JOTA manuscript No.
(will be inserted by the editor)

Inexact Variable Metric Stochastic Block-Coordinate Descent for Regularized Optimization

Ching-pei Lee · Stephen J. Wright

received: date / accepted: date

Abstract Block-coordinate descent is a popular framework for large-scale regularized optimization problems with block-separable structure. Existing methods have several limitations. They often assume that subproblems can be solved exactly at each iteration, which in practical terms usually restricts the quadratic term in the subproblem to be diagonal, thus losing most of the benefits of higher-order derivative information. Moreover, in contrast to the smooth case, non-uniform sampling of the blocks has not yet been shown to improve the convergence rate bounds for regularized problems. This work proposes an inexact randomized block-coordinate descent method based on a regularized quadratic subproblem, in which the quadratic term can vary from iteration to iteration: a "variable metric." We provide a detailed convergence analysis for both convex and nonconvex problems. Our analysis generalizes, to the regularized case, Nesterov's proposal for improving convergence of block-coordinate descent by sampling proportional to the blockwise Lipschitz constants. We improve the convergence rate in the convex case by weakening the dependency on the initial objective value. Empirical results also show that significant benefits accrue from the use of a variable metric.

Keywords Variable metric; Stochastic coordinate descent; Regularized optimization; Inexact method; Arbitrary sampling

Communicated by Alexey F. Izmailov

Ching-pei Lee, Corresponding author
E-mail: leechingpei@gmail.com
Department of Mathematics & Institute for Mathematical Sciences, National University of Singapore, Singapore
Stephen J. Wright
E-mail: swright@cs.wisc.edu
Computer Sciences Department and Wisconsin Institute for Discovery, University of Wisconsin-Madison, Madison, WI, USA

1 Introduction

We consider regularized minimization problems that are a sum of a blockwise Lipschitz-continuously differentiable term and a regularization term that is convex and block-separable but possibly nondifferentiable. Many regularized empirical risk minimization (ERM) problems in machine learning have this structure with each block containing more than one coordinate; see, for example, [1–6].

We describe randomized block-coordinate-descent (BCD) type methods for minimizing this regularized problem, where only one block of variables is updated at each iteration. Moreover, we define subproblems with varying quadratic terms, and use possibly non-uniform sampling to select the block to be updated. To accommodate general quadratic terms and complicated regularizers, we also allow inexactness in computation of the update step.

Methods of this type have been discussed in existing works [7–9], but under various assumptions that may be impractical for some problems. In [7], it is required that the blockwise Lipschitz constants are known, and that the subproblem is solved to optimality, which is usually possible only when the regularizer possesses some simple structure and the quadratic approximation is diagonal. In [8], the quadratic terms are required to be fixed over iterations. The extension described in [9] is close to our framework, but (as they point out) their subproblem solver termination condition may be expensive to check except for specific choices of the regularizer. By contrast, we aim for more general applicability by requiring only that the subproblem is solved inexactly, in a sense defined below in (6), that does not even need to be checked. Moreover, these works consider only uniform sampling for the regularized problem.[1] Since [7] showed possible advantages of non-uniform sampling in the non-regularized case, we wish to consider non-uniform sampling in the regularized setting too. Others studied the cyclic version of the block-coordinate approach under various assumptions [10–13]. (The cyclic variant is significantly slower than the randomized one in the worst case [14].)

This paper contributes to both theory and practice. From the practical angle, we extend randomized BCD for regularized functions to a more flexible framework, involving variable quadratic terms and line searches, recovering existing BCD algorithms as special cases. Knowledge of blockwise Lipschitz constants is not assumed. Our algorithms are thus more practical, applicable to wider problem classes (including nonconvex ones), and significantly faster in practice. The theoretical contributions are as follows.

1. For convex problems, our analysis reflects a phenomenon that is widely observed in practice for BCD on convex problems: a kind of Q-linear convergence in the early stages of the algorithm, until a modest degree of suboptimality is attained. This result can be used to strongly weaken the dependency of the iteration complexity on the initial objective value.

[1] For the special case of non-regularized problem in which the regularizer is not present, works including [8] considered arbitrary samplings.

2. We show that global linear convergence holds under the quadratic growth condition, which is significantly weaker than strong convexity.
3. Our convergence analysis allows arbitrary sampling probabilities for the blocks, and we show that non-uniform distributions can reduce the iteration complexity significantly in some cases.
4. Inexactness in the subproblem solution affects the bounds on the number of iterations of the main algorithm in a benign way. It follows that if approximate solutions can be obtained cheaply for the subproblems, overall running time of the algorithm can be reduced significantly.

Special cases of our algorithm with a diagonal quadratic term extend existing analysis for regularized problems, showing that for the regularized problem, sampling with probability proportional to the value of the blockwise Lipschitz constants enjoys the same improvement of the iteration bound as the non-regularized case (by a factor of the maximum blockwise Lipschitz constant divided by the average of these Lipschitz constants) over uniform sampling. We believe this result to be novel in the regularized setting. The same sampling strategy produces similar advantages for nonconvex problems, an observation that is novel even for the non-regularized case.

We introduce the problem setting, our assumptions, and the proposed algorithm in Section 2. Section 3 provides detailed convergence analysis for various classes of problems, including nonconvex problems and problems for which our algorithm enjoys global linear convergence. The special case of traditional BCD (in which the quadratic terms are multiples of identity matrices) with non-uniform sampling is studied in Section 4. We discuss related works in Section 5 and efficient implementation of our algorithm for a wide class of problems in Section 6. Computational results are shown in Section 7, with concluding remarks in Section 8.

2 The Algorithm

We consider the following regularized minimization problem in this work:

$$\min_x F(x) \coloneqq f(x) + \psi(x), \tag{1}$$

where f is blockwise Lipschitz-continuously differentiable (defined below) but not necessarily convex, and the regularizer ψ is convex, extended-valued, proper, closed, and block-separable, but possibly nondifferentiable. We assume F is lower-bounded and denote the solution set by Ω, which is assumed to be nonempty. For simplicity, we assume $x \in \mathbb{R}^n$, but our methods can be applied to matrix variables, too. We decompose $x \in \mathbb{R}^n$ into N blocks such that

$$x = (x_1, x_2, \ldots, x_N) \in \mathbb{R}^n, \quad x_i \in \mathbb{R}^{n_i}, \quad n_i \in \mathbb{N}, \quad \sum_{i=1}^{N} n_i = n,$$

and assume throughout that the function ψ can be decomposed as

$$\psi(x) = \sum_{i=1}^{N} \psi_i(x_i),$$

where all ψ_i have the properties claimed for ψ above.

For the block-separability of x, we use the column submatrices of the identity denoted by $U_1, U_2, \ldots, U_N$, where $U_i \in \mathbb{R}^{n \times n_i}$ corresponds to the indices in the ith block of x. Thus, we have

$$x_i = U_i^\top x, \quad x = \sum_{i=1}^{N} U_i x_i, \quad \text{and } \nabla_i f = U_i^\top \nabla f.$$

The blockwise Lipschitz-continuously differentiable property is that there exist constants $L_i > 0$, $i = 1, 2, \ldots, N$, such that[2]

$$\|\nabla_i f(x + U_i h) - \nabla_i f(x)\| \leq L_i \|h\|, \quad \forall h \in \mathbb{R}^{n_i}, \quad \forall x \in \mathbb{R}^n. \tag{2}$$

In our description, the following parameters are used extensively.

$$L_{\max} \coloneqq \max_{1 \leq i \leq N} L_i, \quad L_{\text{avg}} \coloneqq \frac{1}{N} \sum_{i=1}^{N} L_i, \quad L_{\min} \coloneqq \min_{1 \leq i \leq N} L_i. \tag{3}$$

The kth iteration of the "exact" version of our approach proceeds as follows. Given the current iterate x^k, we pick a block i, according to some discrete probability distribution over $\{1, \ldots, N\}$, and minimize a quadratic approximation of f plus the function ψ_i for that block, to obtain the update direction d_i^k. That is, we have

$$d_i^{k*} \coloneqq \arg\min_{d_i \in \mathbb{R}^{n_i}} Q_i^k(d_i), \tag{4}$$

where

$$Q_i^k(d_i) \coloneqq \nabla_i f\left(x^k\right)^\top d_i + \tfrac{1}{2} d_i^\top H_i^k d_i + \psi_i\left(x_i^k + d_i\right) - \psi_i\left(x_i^k\right), \tag{5}$$

and $H_i^k \in \mathbb{R}^{n_i \times n_i}$ is some positive-definite matrix that can change over iterations. A backtracking line search along d_i^k is then performed to determine the step.

We focus on the case in which (4) is difficult to solve in closed form, so is solved inexactly by an iterative method, such as coordinate descent, proximal gradient, or their respective accelerated variants. We assume that d_i^k is an η-approximate solution to (4), for some $\eta \in [0, 1[$ fixed over all k and all i, satisfying the following condition:

$$-\eta Q_i^{k*} = \eta\left(Q_i^k(0) - Q_i^{k*}\right) \geq Q_i^k\left(d_i^k\right) - Q_i^{k*}, \tag{6}$$

where $Q_i^{k*} \coloneqq \inf_{d_i} Q_i^k(d_i) = Q_i^k(d_i^{k*})$. Note that the setting $\eta = 0$ corresponds to the special case in which the subproblems are solved exactly. In general, we

[2] We use the Euclidean norm throughout the paper.

Algorithm 1 Inexact variable-metric block-coordinate descent for (1)

1: Given $\beta, \gamma \in]0,1[$, $\eta \in [0,1[$, and $x^0 \in \mathbb{R}^n$;
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: Pick a probability distribution $p_1^k, \ldots, p_N^k > 0$, $\sum_i p_i^k = 1$, and sample i_k accordingly;
4: Compute $\nabla_{i_k} f(x^k)$ and choose a positive-definite $H_{i_k}^k$;
5: Approximately solve (4) for $i = i_k$ to obtain a solution $d_{i_k}^k$ satisfying (6);
6: Compute Δ_{i_k} by (8), with $i = i_k$; Set $\alpha_{i_k}^k \leftarrow 1$;
7: **while** (7) is not satisfied **do**
8: $\alpha_{i_k}^k \leftarrow \beta \alpha_{i_k}^k$;
9: **end while**
10: $x^{k+1} \leftarrow x^k + \alpha_{i_k}^k U_{i_k} d_{i_k}^k$;
11: **end for**

do not need to know the value of η or to verify the condition (6) explicitly; we merely need to know that such a value exists. For example, if the algorithm used to solve (4) has a global Q-linear convergence rate, and if we run this method for a fixed number of iterations, then we know that (6) is satisfied for *some* value $\eta \in [0,1[$, even if we do not know this value explicitly. Further discussions on how to achieve this condition can be found in, for example, [15,16]. Our analysis can be extended easily to variable, adaptive choices of η, which might lead to better iteration complexities, but for the sake of interpretability and simplicity, we fix η independent of k and i in our discussion throughout.

Our algorithm is summarized as Algorithm 1. At the current iterate x^k, a block i_k is chosen according to some discrete probability distribution over $\{1, 2, \ldots, N\}$, with *strict positive* probabilities $p_1^k, p_2^k, \ldots, p_N^k$. For the selected block i_k, we compute the partial gradient $\nabla_{i_k} f(x^k)$ and choose a positive-definite $H_{i_k}^k$, thus defining the subproblem objective (5). The selection of $H_{i_k}^k$ is application-dependent; possible choices include the (generalized) Hessian, [3] its quasi-Newton approximation, and a diagonal approximation to the Hessian. A diagonal damping term may also be added to $H_{i_k}^k$. After finding an approximate solution $d_{i_k}^k$ to (4) that satisfies (6) for some $\eta \in [0,1[$, we conduct a backtracking line search, as in [12]: Given $\beta, \gamma \in]0,1[$, we let $\alpha_{i_k}^k$ be the largest value in $\{1, \beta^1, \beta^2, \ldots\}$ such that the following sufficient decrease condition is satisfied:

$$F\left(x^k + \alpha_{i_k}^k U_{i_k} d_{i_k}^k\right) \leq F\left(x^k\right) + \alpha_{i_k}^k \gamma \Delta_{i_k}^k, \tag{7}$$

where

$$\Delta_i^k \coloneqq \nabla_i f\left(x^k\right)^\top d_i^k + \psi_i\left(x_i^k + d_i^k\right) - \psi_i\left(x_i^k\right). \tag{8}$$

Then the iterate is updated to $x^k + \alpha_{i_k}^k U_{i_k} d_{i_k}^k$.

[3] Since $\nabla_{i_k} f$ is Lipschitz continuous, it is differentiable almost everywhere. Therefore, we can at least define a generalized Hessian as suggested by [17].

3 Convergence Analysis

Our convergence analysis extends that of [16], which can be considered as a special case of our framework in which there is just one block ($N = 1$). Non-trivial modifications are needed to allow for multiple blocks and non-uniform sampling. In the results of this section we often focus on a particular iteration k, but rather than considering the consequences of updating the chosen block i_k at that iteration, we examine what would happen for *all possible* choices of $i = 1, 2, \ldots, N$, if each of these values happened to be chosen as i_k. Since the actual update block i_k is chosen randomly from among these N possibilities, we obtain results about the expected change in F by taking *expectations* over all these hypothetical choices.

The following result tracks [16, Corollary 4] and its proof is therefore omitted. Note that we focus on iteration k, and obtain lower bounds for each possible choice of update block $i = 1, 2, \ldots, N$.

Lemma 3.1 *At the kth iteration, suppose that $H_i^k \succeq m_i I$ for some $m_i > 0$, for $i = 1, 2, \ldots, N$, and that the subproblem solution d_i^k satisfies* (6)*, for each $i = 1, 2, \ldots, N$. Then we have*

$$\Delta_i^k \leq -\frac{1}{1+\sqrt{\eta}} \left(d_i^k\right)^\top H_i^k d_i^k \leq -\frac{m_i}{1+\sqrt{\eta}} \|d_i^k\|^2. \tag{9}$$

Moreover, the backtracking line search procedure in Algorithm 1 terminates finitely, with the step size α_i^k lower bounded by

$$\alpha_i^k \geq \bar{\alpha}_i := \min\left\{1, \frac{2\beta\left(1-\gamma\right) m_i}{L_i\left(1+\sqrt{\eta}\right)}\right\}. \tag{10}$$

The bound $\bar{\alpha}_i$ in (10) is a worst-case guarantee. For properly selected H_i^k (for example, when H_i^k includes true second-order information about f confined to the ith block), the steps will usually be closer to 1 because the last inequality in (9) is typically loose.

We proceed to deal with the cases in which F is convex and not necessarily convex, respectively.

3.1 Convex Case

We first state the optimal set strong convexity condition, proposed in [16], that will be used in showing global linear convergence of Algorithm 1.

Definition 3.1 Given any function F whose minimum value F^* is attainable, and for any x, define $P_\Omega(x)$ to be the (Euclidean-norm) projection of x onto the optimal set Ω. We say that F satisfies the *optimal set strong convexity* (OSSC) condition with parameter $\mu \geq 0$, if for any x and any $\lambda \in [0, 1]$, the following holds.

$$\begin{aligned} &F(\lambda x + (1-\lambda) P_\Omega(x)) \\ &\qquad \leq \lambda F\left(x\right) + \left(1-\lambda\right) F^* - \tfrac{1}{2}\mu\lambda\left(1-\lambda\right)\left\|x - P_\Omega\left(x\right)\right\|^2. \end{aligned} \tag{11}$$

The following technical lemma is crucial for both the convergence rate proofs and for motivating the choice of p_i, $i = 1, 2, \ldots, N$. We will use this result to bound the expected improvement of the objective value over one step, which leads to convergence rates for the algorithm.

Lemma 3.2 *Let f and ψ be convex with F satisfying (11) for some $\mu \geq 0$. At iteration k, we consider matrices $H_i^k \succeq 0$ with $H_i^k \in \mathbb{R}^{n_i \times n_i}$, $i = 1, \ldots, N$, probability distribution $\{p_i^k\}_{i=1}^N > 0$, and step sizes $\{\alpha_i^k\}_{i=1}^N > 0$. We define*

$$\mathcal{P}_k \coloneqq \operatorname{diag}(p_1^k I_{n_1}, \ldots, p_N^k I_{n_N}), \quad \mathcal{A}_k \coloneqq \operatorname{diag}(\alpha_1^k I_{n_1}, \ldots, \alpha_N^k I_{n_N}),$$
$$\mathcal{H}_k \coloneqq \operatorname{diag}(H_1^k, \ldots, H_N^k).$$

Then for Q_i^k defined by (5), the following holds for all $\lambda \in [0,1]$ and all θ such that $0 \leq \theta \leq \alpha_i^k p_i^k$, $i = 1, \ldots, N$:

$$\mathbb{E}_i\left[\alpha_i^k Q_i^{k*} \middle| x^k\right] \leq \theta\lambda\left(F^* - F\left(x^k\right)\right) - \tfrac{1}{2}\mu\theta\lambda(1-\lambda)\left\|x^k - P_\Omega\left(x^k\right)\right\|^2 + \tfrac{1}{2}\theta^2\lambda^2\left(x^k - P_\Omega\left(x^k\right)\right)^\top \mathcal{P}_k^{-1}\mathcal{A}_k^{-1}\mathcal{H}_k\left(x^k - P_\Omega\left(x^k\right)\right). \quad (12)$$

Proof Given any $d \in \mathbb{R}^n$, let $\tilde{d} \coloneqq \mathcal{A}_k\mathcal{P}_k d \in \mathbb{R}^n$. We obtain by change of variables that

$$\begin{aligned}
&\mathbb{E}_i\left[\alpha_i^k Q_i^{k*} \middle| x^k\right] \\
&= \min_d \ \nabla f\left(x^k\right)^\top \mathcal{A}_k\mathcal{P}_k d + \tfrac{1}{2}d^\top\mathcal{H}_k\mathcal{A}_k\mathcal{P}_k d + \sum_{i=1}^N \alpha_i^k p_i^k\left(\psi_i\left(x_i^k + d_i\right) - \psi_i\left(x_i^k\right)\right) \\
&= \min_{\tilde{d}\in\mathbb{R}^n} \ \nabla f\left(x^k\right)^\top \tilde{d} + \tfrac{1}{2}\tilde{d}^\top\mathcal{P}_k^{-1}\mathcal{A}_k^{-1}\mathcal{H}_k\tilde{d} \\
&\qquad + \sum_{i=1}^N \alpha_i^k p_i^k\left(\psi_i\left(x_i^k + \frac{\tilde{d}_i}{\alpha_i^k p_i^k}\right) - \psi_i\left(x_i^k\right)\right) \\
&\leq \min_{\tilde{d}\in\mathbb{R}^n} \ \min_{\theta\in[0,1]\,\text{s.t.}\,\frac{\theta}{\alpha_i^k p_i^k}\leq 1,\forall i} \nabla f\left(x^k\right)^\top\left(\theta\tilde{d}\right) + \tfrac{1}{2}\left(\theta\tilde{d}\right)^\top\mathcal{P}_k^{-1}\mathcal{A}_k^{-1}\mathcal{H}_k\left(\theta\tilde{d}\right) \\
&\qquad + \sum_{i=1}^N \alpha_i^k p_i^k\left(\psi_i\left(x_i^k + \frac{\theta\tilde{d}_i}{\alpha_i^k p_i^k}\right) - \psi_i\left(x_i^k\right)\right), \quad (13)
\end{aligned}$$

where each $\tilde{d}_i \in \mathbb{R}^{n_i}$. In (13), we used the fact that $\theta\tilde{d}$ is also a feasible point for the left-hand side, hence its objective value is no smaller than the minimizer.

Next, from the convexity of f, we have

$$\nabla f\left(x^k\right)^\top\theta\tilde{d} = \theta\left(\nabla f\left(x^k\right)^\top\tilde{d}\right) \leq \theta\left(f\left(x^k + \tilde{d}\right) - f\left(x^k\right)\right),$$

and from $\theta/(\alpha_i^k p_i^k) \leq 1$ for all i and the convexity of ψ, we obtain

$$\begin{aligned}
\psi_i\left(x_i^k + \frac{\theta\tilde{d}_i}{\alpha_i^k p_i^k}\right) &\leq \left(1 - \frac{\theta}{\alpha_i^k p_i^k}\right)\psi_i\left(x_i^k\right) + \frac{\theta}{\alpha_i^k p_i^k}\psi_i\left(x_i^k + \tilde{d}_i\right) \\
&= \frac{\theta}{\alpha_i^k p_i^k}\left(\psi_i\left(x_i^k + \tilde{d}_i\right) - \psi_i\left(x_i^k\right)\right) + \psi_i\left(x_i^k\right).
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
&\min_{\tilde{d}} \left\{ \min_{\theta\in[0,1]\text{ s.t. }\frac{\theta}{\alpha_i^k p_i^k}\le 1,\forall i} \nabla f\left(x^k\right)^\top \left(\theta\tilde{d}\right) + \tfrac{1}{2}\left(\theta\tilde{d}\right)^\top \mathcal{P}_k^{-1}\mathcal{A}_k^{-1}\mathcal{H}_k\left(\theta\tilde{d}\right) \right. \\
&\qquad \left. + \sum_{i=1}^N \alpha_i^k p_i^k \left(\psi_i\left(x_i^k + \frac{\theta\tilde{d}_i}{\alpha_i^k p_i^k}\right) - \psi_i\left(x_i^k\right)\right) \right\} \\
\le\ &\min_{\tilde{d}} \left\{ \min_{\theta\in[0,1]\text{ s.t. }\frac{\theta}{\alpha_i^k p_i^k}\le 1,\forall i} \theta\left(F(x^k+\tilde{d}) - F(x^k)\right) + \frac{\theta^2}{2}\tilde{d}^\top \mathcal{P}_k^{-1}\mathcal{A}_k^{-1}\mathcal{H}_k\tilde{d} \right\} \\
\le\ &\min_{\lambda\in[0,1]} \left\{ \min_{\theta\in[0,1]\text{ s.t. }\frac{\theta}{\alpha_i^k p_i^k}\le 1,\forall i} \theta\left(F\left(x^k + \lambda\left(P_\Omega\left(x^k\right) - x^k\right)\right) - F\left(x^k\right)\right) \right. \\
&\qquad \left. + \frac{\theta^2\lambda^2}{2}\left(P_\Omega\left(x^k\right) - x^k\right)^\top \mathcal{P}_k^{-1}\mathcal{A}_k^{-1}\mathcal{H}_k\left(P_\Omega\left(x^k\right) - x^k\right) \right\}. \qquad (14)
\end{aligned}$$

The result (12) then follows from combining (13), (14), and (11). □

By positive semidefiniteness of H_i^k for all i and all k, (7) implies that

$$\begin{aligned}
F\left(x^k + \alpha_{i_k}^k U_{i_k} d_{i_k}^k\right) - F\left(x^k\right) &\le \gamma\alpha_{i_k}^k \left(\Delta_{i_k}^k + \tfrac{1}{2}\left(d_{i_k}^k\right)^\top H_{i_k}^k d_{i_k}^k\right) \\
&= \gamma\alpha_{i_k}^k Q_{i_k}^k(d_{i_k}^k) \le (1-\eta)\gamma\alpha_{i_k}^k Q_{i_k}^{k*}. \qquad (15)
\end{aligned}$$

Thus Lemma 3.2 can be applied to the right-hand side of this bound to obtain an estimate of the decrease in F at the current step.

Given any x^0, we define

$$R_0 \coloneqq \sup_{x:F(x)\le F(x^0)} \|x - P_\Omega(x)\|. \qquad (16)$$

For the case of general convex problems, we make the assumption that for any x^0, the value of R_0 defined in (16) is finite. We are ready to state results concerning the rate of convergence. Part 1 of the following result shows that when the objective function optimality gap $F(x^k) - F^*$ is above a certain threshold, a linear convergence rate applies. Part 2 identifies an iteration k_0 such that for $k \ge k_0$, and for a fixed probability distribution $\{p_i\}$ for the choice of index to update, a sublinear "$1/k$" convergence rate applies. Part 3 shows that when a fixed probability distribution $\{p_i\}$ is used throughout, an initial linear phase of decrease in the expected objective function optimality gap is followed by a $1/k$ sublinear phase, and the change point of the phase is based on the expected value of $F(x^k) - F^*$ instead, making the iteration complexity calculable.

Theorem 3.1 *Assume that f and ψ are convex and (2) holds. Suppose that at all iterations k of Algorithm 1, and for any choice $i = i_k$ of the update block*

at iteration k, we have that (6) *is satisfied with a fixed $\eta \in [0, 1[$, with H_i^k chosen such that*

$$H_i^k \succeq m_i I, \quad k = 0, 1, \ldots, \tag{17}$$

for some $m_i > 0$ for all i. Then the following are true.

1. *At iteration k, given any probability distribution $\{p_i^k\}_{i=1}^N > 0$ for choosing the update block i_k, denote by $\{\alpha_i^k\}_{i=1}^N > 0$ the step sizes generated by the backtracking line search for each possible choice $i = 1, 2, \ldots, N$. (These step sizes are guaranteed to be bounded away from zero, by Lemma 3.1.) Define*

$$\pi^k \coloneqq \min_{1 \le i \le N} \alpha_i^k p_i^k, \tag{18}$$

and let $\mathcal{P}_k$, $\mathcal{A}_k$, and $\mathcal{H}_k$ be defined as in Lemma 3.2. If

$$F\left(x^k\right) - F^* \ge \left(x^k - P_\Omega\left(x^k\right)\right)^\top \mathcal{P}_k^{-1} \mathcal{A}_k^{-1} \mathcal{H}_k \left(x^k - P_\Omega\left(x^k\right)\right) \pi^k,$$

then the expected improvement in objective optimality gap at this iteration is bounded away from 1, as follows:

$$\frac{\mathbb{E}_{i_k}\left[F\left(x^{k+1}\right) - F^* \middle| x^k\right]}{\left(F\left(x^k\right) - F^*\right)} \le \left(1 - \frac{(1-\eta)\gamma\pi^k}{2}\right). \tag{19}$$

2. *Given $M_i \ge m_i, i = 1, \ldots, N$, and define*

$$\mathcal{M} \coloneqq \operatorname{diag}\left(M_1 I_{n_1}, \ldots, M_N I_{n_N}\right), \quad \bar{\mathcal{A}} \coloneqq \operatorname{diag}\left(\bar{\alpha}_1 I_{n_1}, \ldots, \bar{\alpha}_N I_{n_N}\right), \tag{20}$$

where $\bar{\alpha}_i$ are defined in Lemma 3.1. For a given probability distribution $\{p_i\}_{i=1}^N > 0$, we define

$$\mathcal{P} \coloneqq \operatorname{diag}\left(p_1 I_{n_1}, \ldots, p_N I_{n_N}\right), \quad \bar{\pi} \coloneqq \min_{1 \le i \le N} \bar{\alpha}_i p_i, \tag{21}$$

and let

$$k_0 \coloneqq \arg\min\left\{k : F\left(x^k\right) - F^* < \|\mathcal{P}^{-1}\bar{\mathcal{A}}^{-1}\mathcal{M}\|\bar{\pi}R_0^2\right\}. \tag{22}$$

Suppose that for all $k \ge k_0$, the sampling of i_k follows the distribution $\{p_i\}$, which does not depend on k, and

$$M_i I \succeq H_i^k \succeq m_i I, \quad i = 1, \ldots, N. \tag{23}$$

Then for $k \ge k_0$, the expected objective follows a sublinear convergence rate, as follows:

$$\mathbb{E}_{i_{k_0}, i_{k_0+1}, \ldots, i_{k-1}}\left[F\left(x^k\right) \middle| x^{k_0}\right] - F^* \le \frac{2\|\mathcal{P}^{-1}\bar{\mathcal{A}}^{-1}\mathcal{M}\|R_0^2}{2N + (1-\eta)\gamma(k-k_0)}. \tag{24}$$

3. *Suppose that a fixed probability distribution* $\{p_i\}_{i=1}^N > 0$ *is used throughout to sample the blocks and that* (23) *holds for all* k*. Then, defining*

$$\bar{k}_0 \coloneqq \left\lceil \max\left\{0, \frac{\log \frac{F\left(x^0\right)-F^*}{\left\|\mathcal{P}^{-1}\bar{\mathcal{A}}^{-1}\mathcal{M}\right\|\bar{\pi}R_0^2}}{\log\left(\frac{2}{2-(1-\eta)\gamma\bar{\pi}}\right)}\right\}\right\rceil, \tag{25}$$

(with $\bar{\pi}$ *defined in* (21) *and* $\lceil x \rceil$ *representing the least integer that is larger or equal to* x*), we have for all* $k < \bar{k}_0$ *that the expected objective satisfies*

$$\mathbb{E}_{i_0,\ldots,i_{k-1}}\left[F\left(x^k\right) - F^*\right] \leq \left(1 - \frac{(1-\eta)\gamma\bar{\pi}}{2}\right)^k \left(F\left(x^0\right) - F^*\right), \tag{26}$$

while for all $k \geq \bar{k}_0$*, we have*

$$\mathbb{E}_{i_0,\ldots,i_{k-1}}\left[F\left(x^k\right) - F^*\right] \leq \frac{2\left\|\mathcal{P}^{-1}\bar{\mathcal{A}}^{-1}\mathcal{M}\right\| R_0^2}{2N + (1-\eta)\gamma(k-\bar{k}_0)}. \tag{27}$$

Proof We first prove Part 1. Consider Lemma 3.2. For the general convex case, we have $\mu = 0$ in the OSSC condition (11), so (12) reduces to

$$\begin{aligned} &\mathbb{E}_{i_k}\left[\alpha_{i_k}^k Q_{i_k}^{k*}\middle| x^k\right] \\ &\leq \theta\lambda\left(F^* - F\left(x^k\right)\right) + \frac{\theta^2\lambda^2}{2}\left(x^k - P_\Omega\left(x^k\right)\right)^\top \mathcal{P}_k^{-1}\mathcal{A}_k^{-1}\mathcal{H}_k\left(x^k - P_\Omega\left(x^k\right)\right), \end{aligned} \tag{28}$$

for all $\lambda \in [0,1]$ and all $\theta \in [0, \pi^k]$. Setting $\theta = \pi^k$, we note that the right-hand side of (28) is a strongly convex function of λ for $x \notin \Omega$, so by minimizing explicitly with respect to λ, we obtain

$$\lambda = \min\left\{1, \frac{F\left(x^k\right) - F^*}{\left(x^k - P_\Omega\left(x^k\right)\right)^\top \mathcal{P}_k^{-1}\mathcal{A}_k^{-1}\mathcal{H}_k\left(x^k - P_\Omega\left(x^k\right)\right)\pi^k}\right\}. \tag{29}$$

With this setting of λ, when

$$F\left(x^k\right) - F^* \geq \left(x^k - P_\Omega\left(x^k\right)\right)^\top \mathcal{P}_k^{-1}\mathcal{A}_k^{-1}\mathcal{H}_k\left(x^k - P_\Omega\left(x^k\right)\right)\pi^k,$$

we have $\lambda = 1$ and (28) becomes

$$\mathbb{E}_{i_k}\left[\alpha_{i_k}^k Q_{i_k}^{k*}\middle| x^k\right] \leq \tfrac{1}{2}\pi^k\left(F^* - F\left(x^k\right)\right). \tag{30}$$

By combining (30) and (15), we have proved (19).

Next, we prove Part 2. Consider (28) with $\alpha_{i_k}^k$ replaced by $\bar{\alpha}_{i_k}$ (so that $\mathcal{A}_k$ is replaced by $\bar{\mathcal{A}}$) and $p_{i_k}^k$ replaced by p_{i_k} (so that $\mathcal{P}_k$ is replaced by $\mathcal{P}$). For any $k \geq k_0$, we define

$$\delta_k \coloneqq \mathbb{E}_{i_{k_0},\ldots,i_{k-1}}\left[F\left(x^k\right) - F^*\middle| x^{k_0}\right].$$

By applying the definition (16) and the bound (23) on the right-hand side of the updated (28), and then taking expectations on both sides over $i_{k_0}, \ldots, i_{k-1}$ conditional on x^{k_0}, we have that

$$\mathbb{E}_{i_{k_0},\ldots,i_k}\left[\bar{\alpha}_{i_k} Q_{i_k}^{k*} \middle| x^{k_0}\right] \leq -\theta\lambda\delta_k + \frac{\theta^2\lambda^2}{2}\left\|\mathcal{P}^{-1}\bar{\mathcal{A}}^{-1}\mathcal{M}\right\| R_0^2, \tag{31}$$

for all $\lambda \in [0,1]$ and all $\theta \in [0,\bar{\pi}]$. Setting $\theta = \bar{\pi}$ in (31), we have from (22) that since Algorithm 1 is a descent method, $\delta_k < \bar{\pi}\left\|\mathcal{P}^{-1}\bar{\mathcal{A}}^{-1}\mathcal{M}\right\| R_0^2$, for all $k \geq k_0$. Therefore, we can use

$$\lambda = \frac{\delta_k}{\bar{\pi}\left\|\mathcal{P}^{-1}\bar{\mathcal{A}}^{-1}\mathcal{M}\right\| R_0^2}$$

in (31) to obtain

$$\begin{aligned}\mathbb{E}_{i_{k_0},\ldots,i_k}\left[\bar{\alpha}_{i_k} Q_{i_k}^{k*} \middle| x^{k_0}\right] \leq & -\bar{\pi}\frac{\delta_k^2}{2\bar{\pi}\left\|\mathcal{P}^{-1}\bar{\mathcal{A}}^{-1}\mathcal{M}\right\| R_0^2} \\ = & -\frac{\delta_k^2}{2\left\|\mathcal{P}^{-1}\bar{\mathcal{A}}^{-1}\mathcal{M}\right\| R_0^2}.\end{aligned} \tag{32}$$

Therefore, by taking expectation on (15) over $i_{k_0}, \ldots, i_k$ conditional on x^{k_0}, and using (32), we obtain

$$\delta_{k+1} \leq \delta_k - \frac{(1-\eta)\,\gamma\delta_k^2}{2\left\|\mathcal{P}^{-1}\bar{\mathcal{A}}^{-1}\mathcal{M}\right\| R_0^2}. \tag{33}$$

By dividing both sides of (33) by $\delta_k\delta_{k+1}$ and noting from (7) and Lemma 3.1 that $\{F(x_k)\}$ and therefore $\{\delta_k\}$ is descending, we obtain

$$\frac{1}{\delta_k} \leq \frac{1}{\delta_{k+1}} - \frac{(1-\eta)\,\gamma\delta_k}{2\delta_{k+1}\left\|\mathcal{P}^{-1}\bar{\mathcal{A}}^{-1}\mathcal{M}\right\| R_0^2} \leq \frac{1}{\delta_{k+1}} - \frac{(1-\eta)\,\gamma}{2\left\|\mathcal{P}^{-1}\bar{\mathcal{A}}^{-1}\mathcal{M}\right\| R_0^2}. \tag{34}$$

By summing and telescoping (34), we obtain

$$\frac{1}{\delta_k} \geq \frac{1}{\delta_{k_0}} + (k-k_0)\,\frac{(1-\eta)\,\gamma}{2\left\|\mathcal{P}^{-1}\bar{\mathcal{A}}^{-1}\mathcal{M}\right\| R_0^2}. \tag{35}$$

Finally, note that because $\bar{\alpha}_i \in [0,1]$ for $i = 1, \ldots, N$, (22) implies that

$$\frac{1}{\delta_{k_0}} \geq \frac{1}{\bar{\pi}\left\|\mathcal{P}^{-1}\bar{\mathcal{A}}^{-1}\mathcal{M}\right\| R_0^2} \geq \frac{1}{\min_i p_i\left\|\mathcal{P}^{-1}\bar{\mathcal{A}}^{-1}\mathcal{M}\right\| R_0^2}. \tag{36}$$

Next, it is straightforward that the solution to

$$\min_{p_1,\ldots,p_N} \frac{1}{\min_{1\leq i\leq N} p_i} \quad \text{subject to} \quad \sum_{i=1}^{N} p_i = 1, \;\; p_i \geq 0, \;\; i = 1, \ldots, N$$

is $p_i \equiv 1/N$, and the corresponding objective value is N. Therefore, (36) further implies that

$$\frac{1}{\delta_{k_0}} \geq \frac{N}{\left\|\mathcal{P}^{-1}\bar{\mathcal{A}}^{-1}\mathcal{M}\right\| R_0^2}.$$

By combining this inequality with (35), we obtain (24).

For Part 3, we again start from (28) and replace $\alpha_{i_k}^k$ with $\bar{\alpha}_i$ in (28) to obtain

$$\begin{aligned} &\mathbb{E}_{i_k}\left[\bar{\alpha}_{i_k} Q_{i_k}^{k*} \middle| x^k\right] \\ &\leq \theta\lambda\left(F^* - F\left(x^k\right)\right) + \frac{\theta^2\lambda^2}{2}\left(x^k - P_\Omega\left(x^k\right)\right)^\top \mathcal{P}^{-1}\bar{\mathcal{A}}^{-1}\mathcal{H}_k\left(x^k - P_\Omega\left(x^k\right)\right), \end{aligned} \tag{37}$$

for all $\lambda \in [0,1]$ and all $\theta \in [0, \bar{\pi}]$. By applying (16) and (23), we have

$$\begin{aligned} &\mathbb{E}_{i_k}\left[\bar{\alpha}_{i_k} Q_{i_k}^{k*} \middle| x^k\right] \\ &\leq \theta\lambda\left(F^* - F\left(x^k\right)\right) + \frac{\theta^2\lambda^2}{2}\left\|x^k - P_\Omega\left(x^k\right)\right\| \left\|\mathcal{P}^{-1}\bar{\mathcal{A}}^{-1}\mathcal{M}\right\| \left\|x^k - P_\Omega\left(x^k\right)\right\| \\ &\leq \theta\lambda\left(F^* - F\left(x^k\right)\right) + \frac{\theta^2\lambda^2}{2}\|\mathcal{P}^{-1}\bar{\mathcal{A}}^{-1}\mathcal{M}\|R_0^2. \end{aligned}$$

Now we take expectation over $i_0, \ldots, i_{k-1}$ on both sides of this inequality (noting that the last term on the right-hand side are all constants that do not depend on i_k) to obtain

$$\mathbb{E}_{i_0,\ldots,i_k}\left[\bar{\alpha}_{i_k} Q_{i_k}^{k*}\right] \leq -\theta\lambda\mathbb{E}_{i_0,\ldots,i_{k-1}}\left[F^* - F\left(x^k\right)\right] + \frac{\theta^2\lambda^2}{2}\left\|\mathcal{P}^{-1}\bar{\mathcal{A}}^{-1}\mathcal{M}\right\| R_0^2.$$

By defining

$$\hat{\delta}_k \coloneqq \mathbb{E}_{i_0,\ldots,i_{k-1}}\left[F\left(x^k\right) - F^*\right]$$

and setting $\theta = \bar{\pi}$, we have that

$$\mathbb{E}_{i_0,\ldots,i_k}\left[\bar{\alpha}_{i_k} Q_{i_k}^{k*}\right] \leq -\bar{\pi}\lambda\hat{\delta}_k + \frac{\bar{\pi}^2\lambda^2}{2}\left\|\mathcal{P}^{-1}\bar{\mathcal{A}}^{-1}\mathcal{M}\right\| R_0^2. \tag{38}$$

The minimum of the right-hand side happens when

$$\lambda = \min\left\{1, \frac{\hat{\delta}_k}{\bar{\pi}\|\mathcal{P}^{-1}\bar{\mathcal{A}}^{-1}\mathcal{M}\|R_0^2}\right\}.$$

When the expected function value satisfies

$$\hat{\delta}_k \geq \bar{\pi}\|\mathcal{P}^{-1}\bar{\mathcal{A}}^{-1}\mathcal{M}\|R_0^2, \tag{39}$$

the minimizer is $\lambda = 1$, and the bound becomes

$$\mathbb{E}_{i_0,\ldots,i_k}\left[\bar{\alpha}_{i_k} Q_{i_k}^{k*}\right] \leq -\bar{\pi}\hat{\delta}_k + \frac{\bar{\pi}^2}{2}\left\|\mathcal{P}^{-1}\bar{\mathcal{A}}^{-1}\mathcal{M}\right\| R_0^2 \leq -\frac{\bar{\pi}\hat{\delta}_k}{2}.$$

Now we consider (15) and take expectation over $i_0, \ldots, i_k$ on both sides. Note that $Q_{i_k}^{k*} \leq 0$ so the upper bound is still valid if we replace $\alpha_{i_k}^k$ with $\bar{\alpha}_{i_k}$. Thus we obtain

$$\hat{\delta}_{k+1} - \hat{\delta}_k = \mathbb{E}_{i_0,\ldots,i_k}\left[F\left(x^{k+1}\right) - F\left(x^k\right)\right] \leq (1-\eta)\gamma \mathbb{E}_{i_0,\ldots,i_k}\left[\bar{\alpha}_{i_k} Q_{i_k}^{k*}\right] \quad (40)$$
$$\leq -\frac{(1-\eta)\gamma\bar{\pi}\hat{\delta}_k}{2}.$$

By rearranging the inequality above, we get the linear convergence of

$$\hat{\delta}_{k+1} \leq \hat{\delta}_k \left(1 - \frac{(1-\eta)\gamma\bar{\pi}}{2}\right).$$

Therefore, we always get the bound

$$\hat{\delta}_k \leq \left(1 - \frac{(1-\eta)\gamma\bar{\pi}}{2}\right)^k \left(F(x^0) - F^*\right)$$

until $\hat{\delta}_{k-1} < \bar{\pi}\|\mathcal{P}^{-1}\bar{\mathcal{A}}^{-1}\mathcal{M}\|R_0^2$. Note that $\bar{k}_0$ is obtained as the first value of k such that

$$\left(1 - \frac{(1-\eta)\gamma\bar{\pi}}{2}\right)^k \left(F\left(x^0\right) - F^*\right) \leq \bar{\pi}\|\mathcal{P}^{-1}\bar{\mathcal{A}}^{-1}\mathcal{M}\|R_0^2.$$

Therefore, for $k < \bar{k}_0$, the upper bound in (26) is larger than $\bar{\pi}\|\mathcal{P}^{-1}\bar{\mathcal{A}}^{-1}\mathcal{M}\|R_0^2$, so the rate (26) remains valid. Note that if $\hat{\delta}_k \leq \bar{\pi}\|\mathcal{P}^{-1}\bar{\mathcal{A}}^{-1}\mathcal{M}\|R_0^2$ has already held true for some $k < \bar{k}_0$, clearly this bound is still valid. On the other hand, after $\bar{k}_0$, we are guaranteed that (39) must stop holding. Thus the minimizer for (38) becomes $\lambda = \hat{\delta}_k/(\bar{\pi}\|\mathcal{P}^{-1}\bar{\mathcal{A}}^{-1}\mathcal{M}\|R_0^2)$. We then start from the first inequality of (40) and get

$$\hat{\delta}_{k+1} - \hat{\delta}_k \leq -\frac{(1-\eta)\gamma\hat{\delta}_k^2}{2\|\mathcal{P}^{-1}\bar{\mathcal{A}}^{-1}\mathcal{M}\|R_0^2}.$$

Following the same derivation we had in Part 2, we can get

$$\frac{1}{\hat{\delta}_k} \leq \frac{1}{\hat{\delta}_{k+1}} - \frac{(1-\eta)\gamma}{2\|\mathcal{P}^{-1}\bar{\mathcal{A}}^{-1}\mathcal{M}\|R_0^2}.$$

By summing and telescoping the result above, we get

$$\frac{1}{\hat{\delta}_k} \geq \frac{1}{\hat{\delta}_{\bar{k}_0}} + (k - \bar{k}_0)\frac{(1-\eta)\gamma}{2\|\mathcal{P}^{-1}\bar{\mathcal{A}}^{-1}\mathcal{M}\|R_0^2}.$$

Following the same argument in Part 2, we get the final claim in Part 3. □

The rate indicated by Part 1 of Theorem 3.1 has been observed frequently in practice, and some restricted special cases without a regularizer have been discussed in the literature [18,19]. To our knowledge, ours is the first theoretical result for BCD-type methods on general regularized problems (1). The global convergence bounds in other works depend on $R_0^2 + F(x^0) - F^*$, whereas our results significantly weaken the dependence on the initial objective value.

We can see from Part 2 of Theorem 3.1 that the optimal probability distribution after k_0 iterations is the one for which $\|\mathcal{P}^{-1}\bar{\mathcal{A}}^{-1}\mathcal{M}\|$ is minimized, that is,

$$p_i = \frac{M_i \bar{\alpha}_i^{-1}}{\sum_j M_j \bar{\alpha}_j^{-1}}. \tag{41}$$

It is possible to replace $\bar{\alpha}_i$ and M_i with the values α_i^k and $\|H_i^k\|$ (respectively), to obtain adaptive probabilities and possibly sharper rates, but we fix the probabilities for the sake of more succinct analysis. We discuss in Section 5 some issues relating to the use of adaptive probabilities.

We now consider the case that F satisfies the quadratic growth condition

$$F(x) - F^* \geq \frac{\mu}{2} \|x - P_\Omega(x)\|^2 \tag{42}$$

for some $\mu > 0$. This condition is implied by the OSSC condition (11) but not vice versa. The following theorem shows a global Q-linear convergence result for this case.

Theorem 3.2 *Assume that f and ψ are convex and that (2) and (42) hold for some $L_1, \ldots, L_N, \mu > 0$. Suppose that at the kth iteration of Algorithm 1, (6) is satisfied with some $\eta \in [0, 1[$ and H_i^k is chosen such that (17) holds for some $m_i > 0$ and all $i = 1, 2, \ldots, N$, so that the step sizes α_i^k are all bounded away from 0, as indicated by Lemma 3.1. Then given any probability distribution $\{p_i^k\} > 0$, with π^k defined as in (18), we have that the expected decrease at iteration k is*

$$\frac{\mathbb{E}_{i_k}\left[F\left(x^{k+1}\right) - F^* \middle| x^k\right]}{F\left(x^k\right) - F^*} \leq 1 - (1-\eta)\gamma\rho_k, \tag{43}$$

where ρ_k is bounded below by the following quantities:

$$\frac{\mu}{4\|\mathcal{P}_k^{-1}\mathcal{A}_k^{-1}\mathcal{H}_k\|}, \quad \text{if} \quad \frac{\mu}{2\|\mathcal{P}_k^{-1}\mathcal{A}_k^{-1}\mathcal{H}_k\|\pi^k} \leq 1, \tag{44a}$$

$$\pi^k\left(1 - \frac{\pi^k\|\mathcal{P}_k^{-1}\mathcal{A}_k^{-1}\mathcal{H}_k\|}{\mu}\right), \quad \text{otherwise.} \tag{44b}$$

Proof By (15), (28), the Cauchy-Schwarz inequality, and (42), we have

$$\begin{aligned}
&\mathbb{E}_{i_k}\left[F\left(x^{k+1}\right) - F\left(x^k\right) \mid x^k\right] \\
&\leq \gamma(1-\eta)\mathbb{E}_{i_k}\left[\alpha_{i_k}^k Q_{i_k}^{k*}\right] \\
&\leq \gamma(1-\eta)\theta\left(F\left(x^k\right) - F^*\right)\left(-\lambda + \frac{\theta\lambda^2\left\|\mathcal{P}_k^{-1}\mathcal{A}_k^{-1}\mathcal{H}_k\right\|}{\mu}\right),
\end{aligned} \tag{45}$$

for all $\lambda \in [0,1]$ and all $\theta \in [0, \pi^k]$. By the same argument as in the previous proofs, we let $\theta = \pi^k$. By minimizing the right-hand side of (45) over $\lambda \in [0,1]$, we obtain the two cases (44a) and (44b). □

We can improve on Theorem 3.2 for problems satisfying the OSSC condition (11).

Theorem 3.3 *Assume that f and ψ are convex and that* (2) *and* (11) *hold for some $L_1, \dots, L_N, \mu > 0$. Suppose that at the kth iteration of Algorithm 1,* (6) *is satisfied for some $\eta \in [0,1[$ and H_i^k is chosen such that* (17) *holds for some $m_i > 0$ and all $i = 1, 2, \dots, N$ so that the step sizes α_i^k are all lower bounded away from 0 as indicated by Lemma 3.1. Then given any probability distribution $\{p_i^k\} > 0$, and with π^k defined as in* (18)*, the expected function decrease at iteration k is the same as* (43)*, but with ρ_k lower-bounded by*

$$\rho_k \geq \left(\frac{1}{\pi^k} + \max_i \frac{\left\| H_i^k \right\|}{\mu \alpha_i^k p_i^k} \right)^{-1}. \tag{46}$$

Proof We note by bounding the last term in (12) that

$$\mathbb{E}_i \left[\alpha_i^k Q_i^{k*} \middle| x^k \right] \leq \theta\lambda \left(F^* - F\left(x^k\right) \right) - \tfrac{1}{2}\mu\theta\lambda \left(1-\lambda\right) \left\| x^k - P_\Omega\left(x^k\right) \right\|^2 + \tfrac{1}{2}\theta^2\lambda^2 \left\| x^k - P_\Omega(x^k) \right\|^2 \left\| \mathcal{P}_k^{-1} \mathcal{A}_k^{-1} \mathcal{H}_k \right\|.$$

Thus by setting $\lambda = \mu / (\mu + \|\mathcal{P}_k^{-1}\mathcal{A}_k^{-1}\mathcal{H}_k\|\theta) \in [0,1]$, the last two terms cancel. Then by setting $\theta = \pi^k$, we obtain

$$\begin{aligned} \mathbb{E}_{i_k} \left[\alpha_{i_k} Q_{i_k}^{k*} \middle| x^k \right] &\leq \frac{\mu\theta}{\mu + \left\| \mathcal{P}_k^{-1} \mathcal{A}_k^{-1} \mathcal{H}_k \right\| \theta} \left(F^* - F\left(x^k\right) \right) \\ &= \frac{1}{\frac{1}{\theta} + \frac{\left\| \mathcal{P}_k^{-1} \mathcal{A}_k^{-1} \mathcal{H}_k \right\|}{\mu}} \left(F^* - F\left(x^k\right) \right) \\ &= \frac{1}{\frac{1}{\pi^k} + \max_i \frac{\left\| H_i^k \right\|}{\mu \alpha_i^k p_i^k}} \left(F^* - F\left(x^k\right) \right). \end{aligned} \tag{47}$$

By combining (47) and (15), we obtain the desired result. □

For problems on which Theorem 3.2 or 3.3 holds, Theorem 3.1 is also applicable, and the early linear convergence rate can be faster than the global rates described in Theorems 3.2 and 3.3 (always better than the rate in Theorem 3.2 and for Theorem 3.3 it depends on the value of μ and $\|H_i^k\|$). Thus, we could sharpen the global iteration complexity for problems satisfying the OSSC condition (11) with $\mu > 0$ by using Theorem 3.1. We also notice that the rate in Theorem 3.3 is faster than that in Theorem 3.2, which is why we consider these two conditions separately.

Note too that with knowledge of α_i and $\|H_i^k\|$, we could in principle minimize the expected gap $\mathbb{E}_{i_k}\left[F\left(x^{k+1}\right) - F^* \middle| x^k \right]$ by minimizing the denominator on the right-hand side of (46) and (44) with respect to p_i^k over $p_i^k > 0$

and $\sum_i p_i^k = 1$. Such an approach is not practical except in the special cases discussed in Section 4, as it is unclear how to find α_i^k and $\|H_i^k\|$ in general for the blocks not selected.

Theorems 3.2 and 3.3 suggest that larger step sizes lead to faster convergence. When H_i^k incorporates curvature information of f, empirically we tend to have much larger step sizes than the lower bound predicted in Lemma 3.1, and thus the practical performance of using the Hessian or its approximation usually outperforms using a multiple of the identity as H_i^k.

All the results here can be combined in a standard way with Markov's inequality to get high-probability bounds for the objective value. We omit these results.

3.2 **Nonconvex Case**

When f is not necessarily convex, we cannot use Lemma 3.2 to estimate the expected model decrease at each iteration, and we cannot guarantee convergence to the global optima. Instead, we analyze the convergence of certain measures of stationarity.

The first measure we consider is how fast the optimal objective of the subproblem (4) converges to zero. Since the subproblems are strongly convex, this measure is zero if and only if the optimal solution is the zero vector, implying that the algorithm will not step away from this point. These claims are verified in the following lemma.

Lemma 3.3 *At iteration k, assume that in (4)-(5) we have $H_i^k \succeq m_i I$ for some $m_i > 0$, $i = 1, 2, \ldots, N$, and that (2) is satisfied for some positive values $L_1, L_2, \ldots, L_N$. Then for any positive step sizes $\{\alpha_i^k\}_{i=1}^N > 0$ and any probability distribution $\{p_i^k\}_{i=1}^N > 0$, we have*

$$\mathbb{E}_i[\alpha_i^k Q_i^{k*}] = 0 \;\Leftrightarrow\; Q_i^{k*} = 0,\ i = 1, \ldots, N \;\Leftrightarrow\; 0 \in \partial F\left(x^k\right), \tag{48}$$

where $\partial F(x^k) = \nabla f(x^k) + \partial\psi(x^k)$ is the generalized gradient of F at x^k.

Proof From (9) in Lemma 3.1, by setting $\eta = 0$ we see that for all i and k we have $Q_i^{k*} \leq 0$, proving the first equivalence in (48). To prove the second equivalence, we first notice that since Q_i^k are all strongly convex and $Q_i^k(0) \equiv 0$, $Q_i^{k*} = 0$ if and only if $d_i^{k*} = 0$, where d_i^{k*} is defined in (4). Therefore, it suffices to prove that

$$d_i^{k*} = 0 \quad \Leftrightarrow \quad -\nabla_i f\left(x^k\right) \in \partial\psi_i\left(x_i^k\right),\ \ i = 1, \ldots, N. \tag{49}$$

From optimality of (4), we have

$$-\left(\nabla_i f\left(x^k\right) + H_i^k d_i^{k*}\right) \in \partial\psi_i\left(x_i^k + d_i^{k*}\right). \tag{50}$$

When $d_i^{k*} = 0$, (50) implies that $-\nabla_i f(x^k) \in \partial\psi_i(x_i^k)$. Conversely, if

$$-\nabla_i f(x^k) \in \partial\psi_i(x_i^k). \tag{51}$$

We have from the convexity of ψ_i together with (51) and (50) that

$$\psi_i\left(x_i^k + d_i^{k*}\right) \geq \psi_i\left(x_i^k\right) - \nabla_i f(x^k)^\top d_i^{k*},$$
$$\psi_i\left(x_i^k\right) \geq \psi_i\left(x_i^k + d_i^{k*}\right) - (d_i^{k*})^\top(-\nabla_i f(x^k) - H_i^k d_i^{k*}).$$

By adding these two inequalities, we obtain $0 \geq \left(d_i^{k*}\right)^\top H_i^k d_i^{k*}$, so that $d_i^{k*} = 0$ by the positive definiteness of H_i^k. □

The second measure of convergence is the following:

$$G_k \coloneqq \arg\min_d \nabla f(x^k)^\top d + \tfrac{1}{2} d^\top d + \psi(x^k + d). \tag{52}$$

From Lemma 3.3, it is clear that $G_k = 0$ if and only if $0 \in \partial F(x^k)$, so G_k can serve as an indicator for closeness to stationarity.

We show convergence rates for the two measures proposed above.

Theorem 3.4 *Given any x^0 in Algorithm 1, let $\{\alpha_i^k\}_{i=1}^N > 0$ be the step sizes generated by the line search procedure for $k = 0, 1, 2, \ldots$. If $H_i^k \succeq 0$ for all i and k, we have*

$$\min_{0\leq k\leq T} \left|\mathbb{E}_{i_0,\ldots,i_k}\left[\alpha_{i_k}^k Q_{i_k}^k\left(d_{i_k}^k\right)\right]\right| \leq \frac{F\left(x^0\right) - F^*}{\gamma\left(T+1\right)}, \quad \text{for all } T \geq 0. \tag{53}$$

Moreover, $\mathbb{E}_{i_0,\ldots,i_k}\left[\alpha_{i_k}^k Q_{i_k}^k(d_{i_k}^k)\right] \to 0$ as k approaches infinity.

Proof Taking expectation on (15) over i_k, we obtain

$$\mathbb{E}_{i_k}\left[F\left(x^{k+1}\right)\middle|\, x^k\right] - F\left(x^k\right) \leq \gamma \mathbb{E}_{i_k}\left[\alpha_{i_k} Q_{i_k}^k\left(d_{i_k}^k\right)\middle|\, x^k\right]. \tag{54}$$

By taking expectation on (54) over $i_0, \ldots, i_{k-1}$ and summing over $k = 0, 1, \ldots, T$, and noting from (6) and Lemma 3.1 that $Q_i^k(d_i^k) \leq 0$ for all k and all i, we obtain

$$\begin{aligned}
&\gamma \sum_{k=0}^{T} \left|\mathbb{E}_{i_0,\ldots,i_k}\left[\alpha_{i_k} Q_{i_k}^k\left(d_{i_k}^k\right)\right]\right| \\
=\ & -\gamma \sum_{k=0}^{T} \mathbb{E}_{i_0,\ldots,i_k}\left[\alpha_{i_k} Q_{i_k}^k\left(d_{i_k}^k\right)\right] \\
\leq\ & \sum_{k=0}^{T} \left\{\mathbb{E}_{i_0,\ldots,i_{k-1}}\left[F\left(x^k\right)\right] - \mathbb{E}_{i_0,\ldots,i_k}\left[F\left(x^{k+1}\right)\right]\right\} \\
=\ & F\left(x^0\right) - \mathbb{E}_{i_0,\ldots,i_T}\left[F\left(x^{T+1}\right)\right] \leq\ F\left(x^0\right) - F^*.
\end{aligned} \tag{55}$$

The result now follows from

$$\sum_{k=0}^{T} \left|\mathbb{E}_{i_0,\ldots,i_k}\left[\alpha_{i_k} Q_{i_k}^k\left(d_{i_k}^k\right)\right]\right| \geq (T+1) \min_{0\leq k\leq T} \left|\mathbb{E}_{i_0,\ldots,i_k}\left[\alpha_{i_k} Q_{i_k}^k\left(d_{i_k}^k\right)\right]\right|$$

The result that $\left|\mathbb{E}_{i_0,\ldots,i_k}\left[\alpha_{i_k} Q_{i_k}^k\left(d_{i_k}^k\right)\right]\right| \to 0$ follows from the summability implied by (55). □

Unlike previous results, the convergence speed for the right-hand side of (53) is independent of how accurately the subproblem is solved, the probability distributions for sampling the blocks, and the step sizes. We next consider the second measure (52) and show that its convergence behavior depends on these factors. We need the following lemma from [12].

Lemma 3.4 ([12, Lemma 3]) *Given x^k, assume that H_i^k satisfies* (23) *for some $M_i \geq m_i > 0$ for all i. Then we have*

$$\left\|U_i^\top G_k\right\| \leq \frac{1+\frac{1}{m_i}+\sqrt{1-2\frac{1}{M_i}+\frac{1}{m_i^2}}}{2} M_i \left\|d_i^{k*}\right\|.$$

By combining this lemma with Theorem 3.4, we can show a convergence rate for $\min_{0\leq k\leq T} \mathbb{E}_{i_0,\ldots,i_k} \|G_k\|$.

Corollary 3.1 *Assume that H_i^k satisfies* (23) *for all $k = 0, 1, \ldots$ and all $i = 1, 2, \ldots, N$. Let $\{\alpha_i^k\}_{i=1}^N > 0$ be the step sizes generated by the line search procedure. Then we have*

$$\begin{aligned}&\min_{0\leq k\leq T} \mathbb{E}_{i_0,\ldots,i_{k-1}} \left[\|G_k\|^2\right] \\ &\leq \frac{F\left(x^0\right)-F^*}{2(1-\eta)\gamma(T+1)} \max_{0\leq k\leq T,\ 1\leq i\leq N} \frac{M_i^2\left(1+\frac{1}{m_i}+\sqrt{1-2\frac{1}{M_i}+\frac{1}{m_i^2}}\right)^2}{p_i^k \alpha_i^k m_i}. \end{aligned} \quad (56)$$

Proof We consider Theorem 3.4 and let $\bar{k}$ be the iteration that achieves the minimum on the left-hand side of (53). We have from (6) and Theorem 3.4 that

$$\frac{F\left(x^0\right)-F^*}{\gamma\left(T+1\right)} \geq \left|\mathbb{E}_{i_0,\ldots,i_{\bar{k}}}\left[\alpha_{i_{\bar{k}}}^{\bar{k}} Q_{i_{\bar{k}}}^{\bar{k}}\left(d_{i_{\bar{k}}}^{\bar{k}}\right)\right]\right| \geq -(1-\eta)\mathbb{E}_{i_0,\ldots,i_{\bar{k}}}\left[\alpha_{i_{\bar{k}}}^{\bar{k}} Q_{i_{\bar{k}}}^{\bar{k}*}\right]. \quad (57)$$

Since $H_i^k \succeq m_i I$ from (23) and the ψ_i are convex, we have that for all i and k, the functions Q_i^k are m_i-strongly convex and hence satisfy (42) with $\mu = m_i$. Therefore, we have

$$Q_i^k(0) - Q_i^{k*} = -Q_i^{k*} \geq \frac{m_i}{2}\left\|d_i^{k*}\right\|^2, \quad \text{for all } k \text{ and all } i = 1, 2, \ldots, N. \quad (58)$$

Algorithm 2 Inexact Randomized BCD with Unit Step Size for (1)

Given $\eta \in [0,1[$ and $x^0 \in \mathbb{R}^n$;
for $k = 0, 1, 2, \ldots$ **do**
  Pick a probability distribution $p_1^k, \ldots, p_N^k > 0$, $\sum_i p_i^k = 1$, and sample i_k accordingly;
  Compute $\nabla_{i_k} f(x^k)$ and let $H_{i_k}^k = L_{i_k} I$;
  Approximately solve (4) to obtain a solution $d_{i_k}^k$ satisfying (6);
  $x^{k+1} \leftarrow x^k + U_{i_k} d_{i_k}^k$;
end for

Algorithm 3 Inexact Randomized BCD with Short Step Size for (1)

Given $\eta \in [0,1[$ and $x^0 \in \mathbb{R}^n$;
for $k = 0, 1, 2, \ldots$ **do**
  Pick a probability distribution $p_1^k, \ldots, p_N^k > 0$, $\sum_i p_i^k = 1$, and sample i_k accordingly;
  Compute $\nabla_{i_k} f(x^k)$ and let $H_{i_k}^k = L_{\min} I$;
  Approximately solve (4) to obtain a solution $d_{i_k}^k$ satisfying (6);
  $x^{k+1} \leftarrow x^k + \frac{L_{\min}}{L_{i_k}} U_{i_k} d_{i_k}^k$;
end for

By substituting (58) into (57) and using Lemma 3.4, we obtain

$$\begin{aligned}
&\frac{F\left(x^0\right) - F^*}{(1-\eta)\gamma\,(T+1)} \\
&\geq \frac{1}{2}\sum_{i=1}^{N} p_i^{\bar{k}} \alpha_i^{\bar{k}} m_i \mathbb{E}_{i_0,\ldots,i_{\bar{k}-1}}\left[\left\|d_i^{\bar{k}*}\right\|^2\right] \qquad (59)\\
&\geq 2\sum_{i=1}^{N} \frac{p_i^{\bar{k}} \alpha_i^{\bar{k}} m_i}{M_i^2\left(1+\frac{1}{m_i}+\sqrt{1-2\frac{1}{M_i}+\frac{1}{m_i^2}}\right)^2} \mathbb{E}_{i_0,\ldots,i_{\bar{k}-1}}\left[\left\|U_i^\top G_{\bar{k}}\right\|^2\right] \\
&\geq 2\mathbb{E}_{i_0,\ldots,i_{\bar{k}-1}}\left[\left\|G_{\bar{k}}\right\|^2\right] \min_{1\leq i\leq N} \frac{p_i^{\bar{k}} \alpha_i^{\bar{k}} m_i}{M_i^2\left(1+\frac{1}{m_i}+\sqrt{1-2\frac{1}{M_i}+\frac{1}{m_i^2}}\right)^2}, \qquad (60)
\end{aligned}$$

where in (60), we used the fact that $\|x\|^2 = \sum_{i=1}^N \|U_i^\top x\|^2$ for any $x \in \mathbb{R}^n$. The result (56) is then proved by noting that

$$\min_{0\leq k\leq T} \mathbb{E}_{i_0,\ldots,i_{k-1}} \|G_k\|^2 \leq \mathbb{E}_{i_0,\ldots,i_{\bar{k}-1}} \|G_{\bar{k}}\|^2. \qquad \square$$

Corollary 3.1 reveals that line search can help improve the convergence speed as larger values of α_i^k make the right-hand side of (56) smaller, and non-uniform sampling can possibly lead to faster convergence.

4 Randomized Block Coordinate Descent

Non-uniform sampling in coordinate descent for smooth convex objectives was discussed in [7]. In this section, we extend these results to the regularized objective function (1), using results from Section 3. In the non-regularized case, the update for the ith block described in [7] is $-\nabla_i f(x)/L_i$, which can be viewed as either the solution of

$$\min_{d_i} \quad \nabla_i f(x)^\top d_i + \tfrac{1}{2} L_i d_i^\top d_i$$

with unit step size, or equivalently as the solution of

$$\min_{d_i} \quad \nabla_i f(x)^\top d_i + \tfrac{1}{2} L_{\min} d_i^\top d_i$$

with step size $L_{\min}/L_i$ (so that the step size is no larger than 1). As in [7], we do not consider backtracking, but assume that L_i is available, and thus an appropriate choice for α_i can be made. When these step calculations are adapted to the regularized case (1), as in (4)-(5) with $H_i^k = L_i I$ and $H_i^k = L_{\min} I$, respectively, they lose their equivalence to each other and give different directions. The resulting special cases of Algorithm 1 are shown as Algorithms 2 and 3.

We show in the following result that both approaches achieve a guaranteed decrease in the objective.

Lemma 4.1 *Assume that* (2) *holds, and consider iteration k of Algorithm 1. If the ith block is selected for updating, and $H_i^k \succeq c_i I$ in* (5) *for some $c_i \in]0, L_i]$, then $\hat{\alpha}_i := c_i/L_i$ satisfies*

$$F(x^k + \alpha U_i d_i^k) - F(x^k) \leq \alpha Q_i^k(d_i^k), \text{ for all } d_i^k \in \mathbb{R}^{n_i} \text{ and all } \alpha \in [0, \hat{\alpha}_i]. \quad (61)$$

Proof Because $c_i \in]0, L_i]$, we have $\hat{\alpha}_i = c_i/L_i \in]0, 1]$. Thus from (2) and the convexity of ψ, we have for any $\alpha \in [0, \hat{\alpha}_i]$ that

$$\begin{aligned}
&F\left(x^k + \alpha U_i d_i^k\right) \\
&= f\left(x^k + \alpha U_i d_i^k\right) + \psi\left(x^k + \alpha U_i d_i^k\right) \\
&\leq f\left(x^k\right) + \alpha \nabla_i f\left(x^k\right)^\top d_i^k + \tfrac{1}{2} L_i \alpha^2 \left\|d_i^k\right\|^2 + \alpha\psi\left(x^k + U_i d_i^k\right) + (1-\alpha)\,\psi\left(x^k\right) \\
&= f\left(x^k\right) + \psi\left(x^k\right) + \alpha\left[\nabla_i f(x^k)^\top d_i^k + \tfrac{1}{2} L_i \alpha \left\|d_i^k\right\|^2 + \psi\left(x^k + U_i d_i^k\right) - \psi\left(x^k\right)\right] \\
&\leq F\left(x^k\right) + \alpha Q_i^k\left(d_i^k\right).
\end{aligned}$$

In the last inequality, we used the fact that for the term H_i appearing in $Q_i^k(d_i^k)$, we have

$$H_i \succeq c_i I = \hat{\alpha}_i L_i I \succeq \alpha L_i I. \qquad \square$$

With the help of Lemma 4.1, we can discuss the iteration complexities of randomized BCD (Algorithms 2 and 3) with different sampling strategies. We first consider the interpretation in Algorithm 2, starting from the case in which f is convex. The results below are direct applications of Theorem 3.1.

Corollary 4.1 *Consider Algorithm 2 applied to* (1) *with convex f, and assume that* (2) *holds. The expected objective value satisfies the following.*

1. *With uniform sampling* $p_i^k \equiv 1/N$, *we have the following.*
   1.1. *If* $F\left(x^k\right) - F^* \geq \left(x^k - P_\Omega\left(x^k\right)\right)^\top \mathcal{L}\left(x^k - P_\Omega\left(x^k\right)\right)$, *where*

$$\mathcal{L} \coloneqq \operatorname{diag}\left(L_1 I_{n_1}, \ldots, L_N I_{n_N}\right), \tag{62}$$

   we have

$$\mathbb{E}_{i_k}\left[F\left(x^{k+1}\right) - F^* \mid x^k\right] \leq \left(1 - \frac{(1-\eta)}{2N}\right)\left(F\left(x^k\right) - F^*\right).$$

   1.2. *For all* $k \geq k_0$, *where* $k_0 \coloneqq \arg\min\{k : F\left(x^k\right) - F^* < L_{\max} R_0^2\}$, *we have*

$$\mathbb{E}_{i_{k_0},\ldots,i_{k-1}}\left[F\left(x^k\right) \mid x^{k_0}\right] - F^* \leq \frac{2NL_{\max}R_0^2}{2N + (1-\eta)(k-k_0)}.$$

2. *When* p_i^k *are defined as*

$$p_i^k = \frac{L_i}{NL_{avg}}, \quad i = 1, 2, \ldots, N, \tag{63}$$

   we have the following.
   2.1. *If* $F\left(x^k\right) - F^* \geq L_{\min}\left\|x^k - P_\Omega\left(x^k\right)\right\|^2$, *then*

$$\mathbb{E}_{i_k}\left[F\left(x^{k+1}\right) - F^* \mid x^k\right] \leq \left(1 - \frac{L_{\min}(1-\eta)}{2NL_{avg}}\right)\left(F\left(x^k\right) - F^*\right).$$

   2.2. *For all* $k \geq k_0$, *where* $k_0 \coloneqq \arg\min\{k : F\left(x^k\right) - F^* < L_{\min} R_0^2\}$, *we have*

$$\mathbb{E}_{i_{k_0},\ldots,i_{k-1}}\left[F\left(x^k\right) \mid x^{k_0}\right] - F^* \leq \frac{2NL_{avg}R_0^2}{2N + (1-\eta)(k-k_0)}.$$

The strategy (63) is referred to henceforth as "Lipschitz sampling." In both Algorithms 2 and 3, we have

$$\frac{\|H_i^k\|}{\alpha_i^k} = L_i.$$

Recalling the definitions of M_i from (23) and $\mathcal{M}$ from (20), we have that since H_i^k are fixed over k for all i, both algorithms have $\|H_i^k\| \equiv M_i$. Therefore, (63) matches the optimal probability distribution (41), resulting in

$$\|\mathcal{P}^{-1}\mathcal{A}^{-1}\mathcal{M}\| = NL_{\text{avg}}. \tag{64}$$

We next consider the case in which the OSSC condition (11) holds for some $\mu > 0$.

Corollary 4.2 *Consider Algorithm 2 and assume that* (2) *holds. For problems satisfying* (11) *with* $\mu \in]0, L_{\min}]$*, the iteration complexity for the expected objective value to reach*

$$\mathbb{E}_{i_0,\ldots,i_{k-1}} F\left(x^k\right) - F^* \leq \epsilon$$

for any given $\epsilon > 0$ *is as follows. When* $p_i^k \equiv 1/N$*,* $i = 1, 2, \ldots, N$*, we have complexity*

$$O\left(\frac{NL_{\max}}{(1-\eta)\mu}\log(1/\epsilon)\right),$$

while if the p_i^k *are defined by* (63)*, we have complexity*

$$O\left(\frac{NL_{avg}}{(1-\eta)\mu}\log(1/\epsilon)\right).$$

Proof As shown in Lemma 4.1, this choice of H_i and α_i satisfies (15) with $\gamma = 1$. Thus the case of uniform sampling is directly obtained from Theorem 3.3 and the known fact that for Q-linear convergence rate of $1-\tau$ with $\tau \in]0, 1[$, the iteration complexity for obtaining an ϵ-accurate solution is $O(\tau^{-1}\log(1/\epsilon))$.

For (63), we use (12) to derive a different result. Since $\|\mathcal{P}_k^{-1}\mathcal{A}_k^{-1}\mathcal{H}_k\| = NL_{\text{avg}}$ (from (64)), and by letting $\lambda = 1/2$ and $\theta = \mu/(NL_{\text{avg}})$, (12) leads to

$$\mathbb{E}_{i_k}\left[\alpha_{i_k} Q_{i_k}^{k*} \middle| x^k\right] \leq \frac{\mu}{2NL_{\text{avg}}}\left(F^* - F\left(x^k\right)\right). \tag{65}$$

The remainder of the proof tracks the proof of Theorem 3.3 to get a Q-linear convergence rate. □

When $\eta = 0$ (so that the solutions of the subproblems are exact), the rates in Corollaries 4.1 and 4.2 are similar to Nesterov's result [7] for the non-regularized case with the same sampling strategies, if we interpret this result in the Euclidean norm. The advantage of Lipschitz sampling over uniform sampling is seen clearly. Note that [7] discusses the case of constrained optimization, which can be treated as a special case of regularized optimization. In this special case, Nesterov shows a $O(1/k)$ convergence rate of the objective value when the objective is convex, but the convergence speed depends on $(R_0^2/2 + F(x^0) - F^*)$. Here, we weaken the dependency on the initial objective value by showing linear convergence in the early stages of iteration. The case in which F satisfies (42) can also provide linear convergence for Algorithm 2, but the consequent rates do not suggest clear advantages of the Lipschitz sampling, and the derivations are trivial. We therefore omit these results.

When f is not necessarily convex, Algorithm 2 still benefits from Lipschitz sampling, as we now discuss.

Corollary 4.3 *Consider Algorithm 2 and assume that* (2) *holds. Suppose that a fixed probability distribution is used for the choice of blocks, that is,* $p_i^k \equiv p_i$ *for all* $k \geq 0$ *and all* $i = 1, 2, \ldots, N$*. Then we have that*

$$\min_{0 \leq k \leq T} \mathbb{E}_{i_0, \ldots, i_{k-1}} \left\| G_k \right\|^2 \leq \frac{2(F\left(x^0\right) - F^*)}{(1-\eta)(T+1)} \max_{1 \leq i \leq N} \frac{L_i}{p_i}.$$

Therefore, when uniform sampling is used, we obtain

$$\min_{0 \leq k \leq T} \mathbb{E}_{i_0, \ldots, i_{k-1}} \left\| G_k \right\|^2 \leq \frac{2NL_{\max}(F\left(x^0\right) - F^*)}{(1-\eta)(T+1)},$$

whereas when Lipschitz sampling is used, we obtain

$$\min_{0 \leq k \leq T} \mathbb{E}_{i_0, \ldots, i_{k-1}} \left\| G_k \right\|^2 \leq \frac{2NL_{avg}(F\left(x^0\right) - F^*)}{(1-\eta)(T+1)}.$$

Our result here for the case of uniform sampling is similar to that in [20], but we show that Lipschitz sampling can improve the convergence rate by considering a slightly different measure of stationarity.

We turn now to Algorithm 3, which can also be viewed as an extension of the algorithm in [7] to the regularized problem (1).

Corollary 4.4 *Consider Algorithm 3 and assume that* (2) *holds. Suppose that a fixed probability distribution is used for the choice of blocks, that is,* $p_i^k \equiv p_i$ *for all* $k \geq 0$ *and all* $i = 1, 2, \ldots, N$*. Then the following claims hold.*

1. *For uniform sampling* ($p_i = 1/N$, $i = 1, 2, \ldots, N$)*, we have*

$$\min_{0 \leq k \leq T} \mathbb{E}_{i_0, \ldots, i_{k-1}} \left\| G_k \right\|^2 \leq \frac{2NL_{\max}(F\left(x^0\right) - F^*)}{(1-\eta)(T+1)}.$$

2. *If* f *is convex, then for uniform sampling, we have the following results.*
   2.1. When

$$F(x^k) - F^* \geq (1/L_{\max})(x^k - P_\Omega(x^k))^\top \mathcal{L}(x^k - P_\Omega(x^k)), \tag{66}$$

   where $\mathcal{L}$ *is defined in* (62)*, the convergence of the expected objective value is Q-linear:*

$$\mathbb{E}_{i_k}\left[F\left(x^{k+1}\right) - F^* \middle| \, x^k\right] \leq \left(1 - \frac{(1-\eta)}{2NL_{\max}}\right)\left(F\left(x^k\right) - F^*\right).$$

   2.2. For all $k \geq k_0$*, where* $k_0 \coloneqq \arg\min\{k : F\left(x^k\right) - F^* < R_0^2\}$*, the expected objective follows a sublinear convergence rate*

$$\mathbb{E}_{i_{k_0}, \ldots, i_{k-1}}\left[F\left(x^k\right) \middle| \, x^{k_0}\right] - F^* \leq \frac{2NL_{\max}R_0^2}{2N + (1-\eta)(k-k_0)}.$$

3. *If F satisfies the OSSC condition* (11) *for some $\mu > 0$, then for uniform sampling, we have*

$$\mathbb{E}_{i_k}\left[F(x^{k+1}) - F^* \middle| x^k\right] \leq \left(1 - \frac{(1-\eta)(1+1/\mu)^{-1}}{NL_{\max}}\right)\left(F(x^k) - F^*\right).$$

4. *With p_i chosen from* (63)*, results in Parts 1 and 3 hold, with $L_{\max}$ improved to L_{avg}. For Part 2, for convex f, we obtain the same improvement from $L_{\max}$ to L_{avg} for all rates, but the condition for early linear convergence becomes $F(x^k) - F^* \geq \|x^k - P_\Omega(x^k)\|^2$ rather than* (66)*.*

Whether the OSSC condition (11) holds or not, the bounds indicate a potential improvement of $L_{\max}/L_{\text{avg}}$ in iteration bounds when (63) is used.

An advantage of Algorithm 2 over Algorithm 3 is that when the solution exhibits some partial smoothness structure, Algorithm 2 may be able to identify the low-dimensional manifold on which the solution lies, as it is the case for the cyclic variant described in [21]. We can see that the convergence rate bounds for $\|G_k\|$ are the same in both algorithms, and the convergence in the general convex case after k_0 iterations is the same as well, although the definition of k_0 can be different and the early linear convergence conditions and rates also differ slightly. Thus, except when partial smoothness is present, the convergence behaviors of the two algorithms appear to be similar.

5 Related Work

One of the (serial, deterministic) algorithms considered in our recent paper [16] is a special case of Algorithm 1 with only one block ($N = 1$). The technique for measuring inexactness is borrowed from [16], but the extension described above, to randomized BCD and arbitrary sampling probabilities, requires novel convergence analysis.

The case in which (4) is solved exactly is discussed in [12]. This paper uses the same boundedness condition for the H_i^k as ours, and the blocks can be selected under a cyclic manner (with an arbitrary order), or a Gauss-Southwell fashion. For the cyclic variant, the convergence rate of the special case in which Q forms an upper bound of the objective improvement is further sharpened by [11, 22]. The relaxation to approximate subproblem solutions, with an inexactness criterion different from ours, is analyzed in [10]. The latter paper shows linear or sublinear convergence rates of a certain type, but the relation between the convergence rates and either the measure of inexactness or the choice of H_i^k is unclear. We note too that the cyclic ordering of blocks is inefficient in certain cases: [14] showed that the worst case of cyclic BCD is $O(N^2)$ times slower than the expected rate of randomized BCD.

The Gauss-Southwell variant discussed in [12] can be extended to the inexact case via straightforward modification of the analyses for inexact variable-metric methods (see for example [15,16,23–25]), giving results similar to what we obtain here with uniform sampling. It may be possible to utilize techniques

for single-coordinate descent in [26] to obtain better rates by considering a norm other than the Euclidean norm, as was done in [27], but such extensions are beyond the scope of the current paper.

The special case of Algorithm 2 discussed in Section 4 has received much attention in the literature. As mentioned earlier, the non-regularized case ($\psi \equiv 0$ in (1)) was first analyzed in [7] for convex and strongly convex f. That paper uses a quadratic approximation of f that is invariant over iterations, together with a fixed step size. Since it is relatively easy to solve the subproblem to optimality in the non-regularized case, inexactness is not considered. The sampling strategy of using the probability $p_i = L_i^\alpha / \sum_j L_j^\alpha$ for any $\alpha \in [0, 1]$ was analyzed in [7]. The two extreme cases of $\alpha = 0$ and $\alpha = 1$ correspond to uniform sampling and (63), respectively. The ith block update in either case is $d_i = -\nabla_i f(x)/L_i$, so we obtain from the blockwise Lipschitz continuity of ∇f that

$$\begin{aligned}\mathbb{E}_i \left[f\left(x + U_i d_i\right) - f\left(x\right)\right] &\leq \sum_i p_i f\left(x\right) - \frac{p_i}{2L_i} \left\|\nabla_i f(x)\right\|^2 - f\left(x\right) \\ &\leq -\min_i \frac{p_i}{2L_i} \left\|\nabla f(x)\right\|^2 .\end{aligned}$$

This bound suggests that if we use $p_i = 1/N$, the complexity will be related to $NL_{\max}$, whereas when p_i is proportional to L_i, the complexity is related to the smaller quantity NL_{avg}, consistent with our discussion in Section 4. The case in which ψ is an indicator function of a convex set is also analyzed in [7], with an extension in [28] to convex and strongly convex regularized problems, but both these analyses are limited to Algorithm 2 with uniform sampling. The case in which f in (1) is not necessarily convex is analyzed in [20], again under uniform sampling. Our results allow broader choices of algorithm, and show that non-uniform sampling can accelerate the optimization process.

The special case of Algorithm 2 applied to the dual of convex regularized ERM, where each ψ_i is strongly convex, with non-uniform samplings for the blocks, is analyzed in [29]. Some primal-dual properties of these problems are used to derive the optimal probability distribution for the primal suboptimality. It is unclear how to generalize this analysis to other classes of problems. Our recent work [30] shows a convergence rate of $o(1/k)$ of Algorithm 2 when f is convex, under arbitrary non-uniform sampling of the blocks, and without the assumption of finite R_0^2. However, this work does not show convergence improvement for non-uniform sampling, like the improvement shown above for (63). Moreover, our earlier paper does not address the early linear convergence rates in the convex case.

He et al. [31] consider the case of adaptive probability distributions that change every iteration for sampling the coordinates or the blocks, for an algorithm slightly different from the BCD framework considered here. They show that suitable choices for adaptive probabilities may further improve the convergence. Although our framework allows for adaptive probability distributions as well, most of our convergence results are for fixed probabilities. Moreover,

most works considering adaptive probabilities do not yield an empirical advantage for the adaptive distribution that give better theoretical convergence, because updating the probabilities followed by sampling can incur an additional per-iteration cost of $O(N)$ (and a cost of $O(N^2)$ per "epoch" of n successive iterations). For high-dimensional problems, these works usually rely on heuristics to work in practice; see the discussion in [31] and the references therein.

The paper [8] describes inexact extensions of [7] to convex versions of (1). This paper uses a different inexactness criterion from ours, and their framework fixes H_i^k over all iterations, using small steps based on L_i rather than a line search. Thus, their algorithm requires knowledge of the parameters L_i. In the regularized case of $\psi \neq 0$, their algorithm is compatible only with uniform sampling. [9] allows variable H_i and backtracking line search, but under a different sampling strategy in which a predefined number of blocks is sampled at each iteration from a uniform distribution. The other difference between our algorithm and that of [9] is that their inexactness condition can be expensive to check except for special cases of ψ (see their Remark 5). Our improvements over [9] include (1) an inexactness framework that allows more general ψ, (2) nonuniform sampling that may lead to significant acceleration when additional information is available, (3) sharper convergence rates, and (4) convergence rate results for nonconvex f.

6 Efficient Implementation for Algorithm 1

An important concern in assessing the practicality of Algorithm 1 is whether the operations of partial gradient evaluation and line search can be carried out efficiently, and whether there are natural choices of the variable metrics H_i^k that can be maintained efficiently. In this section and the computational section to follow, we consider problems in which f has the form

$$f(x) = g(Ax) \tag{67}$$

for a given matrix $A \in \mathbb{R}^{\ell \times n}$ and a function $g : \mathbb{R}^\ell \to \mathbb{R}$ that is block-separable, and the evaluation of $g(z)$ costs $O(\ell)$ operations. This structure includes many problems seen in applications, including the regularized ERM problem in machine learning and its Lagrange dual. We also discuss the practicality of nonuniform sampling in this section.

One key to efficient implementation of Algorithm 1 is to maintain explicitly the matrix-vector product Ax, updating it during each step. The updates have the form

$$A(x + U_i d_i) = Ax + A_i d_i,$$

where $d_i \in \mathbb{R}^{n_i}$ is the update to the ith block and $A_i \coloneqq AU_i$ is the column submatrix of A that corresponds to this block. The partial gradient has the form

$$\nabla_i f(x) = A_i^\top \nabla g(Ax),$$

so it can be evaluated at the cost of evaluating ∇g (costs $O(\ell)$ operations as evaluating g costs $O(\ell)$) together with a matrix-vector product involving A_i.

To perform the line search in Algorithm 1, we need to evaluate $\psi_i(x_i+\alpha d_i)$ for each value of α, along with $f(x+\alpha U_i d_i) = g(Ax+\alpha A_i d_i)$. Once $A_i d_i$ has been calculated (once), the marginal cost of performing this operation for each α is the $O(\ell)$ operations needed to calculate $Ax+\alpha A_i d_i$ and the $O(\ell)$ operations needed to evaluate g.

A natural choice for the quadratic term H_i^k in subproblem (5) is the ith diagonal block of the true Hessian, which is

$$[\nabla^2 f(x)]_{ii} = A_i^\top \nabla^2 g(Ax) A_i. \tag{68}$$

(Note that the subscript is the (i,i) block, not the (i,i) entry.) The block-separability of g makes $\nabla^2 g(Ax)$ block-diagonal, and actually diagonal in many applications. Thus the matrix (68) has a particularly simple form. We note moreover that when iterative methods are used to (approximately) minimize (5), we do not need to know this matrix explicitly, but only to be able to compute matrix-vector products of the form $H_i^k v_i$ (for various v_i) efficiently. This operation can be done at the cost of two matrix-vector multiplications involving A_i, together with the (typically $O(\ell)$) cost of multiplying by $\nabla^2 g(Ax)$.

There are two concerns in implementing non-uniform samplings such as the Lipschitz sampling. The first is simply the cost of sampling from a non-uniform distribution, for which a naive method may cost $O(N)$ operations. Fortunately, there are efficient methods such as that proposed in [32] for non-uniform samplings such that given a fixed distribution, after a $O(N)$ cost of initialization, each run costs the same as sampling two points uniformly randomly. Note that the overhead incurred in changing probability distributions $\{p_i^k\}$ between iterations can nullify any efficiencies gained; the sampling can then become the bottleneck especially when the update itself is inexpensive. For completeness, we give details of our implementation of non-uniform sampling in the Appendix.

The second concern is that the cost per iteration is different under different sampling strategies. Especially when the data are sparse, the value of L_i may be positively correlated to the density of the corresponding data point. In this case, sampling according to L_i may increase the cost per iteration significantly. However, if one can estimate each norm $\|H_i\|$, the step sizes, and the cost of updating different blocks in advance, it is not hard to compare the expected cost increase and the expected convergence improvement to decide if non-uniform sampling should be considered. When such information is unavailable or hard to obtain, uniform sampling can still be used.

7 Computational Results

This section reports on the empirical performance of Algorithms 1-3 on three sets of experiments. In the first set of problems, which are convex, we compare uniform sampling and the Lipschitz sampling for the traditional randomized

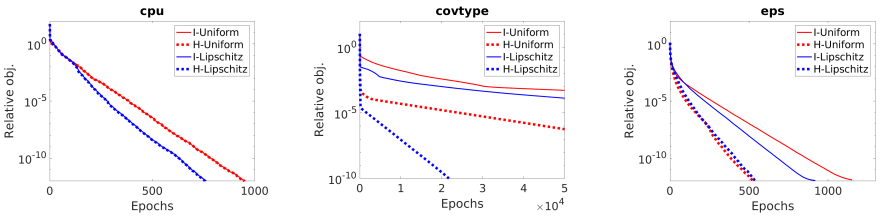

Fig. 1: Comparison of different sampling strategies using fixed step sizes in terms of epochs. The prefix "H" refers to the choice $H_i = L_i I$, while "I" means $H = I$.

Table 1: Data sets used in the LASSO problem.

Data set	#instances	n	$L_{\max}/L_{\text{avg}}$	C
cpusmall_scale	$8,192$	12	1.29	.001
covtype.binary.scale	$581,012$	54	8.58	.001
epsilon_normalized	$400,000$	$2,000$	5.49	.2

BCD approaches discussed in Section 4, on both Algorithms 2 and 3. In the second set of experiments, also on convex objectives, we investigate a version of Algorithm 1 in which the ith diagonal block of the true generalized Hessian is used as H_i^k in (5). In both experiments, we report the relative objective value difference to the optimum, defined as $(F(x) - F^*)/F^*$, where F^* is obtained by running our algorithm with a tight termination condition. The third set of experiment considers a nonconvex problem, and therefore the algorithms are not guaranteed to find F^*. We report the measure $\|G_k\|^2$ of stationarity instead.

7.1 **Traditional Coordinate Descent**

We first illustrate the speedup of Lipschitz sampling over uniform sampling using the simple LASSO problem [33]

$$\min_{x\in\mathbb{R}^n} \quad \frac{C}{2}\sum_{i=1}^{l}(a_i^\top x - b_i)^2 + \|x\|_1, \tag{69}$$

where $(a_i, b_i) \in \mathbb{R}^n \times \mathbb{R}, i = 1, \ldots, l$, are the training data points and $C > 0$ is a parameter to balance the two terms. In the subproblem, each "block" consists of only one coordinate and therefore $n = N$. Note that the corresponding subproblem (4) has a closed-form solution when H is a multiple of identity, so we have $\eta = 0$ in (6).

Our goal here is not to propose an optimal BCD algorithm for (69) but merely to compare sampling strategies. We choose C so that among the final solutions generated by different variants we compare, the sparsest one has a sparsity of around 50%. Statistics of the data sets and the value of C are listed in Table 1. We test both Algorithms 2 and 3, and both uniform and Lipschitz samplings. We present convergence in terms of epochs, where each

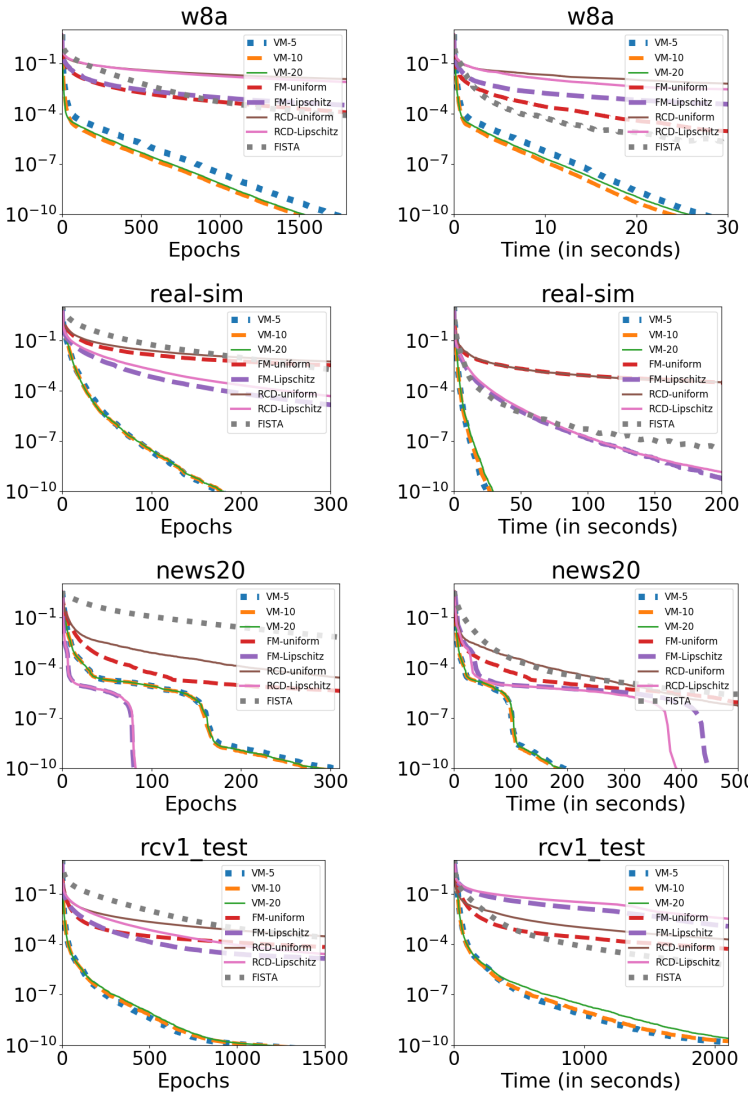

Fig. 2: Comparison of fixed and variable quadratic terms for solving (70) with $C = 1$. Left column: epochs; Right column: running time.

epoch is a group of N successive iterations. Most of the results in Figure 1 show a clear advantage for Lipschitz sampling, consistent with our convergence analysis. The only exception is Algorithm 3 on the data set epsilon, where the two sampling strategies give similar performance. The major reason for this exception is that different sampling strategies identified the correct active set at different stages, and these differences affect the overall convergence behavior.

We also observe that because of the effects of active set identification, Algorithm 2 often outperforms Algorithm 3, but when n is small (as in cpusmall_scale), the two perform quite similarly. Early fast convergence can be observed empirically in all examples, as suggested by Theorem 3.1.

7.2 **Variable Metric Approach**

We show the advantage of using variable quadratic terms H_i^k in (5), in comparison with a fixed term. For this purpose, we consider a group-LASSO regularized squared-hinge loss problem defined by

$$\min_{x\in\mathbb{R}^n} \quad C\sum_{i=1}^{l}\max\left\{1-b_i a_i^\top x, 0\right\}^2 + \sum_{i=1}^{\lceil n/5\rceil}\sqrt{\sum_{j=1}^{\min\{5,n-5(i-1)\}} x_{5(i-1)+j}^2}, \quad (70)$$

where $(a_i, b_i) \in \mathbb{R}^n \times \{-1, 1\}, i = 1, \ldots, l$ are the training data points and $C > 0$ is a parameter to balance the two terms. Each set of five consecutive coordinates is grouped into a single block to form the regularizer. We compare the following algorithms.

- VM-t: our variable metric approach of Algorithm 1, with H being the generalized Hessian with $10^{-10}I$ added to ensure that the condition (23) is satisfied with $m_i > 0$. We use uniform sampling of the blocks and the SpaRSA approach of [34] to solve the subproblem, with $t \in \{5, 10, 20\}$ being the number of SpaRSA iterations applied to each subproblem.
- FM: the fixed metric approach considered in [8]. We use a global upper bound of the generalized Hessian as the fixed metric. As H_i are precomputed, we consider both uniform sampling and the sampling scheme of (41) using the largest eigenvalue of each H_i. We solve each subproblem inexactly using 10 SpaRSA iterations.
- RCD: Algorithm 2 with $\eta = 0$. We use both Lipschitz sampling (63) and uniform sampling.
- FISTA [35]: the accelerated proximal gradient approach that does not exploit the block-separable nature of the regularization term.

The FISTA approach is included as a comparison with state of the art for problems without block separability.

We consider the data sets in Table 2, obtained from the LIBSVM website,[4] and set $C = 1$ in (70). Results are shown in Figure 2. Note that the varying number of SpaRSA iterations used in VM-5, VM-10, and VM-20 have little impact on the convergence in terms of both epochs and running time, and that all these variants are significantly faster than their competitors, showing the advantages of solving the subproblems with variable metrics inexactly. For news20, Lipschitz sampling with both the fixed metric approach and Algorithm 2 are the fastest in terms of epochs, but the running times are much

[4] `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`.

Table 2: Data sets used in the group-LASSO regularization experiment.

Data set	#instances	n
w8a	49,749	300
real-sim	72,309	20,958
news20	19,996	1,355,191
rcv1_test	677,399	47,236

slower than the proposed variable metric approach. The reason is that news20 is a very sparse data set, with the size of the Lipschitz constants highly correlated to the density of each coordinate, making the average number of nonzero elements processed per epoch much higher when Lipschitz sampling is considered.

We also observe that for both the fixed metric approach and Algorithm 2, Lipschitz sampling is always faster than uniform sampling in terms of epochs, confirming our analysis. But in terms of running time, the situation may differ. We also observe that FISTA performs better in running time than in epochs, mainly because it updates the variables and the gradient less frequently, and its memory access is always sequential and therefore faster. Finally, we observe the early linear convergence in the variable metric approach, the fixed metric approach, and Algorithm 2, verifying the result in Theorem 3.1 empirically.

We also notice that although the variable metric approach is the only one that requires line search, it is still the fastest in terms of running time, showing that line search does not occupy a significant portion of the running time.

7.3 **A Nonconvex Problem**

We now consider a nonconvex problem. Following the setting of [36], we consider the smooth biweight loss by [37]:

$$f(x) = C \sum_{i=1}^{l} \phi\left(a_i^\top x - b_i\right), \text{ where } \phi(z) = \frac{z^2}{1+z^2}, \tag{71}$$

for some $C > 0$, with $(a_i, b_i) \in \mathbb{R}^n \times \mathbb{R}$ for $i = 1, \ldots, l$. Through simple calculation, we can see that the Hessian of ϕ is

$$\phi''(z) = -\frac{2\left(3z^2-1\right)}{(z^2+1)^3}.$$

Its value lies in $[-0.5, 2]$, showing that f is nonconvex. For the regularization term, we consider both the ℓ_1 norm used in the first set of experiments and the group-LASSO regularization used in the second set of experiments.

For the ℓ_1-regularized problem, we compare different sampling strategies of RCD. In the previous results, Algorithm 2 tends to perform better than Algorithm 3, so we apply only the former in this experiment. As this nonconvex problem is harder than LASSO, we consider the first two smaller data sets in

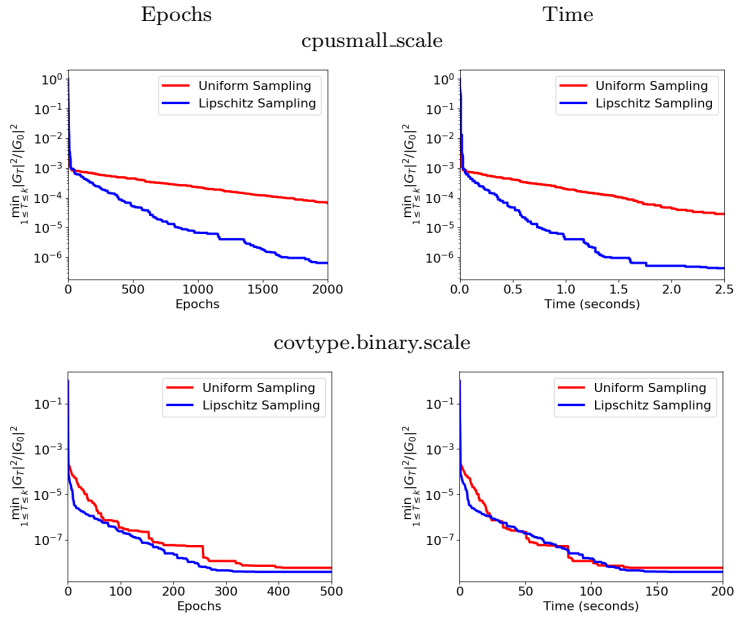

Fig. 3: Comparison of different sampling strategies on the nonconvex biweight problem with ℓ_1 regularization.

Table 1. Results are shown in Figure 3. We see that as predicted by our theory, sampling according to (63) yields faster convergence than uniform sampling.

For the group-LASSO-regularized part, different from the previous experiment, we do not include FISTA in our comparison because it is not applicable to nonconvex problems. The FM approach obtains the global upper bound for the Hessian through using the upper bound 2 for $\phi''(a_i^\top x - b_i)$ for all i. For the VM approach, the Hessian block may be indefinite so we obtain H_i^k by adding a multiple of identity as needed to make it positive definite. In particular, we compute the eigenvalues of the Hessian block, and when the smallest eigenvalue is smaller than 10^{-10}, we add a multiple of identity to H_i^k to make the smallest eigenvalue exactly 10^{-10}, and otherwise we do not modify H_i^k. Note that since the size of each H_i^k is at most 5×5, computing its eigenvalues is cheap. We conduct the comparison using the first three data sets in Table 2. The comparison between the variable metric approach and the fixed metric approach with different samplings is shown in Figures 4. All approaches use 10 SpaRSA iterations for each subproblem. On all three data sets, the variable metric approach converges faster than the fixed metric approach with uniform sampling. As in the previous experiment, Lipschitz sampling has much better convergence on news20 in terms of epochs. An interesting difference is that Lipschitz sampling does not work well on the other two data sets. A further examination indicates that on those two data sets, the Lipschitz sampling

strategy identifies the correct sparsity pattern much later, possibly affecting the convergence behavior.

With regard to running time, the fixed metric approach with uniform sampling tends to be the fastest. The reason is that on this nonconvex problem, the convergence advantage of the variable metric approach is not significant enough to counterbalance the higher per-iteration cost. The less strong convergence advantage is likely from the damping term being added to the variable metric. There are various ways to modify the indefinite Hessian to make it positive definite, but so far there is no conclusion which approach is most effective. Comparing various Hessian modification strategies is an interesting future work.

This set of experiments shows that when we are dealing with nonconvex problems, variable metric approach based on the Hessian might be less effective because of the indefiniteness of the Hessian. On the other hand, Lipschitz sampling has better convergence speed on three out of the five data sets, indicating that when the sparsity pattern identification is not a problem, Lipschitz sampling has better convergence speed.

8 Conclusions

Starting with a strategy for regularized optimization using regularized quadratic subproblems with variable quadratic terms, we have described a stochastic block-coordinate-descent scheme that is well suited to large-scale problems with general structure. We provide detailed iteration complexity analysis, allowing for arbitrary sampling schemes. A special case of our theory extends known results for a sampling strategy based on blockwise Lipschitz constants for randomized gradient-coordinate descent from the non-regularized setting to the regularized problem (1) and from convex problems to nonconvex problems. Computational experiments show empirical advantages for our variable metric approaches.

Acknowledgements This work was supported by NSF awards 1447449, 1628384, 1634579, and 1740707; Subcontracts 3F-30222 and 8F-30039 from Argonne National Laboratory; and Award N660011824020 from the DARPA Lagrange Program. This work was performed mostly when Ching-pei Lee was at the University of Wisconsin-Madison.

References

1. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **68**(1), 49–67 (2006)
2. Meier, L., Van De Geer, S., Bühlmann, P.: The group LASSO for logistic regression. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **70**(1), 53–71 (2008)
3. Crammer, K., Singer, Y.: On the learnability and design of output codes for multiclass problems. Machine Learning **47**(2–3), 201–233 (2002)

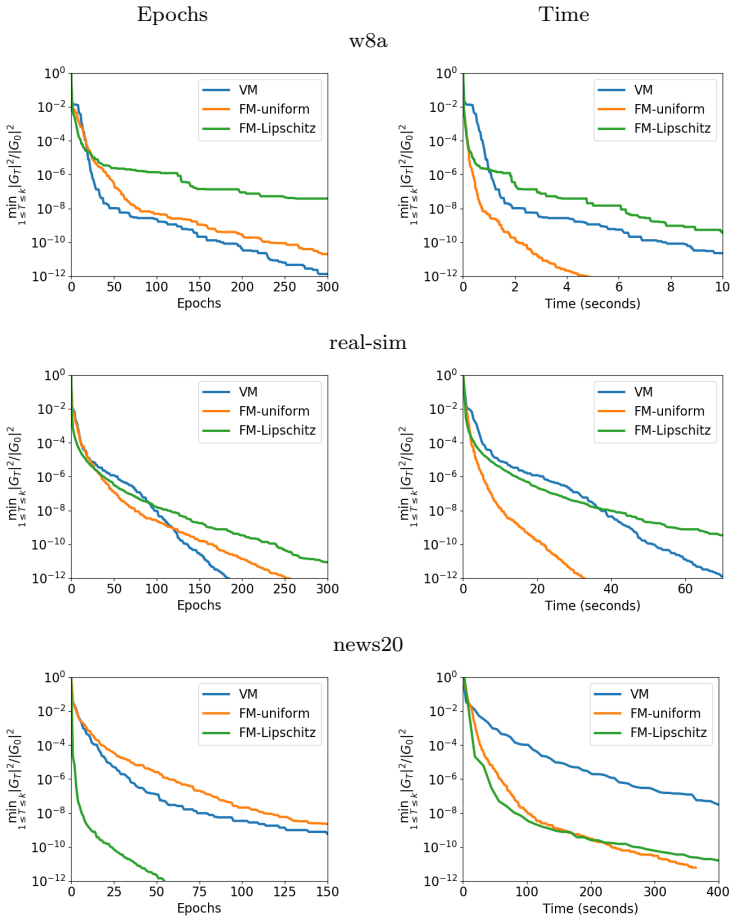

Fig. 4: Comparison of fixed and variable quadratic terms for solving the biweight loss problem with group-LASSO regularization. The y-axis is $\min_{0\le k\le T} \|G_k\|^2/\|G_0\|^2$.

4. Lebanon, G., Lafferty, J.D.: Boosting and maximum likelihood for exponential models. In: Advances in neural information processing systems, pp. 447–454 (2002)
5. Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y.: Large margin methods for structured and interdependent output variables. Journal of machine learning research **6**(Sep), 1453–1484 (2005)
6. Lee, C.p., Lin, C.J.: A study on L2-loss (squared hinge-loss) multi-class SVM. Neural Computation **25**(5), 1302–1323 (2013)
7. Nesterov, Y.: Efficiency of coordinate descent methods on huge-scale optimization problems. SIAM Journal on Optimization **22**(2), 341–362 (2012)
8. Tappenden, R., Richtárik, P., Gondzio, J.: Inexact coordinate descent: complexity and preconditioning. Journal of Optimization Theory and Applications **170**(1), 144–176 (2016)
9. Fountoulakis, K., Tappenden, R.: A flexible coordinate descent method. Computational Optimization and Applications **70**(2), 351–394 (2018)
10. Chouzenoux, E., Pesquet, J.C., Repetti, A.: A block coordinate variable metric forward–backward algorithm. Journal of Global Optimization **66**(3), 457–485 (2016)

11. Sun, R., Hong, M.: Improved iteration complexity bounds of cyclic block coordinate descent for convex problems. In: Advances in Neural Information Processing Systems, pp. 1306–1314 (2015)
12. Tseng, P., Yun, S.: A coordinate gradient descent method for nonsmooth separable minimization. Mathematical Programming **117**(1), 387–423 (2009)
13. Yun, S.: On the iteration complexity of cyclic coordinate gradient descent methods. SIAM Journal on Optimization **24**(3), 1567–1580 (2014)
14. Sun, R., Ye, Y.: Worst-case complexity of cyclic coordinate descent: $O(n^2)$ gap with randomized version. Mathematical Programming pp. 1–34 (2019). Online first.
15. Bonettini, S., Loris, I., Porta, F., Prato, M.: Variable metric inexact line-search-based methods for nonsmooth optimization. SIAM Journal on Optimization **26**(2), 891–921 (2016)
16. Lee, C.p., Wright, S.J.: Inexact successive quadratic approximation for regularized optimization. Computational Optimization and Applications **72**, 641–674 (2019)
17. Hiriart-Urruty, J.B., Strodiot, J.J., Nguyen, V.H.: Generalized hessian matrix and second-order optimality conditions for problems with $C^{1,1}$ data. Applied Mathematics & Optimization **11**(1), 43–56 (1984)
18. Lee, C.p., Wright, S.J.: Random permutations fix a worst case for cyclic coordinate descent. IMA Journal of Numerical Analysis **39**(3), 1246–1275 (2019)
19. Wright, S.J., Lee, C.p.: Analyzing random permutations for cyclic coordinate descent. Mathematics of Computation (2020). To appear..
20. Patrascu, A., Necoara, I.: Efficient random coordinate descent algorithms for large-scale structured nonconvex optimization. Journal of Global Optimization **61**(1), 19–46 (2015)
21. Wright, S.J.: Accelerated block-coordinate relaxation for regularized optimization. SIAM Journal on Optimization **22**(1), 159–186 (2012)
22. Li, X., Zhao, T., Arora, R., Liu, H., Hong, M.: On faster convergence of cyclic block coordinate descent-type methods for strongly convex minimization. Journal of Machine Learning Research **18**(1), 6741–6764 (2017)
23. Scheinberg, K., Tang, X.: Practical inexact proximal quasi-Newton method with global complexity analysis. Mathematical Programming **160**(1-2), 495–529 (2016)
24. Ghanbari, H., Scheinberg, K.: Proximal quasi-Newton methods for regularized convex optimization with linear and accelerated sublinear convergence rates. Computational Optimization and Applications **69**(3), 597–627 (2018)
25. Peng, W., Zhang, H., Zhang, X.: Global complexity analysis of inexact successive quadratic approximation methods for regularized optimization under mild assumptions. Tech. Rep. arXiv:1808.04291, Department of Mathematics, University of Defense Technology, China (2018)
26. Nutini, J., Schmidt, M., Laradji, I., Friedlander, M., Koepke, H.: Coordinate descent converges faster with the gauss-southwell rule than random selection. In: International Conference on Machine Learning, pp. 1632–1641 (2015)
27. Nutini, J., Laradji, I., Schmidt, M.: Let's make block coordinate descent go fast: Faster greedy rules, message-passing, active-set complexity, and superlinear convergence. Tech. rep. (2017). ArXiv:1712.08859
28. Lu, Z., Xiao, L.: On the complexity analysis of randomized block-coordinate descent methods. Mathematical Programming **152**(1-2), 615–642 (2015)
29. Zhao, P., Zhang, T.: Stochastic optimization with importance sampling for regularized loss minimization. In: Proceedings of the 32nd International Conference on Machine Learning (2015)
30. Lee, C.p., Wright, S.J.: First-order algorithms converge faster than $O(1/k)$ on convex problems. In: Proceedings of the 36th International Conference on Machine Learning (2019)
31. He, X., Tappenden, R., Takac, M.: Dual free adaptive minibatch sdca for empirical risk minimization. Frontiers in Applied Mathematics and Statistics **4**, 33 (2018)
32. Walker, A.J.: An efficient method for generating discrete random variables with general distributions. ACM Transactions on Mathematical Software **3**(3), 253–256 (1977)
33. Tibshirani, R.: Regression shrinkage and selection via the LASSO. Journal of the Royal Statistical Society Series B **58**, 267–288 (1996)
34. Wright, S.J., Nowak, R.D., Figueiredo, M.A.T.: Sparse reconstruction by separable approximation. IEEE Transactions on Signal Processing **57**(7), 2479–2493 (2009)

35. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM Journal on Imaging Sciences **2**(1), 183–202 (2009)
36. Carmon, Y., Duchi, J.C., Hinder, O., Sidford, A.: Convex until proven guilty: Dimension-free acceleration of gradient descent on non-convex functions. In: International Conference on Machine Learning, pp. 654–663. JMLR. org (2017)
37. Beaton, A.E., Tukey, J.W.: The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. Technometrics **16**(2), 147–185 (1974)

Appendix: Efficient Implementation of Non-uniform Sampling

We describe our implementation of non-uniform sampling. The $O(N)$ initialization step is described in Algorithm 4. After the initialization, each time to sample a point from the given probability distribution, it takes only 2 independent uniform sampling as described in Algorithm 5.

Algorithm 4 Initialization for non-uniform sampling

Given a probability distribution $p_1, \ldots, p_N > 0$;
$i \leftarrow 1$;
Construct $U \leftarrow \{u \mid p_u > 1/N\}$, $L \leftarrow \{l \mid p_l \leq 1/N\}$;
while $L \neq \phi$ **do**
  Pop an element l from L;
  Pop an element u from U;
  $\text{upper}_i \leftarrow u$, $\text{lower}_i \leftarrow l$, $\text{threshold}_i \leftarrow p_l/(1/N)$;
  $p_u \leftarrow p_u - (1/N - p_l)$;
  if $p_u > 1/N$ **then**
    $U \leftarrow U \cup \{u\}$;
  else
    $L \leftarrow L \cup \{u\}$;
  end if
  $i \leftarrow i + 1$;
end while

Algorithm 5 Non-uniform sampling after initialization by Algorithm 4

Sample i and j independently and uniformly from $\{1, \ldots, N\}$;
if $j/N \geq \text{threshold}_i$ **then**
  Output upper_i;
else
  Output lower_i;
end if