = \mathbb{E}_{\xi_k}\mathbb{E}_{j_k}[\|y^{k+1} - y^k\|^2|\xi_k] = \frac{1}{|J|}\mathbb{E}_{\xi_k}\|y^{k+1} - y^k\|^2. \] (4.12)

From (4.10) and (4.12), taking expectation with respect to \( \xi_k \) on (4.8) implies
\[ \mathbb{E}\Gamma_\alpha(y^{k+1}) \leq \left( 1 - \frac{\alpha\sigma}{|J|(1 + \alpha\sigma)} \right) \mathbb{E}\Gamma_\alpha(y^k) \]
\[ - \left( \frac{1}{2\alpha} - \eta_1 - \eta_2 \right) \mathbb{E}\|y^{k+1} - y^k\|^2 + \frac{\eta_2}{|J|} \sum_{s=k-\tau}^{k-1} \mathbb{E}\|y^{s+1} - y^s\|^2. \] (4.13)

By the choice of \( \alpha \), it follows that \( \alpha \leq \frac{1}{4(\eta_1 + \eta_2)} \). Then it follows from (4.13) that
\[ \mathbb{E}\Gamma_\alpha(y^{k+1}) \leq \left( 1 - \frac{\alpha\sigma}{|J|(1 + \alpha\sigma)} \right) \mathbb{E}\Gamma_\alpha(y^k) \]
\[ - \frac{1}{4\alpha}\mathbb{E}\|y^{k+1} - y^k\|^2 + \frac{\eta_2}{|J|} \sum_{s=k-\tau}^{k} \mathbb{E}\|y^{s+1} - y^s\|^2. \] (4.14)

Let \( V_k := \mathbb{E}\Gamma_\alpha(y^k), w_k := \mathbb{E}\|y^{k+1} - y^k\|^2, a := 1 - \frac{\alpha\sigma}{|J|(1 + \alpha\sigma)}, b := \frac{1}{4\alpha}, c = \frac{\eta_2}{|J|} \). Then (4.14) becomes
\[ V_{k+1} \leq aV_k - bw_k + c \sum_{s=k-\tau}^{k} w_s. \] (4.15)

To employ Lemma 4.1, it remains to determine the stepsize \( \alpha \) such that the following condition holds
\[ \frac{c}{1-a} \frac{1-a^{r+1}}{a^r} \leq b. \]

Let \( \beta := 1 - \frac{1}{|J|} \). Then \( a = \frac{1+\beta\alpha\sigma}{1+\alpha\sigma} \). After some simple calculations, the condition above becomes
\[ \frac{1}{a^r} \leq 1 + \frac{\sigma(1-\beta)}{1+\alpha\sigma} \left( \frac{c}{4} - \alpha \right). \]

Since \( \alpha \leq \frac{c}{8} \) by the choice of \( \alpha \), it suffices to require that
\[ \frac{1}{a^r} = \left( \frac{1+\alpha\sigma}{1+\beta\alpha\sigma} \right)^r \leq 1 + \frac{c\sigma(1-\beta)}{8(1+\alpha\sigma)}. \] (4.16)
Denote $\gamma := \frac{c\sigma(1-\beta)}{8}$ and $z := \alpha\sigma$. Then (4.16) becomes

$$\left(\frac{1+z}{1+\beta z}\right)^\tau \leq 1 + \frac{\gamma}{1+z}.$$  (4.17)

It is not hard to find that there exists $z_0$ such that when $0 < z \leq z_0$, or equivalently $\alpha \leq \frac{z_0}{\sigma}$, (4.17) always holds. Actually, by the monotonicity of $\left(\frac{1+z}{1+\beta z}\right)^\tau$ and $1 + \frac{\gamma}{1+z}$ with respect to $z$, we can choose $z_0$ as the solution to the equation

$$\left(\frac{1+z}{1+\beta z}\right)^\tau = 1 + \frac{\gamma}{1+z}.$$  (4.18)

Collecting all the bounds on $\alpha$, we can conclude that if the stepsize $\alpha$ satisfies (4.4), then the linear convergence result (4.5) follows by Lemma 4.1.

It remains to show the convergence result (4.6). Using the expression of $x_k$ in (3.7) and the Lipschitz continuity property of $\nabla f_i^*$, and letting $\bar{y}_k := P_Y(y_{k-\tau_k}^i)$ for a fixed index $i \in I$, we derive that

$$\|x_k^i - \bar{x}_i\| = \|\nabla f_i^*(- \sum_{j \in J} A_{ji}^*y_j^{k-\tau_k^i}) - \nabla f_i^*(- \sum_{j \in J} A_{ji}^*\bar{y}_j^k)\|
\leq \frac{1}{\mu_i} \sum_{j \in J} \|A_{ji}\| d(y_j^{k-\tau_k^i}, y_j^{k-\tau_k^i})
\leq \frac{\sqrt{\sum_{j \in J} \|A_{ji}\|^2}}{\mu_i} \sqrt{\sum_{j \in J} \|\bar{y}_j^k - y_j^{k-\tau_k^i}\|^2}
\leq \frac{\sqrt{\sum_{j \in J} \|A_{ji}\|^2}}{\mu_i} d(y^{k-\tau_k^i}, Y),$$

where the second inequality follows from the Cauchy inequality. Thus

$$\|x^k - \bar{x}\|^2 = \sum_{i \in I} \|x_i^k - \bar{x}_i\|^2 \leq \sum_{i \in I} \sum_{j \in J} \frac{\|A_{ji}\|^2}{\mu_i^2} d(y^{k-\tau_k^i}, Y).$$  (4.19)

Using (4.5), we have

$$\mathbb{E}[d^2(y^{k-\tau_k^i}, Y)] \leq 2\alpha\mathbb{E}[\Gamma_\alpha(y^{k-\tau_k^i})]
\leq 2\alpha \left(1 - \frac{\alpha\sigma}{|J|(1+\alpha\sigma)^k}\right)^{k-\tau} \Gamma_\alpha(y^0),$$  (4.20)

where we use the fact that $\tau_k^i \leq \tau$ for all $k$ and $i$. Then (4.6) follows from (4.19) and (4.20) immediately. The proof is complete.

5 Application examples

In this section, we present three application examples to illustrate the prospect of our proposed method.
5.1 Best approximation problem

As the first application example, we consider the best approximation problem, i.e., finding the best approximation to a given point $v$ from the intersection of some closed convex sets $\Omega_0, \Omega_i, i \in I$. Mathematically, we solve the following minimization problem

$$
\begin{align*}
\text{minimize} & \quad \frac{1}{2}\|x - v\|^2 \\
\text{subject to} & \quad x \in \bigcap_{i=1}^{m} \Omega_i, x \in \Omega_0.
\end{align*}
$$

(5.1)

Among the iterative algorithms for solving this problem, the (random) Dykstra method is one of the first projection-based algorithms, whose linear convergence was recently established under very mild assumptions; see e.g. [21]. If each iteration point $x^k$ is required to lie in $\Omega_0$, the Dykstra-type methods will not be applicable. Note that the requirement about that $x^k$ lies in $\Omega_0$ is a very natural constraint in many practical cases; e.g. in imaging processing the pixel of denosing/deblurring images has to belong to some certain interval. Interestingly, our proposed algorithm could meet this requirement. To this end, we denote

$$
f(x) := \frac{1}{2}\|x - v\|^2 + \delta_{\Omega_0}(x)
$$

(5.2)

and $g_j(x) := \delta_{\Omega_j}(x)$ for all $j = 1, \cdots, m$. Then problem (5.1) can be reformulated as

$$
\begin{align*}
\text{minimize} & \quad f(x) + \sum_{j=1}^{m} g_j(x),
\end{align*}
$$

so that the proposed RDCIAG method can be applied. Using the expressions of $f(x)$ and $g_j(x)$, it is not hard to verify that

$$
\nabla f^*(y) = P_{\Omega_0}(v + y)
$$

and

$$
\text{prox}_{\alpha g_j^*}(y) = y - \alpha P_{\Omega_j}(\alpha^{-1} y).
$$

Thus the iterative scheme of applying the proposed RDCIAG method to solve problem (5.1) is

$$
\begin{align*}
x^{k+1} & = P_{\Omega_0}(v - \sum_{j \in J} y_j^k), \\
y_j^{k+1} & = y_j^k + \alpha_k x^k - \alpha_k P_{\Omega_j}(\alpha^{-1} y_j^k + x^k), \quad j \in J.
\end{align*}
$$

(5.3)

Theoretically, if the constraints $\Omega_i$ are polyhedral, then both $f$ and $g_j$ are piecewise linear-quadratic and their subdifferentials are polyhedral. Thus the bounded quadratic growth condition holds by Proposition 2.3 and the iterative scheme (5.3) converges linearly by Theorem 4.1. However, since $f$ is nonsmooth, the theoretical results proposed in [21] cannot be applied to analyze the algorithm above.

5.2 Sparse optimization problem

In this subsection, we point out that our proposed algorithm is suitable for solving the augmented $\ell_1$ minimization problem

$$
\begin{align*}
\text{minimize} & \quad \lambda \|x\|_1 + \frac{1}{2}\|x\|^2 \\
\text{subject to} & \quad Ax = b,
\end{align*}
$$

(5.4)
where $A \in \mathbb{R}^{m \times n}$ is a given matrix with rows $a_i$, $b \in \mathbb{R}^m$ is a given vector with entries $b_i$, and $\lambda > 0$ is a regularization parameter. In compressive sensing, the matrix $A$ represents the compressed linear measure and hence the number $m$ of measures is much less than the dimension $n$ of the signal $x$. The well-known algorithm for solving this problem is the linearized Bregman method [18], which is actually the dual gradient descent applied to the Lagrangian dual problem to (5.4). Recently, the authors of [19, 26] proposed (randomized) sparse Kaczmarz algorithms by viewing the linearized Bregman method as a Bregman projection method. If we let $f(x) := \lambda \|x\|_1 + \frac{1}{2} \|x\|^2$, then the randomized sparse Kaczmarz algorithm reads as

$$\begin{cases}
x_{i+1}^k = x_i^k - \frac{(a_i, x_i^k) - b_i}{\|a_i\|^2} \cdot a_i, \\
x_{i+1}^k = \nabla f^*(x_{i+1}^k),
\end{cases} \quad (5.5)$$

where the index $i \in \{1, \cdots, m\}$ is chosen randomly. It was observed (see [26] and its reference) that the randomized sparse Kaczmarz algorithm could be identified as a random dual coordinate descent method applied to the dual objective function

$$\frac{1}{2} \|\nabla f^*(A^T y)\|^2 - \langle b, y \rangle. \quad (5.6)$$

In these dual-type methods only the case of large $m$ is exploited via choosing the index $i \in \{1, \cdots, m\}$ randomly. However, as pointed out previously, $n$ is much larger than $m$ in compressive sensing. Therefore, our proposed algorithm could be applicable to the case of large $m$ and $n$. We now describe this case more clearly. First, we view the constraint $Ax = b$ as an intersection of the hyperplanes $\Omega_j := \{x : \langle a_j, x \rangle = b_j\}, j = 1, \cdots, m$ and write down $f(x) = \sum_{i=1}^n f_i(x_i)$ with $f_i(x_i) = \lambda |x_i| + \frac{1}{2} |x_i|^2$. From this point of view, problem (5.4) can be written as

$$\min \sum_{i=1}^n f_i(x_i) + \sum_{i=1}^m \delta_{\Omega_i}(x). \quad (5.7)$$

Its Fenchel-Rockafellar dual problem reads as

$$\min_{y_j \in \mathbb{R}^n, j=1, \cdots, m} \sum_{i=1}^n f_i^*(- \sum_{j=1}^m y_{ji}) + \sum_{j=1}^m \delta_{\Omega_j}^*(y_j),$$

whose objective function is obviously different from (5.6). Since the objective function is piecewise linear-quadratic and the constraint are polyhedral convex, it then follows that $\partial f_i$ and $\partial \delta_{\Omega_j}$ are polyhedral. Thus, the bounded quadratic growth conditions holds by Proposition 2.3. Applying the proposed algorithm to solve problem (5.7) can get an optimal solution linearly by Theorem 1.1.

### 5.3 Network Utility Maximization

As the last example, we revisit the network utility maximization problem that was discussed in Beck’s book [4]. To recover the existing algorithm, we follow the description of this problem in [4]. Consider a network that consists of a set $S = \{1, 2, \cdots, S\}$ of sources and a set $L = \{1, 2, \cdots, L\}$ of links, where a link $\ell$ has a capacity $c_\ell$. For each source $s \in S$, the set of all links used by source $s$ is denoted by $L(s) \subseteq L$. For a given link $\ell \in L$, the set of all sources that use link $\ell$ is denoted by $S(\ell) \subseteq S$. Then, $S(\ell)$ and $L(s)$ have the relation that $s \in S(\ell)$ if and only if $\ell \in L(s)$. Each
source \( s \in \mathcal{S} \) is associated with a concave utility function \( u_s \), meaning that if source \( s \) sends data at a rate \( x_s \), it gains a utility \( u_s(x_s) \). Assume that the rate \( x_s \) lies in the interval \( I_s = [0, M_s] \) with \( M_s > 0 \) being given. The network utility maximization problem is to allocate optimally the source rates, which is mathematically formulated as

\[
\text{maximize} \quad \sum_{s \in \mathcal{S}} u_s(x_s) \\
\text{subject to} \quad \sum_{s \in \mathcal{S}(\ell)} x_s \leq c_\ell, \quad \ell \in \mathcal{L}, \\
x_s \in I_s, \quad s \in \mathcal{S}.
\] (5.8)

In order to reformulate it as a special case of our framework, we let \( I_s = \mathcal{S}(\ell) \), \( J_s = \mathcal{L} \), \( f_s(x_s) = -u_s(x_s) + \delta_{I_s}(x_s) + \frac{1}{2}\|x_s\|^2 \), \( g_\ell(\cdot) = \delta_{(\infty,c_\ell]}(\cdot) \) with \( \lambda \geq 0 \) being a regularization parameter. Define

\[ A_{\ell s} := \begin{cases} 1, & s \in \mathcal{S}(\ell), \\ 0, & \text{otherwise}, \end{cases} \]

and \( A_\ell x := \sum_{s \in \mathcal{S}} A_{\ell s} x_s = \sum_{s \in \mathcal{S}(\ell)} x_s. \) Then the problem above can be written as

\[
\text{minimize} \quad \sum_{i \in I} f_i(x_i) + \sum_{j \in J} g_j(A_j x).
\] (5.9)

Applying the iterate (3.6) to solve problem (5.9) yields

\[
\begin{align*}
  x^k_i &= \nabla f^*_i(-\sum_{j \in J} A^*_{ji} y_j^{k-\tau^k_j}), \quad i \in I, \\
  y_{j_k}^{k+1} &= \text{prox}_{\lambda_k g^*_j}(y_j^k + \alpha_k \sum_{i \in I} A_{ji} x_i^k), \quad j_k \in J,
\end{align*}
\] (5.10)

where \( j_k \in J \) is uniformly chosen. Note that

\[ \nabla f^*_i(y) = \arg \min_x \{ \langle x, y \rangle - f_i(x) \}, \]

and

\[ \text{prox}_{\lambda_k g^*_j}(y) = [y - \alpha_k c_j]^+. \]

Since \( \sum_{j \in J} A^*_{ji} y_j^{k-\tau^k_j} = \sum_{j \in \mathcal{L}(i)} y_j^{k-\tau^k_j} \), using the notations involved in the original problem, we can rewrite the algorithm (5.10) as

\[
\begin{align*}
  x^k_s &= \arg \min_{x_s \in I_s} \left\{ \frac{1}{2}\|x_s\|^2 - u_s(x_s) + (\sum_{\ell \in \mathcal{L}(s)} y_\ell^{k-\tau^k_\ell}) x_s \right\}, \quad s \in \mathcal{S}, \\
  y_{\ell_k}^{k+1} &= [y_\ell^k + \alpha_k \sum_{s \in \mathcal{S}(\ell)} x_s^k - \alpha_k c_j]^+, \quad \ell \in \mathcal{L},
\end{align*}
\] (5.11)

where \( \ell \in \mathcal{L} \) is uniformly chosen at random. If \( \lambda = 0 \) and \( \tau^k_s \equiv 0 \), and updating \( y_\ell^k \) for all \( \ell \in \mathcal{L} \), then the iterate (5.11) reduces to the dual projected subgradient method presented in [4].

### 6 Conclusions

In this paper, we have proposed a hybrid algorithm by blending the well-known random dual block coordinate descent method and the recently popularized PIAG method to deal with a class of large-scale problems. Based on a newly established generalized descent lemma, the linear convergence
of the proposed algorithm is derived under the bounded error bound condition. Finally, some application examples have been illustrated by modifying or extending several existing algorithms. The proposed algorithm may be accelerated with the help of the restart or inertial accelerated techniques as employed in \cite{21}. Moreover, we would like to generalize the proposed algorithm in non-Euclidean spaces and enhance the linear convergence by using Hölderian error bound conditions. We leave them for future research.

Acknowledgements

This work is supported by the National Science Foundation of China (Nos.11971480, 11771287, 71632007, 11631013, and 11991020), the Beijing Academy of Artificial Intelligence, and the Natural Science Fund of Hunan for Excellent Youth (No.2020JJ3038).

7 Appendix: The missing proofs

The proof of Proposition \ref{21} Let $y \in \partial \psi^{-1}(x)$, which is equivalent to $x \in \partial \psi(y)$. By (2.1), we have
\[ d(y, \partial \psi^{-1}(x)) \leq \kappa \|x - \bar{x}\|, \quad \forall x \in B(\bar{x}, \delta) \cap \partial \psi(y). \]
This means that
\[ d(y, \partial \psi^{-1}(x)) \leq \kappa d(\bar{x}, B(\bar{x}, \delta) \cap \partial \psi(y)). \]
For any $y$ such that $d(\bar{x}, \partial \psi(y)) \leq \delta$, the above inequality implies
\[ d(y, \partial \psi^{-1}(x)) \leq \kappa d(\bar{x}, \partial \psi(y)). \]
Then by \cite{31, Proposition 1}, for any $r > 0$ there exists $k_r > 0$ such that
\[ d(y, \partial \psi^{-1}(x)) \leq \kappa_r d(\bar{x}, \partial \psi(y)), \quad \forall y \in B_r. \]
The proof is complete.

The proof of Proposition \ref{23} (i) Applying \cite{13, Theorem 4.3} to the dual problem (1.4) requires three conditions. The first is that $f^*$ is essentially smooth and the primal problem has a unique minimizer $\bar{x}$, which can be satisfied if $f$ is essentially strictly convex. The second is the BLR property for the pair of sets $\partial g(A\bar{x})$ and $(A^*)^{-1}(-\partial f(\bar{y}))$ to upper estimate $d(y, \mathcal{Y})$, which can be satisfied by condition (2.8).

The third is the firm convexity of the sum functions $f^*$ and $g^*$. Indeed, since $\partial f_i$ is upper Lipschitz continuous at $\bar{x}_i$, by Proposition \ref{21} $\partial f_i^*$ is BMS at $\bar{x}_i$ and then $f_i^*$ is firmly convex with respect to $\bar{x}_i$. In the same way, it follows that $g_i^*$ is firmly convex with respect to $A_j \bar{x}$. Then by \cite{13, Lemma 5}, the sum functions $f^*$ and $g^*$ are firmly convex with respect to $\bar{x}$ and $A\bar{x}$ respectively.

Choosing a compact set $\mathbb{B}_r$ to replace the sublevel set $\{y : D(y) \leq D^* + v\}$ used in the proof of \cite{13, Theorem 4.3}, the desired result follows immediately.

(ii) Noting that for a closed proper convex function $\psi$, $\partial \psi$ is polyhedral if and only if $\partial \psi^*$ is polyhedral, the desired result follows immediately from \cite{13, Corollary 4.4}.
The proof of Proposition 4.1 Recall that \( h_i(y) = f_i^*(- \sum_{j \in J} A_{ji}^* y_j) \). We have

\[
\nabla h_i(y) = (\nabla_j h_i(y))_{j \in J} = \left[ (-A_{ji}) \nabla f_i^*\left(- \sum_{j \in J} A_{ji}^* y_j\right) \right]_{j \in J}.
\]

Since \( f_i \) is strongly convex with modulus \( \mu_i \), \( \nabla f_i^* \) must be \( \frac{1}{\mu_i} \)-Lipschitz continuous. Thereby, we derive that

\[
\|\nabla h_i(y) - \nabla h_i(y')\|^2 = \sum_{j \in J} \|\nabla_j h_i(y) - \nabla_j h_i(y')\|^2 \\
= \sum_{j \in J} \|A_{ji}\| \|\nabla f_i^*(- \sum_{j \in J} A_{ji}^* y_j') - \nabla f_i^*(- \sum_{j \in J} A_{ji}^* y_j')\|^2 \\
\leq \left( \sum_{j \in J} \frac{\|A_{ji}\|^2}{\mu_i^2} \right) \|\nabla f_i^*(- \sum_{j \in J} A_{ji}^* y_j') - \nabla f_i^*(- \sum_{j \in J} A_{ji}^* y_j')\|^2 \\
\leq \left( \sum_{j \in J} \frac{\|A_{ji}\|^2}{\mu_i^2} \right) \|j\| \sum_{j \in J} \|A_{ji}^*\|^2 \|y_j - y'_j\|^2 \\
\leq \left( \sum_{j \in J} \frac{\|A_{ji}\|^2}{\mu_i^2} \right) \|j\| \max_{j \in J} \{\|A_{ji}^*\|^2\} \sum_{j \in J} \|y_j - y'_j\|^2 \\
= \left( \sum_{j \in J} \frac{\|A_{ji}\|^2}{\mu_i^2} \right) \|j\| \max_{j \in J} \{\|A_{ji}^*\|^2\} \|y - y'\|^2,
\]

where the third inequality follows from the Jensen inequality.

The proof of Proposition 4.2 We divide the proof into four steps.

Step 1. We first show that

\[
\mathbb{E}_{j_k} \left[ \sum_{j \in J} g_j^*(y_j^{k+1}) | \xi_k \right] = \frac{|J| - 1}{|J|} \sum_{j \in J} g_j^*(y_j^k) + \frac{1}{|J|} \sum_{j \in J} g_j^*(\bar{y}_j^{k+1}). \tag{7.2}
\]

Actually, denoting \( G(y) := \sum_{j \in J} g_j^*(y_j) \) and taking the conditional expectation over \( j_k \) conditioned on \( \xi_k \), we can derive that

\[
\mathbb{E}_{j_k} \left[ \sum_{j \in J} g_j^*(y_j^{k+1}) | \xi_k \right] = \mathbb{E}_{j_k} \left[ G(y^{k+1}) | \xi_k \right] \\
= \mathbb{E}_{j_k} \left[ G(U_j y_j^{k+1} + \sum_{q \neq j_k} U_q y_q^k) | \xi_k \right] \\
= \frac{1}{|J|} \sum_{j \in J} G(U_j y_j^{k+1} + \sum_{q \neq j} U_q y_q^k) \\
= \frac{1}{|J|} \sum_{j \in J} \left( g_j^*(\bar{y}_j^{k+1}) + \sum_{q \neq j} g_q^*(y_q^k) \right) \\
= \frac{|J| - 1}{|J|} \sum_{j \in J} g_j^*(y_j^k) + \frac{1}{|J|} \sum_{j \in J} g_j^*(\bar{y}_j^{k+1}). \tag{7.3}
\]

Step 2. We show that

\[
\mathbb{E}_{j_k} \left[ \sum_{i \in L} h_i(y^{k+1}) | \xi_k \right] \leq \sum_{i \in L} \left( h_i(y^k) + \frac{1}{|J|} \langle \nabla h_i(y^k), \bar{y}_j^{k+1} - y^k \rangle + \frac{\ell_i}{2 |J|} \|\bar{y}_j^{k+1} - y^k\|^2 \right). \tag{7.4}
\]
By Proposition 4.1, we derive that
\[ h_i(y^{k+1}) = h_i(U_{j_k} \tilde{y}_{j_k}^{k+1} + \sum_{q \neq j_k} U_q y_q^k) \]
\[ \leq h_i(y^k) + \langle \nabla h_i(y^k), U_{j_k} (\tilde{y}_{j_k}^{k+1} - y_{j_k}^k) \rangle + \frac{\ell_i}{2} \| U_{j_k} (\tilde{y}_{j_k}^{k+1} - y_{j_k}^k) \|^2. \]

Taking the conditional expectation over \( j_k \) conditioned on \( \xi_k \), we obtain
\[ \mathbb{E}_{j_k}[h_i(y^{k+1})|\xi_k] \leq h_i(y^k) + \frac{1}{|J|} \sum_{j \in J} \langle \nabla h_i(y^k), U_j (\tilde{y}_j^{k+1} - y_j^k) \rangle + \frac{1}{|J|} \sum_{j \in J} \frac{\ell_i}{2} \| U_j (\tilde{y}_j^{k+1} - y_j^k) \|^2 \]
\[ \leq h_i(y^k) + \frac{1}{|J|} \langle \nabla h_i(y^k), \tilde{y}^{k+1} - y^k \rangle + \frac{1}{|J|} \frac{\ell_i}{2} \| \tilde{y}^{k+1} - y^k \|^2, \]
from which \((7.4)\) follows.

**Step 3.** We show that
\[ \mathbb{E}_{j_k} \left[ \sum_{i \in I} h_i(y^{k+1})|\xi_k \right] \leq \sum_{i \in I} \left( h_i(y^{k+1}) + \frac{|J| - 1}{|J|} \langle \nabla h_i(y^{k+1}), y^{k+1} - \tilde{y}^{k+1} \rangle + \frac{|J| - 1}{|J|} \frac{\ell_i}{2} \| y^{k+1} - \tilde{y}^{k+1} \|^2 \right). \]

(7.5)

By Proposition 4.1, we derive that
\[ h_i(y^{k+1}) = h_i(U_{j_k} \tilde{y}_{j_k}^{k+1} + \sum_{q \neq j_k} U_q y_q^k) \]
\[ \leq h_i(\tilde{y}^{k+1}) + \langle \nabla h_i(\tilde{y}^{k+1}), \sum_{q \neq j_k} U_q (y_q^k - \tilde{y}_q^{k+1}) \rangle + \frac{\ell_i}{2} \| \sum_{q \neq j_k} U_q (y_q^k - \tilde{y}_q^{k+1}) \|^2. \]

Taking the conditional expectation over \( j_k \) conditioned on \( \xi_k \), we obtain
\[ \mathbb{E}_{j_k}[h_i(y^{k+1})|\xi_k] \leq h_i(\tilde{y}^{k+1}) + \frac{|J| - 1}{|J|} \langle \nabla h_i(\tilde{y}^{k+1}), y^{k+1} - \tilde{y}^{k+1} \rangle + \frac{|J| - 1}{|J|} \frac{\ell_i}{2} \| y^{k+1} - \tilde{y}^{k+1} \|^2, \]
(7.6)

from which \((7.5)\) follows.

**Step 4.** From \((7.3) < \frac{|J| - 1}{|J|} + \frac{|J| - 1}{|J|} \times \frac{1}{|J|} \times \sum_{i \in I} \mathbb{E}_{j_k}[h_i(y^{k+1})|\xi_k] \) and using the monotonicity of \( \nabla h_i \), we derive that
\[ \mathbb{E}_{j_k} \left[ \sum_{i \in I} h_i(y^{k+1})|\xi_k \right] \leq \frac{|J| - 1}{|J|} \sum_{i \in I} h_i(y^k) + \frac{1}{|J|} \sum_{i \in I} h_i(\tilde{y}^{k+1}) + \frac{1}{|J|} \frac{|J| - 1}{|J|^2} \| y^{k+1} - \tilde{y}^{k+1} \|^2 \]
\[ + \frac{|J| - 1}{|J|^2} \sum_{i \in I} \langle \nabla h_i(y^k) - \nabla h_i(\tilde{y}^{k+1}), \tilde{y}^{k+1} - y^k \rangle \]
\[ \leq \frac{|J| - 1}{|J|} \sum_{i \in I} h_i(y^k) + \frac{1}{|J|} \sum_{i \in I} h_i(\tilde{y}^{k+1}) + \frac{1}{|J|} \frac{|J| - 1}{|J|^2} \| y^{k+1} - \tilde{y}^{k+1} \|^2. \]

(7.7)

The proof is complete by summing \((7.3)\) and \((7.7)\).
The proof of Lemma 4.2 Since $h_i(x)$ is convex and gradient-Lipschitz-continuous by Proposition 4.1, it follows that

$$h_i(y^{k+1}) \leq h_i(y^{k-\tau_k}) + \langle \nabla h_i(y^{k-\tau_k}), y^{k+1} - y^{k-\tau_k} \rangle + \frac{\ell_i}{2} \| y^{k+1} - y^{k-\tau_k} \|^2$$

$$= h_i(y) + \langle \nabla h_i(y^{k-\tau_k}), y^{k+1} - y \rangle + \frac{\ell_i}{2} \| y^{k+1} - y^{k-\tau_k} \|^2. \quad (7.8)$$

Note that $\tau_k^i \leq \tau$ and $\ell_{max} = \max_{i \in I} \{ \ell_i \}$. Using the following inequality

$$\| v_k - v_j \|^2 = \| \sum_{i=1}^{k-1} (v_{i+1} - v_i) \|^2 \leq (k - j) \sum_{i=1}^{k-1} \| v_{i+1} - v_i \|^2, \quad \forall k > j \geq 1.$$

and summing (7.8) over all $i \in I$, we obtain

$$\sum_{i \in I} h_i(y^{k+1}) \leq \sum_{i \in I} h_i(y) + \langle \sum_{i \in I} \nabla h_i(y^{k-\tau_k}), y^{k+1} - y \rangle + \frac{\ell_{max}|I|(\tau + 1)}{2} \left( \| y^{k+1} - y^k \|^2 + \sum_{s=k-\tau}^{k-1} \| y^{s+1} - y^s \|^2 \right). \quad (7.9)$$

By the optimality of $y_j^{k+1}$ in (7.8) and Fermat’s rule, we have

$$0 \in \sum_{i \in I} \nabla_j h_i(y^{k-\tau_k}) + \alpha^{-1}(y^{k+1} - y_j^k) + \partial g_j^*(y_j^{k+1}). \quad (7.10)$$

This together with the subgradient inequality for the convex function $g_j^*(y_j)$ at $y_j^{k+1}$ implies that

$$g_j^*(y_j^{k+1}) \leq g_j^*(y_j^k) + \langle \sum_{i \in I} \nabla h_i(y^{k-\tau_k}), y_j - y_j^{k+1} \rangle + \alpha^{-1}(y^{k+1} - y_j^k), y_j - y_j^{k+1}). \quad (7.11)$$

Thus, we obtain

$$\sum_{j \in J} g_j^*(y_j^{k+1}) \leq \sum_{j \in J} g_j^*(y_j^k) + \langle \sum_{i \in I} \nabla h_i(y^{k-\tau_k}), y - y^{k+1} \rangle + \alpha^{-1}(y^{k+1} - y^k), y - y^{k+1}). \quad (7.12)$$

Noting $\eta_2 = \frac{\ell_{max}|I|(\tau + 1)}{2}$ and summing up (7.9) and (7.12), we obtain

$$D(y^{k+1}) \leq D(y) + \frac{1}{\alpha} \langle \nabla h_i(y^{k+1} - y^k), y^{k+1} \rangle + \eta_2(\| y^{k+1} - y^k \|^2 + \sum_{s=k-\tau}^{k-1} \| y^{s+1} - y^s \|^2). \quad (7.13)$$

Note that

$$\langle y^{k+1} - y^k, y - y^{k+1} \rangle = \frac{1}{2} \| y - y^k \|^2 - \frac{1}{2} \| y - y^{k+1} \|^2 - \frac{1}{2} \| y^{k+1} - y^k \|^2.$$  

Thereby, we have

$$D(y^{k+1}) \leq D(y) + \frac{1}{2\alpha} \| y - y^k \|^2 - \frac{1}{2\alpha} \| y - y^{k+1} \|^2 - \frac{1}{2\alpha} \| y^{k+1} - y^k \|^2$$

$$+ \eta_2 \left( \| y^{k+1} - y^k \|^2 + \sum_{s=k-\tau}^{k-1} \| y^{s+1} - y^s \|^2 \right), \quad \forall y \in \mathcal{E}_2. \quad (7.14)$$

The descent lemma follows by combining (4.2) and (7.14). This completes the proof.
References

[1] A. Aytekin, H. R. Feyzmahdavian, and M. Johansson. Analysis and implementation of an asynchronous optimization algorithm for the parameter server. *arXiv preprint arXiv:1610.05507*, 2016.

[2] H. H. Bauschke and J. M. Borwein. On the convergence of von Neumann’s alternating projection algorithm for two sets. *Set-Valued Analysis*, 1(2):185–212, 1993.

[3] H. H. Bauschke and P. L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer International Publishing, 2017.

[4] A. Beck. *First-Order Methods in Optimization*. Society for Industrial and Applied Mathematics, 2017.

[5] A. Beck and S. Shtern. Linearly convergent away-step conditional gradient for non-strongly convex functions. *Mathematical Programming*, 164(1-2):1–27, 2017.

[6] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 17(4):183–202, 2009.

[7] D. P. Bertsekas. Incremental aggregated proximal and augmented lagrangian algorithms. *arXiv preprint arXiv:1509.09257*, 2015.

[8] D. P. Bertsekas. *Nonlinear Programming:3rd Edition*. Athena Scientific press, 2016.

[9] D. P. Bertsekas. *Reinforcement Learning and Optimal Control*. Athena Scientific press, 2019.

[10] D. Blatt, A. O. Hero, and H. Gauchman. A convergent incremental gradient method with a constant step size. *SIAM Journal on Optimization*, 18(1):29–51, 2007.

[11] J. Bolte, T. P. Nguyen, J. Peypouquet, and B. W. Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 165(2):471–507, 2017.

[12] D. Drusvyatskiy and A. D. Ioffe. Quadratic growth and critical point stability of semi-algebraic functions. *Mathematical Programming*, 153(2):635–653, 2015.

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

[14] P. Gong and J. Ye. Linear convergence of variance-reduced stochastic gradient without strong convexity. *arXiv preprint arXiv:1406.1102*, 2014.

[15] M. Gurbuzbalaban, A. Ozdaglar, and P. A. Parrilo. On the convergence rate of incremental aggregated gradient algorithms. *SIAM Journal on Optimization*, 27(2):1035–1048, 2017.

[16] M. Gurbuzbalaban, A. Ozdaglar, and P. A. Parrilo. Convergence rate of incremental gradient and newton methods. *SIAM Journal on Optimization*, 29(4):2542–2565, 2019.

[17] J.-B. Hiriart-Urruty and C. Lemaréchal. *Foundations of convex analysis*. Springer-Verlag Publishers, 2004.

[18] M. J. Lai and W. T. Yin. Augmented ℓ1 and nuclear-norm models with a globally linearly convergent algorithm. *SIAM Journal on Imaging Sciences*, 6(2):1059–1091, 2013.

[19] D. A. Lorenz, F. Schöpfer, and S. Wenger. The linearized bregman method via split feasibility problems: Analysis and generalizations. *SIAM Journal on Imaging Sciences*, 7(2):1237–1262.

[20] Z.-Q. Luo and P. Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46(1):157–178, 1993.

[21] I. Necoara and O. Fercoq. Linear convergence of dual coordinate descent on non-polyhedral convex problems. 2019.
[22] W. Peng, H. Zhang, and X. Zhang. Nonconvex proximal incremental aggregated gradient method with linear convergence. *Journal of Optimization Theory and Applications*, 183(1):230–245, 2019.

[23] P. Richtárik and M. Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144:1–38, 2014.

[24] R. T. Rockafellar. *Convex analysis*. Princeton university press, 1970.

[25] F. Schöpfer. Linear convergence of descent methods for the unconstrained minimization of restricted strongly convex functions. *SIAM Journal on Optimization*, 26(3):1883–1911, 2016.

[26] F. Schöpfer and D. A. Lorenz. Linear convergence of the randomized sparse kaczmarz method. *Mathematical Programming*, 1703:509–536, 2019.

[27] M. Teboulle. A simplified view of first order methods for optimization. *Mathematical Programming*, 170:67–96, 2018.

[28] P. Tseng. Approximation accuracy, gradient methods, and error bound for structured convex optimization. *Mathematical Programming*, 125(2):263–295, 2010.

[29] P. Tseng and S. Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117(1-2):387–423, 2009.

[30] N. D. Vanli, M. Gurbuzbalaban, and A. Ozdaglar. Global convergence rate of proximal incremental aggregated gradient methods. *SIAM Journal on Optimization*, 28(2):1282–1300, 2018.

[31] J. J. Ye, X. Yuan, S. Zeng, and J. Zhang. Variational analysis perspective on linear convergence of some first order methods for nonsmooth convex optimization problems. 2018.

[32] H. Zhang. New analysis of linear convergence of gradient-type methods via unifying error bound conditions. *Mathematical Programming*, 180(1):371–416, 2020.

[33] H. Zhang, Y.-H. Dai, L. Guo, and W. Peng. Proximal-like incremental aggregated gradient method with linear convergence under bregman distance growth conditions. *Mathematics of Operations Research*, 2020.

[34] X. Y. Zheng and K. F. Ng. Metric subregularity of piecewise linear multifunctions and applications to piecewise linear multiobjective optimization. *SIAM Journal on Optimization*, 24(1):154–174, 2014.

[35] Z. Zhou and A. M.-C. So. A unified approach to error bounds for structured convex optimization problems. *Mathematical Programming*, 165(2):689–728, 2017.