Received: 10 May 2021 | Revised: 6 July 2021 | Accepted: 13 July 2021

IET Control Theory & Applications

DOI: 10.1049/cth2.12184

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH PAPER

Almost sure convergence of randomised-difference descent algorithm for stochastic convex optimisation

Xiaoxue Geng[1,2] | **Gao Huang**[3] | **Wenxiao Zhao**[1,2]

[1] Key Laboratory of Systems and Control, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, People's Republic of China

[2] School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, People's Republic of China

[3] Department of Automation, Tsinghua University, Beijing, People's Republic of China

Correspondence
Wenxiao Zhao, School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, People's Republic of China.
Email: wxzhao@amss.ac.cn

Funding information
National Key Research and Development Program of China, Grant/Award Number: 2018YFA0703800; National Nature Science Foundation of China, Grant/Award Numbers: 61822312, 62022048; Strategic Priority Research Program of Chinese Academy of Sciences, Grant/Award Number: XDA27000000

Abstract

Stochastic gradient descent algorithm is a classical and useful method for stochastic optimisation. While stochastic gradient descent has been theoretically investigated for decades and successfully applied in machine learning such as training of deep neural networks, it essentially relies on obtaining the unbiased estimates of gradients/subgradients of the objective functions. In this paper, by constructing the randomised differences of the objective function, a gradient-free algorithm, named the stochastic randomised-difference descent algorithm, is proposed for stochastic convex optimisation. Under the strongly convex assumption of the objective function, it is proved that the estimates generated from stochastic randomised-difference descent converge to the optimal value with probability one, and the convergence rates of both the mean square error of estimates and the regret functions are established. Finally, some numerical examples are prsented.

1 | INTRODUCTION

This paper considers the stochastic optimisation problem of the following form

$$\begin{aligned} &\min_{w\in\mathbf{R}^m} F(w) = \mathbb{E}\big[f(w;\xi)\big], \\ &\text{s.t.} \quad w \in \Omega \subset \mathbf{R}^m, \end{aligned} \tag{1}$$

where ξ is a random variable with some probability distribution and Ω is the constraint set. In the case of empirical risk minimisation, $\{\xi_i\}_{i=1}^n$ are randomly pulled from the training set with some distribution and the problem turns to

$$\begin{aligned} &\min_{w\in\mathbf{R}^m} F(w) = \frac{1}{n}\sum_{i=1}^{n} f(w;\xi_i), \\ &\text{s.t.} \quad w \in \Omega \subset \mathbf{R}^m. \end{aligned} \tag{2}$$

Problems (1) and (2) widely arise in the statistical learning area. For example, the logistic regression, the ridge regression, and the Lasso regression etc., fall into problems (1) and (2). The stochastic gradient descent (SGD) algorithm and its variants are simple but efficient methods for solving such problems. The basic recurrence formula of SGD is given by

$$w_{t+1} = \mathcal{P}_\Omega(w_t - \eta_t \nabla f(w_t, \xi_t)), \tag{3}$$

where η_t is the step size of the algorithm (or called as the learning rate), $\mathcal{P}_\Omega(\cdot)$ is the projection operator, and $\{\nabla f(w,\xi_t)\}_{t\geq 0}$ are the unbiased estimates for the gradients (or more general, subgradients) of $F(\cdot)$, i.e.

$$\mathbb{E}\big[\nabla f(w;\xi_t)\big] = \nabla F(w). \tag{4}$$

The class of recurrence formula (3) was first introduced in the pioneer work [1], where it is named as the stochastic

approximation algorithm (SAA), for solving the roots of unknown functions with observations corrupted by noises. Since then, there have been plenty of researches on this kind of algorithm from diverse research areas including statistics, electrical engineering etc. Some earlier theoretical papers focus on improving the convergence rate of the algorithm and achieving the asymptotical efficiency [2, 3]. An averaged-type algorithm of (3) was proposed in ref. [4], in which the algorithm is shown to be almost surely convergent with the highest possible rate of convergence. Other variants of the algorithm as well as different types of convergence analysis methods are well summarised in ref. [5, 6].

Due to its simple and effective nature, SGD has been a very useful tool for solving large-scale machine learning problems such as the training of deep neural networks and there are many studies on SGD in the computer science community. These researches concentrate on relaxing the convexity and the smoothness assumptions on $F(\cdot)$, improving the rate of convergence of estimates, and reducing the fluctuation of estimates etc. In the paper, [7] the convergence rate $\mathcal{O}(\log(T)/T)$ of SGD for strongly convex functions after T rounds is established, while in ref.[8] a modified version of SGD is proposed and the optimal $\mathcal{O}(1/T)$ rate of convergence is obtained. In ref. [9], with an average of the last αT rounds of estimates for arbitrary but fixed $\alpha \in (0, 1)$ the optimal $\mathcal{O}(1/T)$ rate of convergence is achieved for non-smooth objective functions. The uniform boundedness condition on the stochastic gradients required in the above papers is removed in ref. [10]. The variants of SGD include the momentum algorithm [11], Nesterov accelerated gradient [12], Adagrad [13], Adadelta [14], and Adam [15] etc. See also ref. [16] for the most recent progresses on SGD.

All the above algorithms require the unbiased estimates of the gradients of the objective functions, i.e. the formula (4) being an a priori assumption. As shown in refs. [17–19], there are many situations including bandit optimisation [20–23], distribution optimisation [24], adversarial attacks on neural networks [25], simulation-based optimisation [17], and hyper-parameter tuning [26], where the calculation or the measurement of gradients is computationally expensive, or infeasible. For instance, in bandit optimisation of machine learning research, a player tries to minimise a sequence of loss functions, which are generated by an adversary, and the player can only observe the values of the loss functions at the chosen points. In simulation-based optimisation, the objective function under consideration can only be evaluated using repeated simulation. Therefore, the optimisation algorithms without using the gradient information, also called the derivative-free algorithms, are of importance in both application and theoretical investigation. In fact, there are many studies on the derivative-free algorithms. See, e.g. refs. [18–21, 27–30] and references therein. The basic idea of the derivative-free algorithms lies in constructing the differences with the values of the objective functions to replace the gradients in the algorithms and such kind of algorithms can correspondingly be divided into two categories, the deterministic method [18] and the randomised method [18–21, 27–30], according to the ways of construction for the differences. The randomised methods have promising theoretical properties and are easy to be implemented, and thus have received attention from researchers. Perhaps surprisingly, there are still important theoretical gaps left in the investigation of the derivative-free algorithms to the stochastic optimisation problem (2). One of such problems is the high probability convergence of the derivative-free stochastic optimisation algorithms. To the authors' knowledge, all the above mentioned literature aims at building the bounds on the mathematical expectation of the regret functions of the associated algorithms, but whether the estimates generated from the derivative-free algorithms converge to the optimal solution of the convex optimisation problem with high probability remain unclear. To be specific, the difference between this paper and other literature are summarised in Table 1.

TABLE 1 Comparison of this paper to some related works on derivative-free convex optimisation

Reference	Almost sure convergence	Mean square convergence	Average regret bound
[19]	-	-	$\mathcal{O}\left(\frac{1}{t^{1/2}}\right)$
[20]	-	-	$\mathcal{O}\left(\frac{1}{t^{1/2}}\right)$
[21]	-	-	$\mathcal{O}\left(\frac{1}{t^{1/2}}\right)$
[30]	-	-	$\mathcal{O}\left(\frac{1}{t^{1/2}}\right)$
This paper	$\surd$	$\surd$	$\mathcal{O}\left(\frac{1}{t^{1/3}}\right)$

In this paper, we will investigate the almost sure convergence of the derivative-free stochastic optimisation algorithms. Since we aim at a problem different from those considered in refs. [18–21, 27–30], the mathematical analysis given in these papers cannot be directly applied. The contributions of the paper are as follows:

- Different from the variables uniformly distributed over the unit sphere, or the variables with Gaussian distribution, see, e.g. refs. [19, 21, 30], here we introduce a sequence of bounded but non-zero variables for the randomized difference and then construct the derivative-free stochastic optimisation algorithm, named as the stochastic randomised-difference descent (SRDD) algorithm in this paper.
- We consider not only the optimisation constraint set Ω but also the observation noises for the empirical objective functions.
- We prove that, under the strongly convex and smooth assumption of the objective functions, the estimates generated from SRDD converge to the optimal value with high probability, in fact with probability one, and the convergence rates of both the mean square error of estimates and the regret functions are obtained.
- We compare the performance of SGD and SRDD through the experimental studies, which indicate that the performance of SRDD is comparable with that of SGD.

The remainder of this paper is organised as follows. In Section 2, we propose the SRDD algorithm with some technical assumptions. In Section 3, we present the main results and the mathematical proofs. In Section 4 we give some numerical experiments and in Section 5 we make the conclusions.

Notations: Let $\mathbf{R}^m$ be the m-dimensional Euclidean space. The Euclidean norm of vector $x \in \mathbf{R}^m$ is defined by $\|x\|_2 = \sqrt{x^T x}$, where x^T denote the transpose of x. For matrix $A \in \mathbf{R}^{n\times m}$, its 2-norm $\|A\|_2$ is defined by $\|A\|_2 = \sqrt{\lambda_{\max}(A^T A)}$, i.e. the square root of the largest eigenvalue of $A^T A$. For vectors x and y with the same dimension, the inner product $\langle x, y\rangle$ is defined by $\langle x, y\rangle = x^T y$. Denote the domain of a function $f(\cdot)$ by $\mathbf{dom}(f)$. Denote the gradient of function $f(\cdot)$ at x by $\partial f(x)$. The projection operator $\mathcal{P}_\Omega(\cdot)$ on Ω is defined as $\mathcal{P}_\Omega(x) = \arg\min_{y\in\Omega}\|x - y\|_2$.

2 | STOCHASTIC RANDOMISED-DIFFERENCE DESCENT (SRDD) ALGORITHM

In this section, we first introduce the SRDD algorithm, which is motivated by the Kiefer–Wolfowitz algorithm for the stochastic approximation problem [27]. The differences lie in that the empirical objective function and the projection operator are applied here while they are not considered in [27] and the subsequent literature.

Assume that the constraint set Ω is closed and convex in $\mathbf{R}^m$. Denote the optimal solution of Equation (1) by w_* if it exists, i.e.

$$w_* = \arg\min_{w\in\Omega} F(w).$$

Denote the estimate for w_* at iteration t by w_t. Let $\Delta_{t+1} = [\Delta_{t+1}^1, \Delta_{t+1}^2, \dots, \Delta_{t+1}^m]^T \in \mathbf{R}^m$ be the random perturbation vector at iteration t. Denote by $\{\beta_t\}_{t\geqslant 0}$ a positive sequence tending to zero. At each iteration t, we obtain two measurements of the objective function $[y_{t+1}]^+ = f(w_t + \beta_t\Delta_{t+1}; \xi_{t+1}) + [\varepsilon_{t+1}]^+$ and $[y_{t+1}]^- = f(w_t - \beta_t\Delta_{t+1}; \xi_{t+1}) + [\varepsilon_{t+1}]^-$, where $[\varepsilon_{t+1}]^+$ and $[\varepsilon_{t+1}]^-$ are the corresponding observation noises. Denote $\varepsilon_{t+1} \triangleq [\varepsilon_{t+1}]^+ - [\varepsilon_{t+1}]^-$. The randomised difference is given by

$$d_{t+1} = \frac{\left([y_{t+1}]^+ - [y_{t+1}]^-\right)[\Delta_{t+1}]^{-1}}{2\beta_t}, \tag{5}$$

where

$$[\Delta_{t+1}]^{-1} \triangleq \left[\frac{1}{\Delta_{t+1}^1}, \frac{1}{\Delta_{t+1}^2}, \dots, \frac{1}{\Delta_{t+1}^m}\right]^T.$$

Based on the randomised difference given by Equation (5), the SRDD algorithm is formulated as follows.

ALGORITHM 1 Stochastic randomised-difference descent (SRDD) algorithm

Initialise: w_0

Iterate:

for $t = 0, 1, 2, \dots$ **do**

Choose a learning rate η_t and a perturbation step size β_t.

Generate random variables ξ_{t+1}, Δ_{t+1} and ε_{t+1}.

Calculate a randomised difference d_{t+1}.

Update the new estimate:

$$w_{t+\frac{1}{2}} = w_t - \eta_t d_{t+1},$$

$$w_{t+1} = \mathcal{P}_\Omega\left(w_{t+\frac{1}{2}}\right).$$

end for

Remark 1. In fact, the randomised difference (5) is an estimate of the directional derivative of the empirical objective function $f(w, \xi)$. In the following, by using the martingale method [31] we will prove that the estimates generated from Equation (5) converge to the optimal solution of the convex optimisation problem with probability one. As far as we know, the convergence of the derivative-free stochastic convex optimisation algorithms with probability one has not been reported in existing literature, see e.g. refs. [18–21, 27–30] and references therein.

Remark 2. The stochastic approximation algorithm (SAA) proposed by Robbins and Monro in 1950s is a powerful tool in estimating the zero points of functions with noisy observations [1, 6]. Generally speaking, the iteration formulas for SRDD, SGD, and SAA are of the same type. On the other hand, since SRDD and SGD aim at iteratively estimating the minimum point of a convex objective function, the mathematical analysis of SRDD and SGD is different from that of SAA.

We introduce the following assumptions on the objective functions $F(w)$ and $f(w; \xi)$, the random perturbation sequence $\{\Delta_t\}_{t\geqslant 1}$ as well as the learning rate sequence $\{\eta_t\}_{t\geqslant 0}$ and the perturbation step size sequence $\{\beta_t\}_{t\geqslant 0}$.

Assumption 1. Denote the gradient function of $F(w)$ by $\partial F(w)$. There exists a constant $\mu > 0$ such that for all $w_1, w_2 \in \mathbf{dom}(F)$,

$$F(w_1) \geqslant F(w_2) + \langle \partial F(w_2), w_1 - w_2\rangle + \frac{\mu}{2}\|w_1 - w_2\|_2^2.$$

Assumption 2. For any fixed ξ, $f(w; \xi)$ is convex in w and there exists a positive constant K independent of ξ such that

$$\left|f(w_1; \xi) - f(w_2; \xi)\right| \leqslant K\|w_1 - w_2\|_2$$

for all $w_1, w_2 \in \mathbf{dom}(f)$.

Assumption 3. Denote the gradient of $f(\cdot;\xi)$ at w by $\partial f(w;\xi)$. There exists a constant $L > 0$ independent of ξ such that

$$\|\partial f(w_1;\xi) - \partial f(w_2;\xi)\|_2 \leqslant L\|w_1 - w_2\|_2.$$

Remark 3. These assumptions are widely applied in the literature of stochastic convex optimisation, see, e.g.refs. [9, 10, 19]. To be specific, Assumption 1 implies that $F(w)$ is μ-strongly convex and Assumption 2 indicates that $f(w,\xi)$ is Lipschitz continuous with respect to w. By Assumption 3 we suppose that the function f is L-smooth. However, in algorithm design the gradient information of $f(w,\xi)$ is not required. Thus Algorithm 1 is gradient-free.

For each iteration t, define σ-algebras $\mathcal{F}_t \triangleq \sigma\{w_0, \xi_1, \xi_2, \dots, \xi_t, \Delta_1, \Delta_2, \dots, \Delta_t, \varepsilon_1, \varepsilon_2, \dots, \varepsilon_t\}$ and $\mathcal{F}'_t \triangleq \sigma\{\mathcal{F}_t, \xi_{t+1}\}$.

Assumption 4.

(a) Both $\{\xi_t\}_{t\geqslant 1}$ and $\{\Delta_t\}_{t\geqslant 1}$ are i.i.d.[1] sequences and for each $i \neq j$, $i, j = 1, \dots, m$, Δ_t^i is independent of Δ_t^j.

(b) For each t, the entries of Δ_t satisfy

$$|\Delta_t^i| \leqslant a, \quad \left|\frac{1}{\Delta_t^i}\right| \leqslant b, \quad \mathbb{E}\left[\frac{1}{\Delta_t^i}\right] = 0, \quad i = 1, 2, \dots, m,$$

for some constants $a > 0$ and $b > 0$.

(c) $\{\Delta_t\}_{t\geqslant 1}$, $\{\varepsilon_t\}_{t\geqslant 1}$ and $\{\xi_t\}_{t\geqslant 1}$ are mutually independent. Further, $\mathbb{E}[\varepsilon_t|\mathcal{F}_{t-1}] = 0$ and $\sup_{t\geqslant 1}\mathbb{E}\varepsilon_t^2 < \infty$.

Remark 4. In existing literature on the design of the derivative-free algorithm, c.f. refs. [19, 21, 30], the perturbation variables $\{\Delta_t\}_{t\geqslant 1}$ are usually selected as i.i.d. variables uniformly distributed over the unit sphere, or with Gaussian distribution, and correspondingly, the difference is given by

$$d_{t+1} = \frac{\left([y_{t+1}]^+ - [y_{t+1}]^-\right)}{2\beta_t}[\Delta_{t+1}]. \tag{6}$$

Compared with Equation (6), the perturbation signals in Assumption 4 and the design of the difference in Equation (5) are different, and we make no assumptions on the specific probability distribution of $\{\Delta_t\}_{t\geq 0}$. This makes the convergence analysis in this paper different from those given in refs. [19, 21 30]. From the definition of the randomised difference, we can also find that d_{t+1} is not an unbiased estimate of $\nabla F(w_t)$. Thus the convergence results of SGD cannot be directly applied.

Assumption 5.

(a) $\eta_t > 0$, $\sum_{t=0}^{\infty}\eta_t = \infty$, $\sum_{t=0}^{\infty}\eta_t^2 < \infty$.

(b) $\beta_t > 0$, $\beta_t \to 0$, $\sum_{t=0}^{\infty}\eta_t\beta_t < \infty$, $\sum_{t=0}^{\infty}\frac{\eta_t^2}{\beta_t^2} < \infty$.

Remark 5. Choose

$$\eta_t = \frac{1}{t}, \quad \beta_t = \frac{1}{t^\lambda}$$

with $\lambda \in (0, \frac{1}{2})$. It is directly to verify that such choices on η_t and β_t meet the requirements in Assumption 5.

3 | CONVERGENCE OF SRDD ALGORITHM

Before establishing convergence of Algorithm 1, we first introduce an elementary inequality in existing literature.

Lemma 1 ([31] (Nonexpansive property of projection operator)). *Let Ω be a closed convex set in $\mathbf{R}^m$. Then for the projection operator $\mathcal{P}_\Omega(\cdot)$, it holds that*

$$\|\mathcal{P}_\Omega(x_1) - \mathcal{P}_\Omega(x_2)\|_2 \leqslant \|x_1 - x_2\|_2$$

for all $x_1, x_2 \in \mathbf{R}^m$.

Noting that w_* is the optimal solution of Problem (1), we have the following two technical lemmas.

Lemma 2. *Suppose that Assumptions 3 and 4 hold and $v_{t+1} = \partial f(w_t;\xi_{t+1})$. Then*

$$\mathbb{E}\left[\langle v_{t+1} - d_{t+1}, w_t - w_*\rangle|\mathcal{F}_t\right] \leqslant LC_1\beta_t\|w_t - w_*\|_2, \tag{7}$$

where L is the constant given in Assumption 3 and $C_1 = \mathbb{E}\{\|[\Delta_{t+1}]^{-1}\Delta_{t+1}^T\|_2\|\Delta_{t+1}\|_2\}$.

Proof. From the definition of d_{t+1}, it directly follows that

$$\begin{aligned} d_{t+1} &= \frac{\left([y_{t+1}]^+ - [y_{t+1}]^-\right)[\Delta_{t+1}]^{-1}}{2\beta_t} \\ &= \frac{(f(w_t + \beta_t\Delta_{t+1};\xi_{t+1}) - f(w_t - \beta_t\Delta_{t+1};\xi_{t+1}))[\Delta_{t+1}]^{-1}}{2\beta_t} \\ &\quad + \frac{\varepsilon_{t+1}[\Delta_{t+1}]^{-1}}{2\beta_t}. \end{aligned} \tag{8}$$

By Lebourg's mean value theorem, we know that

$$\begin{aligned} &f(w_t + \beta_t\Delta_{t+1};\xi_{t+1}) - f(w_t - \beta_t\Delta_{t+1};\xi_{t+1}) \\ &= \langle\partial f(w_t + \theta_t\beta_t\Delta_{t+1};\xi_{t+1}), 2\beta_t\Delta_{t+1}\rangle \end{aligned}$$

[1] Independent and identically distributed.

for some $\theta_t \in [-1, 1]$.

Denote $\gamma_{t+1} = \partial f(w_t + \theta_t \beta_t \Delta_{t+1}; \xi_{t+1})$. From Equation (8) and Assumption 4, we obtain the following equalities and inequalities,

$$
\begin{aligned}
&\mathbb{E}\left[\langle v_{t+1} - d_{t+1}, w_t - w_*\rangle | \mathcal{F}_t\right] \\
=&\mathbb{E}\left[\left\langle v_{t+1} - \frac{\langle \gamma_{t+1}, 2\beta_t \Delta_{t+1}\rangle \left[\Delta_{t+1}\right]^{-1}}{2\beta_t}, w_t - w_*\right\rangle \middle| \mathcal{F}_t\right] \\
&- \mathbb{E}\left[\left\langle \frac{\varepsilon_{t+1}\left[\Delta_{t+1}\right]^{-1}}{2\beta_t}, w_t - w_*\right\rangle \middle| \mathcal{F}_t\right] \\
=&\mathbb{E}\left[\left\langle v_{t+1} - \left[\Delta_{t+1}\right]^{-1}\Delta_{t+1}^T \gamma_{t+1}, w_t - w_*\right\rangle \middle| \mathcal{F}_t\right] \\
&- \frac{1}{2\beta_t}\mathbb{E}\left[\left\langle \varepsilon_{t+1}\left[\Delta_{t+1}\right]^{-1}, w_t - w_*\right\rangle \middle| \mathcal{F}_t\right] \\
=&\mathbb{E}\left[\left\langle \left(\left[\Delta_{t+1}\right]^{-1}\Delta_{t+1}^T\right)(v_{t+1} - \gamma_{t+1}), w_t - w_*\right\rangle \middle| \mathcal{F}_t\right] \\
&+ \mathbb{E}\left[\left\langle \left(I - \left[\Delta_{t+1}\right]^{-1}\Delta_{t+1}^T\right)v_{t+1}, w_t - w_*\right\rangle \middle| \mathcal{F}_t\right] \\
&- \frac{1}{2\beta_t}\mathbb{E}[\varepsilon_{t+1}|\mathcal{F}_t]\mathbb{E}[\Delta_{t+1}]^{-T}(w_t - w_*) \\
\leqslant&\mathbb{E}\left[\left\|[\Delta_{t+1}]^{-1}\Delta_{t+1}^T\right\|_2 \|v_{t+1} - \gamma_{t+1}\|_2 \|w_t - w_*\|_2 \middle| \mathcal{F}_t\right] \\
&+ \mathbb{E}\left(\mathbb{E}\left[\left\langle \left(I - [\Delta_{t+1}]^{-1}\Delta_{t+1}^T\right)v_{t+1}, w_t - w_*\right\rangle \middle| \mathcal{F}_t'\right] \middle| \mathcal{F}_t\right) \\
\leqslant&\mathbb{E}\left[\left\|[\Delta_{t+1}]^{-1}\Delta_{t+1}^T\right\|_2 \cdot L\beta_t \|\Delta_{t+1}\|_2 \middle| \mathcal{F}_t\right] \cdot \|w_t - w_*\|_2 \\
&+ \mathbb{E}\left(v_{t+1}^T \mathbb{E}\left(I - \Delta_{t+1}[\Delta_{t+1}]^{-T}\right)(w_t - w_*) \middle| \mathcal{F}_t\right),
\end{aligned} \tag{9}
$$

where for the last inequality Assumption 3 is applied.

By Assumption 4, we know that both $\|\Delta_{t+1}\|_2$ and $\|\Delta_{t+1}^{-1}\|_2$ are bounded and

$$\mathbb{E}\left(I - \Delta_{t+1}[\Delta_{t+1}]^{-T}\right) = 0.$$

Then from Equation (9) we obtain

$$\mathbb{E}\left[\langle v_{t+1} - d_{t+1}, w_t - w_*\rangle | \mathcal{F}_t\right] \leqslant LC_1\beta_t \|w_t - w_*\|_2.$$

This finishes the proof. □

Lemma 3. *Under Assumptions 1–4, it follows that $\mathbb{E}(v_{t+1}|\mathcal{F}_t)$ is the gradient of $F(w)$ at w_t and*

$$\mathbb{E}\left[\langle d_{t+1}, w_t - w_*\rangle | \mathcal{F}_t\right] \geqslant \frac{\mu}{2}\|w_t - w_*\|_2^2 - LC_1\beta_t\|w_t - w_*\|_2, \tag{10}$$

where μ is the constant in Assumption 1.

Proof. Note that $v_{t+1} = \partial f(w_t; \xi_{t+1})$. Define

$$\mathbb{E}\left[\langle d_{t+1}, w_t - w_*\rangle | \mathcal{F}_t\right] \triangleq J_t(1) - J_t(2).$$

with

$$
\begin{aligned}
J_t(1) &= \mathbb{E}\left[\langle v_{t+1}, w_t - w_*\rangle | \mathcal{F}_t\right], \\
J_t(2) &= \mathbb{E}\left[\langle v_{t+1} - d_{t+1}, w_t - w_*\rangle | \mathcal{F}_t\right].
\end{aligned}
$$

Since $v_{t+1} = \partial f(w_t; \xi_{t+1})$, by the convexity of $f(w, \xi)$ given in Assumption 2, for any $w' \in \mathbf{dom}(f)$ we have

$$f(w'; \xi_{t+1}) - f(w_t; \xi_{t+1}) \geqslant \langle v_{t+1}, w' - w_t\rangle. \tag{11}$$

Noting that ξ_{t+1} is independent of $\mathcal{F}_t$ and w_t is $\mathcal{F}_t$-measurable, we obtain the following equality

$$F(w') - F(w_t) = \mathbb{E}\left[f(w'; \xi_{t+1}) - f(w_t; \xi_{t+1}) | \mathcal{F}_t\right], \tag{12}$$

from which and Equation (11),

$$
\begin{aligned}
F(w') - F(w_t) &\geqslant \mathbb{E}\left[\langle v_{t+1}, w' - w_t\rangle | \mathcal{F}_t\right] \\
&= \left\langle \mathbb{E}(v_{t+1}|\mathcal{F}_t), w' - w_t\right\rangle.
\end{aligned} \tag{13}
$$

By the convexity of $F(\cdot)$ we know that $\mathbb{E}(v_{t+1}|\mathcal{F}_t)$ is the gradient of $F(w)$ at w_t, i.e. $\mathbb{E}(v_{t+1}|\mathcal{F}_t) = \partial F(w_t)$.

Denote $s_{t+1} \triangleq \mathbb{E}(v_{t+1}|\mathcal{F}_t)$. By Assumption 1, we have

$$F(w_*) - F(w_t) \geqslant \langle s_{t+1}, w_* - w_t\rangle + \frac{\mu}{2}\|w_* - w_t\|_2^2.$$

So for $J_t(1)$, it follows that

$$
\begin{aligned}
J_t(1) &= \mathbb{E}\left[\langle v_{t+1}, w_t - w_*\rangle | \mathcal{F}_t\right] \\
&= \left\langle \mathbb{E}\left(v_{t+1}|\mathcal{F}_t\right), w_t - w_*\right\rangle \\
&= \langle s_{t+1}, w_t - w_*\rangle \\
&\geqslant F(w_t) - F(w_*) + \frac{\mu}{2}\|w_* - w_t\|_2^2 \\
&\geqslant \frac{\mu}{2}\|w_* - w_t\|_2^2
\end{aligned}
$$

because $F(w_t) - F(w_*) \geqslant 0$. The above inequality together with the definition of $J_t(2)$ and the conclusion of Lemma 2 yield the result. □

Theorem 1. *Assume that Assumptions 1–5 hold. Then $\{w_t\}_{t\geqslant 0}$ generated by Algorithm 1 converges to w_*, the optimal solution of (1), with probability one.*

Proof. By Lemma 1, we have

$$\mathbb{E}\left[\|w_{t+1} - w_*\|_2^2 \middle| \mathcal{F}_t\right]$$

$$
\begin{aligned}
&= \mathbb{E}\left[\|\mathcal{P}_{\Omega}(w_t - \eta_t d_{t+1}) - w_*\|_2^2 \middle| \mathcal{F}_t \right] \\
&\leqslant \mathbb{E}\left[\|w_t - \eta_t d_{t+1} - w_*\|_2^2 \middle| \mathcal{F}_t \right] \\
&= \mathbb{E}\left[\|w_t - w_*\|_2^2 \middle| \mathcal{F}_t \right] + \eta_t^2 \mathbb{E}\left[\|d_{t+1}\|_2^2 \middle| \mathcal{F}_t \right] \\
&\quad - 2\eta_t \mathbb{E}\left[\langle d_{t+1}, w_t - w_* \rangle \middle| \mathcal{F}_t \right]. \qquad (14)
\end{aligned}
$$

From the definition of d_{t+1}, we obtain the following inequality:

$$
\begin{aligned}
&\mathbb{E}\left[\|d_{t+1}\|_2^2 \middle| \mathcal{F}_t \right] \\
&= \mathbb{E}\left[\left\| \frac{(f(w_t + \beta_t \Delta_{t+1}; \xi_{t+1}) - f(w_t - \beta_t \Delta_{t+1}; \xi_{t+1}))[\Delta_{t+1}]^{-1}}{2\beta_t} \right.\right. \\
&\quad \left.\left. + \frac{\varepsilon_{t+1}[\Delta_{t+1}]^{-1}}{2\beta_t} \right\|_2^2 \middle| \mathcal{F}_t \right] \\
&\leqslant 2\mathbb{E}\left[\left\| \frac{(f(w_t + \beta_t \Delta_{t+1}; \xi_{t+1}) - f(w_t - \beta_t \Delta_{t+1}; \xi_{t+1}))}{2\beta_t} \right\|_2^2 \right. \\
&\quad \left. \left\| [\Delta_{t+1}]^{-1} \right\|_2^2 \middle| \mathcal{F}_t \right] + 2\mathbb{E}\left[\left\| \frac{\varepsilon_{t+1}[\Delta_{t+1}]^{-1}}{2\beta_t} \right\|_2^2 \middle| \mathcal{F}_t \right].
\end{aligned}
$$

Then, by noting Assumption 2 that $f(w; \xi)$ is Lipschitz continuous with respect to w, it follows that

$$
\begin{aligned}
\mathbb{E}\left[\|d_{t+1}\|_2^2 \middle| \mathcal{F}_t \right] &\leqslant 2K^2 \mathbb{E}\left[\|\Delta_{t+1}\|_2^2 \left\| [\Delta_{t+1}]^{-1} \right\|_2^2 \middle| \mathcal{F}_t \right] \\
&\quad + \frac{1}{2\beta_t^2} \mathbb{E}\left[\left\| \varepsilon_{t+1}[\Delta_{t+1}]^{-1} \right\|_2^2 \middle| \mathcal{F}_t \right]. \qquad (15)
\end{aligned}
$$

Since Δ_{t+1} is independent of both ε_{t+1} and $\mathcal{F}_t$, from the above inequality we obtain

$$
\begin{aligned}
&\mathbb{E}\left[\|d_{t+1}\|_2^2 \middle| \mathcal{F}_t \right] \\
&\leqslant 2K^2 \mathbb{E}\left[\|\Delta_{t+1}\|_2^2 \left\| [\Delta_{t+1}]^{-1} \right\|_2^2 \right] \\
&\quad + \frac{1}{2\beta_t^2} \mathbb{E}[\varepsilon_{t+1}^2 | \mathcal{F}_t] \mathbb{E}\left\| [\Delta_{t+1}]^{-1} \right\|_2^2 \\
&\leqslant 2K^2 C_3 + \frac{C_2}{2\beta_t^2}, \qquad (16)
\end{aligned}
$$

where $C_3 = \mathbb{E}[\|\Delta_{t+1}\|_2^2 \|[\Delta_{t+1}]^{-1}\|_2^2]$ and $C_2 = \mathbb{E}[\varepsilon_{t+1}^2 | \mathcal{F}_t] \cdot \mathbb{E}\|[\Delta_{t+1}]^{-1}\|_2^2$.

Since w_t is measurable with respect to $\mathcal{F}_t$, from Equations (14), (16), and (10) and the elementary inequality $2|x| \leqslant (1 + x^2)$, it yields that

$$
\begin{aligned}
&\mathbb{E}\left[\|w_{t+1} - w_*\|_2^2 \middle| \mathcal{F}_t \right] \\
&\leqslant \|w_t - w_*\|_2^2 - \mu\eta_t \|w_t - w_*\|_2^2 + 2LC_1 \eta_t \beta_t \|w_t - w_*\|_2 \\
&\quad + 2K^2 C_3 \eta_t^2 + \frac{C_2}{2} \frac{\eta_t^2}{\beta_t^2} \\
&\leqslant (1 + LC_1 \eta_t \beta_t) \|w_t - w_*\|_2^2 - \mu\eta_t \|w_t - w_*\|_2^2 \\
&\quad + LC_1 \eta_t \beta_t + 2K^2 C_3 \eta_t^2 + \frac{C_2}{2} \frac{\eta_t^2}{\beta_t^2}. \qquad (17)
\end{aligned}
$$

Since $\sum_{t=0}^{\infty} \eta_t \beta_t < \infty$, $\sum_{t=0}^{\infty} \eta_t^2 < \infty$ and $\sum_{t=0}^{\infty} \frac{\eta_t^2}{\beta_t^2} < \infty$, by applying Theorem A.1 in Appendix, we obtain that

$$
\|w_t - w_*\|_2^2 \xrightarrow[t \to \infty]{} W \geqslant 0 \text{ a.s.} \qquad (18)
$$

and

$$
\sum_{t=0}^{\infty} \mu\eta_t \|w_t - w_*\|_2^2 < \infty \text{ a.s.} \qquad (19)
$$

We proceed to show that $\|w_t - w_*\|_2^2 \to 0$ as $t \to \infty$. Assume the converse, i.e. $W > 0$. Then there exists $\delta > 0$ and $t_0 > 0$ such that for all $t > t_0$ we have $\|w_t - w_*\|_2^2 > \delta$ and hence,

$$
\begin{aligned}
&\sum_{t=0}^{\infty} \mu\eta_t \|w_t - w_*\|_2^2 \\
&= \sum_{t=0}^{t_0} \mu\eta_t \|w_t - w_*\|_2^2 + \sum_{t=t_0+1}^{\infty} \mu\eta_t \|w_t - w_*\|_2^2 \\
&\geqslant \mu\delta \sum_{t=t_0+1}^{\infty} \eta_t = \infty,
\end{aligned}
$$

which contradicts with Equation (19). This proves that $\|w_t - w_*\|_2^2 \to 0$ as $t \to \infty$ with probability one. □

Theorem 2. *Choose $\eta_t = \frac{1}{t}$ and $\beta_t = \frac{1}{t^{\delta}}$ with $0 < \delta < \frac{1}{2}$. Suppose that Assumptions 1–4 hold. Then for Algorithm 1,*

$$
\mathbb{E}\|w_t - w_*\|_2^2 = \mathcal{O}\left(\max\left\{ \frac{1}{t^{\delta}}, \frac{1}{t^{1-2\delta}}, \frac{1}{t^{(1-\epsilon)\mu}} \right\} \right) \qquad (20)
$$

for all t large enough, where μ is the strong convexity parameter specified in Assumption 1 and $\epsilon > 0$ can be arbitrarily small.

Proof. Since $\beta_t \to 0$ as $t \to \infty$, for any fixed $\epsilon > 0$ there exists t_0 sufficiently large such that $LC_1\eta_t\beta_t - \mu\eta_t \leqslant -(1-\epsilon)\mu\eta_t$ for all $t \geqslant t_0$. Then from Equation (17) it follows that for $t \geqslant t_0$,

$$\begin{aligned}
&\mathbb{E}\left[\|w_{t+1} - w_*\|_2^2 \middle| \mathcal{F}_t\right] \\
&\leqslant (1 + LC_1\eta_t\beta_t - \mu\eta_t)\|w_t - w_*\|_2^2 + LC_1\eta_t\beta_t \\
&\quad + 2K^2C_3\eta_t^2 + \frac{C_2}{2}\frac{\eta_t^2}{\beta_t^2} \\
&\leqslant (1 - (1-\epsilon)\mu\eta_t)\|w_t - w_*\|_2^2 + LC_1\eta_t\beta_t \\
&\quad + 2K^2C_3\eta_t^2 + \frac{C_2}{2}\frac{\eta_t^2}{\beta_t^2}.
\end{aligned} \tag{21}$$

By taking expectation to both sides of Equation (21) we obtain

$$\begin{aligned}
&\mathbb{E}\|w_{t+1} - w_*\|_2^2 \\
&\leqslant (1 - (1-\epsilon)\mu\eta_t)\mathbb{E}\|w_t - w_*\|_2^2 + LC_1\eta_t\beta_t \\
&\quad + 2K^2C_3\eta_t^2 + \frac{C_2}{2}\frac{\eta_t^2}{\beta_t^2} \\
&\leqslant \prod_{i=t_0}^{t}(1 - (1-\epsilon)\mu\eta_i)\mathbb{E}\|w_{t_0} - w_*\|_2^2 \\
&\quad + \sum_{j=t_0}^{t}[LC_1\eta_j\beta_j + 2K^2C_3\eta_j^2]\prod_{i=j+1}^{t}(1 - (1-\epsilon)\mu\eta_i) \\
&\quad + \sum_{j=t_0}^{t}\left[\frac{C_2}{2}\frac{\eta_j^2}{\beta_j^2}\right]\prod_{i=j+1}^{t}(1 - (1-\epsilon)\mu\eta_i).
\end{aligned} \tag{22}$$

By using $1 - x \leqslant \mathrm{e}^{-x}$ for $x \geqslant 0$, it follows that

$$\begin{aligned}
\prod_{i=t_0}^{t}(1 - (1-\epsilon)\mu\eta_i) &\leqslant \mathrm{e}^{-(1-\epsilon)\mu\sum_{i=t_0}^{t}\eta_i} \\
&= \mathrm{e}^{-(1-\epsilon)\mu\sum_{i=t_0}^{t}\frac{1}{i}} \\
&\leqslant \mathrm{e}^{-(1-\epsilon)\mu(\log(t+1) - \log t_0)} \\
&= \frac{t_0^{(1-\epsilon)\mu}}{(t+1)^{(1-\epsilon)\mu}}.
\end{aligned} \tag{23}$$

Define $z_t^{(1)} \triangleq \sum_{j=t_0}^{t} LC_1\eta_j\beta_j \prod_{i=j+1}^{t}(1 - (1-\epsilon)\mu\eta_i)$. By (23) we have that

$$\begin{aligned}
z_t^{(1)} &= \sum_{j=t_0}^{t} LC_1\frac{1}{j^{1+\delta}}\prod_{i=j+1}^{t}(1 - (1-\epsilon)\mu\eta_i) \\
&\leqslant LC_1\sum_{j=t_0}^{t}\frac{1}{j^{1+\delta}}\frac{(j+1)^{(1-\epsilon)\mu}}{(t+1)^{(1-\epsilon)\mu}} \\
&\leqslant \frac{2LC_1}{(t+1)^{(1-\epsilon)\mu}}\sum_{j=t_0}^{t}\frac{1}{j^{1+\delta-(1-\epsilon)\mu}} \leqslant \frac{2LC_1}{t^{\delta}}.
\end{aligned} \tag{24}$$

Define $z_t^{(2)} = \sum_{j=t_0}^{t} 2K^2C_3\eta_j^2\prod_{i=j+1}^{t}(1 - (1-\epsilon)\mu\eta_i)$ and $z_t^{(3)} = \sum_{j=t_0}^{t}\frac{C_2}{2}\frac{\eta_j^2}{\beta_j^2}\prod_{i=j+1}^{t}(1 - (1-\epsilon)\mu\eta_i)$. Similar to Equation (24), we obtain that

$$z_t^{(2)} \leqslant \frac{4K^2C_3}{t} \quad \text{and} \quad z_t^{(3)} \leqslant \frac{C_2}{t^{1-2\delta}}. \tag{25}$$

Hence, Equation (20) can be proved by Equations (22)–(25) with the observation that $\epsilon > 0$ can be sufficiently small. □

Remark 6. From Equation (20) it follows that, to obtain the higher convergence rate, the optimal value for δ in $\beta_t = \frac{1}{t^{\delta}}$ is $\frac{1}{3}$, which in turn indicates that

$$\mathbb{E}\|w_t - w_*\|_2^2 = \mathcal{O}\left(\max\left\{\frac{1}{t^{\frac{1}{3}}}, \frac{1}{t^{(1-\epsilon)\mu}}\right\}\right). \tag{26}$$

Next, we define the regret function $R(T)$ as follow:

$$R(T) \triangleq \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}(F(w_t) - F(w_*)),$$

where w_* is the optimal solution of Problem (1).

Theorem 3. *Choose $\eta_t = t^{-1}$ and $\beta_t = t^{-\frac{1}{3}}$. Under Assumptions 1–4, we have that*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\big(F(w_t) - F(w_*)\big) = \mathcal{O}\left(\max\left\{\frac{1}{T^{\frac{1}{3}}}, \frac{1}{T^{(1-\epsilon)\mu}}\right\}\right) \tag{27}$$

for any $\epsilon > 0$.

Proof. By taking expectation to both sides of Equation (14), we have

$$\begin{aligned}
\mathbb{E}\|w_{t+1} - w_*\|_2^2 \leqslant &\mathbb{E}\|w_t - w_*\|_2^2 + \eta_t^2\mathbb{E}\|d_{t+1}\|_2^2 \\
&- 2\eta_t\mathbb{E}[\langle d_{t+1}, w_t - w_*\rangle].
\end{aligned} \tag{28}$$

From Equation (28) we obtain

$$\begin{aligned}&\mathbb{E}[\langle d_{t+1}, w_t - w_*\rangle]\\ \leqslant&\frac{1}{2\eta_t}\left[\mathbb{E}\|w_t - w_*\|_2^2 - \mathbb{E}\|w_{t+1} - w_*\|_2^2\right] + \frac{\eta_t}{2}\mathbb{E}\|d_{t+1}\|_2^2. \quad (29)\end{aligned}$$

By Lemma 2, it follows that

$$\begin{aligned}&\mathbb{E}[\langle v_{t+1}, w_t - w_*\rangle|\mathcal{F}_t]\\ \leqslant&\mathbb{E}[\langle d_{t+1}, w_t - w_*\rangle|\mathcal{F}_t] + LC_1\beta_t\|w_t - w_*\|_2. \quad (30)\end{aligned}$$

By Lemma 3 we know that $\mathbb{E}(v_{t+1}|\mathcal{F}_t) = \partial F(w_t)$ and hence

$$F(w_*) - F(w_t) \geqslant \left\langle \mathbb{E}\left(v_{t+1}|\mathcal{F}_t\right), w_* - w_t\right\rangle,$$

which combining with Equation (30) and the fact that w_t being $\mathcal{F}_t$-measurable yields

$$\begin{aligned}&F(w_t) - F(w_*)\\ \leqslant&\mathbb{E}\left[\langle d_{t+1}, w_t - w_*\rangle|\mathcal{F}_t\right] + LC_1\beta_t\|w_t - w_*\|_2. \quad (31)\end{aligned}$$

By taking expectation to both sides of Equation (31) and then summing up the inequalities from $t = 1, 2, \ldots, T$, also by noting Equation (29) we obtain that

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}(F(w_t) - F(w_*))\\ \leqslant&\frac{1}{T}\sum_{t=1}^{T}\frac{1}{2\eta_t}\left[\mathbb{E}\|w_t - w_*\|_2^2 - \mathbb{E}\|w_{t+1} - w_*\|_2^2\right]\\ &+ \frac{1}{T}\sum_{t=1}^{T}\frac{\eta_t}{2}\mathbb{E}\|d_{t+1}\|_2^2 + \frac{1}{T}\sum_{t=1}^{T}LC_1\beta_t\mathbb{E}\|w_t - w_*\|_2. \quad (32)\end{aligned}$$

From Equation (16), we have

$$\mathbb{E}\|d_{t+1}\|_2^2 \leqslant 2K^2C_3 + \frac{C_2}{2\beta_t^2}. \quad (33)$$

Then from Equations (32) and (33), it follows that

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}(F(w_t) - F(w_*))\\ \leqslant&\frac{1}{T}\sum_{t=1}^{T}\left(\eta_t K^2C_3 + \frac{C_2\eta_t}{4\beta_t^2}\right) + \frac{LC_1}{T}\sum_{t=1}^{T}\beta_t\mathbb{E}\|w_t - w_*\|_2\\ &+ \frac{1}{2T}\sum_{t=2}^{T}\mathbb{E}\|w_t - w_*\|_2^2\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)\\ &+ \frac{1}{2T\eta_1}\mathbb{E}\|w_1 - w_*\|_2^2. \quad (34)\end{aligned}$$

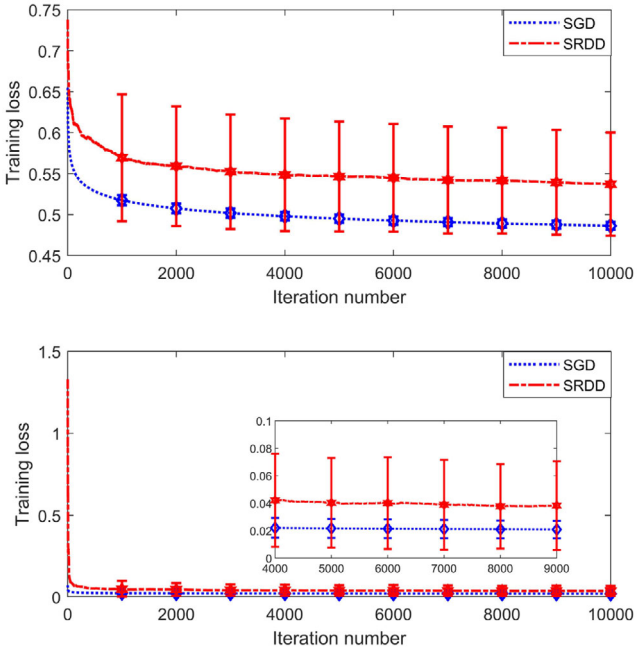

FIGURE 1 Comparison between SGD and SRDD for logistic regression. Topic subfigure: data set `ijcnn1`. Lower subfigure: data set `covtype.binary`

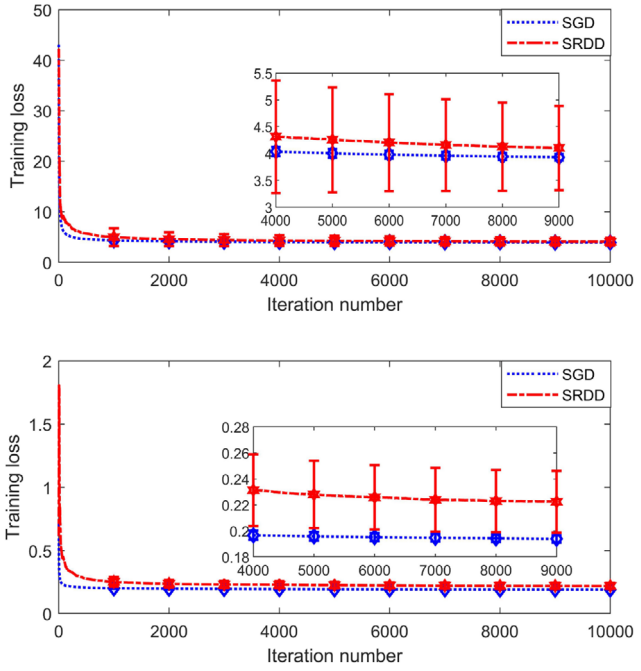

FIGURE 2 Comparison between SGD and SRDD for ridge regression. Topic subfigure: data set `abalone`. Lower subfigure: data set `mg`

Since $\eta_t = t^{-1}$ and $\beta_t = t^{-\frac{1}{3}}$. Noticing Equation (26) and using Jensen's Inequality, from Equation (34) we establish that

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}(F(w_t) - F(w_*))$$

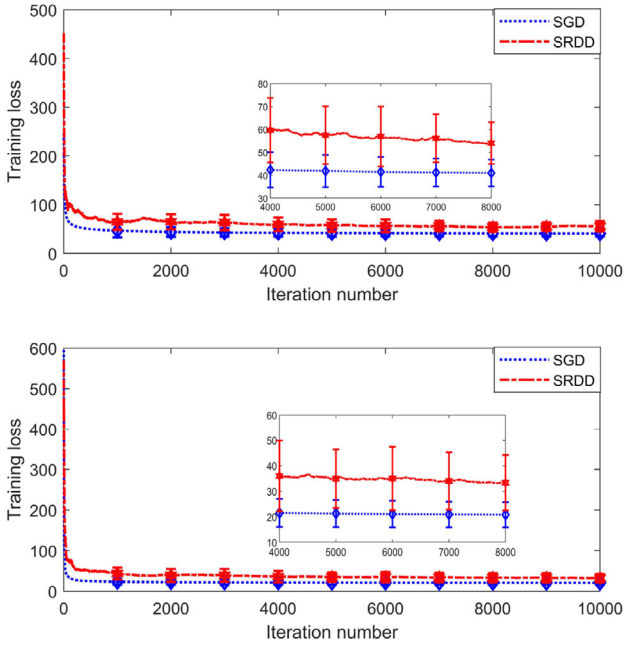

FIGURE 3 Comparison between SGD and SRDD for lasso regression. Topic subfigure: data set `space_ga`. Lower subfigure: data set `mg`

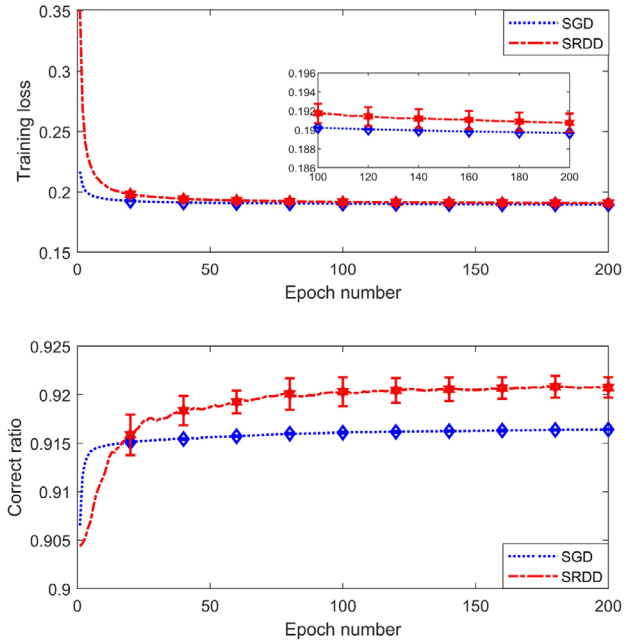

FIGURE 4 Comparison between SGD and SRDD on dataset `ijcnn1` with batch size equal to 64 for both SGD and SRDD for logistic regression. Topic subfigure: training loss. Lower subfigure: ratio of correct classification

$$\leqslant \mathcal{O}\left(\frac{1}{T}\sum_{t=1}^{T}\frac{1}{t}\right)+\mathcal{O}\left(\frac{1}{T}\sum_{t=1}^{T}\frac{1}{t^{\frac{1}{3}}}\right)+\mathcal{O}\left(\frac{1}{T}\sum_{t=1}^{T}\frac{1}{t^{\frac{1}{3}}}\cdot\frac{1}{t^{\frac{1}{6}}}\right)+\mathcal{O}\left(\frac{1}{T}\sum_{t=1}^{T}\frac{1}{t^{\frac{1}{3}}}\cdot\frac{1}{t^{\frac{(1-\epsilon)\mu}{2}}}\right)$$

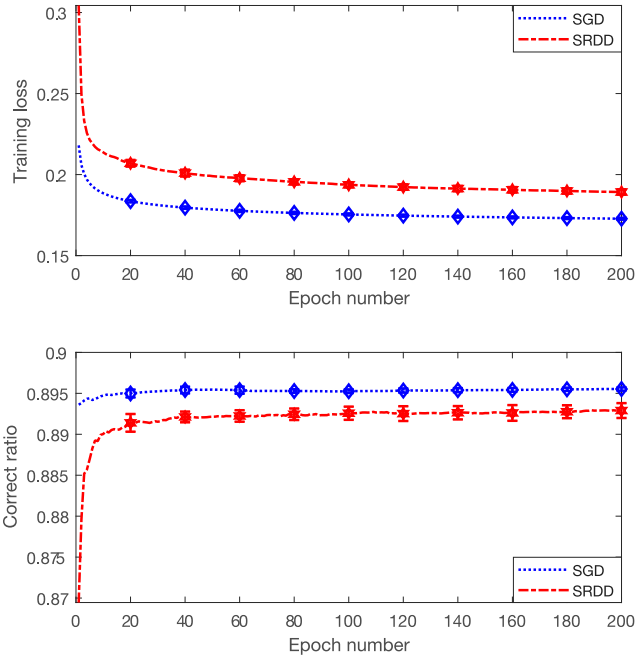

FIGURE 5 Comparison between SGD and SRDD on data set `w1a` with batch size equal to 64 for both SGD and SRDD for logistic regression. Topic subfigure: training loss. Lower subfigure: ratio of correct classification

$$+\mathcal{O}\left(\frac{1}{T}\sum_{t=1}^{T}\frac{1}{t^{(1-\epsilon)\mu}}\right)+\mathcal{O}\left(\frac{1}{T}\right)$$

$$=\mathcal{O}\left(\max\left\{\frac{1}{T^{\frac{1}{3}}},\frac{1}{T^{(1-\epsilon)\mu}}\right\}\right).$$

This finishes the proof. □

4 | NUMERICAL EXPERIMENT

We test the performance of the algorithm with the logistic regression problem

$$\min_{w}\frac{1}{N}\sum_{i=1}^{N}\log(1+e^{-y_i\langle x_i,w\rangle})+\frac{\lambda}{2}\|w\|_2^2,$$

the ridge regression problem

$$\min_{w}\frac{1}{2N}\sum_{i=1}^{N}(x_i^T w-y_i)^2+\frac{\lambda}{2}\|w\|_2^2,$$

and the lasso regression problem

$$\min_{w}\frac{1}{2}\sum_{i=1}^{N}\left(y_i-w_0-\sum\nolimits_{j=1}^{p}x_{ij}w_j\right)^2+\lambda\sum_{j=1}^{p}|w_j|,$$

with $\lambda = 1/N$. The data sets are available from LIBSVM [32].

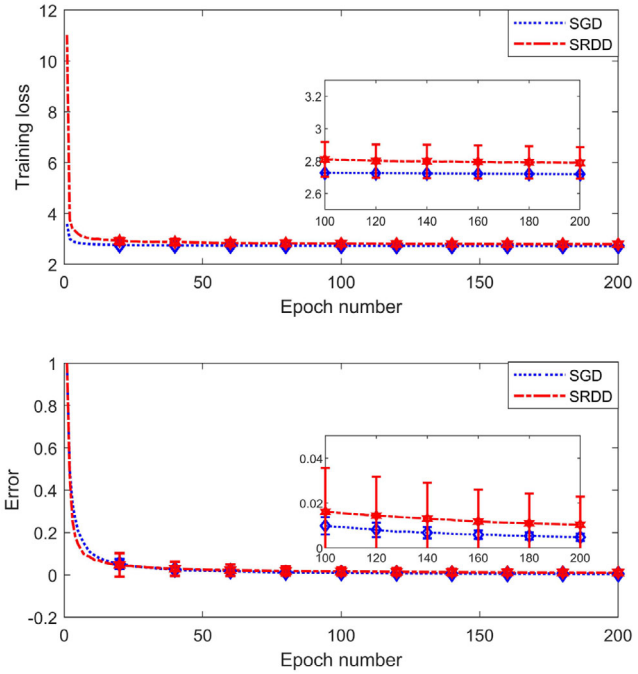

FIGURE 6 Comparison between SGD and SRDD on data set `abalone` with batch size equal to 8 for SGD and 16 for SRDD for ridge regression. Topic subfigure: training loss. Lower subfigure: relative estimation error

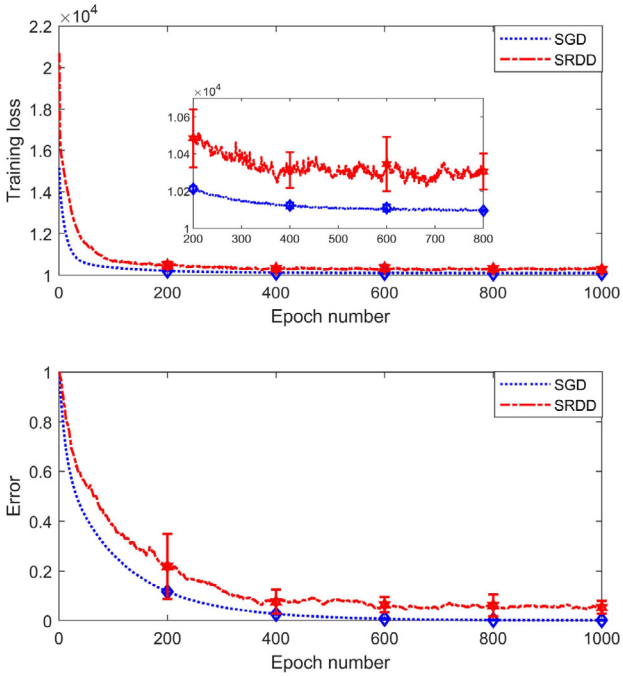

FIGURE 8 Comparison between SGD and SRDD on data set `abalone` with batch size equal to 64 for SGD and SRDD for lasso regression. Topic subfigure: training loss. Lower subfigure: relative estimation error

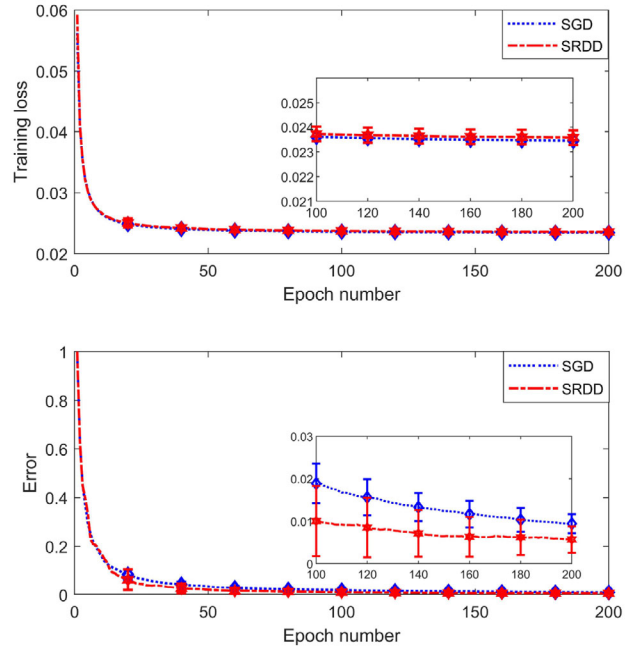

FIGURE 7 Comparison between SGD and SRDD on data set `space_ga` with batch size equal to 4 for both SGD and SRDD for ridge regression. Topic subfigure: training loss. Lower subfigure: relative estimation error

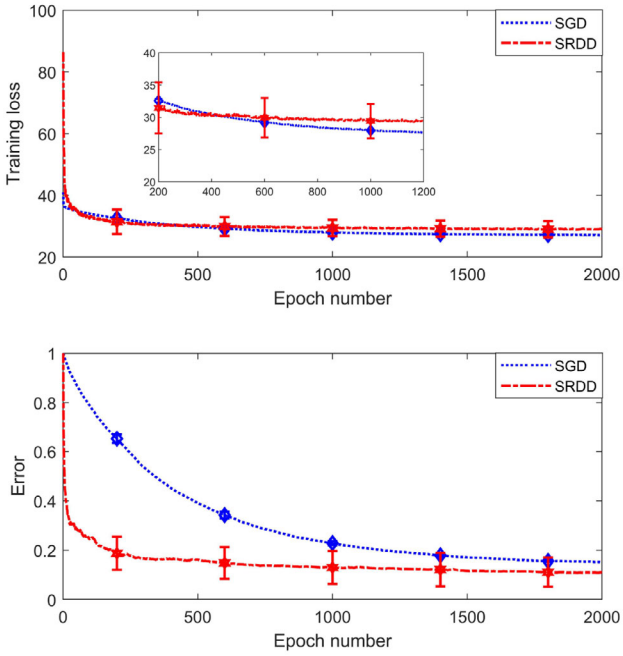

FIGURE 9 Comparison between SGD and SRDD on data set `space_ga` with batch size equal to 4 for SGD and 8 for SRDD for lasso regression. Topic subfigure: training loss. Lower subfigure: relative estimation error

To be specific, we compare the performance of SRDD and SGD through the logistic regression problem with data set `ijcnn1` with N=49990 and data set `covtype.binary` with N=581012, the ridge regression problem with data set `abalone` with N=4177 and data set `mg` with N=1385, and the lasso regression problem with data set `space_ga` with N=3107 and data set `mg` with N=1385. The simulation results are plotted in Figures 1– 3. To further testify the performance of the proposed algorithm, we apply the mini-batch technique to SRDD and SGD and the simulation results are given in Figures 4–9. We have also applied SRDD to the training of neu-

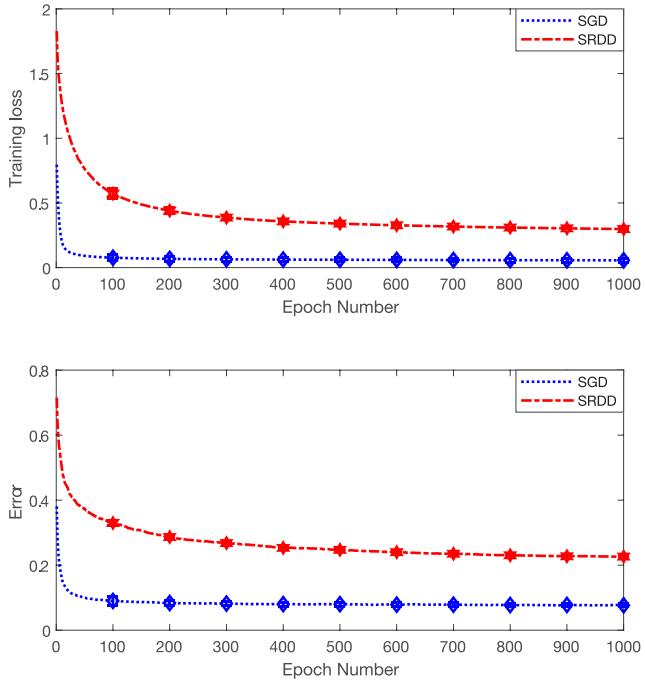

FIGURE 10 Comparison between SGD and SRDD on data set `MNIST` with constant step size. Topic subfigure: training loss. Lower subfigure: test error

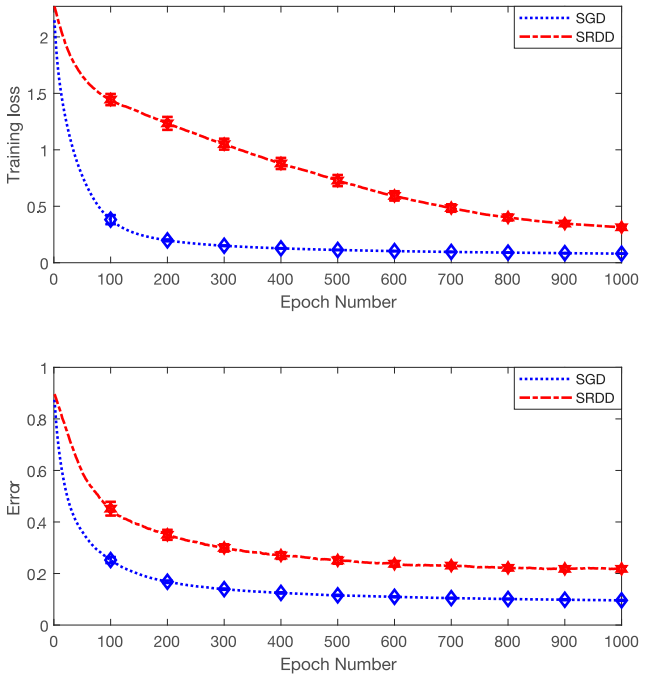

FIGURE 11 Comparison between SGD and SRDD on data set `MNIST` with decreasing step size. Topic subfigure: training loss. Lower subfigure: test error

ral networks and the simulation results are shown in Figures 10 and 11. The simulation details are given below.

Set $\eta_t = 1/t$ for both SGD and SRDD unless otherwise stated and $\beta_t = 1/t^{\frac{1}{3}}$. Choose perturbation vectors $\{\Delta_t\}_{t\geqslant 1}$ to be i.i.d. random variables uniformly distributed over $[-5,-1]\cup[1,5]$ and observation noises of function values $\{\varepsilon_t\}_{t\geqslant 1}$ to be Gaussian i.i.d. $\mathcal{N}(0,1)$. We perform SGD and SRDD using samples uniformly selected from training sets and calculate the values of objective functions at each round t. For each method, we run ten experiments and then obtain the mean and standard deviation of the objective functions. The results are shown in Figure 1 for the logistic regression, in Figure 2 for the ridge regression and in Figure 3 for the lasso regression, where in these figures the horizontal axis denotes the number of iterations and the vertical axis denotes the values of objective functions. We note that the step size is selected as $\eta_t = \frac{1}{t+10}$ in the case of ridge regression.

We further compare the performance of SGD and SRDD for the logistic regression, the ridge regression and the lasso regression using minibatches with data sets `ijcnn1`, `w1a`, `abalone`, and `space_ga` from LIBSVM. For the logistic regression, we calculate the values of objective functions and the ratio of correct classification. While for the ridge regression and the lasso regression, we calculate the values of objective functions and the relative estimation error $\|w_t - w_*\|_2^2/\|w_0 - w_*\|_2^2$. We also perform ten experiments and obtain the mean and standard deviation. The results are shown in Figures 4–9.

We also test the performance of SRDD through neural networks with one hidden layer of 100 nodes. The numbers of neurons in the input layer and the output layer are 784 and 10, respectively. We employ the hyperbolic tangent function and the sigmoid function as activation functions between the input and the hidden layers and between the hidden and the output layers, respectively. The perturbation signals $\{\Delta_t\}_{t\geqslant 1}$ are chosen as i.i.d. variables uniformly distributed over $[-0.7,-0.2]\cup[0.2,0.7]$ and the observation noises of the function values obey a uniform distribution over $[-0.5,0.5]$. With the data set `MNIST`, two experiments with different step sizes are conducted and the results are shown in Figures 10 and 11, in which, the topic subfigures show the trajectories of the training loss and the lower subfigures denote the test errors. Figure 10 shows the experiment results with a constant step size while Figure 11 with a decreasing step size, for both SGD and SRDD.

From the above experiments, we find that, although the performance of SRDD is not always better than SGD, it is rather comparable with that of SGD. In fact, the randomised differences in Equation (5) can be formulated as 'difference=gradient−[gradient−difference]". The decreasing learning rate in this paper is able to eliminate the effect of the additional term "[gradient−difference]". This might be the reason why the performance of SRDD is comparable with SGD.

5 | CONCLUSION

In this paper, we propose a SRDD algorithm for stochastic optimisation. We study the convergence of SRDD for strongly convex functions with probability one and establish the convergence rate. We also compare SRDD with SGD through

experiments. SRDD is suitable for the cases when the unbiased estimates for gradients/subgradients are unavailable. For future research, it is of interest to relax the assumptions for convergence of SRDD and to improve the convergence rate of the regret function.

ACKNOWLEDGMENT

The research of Xiaoxue Geng and Wenxiao Zhao was supported by National Key Research and Development Program of China (2018YFA0703800), the National Nature Science Foundation of China under Grant with No. 61822312 and the Strategic Priority Research Program of Chinese Academy of Sciences under Grant with No. XDA27000000. The research of Gao Huang was supported by the National Nature Science Foundation of China under Grant with No. 62022048.

ORCID

Wenxiao Zhao https://orcid.org/0000-0002-0371-5664

REFERENCES

1. Robbins, H., Monro, S.: A stochastic approximation method. Ann. Math. Stat. 22(3), 400–407 (1951)
2. Sakrison, D.J.: Stochastic approximation: A recursive method for solving regression problems. Adv. Commun. Syst. 2, 51–106 (1966)
3. Fabian, V.: Asymptotically efficient stochastic approximation: The RM case. Ann. Stat. 1(3), 486–495 (1973)
4. Polyak, B.T., Juditsky, A.B.: Acceleration of stochastic approximation by averaging. SIAM J. Control Optim. 30(4), 838–855 (1992)
5. Kushner, H., Yin, G.G.: Stochastic Approximation and Recursive Algorithms and Applications. Second Edition. Springer-Verlag, New York (2003)
6. Chen, H.F., Zhao, W.: Recursive Identification and Parameter Estimation. CRC Press, Singapore (2014)
7. Hazan, E., Agarwal, A., Kale, S.: Logarithmic regret algorithms for online convex optimization. Mach. Learn. 69(2–3), 169–192 (2007)
8. Hazan, E., Kale, S.: Beyond the regret minimization barrier: An optimal algorithm for stochastic strongly-convex optimization. In: Proceedings of the 24th Annual Conference on Learning Theory, pp. 421–436. MIT Press, Cambridge, MA (2011)
9. Rakhlin, A., Shamir, O., Sridharan, K.: Making gradient descent optimal for strongly convex stochastic optimization. arXiv:1109.5647 (2011)
10. Nguyen, L.M., et al.: SGD and Hogwild! convergence without the bounded gradients assumption. arXiv:1802.03801 (2018)
11. Qian, N.: On the momentum term in gradient descent learning algorithms. Neur. Netw. 12(1), 145–151 (1999)
12. Nesterov, Y.: A method for unconstrained convex minimization problem with the rate of convergence $\mathcal{O}(1/k^2)$. Soviet Math. Dokl. (in English) 27, 372–376 (1983); Dokl. Akad. Nauk SSSR (in Russian) 269, 543–547 (1983)
13. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. J. Mach. Learn. Res. 12(7), 2121–2159 (2011)
14. Zeiler, M.D.: Adadelta: An adaptive learning rate method. arXiv:1212.5701 (2012)
15. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv:1412.6980 (2014)
16. Gower, R.M., et al.: SGD: General analysis and improved rates. arXiv:1901.09401 (2019)
17. Spall, J.C.: Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control. Wiley, Hoboken (2005)
18. Conn, A.R., Scheinberg, K., Vicente, L.N.: Introduction to Derivative-Free Optimization. Society for Industrial and Applied Mathematics, Philadelphia, PA (2009)
19. Nesterov, Y., Spokoiny, V.: Random gradient-free minimization of convex functions. Found. Comput. Math. 17(2), 527–566 (2017)
20. Hazan, E., Levy, K.Y.: Bandit convex optimization: towards tight bounds. In: Proceedings of the 28th Conference on Neural Information Processing Systems, pp. 784–792. Curran Associates, Red Hook, NY (2014)
21. Shamir, O.: An optimal algorithm for bandit and zero-order convex optimization with two-point feedback. J. Mach. Learn. Res. 18(1), 1–11 (2017)
22. Flaxman, A.D., Kalai, A.T., McMahan, H.B.: Online convex optimization in the bandit setting: gradient descent without a gradient. arXiv: cs/0408007 (2004)
23. Tan, C., et al.: Gittins index based control policy for a class of pursuit-evasion problems. IET Control Theory Appl. 12(1), 110–118 (2017)
24. Chen, G., Zeng, X., Hong, Y.: Distributed optimization design for solving the stein equation with constraints. IET Control Theory Appl. 13(15), 2492–2499 (2019)
25. Papernot, N., et al.: Practical black-box attacks against machine learning. In: Proc. of the ACM on Asia Conference on Computer and Communications Security, pp. 506–519. ACM Press, New York (2017)
26. Audet, C., Orban, D.: Finding optimal algorithmic parameters using derivative-free optimization. SIAM J. Optim. 17(3), 642–664 (2006)
27. Kiefer, J., Wolfowitz, J.: Stochastic estimation of the maximum of a regression function. Ann. Math. Stat. 23(3), 462–466 (1952)
28. Spall, J.C.: Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. IEEE Trans. Autom. Control 37(3), 332–341 (1992)
29. Chen, H.F., Duncan, T.E., Pasik-Duncan, B.: A Kiefer-Wolfowitz algorithm with randomized differences. IEEE Trans. Autom. Control 44(3), 442–453 (1999)
30. Duchi, J.C., et al.: Optimal rates for zero-order convex optimization: The power of two function evaluations. IEEE Trans. Inf. Theory 61(5), 2788–2806 (2015)
31. Polyak, B.T.: Introduction to Optimization. Optimization Software Inc., New York (1987)
32. Chang, C.-C., Lin, C.-J.: LIBSVM : A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2(27), 1–27 (2011)

How to cite this article: Geng, X., Huang, G., Zhao, W. Almost sure convergence of randomised-difference descent algorithm for stochastic convex optimisation. IET Control Theory Appl. 2021;15:2183–2194. https://doi.org/10.1049/cth2.12184

APPENDIX

In this section we introduce a technical result, which will be used for proving the convergence of SRDD algorithm.

Theorem A.1 ([31]). *Denote $\{\mathcal{F}_k\}_{k\geqslant 0}$ a sequence of nondecreasing sub-σ-algebras in the basic probability space. Let $\{W_k\}_{k\geqslant 0}$, $\{U_k\}_{k\geqslant 0}$ and $\{V_k\}_{k\geqslant 0}$ be nonnegative random sequences and W_k, U_k, and V_k be $\mathcal{F}_k$-measurable for each $k \geqslant 0$. Suppose that $\sum_{k=0}^{\infty}\gamma_k < \infty$,*

$$\mathbb{E}\left[W_{k+1}|\mathcal{F}_k\right] \leqslant (1+\gamma_k)W_k - U_k + V_k$$

holds for all $k \geqslant 0$, and $\sum_{k=0}^{\infty} V_k < \infty$ with probability one. Then we have

$$W_k \to W \geqslant 0 \quad \text{and} \quad \sum_{k=0}^{\infty} U_k < \infty$$

with probability one.