Hindawi
Complexity
Volume 2018, Article ID 2609471, 11 pages
https://doi.org/10.1155/2018/2609471

Research Article

Stochastic Block-Coordinate Gradient Projection Algorithms for Submodular Maximization

Zhigang Li,[1] Mingchuan Zhang ,[2] Junlong Zhu ,[2] Ruijuan Zheng ,[2] Qikun Zhang ,[1] and Qingtao Wu [2]

[1] *School of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou, 450002, China*
[2] *Information Engineering College, Henan University of Science and Technology, Luoyang, 471023, China*

Correspondence should be addressed to Mingchuan Zhang; zhang_mch@haust.edu.cn

Received 25 May 2018; Accepted 26 November 2018; Published 5 December 2018

Academic Editor: Mahardhika Pratama

We consider a stochastic continuous submodular huge-scale optimization problem, which arises naturally in many applications such as machine learning. Due to high-dimensional data, the computation of the whole gradient vector can become prohibitively expensive. To reduce the complexity and memory requirements, we propose a stochastic block-coordinate gradient projection algorithm for maximizing continuous submodular functions, which chooses a random subset of gradient vector and updates the estimates along the positive gradient direction. We prove that the estimates of all nodes generated by the algorithm converge to some stationary points with probability 1. Moreover, we show that the proposed algorithm achieves the tight $((p_{\min}/2)F^{*} - \epsilon)$ approximation guarantee after $O(1/\epsilon^{2})$ iterations for DR-submodular functions by choosing appropriate step sizes. Furthermore, we also show that the algorithm achieves the tight $((\gamma^{2}/(1+\gamma^{2}))p_{\min}F^{*} - \epsilon)$ approximation guarantee after $O(1/\epsilon^{2})$ iterations for weakly DR-submodular functions with parameter γ by choosing diminishing step sizes.

1. Introduction

In this paper, we focus on the submodular function maximization, which has recently attracted significant attention in academia since submodularity is a crucial concept in combinatorial optimization. Furthermore, they have arisen in a variety of areas, such as social sciences, algorithm game theory, signal processing, machine learning, and computer vision. Furthermore, submodular functions have found many applications in the applied mathematics and computer science, such as probabilistic models [1, 2], crowd teaching [3, 4], representation learning [5], data summarization [6], document summarization [7], recommender systems [8], product recommendation [9, 10], sensor placement [11], network monitoring [12, 13], the design of structured norms [14], clustering [15], dictionary learning [16], active learning [17], and the utility maximization in sensor networks [18].

In submodular optimization problems, there exist many polynomial time algorithms for exactly minimizing the submodular functions, such as combinatorial algorithms [19–21]. In addition, there also exist many polynomial time algorithms for approximately maximizing the submodular functions with approximation guarantees, such as the local search and greedy algorithms [22–25]. Despite this progress, these methods use the combinatorial techniques, which have some limitations [26]. For this reason, a new approach is proposed by using multilinear relaxation [27], which can lift the submodular functions optimization problems into the continuous domain. Thus, the continuous optimization techniques are used to minimize exactly or maximize approximately submodular functions in polynomial time. Recently, most literature is devoted to continuous submodular optimization [28–31]. The algorithms cited above need to compute all the (sub)gradients.

However, the computation of all (sub)gradients can become prohibitively expensive when dealing with huge-scale optimization problems, where the decision vectors are high-dimensional. For this reason, coordinate descent method and its variants are proposed for solving efficiently convex optimization problems [32]. At each iteration, the coordinate descent methods only choose one block of variables to update their decision vectors. Thus, they can reduce the memory and complexity requirements at each iteration when dealing with high-dimensional data. Furthermore, coordinate descent methods can be applied in support vector machine [33], large-scale optimization problems [34–37], protein loop closure [38], regression [39], compressed sensing [40], etc. In coordinate descent methods, the choice of search strategy mainly include cyclic coordinate search [41–43] and the random coordinate search [44–46]. In addition, the asynchronous coordinate decent methods are also proposed in recent years [47, 48].

Despite this progress, however, stochastic block-coordinate gradient projection methods for maximizing submodular functions have barely been investigated. To fill this gap, we propose the stochastic block-coordinate gradient projection algorithm to solve stochastic continuous submodular optimization problems, which are introduced in [30]. In order to reduce the complexity and memory requirements at each iteration, we incorporate the block-coordinate decomposition into the stochastic gradient projection in the proposed algorithm. The main contributions of this paper are as follows:

(i) We propose a stochastic block-coordinate gradient projection algorithm for maximizing continuous submodular functions. In the proposed algorithm, each node chooses a random subset of the whole approximation gradient vector and updates its decision vector along gradient ascent direction.

(ii) We show that each node asymptotically converges to some stationary points by the stochastic block-coordinate gradient projection algorithm; i.e., the estimates of all nodes converge to some stationary points with probability 1.

(iii) We investigate the convergence rate of stochastic block-coordinate gradient projection algorithm with approximation guarantee. When the submodular functions are DR-submodular, we prove that the convergence rate of $O(1/\sqrt{T})$ is achieved with $p_{\min}/2$ approximation guarantee. More generally, we show that the convergence rate of $O(1/\sqrt{T})$ is achieved with $p_{\min}(1-\gamma^2/(1+\gamma^2))$ approximation guarantee for weakly DR-submodular functions with parameter γ.

The remainder of this paper is organized as follows. We describe mathematical background in Section 2. We formulate the problem of our interest and propose a stochastic block-coordinate gradient projection algorithm in Section 3. In Section 4, the main results of this paper are stated. The detailed proofs of the main results of the paper are provided in Section 5. The conclusion of the paper is presented in Section 6.

2. Mathematical Background

Given a ground set V, which consists of n elements. If a set function $f : 2^V \longrightarrow \mathbb{R}_+$ satisfies

$$f(A) + f(B) \geq f(A \cap B) + f(A \cup B) \tag{1}$$

for all subsets $A, B \subseteq V$, then the set function f is called submodular. The notation of submodularity is mostly used in discrete domain, but it can be extended to continuous domain [49]. Given a subset of $\mathbb{R}_+^n$, $\mathscr{X} = \prod_{i=1}^n \mathscr{X}_i$, where each set $\mathscr{X}_i$ is a subset of $\mathbb{R}_+$ and is compact. A continuous function $F : \mathscr{X} \longrightarrow \mathbb{R}_+$ is called submodular continuous function if, for all $\mathbf{x}, \mathbf{y} \in \mathscr{X}$, the following inequality

$$F(\mathbf{x}) + F(\mathbf{y}) \geq F(\mathbf{x} \vee \mathbf{y}) + F(\mathbf{x} \wedge \mathbf{y}) \tag{2}$$

holds, where $\mathbf{x} \vee \mathbf{y} := \max\{\mathbf{x}, \mathbf{y}\}$ (coordinate-wise) and $\mathbf{x} \wedge \mathbf{y} := \min\{\mathbf{x}, \mathbf{y}\}$ (coordinate-wise). Moreover, if $\mathbf{x} \preceq \mathbf{y}$, we have $F(\mathbf{x}) \leq F(\mathbf{y})$ for all $\mathbf{x}, \mathbf{y} \in \mathscr{X}$, and then the submodular continuous function F is called monotone on $\mathscr{X}$. Furthermore, a differentiable submodular continuous function F is called *DR-submodular* if, for all $\mathbf{x}, \mathbf{y} \in \mathscr{X}$ such that $\mathbf{x} \preceq \mathbf{y}$, we have $\nabla F(\mathbf{x}) \succeq \nabla F(\mathbf{y})$; i.e., $\nabla F(\cdot)$ is an antitone mapping [29]. When the submodular continuous function F is twice differentiable, the submodular F is submodular if and only if all off-diagonal components of its Hessian matrix are nonpositive [28]; i.e., for all $\mathbf{x} \in \mathscr{X}$,

$$\frac{\partial^2 F(\mathbf{x})}{\partial x_i \partial x_j} \leq 0, \quad \forall i \neq j. \tag{3}$$

Furthermore, if the submodular function F is DR-submodular, then all second-derivatives are nonpositive [29]; i.e., for all $\mathbf{x} \in \mathscr{X}$,

$$\frac{\partial^2 F(\mathbf{x})}{\partial x_i \partial x_j} \leq 0. \tag{4}$$

In addition, the twice differentiability implies that the submodular F is smooth [50]. Moreover, we say that a submodular function F is β-smooth if, $\forall \mathbf{x}, \mathbf{y} \in \mathscr{X}$, we have

$$F(\mathbf{y}) \leq F(\mathbf{x}) + \langle \nabla F(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\beta}{2} \|\mathbf{y} - \mathbf{x}\|^2. \tag{5}$$

Note that the above definition is equivalent to

$$\|\nabla F(\mathbf{x}) - \nabla F(\mathbf{y})\| \leq \beta \|\mathbf{x} - \mathbf{y}\|. \tag{6}$$

Furthermore, a function F is called *weakly DR-submodular* function with parameter γ if

$$\gamma := \inf_{\mathbf{x}, \mathbf{y} \in \mathscr{X}, \mathbf{x} \preceq \mathbf{y}} \inf_{i \in \{1, \dots, n\}} \frac{[\nabla F(\mathbf{x})]_i}{[\nabla F(\mathbf{y})]_i}. \tag{7}$$

More details about weak DR-submodular functions are available in [29].

3. Problem Formulation and Algorithm Design

In this section, we first describe the problem of our interest, and then we design an algorithm to efficiently solve the problem.

In this paper, we focus on the following constrained optimization problem:

$$\max_{\mathbf{x}\in\mathscr{K}} F(\mathbf{x}) := \mathbb{E}_{\xi\sim\mathscr{D}}\left[F_{\xi}(\mathbf{x})\right], \tag{8}$$

where $\mathscr{K}$ denotes the constraint set, $\mathscr{D}$ denotes an unknown distribution, F_ξ is a submodular continuous function for all $\xi \in \mathscr{D}$. Moreover, we assume that the constraint set, $\mathscr{K} = \prod_{i=1}^{n}\mathscr{K}_i \subseteq \mathscr{X}$, is convex, where each $\mathscr{K}_i \subseteq \mathbb{R}_+$ is convex and closed set for all $i = 1,\dots,n$. The problem has recently been introduced in [30]. In addition, we use the notation F^* to denote the optimal value of $F(\mathbf{x})$ for all $\mathbf{x} \in \mathscr{K}$, i.e., $F^* := \max_{\mathbf{x}\in\mathscr{K}}F(\mathbf{x})$. Furthermore, we can see that the function $F(\mathbf{x})$ is submodular function because each function F_ξ is submodular continuous function for all $\xi \in \mathscr{D}$ [28].

To solve problem (8), the projected stochastic gradient methods are a class of efficient algorithms [31]. However, we focus on the case that the decision vectors $\mathbf{x}$ are high-dimensional in this work; i.e., the dimensionality of vectors n is large. The full gradient computations are prohibitive expensive and become computational bottleneck. Therefore, we propose a stochastic block-coordinate gradient method by combining the great features of block-coordinate and stochastic gradient. We assume that the components of decision variables are arbitrarily chosen but fixed for each processor. Furthermore, at each iteration, each processor randomly chooses a subset of (stochastic) gradients, rather than all the (stochastic) gradients. The detailed description of the proposed algorithm is as follows. Starting from an initial value $\mathbf{x} \in \mathscr{K}$, for $t = 0,1,2,\dots$, each $i = 1,\dots,n$ updates its decision variable as

$$x_i(t+1) = \Pi_{\mathscr{K}_i}\left(x_i(t) + q_i(t)\,\alpha(t)\,g_i(t)\right), \tag{9}$$

where $\alpha(t)$ is the step-size, $\Pi_{\mathscr{K}_i}(x)$ denotes the Euclidean projection of x on the set $\mathscr{K}_i$, $q_i(t)$ are independent and identically Bernoulli random variables with $\mathbb{P}(q_i(t) = 1) = p_i$ for all $t = 0,1,2,\dots$ and $i = 1,\dots,n$, and $g_i(t)$ denotes the unbiased estimate of the gradient $\nabla_i F(\mathbf{x}(t))$, which denotes the i-th coordinate in $\nabla F(\mathbf{x}(t))$.

We introduce the following matrix.

$$Q(t) := \operatorname{diag}\{q_1(t),\dots,q_n(t)\} \tag{10}$$

Therefore, we can write relation (9) more compactly as

$$\mathbf{x}(t+1) = \Pi_{\mathscr{K}}\left(\mathbf{x}(t) + \alpha(t)\,Q(t)\,\mathbf{g}(t)\right), \tag{11}$$

where $\mathbf{x}(t) := (x_i(t),\dots,x_n(t))^{\top} \in \mathbb{R}_+^n$, and $\mathbf{g}(t) := (g_1(t),\dots,g_n(t))^{\top} \in \mathbb{R}^n$. Note that the i-th coordinate of $\mathbf{g}(t)$ is missing when $q_i(t) = 0$ at each iteration t, and then the i-th coordinate of $\mathbf{x}(t)$ is not updated. Therefore, a random subset of $\mathbf{x} \in \mathbb{R}_+^n$ is updated at each iteration t. In addition, we use the notation P to denote a diagonal matrix with size $n\times n$; i.e., $P := \operatorname{diag}\{P_{11},\dots,P_{ii},\dots,P_{nn}\}$, where $P_{ii} = p_i$.

Let $\mathscr{F}_t$ denote the history information of all random variables generated by the proposed algorithm (11) up to time t. In this paper, we adopt the following assumption on the random variables $q_i(t)$, which is stated as follows.

Assumption 1. For all i, j, t, s, the random variables $q_i(t)$ and $q_j(s)$ are independent of each other. Furthermore, the random variables $\{q_i(t)\}$ are independent of $\mathscr{F}_{t-1}$ and $\mathbf{g}(t)$ for any decision variables $\mathbf{x} \in \mathscr{F}_{t-1}$.

In addition, we assume that the function F and the sets $\mathscr{K}_i$ satisfy the following conditions.

Assumption 2. Assume that the following properties hold:

(a) The constraint set $\mathscr{K} \subseteq \mathscr{X}$ is convex, and each set $\mathscr{K}_i \subseteq \mathbb{R}_+$ is convex and closed for all $i \in \{1,\dots,n\}$.

(b) The function $F : \mathscr{X} \longrightarrow \mathbb{R}_+$ is monotone and weakly DR-submodular with parameter γ over $\mathscr{X}$.

(c) The function F is differentiable and β-smooth with respect to norm $\|\cdot\|$.

Next, we make the following assumption about stochastic oracle $\mathbf{g}(t)$.

Assumption 3. Assume that the stochastic oracle $\mathbf{g}(t)$ satisfies the following conditions:

$$\mathbb{E}\left[\mathbf{g}(t)\right] = \nabla F(\mathbf{x}(t)) \tag{12}$$

and

$$\mathbb{E}\left[\left\|\mathbf{g}(t) - \nabla F(\mathbf{x}(t))\right\|^2\right] \le \mu^2. \tag{13}$$

The above assumption implies that the stochastic oracle $\mathbf{g}(t)$ is an unbiased estimate of $\nabla F(\mathbf{x}(t))$.

In this section, we first formulate an optimization problem, and then design an optimization method to solve it. Moreover, we also give some standard assumptions to analyze the performance of the proposed method.

4. Main Results

In this section, We first provide the performance of convergence. To this end, we first introduce the definition of a stationary point, which is defined as in [31].

Definition 4. For a vector $\mathbf{x} \in \mathscr{K}$ and a function $F : \mathscr{X} \longrightarrow \mathbb{R}_+$, if $\max_{\mathbf{y}\in\mathscr{K}}\langle \mathbf{y}-\mathbf{x}, \nabla F(\mathbf{x})\rangle \le 0$, then $\mathbf{x}$ is a stationary point of F over $\mathscr{K} \subset \mathscr{X}$.

From Definition 4, the convergence of our proposed algorithm is given in the following theorem.

Theorem 5. *Let Assumptions 1–3 hold. Assume that the set of stationary points is nonempty and $0 < \alpha(t) < 2/\beta$. Moreover, the sequence $\{\mathbf{x}(t)\}$ is generated by the stochastic block-coordinate gradient projection algorithm (11). Then, the iterative sequence $\{\mathbf{x}(t)\}$ converges to some stationary point $\overline{\mathbf{x}} \in \mathscr{K}$ with probability 1.*

The proof can be found in the next section. The above result shows that the iterations converge to some local maximum with probability 1.

Furthermore, when the function $F : \mathcal{X} \longrightarrow \mathbb{R}_+$ is differentiable and DR-submodular, we have the following result.

Theorem 6. *Let Assumptions 1–3 hold. Moreover, assume $\gamma = 1$ in (7) and $\alpha(t) = 1/(\beta + \sqrt{t})$. The sequence $\{\mathbf{x}(t)\}$ is generated by the stochastic block-coordinate gradient projection algorithm (11). Furthermore, the random decision variable $\mathbf{x}(\tau)$ is picked by choosing $\mathbf{x}(1)$, $\mathbf{x}(T)$ with probability $1/2(T-1)$ and the other variables with probability $1/T$. Then, for any random variable for $\tau \in \{1, \ldots, T\}$, we have*

$$\mathbb{E}\left[F(\mathbf{x}(\tau))\right] \geq \frac{p_{\min}}{2} F^* - \frac{1}{2}\left(\frac{\delta^2 \beta}{2T} + \frac{\delta^2}{2\sqrt{T}} + \frac{\mu^2}{\sqrt{T}}\right), \quad (14)$$

where $\delta^2 := \sup_{x,y \in \mathcal{K}} \|\mathbf{x} - \mathbf{y}\|^2$, $p_{\min} := \min_{i \in \{1,\ldots,n\}} p_i$.

The proof can be found in the next section. From the above result, we can see that an objective value in expectation can be obtained after $O(\delta^2\beta/\epsilon + (\delta^2 + \mu^2)/\epsilon^2)$ iterations of the stochastic block-coordinate gradient projection algorithm (11) for any initial value. Moreover, the objective value is at least $(p_{\min}/2)F^* - \epsilon$ for any DR-submodular function.

In addition, when the function $F : \mathcal{X} \longrightarrow \mathbb{R}_+$ is weakly DR-submodular function with parameter γ, we also yield the following result.

Theorem 7. *Let Assumptions 1–3 hold. The sequence $\{\mathbf{x}(t)\}$ is generated by the stochastic block-coordinate gradient projection algorithm (11) with $\alpha(t) = 1/(\beta + \sqrt{t})$. Furthermore, the random decision variable $\mathbf{x}(\tau)$ is picked by choosing $\mathbf{x}(t)$ in $\{\mathbf{x}(1), \ldots, \mathbf{x}(T)\}$ with probability $1/T$. Then, for any for $\tau \in \{1, \ldots, T\}$, we have*

$$\begin{aligned}\mathbb{E}\left[F(\mathbf{x}(\tau))\right] \geq {} & \frac{\gamma^2}{1+\gamma^2}\left(p_{\min} - \frac{1}{\gamma T}\right) F^* \\ & - \frac{\gamma}{1+\gamma^2}\left(\frac{\delta^2\left(\beta + \sqrt{T}\right)}{2T} + \frac{\mu^2}{\sqrt{T}}\right),\end{aligned} \quad (15)$$

where $p_{\min} := \min_{i \in \{1,\ldots,n\}} p_i$, $\delta^2 := \sup_{x,y \in \mathcal{K}} \|\mathbf{x} - \mathbf{y}\|^2$.

The proof can be found in the next section. Note that the stochastic block-coordinate gradient projection algorithm yields an objective value after $O(\delta^2\beta/\epsilon + (\delta^2 + \mu^2)/\epsilon^2)$ iterations from any initial value. Furthermore, the expectation of the objective value is in at least $(\gamma^2/(1+\gamma^2))p_{\min}F^*$ for any weakly DR-submodular function.

5. Performance Analysis

In this section, the detailed proofs of main results are provided. We first analyze the convergence performance of the stochastic block-coordinate gradient projection algorithm.

Proof of Theorem 5. By the Projection Theorem [32], we have

$$\left\langle \mathbf{x} - \Pi_{\mathcal{K}}(\mathbf{x}), \mathbf{z} - \Pi_{\mathcal{K}}(\mathbf{x}) \right\rangle \leq 0 \quad (16)$$

for all $\mathbf{z} \in \mathcal{K}$. Therefore, let $\mathbf{z} = \mathbf{x}(t)$ and $\mathbf{x} = \mathbf{x}(t) + \alpha(t)Q(t)\mathbf{g}(t)$ in inequality (16); we obtain

$$\begin{aligned}&\left\langle \mathbf{x}(t+1) - \mathbf{x}(t) - \alpha(t) Q(t)\mathbf{g}(t), \mathbf{x}(t+1) - \mathbf{x}(t)\right\rangle \\ &\quad \leq 0,\end{aligned} \quad (17)$$

where we have used relation (11). By simple algebraic manipulations, we yield

$$\begin{aligned}&\|\mathbf{x}(t+1) - \mathbf{x}(t)\|^2 \\ &\quad \leq \left\langle \mathbf{x}(t+1) - \mathbf{x}(t), \alpha(t) Q(t)\mathbf{g}(t)\right\rangle.\end{aligned} \quad (18)$$

Furthermore, when $q_i(t) = 0$ for any $i \in \{1, \ldots, n\}$, $x_i(t+1) = x_i(t)$ at each iteration $t \geq 0$. Therefore, we have

$$\left\langle \mathbf{x}(t+1) - \mathbf{x}(t), \alpha(t)(Q(t) - I)\mathbf{g}(t)\right\rangle = 0. \quad (19)$$

From the above relation, we also obtain

$$\begin{aligned}&\left\langle \mathbf{x}(t+1) - \mathbf{x}(t), Q(t)\mathbf{g}(t)\right\rangle \\ &\quad = \left\langle \mathbf{x}(t+1) - \mathbf{x}(t), \mathbf{g}(t)\right\rangle.\end{aligned} \quad (20)$$

Plugging relation (20) into inequality (18), we have

$$\|\mathbf{x}(t+1) - \mathbf{x}(t)\|^2 \leq \left\langle \mathbf{x}(t+1) - \mathbf{x}(t), \alpha(t)\mathbf{g}(t)\right\rangle. \quad (21)$$

Taking conditional expectation in (21), we have

$$\begin{aligned}&\mathbb{E}\left[\|\mathbf{x}(t+1) - \mathbf{x}(t)\|^2 \mid \mathcal{F}_t\right] \\ &\quad \leq \alpha(t)(\nabla F(\mathbf{x}(t)))^{\top} \times \mathbb{E}\left[\mathbf{x}(t+1) - \mathbf{x}(t) \mid \mathcal{F}_t\right],\end{aligned} \quad (22)$$

where we have used $\mathbb{E}[\mathbf{g}(t)] = \nabla F(\mathbf{x}(t))$ in the last inequality. In addition, since the function F is β-smooth, we have

$$\begin{aligned}F(\mathbf{x}(t+1)) \geq {} & F(\mathbf{x}(t)) \\ & + \langle \mathbf{x}(t+1) - \mathbf{x}(t), \nabla F(\mathbf{x}(t))\rangle \\ & - \frac{\beta}{2}\|\mathbf{x}(t+1) - \mathbf{x}(t)\|^2.\end{aligned} \quad (23)$$

Taking conditional expectation on $\mathcal{F}_t$ in (23) and using relation (22), we obtain

$$\begin{aligned}&\mathbb{E}\left[F(\mathbf{x}(t+1)) \mid \mathcal{F}_t\right] \\ &\quad \geq F(\mathbf{x}(t)) \\ &\qquad + \left(\frac{1}{\alpha(t)} - \frac{\beta}{2}\right)\mathbb{E}\left[\|\mathbf{x}(t+1) - \mathbf{x}(t)\|^2 \mid \mathcal{F}_t\right]\end{aligned} \quad (24)$$

for step-size $\alpha(t) \in (0, 2/\beta)$. For brevity, let $\kappa := 1/\alpha(t) - \beta/2$. Inequality (24) implies that

$$\begin{aligned}&F^* - \mathbb{E}\left[F(\mathbf{x}(t+1)) \mid \mathcal{F}_t\right] \\ &\quad \leq F^* - F(\mathbf{x}(t))\kappa \cdot \mathbb{E}\left[\|\mathbf{x}(t+1) - \mathbf{x}(t)\|^2 \mid \mathcal{F}_t\right].\end{aligned} \quad (25)$$

From the definition of F^*, we have $F^* \geq F(\mathbf{x}(t))$ for $t \geq 0$; i.e., the sequence of random variables $\{F^* - F(\mathbf{x}(t))\}$

is nonnegative for all $t \geq 0$. Therefore, according to the *Supermartingale Convergence Theorem* [51], we can see that the sequence $\{F(\mathbf{x}(t))\}$ is convergent with probability 1. Furthermore, we also have

$$\sum_{t=0}^{\infty} \mathbb{E}\left[\|\mathbf{x}(t+1)-\mathbf{x}(t)\|^2 \mid \mathcal{F}_t\right] < \infty \tag{26}$$

with probability 1. From relation (11), inequality (26) implies that

$$\sum_{t=0}^{\infty} \left\|\Pi_{\mathcal{K}}\left(\mathbf{x}(t)+\alpha(t)\nabla F(\mathbf{x}(t))\right)-\mathbf{x}(t)\right\|_P^2 < \infty \tag{27}$$

with probability 1, where $P \coloneqq \operatorname{diag}\{p_1, p_2, \ldots, p_n\}$. Therefore, we obtain that

$$\lim_{t\longrightarrow\infty} \left(\Pi_{\mathcal{K}}\left(\mathbf{x}(t)+\alpha(t)\nabla F(\mathbf{x}(t))\right)-\mathbf{x}(t)\right) = 0 \tag{28}$$

with probability 1. Thus, there exists a subsequence $\{\mathbf{x}(t_\ell)\}$, which converges to $\overline{\mathbf{x}}$. Then, we have

$$\lim_{\ell\longrightarrow\infty} \Pi_{\mathcal{K}}\left(\mathbf{x}(t_\ell)+\alpha(t)\nabla F\left(\mathbf{x}(t_\ell)\right)\right) = \overline{\mathbf{x}}. \tag{29}$$

Since the gradient projection operation is continuous, we have

$$\Pi_{\mathcal{K}}\left(\overline{\mathbf{x}}+\alpha(t)\nabla F(\overline{\mathbf{x}})\right) = \overline{\mathbf{x}} \tag{30}$$

with probability 1. The above relation implies that

$$\Pi_{\mathcal{K}-\{\overline{\mathbf{x}}\}}\left(\alpha(t)\nabla F(\overline{\mathbf{x}})\right) = 0 \tag{31}$$

with probability 1. Then, relation (31) implies that $\max_{\mathbf{z}\in\mathcal{K}}\langle\mathbf{z}-\overline{\mathbf{x}}, \nabla F(\overline{\mathbf{x}})\rangle \leq 0$. Therefore, $\overline{\mathbf{x}}$ is a stationary point of $F(\cdot)$ over $\mathcal{K}$ with probability 1. The statement of the theorem is completely proved. □

To prove Theorem 6, we first present the following lemmas. The first lemma follows from [52], which is stated as follows.

Lemma 8. *For all $\mathbf{z} \in \mathcal{K}$, we have*

$$\left\langle \mathbf{x}-\Pi_{\mathcal{K}}(\mathbf{x}), \mathbf{z}-\Pi_{\mathcal{K}}(\mathbf{x})\right\rangle_P \leq 0 \tag{32}$$

for any diagonal matrix P.

The next lemma is due to [53], which is stated as follows.

Lemma 9. *Assume that a function F is submodular and monotone. Then, we have*

$$\left\langle \mathbf{x}-\mathbf{y}, \nabla F(\mathbf{x})\right\rangle \leq 2F(\mathbf{x}) - F(\mathbf{x}\vee\mathbf{y}) - F(\mathbf{x}\wedge\mathbf{y}) \tag{33}$$

for any points $\mathbf{x}, \mathbf{y} \in \mathcal{K}$.

In addition, we also have the following lemma.

Lemma 10. *Let Assumptions 1–3 hold. The iterative sequence $\{\mathbf{x}(t)\}$ is generated by the stochastic block-coordinate gradient projection algorithm (11) with step-size $\alpha(t) \leq \sigma(t)/(1+\beta\sigma(t))$, where $\sigma(t) > 0$ for $t \geq 0$. Then, we have*

$$\begin{aligned} &F(\mathbf{x}(t+1)) \\ &\quad\geq F(\mathbf{x}(t)) - \frac{\sigma(t)}{2}\left\|\nabla F(\mathbf{x}(t))-\mathbf{g}(t)\right\|^2 \\ &\qquad + p_{\min}\left\langle \mathbf{x}(t+1)-\mathbf{x}(t), Q(t)\mathbf{g}(t)\right\rangle_{P^{-1}} \\ &\qquad - \frac{p_{\min}}{2}\left(\beta+\frac{1}{\sigma(t)}\right)\|\mathbf{x}(t+1)-\mathbf{x}(t)\|_{P^{-1}}^2, \end{aligned} \tag{34}$$

where $p_{\min} \coloneqq \min_{i\in\{1,\ldots,n\}} p_i$.

Proof. In inequality (21), we let $Q(t) = I_k$, where I_k denotes the diagonal matrix with the k-th entry equal to 1 and the other entries equal to 0. Then, we have

$$\begin{aligned} &\frac{1}{\alpha(t)}\left|\Pi_{\mathcal{K}_k}\left(x_k(t)+\alpha(t)g_k(t)\right)-x_k(t)\right|^2 \\ &\quad - \left(\Pi_{\mathcal{K}_k}\left(x_k(t)+\alpha(t)g_k(t)\right)-x_k(t)\right)g_k(t) \\ &\leq 0. \end{aligned} \tag{35}$$

Furthermore, the above relation implies that

$$\begin{aligned} &\left(\Pi_{\mathcal{K}_k}\left(x_k(t)+\alpha(t)g_k(t)\right)-x_k(t)\right)g_k(t) \\ &\quad - \frac{1}{2\alpha(t)}\left|\Pi_{\mathcal{K}_k}\left(x_k(t)+\alpha(t)g_k(t)\right)-x_k(t)\right|^2 \\ &\geq 0. \end{aligned} \tag{36}$$

Therefore, for any k, we obtain

$$\begin{aligned} &\left(\Pi_{\mathcal{K}_k}\left(x_k(t)+\alpha(t)g_k(t)\right)-x_k(t)\right)g_k(t) \\ &\quad - \frac{1}{2\alpha(t)}\left|\Pi_{\mathcal{K}_k}\left(x_k(t)+\alpha(t)g_k(t)\right)-x_k(t)\right|^2 \\ &\geq \frac{p_{\min}}{p_k}\left(\Pi_{\mathcal{K}_k}\left(x_k(t)+\alpha(t)g_k(t)\right)-x_k(t)\right)g_k(t) \\ &\quad - \frac{p_{\min}}{2\alpha(t)p_k}\left|\Pi_{\mathcal{K}_k}\left(x_k(t)+\alpha(t)g_k(t)\right)-x_k(t)\right|^2 \end{aligned} \tag{37}$$

where in the last inequality we have used $p_{\min}/p_k \leq 1$. Since $0 \leq p_i \leq 1$ for all $i \in \{1,\ldots,n\}$, setting $\alpha(t) = 1/\beta$ and following from relation (23), we have

$$\begin{aligned} &F(\mathbf{x}(t+1)) - F(\mathbf{x}(t)) \\ &\quad\geq \left\langle \mathbf{x}(t+1)-\mathbf{x}(t), \nabla F(\mathbf{x}(t))\right\rangle \\ &\qquad - \frac{\beta}{2}\|\mathbf{x}(t+1)-\mathbf{x}(t)\|^2 \\ &\quad= \left\langle \mathbf{x}(t+1)-\mathbf{x}(t), \nabla F(\mathbf{x}(t))-\mathbf{g}(t)\right\rangle \\ &\qquad + \left\langle \mathbf{x}(t+1)-\mathbf{x}(t), \mathbf{g}(t)\right\rangle \end{aligned}$$

$$
\begin{aligned}
&- \frac{\beta}{2} \|\mathbf{x}(t+1) - \mathbf{x}(t)\|^2 \\
&\geq p_{\min} \langle \mathbf{x}(t+1) - \mathbf{x}(t), Q(t)\,\mathbf{g}(t) \rangle_{P^{-1}} \\
&- \frac{p_{\min}}{2} \left(\beta + \frac{1}{\sigma(t)} \right) \|\mathbf{x}(t+1) - \mathbf{x}(t)\|_{P^{-1}}^2 \\
&- \frac{\sigma(t)}{2} \left\| \nabla F(\mathbf{x}(t)) - \mathbf{g}(t) \right\|^2,
\end{aligned} \tag{38}
$$

where the last inequality is obtained by using inequality (37), Young's inequality, and the fact that $x_i(t+1) = x_i(t)$ when $q_i(t) = 0$ for all $i \in \{1, \ldots, n\}$. Moreover, $0 \leq p_i \leq 1$ for all $i \in \{1, \ldots, n\}$. Rearranging the terms in (38), the lemma is obtained completely. □

With Lemmas 8 and 10 in places, we have the following result.

Lemma 11. *Let Assumptions 1–3 hold. The iterative sequence $\{\mathbf{x}(t)\}$ is generated by the stochastic block-coordinate gradient projection algorithm (11) with step-size $\alpha(t) \leq \sigma(t)/(1+\beta\sigma(t))$. Then, for all $t \geq 0$, we have*

$$
\begin{aligned}
\mathbb{E}\left[\langle \mathbf{x}(t) - \mathbf{v}, \mathbf{g}(t) \rangle \mid \mathcal{F}_t \right] &\geq \frac{1}{p_{\min}} F(\mathbf{x}(t)) - \frac{1}{p_{\min}} \\
&\cdot \mathbb{E}\left[F(\mathbf{x}(t+1)) \mid \mathcal{F}_t \right] - \frac{\sigma(t)}{2 p_{\min}} \\
&\cdot \mathbb{E}\left[\left\| \nabla F(\mathbf{x}(t)) - \mathbf{g}(t) \right\|^2 \mid \mathcal{F}_t \right] + \frac{1}{2\alpha(t)} \\
&\cdot \mathbb{E}\left[\|\mathbf{x}(t+1) - \mathbf{v}\|_{P^{-1}}^2 - \|\mathbf{x}(t) - \mathbf{v}\|_{P^{-1}}^2 \mid \mathcal{F}_t \right],
\end{aligned} \tag{39}
$$

where $p_{\min} = \min_{i \in \{1,\ldots,n\}} p_i$.

Proof. From the result in Lemma 10, we have

$$
\begin{aligned}
&F(\mathbf{x}(t+1)) \\
&\quad \geq F(\mathbf{x}(t)) - \frac{\sigma(t)}{2} \left\| \nabla F(\mathbf{x}(t)) - \mathbf{g}(t) \right\|^2 \\
&\quad + p_{\min} \langle \mathbf{v} - \mathbf{x}(t), Q(t)\,\mathbf{g}(t) \rangle_{P^{-1}} \\
&\quad + p_{\min} \langle \mathbf{x}(t+1) - \mathbf{v}, Q(t)\,\mathbf{g}(t) \rangle_{P^{-1}} \\
&\quad - \frac{p_{\min}}{2} \left(\beta + \frac{1}{\sigma(t)} \right) \|\mathbf{x}(t+1) - \mathbf{x}(t)\|_{P^{-1}}^2 .
\end{aligned} \tag{40}
$$

In addition, following on from Lemma 8, we also obtain

$$
\begin{aligned}
&\langle \mathbf{x}(t) + \alpha(t)\,Q(t)\,\mathbf{g}(t) - \mathbf{x}(t+1), \mathbf{v} - \mathbf{x}(t+1) \rangle_{P^{-1}} \\
&\quad \leq 0,
\end{aligned} \tag{41}
$$

which implies that

$$
\begin{aligned}
&\langle \mathbf{x}(t+1) - \mathbf{v}, Q(t)\,\mathbf{g}(t) \rangle_{P^{-1}} \\
&\quad \geq \frac{1}{\alpha(t)} \langle \mathbf{x}(t+1) - \mathbf{v}, \mathbf{x}(t+1) - \mathbf{x}(t) \rangle_{P^{-1}} .
\end{aligned} \tag{42}
$$

Combining inequalities (40) and (42), we yield

$$
\begin{aligned}
&F(\mathbf{x}(t+1)) \\
&\quad \geq F(\mathbf{x}(t)) - \frac{\sigma(t)}{2} \left\| \nabla F(\mathbf{x}(t)) - \mathbf{g}(t) \right\|^2 \\
&\qquad + p_{\min} \langle \mathbf{v} - \mathbf{x}(t), Q(t)\,\mathbf{g}(t) \rangle_{P^{-1}} \\
&\qquad + \frac{p_{\min}}{\alpha(t)} \langle \mathbf{x}(t+1) - \mathbf{v}, \mathbf{x}(t+1) - \mathbf{x}(t) \rangle_{P^{-1}} \\
&\qquad - \frac{p_{\min}}{2} \left(\beta + \frac{1}{\sigma(t)} \right) \|\mathbf{x}(t+1) - \mathbf{x}(t)\|_{P^{-1}}^2 . \\
&\quad = F(\mathbf{x}(t)) - \frac{\sigma(t)}{2} \left\| \nabla F(\mathbf{x}(t)) - \mathbf{g}(t) \right\|^2 \\
&\qquad + p_{\min} \langle \mathbf{v} - \mathbf{x}(t), Q(t)\,\mathbf{g}(t) \rangle_{P^{-1}} \\
&\qquad + \frac{p_{\min}}{2\alpha(t)} \left(\|\mathbf{x}(t+1) - \mathbf{v}\|_{P^{-1}}^2 - \|\mathbf{x}(t) - \mathbf{v}\|_{P^{-1}}^2 \right) \\
&\qquad + \frac{p_{\min}}{2} \left(\frac{1}{\alpha(t)} - \beta - \frac{1}{\sigma(t)} \right) \|\mathbf{x}(t+1) - \mathbf{x}(t)\|_{P^{-1}}^2 \\
&\quad \geq F(\mathbf{x}(t)) - \frac{\sigma(t)}{2} \left\| \nabla F(\mathbf{x}(t)) - \mathbf{g}(t) \right\|^2 \\
&\qquad + p_{\min} \langle \mathbf{v} - \mathbf{x}(t), Q(t)\,\mathbf{g}(t) \rangle_{P^{-1}} \\
&\qquad + \frac{p_{\min}}{2\alpha(t)} \left(\|\mathbf{x}(t+1) - \mathbf{v}\|_{P^{-1}}^2 - \|\mathbf{x}(t) - \mathbf{v}\|_{P^{-1}}^2 \right),
\end{aligned} \tag{43}
$$

where the last inequality is due to $\alpha(t) \leq \sigma(t)/(1+\beta\sigma(t))$. Taking conditional expectation of the above inequality on $\mathcal{F}_t$, we yield

$$
\begin{aligned}
\mathbb{E}\left[F(\mathbf{x}(t+1)) \mid \mathcal{F}_t \right] &\geq \mathbb{E}\left[F(\mathbf{x}(t)) \mid \mathcal{F}_t \right] - \frac{\sigma(t)}{2} \\
&\cdot \mathbb{E}\left[\left\| \nabla F(\mathbf{x}(t)) - \mathbf{g}(t) \right\|^2 \mid \mathcal{F}_t \right] \\
&+ p_{\min} \mathbb{E}\left[\langle \mathbf{v} - \mathbf{x}(t), \mathbf{g}(t) \rangle \mid \mathcal{F}_t \right] + \frac{p_{\min}}{2\alpha(t)} \\
&\cdot \mathbb{E}\left[\|\mathbf{x}(t+1) - \mathbf{v}\|_{P^{-1}}^2 - \|\mathbf{x}(t) - \mathbf{v}\|_{P^{-1}}^2 \mid \mathcal{F}_t \right].
\end{aligned} \tag{44}
$$

Thus, by some algebraic manipulations, inequality (39) is obtained. □

Next, we start to prove Theorem 6.

Proof of Theorem 6. Setting $\mathbf{v} = \mathbf{x}^*$ in Lemma 11, where $\mathbf{x}^*$ is the globally optimal solution for problem (8), i.e., $\mathbf{x}^* := \arg\max_{\mathbf{x} \in \mathcal{K}} F(\mathbf{x})$, we have

$$
\begin{aligned}
\mathbb{E}\left[\langle \mathbf{x}(t) - \mathbf{x}^*, \mathbf{g}(t) \rangle \mid \mathcal{F}_t \right] &\geq \frac{1}{p_{\min}} F(\mathbf{x}(t)) - \frac{1}{p_{\min}} \\
&\cdot \mathbb{E}\left[F(\mathbf{x}(t+1)) \mid \mathcal{F}_t \right] - \frac{\sigma(t)}{2 p_{\min}}
\end{aligned}
$$

$$\cdot \mathbb{E}\left[\left\|\nabla F(\mathbf{x}(t))-\mathbf{g}(t)\right\|^{2} \mid \mathscr{F}_{t}\right]+\frac{1}{2\alpha(t)} \cdot \mathbb{E}\left[\left\|\mathbf{x}(t+1)-\mathbf{x}^{*}\right\|_{P^{-1}}^{2}-\left\|\mathbf{x}(t)-\mathbf{x}^{*}\right\|_{P^{-1}}^{2} \mid \mathscr{F}_{t}\right]. \tag{45}$$

Since

$$\begin{aligned}\left\langle\mathbf{x}(t)-\mathbf{x}^{*}, \mathbf{g}(t)\right\rangle &= \left\langle\mathbf{x}(t)-\mathbf{x}^{*}, \nabla F(\mathbf{x}(t))\right\rangle \\ &+\left\langle\mathbf{x}(t)-\mathbf{x}^{*}, \mathbf{g}(t)-\nabla F(\mathbf{x}(t))\right\rangle,\end{aligned} \tag{46}$$

taking conditional expectation of (46) with respect to $\mathscr{F}_t$, we have

$$\begin{aligned}&\mathbb{E}\left[\left\langle\mathbf{x}(t)-\mathbf{x}^{*}, \nabla F(\mathbf{x}(t))\right\rangle \mid \mathscr{F}_{t}\right] \\ &= \mathbb{E}\left[\left\langle\mathbf{x}(t)-\mathbf{x}^{*}, \mathbf{g}(t)\right\rangle \mid \mathscr{F}_{t}\right] \\ &+\mathbb{E}\left[\left\langle\mathbf{x}(t)-\mathbf{x}^{*}, \mathbf{g}(t)-\nabla F(\mathbf{x}(t))\right\rangle \mid \mathscr{F}_{t}\right] \\ &\geq \mathbb{E}\left[\left\langle\mathbf{x}(t)-\mathbf{x}^{*}, \mathbf{g}(t)-\nabla F(\mathbf{x}(t))\right\rangle \mid \mathscr{F}_{t}\right]+\frac{1}{p_{\min}} \\ &\cdot F(\mathbf{x}(t))-\frac{1}{p_{\min}} \mathbb{E}\left[F(\mathbf{x}(t+1)) \mid \mathscr{F}_{t}\right]-\frac{\sigma(t)}{2 p_{\min}} \\ &\cdot \mathbb{E}\left[\left\|\nabla F(\mathbf{x}(t))-\mathbf{g}(t)\right\|^{2} \mid \mathscr{F}_{t}\right]+\frac{1}{2\alpha(t)} \\ &\cdot \mathbb{E}\left[\left\|\mathbf{x}(t+1)-\mathbf{x}^{*}\right\|_{P^{-1}}^{2}-\left\|\mathbf{x}(t)-\mathbf{x}^{*}\right\|_{P^{-1}}^{2} \mid \mathscr{F}_{t}\right],\end{aligned} \tag{47}$$

which implies that

$$\begin{aligned}&\mathbb{E}\left[F(\mathbf{x}(t+1)) \mid \mathscr{F}_{t}\right] \geq F(\mathbf{x}(t)) \\ &-p_{\min} \mathbb{E}\left[\left\langle\mathbf{x}(t)-\mathbf{x}^{*}, \nabla F(\mathbf{x}(t))\right\rangle \mid \mathscr{F}_{t}\right] \\ &+p_{\min} \mathbb{E}\left[\left\langle\mathbf{x}(t)-\mathbf{x}^{*}, \mathbf{g}(t)-\nabla F(\mathbf{x}(t))\right\rangle \mid \mathscr{F}_{t}\right] \\ &-\frac{\sigma(t)}{2} \mathbb{E}\left[\left\|\nabla F(\mathbf{x}(t))-\mathbf{g}(t)\right\|^{2} \mid \mathscr{F}_{t}\right]+\frac{p_{\min}}{2\alpha(t)} \\ &\cdot \mathbb{E}\left[\left\|\mathbf{x}(t+1)-\mathbf{x}^{*}\right\|_{P^{-1}}^{2}-\left\|\mathbf{x}(t)-\mathbf{x}^{*}\right\|_{P^{-1}}^{2} \mid \mathscr{F}_{t}\right].\end{aligned} \tag{48}$$

Setting $\mathbf{x}=\mathbf{x}(t)$ and $\mathbf{y}=\mathbf{x}^*$ in Lemma 9 and taking condition expectation on $\mathscr{F}_t$, we obtain

$$\begin{aligned}&\mathbb{E}\left[\left\langle\mathbf{x}(t)-\mathbf{x}^{*}, \nabla F(\mathbf{x}(t))\right\rangle \mid \mathscr{F}_{t}\right] \\ &\leq 2\mathbb{E}\left[F(\mathbf{x}(t)) \mid \mathscr{F}_{t}\right]-\mathbb{E}\left[\mathbf{x}(t) \vee \mathbf{x}^{*} \mid \mathscr{F}_{t}\right] \\ &-\mathbb{E}\left[\mathbf{x}(t) \wedge \mathbf{x}^{*} \mid \mathscr{F}_{t}\right].\end{aligned} \tag{49}$$

Thus, plugging inequality (49) into relation (48), we get

$$\begin{aligned}&\mathbb{E}\left[F(\mathbf{x}(t+1)) \mid \mathscr{F}_{t}\right]+\mathbb{E}\left[F(\mathbf{x}(t)) \mid \mathscr{F}_{t}\right] \\ &-p_{\min} F\left(\mathbf{x}^{*}\right) \\ &\geq p_{\min} \mathbb{E}\left[\left\langle\mathbf{x}(t)-\mathbf{x}^{*}, \mathbf{g}(t)-\nabla F(\mathbf{x}(t))\right\rangle \mid \mathscr{F}_{t}\right] \\ &-\frac{\sigma(t)}{2} \mathbb{E}\left[\left\|\nabla F(\mathbf{x}(t))-\mathbf{g}(t)\right\|^{2} \mid \mathscr{F}_{t}\right]+\frac{p_{\min}}{2\alpha(t)} \\ &\cdot \mathbb{E}\left[\left\|\mathbf{x}(t+1)-\mathbf{x}^{*}\right\|_{P^{-1}}^{2}-\left\|\mathbf{x}(t)-\mathbf{x}^{*}\right\|_{P^{-1}}^{2} \mid \mathscr{F}_{t}\right].\end{aligned} \tag{50}$$

Taking expectation in (50) and using some algebraic manipulations, we have

$$\begin{aligned}&\mathbb{E}[F(\mathbf{x}(t+1))]+\mathbb{E}[F(\mathbf{x}(t))]-p_{\min} F\left(\mathbf{x}^{*}\right) \\ &\geq \frac{p_{\min}}{2\alpha(t)} \mathbb{E}\left[\left\|\mathbf{x}(t+1)-\mathbf{x}^{*}\right\|_{P^{-1}}^{2}-\left\|\mathbf{x}(t)-\mathbf{x}^{*}\right\|_{P^{-1}}^{2}\right] \\ &-\frac{\sigma(t)}{2} \mathbb{E}\left[\left\|\nabla F(\mathbf{x}(t))-\mathbf{g}(t)\right\|^{2}\right] \\ &\geq \frac{p_{\min}}{2\alpha(t)} \mathbb{E}\left[\left\|\mathbf{x}(t+1)-\mathbf{x}^{*}\right\|_{P^{-1}}^{2}-\left\|\mathbf{x}(t)-\mathbf{x}^{*}\right\|_{P^{-1}}^{2}\right] \\ &-\frac{\mu^{2}}{2} \sigma(t),\end{aligned} \tag{51}$$

where we have used the relation $\mathbb{E}[\mathbf{g}(t)] = \nabla F(\mathbf{x}(t))$ to obtain the first inequality. Summing both sides of (51) for $t=1,\ldots,T$, we obtain

$$\begin{aligned}&\sum_{t=1}^{T}\left(\mathbb{E}[F(\mathbf{x}(t+1))]+\mathbb{E}[F(\mathbf{x}(t))]-p_{\min} F\left(\mathbf{x}^{*}\right)\right) \\ &\geq \frac{p_{\min}}{2\alpha(T)} \mathbb{E}\left[\left\|\mathbf{x}(T+1)-\mathbf{x}^{*}\right\|_{P^{-1}}^{2}\right]-\frac{p_{\min}}{2\alpha(1)} \\ &\cdot \mathbb{E}\left[\left\|\mathbf{x}(1)-\mathbf{x}^{*}\right\|_{P^{-1}}^{2}\right]-\frac{\mu^{2}}{2} \sum_{t=1}^{T} \sigma(t)+\frac{p_{\min}}{2} \\ &\cdot \sum_{t=1}^{T-1}\left(\frac{1}{\alpha(t)}-\frac{1}{\alpha(t+1)}\right) \mathbb{E}\left[\left\|\mathbf{x}(t+1)-\mathbf{x}^{*}\right\|_{P^{-1}}^{2}\right] \\ &\geq-\frac{\delta^{2}}{2\alpha(1)}+\frac{\delta^{2}}{2} \sum_{t=1}^{T-1}\left(\frac{1}{\alpha(t)}-\frac{1}{\alpha(t+1)}\right)-\frac{\mu^{2}}{2} \\ &\cdot \sum_{t=1}^{T} \sigma(t) \geq-\frac{\delta^{2}}{2\alpha(T)}-\frac{\mu^{2}}{2} \sum_{t=1}^{T} \sigma(t),\end{aligned} \tag{52}$$

where in the last inequality we have used the fact that $\mathbb{E}[\|\mathbf{x}(t)-\mathbf{x}^*\|_{P^{-1}}^2] \leq \delta^2/p_{\min}$ and $1/\alpha(t)-1/\alpha(t+1) \leq 0$ for all $t=1,\ldots,T$. On the other hand, we also have

$$\begin{aligned}&\sum_{t=1}^{T}\left(2\mathbb{E}[F(\mathbf{x}(t))]-p_{\min} F^{*}\right)=\mathbb{E}[F(\mathbf{x}(1))] \\ &-\mathbb{E}[F(\mathbf{x}(T+1))] \\ &+\sum_{t=1}^{T}\left(\mathbb{E}[F(\mathbf{x}(t+1))]+\mathbb{E}[F(\mathbf{x}(t))]-p_{\min} F\left(\mathbf{x}^{*}\right)\right) \\ &\geq-\frac{\delta^{2}}{2\alpha(T)}-\frac{\mu^{2}}{2} \sum_{t=1}^{T} \sigma(t)-\mathbb{E}[F(\mathbf{x}(T+1))],\end{aligned} \tag{53}$$

where $F^* = \max_{\mathbf{x}\in\mathscr{K}} F(\mathbf{x})$ and the last inequality is due to (52). Since $\sigma(t) = 1/\sqrt{t}$, we have

$$\sum_{t=1}^{T} \sigma(t)=\sum_{t=1}^{T} \frac{1}{\sqrt{t}} \leq \int_{0}^{T} \frac{1}{\sqrt{t}} d t=2 \sqrt{T}. \tag{54}$$

Plugging the above inequality into (53) and dividing both sides by $2T$,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[F\left(\mathbf{x}\left(t\right)\right)\right]+\frac{1}{2T}\mathbb{E}\left[F\left(\mathbf{x}\left(T+1\right)\right)\right] \geq \frac{p_{\min}}{2}F^{*}-\frac{\delta^{2}}{4T}\left(\beta+\sqrt{T}\right)-\frac{\mu^{2}}{2\sqrt{T}}, \tag{55}$$

where we have used the fact that $\alpha(T)=1/(\beta+\sqrt{T})$ in the last inequality. Furthermore, the above inequality implies that

$$\begin{aligned}&\frac{1}{T}\sum_{t=2}^{T}\mathbb{E}\left[F\left(\mathbf{x}\left(t\right)\right)\right]\\&\quad+\frac{1}{2T}\left(\mathbb{E}\left[F\left(\mathbf{x}\left(T+1\right)\right)\right]+\mathbb{E}\left[F\left(\mathbf{x}\left(1\right)\right)\right]\right)\\&\quad\geq\frac{p_{\min}}{2}F^{*}-\frac{\delta^{2}}{4T}\left(\beta+\sqrt{T}\right)-\frac{\mu^{2}}{2\sqrt{T}}.\end{aligned} \tag{56}$$

In addition, the sample $\mathbf{x}(\tau)$ is obtained for $\tau\in\{1,\ldots,T\}$ by choosing $\mathbf{x}(1),\mathbf{x}(T+1)$ with probability $1/(2T)$ and the other decision vectors with probability $1/T$; we have

$$\mathbb{E}\left[F\left(\mathbf{x}\left(\tau\right)\right)\right]\geq\frac{p_{\min}}{2}F^{*}-\frac{\delta^{2}}{4T}\left(\beta+\sqrt{T}\right)-\frac{\mu^{2}}{2\sqrt{T}}. \tag{57}$$

Therefore, the theorem is completely proved. □

We now start to prove Theorem 7.

Proof of Theorem 7. From the definition of weakly DR-submodular function, for any $\mathbf{x},\mathbf{y}\in\mathscr{K}$, we obtain

$$\nabla F\left(\mathbf{x}\right)\succeq\gamma\nabla F\left(\mathbf{y}\right) \tag{58}$$

for all $\mathbf{x}\preceq\mathbf{y}$. Recall that the following relation is from [31]; i.e., for any $\mathbf{x},\mathbf{y}\in\mathscr{K}$,

$$\frac{1}{\gamma}\left\langle\mathbf{x}-\mathbf{y},\nabla F\left(\mathbf{x}\right)\right\rangle\leq\left(1-\frac{1}{\gamma^{2}}\right)F\left(\mathbf{x}\right)-F\left(\mathbf{y}\right). \tag{59}$$

Setting $\mathbf{x}=\mathbf{x}(t)$ and $\mathbf{y}=\mathbf{x}^{*}$ in the above inequality, then, following from (48) and taking conditional expectation, we obtain

$$\begin{aligned}&\mathbb{E}\left[F\left(\mathbf{x}\left(t+1\right)\right)\mid\mathscr{F}_{t}\right]+\left(\gamma+\frac{1}{\gamma}-1\right)\\&\quad\cdot\mathbb{E}\left[F\left(\mathbf{x}\left(t\right)\right)\mid\mathscr{F}_{t}\right]-p_{\min}\gamma F\left(\mathbf{x}^{*}\right)\\&\quad\geq p_{\min}\mathbb{E}\left[\left\langle\mathbf{x}\left(t\right)-\mathbf{x}^{*},\mathbf{g}\left(t\right)-\nabla F\left(\mathbf{x}\left(t\right)\right)\right\rangle\mid\mathscr{F}_{t}\right]\\&\quad-\frac{\sigma\left(t\right)}{2}\mathbb{E}\left[\left\|\nabla F\left(\mathbf{x}\left(t\right)\right)-\mathbf{g}\left(t\right)\right\|^{2}\mid\mathscr{F}_{t}\right]+\frac{p_{\min}}{2\alpha\left(t\right)}\\&\quad\cdot\mathbb{E}\left[\left\|\mathbf{x}\left(t+1\right)-\mathbf{x}^{*}\right\|_{P^{-1}}^{2}-\left\|\mathbf{x}\left(t\right)-\mathbf{x}^{*}\right\|_{P^{-1}}^{2}\mid\mathscr{F}_{t}\right].\end{aligned} \tag{60}$$

Taking expectation in (60) with respect to $\mathscr{F}_{t}$, we get

$$\begin{aligned}&\mathbb{E}\left[F\left(\mathbf{x}\left(t+1\right)\right)\right]+\left(\gamma+\frac{1}{\gamma}-1\right)\mathbb{E}\left[F\left(\mathbf{x}\left(t\right)\right)\right]\\&\quad-p_{\min}\gamma F\left(\mathbf{x}^{*}\right)\\&\quad\geq\frac{p_{\min}}{2\alpha\left(t\right)}\mathbb{E}\left[\left\|\mathbf{x}\left(t+1\right)-\mathbf{x}^{*}\right\|_{P^{-1}}^{2}-\left\|\mathbf{x}\left(t\right)-\mathbf{x}^{*}\right\|_{P^{-1}}^{2}\right]\\&\quad-\frac{\sigma\left(t\right)}{2}\mathbb{E}\left[\left\|\nabla F\left(\mathbf{x}\left(t\right)\right)-\mathbf{g}\left(t\right)\right\|^{2}\right]\\&\quad\geq\frac{p_{\min}}{2\alpha\left(t\right)}\mathbb{E}\left[\left\|\mathbf{x}\left(t+1\right)-\mathbf{x}^{*}\right\|_{P^{-1}}^{2}-\left\|\mathbf{x}\left(t\right)-\mathbf{x}^{*}\right\|_{P^{-1}}^{2}\right]\\&\quad-\frac{\mu^{2}}{2}\sigma\left(t\right),\end{aligned} \tag{61}$$

where the last inequality is due to Assumption 3.

Adding the above inequalities for $t=1,\ldots,T$, we have

$$\begin{aligned}&\sum_{t=1}^{T}\left(\mathbb{E}\left[F\left(\mathbf{x}\left(t+1\right)\right)\right]+\left(\gamma+\frac{1}{\gamma}-1\right)\mathbb{E}\left[\mathbf{x}\left(t\right)\right]\right.\\&\quad\left.-p_{\min}\gamma F\left(\mathbf{x}^{*}\right)\right)\geq\frac{p_{\min}}{2\alpha\left(T\right)}\mathbb{E}\left[\left\|\mathbf{x}\left(T+1\right)-\mathbf{x}^{*}\right\|_{P^{-1}}^{2}\right]\\&\quad-\frac{p_{\min}}{2\alpha\left(1\right)}\mathbb{E}\left[\left\|\mathbf{x}\left(1\right)-\mathbf{x}^{*}\right\|_{P^{-1}}^{2}\right]-\frac{\mu^{2}}{2}\sum_{t=1}^{T}\sigma\left(t\right)+\frac{p_{\min}}{2}\\&\quad\cdot\sum_{t=1}^{T-1}\left(\frac{1}{\alpha\left(t\right)}-\frac{1}{\alpha\left(t+1\right)}\right)\mathbb{E}\left[\left\|\mathbf{x}\left(t+1\right)-\mathbf{x}^{*}\right\|_{P^{-1}}^{2}\right]\\&\quad\geq-\frac{\delta^{2}}{2\alpha\left(1\right)}+\frac{\delta^{2}}{2}\sum_{t=1}^{T-1}\left(\frac{1}{\alpha\left(t\right)}-\frac{1}{\alpha\left(t+1\right)}\right)-\frac{\mu^{2}}{2}\\&\quad\cdot\sum_{t=1}^{T}\sigma\left(t\right)\geq-\frac{\delta^{2}}{2\alpha\left(T\right)}-\frac{\mu^{2}}{2}\sum_{t=1}^{T}\sigma\left(t\right),\end{aligned} \tag{62}$$

where the last inequality follows from $\mathbb{E}[\|\mathbf{x}(t)-\mathbf{x}^{*}\|_{P^{-1}}^{2}]\leq\delta^{2}/p_{\min}$ and $1/\alpha(t)-1/\alpha(t+1)\leq 0$ for all $t=1,\ldots,T$. Moreover, since $\sigma(t)=1/\sqrt{t}$ and $\alpha(t)=1/(\beta+1/\sigma(t))$, we obtain

$$\begin{aligned}&\sum_{t=1}^{T}\left(\mathbb{E}\left[F\left(\mathbf{x}\left(t+1\right)\right)\right]+\left(\gamma+\frac{1}{\gamma}-1\right)\mathbb{E}\left[\mathbf{x}\left(t\right)\right]\right.\\&\quad\left.-p_{\min}\gamma F\left(\mathbf{x}^{*}\right)\right)\geq-\frac{\delta^{2}}{2}\left(\beta+\sqrt{T}\right)-\mu^{2}\sqrt{T},\end{aligned} \tag{63}$$

where we have used inequality (54) to obtain the last inequality. On the other hand, we have

$$\begin{aligned}&\sum_{t=1}^{T}\left[\left(\gamma+\frac{1}{\gamma}\right)\mathbb{E}\left[F\left(\mathbf{x}\left(t\right)\right)\right]-p_{\min}\gamma F^{*}\right]\\&\quad=\mathbb{E}\left[F\left(\mathbf{x}\left(1\right)\right)\right]+\sum_{t=1}^{T}\mathbb{E}\left[F\left(\mathbf{x}\left(t+1\right)\right)\right]\end{aligned}$$

$$-\mathbb{E}\left[F\left(\mathbf{x}\left(T+1\right)\right)\right] + \sum_{t=1}^{T}\left[\left(\gamma+\frac{1}{\gamma}-1\right)\mathbb{E}\left[\mathbf{x}\left(t\right)\right]-p_{\min}\gamma F\left(\mathbf{x}^{*}\right)\right]. \tag{64}$$

Plugging inequality (63) into equality (64), we get

$$\begin{aligned}&\sum_{t=1}^{T}\left[\left(\gamma+\frac{1}{\gamma}\right)\mathbb{E}\left[F\left(\mathbf{x}\left(t\right)\right)\right]-p_{\min}\gamma F^{*}\right]\\&\quad\geq-\frac{\delta^{2}}{2}\left(\beta+\sqrt{T}\right)-\mu^{2}\sqrt{T}-F^{*}.\end{aligned} \tag{65}$$

Dividing both sides by $(\gamma+1/\gamma)T$ in (65), we have

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[F\left(\mathbf{x}\left(t\right)\right)\right]\\&\quad\geq\frac{\gamma^{2}p_{\min}}{1+\gamma^{2}}F^{*}\\&\quad-\frac{\gamma}{1+\gamma^{2}}\left(\frac{\delta^{2}\left(\beta+\sqrt{T}\right)}{2T}+\frac{\mu^{2}}{\sqrt{T}}+\frac{F^{*}}{T}\right).\end{aligned} \tag{66}$$

In addition, we obtain the sample $\mathbf{x}(\tau)$ by choosing $\mathbf{x}(t)$ with probability $1/T$. Then, for any $\tau\in\{1,\ldots,T\}$, we have

$$\begin{aligned}&\mathbb{E}\left[F\left(\mathbf{x}\left(\tau\right)\right)\right]\\&\quad\geq\frac{\gamma^{2}p_{\min}}{1+\gamma^{2}}F^{*}\\&\quad-\frac{\gamma}{1+\gamma^{2}}\left(\frac{\delta^{2}\left(\beta+\sqrt{T}\right)}{2T}+\frac{\mu^{2}}{\sqrt{T}}+\frac{F^{*}}{T}\right).\end{aligned} \tag{67}$$

Therefore, the theorem is obtained completely. □

In this section, we proved the main results of the paper in detail. The conclusion of this paper is provided in the next section.

6. Conclusion

In this paper, we have considered a stochastic optimization problem of continuous submodular functions, which is an important problem in many areas such as machine learning and social science. Since the data is high-dimensional, usual algorithms based on the computation of the whole approximate gradient vector, such as stochastic gradient methods, are prohibitive. For this reason, we proposed the stochastic block-coordinate gradient projection algorithm for maximizing submodular functions, which randomly chooses a subset of the approximate gradient vector. Moreover, we studied the convergence performance of the proposed algorithm. We proved that the iterations converge to some stationary points with probability 1 by using the suitable step sizes. Furthermore, we showed that the algorithm achieves a tight $((p_{\min}/2)F^{*}-\epsilon)$ approximation guarantee after $O(1/\epsilon^{2})$ when the submodular functions are DR-submodular and the suitable step sizes are used. More generally, we also showed that the algorithm achieves the tight $((\gamma^{2}/(1+\gamma^{2}))p_{\min}F^{*}-\epsilon)$ after $O(1/\epsilon^{2})$ iterations when the submodular functions are weakly DR-submodular with parameter γ and the appropriate step sizes are used.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grants no. U1604155, no. 61602155, no. 61871430, no. 61772477, and no. 61572445; in part by Henan Science and Technology Innovation Project under Grant no. 174100510010; in part by the basic research projects at the University of Henan Province under Grant no. 19zx010; in part by the Ph.D. Research Fund of the Zhengzhou University of Light Industry; and in part by the Natural Science Foundation of Henan Province under Grant no. 162300410322.

References

[1] J. Djolonga and A. Krause, "From MAP to marginals: Variational inference in Bayesian submodular models," in *Proceedings of the 28th Annual Conference on Neural Information Processing Systems 2014, NIPS 2014*, pp. 244–252, Canada, December 2014.

[2] R. Iyer and J. Bilmes, "Submodular point processes with applications to machine learning," in *Proceedings of the in Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, pp. 388–397, 2015.

[3] A. Singla, I. Bogunovic, A. Karbasi, and A. Krause, "Near-optimally teaching the crowd to classify," in *Proceedings of the 31st International Conference on Machine Learning*, pp. 154–162, 2014.

[4] B. Kim, R. Khanna, and O. Koyejo, "Examples are not enough, learn to criticize! Criticism for interpretability," in *Proceedings of the 30th Annual Conference on Neural Information Processing Systems, NIPS 2016*, pp. 2288–2296, Spain, December 2016.

[5] V. Cevher and A. Krause, "Greedy dictionary selection for sparse representation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 5, pp. 979–988, 2011.

[6] B. Mirzasoleiman, A. Karbasi, R. Sarkar, and A. Krause, "Distributed submodular maximization: Identifying representative elements in massive data," in *Proceedings of the 27th Annual Conference on Neural Information Processing Systems, NIPS 2013*, USA, December 2013.

[7] H. Lin and J. Bilmes, "A class of submodular functions for document summarization," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, ACL-HLT 2011*, pp. 510–520, USA, June 2011.

[8] Y. Yue and C. Guestrin, "Linear submodular bandits and their application to diversified retrieval," in *Advances in Neural Information Processing Systems*, pp. 2483–2491, 2011.

[9] K. El-Arini, G. Veda, D. Shahaf, and C. Guestrin, "Turning down the noise in the blogosphere," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '09*, pp. 289–297, France, July 2009.

[10] B. Mirzasoleiman, A. Badanidiyuru, and A. Karbasi, "FAst coNsTrained submodular maximization: Personalized data summarization," in *Proceedings of the 33rd International Conference on Machine Learning, ICML 2016*, pp. 2042–2054, USA, June 2016.

[11] C. Guestrin, A. Krause, and A. P. Singh, "Near-optimal sensor placements in gaussian processes," in *Proceedings of the ICML 2005: 22nd International Conference on Machine Learning*, pp. 265–272, August 2005.

[12] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. Vanbriesen, and N. Glance, "Cost-effective outbreak detection in networks," in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '07)*, pp. 420–429, New York, NY, USA, August 2007.

[13] R. Gomez, J. Leskovec, and A. Krause, "Inferring networks of diffusion and influence," *ACM Transactions on Knowledge Discovery from Data*, vol. 5, no. 4, pp. 1–37, 2012.

[14] F. Bach, "Structured sparsity-including norms through submodular functions," in *Advances in Neural Information Processing Systems*, pp. 118–126, 2010.

[15] M. Narasimhan, N. Jojic, and J. Bilmes, "Q-clustering," in *Proceedings of the 2005 Annual Conference on Neural Information Processing Systems, NIPS 2005*, pp. 979–986, Canada, December 2005.

[16] A. Das and D. Kempe, "Submodular meets spectral: Greedy algorithms for subset selection, sparse approximation and dictionary selection," in *Proceedings of the 28th International Conference on Machine Learning, ICML 2011*, pp. 1057–1064, USA, July 2011.

[17] D. Golovin and A. Krause, "Adaptive submodularity: theory and applications in active learning and stochastic optimization," *Journal of Artificial Intelligence Research*, vol. 42, pp. 427–486, 2011.

[18] Z. Zheng and N. B. Shroff, "Submodular utility maximization for deadline constrained data collection in sensor networks," *Institute of Electrical and Electronics Engineers Transactions on Automatic Control*, vol. 59, no. 9, pp. 2400–2412, 2014.

[19] S. Iwata, L. Fleischer, and S. Fujishige, "A combinatorial strongly polynomial algorithm for minimizing submodular functions," *Journal of the ACM*, vol. 48, no. 4, pp. 761–777, 2001.

[20] J. B. Orlin, "A faster strongly polynomial time algorithm for submodular function minimization," *Mathematical Programming*, vol. 118, no. 2, Ser. A, pp. 237–251, 2009.

[21] A. Schrijver, "A combinatorial algorithm minimizing submodular functions in strongly polynomial time," *Journal of Combinatorial Theory, Series B*, vol. 80, no. 2, pp. 346–355, 2000.

[22] G. Cornuejols, M. L. Fisher, and G. L. Nemhauser, "Location of bank accounts to optimize float: an analytic study of exact and approximate algorithms," *Management Science*, vol. 23, no. 8, pp. 789–810, 1977.

[23] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functions—I," *Mathematical Programming*, vol. 14, no. 1, pp. 265–294, 1978.

[24] M. L. Fisher, G. L. Nemhauser, and L. A. Wolsey, "An analysis of approximations for maximizing submodular set functions. II," *Mathematical Programming*, no. 8, pp. 73–87, 1978.

[25] U. Feige, V. S. Mirrokni, and J. Vondrák, "Maximizing non-monotone submodular functions," *SIAM Journal on Computing*, vol. 40, no. 4, pp. 1133–1153, 2011.

[26] C. Chekuri, J. Vondrák, and R. Zenklusen, "Submodular function maximization via the multilinear relaxation and contention resolution schemes," *SIAM Journal on Computing*, vol. 43, no. 6, pp. 1831–1879, 2014.

[27] G. Calinescu, C. Chekuri, M. Pál, and J. Vondrák, "Maximizing a monotone submodular function subject to a matroid constraint," *SIAM Journal on Computing*, vol. 40, no. 6, pp. 1740–1766, 2011.

[28] F. Bach, "Submodular functions: from discrete to continuous domains," https://arxiv.org/abs/1511.00394v2.

[29] A. Bian, B. Mirzasoleiman, J. Buhmann, and A. Krause, "Guaranteed non-convex optimization: Submodular maximization over continuous domains," https://arxiv.org/abs/1606.05615.

[30] M. Karimi, M. Lucic, H. Hassani, and A. Krause, "Stochastic submodular maximization: The case of convergence functions," in *Advances in Neural Information Processing Systems*, pp. 6853–6863, 2017.

[31] H. Hassani, M. Soltanolkotabi, and A. Karbasi, "Gradient methods for submodular maximization," https://arxiv.org/abs/1708.03949.

[32] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, Belmont, Mass, USA, 2nd edition, 1999.

[33] K.-W. Chang, C.-J. Hsieh, and C.-J. Lin, "Coordinate descent method for large-scale L2-loss linear support vector machines," *Journal of Machine Learning Research (JMLR)*, vol. 9, pp. 1369–1398, 2008.

[34] S. J. Wright, "Accelerated block-coordinate relaxation for regularized optimization," *SIAM Journal on Optimization*, vol. 22, no. 1, pp. 159–186, 2012.

[35] P. Tseng and S. Yun, "A coordinate gradient descent method for nonsmooth separable minimization," *Mathematical Programming—Series B*, vol. 117, no. 1, pp. 387–423, 2009.

[36] P. Tseng and S. Yun, "Block-coordinate gradient descent method for linearly constrained nonsmooth separable optimization," *Journal of Optimization Theory and Applications*, vol. 140, no. 3, pp. 513–535, 2009.

[37] S. Yun and K.-C. Toh, "A coordinate gradient descent method for l_1-regularized convex minimization," *Computational Optimization and Applications*, vol. 48, no. 2, pp. 273–307, 2011.

[38] A. A. Canutescu and R. L. Dunbrack Jr., "Cyclic coordinate descent: A robotics algorithm for protein loop closure," *Protein Science*, vol. 12, no. 5, pp. 963–972, 2003.

[39] T. T. Wu and K. Lange, "Coordinate descent algorithms for lasso penalized regression," *The Annals of Applied Statistics*, vol. 2, no. 1, pp. 224–244, 2008.

[40] Y. Li and S. Osher, "Coordinate descent optimization for l^1 minimization with application to compressed sensing; a greedy algorithm," *Inverse Problems and Imaging*, vol. 3, no. 3, pp. 487–503, 2009.

[41] Z. Q. Luo and P. Tseng, "On the convergence of the coordinate descent method for convex differentiable minimization," *Journal of Optimization Theory and Applications*, vol. 72, no. 1, pp. 7–35, 1992.

[42] P. Tseng, "Convergence of a block coordinate descent method for nondifferentiable minimization," *Journal of Optimization Theory and Applications*, vol. 109, no. 3, pp. 475–494, 2001.

[43] P.-W. Wang and C.-J. Lin, "Iteration complexity of feasible descent methods for convex optimization," *Journal of Machine Learning Research (JMLR)*, vol. 15, pp. 1523–1548, 2014.

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

[45] P. Richtárik and M. Takác, "Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function," *Mathematical Programming*, vol. 144, no. 1-2, Ser. A, pp. 1–38, 2014.

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

[47] J. Liu, S. J. Wright, C. RÉ, V. Bittorf, and S. Sridhar, "An asynchronous parallel stochastic coordinate descent algorithm," *Journal of Machine Learning Research (JMLR)*, vol. 16, pp. 285–322, 2015.

[48] J. Liu and S. J. Wright, "Asynchronous stochastic coordinate descent: parallelism and convergence properties," *SIAM Journal on Optimization*, vol. 25, no. 1, pp. 351–376, 2015.

[49] S. Fujishige, *Submodular functions and optimization*, vol. 2nd of *Annals of Discrete Mathematics*, North-Holland Publishing, Amsterdam, Netherlands, 2005.

[50] C. Chekuri, T. S. Jayram, and J. Vondrak, "On multiplicative weight updates for concave and submodular function maximization," in *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*, pp. 201–210, ACM.

[51] B. T. Polyak, *Introduction to Optimization*, Optimization Software, New York, NY, USA, 1987.

[52] C. Singh, A. Nedic, and R. Srikant, "Random block-coordinate gradient projection algorithms," in *Proceedings of the 2014 53rd IEEE Annual Conference on Decision and Control, CDC 2014*, pp. 185–190, USA, December 2014.

[53] C. Chekuri, J. Vondrak, and R. Zenklusen, "Submodular function maximization via the multilinear relaxation and contention resolution schemes," in *Proceedings of the Forty-third Annual ACM Symposium on Theory of Computing*, pp. 783–792, ACM, 2011.