Citation
Yu, T.T. and Liu, X.W. and Dai, Y.H. and Sun, J. 2022. A Mini-Batch Proximal Stochastic Recursive Gradient Algorithm with Diagonal Barzilai–Borwein Stepsize. Journal of the Operations Research Society of China.11: pp. 277-307.
http://doi.org/10.1007/s40305-022-00436-2

A mini-batch proximal stochastic recursive gradient algorithm with diagonal Barzilai-Borwein stepsize

Tengteng Yu[1,3], Xin-Wei Liu[2*], Yu-Hong Dai[3] and Jie Sun[2,4]

[1]School of Artificial Intelligence, Hebei University of Technology, Xiping Road No. 5340, Beichen District, 300401, Tianjin, China.

[2]Institute of Mathematics, Hebei University of Technology, Xiping Road No. 5340, Beichen District, 300401, Tianjin, China.

[3]LSEC, ICMSEC, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Zhongguancun East Road No. 55, Haidian District, 100190, Beijing, China.

[4]School of Business, National University of Singapore, 15 Kent Ridge Drive, 119245, Singapore, Singapore.

*Corresponding author(s). E-mail(s): mathlxw@hebut.edu.cn;
Contributing authors: ytt2021@lsec.cc.ac.cn; dyh@lsec.cc.ac.cn; jsun@nus.edu.sg;

Abstract

Many machine learning problems can be formulated as minimizing the sum of a function and a nonsmooth regularization term. Proximal stochastic gradient methods are popular for solving such composition optimization problems. We propose a mini-batch proximal stochastic recursive gradient algorithm SRG-DBB, which incorporates the diagonal Barzilai-Borwein (DBB) stepsize strategy to capture the local geometry of the problem. The linear convergence and complexity of SRG-DBB is analyzed for strongly convex functions. We further establish the linear convergence of SRG-DBB under the non-strong convexity condition. Moreover, it is proved that SRG-DBB converges sublinearly in the convex case. Numerical experiments on standard data sets indicate that the performance of SRG-DBB is better than or comparable to the proximal stochastic recursive gradient algorithm with best-tuned scalar stepsizes or BB stepsizes. Furthermore, SRG-DBB is superior to some advanced mini-batch proximal stochastic gradient methods.

Keywords: Stochastic recursive gradient, proximal gradient algorithm, Barzilai-Borwein method, composite optimization

1 Introduction

We are interested in solving the following problem

$$\min_{w \in \mathbb{R}^d} P(w) = F(w) + R(w), \tag{1}$$

where $F(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w)$, each component function $f_i(w) : \mathbb{R}^d \to \mathbb{R}$ is smooth and convex, and $R(w) : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a relatively simple proper convex function (sometimes referred to as a regularization) but can be non-differentiable. We focus on the case where the sample size n is extremely large, and the scaled proximal operator of $R(w)$ could be computed efficiently.

The formulation (1) appears across a broad range of applications in machine learning [1–3], statistics [4], matrix completion [5], neural networks [6–8], etc. One important instance is the regularized empirical risk minimization (ERM) [1, 4, 9], which involves a collection of training examples $\{(a_i, b_i)\}_{i=1}^{n}$, where $a_i \in \mathbb{R}^d$ is a feature vector and $b_i \in \mathbb{R}$ is the desired response. With the component functions $f_i(w) = \frac{1}{2}(b_i - a_i^T w)^2$, Lasso, ridge regression and elastic net employ the regularization terms $R(w) = \lambda_1\|w\|_1$, $R(w) = \frac{\lambda_2}{2}\|w\|_2^2$ and $R(w) = \lambda_1\|w\|_1 + \frac{\lambda_2}{2}\|w\|_2^2$, respectively, where λ_1 and λ_2 are nonnegative regularization parameters. When considering binary classification problems, one frequently used component function is the logistic loss $f_i(w) = \log(1+\exp(-b_i a_i^T w))$ and $R(w)$ can be any of the aforementioned regularization terms.

Popular methods for solving optimization problems (1) in large-scale setting relay on proximal stochastic approaches. Motivated by the seminal work of Robbins and Monro [10], a proximal stochastic gradient descent (Prox-SGD) method has been developed, which chooses $i_k \in \{1, 2, \ldots, n\}$ uniformly at random and takes the update

$$w_{k+1} = \mathrm{prox}_R^{\eta_k^{-1}\mathbb{I}_d}\big(w_k - \eta_k \nabla f_{i_k}(w_k)\big), \tag{2}$$

where $\eta_k > 0$ is the stepsize (a.k.a. learning rate), $\mathbb{I}_d \in \mathbb{R}^{d\times d}$ is the identity matrix, and $\mathrm{prox}_R^A(w)$ is defined as

$$\mathrm{prox}_R^A(w) = \arg\min_{y \in \mathbb{R}^d} \left\{ \frac{1}{2}\|y - w\|_A^2 + R(y) \right\}. \tag{3}$$

However, due to the large variance of the stochastic gradient introduced by random sampling, Prox-SGD only enjoys a sublinear convergence rate for strongly convex functions. Starting from several prevalent variance reduced stochastic gradient methods such as stochastic average gradient (SAG) [11, 12], stochastic variance reduced gradient (SVRG) [13], and stochastic recursive gradient algorithm (SARAH) [14], some efficient proximal stochastic methods have been developed for solving composite problems. In [15], Xiao and Zhang proposed a proximal variant of SVRG, called Prox-SVRG and proved its linear convergence rate for strongly convex problems. By combining the mini-batch scheme with semi-stochastic gradient descent methd (S2GD)[16], Konečný et al. [17] developed the mS2GD method that achieves better convergence rate and practical performance than Prox-SVRG. A proximal version of SARAH can be found in [18].

Since the stepsize has an important influence on the performance of stochastic gradient methods, many researchers are devoted to designing more efficient scheme of stepsizes. For classical SGD, one frequently employed stepsize strategy in practical computation is

$$\sum_{k=1}^{\infty} \eta_k = \infty \quad \text{and} \quad \sum_{k=1}^{\infty} \eta_k^2 < \infty.$$

However, such a choice often yields sublinear convergence of SGD, see [1] for example. In recent years, using the Barzilai-Borwein (BB) method [19] to automatically calculate stepsizes for SGD and its variants has attracted more and more attention. One great advantage of BB stepsize is that it is able to capture hidden second-order information and is not sensitive to the choice of initial stepsizes, which makes it very promising in practice. See [19–21] and references therein for more details about BB-like methods. One pioneer work in this line is due to Tan et al. [22], who proposed to incorporate the BB stepsize with SGD and SVRG, and got the SGD-BB and SVRG-BB methods. By combining SARAH with the BB method and importance sampling strategy, Liu et al. [23] suggested the SARAH-I-BB method. To solve problem (1), Yu et al. [24] developed a mini-batch proximal stochastic recursive gradient algorithm that incorporates a trust-region-like scheme and BB stepsizes. Inspired by the adaptive metric selection strategy in [25], the authors proposed a new diagonal BB stepsize to update the metric for Prox-SVRG and devised a VM-SVRG method [26].

In this paper, motivated by the diagonal BB stepsize startegy and the success of SARAH in solving problem (1), we propose a mini-batch proximal stochastic recursive gradient method, named SRG-DBB. We present the convergence analysis of SRG-DBB under different conditions, which shows that it converges linearly for strongly convex and non-strongly convex functions. The sublinear convergence of SRG-DBB in the convex case is given. Numerical results for solving regularized logistic regression problems on standard data sets show that the performance of SRG-DBB is better than or comparable to proximal SARAH with best-tuned stepsizes and the proximal variant of

SARAH-BB with different initial stepsizes. Further comparisons between SRG-DBB and some advanced mini-batch proximal stochastic gradient methods demonstrate the efficiency of SRG-DBB.

The rest of this paper is organized as follows. In Section 2 we propose our SRG-DBB method. In Section 3 we prove that SRG-DBB enjoys a linear convergence rate under strong convexity and non-strong convexity conditions, and converges sublinearly under convex condition. Numerical experiments are reported in Section 4. Finally, we draw some conclusions in Section 5.

2 The SRG-DBB method

A formal description of SRG-DBB is given in Algorithm 1.

Algorithm 1 SRG-DBB($\tilde{w}^0, m, b, U_0$)

Input: update frequency m (max # of stochastic steps per outer loop), initial point $\tilde{w}^0 \in \mathbb{R}^d$, initial matrix $U_0 = \eta_0 \mathbb{I}_d$, mini-batch size $b \in \{1, 2, \ldots, n\}$, probability $\Omega = \{q_1, q_2, \ldots, q_n\}$

for $k = 0, 1, \ldots, K-1$ **do**

    $w_1^k = w_0^k = \tilde{w}^k$

    $v_0^k = \nabla F(w_0^k)$

    Choose $t_k \in \{1, 2, \ldots, m\}$ uniformly at random

    for $t = 1, \ldots, t_k$ **do**

        Choose mini-batch $I_t \subseteq \{1, 2, \ldots, n\}$ of size b, where each $i \in I_t$ is chosen from $\{1, 2, \ldots, n\}$ randomly according to Ω

$$v_t^k = \frac{1}{b} \sum_{i \in I_t} \left[(\nabla f_i(w_t^k) - \nabla f_i(w_{t-1}^k))/(q_i n) \right] + v_{t-1}^k \tag{4}$$

        $w_{t+1}^k = \mathrm{prox}_R^{U_k^{-1}}(w_t^k - U_k v_t^k)$

    end for

    $\tilde{w}^{k+1} = w_{t_k+1}^k$

    Compute U_k from (6)

end for

Output: Iterate w_a chosen uniformly at random from $\{\{w_t^k\}_{t=1}^{t_k}\}_{k=0}^{K-1}$

Note that, when $U_k = \alpha_k \mathbb{I}_d$ with α_k being a scalar stepsize, Algorithm 1 is a proximal version of SARAH [14]. It transforms to the stochastic proximal quasi-Newton method for $U_k \approx (\nabla^2 F(w_t^k))^{-1}$ [27, 28].

We will use a diagonal matrix U_k to estimate the second-order information of $F(w)$. In particular, we employ the approach in [26] to compute U_k,

$$\begin{aligned} \min_{u \in \mathbb{R}^d} \quad & \|s_k - U y_k\|_2^2 + \omega \|U - U_{k-1}\|_F^2 \\ \text{s.t.} \quad & \alpha_k^2 \mathbb{I}_d \preceq U \preceq \alpha_k^1 \mathbb{I}_d, \end{aligned} \tag{5}$$

$$U = \mathrm{Diag}(u),$$

where $s_k = \tilde{w}^k - \tilde{w}^{k-1}$, $y_k = \nabla F(\tilde{w}^k) - \nabla F(\tilde{w}^{k-1})$, $\|\cdot\|_F$ is the Frobenius norm and $0 < \alpha_k^2 \leq \alpha_k^1$ are two stepsizes given by users. Clearly, the solution U_k of (5) satisfies the secant equation $s_k = U_k y_k$ in the sense of least squares and is close to the previous matrix U_{k-1} where the closeness is controlled by the hyperparameter $\omega > 0$. In this way, U_k can capture the geometry of the inverse Hessian of $F(w)$, which is different from the one in [25].

Denote $u_k = [u_k^{(1)}, u_k^{(2)}, \ldots, u_k^{(d)}] \in \mathbb{R}^d$ and $U_k = \mathrm{Diag}(u_k) \in \mathbb{R}^{d\times d}$. Problem (5) has a closed-form solution given by

$$u_k^{(j)} = \begin{cases} \alpha_k^2, & \frac{s_k^{(j)} y_k^{(i)} + \omega u_{k-1}^{(j)}}{(y_k^{(j)})^2 + \omega} < \alpha_k^2; \\ \alpha_k^1, & \frac{s_k^{(j)} y_k^{(j)} + \omega u_{k-1}^{(j)}}{(y_k^{(j)})^2 + \omega} > \alpha_k^1; \\ \frac{s_k^{(j)} y_k^{(j)} + \omega u_{k-1}^{(j)}}{(y_k^{(j)})^2 + \omega}, & \text{otherwise.} \end{cases} \tag{6}$$

where $s_k^{(j)}$ and $y_k^{(j)}$ are the j-th elements of s_k and y_k, respectively.

As mentioned before, the BB stepsize is suitable for SGD and its variants. We would like to employ BB-like stepsizes for α_k^1 and α_k^2. Since at most m biased gradient estimators are added to w_0^k for getting w_m^k in the inner loop, we employ the following stepsizes

$$\alpha_k^1 = \frac{2}{m} \cdot \frac{\|s_k\|_2}{s_k^T y_k} \tag{7}$$

and

$$\alpha_k^2 = \frac{2}{m} \cdot \frac{s_k^T y_k}{\|y_k\|_2^2}, \tag{8}$$

where α_k^1 and α_k^2 are two variant of the long BB stepsize $\alpha_k^{BB1} = \frac{\|s_k\|_2}{s_k^T y_k}$ and the short BB stepsize $\alpha_k^{BB2} = \frac{s_k^T y_k}{\|y_k\|_2^2}$ in [19], respectively. In order to guarantee the boundedness of $u_k^{(j)}$ $(k = 0, 1, \ldots, K-1;\ j = 1, 2, \ldots, d)$, we project them into the interval $[\underline{\alpha}, \overline{\alpha}]$ when the objective function is not strongly convex. Here, $0 < \underline{\alpha} \leq \overline{\alpha}$ are two given constants.

We mention that v_t^k in Algorithm 1 is a biased estimate of the full gradient $\nabla F(w_t^k)$, which is the same as SARAH [14] but different from SGD and SVRG types of methods [13, 15]. In fact, it is easy to see that conditioned on $\mathcal{F}_t$, the expectation of v_t^k with respect to I_t is

$$\begin{aligned} \mathbb{E}_{I_t}[v_t^k|\mathcal{F}_t] &= \sum_{i=1}^{n} \frac{\nabla f_i(w_t^k) - \nabla f_i(w_{t-1}^k)}{q_i n} \cdot q_i + v_{t-1}^k \\ &= \nabla F(w_t^k) - \nabla F(w_{t-1}^k) + v_{t-1}^k, \end{aligned}$$

where $\mathcal{F}_t = \sigma(w_0^k, I_1, I_2, \ldots, I_{t-1})$ is the σ-algebra generated by $w_0^k, I_1, I_2, \ldots, I_{t-1}$ and $\mathcal{F}_0 = \mathcal{F}_1 = \sigma(w_0^k)$. As will be seen in Theorems 1 and 2, the simple recursive framework for updating v_t^k yields a non-increasing property and a linear convergence of the inner loop of our SRG-DBB method, which does not hold for Prox-SVRG and mS2GD.

Let $\tilde{\mathbb{E}}[\cdot]$ be the expectation with respect to all random variables. That is, $\tilde{\mathbb{E}}[v_t^k] = \mathbb{E}_{I_1}\mathbb{E}_{I_2}\ldots\mathbb{E}_{I_{t-1}}[v_t^k]$. When taking total expectation and employing the fact $v_0^k = \nabla F(w_0^k)$, it follows that $\tilde{\mathbb{E}}[v_1^k] = \tilde{\mathbb{E}}[\nabla F(w_1^k)] - \tilde{\mathbb{E}}[\nabla F(w_0^k)] + \tilde{\mathbb{E}}[v_0^k] = \tilde{\mathbb{E}}[\nabla F(w_1^k)]$. By induction, we obtain

$$\tilde{\mathbb{E}}[v_t^k] = \tilde{\mathbb{E}}[\nabla F(w_t^k)]. \tag{9}$$

3 Convergence analysis

In order to establish convergence of SRG-DBB in different cases, we make the following two blanket assumptions.

Assumption 1 The regularization $R(w) : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is a lower semi-continuous and convex function. However, it can be non-differentiable. Its effective domain, $\mathrm{dom}(R) = \{w \in \mathbb{R}^d | R(w) < +\infty\}$, is closed.

Assumption 2 Each component function $f_i(w) : \mathbb{R}^d \to \mathbb{R}$ is convex and L_i-smooth, that is, there exists $L_i > 0$ such that

$$\|\nabla f_i(w) - \nabla f_i(v)\|_2 \le L_i \|w - v\|_2, \quad \forall w, v \in \mathrm{dom}(R).$$

Let $L = \frac{1}{n}\sum_{i=1}^n L_i$, from Assumption 2, we know that $F(w)$ is also L-smooth. For simplicity, we denote L_Ω as

$$L_\Omega = \max_{i=1,2,\ldots,n} \frac{L_i}{nq_i},$$

then $L_\Omega \ge L$. It is not difficult to obtain the following result from Assumption 2.

Lemma 1 (Theorem 2.1.5 [29]) Suppose that $f_i(w)$ is convex and L_i-smooth. Then, for any $w, v \in \mathbb{R}^d$, the following inequality holds

$$(\nabla f_i(w) - \nabla f_i(v))^T (w - v) \ge \frac{1}{L_i} \|\nabla f_i(w) - \nabla f_i(v)\|_2^2.$$

Now we generalize some basic properties of proximal mapping to scaled proximal operator. Although they are direct extensions, we have not found the same results in literatures.

Lemma 2 Let $R(w)$ be a proper closed and convex function on $\mathbb{R}^d$. Then $\operatorname{prox}_R^{A^{-1}}(w)$ is a singleton for any $w \in \operatorname{dom}(R)$ and any symmetric positive definite matrix $A \in \mathbb{S}_{++}^{d\times d}$. Furthermore, the following statements are equivalent:

(i) $\mathbf{u} = \operatorname{prox}_R^{A^{-1}}(w)$.
(ii) $A^{-1}(w - \mathbf{u}) \in \partial R(\mathbf{u})$, where ∂R is the subdifferential of R.

Proof The uniqueness of $\operatorname{prox}_R^{A^{-1}}(w)$ can be proved in a similar way as Theorem 6.3 of [30] by noting that A is symmetric positive definite. For the latter part, one can employ the techniques used in the proof of Theorem 6.39 in [30]. We omit the details here. □

Lemma 3 Let $R(w)$ be a proper closed and convex function on $\mathbb{R}^d$. Then, for any $w, v \in \operatorname{dom}(R)$ and any $A \in \mathbb{S}_{++}^{d\times d}$, the following inequality holds

$$\|\operatorname{prox}_R^{A^{-1}}(w) - \operatorname{prox}_R^{A^{-1}}(v)\|_{A^{-1}}^2 \le \|w - v\|_{A^{-1}}^2.$$

Proof We only need to consider the nontrivial case $w \neq v$. Denoting $\mathbf{u} = \operatorname{prox}_R^{A^{-1}}(w)$ and $\mathbf{v} = \operatorname{prox}_R^{A^{-1}}(v)$. It follows from Lemma 2 that

$$A^{-1}(w - \mathbf{u}) \in \partial R(\mathbf{u}), \quad A^{-1}(v - \mathbf{v}) \in \partial R(\mathbf{v}).$$

By the definition of subdifferential, we have

$$\begin{aligned} R(\mathbf{u}) &\ge R(\mathbf{v}) + (A^{-1}(v - \mathbf{v}))^T(\mathbf{u} - \mathbf{v}), \\ R(\mathbf{v}) &\ge R(\mathbf{u}) + (A^{-1}(w - \mathbf{u}))^T(\mathbf{v} - \mathbf{u}). \end{aligned}$$

Summing the above two inequalities to get

$$\begin{aligned} 0 &\ge \left(A^{-1}((v - \mathbf{v}) - (w - \mathbf{u}))\right)^T \left(\mathbf{u} - \mathbf{v}\right) \\ &= \left(A^{-1}((v - w) + (\mathbf{u} - \mathbf{v}))\right)^T \left(\mathbf{u} - \mathbf{v}\right), \end{aligned}$$

which results in,

$$\begin{aligned} \|\mathbf{u} - \mathbf{v}\|_{A^{-1}}^2 &\le (A^{-1}(w - v))^T(\mathbf{u} - \mathbf{v}) \\ &= (A^{-1/2}(w - v))^T(A^{-1/2}(\mathbf{u} - \mathbf{v})) \\ &\le \|A^{-1/2}(w - v)\|_2 \cdot \|A^{-1/2}(\mathbf{u} - \mathbf{v})\|_2, \end{aligned}$$

where the first equality holds due to the symmetry and positive definiteness of A while the last inequality follows from the Cauchy-Schwarz inequality. By squaring the above inequality, we obtain

$$\begin{aligned} \|\mathbf{u} - \mathbf{v}\|_{A^{-1}}^2 \cdot \|\mathbf{u} - \mathbf{v}\|_{A^{-1}}^2 \le& \|A^{-1/2}(w - v)\|_2^2 \cdot \|A^{-1/2}(\mathbf{u} - \mathbf{v})\|_2^2 \\ =& \|w - v\|_{A^{-1}}^2 \cdot \|\mathbf{u} - \mathbf{v}\|_{A^{-1}}^2. \end{aligned}$$

Since $w \neq v$, we know that $\|\mathbf{u} - \mathbf{v}\|_{A^{-1}}^2 \neq 0$. We complete the proof by dividing both sides of the above inequality by $\|\mathbf{u} - \mathbf{v}\|_{A^{-1}}^2$. □

The following theorem shows that our proximal stochastic recursive step $w_{t+1}^k - w_t^k$ decreases in expectation for convex functions.

Theorem 1 *Suppose that Assumptions 1 and 2 hold. Consider v_t^k defined by (4) in SRG-DBB with $0 \prec U_k \preceq 1/L_\Omega \mathbb{I}_d$. Then, in the k-th outer loop, for any $t > 1$, we have*

$$\tilde{\mathbb{E}}\big[\|w_{t+1}^k - w_t^k\|_{U_k^{-1}}^2\big] \leq \tilde{\mathbb{E}}\big[\|w_t^k - w_{t-1}^k\|_{U_k^{-1}}^2\big].$$

Proof Conditioned on $\mathcal{F}_t$, we take expectation with respect to I_t to obtain

$$\begin{aligned}
&\mathbb{E}_{I_t}\big[\|w_{t+1}^k - w_t^k\|_{U_k^{-1}}^2|\mathcal{F}_t\big] \\
&= \mathbb{E}_{I_t}\big[\|\mathrm{prox}_R^{U_k^{-1}}(w_t^k - U_k v_t^k) - \mathrm{prox}_R^{U_k^{-1}}(w_{t-1}^k - U_k v_{t-1}^k)\|_{U_k^{-1}}^2|\mathcal{F}_t\big] \\
&\leq \mathbb{E}_{I_t}\big[\|w_t^k - w_{t-1}^k - U_k(v_t^k - v_{t-1}^k)\|_{U_k^{-1}}^2|\mathcal{F}_t\big] \\
&= \mathbb{E}_{I_t}\big[\|w_t^k - w_{t-1}^k\|_{U_k^{-1}}^2 + \|v_t^k - v_{t-1}^k\|_{U_k}^2 - 2(w_t^k - w_{t-1}^k)^T(v_t^k - v_{t-1}^k)|\mathcal{F}_t\big] \\
&= \|w_t^k - w_{t-1}^k\|_{U_k^{-1}}^2 + \mathbb{E}_{I_t}\big[\|v_t^k - v_{t-1}^k\|_{U_k}^2|\mathcal{F}_t\big] \\
&\quad - 2\mathbb{E}_{I_t}\big[(w_t^k - w_{t-1}^k)^T(\frac{1}{b}\sum_{i\in I_t}\frac{\nabla f_i(w_t^k) - \nabla f_i(w_{t-1}^k)}{q_i n})|\mathcal{F}_t\big] \\
&\leq \|w_t^k - w_{t-1}^k\|_{U_k^{-1}}^2 + \mathbb{E}_{I_t}\big[\|v_t^k - v_{t-1}^k\|_{U_k}^2|\mathcal{F}_t\big] \\
&\quad - 2\mathbb{E}_{I_t}\big[\frac{1}{b}\sum_{i\in I_t}\frac{\|\nabla f_i(w_t^k) - \nabla f_i(w_{t-1}^k)\|_2^2}{q_i n L_i}|\mathcal{F}_t\big] \\
&\leq \|w_t^k - w_{t-1}^k\|_{U_k^{-1}}^2 + \mathbb{E}_{I_t}\big[\|v_t^k - v_{t-1}^k\|_{U_k}^2|\mathcal{F}_t\big] \\
&\quad - \frac{2}{L_\Omega}\mathbb{E}_{I_t}\big[\frac{1}{b}\sum_{i\in I_t}\|\frac{\nabla f_i(w_t^k) - \nabla f_i(w_{t-1}^k)}{q_i n}\|_2^2|\mathcal{F}_t\big] \\
&\leq \|w_t^k - w_{t-1}^k\|_{U_k^{-1}}^2 + \mathbb{E}_{I_t}\big[\|v_t^k - v_{t-1}^k\|_{U_k}^2|\mathcal{F}_t\big] \\
&\quad - \frac{2}{L_\Omega}\mathbb{E}_{I_t}\big[\|\frac{1}{b}\sum_{i\in I_t}\frac{\nabla f_i(w_t^k) - \nabla f_i(w_{t-1}^k)}{q_i n}\|_2^2|\mathcal{F}_t\big] \qquad (10) \\
&\leq \|w_t^k - w_{t-1}^k\|_{U_k^{-1}}^2 + \frac{1}{L_\Omega}\mathbb{E}_{I_t}\big[\|v_t^k - v_{t-1}^k\|_2^2|\mathcal{F}_t\big] \\
&\quad - \frac{2}{L_\Omega}\mathbb{E}_{I_t}\big[\|\frac{1}{b}\sum_{i\in I_t}\frac{\nabla f_i(w_t^k) - \nabla f_i(w_{t-1}^k)}{q_i n}\|_2^2|\mathcal{F}_t\big] \\
&= \|w_t^k - w_{t-1}^k\|_{U_k^{-1}}^2 - \frac{1}{L_\Omega}\mathbb{E}_{I_t}\big[\|v_t^k - v_{t-1}^k\|_2^2|\mathcal{F}_t\big] \\
&\leq \|w_t^k - w_{t-1}^k\|_{U_k^{-1}}^2,
\end{aligned}$$

where the first inequality follows from Lemma 3 and the second inequality uses Lemma 1. The third inequality holds due to $L_\Omega \geq L_i/(nq_i)$ for $i = 1, 2, \dots, n$. In the fourth and fifth inequalities we use the facts that $\mathbb{E}\big[\|z_1 + z_2 + \ldots + z_r\|_2^2\big] \leq$

$r\mathbb{E}\big[\|z_1\|_2^2+\|z_2\|_2^2+\ldots+\|z_r\|_2^2\big]$ with z_j being random variables for $j\in\{1,2,\ldots,r\}$ and $0\prec U_k\preceq 1/L_\Omega\mathbb{I}_d$, respectively. The last equality holds by the definition of v_t^k. We can obtain the desired result by taking total expectation.

□

Let $\mathcal{W}_*$ be the set of optimal solutions of problem (1) and $w_*\in\mathcal{W}_*$. From Theorem 2 in [24], an upper bound on the variance of v_t^k can be given as follows.

Lemma 4 Suppose that Assumptions 1 and 2 hold, and choose $b\in\{1,2,\ldots,n\}$. Consider v_t^k as defined in (4). Then, for any $t>1$, we have

$$\tilde{\mathbb{E}}\big[\|v_t^k-\nabla F(w_t^k)\|_2^2\big]\le\frac{4L_\Omega}{b}\tilde{\mathbb{E}}\big[P(w_t^k)-P(w_*)+P(w_{t-1}^k)-P(w_*)\big].$$

To analyze the convergence of multiple outer loops, we define the following generalization of stochastic gradient mapping

$$g_t^k=U_k^{-1}(w_t^k-w_{t+1}^k)=U_k^{-1}\big(w_t^k-\mathrm{prox}_R^{U_k^{-1}}(w_t^k-U_kv_t^k)\big).\qquad(11)$$

Then the proximal stochastic gradient step in Algorithm 1 can be written as

$$w_{t+1}^k=w_t^k-U_kg_t^k.\qquad(12)$$

Before establishing the convergence of SRG-DBB, we show an upper bound on $P(w)$ by using (11) and (12) in a similar way to Lemma 3.7 in [15]. However, we do not require the strong convexity of $F(w)$ and $R(w)$.

Lemma 5 Suppose that Assumptions 1 and 2 hold, and $0\prec U_k\preceq 1/L_\Omega\mathbb{I}_d$. For any $t\ge 1$, we have

$$(w_*-w_t^k)^Tg_t^k+\frac{1}{2}\|g_t^k\|_{U_k}^2\le P(w_*)-P(w_{t+1}^k)-(w_*-w_{t+1}^k)^T\delta_t^k,$$

where $\delta_t^k=\nabla F(w_t^k)-v_t^k$.

Proof Since

$$w_{t+1}^k=\arg\min_y\big\{R(y)+\frac{1}{2}\|y-(w_t^k-U_kv_t^k)\|_{U_k^{-1}}^2\big\},$$

by Lemma 2, we get

$$U_k^{-1}\big((w_t^k-U_kv_t^k)-w_{t+1}^k\big)\in\partial R(w_{t+1}^k),$$

which implies that there exists $\varphi\in\partial R(w_{t+1}^k)$ such that

$$U_k^{-1}\big(w_{t+1}^k-(w_t^k-U_kv_t^k)\big)+\varphi=0.$$

This together with (12) gives $v_t^k+\varphi=g_t^k$. Then

$$(w_*-w_{t+1}^k)^T(v_t^k+\varphi)=(w_*-w_{t+1}^k)^Tg_t^k.\qquad(13)$$

From the convexity of $F(w)$ and $R(w)$, we get

$$P(w_*) \geq F(w_t^k) + \nabla F(w_t^k)^T (w_* - w_t^k) + R(w_{t+1}^k) + \varphi^T (w_* - w_{t+1}^k). \tag{14}$$

It follows from the L-smoothness of $F(w)$ that

$$\begin{aligned} F(w_t^k) &\geq F(w_{t+1}^k) - \nabla F(w_t^k)^T (w_{t+1}^k - w_t^k) - \frac{L}{2}\|w_{t+1}^k - w_t^k\|_2^2 \\ &\geq F(w_{t+1}^k) - \nabla F(w_t^k)^T (w_{t+1}^k - w_t^k) - \frac{L_\Omega}{2}\|w_{t+1}^k - w_t^k\|_2^2, \end{aligned} \tag{15}$$

where the second inequality is due to the fact $0 < L \leq L_\Omega$. Combining (14) and (15), we have

$$\begin{aligned} P(w_*) \geq\ & F(w_{t+1}^k) - \nabla F(w_t^k)^T (w_{t+1}^k - w_t^k) + \nabla F(w_t^k)^T (w_* - w_t^k) + R(w_{t+1}^k) \\ & + \varphi^T (w_* - w_{t+1}^k) - \frac{L_\Omega}{2}\|w_{t+1}^k - w_t^k\|_2^2 \\ =\ & P(w_{t+1}^k) + \nabla F(w_t^k)^T (w_* - w_{t+1}^k) + \varphi^T (w_* - w_{t+1}^k) - \frac{L_\Omega}{2}\|w_{t+1}^k - w_t^k\|_2^2 \\ \geq\ & P(w_{t+1}^k) + \nabla F(w_t^k)^T (w_* - w_{t+1}^k) + \varphi^T (w_* - w_{t+1}^k) - \frac{1}{2}\|g_t^k\|_{U_k}^2, \end{aligned} \tag{16}$$

where the first equality follows from the definition of $P(w)$ and the last inequality holds by (12) and $0 \prec U_k \preceq 1/L_\Omega \mathbb{I}_d$. Collecting all inner products on the right-hand side of (16), we obtain

$$\begin{aligned} & \nabla F(w_t^k)^T (w_* - w_{t+1}^k) + \varphi^T (w_* - w_{t+1}^k) \\ &= (w_* - w_{t+1}^k)^T (\delta_t^k + v_t^k) + (w_* - w_{t+1}^k)^T \varphi \\ &= (w_* - w_{t+1}^k)^T \delta_t^k + (w_* - w_{t+1}^k)^T (v_t^k + \varphi) \\ &= (w_* - w_{t+1}^k)^T \delta_t^k + (w_* - w_{t+1}^k)^T g_t^k \\ &= (w_* - w_{t+1}^k)^T \delta_t^k + (w_* - w_t^k + w_t^k - w_{t+1}^k)^T g_t^k \\ &= (w_* - w_{t+1}^k)^T \delta_t^k + (w_* - w_t^k)^T g_t^k + (g_t^k)^T U_k g_t^k \\ &= (w_* - w_{t+1}^k)^T \delta_t^k + (w_* - w_t^k)^T g_t^k + \|g_t^k\|_{U_k}^2, \end{aligned} \tag{17}$$

where the first equality follows from the definition of δ_t^k, and the third and fifth equalities are derived from (13) and (12), respectively. Applying (17) to (16), we get

$$P(w_*) \geq P(w_{t+1}^k) + \frac{1}{2}\|g_t^k\|_{U_k}^2 + (w_* - w_{t+1}^k)^T \delta_t^k + (w_* - w_t^k)^T g_t^k.$$

Then the desired result is obtained. □

3.1 Convergence properties for strongly convex case

We analyze the linear convergence of SRG-DBB in the case where $P(w)$ is strongly convex.

Assumption 3 The objective function $P(w)$ is μ-strongly convex, that is, there exists $\mu > 0$ such that for all $w \in \mathrm{dom}(R)$ and $\upsilon \in \mathbb{R}^d$,

$$P(\upsilon) \geq P(w) + \xi^T(\upsilon - w) + \frac{\mu}{2}\|\upsilon - w\|_2^2,\ \forall\ \xi \in \partial P(w).$$

Assumptions 1, 2 and 3 are often satisfied by objective functions in machine learning, such as ridge regression and elastic net regularization logistic regression. Moreover, w_* is unique when $P(w)$ is strongly convex.

3.1.1 Linear convergence

The following theorem shows that our proximal stochastic recursive step has a linear convergence rate for strongly convex functions.

Theorem 2 *Suppose that Assumptions 1 and 2 hold, $F(w)$ is μ_F-strongly convex and $0 \prec U_k \preceq 2/L_\Omega \mathbb{I}_d$. Then, in the k-th outer loop, for any $t > 1$, we have*

$$\tilde{\mathbb{E}}\big[\|w_{t+1}^k - w_t^k\|_{U_k^{-1}}^2\big] \le \left(1 - (\mu_F^2 u_k^{\min})(\frac{2}{L_\Omega} - u_k^{\max})\right) \tilde{\mathbb{E}}\big[\|w_t^k - w_{t-1}^k\|_{U_k^{-1}}^2\big],$$

where $u_k^{\max} = \max_j\{u_k^{(j)}\}$ and $u_k^{\min} = \min_j\{u_k^{(j)}\}$.

Proof The inequality (10) in Theorem 1 indicates that

$$\begin{aligned}
&\mathbb{E}_{I_t}\big[\|w_{t+1}^k - w_t^k\|_{U_k^{-1}}^2 |\mathcal{F}_t\big] \\
&\le \|w_t^k - w_{t-1}^k\|_{U_k^{-1}}^2 + \mathbb{E}_{I_t}\big[\|v_t^k - v_{t-1}^k\|_{U_k}^2 |\mathcal{F}_t\big] - \frac{2}{L_\Omega}\mathbb{E}_{I_t}\big[\|v_t^k - v_{t-1}^k\|_2^2 |\mathcal{F}_t\big] \\
&\le \|w_t^k - w_{t-1}^k\|_{U_k^{-1}}^2 + (u_k^{\max} - \frac{2}{L_\Omega})\mathbb{E}_{I_t}\big[\|v_t^k - v_{t-1}^k\|_2^2 |\mathcal{F}_t\big] \\
&\le \|w_t^k - w_{t-1}^k\|_{U_k^{-1}}^2 + (u_k^{\max} - \frac{2}{L_\Omega})\|\nabla F(w_t^k) - \nabla F(w_{t-1}^k)\|_2^2 \\
&\le \|w_t^k - w_{t-1}^k\|_{U_k^{-1}}^2 + \mu_F^2(u_k^{\max} - \frac{2}{L_\Omega})\|w_t^k - w_{t-1}^k\|_2^2 \\
&\le \|w_t^k - w_{t-1}^k\|_{U_k^{-1}}^2 + \mu_F^2 u_k^{\min}(u_k^{\max} - \frac{2}{L_\Omega})\|w_t^k - w_{t-1}^k\|_{U_k^{-1}}^2 \\
&= \big(1 - \mu_F^2 u_k^{\min}(\frac{2}{L_\Omega} - u_k^{\max})\big)\|w_t^k - w_{t-1}^k\|_{U_k^{-1}}^2.
\end{aligned}$$

Here, the second inequality holds due to $U_k \preceq u_k^{\max}\mathbb{I}_d$, and the third inequality uses $\|\nabla F(w_t^k) - \nabla F(w_{t-1}^k)\|_2^2 = \|\mathbb{E}_{I_t}[v_t^k - v_{t-1}^k|\mathcal{F}_t]\|_2^2 \le \mathbb{E}_{I_t}[\|v_t^k - v_{t-1}^k\|_2^2|\mathcal{F}_t]$, because it holds that $\mathbb{E}[\|z - \mathbb{E}[z]\|_2^2] = \mathbb{E}[\|z\|_2^2] - \|\mathbb{E}[z]\|^2 \ge 0$ for random vector $z \in \mathbb{R}^d$. Notice that $u_k^{\max} - 2/L_\Omega \le 0$ since $U_k \preceq 2/L_\Omega\mathbb{I}_d$. In the fourth inequality we use the fact that $\mu_F\|w_t^k - w_{t-1}^k\|_2 \le \|\nabla F(w_t^k) - \nabla F(w_{t-1}^k)\|_2$, which can be deduced from the strong convexity of $F(w)$. The last inequality is due to the definition of $u_k^{\min}$. By taking total expectation, we obtain the desired result. □

The following theorem establishes the linear convergence of SRG-DBB under the strongly convex condition.

Theorem 3 *Suppose that Assumptions 1, 2 and 3 hold, and choose $b \in \{1, 2, \ldots, n\}$. Assume that $0 \prec U_k \preceq 1/L_\Omega\mathbb{I}_d$, $8L_\Omega u_k^{\max}/b < 1$, and m is chosen so that*

$$\rho_k = \frac{1}{m\mu u_k^{\min}\big(1 - 8L_\Omega u_k^{\max}/b\big)} + \frac{4L_\Omega u_k^{\max}}{mb\big(1 - 8L_\Omega u_k^{\max}/b\big)} < 1.$$

Then, SRG-DBB converges linearly in expectation

$$\tilde{\mathbb{E}}\big[P(\tilde{w}^{k+1}) - P(w_*)\big] \le \rho_k\tilde{\mathbb{E}}\big[P(\tilde{w}^k) - P(w_*)\big].$$

Proof From the update rule (12), we obtain that, for any $t \geq 1$,

$$\begin{aligned}\|w_{t+1}^k - w_*\|_{U_k^{-1}}^2 &= \|w_t^k - U_k g_t^k - w_*\|_{U_k^{-1}}^2 \\ &= \|w_t^k - w_*\|_{U_k^{-1}}^2 - 2(w_t^k - w_*)^T g_t^k + \|g_t^k\|_{U_k}^2 \\ &\leq \|w_t^k - w_*\|_{U_k^{-1}}^2 - 2\big(P(w_{t+1}^k) - P(w_*)\big) + 2(w_{t+1}^k - w_*)^T \delta_t^k,\end{aligned} \tag{18}$$

where the last inequality uses Lemma 5. In order to provide an upper bound on the quantity $2(w_{t+1}^k - w_*)^T \delta_t^k$, we need the following notation

$$\bar{w}_{t+1}^k = \mathrm{prox}_R^{U_k^{-1}}(w_t^k - U_k \nabla F(w_t^k)), \tag{19}$$

which is independent of the random variable I_t. Then we get

$$\begin{aligned}&2(w_{t+1}^k - w_*)^T \delta_t^k \\ &= 2(w_{t+1}^k - \bar{w}_{t+1}^k)^T \delta_t^k + 2(\bar{w}_{t+1}^k - w_*)^T \delta_t^k \\ &\leq 2\|\delta_t^k\|_{U_k} \|w_{t+1}^k - \bar{w}_{t+1}^k\|_{U_k^{-1}} + 2(\bar{w}_{t+1}^k - w_*)^T \delta_t^k \\ &\leq 2\|\delta_t^k\|_{U_k} \|(w_t^k - U_k v_t^k) - (w_t^k - U_k \nabla F(w_t^k))\|_{U_k^{-1}} + 2(\bar{w}_{t+1}^k - w_*)^T \delta_t^k \\ &\leq 2u_k^{\max} \|\delta_t^k\|_2^2 + 2(\bar{w}_{t+1}^k - w_*)^T \delta_t^k,\end{aligned} \tag{20}$$

where the first equality uses the fact that $w^T v \leq \|w\|_A \cdot \|v\|_{A^{-1}}$ with any positive definite matrix A, the second inequality holds due to Lemma 3, and the last inequality follows from the definitions of $u_k^{\max}$ and δ_t^k. Combining (20) with (18), we obtain

$$\begin{aligned}\|w_{t+1}^k - w_*\|_{U_k^{-1}}^2 \leq \|w_t^k - w_*\|_{U_k^{-1}}^2 - 2\big(P(w_{t+1}^k) - P(w_*)\big) \\ + 2u_k^{\max} \|\delta_t^k\|_2^2 + 2(\bar{w}_{t+1}^k - w_*)^T \delta_t^k\end{aligned} \tag{21}$$

Since both $\bar{w}_{t+1}^k$ and w_* are independent of I_t and the history of random variables w_0^k, I_1, I_2, …, I_{t-1}, and $\tilde{\mathbb{E}}[\delta_t^k] = 0$, we have

$$\tilde{\mathbb{E}}\big[(\bar{w}_{t+1}^k - w_*)^T \delta_t^k\big] = 0.$$

By taking total expectation and applying Lemma 4 to (21), we obtain

$$\begin{aligned}&\tilde{\mathbb{E}}\big[\|w_{t+1}^k - w_*\|_{U_k^{-1}}^2\big] \\ &\leq \tilde{\mathbb{E}}\big[\|w_t^k - w_*\|_{U_k^{-1}}^2\big] - 2\tilde{\mathbb{E}}\big[P(w_{t+1}^k) - P(w_*)\big] + 2u_k^{\max} \tilde{\mathbb{E}}\big[\|\delta_t^k\|_2^2\big] \\ &\leq \tilde{\mathbb{E}}\big[\|w_t^k - w_*\|_{U_k^{-1}}^2\big] - 2\tilde{\mathbb{E}}\big[P(w_{t+1}^k) - P(w_*)\big] + \frac{8L_\Omega u_k^{\max}}{b} \tilde{\mathbb{E}}\big[P(w_t^k) - P(w_*)\big] \\ &\quad + \frac{8L_\Omega u_k^{\max}}{b} \tilde{\mathbb{E}}\big[P(w_{t-1}^k) - P(w_*)\big].\end{aligned} \tag{22}$$

Notice that $v_1^k = v_0^k$ and $\delta_1^k = \nabla F(w_1^k) - v_1^k = \nabla F(\tilde{w}^k) - v_0^k = 0$ since $w_1^k = w_0^k = \tilde{w}^k$ and $v_0^k = \nabla F(\tilde{w}^k)$. It follows from (18) that

$$\|w_2^k - w_*\|_{U_k^{-1}}^2 \leq \|w_1^k - w_*\|_{U_k^{-1}}^2 - 2\big((P(w_2^k) - P(w_*)\big). \tag{23}$$

Summing (22) over $t = 2, \ldots, m$ and taking into account (23), we get

$$\tilde{\mathbb{E}}\big[\|w_{m+1}^k - w_*\|_{U_k^{-1}}^2\big] + 2\tilde{\mathbb{E}}\big[P(w_{m+1}^k) - P(w_*)\big]$$

$$\begin{aligned}
&+2\big(1-\frac{4L_\Omega u_k^{\max}}{b}\big)\sum_{t=2}^{m}\tilde{\mathbb{E}}\big[P(w_t^k)-P(w_*)\big]\\
&\le \tilde{\mathbb{E}}\big[\|w_1^k-w_*\|^2_{U_k^{-1}}\big]+\frac{8L_\Omega u_k^{\max}}{b}\tilde{\mathbb{E}}\big[P(w_1^k)-P(w_*)\big]\\
&+\frac{8L_\Omega u_k^{\max}}{b}\sum_{t=2}^{m-1}\tilde{\mathbb{E}}\big[P(w_t^k)-P(w_*)\big]\\
&\le \tilde{\mathbb{E}}\big[\|w_1^k-w_*\|^2_{U_k^{-1}}\big]+\frac{8L_\Omega u_k^{\max}}{b}\tilde{\mathbb{E}}\big[P(w_1^k)-P(w_*)\big]\\
&+\frac{8L_\Omega u_k^{\max}}{b}\sum_{t=2}^{m}\tilde{\mathbb{E}}\big[P(w_t^k)-P(w_*)\big],
\end{aligned}\tag{24}$$

where the last inequality uses the fact that $P(w)\ge P(w_*)$ for any $w\in\mathbb{R}^d$. Rearranging terms of (24), this yields

$$\begin{aligned}
&\tilde{\mathbb{E}}\big[\|w_{m+1}^k-w_*\|^2_{U_k^{-1}}\big]+2\tilde{\mathbb{E}}\big[P(w_{m+1}^k)-P(w_*)\big]\\
&\quad+2\big(1-\frac{8L_\Omega u_k^{\max}}{b}\big)\sum_{t=2}^{m}\tilde{\mathbb{E}}\big[P(w_t^k)-P(w_*)\big]\\
&\quad\le \tilde{\mathbb{E}}\big[\|w_1^k-w_*\|^2_{U_k^{-1}}\big]+\frac{8L_\Omega u_k^{\max}}{b}\tilde{\mathbb{E}}\big[P(w_1^k)-P(w_*)\big],
\end{aligned}\tag{25}$$

Since $2(1-8L_\Omega u_k^{\max}/b)<2$, $\tilde{\mathbb{E}}\big[\|w_{m+1}^k-w_*\|^2_{U_k^{-1}}\big]\ge 0$, and $w_1^k=\tilde{w}^k$, we obtain

$$\begin{aligned}
&2\big(1-\frac{8L_\Omega u_k^{\max}}{b}\big)\sum_{t=2}^{m+1}\tilde{\mathbb{E}}\big[P(w_t^k)-P(w_*)\big]\\
&\le \|\tilde{w}^k-w_*\|^2_{U_k^{-1}}+\frac{8L_\Omega u_k^{\max}}{b}\tilde{\mathbb{E}}\big[P(\tilde{w}^k)-P(w_*)\big]\\
&\le \frac{1}{u_k^{\min}}\cdot\|\tilde{w}^k-w_*\|_2^2+\frac{8L_\Omega u_k^{\max}}{b}\tilde{\mathbb{E}}\big[P(\tilde{w}^k)-P(w_*)\big]\\
&\le \left(\frac{2}{\mu u_k^{\min}}+\frac{8L_\Omega u_k^{\max}}{b}\right)\tilde{\mathbb{E}}\big[P(\tilde{w}^k)-P(w_*)\big],
\end{aligned}$$

where the second inequality holds by the definition of $u_k^{\min}$ and in the last inequality we use the fact that $\|\tilde{w}^k-w_*\|_2^2\le 2/\mu(P(\tilde{w}^k)-P(w_*))$, which can be deduced from the strong convexity of $P(w)$. By the definition of $\tilde{w}^{k+1}$ in Algorithm 1, we have $\tilde{\mathbb{E}}[P(\tilde{w}^{k+1})]=(1/m)\sum_{t=1}^{m}\tilde{\mathbb{E}}[P(w_{t+1}^k)]$. Then the following inequality holds

$$\begin{aligned}
&2m\big(1-\frac{8L_\Omega u_k^{\max}}{b}\big)\tilde{\mathbb{E}}\big[P(\tilde{w}^{k+1})-P(w_*)\big]\\
&\le\left(\frac{2}{\mu u_k^{\min}}+\frac{8L_\Omega u_k^{\max}}{b}\right)\tilde{\mathbb{E}}\big[P(\tilde{w}^k)-P(w_*)\big].
\end{aligned}$$

Dividing both sides of the above inequality by $2m\big(1-8L_\Omega u_k^{\max}/b\big)$ and using the definition of ρ_k, we arrive at

$$\tilde{\mathbb{E}}\big[P(\tilde{w}^{k+1})-P(w_*)\big]\le\rho_k\tilde{\mathbb{E}}\big[P(\tilde{w}^k)-P(w_*)\big].$$

Then the desired result is proved.

□

3.1.2 Comparisons of complexity

In order to achieve an ϵ-accuracy, i.e.,

$$\tilde{\mathbb{E}}\big[P(\tilde{w}^{k+1}) - P(w_*)\big] \leq \epsilon,$$

from Theorem 3, we know that the number of outer loops should be set to $O(\log(1/\epsilon))$. Let $u_k^{\min} = u_k^{\max} = \eta = \theta b/L_\Omega$ with $0 < \theta < 1/8$. Then we have

$$\rho = \frac{\kappa}{mb\theta(1-8\theta)} + \frac{4\theta}{m(1-8\theta)}, \tag{26}$$

where $\kappa = L_\Omega/\mu$ is the condition number of the objective function. When setting $\theta = 1/16$ and $m = \max\{64\kappa/b, 4\}$, by (26), it is easy to obtain that $\rho \leq 5/8$. Since SRG-DBB requires at most $n + 2bm$ component gradient computation in each outer loop, the overall workload of SRG-DBB is

$$O((n+\kappa)\log(\frac{1}{\epsilon})).$$

Table 1 shows the comparison results of complexity of the existing methods and SRG-DBB under the strong convexity condition. Inequality $n+\kappa \leq n\sqrt{\kappa} \leq n\kappa$ implies that the complexity of SRG-DBB is lower than ISTA and FISTA. It is easy to see that the complexity of SRG-DBB is the same as Prox-SVRG, mS2GD and SARAH, and is lower than Prox-SGD.

Table 1 Complexity of different methods

Methods	Complexity
ISTA	$O(n\kappa\log(\frac{1}{\epsilon}))$
FISTA [31]	$O(n\sqrt{\kappa}\log(\frac{1}{\epsilon}))$
Prox-SGD	$O(\frac{1}{\epsilon})$
Prox-SVRG	$O((n+\kappa)\log(\frac{1}{\epsilon}))$
mS2GD	$O((n+\kappa)\log(\frac{1}{\epsilon}))$
SARAH	$O((n+\kappa)\log(\frac{1}{\epsilon}))$
SRG-DBB	$O((n+\kappa)\log(\frac{1}{\epsilon}))$

3.2 Convergence properties for non-strongly convex case

We establish linear convergence of our SRG-DBB method under quadratic growth condition (QGC) [32], which is stated as follows:

$$P(w) - P_* \geq \frac{\nu}{2}\|w - \hat{w}\|_2^2, \ \forall w \in \mathbb{R}^d, \tag{27}$$

where $\nu > 0$, $\hat{w}$ is the projection of w onto $\mathcal{W}_*$ and P_* represents the optimal value of (1).

QGC is weaker than the strongly convex condition. For example, the ℓ_1-regularized least squares problems and logistic regression problems satisfy QGC [33], however, they are not strongly convex when the data matrix does not have full column rank. It is shown that a nonsmooth convex function satisfying QGC meets the proximal Polyak-Łojasiewicz inequality [32]. The authors of [34] deduced the equivalence among QGC, the extended restricted strongly convex property (eRSC) and the extended global error bound property (eGEB).

Theorem 4 *Suppose that Assumptions 1 and 2 hold, problem (1) satisfies QGC inequality with $\nu > 0$, and choose $b \in \{1, 2, \dots, n\}$. Further assume that $0 \prec U_k \preceq 1/L_\Omega \mathbb{I}_d$, $8L_\Omega u_k^{\max}/b < 1$, and m is chosen so that*

$$\hat{\rho}_k = \frac{1}{m\nu u_k^{\min}\big(1 - 8L_\Omega u_k^{\max}/b\big)} + \frac{4L_\Omega u_k^{\max}}{mb\big(1 - 8L_\Omega u_k^{\max}/b\big)} < 1.$$

Then, SRG-DBB achieves a linear convergence rate in expectation

$$\tilde{\mathbb{E}}\big[P(\tilde{w}^{k+1}) - P_*\big] \le \hat{\rho}_k \tilde{\mathbb{E}}\big[P(\tilde{w}^k) - P_*\big].$$

Proof Let $\hat{w}_t^k$ be the projection of w_t^k onto $\mathcal{W}_*$, i.e.,

$$\hat{w}_t^k = \Pi_{\mathcal{W}_*}(w_t^k) = \arg\min_w \{w \in \mathcal{W}_* : \|w_t^k - w\|^2_{U_k^{-1}}\}.$$

Then $\hat{w}_t^k, \hat{w}_{t+1}^k \in \mathcal{W}_*$, which together with (12) implies that, for $t \ge 1$,

$$\begin{aligned}
\|w_{t+1}^k - \hat{w}_{t+1}^k\|^2_{U_k^{-1}} &\le \|w_{t+1}^k - \hat{w}_t^k\|^2_{U_k^{-1}} \\
&= \|w_t^k - U_k g_t^k - \hat{w}_t^k\|^2_{U_k^{-1}} \\
&= \|w_t^k - \hat{w}_t^k\|^2_{U_k^{-1}} + 2(\hat{w}_t^k - w_t^k)^T g_t^k + \|g_t^k\|^2_{U_k} \\
&\le \|w_t^k - \hat{w}_t^k\|^2_{U_k^{-1}} + 2(P_* - P(w_{t+1}^k)) - 2(\hat{w}_t^k - w_{t+1}^k)^T \delta_t^k,
\end{aligned}$$

where the first inequality holds due to the positive definiteness of U_k, and the last inequality is the application of Lemma 5 with $\hat{w}_t^k \in \mathcal{W}_*$.

Similarly to the proof of (23)-(25) in Theorem 3, we obtain

$$\begin{aligned}
&2\big(1 - \frac{8L_\Omega u_k^{\max}}{b}\big) \sum_{t=2}^{m+1} \tilde{\mathbb{E}}\big[P(w_t^k) - P_*\big] \\
&\le \|\tilde{w}^k - \hat{w}_1^k\|^2_{U_k^{-1}} + \frac{8L_\Omega u_k^{\max}}{b} \tilde{\mathbb{E}}\big[P(\tilde{w}^k) - P_*\big] \\
&\le \frac{1}{u_k^{\min}} \cdot \|\tilde{w}^k - \hat{w}_1^k\|_2^2 + \frac{8L_\Omega u_k^{\max}}{b} \tilde{\mathbb{E}}\big[P(\tilde{w}^k) - P_*\big].
\end{aligned} \tag{28}$$

The definition of $\tilde{w}^{k+1}$ implies that

$$\tilde{\mathbb{E}}[P(\tilde{w}^{k+1})] = \frac{1}{m} \sum_{t=1}^{m} \tilde{\mathbb{E}}[P(w_{t+1}^k)].$$

Considering QGC with $w = \tilde{w}^k$, $\tilde{w}^k = w_1^k$ and $\hat{w}_1^k = \Pi_{\mathcal{W}_*}(w_1^k) \in \mathcal{W}_*$, we have

$$P(\tilde{w}^k) - P_* = P(\tilde{w}^k) - P(\hat{w}_1^k) \geq \frac{\nu}{2}\|\tilde{w}^k - \hat{w}_1^k\|_2^2,$$

which together with (28) yields

$$\begin{aligned} &2m\big(1 - \frac{8L_\Omega u_k^{\max}}{b}\big)\tilde{\mathbb{E}}\big[P(\tilde{w}^{k+1}) - P_*\big] \\ \leq &\left(\frac{2}{\nu u_k^{\min}} + \frac{8L_\Omega u_k^{\max}}{b}\right)\tilde{\mathbb{E}}\big[P(\tilde{w}^k) - P_*\big]. \end{aligned}$$

Dividing both sides of the above inequality by $2m(1 - 8L_\Omega u_k^{\max}/b)$, and considering the definition of $\hat{\rho}_k$, we arrive at

$$\tilde{\mathbb{E}}\big[P(\tilde{w}^{k+1}) - P_*\big] \leq \hat{\rho}_k\tilde{\mathbb{E}}\big[P(\tilde{w}^k) - P_*\big].$$

□

3.3 Convergence properties for convex case

We study the convergence of SRG-DBB for convex nonsmooth functions. Next lemma presents a new 3-point property which generalizes the one in [35].

Lemma 6 (generalized 3-point property) Suppose that $R : \mathbb{R}^d \to \mathbb{R}$ is lower semicontinuous convex (but possibly nondifferentiable) and $w' = \mathrm{prox}_R^{A^{-1}}(w)$ with $A \in \mathbb{S}_{++}^{d\times d}$. Then, for any $z \in \mathbb{R}^d$, we have the following inequality

$$R(w') + \frac{1}{2}\|w' - w\|_{A^{-1}}^2 \leq R(z) + \frac{1}{2}\|z - w\|_{A^{-1}}^2 - \frac{1}{2}\|w' - z\|_{A^{-1}}^2.$$

Proof Since $w' = \mathrm{prox}_R^{A^{-1}}(w) = \arg\min_z\{R(z) + \frac{1}{2}\|z - w\|_{A^{-1}}^2\}$, there exists $\varpi \in \partial R(w')$ such that

$$\varpi + A^{-1}(w' - w) = 0.$$

By direct expansion, we have

$$\begin{aligned} \frac{1}{2}\|z - w\|_{A^{-1}}^2 = &\frac{1}{2}\|z - w'\|_{A^{-1}}^2 + \frac{1}{2}\|w' - w\|_{A^{-1}}^2 \\ &+ (z - w')^T A^{-1}(w' - w), \quad \forall z \in \mathbb{R}^d. \end{aligned}$$

Using the above two relations and the convexity of $R(z)$, we conclude that

$$\begin{aligned} &R(z) + \frac{1}{2}\|z - w\|_{A^{-1}}^2 \\ = &R(z) + \frac{1}{2}\|z - w'\|_{A^{-1}}^2 + \frac{1}{2}\|w' - w\|_{A^{-1}}^2 + (z - w')^T A^{-1}(w' - w) \\ \geq &R(w') + \varpi^T(z - w') + \frac{1}{2}\|z - w'\|_{A^{-1}}^2 + \frac{1}{2}\|w' - w\|_{A^{-1}}^2 + (z - w')^T A^{-1}(w' - w) \\ = &R(w') + \frac{1}{2}\|z - w'\|_{A^{-1}}^2 + \frac{1}{2}\|w' - w\|_{A^{-1}}^2. \end{aligned}$$

□

Lemma 7 Suppose that $R : \mathbb{R}^d \to \mathbb{R}$ is lower semicontinuous convex (but possibly nondifferentiable) and

$$w' = \mathrm{prox}_R^{A^{-1}}(w - A\zeta) \tag{29}$$

with $A \in \mathbb{S}_{++}^{d\times d}$ and $\zeta \in \mathbb{R}^d$. Then, the following inequality holds

$$R(w') \le R(z) + (z - w')^T\zeta + \frac{1}{2}\left[\|z - w\|_{A^{-1}}^2 - \|w' - w\|_{A^{-1}}^2 - \|w' - z\|_{A^{-1}}^2\right] \tag{30}$$

for all $z \in \mathbb{R}^d$.

Proof By applying Lemma 6 to (29), we get

$$\begin{aligned}
&R(w') + (w' - w)^T\zeta + \frac{1}{2}\|w' - w\|_{A^{-1}}^2 + \frac{1}{2}\|\zeta\|_A^2 \\
&= R(w') + \frac{1}{2}\|w' - (w - A\zeta)\|_{A^{-1}}^2 \\
&\le R(z) + \frac{1}{2}\|z - (w - A\zeta)\|_{A^{-1}}^2 - \frac{1}{2}\|w' - z\|_{A^{-1}}^2 \\
&= R(z) + (z - w)^T\zeta + \frac{1}{2}\|z - w\|_{A^{-1}}^2 + \frac{1}{2}\|\zeta\|_A^2 - \frac{1}{2}\|w' - z\|_{A^{-1}}^2.
\end{aligned}$$

□

Lemma 8 Consider $P(w)$ as defined in (1). Suppose that Assumptions 1 and 2 hold. Then, for w' defined by (29), the following inequality holds

$$\begin{aligned}
P(w') \le P(z) + (w' - z)^T(\nabla F(w) - \zeta) - \frac{1}{2}\|w' - z\|_{A^{-1}}^2 \\
+\frac{1}{2}\|w' - w\|_{(L_\Omega \mathbb{I}_d - A^{-1})}^2 + \frac{1}{2}\|z - w\|_{(L_\Omega \mathbb{I}_d + A^{-1})}^2,
\end{aligned}$$

for all $z \in \mathbb{R}^d$.

Proof From the L-smoothness of F and $L \le L_\Omega$, we obtain

$$\begin{aligned}
F(w') \le F(w) + \nabla F(w)^T(w' - w) + \frac{L_\Omega}{2}\|w' - w\|_2^2, \\
F(w) \le F(z) + \nabla F(w)^T(w - z) + \frac{L_\Omega}{2}\|w - z\|_2^2.
\end{aligned}$$

By summing the above two inequalities, we have

$$F(w') \le F(z) + \nabla F(w)^T(w' - z) + \frac{L_\Omega}{2}\|w' - w\|_2^2 + \frac{L_\Omega}{2}\|w - z\|_2^2. \tag{31}$$

Summing (30) and (31), we get

$$\begin{aligned}
P(w') \le P(z) + (w' - z)^T(\nabla F(w) - \zeta) - \frac{1}{2}\|w' - z\|_{A^{-1}}^2 \\
+\frac{1}{2}\|w' - w\|_{(L_\Omega \mathbb{I}_d - A^{-1})}^2 + \frac{1}{2}\|z - w\|_{(L_\Omega \mathbb{I}_d + A^{-1})}^2,
\end{aligned}$$

which completes our proof. □

In order to derive an upper bound on the variance of v_t^k in the mini-batch setting, we first show the result in the case where $b = 1$.

Lemma 9 Suppose that Assumption 1 holds. Consider v_t^k as defined in (4) with $b = 1$, i.e.,

$$v_t^k = \frac{\nabla f_{i_t}(w_t^k) - \nabla f_{i_t}(w_{t-1}^k)}{nq_{i_t}} + v_{t-1}^k. \tag{32}$$

Then the following inequality holds

$$\tilde{\mathbb{E}}[\|v_t^k - \nabla F(w_t^k)\|_2^2] \leq L_\Omega^2 \tilde{\mathbb{E}}[\|w_t^k - w_{t-1}^k\|_2^2], \quad \forall t \geq 1.$$

Proof Consider v_t^k defined in (32). Conditioned on $\mathcal{F}_t = \sigma(w_0^k, i_1, \ldots, i_{t-1})$, we take expectation with respect to i_t and obtain

$$\mathbb{E}_{i_t}\Big[\frac{\nabla f_{i_t}(w_t^k)}{nq_{i_t}} | \mathcal{F}_t\Big] = \sum_{i=1}^{n} \frac{q_i}{nq_i} \nabla f_i(w_t^k) = \nabla F(w_t^k). \tag{33}$$

Similarly we have

$$\mathbb{E}_{i_t}\Big[\frac{\nabla f_{i_t}(w_{t-1}^k)}{nq_{i_t}} | \mathcal{F}_t\Big] = \nabla F(w_{t-1}^k). \tag{34}$$

Then we obtain

$$\begin{aligned}
&\mathbb{E}_{i_t}\Big[\|v_t^k - \nabla F(w_t^k)\|_2^2 | \mathcal{F}_t\Big] \\
&= \mathbb{E}_{i_t}\Big[\| \frac{\nabla f_{i_t}(w_t^k) - \nabla f_{i_t}(w_{t-1}^k)}{nq_{i_t}} \\
&- \big(\nabla F(w_t^k) - \nabla F(w_{t-1}^k)\big) + \big(v_{t-1}^k - \nabla F(w_{t-1}^k)\big)\|_2^2 | \mathcal{F}_t\Big] \\
&= \mathbb{E}_{i_t}\Big[\| \frac{\nabla f_{i_t}(w_t^k) - \nabla f_{i_t}(w_{t-1}^k)}{nq_{i_t}} \|_2^2 | \mathcal{F}_t\Big] \\
&\quad - \|\nabla F(w_t^k) - \nabla F(w_{t-1}^k)\|_2^2 + \|v_{t-1}^k - \nabla F(w_{t-1}^k)\|_2^2 \\
&= \mathbb{E}_{i_t}\Big[\| \frac{\nabla f_{i_t}(w_t^k) - \nabla f_{i_t}(w_{t-1}^k)}{nq_{i_t}} \|_2^2 | \mathcal{F}_t\Big] \\
&\quad - 2\big(\nabla F(w_t^k) - v_{t-1}^k\big)^T \big(v_{t-1}^k - \nabla F(w_{t-1}^k)\big) - \|\nabla F(w_t^k) - v_{t-1}^k\|_2^2 \\
&\leq \mathbb{E}_{i_t}\Big[\| \frac{\nabla f_{i_t}(w_t^k) - \nabla f_{i_t}(w_{t-1}^k)}{nq_{i_t}} \|_2^2 | \mathcal{F}_t\Big] - 2\big(\nabla F(w_t^k) - v_{t-1}^k\big)^T \big(v_{t-1}^k - \nabla F(w_{t-1}^k)\big),
\end{aligned}$$

where the second equality follows from (33) and (34).

Taking total expectation, this yields

$$\begin{aligned}
\tilde{\mathbb{E}}\big[\|v_t^k - \nabla F(w_t^k)\|_2^2\big] &\leq \tilde{\mathbb{E}}\big[\| \frac{\nabla f_{i_t}(w_t^k) - \nabla f_{i_t}(w_{t-1}^k)}{nq_{i_t}} \|_2^2\big] \\
&\leq \tilde{\mathbb{E}}\big[\frac{L_{i_t}^2}{n^2 q_{i_t}^2} \|w_t^k - w_{t-1}^k\|_2^2\big] \\
&\leq L_\Omega^2 \tilde{\mathbb{E}}\big[\|w_t^k - w_{t-1}^k\|_2^2\big],
\end{aligned}$$

where the first inequality holds due to (9), the second inequality follows from the smoothness of f_i, and the last inequality is due to the fact that $L_\Omega \geq L_i/(nq_i)$ for $i = 1, 2, \ldots, n$. $\square$

The following lemma provides an upper bound on v_t^k, which looks similar to the Lemma 3 in the appendix of [36], but they are essentially different due to the update rule of v_t^k.

Lemma 10 Suppose that Assumption 1 holds and choose $b \in \{1, 2, \dots, n\}$. Consider v_t^k as defined in (4). Then, for any $t \geq 1$, the following inequality holds

$$\tilde{\mathbb{E}}[\|v_t^k - \nabla F(w_t^k)\|_2^2] \leq \frac{L_\Omega^2}{b}\tilde{\mathbb{E}}[\|w_t^k - w_{t-1}^k\|_2^2].$$

Proof We define

$$G_i = (\nabla f_i(w_t^k) - \nabla f_i(w_{t-1}^k))/(nq_i) + v_{t-1}^k.$$

Then v_t^k in (4) can be written as

$$v_t^k = \frac{1}{b}\sum_{i \in I_t}\Big(\frac{\nabla f_i(w_t^k) - \nabla f_i(w_{t-1}^k)}{nq_i} + v_{t-1}^k\Big) = \frac{1}{b}\sum_{i \in I_t} G_i.$$

Conditioned on $\mathcal{F}_t = \sigma(w_0^k, I_1, \dots, I_{t-1})$, we take expectation with respect to I_t and get

$$\begin{aligned}
&\mathbb{E}_{I_t}\big[\|v_t^k - \nabla F(w_t^k)\|_2^2|\mathcal{F}_t\big]\\
=&\frac{1}{b^2}\mathbb{E}_{I_t}\big[\|\sum_{i \in I_t}(G_i - \nabla F(w_t^k))\|_2^2|\mathcal{F}_t\big]\\
=&\frac{1}{b^2}\mathbb{E}_{I_t}\big[\|\sum_{i \in S_1}(G_i - \nabla F(w_t^k)) + (G_{I_t/S_1} - \nabla F(w_t^k))\|_2^2|\mathcal{F}_t\big]\\
=&\frac{1}{b^2}\mathbb{E}_{I_t}\big[\|\sum_{i \in S_1}(G_i - \nabla F(w_t^k))\|_2^2|\mathcal{F}_t\big] + \frac{1}{b^2}\mathbb{E}_{I_t}\big[\|G_{I_t/S_1} - \nabla F(w_t^k)\|_2^2|\mathcal{F}_t\big]\\
&+\frac{2}{b^2}\mathbb{E}_{I_t}\big[\big(\sum_{i \in S_1}(G_i - \nabla F(w_t^k))\big)^T\big(G_{I_t/S_1} - \nabla F(w_t^k)\big)|\mathcal{F}_t\big],
\end{aligned}$$

where $S_1 \subset I_t$ and the number of elements in the set I_t/S_1 is 1. By taking total expectation and applying the above inequality recursively, we obtain

$$\begin{aligned}
\tilde{\mathbb{E}}[\|v_t^k - \nabla F(w_t^k)\|_2^2] =&\frac{1}{b^2}\tilde{\mathbb{E}}\big[\|\sum_{i \in S_1}(G_i - \nabla F(w_t^k))\|_2^2\big] + \frac{1}{b^2}\tilde{\mathbb{E}}\big[\|G_{I_t/S_1} - \nabla F(w_t^k)\|_2^2\big]\\
=&\frac{1}{b^2}\sum_{i \in I_t}\tilde{\mathbb{E}}\big[\|G_i - \nabla F(w_t^k)\|_2^2\big]\\
\leq&\frac{L_\Omega^2}{b}\tilde{\mathbb{E}}\big[\|w_t^k - w_{t-1}^k\|_2^2\big],
\end{aligned}$$

where the first equality holds due to the fact $\tilde{\mathbb{E}}[G_i] = \tilde{\mathbb{E}}[\nabla F(w_t^k)]$, which follows from (9) with $b = 1$. In the last inequality we use Lemma 9. $\square$

To establish the convergence of SRG-DBB under convexity condition, we need the following notation of gradient mapping

$$\mathcal{G}_{A^{-1}}(w) = A^{-1}\Big(w - \mathrm{prox}_R^{A^{-1}}\big(w - A\nabla F(w)\big)\Big), \tag{35}$$

where A is a symmetric positive definite matrix. Note that when $R(w)$ is a constant function, the gradient mapping can be reduced to $\mathcal{G}_{A^{-1}}(w) = \nabla F(w)$. It is not difficult to show that $\mathcal{G}_{A^{-1}}(w) = 0$ if and only if w is a solution of problem (1).

Theorem 5 *Suppose that Assumptions 1 and 2 hold, and $0 \prec U_k \preceq 1/(3L_\Omega)\mathbb{I}_d$. Let $c_{t_k+1} = 0$ and $c_t^k = c_{t+1}^k + (u_k^{\max})^2 L_\Omega^2/(2b)$. Then, for the output w_a of Algorithm 1, we have*

$$\tilde{\mathbb{E}}[\|\mathcal{G}_{U_k^{-1}}(w_a)\|_{U_k}^2] \leq \frac{6(P(\tilde{w}^0) - P(w_*))}{T},$$

where $T = \sum_{k=0}^{K-1} t_k$.

Proof By applying Lemma 8 to the proximal full gradient update defined in (19) (with $w' = \bar{w}_{t+1}^k$, $w = z = w_t^k$, $A = U_k$ and $\zeta = \nabla F(w_t^k)$), and taking total expectation, we have

$$\tilde{\mathbb{E}}[P(\bar{w}_{t+1}^k)] \leq \tilde{\mathbb{E}}[P(w_t^k) + \|\bar{w}_{t+1}^k - w_t^k\|^2_{(\frac{L_\Omega}{2}\mathbb{I}_d - U_k^{-1})}]. \tag{36}$$

Recalling that the iterates of Algorithm 1 are computed by

$$w_{t+1}^k = \mathrm{prox}_R^{U_k^{-1}}(w_t^k - U_k v_t^k).$$

Again by applying Lemma 8 to the above update equation (with $w' = w_{t+1}^k$, $z = \bar{w}_{t+1}^k$, $w = w_t^k$, $A = U_k$ and $\zeta = v_t^k$) and taking total expectation, we have

$$\begin{aligned}
\tilde{\mathbb{E}}[P(w_{t+1}^k)] \leq \tilde{\mathbb{E}}[&P(\bar{w}_{t+1}^k) + \frac{1}{2}\|\bar{w}_{t+1}^k - w_t^k\|^2_{(L_\Omega\mathbb{I}_d + U_k^{-1})} \\
&+ \frac{1}{2}\|w_{t+1}^k - w_t^k\|^2_{(L_\Omega\mathbb{I}_d - U_k^{-1})} - \frac{1}{2}\|w_{t+1}^k - \bar{w}_{t+1}^k\|^2_{U_k^{-1}} \\
&+ (w_{t+1}^k - \bar{w}_{t+1}^k)^T(\nabla F(w_t^k) - v_t^k)].
\end{aligned} \tag{37}$$

By summing (36) and (37), we obtain

$$\begin{aligned}
\tilde{\mathbb{E}}[P(w_{t+1}^k)] \leq \tilde{\mathbb{E}}[&P(w_t^k) + \|\bar{w}_{t+1}^k - w_t^k\|^2_{(L_\Omega\mathbb{I}_d - \frac{1}{2}U_k^{-1})} \\
&+ \frac{1}{2}\|w_{t+1}^k - w_t^k\|^2_{(L_\Omega\mathbb{I}_d - U_k^{-1})} - \frac{1}{2}\|w_{t+1}^k - \bar{w}_{t+1}^k\|^2_{U_k^{-1}} \\
&+ (w_{t+1}^k - \bar{w}_{t+1}^k)^T(\nabla F(w_t^k) - v_t^k)].
\end{aligned} \tag{38}$$

Let $\Gamma = (w_{t+1}^k - \bar{w}_{t+1}^k)^T(\nabla F(w_t^k) - v_t^k)$. The expectation on Γ can be bounded above by

$$\begin{aligned}
\tilde{\mathbb{E}}[\Gamma] &\leq \frac{1}{2}\tilde{\mathbb{E}}[\|w_{t+1}^k - \bar{w}_{t+1}^k\|^2_{U_k^{-1}}] + \frac{1}{2}\tilde{\mathbb{E}}[\|\nabla F(w_t^k) - v_t^k\|^2_{U_k}] \\
&\leq \frac{1}{2}\tilde{\mathbb{E}}[\|w_{t+1}^k - \bar{w}_{t+1}^k\|^2_{U_k^{-1}}] + \frac{u_k^{\max} L_\Omega^2}{2b}\tilde{\mathbb{E}}[\|w_t^k - w_{t-1}^k\|_2^2],
\end{aligned}$$

where in the first inequality we use Cauchy-Schwarz and Young's inequality, and the second inequality follows from the definition of $u_k^{\max}$ and Lemma 10. We substitute the upper bound on Γ in (38) and then obtain

$$\tilde{\mathbb{E}}[P(w_{t+1}^k)] \leq \tilde{\mathbb{E}}[P(w_t^k) + \|\bar{w}_{t+1}^k - w_t^k\|^2_{(L_\Omega\mathbb{I}_d - \frac{1}{2}U_k^{-1})}$$

$$+\frac{1}{2}\|w_{t+1}^k - w_t^k\|^2_{(L_\Omega \mathbb{I}_d - U_k^{-1})} + \frac{u_k^{\max} L_\Omega^2}{2b}\|w_t^k - w_{t-1}^k\|_2^2]. \quad (39)$$

In order to further analyze (39), we need the following auxiliary function

$$\Upsilon(w_{t+1}^k) = \tilde{\mathbb{E}}[P(w_{t+1}^k) + c_{t+1}^k\|w_{t+1}^k - w_t^k\|^2_{U_k^{-1}}], \quad (40)$$

where $c_{t_k+1}^k = 0$ and $c_t^k = c_{t+1}^k + (u_k^{\max})^2 L_\Omega^2/(2b)$. Then $\Upsilon(w_{t+1}^k)$ can be bounded above by

$$\begin{aligned}
\Upsilon(w_{t+1}^k) &= \tilde{\mathbb{E}}[P(w_{t+1}^k) + c_{t+1}^k\|w_{t+1}^k - w_t^k\|^2_{U_k^{-1}}] \\
&\le \tilde{\mathbb{E}}[P(w_{t+1}^k) + c_{t+1}^k\|w_t^k - w_{t-1}^k\|^2_{U_k^{-1}}] \\
&\le \tilde{\mathbb{E}}[P(w_t^k) + \|\bar{w}_{t+1}^k - w_t^k\|^2_{(L_\Omega \mathbb{I}_d - \frac{1}{2}U_k^{-1})} \\
&\quad + c_{t+1}^k\|w_t^k - w_{t-1}^k\|^2_{U_k^{-1}} + \frac{u_k^{\max} L_\Omega^2}{2b}\|w_t^k - w_{t-1}^k\|_2^2] \\
&\le \tilde{\mathbb{E}}[P(w_t^k) + \|\bar{w}_{t+1}^k - w_t^k\|^2_{(L_\Omega \mathbb{I}_d - \frac{1}{2}U_k^{-1})} \\
&\quad + (c_{t+1}^k + \frac{(u_k^{\max})^2 L_\Omega^2}{2b})\|w_t^k - w_{t-1}^k\|^2_{U_k^{-1}}] \\
&= \Upsilon(w_t^k) + \tilde{\mathbb{E}}[\|\bar{w}_{t+1}^k - w_t^k\|^2_{(L_\Omega \mathbb{I}_d - \frac{1}{2}U_k^{-1})}], \qquad (41)
\end{aligned}$$

where the first inequality follows from Theorem 1, and the second inequality holds by (39) and $0 \prec U_k \preceq 1/(3L_\Omega)\mathbb{I}_d \prec 1/L_\Omega \mathbb{I}_d$. The last inequality holds by the definition of $u_k^{\max}$ and the last equality is due to the definitions of c_t^k and $\Upsilon(w_t^k)$. By summing (41) over $t = 1, \ldots, t_k$, we get

$$\Upsilon(w_{t_k+1}^k) \le \Upsilon(w_1^k) + \sum_{t=1}^{t_k}\tilde{\mathbb{E}}[\|\bar{w}_{t+1}^k - w_t^k\|^2_{(L_\Omega \mathbb{I}_d - \frac{1}{2}U_k^{-1})}]. \quad (42)$$

By the fact $c_{t_k+1}^k = 0$ and the definition of $\tilde{w}^{k+1}$, we have

$$\Upsilon(w_{t_k+1}^k) = \tilde{\mathbb{E}}[P(w_{t_k+1}^k)] = \tilde{\mathbb{E}}[P(\tilde{w}^{k+1})].$$

Since $w_1^k = w_0^k = \tilde{w}^k$, we know that $\Upsilon(w_1^k) = \tilde{\mathbb{E}}[P(w_1^k)] = \tilde{\mathbb{E}}[P(\tilde{w}^k)]$. It follows from (42) that

$$\tilde{\mathbb{E}}[P(\tilde{w}^{k+1})] \le \tilde{\mathbb{E}}[P(\tilde{w}^k)] + \sum_{t=1}^{t_k}\tilde{\mathbb{E}}[\|\bar{w}_{t+1}^k - w_t^k\|^2_{(L_\Omega \mathbb{I}_d - \frac{1}{2}U_k^{-1})}]. \quad (43)$$

By summing (43) over $k = 0, \ldots, K-1$ and rearranging terms, we obtain

$$\sum_{k=0}^{K-1}\sum_{t=1}^{t_k}\tilde{\mathbb{E}}[\|\bar{w}_{t+1}^k - w_t^k\|^2_{(\frac{1}{2}U_k^{-1} - L_\Omega \mathbb{I}_d)}] \le P(\tilde{w}^0) - P(\tilde{w}^K) \le P(\tilde{w}^0) - P(w_*), \quad (44)$$

where in the second inequality we use the fact that $P(\tilde{w}^k) \ge P(w_*)$ for all $k \in \{0, 1, \ldots, K\}$.

From (35) and (19), it follows that

$$\mathcal{G}_{U_k^{-1}}(w_t^k) = U_k^{-1}\left(w_t^k - \operatorname{prox}_R^{U_k^{-1}}\left(w_t^k - U_k \nabla F(w_t^k)\right)\right) = U_k^{-1}(w_t^k - \bar{w}_{t+1}^k).$$

By $0 \prec U_k \preceq 1/(3L_\Omega)\mathbb{I}_d$, we have

$$\|\bar{w}_{t+1}^k - w_t^k\|^2_{(\frac{1}{2}U_k^{-1} - L_\Omega \mathbb{I}_d)} = \|U_k \mathcal{G}_{U_k^{-1}}(w_t^k)\|^2_{(\frac{1}{2}U_k^{-1} - L_\Omega \mathbb{I}_d)}$$

$$
\begin{aligned}
&= \mathcal{G}_{U_k^{-1}}(w_t^k)^T U_k^T (\frac{1}{2} U_k^{-1} - L_\Omega \mathbb{I}_d) U_k \mathcal{G}_{U_k^{-1}}(w_t^k) \\
&\geq \mathcal{G}_{U_k^{-1}}(w_t^k)^T U_k^T (\frac{1}{6} U_k^{-1}) U_k \mathcal{G}_{U_k^{-1}}(w_t^k) \\
&= \frac{1}{6} \|\mathcal{G}_{U_k^{-1}}(w_t^k)\|_{U_k}^2.
\end{aligned}
$$

Combining the above inequality with (44), we get

$$
\sum_{k=0}^{K-1} \sum_{t=1}^{t_k} \frac{1}{6} \tilde{\mathbb{E}}[\|\mathcal{G}_{U_k^{-1}}(w_t^k)\|_{U_k}^2] \leq P(\tilde{w}^0) - P(w_*). \tag{45}
$$

Then we obtain the desired result by the definitions of w_a and T. □

4 Numerical experiments

In this section, we present experimental results on the following elastic net regularized logistic regression problem

$$
\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-b_i a_i^T w)) + \frac{\lambda_2}{2} \|w\|_2^2 + \lambda_1 \|w\|_1, \tag{46}
$$

which is usually employed in machine learning for binary classification. All the tests were performed with $R(w) = \lambda_1 \|w\|_1$ and

$$
f_i(w) = \log(1 + \exp(-b_i a_i^T w)) + \frac{\lambda_2}{2} \|w\|_2^2.
$$

Four publicly available data sets ijcnn1, rcv1, real-sim and covtype, which can be downloaded from the LIBSVM website [1], were tested. Table 2 lists the detailed information of these four data sets, including their sizes n, dimensions d, and Lipschitz constants L. Moreover, the values of regularization parameters λ_1 and λ_2 used in our experiments are also listed in Table 2. Notice that the choices of regularization parameters are typical in machine learning benchmarks to obtain good classification performance, see [15] for example.

Table 2 Data sets and parameters used in numerical experiments

Data sets	n	d	λ_2	λ_1	L
ijcnn1	49,990	22	10^{-4}	10^{-5}	0.9842
rcv1	20,242	47,236	10^{-4}	10^{-5}	0.2501
real-sim	72,309	20,958	10^{-4}	10^{-5}	0.2501
covtype	581,012	54	10^{-5}	10^{-4}	1.9040

For fair comparison, all methods were implemented in Matlab 2018b, and the experiments were conducted on a laptop with an Intel Core i7, 1.80 GHz processor and 16 GB of RAM running Windows 10 system. In Figures 1-3, the

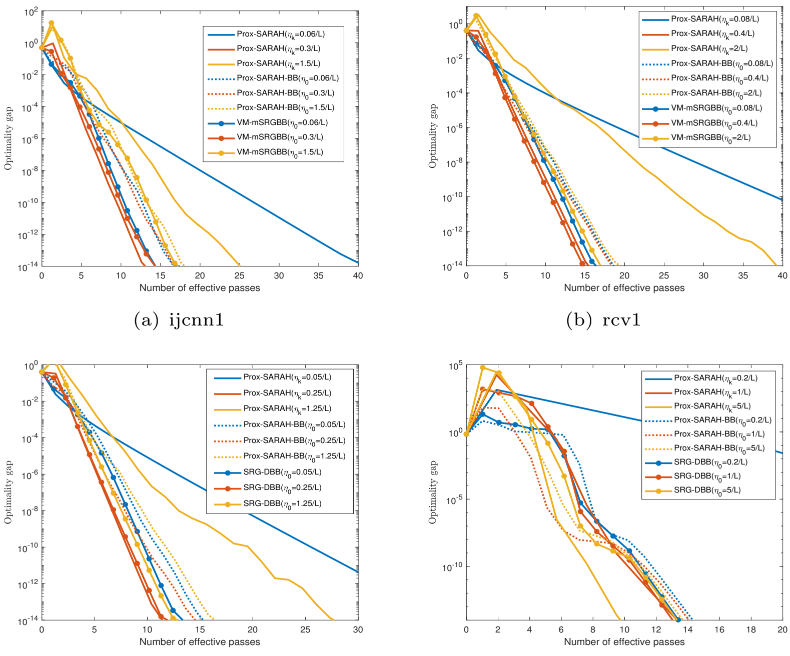
[1] https://www.csie.ntu.edu.tw/~cjlin/libsvm

x-axis is the number of effective passes over the data, where the evaluation of n component gradients counts as one effective pass. The y-axis with "optimality gap" denotes the value $P(\tilde{w}^k) - P(w_*)$ with w_* obtained by running the proximal SARAH with best-tuned fixed stepsizes.

4.1 Comparison with proximal variants of SARAH and SARAH-BB

This subsection presents the results of SRG-DBB with $b = 1$ for solving (46) on the four data sets listed in Table 2. Proximal SARAH (Prox-SARAH) and the proximal version of SARAH-BB (Prox-SARAH-BB) were also run for comparison. Notice that the SARAH-BB method is proposed to solve problem (1) with $R(w) = 0$. In order to solve the nonsmooth problem (46), the proximal operator was incorporated to obtain the Prox-SARAH-BB method. The best-tuned m were employed by Prox-SARAH and Prox-SARAH-BB.

It can be seen from Figure 1 that SRG-DBB often performs better than Prox-SARAH with different initial stepsizes. Unlike Prox-SARAH, SRG-DBB is not sensitive to the choice of initial stepsize, which would save much time on choosing initial stepsize so that it has promising potential in practice. Moreover, for different initial stepsizes, SRG-DBB performs better than Prox-SARAH-BB.



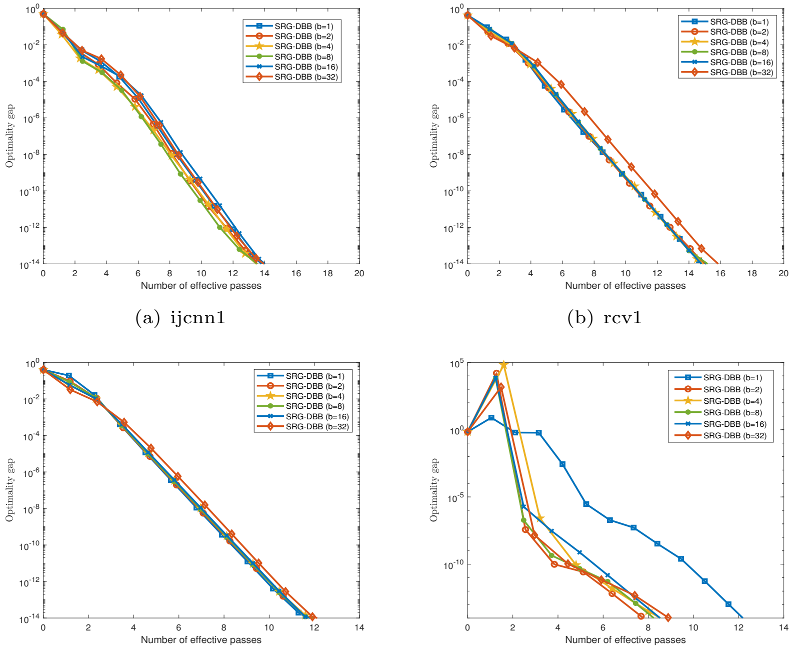
(a) ijcnn1 (b) rcv1

(c) real-sim (d) covtype

Fig. 1 Comparison of SRG-DBB, Prox-SARAH and Prox-SARAH-BB with different initial stepsizes

4.2 Properties of SRG-DBB with different mini-batch sizes

Figure 2 illustrates the results of SRG-DBB under various mini-batch sizes b on the four data sets. We can see that compared with $b = 1$, SRG-DBB has better or same performance by increasing the mini-batch size to $b = 2, 4, 8, 16, 32$.



(a) ijcnn1 (b) rcv1

(c) real-sim (d) covtype

Fig. 2 Comparison of SRG-DBB with different mini-batch sizes

4.3 Comparison with other algorithms

In this part, we conduct experiments on SRG-DBB with $b = 4$ in comparison with Prox-SVRG in [15] and four modern mini-batch proximal stochastic gradient methods, which are specified as follows.

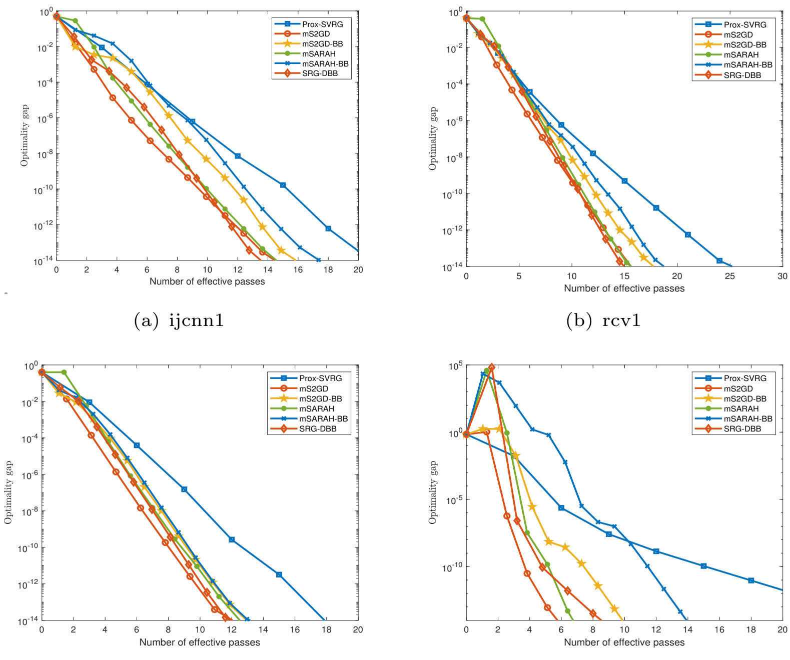
(1) mS2GD: mS2GD is a mini-batch proximal version of S2GD [16] to deal with nonsmooth problems. In mS2GD, a constant stepsize was used.
(2) mS2GD-BB: mS2GD-BB uses (7) to compute stepsizes for mS2GD.
(3) mSARAH: mSARAH is a mini-batch proximal variant of stochastic recursive gradient algorithm proposed in [14]. In mSARAH, a constant stepsize was used.
(4) mSARAH-BB: mSARAH-BB is a mini-batch variant of SARAH-BB [23].

Parameters suggested in [15] were used by Prox-SVRG. For the above four methods, we set $b = 4$. The best choices of parameters employed by SRG-DBB and the compared five methods are given in Table 3, including m for mS2GD, mS2GD-BB, mSARAH, mSARAH-BB and SRG-DBB, as well as the best-tuned stepsize η for mS2GD and mSARAH.

Table 3 Best choices of parameters for the methods

Parameter	ijcnn1	rcv1	real-sim	covtype
mS2GD (m, η)	$(\frac{1.0}{L}, 0.06n)$	$(\frac{1.5}{L}, 0.11n)$	$(\frac{0.7}{L}, 0.07n)$	$(\frac{21}{L}, 0.07n)$
mS2GD-BB	$0.06n$	$0.03n$	$0.02n$	$0.01n$
mSARAH (m, η)	$(\frac{1.1}{L}, 0.06n)$	$(\frac{1.6}{L}, 0.13n)$	$(\frac{1.0}{L}, 0.1n)$	$(\frac{25}{L}, 0.07n)$
mSARAH-BB	$0.06n$	$0.03n$	$0.02n$	$0.01n$
SRG-DBB	$0.04n$	$0.08n$	$0.04n$	$0.15n$

Figure 3 demonstrates that our SRG-DBB is better than or competitive with the compared algorithms on the four data sets.



(a) ijcnn1

(b) rcv1

(c) real-sim

(d) covtype

Fig. 3 Comparison of SRG-DBB and other modern methods

5 Conclusion

Based on a diagonal BB stepsize, we proposed a mini-batch proximal stochastic recursive gradient method named SRG-DBB to minimize the composition of two convex functions. Linear convergence of SRG-DBB was established in strongly convex and non-strongly convex cases, respectively. We further analyzed the sublinear convergence of SRG-DBB for the general convex function. Numerical comparisons of SRG-DBB and recent successful variance reduced stochastic gradient methods on some real data sets highly suggest the potential benefits of our SRG-DBB method for composition optimization problems arising in machine learning. Due to the popularity of deep learning, the nonsmooth nonconvex problems have attracted more and more attention. It would be interesting to explore the convergence of the SRG-DBB algorithm in the nonconvex case.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (Grant Nos. 11671116, 11701137, 12071108, 11631013, 11991020 and 12021001), the Major Research Plan of the NSFC (No. 91630202), Beijing Academy of Artificial Intelligence (BAAI), and Natural Science Foundation of Hebei Province (Grant No. A2021202010).

Declarations

- The authors declare that they have no conflict of interest.
- Not applicable, because this article does not contain any studies with human or animal subjects.

References

[1] Bottou, L., Curtis, F.E., Nocedal, J.: Optimization methods for large-scale machine learning. SIAM Rev. **60**(2), 223–311 (2018)

[2] Shalev-Shwartz, S., Ben-David, S.: Understanding Machine Learning: From theory to algorithms. Cambridge University Press, NY, USA (2014)

[3] Sra, S., Nowozin, S., Wright, S.J.: Optimization for Machine Learning. MIT Press, Cambridge, London, England (2012)

[4] Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Science & Business Media, NY, USA (2009)

[5] Recht, B., Ré, C.: Parallel stochastic gradient algorithms for large-scale matrix completion. Math. Prog. Comp. **5**(2), 201–226 (2013)

[6] Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y.: Deep Learning. MIT Press, Cambridge, London, England (2016)

[7] Li, X.L.: Preconditioned stochastic gradient descent. IEEE T. Neur. Net. Lear. **29**(5), 1454–1466 (2017)

[8] Zhang, S., Choromanska, A.E., LeCun, Y.: Deep learning with elastic averaging SGD. In: Advances in Neural Information Processing Systems, pp. 685–693 (2015)

[9] Jin, X.B., Zhang, X.Y., Huang, K., Geng, G.G.: Stochastic conjugate gradient algorithm with variance reduction. IEEE T. Neur. Net. Lear. **30**(5), 1360–1369 (2018)

[10] Robbins, H., Monro, S.: A stochastic approximation method. Ann. Math. Stat. **22**(3), 400–407 (1951)

[11] Roux, N.L., Schmidt, M., Bach, F.R.: A stochastic gradient method with an exponential convergence rate for finite training sets. In: Advances in Neural Information Processing Systems, pp. 2663–2671 (2012)

[12] Schmidt, M., Le Roux, N., Bach, F.: Minimizing finite sums with the stochastic average gradient. Math. Program. **162**(1-2), 83–112 (2017)

[13] Johnson, R., Zhang, T.: Accelerating stochastic gradient descent using predictive variance reduction. In: Advances in Neural Information Processing Systems, pp. 315–323 (2013)

[14] Nguyen, L.M., Liu, J., Scheinberg, K., Takáč, M.: SARAH: A novel method for machine learning problems using stochastic recursive gradient. In: Proceedings of the 34th International Conference on Machine, pp. 2613–2621 (2017)

[15] Xiao, L., Zhang, T.: A proximal stochastic gradient method with progressive variance reduction. SIAM J. Optim. **24**(4), 2057–2075 (2014)

[16] Konečnỳ, J., Richtárik, P.: Semi-stochastic gradient descent methods. Front. Appl. Math. Stat. **3**(9), 1–14 (2017)

[17] Konečnỳ, J., Liu, J., Richtárik, P., Takáč, M.: Mini-batch semi-stochastic gradient descent in the proximal setting. IEEE JSTSP **10**(2), 242–255 (2015)

[18] Pham, N.H., Nguyen, L.M., Phan, D.T., Tran-Dinh, Q.: ProxSARAH: An efficient algorithmic framework for stochastic composite nonconvex optimization. J. Mach. Learn. Res. **21**(110), 1–48 (2020)

[19] Barzilai, J., Borwein, J.M.: Two-point step size gradient methods. IMA J. Numer. Anal. **8**(1), 141–148 (1988)

[20] Dai, Y.H., Huang, Y., Liu, X.W.: A family of spectral gradient methods for optimization. Comput. Optim. Appl. **74**(1), 43–65 (2019)

[21] Fletcher, R.: On the Barzilai–Borwein method. In: Qi, L., Teo, K., Yang, X. (eds.) Optimization and Control with Applications, vol 96, pp. 235–256. Springer, Boston, USA (2005)

[22] Tan, C., Ma, S., Dai, Y.H., Qian, Y.: Barzilai-Borwein step size for stochastic gradient descent. In: Advances in Neural Information Processing Systems, pp. 685–693 (2016)

[23] Liu, Y., Wang, X., Guo, T.: A linearly convergent stochastic recursive gradient method for convex optimization. Optim. Lett. **14**, 2265–2283 (2020)

[24] Yu, T., Liu, X.W., Dai, Y.H., Sun, J.: A minibatch proximal stochastic recursive gradient algorithm using a trust-region-like scheme and Barzilai–Borwein stepsizes. IEEE T. Neur. Net. Lear. **32**(10), (2021)

[25] Park, Y., Dhar, S., Boyd, S., Shah, M.: Variable metric proximal gradient method with diagonal Barzilai–Borwein stepsize. In: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3597–3601 (2020)

[26] Yu, T., Liu, X.W., Dai, Y.H., Sun, J.: Variable metric proximal stochastic variance reduced gradient methods for nonconvex nonsmooth optimization. J. Ind. Manag. Optim. (2021). https://www.aimsciences.org/article/doi/10.3934/jimo.2021084

[27] Wang, X., Wang, S., Zhang, H.: Inexact proximal stochastic gradient method for convex composite optimization. Comput. Optim. Appl. **68**(3), 579–618 (2017)

[28] Wang, X., Wang, X., Yuan, Y.X.: Stochastic proximal quasi-newton methods for non-convex composite optimization. Optim. Method Softw. **34**(5), 922–948 (2019)

[29] Nesterov, Y.: Introductory Lectures on Convex Programming. Springer, Boston, MA, USA (1998)

[30] Beck, A.: First-order Methods in Optimization. SIAM, Philadelphia, PA, USA (2017)

[31] Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**(1), 183–202 (2009)

[32] Karimi, H., Nutini, J., Schmidt, M.: Linear convergence of gradient and proximal-gradient methods under the Polyak-Łojasiewicz condition. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 795–811 (2016)

[33] Gong, P., Ye, J.: Linear convergence of variance-reduced stochastic gradient without strong convexity. arXiv preprint. https://arxiv.org/abs/1406.1102 (2014). Accessed 4 June 2014

[34] Zhang, H.: The restricted strong convexity revisited: analysis of equivalence to error bound and quadratic growth. Optim. Lett. **11**(4), 817–833 (2017)

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

[36] Reddi, S.J., Sra, S., Poczos, B., Smola, A.J.: Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. In: Advances in Neural Information Processing Systems, pp. 1145–1153 (2016)