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

A linearly convergent stochastic recursive gradient method for convex optimization

Yan Liu [1,2] · Xiao Wang [1,2] · Tiande Guo [1,2]

Received: date / Accepted: date

Abstract The stochastic recursive gradient algorithm (SARAH) [8] attracts much interest recently. It admits a simple recursive framework for updating stochastic gradient estimates. Motivated by this, in this paper, we propose a SARAH-I method incorporating importance sampling, whose linear convergence rate of the sequence of distances between iterates and the optima set is proven under both strong convexity and non-strong convexity conditions. Further, we propose to use the Barzilai-Borwein (BB) method to automatically compute step sizes for SARAH-I, named as SARAH-I-BB, and we establish its convergence and complexity properties in different cases. Finally numerical tests are reported to indicate promising performances of SARAH-I-BB.
Keywords: Stochastic optimization, stochastic gradient, BB method, linear convergence rate, complexity

1 Introduction

In the context of large scale machine learning, the following type of optimization problems is widely considered:

$$\min_{w\in\mathbb{R}^d} \quad P(w) \equiv \frac{1}{n}\sum_{i=1}^{n} f_i(w), \tag{1}$$

Yan Liu
E-mail: liuyan23ucas@outlook.com

✉Xiao Wang
E-mail: wangxiao@ucas.ac.cn

Tiande Guo
E-mail: tdguo@ucas.ac.cn

1 School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing, 100049, China
2 Key Laboratory of Big Data Mining and Knowledge Management, University of Chinese Academy of Sciences, Beijing, 100049, China

where n is the sample size, and each f_i, $i \in \{1, 2, \ldots, n\}$ is first-order continuously differentiable.

The problem (1) is challenging when n is extremely large. Because the exact full gradient information is not easy to obtain, exact gradient-based methods are impractical and prohibited. Stochastic gradient descent (SGD) method, however, traced back to the seminal work by [1], has become the dominating approach for solving (1). In the t-th iteration, SGD picks an index $i \in [n]$ at random, and updates the iterate w_t by

$$w_{t+1} = w_t - \eta_t \nabla f_i(w_t),$$

where $\eta_t > 0$ is the step size, and $\nabla f_i(w_t)$ denotes the sample gradient.

A surge of methods to improve the performance of SGD have been developed. One type of most prevalent methods are gradient aggregation algorithms [3], such as SAG [4][5], SAGA [6]. They compute a stochastic gradient as an average of stochastic gradients evaluated at previous iterates. Then they store previous stochastic gradients at the expense of memory. SVRG [7] has two loops, with a full gradient computed in the outer loop (each outer iteration is called an epoch) and lower variance stochastic gradients computed in the inner loop. S2GD [14] runs a random number of stochastic gradients, following a geometric law, in each epoch. Batching SVRG [13] chooses a large set of batch samples to approximate full gradient in every outer loop. Gradient estimator in algorithms mentioned above are unbiased. There are also some biased estimators with nice performances. SARAH [8] and iSARAH [9] admit a simple recursive framework for updating stochastic gradient estimates. A new technique named SPIDER [11] was proposed to find an approximate stationary piont for non-convex stochastic optimization problem, which outperforms existing algorithms of the same type. As an improved SPIDER scheme, SpiderBoost [12] allows much more flexibility for choosing parameters. Gradient aggregation algorithms mentioned above are now widely used in the machine learning community for solving (1). They can achieve linear convergence rate in the strong convexity case.

In practical computation, the performance of SGD is affected by the step size. One common strategy is to adopt diminishing step sizes that satisfy

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

AdaGrad [16] and Adam [17] adaptively select the step size for every component based on the sum of the squares of the past gradients. As is well known, the Barzilai-Borwein (BB) approach [21] has the advantage of adaptively updating the step size and also can capture the hidden second order information, thus achieves very promising performance when solving optimization problems. SVRG-BB [19] incorporates the BB step size to automatically compute step size for SVRG. SA-GD [15] computes an adaptive step size based on local norm at the current point for self-concordant functions. Adaptive batch SGD

[18] considers a backtracking variant of SGD that adaptively tunes the step size and they derive an adaptive step size using the BB estimate too.

Another factor to affect the performance of SGD is how to sample in training. The advantage of uniform sampling of data lies in that the sampled stochastic gradient is an unbiased estimate of the true gradient (full gradient). But it may have a rather high variance which negatively affects the convergence of optimization procedure. Stochastic optimization with importance sampling can improve the convergence rate by reducing the stochastic variance. Such strategy is widely used in many works, such as Proximal SVRG (Prox-SVRG) [24] and Proximal Stochastic Dual Coordinate Ascent (prox-SDCA) [25].

Contributions Our contributions in this paper lie in the following several folds.

1) We study a SARAH-I method with importance sampling in which the full gradient of the last iterate in each inner loop is calculated. We establish the linear convergence rate of the sequence of distances between iterates and the minimizer when solving strongly convex problem (1). Similar algorithm was also considered in [8]. However, no convergence analysis was given there.
2) We further prove the linear convergence of SARAH-I for non-strongly convex optimization. Under restricted secant inequality [28], we prove the expectation of the distance of iterates to the optimal solution set is linearly convergent. As far as we know, this is the first property related with SARAH in convex settings in literatures.
3) SARAH-I-BB for strongly and non-strongly convex optimization, which is inspired by SVRG-BB [19]. We give the convergence analysis and complexity analysis of SARAH-I-BB. Numerical experiments to demonstrate the validity of SARAH-I-BB for strongly convex function is performed. The numerical results indicate that SARAH-I-BB is comparable to SARAH-I with best-tuned step sizes.

Organizations An outline of this paper is as follows. In Section 2, we present the SARAH-I algorithm with importance sampling strategy. The convergence analysis for strongly convex function are stated in Section 3. Then we analyze its convergence properties in non-strongly convex case in Section 4. We propose a SARAH-I-BB method by using BB approach to adaptively update step sizes in Section 5. Numerical experiments are then reported in Section 6. Finally, we draw some conclusions in Section 7.

2 The SARAH-I method

The original SARAH [8] method updates the stochastic step direction v_t recursively by adding and subtracting component gradients to and from the previous v_{t-1} in the inner loop. It contains outer loops to compute full gradient. We consider here the random sampling from a general distribution $Q \sim \{q_1, q_2, \dots q_n\}$, which is more flexible than the uniform sampling scheme. The key step of

SARAH-I is to compute

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

Conditioned on w_{t-1}, we take expectation on v_t with respect to i_t obtaining

$$\begin{aligned}\mathbb{E}[v_t] &= \mathbb{E}[\frac{\nabla f_{i_t}(w_t) - \nabla f_{i_t}(w_{t-1})}{nq_{i_t}} + v_{t-1}] \\ &= \sum_{i=1}^{n} \frac{q_{i_t}}{nq_{i_t}} \left(\nabla f_{i_t}(w_t) - \nabla f_{i_t}(w_{t-1})\right) + v_{t-1} \\ &= \nabla P(w_t) - \nabla P(w_{t-1}) + v_{t-1}\end{aligned} \tag{2}$$

We can see that SARAH is different from the aforementioned algorithms since it has a biased estimator of gradient by (2). The pseudocode is outlined as Algorithm 1.

Algorithm 1 SARAH-I

Parameters: update frequency m, step size $\eta > 0$, initial point $\tilde{w}_0$

1: **for** $k = 1, 2, \ldots$ **do**
2: $w_0 = \tilde{w}_{k-1}$
3: $v_0 = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\tilde{w}_0)$
4: $w_1 = w_0 - \eta v_0$
5: **for** $t = 1, \ldots, m-1$ **do**
6: Randomly pick $i_t \in \{1, \ldots, n\}$ according to Q
7: $v_t = (\nabla f_{i_t}(w_t) - \nabla f_{i_t}(w_{t-1}))/(nq_{i_t}) + v_{t-1}$
8: $w_{t+1} = w_t - \eta v_t$
9: **end for**
10: $\tilde{w}_k = w_m$
11: **end for**

Notice that in step 10 it computes $\tilde{w}$ as the last iterate of each inner iteration, which is the most important difference from the original SARAH [8]. Step 10 seems a more reasonable choice, because the latest information in each inner loop is used. SARAH instead sets $\tilde{w}$ as a uniformly randomly picked iterate from inner iteration and enjoys linear convergence rate for strongly convex optimization. We here cite the convergence property of SARAH as follows.

Theorem 1 *(Theorem 4 [8]) Suppose that f_i, $i = 1, \ldots, n$, is L-smooth and convex, and P is μ-strongly convex. Consider SARAH with η and m such that*

$$\delta \triangleq \frac{1}{\mu\eta(m+1)} + \frac{\eta L}{2 - \eta L} < 1.$$

Then, we have

$$\mathbb{E}[\|\nabla P(\tilde{w}_k)\|^2] \leqslant \delta^k \|\nabla P(\tilde{w}_0)\|^2.$$

A practical variant SARAH+ with $\tilde{w}$ computed same as step 10 is also considered in [8], however, no convergence analysis is given there. [19] studies the same strategy but integrated in SVRG framework [7], named as SVRG-I. But in order to analyze the convergence property of SARAH-I, it is non-trivial to use analysis techniques from SVRG-I and any existing stochastic recursive gradient algorithm.

To establish some basic properties of SARAH-I under the convexity setting, we first make the following assumption used throughout this paper.

Assumption 1 *Each f_i, $i = 1 \ldots, n$, is convex and first-order continuously differentiable. And the gradient of each component function f_i is L-Lipschitz continuous, i.e.,*

$$\|\nabla f_i(w) - \nabla f_i(w')\|_2 \leqslant L\|w - w'\|_2, \quad \forall w, w' \in \mathbb{R}^d.$$

Under this assumption, it is easy to see that $\nabla P(w)$ is also L-Lipschitz continuous:

$$\|\nabla P(w) - \nabla P(w')\|_2 \leqslant L\|w - w'\|_2, \quad \forall w, w' \in \mathbb{R}^d.$$

For simplicity, we denote L_Q as

$$L_Q = \max_i \frac{L}{nq_i}.$$

Then it is easy to find that $L_Q \geqslant L$. The following property is a straightforward result of the Assumption 1.

Lemma 1 *(Theorem 2.1.5 [2]). Suppose that f is convex and ∇f is L-Lipschitz continuous. Then for any $w, w' \in \mathbb{R}^d$,*

$$(\nabla f(w) - \nabla f(w'))^T (w - w') \geqslant \frac{1}{L}\|\nabla f(w) - \nabla f(w')\|^2,$$

Lemma 2 *Suppose that Assumption 1 holds and $\eta < \frac{2}{L_Q}$. Then in the k-th epoch of SARAH-I, for all $t \geq 1$ we have*

$$\mathbb{E}[\|v_t\|^2] \leqslant \mathbb{E}[\|v_{t-1}\|^2],$$

where the expectation is taken with respect to all the variables generated in the k-th epoch.

Proof For all $t \geqslant 1$, we have

$$\begin{aligned}
\mathbb{E}[\|v_t\|^2] &= \mathbb{E}[\left\| v_{t-1} - \frac{1}{nq_{i_t}} \left(\nabla f_{i_t}(w_{t-1}) - \nabla f_{i_t}(w_t)\right)\right\|^2] \\
&= \mathbb{E}[\|v_{t-1}\|^2] + \mathbb{E}[\frac{1}{(nq_{i_t})^2}\|\nabla f_{i_t}(w_{t-1}) - \nabla f_{i_t}(w_t)\|^2] \\
&\quad - \mathbb{E}[\frac{2}{\eta} \cdot \frac{1}{nq_{i_t}} (\nabla f_{i_t}(w_{t-1}) - \nabla f_{i_t}(w_t))^T (w_{t-1} - w_t)]
\end{aligned}$$

$$
\begin{aligned}
&\leqslant \mathbb{E}[\|v_{t-1}\|^2] + \mathbb{E}[\frac{1}{(nq_{i_t})^2}\|\nabla f_{i_t}(w_{t-1}) - \nabla f_{i_t}(w_t)\|^2] \\
&\quad - \mathbb{E}[\frac{2}{\eta L}\cdot\frac{1}{nq_{i_t}}\|\nabla f_{i_t}(w_{t-1}) - \nabla f_{i_t}(w_t)\|^2] \\
&\leqslant \mathbb{E}[\|v_{t-1}\|^2] + \left(1 - \frac{2}{\eta L_Q}\right)\mathbb{E}[\|\frac{1}{nq_{i_t}}(\nabla f_{i_t}(w_{t-1}) - \nabla f_{i_t}(w_t))\|^2] \\
&\leqslant \mathbb{E}[\|v_{t-1}\|^2] + \left(1 - \frac{2}{\eta L_Q}\right)\mathbb{E}[\|v_t - v_{t-1}\|^2].
\end{aligned}
$$

where the first inequality uses the Lemma 1. Notice that $1 - \frac{2}{\eta L_Q} < 0$ since $\eta < \frac{2}{L_Q}$, therefore $\mathbb{E}[\|v_t\|^2]$ is decreasing.

Lemma 3 *Suppose that Assumption 1 holds and $\eta \leqslant \frac{1}{L_Q}$. Then in the k-th epoch of SARAH-I for all $t \geq 1$,*

$$\mathbb{E}[\|\nabla P(w_t) - v_t\|^2] \leq \eta L_Q^3 \mathbb{E}[\|w_0 - w_*\|^2]. \tag{3}$$

Proof Following from the proof of Lemma 2, we have

$$\mathbb{E}[\|v_t\|^2] \leqslant \mathbb{E}[\|v_{t-1}\|^2] + \left(1 - \frac{2}{\eta L_Q}\right)\mathbb{E}[\|v_t - v_{t-1}\|^2],$$

which implies that

$$\mathbb{E}[\|v_t - v_{t-1}\|^2] \leqslant \frac{\eta L_Q}{2 - \eta L_Q}[\mathbb{E}[\|v_{t-1}\|^2] - \mathbb{E}[\|v_t\|^2]] \leqslant \eta L_Q[\mathbb{E}[\|v_{t-1}\|^2] - \mathbb{E}[\|v_t\|^2]],$$

when $\eta \leqslant \frac{1}{L_Q}$. By summing the above inequality over $j = 1, 2, \ldots, t$ $(t \geqslant 1)$, we have

$$\sum_{j=1}^{t}\mathbb{E}[\|v_j - v_{j-1}\|^2] \leqslant \eta L_Q[\mathbb{E}[\|v_0\|^2] - \mathbb{E}[\|v_t\|^2]],$$

Notice that Lemma 2 in [8] shows that

$$\mathbb{E}[\|\nabla P(w_t) - v_t\|^2] = \sum_{j=1}^{t}\mathbb{E}[\|v_j - v_{j-1}\|^2] - \sum_{j=1}^{t}\mathbb{E}[\|\nabla P(w_j) - \nabla P(w_{j-1})\|^2].$$

Consequently, It implies (3) from the fact that

$$\|v_0\|^2 = \|\nabla P(w_0)\|^2 \leqslant L^2\|w_0 - w_*\|^2 \leqslant L_Q^2\|w_0 - w_*\|^2.$$

3 SARAH-I for Strongly Convex Optimization

We analyze the linear convergence of SARAH-I when P is strongly convex in this section. That is we assume that each f_i is convex and the objective function $P(w)$ is μ-strongly convex, i.e.,

$$P(w) \geqslant P(w') + \nabla P(w')^T (w - w') + \frac{\mu}{2} \|w - w'\|^2, \quad \forall w, w' \in \mathbb{R}^d$$

Denote w_* as the optimal solution of (1). Then due to the strong convexity of P, w_* is unique. We now cite the following useful lemma.

Lemma 4 *(Theorem 2.1.12 [2]) Suppose that P is μ-strongly convex and $\nabla P(w)$ is Lipschitz continuous with the constant L, then for any $w, w' \in \mathbb{R}^d$ we have*

$$(P(w) - P(w'))^T (w - w') \geqslant \frac{\mu L}{\mu + L} \|w - w'\|^2 + \frac{1}{\mu + L} \|\nabla P(w) - \nabla P(w')\|^2.$$

Lemma 5 *Suppose that Assumption 1 holds and $\eta < \frac{2}{L_Q}$. If P is μ-strongly convex, then in the k-th epoch of SARAH-I, we have for all $t \geq 1$,*

$$\mathbb{E}[\|v_t\|^2] \leqslant [1 - \left(\frac{2}{\eta L_Q} - 1\right) \mu^2 \eta^2] \mathbb{E}[\|v_{t-1}\|^2] \leqslant [1 - \left(\frac{2}{\eta L_Q} - 1\right) \mu^2 \eta^2]^t \mathbb{E}[\|v_0\|^2]. \tag{4}$$

Proof The proof of Lemma 2 indicates that

$$\begin{aligned}
\mathbb{E}[\|v_t\|^2] &\leqslant \|v_{t-1}\|^2 + \left(1 - \frac{2}{\eta L_Q}\right) \mathbb{E}[\|v_t - v_{t-1}\|^2] \\
&\leqslant \|v_{t-1}\|^2 + \left(1 - \frac{2}{\eta L_Q}\right) \|\nabla P(w_t) - \nabla P(w_{t-1})\|^2 \\
&\leqslant \|v_{t-1}\|^2 + \left(1 - \frac{2}{\eta L_Q}\right) \mu^2 \eta^2 \|v_{t-1}\|^2.
\end{aligned}$$

The first inequality uses

$$\|\nabla P(w_t) - \nabla P(w_{t-1})\|^2 = \|\mathbb{E}[v_t - v_{t-1}]\|^2 \leqslant \mathbb{E}[\|v_t - v_{t-1}\|^2],$$

because for any random vector $\zeta \in \mathbb{R}^d$, it holds that $\mathbb{E}\|\zeta - \mathbb{E}\zeta\|^2 = \mathbb{E}\|\zeta\|^2 - \|\mathbb{E}\zeta\|^2 \geqslant 0$. The last inequality follows by the strong convexity of P and the fact that $w_t = w_{t-1} - \eta v_{t-1}$. Then recursively we obtain (4).

Lemma 6 *Suppose that Assumption 1 holds and $\eta \leqslant \frac{1}{L_Q}$. If P is μ-strongly convex, then in the k-th epoch of SARAH-I for all $t \geq 1$,*

$$\left(\mathbb{E}[\|w_t - w_*\|^2]\right)^{\frac{1}{2}} \leqslant \left(1 + \frac{\eta^{1/2} L_Q^{3/2}}{\mu}\right) \left(\mathbb{E}[\|w_0 - w_*\|^2]\right)^{\frac{1}{2}}.$$

Proof For ease of notations, we let $v_t = \nabla P(w_t) + e_t$. Then we get

$$\begin{aligned}
\|w_t - w_*\|^2 &= \|w_{t-1} - \eta v_{t-1} - w_*\|^2 \\
&= \|w_{t-1} - \eta(\nabla P(w_{t-1}) + e_{t-1}) - w_* + \eta \nabla P(w_*)\|^2 \\
&= \|w_{t-1} - w_* - \eta(\nabla P(w_{t-1}) - \nabla P(w_*))\|^2 + \eta^2 \|e_{t-1}\|^2 \\
&\quad - 2\eta e_{t-1}^T (w_{t-1} - w_* - \eta(\nabla P(w_{t-1}) - \nabla P(w_*))) \\
&\leqslant \|w_{t-1} - w_* - \eta(\nabla P(w_{t-1}) - \nabla P(w_*))\|^2 + \eta^2 \|e_{t-1}\|^2 \\
&\quad + 2\eta \|e_{t-1}\| \|w_{t-1} - w_* - \eta(\nabla P(w_{t-1}) - \nabla P(w_*))\|,
\end{aligned}$$

where we use $\nabla P(w_*) = 0$ in the second equality and the Cauchy-Schwartz inequality in the last inequality.

We now need to bound $\|w_{t-1} - w_* - \eta(\nabla P(w_{t-1}) - \nabla P(w_*))\|$ to get the final result. Notice that

$$\begin{aligned}
&\|w_{t-1} - w_* - \eta(\nabla P(w_{t-1}) - \nabla P(w_*))\|^2 \\
&= \|w_{t-1} - w_*\|^2 + \eta^2 \|\nabla P(w_{t-1}) - \nabla P(w_*)\|^2 \\
&\qquad - 2\eta (\nabla P(w_{t-1}) - \nabla P(w_*))^T (w_{t-1} - w_*) \\
&\leqslant \|w_{t-1} - w_*\|^2 + \eta^2 \|\nabla P(w_{t-1}) - \nabla P(w_*)\|^2 \\
&\qquad - 2\eta \left(\frac{1}{\mu + L} \|\nabla P(w_{t-1}) - \nabla P(w_*)\|^2 + \frac{\mu L}{\mu + L} \|w_{t-1} - w_*\|^2 \right) \\
&= \left(1 - 2\eta \cdot \frac{\mu L}{\mu + L}\right) \|w_{t-1} - w_*\|^2 + \eta \left(\eta - \frac{2}{\mu + L}\right) \|\nabla P(w_{t-1}) - \nabla P(w_*)\|^2 \\
&\leqslant \left(1 - 2\eta \cdot \frac{\mu L}{\mu + L}\right) \|w_{t-1} - w_*\|^2 + \eta \mu^2 \left(\eta - \frac{2}{\mu + L}\right) \|w_{t-1} - w_*\|^2 \\
&= (1 - \mu\eta)^2 \|w_{t-1} - w_*\|^2,
\end{aligned}$$

where the first inequality uses Lemma 4 and the second inequality uses the truth of $\eta \leqslant \frac{1}{L_Q} \leqslant \frac{1}{L} < \frac{2}{\mu + L}$ and the strong convexity of $P(w)$. Thus, by taking expectation on both side for the above inequality, we have

$$\begin{aligned}
&\mathbb{E}[\|w_t - w_*\|^2] \\
&\leqslant (1 - \mu\eta)^2 \mathbb{E}[\|w_{t-1} - w_*\|^2] + \eta^2 \mathbb{E}[\|e_{t-1}\|^2] \\
&\qquad + 2\eta (1 - \mu\eta) \mathbb{E}[\|e_{t-1}\| \|w_{t-1} - w_*\|] \\
&\leqslant (1 - \mu\eta)^2 \mathbb{E}[\|w_{t-1} - w_*\|^2] + \eta^2 \mathbb{E}[\|e_{t-1}\|^2] \\
&\qquad + 2\eta (1 - \mu\eta) \left(\mathbb{E}[\|w_{t-1} - w_*\|^2]\right)^{\frac{1}{2}} \left(\mathbb{E}[\|e_{t-1}\|^2]\right)^{\frac{1}{2}} \\
&\leqslant \left((1 - \mu\eta) \left(\mathbb{E}[\|w_{t-1} - w_*\|^2]\right)^{\frac{1}{2}} + \eta \left(\mathbb{E}[\|e_{t-1}\|^2]\right)^{\frac{1}{2}}\right)^2,
\end{aligned}$$

which implies

$$\left(\mathbb{E}[\|w_t - w_*\|^2]\right)^{\frac{1}{2}} \leqslant (1 - \mu\eta) \left(\mathbb{E}[\|w_{t-1} - w_*\|^2]\right)^{\frac{1}{2}} + \eta \left(\mathbb{E}[\|e_{t-1}\|^2]\right)^{\frac{1}{2}}.$$

Applying the bound recursively yields

$$\left(\mathbb{E}[\|w_t-w_*\|^2]\right)^{\frac{1}{2}} \leqslant (1-\mu\eta)^t\left(\mathbb{E}[\|w_0-w_*\|^2]\right)^{\frac{1}{2}}+\sum_{j=1}^{t}(1-\mu\eta)^{t-j}\eta\left(\mathbb{E}[\|e_j\|^2]\right)^{\frac{1}{2}}.$$

Consequently, we infer that

$$\begin{aligned}
&\left(\mathbb{E}[\|w_t-w_*\|^2]\right)^{\frac{1}{2}}\\
&\leqslant (1-\mu\eta)^t\left(\mathbb{E}[\|w_0-w_*\|^2]\right)^{\frac{1}{2}}+\sum_{j=1}^{t}(1-\mu\eta)^{t-j}\eta\left(\mathbb{E}[\|e_j\|^2]\right)^{\frac{1}{2}}\\
&\leqslant (1-\mu\eta)^t\left(\mathbb{E}[\|w_0-w_*\|^2]\right)^{\frac{1}{2}}+\sum_{j=1}^{t}(1-\mu\eta)^{t-j}\eta\left(\eta L_Q^3\right)^{\frac{1}{2}}\left(\mathbb{E}[\|w_0-w_*\|^2]\right)^{\frac{1}{2}}\\
&\leqslant (1-\mu\eta)^t\left(\mathbb{E}[\|w_0-w_*\|^2]\right)^{\frac{1}{2}}+\frac{1}{\mu\eta}\eta\left(\eta L_Q^3\right)^{\frac{1}{2}}\left(\mathbb{E}[\|w_0-w_*\|^2]\right)^{\frac{1}{2}}\\
&\leqslant \left(1+\frac{\eta^{1/2}L_Q^{3/2}}{\mu}\right)\left(\mathbb{E}[\|w_0-w_*\|^2]\right)^{\frac{1}{2}}
\end{aligned}$$

where the second inequality follows from Lemma 3.

Lemma 7 *Suppose that Assumption 1 holds and $\eta \leqslant \frac{1}{L_Q}$. If P is μ-strongly convex, then in the k-th epoch of SARAH-I, for all $t \geq 1$, we have*

$$\mathbb{E}[(w_t-w_*)^T(\nabla P(w_t)-v_t)] \leqslant \left(1+\frac{\eta^{1/2}L_Q^{3/2}}{\mu}\right)\cdot\eta^{1/2}L_Q^{3/2}\left(\mathbb{E}[\|w_0-w_*\|^2]\right).$$

Proof We are now in a position to prove

$$\begin{aligned}
&\mathbb{E}[(w_t-w_*)^T(\nabla P(w_t)-v_t)]\\
&\leqslant \mathbb{E}[\|w_t-w_*\|\,\|\nabla P(w_t)-v_t\|]\\
&\leqslant \left(\mathbb{E}[\|w_t-w_*\|^2]\right)^{\frac{1}{2}}\left(\mathbb{E}[\|\nabla P(w_t)-v_t\|^2]\right)^{\frac{1}{2}}\\
&\leqslant \left(1+\frac{\eta^{1/2}L_Q^{3/2}}{\mu}\right)\left(\mathbb{E}[\|w_0-w_*\|^2]\right)^{\frac{1}{2}}\cdot\left(\eta L_Q^3\right)^{\frac{1}{2}}\left(\mathbb{E}[\|w_0-w_*\|^2]\right)^{\frac{1}{2}}\\
&\leqslant \left(1+\frac{\eta^{1/2}L_Q^{3/2}}{\mu}\right)\cdot\eta^{1/2}L_Q^{3/2}\left(\mathbb{E}[\|w_0-w_*\|^2]\right).
\end{aligned}$$

We now achieve the linear convergence rate of SARAH-I in the strongly convex case as follow.

Theorem 2 *Suppose that Assumption 1 holds and $\eta \leqslant \frac{1}{L_Q}$. If P is μ-strongly convex, then for all $k \geq 1$, we have*

$$\mathbb{E}[\|\tilde{w}_{k+1}-w^*\|^2] \leqslant \sigma\|\tilde{w}_k-w_*\|^2,$$

where

$$\sigma := (1-2\mu\eta)^m + \frac{\eta^{1/2}L_Q^{3/2}}{\mu} + \frac{\eta L_Q^3}{\mu^2} + \frac{\eta L_Q^2}{2\mu}. \tag{5}$$

Consequently, if m and η are chosen such that $\sigma < 1$, then SARAH-I converges linearly in expectation, namely,

$$\mathbb{E}[\|\tilde{w}_k - w_*\|^2] \leqslant \sigma^k \|\tilde{w}_0 - w_*\|^2 .$$

Here, the expectation is taken with respect to random variables generated in the whole algorithm.

Proof In the k-th epoch of SARAH-I, we have

$$\begin{aligned}
&\mathbb{E}[\|w_{t+1} - w_*\|^2] \\
&= \mathbb{E}[\|w_t - \eta v_t - w_*\|^2] \\
&= \mathbb{E}[\|w_t - w_*\|^2] - 2\eta(w_t - w_*)^T\mathbb{E}[v_t] + \eta^2\mathbb{E}[\|v_t\|^2] \\
&= \mathbb{E}[\|w_t - w_*\|^2] - 2\eta[(w_t - w_*)^T\nabla P(w_t)] \\
&\qquad + 2\eta\mathbb{E}[(w_t - w_*)^T(\nabla P(w_t) - v_t)] + \eta^2\mathbb{E}[\|v_t\|^2] \\
&\leqslant \mathbb{E}[\|w_t - w_*\|^2] - 2\eta[(w_t - w_*)^T\nabla P(w_t)] \\
&\qquad + 2\eta\mathbb{E}[(w_t - w_*)^T(\nabla P(w_t) - v_t)] + \eta^2\mathbb{E}[\|\nabla P(w_0)\|^2] \\
&\leqslant \mathbb{E}[\|w_t - w_*\|^2] - 2\eta[(w_t - w_*)^T\nabla P(w_t)] \\
&\qquad + 2\left(1 + \frac{\eta^{1/2}L_Q^{3/2}}{\mu}\right)\cdot\eta^{3/2}L_Q^{3/2}\|w_0 - w_*\|^2 + L_Q^2\eta^2\|w_0 - w_*\|^2 \\
&\leqslant (1-2\mu\eta)\,\mathbb{E}[\|w_t - w_*\|^2] + \left(2\left(1 + \frac{\eta^{1/2}L_Q^{3/2}}{\mu}\right)\cdot\eta^{3/2}L_Q^{3/2} + L_Q^2\eta^2\right)\|w_0 - w_*\|^2 ,
\end{aligned}$$

where the first inequality applies the Lemma 5 and the second inequality is on account of Lemma 7, Lipschitz continuity of ∇P and $L_Q \geqslant L$. The last inequality uses the strong convexity of P. By recursively applying the above inequality over t, and noting that $\tilde{w}_k = w_0$ and $\tilde{w}_{k+1} = w_m$, we obtain

$$\begin{aligned}
&\mathbb{E}[\|\tilde{w}_{k+1} - w_*\|^2] \\
&\leqslant (1-2\mu\eta)^m\|\tilde{w}_k - w_*\|^2 \\
&\qquad + \left(2\left(1 + \frac{\eta^{1/2}L_Q^{3/2}}{\mu}\right)\cdot\eta^{3/2}L_Q^{3/2} + L_Q^2\eta^2\right)\sum_{j=0}^{m-1}(1-2\mu\eta)^j\|\tilde{w}_k - w_*\|^2 \\
&\leqslant [(1-2\mu\eta)^m + \frac{\eta^{1/2}L_Q^{3/2}}{\mu} + \frac{\eta L_Q^3}{\mu^2} + \frac{\eta L_Q^2}{2\mu}]\|\tilde{w}_k - w_*\|^2 \\
&= \sigma_k\|\tilde{w}_k - w_*\|^2 .
\end{aligned}$$

which completes the proof.

4 SARAH-I for Non-Strongly Convex Optimization

The strong convexity of the objective function has been the standard assumption for proving linear convergence of stochastic first-order methods in recent years, but this assumption does not hold for many problems in practice. Investigators propose some strictly weaker concepts, such as *weak strong convexity* (WSC) [30], *Polyak-ojasiewicz inequality* (PL) [26], *the error bound* (EB) [27], *the quadratic growth* (QG) [31] and *restricted secant inequality* (RSI) [28]. We now state them as follows. Here they all involve some constant $\nu > 0$. We denote W^* as the set of optimal solutions of problem (1) and w^{proj} as the projection of w onto W^*. It is simple to see that $\nabla P\left(w^{proj}\right) = 0$. We use P^* to represent the optimal value of (1). In the four conditions, we suppose that P is first-order continuously differentiable and ∇P is L-Lipschitz continuous.

1. **Polyak-ojasiewicz inequality** (PL):

$$\frac{1}{2}\|\nabla P(w)\|^2 \geqslant \nu\left(P(w) - P^*\right), \quad \forall w.$$

2. **Error Bound** (EB):

$$\|\nabla P(w)\| \geqslant \nu\left\|w - w^{proj}\right\|^2, \quad \forall w.$$

3. **Quadratic Growth** (QG):

$$P(w) - P^* \geqslant \frac{\nu}{2}\left\|w - w^{proj}\right\|^2, \quad \forall w.$$

4. **Restricted Secant Inequality** (RSI):

$$\left(\nabla P(w) - \nabla P\left(w^{proj}\right)\right)^T\left(w - w^{proj}\right) \geqslant \nu\left\|w - w^{proj}\right\|^2, \quad \forall w.$$

It is shown in [29] that when P is convex, all the above conditions are equivalent. Therefore, we only need to analyze the properties of SARAH-I when P is convex and satisfies RSI with $\nu > 0$.

Lemma 8 *(Lemma 3 [28]) If P is convex, $\nabla P(w)$ is L-Lipschitz continuous and P satisfies RSI with $\nu > 0$, then for any $\alpha \in [0, 1]$,*

$$\left(\nabla P(w) - \nabla P\left(w^{proj}\right)\right)^T\left(w - w^{proj}\right) \geqslant \frac{\alpha}{L}\left\|\nabla P(w) - \nabla P\left(w^{proj}\right)\right\|^2 + (1-\alpha)\nu\left\|w - w^{proj}\right\|^2.$$

Lemma 9 *Suppose that Assumption 1 holds and P is convex and satisfies RSI with $\nu > 0$. Let $\eta \leqslant \frac{1}{L_Q}$, then in the k-th epoch of SARAH-I, we have for all $t \geq 1$,*

$$\left(\mathbb{E}[\left\|w_t - w_t^{proj}\right\|^2]\right)^{\frac{1}{2}} \leqslant \left(1 + \frac{2\eta^{1/2}L_Q^{3/2}}{\nu}\right)\left(\mathbb{E}[\left\|w_0 - w_0^{proj}\right\|^2]\right)^{\frac{1}{2}}.$$

Proof Let $v_t = \nabla P(w_t) + e_t$, then we get

$$
\begin{aligned}
&\left\|w_t - w_t^{proj}\right\|^2 \\
&\leqslant \left\|w_t - w_{t-1}^{proj}\right\|^2 \\
&= \left\|w_{t-1} - \eta v_{t-1} - w_{t-1}^{proj}\right\|^2 \\
&= \left\|w_{t-1} - \eta\left(\nabla P(w_{t-1}) + e_{t-1}\right) - w_{t-1}^{proj} + \eta\nabla P\left(w_{t-1}^{proj}\right)\right\|^2 \\
&= \left\|w_{t-1} - w_{t-1}^{proj} - \eta\left(\nabla P(w_{t-1}) - \nabla P\left(w_{t-1}^{proj}\right)\right)\right\|^2 + \eta^2\|e_{t-1}\|^2 \\
&\qquad\qquad - 2\eta e_{t-1}^T\left(w_{t-1} - w_{t-1}^{proj} - \eta\left(\nabla P(w_{t-1}) - \nabla P\left(w_{t-1}^{proj}\right)\right)\right) \\
&\leqslant \left\|w_{t-1} - w_{t-1}^{proj} - \eta\left(\nabla P(w_{t-1}) - \nabla P\left(w_{t-1}^{proj}\right)\right)\right\|^2 + \eta^2\|e_{t-1}\|^2 \\
&\qquad\qquad + 2\eta\|e_{t-1}\|\left\|w_{t-1} - w_{t-1}^{proj} - \eta\left(\nabla P(w_{t-1}) - \nabla P\left(w_{t-1}^{proj}\right)\right)\right\|,
\end{aligned}
$$

where we use the fact $\nabla P\left(w^{proj}\right) = 0$ in the second equality and the Cauchy-Schwartz inequality in the last inequality. Notice that

$$
\begin{aligned}
&\left\|w_{t-1} - w_{t-1}^{proj} - \eta\left(\nabla P(w_{t-1}) - \nabla P\left(w_{t-1}^{proj}\right)\right)\right\|^2 \\
&= \left\|w_{t-1} - w_{t-1}^{proj}\right\|^2 + \eta^2\left\|\nabla P(w_{t-1}) - \nabla P\left(w_{t-1}^{proj}\right)\right\|^2 \\
&\qquad\qquad - 2\eta\left(\nabla P(w_{t-1}) - \nabla P\left(w_{t-1}^{proj}\right)\right)^T\left(w_{t-1} - w_{t-1}^{proj}\right) \\
&\leqslant \left\|w_{t-1} - w_{t-1}^{proj}\right\|^2 + \eta^2\left\|\nabla P(w_{t-1}) - \nabla P\left(w_{t-1}^{proj}\right)\right\|^2 \\
&\qquad\qquad - 2\eta\left(\frac{1}{2L}\left\|\nabla P(w_{t-1}) - \nabla P\left(w_{t-1}^{proj}\right)\right\|^2 + \frac{1}{2}\nu\left\|w_{t-1} - w_{t-1}^{proj}\right\|^2\right) \\
&= (1 - \nu\eta)\left\|w_{t-1} - w_{t-1}^{proj}\right\|^2 + \eta\left(\eta - \frac{1}{L}\right)\left\|\nabla P(w_{t-1}) - \nabla P\left(w_{t-1}^{proj}\right)\right\|^2 \\
&\leqslant (1 - \nu\eta)\left\|w_{t-1} - w_{t-1}^{proj}\right\|^2,
\end{aligned}
$$

where the first inequality follows from Lemma 8 with $\alpha = \frac{1}{2}$ and the last inequality follows from $\eta \leqslant \frac{1}{L_Q} \leqslant \frac{1}{L}$. Thus, taking expectation on both side of the above inequality yields

$$
\begin{aligned}
&\mathbb{E}[\left\|w_t - w_t^{proj}\right\|^2] \\
&\leqslant (1 - \eta)\,\mathbb{E}[\left\|w_{t-1} - w_{t-1}^{proj}\right\|^2] + \eta^2\mathbb{E}[\|e_{t-1}\|^2] \\
&\qquad\qquad + 2\eta(1 - \nu\eta)^{\frac{1}{2}}\mathbb{E}[\|e_{t-1}\|\left\|w_{t-1} - w_{t-1}^{proj}\right\|]
\end{aligned}
$$

$$
\begin{aligned}
&\leqslant (1-\nu\eta)\,\mathbb{E}[\left\|w_{t-1}-w_{t-1}^{proj}\right\|^2]+\eta^2\mathbb{E}[\|e_{t-1}\|^2]\\
&\qquad\qquad +2\eta(1-\nu\eta)^{\frac{1}{2}}\left(\mathbb{E}[\left\|w_{t-1}-w_{t-1}^{proj}\right\|^2]\right)^{\frac{1}{2}}\left(\mathbb{E}[\|e_{t-1}\|^2]\right)^{\frac{1}{2}}\\
&\leqslant \left((1-\nu\eta)^{\frac{1}{2}}\left(\mathbb{E}[\left\|w_{t-1}-w_{t-1}^{proj}\right\|^2]\right)^{\frac{1}{2}}+\eta\Big(\mathbb{E}[\|e_{t-1}\|^2]\Big)^{\frac{1}{2}}\right)^2,
\end{aligned}
$$

which implies

$$
\left(\mathbb{E}[\left\|w_t-w_t^{proj}\right\|^2]\right)^{\frac{1}{2}}\leqslant(1-\nu\eta)^{\frac{1}{2}}\left(\mathbb{E}[\left\|w_{t-1}-w_{t-1}^{proj}\right\|^2]\right)^{\frac{1}{2}}+\eta\Big(\mathbb{E}[\|e_{t-1}\|^2]\Big)^{\frac{1}{2}}.
$$

Applying the above bound recursively yields

$$
\left(\mathbb{E}\Big\|w_t-w_t^{proj}\Big\|^2]\right)^{\frac{1}{2}}\leqslant(1-\nu\eta)^{\frac{t}{2}}\left(\mathbb{E}[\Big\|w_0-w_0^{proj}\Big\|^2]\right)^{\frac{1}{2}}+\sum_{j=1}^{t}(1-\nu\eta)^{\frac{t-j}{2}}\eta\Big(\mathbb{E}[\|e_j\|^2]\Big)^{\frac{1}{2}}.
$$

Consequently, we can infer that

$$
\begin{aligned}
&\left(\mathbb{E}[\left\|w_t-w_t^{proj}\right\|^2]\right)^{\frac{1}{2}}\\
&\leqslant(1-\nu\eta)^{\frac{t}{2}}\left(\mathbb{E}[\left\|w_0-w_0^{proj}\right\|^2]\right)^{\frac{1}{2}}+\sum_{j=1}^{t}(1-\nu\eta)^{\frac{t-j}{2}}\eta\Big(\mathbb{E}[\|e_j\|^2]\Big)^{\frac{1}{2}}.\\
&\leqslant(1-\nu\eta)^{\frac{t}{2}}\left(\mathbb{E}[\Big\|w_0-w_0^{proj}\Big\|^2]\right)^{\frac{1}{2}}+\sum_{j=1}^{t}(1-\nu\eta)^{\frac{t-j}{2}}\eta\big(\eta L_Q^3\big)^{\frac{1}{2}}\left(\mathbb{E}[\Big\|w_0-w_0^{proj}\Big\|^2]\right)^{\frac{1}{2}},\\
&\leqslant(1-\nu\eta)^{\frac{t}{2}}\left(\mathbb{E}[\left\|w_0-w_0^{proj}\right\|^2]\right)^{\frac{1}{2}}+\frac{2}{\nu\eta}\eta\big(\eta L_Q^3\big)^{\frac{1}{2}}\left(\mathbb{E}[\Big\|w_0-w_0^{proj}\Big\|^2]\right)^{\frac{1}{2}}\\
&\leqslant\left(1+\frac{2\eta^{1/2}L_Q^{3/2}}{\nu}\right)\left(\mathbb{E}[\Big\|w_0-w_0^{proj}\Big\|^2]\right)^{\frac{1}{2}},
\end{aligned}
$$

where the second inequality follows from Lemma 2 and Lemma 3.

Lemma 10 *Suppose that Assumption 1 holds and P is convex and satisfies RSI with $\nu>0$. If $\eta\leqslant\frac{1}{L_Q}$, then in the k-th epoch of SARAH-I (Algorithm 1), for all $t\geq 1$, we have*

$$
\mathbb{E}[\Big(w_t-w_t^{proj}\Big)^T(\nabla P\,(w_t)-v_t)]\leqslant\left(1+\frac{2\eta^{1/2}L_Q^{3/2}}{\nu}\right)\cdot\eta^{1/2}L_Q^{3/2}\left(\mathbb{E}[\Big\|w_0-w_0^{proj}\Big\|^2]\right).
$$

Proof According to Lemma 6 and 9, we have

$$\begin{aligned}
&\mathbb{E}[\left(w_t - w_t^{proj}\right)^T \left(\nabla P\left(w_t\right) - v_t\right)] \\
&\leqslant \mathbb{E}[\left\|w_t - w_t^{proj}\right\| \left\|\nabla P\left(w_t\right) - v_t\right\|] \\
&\leqslant \left(\mathbb{E}[\left\|w_t - w_t^{proj}\right\|^2]\right)^{\frac{1}{2}} \left(\mathbb{E}[\left\|\nabla P\left(w_t\right) - v_t\right\|^2]\right)^{\frac{1}{2}} \\
&\leqslant \left(1 + \frac{2\eta^{1/2} L_Q^{3/2}}{\nu}\right) \left(\mathbb{E}[\left\|w_0 - w_0^{proj}\right\|^2]\right)^{\frac{1}{2}} \cdot \left(\eta L_Q^3\right)^{\frac{1}{2}} \left(\mathbb{E}[\left\|w_0 - w_0^{proj}\right\|^2]\right)^{\frac{1}{2}} \\
&\leqslant \left(1 + \frac{2\eta^{1/2} L_Q^{3/2}}{\nu}\right) \cdot \eta^{1/2} L_Q^{3/2} \left(\mathbb{E}[\left\|w_0 - w_0^{proj}\right\|^2]\right).
\end{aligned}$$

Theorem 3 *Suppose that Assumption 1 holds and P is convex and satisfies RSI with $\nu > 0$. If $\eta \leqslant \frac{1}{L_Q}$, then for we have all $k \geq 1$,*

$$\mathbb{E}[\left\|\tilde{w}_{k+1} - \tilde{w}_{k+1}^{proj}\right\|^2] \leqslant \tilde{\sigma} \mathbb{E}[\left\|\tilde{w}_k - \tilde{w}_k^{proj}\right\|^2],$$

where

$$\tilde{\sigma} := (1 - 2v\eta)^m + \frac{\eta^{1/2} L_Q^{3/2}}{\nu} + \frac{2\eta L_Q^3}{\nu^2} + \frac{\eta L_Q^2}{2\nu}. \tag{6}$$

Consequently, if m and η are chosen such that $\tilde{\sigma} < 1$, then SARAH-I converges linearly in expectation:

$$\mathbb{E}[\left\|\tilde{w}_k - \tilde{w}_k^{proj}\right\|^2] \leqslant \tilde{\sigma}^k \left\|\tilde{w}_0 - \tilde{w}_0^{proj}\right\|^2.$$

Proof In the k-th epoch of SARAH-I, we have

$$\begin{aligned}
&\mathbb{E}[\left\|w_{t+1} - w_{t+1}^{proj}\right\|^2] \\
&\leqslant \mathbb{E}[\left\|w_{t+1} - w_t^{proj}\right\|^2] \\
&= \mathbb{E}[\left\|w_t - \eta v_t - w_t^{proj}\right\|^2] \\
&= \mathbb{E}[\left\|w_t - w_t^{proj}\right\|^2] - 2\eta \left(w_t - w_t^{proj}\right)^T \mathbb{E}[v_t] + \eta^2 \mathbb{E}[\left\|v_t\right\|^2] \\
&= \mathbb{E}[\left\|w_t - w_t^{proj}\right\|^2] - 2\eta [\left(w_t - w_t^{proj}\right)^T \nabla P\left(w_t\right)] \\
&\qquad + 2\eta \mathbb{E}[\left(w_t - w_t^{proj}\right)^T \left(\nabla P\left(w_t\right) - v_t\right)] + \eta^2 \mathbb{E}[\left\|v_t\right\|^2] \\
&\leqslant \mathbb{E}[\left\|w_t - w_t^{proj}\right\|^2] - 2\eta [\left(w_t - w_t^{proj}\right)^T \nabla P\left(w_t\right)] \\
&\qquad + 2\eta \mathbb{E}[\left(w_t - w_t^{proj}\right)^T \left(\nabla P\left(w_t\right) - v_t\right)] + \eta^2 \mathbb{E}[\left\|v_0\right\|^2]
\end{aligned}$$

$$
\begin{aligned}
&\leqslant \mathbb{E}[\left\|w_t - w_t^{proj}\right\|^2] - 2v\eta\left\|w_t - w_t^{proj}\right\|^2 \\
&\qquad + 2\left(1 + \frac{2\eta^{1/2}L_Q^{3/2}}{\nu}\right)\cdot \eta^{3/2}L_Q^{3/2}\left\|w_0 - w_0^{proj}\right\|^2 + L^2\eta^2\left\|w_0 - w_0^{proj}\right\|^2 \\
&\leqslant (1-2\nu\eta)\,\mathbb{E}[\left\|w_t - w_t^{proj}\right\|^2] \\
&\qquad + 2\left(1 + \frac{2\eta^{1/2}L_Q^{3/2}}{\nu}\right)\cdot \eta^{3/2}L_Q^{3/2}\left\|w_0 - w_0^{proj}\right\|^2 + L_Q^2\eta^2\left\|w_0 - w_0^{proj}\right\|^2,
\end{aligned}
$$

where the second inequality follows from $\mathbb{E}[\|v_t\|^2] \leqslant \|v_{t-1}\|^2$, for all $t > 0$ from proof of Lemma 5, the third inequality follows from RSI, the Lipschitz continuity of $\nabla P(w)$ and Lemma 10, and the last inequality uses the fact $L \leqslant L_Q$. By recursively applying the above inequality over t, and noting that $\tilde{w}_t = w_0$ and $\tilde{w}_{t+1} = w_m$, we obtain

$$
\begin{aligned}
&\mathbb{E}[\left\|\tilde{w}_{k+1} - \tilde{w}_{k+1}^{proj}\right\|^2] \\
&\leqslant (1-2\nu\eta)^m\left\|\tilde{w}_k - \tilde{w}_k^{proj}\right\|^2 \\
&\qquad + \left(2\left(1 + \frac{2\eta^{1/2}L_Q^{3/2}}{\nu}\right)\cdot \eta^{3/2}L_Q^{3/2} + L_Q^2\eta^2\right)\sum_{j=0}^{m-1}(1-2\nu\eta)^j\left\|\tilde{w}_k - \tilde{w}_k^{proj}\right\|^2 \\
&= \tilde{\sigma}_k\left\|\tilde{w}_k - \tilde{w}_k^{proj}\right\|^2.
\end{aligned}
$$

This completes the proof.

5 The SARAH-I Method with Barzilai-Borwein step sizes

The Barzilai-Borwein (BB) [21][22] method fits a quadratic model to the objective in each iteration, and a step size is proposed that is optimal for the local quadratic model. Consider the generic unconstrained optimization problem

$$\min_{w\in\mathbb{R}^d} f(w),$$

where $f(w)$ is first-order continuously differentiable. The standard BB method updates the iterates through

$$w_{k+1} = w_k - \eta_k^{-1}\nabla f(w_k),$$

where η_k is the so-called BB step size. Here η_k is introduced such that ηI is an approximation to the Hessian of f at w_k, so it follows a certain quasi-Newton property. It is normally computed through solving the following problem:

$$\min_{\eta}\left\|\frac{1}{\eta_k}s_k - y_k\right\|_2 \quad \text{or} \quad \min_{\eta}\|s_k - \eta_k y_k\|_2$$

where $s_k = w_k - w_{k-1}$ and $y_k = \nabla f(x_k) - \nabla f(x_{k-1})$. Then it yields that

$$\eta_k^{BB1} = \frac{s_k^T s_k}{s_k^T y_k} \quad \text{or} \quad \eta_k^{BB2} = \frac{s_k^T y_k}{y_k^T y_k}.$$

When $s_k^T y_k > 0$, it is easy to obtain $\alpha_k^{BB1} \geq \alpha_k^{BB2}$ which means α_k^{BB1} is a more aggressive step size to decrease the objective function. In literatures, such as [23], it has been discussed that α_k^{BB1} is superior to α_k^{BB2}. Recently, however, [20] studies some numerical examples which show that α_k^{BB1} may not be the best choice. Thus they propose a new family of spectral gradient methods which compute the step size based on the convex combination of η_k^{BB1} and η_k^{BB2}, i.e.,

$$\eta_k = \tau \eta_k^{BB1} + (1-\tau)\, \eta_k^{BB2}$$

where $\tau \in [0, 1]$. This whole family of step sizes shares the same properties. For details, interested readers are referred to [20]. We now propose to incorporate the above step size to SARAH-I, which leads to the SARAH-I-BB method.

5.1 SARAH-I-BB Method

In this section, we propose the SARAH-I-BB algorithm by employing the BB approach to calculate step sizes. We now describe its framework in the following Algorithm 2.

Algorithm 2 The SARAH-I-BB Method

Parameters: update frequency m, initial step size $\eta_1 > 0$, initial point $\tilde{w}_0$

for $k = 1, 2, \ldots$ **do**
  $w_0 = \tilde{w}_{k-1}$
  $v_0^k = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(w_0)$
  if k > 1 **then**
    calculate η_k using the BB method
  end if
  $w_1 = w_0 - \eta_k v_0^k$
  for $t = 1, \ldots, m-1$ **do**
    Randomly pick $i_t \in \{1, \ldots, n\}$ randomly according to Q
    $v_t = (\nabla f_{i_t}(w_t) - \nabla f_{i_t}(w_{t-1}))/(n q_{i_t}) + v_{t-1}$
    $w_{t+1} = w_t - \eta_k v_t$
  end for
  $\tilde{w}_k = w_m$
end for

In SARAH-I-BB, we use two successive outer iterates to calculate BB step sizes, i.e.,

$$\eta_k^1 = \frac{\|\tilde{w}_k - \tilde{w}_{k-1}\|^2}{(\tilde{w}_k - \tilde{w}_{k-1})^T \left(v_0^k - v_0^{k-1}\right)} \quad \text{and} \quad \eta_k^2 = \frac{(\tilde{w}_k - \tilde{w}_{k-1})^T \left(v_0^k - v_0^{k-1}\right)}{\left\|v_0^k - v_0^{k-1}\right\|^2}.$$

5.2 SARAH-I-BB Method for Strongly Convex Optimization

In this case, we calculate η_k through

$$\eta_k = \frac{1}{m} \cdot \left(\tau \eta_k^1 + (1-\tau)\, \eta_k^2\right). \tag{7}$$

The following theorem shows that the linear convergence rate of SARAH-I-BB could be achieved, provided that the inner iteration number m is sufficiently large.

Theorem 4 *Suppose that Assumption 1 holds. Suppose that P is μ-strongly convex. Given $\theta = \left(1 - e^{-2\mu/L_Q}\right)/4$. It is obviously that $\theta \in (0, 1/4)$. Then if m is chosen such that*

$$m \geqslant \frac{L_Q^3}{\theta^2 \mu^3},$$

SARAH-I-BB converges linearly in expectation, i.e.,

$$\mathbb{E}[\|\tilde{w}_k - w_*\|^2] \leqslant (1-\theta)^k \|\tilde{w}_0 - w_*\|^2.$$

Proof From the Lipschitz continuity of $\nabla P(w)$, it is easy to obtain that

$$\eta_k^{BB1} \geqslant \frac{\|\tilde{w}_k - \tilde{w}_{k-1}\|^2}{L\|\tilde{w}_k - \tilde{w}_{k-1}\|^2} = \frac{1}{L},$$

$$\eta_k^{BB2} \geqslant \frac{\left\|v_0^k - v_0^{k-1}\right\|^2}{L\left\|v_0^k - v_0^{k-1}\right\|^2} = \frac{1}{L},$$

Thus we obtain the lower bound of η_k:

$$\eta_k = \frac{1}{m} \cdot \left(\tau \eta_k^{BB1} + (1-\tau)\, \eta_k^{BB2}\right) \geqslant \frac{1}{m} \cdot \left(\tau \frac{1}{L} + (1-\tau)\, \frac{1}{L}\right) = \frac{1}{mL}.$$

Meanwhile, the strong convexity of P indicates that $\eta_k \leqslant 1/m\mu$. Therefore, the coefficient σ defined in (5) can be bounded by

$$\begin{aligned}
\sigma &= (1 - 2\mu\eta)^m + \frac{\eta^{1/2} L_Q^{3/2}}{\mu} + \frac{\eta L_Q^3}{\mu^2} + \frac{\eta L_Q^2}{2\mu} \\
&\leqslant \exp\left\{-\frac{2\mu}{mL} \cdot m\right\} + \frac{L_Q^{3/2}}{m^{1/2}\mu^{3/2}} + \frac{L_Q^3}{m\mu^3} + \frac{L_Q^2}{2m\mu^2} \\
&\leqslant \exp\left\{-\frac{2\mu}{L_Q}\right\} + \frac{L_Q^{3/2}}{m^{1/2}\mu^{3/2}} + \frac{L_Q^3}{m\mu^3} + \frac{L_Q^2}{2m\mu^2} \\
&< 1 - 4\theta + \theta + \theta + \theta \\
&= 1 - \theta.
\end{aligned}$$

This completes the proof.

We now obtain the computational complexity of SARAH-I-BB with respect to the total number of component gradient evaluations to achieve the ε-accurate solution $\tilde{w}_T$ satisfying $\mathbb{E}[\|\tilde{w}_T - w_*\|^2] < \varepsilon$.

Corollary 1 *Suppose that Assumption 1 holds. Suppose that P is μ-strongly convex. Then the computational complexity of SARAH-I-BB to achieve an ε-accurate solution is $\mathcal{O}\left(\left(n + \frac{L^3}{\theta^2\mu^3}\right)\log\left(\frac{1}{\varepsilon}\right)\right)$.*

Proof To obtain $\mathbb{E}[\|\tilde{w}_k - w_*\|^2] < \varepsilon$, it suffices to require

$$\mathbb{E}[\|\tilde{w}_k - w_*\|^2] \leqslant (1-\theta)^k \|\tilde{w}_0 - w_*\|^2 < \varepsilon,$$

which implies that

$$k > \log\left(\frac{\|\tilde{w}_0 - w_*\|^2}{\varepsilon}\right) \Big/ \log\left(\frac{1}{1-\theta}\right).$$

Consequently, the total number of component gradient evaluations is

$$(n+2m)\,k = \mathcal{O}\left(\left(n + \frac{\kappa^3}{\theta^2}\right)\log\left(\frac{1}{\varepsilon}\right)\right).$$

5.3 SARAH-I-BB Method for Non-Strongly Convex Optimization

When P is non-strongly convex, it could not be guaranteed that η_k defined through (7) is uniformly upper bounded. We thus in this case calculate η_k through

$$\eta_k = \frac{1}{m} \cdot \min\left\{\tau\eta_k^{BB1} + (1-\tau)\,\eta_k^{BB2}, \frac{1}{\rho}\right\},$$

where $\rho < L \leqslant L_Q$.

Theorem 5 *Suppose that Assumption 1 holds. Suppose that P is convex and satisfies RSI with $\nu > 0$. Given $\theta = \left(1 - e^{-2\nu/L_Q}\right)/4$. It is obviously that $\theta \in (0, 1/4)$. In SARAH-I-BB, if m is chosen such that*

$$m \geqslant \frac{L_Q^3}{\rho\theta^2\nu^2},$$

then SARAH-I-BB (Algorithm 2) converges linearly in expectation:

$$\mathbb{E}[\left\|\tilde{w}_k - \tilde{w}_k^{proj}\right\|^2] \leqslant (1-\theta)^k \left\|\tilde{w}_0 - \tilde{w}_0^{proj}\right\|^2.$$

Proof Using the Lipschitz continuity of $\nabla P(w)$, it is easy to obtain $\frac{1}{mL} \leqslant \eta_k \leqslant \frac{1}{m\rho}$. Therefore, $\tilde{\sigma}$ defined in (6) can be bounded by:

$$\begin{aligned}\tilde{\sigma} &= (1-2v\eta)^m + \frac{\eta^{1/2}L_Q^{3/2}}{\nu} + \frac{2\eta L_Q^3}{\nu^2} + \frac{\eta L_Q^2}{2\nu} \\ &\leqslant \exp\{-\frac{2\nu}{mL}\cdot m\} + \frac{L_Q^{3/2}}{m^{1/2}\rho^{1/2}\nu} + \frac{2L_Q^3}{m\rho\nu^2} + \frac{L_Q^2}{2m\rho\nu} \\ &\leqslant \exp\{-\frac{2\nu}{L}\} + \frac{L_Q^{3/2}}{m^{1/2}\rho^{1/2}\nu} + \frac{2L_Q^3}{m\rho\nu^2} + \frac{L_Q^2}{2m\rho\nu} \\ &< 1-4\theta+\theta+\theta+\theta = 1-\theta.\end{aligned}$$

This completes the proof.

We now obtain the computational complexity with respect to total number of component gradient evaluations to achieve an ε-accurate solution.

Corollary 2 *Suppose that Assumption 1 holds. Suppose that P is convex and satisfies RSI with $\nu > 0$. The total complexity of SARAH-I-BB 2 to achieve the ε-accurate solution is $\mathcal{O}\left(\left(n + \frac{L_Q^3}{\rho\theta^2\nu^2}\right)\log\left(\frac{1}{\varepsilon}\right)\right)$.*

Proof Following from Theorem 5, to obtain$\mathbb{E}[\left\|\tilde{w}_k - \tilde{w}_k^{proj}\right\|^2] < \varepsilon$, it needs to require

$$\mathbb{E}[\left\|\tilde{w}_k - \tilde{w}_k^{proj}\right\|^2] \leqslant (1-\theta)^k \left\|\tilde{w}_0 - \tilde{w}_0^{proj}\right\|^2 < \varepsilon,$$

which implies

$$k > \log\left(\frac{\left\|\tilde{w}_0 - \tilde{w}_0^{proj}\right\|^2}{\varepsilon}\right) \Big/ \log\left(\frac{1}{1-\theta}\right).$$

Then the total complexity can be obtained as

$$(n+2m)\,k = \mathcal{O}\left(\left(n + \frac{L_Q^3}{\rho\theta^2\nu^2}\right)\log\left(\frac{1}{\varepsilon}\right)\right).$$

6 Numerical Experiments

In this section, we present our numerical experiments. We study the binary classification problem with f_i being the ℓ_2-regularized logistic regression

$$f_i(w) = \log\left(1+\exp\left(-y_i x_i^T w\right)\right) + \frac{\lambda}{2}\|w\|^2$$

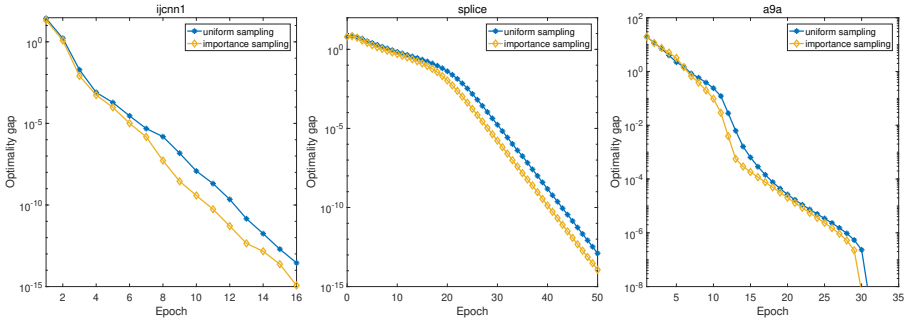
on datasets *ijcnn1*, *splice*, *a9a*, *real-sim*, *rcv1* as Table1[1].

[1] All datasets are available at https://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/

Dataset	ijcnn1	splice	a9a	real-sim	rcv1
n	49990	1000	32561	72309	20242
d	22	60	123	20958	47236

Table 1: Data information

We set $\tau = 0.5$ in SARAH-I-BB. Our numerical experiments are implemented by MATLAB 9.1, CPU version is i7-6700K. For both SARAH-I and SARAH-I-BB, we set $m = \mathcal{O}(n)$ and $\lambda = \frac{1}{n}$ which are commonly used in practice[8]. In following figures, x-axis with "epoch" means the number of epochs, i.e., the number of outer loops in both algorithms and y-axis with "optimality gap" denotes the value $P(\tilde{w}_k) - P(w_*)$ where w_* is obtained by running L-BFGS with backtracking [32].



(a) (b) (c)

Fig. 1: Comparison of SARAH-I with importance sampling and with uniform sampling

In Fig. 1, we test the effect of the importance sampling strategy on SARAH-I. So we compare SARAH-I with importance sampling and with uniform sampling. Numerical experiments are tested on datasets *ijcnn1, splice, a9a*. When adopting importance sampling, we calculate $Q \sim \{q_1, q_2, \dots q_n\}$ same as [25], setting $\phi_i(w) = \log(1 + \exp(-y_i x_i^T w))$, then we have $\nabla\phi_i(w) = -\frac{\exp(-y_i x_i^T w)}{1+\exp(-y_i x_i^T w)} y_i x_i^T$. Since $0 < \frac{\exp(-y_i x_i^T w)}{1+\exp(-y_i x_i^T w)} < 1$ and $y_i = \{-1, 1\}$, so we have $\|\nabla\phi_i(w)\| \leqslant \|x_i\|$. Then we calculate q_i as follow

$$q_i = \frac{\|x_i\|}{\sum_{j=1}^n \|x_j\|}.$$

We can see from Fig. 1 that SARAH-I converges faster with importance sampling being incorporated.

Fig. 2 shows the comparison results between SARAH-I and SARAH-I-BB on datasets *ijcnn1, real-sim, rcv1*. We used three different step sizes for SARAH-I, and three different initial step sizes for SARAH-I-BB. In the lower

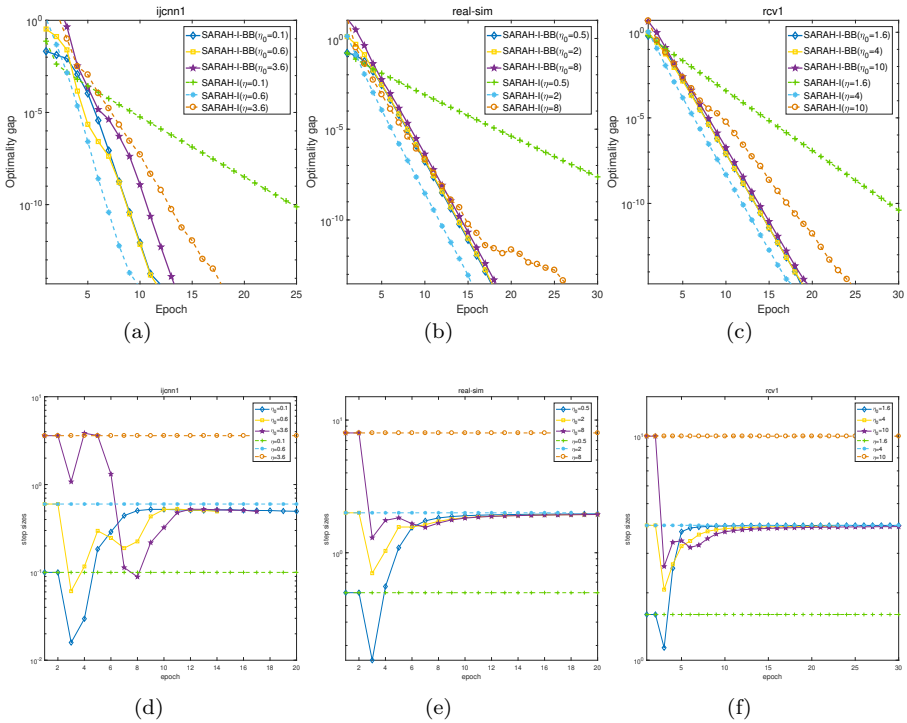


Fig. 2: Comparison of SARAH-I-BB and SARAH-I with constant step sizes on different datasets. The solid lines stand for SARAH-I-BB with different η_0. The dashed lines stand for SARAH-I with different η. In these experiments, we choose $m = n/2$.

subfigures, the y-axis represents the step sizes. In all the six subfigures, the dashed lines correspond to SARAH-I with fixed step size. The solid lines correspond to SARAH-I-BB with different initial step sizes η_0. Moreover, the dashed lines in light blue color always represent SARAH-I with best-tuned fixed step size. The solid lines with blue, yellow, purple colors in Fig. 2(a) and 2(d) correspond to $\eta_0 = 0.1$, 0.6 and 3.6, respectively. The solid lines with blue, yellow, purple colors in Fig. 2(b) and 2(e) correspond to $\eta_0 = 0.5$, 2 and 8, respectively. The solid lines with blue, yellow, purple colors in Fig. 2(c) and 2(f) correspond to $\eta_0 = 1.6$, 4 and 10, respectively. The dashed lines with green, light blue and orange colors correspond to the consistent fixed step sizes.

From Fig. 2(a), 2(b) and 2(c), we can see that SARAH-I-BB performs as well as SARAH-I with best-tuned step sizes on different datasets. SARAH-I-BB outperforms SARAH-I with the two other choices of step sizes. It can be seen from Fig. 2(d), 2(e) and 2(f) that, different initial step sizes have little effect on performance of SARAH-I-BB, which indicates that SARAH-I-BB is not sensitive to the choice of η_0. And the step sizes computed by SARAH-I-

BB always converge to the best-tuned step size. All in all, SARAH-I-BB can compute the best step size automatically in practice.

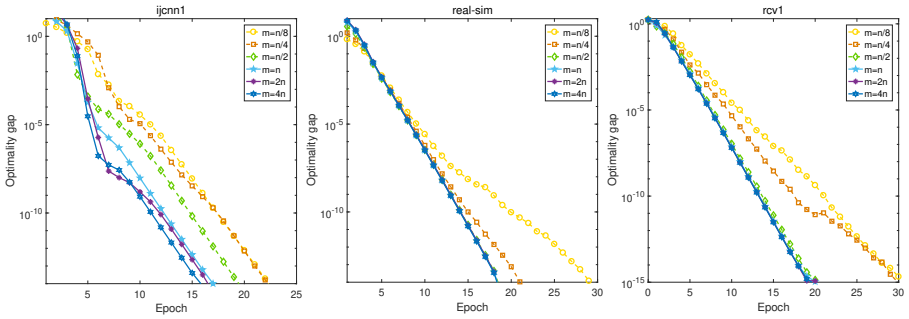


Fig. 3: Comparison of different number of inner loops m of SARAH-I-BB

Fig. 3 shows that the effects of different inner iteration numbers m on the performance of SARAH-I-BB. Fig. 3(a), 3(b) and 3(c) corresponds to dataset *ijcnn1*, *real-sim* and *rcv1*, respectively. We choose different m as $n/8$, $n/4$, $n/2$, n, $2n$, $4n$, respectively. The dashed lines in yellow, orange and green correspond to $n/8$, $n/4$, $n/2$, respectively. And the solid lines with light blue, purple, dark blue correspond to n, $2n$, $4n$, respectively. The Figure indicate that larger m is better for algorithmic performance within certain limits, but SARAH-I-BB is insensitive to the choice of m. SARAH-I-BB is more robust in terms of the number of inner loops.

Table 2 reports the CPU time of every epoch for all the three datasets with respect to different settings of m . We can see that as m increases, CPU time increases by almost corresponding multiples. Fig 4 is more intuitive to illustrate the rate of increase of CPU time is raising with the increase of m.

m \ dataset	ijcnn1	rcv1	real-sim
$n/8$	0.3294	49.2684	175.6825
$n/4$	0.6050	60.4725	228.3427
$n/2$	1.1365	77.1907	333.2679
n	2.2568	117.2433	551.2994
$2n$	4.2679	197.8093	1.0024e+03
$4n$	7.9216	362.8992	1.8432e+03

Table 2: CPU Time (s) of every epoch

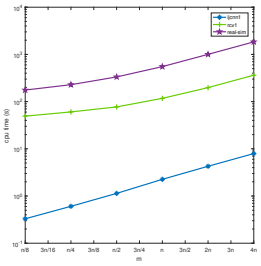


Fig. 4: CPU Time

7 Conclusion

In this paper, we studied the SARAH-I method with importance sampling strategy and each outer iterate being set as the last inner iterate. We proved the linear convergence of SARAH-I in strongly convex case. And we also established its linear convergence in non-strongly convex case under RSI condition. Moreover, we proposed to use the BB method to calculate the step sizes for SARAH-I and provided complexity analysis of SARAH-I-BB in both strongly convex and non-strongly convex cases. At last we reported some preliminary test results which revealed competitive performances of proposed algorithms.

Acknowledgement

This work is supported by National Natural Science Foundation of China under Grants 11871453, 11731013, 11571014 and Young Elite Scientists Sponsorship Program by China Association for Science and Technology. Part of research by Xiao Wang was done during her working as a research fellow in the Hong Kong Polytechnic University. The authors would like to thank Professor Shiqian Ma for his valuable suggestions.

References

1. Robbins, H., Monro, S.: A stochastic approximation method. Annals of Mathematical Statistics, 22(3), 400-407 (1951)
2. Nesterov, Y.: Introductory lectures on convex optimization: A basic course. Springer, Boston, MA (2014)
3. Bottou, L., Curtis, F. E., Nocedal, J.: Optimization methods for large-scale machine learning. SIAM Review, 60(2), 223-311 (2018)
4. Roux, N. L., Schmidt, M., Bach, F. R.: A stochastic gradient method with an exponential convergence rate for finite training sets. In: Neural Information Processing Systems, pp. 2663-2671 (2013)
5. Schmidt, M. W., Roux, N. L., Bach, F.: Minimizing finite sums with the stochastic average gradient. Mathematical Programming, 162(1), 83-112 (2017)
6. Defazio, A., Bach, F., Lacoste-Julien, S.: SAGA: a fast incremental gradient method with support for non-strongly convex composite objectives. In: Neural Information Processing Systems, pp. 1646-1654 (2014)
7. Johnson, R., Zhang, T.: Accelerating stochastic gradient descent using predictive variance reduction. In: Neural Information Processing Systems, pp. 315-323 (2013)
8. Nguyen, L. M., Liu, J., Scheinberg, K., Takac, M.: SARAH: a novel method for machine learning problems using stochastic recursive gradient. In: Neural Information Processing Systems, pp. 2613-2621 (2017).
9. Nguyen, L. M., Scheinberg, K, Taká, M.: Inexact SARAH algorithm for stochastic optimization. arXiv:1811.10105. (2018)
10. Nguyen, L. M., Scheinberg, K, Taká, M.: Stochastic recursive gradient algorithm for nonconvex optimization. arXiv:1705.07261. (2017)
11. Fang, C., Li, C. J., Lin, Z., Zhang, T.: Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In: Neural Information Processing Systems, pp. 687-697 (2018)
12. Wang, Z., Ji, K., Zhou, Y., Liang, Y., Tarokh, V.: SpiderBoost: A class of faster variance-reduced algorithms for nonconvex optimization. arXiv:1810.10690. (2018)

13. Babanezhad, R., Ahmed, M. O., Virani, A., Schmidt, M. W., Konecný, J., Sallinen, S.: Stop wasting my gradients: practical SVRG. In: Neural Information Processing Systems, pp. 2251-2259 (2015)
14. Konecný, J., Richtarik, P.: Semi-stochastic gradient descent methods. Frontiers in Applied Mathematics and Statistics, 3, 9 (2017)
15. Zhou, C., Gao, W., Goldfarb, D.: Stochastic adaptive quasi-Newton methods for minimizing expected values. In: International Conference on Machine Learning, pp. 4150-4159 (2017)
16. Duchi, J. C., Hazan, E., Singer, Y.J.: Adaptive subgradient methods for online learning and stochastic optimization, In: Journal of Machine Learning Research, pp. 2121-2159, (2011)
17. Kingma, D. P., Ba, J.: Adam: a method for stochastic optimization. In: International Conference on Learning Representations, pp. 1-13 (2015)
18. De, S., Yadav, A. K., Jacobs, D. W., Goldstein, T.: Automated inference with adaptive batches. In: International Conference on Artificial Intelligence and Statistics, pp. 1504-1513 (2017)
19. Tan, C., Ma, S., Dai, Y. H., Qian, Y.: Barzilai-Borwein step size for stochastic gradient descent. In: Neural Information Processing Systems, pp. 685-693 (2016)
20. Dai, Y. H., Huang, Y., Liu, X. W.: A family of spectral gradient methods for optimization. arXiv:1812.02974 (2018)
21. Barzilai, J., Borwein, J.M.: Two-point step size gradient methods. IMA Journal of Numerical Analysis, 8(1), 141-148 (1988)
22. Dai, Y. H., Liao, L. Z.: Rlinear convergence of the Barzilai and Borwein gradient method. IMA Journal of Numerical Analysis, 22(1), 1-10 (2002)
23. Fletcher, R.: On the barzilai-borwein method. Optimization and control with applications, 235-256 (2005)
24. Xiao, L., Zhang, T.: A proximal stochastic gradient method with progressive variance reduction. SIAM Journal on Optimization. 24(4), 2057-2075 (2014)
25. Zhao, P., Zhang, T.: Stochastic optimization with importance sampling for regularized loss minimization. In : International Conference on Machine Learning, pp. 1-9 (2015)
26. Polyak, B. T.: Gradient methods for the minimisation of functionals. USSR Computational Mathematics and Mathematical Physics, 3(4), 864-878 (1963)
27. Luo, Z. Q., Tseng, P.: Error bounds and convergence analysis of feasible descent methods: a general approach. Annals of Operations Research, 46(1), 157-178 (1993)
28. Zhang, H., Cheng, L.: Restricted strong convexity and its applications to convergence analysis of gradient-type methods in convex optimization. Optimization Letters, 9(5), 961-979 (2015)
29. Karimi, H., Nutini, J., Schmidt, M.: Linear convergence of gradient and proximal-gradient methods under the polyak-ojasiewicz condition. In: Machine Learning and Knowledge Discovery in Databases, pp. 795-811 (2016)
30. Necoara, I., Nesterov, Y., Glineur, F.: Linear convergence of first order methods for non-strongly convex optimization. Mathematical Programming, 1-39 (2018)
31. Anitescu, M.: Degenerate nonlinear programming with a quadratic growth condition. SIAM Journal on Optimization, 10(4), 1116-1135 (2000)
32. Nocedal, J., Wright, S.: Numerical optimization. Springer, New York, NY (2006)