ON THE CONVERGENCE OF RANDOMIZED BREGMAN COORDINATE DESCENT FOR NON-LIPSCHITZ COMPOSITE PROBLEMS

Tianxiang Gao[‡], *Songtao Lu*[†], *Jia Liu*[*], *and Chris Chu*[‡]

[‡]Department of Electrical and Computer Engineering, Iowa State University, Ames, IA, 50010, USA
[†]IBM Research AI, IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598, USA
[*]Department of Computer Science, Iowa State University, Ames, IA, 50010, USA

ABSTRACT

We propose a new *randomized Bregman (block) coordinate descent* (RBCD) method for minimizing a composite problem, where the objective function could be either convex or nonconvex, and the smooth part are freed from the global Lipschitz-continuous (partial) gradient assumption. Under the notion of relative smoothness based on the Bregman distance, we prove that every limit point of the generated sequence is a stationary point. Further, we show that the iteration complexity of the proposed method is $\mathcal{O}(n\varepsilon^{-2})$ to achieve ϵ-stationary point, where n is the number of blocks of coordinates. If the objective is assumed to be convex, the iteration complexity is improved to $\mathcal{O}(n\epsilon^{-1})$. If, in addition, the objective is strongly convex (relative to the reference function), the global linear convergence rate is recovered. We also present the accelerated version of the RBCD method, which attains an $\mathcal{O}(n\varepsilon^{-1/\gamma})$ iteration complexity for the convex case, where the scalar $\gamma \in [1, 2]$ is determined by the *generalized translation variant* of the Bregman distance. Convergence analysis without assuming the global Lipschitz-continuous (partial) gradient sets our results apart from the existing works in the composite problems.

Index Terms— Bregman distance, Non-Lipschitz, Coordinate Descent, Convex and Nonconvex Optimization

1. INTRODUCTION

In this paper, we consider an optimization problem as follows

$$\underset{\mathbf{x}}{\text{minimize}} \;\; F(\mathbf{x}) \equiv f(\mathbf{x}) + r(\mathbf{x}), \tag{1}$$

where r has block separable structure. More specifically, we have

$$r(\mathbf{x}) = \sum_{i=1}^{n} r_i(\mathbf{x}_i), \tag{2}$$

where $\mathbf{x}_i$ denotes a subvector of $\mathbf{x}$ with dimension N_i such that $\sum_{i=1}^{n} N_i = N$, and each r_i is a (possibly nonsmooth) convex function.

Due to the block separable structure, Problem (1) can be solved by *(block) coordinate descent* (CD) methods or their variants, especially in the large scale optimization problems. Roughly speaking, these methods are based on the strategy of selecting one coordinate/block of variables at each iteration using some index selection procedure (*e.g.*, cyclic, greedy, randomized). This often dramatically reduces the computational complexity of the algorithms per iteration as well as memory storage, making these methods simple and salable. See for instance [1, 2, 3, 4] and references therein, as well as the recent comprehensive review paper [5].

A widely used assumption in showing the convergence of CD methods in the literature is that the (partial) gradient of f is globally Lipschitz-continuous. However, this could be a restrictive assumption violated in diverse applications in practice, such as matrix factorization [6, 7], tensor decomposition [8], matrix/tensor completion [9], Poisson likelihood models [10], etc. Although this assumption may be relaxed by adopting conventional line search methods, the efficiency and computational complexity of the first-order method are unavoidably distorted, especially when the size of the problem is large. In fact, this longstanding issue also appears in the classical *proximal gradient descent* (PGD) method. Fortunately, this issue is solved in [11, 12, 13]. They develop a new framework called *Bregman proximal gradient* (BPG) method that adapts the geometry of f by the Bregman distance. In such a way, the decrease of the objective value can be still quantified. As a result, they are able to characterize the convergence behavior of BPG for minimizing convex composite problems without assuming globally Lipschitz-continuous gradient of the objective function. Further, this framework has been extended to the case of nonconvex optimization in [14].

Despite the crucial issue is solved in PGD-type methods, there are only few results on CD-type methods. A *cyclic Bregman coordinate descent* (CBCD) method has been proposed in [15, 16, 17], but no rates are given. In [18], the authors provide a convergence rate results using randomized (block) coordinate selection strategy in the special case where F is smooth convex and $r \equiv 0$. To the best of our knowledge, how to deal with this crucial issue is still an open problem, when using CD methods to solve a nonsmooth and convex/nonconvex Problem (1). In this paper, we bridge this gap by proposing a *randomized Bregman (block) coordinate descent* (RBCD) method. In each iteration, a (block) coordinate is selected uniformly at random, and updated using the *Bregman proximal mapping*, while the rest of blocks are fixed. The main contributions are summarized as follows.

1. We propose a randomized Bregman (block) coordinate descent (RBCD) method to solve the composite problem where the smooth part does not have the global Lipschitz-continuous (partial) gradient property.
2. By adapting the relative smoothness framework, we establish a rigorous convergence rate analysis of the RBCD method, showing that the convergence rate to an ε-stationary point is $\mathcal{O}(n\varepsilon^{-2})$ if F is nonconvex, where n is the number of iterations.
3. If F is convex, RBCD achieves the global sublinear convergence rate of $\mathcal{O}(n\varepsilon^{-1})$. The global linear convergence rate is obtained if f is (relative) strongly convex.

This work is supported in part by NSF grants CAREER CNS-2110259, CIF-2110252, ECCS-1818791, CNS-2102233, ONR grant ONR N00014-17-1-2417, and a Google Faculty Research Award.

4. The RBCD method can also be accelerated in the relative smoothness setting. The iteration complexity of $\mathcal{O}(n\varepsilon^{-1/\gamma})$ can be obtained through the notion of *generalized translation variant* (explained in the latter section) of the Bregman distance.

2. PRELIMINARIES

Notation. Throughout this paper, we use bold upper case letters to denote matrices (*e.g.*, $\mathbf{X}$), bold lower case letters to denote vectors (*e.g.*, $\mathbf{x}$), and Calligraphic letters (*e.g.*, $\mathcal{X}$) are used to denote sets. For a function f, $\nabla f(\mathbf{x})$ denotes its the gradient, and $\nabla_i f(\mathbf{x})$ is the i-th partial gradient. Let $f_i(\mathbf{x}_i)$ be the function with respect to the i-th block, while the rest of blocks are fixed. Then $\nabla_i f(\mathbf{x}) = \nabla f_i(\mathbf{x}_i)$. If f is not differentiable, ∂f denotes the subdifferential of f.

Given a convex function ϕ, the *Bregman proximal mapping* of ϕ at a point $\mathbf{x}$ is defined as $T_\phi(\mathbf{x}) = \text{argmin}_{\mathbf{u}} \phi(\mathbf{u}) + D_h(\mathbf{u}, \mathbf{x})$, where $D_h(\mathbf{u}, \mathbf{x}) = h(\mathbf{u}) - h(\mathbf{x}) - \langle \nabla h(\mathbf{x}), \mathbf{u} - \mathbf{x} \rangle$ is the *Bregman distance* with the reference convex function h. This mapping is well-defined since the functions ϕ and h are convex. The convexity of h also implies $D_h(\mathbf{x}, \mathbf{y}) \geq 0, \forall \mathbf{x}, \mathbf{y}$. If, in addition, h is strictly convex, $D_h(\mathbf{x}, \mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$. In the rest of this paper, we assume h is strictly convex. Note that $D_h(\mathbf{x}, \mathbf{y})$ is not symmetric in general. Therefore, we use *symmetric coefficient* $\theta(h) = \inf_{\mathbf{x} \neq \mathbf{y}} \{D_h(\mathbf{x}, \mathbf{y}) / D_h(\mathbf{y}, \mathbf{x})\}$ to measure the symmetry. When $\phi = \delta_{\mathbf{x}}$, the Bregman proximal mapping reduces to the *Bregman projection* $P^h_{\mathcal{X}}(\mathbf{x}) = \text{argmin}\{D_h(\mathbf{u}, \mathbf{x}) : \mathbf{u} \in \mathcal{X}\}$.

Our goal is to solve the optimization (1) and the following reasonable assumptions are made throughout this paper.

Assumption 1.

(i) *f is continuously differentiable.*

(ii) *r is convex, block separable, proper and loser semi-continuous.*

(iii) $F^* = \inf_{\mathbf{x}} F(\mathbf{x}) > -\infty$.

An estimate $\mathbf{x}$ is said to be a *stationary point* of F if it satisfies

$$0 \in \partial F \equiv \nabla f(\mathbf{x}) + \partial r(\mathbf{x}). \tag{3}$$

Due to the page limit, all the proofs in details are omitted in this paper and will be included in the journal version [19].

3. RANDOMIZED BREGMAN COORDINATE DESCENT

In this section, we introduce the *randomized Bregman (block) coordinate descent* (RBCD) method for solving problem (1). Given the current estimate $\mathbf{x}$, the i-th block of coordinates is selected uniformly at random, then the new estimate $\mathbf{x}^+$ is updated as follows

$$\mathbf{x}_i^+ = T_i(\mathbf{x}), \quad \text{and} \quad \mathbf{x}_j^+ = \mathbf{x}_j, \forall j \neq i, \tag{4}$$

where, for some stepsize α, the vector $T_i(\mathbf{x})$ is defined as

$$T_i(\mathbf{x}) = \text{argmin}_{\mathbf{u}_i} \langle \nabla_i f(\mathbf{x}), \mathbf{u}_i - \mathbf{x}_i \rangle + \tfrac{1}{\alpha} D_h(\mathbf{u}_i, \mathbf{x}_i) + r_i(\mathbf{u}_i). \tag{5}$$

Note that we drop the index i in D_{h_i} to simplify the notation. The algorithm is summarized in Algorithm 1. Here the stepsize α can be determined by a conventional line search method and the global convergence results can be established. However, line search methods are usually expensive since this subroutine requires to evaluate the objective function multiple times to ensure the sufficient descent in the objective value. To establish convergence results for a CD-type method with a constant stepsize, the common assumption is that $\nabla f(\mathbf{x})$ (or $\nabla_i f(\mathbf{x})$) is globally Lipschitz-continuous [1, 2, 20]. However, this assumption may be restrict to some modern optimization problems. See for instances [6, 8, 9, 10] and reference therein. In the following section, we review the notion of *relative smoothness* introduced in [11, 12, 13]. This notion allows us to establish the convergence results for RBCD method without the assumption of global Lipschitz-continuous gradient.

Algorithm 1: RBCD Method.

```
Choose $\mathbf{x}^0$ and stepsize $\alpha$.
for $k = 1, 2 \cdots$ do
    Choose $i_k \in \{1, 2, \cdots, n\}$ uniformaly at random
    Compute $T_{i_k}(\mathbf{x}^k)$ from (5)
    Update $\mathbf{x}^{k+1}$ by (4)
end
```

4. CONVERGENCE ANALYSES

Definition 1 (Relative Smoothness). *[13, Definition 1.1] A pair of functions (g, h) is said to be relatively smooth if h is convex and there exists a scalar $L > 0$ such that $Lh - g$ is convex.*

The relative smoothness nicely translates the Bregman distance to produce a non-Lipschitz descent lemma [13, 12].

Lemma 1. *[12, Lemma 1] The pair of functions (g, h) is relatively smooth if and only if for all $\mathbf{x}$ and $\mathbf{y}$, it holds that*

$$g(\mathbf{y}) - g(\mathbf{x}) - \langle \nabla g(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \leq L D_h(\mathbf{y}, \mathbf{x}). \tag{6}$$

Note that if $h = \frac{1}{2}\| \cdot \|^2$, the classical descent lemma is recovered, *i.e.*, $g(\mathbf{y}) - g(\mathbf{x}) - \langle \nabla g(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \leq \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2$. To use Lemma 1, we additionally make the following assumptions.

Assumption 2. *The functions (f_i, h_i) are relatively smooth with constants $L_i > 0, \forall i$ and let $L = \max_i\{L_i\}$.*

With the relative smoothness between (f_i, h_i), the following result shows the basic descent property of the proposed method.

Lemma 2. *For any $\mathbf{x}$, and any $i \in \{1, 2, \cdots, n\}$, let $\mathbf{x}^+$ to be defined as in E.q. (4) and stepsize $\alpha = \frac{1+\theta_i}{2L_i}$. Then we have*

$$F(\mathbf{x}^+) \leq F(\mathbf{x}) - L_i D_h(T_i(\mathbf{x}), \mathbf{x}_i). \tag{7}$$

where $\theta_i = \theta(h_i)$. In other words, the sufficient descent in the objective value of F is guaranteed.

Since only one block is selected and updated per iteration, the quantity $D_h(\mathbf{x}^+, \mathbf{x})$ introduced in [13, 12] cannot be used to measure the optimality of the RBCD method. Given an estimate $\mathbf{x}$, we introduce the reference function H and the corresponding Bregman mapping as follows:

$$H(\mathbf{x}) = \sum_{i=1}^{n} L_i h_i(\mathbf{x}_i), \tag{8}$$

$$T(\mathbf{x}) = \text{argmin}_{\mathbf{u}} \langle \nabla f(\mathbf{x}), \mathbf{u} - \mathbf{x} \rangle + D_H(\mathbf{u}, \mathbf{x}) + r(\mathbf{u}). \tag{9}$$

Based on this mapping, the following result shows that the quantity $D_H(T(\mathbf{x}), \mathbf{x})$ can be used to measure the optimality of F.

Lemma 3. *A vector $\mathbf{x}$ is a stationary point of F if and only if $D_H(T(\mathbf{x}), \mathbf{x}) = 0$.*

Clearly, when F is convex, then the current estimate $\mathbf{x}$ is a global minimum if $D_H(T(\mathbf{x}), \mathbf{x}) = 0$.

4.1. Convex case

We use $\mathbb{E}_i$ (or $\mathbb{E}_{i_k}$) to denote the expectation with respect to a single random variable i (or i_k). We use $\mathbb{E}$ to denote the expectation with respect to all random variables $\{i_0, i_1, \cdots\}$.

For simplicity, we here use the *relative strongly convexity* introduced in [13], which is similar to the relative smoothness.

Definition 2. *A function g is μ-strongly convex relative to h if for any $\mathbf{x}$ and $\mathbf{y}$, there exists a scalar $\mu \geq 0$ such that*

$$g(\mathbf{y}) \geq g(\mathbf{x}) + \langle \nabla g(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \mu D_h(\mathbf{y}, \mathbf{x}). \tag{10}$$

Note that if $\mu = 0$, the classical convexity for a smooth function g is recovered. Moreover, when $h = \frac{1}{n} \| \cdot \|$, the classical strongly convexity is recovered. In the rest of this subsection, we assume f is strongly convex relative to H.

Assumption 3. *f is μ-strongly convex relative to H, i.e., there exists a scalar $\mu \geq 0$ such that for every $\mathbf{y}$ and $\mathbf{x}$*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \mu D_H(\mathbf{y}, \mathbf{x}). \tag{11}$$

Since r is assumed to be convex, the function F is also μ-strongly convex relative to H, *i.e.*,

$$F(\mathbf{y}) \geq F(\mathbf{x}) + \langle \mathbf{v}, \mathbf{y} - \mathbf{x} \rangle + \mu D_H(\mathbf{y}, \mathbf{x}), \tag{12}$$

for some $\mathbf{v} \in \partial F(\mathbf{x})$. Moreover, by Assumption 2, we have

$$f(T_i(\mathbf{x})) \leq f(\mathbf{x}) + \langle \nabla_i f(\mathbf{x}), T_i(\mathbf{x}) - \mathbf{x}_i \rangle + L_i D_h(T_i(\mathbf{x}), \mathbf{x}_i). \tag{13}$$

Substituting $\mathbf{y} = T_i(\mathbf{x})$ in E.q.(11) and combing it with the inequality (13), we immediately obtain that $\mu \leq 1$.

The following lemma provides the key inequalities used to prove the convergence results of the RBCD method.

Lemma 4. *For any vector $\mathbf{x}$, let $\mathbf{x}^+$ to be defined as in E.q.(4) by picking up $i \in \{1, 2, \cdots, n\}$ uniformly at random. Set stepsize $\alpha = \frac{1+\theta_i}{2L_i}$. For any vector $\mathbf{u}$, the expectation of $F(\mathbf{x}^+)$ satisfies*

$$\mathbb{E}_i[F(\mathbf{x}^+)] \leq \frac{1}{n}\Big[(n-1)F(\mathbf{x}) + F(\mathbf{u}) + (1-\mu)D_H(\mathbf{u}, \mathbf{x}) - D_H(\mathbf{u}, T(\mathbf{x}))\Big], \tag{14}$$

and the expectation of $D_H(\mathbf{x}^+, \mathbf{x})$ satisfies

$$\mathbb{E}_i[D_H(\mathbf{u}, \mathbf{x}^+)] = \frac{n-1}{n} D_H(\mathbf{u}, \mathbf{x}) + \frac{1}{n} D_H(\mathbf{u}, T(\mathbf{x})). \tag{15}$$

Theorem 1. *Let $\{\mathbf{x}^k\}$ be the sequence generated by Algorithm 1. Then for any $k \geq 0$, the iterates $\mathbf{x}^k$ satisfies*

$$\mathbb{E}[F(\mathbf{x}^k) - F(\mathbf{x}^*)] \leq \tfrac{n}{n+k}\Big(F(\mathbf{x}^*) - F(\mathbf{x}^0) + D_H(\mathbf{x}^*, \mathbf{x}^0)\Big). \tag{16}$$

Further, if f is μ-strongly convex relative to H, i.e., $\mu > 0$, then

$$\mathbb{E}[F(\mathbf{x}^k) - F(\mathbf{x}^*)] \leq \Big(1 - \tfrac{(1+\theta)\mu}{n(1+\theta\mu)}\Big)^k \Big(F(\mathbf{x}^0) - F(\mathbf{x}^*) + D_H(\mathbf{x}^*, \mathbf{x}^0)\Big), \tag{17}$$

where $\theta = \min_i \{\theta_i\}$.

Therefore, if F is convex, the sequence $\{\mathbf{x}^k\}$ converges to a global minimum at the rate of $\mathcal{O}(\frac{n}{n+k})$. Further, the classical linear convergence rate is obtained if f is strongly convex (relative to H).

4.2. Nonconvex case

If F is nonconvex, the following result shows the descent property of the proposed method in terms of the optimality gap $D_H(T(\mathbf{x}), \mathbf{x})$.

Lemma 5. *For any $\mathbf{x}$, let $\mathbf{x}^+$ to be defined as in E.q.(4) by picking up the index i uniformly at random. Let $\alpha = \frac{1+\theta_i}{2L_i}$. Then the following inequality holds:*

$$\mathbb{E}_i[F(\mathbf{x}^+)] \leq F(\mathbf{x}) - \frac{1}{n} D_H(T(\mathbf{x}), \mathbf{x}). \tag{18}$$

Theorem 2. *Let $\{\mathbf{x}^k\}$ to be the sequence generated by Algorithm 1. Let stepsize $\alpha^k = \frac{1+\theta_{i_k}}{2L_{i_k}}$, then*

(i) The sequence $\{F(\mathbf{x}^k)\}$ is non-increasing.

(ii) $\sum_{l=0}^{\infty} \mathbb{E}[D_H(T(\mathbf{x}^l), \mathbf{x}^l)] < \infty$, and hence the sequence $\{\mathbb{E}[D_H(T(\mathbf{x}^l), \mathbf{x}^l)]\}$ converges to zero.

(iii) $\forall k \geq 0$, we obtain

$$\min_{0 \leq l \leq k} \mathbb{E}\left[D_H(T(\mathbf{x}^l), \mathbf{x}^l)\right] \leq \tfrac{n}{k+1}(F(\mathbf{x}^0) - F^*), \tag{19}$$

where $F^* = \inf F(\mathbf{x}) > -\infty$.

(iv) Every limit point of $\{\mathbf{x}^k\}$ is a stationary point.

Suppose H is σ-strongly convex with respect to the Euclidean norm $\| \cdot \|$. Then we have $D_H(\mathbf{y}, \mathbf{x}) \geq \frac{\sigma}{2} \|\mathbf{y} - \mathbf{x}\|^2$. Combining the strongly convexity of H with the Theorem 2, we immediately obtain the following convergence rate result

$$\min_{0 \leq l \leq k} \mathbb{E}\|T(\mathbf{x}^l) - \mathbf{x}^l\|^2 \leq \tfrac{2n}{\sigma(k+1)}(F(\mathbf{x}^0) - F^*). \tag{20}$$

Therefore, the sequence $\{\mathbf{x}^k\}$ converges to a stationary point at the rate of $\mathcal{O}(\frac{\sqrt{n}}{\sqrt{k}})$. In another word, to obtain an ε-stationary point, *i.e.*, $\|T(\mathbf{x}) - \mathbf{x}\| \leq \varepsilon$, the RBCD method needs to run $\mathcal{O}(\frac{n}{\varepsilon^2})$ iterations.

5. ACCELERATED RANDOMIZED BREGMAN COORDINATE DESCENT

In this section, we restrict ourselves to the unconstrained smooth minimization problem

$$\underset{\mathbf{x}}{\text{minimize}} \;\; f(\mathbf{x}), \tag{21}$$

where f is convex and satisfies Assumption 1.

The accelerated randomized Bregman coordinate descent (ARBCD) method is given in Algorithm 2. At the k-th iteration, the ARBCD method selects a coordinate i_k uniformly at random, and generates the three vectors $\mathbf{y}^k$, $\mathbf{z}^{k+1}$, and $\mathbf{x}^{k+1}$, where

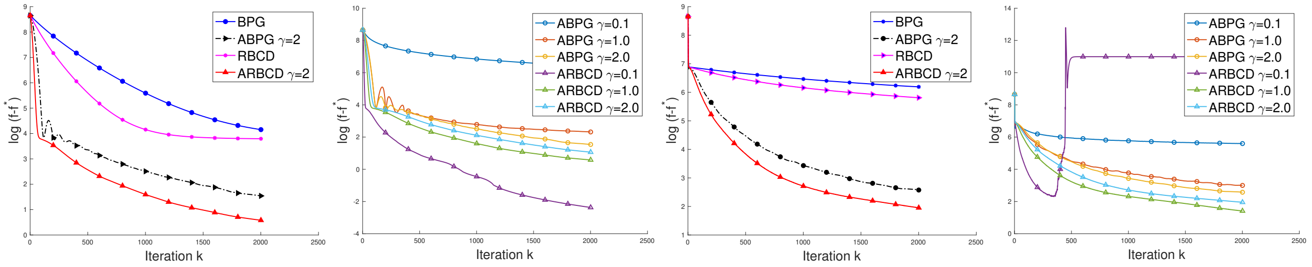
$$\mathbf{z}^{k+1} = \text{argmin}_u \langle \nabla_{i_k} f(\mathbf{y}^k), \mathbf{u}_{i_k} - \mathbf{y}^k_{i_k} \rangle + (n\beta_k)^{\gamma-1} D_H(\mathbf{u}, \mathbf{z}^k). \tag{22}$$

Note that Step 1 and 3 of Algorithm 2 need $\mathcal{O}(n)$ operations, while $\mathcal{O}(1)$ operations are usually expected in a general coordinate descent method. Due to the space limit, we will introduce an efficient implementation of the ARBCD method in our journal version that only needs $\mathcal{O}(1)$ operations at each iteration.

One of the challenges to establish the convergence results is from the nature of Bregman distances, that is, a Bregman distance does not hold the *homogeneous translation invariant*, *i.e.*,

$$\|\mathbf{u} + \theta(\mathbf{v} - \mathbf{w}) - \mathbf{u}\| = |\theta| \, \|\mathbf{v} - \mathbf{w}\|, \quad \forall \theta, \mathbf{u}, \mathbf{v}, \mathbf{w}. \tag{23}$$

To handle this issue, [21] introduces the notion of *triangle scaling property* (TSP). In contrast, we introduce the notion of *generalized translation invariant* (GTI), and show it is equivalent to triangle scaling property, when restricting $\theta \in [0, 1]$.



(a) Performance of algorithms.

(b) Effectiveness of γ.

(c) Performance of algorithms.

(d) Effectiveness of γ.

Fig. 1: Poisson inverse problem: synthetic dataset with $M = 500$ and $N = 500$. (a)-(b) are the results for (27) and (c)-(d) are for (28).

Algorithm 2: Accelerated Randomized Bregman (Block) Coordinate Descent (ARBCD).

```
Input: initial $\mathbf{x}_0$ and $\gamma \geq 1$
Initialize: $\mathbf{z}^0 = \mathbf{x}^0$ and $\beta_0 = 1$
for $k = 1, 2 \cdots$ do
    1. $\mathbf{y}^k = (1 - \beta_k)\mathbf{x}^k + \beta_k \mathbf{z}^k$
    2. Choose $i_k \in \{1, 2, \cdots, n\}$ uniformaly at random
    3. Compute $\mathbf{z}^{k+1}$ by Eq.(22)
    4. $\mathbf{x}^{k+1} = \mathbf{y}^k + n\beta_k(\mathbf{z}^{k+1} - \mathbf{z}^k)$
    5. Choose $\beta_{k+1} \in (0, 1]$ such that $\frac{1-\beta_{k+1}}{\beta_{k+1}^{\gamma}} \leq \frac{1}{\beta_k^{\gamma}}$
end
```

Definition 3 (Generalized Translation Invariant)**.** *The Bregman distance defined with a convex reference function h has the generalized translation invariant property if there exists some scalar $\gamma > 0$ such that for all* $\mathbf{u}, \mathbf{v}, \mathbf{w}$

$$D_h(\mathbf{u} + \theta(\mathbf{v} - \mathbf{w}), \mathbf{u}) \leq |\theta|^{\gamma} D_h(\mathbf{v}, \mathbf{w}). \quad \forall \theta \in \mathbf{R}. \tag{24}$$

Lemma 6. *The Bregman distance has the generalized translation invariant with $\theta \in [0, 1]$ iff it holds the triangle scaling property.*

Note that the GTI is more general since TSP needs $\theta \in [0, 1]$, but GTI holds for all $\theta \in \mathbf{R}$. To use the notion of GTI, we make the following assumption.

Assumption 4. *The Bregman distances $D_h(\cdot, \cdot)$ have the generalized translation invariant with the constant $\gamma > 0$, $\forall i$.*

Using the notion of GTI, we will show that the ARBCD method converges with a sublinear rate of $\mathcal{O}(n^{\gamma} k^{-\gamma})$. The key relationship between two consecutive iterates in Algorithm 2 is established in the following lemma.

Lemma 7. *Suppose Assumptions 1, 2, and 4 hold. For any vector* $\mathbf{u}$*, the sequences generated by Algorithm 2 satisfy, for all $k \geq 0$,*

$$\begin{aligned}&\mathbb{E}_{i_k}\left[\frac{1-\beta_{k+1}}{\beta_{k+1}^{\gamma}}(f(\mathbf{x}^{k+1}) - f(\mathbf{u})) + n^{\gamma} D_H(\mathbf{u}, \mathbf{z}^{k+1})\right] \\ \leq &\frac{1-\beta_k}{\beta_k^{\gamma}}(f(\mathbf{x}^k) - f(\mathbf{u})) + n^{\gamma} D_H(\mathbf{u}, \mathbf{z}^k).\end{aligned} \tag{25}$$

The following lemma introduces a sequence $\{\beta_k\}$ that satisfies the condition in Step 4 of Algorithm 2.

Lemma 8. *The sequence $\beta_k = \frac{\gamma}{k+\gamma}$ satisfies $\frac{\beta_{k+1}-1}{\beta_{k+1}^{\gamma}} \leq \frac{1}{\beta_k^{\gamma}}, \forall k$.*

Theorem 3. *Suppose Assumptions 1, 2, and 4 hold. If $\beta_k = \frac{\gamma}{k+\gamma}$ for all $k \geq 0$, then the following inequality holds, for any* $\mathbf{u}$*,*

$$\mathbb{E}\left[f(\mathbf{x}^{k+1}) - f(\mathbf{u})\right] \leq \left(\frac{n\gamma}{k+\gamma}\right)^{\gamma} D_H(\mathbf{u}, \mathbf{x}^0), \ \forall k \geq 0. \tag{26}$$

6. NUMERICAL EXPERIMENTS

To showcase the strength of the proposed methods, we consider an application of Poisson inverse problem or relative-entropy nonnegative regression.

A large number of problems in nuclear medicine, night vision, astronomy and hyperspectral imaging can be formulated as Poisson inversion problems [22, 23, 24, 25, 26] in the following form

$$\underset{\mathbf{x} \geq 0}{\text{minimize}} \ \ f(\mathbf{x}) \equiv D_{\text{KL}}(\mathbf{b}, \mathbf{A}\mathbf{x}). \tag{27}$$

where $\mathbf{A} \in \mathbf{R}_+^{M \times N}$ is nonnegative observation matrix and $\mathbf{b} \in \mathbf{R}_+^M$ is noisy measurement. To apply the proposed methods, we applied the Burg's entropy as the reference function. Then we can show the functions (f_i, h_i) are relatively smooth with any scalar L_i satisfying $L_i \geq \|\mathbf{b}\|_1 = \sum_{i=1}^{M} \mathbf{b}_i$.

Anther broadly used formulation to solve the Poisson inverse problem is to minimize $D_{\text{KL}}(\mathbf{A}\mathbf{x}, \mathbf{b})$ [25], *i.e.*,

$$\underset{\mathbf{x} \geq 0}{\text{minimize}} \ \ f(\mathbf{x}) \equiv D_{\text{KL}}(\mathbf{A}\mathbf{x}, \mathbf{b}). \tag{28}$$

In this case, we choose the Boltzman-Shannon entropy $h(\mathbf{x}) = \mathbf{x} \log \mathbf{x}$ as the reference function. We can show (f_i, h_i) are relatively smooth with any scalar L_i satisfying $L_i \geq \sum_{i=1}^{M} \mathbf{a}_{ij}$.

We compare the proposed algorithms RBCD and ARBCD with two state-of-the-art algorithms: Bregman Proximal Gradient (BPG) method [12] and accelerated Bregman Proximal Gradient (ABPG) [21] method. All algorithms are implemented in Matlab code.

In Figure 1(a) and (c), we can see that the RBCD method is only slightly better than the BPG method, because BPG and RBCD methods use the same stepsize $\alpha_k = \frac{1}{2\|b\|_1}$, while the ARBCD method is faster than the other methods. ARBCD method can be even faster if we select γ smaller which are shown in Figure 1(c) and (d). If γ is too small, however, ARBCD method could diverge which is shown in Figure 1(d), since D_{IS} does not hold GTI or TSP property.

7. REFERENCES

[1] Yu Nesterov, "Efficiency of coordinate descent methods on huge-scale optimization problems," *SIAM Journal on Optimization*, vol. 22, no. 2, pp. 341–362, 2012.

[2] Zhaosong Lu and Lin Xiao, "On the complexity analysis of randomized block-coordinate descent methods," *Mathematical Programming*, vol. 152, no. 1-2, pp. 615–642, 2015.

[3] Julie Nutini, Mark Schmidt, Issam Laradji, Michael Friedlander, and Hoyt Koepke, "Coordinate descent converges faster with the gauss-southwell rule than random selection," in *Proceedings of International Conference on Machine Learning*, 2015, pp. 1632–1641.

[4] Amir Beck and Luba Tetruashvili, "On the convergence of block coordinate descent type methods," *SIAM Journal on Optimization*, vol. 23, no. 4, pp. 2037–2060, 2013.

[5] Stephen J Wright, "Coordinate descent algorithms," *Mathematical Programming*, vol. 151, no. 1, pp. 3–34, 2015.

[6] Daniel D Lee and H Sebastian Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788, 1999.

[7] Tianxiang Gao, Sigurdur Olofsson, and Songtao Lu, "Minimum-volume-regularized weighted symmetric nonnegative matrix factorization for clustering," in *2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2016, pp. 247–251.

[8] Yong-Deok Kim and Seungjin Choi, "Nonnegative tucker decomposition," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[9] Yangyang Xu, Wotao Yin, Zaiwen Wen, and Yin Zhang, "An alternating direction algorithm for matrix completion with nonnegative factors," *Frontiers of Mathematics in China*, vol. 7, no. 2, pp. 365–384, 2012.

[10] Niao He, Zaid Harchaoui, Yichen Wang, and Le Song, "Fast and simple optimization for poisson likelihood models," *arXiv preprint arXiv:1608.01264*, 2016.

[11] Benjamin Birnbaum, Nikhil R Devanur, and Lin Xiao, "Distributed algorithms via gradient descent for fisher markets," in *Proceedings of the 12th ACM conference on Electronic commerce*. ACM, 2011, pp. 127–136.

[12] Heinz H Bauschke, Jérôme Bolte, and Marc Teboulle, "A descent lemma beyond lipschitz gradient continuity: first-order methods revisited and applications," *Mathematics of Operations Research*, 2016.

[13] Haihao Lu, Robert M Freund, and Yurii Nesterov, "Relatively smooth convex optimization by first-order methods, and applications," *SIAM Journal on Optimization*, vol. 28, no. 1, pp. 333–354, 2018.

[14] Jérôme Bolte, Shoham Sabach, Marc Teboulle, and Yakov Vaisbourd, "First order methods beyond convexity and lipschitz gradient continuity with applications to quadratic inverse problems," *SIAM Journal on Optimization*, vol. 28, no. 3, pp. 2131–2151, 2018.

[15] Masoud Ahookhosh, Le Thi Khanh Hien, Nicolas Gillis, and Panagiotis Patrinos, "Multi-block bregman proximal alternating linearized minimization and its application to sparse orthogonal nonnegative matrix factorization," *arXiv preprint arXiv:1908.01402*, 2019.

[16] Xiangfeng Wang, Xiaoming Yuan, Shangzhi Zeng, Jin Zhang, and Jinchuan Zhou, "Block coordinate proximal gradient method for nonconvex optimization problems: Convergence analysis," 2018.

[17] Tianxiang Gao, Songtao Lu, Jia Liu, and Chris Chu, "Leveraging two reference functions in block bregman proximal gradient descent for non-convex and non-lipschitz problems," *arXiv preprint arXiv:1912.07527*, 2019.

[18] Filip Hanzely and Peter Richtárik, "Fastest rates for stochastic mirror descent methods," *arXiv preprint arXiv:1803.07374*, 2018.

[19] Tianxiang Gao, Songtao Lu, Jia Liu, and Chris Chu, "Randomized bregman coordinate descent methods for non-lipschitz optimization," *arXiv preprint arXiv:2001.05202*, 2020.

[20] Silvia Bonettini, Marco Prato, and Simone Rebegoldi, "A cyclic block coordinate descent method with generalized gradient projections," *Applied Mathematics and Computation*, vol. 286, pp. 288–300, 2016.

[21] Filip Hanzely, Peter Richtarik, and Lin Xiao, "Accelerated bregman proximal gradient methods for relatively smooth convex optimization," *arXiv preprint arXiv:1808.03045*, 2018.

[22] Imre Csiszar et al., "Why least squares and maximum entropy? an axiomatic approach to inference for linear inverse problems," *The annals of statistics*, vol. 19, no. 4, pp. 2032–2066, 1991.

[23] Arthur P Dempster, Nan M Laird, and Donald B Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[24] Heinz Werner Engl, Martin Hanke, and Andreas Neubauer, *Regularization of inverse problems*, vol. 375, Springer Science & Business Media, 1996.

[25] Mario Bertero, Patrizia Boccacci, Gabriele Desiderà, and Giuseppe Vicidomini, "Image deblurring with poisson data: from cells to galaxies," *Inverse Problems*, vol. 25, no. 12, pp. 123006, 2009.

[26] Tianxiang Gao and Chris Chu, "Did: distributed incremental block coordinate descent for nonnegative matrix factorization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018, vol. 32.

A. APPENDIX

A.1. Proof of Lemma 2

Proof. From the relative smoothness, we obtain

$$f(\mathbf{x}^+) \leq f(\mathbf{x}) + \langle \nabla_i f(\mathbf{x}), T_i(\mathbf{x}) - \mathbf{x}_i \rangle + L_i D_h(T_i(\mathbf{x}), \mathbf{x}_i). \tag{29}$$

From the optimality of $T_i(\mathbf{x})$ in (5), we have

$$\nabla_i f(\mathbf{x}) + \frac{1}{\alpha} \left(\nabla h_i(T_i(\mathbf{x}) - \nabla h_i(\mathbf{x}_i)\right) + \mathbf{v}_i^+ = 0,$$

for some $\mathbf{v}_i^+ \in \partial r_i(T_i(\mathbf{x}))$. The convexity of r_i implies

$$\begin{aligned} r_i(\mathbf{x}_i) \geq & r_i(T_i(\mathbf{x})) + \langle \mathbf{v}_i^+, \mathbf{x}_i - T_i(\mathbf{x}) \rangle \\ = & r_i(T_i(\mathbf{x})) - \langle \nabla_i f(\mathbf{x}) + \frac{1}{\alpha} \left(\nabla h_i(T_i(\mathbf{x}) - \nabla h_i(\mathbf{x}_i)\right), \mathbf{x}_i - T_i(\mathbf{x}) \rangle \\ = & r_i(T_i(\mathbf{x})) - \langle \nabla_i f(\mathbf{x}), \mathbf{x}_i - T_i(\mathbf{x}) \rangle + \frac{1}{\alpha} \left(D_h(\mathbf{x}_i, T_i(\mathbf{x})) + D_h(T_i(\mathbf{x}), \mathbf{x}_i)\right) \end{aligned} \tag{30}$$

Combining the equations (29) and (30) yields

$$\begin{aligned} f(\mathbf{x}^+) + r_i(T_i(\mathbf{x})) \leq & f(\mathbf{x}) + r_i(\mathbf{x}_i) + L_i D_h(T_i(\mathbf{x}), \mathbf{x}_i) - \frac{1}{\alpha} \left(D_h(\mathbf{x}_i, T_i(\mathbf{x})) + D_h(T_i(\mathbf{x}), \mathbf{x}_i)\right) \\ \leq & f(\mathbf{x}) + r_i(\mathbf{x}_i) - \left(\frac{1+\theta}{\alpha} - L_i\right) D_h(T_i(\mathbf{x}), \mathbf{x}_i), \end{aligned}$$

where the second inequality is due to $D_h(\mathbf{x}_i, T_i(\mathbf{x})) \geq \theta D_h(T_i(\mathbf{x}), \mathbf{x}_i)$. Since $\mathbf{x}_j^+ = \mathbf{x}_j \ \forall i \neq j$, we obtain

$$F(\mathbf{x}^+) \leq F(\mathbf{x}) - \left(\frac{1+\theta}{\alpha} - L_i\right) D_h(T_i(\mathbf{x}), \mathbf{x}_i).$$

□

A.2. Proof of Lemma 3

Proof. ($\Longrightarrow$). Suppose $\mathbf{x}$ is a stationary point. Then we have

$$\nabla f(\mathbf{x}) + \mathbf{v} = 0,$$

for some $\mathbf{v} \in \partial r(\mathbf{x})$. From the convexity of r, it follows that for any vector $\mathbf{u}$

$$r(\mathbf{u}) \geq r(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{u} - \mathbf{x} \rangle. \tag{31}$$

By the optimality of (9), we obtain

$$\nabla f(\mathbf{x}) + \nabla H(T(\mathbf{x})) - \nabla H(\mathbf{x}) + \mathbf{v}^+ = 0, \tag{32}$$

for some $\mathbf{v}^+ \in \partial r(T(\mathbf{x}))$. It follows that

$$r(\mathbf{x}) \geq r(T(\mathbf{x})) - \langle \nabla f(\mathbf{x}), \mathbf{x} - T(\mathbf{x}) \rangle - \langle \nabla H(T(\mathbf{x})) - \nabla H(\mathbf{x}), \mathbf{x} - T(\mathbf{x}) \rangle. \tag{33}$$

Let $\mathbf{u} = T(\mathbf{x})$ and combine the equations (31) and (33). Then we obtain

$$0 \geq D_H(\mathbf{x}, T(\mathbf{x})) + D_H(T(\mathbf{x}), \mathbf{x}).$$

Since $D_H(\mathbf{x}, T(\mathbf{x})), D_H(T(\mathbf{x}), \mathbf{x}) \geq 0$, we obtain $D_H(T(\mathbf{x}), \mathbf{x}) = 0$.

($\Longleftarrow$). Suppose $D_H(T(\mathbf{x}), \mathbf{x}) = 0$. The (strict) convexity of H implies $T(\mathbf{x}) = \mathbf{x}$. From (32), we obtain

$$0 \in \nabla f(\mathbf{x}) + \partial r(\mathbf{x}),$$

which indicates $\mathbf{x}$ is a stationary point. □

A.3. Proof of Lemma 4

Proof. Since each block i is selected uniformly at random, we have

$$\begin{aligned}
\mathbb{E}_i[F(\mathbf{x}^+)] =& \sum_{i=1}^{n} \frac{1}{n} F(\mathbf{x}^+) \\
=& \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{x}^+) + r(\mathbf{x}^+) \\
\overset{(a)}{\le}& \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{x}) + \langle \nabla_i f(\mathbf{x}), T_i(\mathbf{x}) - \mathbf{x}_i \rangle + L_i D_h(T_i(\mathbf{x}), \mathbf{x}_i) + r_i(T_i(\mathbf{x})) + \sum_{j \neq i} r_j(\mathbf{x}_j) \\
\overset{(b)}{\le}& \frac{1}{n} \left[n f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), T(\mathbf{x}) - \mathbf{x} \rangle + D_H(T(\mathbf{x}), \mathbf{x}) + r(T(\mathbf{x})) + (n-1) r(\mathbf{x}) \right] \\
=& \frac{1}{n} \left[(n-1) F(\mathbf{x}) + f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), T(\mathbf{x}) - \mathbf{x} \rangle + D_H(T(\mathbf{x}), \mathbf{x}) + r(T(\mathbf{x})) \right] \\
\overset{(c)}{=}& \frac{1}{n} \left[(n-1) F(\mathbf{x}) + f(\mathbf{u}) - \mu D_H(\mathbf{u}, \mathbf{x}) + D_H(T(\mathbf{x}), \mathbf{x}) + r(\mathbf{u}) + \langle \nabla H(T(\mathbf{x})) - \nabla H(\mathbf{x}), \mathbf{u} - T(\mathbf{x}) \rangle \right] \\
\overset{(d)}{=}& \frac{1}{n} \left[(n-1) F(\mathbf{x}) + F(\mathbf{u}) + (1-\mu) D_H(\mathbf{u}, \mathbf{x}) - D_H(\mathbf{u}, T(\mathbf{x})) \right]
\end{aligned}$$

where (a) follows from the relative smoothness of (f_i, h_i); (b) uses the fact of $T_i(\mathbf{x}) = T(\mathbf{x})_i$; (c) is based on the convexity of f and r; (d) uses the the fact of $\langle \nabla h(\mathbf{z}) - \nabla h(\mathbf{x}), \mathbf{y} - \mathbf{z} \rangle = D_h(\mathbf{y}, \mathbf{x}) - D_h(\mathbf{y}, \mathbf{z}) - D_h(\mathbf{z}, \mathbf{x})$.

For any vector $\mathbf{u}$, we have

$$\begin{aligned}
D_H(\mathbf{u}, \mathbf{x}^+) =& L_i D_h(\mathbf{u}_i, T_i(\mathbf{x})) + \textstyle\sum_{j \neq i} L_j D_h(\mathbf{u}_j, \mathbf{x}_j) \\
=& L_i D_h(\mathbf{u}_i, T_i(\mathbf{x})) - L_i D_h(\mathbf{u}_i, \mathbf{x}_i) + D_H(\mathbf{u}, \mathbf{x})
\end{aligned} \tag{34}$$

Taking the expectation of Eq.(34) with respect to i yields

$$\begin{aligned}
\mathbb{E}_i[D_H(\mathbf{u}, \mathbf{x}^+)] =& \mathbb{E}_i \left[D_H(\mathbf{u}, \mathbf{x}) - L_i D_h(\mathbf{u}_i, \mathbf{x}_i) + L_i D_h(\mathbf{u}_i, T_i(\mathbf{x})) \right] \\
=& \sum_{i=1}^{n} \frac{1}{n} \left[D_H(\mathbf{u}, \mathbf{x}) - L_i D_h(\mathbf{u}_i, \mathbf{x}_i) + L_i D_h(\mathbf{u}_i, T_i(\mathbf{x})) \right] \\
=& \frac{1}{n} \left[n D_H(\mathbf{u}, \mathbf{x}) - D_H(\mathbf{u}, \mathbf{x}) + D_H(\mathbf{u}, T(\mathbf{x})) \right] \\
=& D_H(\mathbf{u}, \mathbf{x}) - \frac{1}{n} \left[D_H(\mathbf{u}, \mathbf{x}) - D_H(\mathbf{u}, T(\mathbf{x})) \right]
\end{aligned}$$

□

A.4. Proof of Theorem 1

Proof. Combining (15) with (14), let $\mathbf{u} = \mathbf{x}^*$, and we have

$$\mathbb{E}_i[F(\mathbf{x}^+) + D_H(\mathbf{x}^*, \mathbf{x}^+)] \le \frac{n-1}{n} F(\mathbf{x}) + \frac{1}{n} F(\mathbf{x}^*) + \left(1 - \frac{\mu}{n}\right) D_H(\mathbf{x}^*, \mathbf{x}) \tag{35}$$

$$\le \frac{n-1}{n} F(\mathbf{x}) + \frac{1}{n} F(\mathbf{x}^*) + D_H(\mathbf{x}^*, \mathbf{x}). \tag{36}$$

Taking the expectation of (36) with respect to $\{i_0, i_1, \cdots\}$ yields

$$\mathbb{E}[F(\mathbf{x}^+)] \le \mathbb{E}\left[F(\mathbf{x}) + D_H(\mathbf{x}^*, \mathbf{x}) - D_H(\mathbf{x}^*, \mathbf{x}^+) - \frac{1}{n} \left(F(\mathbf{x}) - F(\mathbf{x}^*) \right) \right].$$

Summing over $l = 0, 1, \cdots, k-1$ yields

$$\begin{aligned}
\mathbb{E}[F(\mathbf{x}^k)] \le& F(\mathbf{x}^0) + D_H(\mathbf{x}^*, \mathbf{x}^0) - \mathbb{E}[D_H(\mathbf{x}^*, \mathbf{x}^k)] - \frac{1}{n} \sum_{l=0}^{k-1} \mathbb{E}\left[F(\mathbf{x}^l) - F(\mathbf{x}^*) \right] \\
\le& F(\mathbf{x}^0) + D_H(\mathbf{x}^*, \mathbf{x}^0) - \frac{1}{n} \sum_{l=0}^{k-1} \mathbb{E}\left[F(\mathbf{x}^l) - F(\mathbf{x}^*) \right] \\
\le& F(\mathbf{x}^0) + D_H(\mathbf{x}^*, \mathbf{x}^0) - \frac{k}{n} \mathbb{E}\left[F(\mathbf{x}^{k+1}) - F(\mathbf{x}^*) \right],
\end{aligned}$$

where the last inequality is because $\{F(\mathbf{x}^l)\}$ is a descent sequence. Subtracting $F(\mathbf{x}^*)$ on both sides and rearrange yields

$$\frac{n+k}{n}\mathbb{E}[F(\mathbf{x}^k) - F(\mathbf{x}^*)] \le F(\mathbf{x}^*) - F(\mathbf{x}^0) + D_H(\mathbf{x}^*, \mathbf{x}^0).$$

Dividing both sides by $\frac{n+k}{n}$ yields the desired result.

If f is μ-strongly convex relative to H, we have

$$\mathbb{E}_i[F(\mathbf{x}^+) + D_H(\mathbf{x}^*, \mathbf{x}^+)] \le \frac{n-1}{n}F(\mathbf{x}) + \frac{1}{n}F(\mathbf{x}^*) + \left(1 - \frac{\mu}{n}\right) D_H(\mathbf{x}^*, \mathbf{x}).$$

Subtracting $F(\mathbf{x}^*)$ on the both sides and rearrange yields

$$\mathbb{E}_i[F(\mathbf{x}^+) - F(\mathbf{x}^*) + D_H(\mathbf{x}^*, \mathbf{x}^+)] \le F(\mathbf{x}) - F(\mathbf{x}^*) + D_H(\mathbf{x}^*, \mathbf{x}) - \frac{1}{n}F(\mathbf{x}) - F(\mathbf{x}^*) + \mu D_H(\mathbf{x}^*, \mathbf{x}). \tag{37}$$

The relative strongly convexity of F implies

$$F(\mathbf{x}) - F(\mathbf{x}^*) + \mu D_H(\mathbf{x}^*, \mathbf{x}) \ge \mu D_H(\mathbf{x}, \mathbf{x}^*) + \mu D_H(\mathbf{x}^*, \mathbf{x}) \ge (1+\theta)\mu D_H(\mathbf{x}^*, \mathbf{x}).$$

Define

$$\beta = \frac{(1+\theta)\mu}{1+\theta\mu}. \tag{38}$$

Clearly, we have $\beta \le 1$ since $\mu \le 1$. Then

$$\begin{aligned} F(\mathbf{x}) - F(\mathbf{x}^*) + \mu D_H(\mathbf{x}^*, \mathbf{x}) \ge& \beta(F(\mathbf{x}) - F(\mathbf{x}^*) + \mu D_H(\mathbf{x}^*, \mathbf{x})) + (1-\beta)(1-\theta)\mu D_H(\mathbf{x}^*, \mathbf{x}) \\ =& \beta(F(\mathbf{x}) - F(\mathbf{x}^*) + D_H(\mathbf{x}^*, \mathbf{x})). \end{aligned}$$

Combining the inequality above with (37) yields

$$\mathbb{E}_i[F(\mathbf{x}^+) - F(\mathbf{x}^*) + D_H(\mathbf{x}^*, \mathbf{x}^+)] \le \left(1 - \frac{\beta}{n}\right)(F(\mathbf{x}) - F(\mathbf{x}^*) + D_H(\mathbf{x}^*, \mathbf{x}))$$

Taking the expectation with respect to $\{i_0, i_1, \cdots\}$ on the both sides of the relation above, we have

$$\mathbb{E}[F(\mathbf{x}^k) - F(\mathbf{x}^*) + D_H(\mathbf{x}^*, \mathbf{x}^k)] \le \left(1 - \frac{\beta}{n}\right)^k \left(F(\mathbf{x}^0) - F(\mathbf{x}^*) + D_H(\mathbf{x}^*, \mathbf{x}^0)\right).$$

Dropping $D_H(x^*, \mathbf{x}^k)$ on the left hand yields the desired result. □

A.5. Proof of Lemma 5

Proof. Taking the expectation of (7) with respect to i yields

$$\begin{aligned} E_i[F(\mathbf{x}^+)] &\le F(\mathbf{x}) - E_i[L_i D_h(T_i(\mathbf{x}), \mathbf{x}_i)] \\ &= F(\mathbf{x}) - \sum_{i=1}^n \frac{1}{n} L_i D_h(T_i(\mathbf{x}), \mathbf{x}_i) \\ &= F(\mathbf{x}) - \frac{1}{n}\sum_{i=1}^n L_i D_h(T_i(\mathbf{x}), \mathbf{x}_i) \\ &\overset{(a)}{=} F(\mathbf{x}) - \frac{1}{n}\sum_{i=1}^n L_i D_h(T(\mathbf{x})_i, \mathbf{x}_i) \\ &= F(\mathbf{x}) - \frac{1}{n} D_H(T(\mathbf{x}), \mathbf{x}), \end{aligned}$$

where (a) is because $T_i(\mathbf{x}) = T(\mathbf{x})_i$. □

A.6. Proof of Theorem 2

Proof. (i). The result is directly obtained from Lemma 5.
(ii). Taking the expectation of (18) with respect to all variables and rearranging yields

$$\mathbb{E}\left[D_H(T(\mathbf{x}^l),\mathbf{x}^l)\right] \leq n\mathbb{E}\Big(F(\mathbf{x}^l) - F(\mathbf{x}^{l+1})\Big).$$

Taking the telescopic sum of the above inequality for $l = 0, 1, \cdots, k$ gives us

$$\sum_{l=0}^{k}\mathbb{E}\left[D_H(T(\mathbf{x}^l),\mathbf{x}^l)\right] \leq n\left(F(\mathbf{x}^0) - \mathbb{E}[F(\mathbf{x}^{K+1})]\right) \leq n\left(F(\mathbf{x}^0) - F^*\right). \tag{39}$$

Since F is lower bounded, taking the limit $k \to \infty$ yields the desired result.
(iii). The inequality (39) further implies that

$$(k+1)\min_{0\leq l\leq k}\mathbb{E}\left[D_H(T(\mathbf{x}^l),\mathbf{x}^l)\right] \leq \sum_{l=0}^{k}\mathbb{E}\left[D_H(T(\mathbf{x}^l),\mathbf{x}^l)\right] \leq n(F(\mathbf{x}^0) - F^*).$$

Dividing $k+1$ on both sides gives us the desired result.

(iv) Let x^* to be a limit point of $\{\mathbf{x}^k\}$ and there exists a subsequence $\{\mathbf{x}^{k_p}\}$ such that $\mathbf{x}^{k_p} \to \mathbf{x}^*$ as $p \to \infty$. Since the functions r_i are lower semi-continuous, we have for all i,

$$\liminf_{p\to\infty} r_i(\mathbf{x}_i^{k_p}) \geq r_i(\mathbf{x}_i^*). \tag{40}$$

At the k-th iteration, suppose the index i is selected, then the convexity of r_i implies that

$$r_i(\mathbf{x}_i^{k+1}) \leq r_i(\mathbf{x}_i^*) + \langle \nabla_i f(\mathbf{x}^k) + \nabla h_i(\mathbf{x}_i^{k+1}) - \nabla h_i(\mathbf{x}_i^k), \mathbf{x}_i^* - \mathbf{x}_i^{k+1}\rangle$$

Let $\{\mathbf{x}^{k_q}\}$ be the subsequence of $\{\mathbf{x}^{k_p}\}$ such that the index i is selected. Choosing $k = k_q - 1$ in the above inequality, and letting $q \to$ yields

$$\limsup_{q\to\infty} r_i(\mathbf{x}_i^{k_q}) \leq r_i(\mathbf{x}_i^*), \tag{41}$$

where we use the facts $\mathbf{x}^{k_q} \to \mathbf{x}^*$ as $q \to \infty$. Thus, combining (41) with (40), we have

$$\lim_{q\to\infty} r_i(\mathbf{x}_i^{k_q}) = r_i(\mathbf{x}_i^*).$$

Since i is selected arbitrarily, we have

$$\lim_{p\to\infty} r_i(\mathbf{x}_i^{k_p}) = r_i(\mathbf{x}_i^*), \quad \forall i.$$

Furthermore, by the continuity of f, we obtain

$$\lim_{p\to\infty} F(\mathbf{x}^{k_p}) = \lim_{p\to\infty}\left\{f(\mathbf{x}^{k_p}) + \sum_{i=1}^{n} r_i(\mathbf{x}^{k_p})\right\} = f(\mathbf{x}^*) + \sum_{i=1}^{n} r_i(\mathbf{x}_i^*) = F(\mathbf{x}^*).$$

From (ii) and Lemma 3, it follows that $\mathbf{x}^*$ is a stationary point of F. □

A.7. Proof of Lemma 6

Proof. $\Longrightarrow$. Suppose the Bregman distance $D_h(\cdot,\cdot)$ holds the generalized translation variant, and let $\mathbf{u} = (1-\theta)\mathbf{x} + \theta\mathbf{w}$ for any $\mathbf{x}$. Then we have

$$D_h((1-\theta)\mathbf{x} + \theta\mathbf{v}, (1-\theta)\mathbf{x} + \theta\mathbf{w}) \leq |\theta|^{\gamma} D_h(\mathbf{v},\mathbf{w}), \quad \forall\theta \in \mathbf{R}.$$

Since the above inequality holds for all θ, it must hold for $\theta \in [0,1]$.

$\Longleftarrow$. Suppose the triangle scaling property holds. Let $\mathbf{y} = (1-\theta)\mathbf{u} + \theta\mathbf{w}$, then we have

$$D_h(\mathbf{y} + \theta(\mathbf{v} - \mathbf{w}), \mathbf{y}) \leq \theta^{\gamma} D_h(\mathbf{v},\mathbf{w}), \quad \forall\theta \in [0,1]. \tag{42}$$

Therefore, the generalized translation invariant holds for $\theta \in [0,1]$. □

A.8. Proof of Lemma 7

Proof. With simple algebra operations, we have

$$\mathbf{x}^{k+1} - \mathbf{y}^k = n\left[\beta_k(\mathbf{z}^{k+1} - \mathbf{y}^k) + (1-\beta_k)(\mathbf{x}^k - \mathbf{y}^k)\right]. \tag{43}$$

Based on the relation in E.q. (**??**), we know $\mathbf{x}^{k+1}$ and $\mathbf{y}^k$ satisfy the relative smoothness property since they are only one coordinate difference from each other. Therefore, we obtain

$$\begin{aligned}
f(\mathbf{x}^{k+1}) \leq & f(\mathbf{y}^k) + \langle \nabla_{i_k} f(\mathbf{y}^k), \mathbf{x}_{i_k}^{k+1} - \mathbf{y}_{i_k}^k \rangle + L_{i_k} D_h(\mathbf{x}_{i_k}^{k+1}, \mathbf{y}_{i_k}^k) \\
= & f(\mathbf{y}^k) + \langle \nabla_{i_k} f(\mathbf{y}^k), \mathbf{x}_{i_k}^{k+1} - \mathbf{y}_{i_k}^k \rangle + L_{i_k} D_h(\mathbf{y}_{i_k} + n\beta_k(\mathbf{z}_{i_k}^{k+1} - \mathbf{z}_{i_k}^k), \mathbf{y}_{i_k}^k) \\
\overset{(i)}{\leq} & f(\mathbf{y}^k) + \langle \nabla_{i_k} f(\mathbf{y}^k), \mathbf{x}_{i_k}^{k+1} - \mathbf{y}_{i_k}^k \rangle + (n\beta_k)^\gamma L_{i_k} D_h((\mathbf{z}_{i_k}^{k+1}, \mathbf{z}_{i_k}^k)) \\
\overset{(ii)}{=} & f(\mathbf{y}^k) + n\beta_k \langle \nabla_{i_k} f(\mathbf{y}^k), \mathbf{z}_{i_k}^{k+1} - \mathbf{y}_{i_k}^k \rangle + n(1-\beta_k) \langle \nabla_{i_k} f(\mathbf{y}^k), \mathbf{x}_{i_k}^k - \mathbf{y}_{i_k}^k \rangle + (n\beta_k)^\gamma L_{i_k} D_h((\mathbf{z}_{i_k}^{k+1}, \mathbf{z}_{i_k}^k)) \\
\overset{(iii)}{=} & \beta_k \left[f(\mathbf{y}^k) + n\langle \nabla_{i_k} f(\mathbf{y}^k), \tilde{\mathbf{z}}_{i_k}^{k+1} - \mathbf{y}_{i_k}^k \rangle\right] + (1-\beta_k)\left[f(\mathbf{y}^k) + n\langle \nabla_{i_k} f(\mathbf{y}^k), \mathbf{x}_{i_k}^k - \mathbf{y}_{i_k}^k \rangle\right] + (n\beta_k)^\gamma L_{i_k} D_h((\tilde{\mathbf{z}}_{i_k}^{k+1}, \mathbf{z}_{i_k}^k)),
\end{aligned}$$

where (i) is using the generalized transition invariant, (ii) is due to E.q. (43), and (iii) is due to E.q. (**??**). Taking the expectation with respect to i_k on both sides yields for all $\mathbf{u}$

$$\begin{aligned}
\mathbb{E}_{i_k} f(\mathbf{x}^{k+1}) \leq & \beta_k \left[f(\mathbf{y}^k) + \langle \nabla f(\mathbf{y}^k), \tilde{\mathbf{z}}^{k+1} - \mathbf{y}^k \rangle\right] + (1-\beta_k)\left[f(\mathbf{y}^k) + \langle \nabla f(\mathbf{y}^k), \mathbf{x}^k - \mathbf{y}^k \rangle\right] + n^{\gamma-1}\beta_k^\gamma D_H(\tilde{\mathbf{z}}^{k+1}, \mathbf{z}^k) \\
\overset{(i)}{\leq} & (1-\beta_k) f(\mathbf{x}^k) + \beta_k \left[f(\mathbf{y}^k) + \langle \nabla f(\mathbf{y}^k), \tilde{\mathbf{z}}^{k+1} - \mathbf{y}^k \rangle + (n\beta_k)^{\gamma-1} D_H(\tilde{\mathbf{z}}^{k+1}, \mathbf{z}^k)\right] \\
\overset{(ii)}{\leq} & (1-\beta_k) f(\mathbf{x}^k) + \beta_k \left[f(\mathbf{y}^k) + \langle \nabla f(\mathbf{y}^k), \mathbf{u} - \mathbf{y}^k \rangle + (n\beta_k)^{\gamma-1} D_H(\mathbf{u}, \mathbf{z}^k) - (n\beta_k)^{\gamma-1} D_H(\mathbf{u}, \tilde{\mathbf{z}}^{k+1})\right] \\
\overset{(iii)}{\leq} & (1-\beta_k) f(\mathbf{x}^k) + \beta_k \left[f(\mathbf{u}) + (n\beta_k)^{\gamma-1} D_H(\mathbf{u}, \mathbf{z}^k) - (n\beta_k)^{\gamma-1} D_H(\mathbf{u}, \tilde{\mathbf{z}}^{k+1})\right],
\end{aligned}$$

where (i) is due to the convexity of f, (ii) is due to the definition of $\tilde{z}^{k+1}$ in E.q. (**??**), and (iii) is due to the convexity of f. Subtracting $f(\mathbf{u})$ on both sides gives us

$$\mathbb{E}_{i_k} f(\mathbf{x}^{k+1}) - f(\mathbf{u}) \leq (1-\beta_k)(f(\mathbf{x}^k) - f(\mathbf{u})) + n^{\gamma-1}\beta_k^\gamma D_H(\mathbf{u}, \mathbf{z}^k) - n^{\gamma-1}\beta_k^\gamma D_H(\mathbf{u}, \tilde{\mathbf{z}}^{k+1}).$$

Dividing β_k^γ on both sides, we have

$$\frac{1}{\beta_k^\gamma}\mathbb{E}_{i_k}\left[f(\mathbf{x}^{k+1}) - f(\mathbf{u})\right] \leq \frac{1-\beta_k}{\beta_k^\gamma}(f(\mathbf{x}^k) - f(\mathbf{u})) + n^{\gamma-1} D_H(\mathbf{u}, \mathbf{z}^k) - n^{\gamma-1} D_H(\mathbf{u}, \tilde{\mathbf{z}}^{k+1}). \tag{44}$$

Taking the expectation of $D_H(\mathbf{u}, \mathbf{z}^{k+1})$ with respect to $\mathbb{E}_{i_k}$ yields

$$\begin{aligned}
\mathbb{E}_{i_k}[D_H(\mathbf{u}, \mathbf{z}^{k+1})] = & \mathbb{E}_{i_k}\left[D_H(\mathbf{u}, \mathbf{z}^k) - L_{i_k} D_h(\mathbf{u}_{i_k}, \mathbf{x}_{i_k}^k) + L_{i_k} D_h(\mathbf{u}_{i_k}, \tilde{\mathbf{z}}_{i_k}^{k+1})\right] \\
= & \sum_{i_k=1}^{n} \frac{1}{n}\left[D_H(\mathbf{u}, \mathbf{z}^k) - L_{i_k} D_h(\mathbf{u}_{i_k}, \mathbf{z}_{i_k}^k) + L_{i_k} D_h(\mathbf{u}_{i_k}, \tilde{\mathbf{z}}_{i_k}^{k+1})\right] \\
= & \frac{1}{n}\left[n D_H(\mathbf{u}, \mathbf{z}^k) - D_H(\mathbf{u}, \mathbf{z}^k) + D_H(\mathbf{u}, \tilde{\mathbf{z}}^{k+1})\right] \\
= & D_H(\mathbf{u}, \mathbf{z}^k) - \frac{1}{n}\left[D_H(\mathbf{u}, \mathbf{z}^k) - D_H(\mathbf{u}, \tilde{\mathbf{z}}^{k+1})\right].
\end{aligned}$$

Multiplying both sides by n^γ, we obtain

$$n^\gamma \mathbb{E}_{i_k}[D_H(\mathbf{u}, \mathbf{z}^{k+1})] = n^\gamma D_H(\mathbf{u}, \mathbf{z}^k) - n^{\gamma-1}\left[D_H(\mathbf{u}, \mathbf{z}^k) - D_H(\mathbf{u}, \tilde{\mathbf{z}}^{k+1})\right] \tag{45}$$

Combining (45) with (44), we have

$$\mathbb{E}_{i_k}\left[\frac{1}{\beta_k^\gamma}(f(\mathbf{x}^{k+1}) - f(\mathbf{u})) + n^\gamma D_H(\mathbf{u}, \mathbf{z}^{k+1})\right] \leq \frac{1-\beta_k}{\beta_k^\gamma}(f(\mathbf{x}^k) - f(\mathbf{u})) + n^\gamma D_H(\mathbf{u}, \mathbf{z}^k) \tag{46}$$

Finally applying the condition in Step 4 of Algorithm 2 yields the desired result. □

A.9. Proof of Theorem 3

Proof. Taking the expectation with respect to $\{i_0, i_1, \cdots, \}$ yields

$$\mathbb{E}\left[\frac{1-\beta_{k+1}}{\beta_{k+1}^{\gamma}}(f(\mathbf{x}^{k+1})-f(\mathbf{u}))+n^{\gamma}D_H(\mathbf{u},\mathbf{z}^{k+1})\right] \leq \mathbb{E}\left[\frac{1-\beta_k}{\beta_k^{\gamma}}(f(\mathbf{x}^k)-f(\mathbf{u}))+n^{\gamma}D_H(\mathbf{u},\mathbf{z}^k)\right]. \tag{47}$$

The direct consequence of E.q. (47) is, for any $\mathbf{u}$,

$$\mathbb{E}\left[\frac{1-\beta_{k+1}}{\beta_{k+1}^{\gamma}}(f(\mathbf{x}^{k+1})-f(\mathbf{u}))+n^{\gamma}D_H(\mathbf{u},\mathbf{z}^k)\right] \leq \frac{1-\beta_0}{\beta_0^{\gamma}}(f(\mathbf{x}^0)-f(\mathbf{u}))+n^{\gamma}D_H(\mathbf{u},\mathbf{z}^0).$$

Using $D_H(\mathbf{u},\mathbf{z}^{k+1}) \geq 0$, and the initialization $\beta_0 = 1$ and $\mathbf{z}^0 = \mathbf{x}^0$, we obtain

$$\mathbb{E}\left[\frac{1-\beta_{k+1}}{\beta_{k+1}^{\gamma}}(f(\mathbf{x}^{k+1})-f(\mathbf{u}))\right] \leq n^{\gamma}D_H(\mathbf{u},\mathbf{x}^0),$$

which implies

$$\mathbb{E}\left[f(\mathbf{x}^{k+1})-f(\mathbf{u})\right] \leq n^{\gamma}\beta_k^{\gamma}D_H(\mathbf{u},\mathbf{x}^0) = \left(\frac{n\gamma}{k+\gamma}\right)^{\gamma}D_H(\mathbf{u},\mathbf{x}^0).$$

□

A.10. Proof of Proposition ??

Proof. It is straightforward to see that $\mathbf{x}^0 = \mathbf{y}^0 = \mathbf{z}^0 = \mathbf{v}^0$. Suppose the recursive hypotheses hold for the k-th iteration. From the optimality of E.q. (**??**), we have

$$\begin{aligned}
&\langle \nabla_{i_k} f(\beta_k^{\gamma}\mathbf{u}^k+\mathbf{v}^k), \mathbf{d}_{i_k}^k\rangle + (n\beta_k)^{\gamma-1}L_{i_k}D_h(\mathbf{v}_{i_k}^k+\mathbf{d}_{i_k}^k, \mathbf{v}_{i_k}^k)\\
\overset{(i)}{\leq}&\langle \nabla_{i_k} f(\beta_k^{\gamma}\mathbf{u}^k+\mathbf{v}^k), \mathbf{z}_{i_k}^{k+1}-\mathbf{z}_{i_k}^k\rangle + (n\beta_k)^{\gamma-1}L_{i_k}D_h(\mathbf{z}_{i_k}^{k+1}, \mathbf{v}_{i_k}^k)\\
\overset{(ii)}{=}&\langle \nabla_{i_k} f(\mathbf{y}^k), \mathbf{z}_{i_k}^{k+1}-\mathbf{z}_{i_k}^k\rangle + (n\beta_k)^{\gamma-1}L_{i_k}D_h(\mathbf{z}_{i_k}^{k+1}, \mathbf{z}_{i_k}^k),
\end{aligned} \tag{48}$$

where (i) is due to the optimality, and (ii) is due to the recursive hypotheses. Similarly, from the optimality of E.q. (22), we obtain

$$\begin{aligned}
&\langle \nabla_{i_k} f(\mathbf{y}^k), \mathbf{z}_{i_k}^{k+1}-\mathbf{z}_{i_k}^k\rangle + (n\beta_k)^{\gamma-1}L_{i_k}D_h(\mathbf{z}_{i_k}^{k+1}, \mathbf{z}_{i_k}^k)\\
\overset{(i)}{\leq}&\langle \nabla_{i_k} f(\mathbf{y}^k), \mathbf{d}_{i_k}^k\rangle + (n\beta_k)^{\gamma-1}L_{i_k}D_h(\mathbf{z}_{i_k}^k+d_{i_k}^k, \mathbf{z}_{i_k}^k)\\
\overset{(ii)}{=}&\langle \nabla_{i_k} f(\beta_k^{\gamma}\mathbf{u}^k+\mathbf{v}^k), \mathbf{d}_{i_k}^k\rangle + (n\beta_k)^{\gamma-1}L_{i_k}D_h(\mathbf{v}_{i_k}^k+\mathbf{d}_{i_k}^k, \mathbf{v}_{i_k}^k),
\end{aligned} \tag{49}$$

where (i) is due to the optimality, and (ii) is due to the recursive hypotheses. Combing (48) and (49) yields

$$\mathbf{z}_{i_k}^{k+1} = \mathbf{z}_{i_k}^k + \mathbf{d}_{i_k}^k = \mathbf{v}_{i_k}^k + \mathbf{d}_{i_k}^k = \mathbf{v}_{i_k}^{k+1},$$

or equivalently

$$\mathbf{z}^{k+1} = \mathbf{v}^{k+1}.$$

From Step 3 of Algorithm **??**, we have

$$\mathbf{u}^{k+1} = \mathbf{u}^k - \frac{1-n\beta_k}{\beta_k^{\gamma}}(\mathbf{v}^{k+1}-\mathbf{v}^k). \tag{50}$$

Then, we have

$$\begin{aligned}
\beta_k^{\gamma}\mathbf{u}^{k+1}+\mathbf{v}^{k+1} \overset{(i)}{=}&\beta_k^{\gamma}\left(\mathbf{u}^k - \frac{1-n\beta_k}{\beta_k^{\gamma}}(\mathbf{v}^{k+1}-\mathbf{v}^k)\right)+\mathbf{v}^{k+1}\\
=&\beta_k^{\gamma}\mathbf{u}^k-(1-n\beta_k)(\mathbf{v}^{k+1}-\mathbf{v}^k)+\mathbf{v}^{k+1}\\
=&\beta_k^{\gamma}\mathbf{u}^k+\mathbf{v}^k+n\beta_k(\mathbf{v}^{k+1}-\mathbf{v}^k)\\
\overset{(ii)}{=}&\mathbf{y}^k+n\beta_k(\mathbf{z}^{k+1}-\mathbf{z}^k)\\
=&\mathbf{x}^{k+1},
\end{aligned}$$

where (i) is due to E.q. (50) and (ii) is due to the recursive hypotheses.

Finally, we have

$$\begin{aligned}
\beta_{k+1}^{\gamma}\mathbf{u}^{k+1}+\mathbf{v}^{k+1} \overset{(i)}{=}& \frac{\beta_{k+1}^{\gamma}}{\beta_{k}^{\gamma}}(\mathbf{x}^{k+1}-\mathbf{v}^{k+1})+\mathbf{v}^{k+1}\\
\overset{(ii)}{=}&(1-\beta_{k+1})(\mathbf{x}^{k+1}-\mathbf{v}^{k+1})+\mathbf{v}^{k+1}\\
=&(1-\beta_{k+1})\mathbf{x}^{k+1}+\beta_{k+1}\mathbf{v}^{k+1}\\
\overset{(iii)}{=}&(1-\beta_{k+1})\mathbf{x}^{k+1}+\beta_{k+1}\mathbf{z}^{k+1}\\
=&\mathbf{y}^{k+1},
\end{aligned}$$

where (i) and (iii) is due to recursive hypotheses, and (ii) is due to Step 4 of Algorithm **??**. □