Stochastic Variance Reduced Gradient Method Embedded with Positive Defined Stabilized Barzilai-Borwein

Weijuan Shi, Adibah Shuib and Zuraida Alwadood

Abstract—**Machine learning (ML) is evolving rapidly and has made many theoretical breakthroughs while widely applied in various fields. ML allows systems the ability to access data and use it to enable computers to execute cognitive processes such as learning and improving from previous experiences and solving complicated issues. Many first-order stochastic optimization methods have been used to solve the optimization model of ML. These algorithms adopt Barzilai-Borwein (BB) step size instead of fixed or diminishing step size to improve performance. However, the BB step size format involves fractional calculation, which inevitably leads to a zero denominator, especially when the objective function is non-convex. The BB technique will be violated if the denominator is near 0 or even negative. To improve the computation of the step size, a Positive Defined Stabilized Barzilai-Borwein (PDSBB) approach is introduced in this paper. Integrating PDSBB with the stochastic variance reduced gradient (SVRG) approach, a new method SVRG-PDSBB is proposed. Numerical experiments have shown that the new algorithm has stabilized the performance of the new step size, which successfully avoiding zero denominators and effectively solving the common problems in machine learning. The convergence of SVRG-PDSBB is theoretically and numerically proven, and the effectiveness of the new algorithm is shown by comparison with other algorithms.**

Index Terms—**BB, PDSBB, stochastic optimization, SVRG, machine learning**

I. Introduction

MACHINE learning (ML) is evolving rapidly and has made many theoretical breakthroughs while widely applied in various fields. ML is an artificial intelligence (AI) application that provides systems with the ability to access data and use them to enable machines to perform cognitive functions by learning and improving from past experiences. ML solves complex problems and enables the analysis of massive quantities of data.

Manuscript received March 26, 2023; revised October 12, 2023.

This work was supported in part by Hunan Province educational innovation subject in 2022 under Grant HNJG-2022-1110.

Weijuan Shi is a PhD candidate of Universiti Teknologi MARA, Shah Alam, Selangor, 40450, Malaysia. (e-mail: shiweijuan2007@sina.com).

Adibah Shuib is an Associate Professor at School of Mathematical Sciences, College of Computing, Informatics and Mathematics (KPPIM), Al-Khawarizmi Building, Universiti Teknologi MARA (UiTM), 40450 Shah Alam, Selangor Darul Ehsan, Malaysia. (Phone: +60192622348; e-mail: adibah253@uitm.edu. my).

Zuraida Alwadood is a senior lecturer at School of Mathematical Sciences, College of Computing, Informatics and Mathematics (KPPIM), Al-Khawarizmi Building, Universiti Teknologi MARA, 40450 Shah Alam, Selangor Darul Ehsan, Malaysia. (e-mail: zuraida794@uitm.edu.my).

First-order stochastic optimization algorithm is widely used to solve machine learning models because of its high efficiency, reasonably fast convergence speed, practical objective function, and easy implementation. However, the traditional first-order optimization algorithm encountered various problems. On one hand, the explosive growth of data and the continuous increase of parameters in ML models such as deep neural networks have led the traditional deterministic numerical optimization algorithm to a problem of the excessive computation process as well as a slower convergence rate. In addition, the first-order algorithm analysis discussed in numerical optimization is often based on the worst-case computational complexity.

Researchers have been working on developing a more efficient forecasting model than the previous ones [1]. Reference [2] proposed the Stochastic Gradient Descent (SGD) techniques. It has turned into a central part in science and designing, for example, in measurements, Artificial Intelligence (AI), signal/picture handling, opposite issues, and others. In SGD, a manually fixed step size (learning rate) or a decreasing step size is normally used. In reality, these two ways can be time-consuming.

Reference [3] introduced the accelerated mini-batch proximal stochastic variance reduction gradient (AccProx-SVRG) approach, which combines Nesterov's acceleration with the accelerated mini-batch comparable to Nesterov's acceleration method in a mini-batch environment. The accelerated efficient mini-batch SVRG (AMSVRG) has been demonstrated to be capable of achieving a speedy combination complication for general curved and ideal forceful stated difficulties. In addition to this, Accelerated Stochastic Gradient Descent (ASGD) approaches have used either a best-tuned step size or a decreasing step size [3]-[5]. In general, numerous step-size sequence methods have been developed in previous research.

Step size is a crucial issue in using ML stochastic optimization, particularly in the first-order stochastic optimization algorithm. The processing speed of a large amount of data will be directly affected by rapid updating of the step size. ML has significant theoretical and practical implications. Choosing an appropriate step size is still a significant obstacle in ML stochastic optimization and thus must be actively studied to support any advancements in overcoming this issue.

Several past research have studied step size difficulties to optimize the mini-batch and produced impressive results. Reference [6] proposed a new mini-batch approach known as mS2GD-BB by inserting the Barzilai-Borwein (BB). They have proven that mS2GD-BB converged linearly, as expected for non-smooth and highly convex functions. Furthermore, by generating a quickly updated step size sequence, mS2GD-BB beats some of the most cutting-edge logistic regression systems.

Reference [7] incorporate the BB method to automatically compute step size for Acc-Prox-SVRG method to obtain a new accelerated method known as the Acc-Prox-SVRG-BB. The convergence of the Acc-Prox-SVRG-BB is proven in which its complexity achieves the same level as the best known stochastic gradient methods to make it comparable with the best-known stochastic gradient (SG) methods.

The BB or Random Barzilai-Borwein (RBB) formulae are used to calculate step size, and the technique proposed by [7]-[8] does not avoid the denominator from being close to zero. In ML, the optimization problem usually considered $f_1, f_2, \cdots, f_n$ to be a sequence of vector functions from $\mathbb{R}^d \to \mathbb{R}$. The goal is to minimize the objective function (1):

$$\min_{w \in R^d} F(w) = \frac{1}{n}\sum_{i=1}^{n} f_i(w) \tag{1}$$

where n is the model size, w represents the parameter, and $f_i(w)$ is a sequence of loss functions that evaluate the cost of the current parameter w. Each $f_i : \mathbb{R}^d \to \mathbb{R}$ is the cost function for the first sample of data.

SGD, SVRG, and mS2GD are only some examples of first-order stochastic optimization techniques that have been utilized to solve ML optimization models. However, each approach uses a constant or decreasing step size that is frequently inappropriate, impractical, and time-consuming. The BB method was incorporated into the SGD and SVRG, which produced two new algorithms, SGD-BB and SVRG-BB [9]. The SVRG-BB algorithm's convergence was analyzed and these two algorithms were used to solve the objective function as a smooth function. The first-order optimization algorithms are explicitly involved in choosing the movement and step size in the search space using the first derivative (gradient). The step size acts as a hyperparameter for the search space.

Our study aims to solve the problem of choosing an appropriate step size that will result in faster convergence rate in ML stochastic optimization by introducing a newly proposed positive defined stabilized Barzilai-Borwein (PDSBB) step size. This paper discusses on PDSBB algorithm and the computational results which include the comparison of performance with existing SG algorithms.

II. SVRG Algorithm with PDSBB Step Size

The BB method is also known as the two-point step gradient method [10]. This method is mainly used in solving nonlinear optimization problems. Compared to the traditional quasi-Newton method, BB only needs a small amount of calculation to satisfy the quasi-Newton property. Satisfying the quasi-Newton property is referred to as satisfying the secant equation. Suppose that problem (2) needs to be solved:

$$\min f(w) \tag{2}$$

where $f(w)$ is differentiable. The iterative formula of the quasi-Newton method for problem (2) is,

$$w_{t+1} = w_t - B_t^{-1}\nabla f(w_t) \tag{3}$$

where B_t is an estimate of the Hessian matrix of f at w_t.

We use the scalar matrix $B_t = \frac{1}{\eta_t} I$ to approximate Hessian matrix ($\eta_t > 0$) and substitute it into the secant equation Let $s_t = w_t - w_{t-1}$ and $y_t = \nabla f(w_t) - \nabla f(w_{t-1}), t > 1$, then the secant equation is $B_t s_t = y_t$. By solving the residual of the secant equation,

$$\min \left\| \left(\frac{1}{n_t s_t} - y_t\right) \right\|^2 \tag{4}$$

the BB step size can be obtained as:

$$\eta_t^{BB1} = \frac{\|s_t\|^2}{s_t^T y_t} \tag{5}$$

Another form of BB step size is

$$\eta_t^{BB2} = \frac{s_t^T y_t}{\|y_t\|^2} \tag{6}$$

By solving (6), the solution is

$$\min \|s_t - \eta_t y_t\|^2 \tag{7}$$

A small search space will take a long time and will be trapped at local optima. A big search space will lead to zig-zag or bounces throughout the search region, missing the optima totally. This study will create a unique dynamic adaptive step size based on modifying the BB technique to automatically compute step size, motivated by the need to stabilize the BB approach. The new PDSBB step size is intended to address the issue of the denominator of BB close to 0. When the estimated step size is fewer than a given positive parameter, the condition will be satisfied, and the PDSBB method will automatically select the average of the past n step size as a new step size. The specific description is as follows:

Firstly, calculate $\eta_t = \eta_t^{BB1}$, which is the BB step size. Secondly, compare the denominator and the given positive parameter ε (in this paper, we set $\varepsilon = 10^{-4}$). If $s_t^T y_t < \varepsilon$, set

$$\eta_t = \frac{1}{t}\sum_{i=0}^{t-1}\eta_i \tag{8}$$

That is

$$\eta_t = \begin{cases} \eta_t^{BB1}, s_t^T y_t > \varepsilon \\ \eta_t = \frac{1}{t}\sum_{i=0}^{t-1}\eta_i, s_t^T y_t < \varepsilon \end{cases} \tag{9}$$

The pseudo-code for PDSBB is given in Table I:

TABLE I
Algorithm of Positive Defined Stabilized Barzilai -Borwein (PDSBB) method

Evaluate g_0 and g_1, given $\varepsilon = 0.001 > 0$
For $k = 0,1,...$ **do**
If $g_k = 0$ then stop.
Set $s_{k-1} \leftarrow w_k - w_{k-1}$ and $y_{k-1} \leftarrow g_k - g_{k-1}$
Compute η_k by formula (9)
set $x_{k+1} \leftarrow g_k - g_{k-1}$ and evaluate g_{k+1}.
end for

The PDSBB method automatically adjusts the step size and stabilizes the step size to an appropriate value (best-tuned step size). The step size in the PDSBB algorithm is divided by updating the frequency m, and then combining it with the SVRG algorithm proposed by [11]. Note that SVRG determines the best step size by comparing several runs where for each run a fixed step-size is set manually. The new algorithm called the stochastic variance reduction gradient with positive defined stabilized Barzilai-Borwein (SVRG-PDSBB) is shown in Table II.

One may notice that if we set $\eta_k = \eta$ in SVRG-PDSBB, then the algorithm turns to SVRG.

TABLE II
ALGORITHM OF SVRG WITH PDSBB STEP SIZE (SVRG-PDSBB)

Parameters: update frequency m, step size η, initial point $\tilde{w}_0$, small positive ε
For $k = 0,1,...$ **do**

$$g_k = \frac{1}{n}\sum_{i=1}^{n}\nabla f_i(\tilde{w}_k)$$

if $k > 0$ then

$$\gamma_k = \frac{1}{m}\|\tilde{w}_k - \tilde{w}_{k-1}\|_2^2 / (\tilde{w}_k - \tilde{w}_{k-1})^T(g_k - g_{k-1})$$

if $\gamma_k > \varepsilon$

$$\eta_k = \gamma_k$$

else

$$\eta_k = \frac{1}{k}\sum_{t=0}^{k-1}\eta_i$$

end if

$$w_0 = \tilde{w}_k$$

for $t = 0,1,...,m-1$ do
Randomly pick $i_t \in \{1,...,n\}$

$$w_{t+1} = w_t - \eta_k(\nabla f_{i_t}(w_k) - \nabla f_{i_t}(\tilde{w}_k) + g_k)$$

end for

$$\tilde{w}_{t+1} = w_m$$

end for

III. CONVERGENCE ANALYSIS

In this section, the proof of the convergence of the SVRG-PDSBB system is presented. The following assumptions and lemma are provided in which most ML models meet these assumptions. Reference [12]-[13] provided further details on these assumptions and lemma.

Assumption 1: The objective function $F(w)$ is μ-strongly convex, which is

$$\begin{aligned}&F(w)\\ &\geq F(w) + \nabla F(w)^T(v-w) + \frac{\mu}{2}\|w-v\|_2^2, \forall w,v \in \mathbb{R}^d\end{aligned} \tag{10}$$

Assumption 2: The gradient of f is L-Lipschitz continuous, which is

$$\|\nabla f_i(w) - \nabla f_i(v)\|_2 \leq L\|w-v\|_2, \forall w,v \in R^d \tag{11}$$

Lemma 1: $f(w): R^d \to R$ is convex and its gradient is L-Lipschitz continuous, then

$$\|\nabla f(w) - \nabla f(v)\|_2^2 \leq L\|w-v\|^T(\nabla f(w) - \nabla f(v)), \forall w,v \in R^d \tag{12}$$

Theorem 1: Define

$$\alpha_k := (1-2\eta_k\mu(1-\eta_k L))^m + \frac{4\eta_k L^2}{\mu(1-\eta_k L)} \tag{13}$$

then for SVRG-PDSBB, the following inequality for the $k-th$ epoch is held:

$$E\|\tilde{w}_{t+1} - w_k\|_2^2 < \alpha_k\|\tilde{w}_k - w^*\|_2^2 \tag{14}$$

where w^* is the optimal solution to function (1).

Similar to [14], Theorem 1 is proven as follows:

Proof: For the $k-th$ epoch of SVRG-PDSBB, let $v_{i_t}^t = \nabla f_{i_t}(w_t) - \nabla f_{i_t}(\tilde{w}_k) + \nabla F(\tilde{w}_k)$, then

$$E\|v_{i_t}^t\|_2^2 \leq 2L(w_t - w^*)^T\nabla F(w_t) + 8L^2\|\tilde{w}_k - w^*\|_2^2 \tag{15}$$

Next, the distance of w_{t+1} to w^* is calculated by

$$\begin{aligned}&E\|w_{t+1} - w^*\|_2^2\\ &\leq (1-2\eta_k\mu(1-\eta L))\|w_t - w^*\|_2^2 + 8\eta_k^2L^2\|\tilde{w}_k - w^*\|_2^2\end{aligned} \tag{16}$$

Since $\tilde{w}_k = w_0$ and $\tilde{w}_{k+1} = w_m$, we can get

$$\begin{aligned}&E\|w_{t+1} - w^*\|_2^2\\ &< \left[(1-2\eta_k\mu(1-\eta_k L))^m + \frac{4\eta_k L^2}{\mu(1-\eta_k L)}\right]\|w_k - w^*\|_2^2\\ &= \alpha_k\|w_k - w^*\|_2^2\end{aligned} \tag{17}$$

Theorem 2: Denote $\theta = (1-e^{-2\mu/L})/2$, then it is obvious that $\theta \in (0, 1/2)$. In SVRG-PDSBB, choosing m such that

$$m > \max\left\{\frac{2}{\log(1-2\theta) + 2\mu/L}, \frac{4L^2}{\theta\mu^2} + \frac{L}{\mu},\right\} \tag{18}$$

will make the SVRG-PDSBB to converge linearly as expected

$$E\|\tilde{w}_k - w^*\|_2^2 < (1-\theta)^2\|\tilde{w}_0 - w^*\|_2^2 \tag{19}$$

Proof: We choose $\eta_k = \frac{1}{m}\frac{\|s_k\|^2}{s_k^T y_k}$ *or* $\eta_k = \frac{1}{k}\sum_{i=0}^{k-1}\eta_i$. Using the strong convexity of $F(w)$, it can be derived that the upper bound of the PDSBB step size in the SVRG-PDSBB method is:

$$\begin{aligned}\eta_k &= \frac{1}{m}\bullet\frac{\|\tilde{w}_k - \tilde{w}_{k-1}\|^2}{(\tilde{w}_k - \tilde{w}_{k-1})^T(g_k - g_{k-1})}\\ &\leq \frac{1}{m}\bullet\frac{\|\tilde{w}_k - \tilde{w}_{k-1}\|^2}{(\tilde{w}_k - \tilde{w}_{k-1})^T(g_k - g_{k-1})}\\ &\leq \frac{1}{m}\bullet\frac{\|\tilde{w}_k - \tilde{w}_{k-1}\|^2}{\mu\|\tilde{w}_k - \tilde{w}_{k-1}\|^2}\\ &= \frac{1}{m\mu}\end{aligned} \tag{20}$$

or

$$\eta_k = \frac{1}{k}\sum_{i=0}^{k-1}\eta_i < \frac{1}{k}\bullet k\bullet\frac{1}{m\mu} = \frac{1}{m\mu} \tag{21}$$

Thus, the upper bound of the PDSBB step size is $\frac{1}{m\mu}$.

Similarly, using the $L-$Lipschitz endurance of $\nabla F(x)$, it is known that $\eta_k > \frac{1}{mL}$. Therefore,

$$\alpha_k = (1-2\eta_k\mu(1-\eta_k L))^m + \frac{4\eta_k L^2}{\mu(1-\eta_k L)}$$

$$\leq (1-\frac{2\mu}{mL}(1-\frac{L}{m\mu}))^m + \frac{4L^2}{m\mu^2(1-L/(m\mu))}$$

$$\leq \exp\left(-\frac{2\mu}{mL}(1-\frac{L}{m\mu})\cdot m\right) + \frac{4L^2}{m\mu^2(1-L/(m\mu))} \quad (22)$$

$$= \exp\left(-\frac{2\mu}{L}+\frac{2}{m}\right) + \frac{4L^2}{m\mu^2 - L\mu}$$

Substituting (18) into the inequality (22),

$$\alpha_k < \exp\left(-\frac{2\mu}{L} + \log(1-2\theta) + 2\mu/L\right)$$
$$+\frac{4L^2}{\frac{4L^2}{\theta} + L\mu - L\mu} = 1-2\theta+\theta = 1-\theta \quad (23)$$

The conclusion is established by applying Theorem 1.

IV. Numerical Experiments

In this section, the proposed PDSBB algorithm is verified by numerical experiments, and the results are compared to the deterministic algorithm of SVRG and the stochastic optimization algorithm of SVRG-BB. In the numerical experiments, Logistic Regression (LR) model with l_2 − norm regularization is chosen with the objective function of:

$$\min_w F(w) = \frac{1}{n}\sum_{i=1}^{n}\log[1+\exp(-b_i a_i^T w)] + \frac{\lambda}{2}\|w\|_2^2 \quad (24)$$

where $a_i \in R^d$ are the feature vector, $b_i \in \{\pm 1\}$ are the class label of the $i-th$ sample and $\lambda > 0$ is a weighting parameter. Table III presents the details of computational experiments concerning the LR using three data sets.

TABLE III
DATA AND MODEL INFORMATION OF THE EXPERIMENTS

Data sets	n	d	λ
rcv1.binary	20,242	47,236	10^{-5}
w8a	49,749	300	10^{-4}
ijcnn1	49,990	22	10^{-4}

Note: n represents the number of samples and d represents the data dimension

In the numerical experiments, the following statements are considered:

1) whether or not these algorithms can realize the identical side by side of sub-optimality as in existing algorithms.
2) whether or not the algorithms are sensitive to the choice of initial step sizes.

Fig. 1, Fig. 2, and Fig. 3 depict the step size results on data sets ijcnn1, w8a, and rcv1, respectively where η_{PDSBB0} indicates the initial step size of SVRG-PDSBB, η_{BB0} denotes the first step size of the algorithm SVRG-BB, and η presents the fixed step size of the algorithm SVRG. We use dashed lines with different markers to represent varying step sizes for the algorithm SVRG, dotted lines with different markers to stand for the step size results for the algorithm SVRG-BB with different initial step sizes, and solid lines with different markers to denote the step size results for the algorithm SVRG-PDSBB with different initial step sizes. For example, solid line with circles represents the step size performance of the algorithm SVRG-PDSBB when the initial step size is 0.1.

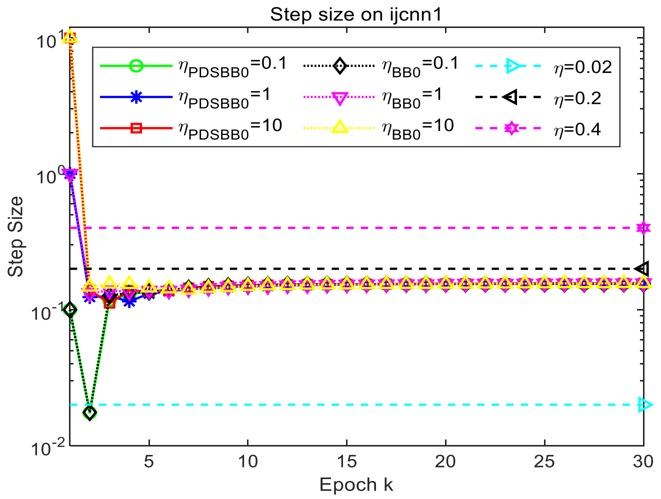

Fig. 1. Step size results of SVRG-PDSBB, SVRG-BB and SVRG on data set ijcnn1 with different initial steps.

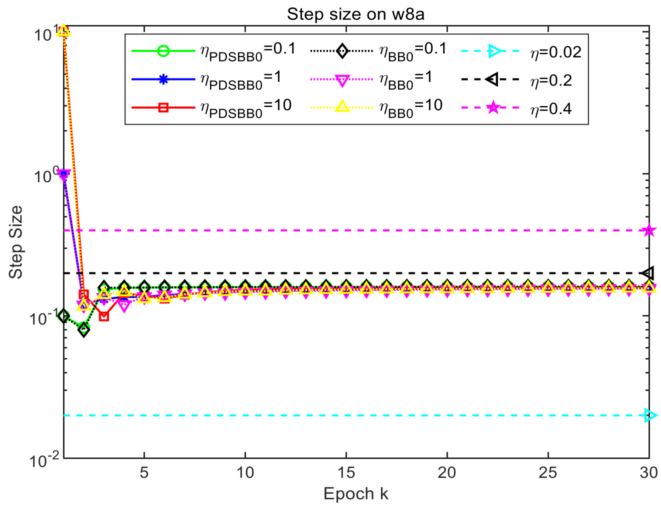

Fig. 2. Step size results of SVRG-PDSBB, SVRG-BB and SVRG on data set w8a with different initial steps.

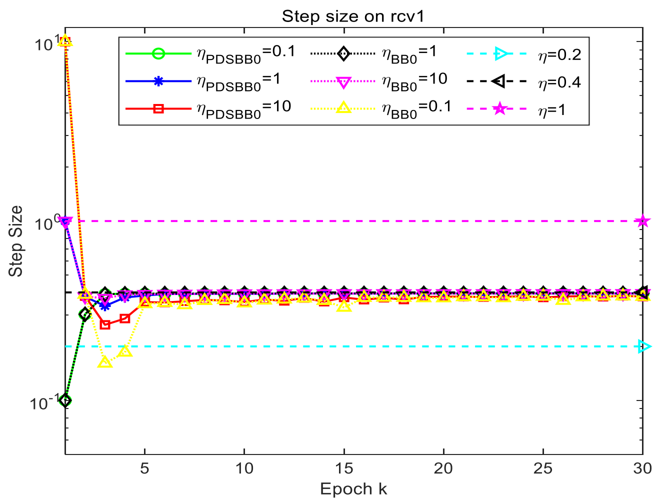

Fig. 3. Step size results of SVRG-PDSBB, SVRG-BB and SVRG on data set ijcnn1 with different initial steps.

The second dash line always denotes the best-tuned step size of the SVRG. The x-axis represents the number of epochs, k, which corresponds to the number of outer loops in Algorithm 2. The y-axis denotes the step size η_k. For SVRG-PDSBB and SVRG-BB methods, the choice of the initial step sizes are 0.1,1, and 10, respectively, as shown in Fig. 1, Fig. 2, and Fig. 3.

On the other hand, different step sizes for different data sets are chosen for SVRG. For example, a step size of η =0.02, 0.2, and 0.4 are chosen for data sets ijcnn1 and w8a, and a step size of η =0.2, 0.4, and 1 are chosen for data set rcv1.

Regardless of the initial step size, after several epochs, on all data sets the newly proposed SVRG-PDSBB converges to the best-tuned step sizes.

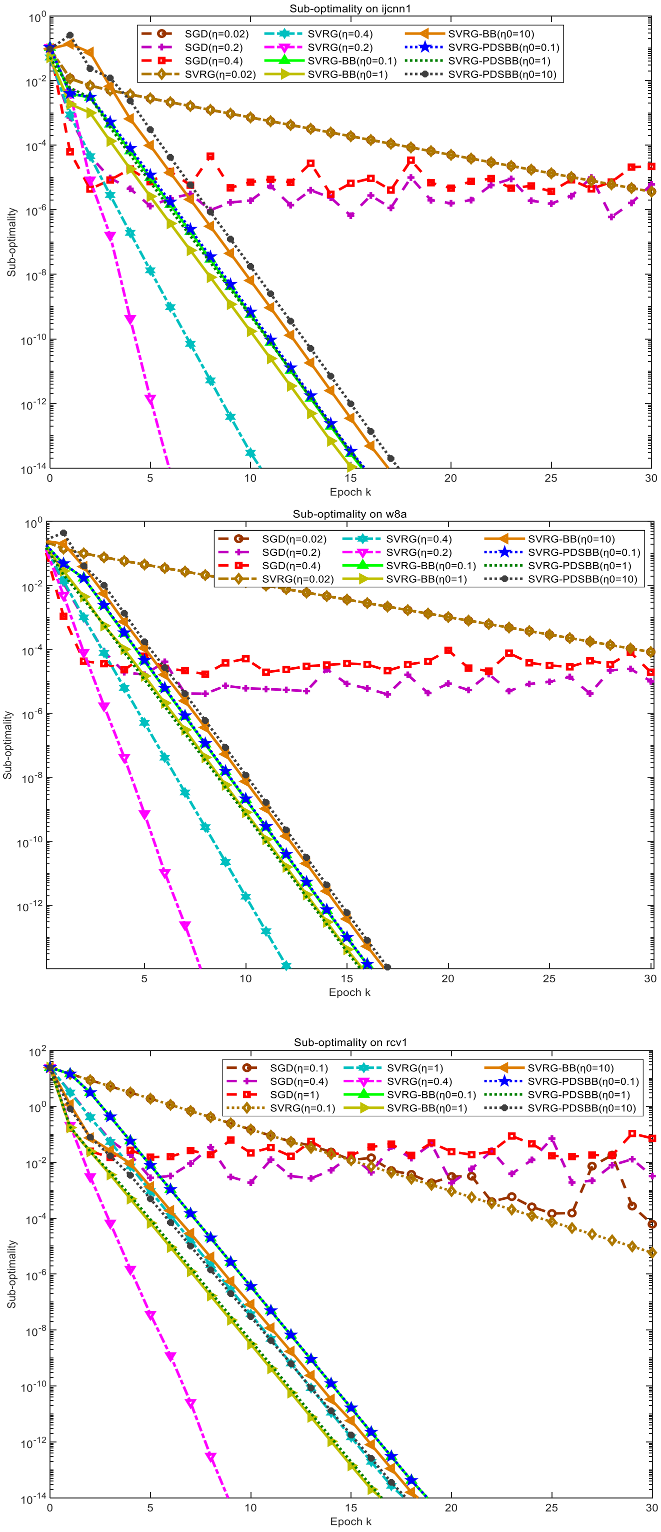

Fig. 4. Sub-optimality results on ijcnn1, w8a, and rcv1 with different initial step size.

In Fig. 4, the x-axis signifies the epochs, k, that denotes the number of outer loops in Algorithm 2. The y-axis represents sub-optimality $F(\tilde{w}_k)-F(w^*)$ on each data set with different initial step sizes. As compared to SVRG-BB algorithms, our new proposed algorithm (SVRG-PDSBB) has achieved similar sub-optimality performance results on all data sets rcv1, w8a and ijcnn1.

From all the subplots in Figure 4, it can be seen that SVRG-PDSBB has achieved suboptimality of 10^{-14} within approximately 15 epochs with different initial step sizes. This demonstrates the effectiveness of the proposed algorithm. All sub-figures in Fig. 4 also show that SVRG-PDSBB is reaching the same level of sub-optimality as SVRG for the best-tuned step size except with slightly larger epochs. However, SVRG-PDSBB outperforms SGD with all choices of step size. In addition, the step size for SVRG is adjusted manually, while the step size for SVRG-PDSBB is adjusted automatically. Thus, SVRG-PDSBB performs better and more practical than SVRG as it automatically generates the optimal step size when running algorithms.

TABLE IV
ACCURACY RATE ON DIFFERENT DATA SETS

Data sets	ALGORITHM	Initial Step size η/η_0	Accuracy
ijcnn1	SGD	0.1	0.7555
	SVRG		0.7360
	SVRG-BB		0.9680
	SVRG-PDSBB		0.9653
	SGD	0.2	0.7019
	SVRG		0.7067
	SVRG-BB		0.9866
	SVRG-PDSBB		0.9865
	SGD	0.4	0.6785
	SVRG		0.6743
	SVRG-BB		0.9877
	SVRG-PDSBB		0.9855
w8a	SGD	0.1	0.7723
	SVRG		0.7485
	SVRG-BB		0.9647
	SVRG-PDSBB		0.9877
	SGD	0.2	0.7090
	SVRG		0.7025
	SVRG-BB		0.9859
	SVRG-PDSBB		0.9872
	SGD	0.4	0.6613
	SVRG		0.6636
	SVRG-BB		0.9858
	SVRG-PDSBB		0.9871
rcv1	SGD	0.1	0.7525
	SVRG		0.7432
	SVRG-BB		0.9871
	SVRG-PDSBB		0.9665
	SGD	0.2	0.7245
	SVRG		0.7168
	SVRG-BB		0.9875
	SVRG-PDSBB		0.9868
	SGD	0.4	0.6853
	SVRG		0.7004
	SVRG-BB		0.9865
	SVRG-PDSBB		0.9868

The accuracy rate for algorithm SGD, SVRG, SVRG-BB, SVRG-PDSBB on different data sets with different step size (initial step size).

To observe the final classification accuracy of four algorithms (SGD, SVRG, SVRG-BB, and SVRG-PDSBB), we let η be the fix step size for SGD and SVRG while η_0, the initial step size for SVRG-BB and SVRG-PDSBB, is set as 0.1, 0.2, and 0.4, when testing the algorithms separately on the three data sets. In the last column of Table IV, it can be seen that overall, SGD and SVRG have almost similar final classification accuracies (between 0.66 to 0.77) for each dataset with a fixed step size, while SVRG-BB and SVRG-PDSBB have higher final classification accuracies (mostly above 0.98). From Table 4, we can conclude that the newly proposed algorithm SVRG-PDSBB is effective and significantly improves the final classification accuracy of the original SVRG algorithm when solving the optimization problem described in (24). Based on numerical experiments, the final classification accuracy of the SVRG-PDSBB is consistent with that of SVRG-BB across all data sets.

V. Conclusion

Conferring to the shortcomings of the fixed, decreasing, and existing BB step size, this study has proposed a new step-size called positive defined stabilized Barzilai-Borwein step size (PDSBB) that prevents the denominator from becoming close to zero or even negative. The new stabilized step size is combined with the existing algorithm SVRG to form the SVRG-PDSBB algorithm. By comparing with other algorithms like SVRG-BB and SVRG, the effectiveness of the new algorithm SVRG-PDSBB is proven theoretically and numerically.

Data Availability Statement: The three standard actual data sets used in this study were obtained from the LIBSVM website, http://www.csie.ntu.edu.tw/~cjlin/libsvm.

References

[1] S. Verma, S. P. Sahu, and T. P. Sahu, "Stock Market Forecasting with Different Input Indicators using Machine Learning and Deep Learning Techniques : A Review," *Engineering Letters*, vol. 31, no. 1, pp. 213–229, 2023.

[2] H. Robbins and S. Monro, "A Stochastic Approximation Method," *The Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 400–407, 1951.

[3] A. Nitanda, "Stochastic proximal gradient descent with acceleration techniques," *Advances in Neural Information Processing Systems*, vol. 2, no. January, pp. 1574–1582, 2014.

[4] A. Nitanda, "Accelerated stochastic gradient descent for minimizing finite sums," in *Proc. 19th International Conf. Artificial Intelligence and Statistics*, Cadiz, 2016, vol. 51, pp. 195–203.

[5] S. Ghadimi and G. Lan, "Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization, II: Shrinking procedures and optimal algorithms," *SIAM Journal on Optimization*, vol. 23, no. 4, pp. 2061–2089, 2013.

[6] Z. Yang, Z. Chen, and C. Wang, "Accelerating mini-batch SARAH by step size rules," *Information Sciences*, vol. 558, pp. 157–173, 2021.

[7] Z. Yang, C. Wang, Z. Zhang, and J. Li, "Accelerated stochastic gradient descent with step size selection rules," *Signal Processing*, vol. 159, pp. 171–186, 2019.

[8] S. De, A. Yadav, D. Jacobs, and T. Goldstein, "Automated inference with adaptive batches," in *Proc. 20th International Conf. Artificial Intelligence and Statistics*, Fort Lauderdale, FL, 2017, pp. 1504-1513.

[9] Z. Yang, C. Wang, Z. Zhang, and J. Li, "Random Barzilai–Borwein step size for mini-batch algorithms," *Engineering Applications of Artificial Intelligence*, vol. 72, pp. 124–135, 2018.

[10] J. BARZILAI and J. (Jon) Borwein, "Two-Point Step Size Gradient Methods," *IMA Journal of Numerical Analysis*, vol. 8, pp. 141–148, Jan. 1988.

[11] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," in *Proc. the Twenty-Seventh Annual Conf. on Neural Information Processing Systems*, Nevada, 2013, pp. 1–9.

[12] N. Le Roux, M. Schmidt, and F. Bach, "A stochastic gradient method with an exponential convergence rate for finite training sets," in *Proc. the Twenty-Seventh Annual Conf. on Neural Information Processing Systems*, Nevada, 2012, pp. 2663–2671.

[13] J. J. Moreau, "Fonctions convexes duales et points proximaux dans un espace hilbertien," *Comptes rendus de l'Académie des Sciences*, vol. 255, pp. 2897–2899, 1962.

[14] C. Tan, S. Ma, Y. H. Dai, and Y. Qian, "Barzilai-borwein step size for stochastic gradient descent," in *Proc. The Thirtieth Annual Conf. on Neural Information Processing Systems*, Barcelona, 2016, pp. 685–693.

Weijuan Shi is a PhD candidate in Mathematics of Universiti Teknologi MARA (UiTM), Malaysia. Her main research interests include stochastic optimization and machine learning. She is a lecturer at College of Mathematics and Finance, Hunan University of Humanities, Science and Technology since 2014.

Adibah Shuib is the Main Supervisor of the first author Weijuan Shi. She is an Associate Professor at Universiti Teknologi MARA (UiTM), Malaysia. She received her PhD (Mathematics and Statistics) with specialization in Operations Research from University of Birmingham, United Kingdom. Her expertise includes Mathematical Programming / Optimization Modeling Network Optimization in Transportation, Logistics and Supply Chain.

Zuraida Alwadood is the co-supervisor of the first author Weijuan Shi. She received her PhD degree in Information Technology and Quantitative Sciences from the Universiti Teknologi MARA (UiTM), Malaysia. Her expertise is in Mathematical Programming/Modelling.