STOCHASTIC BLOCK MIRROR DESCENT METHODS FOR NONSMOOTH AND STOCHASTIC OPTIMIZATION *

CONG D. DANG † AND GUANGHUI LAN ‡

Abstract. In this paper, we present a new stochastic algorithm, namely the stochastic block mirror descent (SBMD) method for solving large-scale nonsmooth and stochastic optimization problems. The basic idea of this algorithm is to incorporate the block-coordinate decomposition and an incremental block averaging scheme into the classic (stochastic) mirror-descent method, in order to significantly reduce the cost per iteration of the latter algorithm. We establish the rate of convergence of the SBMD method along with its associated large-deviation results for solving general nonsmooth and stochastic optimizaiton problems. We also introduce different variants of this method and establish their rate of convergence for solving strongly convex, smooth, and composite optimization problems, as well as certain nonconvex optimization problems. To the best of our knowledge, all these developments related to the SBMD methods are new in the stochastic optimization literature. Moreover, some of our results also seem to be new for block coordinate descent methods for deterministic optimization.

Keywords. Stochastic Optimization, Mirror Descent, Block Coordinate Descent, Nonsmooth Optimization, Stochastic Composite Optimization, Metric Learning

AMS subject classifications. 62L20, 90C25, 90C15, 68Q25

1. Introduction.

The basic problem of interest in the paper is the stochastic programming (SP) problem given by

$$f^* := \min_{x\in X}\{f(x) := \mathbb{E}[F(x,\xi)]\}. \tag{1.1}$$

Here $X \in \mathbb{R}^n$ is a closed convex set, ξ is a random variable with support $\Xi \subseteq \mathbb{R}^d$ and $F(\cdot,\xi) : X \to \mathbb{R}$ is continuous for every $\xi \in \Xi$. In addition, we assume that X has a block structure, i.e.,

$$X = X_1 \times X_2 \times \cdots \times X_b, \tag{1.2}$$

where $X_i \subseteq \mathbb{R}^{n_i}$, $i = 1,\ldots,b$, are closed convex sets with $n_1 + n_2 + \ldots + n_b = n$.

The last few years have seen a resurgence of interest in the block coordinate descent (BCD) method for solving problems with X given in the form of (1.2). In comparision with regular first-order methods, each iteration of these methods updates only one block of variables. In particular, if each block consists of only one variable (i.e., $n_i = 1, i = 1,\ldots,b$), then the BCD method becomes the simplest coordinate descent (CD) method. Although simple, these methods are found to be effective in solving huge-scale problems with n as big as 10^8-10^{12} (see, e.g., [28, 19, 29, 33, 4]), and hence are very useful for dealing with high-dimensional problems, especially those from large-scale data analysis applications. While earlier studies on the BCD method were focused on their asymptotical convergence behaviour (see, e.g., [22, 38] and also [39, 40]), much recent effort has been directed to the complexity analysis of these types of methods (see [28, 19, 35, 33, 4]. In particular, Nesterov [28] was among the first (see also Leventhal and Lewis [19], and Shalev-Shwartz and Tewari [35]) to analyze the iteration complexity of a randomized BCD method for minimizing smooth convex functions. More recently, the BCD methods were further enhanced by Richtárik and Takáč [33], Beck and Tetruashvili [4], Lu and Xiao [21], etc. We refer to [33] for an excellent review on the earlier developments of BCD methods.

However, to the best of our knowledge, most current BCD methods were designed for solving determinisitc optimization problems. One possible approach for solving problem (1.1), based on existing BCD methods and the sample average approximation (SAA) [36], can be described as follows. For a given set of i.i.d. samples (dataset) $\xi_k, k = 1,\ldots,N$, of ξ, we first approximate $f(\cdot)$ in (1.1) by $\tilde{f}(x) := \frac{1}{N}\sum_{k=1}^N F(x,\xi_k)$ and then apply the BCD methods to $\min_{x\in X}\tilde{f}(x)$. Since ξ_k, $k = 1,\ldots,N$, are fixed a priori, by recursively updating the (sub)gradient of $\tilde{f}$ (see [28, 29]), the iteration cost of the BCD method can be considerably smaller than that

*September, 2013. This research was partially supported by NSF grants CMMI-1000347, CMMI-1254446, DMS-1319050, and ONR grant N00014-13-1-0036.

†Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL, 32611. (email: congdd@ufl.edu).

‡Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL, 32611. (email: glan@ise.ufl.edu).

of the gradient descent methods. However, the above SAA approach is also known for the following drawbacks: a) the high memory requirement to store ξ_k, $k = 1, \ldots, N$; b) the high dependence (at least linear) of the iteration cost on the sample size N, which can be expensive when dealing with large datasets; and c) the difficulty to apply the approach to the on-line setting where one needs to update the solution whenever a new piece of data ξ_k is collected.

A different approach to solve problem (1.1) is called stochastic approximation (SA), which was initially proposed by Robbins and Monro [34] in 1950s for solving strongly convex SP problems (see also [31, 32]). The SA method has also attracted much interest recently (see, e.g., [24, 14, 30, 7, 8, 20, 9, 23, 10, 11, 13, 27, 41]). In particular, Nemirovski et. al. [24] presented a properly modified SA approach, namely, the mirror descent SA for solving general nonsmooth convex SP problems. Lan [14] introduced a unified optimal SA method for smooth, nonsmooth and stochastic optimization (see also [7, 8] for a more general framework). Ghadimi and Lan [9] presented novel SA methods for nonconvex optimization (see also [10]). Related methods, based on dual averaging, have been studied in [11, 13, 27, 41]. Note that all these SA algorithms only need to access one single ξ_k at each iteration, and hence does not require much memory. In addition, their iteration cost is independent of the sample size N. However, since these algorithms need to update the whole vector x at each iteration, their iteration cost can strongly depend on n unless the problem is very sparse (see, e.g., [30]). In addition, it is unclear whether the SA methods can benefit from the recursive updating as in the BCD methods, since the samples ξ_k used in different iterations are supposed to be independent.

Our main goal in this paper is to present a new class of stochastic methods, referered to as the stochastic block mirror descent (SBMD) methods, by incorporating the aforementioned block-coordinate decomposition into the classic (stochastic) mirror descent method ([25, 3, 24]). Our study has been mainly motivated by solving an important class of SP problems with $F(x, \xi) = \psi(Bx, \xi)$, where B is a certain linear operator and ψ is a relatively simple function. These problems arise from many machine learning applications, where ψ is a loss function and B denotes a certain basis (or dictionary) obtained by, e.g., metric learning (e.g., [42]). Each iteration of existing SA methods would require $\mathcal{O}(n^2)$ arithmetic operations to compute Bx and becomes prohibitive if n exceeds 10^6. On the other hand, by using block-coordinate decomposition with $n_i = 1$, the iteration cost of the SBMD algorithms can be significantly reduced to $\mathcal{O}(n)$, which can be further reduced if B and ξ_k are sparse (see Subsection 2.1 for more discussions). Our development has also been motivated by the situation when the bottleneck of the mirror descent method exists in the projection (or prox-mapping) subproblems (see (2.5)). In this case, we can also significantly reduce the iteration cost by using the block-coordinate decomposition, since each iteration of the SBMD method requires only one projection over X_i for some $1 \le i \le b$, while the mirror descent method needs to perform the projectons over X_i for all $1 \le i \le b$.

Our contribution in this paper mainly lies in the following aspects. Firstly, we introduce the block decompsition into the classic mirror descent method for solving general nonsmooth optimization problems. Each iteration of this algorithm updates one block of the search point along a stochastic (sub)gradient $G_{i_k}(x_k, \xi_k)$. Here, the index i_k is randomly chosen and $G(x, \xi)$ is an unbiased estimator of the subgradient of $f(\cdot)$, i.e.,

$$\mathbb{E}[G(x, \xi)] = g(x) \in \partial f(x), \quad \forall x \in X. \tag{1.3}$$

In addition, in order to compute the output of the algorithm, we introduce an *incremental block averaging* scheme, which updates only one block of the weighted sum of the search points in each iteration. We demonstrate that if $f(\cdot)$ is a general nonsmooth convex function, then the number of iterations performed by the SBMD method to find an ϵ-solution of (1.1), i.e., a point $\bar{x} \in X$ such that (s.t.) $\mathbb{E}[f(\bar{x}) - f^*] \le \epsilon$, can be bounded by $\mathcal{O}(b/\epsilon^2)$. Here the expectation is taken w.r.t. the random elements $\{i_k\}$ and $\{\xi_k\}$. In addition, if $f(\cdot)$ is strongly convex, then the number of iterations performed by the SBMD method (with a different stepsize policy and averaging scheme) to find an ϵ-solution of (1.1) can be bounded by $\mathcal{O}(b/\epsilon)$. We also derive the large-deviation results associated with these rates of convergence for the SBMD algorithm. Secondly, we consider a special class of convex stochastic composite optimization problems given by

$$\phi^* := \min_{x \in X} \{\phi(x) := f(x) + \chi(x)\}. \tag{1.4}$$

Here $\chi(\cdot)$ is a relatively simple convex function and $f(\cdot)$ defined in (1.1) is a smooth convex function with

Lipschitz-continuous gradients $g(\cdot)$. We show that, by properly modifying the SBMD method, we can significantly improve the aforementioned complexity bounds in terms of their dependence on the Lipchsitz constants of $g(\cdot)$. We show that the complexity bounds can be further improved if $f(\cdot)$ is strongly convex. Thirdly, we generalize our study to a class of nonconvex stochastic composite optimization problems in the form of (1.4), but with $f(\cdot)$ being possibly nonconvex. Instead of using the aforementioned incremental block averaging, we incorporate a certain randomization scheme to compute the output of the algorithm. We also establish the complexity of this algorithm to generate an approximate stationary point for solving problem (1.4).

While this paper focuses on stochastic optimization, it is worth noting that some of our results also seem to be new in the literature for the BCD methods for deterministic optimization. Firstly, currently the only BCD-type methods for solving general nonsmooth CP problems are based on the subgradient methods without involving averaging, e.g., those by Polak and a constrained version by Shor (see Nesterov [29]). Our development shows that it is possible to develop new BCD type methods involving different averaging schemes for convex optimization. Secondly, the large-deviation result for the BCD methods for general nonsmooth problems and the $\mathcal{O}(b/\epsilon)$ complexity result for the BCD methods for general nonsmooth strongly convex problems are new in the literature. Thirdly, it appears to us that the complexity for solving nonconvex optimization by the BCD methods has not been studied before in the literature.

This paper is organized as follows. After reviewing some notations in Section 1.1, we present the basic SBMD algorithm for general nonsmooth optimization and discuss its convergence properties in Section 2. A variant of this algorithm for solving convex stochastic composite optimization problems, along with its complexity analysis are developed in Section 3. A generalization of this algorithm for solving nonconvex stochastic composite optimization is presented in Section 4. Finally some brief concluding remarks are given in Section 5.

1.1. Notation and terminology. Let $\mathbb{R}^{n_i}$, $i = 1, \ldots, b$, be Euclidean spaces equipped with inner product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|_i$ ($\| \cdot \|_{i,*}$ be the congugate) such that $\sum_{i=1}^b n_i = n$. Let I_n be the identity matrix in $\mathbb{R}^n$ and $U_i \in \mathbb{R}^{n \times n_i}, i = 1, 2, \ldots, b$, be the set of matrices satisfying

$$(U_1, U_2, \ldots, U_b) = I_n.$$

For a given $x \in \mathbb{R}^n$, we denote its i-th block by $x^{(i)} = U_i^T x$, $i = 1, \ldots, b$. Note that

$$x = U_1 x^{(1)} + \ldots + U_b x^{(b)}.$$

Moreover, we define

$$\|x\|^2 = \|x^{(1)}\|_1^2 + \ldots + \|x^{(b)}\|_b^2.$$

and denote its conjugate by $\|y\|_*^2 = \|y^{(1)}\|_{1,*}^2 + \ldots + \|y^{(b)}\|_{b,*}^2$.

Let X be defined in (1.2) and $f : X \to \mathbb{R}$ be a closed convex function. For any $x \in X$, let $G(x, \xi)$ be a stochastic subgradient of $f(\cdot)$ such that (1.3) holds. We denote the partial stochastic subgradient of $f(\cdot)$ by $G_i(x, \xi) = U_i^T G(x, \xi)$, $i = 1, 2, ..., b$.

2. The SBMD methods for nonsmooth convex optimization. In this section, we present the stochastic block coordinate descent method for solving stochastic nonsmooth convex optimization problems and discuss its convergence properties. More specifically, we present the basic scheme of the SBMD method in Subsection 2.1 and discuss its convergence properties for solving general nonsmooth and strongly convex nonsmooth problems in Subsections 2.2 and 2.3, respectively.

Throughout this section, we assume that $f(\cdot)$ in (1.1) is convex and its stochastic subgradients satisfy, in addition to (1.3), the following condition:

$$\mathbb{E}[\|G_i(x, \xi)\|_{i,*}^2] \le M_i^2, \ i = 1, 2, ..., b. \tag{2.1}$$

Clearly, by (1.3) and (2.1), we have

$$\|g_i(x)\|_{i,*}^2 = \|\mathbb{E}[G_i(x, \xi)]\|_{i,*}^2 \le \mathbb{E}[\|G_i(x, \xi)\|_{i,*}^2] \le M_i^2, \ i = 1, 2, ..., b, \tag{2.2}$$

and

$$\|g(x)\|_*^2 = \sum_{i=1}^{b} \|g_i(x)\|_{i,*}^2 \le \sum_{i=1}^{b} M_i^2. \tag{2.3}$$

2.1. The SBMD algorithm for nonsmooth problems. We present a general scheme of the SBMD algorithm, based on Bregman's divergence, to solve stochastic convex optimization problems.

Recall that a function $\omega_i : X_i \to R$ is a distance generating function [24] with modulus α_i with respect to $\|\cdot\|_i$, if ω is continuously differentiable and strongly convex with parameter α_i with respect to $\|\cdot\|_i$. Without loss of generality, we assume throughout the paper that $\alpha_i = 1$ for any $i = 1, \ldots, b$. Therefore, we have

$$\langle x - z, \nabla\omega_i(x) - \nabla\omega_i(z)\rangle \ge \|x - z\|_i^2 \quad \forall x, z \in X_i.$$

The prox-function associated with ω_i is given by

$$V_i(z, x) = \omega_i(x) - [\omega_i(z) + \langle \omega_i'(z), x - z\rangle] \quad \forall x, z \in X_i. \tag{2.4}$$

The prox-function $V_i(\cdot, \cdot)$ is also called the Bregman's distance, which was initially studied by Bregman [5] and later by many others (see [1, 2, 37] and references therein). For a given $x \in X_i$ and $y \in \mathbb{R}^{n_i}$, we define the prox-mapping as

$$P_i(v, y, \gamma) = \arg\min_{u \in X_i} \langle y, u\rangle + \frac{1}{\gamma} V_i(u, v). \tag{2.5}$$

Suppose that the set X_i is bounded, the distance generating function ω_i also gives rise to the following characteristic entity that will be used frequently in our convergence analysis:

$$\mathcal{D}_i \equiv \mathcal{D}_{\omega_i, X_i} := \left(\max_{x \in X_i} \omega_i(x) - \min_{x \in X_i} \omega_i(x)\right)^{\frac{1}{2}}. \tag{2.6}$$

Let $x_1^{(i)} = \operatorname{argmin}_{x \in X_i} \omega_i(x)$, $i = 1, \ldots, b$. We can easily see that for any $x \in X$,

$$V_i(x_1^{(i)}, x^{(i)}) = \omega_i(x^{(i)}) - \omega_i(x_1^{(i)}) - \langle \nabla\omega_i(x_1^{(i)}), x^{(i)} - x_1^{(i)}\rangle \le \omega_i(x^{(i)}) - \omega_i(x_1^{(i)}) \le \mathcal{D}_i, \tag{2.7}$$

which, in view of the strong convexity of ω_i, also implies that $\|x_1^{(i)} - x^{(i)}\|_i^2/2 \le \mathcal{D}_i$. Therefore, for any $x, y \in X$, we have

$$\|x^{(i)} - y^{(i)}\|_i \le \|x^{(i)} - x_1^{(i)}\|_i + \|x_1^{(i)} - y^{(i)}\|_i \le 2\sqrt{2\mathcal{D}_i}, \tag{2.8}$$

$$\|x - y\| = \sqrt{\sum_{i=1}^{b} \|x^{(i)} - y^{(i)}\|_i^2} \le 2\sqrt{2\sum_{i=1}^{b} \mathcal{D}_i}. \tag{2.9}$$

With the above definition of the prox-mapping, we can formally describe the stochastic block mirror

descent (SBMD) method as follows.

Algorithm 1 The Stochastic Block Mirror Descent (SBMD) Algorithm

Let $x_1 \in X$, stepsizes $\{\gamma_k\}_{k\geq 1}$, weights $\{\theta_k\}_{k\geq 1}$, and probabilities $p_i \in [0,1]$, $i = 1,\ldots,b$, s.t. $\sum_{i=1}^b p_i = 1$ be given. Set $s_1 = 0$, and $u_i = 1$ for $i = 1,\ldots,b$.

for $k = 1,\ldots,N$ **do**

1. Generate a random variable i_k according to

$$\text{Prob}\,\{i_k = i\} = p_i, \quad i = 1,\ldots,b. \tag{2.10}$$

2. Update $s_k^{(i)}$, $i = 1,\ldots,b$, by

$$s_{k+1}^{(i)} = \begin{cases} s_k^{(i)} + x_k^{(i)} \sum_{j=u_{i_k}}^k \theta_j & i = i_k, \\ s_k^{(i)} & i \neq i_k, \end{cases} \tag{2.11}$$

and then set $u_{i_k} = k+1$.

3. Update $x_k^{(i)}$, $i = 1,\ldots,b$, by

$$x_{k+1}^{(i)} = \begin{cases} P_i(x_k^{(i)}, G_{i_k}(x_k,\xi_k), \gamma_k) & i = i_k, \\ x_k^{(i)} & i \neq i_k. \end{cases} \tag{2.12}$$

end for

Output: Set $s_{N+1}^{(i)} = s_{N+1}^{(i)} + x_N^{(i)} \sum_{j=u_i}^N \theta_j$, $i = 1,\ldots,b$, and $\bar{x}_N = s_{N+1} / \sum_{k=1}^N \theta_k$.

We now add a few remarks about the SBMD algorithm stated above. Firstly, each iteration of the SBMD method recursively updates the search point x_k based on the partial stochastic subgradient $G_{i_k}(x_k,\xi_k)$. In addition, rather than taking the average of $\{x_k\}$ in the end of algorithm as in the mirror-descent method, we introduce an incremental block averaging scheme to compute the output of the algorithm. More specifically, we use a summation vector s_k to denote the weighted sum of x_k's and the index variables u_i, $i = 1,\ldots,b$, to record the latest iteration when the i-th block of s_k is updated. Then in (2.11), we add up the i_k-th block of s_k with $x_k \sum_{j=i_k}^k \theta_j$, where $\sum_{j=i_k}^k \theta_j$ is often given by explicit formula and hence easy to compute. It can be checked that by using this averaging scheme, we have

$$\bar{x}_N = \left(\sum_{k=1}^N \theta_k\right)^{-1} \sum_{k=1}^N (\theta_k x_k). \tag{2.13}$$

Secondly, observe that in addition to (2.11) and (2.12), each iteration of the SBMD method involves the computation of G_{i_k}. Whenever possible, we should update G_{i_k} recursively in order to reduce the iteration cost of the SBMD algorithm. Consider an important class of SP problems with the objective function

$$f(x) = \mathbb{E}[\psi(Bx - q, \xi)] + \chi(x),$$

where $\psi(\cdot)$ and $\chi(\cdot)$ are relatively simple functions, $q \in \mathbb{R}^n$, and $B \in \mathbb{R}^{n\times n}$. For the sake of simplicity, let us also assume that $n_1 = \ldots = n_b = 1$. For example, in the well-known support vector machine (SVM) problem, we have $\psi(y) = \max\{\langle y,\xi\rangle, 0\}$ and $\chi(x) = \|x\|_2^2/2$. In order to compute the full vector $G(x_k,\xi_k)$, we need $\mathcal{O}(n^2)$ arithmetic operations to compute the vector $Bx_k - q$, which majorizes other arithmetic operations if ψ and χ are simple. On the other hand, by recursively updating the vector $y_k = Bx_k$ in the SBMD method, we can significantly reduce the iteration cost from $\mathcal{O}(n^2)$ to $\mathcal{O}(n)$. This bound can be further reduced if both

ξ_k and B are sparse (i.e., the vector ξ_k and each row vector of B contain just a few nonzeros). The above example can be generalized to the case when B has $r \times b$ blocks denoted by $B_{i,j} \in \mathbb{R}^{m_i \times n_j}$, $1 \le i \le r$ and $1 \le j \le b$, and each block row $B_i = (B_{i,1}, \ldots, B_{i,b})$, $i = 1, \ldots, r$, is block-sparse (see [29] for some related discussion).

Thirdly, observe that the above SBMD method is conceptual only because we have not yet specified the selection of the stepsizes $\{\gamma_k\}$, the weights $\{\theta_k\}$, and the probabilities $\{p_i\}$. We will specify these parameters after establishing some basic convergence properties of this method.

2.2. Convergence properties of SBMD for nonsmooth problems. In this subsection, we discuss the main convergence properties of the SBMD method for solving general nonsmooth convex problems.

THEOREM 2.1. *Let $\bar{x}_N$ be the output of the SBMD algorithm and suppose that*

$$\theta_k = \gamma_k, \quad k = 1, \ldots, N. \tag{2.14}$$

Then we have, for any $N \ge 1$ and $x \in X$,

$$\mathbb{E}[f(\bar{x}_N) - f(x)] \le \left(\sum_{k=1}^N \gamma_k\right)^{-1} \left[\sum_{i=1}^b p_i^{-1} V_i(x_1^{(i)}, x^{(i)}) + \frac{1}{2}\sum_{k=1}^N \gamma_k^2 \sum_{i=1}^b M_i^2\right], \tag{2.15}$$

where the expectation is taken with respect to (w.r.t.) $\{i_k\}$ and $\{\xi_k\}$.

Proof. For simplicity, let us denote $V_i(z,x) \equiv V_i(z^{(i)}, x^{(i)})$, $g_{i_k} \equiv g^{(i_k)}(x_k)$ (c.f. (1.3)) and $V(z,x) = \sum_{i=1}^b p_i^{-1} V_i(z,x)$. Also let us denote $\zeta_k = (i_k, \xi_k)$ and $\zeta_{[k]} = (\zeta_1, \ldots, \zeta_k)$. By the optimality condition of (2.5) (e.g., Lemma 1 of [24]) and the definition of $x_k^{(i)}$ in (2.12), we have

$$V_{i_k}(x_{k+1}, x) \le V_{i_k}(x_k, x) + \gamma_k \left\langle G_{i_k}(x_k, \xi_k), U_{i_k}^T(x - x_k)\right\rangle + \frac{1}{2}\gamma_k^2 \left\|G_{i_k}(x_k, \xi_k)\right\|_{i_k,*}^2.$$

Using this observation, we have, for any $k \ge 1$ and $x \in X$,

$$\begin{aligned} V(x_{k+1}, x) &= \sum_{i \ne i_k} p_i^{-1} V_i(x_k, x) + p_{i_k}^{-1} V_{i_k}(x_{k+1}, x) \\ &\le \sum_{i \ne i_k} p_i^{-1} V_i(x_k, x) + p_{i_k}^{-1}\left[V_{i_k}(x_k, x) + \gamma_k \left\langle G_{i_k}(x_k, \xi_k), U_{i_k}^T(x - x_k)\right\rangle + \tfrac{1}{2}\gamma_k^2 \left\|G_{i_k}(x_k, \xi_k)\right\|_{i_k,*}^2\right] \\ &= V(x_k, x) + \gamma_k p_{i_k}^{-1} \left\langle U_{i_k} G_{i_k}(x_k, \xi_k), x - x_k\right\rangle + \tfrac{1}{2}\gamma_k^2 p_{i_k}^{-1} \left\|G_{i_k}(x_k, \xi_k)\right\|_{i_k,*}^2 \\ &= V(x_k, x) + \gamma_k \langle g(x_k), x - x_k\rangle + \gamma_k \delta_k + \tfrac{1}{2}\gamma_k^2 \bar{\delta}_k, \end{aligned} \tag{2.16}$$

where

$$\delta_k := \langle p_{i_k}^{-1} U_{i_k} G_{i_k}(x_k, \xi_k) - g(x_k), x - x_k\rangle \ \text{ and } \ \bar{\delta}_k := p_{i_k}^{-1} \left\|G_{i_k}(x_k, \xi_k)\right\|_{i_k,*}^2. \tag{2.17}$$

It then follows from (2.16) and the convexity of $f(\cdot)$ that, for any $k \ge 1$ and $x \in X$,

$$\gamma_k[f(x_k) - f(x)] \le \gamma_k \langle g(x_k), x_k - x\rangle \le V(x_k, x) - V(x_{k+1}, x) + \gamma_k \delta_k + \frac{1}{2}\gamma_k^2 \bar{\delta}_k.$$

By using the above inequalities, the convexity of $f(\cdot)$, and the fact that $\bar{x}_N = \sum_{k=1}^N (\gamma_k x_k) / \sum_{k=1}^N \gamma_k$ due to (2.13) and (2.14), we conclude that for any $N \ge 1$ and $x \in X$,

$$\begin{aligned} f(\bar{x}_N) - f(x) &\le \left(\sum_{k=1}^N \gamma_k\right)^{-1} \textstyle\sum_{k=1}^N \gamma_k \left[f(x_k) - f(x)\right] \\ &\le \left(\sum_{k=1}^N \gamma_k\right)^{-1} \left[V(x_1, x) + \sum_{k=1}^N \left(\gamma_k \delta_k + \tfrac{1}{2}\gamma_k^2 \bar{\delta}_k\right)\right]. \end{aligned} \tag{2.18}$$

Now, observe that by (1.3) and (2.10),

$$\begin{aligned}\mathbb{E}_{\zeta_k}\left[p_{i_k}^{-1}\langle U_{i_k}G_{i_k}, x-x_k\rangle|\zeta_{[k-1]}\right] &= \sum_{i=1}^b \mathbb{E}_{\xi_k}\left[\langle U_iG_i(x_k,\xi_k), x-x_k\rangle|\zeta_{[k-1]}\right]\\ &= \sum_{i=1}^b \langle U_ig_i(x_k), x-x_k\rangle = \langle g(x_k), x-x_k\rangle,\end{aligned}$$

and hence that

$$\mathbb{E}[\delta_k|\zeta_{k-1}] = 0. \tag{2.19}$$

Also, by (2.10) and (2.1),

$$\mathbb{E}\left[p_{i_k}^{-1}\left\|G_{i_k}(x_k,\xi_k)\right\|_{i_k,*}^2\right] = \sum_{i=1}^b p_ip_i^{-1}\left\|G_i(x_k,\xi_k)\right\|_{i,*}^2 \le \sum_{i=1}^b M_i^2. \tag{2.20}$$

Our result in (2.15) then immediately follows by taking expectation on both sides of (2.18), and using the previous observations in (2.19) and (2.20). ∎

Below we provide a few specialized convergence results for the SBMD algorithm after properly selecting $\{p_i\}$, $\{\gamma_k\}$, and $\{\theta_k\}$.

COROLLARY 2.2. *Suppose that $\{\theta_k\}$ in Algorithm 1 are set to (2.14).*

a) If X is bounded, and $\{p_i\}$ and $\{\gamma_k\}$ are set to

$$p_i = \frac{\sqrt{\mathcal{D}_i}}{\sum_{i=1}^b \sqrt{\mathcal{D}_i}}, \quad i=1,\dots,b, \quad \text{and} \quad \gamma_k = \gamma \equiv \frac{\sqrt{2}\sum_{i=1}^b \sqrt{\mathcal{D}_i}}{\sqrt{N\sum_{i=1}^b M_i^2}}, \quad k=1,\dots,N, \tag{2.21}$$

then

$$\mathbb{E}[f(\bar{x}_N) - f(x)] \le \sqrt{\frac{2}{N}}\sum_{i=1}^b \sqrt{\mathcal{D}_i}\sqrt{\sum_{i=1}^b M_i^2} \quad \forall x \in X. \tag{2.22}$$

b) If $\{p_i\}$ and $\{\gamma_k\}$ are set to

$$p_i = \frac{1}{b}, \quad i=1,\dots,b, \quad \text{and} \quad \gamma_k = \gamma \equiv \frac{\sqrt{2b}\tilde{D}}{\sqrt{N\sum_{i=1}^b M_i^2}}, \quad k=1,\dots,N, \tag{2.23}$$

for some $\tilde{D} > 0$, then

$$\mathbb{E}[f(\bar{x}_N) - f(x)] \le \sqrt{\sum_{i=1}^b M_i^2}\left(\frac{\sum_{i=1}^b V_i(x_1^{(i)}, x_*^{(i)})}{\tilde{D}} + \tilde{D}\right)\frac{\sqrt{b}}{\sqrt{2N}} \quad \forall x \in X. \tag{2.24}$$

Proof. We show part a) only, since part b) can be proved similarly. Note that by (2.7) and (2.21), we have

$$\sum_{i=1}^b p_i^{-1}V_i(x_1^{(i)}, x^{(i)}) \le \sum_{i=1}^b p_i^{-1}\mathcal{D}_i \le \left(\sum_{i=1}^b \sqrt{\mathcal{D}_i}\right)^2.$$

Using this observation, (2.15), and (2.21), we have

$$\mathbb{E}[f(\bar{x}_N) - f(x_*)] \le (N\gamma)^{-1}\left[\left(\sum_{i=1}^b \sqrt{\mathcal{D}_i}\right)^2 + \frac{N\gamma^2}{2}\sum_{i=1}^b M_i^2\right] = \sqrt{\frac{2}{N}}\sum_{i=1}^b \sqrt{\mathcal{D}_i}\sqrt{\sum_{i=1}^b M_i^2}.$$

■

A few remarks about the results obtained in Theorem 2.1 and Corollary 2.2 are in place. First, the parameter setting in (2.21) only works for the case when X is bounded, while the one in (2.23) also applies to the case when X is unbounded or when the bounds $\mathcal{D}_i$, $i = 1, \ldots, b$, are not available. It can be easily seen that the optimal choice of $\tilde{D}$ in (2.24) would be $\sqrt{\sum_{i=1}^b V_i(x_1^{(i)}, x_*^{(i)})}$. In this case, (2.24) reduces to

$$\mathbb{E}[f(\bar{x}_N) - f(x)] \le \sqrt{2\sum_{i=1}^b M_i^2}\sqrt{\sum_{i=1}^b V_i(x_1^{(i)}, x_*^{(i)})}\frac{\sqrt{b}}{\sqrt{N}} \le \sqrt{2\sum_{i=1}^b M_i^2}\sqrt{\sum_{i=1}^b \mathcal{D}_i}\frac{\sqrt{b}}{\sqrt{N}}, \tag{2.25}$$

where the second inequality follows from (2.7). It is interesting to note the difference between the above bound and (2.22). Specifically, the bound obtained in (2.22) by using a non-uniform distribution $\{p_i\}$ always minorizes the one in (2.25) by the Cauchy-Schwartz inequality.

Second, observe that in view of (2.22), the total number of iterations required by the SBMD method to find an ϵ-solution of (1.1) can be bounded by

$$2\left(\sum_{i=1}^b \sqrt{\mathcal{D}_i}\right)^2 \left(\sum_{i=1}^b M_i^2\right)\frac{1}{\epsilon^2}. \tag{2.26}$$

Also note that the iteration complexity of the mirror-descent SA algorithm employed with the same $\omega_i(\cdot)$, $i = 1, \ldots, b$, is given by

$$2\sum_{i=1}^b \mathcal{D}_i \left(\sum_{i=1}^b M_i^2\right) \frac{1}{\epsilon^2}. \tag{2.27}$$

Clearly, the bound in (2.26) can be larger, up to a factor of b, than the one in (2.27). Therefore, the total arithmetic cost of the SBMD algorithm will be comparable to or smaller than that of the mirror descent SA, if its iteration cost is smaller than that of the latter algorithm by a factor of $\mathcal{O}(b)$.

Third, in Corollary 2.2 we have used a constant stepsize policy where $\gamma_1 = \ldots = \gamma_N$. However, it should be noted that variable stepsize policies, e.g., those similar to [24], can also be used in the SBMD method.

2.3. Convergence properties of SBMD for nonsmooth strongly convex problems. In this subsection, we assume that the objective function $f(\cdot)$ in (1.1) is strongly convex, i.e., $\exists\ \mu > 0$ s.t.

$$f(y) \ge f(x) + \langle g(x), y - x\rangle + \frac{\mu}{2}\|y - x\|^2 \quad \forall\, x, y \in X. \tag{2.28}$$

In order to establish the convergence of the SBMD algorithm for solving strongly convex problems, we need to assume that the prox-functions $V_i(\cdot,\cdot)$, $i = 1, \ldots, b$, satisfy a quadratic growth condition (e.g., [12, 7, 8]):

$$V_i(z^{(i)}, x^{(i)}) \le \frac{Q}{2}\|z^{(i)} - x^{(i)}\|_i^2 \quad \forall z^{(i)}, x^{(i)} \in X_i, \tag{2.29}$$

for some $Q > 0$. In addition, we need to assume that the probability distribution of i_k is uniform, i.e.,

$$p_1 = p_2 = \ldots = p_b = \frac{1}{b}. \tag{2.30}$$

Before proving the convergence of the SBMD algorithm for solving strongly convex problems, we first state a simple technical result obtained by slightly modifying Lemma 3 of [15].

LEMMA 2.3. *Let $a_k \in (0, 1]$, $k = 1, 2, \ldots$, be given. Also let us denote*

$$A_k := \begin{cases} 1 & k = 1 \\ (1 - a_k)\, A_{k-1} & k \ge 2. \end{cases} \tag{2.31}$$

Suppose that $A_k > 0$ for all $k \geq 2$ and that the sequence $\{\Delta_k\}$ satisfies

$$\Delta_{k+1} \leq (1 - a_k)\Delta_k + B_k, \quad k = 1, 2, \ldots. \tag{2.32}$$

Then, we have $\Delta_{k+1}/A_k \leq (1 - a_1)\Delta_1 + \sum_{i=1}^{k}(B_i/A_i)$.

We are now ready to describe the main convergence properties of the SBMD algorithm for solving nonsmooth strongly convex problems.

THEOREM 2.4. *Suppose that (2.28), (2.29), and (2.30) hold. If*

$$\gamma_k \leq \frac{bQ}{\mu} \tag{2.33}$$

and

$$\theta_k = \frac{\gamma_k}{\Gamma_k} \quad \textit{with} \quad \Gamma_k = \begin{cases} 1 & k = 1 \\ \Gamma_{k-1}(1 - \frac{\gamma_k \mu}{bQ}) & k \geq 2, \end{cases} \tag{2.34}$$

then, for any $N \geq 1$ and $x \in X$, we have

$$\mathbb{E}[f(\bar{x}_N) - f(x)] \leq \left(\sum_{k=1}^{N} \theta_k\right)^{-1} \left[(b - \frac{\gamma_1 \mu}{Q}) \sum_{i=1}^{b} V_i(x_1^{(i)}, x^{(i)}) + \frac{1}{2} \sum_{k=1}^{N} \gamma_k \theta_k \sum_{i=1}^{b} M_i^2\right]. \tag{2.35}$$

Proof. For simplicity, let us denote $V_i(z, x) \equiv V_i(z^{(i)}, x^{(i)})$, $g_{i_k} \equiv g^{(i_k)}(x_k)$, and $V(z, x) = \sum_{i=1}^{b} p_i^{-1} V_i(z, x)$. Also let us denote $\zeta_k = (i_k, \xi_k)$ and $\zeta_{[k]} = (\zeta_1, \ldots, \zeta_k)$, and let δ_k and $\bar{\delta}_k$ be defined in (2.17). By (2.29) and (2.30), we have

$$V(z, x) = b \sum_{i=1}^{b} V_i(z^{(i)}, x^{(i)}) \leq \frac{bQ}{2} \sum_{i=1}^{b} \|z^{(i)} - x^{(i)}\|_i^2 = \frac{bQ}{2} \|z - x\|^2. \tag{2.36}$$

Using this observation, (2.16), and (2.28), we obtain

$$\begin{aligned} V(x_{k+1}, x) &\leq V(x_k, x) + \gamma_k \langle g(x_k), x - x_k \rangle + \gamma_k \delta_k + \frac{1}{2} \gamma_k^2 \bar{\delta}_k \\ &\leq V(x_k, x) + \gamma_k \left[f(x) - f(x_k) - \frac{\mu}{2} \|x - x_k\|^2\right] + \gamma_k \delta_k + \frac{1}{2} \gamma_k^2 \bar{\delta}_k \\ &\leq \left(1 - \frac{\gamma_k \mu}{bQ}\right) V(x_k, x) + \gamma_k \left[f(x) - f(x_k)\right] + \gamma_k \delta_k + \frac{1}{2} \gamma_k^2 \bar{\delta}_k, \end{aligned}$$

which, in view of Lemma 2.3 (with $a_k = 1 - \gamma_k \mu/(bQ)$ and $A_k = \Gamma_k$), then implies that

$$\frac{1}{\Gamma_N} V(x_{N+1}, x) \leq \left(1 - \frac{\gamma_1 \mu}{bQ}\right) V(x_1, x) + \sum_{k=1}^{N} \Gamma_k^{-1} \gamma_k \left[f(x) - f(x_k) + \delta_k + \frac{1}{2} \gamma_k^2 \bar{\delta}_k\right]. \tag{2.37}$$

Using the fact that $V(x_{N+1}, x) \geq 0$ and (2.34), we conclude from the above relation that

$$\sum_{k=1}^{N} \theta_k [f(x_k) - f(x)] \leq \left(1 - \frac{\gamma_1 \mu}{bQ}\right) V(x_1, x) + \sum_{k=1}^{N} \theta_k \delta_k + \frac{1}{2} \sum_{k=1}^{N} \gamma_k \theta_k \bar{\delta}_k. \tag{2.38}$$

Taking expectation on both sides of the above inequality, and using relations (2.19) and (2.20), we obtain

$$\sum_{k=1}^{N} \theta_k \mathbb{E}[f(x_k) - f(x)] \leq \left(1 - \frac{\gamma_1 \mu}{bQ}\right) V(x_1, x) + \frac{1}{2} \sum_{k=1}^{N} \gamma_k \theta_k \sum_{i=1}^{b} M_i^2,$$

which, in view of (2.13), (2.30) and the convexity of $f(\cdot)$, then clearly implies (2.35). ∎

Below we provide a specialized convergence result for the SBMD method to solve nonsmooth strongly convex problems after properly selecting $\{\gamma_k\}$.

COROLLARY 2.5. *Suppose that (2.28), (2.29) and (2.30) hold. If $\{\theta_k\}$ are set to (2.34) and $\{\gamma_k\}$ are set to*

$$\gamma_k = \frac{2bQ}{\mu(k+1)}, \quad k = 1, \ldots, N, \tag{2.39}$$

then, for any $N \geq 1$ and $x \in X$, we have

$$\mathbb{E}[f(\bar{x}_N) - f(x)] \leq \frac{2bQ}{\mu(N+1)} \sum_{i=1}^{b} M_i^2. \tag{2.40}$$

Proof. It can be easily seen from (2.34) and (2.39) that

$$\Gamma_k = \frac{2}{k(k+1)}, \quad \theta_k = \frac{\gamma_k}{\Gamma_k} = \frac{bkQ}{\mu}, \quad b - \frac{\gamma_1 \mu}{Q} = 0, \tag{2.41}$$

$$\sum_{k=1}^{N} \theta_k = \frac{bQN(N+1)}{2\mu}, \quad \sum_{k=1}^{N} \gamma_k \theta_k \leq \frac{2b^2Q^2N}{\mu^2}, \tag{2.42}$$

and

$$\sum_{k=1}^{N} \theta_k^2 = \frac{b^2Q^2}{\mu^2} \frac{N(N+1)(2N+1)}{6} \leq \frac{b^2Q^2}{\mu^2} \frac{N(N+1)^2}{3}. \tag{2.43}$$

Hence, by (2.35),

$$\mathbb{E}[f(\bar{x}_N) - f(x)] \leq \frac{1}{2} \left(\sum_{k=1}^{N} \theta_k \right)^{-1} \sum_{k=1}^{N} \gamma_k \theta_k \sum_{i=1}^{b} M_i^2 \leq \frac{2bQ}{\mu(N+1)} \sum_{i=1}^{b} M_i^2.$$

∎

In view of (2.40), the number of iterations performed by the SBMD method to find an ϵ-solution for nonsmooth strongly convex problems can be bound by

$$\frac{2b}{\mu\epsilon} \sum_{i=1}^{b} M_i^2,$$

which is comparable to the optimal bound obtained in [12, 7, 23] (up to a constant factor b). To the best of our knowledge, no such complexity results have been obtained before for BCD type methods in the literature.

2.4. Large-deviation properties of SBMD for nonsmooth problems. Our goal in this subsection is to establish the large-deviation results associated with the SBMD algorithm under the following "light-tail" assumption about the random variable ξ:

$$\mathbb{E}\left\{ \exp\left[\|G_i(x,\xi)\|_{i,*}^2 / M_i^2 \right] \right\} \leq \exp(1), \ i = 1, 2, ..., b. \tag{2.44}$$

It can be easily seen that (2.44) implies (2.1) by Jensen's inequality. It should be pointed out that the above "light-tail" assumption is alway satisified for determinisitc problems with bounded subgradients.

For the sake of simplicity, we only consider the case when the random variables $\{i_k\}$ in the SBMD agorithm are uniformly distributed, i.e., relation (2.30) holds. The following result states the large-deviation properties of the SBMD algorithm for solving general nonsmooth problems without assuming strong convexity.

THEOREM 2.6. *Suppose that Assumptions (2.44) and (2.30) holds. Also assume that X is bounded.*

a) For solving general nonsmooth CP problems (i.e., (2.14) holds), we have

$$\begin{aligned}\text{Prob}\Bigg\{f(\bar{x}_N)-f(x)\geq b\left(\sum_{k=1}^{N}\gamma_k\right)^{-1}\Bigg[\sum_{i=1}^{b}V_i(x_1^{(i)},x^{(i)})+\bar{M}^2\sum_{k=1}^{N}\gamma_k^2\\ +\lambda\bar{M}^2\left(\sum_{k=1}^{N}\gamma_k^2+32\,b\sum_{i=1}^{b}\mathcal{D}_i\sqrt{\sum_{k=1}^{N}\gamma_k^2}\right)\Bigg]\Bigg\}\leq\exp\left(-\lambda^2/3\right)+\exp(-\lambda),\end{aligned}\tag{2.45}$$

for any $N\geq 1$, $x\in X$ and $\lambda>0$, where $\bar{M}=\max_{i=1,\ldots,b}M_i$.

b) For solving strongly convex problems (i.e., (2.28), (2.29), (2.33), and (2.34) hold), we have

$$\begin{aligned}\text{Prob}\Bigg\{f(\bar{x}_N)-f(x)\geq b\left(\sum_{k=1}^{N}\theta_k\right)^{-1}\Bigg[(b-\tfrac{\gamma_1\mu}{Q})\sum_{i=1}^{b}V_i(x_1^{(i)},x^{(i)})+\bar{M}^2\sum_{k=1}^{N}\gamma_k\theta_k\\ +\lambda\bar{M}^2\left(\sum_{k=1}^{N}\gamma_k\theta_k+32b\sum_{i=1}^{b}\mathcal{D}_i\sqrt{\sum_{k=1}^{N}\theta_k^2}\right)\Bigg]\Bigg\}\leq\exp\left(-\lambda^2/3\right)+\exp(-\lambda),\end{aligned}\tag{2.46}$$

for any $N\geq 1$, $x\in X$ and $\lambda>0$.

Proof. We first show part a). Note that by (2.44), the concavity of $\phi(t)=\sqrt{t}$ for $t\geq 0$ and the Jensen's inequality, we have, for any $i=1,2,...,b$,

$$\mathbb{E}\left\{\exp\left[\|G_i(x,\xi)\|_{i,*}^2/(2M_i^2)\right]\right\}\leq\sqrt{\mathbb{E}\left\{\exp\left[\|G_i(x,\xi)\|_{i,*}^2/M_i^2\right]\right\}}\leq\exp(1/2).\tag{2.47}$$

Also note that by (2.19), δ_k, $k=1,\ldots,N$, is the martingale-difference. In addition, denoting $\mathcal{M}^2\equiv 32\,b^2\bar{M}^2\sum_{i=1}^{b}\mathcal{D}_i$, we have

$$\begin{aligned}\mathbb{E}[\exp\left(\mathcal{M}^{-2}\delta_k^2\right)]&\leq\sum_{i=1}^{b}p_i\mathbb{E}\left[\exp\left(\mathcal{M}^{-2}\|x-x_k\|^2\,\|p_i^{-1}U_i^TG_i-g(x_k)\|_*^2\right)\right] && \text{(by (2.10), (2.17))}\\ &\leq\sum_{i=1}^{b}p_i\mathbb{E}\left\{\exp\left[2\mathcal{M}^{-2}\|x-x_k\|^2\left(b^2\|G_i\|_*^2+\|g(x_k)\|_*^2\right)\right]\right\} && \text{(by definition of } U_i \text{ and (2.30))}\\ &\leq\sum_{i=1}^{b}p_i\mathbb{E}\left\{\exp\left[16\mathcal{M}^{-2}\left(\sum_{i=1}^{b}\mathcal{D}_i\right)\left(b^2\|G_i\|_*^2+\sum_{i=1}^{b}M_i^2\right)\right]\right\} && \text{(by (2.3) and (2.9))}\\ &\leq\sum_{i=1}^{b}p_i\mathbb{E}\left\{\exp\left[\frac{b^2\|G_i\|_*^2+\sum_{i=1}^{b}M_i^2}{2b^2\bar{M}^2}\right]\right\} && \text{(by definition of } \mathcal{M})\\ &\leq\sum_{i=1}^{b}p_i\mathbb{E}\left\{\exp\left[\frac{\|G_i\|_*^2}{2M_i^2}+\frac{1}{2}\right]\right\}\leq\exp(1). && \text{(by (2.47))}\end{aligned}$$

Therefore, by the well-known large-deviation theorem on the Martingale-difference (see, e.g., Lemma 2 of [18]), we have

$$\text{Prob}\left\{\sum_{k=1}^{N}\gamma_k\delta_k\geq\lambda\mathcal{M}\sqrt{\sum_{k=1}^{N}\gamma_k^2}\right\}\leq\exp(-\lambda^2/3).\tag{2.48}$$

Also observe that under Assumption (2.44),

$$\begin{aligned}
\mathbb{E}\left[\exp\left(\bar{\delta}_k/(b\bar{M}^2)\right)\right] &\le \sum_{i=1}^{b} p_i \mathbb{E}\left[\exp\left(\|G_i(x_k,\xi_k)\|_{i,*}^2/\bar{M}^2\right)\right] && \text{(by (2.10), (2.17), (2.30))}\\
&\le \sum_{i=1}^{b} p_i \mathbb{E}\left[\exp\left(\|G_i(x_k,\xi_k)\|_{i,*}^2/M_i^2\right)\right] && \text{(by definition of } \bar{M})\\
&\le \sum_{i=1}^{b} p_i \exp(1) = \exp(1). && \text{(by (2.1))}
\end{aligned}$$

Setting $\pi_k = \gamma_k^2/\sum_{k=1}^N \gamma_k^2$, we have $\exp\left\{\sum_{k=1}^N \pi_k \bar{\delta}_k/(b\bar{M}^2)\right\} \le \sum_{k=1}^N \pi_k \exp\{\bar{\delta}_k/(b\bar{M}^2)\}$. Using these previous two inequalities, we have

$$\mathbb{E}\left[\exp\left\{\sum_{k=1}^N \gamma_k^2\bar{\delta}_k/(b\bar{M}^2\sum_{k=1}^N \gamma_k^2)\right\}\right] \le \exp\{1\}.$$

It then follows from Markov's inequality that

$$\forall \lambda \ge 0 : \text{Prob}\left\{\sum_{k=1}^N \gamma_k^2\bar{\delta}_k > (1+\lambda)(b\bar{M}^2)\sum_{k=1}^N \gamma_k^2\right\} \le \exp\{-\lambda\}. \tag{2.49}$$

Combining (2.18), (2.48) and (2.49), we obtain (2.45).

The probabilistic bound in (2.46) follows from (2.38) and an argument similar to the one used in the proof of (2.45), and hence the details are skipped.

■

We now provide some specialized large-deviation results for the SBMD algorithm with different selections of $\{\gamma_k\}$ and $\{\theta_k\}$.

COROLLARY 2.7. *Suppose that (2.44) and (2.30) hold. Also assume that X is bounded.*

a) If $\{\theta_k\}$ and $\{\gamma_k\}$ are set to (2.14) and (2.23) for general nonsmooth problems, then we have

$$\begin{aligned}
&\text{Prob}\left\{f(\bar{x}_N) - f(x) \ge \frac{b\sqrt{\sum_{i=1}^b M_i^2}}{\sqrt{2Nb\tilde{D}^2}}\left(2b\tilde{D}^2 + \sum_{i=1}^b \mathcal{D}_i + 2\lambda b\tilde{D}^2\right) + \frac{32\lambda b^{\frac{5}{2}}\bar{M}^2\sum_{i=1}^b \mathcal{D}_i}{\sqrt{Nb\tilde{D}^2}}\right\}\\
&\le \exp\left(-\lambda^2/3\right) + \exp(-\lambda)
\end{aligned} \tag{2.50}$$

for any $x \in X$ and $\lambda > 0$.

b) If $\{\theta_k\}$ and $\{\gamma_k\}$ are set to (2.34) and (2.39) for strongly convex problems, then we have

$$\text{Prob}\left\{f(\bar{x}_N) - f(x) \ge \frac{4(1+\lambda)b^2\bar{M}^2Q}{(N+1)\mu} + \frac{64\lambda b^2\bar{M}^2\sum_{i=1}^b \mathcal{D}_i}{\sqrt{3N}}\right\} \le \exp\left(-\lambda^2/3\right) + \exp(-\lambda) \tag{2.51}$$

for any $x \in X$ and $\lambda > 0$.

Proof. Note that by (2.7), we have $\sum_{i=1}^b V_i(x_1^{(i)}, x^{(i)}) \le \sum_{i=1}^b \mathcal{D}_i$. Also by (2.23), we have

$$\sum_{k=1}^N \gamma_k = \left(\frac{2Nb\tilde{D}^2}{\sum_{i=1}^b M_i^2}\right)^{\frac{1}{2}} \quad \text{and} \quad \sum_{k=1}^N \gamma_k^2 = \frac{2b\tilde{D}^2}{\sum_{i=1}^b M_i^2}.$$

Using these identities and (2.45), we conclude that

$$\begin{aligned}
&\text{Prob}\left\{f(\bar{x}_N) - f(x) \ge b\left(\frac{\sum_{i=1}^b M_i^2}{2Nb\tilde{D}^2}\right)^{\frac{1}{2}}\left[\sum_{i=1}^b \mathcal{D}_i + 2b\tilde{D}^2\bar{M}^2\left(\sum_{i=1}^b M_i^2\right)^{-1}\right.\right.\\
&\left.\left.+\lambda\bar{M}^2\left(2b\tilde{D}^2\left(\sum_{i=1}^b \bar{M}_i^2\right)^{-1} + 32\sqrt{2}\,b^{\frac{3}{2}}\tilde{D}\sum_{i=1}^b \mathcal{D}_i\left(\sum_{i=1}^b \bar{M}_i^2\right)^{-\frac{1}{2}}\right)\right]\right\} \le \exp\left(-\lambda^2/3\right) + \exp(-\lambda).
\end{aligned}$$

Using the fact that $\bar{M}^2 \leq \sum_{i=1}^b M_i^2$ and simplifying the above relation, we obtain (2.50). Similarly, relation (2.51) follows directly from (2.46) and a few bounds in (2.41), (2.42) and (2.43). ∎

We now add a few remarks about the results obtained in Theorem 2.6 and Corollary 2.7. Firstly, observe that by (2.48), the number of iterations required by the SBMD method to find an (ϵ, λ)-solution of (1.1), i.e., a point $\bar{x} \in X$ s.t. $\text{Prob}\{f(\bar{x}) - f^* \geq \epsilon\} \leq \lambda$ can be bounded by

$$\mathcal{O}\left(\frac{\log^2(1/\lambda)}{\epsilon^2}\right)$$

after disregarding a few constant factors. To the best of our knowledge, now such large-deviation results have been obtained before for the BCD methods for solving general nonsmooth CP problems, although similar results have been established for solving smooth problems or some composite problems [28, 33].

Secondly, it follows from (2.46) that the number of iterations performed by the SBMD method to find an (ϵ, λ)-solution for nonsmooth strongly convex problems, after disregarding a few constant factors, can be bounded by $\mathcal{O}\left(\log^2(1/\lambda)/\epsilon^2\right)$, which is about the same as the one obtained for solving nonsmooth problems without assuming convexity. It should be noted, however, that this bound can be improved to $\mathcal{O}\left(\log(1/\lambda)/\epsilon\right)$, for example, by incorporating a domain shrinking procedure [8].

3. The SBMD algorithm for convex composite optimization. In this section, we present a variant of the SBMD algorithm which can make use of the smoothness properties of the objective function of an SP problem. More specifically, we consider convex composite optimization problems given in the form of (1.4), where $f(\cdot)$ is smooth and its gradients $g(\cdot)$ satisfy

$$\|g_i(x + U_i\rho_i) - g_i(x)\|_i \leq L_i\|\rho_i\|_i \ \ \forall\ \rho_i \in \mathbb{R}^{n_i},\ \ i = 1, 2, ..., b. \tag{3.1}$$

It then follows that

$$f(x + U_i\rho_i) \leq f(x) + \langle g_i(x), \rho_i\rangle + \frac{L_i}{2}\|\rho_i\|_i^2 \ \ \forall \rho_i \in \mathbb{R}^{n_i}, x \in X. \tag{3.2}$$

The following assumption is made throughout this section.

ASSUMPTION 1. *The function $\chi(\cdot)$ is block separable, i.e., $\chi(\cdot)$ can be decomposed as*

$$\chi(x) = \sum_{i=1}^n \chi_i(x^{(i)}) \ \ \forall x \in X. \tag{3.3}$$

where $\chi_i : \mathbb{R}^{n_i} \to \mathbb{R}$ are closed and convex.

Let $V_i(\cdot, \cdot)$ defined in (2.4). For a given $x \in X_i$ and $y \in R^{n_i}$, we define the composite prox-mapping as

$$\mathcal{P}_i(x, y, \gamma) := \text{argmin}_{z \in X_i}\langle y, z - x\rangle + \frac{1}{\gamma}V_i(z, x) + \chi_i(x). \tag{3.4}$$

Clearly, if $\chi(x) = 0$ for any $x \in X$, then problem (1.4) becomes a smooth optimization problem and the composite prox-mapping (3.4) reduces to (2.5).

We are now ready to describe a variant of the SBMD algorithm for solving smooth and composite problems.

Algorithm 2 A variant of SBMD for convex stochastic composite optimization

Let $x_1 \in X$, stepsizes $\{\gamma_k\}_{k\geq 1}$, weights $\{\theta_k\}_{k\geq 1}$, and probabilities $p_i \in [0,1]$, $i = 1,\ldots,b$, s.t. $\sum_{i=1}^b p_i = 1$ be given. Set $s_1 = 0$, $u_i = 1$ for $i = 1,\ldots,b$, and $\theta_1 = 0$.

for $k = 1,\ldots,N$ **do**

1. Generate a random variable i_k according to (2.10).
2. Update $s_k^{(i)}$, $i = 1,\ldots,b$, by (2.11) and then set $u_{i_k} = k+1$.
3. Update $x_k^{(i)}$, $i = 1,\ldots,b$, by

$$x_{k+1}^{(i)} = \begin{cases} \mathcal{P}_{i_k}(x_k^{(i)}, G_{i_k}(x_k,\xi_k), \gamma_k) & i = i_k, \\ x_k^{(i)} & i \neq i_k. \end{cases} \tag{3.5}$$

end for

Output: Set $s_{N+1}^{(i)} = s_{N+1}^{(i)} + x_{N+1}^{(i)} \sum_{j=u_i}^{N+1} \theta_j$, $i = 1,\ldots,b$, and $\bar{x}_N = s_{N+1} / \sum_{k=1}^{N+1} \theta_k$.

A few remarks about the above variant of SBMD algorithm for composite convex problem in place. Firstly, similar to Algorithm 1, $G(x_k,\xi_k)$ is an unbiased estimator of $g(x_k)$ (i.e., (1.3) holds). Moreover, in order to know exactly the effect of stochastic noises in $G(x_k,\xi_k)$, we assume that for some $\sigma_i \geq 0$,

$$E[\|G_i(x,\xi) - g_i(x)\|_{i,*}^2] \leq \sigma_i^2,\ i = 1,\ldots,b. \tag{3.6}$$

Clearly, if $\sigma_i = 0$, $i = 1,\ldots,b$, then the problem is deterministic. For notational convenience, we also denote

$$\sigma := \Big(\sum_{i=1}^b \sigma_i^2\Big)^{\frac{1}{2}}. \tag{3.7}$$

Secondly, observe that the way we compute the output $\bar{x}_N$ in Algorithm 2 is slightly different from Algorithm 1. In particular, we set $\theta_1 = 0$ and compute $\bar{x}_N$ of Algorithm 2 as a weighted average of the search points $x_2, ..., x_{N+1}$, i.e.,

$$\bar{x}_N = \left(\sum_{k=2}^{N+1} \theta_k\right)^{-1} s_{N+1} = \left(\sum_{k=2}^{N+1} \theta_k\right)^{-1} \sum_{k=2}^{N+1} (\theta_k x_k), \tag{3.8}$$

while the output of Algorithm 1 is taken as a weighted average of $x_1, ..., x_N$.

Thirdly, it can be easily seen from (2.7), (3.1), and (3.6) that if X is bounded, then

$$\begin{aligned} \mathbb{E}[\|G_i(x,\xi)\|_{i,*}^2] &\leq 2\|g_i(x)\|_{i,*}^2 + 2\mathbb{E}\|G_i(x,\xi) - g_i(x)\|_{i,*}^2] \leq 2\|g_i(x)\|_{i,*}^2 + 2\sigma_i^2 \\ &\leq 2\left[2\|g_i(x) - g_i(x_1)\|_{i,*}^2 + 2\|g_i(x_1)\|_{i,*}^2\right] + 2\sigma_i^2 \\ &\leq 4L_i^2\|x - x_1\|_{i,*}^2 + 4\|g_i(x_1)\|_{i,*}^2 + 2\sigma_i^2 \\ &\leq 8L_i^2\mathcal{D}_i + 4\|g_i(x_1)\|_{i,*}^2 + 2\sigma_i^2, \quad i = 1,\ldots,b. \end{aligned} \tag{3.9}$$

Hence, we can directly apply Algorithm 1 in the previous section to problem (1.4), and its rate of convergence is readily given by Theorem 2.1 and 2.4. However, in this section we will show that by properly selecting $\{\theta_k\}$, $\{\gamma_k\}$, and $\{p_i\}$ in the above variant of the SBMD algorithm, we can significantly improve the dependence of the rate of convergence of the SBMD algorithm on the Lipschitz constants L_i, $i = 1,\ldots,b$.

We first discuss the main convergence properties of Algorithm 2 for convex stochastic composite optimization without assuming strong convexity.

THEOREM 3.1. *Suppose that $\{i_k\}$ in Algorithm 2 are uniformly distributed, i.e.,* (2.30) *holds. Also assume that $\{\gamma_k\}$ and $\{\theta_k\}$ are chosen such that for any $k \geq 1$,*

$$\gamma_k \leq \frac{1}{2\bar{L}} \quad \textit{with} \quad \bar{L} := \max_{i=1,\ldots,b} L_i, \tag{3.10}$$

$$\theta_{k+1} = b\gamma_k - (b-1)\gamma_{k+1}. \tag{3.11}$$

Then, under Assumption (1.3) *and* (3.6)*, we have, for any $N \geq 2$,*

$$E[\phi(\bar{x}_N) - \phi(x^*)] \leq \Big(\sum_{k=2}^{N+1} \theta_k\Big)^{-1} \left[(b-1)\gamma_1[\phi(x_1) - \phi(x^*)] + b\sum_{i=1}^{b} V_i(x_1, x^*) + \sigma^2 \sum_{k=1}^{N} \gamma_k^2\right], \tag{3.12}$$

where x^* is an arbitrary solution of problem (1.4) and σ is defined in (3.7).

Proof. For simplicity, let us denote $V_i(z,x) \equiv V_i(z^{(i)}, x^{(i)})$, $g_{i_k} \equiv g^{(i_k)}(x_k)$, and $V(z,x) = \sum_{i=1}^{b} p_i^{-1} V_i(z,x)$. Also denote $\zeta_k = (i_k, \xi_k)$ and $\zeta_{[k]} = (\zeta_1, \ldots, \zeta_k)$, and let $\delta_{i_k} = G_{i_k}(x_k, \xi_k) - g_{i_k}(x_k)$ and $\rho_{i_k} = U_{i_k}^T(x_{k+1} - x_k)$. By the definition of $\phi(\cdot)$ in (1.4) and (3.2), we have

$$\begin{aligned}\phi(x_{k+1}) &\leq f(x_k) + \langle g_{i_k}(x_k), \rho_{i_k}\rangle + \frac{L_{i_k}}{2}\|\rho_{i_k}\|_{i_k}^2 + \chi(x_{k+1}) \\ &= f(x_k) + \langle G_{i_k}(x_k), \rho_{i_k}\rangle + \frac{L_{i_k}}{2}\|\rho_{i_k}\|_{i_k}^2 + \chi(x_{k+1}) - \langle \delta_{i_k}, \rho_{i_k}\rangle. \end{aligned} \tag{3.13}$$

Moreover, it follows from the optimality condition of (3.4) (see, e.g., Lemma 1 of [14]) and (3.5) that

$$\begin{aligned}\langle G_{i_k}(x_k, \xi_k), \rho_{i_k}\rangle + \chi_{i_k}(x_{k+1}^{(i_k)}) \leq \langle G_{i_k}(x_k, \xi_k), x^{(i_k)} - x_k^{(i_k)}\rangle + \chi_{i_k}(x^{(i_k)}) \\ + \frac{1}{\gamma_k}\left[V_{i_k}(x_k, x) - V_{i_k}(x_{k+1}, x) - V_{i_k}(x_{k+1}, x_k)\right].\end{aligned}$$

Combining the above two inequalities and using (3.3), we obtain

$$\begin{aligned}\phi(x_{k+1}) \leq f(x_k) + \left\langle G_{i_k}(x_k, \xi_k), x^{(i_k)} - x_k^{(i_k)}\right\rangle + \chi_{i_k}(x^{(i_k)}) + \frac{1}{\gamma_k}\left[V_{i_k}(x_k, x) - V_{i_k}(x_{k+1}, x) - V_{i_k}(x_{k+1}, x_k)\right] \\ + \frac{L_{i_k}}{2}\|\rho_{i_k}\|_{i_k}^2 + \sum_{i \neq i_k} \chi_i(x_{k+1}^{(i)}) - \langle \delta_{i_k}, \rho_{i_k}\rangle.\end{aligned} \tag{3.14}$$

Noting that by the strong convexity of $\omega_i(\cdot)$, the Young's inequality, and (3.10), we have

$$\begin{aligned}-\frac{1}{\gamma_k} V_{i_k}(x_{k+1}, x_k) + \frac{L_{i_k}}{2}\|\rho_{i_k}\|_{i_k}^2 - \langle \delta_{i_k}, \rho_{i_k}\rangle &\leq -\left(\frac{1}{2\gamma_k} - \frac{L_{i_k}}{2}\right)\|\rho_{i_k}\|_{i_k}^2 - \langle \delta_{i_k}, \rho_{i_k}\rangle \\ &\leq \frac{\gamma_k \|\delta_{i_k}\|_*^2}{2(1-\gamma_k L_{i_k})} \leq \frac{\gamma_k \|\delta_{i_k}\|_*^2}{2(1-\gamma_k \bar{L})} \leq \gamma_k \|\delta_{i_k}\|_*^2.\end{aligned}$$

Also observe that by the definition of x_{k+1} in (3.5), (2.12), and the definition of $V(\cdot,\cdot)$, we have $\sum_{i \neq i_k} \chi_i(x_{k+1}^{(i)}) = \sum_{i \neq i_k} \chi_i(x_k^{(i)})$ and $V_{i_k}(x_k, x) - V_{i_k}(x_{k+1}, x) = [V(x_k, x) - V(x_{k+1}, x)]/b$. Using these observations, we conclude from (3.14) that

$$\begin{aligned}\phi(x_{k+1}) \leq f(x_k) + \left\langle G_{i_k}(x_k, \xi_k), x^{(i_k)} - x_k^{(i_k)}\right\rangle + \frac{1}{b\gamma_k}\left[V(x_k, x) - V(x_{k+1}, x)\right] \\ + \gamma_k \|\delta_{i_k}\|_*^2 + \sum_{i \neq i_k} \chi_i(x_k^{(i)}) + \chi_{i_k}(x^{(i_k)}).\end{aligned} \tag{3.15}$$

Now noting that

$$\mathbb{E}_{\zeta_k}\left[\left\langle G_{i_k}(x_k,\xi_k), x^{(i_k)} - x_k^{(i_k)}\right\rangle |\zeta_{[k-1]}\right] = \frac{1}{b}\sum_{i=1}^{b}\mathbb{E}_{\xi_k}\left[\left\langle G_i(x_k,\xi_k), x^{(i)} - x_k^{(i)}\right\rangle |\zeta_{[k-1]}\right]$$
$$= \frac{1}{b}\langle g(x_k), x - x_k\rangle \le \frac{1}{b}[f(x) - f(x_k)], \tag{3.16}$$

$$\mathbb{E}_{\zeta_k}\left[\|\delta_{i_k}\|_*^2 |\zeta_{[k-1]}\right] = \frac{1}{b}\sum_{i=1}^{b}\mathbb{E}_{\xi_k}\left[\|G_i(x_k,\xi_k) - g_i(x_k)\|_{i,*}^2 |\zeta_{[k-1]}\right] \le \frac{1}{b}\sum_{i=1}^{b}\sigma_i^2 = \frac{\sigma^2}{b}, \tag{3.17}$$

$$\mathbb{E}_{\zeta_k}\left[\sum_{i\neq i_k}\chi_i(x_k^{(i)})|\zeta_{[k-1]}\right] = \frac{1}{b}\sum_{j=1}^{b}\sum_{i\neq j}\chi_i(x_k^{(i)}) = \frac{b-1}{b}\chi(x_k), \tag{3.18}$$

$$\mathbb{E}_{\zeta_k}\left[\chi_{i_k}(x^{(i_k)})|\zeta_{[k-1]}\right] = \frac{1}{b}\sum_{i=1}^{b}\chi_i(x^{(i)} = \frac{1}{b}\chi(x), \tag{3.19}$$

we conclude from (3.15) that

$$\mathbb{E}_{\zeta_k}\left[\phi(x_{k+1}) + \frac{1}{b\gamma_k V(x_{k+1},x)}|\zeta_{[k-1]}\right] \le f(x_k) + \frac{1}{b}[f(x) - f(x_k)] + \frac{1}{b}\chi(x) + \frac{1}{b\gamma_k}[V(x_k,x)]$$
$$+ \frac{\gamma_k}{b}\sigma^2 + \frac{b-1}{b}\chi(x_k) + \frac{1}{b}\chi(x)$$
$$= \frac{b-1}{b}\phi(x_k) + \frac{1}{b}\phi(x) + \frac{1}{b\gamma_k}[V(x_k,x) - V(x_{k+1},x)] + \frac{\gamma_k}{b}\sigma^2,$$

which implies that

$$b\gamma_k\mathbb{E}[\phi(x_{k+1}) - \phi(x)] + \mathbb{E}[V(x_{k+1},x)] \le (b-1)\gamma_k\mathbb{E}[\phi(x_k) - \phi(x)] + \mathbb{E}[V(x_k,x)] + \gamma_k^2\sigma^2. \tag{3.20}$$

Now, summing up the above inequalities (with $x = x^*$) for $k = 1,\ldots,N$, and noting that $\theta_{k+1} = b\gamma_k - (b-1)\gamma_{k+1}$, we obtain

$$\sum_{k=2}^{N}\theta_k\mathbb{E}[\phi(x_k)-\phi(x^*)]+b\gamma_N\mathbb{E}[\phi(x_{N+1})-\phi(x^*)]+\mathbb{E}[V(x_{N+1},x)] \le (b-1)\gamma_1[\phi(x_1)-\phi(x^*)]+V(x_1,x^*)+\sigma^2\sum_{k=1}^{N}\gamma_k^2,$$

Using the above inequality and the facts that $V(\cdot,\cdot) \ge 0$ and $\phi(x_{N+1}) \ge \phi(x^*)$, we conclude

$$\sum_{k=2}^{N+1}\theta_k\mathbb{E}[\phi(x_k) - \phi(x^*)] \le (b-1)\gamma_1[\phi(x_1) - \phi(x^*)] + V(x_1,x^*) + \sigma^2\sum_{k=1}^{N}\gamma_k^2,$$

which, in view of (3.7), (3.8) and the convexity of $\phi(\cdot)$, clearly implies (3.12). ∎

The following corollary describes a specialized convergence result of Algorithm 2 for solving convex stochastic composite optimization problems after properly selecting $\{\gamma_k\}$.

COROLLARY 3.2. *Suppose that $\{p_i\}$ in Algorithm 2 are set to* (2.30). *Also assume that $\{\gamma_k\}$ are set to*

$$\gamma_k = \gamma = \min\left\{\frac{1}{2\bar{L}}, \frac{\tilde{D}}{\sigma}\sqrt{\frac{b}{N}}\right\} \tag{3.21}$$

for some $\tilde{D} > 0$, and $\{\theta_k\}$ are set to (3.11). Then, under Assumptions (1.3) *and* (3.6), *we have*

$$\mathbb{E}\left[\phi(\bar{x}_N) - \phi(x^*)\right] \le \frac{(b-1)[\phi(x_1) - \phi(x^*)]}{N} + \frac{2b\bar{L}\sum_{i=1}^{b}V_i(x_1,x^*)}{N}$$
$$+ \frac{\sigma\sqrt{b}}{\sqrt{N}}\left[\frac{\sum_{i=1}^{b}V_i(x_1,x^*)}{\tilde{D}} + \tilde{D}\right]. \tag{3.22}$$

where x^* is the optimal solution of problem (1.4).

Proof. It follows from (3.11) and (3.21) that $\theta_k = \gamma_k = \gamma$, $k = 1, \ldots, N$. Using this observation and Theorem 3.1, we obtain

$$\mathbb{E}\left[\phi(\bar{x}_N) - \phi(x^*)\right] \le \frac{(b-1)[\phi(x_1) - \phi(x^*)]}{N} + \frac{b\sum_{i=1}^b V_i(x_1, x^*)}{N\gamma} + \gamma\sigma^2,$$

which, in view of (3.21), then implies (3.22). ∎

We now add a few remarks about the results obtained in Corollary 3.2. First, in view of (3.22), an optimal selection of $\tilde{D}$ would be $\sqrt{\sum_{i=1}^b V_i(x_1, x^*)}$. In this case, (3.22) reduces to

$$\begin{aligned}\mathbb{E}\left[\phi(\bar{x}_N) - \phi(x^*)\right] &\le \frac{(b-1)[\phi(x_1) - \phi(x^*)]}{N} + \frac{2b\bar{L}\sum_{i=1}^b V_i(x_1, x^*)}{N} + \frac{2\sigma\sqrt{b}\sqrt{\sum_{i=1}^b \mathcal{D}_i}}{\sqrt{N}} \\ &\le \frac{(b-1)[\phi(x_1) - \phi(x^*)]}{N} + \frac{2b\bar{L}\sum_{i=1}^b \mathcal{D}_i}{N} + \frac{2\sigma\sqrt{b}\sqrt{\sum_{i=1}^b \mathcal{D}_i}}{\sqrt{N}}. \end{aligned} \tag{3.23}$$

Second, if we directly apply Algorithm 1 to problem (1.4), then, in view of (2.25) and (3.9), we have

$$\begin{aligned}\mathbb{E}[\phi(\bar{x}_N) - \phi(x^*)] &\le 2\sqrt{\sum_{i=1}^b \left[4L_i^2\mathcal{D}_i + 2\|g_i(x_1)\|_{i,*}^2 + \sigma_i^2\right]}\frac{\sqrt{b}\sqrt{\sum_{i=1}^b \mathcal{D}_i}}{\sqrt{N}} \\ &\le \frac{4b\bar{L}\sum_{i=1}^b \mathcal{D}_i}{\sqrt{N}} + 2\sqrt{\sum_{i=1}^b \left(2\|g_i(x_1)\|_{i,*}^2 + \sigma_i^2\right)}\frac{\sqrt{b}\sqrt{\sum_{i=1}^b \mathcal{D}_i}}{\sqrt{N}}. \end{aligned} \tag{3.24}$$

Clearly, the bound in (3.23) has a much weaker dependence on the Lipschitz constant $\bar{L}$ than the one in (3.24). In particular, we can see that $\bar{L}$ can be as large as $\mathcal{O}(\sqrt{N})$ without affecting the bound in (3.23), after disregarding some other constant factors.

In the remaining part of this section, we consider the case when the objective function is strongly convex, i.e., the function $f(\cdot)$ in (1.4) satisfies (2.28). Similar to the previous section, we also assume that the prox-functions $V_i(\cdot,\cdot)$, $i = 1, \ldots, b$, satisfy the quadratic growth condition (2.29). The following theorem describes some convergence properties of the SBMD algorithm for solving strongly convex composite problems.

THEOREM 3.3. *Suppose that (2.28), (2.29), and (2.30) hold. Also assume that the parameters $\{\gamma_k\}$ and $\{\theta_k\}$ are chosen such that for any $k \ge 1$,*

$$\gamma_k \le \min\left\{\frac{1}{2\bar{L}}, \frac{bQ}{\mu}\right\}, \tag{3.25}$$

$$\theta_{k+1} = \frac{b\gamma_k}{\Gamma_k} - \frac{(b-1)\gamma_{k+1}}{\Gamma_{k+1}} \quad \textit{with} \quad \Gamma_k = \begin{cases} 1 & k = 1 \\ \Gamma_{k-1}(1 - \frac{\gamma_k\mu}{bQ}) & k \ge 2. \end{cases} \tag{3.26}$$

Then, for any $N \ge 2$, we have

$$\mathbb{E}[\phi(\bar{x}_N) - \phi(x^*)] \le \left[\sum_{k=2}^{N+1}\theta_k\right]^{-1}\left[\left(b - \mu\gamma_1 Q\right)\sum_{i=1}^b V_i(x_1, x^*) + (b-1)\gamma_1[\phi(x_1) - \phi(x^*)] + \sum_{k=1}^N \frac{\gamma_k^2}{\Gamma_k}\sigma^2\right], \tag{3.27}$$

where x^* is the optimal solution of problem (1.4).

Proof. Observe that by the strong convexity of $f(\cdot)$, the relation in (3.16) can be strengthened to

$$\mathbb{E}_{\zeta_k}\left[\left\langle G_{i_k}(x_k,\xi_k), x^{(i_k)} - x_k^{(i_k)}\right\rangle |\zeta_{[k-1]}\right] = \frac{1}{b}\langle g(x_k), x - x_k\rangle \le \frac{1}{b}[f(x) - f(x_k) - \frac{\mu}{2}\|x - x_k\|^2].$$

Using this observation, (3.17), (3.18), and (3.19), we conclude from (3.15) that

$$\begin{aligned}\mathbb{E}_{\zeta_k}\left[\phi(x_{k+1}) + \frac{1}{b\gamma_k}V(x_{k+1},x)|\zeta_{[k-1]}\right] &\le f(x_k) + \frac{1}{b}\Big[f(x) - f(x_k) - \frac{\mu}{2}\|x - x_k\|^2\Big] + \frac{1}{b\gamma_k}V(x_k,x)\\ &\quad + \frac{\gamma_k}{b}\sigma^2 + \frac{b-1}{b}\chi(x_k) + \frac{1}{b}\chi(x)\\ &\le \frac{b-1}{b}\phi(x_k) + \frac{1}{b}\phi(x) + \Big(\frac{1}{b\gamma_k} - \frac{\mu}{b^2Q}\Big)V(x_k,x) + \frac{\gamma_k}{b}\sigma^2,\end{aligned}$$

where the last inequality follows from (2.36). By taking expectation w.r.t. $\xi_{[k-1]}$ on both sides of the above inequality, we conclude that, for any $k \ge 1$,

$$\mathbb{E}[V(x_{k+1},x^*)] \le \Big(1 - \frac{\mu\gamma_k}{bQ}\Big)\mathbb{E}[V(x_k,x^*)] + (b-1)\gamma_k\mathbb{E}[\phi(x_k) - \phi(x^*)] - b\gamma_k\mathbb{E}\left[\phi(x_{k+1}) - \phi(x^*)\right] + \gamma_k^2\sigma^2,$$

which, in view of Lemma 2.3 (with $a_k = 1 - \gamma_k\mu/(bQ)$ and $A_k = \Gamma_k$ and $B_k = (b-1)\gamma[\phi(x_k) - \phi(x^*)] - b\gamma_k E\left[\phi(x_{k+1}) - \phi(x^*)\right] + \gamma_k^2\sigma^2$), then implies that

$$\begin{aligned}\frac{1}{\Gamma_N}[V(x_{k+1},x^*)] &\le (1 - \frac{\mu\gamma_1}{bQ})V(x_1,x^*) + (b-1)\sum_{k=1}^N \frac{\gamma_k}{\Gamma_k}[\phi(x_k) - \phi(x^*)]\\ &\quad - b\sum_{k=1}^N \frac{\gamma_k}{\Gamma_k}[\phi(x_{k+1}) - \phi(x^*)] + \sum_{k=1}^N \frac{\gamma_k^2}{\Gamma_k}\sigma^2\\ &\le (1 - \frac{\mu\gamma_1}{bQ})V(x_1,x^*) + (b-1)\gamma_1[\phi(x_1) - \phi(x^*)]\\ &\quad - \sum_{k=2}^{N+1}\theta_k[\phi(x_k) - \phi(x^*)] + \sum_{k=1}^N \frac{\gamma_k^2}{\Gamma_k}\sigma^2,\end{aligned}$$

where the last inequality follows from (3.26) and the fact that $\phi(x_{N+1}) - \phi(x^*) \ge 0$. Noting that $V(x_{N+1},x^*) \ge 0$, we conclude from the above inequality that

$$\sum_{k=2}^{N+1}\theta_k\mathbb{E}[\phi(x_k) - \phi(x^*)] \le (1 - \frac{\mu\gamma_1}{bQ})V(x_1,x^*) + (b-1)\gamma_1[\phi(x_1) - \phi(x^*)] + \sum_{k=1}^N \frac{\gamma_k^2}{\Gamma_k}\sigma^2.$$

Our result immediately follows from the above inequality, the convexity of $\phi(\cdot)$, and (3.8). ∎

Below we specialize the rate of convergence of the SBMD method for solving strongly convex composite problems with a proper selection of $\{\gamma_k\}$.

COROLLARY 3.4. *Suppose that (2.28), (2.29), and (2.30) hold. Also assume that $\{\theta_k\}$ are set to (3.26) and*

$$\gamma_k = 2bQ/(\mu(k+k_0)) \quad \forall k \ge 1, \tag{3.28}$$

where

$$k_0 := \left\lfloor \frac{4bQ\bar{L}}{\mu} \right\rfloor.$$

Then, for any $N \geq 2$, we have

$$\mathbb{E}[\phi(\bar{x}_N) - \phi(x^*)] \leq \frac{\mu Q k_0^2}{N(N+1)} \sum_{i=1}^{b} V_i(x_1, x^*) + \frac{2Q(b-1)k_0}{N(N+1)}[\phi(x_1) - \phi(x^*)] + \frac{4b\sigma^2}{\mu Q(N+1)}, \tag{3.29}$$

where x^* is the optimal solution of problem (1.4).

Proof. We can check that

$$\gamma_k = \frac{2bQ}{\mu(k + \left\lfloor \frac{4bQ\bar{L}}{\mu} \right\rfloor)} \leq \frac{1}{2\bar{L}}.$$

It can also be easily seen from the definition of γ_k and (3.26) that

$$\Gamma_k = \frac{k_0(k_0+1)}{(k+k_0)(k+k_0-1)}, \quad 1 - \frac{\gamma_1 \mu}{bQ} = \frac{k_0 - 1}{k_0 + 1}, \quad \forall k \geq 1, \tag{3.30}$$

$$\theta_k = \frac{b\gamma_k}{\Gamma_k} - \frac{(b-1)\gamma_{k+1}}{\Gamma_{k+1}} = \frac{2bkQ + 2bQ(k_0 - b)}{\mu k_0(k_0+1)} \geq \frac{2bk}{\mu Q k_0(k_0+1)}, \tag{3.31}$$

and hence that

$$\sum_{k=2}^{N+1} \theta_k \geq \frac{bQN(N+1)}{\mu k_0(k_0+1)}, \quad \sum_{k=1}^{N} \frac{\gamma_k^2}{\Gamma_k} = \frac{4b^2Q^2}{\mu^2 k_0(k_0+1)} \sum_{k=1}^{N} \frac{k + k_0 - 1}{k + k_0} \leq \frac{4Nb^2Q^2}{\mu^2 k_0(k_0+1)}. \tag{3.32}$$

By using the above observations and (3.27), we have

$$\begin{aligned}
\mathbb{E}[\phi(\bar{x}_N) - \phi(x^*)] &\leq \left(\sum_{k=2}^{N+1} \theta_k \right)^{-1} \left[\left(1 - \frac{\mu\gamma_1}{bQ}\right) V(x_1, x^*) + (b-1)\gamma_1[\phi(x_1) - \phi(x^*)] + \sum_{k=1}^{N} \frac{\gamma_k^2}{\Gamma_k}\sigma^2 \right] \\
&\leq \frac{\mu k_0(k_0+1)}{bQN(N+1)} \left[\frac{k_0 - 1}{k_0 + 1} V(x_1, x^*) + \frac{2b(b-1)Q}{\mu(k_0+1)}[\phi(x_1) - \phi(x^*)] + \frac{4Nb^2Q^2\sigma^2}{\mu^2 k_0(k_0+1)} \right] \\
&\leq \frac{\mu k_0^2}{bQN(N+1)} V(x_1, x^*) + \frac{2(b-1)k_0}{N(N+1)}[\phi(x_1) - \phi(x^*)] + \frac{4bQ\sigma^2}{\mu(N+1)},
\end{aligned}$$

where the second inequality follows (3.30), (3.31) and (3.32). ∎

It is interesting to observe that, in view of (3.29) and the definition of k_0, the Lipschitz constant $\bar{L}$ can be as large as $\mathcal{O}(\sqrt{N})$ without affecting the rate of convergence of the SBMD algorithm, after disregarding other constant factors, for solving strongly convex stochastic composite optimization problems.

4. SBMD Algorithm for nonconvex composite optimization. In this section we still consider composite optimization problems given in the form of (1.4). However, we assume that the smooth component $f(\cdot)$ is not necessarily convex, while the nonsmooth component $\chi(\cdot)$ is still convex and separable (i.e., (3.3) holds). In addition, we assume that the prox-functions satisfy the quadratic growth condition in (2.29). Our goal is to show that the SBMD algorithm, when employed with a certain randomization scheme, can also be used to solve these nonconvex stochastic composite problems.

In order to discuss the convergence of the SBMD algorithm for solving nonconvex composite problems, we need to first define an appropriate termination criterion. Note that if $X = \mathbb{R}^n$ and $\chi(x) = 0$, then a natural way to evaluate the quality of a candidate solution x will be $\|\nabla f(x)\|$. For more general nonconvex composite problems, we introduce the notion of composite projected gradient so as to evaluate the quality of a candidate

solution (see [26, 16, 17, 6, 10] for some related discussions). More specifically, for a given $x \in X$, $y \in \mathbb{R}^n$ and a constant $\gamma > 0$, we define $\mathcal{G}(x, y, \gamma) \equiv (\mathcal{G}_1(x, y, \gamma), \ldots, \mathcal{G}_b(x, y, \gamma))$ by

$$\mathcal{G}_i(x, y, \gamma) := \frac{1}{\gamma}[U_i^T x - \mathcal{P}_i(U_i^T x, U_i^T y, \gamma)], \quad i = 1, \ldots, b, \tag{4.1}$$

where $\mathcal{P}_i$ is defined in (3.4). In particular, if $y = g(x)$, then we call $\mathcal{G}(x, g(x), \gamma)$ the composite projected gradient of x w.r.t. γ. It can be easily seen that $\mathcal{G}(x, g(x), \gamma) = g(x)$ when $X = \mathbb{R}^n$ and $\chi(x) = 0$. Proposition 4.1 below relates the composite projected gradient to the first-order optimality condition of the composite problem under a more general setting.

PROPOSITION 4.1. *Let $x \in X$ be given and $\mathcal{G}(x, y, \gamma)$ be defined as in (4.1) for some $\gamma > 0$. Also let us denote $x^+ := x - \gamma\mathcal{G}(x, g(x), \gamma)$. Then there exists $p_i \in \partial\chi_i(U_i^T x^+)$ s.t.*

$$U_i^T g(x^+) + p_i \in -\mathcal{N}_{X_i}(U_i^T x^+) + \mathcal{B}_i\left((L_i + Q\gamma)\|\mathcal{G}(x, g(x), \gamma)\|_i\right), \quad i = 1, \ldots, b, \tag{4.2}$$

where $\mathcal{B}_i(\epsilon) := \{v \in \mathbb{R}^{n_i} : \|v\|_{i,*} \le \epsilon\}$ and $\mathcal{N}_{X_i}$ denotes the normal cone of X_i at $U_i^T x$.

Proof. By the definition of x^+, (3.4), and (4.1), we have $U_i^T x^+ = \mathcal{P}_i(U_i^T x, U_i^T g(x), \gamma)$. Using the above relation and the optimality condition of (3.4), we conclude that there exists $p_i \in \partial\chi_i(U_i^T x^+)$ s.t.

$$\langle U_i^T g(x) + \frac{1}{\gamma}\left[\nabla\omega_i(U_i^T x^+) - \nabla\omega_i(U_i^T x)\right] + p_i, u - U_i^T x^+\rangle \ge 0, \quad \forall u \in X_i.$$

Now, denoting $\zeta = U_i^T[g(x) - g(x^+) + \frac{1}{\gamma}\left[\nabla\omega_i(U_i^T x^+) - \nabla\omega_i(U_i^T x^+)\right]$, we conclude from the above relation that $U_i^T g(x^+) + p_i + \zeta \in -\mathcal{N}_{X_i}(U_i^T x^+)$. Also noting that, by $\|U_i^T[g(x^+) - g(x)]\|_{i,*} \le L_i\|U_i^T(x^+ - x)\|_i$ and $\|\nabla\omega_i(U_i^T x^+) - \nabla\omega_i(U_i^T x^+)\|_{i,*} \le Q\|U_i^T(x^+ - x)\|_i$,

$$\begin{aligned}\|\zeta\|_{i,*} &\le \left(L_i + \frac{Q}{\gamma}\right)\|U_i^T(x^+ - x)\|_i = \left(L_i + \frac{Q}{\gamma}\right)\gamma\|U_i^T\mathcal{G}(x, g(x), \gamma)\|_i \\ &= (L_i + Q\gamma)\|U_i^T\mathcal{G}(x, g(x), \gamma)\|_i.\end{aligned}$$

Relation (4.2) then immediately follows from the above two relations. ∎

A common practice in the gradient descent methods for solving nonconvex problems (for the simple case when $X = \mathbb{R}^n$ and $\chi(x) = 0$) is to choose the output solution $\bar{x}_N$ so that

$$\|g(\bar{x}_N)\|_* = \min_{k=1,\ldots,N}\|g(x_k)\|_*, \tag{4.3}$$

where x_k, $k = 1, \ldots, N$, is the trajectory generated by the gradient descent method (see, e.g., [26]). However, such a procedure requires the computation of the whole vector $g(x_k)$ at each iteration and hence can be expensive if n is large. In this section, we address this problem by introducing a randomization scheme into the SBMD algorithm as follows. Instead of taking the best solution from the trajectory as in (4.3), we randomly select $\bar{x}_N$ from $x_1, \ldots, x_N$ according to a certain probability distribution. The basic scheme of this algorithm is described as follows.

Algorithm 3 The Nonconvex SBMD Algorithm

Let $x_1 \in X$, stepsizes $\{\gamma_k\}_{k\geq 1}$ s.t. $\gamma_k < 2/L_i$, $i = 1, \ldots, b$, and probabilities $p_i \in [0,1]$, $i = 1, \ldots, b$, s.t. $\sum_{i=1}^b p_i = 1$ be given.

for $k = 1, \ldots, N$ **do**

1. Generate a random variable i_k according to (2.10).
2. Compute the partial (stochastic) gradient G_{i_k} of $f(\cdot)$ at x_k satisfying

$$\mathbb{E}[G_{i_k}] = U_{i_k}^T g(x_k) \quad \text{and} \quad \mathbb{E}[\|G_{i_k} - U_{i_k}^T g(x_k)\|_{i_k,*}] \leq \bar{\sigma}_k^2, \tag{4.4}$$

and update x_k by (3.5).

end for

Set $\bar{x}_N = x_R$ randomly according to

$$\text{Prob}\,(R = k) = \frac{\gamma_k \min\limits_{i=1,\ldots,b} p_i \left(1 - \frac{L_i}{2}\gamma_k\right)}{\sum_{k=1}^N \gamma_k \min\limits_{i=1,\ldots,b} p_i \left(1 - \frac{L_i}{2}\gamma_k\right)}, \quad k = 1, ..., N. \tag{4.5}$$

We add a few remarks about the above nonconvex SBMD algorithm. Firstly, observe that we have not yet specified how the gradient G_{i_k} is computed. If the problem is deterministic, then we can simply set $G_{i_k} = U_{i_k}^T g(x_k)$ and $\bar{\sigma}_k = 0$. However, if the problem is stochastic, then the computation of G_{i_k} is a little complicated and we cannot simply set $G_{i_k} = U_{i_k}^T \nabla F(x_k, \xi_k)$ (see Corollary 4.5).

Before establishing the convergence properties of the above nonconvex SBMD algorithm, we will first present a technical result which summarizes some important properties about the composite prox-mapping and projected gradient. Note that this result generalizes Lemma 1 and 2 in [10].

LEMMA 4.2. *Let x_{k+1} be defined in (3.5), and denote $\mathcal{G}_k \equiv \mathcal{G}_X(x_k, g(x_k), \gamma_k)$ and $\tilde{\mathcal{G}}_k \equiv \mathcal{G}_{i_k}(x_k, U_{i_k} G_{i_k}, \gamma_k)$. We have*

$$\langle G_{i_k}, \tilde{\mathcal{G}}_k \rangle \geq \|\tilde{\mathcal{G}}_k\|^2 + \frac{1}{\gamma_k}\left[\chi(x_{k+1}) - \chi(x_k)\right], \tag{4.6}$$

$$\|\tilde{\mathcal{G}}_k - U_{i_k}^T \mathcal{G}_k\|_{i_k} \leq \|G_{i_k} - U_{i_k} g(x_k)\|_{i_k,*}. \tag{4.7}$$

Proof. By the optimality condition of (3.4) and the definition of x_{k+1} in (3.5), there exists $p \in \partial\chi_{i_k}(x_{k+1})$ such that

$$\langle G_{i_k} + \frac{1}{\gamma_k}\left[\nabla\omega_{i_k}(U_{i_k}^T x_{k+1}) - \nabla\omega_{i_k}(U_{i_k}^T x_k)\right] + p, \frac{1}{\gamma_k}(u - U_{i_k}^T x_{k+1})\rangle \geq 0, \quad \forall u \in X_{i_k}. \tag{4.8}$$

Letting $u = U_{i_k}^T x_k$ in the above inequality and re-arranging terms, we obtain

$$\begin{array}{rcl} \langle G_{i_k}, \frac{1}{\gamma_k} U_{i_k}^T(x_k - x_{k+1})\rangle & \geq & \frac{1}{\gamma_k^2}\langle \nabla w_{i_k}(U_{i_k}^T x_{k+1}) - \nabla w_{i_k}(U_{i_k}^T x_k), U_{i_k}^T(x_{k+1} - x_k)\rangle + \langle p, \frac{1}{\gamma_k} U_{i_k}^T(x_k - x_{k+1})\rangle \\ & \geq & \frac{1}{\gamma_k^2}\langle \nabla w_{i_k}(U_{i_k}^T x_{k+1}) - \nabla w_{i_k}(U_{i_k}^T x_k), U_{i_k}^T(x_{k+1} - x_k)\rangle \\ & & + \frac{1}{\gamma_k}\left[\chi_{i_k}(U_{i_k}^T x_{k+1}) - \chi_{i_k}(U_{i_k}^T x_k)\right] \\ & \geq & \frac{1}{\gamma_k^2}\|U_{i_k}^T(x_{k+1} - x_k)\|^2 + \frac{1}{\gamma_k}\left[\chi_{i_k}(U_{i_k}^T x_{k+1}) - \chi_{i_k}(U_{i_k}^T x_k)\right] \\ & = & \frac{1}{\gamma_k^2}\|U_{i_k}^T(x_{k+1} - x_k)\|^2 + \frac{1}{\gamma_k}[\chi(x_{k+1}) - \chi(x_k)], \end{array} \tag{4.9}$$

where the second and third inequalities, respectively, follow from the convexity of χ_{i_k} and the strong convexity of ω, and the last identity follows from the definition of x_{k+1} and the separability assumption about χ in (3.3). The above inequality, in view of the fact that $\gamma_k \tilde{\mathcal{G}}_k = U_{i_k}^T(x_k - x_{k+1})$ due to (4.1) and (3.5), then implies (4.6).

Now we show that (4.7) holds. Let us denote $x_{k+1}^+ = x_k - \gamma_k \mathcal{G}_k$. By the optimality condition of (3.4) and the definition of $\mathcal{G}_k$, we have, for some $q \in \partial \chi_{i_k}(x_{k+1}^+)$,

$$\langle U_{i_k}^T g(x_k) + \frac{1}{\gamma_k}\left[\nabla \omega_{i_k}(U_{i_k}^T x_{k+1}^+) - \nabla \omega_{i_k}(U_{i_k}^T x_k)\right] + q, \frac{1}{\gamma_k}(u - U_{i_k}^T x_{k+1}^+)\rangle \geq 0, \quad \forall u \in X_{i_k}. \tag{4.10}$$

Letting $u = U_{i_k}^T x_{k+1}^+$ in (4.8) and using an argument similar to (4.9), we have

$$\begin{array}{rcl} \langle G_{i_k}, \frac{1}{\gamma_k} U_{i_k}^T (x_{k+1}^+ - x_{k+1})\rangle & \geq & \frac{1}{\gamma_k^2}\langle \nabla w_{i_k}(U_{i_k}^T x_{k+1}) - \nabla w_{i_k}(U_{i_k}^T x_k), U_{i_k}^T(x_{k+1} - x_{k+1}^+)\rangle \\ & & + \frac{1}{\gamma_k}\left[\chi_{i_k}(U_{i_k}^T x_{k+1}) - \chi_{i_k}(U_{i_k}^T x_{k+1}^+)\right]. \end{array}$$

Similarly, letting $u = U_{i_k}^T x_{k+1}$ in (4.10), we have

$$\begin{array}{rcl} \langle U_{i_k}^T g(x_k), \frac{1}{\gamma_k} U_{i_k}^T (x_{k+1} - x_{k+1}^+)\rangle & \geq & \frac{1}{\gamma_k^2}\langle \nabla w_{i_k}(U_{i_k}^T x_{k+1}^+) - \nabla w_{i_k}(U_{i_k}^T x_k), U_{i_k}^T(x_{k+1}^+ - x_{k+1})\rangle \\ & & + \frac{1}{\gamma_k}\left[\chi_{i_k}(U_{i_k}^T x_{k+1}^+) - \chi_{i_k}(U_{i_k}^T x_{k+1})\right]. \end{array}$$

Summing up the above two inequalities, we obtain

$$\begin{array}{rcl} \langle G_{i_k} - U_{i_k}^T g(x_k), U_{i_k}^T(x_{k+1}^+ - x_{k+1})\rangle & \geq & \frac{1}{\gamma_k}\langle \nabla w_{i_k}(U_{i_k}^T x_{k+1}) - \nabla w_{i_k}(U_{i_k}^T x_{k+1}^+), U_{i_k}^T(x_{k+1} - x_{k+1}^+)\rangle \\ & \geq & \frac{1}{\gamma_k}\|U_{i_k}^T(x_{k+1} - x_{k+1}^+)\|_{i_k}^2, \end{array}$$

which, in view of the Cauchy-Schwarz inequality, then implies that

$$\frac{1}{\gamma_k}\|U_{i_k}^T(x_{k+1} - x_{k+1}^+)\|_{i_k} \leq \|G_{i_k} - U_{i_k}^T g(x_k)\|_{i_k,*}.$$

Using the above relation and (4.1), we have

$$\begin{array}{rcl} \|\tilde{\mathcal{G}}_k - U_{i_k}^T \mathcal{G}_k\|_{i_k} & = & \|\frac{1}{\gamma_k} U_{i_k}^T(x_k - x_{k+1}) - \frac{1}{\gamma_k} U_{i_k}^T(x_k - x_{k+1}^+)\|_{i_k} \\ & = & \frac{1}{\gamma_k}\|U_{i_k}^T(x_{k+1}^+ - x_{k+1})\|_{i_k} \leq \|G_{i_k} - U_{i_k}^T g(x_k)\|_{i_k,*}. \end{array}$$

■

We are now ready to describe the main convergence properties of the nonconvex SBMD algorithm.

THEOREM 4.3. *Let $\bar{x}_N = x_R$ be the output of the nonconvex SBMD algorithm. We have*

$$\mathbb{E}[\|\mathcal{G}_X(x_R, g(x_R), \gamma_R)\|^2] \leq \frac{\phi(x_1) - \phi^* + 2\sum_{k=1}^N \gamma_k \bar{\sigma}_k^2}{\sum_{k=1}^N \gamma_k \min\limits_{i=1,\ldots,b} p_i\left(1 - \frac{L_i}{2}\gamma_k\right)} \tag{4.11}$$

for any $N \geq 1$, where the expectation is taken w.r.t. i_k, G_{i_k}, and R.

Proof. Denote $g_k \equiv g(x_k)$, $\delta_k \equiv G_{i_k} - U_{i_k}^T g_k$, $\mathcal{G}_k \equiv \mathcal{G}_X(x_k, g_k, \gamma_k)$, and $\tilde{\mathcal{G}}_k \equiv \mathcal{G}_{i_k}(x_k, U_{i_k} G_{i_k}, \gamma_k)$ for any $k \geq 1$. Note that by (3.5) and (4.1), we have $x_{k+1} - x_k = -\gamma_k U_{i_k} \tilde{\mathcal{G}}_k$. Using this observation and (3.2), we have, for any $k = 1, \ldots, N$,

$$\begin{array}{rcl} f(x_{k+1}) & \leq & f(x_k) + \langle g_k, x_{k+1} - x_k\rangle + \frac{L_{i_k}}{2}\|x_{k+1} - x_k\|^2 \\ & = & f(x_k) - \gamma_k\langle g_k, U_{i_k}\tilde{\mathcal{G}}_k\rangle + \frac{L_{i_k}}{2}\gamma_k^2\|\tilde{\mathcal{G}}_k\|_{i_k}^2 \\ & = & f(x_k) - \gamma_k\langle G_{i_k}, \tilde{\mathcal{G}}_k\rangle + \frac{L_{i_k}}{2}\gamma_k^2\|\tilde{\mathcal{G}}_k\|_{i_k}^2 + \gamma_k\langle \delta_k, \tilde{\mathcal{G}}_k\rangle. \end{array}$$

Using the above inequality and Lemma 4.2, we obtain

$$f(x_{k+1}) \leq f(x_k) - \left[\gamma_k\|\tilde{\mathcal{G}}_k\|_{i_k}^2 + \chi(x_{k+1}) - \chi(x_k)\right] + \frac{L_{i_k}}{2}\gamma_k^2\|\tilde{\mathcal{G}}_k\|_{i_k}^2 + \gamma_k\langle \delta_k, \tilde{\mathcal{G}}_k\rangle,$$

which, in view of the fact that $\phi(x) = f(x) + \chi(x)$, then implies that

$$\phi(x_{k+1}) \le \phi(x_k) - \gamma_k \left(1 - \frac{L_{i_k}}{2}\gamma_k\right) \|\tilde{\mathcal{G}}_k\|_{i_k}^2 + \gamma_k \langle \delta_k, \tilde{\mathcal{G}}_k \rangle. \tag{4.12}$$

Also observe that by (4.7), the definition of $\tilde{\mathcal{G}}_k$, and the fact $U_{i_k}^T \mathcal{G}_k = \mathcal{G}_{X_{i_k}}(x_k, U_{i_k}^T g_k, \gamma_k)$,

$$\|\tilde{\mathcal{G}}_k - U_{i_k}^T \mathcal{G}_k\|_{i_k} \le \|G_{i_k} - U_{i_k}^T g_k\|_{i_k,*} = \|\delta_k\|_{i_k,*},$$

and hence that

$$\begin{aligned}
\|U_{i_k}^T \mathcal{G}_k\|_{i_k}^2 &= \|\tilde{\mathcal{G}}_k + U_{i_k}^T \mathcal{G}_k - \tilde{\mathcal{G}}_k\|_{i_k}^2 \le 2\|\tilde{\mathcal{G}}_k\|_{i_k}^2 + 2\|U_{i_k}^T \mathcal{G}_k - \tilde{\mathcal{G}}_k\|_{i_k} \\
&\le 2\|\tilde{\mathcal{G}}_k\|^2 + 2\|\delta_k\|_{i_k,*}^2, \\
\langle \delta_k, \tilde{\mathcal{G}}_k \rangle &= \langle \delta_k, U_{i_k}^T \mathcal{G}_k \rangle + \langle \delta_k, \tilde{\mathcal{G}}_k - U_{i_k}^T \mathcal{G}_k \rangle \le \langle \delta_k, U_{i_k}^T \mathcal{G}_k \rangle + \|\delta_k\|_{i_k,*} \|\tilde{\mathcal{G}}_k - U_{i_k}^T \mathcal{G}_k\|_{i_k} \\
&\le \langle \delta_k, U_{i_k}^T \mathcal{G}_k \rangle + \|\delta_k\|_{i_k,*}^2.
\end{aligned}$$

By using the above two bounds and (4.12), we obtain

$$\phi(x_{k+1}) \le \phi(x_k) - \gamma_k \left(1 - \frac{L_{i_k}}{2}\gamma_k\right) \left(\frac{1}{2}\|U_{i_k}^T \mathcal{G}_k\|^2 - \|\delta_k\|^2\right) + \gamma_k \langle \delta_k, U_{i_k}^T \mathcal{G}_k \rangle + \gamma_k \|\delta_k\|_{i_k,*}^2$$

for any $k = 1, \ldots, N$. Summing up the above inequalities and re-arranging the terms, we obtain

$$\begin{aligned}
\textstyle\sum_{k=1}^N \frac{\gamma_k}{2} \left(1 - \frac{L_{i_k}}{2}\gamma_k\right) \|U_{i_k}^T \mathcal{G}_k\|^2 &\le \phi(x_1) - \phi(x_{k+1}) + \textstyle\sum_{k=1}^N \left[\gamma_k \langle \delta_k, U_{i_k}^T \mathcal{G}_k \rangle + \gamma_k \|\delta_k\|_{i_k,*}^2\right] \\
&\quad + \textstyle\sum_{k=1}^N \gamma_k \left(1 - \frac{L_{i_k}}{2}\gamma_k\right) \|\delta_k\|_{i_k,*}^2 \\
&\le \phi(x_1) - \phi^* + \textstyle\sum_{k=1}^N \left[\gamma_k \langle \delta_k, U_{i_k}^T \mathcal{G}_k \rangle + 2\gamma_k \|\delta_k\|_{i_k,*}^2\right],
\end{aligned}$$

where the last inequality follows from the facts that $\phi(x_{k+1}) \ge \phi^*$ and $L_{i_k}\gamma_k^2 \|\delta_k\|_{i_k,*}^2 \ge 0$. Now denoting $\zeta_k = G_{i_k}$, $\zeta_{[k]} = \{\zeta_1, \ldots, \zeta_k\}$ and $i_{[k]} = \{i_1, \ldots, i_k\}$, taking expectation on both sides of the above inequality w.r.t. $\zeta_{[N]}$ and $i_{[N]}$, and noting that by (2.30) and (4.4),

$$\begin{aligned}
\mathbb{E}_{\zeta_k}\left[\langle \delta_k, U_{i_k}^T \mathcal{G}_k \rangle | i_{[k]}, \zeta_{[k-1]}\right] &= \mathbb{E}_{\zeta_k}\left[\langle G_{i_k} - U_{i_k}^T g_k, U_{i_k}^T \mathcal{G}_k \rangle | i_{[k]}, \zeta_{[k-1]}\right] = 0, \\
\mathbb{E}_{\zeta_{[N]}, i_{[N]}}[\|\delta_k\|_{i_k,*}^2] &\le \bar{\sigma}_k^2, \\
\mathbb{E}_{i_k}\left[\left(1 - \frac{L_{i_k}}{2}\gamma_k\right) \|U_{i_k}^T \mathcal{G}_k\|^2 | \zeta_{[k-1]}, i_{[k-1]}\right] &= \sum_{i=1}^b p_i \left(1 - \frac{L_i}{2}\gamma_k\right) \|U_i^T \mathcal{G}_k\|^2 \\
&\ge \sum_{i=1}^b \|U_i^T \mathcal{G}_k\|^2 \min_{i=1,\ldots,b} p_i \left(1 - \frac{L_i}{2}\gamma_k\right) \\
&= \|\mathcal{G}_k\|^2 \min_{i=1,\ldots,b} p_i \left(1 - \frac{L_i}{2}\gamma_k\right),
\end{aligned}$$

we conclude that

$$\sum_{k=1}^N \gamma_k \min_{i=1,\ldots,b} p_i \left(1 - \frac{L_i}{2}\gamma_k\right) \mathbb{E}_{\xi_{[N]}, i_{[N]}}\left[\|\mathcal{G}_k\|^2\right] \le \phi(x_1) - \phi^* + 2\sum_{k=1}^N \gamma_k \bar{\sigma}_k^2.$$

Dividing both sides of the above inequality by $\sum_{k=1}^N \min_{i=1,\ldots,b} p_i \left(1 - \frac{L_i}{2}\gamma_k\right)$, and using the probability distribution of R given in (4.5), we obtain (4.11). ∎

We now discuss some consequences for Theorem 4.3. More specifically, we discuss the rate of convergence of the nonconvex SBMD algorithm for solving deterministic and stochastic problems, respectively, in Corollaries 4.4 and 4.5.

COROLLARY 4.4. *Consider the deterministic case when $\bar{\sigma}_k = 0$, $k = 1, \ldots, N$, in (4.4). Suppose that the random variable $\{i_k\}$ are uniformly distributed (i.e., (2.30) holds). If $\{\gamma_k\}$ are set to*

$$\gamma_k = \frac{1}{\bar{L}}, k = 1, \ldots, N, \tag{4.13}$$

where $\bar{L}$ is defined in (3.10), then we have, for any $N \geq 1$,

$$\mathbb{E}[\|\mathcal{G}_X(x_R, g(x_R), \gamma_R)\|^2] \leq \frac{2b\bar{L}[\phi(x_1) - \phi^*]}{N}. \tag{4.14}$$

Proof. Noting that by our assumptions about p_i and (4.13), we have

$$\min_{i=1,\ldots,b} p_i \left(1 - \frac{L_i}{2}\gamma_k\right) = \frac{1}{b} \min_{i=1,\ldots,b} \left(1 - \frac{L_i}{2}\gamma_k\right) \geq \frac{1}{2b}, \tag{4.15}$$

which, in view of (4.11) and the fact that $\bar{\sigma}_k = 0$, then implies that, for any $N \geq 1$,

$$\mathbb{E}[\|\mathcal{G}_X(x_R, g(x_R), \gamma_R)\|^2] \leq \frac{2b[\phi(x_1) - \phi^*]}{N} \frac{1}{\bar{L}} = \frac{2b\bar{L}[\phi(x_1) - \phi^*]}{N}.$$

■

Now, let us consider the stochastic case when $f(\cdot)$ is given in the form of expectation (see (1.1)). Suppose that the norms $\|\cdot\|_i$ are inner product norms in $\mathbb{R}^{n_i}$ and that

$$\mathbb{E}[\|U_i \nabla F(x, \xi) - g_i(x)\|] \leq \sigma \quad \forall x \in X \tag{4.16}$$

for any $i = 1, \ldots, b$. Also assume that G_{i_k} is computed by using a mini-batch approach with size T_k, i.e.,

$$G_{i_k} = \frac{1}{T_k} \sum_{t=1}^{T_k} U_{i_k} \nabla F(x_k, \xi_{k,t}), \tag{4.17}$$

for some $T_k \geq 1$, where $\xi_{k,1}, \ldots, \xi_{k,T_k}$ are i.i.d. samples of ξ.

COROLLARY 4.5. *Assume that the random variables $\{i_k\}$ are uniformly distributed (i.e., (2.30) holds). Also assume that G_{i_k} is computed by (4.17) for $T_k = T$ and that $\{\gamma_k\}$ are set to (4.13). Then we have*

$$\mathbb{E}[\|\mathcal{G}_X(x_R, g(x_R), \gamma_R, h)\|^2] \leq \frac{2b\bar{L}[\phi(x_1) - \phi^*]}{N} + \frac{4\sum_{i=1}^b \sigma_i^2}{T} \tag{4.18}$$

for any $N \geq 1$, where $\bar{L}$ is defined in (3.10).

Proof. Denote $\delta_{k,t} \equiv U_{i_k}[\nabla F(x_k, \xi_{k,t}) - g(x_k)]$ and $S_t = \sum_{i=1}^t \delta_{k,i}$. Noting that $\mathbb{E}[\langle S_{t-1}, \delta_{k,t}\rangle | S_{t-1}] = 0$ for all $t = 1, \ldots, T_k$, we have

$$\begin{aligned}\mathbb{E}[\|S_{T_k}\|^2] &= \mathbb{E}\left[\|S_{T_k-1}\|^2 + 2\langle S_{T_k-1}, \delta_{k,T_k}\rangle + \|\delta_{k,T_k}\|^2\right] \\ &= \mathbb{E}[\|S_{T_k-1}\|^2] + \mathbb{E}[\|\delta_{k,T_k}\|^2] = \ldots = \sum_{t=1}^{T_k} \|\delta_{k,t}\|^2,\end{aligned}$$

which together with (4.17) then imply that the conditions in (4.4) hold with $\bar{\sigma}_k^2 = \sigma^2/T_k$. It then follows from the previous observation and (4.11) that

$$\begin{aligned}\mathbb{E}[\|\mathcal{G}_X(x_R, g(x_R), \gamma_R, h)\|^2] &\leq \frac{2b[\phi(x_1) - \phi^*]}{\frac{N}{\bar{L}}} + \frac{4b}{N}\sum_{k=1}^N \frac{\sigma^2}{T_k} \\ &\leq \frac{2b\bar{L}[\phi(x_1) - \phi^*]}{N} + \frac{4b\sigma^2}{T}.\end{aligned}$$

■

In view of Corollary 4.5, in order to find an ϵ solution of problem (1.4), we need to have

$$N = \mathcal{O}\left(\frac{b\bar{L}}{\epsilon}[\phi(x_1) - \phi^*]\right) \quad \text{and} \quad T = \mathcal{O}\left(\frac{b\sigma^2}{\epsilon}\right), \tag{4.19}$$

which implies that the total number of samples of ξ required can be bounded by

$$\mathcal{O}\left(b^2\bar{L}\sigma^2[\phi(x_1) - \phi^*]/\epsilon^2\right).$$

The previous bound is comparable, up to a constant factor b^2, to those obtained in [9, 10] for solving nonconvex SP problems without using block decomposition. Note that it is possible to derive and improve the large-deviation results associated with the above complexity results, by using a two-phase procedure similar to those in [9, 10]. However, the development of these results are more involved and hence the details are skipped.

5. Conclusions. In this paper, we study a new class of stochastic algorithms, namely the SBMD methods, by incorporating the block decomposition and an incremental block averaging scheme into the classic mirror-descent method, for solving different convex stochastic optimization problems, including general nonsmooth, smooth, composite and strongly convex problems. We establish the rate of convergence of these algorithms and show that their iteration cost can be considerably smaller than that of the mirror-descent methods. We also develop a nonconvex SBMD algorithm and establish its worst-case complexity for solving nonconvex stochastic composite optimization problems, by replacing the incremental block averaging scheme with a randomization scheme to compute the output solution. While this paper focuses on stochastic optimization, some of our results are also new in BCD type methods for deterministic optimization, which include the incorporation of new averaging/randomization schemes for computing the output solution, the derivation of large-deviation results for nonsmooth optimization and the analysis of the rate of convergence for nonsmooth strongly convex problems and the nonconvex composite optimization problems.

Acknowledgement: The authors would like to thank Professors Stephen J. Wright and Yurri Nesterov for their encouragement to study this topic.

REFERENCES

[1] A. Auslender and M. Teboulle. Interior gradient and proximal methods for convex and conic optimization. *SIAM Journal on Optimization*, 16:697–725, 2006.

[2] H.H. Bauschke, J.M. Borwein, and P.L. Combettes. Bregman monotone optimization algorithms. *SIAM Journal on Controal and Optimization*, 42:596–636, 2003.

[3] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31:167–175, 2003.

[4] A. Beck and L. Tetruashvili. On the convergence of block coordinate descent type methods. Technical report. submitted to *SIAM Journal on Optimization*.

[5] L.M. Bregman. The relaxation method of finding the common point convex sets and its application to the solution of problems in convex programming. *USSR Comput. Math. Phys.*, 7:200–217, 1967.

[6] C. D. Dang and G. Lan. On the convergence properties of non-euclidean extragradient methods for variational inequalities with generalized monotone operators. Manuscript, Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611, USA, April 2012. Available on http://www.optimization-online.org/.

[7] S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, I: a generic algorithmic framework. *SIAM Journal on Optimization*, 22:1469–1492, 2012.

[8] S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, II: shrinking procedures and optimal algorithms. *SIAM Journal on Optimization*, 2013. to appear.

[9] S. Ghadimi and G. Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. Technical report, Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611, USA, June 2012. *SIAM Journal on Optimization* (under second-round review).

[10] S. Ghadimi, G. Lan, and H. Zhang. Mini-batch stochastic approximation methods for constrained nonconvex stochastic programming. Manuscript, Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611, USA, August 2013.

[11] A. Juditsky, A. Nazin, A. B. Tsybakov, and N. Vayatis. Recursive aggregation of estimators via the mirror descent algorithm with average. *Problems of Information Transmission*, 41:n.4, 2005.

[12] A. Juditsky and Y. E. Nesterov. Primal-dual subgradient methods for minimizing uniformly convex functions. Manuscript.
[13] A. Juditsky, P. Rigollet, and A. B. Tsybakov. Learning by mirror averaging. *Annals of Statistics*, 36:2183–2206, 2008.
[14] G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1):365–397, 2012.
[15] G. Lan. Bundle-level type methods uniformly optimal for smooth and non-smooth convex optimization. Manuscript, Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611, USA, January 2013. Revision submitted to *Mathematical Programming*.
[16] G. Lan and R. D. C. Monteiro. Iteration-complexity of first-order penalty methods for convex programming. *Mathematical Programming*, 138:115–139, 2013.
[17] G. Lan and R. D. C. Monteiro. Iteration-complexity of first-order augmented lagrangian methods for convex programming. Manuscript, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA, May 2009. *Mathematical Programming* (under revision).
[18] G. Lan, A. S. Nemirovski, and A. Shapiro. Validation analysis of mirror descent stochastic approximation method. *Mathematical Programming*, 134:425–458, 2012.
[19] D. Leventhal and A. S. Lewis. Randomized methods for linear constraints: Convergence rates and conditioning. *Mathematics of Operations Research*, 35:641–654, 2010.
[20] Q. Lin, X. Chen, and J. Peña. A sparsity preserving stochastic gradient method for composite optimization. Manuscript, Carnegie Mellon University, PA 15213, April 2011.
[21] Z. Lu and L. Xiao. On the complexity analysis of randomized block-coordinate descent methods. Manuscript, 2013.
[22] Z.Q. Luo and P. Tseng. On the convergence of a matrix splitting algorithm for the symmetric monotone linear complementarity problem. *SIAM Journal on Control and Optimization*, 29:037 – 1060, 1991.
[23] A. Nedić. On stochastic subgradient mirror-descent algorithm with weighted averaging. 2012.
[24] A. S. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19:1574–1609, 2009.
[25] A. S. Nemirovski and D. Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience Series in Discrete Mathematics. John Wiley, XV, 1983.
[26] Y. E. Nesterov. *Introductory Lectures on Convex Optimization: a basic course*. Kluwer Academic Publishers, Massachusetts, 2004.
[27] Y. E. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120:221–259, 2006.
[28] Y. E. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. Technical report, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain, Feburary 2010.
[29] Y. E. Nesterov. Subgradient methods for huge-scale optimization problems. Technical report, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain, Feburary 2012.
[30] F. Niu, B. Recht, C. Ré, and S. J. Wright. Hogwild!: A lock-free approach to parallelizing stochastic gradient descent. Manuscript, Computer Sciences Department, University of Wisconsin-Madison, 1210 W Dayton St, Madison, WI 53706, 2011.
[31] B.T. Polyak. New stochastic approximation type procedures. *Automat. i Telemekh.*, 7:98–107, 1990.
[32] B.T. Polyak and A.B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM J. Control and Optimization*, 30:838–855, 1992.
[33] P. Richtárik and M. Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 2012. to appear.
[34] H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.
[35] S. Shalev-Shwartz and A. Tewari. Stochastic methods for l_1 regularized loss minimization. Manuscript, 2011. Submitted to *Journal of Machine Learning Research*.
[36] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, Philadelphia, 2009.
[37] M. Teboulle. Convergence of proximal-like algorithms. *SIAM Journal on Optimization*, 7:1069–1083, 1997.
[38] P. Tseng. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, 109:475494, 2001.
[39] P. Tseng and S. Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Mathematical Programming*, 117:387–423, 2009.
[40] S. J. Wright. Accelerated block-coordinate relaxation for regularized optimizations. Manuscript, University of Wisconsin-Madison, Madison, WI, 2010.
[41] L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *J. Mach. Learn. Res.*, pages 2543–2596, 2010.
[42] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell. Distance metric learning, with application to clustering with side-information. In *Advances in Neural Information Processing Systems 15*, pages 505–512. MIT Press, 2002.