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 optimization 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

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

Here \( X \subseteq \mathbb{R}^n \) is a closed convex set, \( \xi \) 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.,

\[ X = X_1 \times X_2 \times \cdots \times X_b, \] (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 comparison 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 classical 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., [29, 20, 30, 34, 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 asymptotically convergence behaviour (see, e.g., [23, 40] and also [41, 42]), much recent effort has been directed to the complexity analysis of these types of methods (see [29, 20, 36, 34, 4]). In particular, Nesterov [29] was among the first (see also Leventhal and Lewis [20], and Shalev-Shwartz and Tewari [36]) to analyze...

---

*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).
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 [34], Beck and Tetruashvili [4], Lu and Xiao [22], etc. We refer to [34] 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 deterministic optimization problems. One possible approach for solving problem (1.1), based on existing BCD methods and the sample average approximation (SAA) [38], can be described as follows. For a given set of i.i.d. samples (dataset) \( \xi_k, k = 1, \ldots, N \), of \( \xi \), 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 \( \xi_k, k = 1, \ldots, N \), are fixed a priori, by recursively updating the (sub)gradient of \( \tilde{f} \) (see [29, 30]), the iteration cost of the BCD method can be considerably smaller than that of the gradient descent methods. However, the above SAA approach is also known for the following drawbacks: a) the high memory requirement to store \( \xi_k, k = 1, \ldots, N \); 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 \( \xi_k \) is collected.

A different approach to solve problem (1.1) is called stochastic approximation (SA), which was initially proposed by Robbins and Monro [35] in 1950s for solving strongly convex SP problems (see also [32, 33]). The SA method has also attracted much interest recently (see, e.g., [25, 15, 31, 8, 9, 21, 10, 24, 11, 12, 14, 28, 43]). In particular, Nemirovski et. al. [25] presented a properly modified SA approach, namely, the mirror descent SA for solving general nonsmooth convex SP problems. Lan [15] introduced a unified optimal SA method for smooth, nonsmooth and stochastic optimization (see also [8, 9] for a more general framework). Ghadimi and Lan [10] presented novel SA methods for nonconvex optimization (see also [11]). Related methods, based on dual averaging, have been studied in [12, 14, 28, 43]. Note that all these SA algorithms only need to access one single \( \xi_k \) 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., [31]). In addition, it is unclear whether the SA methods can benefit from the recursive updating as in the BCD methods, since the samples \( \xi_k \) used in different iterations are supposed to be independent.

Our main goal in this paper is to present a new class of stochastic methods, referred to as the stochastic block mirror descent (SBMD) methods, by incorporating the aforementioned block-coordinate decomposition into the classic (stochastic) mirror descent method ([26, 3, 25]). Our study has been mainly motivated by solving an important class of SP problems with \( F(x, \xi) = \psi(Bx - q, \xi) \), where \( B \) is a certain linear operator and \( \psi \) is a relatively simple function. These problems arise from many machine learning applications, where \( \psi \) is a loss function and \( B \in \mathbb{R}^{m \times n} \) denotes a certain basis (or dictionary) obtained by, e.g., metric learning (e.g., [44]). Each iteration of existing SA methods would require \( \mathcal{O}(mn) \) arithmetic operations to compute \( Bx \) and becomes prohibitive if \( mn \) exceeds \( 10^{12} \). On the other hand, by using block-coordinate decomposition with \( n_k = 1 \), the iteration cost of the SBMD algorithms can be significantly reduced to \( \mathcal{O}(m) \), which can be further reduced if \( B \) and \( \xi_k \) 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 the sense that \( X \) is decomposable and the computation of the projection over \( X \) is more expensive than that of gradient. 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 \( 1 \leq i \leq b \), while the mirror descent method needs to perform the projections over \( X_i \) for all \( 1 \leq i \leq b \). It should be noted, however, that our algorithm does not apply to the situation when a decomposition of \( X \) is not available.

Our contribution in this paper mainly lies in the following aspects. Firstly, we introduce the block decomposition 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 \( \epsilon \)-solution of (1.1), i.e., a point \( \bar{x} \in X \) such that (s.t.) \( \mathbb{E}[f(\bar{x}) - f^*] \leq \epsilon \), can be bounded by \( O(b/\epsilon^2) \). 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 \( \epsilon \)-solution of (1.1) can be bounded by \( O(b/\epsilon) \). 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 Lipschitz 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 convex programming (CP) problems are based on the subgradient methods without involving averaging, e.g., those by Polak and a constrained version by Shor (see Nesterov [30]). 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 $O(b^3/\epsilon)$ 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 and some brief concluding remarks are given respectively in Section 5 and Section 6. Finally numerical results and some brief concluding remarks are given respectively in Section 5 and Section 6.

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 conjugate) such that $\sum_{i=1}^b n_i = n$. Let $I_n$ be the identity matrix in $\mathbb{R}^n$ and $U_i \in \mathbb{R}^{n_i \times n_i}$, $i = 1, 2, \ldots, b$, 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 $\| x \|_2^2 = \| x^{(1)} \|_2^2 + \ldots + \| x^{(b)} \|_2^2$.

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 $i$-th block of the stochastic subgradient of $f(\cdot)$ by $G_i(x, \xi) = U_i^T G(x, \xi)$, $i = 1, 2,\ldots, b$.

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:

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

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

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

and

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

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 $\omega_i : X_i \to \mathbb{R}$ is a distance generating function [25] with modulus $\alpha_i$ with respect to $\| \cdot \|_i$, if $\omega_i$ is continuously differentiable and strongly convex with parameter $\alpha_i$ with respect to $\| \cdot \|_i$. Without loss of generality, we assume
throughout the paper that $\alpha_i = 1$ for any $i = 1, \ldots, b$, because we can always rescale $\omega(x)$ to $\bar{\omega}(x) = \omega(x)/\alpha_i$ in case $\alpha_i \neq 1$. Therefore, we have
\[
\langle x - z, \nabla \omega_i(x) - \nabla \omega_i(z) \rangle \geq \|x - z\|^2_2 \quad \forall x, z \in X_i.
\]
The prox-function associated with $\omega_i$ is given by
\[
V_i(z, x) = \omega_i(x) - [\omega_i(z) + \langle \nabla \omega_i(z), x - z \rangle] \quad \forall x, z \in X_i.
\] (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, 39] and references therein).

For the sake of notational convenience, sometimes we simply denote $D_{\omega_i, X_i}$ by $D_i$.

Let $x^{(i)}_1 = \arg\min_{x \in X_i} \omega_i(x)$, $i = 1, \ldots, b$. We can easily see that for any $x \in X$,
\[
V_i(x^{(i)}_1, x^{(i)}) = \omega_i(x^{(i)}_1) - \omega_i(x^{(i)}) - \langle \nabla \omega_i(x^{(i)}_1), x^{(i)} - x^{(i)}_1 \rangle \leq \omega_i(x^{(i)}) - \omega_i(x^{(i)}_1) \leq D_i,
\] (2.7)
which, in view of the strong convexity of $\omega_i$, also implies that $\|x^{(i)}_1 - x^{(i)}\|^2_2 / 2 \leq D_i$.
Therefore, for any $x, y \in X$, we have
\[
\|x^{(i)} - y^{(i)}\|_i \leq \|x^{(i)} - x^{(i)}_1\|_i + \|x^{(i)}_1 - y^{(i)}\|_i \leq 2\sqrt{2D_i},
\] (2.8)
\[
\|x - y\| = \sqrt{\sum_{i = 1}^b \|x^{(i)} - y^{(i)}\|_i^2} \leq 2\sqrt{\sum_{i = 1}^b D_i}.
\] (2.9)

With the above definition of the prox-mapping, we can formally describe the stochastic block mirror descent (SBMD) method as in Algorithm 1.

We now add a few remarks about the SBMD algorithm stated above. Firstly, note that in this algorithm, the random variables $\xi_k$ and $i_k$, $k = 1, \ldots, b$, are assumed to be independent of each other. Secondly, each iteration of the SBMD method recursively updates the search point $x_k$ based on $G_{i_k}(x_k, \xi_k)$, the $i_k$-th block of the stochastic subgradient. 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 $x_k \sum_{j = i_k}^{i_k} \theta_j$, where $\sum_{j = i_k} \theta_j$ is often given by explicit formula and hence easy to compute. It can be checked that by using this averaging scheme, we have
\[
x_N = \left( \sum_{k = 1}^N \theta_k \right)^{-1} \sum_{k = 1}^N (\theta_k x_k).
\] (2.13)
Algorithm 1 The Stochastic Block Mirror Descent (SBMD) Algorithm

Let $x_1 \in X$, positive step sizes $\{\gamma_k\}_{k \geq 1}$, nonnegative 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 $s^{(i)}_k$, $i = 1, \ldots, b$, by
\[
s^{(i)}_{k+1} = \begin{cases} 
  s^{(i)}_k + x^{(i)}_k \sum_{j=u_{i_k}}^{b} \theta_j & i = i_k, \\
  s^{(i)}_k & i \neq i_k,
\end{cases} \tag{2.11}
\]

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

3. Update $x^{(i)}_k$, $i = 1, \ldots, b$, by
\[
x^{(i)}_{k+1} = \begin{cases} 
  P_i(x^{(i)}_k, G_{ik}(x_k, \xi_k), \gamma_k) & i = i_k, \\
  x^{(i)}_k & i \neq i_k.
\end{cases} \tag{2.12}
\]

end for

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

Thirdly, observe that in addition to (2.11) and (2.12), each iteration of the SBMD method involves the computation of $G_{ik}$. Whenever possible, we should update $G_{ik}$ 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 $B \in \mathbb{R}^{m \times n}$. 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 $\psi(y) = \max \left\{ \langle y, \xi \rangle, 0 \right\}$ and $\chi(x) = \|x\|^2_2/2$. In order to compute the full vector $G(x_k, \xi_k)$, we need $O(mn)$ arithmetic operations to compute the vector $Bx_k-q$, which majorizes other arithmetic operations if $\psi$ and $\chi$ 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 $O(mn)$ to $O(m)$. This bound can be further reduced if both $\xi_k$ and $B$ are sparse (i.e., the vector $\xi_k$ 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 $B_{i,j} \in \mathbb{R}^{m_i \times n_i}$, $1 \leq i \leq r$ and $1 \leq j \leq b$, and each block row $B_i = (B_{i,1}, \ldots, B_{i,b})$, $i = 1, \ldots, r$, is block-sparse (see [30] for some related discussion).

Finally, 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

$$\theta_k = \gamma_k, \quad k = 1, \ldots, N.$$  

(2.14)

Then we have, for any $N \geq 1$,

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

(2.15)

where $x_*$ is an arbitrary solution of (1.1) and the expectation is taken with respect to (w.r.t.) $\{i_k\}$ and $\{\xi_k\}$.

**Proof.** For simplicity, let us denote $V_i(z, x) \equiv V_i(z^{(i)}, x^{(i)}), g_i(x) = g^{(i)(x)}$ (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 [25]) and the definition of $x_i^{(i)}$ in (2.12), we have

$$V_i(x_{k+1}, x) \leq V_i(x_k, x) + \gamma_k \langle G_i(x_k, \xi_k), U^T_i(x - x_k) \rangle + \frac{1}{2} \gamma_k^2 \|G_i(x_k, \xi_k)\|_{x_k, *}^2.$$  

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

$$V(x_{k+1}, x) = \sum_{i \neq i_k} p_i^{-1} V_i(x_k, x) + p_{i_k}^{-1} V_{i_k}(x_{k+1}, x)$$

$$\leq \sum_{i \neq i_k} p_i^{-1} V_i(x_k, x) + p_{i_k}^{-1} \left[ V_i(x_k, x) + \gamma_k \langle G_i(x_k, \xi_k), U^T_i(x - x_k) \rangle + \frac{1}{2} \gamma_k^2 \|G_i(x_k, \xi_k)\|_{x_k, *}^2 \right]$$

$$= V(x_k, x) + \gamma_k p_{i_k}^{-1} \langle U_i G_i(x_k, \xi_k), x - x_k \rangle + \frac{1}{2} \gamma_k^2 p_{i_k}^{-1} \|G_i(x_k, \xi_k)\|_{x_k, *}^2 + \gamma_k \delta_k + \frac{1}{2} \gamma_k^2 \delta_k,$$  

(2.16)

where

$$\delta_k := \langle p_{i_k}^{-1} U_i G_i(x_k, \xi_k) - g(x_k), x - x_k \rangle \quad \text{and} \quad \delta_k := p_{i_k}^{-1} \|G_i(x_k, \xi_k)\|_{x_k, *}^2.$$  

(2.17)

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

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

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 $N \geq 1$ and $x \in X$,

$$f(\bar{x}_N) - f(x) \leq \left( \sum_{k=1}^{N} \gamma_k \right)^{-1} \sum_{k=1}^{N} \gamma_k [f(x_k) - f(x)]$$

$$\leq \left( \sum_{k=1}^{N} \gamma_k \right)^{-1} \left[ V(x_1, x) + \sum_{k=1}^{N} (\gamma_k \delta_k + \frac{1}{2} \gamma_k^2 \delta_k) \right].$$  

(2.18)

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

$$\mathbb{E}_{\zeta_k} \left[ p_{i_k}^{-1} \langle U_i G_i(x_k, x - x_k) \rangle_{[k-1]} \right] = \sum_{i=1}^{b} \mathbb{E}_{\zeta_k} \left[ \langle U_i G_i(x_k, \xi_k), x - x_k \rangle_{[k-1]} \right]$$

$$= \sum_{i=1}^{b} \langle U_i g_i(x_k), x - x_k \rangle = \langle g(x_k), x - x_k \rangle,$$
and hence that using the independence between \(i_k\) and \(\xi_k\), we have

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

Also, by (2.1) and (2.10),

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

Our result in (2.15) then immediately follows by taking expectation on both sides of (2.18), replacing \(x\) by \(x^*\), 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) and \(x^*\) is an arbitrary solution of (1.1).

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

\[
p_i = \frac{\sqrt{D_i}}{\sum_{i=1}^{b} \sqrt{D_i}}, \quad i = 1, \ldots, b, \quad \text{and} \quad \gamma_k = \gamma = \frac{\sqrt{2} \sum_{i=1}^{b} \sqrt{D_i}}{\sqrt{N \sum_{i=1}^{b} M_i^2}}, \quad k = 1, \ldots, N, \tag{2.21}
\]

then

\[
\mathbb{E}[f(\bar{x}_N) - f(x^*)] \leq \sqrt{\frac{2}{N} \sum_{i=1}^{b} \sqrt{D_i} \sum_{i=1}^{b} M_i^2}. \tag{2.22}
\]

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

\[
p_i = \frac{1}{b}, \quad i = 1, \ldots, b, \quad \text{and} \quad \gamma_k = \gamma = \frac{\sqrt{2b\hat{D}}}{\sqrt{N \sum_{i=1}^{b} M_i^2}}, \quad k = 1, \ldots, N, \tag{2.23}
\]

for some \(\hat{D} > 0\), then

\[
\mathbb{E}[f(\bar{x}_N) - f(x^*)] \leq \sqrt{\sum_{i=1}^{b} M_i^2 \left( \frac{\sum_{i=1}^{b} V_i(x_1^{(i)}, x_2^{(i)})}{\hat{D}} + \hat{D} \right) \frac{\sqrt{b}}{\sqrt{2N}}}. \tag{2.24}
\]

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

\[
\sum_{i=1}^{b} p_i^{-1} V_i(x_1, x_2) \leq \sum_{i=1}^{b} p_i^{-1} D_i = \left( \sum_{i=1}^{b} \sqrt{D_i} \right)^2.
\]

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

\[
\mathbb{E}[f(\bar{x}_N) - f(x^*)] \leq (N\gamma)^{-1} \left[ \left( \sum_{i=1}^{b} \sqrt{D_i} \right)^2 + \frac{N}{2} \sum_{i=1}^{b} M_i^2 \right] = \sqrt{\frac{2}{N} \sum_{i=1}^{b} \sqrt{D_i} \sum_{i=1}^{b} M_i^2}.
\]
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 \( D_i, i = 1, \ldots, b, \) are not available. It can be easily seen that the 
optimal choice of \( \tilde{D} \) in (2.24) would be \( \sqrt{\sum_{i=1}^{b} V_i(x_1, x_*)} \). In this case, (2.24) reduces 
to
\[
\mathbb{E}[f(\bar{x}_N) - f(x_*)] \leq \sqrt{2 \sum_{i=1}^{b} M_i^2} \sqrt{\sum_{i=1}^{b} V_i(x_1, x_*)} \frac{\sqrt{b}}{\sqrt{N}} \leq \sqrt{2 \sum_{i=1}^{b} M_i^2} \sqrt{\sum_{i=1}^{b} D_i} \frac{\sqrt{b}}{\sqrt{N}},
\]
(2.25)
where the second inequality follows from (2.7). It is interesting to note the differ-
ence 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 \( \epsilon \)-solution of (1.1) can be bounded by
\[
2 \left( \sum_{i=1}^{b} \sqrt{D_i} \right)^2 \left( \sum_{i=1}^{b} M_i^2 \right) \frac{1}{\epsilon^2}.
\]
(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
\[
2 \sum_{i=1}^{b} D_i \left( \sum_{i=1}^{b} M_i^2 \right) \frac{1}{\epsilon^2}.
\]
(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 \( O(b) \).

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 
[25], 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.
\[
f(y) \geq f(x) + \langle g(x), y - x \rangle + \frac{\mu}{2} \|y - x\|^2 \quad \forall x, y \in X.
\]
(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., [13, 8, 9]):
\[
V_i(z^{(i)}, x^{(i)}) \leq \frac{Q}{2} \|z^{(i)} - x^{(i)}\|_i^2 \quad \forall z^{(i)}, x^{(i)} \in X_i,
\]
(2.29)
for some \( Q > 0 \). In addition, for the sake of simplicity, we assume that the probability 
distribution of \( i_k \) is uniform, i.e.,
\[
p_1 = p_2 = \ldots = p_b = \frac{1}{b}.
\]
(2.30)
It should be noted, however, that our analysis can be easily adapted to the case when 
\( i_k \) is non-uniform.

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 [16].

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

\[
A_k := \begin{cases} 
1 & k = 1 \\
(1 - a_k) A_{k-1} & k \geq 2.
\end{cases}
\]

(2.31)

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

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

(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

\[
\gamma_k \leq \frac{bQ}{\mu}
\]

and

\[
\theta_k = \frac{\gamma_k}{\Gamma_k} \quad \text{with} \quad \Gamma_k = \begin{cases} 
1 & k = 1 \\
\Gamma_{k-1}(1 - \frac{\gamma_k}{\mu Q}) & k \geq 2,
\end{cases}
\]

(2.34)

then, for any \( N \geq 1 \), we have

\[
E[f(\bar{x}_N) - f(x^*_s)] \leq \left( \sum_{k=1}^{N} \theta_k \right)^{-1} \left[ \left( b - \frac{\gamma_k \mu}{Q} \right) \sum_{i=1}^{b} V_i(x^{(i)}_s, x^{(i)}_s) + \frac{1}{2} \sum_{k=1}^{N} \gamma_k \theta_k \sum_{i=1}^{b} M_i^2 \right],
\]

(2.35)

where \( x^*_s \) is the optimal solution of (1.1).

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

\[
V(z, x) = b \sum_{i=1}^{b} V_i(z^{(i)}, x^{(i)}) \leq \frac{bQ}{2} \sum_{i=1}^{b} \|z^{(i)} - x^{(i)}\|^2 = \frac{bQ}{2} \|z - x\|^2.
\]

(2.36)

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

\[
V(x_{k+1}, x) \leq V(x_k, x) + \gamma_k (g(x_k), x - x_k) + \gamma_k \delta_k + \frac{1}{2} \gamma_k^2 \delta_k
\]

\[
\leq V(x_k, x) + \gamma_k \left[ f(x) - f(x_k) - \frac{\mu}{2} \|x - x_k\|^2 \right] + \gamma_k \delta_k + \frac{1}{2} \gamma_k^2 \delta_k
\]

\[
\leq \left( 1 - \frac{\gamma_k \mu}{bQ} \right) V(x_k, x) + \gamma_k [f(x) - f(x_k)] + \gamma_k \delta_k + \frac{1}{2} \gamma_k^2 \delta_k.
\]
which, in view of Lemma 2.3 (with \( a_k = \gamma_k \mu / (bQ) \) and \( A_k = \Gamma_k \)), then implies that

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

(2.37)

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

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

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

\[
\gamma_k = \frac{2bQ}{\mu(k+1)}, \quad k = 1, \ldots, N,
\]

(2.39)

then, for any \( N \geq 1 \), we have

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

(2.40)

where \( x^* \) is the optimal solution of (1.1).

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

\[
\Gamma_k = \frac{2}{k(k+1)}, \quad \theta_k = \frac{\gamma_k}{\Gamma_k} = \frac{bQ}{\mu}, \quad b - \frac{\gamma_1 \mu}{Q} = 0,
\]

(2.41)

\[
\sum_{k=1}^{N} \theta_k = \frac{bQ(N+1)}{2\mu}, \quad \sum_{k=1}^{N} \gamma_k \theta_k \leq \frac{2b^2 Q^2 N}{\mu^2},
\]

(2.42)

and

\[
\sum_{k=1}^{N} \theta_k^2 = \frac{b^2 Q^2 N(N+1)(2N+1)}{6} \leq \frac{b^2 Q^2 N(N+1)^2}{3}.
\]

(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 \(\epsilon\)-solution for nonsmooth strongly convex problems can be bound by
\[
\frac{2bQ}{\mu\epsilon} \sum_{i=1}^{b} M_i^2,
\]
which is comparable to the optimal bound obtained in [13, 8, 24] (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 \(\xi\):
\[
E\left\{ \exp\left[ \frac{\|G_i(x, \xi)\|^2}{M_i^2} \right] \right\} \leq \exp(1), \ i = 1, 2, ..., b.
\]
(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 always satisfied for deterministic problems with bounded subgradients.

For the sake of simplicity, we only consider the case when the random variables \(\{i_k\}\) in the SBMD algorithm 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.

**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
\[
\text{Prob}\left\{ f(\bar{x}_N) - f(x_\ast) \geq b \left( \sum_{k=1}^{N} \gamma_k \right)^{-1} \left[ \sum_{i=1}^{b} V_i(x_{1}^{(i)}, x_{\ast}^{(i)}) + \frac{1}{2} \bar{M}^2 \sum_{k=1}^{N} \gamma_k^2 \right] \right\} \leq \exp\left( -\lambda^2/3 \right) + \exp(-\lambda),
\]
for any \(N \geq 1\) and \(\lambda > 0\), where \(\bar{M} = \max_{i=1, \ldots, b} M_i\) and \(x_\ast\) is an arbitrary solution of (1.1).

b) For solving strongly convex problems (i.e., (2.28), (2.29), (2.33), and (2.34) hold), we have
\[
\text{Prob}\left\{ f(\bar{x}_N) - f(x_\ast) \geq b \left( \sum_{k=1}^{N} \theta_k \right)^{-1} \left[ (1 - \gamma_k \mu Q) \sum_{i=1}^{b} V_i(x_{1}^{(i)}, x_{\ast}^{(i)}) + \frac{1}{2} \bar{M}^2 \sum_{k=1}^{N} \gamma_k \theta_k \right] \right\} \leq \exp\left( -\lambda^2/3 \right) + \exp(-\lambda),
\]
for any \(N \geq 1\) where \(x_\ast\) is the optimal solution of (1.1).
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 \( i = 1, 2, ..., b, \)

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

(2.47)

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

\[
\mathbb{E}[\exp (M^{-2}\delta_k^2)] \leq \sum_{i=1}^{b} \frac{1}{b} \mathbb{E} \left\{ \exp \left( M^{-2}\|x - x_k\|^2 \|bU_i^T G_i - g(x_k)\|_s^2 \right) \right\} \quad \text{(by (2.10), (2.17))}
\]

\[
\leq \sum_{i=1}^{b} \frac{1}{b} \mathbb{E} \left\{ \exp \left( 16M^{-2} \left( \sum_{i=1}^{b} \mathcal{D}_i \right) \left( b^2 \|G_i\|_{i,s}^2 + \sum_{i=1}^{b} M_i^2 \right) \right) \right\} \quad \text{(by (2.3) and (2.9))}
\]

\[
\leq \sum_{i=1}^{b} \frac{1}{b} \mathbb{E} \left\{ \exp \left[ \frac{b^2 \|G_i\|_{i,s}^2 + \sum_{i=1}^{b} M_i^2}{2b^2M^2} \right] \right\} \quad \text{(by definition of} \ M)\]

\[
\leq \sum_{i=1}^{b} \frac{1}{b} \mathbb{E} \left\{ \exp \left[ \frac{\|G_i\|_{i,s}^2}{2M_i^2} + \frac{1}{2} \right] \right\} \leq \exp(1). \quad \text{(by (2.47))}
\]

Therefore, by the large-deviation theorem on the Martingale-difference (see, e.g., Lemma 2 of [19]), we have

\[
\text{Prob} \left\{ \sum_{k=1}^{N} g_k \delta_k \geq \lambda M \sqrt{\sum_{k=1}^{N} g_k^2} \right\} \leq \exp(-\lambda^2/3). \quad \text{(2.48)}
\]

Also observe that under Assumption (2.44),

\[
\mathbb{E} \left[ \exp \left( \bar{\delta}_k / (bM^2) \right) \right] \leq \sum_{i=1}^{b} \frac{1}{b} \mathbb{E} \left[ \exp \left( \|G_i(x_k, \xi_k)\|_{i,s}^2 / M_i^2 \right) \right] \quad \text{(by (2.10), (2.17), (2.30))}
\]

\[
\leq \sum_{i=1}^{b} \frac{1}{b} \mathbb{E} \left[ \exp \left( \|G_i(x_k, \xi_k)\|_{i,s}^2 / M_i^2 \right) \right] \quad \text{(by definition of} \ M)\]

\[
\leq \sum_{i=1}^{b} \frac{1}{b} \exp(1) = \exp(1). \quad \text{(by (2.1))}
\]

Setting \( \pi_k = \gamma_k^2 / \sum_{k=1}^{N} \gamma_k^2, \) we have \( \exp \left\{ \sum_{k=1}^{N} \pi_k \delta_k / (bM^2) \right\} \leq \sum_{k=1}^{N} \pi_k \exp \{\delta_k / (bM^2)\}. \)

Using these previous two inequalities, we have

\[
\mathbb{E} \left[ \exp \left( \sum_{k=1}^{N} \gamma_k^2 \bar{\delta}_k / (bM^2 \sum_{k=1}^{N} \gamma_k^2) \right) \right] \leq \exp(1).
\]
It then follows from Markov’s inequality that $\forall \lambda \geq 0,$

$$\text{Prob} \left\{ \sum_{k=1}^{N} \gamma_k^2 \delta_k > (1 + \lambda)(bM^2) \sum_{k=1}^{N} \gamma_k^2 \right\}$$

$$= \text{Prob} \left\{ \exp \left\{ \sum_{k=1}^{N} \gamma_k^2 \delta_k / (bM^2 \sum_{k=1}^{N} \gamma_k^2) \right\} > \exp \{1 + \lambda\} \right\}$$

$$\leq \frac{\text{Exp} \left\{ \exp \left\{ \sum_{k=1}^{N} \gamma_k^2 \delta_k / (bM^2 \sum_{k=1}^{N} \gamma_k^2) \right\} \right\}}{\exp \{1 + \lambda\}} \leq \frac{\exp \{1\}}{\exp \{1 + \lambda\}} = \exp \{-\lambda\}. \quad (2.49)$$

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), (2.23) and $\tilde{D} = \sqrt{\sum_{i=1}^{b} V_i(x_1^{(i)}, x_2^{(i)})}$ for general nonsmooth problems, then we have

$$\text{Prob} \left\{ f(\bar{x}_N) - f(x_*) \geq \frac{\sqrt{b \sum_{i=1}^{b} M_i^2 \sum_{i=1}^{b} D_i}}{\sqrt{2N}} \left[ 1 + (1 + \lambda) \frac{M_*^2}{m^2} + 8 \lambda \frac{M_*}{m} \right] \right\}$$

$$\leq \exp (-\lambda^2/3) + \exp(-\lambda) \quad (2.50)$$

for any $\lambda > 0$, where $\bar{m} = \min_{i=1,...,b} M_i$ and $x_*$ is an arbitrary solution of (1.1).

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

$$\text{Prob} \left\{ f(\bar{x}_N) - f(x_*) \geq \frac{2b(M_2 - M_1) \sum_{i=1}^{b} M_i^2 \tilde{y}_i^2}{(N+1)\mu} + \frac{8 \sqrt{\lambda M_* \sqrt{b \sum_{i=1}^{b} M_i^2 \sum_{i=1}^{b} D_i}}}{m \sqrt{bN}} \right\}$$

$$\leq \exp (-\lambda^2/3) + \exp(-\lambda) \quad (2.51)$$

for any $\lambda > 0$, where $x_*$ is the optimal solution of (1.1).

**Proof.** Note that by (2.7), we have $\sum_{i=1}^{b} V_i(x_1^{(i)}, x_2^{(i)}) \leq \sum_{i=1}^{b} D_i$. Also by (2.23), we have

$$\sum_{k=1}^{N} \gamma_k = \left( \frac{2NbD^2}{\sum_{i=1}^{b} M_i^2} \right)^{\frac{1}{2}} \text{ and } \sum_{k=1}^{N} \gamma_k^2 = \frac{2bD^2}{\sum_{i=1}^{b} M_i^2}.$$

Using these identities and (2.45), we conclude that

$$\text{Prob} \left\{ f(\bar{x}_N) - f(x_*) \geq b \left( \frac{\sum_{i=1}^{b} M_i^2}{2bD^2} \right)^{\frac{1}{2}} \left[ \sum_{i=1}^{b} V_i(x_1^{(i)}, x_2^{(i)}) + bD^2 M_2^{-1} \right]^{-1}$$

$$+ \lambda M \left( 2bD^2 \left( \sum_{i=1}^{b} M_i^2 \right)^{-1} + 8 b^2 D \left( \sum_{i=1}^{b} D_i \left( \sum_{i=1}^{b} M_i^2 \right)^{-\frac{1}{2}} \right) \right) \right\} \leq \exp (-\lambda^2/3) + \exp(-\lambda).$$
Using the fact that $\sum_{i=1}^{b} M_i^2 \geq \bar{m}^2$ and $\hat{D} = \sqrt{\sum_{i=1}^{b} V_i(x_1^{(i)}, x_2^{(i)})}$ to simplify 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 (2.50), the number of iterations required by the SBMD method to find an $(\epsilon, \Lambda)$-solution of (1.1), i.e., a point $\bar{x} \in X$ s.t. $\Pr\{f(\bar{x}) - f^* \geq \epsilon\} \leq \Lambda$ can be bounded by

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

after disregarding a few constant factors. To the best of our knowledge, no such large-deviation results have been obtained before for the BCD methods for solving general nonsmooth CP problems, although, in case $b = 1$, similar results have been established for solving smooth problems or some composite problems [29, 34].

Secondly, it follows from (2.51) that the number of iterations performed by the SBMD method to find an $(\epsilon, \Lambda)$-solution for nonsmooth strongly convex problems, after disregarding a few constant factors, can be bounded by $O\left(\frac{b \log^2(1/\Lambda)/\epsilon^2}{\epsilon^2}\right)$, which, in case $b = 1$, 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 $O\left(\frac{b \log(1/\Lambda)/\epsilon}{\epsilon}\right)$, for example, by incorporating a domain shrinking procedure [9].

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

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

It then follows that

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

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

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

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 $y \in \mathbb{R}^{n_i}$, we define the composite prox-mapping as

$$\mathcal{P}_i(x, y, \gamma) := \arg\min_{z \in X_i} \langle y, z - x \rangle + \frac{1}{\gamma} V_i(z, x) + \chi_i(x). \quad (3.4)$$
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 $x_1 \in X$, positive stepsizes $\{\gamma_k\}_{k \geq 1}$, nonnegative 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 $u_{ik} = k + 1$.
3. Update $x_k^{(i)}$, $i = 1, \ldots, b$, by

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

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 $\sigma_i \geq 0$,

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

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

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

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 $x_2, \ldots, x_{N+1}$, 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 $x_1, \ldots, x_N$.

Thirdly, it can be easily seen from (2.7), (3.1), and (3.6) that if $X$ is bounded,
\[ E[\|G_i(x, \xi)\|_{i,*}^2] \leq 2\|g_i(x)\|_{i,*}^2 + 2E[\|G_i(x, \xi) - g_i(x)\|_{i,*}^2] \leq 2\|g_i(x)\|_{i,*}^2 + 2\sigma_i^2 \]
\[ \leq 2 \left[ 2\|g_i(x) - g_i(x_1)\|_{i,*}^2 + 2\|g_i(x_1)\|_{i,*}^2 \right] + 2\sigma_i^2 \]
\[ \leq 2 \left[ 2\|g(x) - g(x_1)\|_{i,*}^2 + 2\|g_i(x_1)\|_{i,*}^2 \right] + 2\sigma_i^2 \]
\[ \leq 4 \left( \sum_{i=1}^{b} L_i \right)^2 \|x - x_1\|^2 + 4\|g_i(x_1)\|_{i,*}^2 + 2\sigma_i^2 \]
\[ \leq 8\delta^2 L^2 \sum_{i=1}^{b} D_i + 4\|g_i(x_1)\|_{i,*}^2 + 2\sigma_i^2, \quad i = 1, \ldots, b, \quad (3.9) \]

where \( L := \max_{i=1, \ldots, b} L_i \) and the fourth inequality follows from the fact that \( g \) is Lipschitz continuous with constant \( \sum_{i=1}^{b} L_i \) (see Lemma 2 of [29]). 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 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. Hence, 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 \( k \geq 1 \),

\[ \gamma_k \leq \frac{1}{2L}, \quad (3.10) \]
\[ \theta_{k+1} = b\gamma_k - (b - 1)\gamma_{k+1}, \quad (3.11) \]

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

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

where \( x_* \) is an arbitrary solution of problem (1.4) and \( \sigma \) is defined in (3.7).

**Proof.** For simplicity, let us denote \( V_i(z, x) \equiv V_i(z^{(i)}, x^{(i)}), g_{ik} \equiv g^{(ik)}(x_k) \), 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 \( \delta_{ik} = G_{ik}(x_k, \xi_k) - g_{ik}(x_k) \) and \( \rho_{ik} = U_{ik}^T (x_{k+1} - x_k) \). By the definition of \( \phi(\cdot) \) in (1.4) and (3.2), we have

\[ \phi(x_{k+1}) \leq f(x_k) + \langle g_{ik}(x_k), \rho_{ik} \rangle + \frac{L_{ik}}{2} \|\rho_{ik}\|_{ik}^2 + \chi(x_{k+1}) \]
\[ = f(x_k) + \langle G_{ik}(x_k, \xi_k), \rho_{ik} \rangle + \frac{L_{ik}}{2} \|\rho_{ik}\|_{ik}^2 + \chi(x_{k+1}) - \langle \delta_{ik}, \rho_{ik} \rangle. \quad (13.13) \]

Moreover, it follows from the optimality condition of (3.4) (see, e.g., Lemma 1 of [15]) and (3.5) that, for any \( x \in X \),

\[ \langle G_{ik}(x_k, \xi_k), \rho_{ik} \rangle + \chi_{ik}(x_{k+1}) \leq \langle G_{ik}(x_k, \xi_k), x_{k+1}^{(ik)} - x_k^{(ik)} \rangle + \chi_{ik}(x^{(ik)}) \]
\[ + \frac{1}{\gamma_k} [V_{ik}(x_k, x) - V_{ik}(x_{k+1}, x) - V_{ik}(x_{k+1}, x_k)]. \]

17
Combining the above two inequalities and using (3.3), we obtain

\[
\phi(x_{k+1}) \leq f(x_k) + \left\langle G_{ik}(x_k, \xi_k), x^{(ik)} - x_k^{(ik)} \right\rangle + \chi_{ik}(x^{(ik)}) \\
+ \frac{1}{\gamma_k} \left[ V_{ik}(x_k, x) - V_{ik}(x_{k+1}, x) - V_{ik}(x_k, x_{k+1}) \right] \\
+ \frac{L_{ik}}{2} \| \rho_{ik} \|^2 + \sum_{i \neq ik} \chi_i(x^{(i)}_k) - \left\langle \delta_{ik}, \rho_{ik} \right\rangle.
\]

(3.14)

By the strong convexity of \( \omega(t) \) and (3.10), using the simple inequality that \( bu - \frac{a^2}{2b} \leq \frac{b^2}{2a}, \forall a > 0 \), we have

\[
-\frac{1}{\gamma_k} V_{ik}(x_{k+1}, x_k) + \frac{L_{ik}}{2} \| \rho_{ik} \|^2 - \left\langle \delta_{ik}, \rho_{ik} \right\rangle \leq -\left( \frac{1}{2\gamma_k} - \frac{L_{ik}}{2} \right) \| \rho_{ik} \|^2 - \left\langle \delta_{ik}, \rho_{ik} \right\rangle \\
\leq \frac{\gamma_k \| \delta_{ik} \|^2}{2(1 - \gamma_k L_{ik})} \leq \frac{\gamma_k \| \delta_{ik} \|^2}{2(1 - \gamma_k L)} \leq \gamma_k \| \delta_{ik} \|^2.
\]

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

\[
\phi(x_{k+1}) \leq f(x_k) + \left\langle G_{ik}(x_k, \xi_k), x^{(ik)} - x_k^{(ik)} \right\rangle + \frac{1}{b \gamma_k} \left[ V(x_k, x) - V(x_{k+1}, x) \right] \\
+ \gamma_k \| \delta_{ik} \|^2 + \sum_{i \neq ik} \chi_i(x^{(i)}_k) + \chi_{ik}(x^{(ik)}).
\]

(3.15)

Now noting that

\[
\mathbb{E}_{\xi_k} \left[ \left\langle G_{ik}(x_k, \xi_k), x^{(ik)} - x_k^{(ik)} \right\rangle \mid \xi_{k-1} \right] = \frac{1}{b} \sum_{i=1}^b \mathbb{E}_{\xi_k} \left[ \left\langle G_i(x_k, \xi_k), x^{(i)} - x_k^{(i)} \right\rangle \mid \xi_{k-1} \right] \\
= \frac{1}{b} \left\langle g(x_k), x - x_k \right\rangle \leq \frac{1}{b} \| f(x) - f(x_k) \|,
\]

(3.16)

\[
\mathbb{E}_{\xi_k} \left[ \| \delta_{ik} \|^2 \mid \xi_{k-1} \right] = \frac{1}{b} \sum_{i=1}^b \mathbb{E}_{\xi_k} \left[ \| G_i(x_k, \xi_k) - g_i(x_k) \|_{\xi_i}^2 \mid \xi_{k-1} \right] \\
\leq \frac{1}{b} \sum_{i=1}^b \sigma_i^2 = \frac{\sigma^2}{b},
\]

(3.17)

\[
\mathbb{E}_{\xi_k} \left[ \sum_{i \neq ik} \chi_i(x^{(i)}_k) \mid \xi_{k-1} \right] = \frac{1}{b} \sum_{i=1}^b \sum_{j=1}^i \chi_i(x^{(i)}_k) = \frac{b-1}{b} \chi(x_k),
\]

(3.18)

\[
\mathbb{E}_{\xi_k} \left[ \chi_{ik}(x^{(ik)}) \mid \xi_{k-1} \right] = \frac{1}{b} \sum_{i=1}^b \chi_i(x^{(i)}) = \frac{1}{b} \chi(x),
\]

(3.19)
we conclude from (3.15) that
\[
\mathbb{E}_k\left[\phi(x_{k+1}) + \frac{V(x_{k+1}, x)}{b\gamma_k} \psi_{k-1}\right] \leq f(x_k) + \frac{1}{b} [f(x) - f(x_k)] + \frac{1}{b} \chi(x)
\]
\[
+ \frac{V(x_k, x)}{b\gamma_k} + \frac{\gamma_k}{b} \sigma^2 + b \frac{b-1}{b} \chi(x_k) = \frac{b-1}{b} \phi(x_k) + \frac{1}{b} \psi(x) + \frac{V(x_k, x)}{b\gamma_k} + \frac{\gamma_k}{b} \sigma^2,
\]
which implies that
\[
b\gamma_k \mathbb{E}[\phi(x_{k+1}) - \phi(x)] + \mathbb{E}[V(x_{k+1}, x)] \leq (b-1) \gamma_k \mathbb{E}[\phi(x_k) - \phi(x)] + \mathbb{E}[V(x_{k+1}, x)] + \gamma_k^2 \sigma^2.
\]
(3.20)
Now, summing up the above inequalities (with \(x = x_\ast\)) for \(k = 1, \ldots, N\), and noting that \(\theta_{k+1} = b\gamma_k - (b-1)\gamma_{k+1}\), we obtain
\[
\sum_{k=2}^{N+1} \theta_k \mathbb{E}[\phi(x_k) - \phi(x_\ast)] + b\gamma_k \mathbb{E}[\phi(x_{N+1}) - \phi(x_\ast)] + \mathbb{E}[V(x_{N+1}, x_\ast)]
\]
\[
\leq (b-1) \gamma_1 [\phi(x_1) - \phi(x_\ast)] + V(x_1, x_\ast) + \sigma^2 \sum_{k=1}^{N} \gamma_k^2,
\]
Using the above inequality and the facts that \(V(\cdot, \cdot) \geq 0\) and \(\phi(x_{N+1}) \geq \phi(x_\ast)\), we conclude
\[
\sum_{k=2}^{N+1} \theta_k \mathbb{E}[\phi(x_k) - \phi(x_\ast)] \leq (b-1) \gamma_1 [\phi(x_1) - \phi(x_\ast)] + V(x_1, x_\ast) + \sigma^2 \sum_{k=1}^{N} \gamma_k^2,
\]
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
\[
\gamma_k = \gamma = \min \left\{ \frac{1}{2L}, \frac{\hat{D}}{\sigma \sqrt{N}} \right\}
\]
(3.21)
for some \(\hat{D} > 0\), and \(\{\theta_k\}\) are set to (3.11). Then, under Assumptions (1.3) and (3.6), we have
\[
\mathbb{E}[\phi(x_N) - \phi(x_\ast)] \leq \frac{(b-1)[\phi(x_1) - \phi(x_\ast)]}{N} + \frac{2b\hat{D} \sum_{i=1}^{b} V_i(x_1^{(i)}, x_\ast^{(i)})}{N} + \frac{b\sigma \sqrt{N}}{\sqrt{N}} \left[ \sum_{i=1}^{b} V_i(x_1^{(i)}, x_\ast^{(i)}) \right] + \frac{\hat{D}}{N}.
\]
(3.22)
where \(x_\ast\) is an arbitrary 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
\[
\mathbb{E}[\phi(x_N) - \phi(x_\ast)] \leq \frac{(b-1)[\phi(x_1) - \phi(x_\ast)]}{N} + \frac{b \sum_{i=1}^{b} V_i(x_1^{(i)}, x_\ast^{(i)})}{N\gamma} + \gamma \sigma^2,
\]

\[19\]
Then, for any (3.9), we have

\[
\text{Second, if we directly apply Algorithm 1 to problem (1.4), then, in view of (2.25) and (3.22), an optimal selection of } \tilde{D} \text{ would be } \sqrt{\sum_{i=1}^{b} V_i(x_1^{(i)}, x_*^{(i)})}. \text{ In this case, (3.22) reduces to}
\]

\[
\mathbb{E} [\phi(\bar{x}_N) - \phi(x_*)] \leq \frac{(b-1)[\phi(x_1) - \phi(x_*)]}{N} + 2b\tilde{L}\sum_{i=1}^{b} V_i(x_1^{(i)}, x_*^{(i)}) + \frac{2\sigma^2}{\sqrt{N}} \sqrt{\sum_{i=1}^{b} D_i}.
\]

(3.23)

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

\[
\mathbb{E} [\phi(\bar{x}_N) - \phi(x_*)] \leq 2 \sqrt{\sum_{i=1}^{b} \left[ 4b^2 \tilde{L}^2 \left( \sum_{i=1}^{b} D_i \right) + 2\|g_i(x_1)\|_{i,*}^2 + \sigma_i^2 \right] \sqrt{\sum_{i=1}^{b} \frac{D_i}{N}}}.
\]

(3.24)

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

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 \( k \geq 1 \),

\[
\gamma_k \leq \min \left\{ \frac{1}{2\tilde{L}}, \frac{bQ}{\mu} \right\},
\]

(3.25)

\[
\theta_{k+1} = \frac{b\gamma_k}{\Gamma_k} - \frac{(b-1)\gamma_{k+1}}{\Gamma_{k+1}} \quad \text{with} \quad \Gamma_k = \begin{cases} 1 & k = 1 \\ \Gamma_{k-1}(1 - \frac{\gamma_k \mu}{bQ}) & k \geq 2. \end{cases}
\]

(3.26)

Then, for any \( N \geq 2 \), we have

\[
\mathbb{E} [\phi(\bar{x}_N) - \phi(x_*)] \leq \left( \sum_{k=2}^{N+1} \theta_k \right)^{-1} \left[ (b - \mu \gamma_1 Q) \sum_{i=1}^{b} V_i(x_1^{(i)}, x_*^{(i)}) + (b-1)\gamma_1 [\phi(x_1) - \phi(x_*)] + \sum_{k=1}^{N} \frac{\tilde{L}}{\Gamma_k} \sigma_i^2 \right],
\]

(3.27)
where \( x_* \) is the optimal solution of problem (1.4).

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

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

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

\[
\mathbb{E}_{\xi_k} \left[ \phi(x_{k+1}) + \frac{1}{b\gamma_k} V(x_{k+1}, x) \mid \xi_{[k-1]} \right] \leq f(x_k) + \frac{1}{b} \left[ f(x) - f(x_k) - \frac{\mu}{2} \|x - x_k\|^2 \right] + \frac{1}{b\gamma_k} V(x_k, x) + \frac{2\gamma_k}{b} \sigma^2 + \sum_{i=1}^{k} \left( b Q_{\gamma_k} \right) V(x_k, x) + \frac{\gamma_k}{b} \sigma^2,
\]

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

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

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

\[
\frac{1}{\Gamma_N} \mathbb{E}[V(x_{k+1}, x_*)] \leq \left( 1 - \frac{\mu \gamma_1}{bQ} \right) \mathbb{E}[V(x_1, x_*]) + (b - 1) \sum_{k=1}^{N} \frac{\gamma_k}{\Gamma_k} (\phi(x_k) - \phi(x_*)) - b \sum_{k=1}^{N} \frac{\gamma_k}{\Gamma_k} (\phi(x_{k+1}) - \phi(x_*)] + \sum_{k=1}^{N} \frac{\gamma_k^2}{\Gamma_k} \sigma^2
\]

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

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

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

\[
\gamma_k = 2bQ/(\mu (k + k_0)) \quad \forall k \geq 1,
\]

(3.28)
where

\[ k_0 := \left\lceil \frac{4bQ\bar{L}}{\mu} \right\rceil. \]

Then, for any \( N \geq 2 \), we have

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

(3.29)

where \( x_* \) is the optimal solution of problem (1.4).

Proof. We can check that

\[
\gamma_k = \frac{2bQ}{\mu(k + \left\lceil \frac{4bQ\bar{L}}{\mu} \right\rceil)} \leq \frac{1}{2\bar{L}}.
\]

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

\[
\Gamma_k = \frac{k_0(k_0 + 1)}{(k + k_0)(k + k_0 - 1)} - \frac{1}{bQ} \leq \frac{1 - \frac{\gamma_k \mu}{bQ}}{k_0 + 1}, \quad \forall k \geq 1,
\]

(3.30)

\[
\theta_k = \frac{b\gamma_k}{\Gamma_k} = \frac{(b - 1)\gamma_k + 1}{\Gamma_{k+1}} = \frac{2bkQ + 2bQ(k_0 - b)}{\mu k_0(k_0 + 1)} \geq \frac{2b\mu}{\mu k_0(k_0 + 1)},
\]

(3.31)

where the relation \( k_0 \geq b \) follows from the definition of \( k_0 \) and the fact that \( Q \geq \alpha = 1 \) and \( \bar{L} \geq \mu \). And hence that

\[
\sum_{k=2}^{N+1} \theta_k \geq \frac{bQ(N+1)}{\mu k_0(k_0 + 1)}, \quad \sum_{k=1}^{N} \frac{\gamma_k^2}{\Gamma_k} = \frac{4b^2Q^2}{\mu^2 k_0(k_0 + 1)} \sum_{k=1}^{N} \frac{k + k_0 - 1}{k + k_0} \leq \frac{4b^2Q^2}{\mu^2 k_0(k_0 + 1)},
\]

(3.32)

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

\[
\mathbb{E}[\phi(\bar{x}_N) - \phi(x_*)] \leq \left( \sum_{k=2}^{N+1} \theta_k \right)^{-1} \left[ (1 - \frac{\gamma_k}{\Gamma_k}) V(x_1, x_*) + (b - 1)\gamma_k [\phi(x_1) - \phi(x_*)] + \sum_{k=1}^{N} \frac{\gamma_k^2}{\Gamma_k} \sigma^2 \right].
\]

\[
\leq \frac{\mu k_0(k_0 + 1)}{bQ N(N+1)} \left[ \frac{k_0 - 1}{k_0 + 1} V(x_1, x_*) + \frac{2b(b - 1)Q}{\mu(k_0 + 1)} [\phi(x_1) - \phi(x_*)] + \frac{4b^2Q^2}{\mu^2 k_0(k_0 + 1)} \sigma^2 \right]
\]

\[
\leq \frac{\mu k_0^2}{bQ N(N+1)} V(x_1, x_*) + \frac{2(b - 1)k_0}{N(N+1)} [\phi(x_1) - \phi(x_*)] + \frac{4bQ^2}{\mu N(N+1)} \sigma^2,
\]

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 (see [27, 17, 18, 6, 11] for some related discussions). 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 [27, 17, 18, 6, 11] for some related discussions). More specifically, for a given \( x \in X \), \( y \in \mathbb{R}^n \) and a constant \( \gamma > 0 \), we define \( \mathcal{G}(x, y, \gamma) = (\mathcal{G}_1(x, y, \gamma), \ldots, \mathcal{G}_b(x, y, \gamma)) \) by

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

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. \( \gamma \). 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 \( p_i \in \partial \chi_i(U^T_i x^+) \) s.t.

\[
U^T_i g(x^+) + p_i \in -N_{X_i}(U^T_i x^+) + \mathcal{B}_i((L_i \gamma + Q) \| \mathcal{G}(x, g(x), \gamma) \|), \quad i = 1, \ldots, b, \tag{4.2}
\]

where \( \mathcal{B}_i(e) := \{ v \in \mathbb{R}^{n_i} : \| v \|_{i,*} \leq e \} \) and \( N_{X_i}(U^T_i x^+) \) denotes the normal cone of \( X_i \) at \( U^T_i x^+ \).

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

\[
(U^T_i g(x) + \frac{1}{\gamma} \left[ \nabla \omega_i(U^T_i x^+) - \nabla \omega_i(U^T_i x) \right] + p_i, u - U^T_i x^+) \geq 0, \quad \forall u \in X_i.
\]

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

\[
\| \zeta \|_{i,*} \leq \left( L_i + \frac{Q}{\gamma} \right) \| U^T_i (x^+ - x) \|_i = \left( L_i + \frac{Q}{\gamma} \right) \| U^T_i \mathcal{G}(x, g(x), \gamma) \|_i = (L_i \gamma + Q) \| U^T_i \mathcal{G}(x, g(x), \gamma) \|_i.
\]

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., [27]). 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 \( \tilde{x}_N \) 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 \( i_k \)-th block of the (stochastic) gradient \( G_{i_k} \) of \( f(\cdot) \) at \( x_k \) satisfying

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

and update \( x_k \) by (3.5).

**end for**

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

\[
\text{Prob}(R = k) = \frac{\gamma_k \min_{i \in [1, b]} p_i (1 - \frac{L_i}{2} \gamma_k)}{\sum_{k=1}^N \gamma_k \min_{i \in [1, b]} p_i (1 - \frac{L_i}{2} \gamma_k)}, \quad k = 1, \ldots, N.
\]

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 \( \tilde{\sigma}_k = 0 \). 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). Secondly, the probability of choosing \( x_R, \ R = 1, \ldots, N, \) as the output solution is given by (4.5). Such a randomization scheme was shown to be critical to establish the complexity for nonconvex stochastic optimization in [10]. Related randomization schemes (with different specification of the probability distribution) have also been discussed by a few researchers for solving stochastic convex optimization problems in the literature (see, e.g., [37]).

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 [11].

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

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

\[
\|\tilde{\mathcal{G}}_k - U_{i_k}^T \mathcal{G}_k\|_{i_k} \leq \|G_{i_k} - U_{i_k} g(x_k)\|_{i_k, \ast}.
\]
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
\[
\langle G_{i_k} + \frac{1}{\gamma_k} [\nabla \omega_{i_k}(U_{i_k}^T x_{k+1}) - \nabla \omega_{i_k}(U_{i_k}^T x_k)] + p, \frac{1}{\gamma_k} (u - U_{i_k}^T x_{k+1}) \rangle \geq 0, \quad \forall u \in X_{i_k}.
\] (4.8)

Letting \( u = U_{i_k}^T x_k \) in the above inequality and re-arranging terms, we obtain
\[
\langle G_{i_k}, \frac{1}{\gamma_k} U_{i_k}^T (x_k - x_{k+1}) \rangle \geq \frac{1}{\gamma_k} \langle \nabla \omega_{i_k}(U_{i_k}^T x_{k+1}) - \nabla \omega_{i_k}(U_{i_k}^T x_k), U_{i_k}^T (x_{k+1} - x_k) \rangle + \langle p, \frac{1}{\gamma_k} U_{i_k}^T (x_{k+1} - x_k) \rangle \\
\geq \frac{1}{\gamma_k} \langle \nabla \omega_{i_k}(U_{i_k}^T x_{k+1}) - \nabla \omega_{i_k}(U_{i_k}^T x_k), U_{i_k}^T (x_{k+1} - x_k) \rangle + \langle \gamma_k \xi_k (U_{i_k}^T x_{k+1} - \chi_{i_k}(U_{i_k}^T x_k)) \rangle \\
\geq \frac{1}{\gamma_k} \| U_{i_k}^T (x_{k+1} - x_k) \|^2 + \frac{1}{\gamma_k} \| \chi_{i_k}(U_{i_k}^T x_{k+1} - \chi_{i_k}(U_{i_k}^T x_k)) \|.\]
(4.9)

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

Now we show that (4.7) holds. Let us denote \( x_{k+1} = x_k - \gamma_k \hat{G}_k \). By the optimality condition of (3.4) and the definition of \( G_{i_k} \), we have, for some \( q \in \partial \chi_{i_k}(x_{k+1}^+), \)
\[
\langle U_{i_k}^T g(x_k) + \frac{1}{\gamma_k} [\nabla \omega_{i_k}(U_{i_k}^T x_{k+1}) - \nabla \omega_{i_k}(U_{i_k}^T x_k)] + q, \frac{1}{\gamma_k} (u - U_{i_k}^T x_{k+1}) \rangle \geq 0, \quad \forall u \in X_{i_k}.
\] (4.10)

Letting \( u = U_{i_k}^T x_{k+1}^+ \) in (4.8) and using an argument similar to (4.9), we have
\[
\langle G_{i_k}, \frac{1}{\gamma_k} U_{i_k}^T (x_{k+1}^+ - x_k) \rangle \geq \frac{1}{\gamma_k} \langle \nabla \omega_{i_k}(U_{i_k}^T x_{k+1}^+) - \nabla \omega_{i_k}(U_{i_k}^T x_k), U_{i_k}^T (x_{k+1}^+ - x_k) \rangle + \frac{1}{\gamma_k} \| \chi_{i_k}(U_{i_k}^T x_{k+1}^+) - \chi_{i_k}(U_{i_k}^T x_k) \|.\]

Similarly, letting \( u = U_{i_k}^T x_{k+1} \) in (4.10), we have
\[
\langle U_{i_k}^T g(x_k), \frac{1}{\gamma_k} U_{i_k}^T (x_{k+1} - x_k) \rangle \geq \frac{1}{\gamma_k} \langle \nabla \omega_{i_k}(U_{i_k}^T x_{k+1}) - \nabla \omega_{i_k}(U_{i_k}^T x_k), U_{i_k}^T (x_{k+1} - x_k) \rangle + \frac{1}{\gamma_k} \| \chi_{i_k}(U_{i_k}^T x_{k+1}) - \chi_{i_k}(U_{i_k}^T x_k) \|.\]

Summing up the above two inequalities, we obtain
\[
\langle G_{i_k} - U_{i_k}^T g(x_k), U_{i_k}^T (x_{k+1}^+ - x_k) \rangle \geq \frac{1}{\gamma_k} \langle \nabla \omega_{i_k}(U_{i_k}^T x_{k+1}^+) - \nabla \omega_{i_k}(U_{i_k}^T x_k), U_{i_k}^T (x_{k+1}^+ - x_k) \rangle \geq \frac{1}{\gamma_k} \| U_{i_k}^T (x_{k+1} - x_k) \|^2,
\]
which, in view of the Cauchy-Schwarz inequality, then implies that
\[
\frac{1}{\gamma_k} \| U_{i_k}^T (x_{k+1} - x_k) \|^2 \leq \| G_{i_k} - U_{i_k}^T g(x_k) \|_{i_k}^2.
\]

Using the above relation and (4.1), we have
\[
\| \hat{G}_k - U_{i_k}^T G_k \|_{i_k} = \| \frac{1}{\gamma_k} U_{i_k}^T (x_k - x_{k+1}) - \frac{1}{\gamma_k} U_{i_k}^T (x_k - x_{k+1}) \|_{i_k} \\
= \frac{1}{\gamma_k} \| U_{i_k}^T (x_{k+1}^+ - x_k) \|_{i_k} \leq \| G_{i_k} - U_{i_k}^T g(x_k) \|_{i_k}.
\]

\[ \blacksquare \]
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

\[
\mathbb{E}[[G(x_R, g(x_R), \gamma_R)]^2] \leq \frac{\phi(x_1) - \phi^* + 2 \sum_{k=1}^{N} \gamma_k \gamma_k^2}{\sum_{k=1}^{N} \gamma_k \min_{i=1, \ldots, l} p_i (1 - \frac{L_i}{2} \gamma_k)}
\]

for any \( N \geq 1 \), where the expectation is taken w.r.t. \( i, k, G_{ik}, \) and \( R \).

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

\[
f(x_{k+1}) \leq f(x_k) + \langle g_k, x_{k+1} - x_k \rangle + \frac{L_{ik}}{2} \| x_{k+1} - x_k \|^2
\]

Using the above inequality and Lemma 4.2, we obtain

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

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

\[
\phi(x_{k+1}) \leq \phi(x_k) - \gamma_k \left( 1 - \frac{L_{ik}}{2} \gamma_k \right) \| \tilde{G}_k \|^2 + \gamma_k \langle \delta_k, \tilde{G}_k \rangle.
\]

Also observe that by (4.7), the definition of \( \tilde{G}_k \), and the fact \( U_{ik}^T G_k = G_{ik}(x_k, g_k, \gamma_k) \),

\[
\| \tilde{G}_k - U_{ik} G_k \|_{i_k, \ast} \leq \| G_{ik} - U_{ik}^T g_k \|_{i_k, \ast} = \| \delta_k \|_{i_k, \ast},
\]

and hence that

\[
\| U_{ik}^T G_k \|_{i_k}^2 = \| \tilde{G}_k + U_{ik}^T g_k - \tilde{G}_k \|_{i_k}^2 \leq 2 \| \tilde{G}_k \|^2 + 2 \| U_{ik}^T G_k - \tilde{G}_k \|^2
\]

\[
\leq 2 \| \tilde{G}_k \|^2 + 2 \| \delta_k \|_{i_k, \ast}^2,
\]

\[
\langle \delta_k, \tilde{G}_k \rangle = \langle \delta_k, U_{ik}^T G_k \rangle + \langle \delta_k, \tilde{G}_k - U_{ik}^T G_k \rangle \leq \langle \delta_k, U_{ik}^T G_k \rangle + \| \delta_k \|_{i_k, \ast} \| \tilde{G}_k - U_{ik}^T G_k \|_{i_k}
\]

\[
\leq \langle \delta_k, U_{ik}^T G_k \rangle + \| \delta_k \|_{i_k, \ast}.
\]

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

\[
\phi(x_{k+1}) \leq \phi(x_k) - \gamma_k \left( 1 - \frac{L_{ik}}{2} \gamma_k \right) \left( \frac{1}{2} \| U_{ik}^T G_k \|_{i_k}^2 - \| \delta_k \|_{i_k, \ast}^2 \right) + \gamma_k \langle \delta_k, U_{ik}^T G_k \rangle + \gamma_k \| \delta_k \|_{i_k, \ast}^2,
\]

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

\[
\sum_{k=1}^{N} \frac{L_{ik}}{2} \gamma_k \| U_{ik}^T G_k \|_{i_k}^2 \leq \phi(x_1) - \phi^* + \sum_{k=1}^{N} \gamma_k \langle \delta_k, U_{ik}^T G_k \rangle + \gamma_k \| \delta_k \|_{i_k, \ast}^2
\]

\[
+ \sum_{k=1}^{N} \gamma_k \left( 1 - \frac{L_{ik}}{2} \gamma_k \right) \| \delta_k \|_{i_k, \ast}^2
\]

\[
\leq \phi(x_1) - \phi^* + \sum_{k=1}^{N} \gamma_k \langle \delta_k, U_{ik}^T G_k \rangle + 2 \gamma_k \| \delta_k \|_{i_k, \ast}^2,
\]

where the last inequality follows from the facts that \( \phi(x_{k+1}) \geq \phi^* \) and \( L_{ik} \gamma_k^2 \| \delta_k \|_{i_k,s}^2 \geq 0 \). Now denoting \( \zeta_k = G_{ik} \), \( \zeta_{[k]} = \{ \zeta_1, \ldots, \zeta_k \} \) 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 (2.10) and (4.4),

\[
\mathbb{E}_{\zeta_k} \left[ \langle \delta_k, U^T_{ik} G_k \rangle | i_{[k]}, \zeta_{[k-1]} \right] = \mathbb{E}_{\zeta_k} \left[ \langle G_{ik} - U^T_{ik} g_k, U^T_{ik} G_k \rangle | i_{[k]}, \zeta_{[k-1]} \right] = 0,
\]

\[
\mathbb{E}_{i_N} \left[ \left(1 - \frac{L_{ik}}{2} \gamma_k \right) \|U^T_{ik} G_k\|_{i_{[k-1]}, i_{[k]}}^2 \right] = \sum_{i=1}^{b} p_i \left( 1 - \frac{L_{i}}{2} \gamma_k \right) \|U^T_{i} G_k\|^2 \geq \left( \sum_{i=1}^{b} \|U^T_{i} G_k\|^2 \right) \min_{i=1, \ldots, b} p_i \left( 1 - \frac{L_{i}}{2} \gamma_k \right) = \|G_k\|^2 \min_{i=1, \ldots, b} p_i \left( 1 - \frac{L_{i}}{2} \gamma_k \right),
\]

we conclude that

\[
\sum_{k=1}^{N} \gamma_k \min_{i=1, \ldots, b} p_i \left( 1 - \frac{L_{i}}{2} \gamma_k \right) \mathbb{E}_{\zeta_{[N]}, i_{[N]}} \left[ \|G_k\|^2 \right] \leq \phi(x_1) - \phi^* + 2 \sum_{k=1}^{N} \gamma_k \sigma_k^2.
\]

Dividing both sides of the above inequality by \( \sum_{k=1}^{N} \gamma_k \min_{i=1, \ldots, b} p_i \left( 1 - \frac{L_{i}}{2} \gamma_k \right) \), 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 \( \sigma_k = 0 \), \( k = 1, \ldots, N \), in (4.4). Suppose that the random variable \( \{ i_k \} \) are uniformly distributed (i.e., (2.30) holds). If \( \{ \gamma_k \} \) are set to

\[
\gamma_k = \frac{1}{\bar{L}}, k = 1, \ldots, N,
\]

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

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

**Proof.** By our assumptions about \( p_i \) and (4.13), we have

\[
\min_{i=1, \ldots, b} p_i \left( 1 - \frac{L_{i}}{2} \gamma_k \right) = \frac{1}{b} \min_{i=1, \ldots, b} \left( 1 - \frac{L_{i}}{2} \gamma_k \right) \geq \frac{1}{2b},
\]

which, in view of (4.11) and the fact that \( \sigma_k = 0 \), then implies that, for any \( N \geq 1 \),

\[
\mathbb{E}[\|G(x_R, g(x_R), \gamma_R)\|^2] \leq \frac{2b [\phi(x_1) - \phi^*]}{N} = \frac{2b \bar{L} [\phi(x_1) - \phi^*]}{N}.
\]
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 \( \mathbb{R}^n \) 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 \( G_{ik} \) is computed by using a mini-batch approach with size \( T_k \), i.e.,
\[
G_{ik} = \frac{1}{T_k} \sum_{t=1}^{T_k} U_{ik} \nabla F(x_k, \xi_{k,t}), \tag{4.17}
\]
for some \( T_k \geq 1 \), where \( \xi_{k,1}, \ldots, \xi_{k,T_k} \) are i.i.d. samples of \( \xi \).

**Corollary 4.5.** Assume that the random variables \( \{i_k\} \) are uniformly distributed (i.e., (2.30) holds). Also assume that \( G_{ik} \) is computed by (4.17) for \( T_k = T \) and that \( \{\gamma_k\} \) are set to (4.13). Then we have
\[
\mathbb{E}[\|G_{R}(x_R, g(x_R), \gamma_R)\|^2] \leq \frac{2bL[\phi(x_1) - \phi^*]}{N} + \frac{4b\sigma^2}{T} \tag{4.18}
\]
for any \( N \geq 1 \), where \( \tilde{L} \) is defined in (3.10).

**Proof.** Denote \( \delta_{k,t} = U_{ik} [\nabla F(x_k, \xi_{k,t}) - g(x_k)] \) and \( S_t = \sum_{i=1}^{t} \delta_{k,i} \). Noting that
\[
\mathbb{E}[\|S_{t-1}, \delta_{k,t}, |S_{t-1}| = 0 \text{ for all } t = 1, \ldots, T_k, \text{ we have}
\]
\[
\mathbb{E}[\|S_{T_k}\|^2] = \mathbb{E}[\|S_{T_k} - S_{T_k-1}\|^2 + 2 \langle S_{T_k-1}, \delta_{k,T_k} \rangle + \|\delta_{k,T_k}\|^2]
\leq \mathbb{E}[\|S_{T_k} - S_{T_k-1}\|^2] + \mathbb{E}[\|\delta_{k,T_k}\|^2] = \ldots = \sum_{t=1}^{T_k} \|\delta_{k,t}\|^2,
\]
which together with (4.17) then imply that the conditions in (4.4) hold with \( \bar{\sigma}^2_k = \sigma^2/T_k \). It then follows from the previous observation and (4.11) that
\[
\mathbb{E}[\|G_{R}(x_R, g(x_R), \gamma_R)\|^2] \leq \frac{2b[\phi(x_1) - \phi^*]}{N} + \frac{4b\sum_{k=1}^{N} \sigma^2}{T_k}
\leq \frac{2b\tilde{L}[\phi(x_1) - \phi^*]}{N} + \frac{4b\sigma^2}{T}.
\]

In view of Corollary 4.5, in order to find an \( \epsilon \) solution of problem (1.4), we need to have
\[
N = \mathcal{O}\left( \frac{b\tilde{L}}{\epsilon} [\phi(x_1) - \phi^*] \right) \quad \text{and} \quad T = \mathcal{O}\left( \frac{b\sigma^2}{\epsilon} \right), \tag{4.19}
\]
which implies that the total number of samples of \( \xi \) required can be bounded by
\[
\mathcal{O}\left( \frac{b^2 \tilde{L}^2 \sigma^2 [\phi(x_1) - \phi^*]}{\epsilon^2} \right).
\]
The previous bound is comparable, up to a constant factor \( b^2 \), to those obtained in [10, 11] 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 [10, 11]. However, the development of these results are more involved and hence the details are skipped.
5. Numerical Results. In this section we report the results from our preliminary numerical experiments for solving the following support vector machine problem:

\[
\min_{x \in [-R,R]^n} \mathbb{E}_{u,v} \left[ \max \left\{ 1 - \langle Bx, u \rangle v, 0 \right\} + \rho \| x \|_1 \right], \quad (5.1)
\]

where \( B \) denotes a basis and \( \rho \) is a certain user-defined parameter. Our goal is to show that the SBMD algorithm is implementable, and work in practice through the comparison with the Mirror Descent SA (MDSA) algorithm (see [25, 19]).

In our first experiment, the basis \( B \) is randomly generated according to \( B^T = Q \Lambda \hat{\Sigma} \), where \( \Lambda \) is a positive diagonal matrix and \( QQ^T = Q^T Q = I \). In particular, the diagonal entries of \( \Lambda \) are uniformly distributed over \([0,1]\), and \( Q \) is obtained from the eigenvalue decomposition of a positive semidefinite matrix \( H = LL^T \), where \( L \) is uniformly distributed over \([0,1]^{n \times n}\). In addition, we assume that \( u \) is uniformly distributed over \([0,1]^n \) and \( v \) is computed by \( v = \text{sign}(\langle B\hat{x}, u \rangle) \). First, we use an initial i.i.d. sample of size \( m = 100 \) for the random variable \((u,v)\) to estimate problem parameters, namely \( M_i, i = 1, \ldots, b \), which are required to specify the stepsizes \( \gamma_k \) in both SBMD and MDSA. The selection of these stepsizes are shown in Table 5. Second, we fix an i.i.d. sample of size \( s = 1,000 \) for the random variable \((u,v)\), which is used in these two algorithms to obtain the approximate solution of (5.1). Finally, we use another i.i.d. sample of size \( S = 10,000 \) to evaluate quality of the solutions generated by these algorithms.

In our second experiment, the basis \( B \) is obtained by metric learning. In particular, the basis is defined as \( B^T = Q \Lambda \hat{\Sigma} \), where \( \Lambda \) is a diagonal matrix whose diagonal elements are eigenvalues of \( \hat{X} \) which is a positive semidefinite learned metric obtained by solving

\[
\min_{X \in S} f(X) = \mathbb{E}[\| \text{tr}(X(a-a')^T(a-a')^T - b) \|],
\]

where \( S = \{X|\text{tr}(X) \leq c, X \succeq 0, \} \) and the expectation is taken with respect to \( \{a,a',b\} \) and \( c \in \mathbb{R}^n_+ \). We consider four data sets: w7a, Ionosphere, Mushroom and Sonar scale which can be download from the UC Irvine Machine Learning Repository at http://archive.ics.uci.edu/ml/datasets.html. Each data set is divided into there subsets: estimation, training and testing, similarly to the three steps for our first experiment described above. While problem parameters are estimated by using a sample size \( m = 20 \), the solution quality is evaluated by using another sample size \( S = 100 \) (except \( S = 201 \) for the data set Ionosphere), and the solutions are generated by using the rest of samples in these datasets.

In both experiments, we consider the case when each block has only one component, i.e., \( b = n \). The user-defined parameter is set to \( \rho = 1 \) and the radius of the box constraints is set to \( R = 10 \). The initial solution is randomly selected from the box \([-10,10]^n\). Following the schemes suggested by [34] to compare coordinate descent algorithms with the full gradient algorithm, we run MDSA and SBMD for \( N = 1000 \) and \( N = 1000b \), respectively. In view of the fact that these algorithms are stochastic, we run each one of them for 50 times, and report the average of the estimated objective values for the obtained solutions in Tables 1 and 2. All these algorithms are implemented in MATLAB R2009b on a Core i5 3.1 Ghz computer. From these experiments, we can clearly see that the performance of the SBMD algorithm is comparable to or even slightly better than that of the MDSA algorithm.

6. Conclusions. In this paper, we study a new class of stochastic algorithms, namely the SBMD methods, by incorporating the block decomposition and an incre-
Table 5.1
Stepsize parameters for the SVM problem (5.1) with randomly generated data

<table>
<thead>
<tr>
<th>Parameters</th>
<th>n = 100</th>
<th>n = 200</th>
<th>n = 300</th>
<th>n = 500</th>
</tr>
</thead>
<tbody>
<tr>
<td>γ_k for SBMD</td>
<td>0.1794</td>
<td>0.1855</td>
<td>0.0241</td>
<td>0.0084</td>
</tr>
<tr>
<td>γ_k for MDSA</td>
<td>0.1581</td>
<td>0.1581</td>
<td>0.0214</td>
<td>0.0074</td>
</tr>
</tbody>
</table>

Table 5.2
Objective values for the SVM problem (5.1) with randomly generated data

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>n = 100</th>
<th>n = 200</th>
<th>n = 300</th>
<th>n = 500</th>
</tr>
</thead>
<tbody>
<tr>
<td>SBMD</td>
<td>8.4805</td>
<td>8.5179</td>
<td>62.7181</td>
<td>280.0990</td>
</tr>
<tr>
<td>MDSA</td>
<td>10.5458</td>
<td>13.0168</td>
<td>72.2237</td>
<td>321.7414</td>
</tr>
</tbody>
</table>

mental 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.

While our paper discusses general nonsmooth problems (see Section 2), Fercoq and Richtárik [7] recently developed a novel block coordinate descent method for solving a certain class of nonsmooth problems with a bilinear saddle-point structure. Since both complexity results obtained in [7] and Section 2 are of the same order in terms of their dependence on ϵ and b, it will be interesting to see if we can further improve these algorithms by exploring the aforementioned bilinear saddle-point structure. It should be noted that Fercoq and Richtárik also consider parallelization in [7]. It will be interesting to see if similar ideas can be used to improve the performance of SBMD type algorithms.

Acknowledgement: The authors would like to thank Professors Stephen J. Wright and Yurii Nesterov for their encouragement and inspiring discussions in the study of this topic.

REFERENCES

[5] L.M. Bregman. The relaxation method of finding the common point convex sets and its appli-
Table 5.3  
Stepsize parameters for the SVM problem (5.1) with real data

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Ionosphere $n = 34$</th>
<th>Sonar scale $n = 60$</th>
<th>Mushroom $n = 112$</th>
<th>w7a $n = 300$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\gamma_k$ for SBMD</td>
<td>0.2089</td>
<td>0.1813</td>
<td>0.1784</td>
<td>0.1812</td>
</tr>
<tr>
<td>$\gamma_k$ for MDSA</td>
<td>0.1554</td>
<td>0.1562</td>
<td>0.1543</td>
<td>0.1581</td>
</tr>
</tbody>
</table>

Table 5.4  
Objective values for the SVM problem (5.1) with real data

<table>
<thead>
<tr>
<th>Algorithms</th>
<th>Ionosphere $n = 34$</th>
<th>Sonar scale $n = 60$</th>
<th>Mushroom $n = 112$</th>
<th>w7a $n = 300$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SBMD</td>
<td>1.9881</td>
<td>3.0044</td>
<td>7.1538</td>
<td>13.1216</td>
</tr>
<tr>
<td>MDSA</td>
<td>2.2533</td>
<td>3.0022</td>
<td>7.0678</td>
<td>13.3233</td>
</tr>
</tbody>
</table>


