Randomized Block Subgradient Methods for Convex Nonsmooth and Stochastic Optimization

Qi Deng * qdeng@ufl.edu
Guanghui Lan† glan@ise.ufl.edu
Anand Rangarajan* anand@cise.ufl.edu

Abstract

Block coordinate descent methods and stochastic subgradient methods have been extensively studied in optimization and machine learning. By combining randomized block sampling with stochastic subgradient methods based on dual averaging ([22, 36]), we present stochastic block dual averaging (SBDA)—a novel class of block subgradient methods for convex nonsmooth and stochastic optimization. SBDA requires only a block of subgradients and updates blocks of variables and hence has significantly lower iteration cost than traditional subgradient methods. We show that the SBDA-based methods exhibit the optimal convergence rate for convex nonsmooth stochastic optimization. More importantly, we introduce randomized stepsize rules and block sampling schemes that are adaptive to the block structures, which significantly improves the convergence rate w.r.t. the problem parameters. This is in sharp contrast to recent block subgradient methods applied to nonsmooth deterministic or stochastic optimization ([3, 24]). For strongly convex objectives, we propose a new averaging scheme to make the regularized dual averaging method optimal, without having to resort to any accelerated schemes.

1 Introduction

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

$$\min_{x\in X} \phi(x), \tag{1}$$

where the feasible set X is embedded in Euclidean space $\mathbb{R}^N$ for some integer $N > 0$. Letting $N_1, N_2, \ldots, N_n$ be n positive integers such that $\sum_{i=1}^n N_i = N$, we assume X can be partitioned as $X = X_1 \times X_2 \times \ldots X_n$, where each $X_i \subseteq \mathbb{R}^{N_i}$. We denote $x \in X$, by $x = x^{(1)} \times x^{(2)} \ldots \times x^{(n)}$ where $x^{(i)} \in X_i$. The objective $\phi(x)$ consists of two parts: $\phi(x) = f(x) + \omega(x)$. We stress that both $f(x)$ and $\omega(x)$ can be nonsmooth. $\omega(x)$ is a convex function with block separable structure: $\omega(x) = \sum_{i=1}^n \omega_i(x_i)$, where each $\omega_i : X_i \to \mathbb{R}$ is convex and relatively simple. In composite optimization or regularized learning, the term $\omega(x)$ imposes solutions with certain preferred structures. Common examples of $\omega(\cdot)$ include the ℓ_1 norm or squared ℓ_2 norm regularizers. $f(x)$ is a general convex function. In many important statistical learning problems, $f(x)$ has the form of $f(x) = \mathbf{E}_\xi[F(x,\xi)]$, where $F(x,\xi)$ is a convex loss function of $x \in X$ with ξ representing sampled data. When it is difficult to evaluate $f(x)$ exactly, as in batch learning or sample average approximation (SAA), $f(x)$ is approximated with finite data. Firstly, a large number of samples $\xi_1, \xi_2, \ldots, \xi_m$ are drawn, and then $f(x)$ is approximated by $\tilde{f}(x) = \frac{1}{m}\sum_{i=1}^m F(x,\xi_i)$, with the alternative problem:

*Department of Computer and Information Science and Engineering, University of Florida, FL, 32611
†Department of Industrial and Systems Engineering, University of Florida, FL, 32611

$$\min_{x \in X} \tilde{\phi}(x) := \tilde{f}(x) + \omega(x). \tag{2}$$

However, although classic first order methods can provide accurate solutions to (2), the major drawback of these approaches is the poor scalability to large data. First order deterministic methods require full information of the (sub)gradient and scan through the entire dataset many times, which is prohibitive for applications where scalability is paramount. In addition, due to the statistical nature of the problem, solutions with high precision may not even be necessary.

To solve the aforementioned problems, stochastic methods—stochastic (sub)gradient descent (SGD) or block coordinate descent (BCD) have received considerable attention in the machine learning community. Both of them confer new advantages in the trade offs between speed and accuracy. Compared to deterministic and full (sub)gradient methods, they are easier to implement, have much lower computational complexity in each iteration, and often exhibit sufficiently fast convergence while obtaining practically good solutions.

SGD was first studied in [29] in the 1950s, with the emphasis mainly on solving strongly convex problems; specifically it only needs the gradient/subgradient on a few data samples while iteratively updating all the variables. In the approach of online learning or stochastic approximation (SA), SGD directly works on the objective (1), and obtains convergence independent of the sample size. While early work emphasizes asymptotic properties, recent work investigate complexity analysis of convergence. Many works ([21, 13, 33, 26, 6, 2, 8]) investigate the optimal SGD under various conditions. Proximal versions of SGD, which explicitly incorporate the regularizer $\omega(x)$, have been studied, for example in [15, 4, 5, 36].

The study of BCD also has a long history. BCD was initiated in [18, 19], but the application of BCD to linear systems dates back to even earlier (for example see the Gauss-Seidel method in [7]). It works on the approximated problem (2) and makes progress by reducing the original problem into subproblems using only a single block coordinate of the variable at a time. Recent works [23, 28, 30, 17] study BCD with random sampling (RBCD) and obtain non-asymptotic complexity rates. For the regularized learning problem as in (2), RBCD on the dual formulation has been proposed [31, 11, 32]. Although most of the work on BCD focuses on smooth (composite) objectives, some recent work ([3, 37, 35, 39]) seeks to extend the realm of BCD in various ways. The works in [24, 3] discuss (block) subgradient methods for nonsmooth optimization. Combining the ideas of SGD and BCD, the works in [3, 37, 35, 39, 27] employ sampling of both features and data instances in BCD.

In this paper, we propose a new class of block subgradient methods, namely, stochastic block dual averaging (SBDA), for solving nonsmooth deterministic and stochastic optimization problems. Specifically, SBDA consists of a new dual averaging step incorporating the average of all past (stochastic) block subgradients and variable updates involving only block components. We bring together two strands of research, namely, the dual averaging algorithm (DA) [36, 22] which was studied for nonsmooth optimization and randomized coordinate descent (RCD) [23], employed for smooth deterministic problems. Our main contributions consist of the following:

- Two types of SBDA have been proposed for different purposes. For regularized learning, we propose SBDA-u which performs uniform random sampling of blocks. For more general nonsmooth learning problems, we propose SBDA-r which applies an optimal sampling scheme with improved convergence. Compared with existing subgradient methods for nonsmooth and stochastic optimization, both SBDA-u and SBDA-r have significantly lower iteration cost when the computation of block subgradients and block updates are convenient.

- We contribute a novel scheme of randomized stepsizes and optimized sampling strategies which

are truly adaptive to the block structures. Selecting block-wise stepsizes and optimal block sampling have been critical issues for speeding up BCD for smooth regularized problems, please see [23, 25, 31, 28] for some recent advances. For nonsmooth or stochastic optimization, the most closely related work to ours are [3, 24] which do not apply block-wise stepsizes. To the best of our knowledge, this is the *first time* block subgradient methods with block adaptive stepsizes and optimized sampling have been proposed for nonsmooth and stochastic optimization.

- We provide new theoretical guarantees of convergence of SBDA methods. SBDA obtains the optimal rate of convergence for general convex problems, matching the state of the art results in the literature of stochastic approximation and online learning. More importantly, SBDA exhibits a significantly improved convergence rate w.r.t. the problem parameters. When the regularizer $\omega(x)$ is strongly convex, our analysis provides a simple way to make the regularized dual averaging methods in [36] optimal. We show an *aggressive* weighting is sufficient to obtain $O\left(\frac{1}{T}\right)$convergence where T is the iteration count, without the need for any accelerated schemes. This appears to be a new result for simple dual averaging methods.

Related work Extending BCD to the realm of nonsmooth and stochastic optimization has been of interest lately. Efficient subgradient methods for a class of nonsmooth problems has been proposed in [24]. However, to compute the stepsize, the block version of this subgradient method requires computation of the entire subgradient and knowledge of the optimal value; hence, it may be not efficient in a more general setting. The methods in [3, 24] employ stepsizes that are not adaptive to the block selection and have therefore suboptimal bounds to our work. For SA or online learning, SBDA applies double sampling of both blocks and data. A similar approach has also been employed for new stochastic methods in some very recent work ([3, 39, 35, 27, 37]). It should be noted here that if the assumptions are strengthened, namely, in the batch learning formulation, and if $\tilde{\phi}$ is smooth, it is possible to obtain a linear convergence rate $O\left(e^{-T}\right)$. Nesterov's randomized block coordinate methods [23, 28] consider different stepsize rules and block sampling but only for smooth objectives with possible nonsmooth regularizers. Recently, nonuniform sampling in BCD has been addressed in [25, 38, 16] and shown to have advantages over uniform sampling. Although our work discusses block-wise stepsizes and nonuniform sampling as well, we stress the nonsmooth objectives that appear in deterministic and stochastic optimization . The proposed algorithms employ very different proof techniques, thereby obtaining different optimized sampling distributions.

Outline of the results.

We introduce two versions of SBDA that are appropriate in different contexts. The first algorithm, SBDA with uniform block sampling (SBDA-u) works for a class of convex composite functions, namely, $\omega(x)$ is explicate in the proximal step. When $\omega(x)$ is a general convex function, for example, the sparsity regularizer $\|x\|_1$, we show that SBDA-u obtains the convergence rate of $O\left(\frac{\sqrt{n}\sum_i^n\sqrt{M_i^2 D_i}}{\sqrt{T}}\right)$, which improves the rate of $O\left(\frac{\sqrt{n}\sqrt{\sum_i^n M_i^2}\cdot\sqrt{\sum_i^n D_i}}{\sqrt{T}}\right)$ by SBMD. Here $\{M_i\}$ and $\{D_i\}$ are some parameters associated with the blocks of coordinates to be specified later. When $\omega(x)$ is a strongly convex function, by using a more aggressive scheme to be later specified, SBDA-u obtains the optimal rate of $O\left(\frac{n\sum_i M_i^2}{\lambda T}\right)$, matching the result from SBMD. In addition, for general convex problems in which $\omega(x)=0$, we propose a variant of SBDA with nonuniform random sampling (SBDA-r) which achieves an improved convergence rate $O\left(\frac{\left(\sum_{j=1}^n M_j^{2/3} D_j^{1/3}\right)^{3/2}}{\sqrt{T}}\right)$. These

Algorithm	Objective	Complexity
SBDA-u	Convex composite	$O\left(\frac{\sqrt{n}\sum_i^n \sqrt{M_i^2 D_i}}{\sqrt{T}}\right)$
SBDA-u	Strongly convex composite	$O\left(\frac{n\sum_i M_i^2}{\lambda T}\right)$
SBDA-r	Convex nonsmooth	$O\left(\frac{\left(\sum_{j=1}^n M_j^{2/3} D_j^{1/3}\right)^{3/2}}{\sqrt{T}}\right)$

Table 1: Iteration complexity of our SBDA algorithms.

computational results are summarized in Table (1).

Structure of the Paper The paper proceeds as follows. Section 2 introduces the notation used in this paper. Section 3 presents and analyzes SBDA-u. Section 4 presents SBDA-r, and discusses optimal sampling and its convergence. Experimental results to demonstrate the performance of SBDA are provided in section 6. Section 7 draws conclusion and comments on possible future directions.

2 Preliminaries

Let $\mathbb{R}^N$ be a Euclidean vector space, $N_1, N_2, \ldots N_n$ be n positive integers such that $N_1 + \ldots N_n = N$. Let I be the identity matrix in $\mathbb{R}^{N\times N}$, U_i be a $N \times N_i$-dim matrix such that

$$I = [U_1 U_2 \ldots U_n].$$

For each $x \in \mathbb{R}^N$, we have the decomposition: $x = U_1 x^{(1)} + U_2 x^{(2)} + \ldots + U_n x^{(n)}$, where $x^{(i)} \in \mathbb{R}^{N_i}$.

Let $\|\cdot\|_{(i)}$ denote the norm on the $\mathbb{R}^{N_i}$, and $\|\cdot\|_{(i),*}$ be the induced dual norm. We define the norm $\|\cdot\|$ in $\mathbb{R}^N$ by: $\|x\|^2 = \sum_{i=1}^n \|x^{(i)}\|_{(i)}^2$ and its dual norm: $\|\cdot\|_*$ by $\|x\|_*^2 = \sum_{i=1}^n \|x^{(i)}\|_{(i),*}^2$

Let $d_i : X_i \to \mathbb{R}$ be a distance transform function with modulus $\|\cdot\|_{(i)}$ with respect to ρ.. $d_i(\cdot)$ is continuously differentiable and strongly convex:

$$d_i(\alpha x + (1-\alpha) y) \le \alpha d_i(x) + (1-\alpha) d_i(y) - \frac{1}{2}\rho\alpha(1-\alpha)\|x-y\|_{(i)}^2, \quad x, y \in X_i,$$

$i = 1, 2, ..., n$.

Let us assume there exists a solution $x^* \in X$ to the problem (1) , and

$$d_i(x^{*(i)}) \le D_i < \infty, \quad i = 1, 2, \ldots n, \tag{3}$$

Without loss of generality, we assume $d_i(\cdot)$ is nonnegative, and write

$$d(x) = \sum_i^n d_i\left(x^{(i)}\right) \tag{4}$$

for simplicity. Further more, we define the Bregman divergence associated with $d_i(\cdot)$ by

$$\mathcal{V}_i(z, x) = d_i(x) - d_i(z) - \langle \nabla_i d(z), x - z\rangle, \quad z, x \in X_i.$$

and $\mathcal{V}(z, x) = \sum_i^n \mathcal{V}_i\left(z^{(i)}, x^{(i)}\right)$.

We denote $f(x) = E_\xi [F(x,\xi)]$, and let $G(x,\xi)$ be a subgradient of $F(x,\xi)$, and $g(x) = E_\xi [G(x,\xi)] \in \partial f(x)$ be a subgradient of $f(x)$. Let $g^{(i)}(\cdot)$, $G^{(i)}(x,\xi)$ denote their i-th block components , for $i = 1, 2, \ldots, n$. Throughout the paper, we assume the (stochastic) block coordinate subgradient of f satisfying:

$$\|g^{(i)}(x)\|_{(i),*}^2 = \mathbf{E}^2 \left[\|G^{(i)}(x,\xi)\|_{(i),*}\right] \leq \mathbf{E}\left[\|G^{(i)}(x,\xi)\|_{(i),*,}^2\right] \leq M_i^2, \quad \forall x \in X \tag{5}$$

for $i = 1, 2, \ldots, n$. Note that although we make assumptions of stochastic objective , the following analysis and conclusions naturally extend to deterministic optimization. To see that, we can simply assume $g(x) \equiv G(x,\xi)$, and $f(x) \equiv F(x,\xi)$, for any ξ.

Before introducing the main convergence properties, we first summarize several useful results in the following lemmas. Lemma 1, 2, and 3 slightly generalize the results in [34, 14], [22], and [13] respectively; their proofs are left in Appendix.

Lemma 1. *Let $f(\cdot)$ be a lower semicontinuous convex function and $d(\cdot)$ be defined by (4). If*

$$z = \arg\min_x \Psi(x) \coloneqq f(x) + d(x),$$

then

$$\Psi(x) \geq \Psi(z) + \mathcal{V}(z,x), \quad \forall x \in X.$$

Moreover, if $f(x)$ is λ-strongly convex with norm $\|\cdot\|_{(i)}$, and $x = z + U_i y \in X$ where $y \in X_i$, $z \in X$, then

$$\Psi(x) \geq \Psi(z) + \mathcal{V}(z,x) + \frac{\lambda}{2}\|y\|_{(i)}^2, \quad \forall x \in X.$$

Lemma 2. *Let $\Psi : X \to \mathbb{R}$ be convex, block separable, and ρ_i-strongly convex with modulus ρ_i w.r.t. $\|\cdot\|_{(i)}$, $\rho_i > 0$, $1 \leq i \leq n$, and $g \in \mathbb{R}^N$. If*

$$x_0 \in \arg\min_{x \in X} \{\Psi(x)\}, \text{ and } z \in \arg\min_{x \in X} \left\{\left\langle U_i g^{(i)}, x \right\rangle + \Psi(x)\right\},$$

then

$$\left\langle U_i g^{(i)}, x_0 \right\rangle + \Psi(x_0) \leq \left\langle U_i g^{(i)}, z \right\rangle + \Psi(z) + \frac{1}{2\rho_i}\|g\|_{(i),*}^2.$$

Lemma 3. *If f satisfies the assumption (5), let $x = z + U_i y \in X$ where $y \in X_i$, $x \in X$, then*

$$f(z) \leq f(x) + \langle g^{(i)}(x), y\rangle + 2M_i\|y\|_{(i)}. \tag{6}$$

3 Uniformly randomized SBDA (SBDA-u)

In this section, we describe uniformly randomized SBDA (SBDA-u) for the composite problem (1). We consider the formulation proposed in [36], since it incorporates the regularizers for composite problems. The main update of the DA algorithm has the form

$$x_{t+1} = \arg\min_{x \in X} \left\{\sum_{s=1}^{t} \langle G_s, x\rangle + t\omega(x) + \beta_t d(x)\right\}, \tag{7}$$

where $\{\beta_t\}$ is a parameter sequence and G_s is shorthand for $G(x_s, \xi_s)$, and $d(x)$ is a strongly convex proximal function. When $\omega(x) = 0$, this reduces to a version of Nesterov's primal-dual subgradient method [22].

Let $\bar{G} = \sum_{s=0}^{t} \alpha_s U_{i_s} G^{(i_s)}(x_s, \xi_s)$, where $\{\alpha_t\}$ is a sequence of positive values, $\{i_t\}$ is a sequence of sampled indices. The main iteration step of SBDA has the form

$$x_{t+1}^{(i_t)} = \arg\min_{x \in X_{i_t}} \left\{ \langle \bar{G}^{(i_t)}, x \rangle + l_t^{(i_t)} \omega_{i_t}(x) + \gamma_t^{(i_t)} d_{i_t}(x) \right\}, \tag{8}$$

and $x_{t+1}^{(i)} = x_t^{(i)}$, $i \neq i_t$.

We highlight two important aspects of the proposed iteration (8). Firstly, the update in (8) incorporates the past randomly sampled block (stochastic) subgradients $\left\{G^{(i_t)}(x_t, \xi_t)\right\}$, rather than the full (stochastic) subgradients. Meanwhile, the update of the primal variable is restricted to the same block (i_t), leaving the other blocks untouched. Such block decomposition significantly reduces the iteration cost of the dual averaging method when the block-wise operation is convenient. Secondly, (8) employs a novel randomized stepsize sequence $\{\gamma_t\}$ where $\gamma_t \in \mathbb{R}^n$. More specifically, γ_t depends not only on the iteration count t, but also on the block index i_t. $\{\gamma_t\}$ satisfies the assumptions,

$$\gamma_t^{(j)} = \gamma_{t-1}^{(j)}, j \neq i_t, \text{ and } \gamma_t^{(j)} \geq \gamma_{t-1}^{(j)}, j = i_t. \tag{9}$$

The most important aspect of (9) is that stepsizes can be specified for each block of coordinates, thereby allowing for aggressive descent. As will be shown later, the rate of convergence, in terms of the problem parameters, can be significantly reduced by properly choosing these control parameters. In addition, we allow the sequence $\{\alpha_t\}$ and the related $\{l_t\}$ to be variable, hence offer the opportunity of different averaging schemes in composite settings. To summarize, the full SBDA-u is described in Algorithm 1.

Input: convex composite function $\phi(x) = f(x) + \omega(x)$, a sequence of samples $\{\xi_t\}$;
initialize $\alpha_{-1} \in \mathbb{R}$, $\gamma_{-1} \in \mathbb{R}^n$, $l_{-1} = \mathbf{0}^n$, $\bar{G} = \mathbf{0}^N$, $x_0 = \arg\min_{x \in X} \sum_{i=1}^{n} \gamma_{-1}^{(i)} d_i\left(x^{(i)}\right)$;
for $t = 0, 1, \ldots, T-1$ **do**
  sample a block $i_t \in \{1, 2, \ldots, n\}$ with uniform probability $\frac{1}{n}$;
  set $\gamma_t^{(i)}$, $i = 1, 2, \ldots, n$;
  set $l_t^{(i_t)} = l_{t-1}^{(i_t)} + \alpha_t$ and $l_t^{(j)} = l_{t-1}^{(j)}$ for $j \neq i_t$;
  update $\bar{G}$: $\bar{G} = \bar{G} + \alpha_t U_{i_t} G^{(i_t)}(x_t, \xi_t)$;
  update $x_{t+1}^{(i_t)} = \arg\min_{x \in X_{i_t}} \left\{ \langle \bar{G}^{(i_t)}, x \rangle + l_t^{(i_t)} \omega_{i_t}(x) + \gamma_t^{(i_t)} d_{i_t}(x) \right\}$;
  $x_{t+1}^{(j)} = x_t^{(j)}$, for $j \neq i_t$;
end
Output: $\bar{x} = \left[\sum_{t=1}^{T} \left(\alpha_{t-1} - \frac{n-1}{n}\alpha_t\right) x_t\right] / \sum_{t=1}^{T} \left(\alpha_{t-1} - \frac{n-1}{n}\alpha_t\right)$;
Algorithm 1: Uniformly randomized stochastic block dual averaging (SBDA-u) method.

The following theorem illustrates an important relation to analyze the convergence of SBDA-u. Throughout the analysis we assume the simple function $\omega(x)$ is λ-strongly convex with modulus λ, where $\lambda \geq 0$.

Theorem 4. *In algorithm 1, if the sequence $\{\gamma_t\}$ satisfies the assumption (9) , then for any $x \in X$,*

we have

$$\sum_{t=1}^{T}\left(\alpha_{t-1}-\frac{n-1}{n}\alpha_t\right)\mathbf{E}\left[\phi\left(x_t\right)-\phi\left(x\right)\right] \leq \alpha_0\frac{n-1}{n}\left[\phi\left(x_0\right)-\phi\left(x\right)\right]+\sum_{i=1}^{n}\mathbf{E}\left[\gamma_{T-1}^{(i)}\right]d_i\left(x\right) + \sum_{i=1}^{n}\frac{5M_i^2}{n}\sum_{t=0}^{T-1}\mathbf{E}\left[\frac{\alpha_t^2}{\left(\gamma_{t-1}^{(i)}\rho+l_{t-1}^{(i)}\lambda\right)}\right]. \tag{10}$$

Proof. Firstly, to simplify the notation, when there is no ambiguity, we use the terms $\omega_i\left(x\right)$ and $\omega_i\left(x^{(i)}\right)$, $d_i\left(x\right)$ and $d_i\left(x^{(i)}\right)$, $\mathcal{V}_i\left(x,y\right)$ and $\mathcal{V}_i\left(x^{(i)},y^{(i)}\right)$ interchangeably. In addition, we denote $\omega_{i^c}\left(x\right)=\omega\left(x\right)-\omega_i\left(x\right)$, and an auxiliary function by

$$\Psi_t\left(x\right)=\begin{cases}\sum_{s=0}^{t}\alpha_s\left[F\left(x_s,\xi_s\right)+\left\langle G_s,x-x_s\right\rangle_{(i_s)}+\omega_{i_s}\left(x\right)\right]+\sum_{i=1}^{n}\gamma_t^{(i)}d_i\left(x^{(i)}\right), & t\geq 0\\ \sum_{i=1}^{n}\gamma_t^{(i)}d_i\left(x^{(i)}\right) & t=-1\end{cases}. \tag{11}$$

It can be easily seen from the definition that x_{t+1} is the minimizer of the problem $\min_{x\in X}\Psi_t\left(x\right)$. Moreover, by the assumption on $\{\gamma_t\}$, we obtain

$$\Psi_t\left(x\right)-\Psi_{t-1}\left(x\right)\geq\alpha_t\left[F\left(x_t,\xi_t\right)+\left\langle G_t,x-x_t\right\rangle_{(i_t)}+\omega_{i_t}\left(x\right)\right]. \quad t=0,1,2,\ldots \tag{12}$$

Applying Lemma 3 and the property equation (12) at $x=x_{t+1}$, we have

$$\begin{aligned}\phi\left(x_{t+1}\right) &\leq f\left(x_t\right)+\left\langle g_t,x_{t+1}-x_t\right\rangle+2M_{i_t}\|x_{t+1}-x_t\|_{(i_t)}+\omega\left(x_{t+1}\right)\\ &= F\left(x_t,\xi_t\right)+\left\langle G_t,x_{t+1}-x_t\right\rangle_{(i_t)}+2M_{i_t}\|x_{t+1}-x_t\|_{(i_t)}\\ &\quad +f\left(x_t\right)-F\left(x_t,\xi_t\right)+\left\langle g_t-G_t,x_{t+1}-x_t\right\rangle_{(i_t)}+\omega\left(x_{t+1}\right)\\ &\leq \underbrace{\frac{1}{\alpha_t}\left[\Psi_t\left(x_{t+1}\right)-\Psi_{t-1}\left(x_{t+1}\right)+\frac{\gamma_{t-1}^{(i_t)}\rho+l_{t-1}^{(i_t)}\lambda}{2}\|x_{t+1}-x_t\|_{(i_t)}^2\right]}_{\Delta_1}\\ &\quad +f\left(x_t\right)-F\left(x_t,\xi_t\right)+\omega_{i_t^c}\left(x_{t+1}\right)\\ &\quad \underbrace{+\left\langle g_t-G_t,x_{t+1}-x_t\right\rangle_{(i_t)}-\frac{\gamma_{t-1}^{(i_t)}\rho+l_{t-1}^{(i_t)}\lambda}{2\alpha_t}\|x_{t+1}-x_t\|_{(i_t)}^2+2M_{i_t}\|x_{t+1}-x_t\|_{(i_t)}}_{\Delta_2}.\end{aligned}$$

We proceed with the analysis by separately taking care of Δ_1 and Δ_2. We first provide a concrete bound on Δ_1. Applying Lemma 1 for $\Psi=\Psi_{t-1}$ with x_t being the optimal point $x=x_{t+1}$, we obtain

$$\Psi_{t-1}\left(x_{t+1}\right)\geq\Psi_{t-1}\left(x_t\right)+\sum_{i=1}^{n}\gamma_{t-1}^{(i)}\mathcal{V}_i\left(x_t,x_{t+1}\right)+\frac{l_{t-1}^{(i_t)}\lambda}{2}\|x_t-x_{t+1}\|_{(i_t)}^2. \tag{13}$$

In view of (13) and the assumption $\mathcal{V}_i\left(x_t,x_{t+1}\right)\geq\frac{\rho}{2}\|x_{t+1}-x_t\|_{(i)}^2$, we obtain an upper bound on Δ_1: $\Delta_1\leq\Psi_t\left(x_{t+1}\right)-\Psi_{t-1}\left(x_t\right)$. On the other hand, from the Cauchy-Schwarz inequality, we have $\left\langle g_t-G_t,x_{t+1}-x_t\right\rangle_{(i_t)}\leq\|g_t-G_t\|_{(i_t),*}\cdot\|x_{t+1}-x_t\|_{(i_t)}$. Then

$$\Delta_2\leq\|x_{t+1}-x_t\|_{(i_t)}\cdot\left(\|g_t-G_t\|_{(i_t)}+2M_{i_t}\right)-\frac{\gamma_{t-1}^{(i_t)}\rho+l_{t-1}^{(i_t)}\lambda}{2\alpha_t}\|x_{t+1}-x_t\|_{(i_t)}^2.$$

The right side of the above inequality is a quadratic function of $\|x_{t+1}-x_t\|_{(i_t)}$. By maximizing it, we obtain

$$\Delta_2 \leq \frac{\alpha_t \left(\|g_t - G_t\|_{(i_t)} + 2M_{i_t}\right)^2}{2\left(\gamma_{t-1}^{(i_t)}\rho + l_{t-1}^{(i_t)}\lambda\right)}.$$

In view of these bounds on Δ_1 and Δ_2, and the fact that $\omega_{i_t^c}(x_t) = \omega_{i_t^c}(x_{t+1})$, we have

$$\begin{aligned}\alpha_t\phi(x_{t+1}) &\leq \Psi_t(x_{t+1}) - \Psi_{t-1}(x_t) + \alpha_t\left[f(x_t) - F(x_t,\xi_t) + \omega_{i_t^c}(x_t)\right] \\ &\quad + \frac{\alpha_t^2\left(\|g_t - G_t\|_{(i_t)} + 2M_i\right)^2}{\gamma_{t-1}^{(i_t)}\rho + l_{t-1}^{(i_t)}\lambda}. \end{aligned} \tag{14}$$

Summing up the above for $t = 0, 1, \ldots, T-1$, and observing that $\Psi_{-1} \geq 0$, $d_i(x_0) \geq 0$ $(1 \leq i \leq n)$, we obtain

$$\begin{aligned}\sum_{t=0}^{T-1}\alpha_t\phi(x_{t+1}) &\leq \Psi_{T-1}(x_T) + \sum_{t=0}^{T-1}\frac{\alpha_t^2\left(\|g_t - G_t\|_{(i_t)} + 2M_i\right)^2}{\gamma_{t-1}^{(i_t)}\rho + l_{t-1}^{(i_t)}\lambda} \\ &\quad + \sum_{t=0}^{T-1}\alpha_t\left[f(x_t) - F(x_t,\xi_t) + \omega_{i_t^c}(x_t)\right]. \end{aligned} \tag{15}$$

Due to the optimality of x_T, for $x \in X$, we have

$$\begin{aligned}\Psi_{T-1}(x_T) &\leq \Psi_{T-1}(x) \\ &= \sum_{t=0}^{T-1}\alpha_t\left[f(x_t) + \frac{1}{n}\langle g_t, x - x_t\rangle + \omega_{i_t}(x^*)\right] + \sum_{i=1}^{n}\gamma_{T-1}^{(i)}d_i(x) \\ &\quad + \sum_{t=0}^{T-1}\alpha_t\left[\langle G_t, x - x_t\rangle_{(i_t)} - \frac{1}{n}\langle g_t, x - x_t\rangle\right] \\ &\leq \sum_{t=0}^{T-1}\alpha_t\left[\frac{n-1}{n}f(x_t) + \frac{1}{n}f(x) + \omega_{i_t}(x)\right] + \sum_{i=1}^{n}\gamma_{T-1}^{(i)}d_i(x) \\ &\quad + \sum_{t=0}^{T-1}\alpha_t\left[\langle G_t, x - x_t\rangle_{(i_t)} - \frac{1}{n}\langle g_t, x - x_t\rangle\right], \end{aligned} \tag{16}$$

where the last inequality follows from the convexity of f:$\langle g_t, x - x_t\rangle \leq f(x) - f(x_t)$. Putting (15) and (16) together yields

$$\begin{aligned}\sum_{t=0}^{T-1}\alpha_t\phi(x_{t+1}) &\leq \sum_{t=0}^{T-1}\alpha_t\left[\frac{n-1}{n}\phi(x_t) + \frac{1}{n}\phi(x)\right] + \sum_{i=1}^{n}\gamma_{T-1}^{(i)}d_i(x) \\ &\quad + \sum_{0}^{T-1}\frac{\alpha_t^2\left(\|g_t - G_t\|_{(i_t)} + 2M_i\right)^2}{2\left(\gamma_{t-1}^{(i_t)}\rho + l_{t-1}^{(i_t)}\lambda\right)} + \delta_T, \end{aligned} \tag{17}$$

where δ_T is defined by

$$\begin{aligned}\delta_T &= \sum_{t=0}^{T-1}\alpha_t\left[\langle G_t, x - x_t\rangle_{(i_t)} - \frac{1}{n}\langle g_t, x - x_t\rangle + f(x_t) - F(x_t,\xi_t)\right] \\ &\quad + \sum_{t=0}^{T-1}\alpha_t\left[\omega_{i_t^c}(x_t) - \frac{n-1}{n}\omega(x_t) + \omega_{i_t}(x) - \frac{1}{n}\omega(x)\right]. \end{aligned} \tag{18}$$

In (17), subtracting $\sum_{t=0}^{T-1}\phi(x)$, and then $\frac{n-1}{n}\sum_{t=1}^{T}\alpha_t[\phi(x_t)-\phi(x)]$ on both sides , one has

$$\sum_{t=1}^{T}\left(\alpha_{t-1}-\frac{n-1}{n}\alpha_t\right)[\phi(x_t)-\phi(x)] \leq \frac{n-1}{n}\alpha_0[\phi(x_0)-\phi(x)]+\sum_{i=1}^{n}\gamma_{T-1}^{(i)}d_i(x) + \delta_T+\sum_{0}^{T-1}\frac{\alpha_t^2\left(\|g_t-G_t\|_{(i_t)}+2M_{i_t}\right)^2}{2\left(\gamma_{t-1}^{(i_t)}\rho+l_{t-1}^{(i_t)}\lambda\right)}. \quad (19)$$

Now let us take the expectation on both sides of (19). Firstly, taking the expectation with respect to i_t, for $t=0,1,...,T-1$, we have $\mathbf{E}_{i_t}\left[\langle G_t,x^*-x_t\rangle_{(i_t)}\right]=\frac{1}{n}\langle G_t,x^*-x_t\rangle$, and $\mathbf{E}_{i_t}\left[\omega_{i_t^c}(x_t)\right]=\omega(x_t)-\mathbf{E}_{i_t}[\omega_{i_t}(x_t)]=\frac{n-1}{n}\omega(x_t)$. Moreover, by the assumptions $\mathbf{E}_{\xi_t}[F(x_t,\xi_t)]=f(x_t)$, $\mathbf{E}_{\xi_t}[G(x_t,\xi_t)]=g(x_t)$. Together with the definition (18), we see $E[\delta_t]=0$. In addition, from the Cauchy-Schwarz inequality, we have $\left(\|g_t-G_t\|_{(i_t)}+2M_{i_t}\right)^2\leq 2\left(\|g_t-G_t\|_{(i_t)}^2+4M_{i_t}^2\right)$, and the expectation $E_{\xi_t}\left[\|g_t-G_t\|_{(i_t)}^2\right]\leq E_{\xi_t}\|G_t\|_{(i_t)}^2\leq M_{i_t}^2$. Furthermore, since ξ_t is independent of γ_{t-1} and l_{t-1}, we have

$$\begin{aligned}\mathbf{E}\left[\frac{\left(\|g_t-G_t\|_{(i_t)}+2M_i\right)^2}{\gamma_{t-1}^{(i_t)}\rho+l_{t-1}^{(i_t)}\lambda}\right] &\leq \mathbf{E}\left[\mathbf{E}_{\xi_t}\left(\frac{2\left(\|g_t-G_t\|_{(i_t)}^2+4M_{i_t}^2\right)}{\gamma_{t-1}^{(i_t)}\rho+l_{t-1}^{(i_t)}\lambda}\right)\right]\\ &\leq \mathbf{E}\left[\left(\frac{10M_{i_t}^2}{\gamma_{t-1}^{(i_t)}\rho+l_{t-1}^{(i_t)}\lambda}\right)\right]\\ &= \sum_{i=1}^{n}\mathbf{E}\left[\frac{10M_i^2}{n\left(\gamma_{t-1}^{(i)}\rho+l_{t-1}^{(i)}\lambda\right)}\right].\end{aligned}$$

Using these results, we obtain

$$\sum_{t=1}^{T}\left(\alpha_{t-1}-\frac{n-1}{n}\alpha_t\right)\mathbf{E}[\phi(x_t)-\phi(x)] \leq \alpha_0\frac{n-1}{n}[\phi(x_0)-\phi(x)]+\sum_{i=1}^{n}\mathbf{E}\left[\gamma_{T-1}^{(i)}\right]d_i(x) + \sum_{i=1}^{n}\frac{5M_i^2}{n}\sum_{t=0}^{T-1}\mathbf{E}\left[\frac{\alpha_t^2}{\left(\gamma_{t-1}^{(i)}\rho+l_{t-1}^{(i)}\lambda\right)}\right].$$

□

In Theorem 4 we presented some general convergence properties of SBDA-u for both stochastic convex and strongly convex functions. It should be noted that the right side of (10) employs expectations since both $\{\gamma_t\}$ and $\{l_t\}$ can be random. In the sequel, we describe more specialized convergence rates for both cases. Let us take $x=x^*$ and use the assumption (3) throughout the analysis.

Convergence rate when $\omega(x)$ is a simple convex function

Firstly, we consider a constant stepsize policy and assume that $\gamma_t^{(i)}$ depends on i and T where T is the iteration number. More specifically, let $\alpha_t\equiv 1$, and $\gamma_t^{(i)}\equiv\beta_i$ for some $\beta_i>0$,$1\leq i\leq n$,

$-1 \le t \le T$. Then $\mathbf{E}\left[\frac{\alpha_t^2}{\gamma_{t-1}^{(i)}\rho}\right] = \frac{1}{\rho\beta_i}$, for $1 \le i \le n$, and hence

$$\sum_{t=1}^{T} \mathbf{E}\left[\phi\left(x_t\right) - \phi\left(x^*\right)\right] \le \left(n-1\right)\left[\phi\left(x_0\right) - \phi\left(x^*\right)\right] + n\sum_{i=1}^{n}\beta_i D_i + T\sum_{i=1}^{n}\frac{5M_i^2}{\rho\beta_i}.$$

Let us choose $\beta_i = \sqrt{\frac{5TM_i^2}{n\rho D_i}}$ for $i = 1, 2, \ldots, p$, to optimize the above function. We obtain an upper bound on the error term:

$$\sum_{t=1}^{T} \mathbf{E}\left[\phi\left(x_t\right) - \phi\left(x^*\right)\right] \le \left(n-1\right)\left[\phi\left(x_0\right) - \phi\left(x^*\right)\right] + 2\sqrt{\frac{5Tn}{\rho}}\sum_{i=1}^{n}\sqrt{M_i^2 D_i}.$$

If we use the average point $\bar{x} = \sum_{t=1}^{T} x_t/T$ as the output, we obtain the expected optimization error:

$$\mathbf{E}\left[\phi\left(\bar{x}\right) - \phi\left(x^*\right)\right] \le \frac{n-1}{T}\left[\phi\left(x_0\right) - \phi\left(x^*\right)\right] + \frac{2\sqrt{5n}\left[\sum_{i=1}^{n}\sqrt{M_i^2 D_i}\right]}{\sqrt{\rho}\sqrt{T}}.$$

In addition, we can also choose varying stepsizes without knowing ahead the iteration number T. Differing from traditional stepsize policies where γ_t is usually associated with t, here $\left\{\gamma_t^{(i)}\right\}$ is a random sequence dependent on both t and i_t. In order to establish the convergence rate with such a randomized γ_t, we first state a useful technical result.

Lemma 5. *Let p be a real number with $0 < p < 1$, $\{a_s\}$ and $\{b_t\}$ be sequences of nonnegative numbers satisfying the relation:*

$$a_t = pb_t + \left(1-p\right)a_{t-1}, \quad t = 1, 2, \ldots$$

Then

$$\sum_{s=0}^{t} a_s \le \sum_{s=1}^{t} b_s + \frac{a_0}{p}.$$

We first let $\alpha_t \equiv 1$, and define $\{\gamma_t\}$ recursively as

$$\gamma_t^{(i)} = \begin{cases} u_i\sqrt{t+1} & i = i_t \\ \gamma_{t-1}^{(i)} & i \ne i_t \end{cases},$$

for some $u_i > 0$, $i = 1, 2, ..., n$, $t = 0, 1, 2, \ldots, T$. From this definition, we obtain

$$\mathbf{E}\left[\frac{1}{\gamma_{t-1}^{(i)}}\right] = \frac{1}{n}\frac{1}{u_i\sqrt{t}} + \frac{n-1}{n}\mathbf{E}\left[\frac{1}{\gamma_{t-2}^{(i)}}\right].$$

Observing the fact that $\sum_{\tau=1}^{t}\frac{1}{\sqrt{\tau}} \le \int_1^{t+1}\frac{1}{\sqrt{x}}dx = 2\sqrt{t+1}$ and applying Lemma 5 with $a_t = \mathbf{E}\left[\frac{1}{\gamma_{t-1}^{(i)}}\right]$ and $b_t = \frac{1}{u_i\sqrt{t}}$, we have

$$\sum_{\tau=0}^{t}\mathbf{E}\left[\frac{1}{\gamma_{\tau-1}^{(i)}}\right] \le \frac{1}{u_i}\sum_{\tau=1}^{t}\frac{1}{\sqrt{\tau}} + \frac{n}{\gamma_{-1}^{(i)}} \le \frac{2\sqrt{t+1}}{u_i} + \frac{n}{\gamma_{-1}^{(i)}}.$$

Hence

$$\sum_{t=0}^{T-1}\mathbf{E}\left[\frac{1}{\gamma_{t-1}^{(i)}\rho}\right]\leq\frac{1}{\rho}\left[\frac{2\sqrt{T}}{u_i}+\frac{n}{\gamma_{-1}^{(i)}}\right],\quad i=1,2,\dots n. \tag{20}$$

With respect to (20) and Theorem 1 , we obtain

$$\sum_{i=1}^{n}\mathbf{E}\left[\gamma_{T-1}^{(i)}\right]d_i\left(x^*\right)+\sum_{t=0}^{T-1}\sum_{i=1}^{n}\mathbf{E}\left[\frac{5\alpha_t^2M_i^2}{n\gamma_{t-1}^{(i)}\rho}\right]\leq\sum_{i=1}^{n}u_i\sqrt{T}D_i+\sum_{i=1}^{n}\left\{\frac{5M_i^2}{n\rho}\left[\frac{2\sqrt{T}}{u_i}+\frac{n}{\gamma_{-1}^{(i)}}\right]\right\}.$$

Choosing $u_i=\sqrt{\frac{10M_i^2}{n\rho D_i}}$, we have

$$\mathbf{E}\left[\phi\left(\bar{x}\right)-\phi\left(x^*\right)\right]\leq\frac{n-1}{T}\left[\phi\left(x_0\right)-\phi\left(x^*\right)\right]+\sum_{i=1}^{n}\frac{5nM_i^2}{\rho\gamma_{-1}^{(i)}T}+\frac{2\sum_{i=1}^{n}\sqrt{10nM_i^2D_i}}{\sqrt{\rho}\sqrt{T}}.$$

We summarize the results in the following corollary:

Corollary 6. *In algorithm 1, let $T>0$, $\bar{x}$ be the average point $\bar{x}=\sum_{t=1}^{T}x_t/T$, and $\alpha_t\equiv1$.*

1. If $\gamma_t^{(i)}=\sqrt{\frac{5TM_i^2}{n\rho D_i}}$, for $t=0,1,2,...,T-1$, $i=1,2,...,n$, then

$$\mathbf{E}\left[\phi\left(\bar{x}\right)-\phi\left(x^*\right)\right]\leq\frac{\left(n-1\right)\left[\phi\left(x_0\right)-\phi\left(x^*\right)\right]}{T}+\frac{2\sum_{i=1}^{n}\sqrt{5nM_i^2D_i}}{\sqrt{\rho}\sqrt{T}};$$

2. If $\gamma_t^{(i)}=\begin{cases}\sqrt{\frac{10M_i^2(t+1)}{n\rho D_i}} & \text{if } i=i_t\\ \gamma_{t-1}^{(i)} & \text{o.w.}\end{cases}$, for $t=0,1,2,...,T-1$, and $\gamma_{-1}^{(i)}=\sqrt{\frac{10M_i^2}{n\rho D_i}}$, $i=1,2,...,n$, then

$$\mathbf{E}\left[\phi\left(\bar{x}\right)-\phi\left(x^*\right)\right]\leq\frac{n-1}{T}\left[\phi\left(x_0\right)-\phi\left(x^*\right)\right]+\sum_{i=1}^{n}\frac{5nM_i^2}{\rho\gamma_{-1}^{(i)}T}+\frac{2\sum_{i=1}^{n}\sqrt{10nM_i^2D_i}}{\sqrt{\rho}\sqrt{T}}.$$

Corollary 6 provides both constant and adaptive stepsizes and SBDA-u obtains a rate of convergence of $O\left(1/\sqrt{T}\right)$for both, which matches the optimal rate for nonsmooth stochastic approximation [please see (2.48) in [21]]. In the context of nonsmooth deterministic problem, it also matches the convergence rate of the subgradient method. However, it is more interesting to compare this with the convergence rate of BCD methods [please see, for example, Corollary 2.2 part b) in [3]]. Ignoring the higher order terms, their convergence rate reads: $O\left(\frac{\sqrt{\sum_{i=1}^{n}M_i^2}}{\sqrt{T}}\sqrt{n\sum_{i=1}^{n}D_i}\right)$. Although the rate of $O\left(1/\sqrt{T}\right)$ is unimprovable, it can be seen (using the Cauchy-Schwarz inequality) that

$$\sum_{i=1}^{n}\sqrt{M_i^2D_i}\leq\sqrt{\sum_{i=1}^{n}M_i^2}\sqrt{\sum_{i=1}^{n}D_i},$$

with the equality holding if and only if the ratio M_i^2/D_i is equal to some positive constant, $1\leq i\leq n$. However, if this ratio is very different in each coordinate block, SBDA-u is able to obtain a much tighter bound. To see this point, consider the sequences $\{M_i\}$ and $\{D_i\}$ such that k items in $\{M_i\}$ are $O(\tilde{M})$ for some integer k, $0<k\ll n$, while the rest are $o\left(1/n\right)$ and D_i is uniformly bounded by $\tilde{D}$, $1\leq i\leq n$. Then the constant in SBDA-u is $O(\sqrt{n}k\tilde{M}\sqrt{\tilde{D}})$ while the one in SBMD is $O(n\sqrt{k}\tilde{M}\sqrt{\tilde{D}})$, which is $\sqrt{n/k}$ times larger.

Convergence rate when $\omega(x)$ is strongly convex

In this section, we investigate the convergence of SBDA-u when $\omega(x)$ is strongly convex with modulus λ, $\lambda > 0$. More specifically, we consider two averaging schemes and stepsize selections. In the first approach, we apply a simple averaging scheme similar to [36]. By setting $\alpha_t \equiv 1$, all the past stochastic block subgradients are weighted equally. In the second approach we apply a more aggressive weighting scheme, which puts more weights on the later iterates.

To prove the convergence of SBDA-u when $\omega(x)$ is strongly convex, we introduce in the following lemma, a useful "coupling" property for Bernoulli random variables:

Lemma 7. *Let r_1, r_2, r_3 be i.i.d. samples from Bernoulli(p), $0 < p < 1$, $a, b > 0$, and any x, such that $0 \leq x \leq a$, then*

$$\mathbf{E}\left[\frac{1}{r_1 x + r_2(a-x) + b}\right] \leq \mathbf{E}\left[\frac{1}{r_3 a + b}\right]. \tag{21}$$

In the next corollary, we derive these specific convergence rates for strongly convex problems.

Corollary 8. *In algorithm 1: if $\omega(x)$ is λ-strongly convex with modulus $\lambda > 0$, then*

1. if $\alpha_t \equiv 1$, $\gamma_t^{(i)} = \lambda/\rho$, for $t = 0, 1, 2, ..., T-1$, and $\bar{x} = \sum_{t=1}^{T} x_t/T$, then

$$\begin{aligned}\mathbf{E}\left[\phi(\bar{x}) - \phi(x^*)\right] \quad \leq \quad & \frac{(n-1)\left[\phi(x_0) - \phi(x^*)\right] + n\lambda/\rho \sum_{i=1}^{n} d_i(x^*)}{T} \\ & + \frac{5n\left(\sum_{i=1}^{n} M_i^2\right)\log(T+1)}{\lambda T}.\end{aligned}$$

2. if $\alpha_t = n + t$, for $t = 0, 1, 2, \ldots$, and $\alpha_{-1} = 0$, $\gamma_t^{(i)} = \lambda(2n+T)/\rho$, for $t = 0, 1, 2, ..., T-1$, then

$$\begin{aligned}\mathbf{E}\left[\phi(\bar{x}) - \phi(x^*)\right] \quad \leq \quad & \frac{2n(n-1)\left[\phi(x_0) - \phi(x^*)\right] + 2n(2n+T)\lambda/\rho \sum_{i}^{n} d_i(x^*)}{T(T+1)} \\ & + \frac{10n\left(\sum_{i=1}^{n} M_i^2\right)}{\lambda(T+1)}\left[1 + \frac{n + (n+1)\log T}{T}\right].\end{aligned}$$

Proof. In part 1), let $\alpha_t \equiv 1$, $\gamma_t^{(i)} \equiv \lambda/\rho$, it can be observed that $l_{t-1}^{(i)} \sim \text{Binomial}\left(t, \frac{1}{n}, \frac{n-1}{n}\right)$ $t \geq 0$, we have

$$\begin{aligned}\mathbf{E}\left[\frac{1}{l_{t-1}^{(i)}\lambda + \gamma_{t-1}^{(i)}\rho}\right] &= \sum_{i=0}^{t}\binom{t}{i}\left(\frac{1}{n}\right)^i\left(\frac{n-1}{n}\right)^{t-i}\frac{1}{\lambda(i+1)} \\ &= \frac{n}{\lambda(t+1)}\sum_{i=0}^{t}\binom{t+1}{i+1}\left(\frac{1}{n}\right)^{i+1}\left(\frac{n-1}{n}\right)^{t-i} \\ &= \frac{n}{\lambda(t+1)}\left[1 - \left(\frac{n-1}{n}\right)^{t+1}\right] \\ &\leq \frac{n}{\lambda(t+1)}.\end{aligned}$$

Observing the fact that $\sum_{\tau=0}^{t}\frac{1}{\tau+1} \leq \int_1^{t+2}\frac{1}{x}dx \leq \log(t+2)$, we obtain

$$\mathbf{E}\left[\phi(\bar{x}) - \phi(x^*)\right] \quad \leq \quad \frac{(n-1)\left[\phi(x_0) - \phi(x^*)\right] + \lambda n/\rho \sum_{i=1}^{n} d_i(x^*)}{T} + \frac{5n\left(\sum_{i=1}^{n} M_i^2\right)\log(T+1)}{\lambda T}.$$

In part 2), let $\alpha_t = n + t$, for $t = 0, 1, 2, \ldots$, and $\alpha_{-1} = 0$, $\gamma_t^{(i)} = \lambda (2n + T) / \rho$, then for $t \geq 0$, for any fixed i, let $r_s = \mathbf{1}_{i_s = i}$, hence $r_s \sim \text{Bernoulli}(p)$. In addition, we assume a sequence of ghost i.i.d. samples $\{r'_s\}_{0 \leq s \leq T}$. For $t > 0$,

$$\begin{aligned}
\mathbf{E}\left[\frac{1}{l_t^{(i)}\lambda + \gamma_t^{(i)}\rho}\right] &= \mathbf{E}\left[\frac{1}{\lambda \sum_{s=0}^{t} r_s (n+s) + \gamma_t^{(i)}\rho}\right] \\
&\leq \mathbf{E}\left[\frac{1}{\lambda \sum_{s=0}^{\lceil t/2 \rceil - 1} \{r_s (n+s) + r_{t-s}(n+t-s)\} + \lambda(2n+T)}\right] \\
&\leq \mathbf{E}\left[\frac{1}{\lambda (2n+t)\left(\sum_{s=0}^{\lceil t/2 \rceil - 1} r'_s + 1\right)}\right] \\
&\leq \frac{n}{\lambda (2n+t)(\max\{\lceil t/2 \rceil, 1\})} \qquad (22)
\end{aligned}$$

where the second inequality follows from the independence of $\{r_s\}$ and $\{r'_s\}$ and the coupling property in Lemma 7. It can be seen that the conclusion in (22) holds when $t = -1, 0$ as well. Hence

$$\begin{aligned}
\sum_{t=0}^{T-1} \mathbf{E}\left[\frac{\alpha_t^2}{\gamma_{t-1}^{(i)}\rho + l_{t-1}^{(i)}\lambda}\right] &\leq \sum_{t=0}^{T-1} \frac{(n+t)^2}{\lambda (2n+t-1)(\max\{\lceil (t-1)/2 \rceil, 1\})} \\
&\leq \sum_{t=0}^{T-1} \frac{(n+t)}{\lambda (\max\{\lceil (t-1)/2 \rceil, 1\})} \\
&= \frac{2n+1}{\lambda} + \sum_{t=2}^{T-1} \frac{(n+t)}{\lambda \lceil (t-1)/2 \rceil} \\
&\leq \frac{2n+1}{\lambda} + \sum_{t=2}^{T-1} \frac{2(n+t)}{\lambda (t-1)} \\
&= \frac{2n+2T-1}{\lambda} + \sum_{t=2}^{T-1} \frac{2(n+1)}{\lambda (t-1)} \\
&\leq \frac{2n+2T}{\lambda} + \frac{2(n+1)}{\lambda} \int_1^{T-1} \frac{1}{x} dx \\
&\leq \frac{2}{\lambda}(n + T + (n+1)\log T).
\end{aligned}$$

Let $\bar{x} = \frac{\sum_{t=1}^{T} t x_t}{\sum_{t=1}^{T} t}$ be the weighted average point, then

$$\begin{aligned}
\mathbf{E}[\phi(\bar{x}) - \phi(x^*)] &\leq \frac{2n(n-1)[\phi(x_0) - \phi(x^*)] + 2n(2n+T)\lambda/\rho \sum_i^n D_i}{T(T+1)} \\
&\quad + \frac{10n\left(\sum_{i=1}^n M_i^2\right)}{\lambda (T+1)}\left[1 + \frac{n + (n+1)\log T}{T}\right].
\end{aligned}$$

□

For nonsmooth and strongly convex objectives, we presented two options to select $\{\alpha_t\}$ and $\{\gamma_t\}$. These results seem to provide new insights on the dual averaging approach as well. To see this,

we consider SBDA-u when $n = 1$. In the first scheme, when $\alpha_t \equiv 1$, the convergence rate of $O(\log T/T)$ is similar to the one in [36]. In the second scheme of Corollary 8, it shows that regularized dual averaging methods can be easily improved to be optimal while being equipped with a more aggressive averaging scheme. Our observation suggests an alternative with rate $O(1/T)$ to the more complicated accelerated scheme ([6, 2]). Such results seems new to the world of simple averaging methods, and is on par with the recent discoveries for stochastic mirror descent methods ([20, 3, 8, 26, 12]).

4 Nonuniformly randomized SBDA (SBDA-r)

In this section we consider the general nonsmooth convex problem when $\omega(x) = 0$ or $\omega(x)$ is lumped into $f(\cdot)$:

$$\min_{x \in X} \phi(x) = f(x),$$

and show a variant of SBDA in which block coordinates are sampled non-uniformly. More specifically, we assume the block coordinates are i.i.d. sampled from a discrete distribution $\{p_i\}_{1 \le i \le n}$, $0 < p_i < 1$, $1 \le i \le n$. We describe in Algorithm 2 the nonuniformly randomized stochastic block dual averaging method (SBDA-r).

Input: convex function f, sequence of samples $\{\xi_t\}$, distribution $\{p_i\}_{1 \le i \le n}$;
initialize $\alpha_0 \in \mathbb{R}$, $\gamma_{-1} \in \mathbb{R}^n$,$\bar{G} = \mathbf{0}^N$ and $x_0 = \arg\min_{x \in X} \sum_{i=1}^n \frac{\gamma_{-1}^{(i)}}{p_i} d_i\left(x^{(i)}\right)$;
for $t = 0, 1, \ldots, T-1$ **do**
  sample a block $i_t \in \{1, 2, \ldots, n\}$ with probability $\text{Prob}(i_t = i) = p_i$;
  set $\gamma_t^{(i)}$, $i = 1, 2, ..., n$;
  receive sample ξ_t and update $\bar{G}$: $\bar{G} = \bar{G} + \frac{\alpha_t}{p_{i_t}} U_i G^{(i_t)}(x_t, \xi_t)$;
  update $x_{t+1}^{(i_t)} = \arg\min_{x \in X_{i_t}} \left\{ \left\langle \bar{G}^{(i_t)}, x \right\rangle + \frac{\gamma_t^{(i_t)}}{p_{i_t}} d_{i_t}(x) \right\}$;
  set $x_{t+1}^{(j)} = x_t^{(j)}$, $j \neq i_t$;
end
Output: $\bar{x} = \left(\sum_{t=0}^T \alpha_t x_t\right) / \left(\sum_{t=0}^T \alpha_t\right)$;

Algorithm 2: Nonuniformly randomized stochastic block dual averaging (SBDA-r) method

In the next theorem, we present the main convergence property of SBDA-r, which expresses the bound of the expected optimization error as a joint function of the sampling distribution $\{p_i\}$, and the sequences $\{\alpha_t\}$, $\{\gamma_t\}$.

Theorem 9. *In algorithm 2, let $\{x_t\}$ be the generated solutions and x^* be the optimal solution, $\{\alpha_t\}$ be a sequence of positive numbers, $\{\gamma_t\}$ be a sequence of vectors satisfying the assumption (9) . Let $\bar{x} = \frac{\sum_{t=0}^T \alpha_t x_t}{\sum_{t=0}^T \alpha_t}$ be the average point, then*

$$\mathbf{E}\left[f(\bar{x}) - f(x)\right] \le \frac{1}{\sum_{t=0}^T \alpha_t} \left\{ \sum_{t=0}^T \sum_{i=1}^n \mathbf{E}\left[\frac{\alpha_t^2 \|G_t\|_{(i),*}^2}{2\rho\gamma_{t-1}^{(i)}} \right] + \sum_{i=1}^n \frac{\mathbf{E}\left[\gamma_T^{(i)}\right]}{p_i} d_i(x) \right\}. \tag{23}$$

Proof. For the sake of simplicity, let us denote $A_t = \sum_{\tau=0}^t \alpha_\tau$, for $t = 0, 1, 2 \ldots$. Based on the convexity of f, we have $f\left(\frac{\sum_{t=0}^T \alpha_t x_t}{A_T}\right) \le \frac{\sum_{t=0}^T \alpha_t f(x_t)}{A_T}$ and $f(x_t) \le f(x) + \langle g_t, x_t - x \rangle$ for $x \in X$. Then

$$A_T\left[f\left(\bar{x}\right)-f\left(x\right)\right] \leq \sum_{t=0}^{T}\alpha_t\left\langle g_t, x_t - x\right\rangle$$
$$\leq \underbrace{\sum_{t=0}^{T}\frac{\alpha_t}{p_{i_t}}\left\langle U_{i_t}G_t^{(i_t)}, x_t - x\right\rangle}_{\Delta_1} + \underbrace{\sum_{t=0}^{T}\alpha_t\left\langle g_t - \frac{1}{p_{i_t}}U_{i_t}G_t^{(i_t)}, x_t - x\right\rangle}_{\Delta_2}. \tag{24}$$

It suffices to provide precise bounds on the expectation of Δ_1, Δ_2 separately.

We define the auxiliary function

$$\Psi_t\left(x\right) = \begin{cases} \sum_{s=0}^{t}\frac{\alpha_s}{p_{i_s}}\left\langle U_{i_s}G_s^{(i_s)}, x\right\rangle + \sum_{i=1}^{n}\frac{\gamma_t^{(i)}}{p_i}d_i\left(x^{(i)}\right) & t \geq 0 \\ \sum_{i=1}^{n}\frac{\gamma_t^{(i)}}{p_i}d_i\left(x^{(i)}\right) & t = -1 \end{cases}.$$

Thus

$$\begin{aligned} \Psi_t\left(x_{t+1}\right) &= \min_x \Psi_t\left(x\right) \\ &\geq \min_x\left\{\sum_{s=0}^{t}\frac{\alpha_s}{p_{i_s}}\left\langle U_{i_t}G_s^{(i_t)}, x\right\rangle + \sum_{i=1}^{n}\frac{\gamma_{t-1}^{(i)}}{p_i}d_i\left(x^{(i)}\right)\right\} \\ &= \min_x\left\{\frac{\alpha_t}{p_{i_t}}\left\langle U_{i_t}G_t^{(i_t)}, x\right\rangle + \Psi_{t-1}\left(x\right)\right\} \end{aligned} \tag{25}$$

The first inequality follows from the property (9). Next, using (25) and Lemma 2, we obtain

$$\frac{\alpha_t}{p_{i_t}}\left\langle U_{i_t}G_t^{(i_t)}, x_t\right\rangle \leq \Psi_t\left(x_{t+1}\right) - \Psi_{t-1}\left(x_t\right) + \frac{\alpha_t^2}{2\rho p_{i_t}\gamma_{t-1}^{(i_t)}}\|G_t\|_{(i_t),*}^2.$$

Summing up the above inequality for $t = 0, \ldots, T$, we have

$$\sum_{t=0}^{T}\frac{\alpha_t}{p_{i_t}}\left\langle U_{i_t}G_t^{(i_t)}, x_t\right\rangle \leq \Psi_T\left(x_{T+1}\right) - \Psi_{-1}\left(x_0\right) + \sum_{t=0}^{T}\frac{\alpha_t^2}{2\rho p_{i_t}\gamma_{t-1}^{(i_t)}}\|G_t\|_{(i_t),*}^2. \tag{26}$$

Moreover, by the optimality of x_{T+1} in solving $\min_x \Psi_T\left(x\right)$, for all $x \in X$, we have

$$\Psi_T\left(x_{T+1}\right) \leq \sum_{t=0}^{T}\frac{\alpha_t}{p_{i_t}}\left\langle U_{i_t}G_t^{(i_t)}, x\right\rangle + \sum_{i=1}^{n}\frac{\gamma_T^{(i)}}{p_i}d_i\left(x\right). \tag{27}$$

Putting (26) and (27) together, and using the fact that $\Psi_{-1}\left(x_0\right) \geq 0$, we obtain:

$$\Delta_1 \leq \sum_{i=1}^{n}\frac{\gamma_T^{(i)}}{p_i}d_i\left(x\right) + \sum_{t=0}^{T}\frac{\alpha_t^2}{2\rho p_{i_t}\gamma_{t-1}^{(i_t)}}\|G_t\|_{(i_t),*}^2.$$

For each t, taking expectation w.r.t. i_t, we have

$$\begin{aligned}\mathbf{E}\left[\frac{\alpha_t^2}{2\rho p_{i_t}\gamma_{t-1}^{(i_t)}}\|G_t\|_{(i_t),*}^2\right] &= \mathbf{E}\left[\mathbf{E}_{i_t}\left[\frac{\alpha_t^2}{2\rho p_{i_t}\gamma_{t-1}^{(i_t)}}\|G_t\|_{(i_t),*}^2\right]\right] \\ &= \sum_{i=1}^n \mathbf{E}\left[\frac{\alpha_t^2}{2\rho\gamma_{t-1}^{(i)}}\|G_t\|_{(i),*}^2\right].\end{aligned}$$

As a consequence, one has

$$\mathbf{E}\left[\Delta_1\right] \leq \sum_{i=1}^n \frac{\mathbf{E}\left[\gamma_T^{(i)}\right]}{p_i} d_i\left(x\right) + \sum_{t=0}^T \sum_{i=1}^n \mathbf{E}\left[\frac{\alpha_t^2\|G_t\|_{(i),*}^2}{2\rho\gamma_{t-1}^{(i)}}\right]. \tag{28}$$

In addition, taking the expectation with respect to i_t, ξ_t and noting that $\mathbf{E}_{\xi_t,i_t}\left[\frac{1}{p_{i_t}}U_{i_t}G_t\right] - g_t = \mathbf{E}_{\xi_t}\left[G_t\right] - g_t = 0$, we obtain

$$\mathbf{E}\left[\Delta_2\right] = 0. \tag{29}$$

In view of (28) and (29), we obtain the bound on the expected optimization error:

$$\mathbf{E}\left[f\left(\bar{x}\right) - f\left(x\right)\right] \leq \frac{1}{\sum_{t=0}^T \alpha_t}\left\{\sum_{t=0}^T\sum_{i=1}^n \mathbf{E}\left[\frac{\alpha_t^2\|G_t\|_{(i),*}^2}{2\rho\gamma_{t-1}^{(i)}}\right] + \sum_{i=1}^n \frac{\mathbf{E}\left[\gamma_T^{(i)}\right]}{p_i} d_i\left(x\right)\right\}.$$

□

Block Coordinates Sampling and Analysis

In view of Theorem 4, the obtained upper bound can be conceived as a joint function of probability mass $\{p_i\}$, and the control sequences $\{\alpha_t\}$, $\{\gamma_t\}$. Firstly, throughout this section, let $x = x^*$ and assume

$$\alpha_t = 1, \quad t = 0, 1, 2, \ldots. \tag{30}$$

Naturally, we can choose the distribution and stepsizes by optimizing the bound

$$\min_{\{\gamma_t\},p} \mathcal{L}(\{\gamma_t\}, p) = \sum_{t=0}^T\sum_{i=1}^n \mathbf{E}\left[\frac{M_i^2}{2\rho\gamma_{t-1}^{(i)}}\right] + \sum_i^n \frac{\mathbf{E}[\gamma_T^{(i)}]}{p_i} D_i. \tag{31}$$

This is a joint problem on two groups of variables. Let us first discuss how to choose $\{\gamma_t\}$ for any fixed p_i. Let us assume p_i has the form: $p_i = \frac{M_i^a D_i^b}{C_{a,b}}, \quad i = 1, 2, \ldots, n$, where $a, b \geq 0$, and define $C_{a,b} = \sum_{i=1}^n M_i^a D_i^b$. We derive two stepsizes rules, depending on whether the iteration number T is known or not. We assume $\gamma_t^{(i)} = \beta_i$, for some constant β_i, $i = 1, 2, \ldots n$, $t = 1, 2, ..., T$. The equivalent problem with p, β, has the form

$$\min_{p,\beta} \mathcal{L}(p, \beta) = \sum_{i=1}^n \frac{(T+1)M_i^2}{2\rho\beta_i} + \sum_i^n \frac{\beta_i}{p_i} D_i. \tag{32}$$

By optimizing w.r.t. β, we obtain the optimal solutions

$$\gamma_t^{(i)} = \beta_i = \sqrt{\frac{(1+T)p_i M_i^2}{2\rho D_i}}. \tag{33}$$

In addition, we can also select stepsizes without assuming the iteration number T. Let us denote

$$\gamma_t^{(i)} = \begin{cases} \sqrt{t+1}u_i & \text{if } i = i_t, \\ \gamma_{t-1}^{(i)} & \text{otherwise,} \end{cases} \tag{34}$$

for some unspecified u_i, $1 \le i \le n$. Applying Lemma 5 with $a_t = E\left[\frac{1}{\gamma_{t-1}^{(i)}}\right]$, $b_t = \frac{1}{u_i\sqrt{t}}$, we have

$$\sum_{t=0}^{T} \mathbf{E}\left[\frac{1}{\gamma_{t-1}^{(i)}}\right] \le \sum_{t=1}^{T} \frac{1}{u_i\sqrt{t}} + \frac{1}{\gamma_{-1}^{(i)}p_i} \le 2\frac{\sqrt{T+1}}{u_i} + \frac{1}{\gamma_{-1}^{(i)}p_i}.$$

In view of the above analysis, we can relax the problem to the following:

$$\min_{p,u} \sum_{i=1}^{n} \left[\frac{M_i^2\sqrt{T+1}}{\rho u_i} + \frac{u_i\sqrt{T+1}}{p_i}D_i + \frac{M_i^2}{2\rho\gamma_{-1}^{(i)}p_i}\right].$$

Note that the third term above is $o\left(\sqrt{T}\right)$ and hence can be ignored for the sake of simplicity. Thus we have the approximate problem

$$\min_{p,u} \sum_{i=1}^{n} \left[\frac{M_i^2\sqrt{T+1}}{\rho u_i} + \frac{u_i\sqrt{T+1}}{p_i}D_i\right], \tag{35}$$

we apply the similar analysis and obtain $u_i = \sqrt{\frac{p_i M_i^2}{\rho D_i}}$ and hence the second stepsize rule

$$\gamma_t^{(i)} = \begin{cases} \sqrt{\frac{(t+1)p_i M_i^2}{\rho D_i}} & \text{if } i = i_t \\ \gamma_{t-1}^{(i)} & \text{otherwise} \end{cases}, \quad t \ge 0. \tag{36}$$

We have established the relation between the optimized sampling probability and stepsizes. Now we are ready to discuss specific choices of the probability distribution. Firstly, the simplest way is to set

$$p_i = \frac{1}{n}, \quad i = 1, 2, \ldots, n, \tag{37}$$

which implies that SBDA-r reduces to the uniform sampling method SBDA-u with the obtained stepsizes entirely similar to the ones we derived earlier. However, from the above analysis, it is possible to choose the sampling distribution properly and obtain a further improved convergence rate. Next we show how to obtain the optimal sampling and stepsize policies from solving the joint problem (31). We first describe an important property in the following lemma.

Lemma 10. *Let $\mathcal{S}^n$ be the n-dimensional simplex. The optimal solution x^*, y^* of the nonlinear problem $\min_{x\in\mathbb{R}^n_{++}, y\in\mathcal{S}^n} \sum_{i=1}^n \left[\frac{a_i}{x_i} + \frac{x_i}{b_i y_i}\right]$ where $a_i, b_i > 0$, $1 \le i \le n$ is*

$$y_i^* = (a_i/b_i)^{\frac{1}{3}}W, \text{ and } x_i^* = a_i^{\frac{2}{3}}b_i^{\frac{1}{3}}\sqrt{W},$$

where $i = 1, 2, \ldots n$ and $W = \left(\sum_i^n (a_i/b_i)^{\frac{1}{3}}\right)^{-1}$.

Applying Lemma 10 to the problem (32) , we obtain the optimal sampling probability

$$p_i = M_i^{\frac{2}{3}}D_i^{\frac{1}{3}}/C, \quad i = 1, 2, \ldots n \tag{38}$$

where C is the normalizing constant. This is also the optimal probability in problem (35). In view of these results, we obtain the specific convergence rates in the following corollary:

Corollary 11. *In algorithm 2, let* $\alpha_t = 1$, $t \geq 0$. *Denote* $C = \left(\sum_{j=1}^{n} M_j^{2/3} D_j^{1/3}\right)$, *with block coordinates sampled from distribution (38). Then:*

1. if $\gamma_t^{(i)} = \sqrt{\frac{(1+T)}{2\rho C}} M_i^{4/3} D_i^{-1/3}$, $t \geq -1$, $i = 1, 2, \ldots, n$, then

$$\mathbf{E}\left[f\left(\bar{x}\right) - f\left(x^*\right)\right] \leq \frac{\sqrt{2}}{\sqrt{\rho}} \frac{C^{3/2}}{\sqrt{T+1}}. \tag{39}$$

2. if $\gamma_{-1}^{(i)} = \sqrt{\frac{1}{\rho C}} M_i^{4/3} D_i^{-1/3}$and $\gamma_t^{(i)} = \begin{cases} \sqrt{\frac{(t+1)}{\rho C}} M_i^{4/3} D_i^{-1/3} & \text{if } i = i_t, \\ \gamma_{t-1}^{(i)} & \text{o.w.} \end{cases}$, $t \geq 0$, $i = 1, 2, \ldots, n$,

   then

$$\mathbf{E}\left[f\left(\bar{x}\right) - f\left(x^*\right)\right] \leq \frac{C^{3/2}}{\sqrt{\rho}} \left[\frac{2}{\sqrt{T+1}} + \frac{1}{2\left(T+1\right)}\right]. \tag{40}$$

Proof. It remains to plug the value of $\{\gamma_t\}$, p back into $\mathcal{L}\left(,\right)$. □

It is interesting to compare the convergence properties of SBDA-r with that of SBDA-u and SBMD. SBDA with uniform sampling of block coordinates only yields suboptimal dependence on the multiplicative constants. Nevertheless, the rate can be further improved by employing optimal nonuniform sampling. To develop further intuition, we relate the two rates of convergence with the help of HÃ¶lder's inequality:

$$\left[\sum_{i=1}^{n}\left(M_i^{2/3} D_i^{1/3}\right)\right]^{3/2} \leq \left\{\left[\sum_{i=1}^{n}\left(M_i^{2/3} D_i^{1/3}\right)^{3/2}\right]^{2/3} \cdot \left[\sum_{i=1}^{n} 1^3\right]^{1/3}\right\}^{3/2} = \sum_{i=1}^{n}\left(M_i\sqrt{D_i}\right) \cdot \sqrt{n}.$$

The inequality is tight if and only if for some constant $c > 0$ and i, $1 \leq i \leq n$: $M_i\sqrt{D_i} = c$. In addition, we compare SBDA-r with a nonuniform version of SBMD[1], which obtains $O\left(\frac{\sqrt{\sum_{i=1}^{n} M_i^2} \cdot \sum_{i=1}^{n} \sqrt{D_i}}{\sqrt{T}}\right)$, assuming blocks are sampled based on the distribution $p_i \propto \sqrt{D_i}$. Again, applying HÃ¶lder's inequality, we have

$$\left[\sum_{i=1}^{n}\left(M_i^{2/3} D_i^{1/3}\right)\right]^{3/2} \leq \left\{\left[\sum_{i=1}^{n}\left(M_i^{2/3}\right)^3\right]^{1/3} \cdot \left[\sum_{i=1}^{n}\left(D_i^{1/3}\right)^{3/2}\right]^{2/3}\right\}^{3/2} = \sqrt{\sum_{i=1}^{n} M_i^2} \cdot \sum_{i=1}^{n}\sqrt{D_i}.$$

In conclusion, SBDA-r, equipped with an optimized block sampling scheme, obtains the *best* iteration complexity among all the block subgradient methods.

5 Experiments

In this section, we examine the theoretical advantages of SBDA through several preliminary experiments. For all the algorithms compared, we estimate the parameters and tune the best stepsizes using separate validation data. We first investigate the performance of SBDA on nonsmooth deterministic problems by comparing its performance against other nonsmooth algorithms. We compare with the following algorithms: SM1 and SM2 are subgradient mirror decent methods with stepsizes $\gamma_1 \propto \frac{1}{\sqrt{t}}$ and $\gamma_2 \propto \frac{1}{\|g(x)\|}$ respectively. Finally, SGD is stochastic mirror descent and SDA a

[1]See Corollary 2.2, part a) of [3]

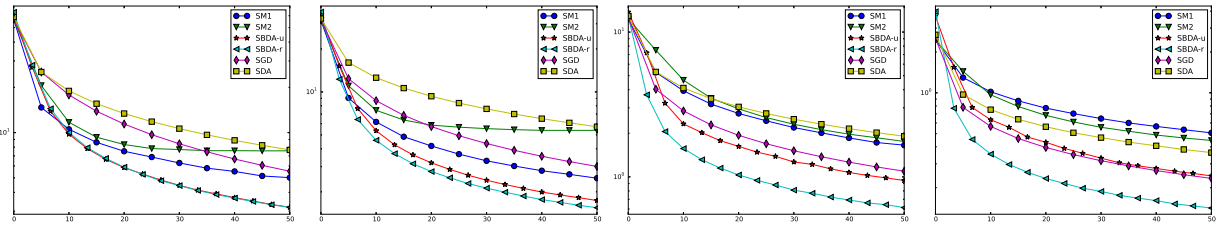


Figure 1: Tests on ℓ_1 regression.

stochastic subgradient dual averaging method. We study the problem of robust linear regression (ℓ_1 regression) with the objective $\phi(x) = \frac{1}{m}\sum_{i=1}^{m}\left|b_i - a_i^T x\right|$. The optimal solution x^* and each a_i are generated from $\mathcal{N}(0, I_{n\times n})$. In addition, we define a scaling vector $s \in \mathbb{R}^n$ and S a diagonal matrix s.t. $S_{ii} = s_i$. We let $b = (AS)x^* + \sigma$, where $A = [a_2, a_2, \ldots, a_m]^T \in \mathbb{R}^{m\times n}$, and the noise $\sigma \sim \mathcal{N}(0, \rho I)$. We set $\rho = 0.01$ and $m = n = 5000$.

We plot the optimization objective with the number of passes of the dataset in Figure 1, for four different choices of s. In the first test case (leftmost subfigure), we let $s = [1, 1, \ldots, 1]^T$ so that columns of A correspond to uniform scaling. We find that SBDA-u and SBDA-r have slightly better performance than the other algorithms while exhibiting very similar performance. In the next three cases, s is generated from the distribution $p(x; a) = a(1-x)^{a-1}$, $0 \le x \le 1$, $a > 0$. We set $a = 1, 5, 30$ respectively. Employing a large a ensures that the bounds on the norms of block subgradients follow the power law. We observe that stochastic methods outperform the deterministic methods, and SBDA-based algorithms have comparable and often better performance than SGD algorithms. In particular, SBDA-r exhibits the best performance, which clearly shows the advantage of SBDA with the nonuniform sampling scheme.

Next, we examine the performance of SBDA for online learning and stochastic approximation. We conduct simulated experiments on the problem: $\phi(x) = \mathbf{E}_{a,b}\left[(b - \langle La, x\rangle)^2\right]$, where the aim is to fit linear regression under a linear transform L. The transform matrix $L \in M^{n\times n}$ is generated as follows: we first sample a matrix $\tilde{L}$ for which each entry $\tilde{L}_{i,j} \sim \mathcal{N}(0,1)$. L is obtained from $\tilde{L}$ with 90% of the rows being randomly rescaled by a factor ρ. To obtain the optimal solution x^*, we first generate a random vector from the distribution $\mathcal{N}(0, I_{n\times n})$ and then truncate each coordinate in $[-1, 1]$. Simulated samples are generated according to $b = \langle La, x^*\rangle + \varepsilon$ where $\varepsilon \in \mathcal{N}(0, 0.01 I_{n\times n})$. We let $n = 200$, and generate 3000 independent samples for training and 10000 independent samples for testing.

To compare the performances of these algorithms under various conditions, we tune the parameter ρ in $[1, 0.1, 0.05, 0.01]$. As can be seen from above, ρ affects the estimation of block-wise parameters $\{M_i\}$. In Figure 2, we show the objective function for the average of 20 runs. The experimental results show the advantages of SBDA over SBMD. When $\rho = 1$, SBDA-u, SBDA-r, and SBMD have the same theoretical convergence rate, and exhibit similar performance. However, as ρ decreases, the "importance" of 90% of the blocks is diminishing and we find SBDA-u and SBDA-r both outperform SBMD. Moreover, SBDA-r seems to perform the best, suggesting the advantage of our proposed stepsize and sampling schemes which are adaptive to the block structures. These observations lends empirical support to our theoretical analysis.

Our next experiment considers online ℓ_1 regularized linear regression (Lasso):

$$\min_{w\in\mathbb{R}^n} \frac{1}{2}\mathbf{E}_{(y,x)}\left[\left(y - w^T x\right)^2\right] + \lambda\|w\|_1 \tag{41}$$

While linear regression has been well studied in the literature, recent work is interested in efficient regression algorithms under different adversarial circumstances [1, 9, 10]. Under the assumptions

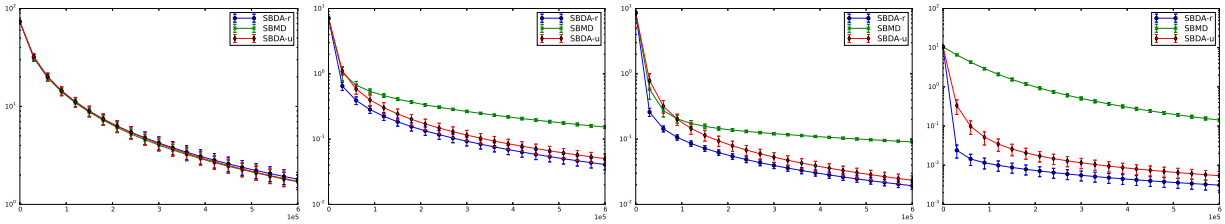
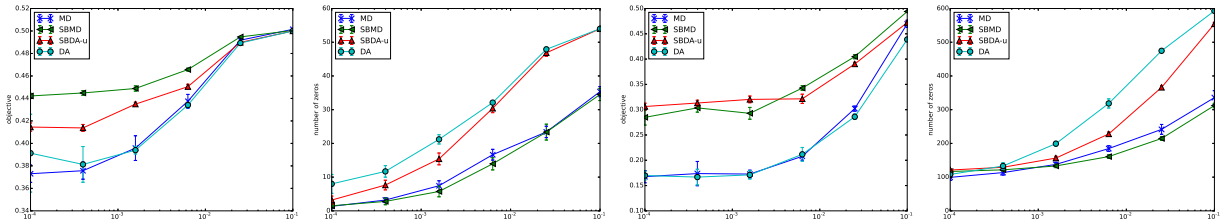


Figure 2: Tests on linear regression, Left to right: $\rho = 1, 0.1, 0.05, 0.01$.



(a) Test on covtype dataset.

(b) Test on mnist dataset.

Figure 3: Tests on online lasso with limited budgets

of limit budgets, the learner only partially observes the features for each incoming instance, but is allowed to choose the sampling distribution of the features. In addition, we explicitly enforce the ℓ_1 penalty, expecting to learn a sparse solution that effectively reduces testing cost. To apply stochastic methods, we estimate the stochastic coordinate gradient of the least squares loss. For the sake of simplicity, we assume for each input sample instance (y, x), two features (i_t, j_t) are revealed. When we sample one coordinate j_t from some distribution $\{p_j\}$, then $\frac{1}{p_{j_t}} w^{(j_t)} x^{(j_t)}$ is an unbiased estimator of $w^T x$. Hence the defined value $G^{(i_t)} = \frac{1}{p_{j_t}} x^{(i_t)} x^{(j_t)} w^{(j_t)} - y x^{(i_t)}$ is an unbiased estimator of the i_t-th coordinate gradient.

We adapt both SBMD and SBDA-u to these problems and conduct the experiments on datasets covtype and mnist (digit "3 vs 5"). We also implement MD (composite mirror descent) and DA (regularized dual averaging method). For all the methods, the training uses the same total number of features. However, SBMD and SBDA-u obtain features sampled using a uniform distribution; both MD and DA have "unfair" access to observe full feature vectors and therefore have the advantages of lower variance. We plot in Figures 3a and 3b, the optimization error and sparsity patterns with respect to the penalty weights λ on the two datasets. It can be seen that SBDA-u has comparable and often better optimization accuracy than SBMD. In addition, we also plot the sparsity patterns for different values of λ. It can be seen that SBDA-u is very effective in enhancing sparsity, more efficient than SBMD, MD, and comparable to DA which doesn't have such budget constraints.

6 Discussion

In this paper we introduced SBDA, a new family of block subgradient methods for nonsmooth and stochastic optimization, based on a novel extension of dual averaging methods. We specialized SBDA-u for regularized problems with nonsmooth or strongly convex regularizers, and SBDA-r for general nonsmooth problems. We proposed novel randomized stepsizes and optimal sampling schemes which are truly block adaptive, and thereby obtain a set of sharper bounds. Experiments demonstrate the advantage of SBDA methods compared with subgradient methods on nonsmooth

deterministic and stochastic optimization. In the future, we will extend SBDA to an important class of regularized learning problems consisting of the finite sum of differentiable losses. On such problems, recent work [31, 32] shows efficient BCD convergence at linear rate. The works in [39, 35] propose randomized BCD methods that sample both primal and dual variables. However both methods apply conservative stepsizes which take the maximum of the block Lipschitz constant. It would be interesting to see whether our techniques of block-wise stepsizes and nonuniform sampling can be applied in this setting as well to obtain improved performance.

7 Appendix

Proof of Lemma 1

Proof. The first part comes from [34]. Let $g(z)$ denote any subgradient of f at z. Since $f(x)$ is strongly convex, we have $f(x) \geq f(z) + \langle U_i g^{(i)}(z), x - z\rangle + \frac{\lambda}{2}\|x - z\|_{(i)}^2$. By the definition of z and optimality condition, we have $g^{(i)}(z) = -\nabla_i d(z)$. Thus

$$f(x) + \langle \nabla_i d(z), x - z\rangle \geq f(z) + \frac{\lambda}{2}\|x - z\|_{(i)}^2.$$

It remains to apply the definition $x = z + U_i y$ and $\mathcal{V}(z, x) = d(x) - d(z) - \langle \nabla d(z), x - z\rangle$. □

Proof of Lemma 2

Proof. Let $h(y) = \max_{x \in X}\{\langle y, x\rangle - \Psi(x)\}$, since $\Psi(\cdot)$ is strongly convex and separable, $h(\cdot)$ is convex and differentiable and its i-th block gradient $\nabla_i h(\cdot)$ is $\frac{1}{\rho_i}$-smooth . Moreover, we have $\nabla h(0) = x_0$ by the definition of x_0. Thus

$$h\left(-U_i g^{(i)}\right) \leq h(0) + \left\langle x_0, -U_i g^{(i)}\right\rangle + \frac{1}{2\rho_i}\|g\|_{(i),*}^2.$$

It remains to plug in the definition of $h(\cdot)$, z, x_0. □

Proof of Lemma 3

Conjecture. *By convexity of $f(\cdot)$, we have $f(z) \leq f(x) + \langle g(z), z - x\rangle$. In addition,*

$$\begin{aligned}
\langle g(z), z - x\rangle &= \langle g(x), z - x\rangle + \langle g(z) - g(x), z - x\rangle \\
&= \langle g^{(i)}(x), y\rangle_{(i)} + \langle g^{(i)}(z) - g^{(i)}(x), y\rangle_{(i)} \\
&\leq \langle g^{(i)}(x), y\rangle_{(i)} + \|g^{(i)}(z) - g^{(i)}(x)\|_{(i),*} \cdot \|y\|_{(i)}.
\end{aligned}$$

The second equation follows from the relation between x, y, z and the last one from the Cauchy-Schwarz inequality. Finally the conclusion directly follows from (5).

Proof of Lemma 5

Proof. Let $A_t = \sum_{s=0}^t a_t$, $B_t = \sum_{s=1}^t b_t$. It is equivalent to show $A_t \leq B_t + \frac{A_0}{p}$. Then

$$
\begin{aligned}
A_t &= pB_t + A_0 + (1-p)\, A_{t-1} \\
&= [p + (1-p)]\, [pB_{t-1} + A_0] + (1-p)^2\, A_{t-2} \\
&= \left[p + (1-p) + (1-p)^2\right] [pB_{t-1} + A_0] + (1-p)^3\, A_{t-3} \\
&= ... \\
&\leq [pB_t + A_0] \left[\sum_{s=0}^{t} (1-p)^s\right].
\end{aligned}
$$

The last inequality follows from the assumption that $B_t \geq B_s$ where $0 \leq s \leq t$ and $A_0 = a_0$. It remains to apply the inequality $\sum_{s=0}^{t} (1-p)^s \leq \sum_{s=0}^{\infty} (1-p)^s = \frac{1}{p}$. □

Proof of Lemma 7

Proof. If $r_1, r_2, r_3 \sim \text{Bernoulli}\,(p)$, $c > 0$, $0 < p < 1$,

$$
\begin{aligned}
\mathbf{E}\left[\frac{1}{r_1 x + r_2\,(a-x) + b}\right] &= \frac{(1-p)^2}{b} + \frac{p\,(1-p)}{a-x+b} + \frac{p\,(1-p)}{x+b} + \frac{p^2}{a+b} \\
&\leq \frac{(1-p)^2}{b} + \frac{p\,(1-p)}{a+b} + \frac{p\,(1-p)}{b} + \frac{p^2}{a+b} \\
&= \frac{1-p}{b} + \frac{p}{a+b} \\
&= \mathbf{E}\left[\frac{1}{r_3 a + b}\right].
\end{aligned}
$$

To see the first inequality, let $f(x) = \frac{A}{x+c} + \frac{B}{a-x+c}$, where $A, B > 0$, it can be seen that $f(\cdot)$ is convex in $[0, a]$, then $\max_{x\in[0,a]} f(x) = \max\{f(0), f(a)\}$. □

Proof of Lemma 10

Proof. Let x^*, y^* be the optimal solution of $\min_{x,y} \mathcal{L}\,(x, y, a, b)$. We consider two subproblems. Firstly, $x^* = \arg\min_x \mathcal{L}\,(x, y^*, a, b)$. Since $\frac{a_i}{x_i} + \frac{x_i}{b_i y_i^*} \geq 2\sqrt{\frac{a_i}{b_i y_i^*}}$, at optimality

$$
\frac{a_i}{x_i^*} = \frac{x_i^*}{b_i y_i^*}. \tag{42}
$$

On the other hand, y^* is the minimizer of the problem $\min_y \mathcal{L}\,(x^*, y, a, b)$. Applying the Cauchy-Schwarz inequality to $\mathcal{L}\,(x^*, y, a, b)$, we obtain

$$
\sum_{i=1}^{n} \frac{x_i^*}{b_i y_i} \cdot 1 = \sum_{i=1}^{n} \frac{x_i^*}{b_i y_i} \sum_{1}^{n} y_i \geq \sum_{i}^{n} \sqrt{\frac{x_i^*}{b_i y_i}} \sqrt{y_i} = \sum_{i}^{n} \sqrt{x_i^*}.
$$

At optimality, the equality holds for some scalar $C > 0$,

$$
\frac{x_i^*}{b_i y_i^*} = C y_i^*, \quad i = 1, 2, \ldots, n. \tag{43}
$$

It remains to solve the equations (42) and (43) with the simplex constraint on y. □

References

[1] N. Cesa-Bianchi, S. Shalev-Shwartz, and O. Shamir. Efficient learning with partially observed attributes. *The Journal of Machine Learning Research (JMLR)*, 12:2857–2878, 2011.

[2] X. Chen, Q. Lin, and J. Pena. Optimal regularized dual averaging methods for stochastic optimization. In *Advances in Neural Information Processing Systems (NIPS) 25*, 2012.

[3] C. D. Dang and G. Lan. Stochastic block mirror descent methods for nonsmooth and stochastic optimization. *SIAM Journal on Optimization*, 25(2):856–881, 2015.

[4] J. Duchi and Y. Singer. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research (JMLR)*, 10:2899–2934, 2009.

[5] J. C. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. In *The 23rd Conference on Learning Theory (COLT)*, 2010.

[6] 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(4):1469–1492, 2012.

[7] L. A. Hageman and D. M. Young. *Applied iterative methods*. Courier Corporation, 2012.

[8] E. Hazan and S. Kale. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *The Journal of Machine Learning Research (JMLR)*, 15(1):2489–2512, 2014.

[9] E. Hazan and T. Koren. Linear regression with limited observation. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, 2012.

[10] D. Kukliansky and O. Shamir. Attribute efficient linear regression with data-dependent sampling. *arXiv preprint arXiv:1410.6382*, 2014.

[11] S. Lacoste-Julien, M. Jaggi, M. Schmidt, and P. Pletscher. Block-coordinate Frank-Wolfe optimization for structural SVMs. In *Proceedings of The 30th International Conference on Machine Learning (ICML)*, 2013.

[12] S. Lacoste-Julien, M. Schmidt, and F. Bach. A simpler approach to obtaining an $O(1/t)$ convergence rate for the projected stochastic subgradient method. *arXiv preprint arXiv:1212.2002*, 2012.

[13] G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1-2):365–397, 2012.

[14] G. Lan, Z. Lu, and R. D. Monteiro. Primal-dual first-order methods with {\ mathcal {O}(1/\ epsilon)} iteration-complexity for cone programming. *Mathematical Programming*, 126(1):1–29, 2011.

[15] J. Langford, L. Li, and T. Zhang. Sparse online learning via truncated gradient. *Journal of Machine Learning Research (JMLR)*, 10:719–743, 2009.

[16] Y. T. Lee and A. Sidford. Efficient accelerated coordinate descent methods and faster algorithms for solving linear systems. In *Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science*, pages 147–156, 2013.

[17] Z. Lu and L. Xiao. On the complexity analysis of randomized block-coordinate descent methods. *Mathematical Programming*, pages 1–28, 2014.

[18] 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(5):1037–1060, 1991.

[19] Z. Q. Luo and P. Tseng. On the convergence of the coordinate descent method for convex differentiable minimization. *Journal of Optimization Theory and Applications*, 72(1):7–35, Jan. 1992.

[20] A. Nedic and S. Lee. On stochastic subgradient mirror-descent algorithm with weighted averaging. *SIAM Journal on Optimization*, 24(1):84–107, 2014.

[21] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, Jan. 2009.

[22] Y. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):221–259, 2009.

[23] Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

[24] Y. Nesterov. Subgradient methods for huge-scale optimization problems. *Mathematical Programming*, 146(1-2):275–297, 2014.

[25] Z. Qu and P. Richtárik. Coordinate descent with arbitrary sampling, I: Algorithms and complexity. *arXiv preprint arXiv:1412.8060*, 2014.

[26] A. Rakhlin, O. Shamir, and K. Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 449–456, 2012.

[27] S. Reddi, A. Hefny, C. Downey, A. Dubey, and S. Sra. Large-scale randomized-coordinate descent methods with non-separable linear constraints. *arXiv preprint arXiv:1409.2617*, 2014.

[28] P. Richtárik and M. Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1-2):1–38, 2014.

[29] H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

[30] S. Shalev-Shwartz and A. Tewari. Stochastic methods for ℓ_1-regularized loss minimization. *The Journal of Machine Learning Research (JMLR)*, 12:1865–1892, 2011.

[31] S. Shalev-Shwartz and T. Zhang. Stochastic dual coordinate ascent methods for regularized loss minimization. *arXiv preprint arXiv:1209.1873*, 2012.

[32] S. Shalev-Shwartz and T. Zhang. Accelerated proximal stochastic dual coordinate ascent for regularized loss minimization. *arXiv preprint arXiv:1309.2375*, 2013.

[33] Y. Singer and J. C. Duchi. Efficient learning using forward-backward splitting. In *Advances in Neural Information Processing Systems (NIPS) 22*, pages 495–503, 2009.

[34] P. Tseng. On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal on Optimization*, 2008.

[35] H. Wang and A. Banerjee. Randomized block coordinate descent for online and stochastic optimization. *arXiv preprint arXiv:1407.0107*, 2014.

[36] L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *The Journal of Machine Learning Research (JMLR)*, 11:2543–2596, 2010.

[37] Y. Xu and W. Yin. Block stochastic gradient iteration for convex and nonconvex optimization. *arXiv preprint arXiv:1408.2597*, 2014.

[38] P. Zhao and T. Zhang. Stochastic optimization with importance sampling. *arXiv preprint arXiv:1401.2753*, 2014.

[39] T. Zhao, M. Yu, Y. Wang, R. Arora, and H. Liu. Accelerated mini-batch randomized block coordinate descent method. In *Advances in Neural Information Processing Systems (NIPS) 27*, 2014.