Hindawi
Mathematical Problems in Engineering
Volume 2019, Article ID 8095054, 10 pages
https://doi.org/10.1155/2019/8095054

*Research Article*

# A Stochastic Trust Region Method for Unconstrained Optimization Problems

**Ningning Li,[1] Dan Xue,[1] Wenyu Sun,[2] and Jing Wang[3]**

[1]*School of Mathematics and Statistics, Qingdao University, Qingdao 266071, China*
[2]*School of Mathematical Science, Nanjing Normal University, Nanjing 210023, China*
[3]*Qingdao Haier Smart Technology R&D Co., Ltd., Qingdao 266103, China*

Correspondence should be addressed to Dan Xue; wtxuedan@126.com

Received 6 November 2018; Revised 7 January 2019; Accepted 9 January 2019; Published 3 February 2019

Academic Editor: Paolo Manfredi



In this paper, a stochastic trust region method is proposed to solve unconstrained minimization problems with stochastic objectives. Particularly, this method can be used to deal with nonconvex problems. At each iteration, we construct a quadratic model of the objective function. In the model, stochastic gradient is used to take the place of deterministic gradient for both the determination of descent directions and the approximation of the Hessians of the objective function. The behavior and the convergence properties of the proposed method are discussed under some reasonable conditions. Some preliminary numerical results show that our method is potentially efficient.

## 1. Introduction

The computational complexity of learning algorithm becomes the critical limiting factor when one envisions very large datasets. This contribution advocates many stochastic optimization algorithms for large-scale problems. To a certain extent, stochastic optimization is effective to solve the problems in stochastic systems. A number of industrial, biological, engineering, and economic problems can be considered as stochastic systems, for example, area of communication, gene, signal processing, geography, civil engineering, aerospace, and banking. Particularly, stochastic optimization algorithms are used to solve the problem of optimizing an objective function over a set of feasible values in situations where the objective function is defined as an expectation over a set of random functions. To be more precise, consider an optimization variable $x \in R^n$ and a random variable $\theta \in \Theta \subseteq R^p$ that determines the choice of a function $f(x,\theta) : R^{n\times p} \longrightarrow R$. Stochastic optimization problems considered in this paper entail determination of the argument $x^*$ that minimizes the expected value $F(x) := \mathrm{E}_\theta[f(x,\theta)]$:

$$x^* := \arg\min_x \mathrm{E}_\theta \left[ f(x,\theta) \right] := \arg\min F(x). \tag{1}$$

We refer to $f(x,\theta)$ as the random or instantaneous functions and to $F(x) := \mathrm{E}_\theta[f(x,\theta)]$ as the average function. We assume that the instantaneous function $f(x,\theta)$ is continuously differentiable for all $\theta$ from which it follows that the average function $F(x)$ is also continuously differentiable. Problems having the form in (1) are often used in machine learning [1–3] as well as in optimal resource allocation in wireless systems [4, 5].

Many descent algorithms can be used for solving such problems as (1). However, descent methods require exact gradient of the objective function, i.e., $\nabla_x F(x) = \mathrm{E}_\theta[\nabla_x f(x,\theta)]$, which is impracticable generally as a result of the very large dataset. Stochastic gradient descent (SGD) methods overcome this obstacle by using gradient approximation which is based on small data samples and are regarded as the workhorse methodology to solve large-scale stochastic optimization problems [3, 6–9]. But practical appeal of SGD methods remains limited, and the number of iterations required to approximate optimal arguments can be prohibitive in high dimensional problems. In fact, SGD inherits slow convergence from its use of gradients which is exasperated by their substitution with stochastic approximation; consecutive stochastic gradients may vary a lot or even point in opposite directions. Several accelerations of SGD have

been presented, and lots of recent works have focused on reducing randomness in SGD by combining gradient and stochastic gradient or updating descent direction so that they become more close to gradient gradually. For example, the Stochastic Average Gradient (SAG) method achieves a faster convergence rate than SGD by incorporating a memory of previous gradient values, which often outperform existing SGD methods and were shown to exhibit linear convergence [10]. The Semistochastic Gradient (S2GD) method in [11] runs for one or several epochs in each of which a single full gradient and a random number of stochastic gradients are computed, following a geometric law. These algorithms end up showing a faster asymptotic convergence rate than SGD [12]; however, practical numerical experiments show that reducing randomness is of no use for problems with a challenging curvature profile.

To overcome problems with the objective function's curvature, one may consider stochastic version of Newton's method. But computing approximation of Newton steps is difficult except in problems with specific structures because of the challenging curvature [13]. Recourse to quasi-Newton methods, regularized stochastic BFGS (RES) method, is proposed. RES utilizes stochastic gradients in lieu of deterministic gradients for determining descent directions and approximating the objective function's curvature. As we can see in [14], RES has been shown to outperform SGD in large-dimensional problems or ill-conditioned functions. However, RES inherits the disadvantage that the Hessian matrix of objective functions must be positive definite which implies that the strongly convex objective functions are indispensable to guarantee provable convergence. Thus, RES can only be used to solve strongly convex stochastic optimization problems. But there are lots of nonconvex optimization problems in practical applications. So, in this paper, we would like to exploit an effective method for solving more general stochastic optimization problems, i.e., the objective functions can be nonconvex.

The trust region methods are deemed to be invaluable tools for solving nonlinear and nonconvex optimization problems [15–17]. The main idea of typical trust region method is to construct a quadratic model and choose a step to be the approximate minimizer of the quadratic model in the trust region around the current point in which the model is trusted to be adequate to the objective function. In the iterative procedures, the trust region radii are adjusted depending on the agreement between the model functions and the objective functions.

In this paper, we develop a stochastic version of trust region method, where the Hessian matrices are approximated by regularized stochastic BFGS method [14]. Similar to classical trust region methods, we use a suitable quadratic model to replace the complex objective function. In the model, deterministic gradients are replaced by stochastic gradients and exact Hessian matrices of the objective function are approximated by its stochastic approximations. Due to the fact that stochastic gradients are computable at manageable cost, stochastic trust region method is feasible in practice. Convergence theory and numerical results verify the reliability of our method to deal with both convex and nonconvex optimization problems.

The rest of the paper is organized as follows. The new algorithm is illustrated in Section 2. In Section 3, convergence of the algorithm is established under suitable conditions. Some preliminary numerical results are reported in Section 4. Finally, some conclusions are given in Section 5.

## 2. Stochastic Trust Region Algorithm

Considering that the objective function $F(x)$ is continuously twice differentiable and further assuming that the instantaneous functions have finite gradients, it follows that the gradients of $F(x)$ are given by

$$g(x) \coloneqq \nabla F(x) = \mathrm{E}_\theta \left[\nabla f(x,\theta)\right]. \tag{2}$$

When the number of functions $f(x,\theta)$ is large, as is the case in most problems of practical interest, exact evaluation of the gradient $g(x)$ is impractical. This motivates the use of stochastic gradients in lieu of exact gradients. In addition, the Hessian matrices of $F(x)$ are given by

$$H(x) \coloneqq \nabla^2 F(x) = \mathrm{E}_\theta \left[\nabla^2 f(x,\theta)\right]. \tag{3}$$

More precisely, consider a given set of $L$ realizations $\widetilde{\theta} = [\theta_1; \cdots; \theta_L]$ and define the stochastic gradient of $F(x)$ at given samples $\widetilde{\theta}$ as

$$\widehat{g}\left(x,\widetilde{\theta}\right) \coloneqq \frac{1}{L}\sum_{l=1}^{L} \nabla f\left(x,\theta_l\right). \tag{4}$$

Consider the trust region subproblem

$$\begin{aligned} \min_s \quad & \widehat{q}^{(k)}(s) = F\left(x_k\right) + \widehat{g}\left(x_k,\widetilde{\theta}\right)^T s + \frac{1}{2} s^T \widehat{G}_k s \\ s.t. \quad & \|s\|_2 \le h_k. \end{aligned} \tag{5}$$

where $s = x - x_k$, $h_k \in (0,\overline{h})$ is the trust region radius, and $\widehat{G}_k$ is a stochastic approximation of Hessian matrix of $F(x)$ (see Remark 2 for details). The model $\widehat{q}^{(k)}(s)$ is minimized approximately to produce a step $s_k$ which is subject to having the norm less than or equal to $h_k$. In other words, the solution of this subproblem $s_k$ represents a step toward minimizing the model $\widehat{q}^{(k)}(s)$ of the objective function $F(x)$ at $x_k$. The radio $\widehat{r}_k$ of the actual reduction of the objective function $ared_k = F(x_k) - F(x_k + s_k)$ to the predicted reduction $pred_k = \widehat{q}^{(k)}(0) - \widehat{q}^{(k)}(s_k)$ is calculated. If $\widehat{r}_k \ge \eta_1$, where $\eta_1 \in (0,1)$, then $s_k$ is accepted, which is called a successful step. Let $x_{k+1} = x_k + s_k$; in this case, the trust region radius is increased or remains unchanged. Otherwise, $s_k$ is rejected, which is called an unsuccessful step in which $x_k$ remains unchanged and the trust region radius is reduced.

*Algorithm 1* (stochastic trust region). *Initialization*. The variable is $x_0$. The symmetric matrix is $\widehat{G}_0$. The initial radius is $h_0 \in (0,\overline{h})$, where $\overline{h} > 0$, and choose constant $0 < \gamma_1 < 1 < \gamma_2$, $0 < \eta_1 \le \eta_2 \le 1$.

*Step 1.* Acquire $L$ independent samples $\widetilde{\theta}_k = [\theta_{k1}; \cdots; \theta_{kL}]$, and calculate

$$\widehat{g}\left(x_k, \widetilde{\theta}_k\right) \coloneqq \frac{1}{L}\sum_{l=1}^{L} \nabla f\left(x_k, \theta_{kl}\right). \tag{6}$$

*Step 2.* Solve trust region subproblem (5), giving a trial step $s_k$.

*Step 3.* Compute

$$\widehat{r}_k = \frac{\text{ared}_k}{\text{pred}_k}, \tag{7}$$

where $\text{ared}_k = F(x_k) - F(x_k + s_k)$ and $\text{pred}_k = \widehat{q}^{(k)}(0) - \widehat{q}^{(k)}(s_k)$. If $\widehat{r}_k \geq \eta_1$, set $x_{k+1} = x_k + s_k$; otherwise, set $x_{k+1} = x_k$.

*Step 4.* Update the trust region radius.

If $\widehat{r}_k < \eta_1$, set $h_{k+1} = \gamma_1 h_k$;

If $\widehat{r}_k > \eta_2$ and $\|s_k\| = h_k$, set $h_{k+1} = \min\{\gamma_2 h_k, \overline{h}\}$;

Otherwise, $h_{k+1} = h_k$.

*Step 5.* Update Hessian approximation matrix to acquire $\widehat{G}_{k+1}$ when $\widehat{r}_k \geq \eta_1$; otherwise, keep the matrix unchanged. Set $k \coloneqq k + 1$; go to Step 2.

*Remark 2.* The matrix $\widehat{G}_{k+1}$ can be updated by regularized stochastic BFGS formula [14] as follows:

$$\widehat{G}_{k+1} = \widehat{G}_k + \frac{\widetilde{r}_k \widetilde{r}_k^T}{{v_k}^T \widetilde{r}_k} - \frac{\widehat{G}_k v_k {v_k}^T \widehat{G}_k}{{v_k}^T \widehat{G}_k v_k} + \delta I, \tag{8}$$

where $\delta > 0$ is a constant, $v_k = x_{k+1} - x_k$, $\widetilde{r}_k = \widehat{g}(x_{k+1}, \widetilde{\theta}_k) - \widehat{g}(x_k, \widetilde{\theta}_k) - \delta v_k$ denote the variable and corrected stochastic gradient variation at time $k$. The addition of the regularization term $\delta I$ and the corrected stochastic gradient variation $\widetilde{r}_k$ avoid the near-singularity problems of more straightforward extensions.

## 3. Convergence Analysis

In this section, we prove that the iterative sequence generated by Algorithm 1 is convergent. For the subsequent analysis, we define the instantaneous objective function associated with samples $\widetilde{\theta} = [\theta_1; \cdots; \theta_L]$ as

$$\widehat{f}\left(x, \widetilde{\theta}\right) \coloneqq \frac{1}{L}\sum_{l=1}^{L} f\left(x, \theta_l\right). \tag{9}$$

The definition of the instantaneous objective function $\widehat{f}(x, \widetilde{\theta})$ in association with the fact that $F(x) \coloneqq \mathrm{E}_\theta[f(x, \theta)]$ implies

$$F(x) = \mathrm{E}_\theta\left[\widehat{f}\left(x, \widetilde{\theta}\right)\right]. \tag{10}$$

In order to prove the global convergence, we make the following assumptions.

**A1** The instantaneous objective function $\widehat{f}(x, \widetilde{\theta})$ is continuously twice differentiable.

**A2** The level set $L = \{x | F(x) \leq F(x_0)\}$ is bounded. Moreover, the function $F(x)$ is bounded below in $L$.

**A3** Since the stochastic gradient $\widehat{g}(x_k, \widetilde{\theta}_k)$ is an unbiased estimator of $\nabla F(x_k)$ in the sense of $\mathrm{E}_\theta[\widehat{g}(x_k, \widetilde{\theta}_k) | x_k] = \nabla F(x_k)$, there exists a positive constant $N$ such that, for all $k$, it holds

$$\left\|\nabla F\left(x_k\right) - \widehat{g}\left(x_k, \widetilde{\theta}_k\right)\right\| \leq N. \tag{11}$$

**A4** There exists a $\rho > 0$, such that $\|\widehat{G}_k\| \leq \rho$ for all $k$.

As a consequence of A1, the function $F(x)$ is also continuously twice differentiable owing to the linearity of the expected operation and the expression in (10).

**Lemma 3.** *Assume that A1, A3, and A4 hold; if $s_k$ is a solution or approximate solution of subproblem (5), then we have*

$$\begin{aligned} &\mathrm{E}_\theta\left[\widehat{q}^{(k)}(0) - \widehat{q}^{(k)}\left(s_k\right) \mid x_k\right] \\ &\geq C_1 \mathrm{E}_\theta\left[\left\|\widehat{g}\left(x_k, \widetilde{\theta}_k\right)\right\| \min\left\{h_k, \frac{\left\|\widehat{g}\left(x_k, \widetilde{\theta}_k\right)\right\|}{\rho}\right\} \mid x_k\right], \end{aligned} \tag{12}$$

*where $C_1 > 0$ is a constant and $\mathrm{E}_\theta[\widehat{q}^{(k)}(0) - \widehat{q}^{(k)}(s_k) | x_k]$ denotes the conditional expectation of $\text{pred}_k = \widehat{q}^{(k)}(0) - \widehat{q}^{(k)}(s_k)$ with giving $x_k$.*

*Proof.* Considering subproblem (5), the Cauchy point can be expressed by

$$s_k{}^c = -\Gamma_k h_k \frac{\widehat{g}\left(x_k, \widetilde{\theta}_k\right)}{\left\|\widehat{g}\left(x_k, \widetilde{\theta}_k\right)\right\|}, \tag{13}$$

where

$$\Gamma_k = \begin{cases} 1, & \text{else.} \\ \min\left\{\dfrac{\left\|\widehat{g}\left(x_k, \widetilde{\theta}_k\right)\right\|^3}{h_k \widehat{g}\left(x_k, \widetilde{\theta}_k\right)^T \widehat{G}_k \widehat{g}\left(x_k, \widetilde{\theta}_k\right)}, 1\right\}, & \text{if } \widehat{g}\left(x_k, \widetilde{\theta}_k\right)^T \widehat{G}_k \widehat{g}\left(x_k, \widetilde{\theta}_k\right) > 0. \end{cases} \tag{14}$$

*Case 1.* When $\widehat{g}(x_k,\widetilde{\theta}_k)^T\widehat{G}_k\widehat{g}(x_k,\widetilde{\theta}_k)\le 0$, it follows from (14) that $\Gamma_k=1$; thus, we have

$$\begin{aligned}
&\mathrm{E}_\theta\left[\widehat{q}^{(k)}(0)-\widehat{q}^{(k)}\left(s_k{}^c\right)\mid x_k\right]=\widehat{q}^{(k)}(0)\\
&\quad-\mathrm{E}_\theta\left[\widehat{q}^{(k)}\left(-\frac{h_k}{\left\|\widehat{g}\left(x_k,\widetilde{\theta}_k\right)\right\|}\widehat{g}\left(x_k,\widetilde{\theta}_k\right)\right)\mid x_k\right]\\
&\quad=F(x_k)-F(x_k)+\mathrm{E}_\theta\left[h_k\left\|\widehat{g}\left(x_k,\widetilde{\theta}_k\right)\right\|\right.\\
&\quad\left.-\frac{h_k{}^2}{2\left\|\widehat{g}\left(x_k,\widetilde{\theta}_k\right)\right\|^2}\widehat{g}\left(x_k,\widetilde{\theta}_k\right)^T\widehat{G}_k\widehat{g}\left(x_k,\widetilde{\theta}_k\right)\mid x_k\right]\\
&\quad=\mathrm{E}_\theta\left[h_k\left\|\widehat{g}\left(x_k,\widetilde{\theta}_k\right)\right\|\mid x_k\right]-\mathrm{E}_\theta\left[\frac{h_k{}^2}{2\left\|\widehat{g}\left(x_k,\widetilde{\theta}_k\right)\right\|^2}\right.\\
&\quad\left.\cdot\,\widehat{g}\left(x_k,\widetilde{\theta}_k\right)^T\widehat{G}_k\widehat{g}\left(x_k,\widetilde{\theta}_k\right)\mid x_k\right]\\
&\quad\ge\mathrm{E}_\theta\left[h_k\left\|\widehat{g}\left(x_k,\widetilde{\theta}_k\right)\right\|\mid x_k\right]\ge\mathrm{E}_\theta\left[\left\|\widehat{g}\left(x_k,\widetilde{\theta}_k\right)\right\|\right.\\
&\quad\left.\cdot\min\left\{h_k,\frac{\left\|\widehat{g}\left(x_k,\widetilde{\theta}_k\right)\right\|}{\rho}\right\}\mid x_k\right].
\end{aligned}\tag{15}$$

*Case 2.* When $\widehat{g}(x_k,\widetilde{\theta}_k)^T\widehat{G}_k\widehat{g}(x_k,\widetilde{\theta}_k)>0$ and $\|\widehat{g}(x_k,\widetilde{\theta}_k)\|^3/h_k\widehat{g}(x_k,\widetilde{\theta}_k)^T\widehat{G}_k\widehat{g}(x_k,\widetilde{\theta}_k)\le 1$, it follows from (14) that $\Gamma_k=\|\widehat{g}(x_k,\widetilde{\theta}_k)\|^3/h_k\widehat{g}(x_k,\widetilde{\theta}_k)^T\widehat{G}_k\widehat{g}(x_k,\widetilde{\theta}_k)$; then, we have

$$\begin{aligned}
&\mathrm{E}_\theta\left[\widehat{q}^{(k)}(0)-\widehat{q}^{(k)}\left(s_k{}^c\right)\mid x_k\right]=\widehat{q}^{(k)}(0)\\
&\quad-\mathrm{E}_\theta\left[\widehat{q}^{(k)}\left(-\frac{\left\|\widehat{g}\left(x_k,\widetilde{\theta}_k\right)\right\|^3}{h_k\widehat{g}\left(x_k,\widetilde{\theta}_k\right)^T\widehat{G}_k\widehat{g}\left(x_k,\widetilde{\theta}_k\right)}\right.\right.\\
&\quad\left.\left.\cdot\frac{h_k}{\left\|\widehat{g}\left(x_k,\widetilde{\theta}_k\right)\right\|}\right)\mid x_k\right]\\
&\quad=\mathrm{E}_\theta\left[\frac{\left\|\widehat{g}\left(x_k,\widetilde{\theta}_k\right)\right\|^4}{\widehat{g}\left(x_k,\widetilde{\theta}_k\right)^T\widehat{G}_k\widehat{g}\left(x_k,\widetilde{\theta}_k\right)}\right.\\
&\quad-\frac{\widehat{g}\left(x_k,\widetilde{\theta}_k\right)^T\widehat{G}_k\widehat{g}\left(x_k,\widetilde{\theta}_k\right)}{2}\\
&\quad\left.\cdot\frac{\left\|\widehat{g}\left(x_k,\widetilde{\theta}_k\right)\right\|^4}{\left[\widehat{g}\left(x_k,\widetilde{\theta}_k\right)^T\widehat{G}_k\widehat{g}\left(x_k,\widetilde{\theta}_k\right)\right]^2}\mid x_k\right]\\
&\quad=\mathrm{E}_\theta\left[\frac{\left\|\widehat{g}\left(x_k,\widetilde{\theta}_k\right)\right\|^4}{2\widehat{g}\left(x_k,\widetilde{\theta}_k\right)^T\widehat{G}_k\widehat{g}\left(x_k,\widetilde{\theta}_k\right)}\mid x_k\right]\\
&\quad\ge\mathrm{E}_\theta\left[\frac{\left\|\widehat{g}\left(x_k,\widetilde{\theta}_k\right)\right\|^4}{2\left\|\widehat{g}\left(x_k,\widetilde{\theta}_k\right)\right\|\left\|\widehat{G}_k\right\|\left\|\widehat{g}\left(x_k,\widetilde{\theta}_k\right)\right\|}\mid x_k\right]\\
&\quad=\mathrm{E}_\theta\left[\frac{\left\|\widehat{g}\left(x_k,\widetilde{\theta}_k\right)\right\|^2}{2\left\|\widehat{G}_k\right\|}\mid x_k\right]\ge\frac{1}{2}\mathrm{E}_\theta\left[\left\|\widehat{g}\left(x_k,\widetilde{\theta}_k\right)\right\|\right.\\
&\quad\left.\cdot\min\left\{h_k,\frac{\left\|\widehat{g}\left(x_k,\widetilde{\theta}_k\right)\right\|}{\left\|\widehat{G}_k\right\|}\right\}\mid x_k\right]\ge\frac{1}{2}\\
&\quad\cdot\mathrm{E}_\theta\left[\left\|\widehat{g}\left(x_k,\widetilde{\theta}_k\right)\right\|\min\left\{h_k,\frac{\left\|\widehat{g}\left(x_k,\widetilde{\theta}_k\right)\right\|}{\rho}\right\}\mid x_k\right].
\end{aligned}\tag{16}$$

*Case 3.* When $\widehat{g}(x_k,\widetilde{\theta}_k)^T\widehat{G}_k\widehat{g}(x_k,\widetilde{\theta}_k)>0$ and $\|\widehat{g}(x_k,\widetilde{\theta}_k)\|^3/h_k\widehat{g}(x_k,\widetilde{\theta}_k)^T\widehat{G}_k\widehat{g}(x_k,\widetilde{\theta}_k)>1$, it follows from (14) that $\Gamma_k=1$, and we have

$$\begin{aligned}
&\mathrm{E}_\theta\left[\widehat{q}^{(k)}(0)-\widehat{q}^{(k)}\left(s_k{}^c\right)\mid x_k\right]=\widehat{q}^{(k)}(0)\\
&\quad-\mathrm{E}_\theta\left[\widehat{q}^{(k)}\left(-\frac{h_k}{\left\|\widehat{g}\left(x_k,\widetilde{\theta}_k\right)\right\|}\widehat{g}\left(x_k,\widetilde{\theta}_k\right)\right)\mid x_k\right]\\
&\quad=\mathrm{E}_\theta\left[h_k\left\|\widehat{g}\left(x_k,\widetilde{\theta}_k\right)\right\|\right.\\
&\quad\left.-\frac{h_k{}^2}{2\left\|\widehat{g}\left(x_k,\widetilde{\theta}_k\right)\right\|^2}\widehat{g}\left(x_k,\widetilde{\theta}_k\right)^T\widehat{G}_k\widehat{g}\left(x_k,\widetilde{\theta}_k\right)\mid x_k\right]\\
&\quad\ge\mathrm{E}_\theta\left[h_k\left\|\widehat{g}\left(x_k,\widetilde{\theta}_k\right)\right\|\right.\\
&\quad\left.-\frac{h_k{}^2}{2\left\|\widehat{g}\left(x_k,\widetilde{\theta}_k\right)\right\|^2}\frac{\left\|\widehat{g}\left(x_k,\widetilde{\theta}_k\right)\right\|^3}{h_k}\mid x_k\right]=\frac{1}{2}\\
&\quad\cdot\mathrm{E}_\theta\left[h_k\left\|\widehat{g}\left(x_k,\widetilde{\theta}_k\right)\right\|\mid x_k\right]\ge\frac{1}{2}\\
&\quad\cdot\mathrm{E}_\theta\left[\left\|\widehat{g}\left(x_k,\widetilde{\theta}_k\right)\right\|\min\left\{h_k,\frac{\left\|\widehat{g}\left(x_k,\widetilde{\theta}_k\right)\right\|}{\rho}\right\}\mid x_k\right].
\end{aligned}\tag{17}$$

With the observation of these cases above, we can come to the conclusion that, for all, the Cauchy point $s_k{}^c$ of (5) satisfies

$$\begin{aligned}
&\mathrm{E}_\theta\left[\widehat{q}^{(k)}(0)-\widehat{q}^{(k)}\left(s_k{}^c\right)\mid x_k\right]\\
&\quad=\widehat{q}^{(k)}(0)-\mathrm{E}_\theta\left[\widehat{q}^{(k)}\left(s_k{}^c\right)\mid x_k\right]\\
&\quad\ge\frac{1}{2}\mathrm{E}_\theta\left[\left\|\widehat{g}\left(x_k,\widetilde{\theta}_k\right)\right\|\min\left\{h_k,\frac{\left\|\widehat{g}\left(x_k,\widetilde{\theta}_k\right)\right\|}{\rho}\right\}\mid x_k\right].
\end{aligned}\tag{18}$$

Assume that ${s_k}^e$ is the exact solution of subproblem (5); we can draw the following inequalities from the property of Cauchy point that

$$\mathrm{E}_\theta\left[\widehat{q}^{(k)}\left({s_k}^e\right) \mid x_k\right] \le \mathrm{E}_\theta\left[\widehat{q}^{(k)}\left({s_k}^c\right) \mid x_k\right]. \tag{19}$$

In combination with inequality (18) above, we can write

$$\begin{aligned}
&\mathrm{E}_\theta\left[\widehat{q}^{(k)}(0) - \widehat{q}^{(k)}\left({s_k}^e\right) \mid x_k\right] \\
&= \widehat{q}^{(k)}(0) - \mathrm{E}_\theta\left[\widehat{q}^{(k)}\left({s_k}^e\right) \mid x_k\right] \\
&\ge \widehat{q}^{(k)}(0) - \mathrm{E}_\theta\left[\widehat{q}^{(k)}\left({s_k}^c\right) \mid x_k\right] \\
&\ge \frac{1}{2}\mathrm{E}_\theta\left[\left\|\widehat{g}\left(x_k, \widetilde{\theta}_k\right)\right\| \min\left\{h_k, \frac{\left\|\widehat{g}\left(x_k, \widetilde{\theta}_k\right)\right\|}{\rho}\right\} \mid x_k\right].
\end{aligned} \tag{20}$$

Further assume that $s_k$ is the approximate solution; then, there exists a constant $C_2 \in (0, 1]$ satisfying

$$\begin{aligned}
&\mathrm{E}_\theta\left[\widehat{q}^{(k)}(0) - \widehat{q}^{(k)}(s_k) \mid x_k\right] \\
&\quad \ge C_2 \mathrm{E}_\theta\left[\widehat{q}^{(k)}(0) - \widehat{q}^{(k)}\left({s_k}^e\right) \mid x_k\right].
\end{aligned} \tag{21}$$

If we set $C_1 = (1/2)C_2$, then we have

$$\begin{aligned}
&\mathrm{E}_\theta\left[\widehat{q}^{(k)}(0) - \widehat{q}^{(k)}(s_k) \mid x_k\right] \ge C_2 \mathrm{E}_\theta\left[\widehat{q}^{(k)}(0)\right. \\
&\left. - \widehat{q}^{(k)}\left({s_k}^c\right) \mid x_k\right] \ge C_1 \mathrm{E}_\theta\left[\left\|\widehat{g}\left(x_k, \widetilde{\theta}_k\right)\right\|\right. \\
&\left. \cdot \min\left\{h_k, \frac{\left\|\widehat{g}\left(x_k, \widetilde{\theta}_k\right)\right\|}{\rho}\right\} \mid x_k\right].
\end{aligned} \tag{22}$$

Lemma 3 implies that a sufficient decrement of the model $\widehat{q}^{(k)}(s)$ on average is guaranteed under assumptions A1 and A4. □

**Lemma 4.** *If A1, A3, and A4 hold true, we have*

$$\begin{aligned}
&\mathrm{E}_\theta\left[\left|F\left(x_k + s_k\right) - \widehat{q}^{(k)}(s_k)\right| \mid x_k\right] \\
&\quad \le N\mathrm{E}_\theta\left[\|s_k\| \mid x_k\right] + \frac{1}{2}\rho\mathrm{E}_\theta\left[\|s_k\|^2 \mid x_k\right] \\
&\qquad + \mathrm{E}_\theta\left[c\left(\|s_k\|\right)\|s_k\| \mid x_k\right],
\end{aligned} \tag{23}$$

*where $c(\|s_k\|)$ is a function of $s_k$ and decreases with the decrease of $s_k$.*

*Proof.* From Taylor theorem, we have

$$\begin{aligned}
F\left(x_k + s_k\right) &= F\left(x_k\right) + \nabla F\left(x_k\right)^T s_k \\
&\quad + \int_0^1 \left[\nabla F\left(x_k + ts_k\right) - \nabla F\left(x_k\right)\right]^T s_k dt.
\end{aligned} \tag{24}$$

Then, observing (24) and the model $\widehat{q}^{(k)}(s)$, we have

$$\begin{aligned}
&\left|F\left(x_k + s_k\right) - \widehat{q}^{(k)}\left(s_k\right)\right| = \left|\nabla F\left(x_k\right)^T s_k\right. \\
&\quad - \widehat{g}\left(x_k, \widetilde{\theta}_k\right)^T s_k - \frac{1}{2}{s_k}^T \widehat{G}_k s_k \\
&\quad \left. + \int_0^1 \left[\nabla F\left(x_k + ts_k\right) - \nabla F\left(x_k\right)\right]^T s_k dt\right| \le \left\|\nabla F\left(x_k\right)\right. \\
&\quad \left. - \widehat{g}\left(x_k, \widetilde{\theta}_k\right)\right\|\|s_k\| + \frac{1}{2}\rho\|s_k\|^2 \\
&\quad + \left|\int_0^1 \left[\nabla F\left(x_k + ts_k\right) - \nabla F\left(x_k\right)\right]^T s_k dt\right|.
\end{aligned} \tag{25}$$

Taking expectation with $x_k$ given in both sides of (25) and applying (11), we can write

$$\begin{aligned}
&\mathrm{E}_\theta\left[\left|F\left(x_k + s_k\right) - \widehat{q}^{(k)}\left(s_k\right)\right| \mid x_k\right] \\
&\quad \le \mathrm{E}_\theta\left[\left\|\nabla F\left(x_k\right) - \widehat{g}\left(x_k, \widetilde{\theta}_k\right)\right\|\|s_k\| + \frac{1}{2}\rho\|s_k\|^2\right. \\
&\quad \left. + \left|\int_0^1 \left[\nabla F\left(x_k + ts_k\right) - \nabla F\left(x_k\right)\right]^T s_k dt\right| \mid x_k\right] \\
&\quad \le N\mathrm{E}_\theta\left[\|s_k\| \mid x_k\right] + \frac{1}{2}\rho\mathrm{E}_\theta\left[\|s_k\|^2 \mid x_k\right] \\
&\quad + \mathrm{E}_\theta\left[c\left(\|s_k\|\right)\|s_k\| \mid x_k\right].
\end{aligned} \tag{26}$$

□

**Lemma 5.** *If the same assumptions as in Lemma 3 hold, $\widehat{g}(x_k, \widetilde{\theta}_k) \ne 0$ and $\widetilde{\Delta}$ is a small tolerance. Then, there exist infinite number of k which satisfy $h_{k+1} \ge h_k$.*

*Proof.* From the definition of $\hat{r}_k = (F(x_k) - F(x_k + s_k))/(\widehat{q}^{(k)}(0) - \widehat{q}^{(k)}(s_k))$ and properties of expectation, we can define

$$\mathrm{E}_\theta\left[\hat{r}_k \mid x_k\right] = \frac{\mathrm{E}_\theta\left[F\left(x_k\right) - F\left(x_k + s_k\right) \mid x_k\right]}{\mathrm{E}_\theta\left[\widehat{q}^{(k)}(0) - \widehat{q}^{(k)}\left(s_k\right) \mid x_k\right]}. \tag{27}$$

Further assume $\|\widehat{g}(x_k, \widetilde{\theta}_k)\| \ge \varepsilon > 0$ and there exists a positive constant $\omega \in [0, 1]$ such that $\|s_k\| \le \omega h_k$, and it follows from assumptions and Lemma 3 and Lemma 4 that

$$\begin{aligned}
\mathrm{E}_\theta\left[\left|\hat{r}_k - 1\right| \mid x_k\right] &= \left|\mathrm{E}_\theta\left[\hat{r}_k \mid x_k\right] - 1\right| = \frac{\left|\mathrm{E}_\theta\left[F\left(x_k\right) - F\left(x_k + s_k\right) \mid x_k\right] - \mathrm{E}_\theta\left[\widehat{q}^{(k)}(0) - \widehat{q}^{(k)}\left(s_k\right) \mid x_k\right]\right|}{\left|\mathrm{E}_\theta\left[\widehat{q}^{(k)}(0) - \widehat{q}^{(k)}\left(s_k\right) \mid x_k\right]\right|} \\
&\le \frac{\mathrm{E}_\theta\left[\left|F\left(x_k + s_k\right) - \widehat{q}^{(k)}\left(s_k\right)\right| \mid x_k\right]}{\mathrm{E}_\theta\left[\widehat{q}^{(k)}(0) - \widehat{q}^{(k)}\left(s_k\right) \mid x_k\right]}
\end{aligned}$$

$$\leq \frac{N\mathrm{E}_\theta\left[\left\|s_k\right\| \mid x_k\right] + (1/2)\,\rho\mathrm{E}_\theta\left[\left\|s_k\right\|^2 \mid x_k\right] + \mathrm{E}_\theta\left[c\left(\left\|s_k\right\|\right)\left\|s_k\right\| \mid x_k\right]}{C_1\mathrm{E}_\theta\left[\left\|\widehat{g}\left(x_k,\widetilde{\theta}_k\right)\right\| \min\left\{h_k, \left\|\widehat{g}\left(x_k,\widetilde{\theta}_k\right)\right\|/\rho\right\} \mid x_k\right]}$$
$$\leq \frac{\omega h_k\left(N + (1/2)\,\rho\omega h_k + c\left(\left\|s_k\right\|\right)\right)}{C_1\varepsilon\min\left\{h_k, \varepsilon/\rho\right\}}. \tag{28}$$

We can select $\widetilde{\Delta}$ small enough to satisfy $h_k \leq \widetilde{\Delta} \leq \varepsilon/\rho$, and

$$N + \frac{1}{2}\rho\omega h_k + c\left(\left\|s_k\right\|\right) \leq \left(1-\eta_1\right)\frac{C_1\varepsilon}{\omega}. \tag{29}$$

Thus, we have $\mathrm{E}_\theta[|\widehat{r}_k - 1| \mid x_k] \leq 1-\eta_1$; this result is shown that $\widehat{r}_k$ is closer to 1 than $\eta_1$ on average which implies by Algorithm 1 that there are infinite numbers of $k$ which satisfy $h_{k+1} \geq h_k$. □

**Theorem 6.** *Considering the new algorithm defined above, suppose that A1-A4 hold, and the sequence of iterates generated by Algorithm 1 satisfies*

$$\lim_{k\longrightarrow\infty}\inf\left\|\widehat{g}\left(x_k,\widetilde{\theta}_k\right)\right\| = 0, \tag{30}$$

*over realizations of the random samples* $\{\widetilde{\theta}_k\}_{k=0}^{\infty}$.

*Proof.* Here we use a contradiction with Lemma 5 to prove (30). For that purpose, assume that there exist $\varepsilon > 0$ and a positive index set $K$ such that for every $k \in K$ we have $\|\widehat{g}(x_k,\widetilde{\theta}_k)\| \geq \varepsilon > 0$. At the same time, it is assumed that there are infinite number of successful iterations, and we can obtain the following inequality from Algorithm 1:

$$F\left(x_k\right) - F\left(x_k + s_k\right) \geq \eta_1\left[\widehat{q}^{(k)}(0) - \widehat{q}^{(k)}\left(s_k\right)\right], \tag{31}$$

taking expectation with $x_k$ given in both sides of (31) we have

$$\begin{aligned}
&\mathrm{E}_\theta\left[\left|F\left(x_k\right) - F\left(x_k + s_k\right)\right| \mid x_k\right]\\
&\geq \eta_1\mathrm{E}_\theta\left[\widehat{q}^{(k)}(0) - \widehat{q}^{(k)}\left(s_k\right) \mid x_k\right]\\
&\geq \eta_1 C_1\mathrm{E}_\theta\left[\left\|\widehat{g}\left(x_k,\widetilde{\theta}_k\right)\right\| \min\left\{h_k, \frac{\left\|\widehat{g}\left(x_k,\widetilde{\theta}_k\right)\right\|}{\rho}\right\} \mid x_k\right]. \qquad (32)\\
&\geq \eta_1 C_1\varepsilon\min\left\{h_k, \frac{\varepsilon}{\rho}\right\}.
\end{aligned}$$

Considering assumption A3, as $k$ approaches infinity, the function $F(x)$ bounded below implies that the right side of (32) is tending to zero; thus, we have

$$\lim_{k\longrightarrow\infty} h_k = 0. \tag{33}$$

But the result in (33) is in contradiction with Lemma 5. Further assume there are finite number of successful iterations; that is, the iterations are unsuccessful for large $k$; then, the trust region radius is reduced, i.e. $\lim_{k\longrightarrow\infty} h_k = 0$. In this case, we still get a contradiction with Lemma 5. Hence, we obtain (30). □

*Remark 7.* We note that the stochastic gradient $\widehat{g}(x_k,\widetilde{\theta}_k)$ is an unbiased estimator of $\nabla F(x_k)$, i.e., $\mathrm{E}_\theta[\widehat{g}(x_k,\widetilde{\theta}_k) \mid x_k] = \nabla F(x_k)$. Thus, we can also obtain $\lim_{k\longrightarrow\infty}\inf\|\nabla F(x_k)\| = 0$ from the result in Theorem 6.

## 4. Numerical Experiments

In this section, we apply the proposed stochastic trust region method to solve convex and nonconvex problems with stochastic objectives and also compare it with SGD in terms of convergence time and central processing unit (CPU) runtime. Both methods will be tested on the following problems: problem 1 and 2 are convex stochastic optimization problems while problem 3 is nonconvex.

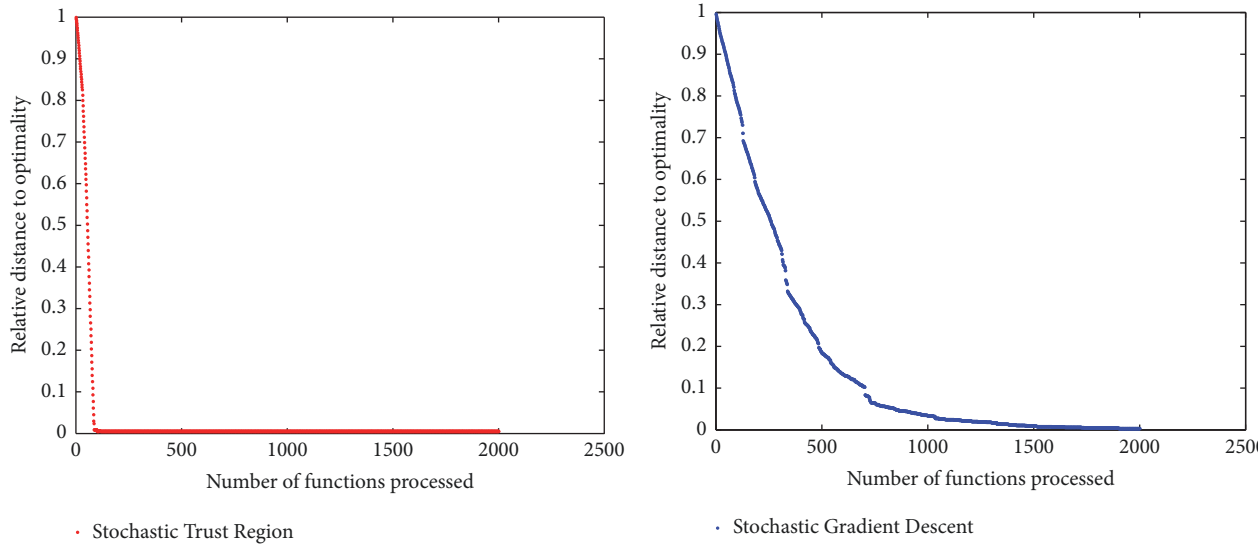
*4.1. Example 1: Standard Quadratic Function.* We use a stochastic quadratic objective function as a test case. In particular, consider a positive definite diagonal matrix $A \in R^{n\times n}$, a vector $b \in R^n$, a random vector $\theta \in R^n$, and diagonal matrix $\mathrm{diag}(\theta)$ defined by $\theta$. The function $F(x)$ is defined as

$$\begin{aligned}
F(x) &\coloneqq \mathrm{E}_\theta\left[f(x,\theta)\right]\\
&\coloneqq \mathrm{E}_\theta\left[\frac{1}{2}x^T\left(A + A\,\mathrm{diag}(\theta)\right)x + b^T x\right].
\end{aligned} \tag{34}$$

In (34), the vector $\theta$ is chosen uniformly at random from the $n$-dimensional box $\Theta = [-\theta_0,\theta_0]^n$ for some given constant $\theta_0 < 1$. The linear term $b^T x$ is added so that the instantaneous functions $f(x,\theta)$ have different minima which are (almost surely) different from the minimum of the average function $F(x)$. The quadratic term is chosen so that the condition number of $F(x)$ is the condition number of $A$. Indeed, since $\mathrm{E}_\theta[\theta] = 0$, the average function in (34) can be written as $F(x) = (1/2)x^T Ax + b^T x$. The parameter $\theta_0$ controls the variability of the instantaneous functions $f(x,\theta)$. For small $\theta_0 \approx 0$, the instantaneous functions are close to each other and to the average function. For large $\theta_0 \approx 1$, the instantaneous functions vary over a large range. Observe that we can write the optimum argument as $x^* = A^{-1}b$ for comparison against iterates $x_k$. Further consider a given $\sigma$ and study the convergence metric

$$\Gamma \coloneqq \min_k\left\{k : \frac{\left\|x_k - x^*\right\|}{\left\|x^*\right\|} \leq \sigma\right\}, \tag{35}$$

which represents the time needed to achieve a given relative distance to optimality. To study the effect of the problem's




(a) 1 (b) 2

Figure 1: Convergence of the stochastic trust region method and SGD for function (35). Relative distance to opotimality $\|x_k - x^*\|/\|x^*\|$ is shown with respect to the number of the stochastic function processed. For our method, the number of iterations required to achieve a certain accuracy is smaller than the corresponding number for SGD.

condition number, we generate instances of (34) by choosing $b$ uniformly at random from the box $[0, 1]^n$ and the matrix $A$ as diagonal with elements $a_{ii}$ uniformly drawn from the discrete set $\{1, 10^{-1}, \ldots, 10^{-\xi}\}$. This choice of $A$ yields problems with condition number $10^{\xi}$. Representative runs of stochastic trust region method and SGD for $n = 50$, $\theta_0 = 0.5$ and $\xi = 3$ are shown in Figure 1. For the stochastic trust region method and SGD run, the stochastic gradients $\widehat{g}(x, \widetilde{\theta})$ in (4) are computed as an average of $L = 10$ realizations. Moreover, we set $\gamma_1 = 0.95$, $\gamma_2 = 2$, and $\overline{h} = 50$ for our method due to the randomness of instantaneous functions $f(x, \widetilde{\theta})$ and the large condition number.

As expected for a problem with large condition number, the condition number of $F(x)$ is $10^3$ since we are using $\xi = 3$, and the stochastic trust region method is much faster than SGD. After $k = 1892$, the distance to optimality for the SGD iterate is $\|x_k - x^*\|/\|x^*\| = 4.2 \times 10^{-3}$. Comparable accuracy for stochastic trust region method is achieved after $k = 105$ iterations. Conversely, upon processing $k = 2 \times 10^3$ random functions, our method achieves accuracy $\|x_k - x^*\|/\|x^*\| = 1.7 \times 10^{-3}$.

*4.2. Example 2: Extended Powell Singular Function.* Powell singular function (PSF) is also known as Powell Quartic Function. The Hessian matrix at minimizer is doubly singular; thus, PSF is a severe test problem. To analyse the numerical performance of our algorithm, for a random vector $\theta \in R^{n/4}$, a variable $x \in R^n$, we develop stochastic Extended PSF as the test function as follows [18]:

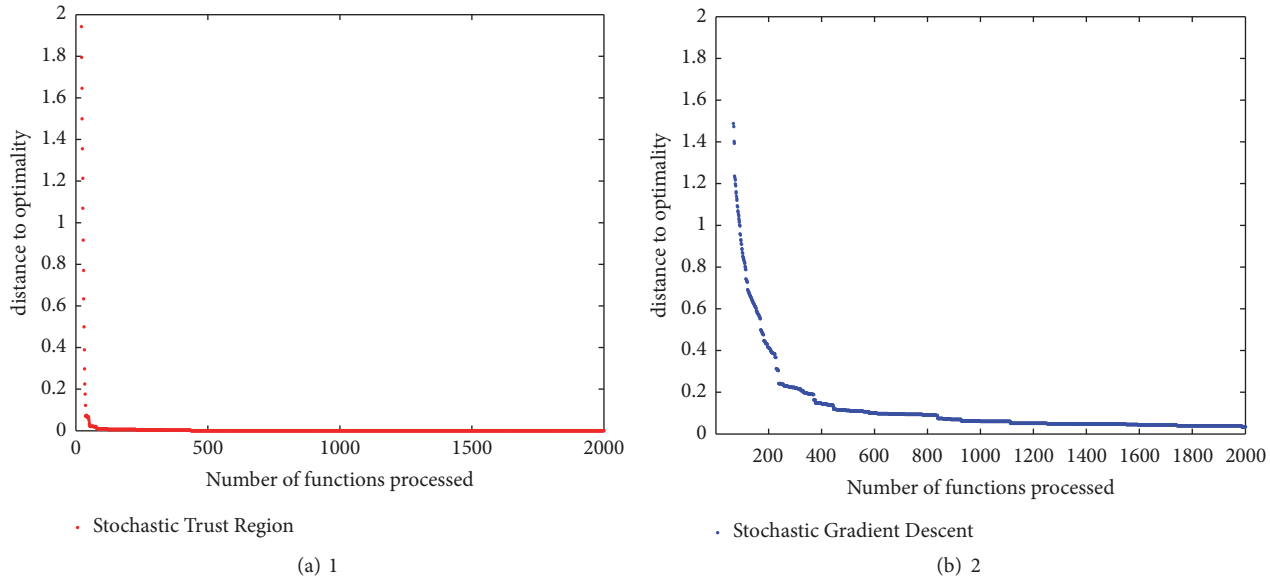
$$F(x) = \mathrm{E}_\theta [f(x, \theta)] = \mathrm{E}_\theta \left[ \sum_{i=1}^{n/4} (1 + \theta_i) \cdot \left( \left(x_{4i-3} + 10x_{4i-2}\right)^2 + 5\left(x_{4i-1} - x_{4i}\right)^2 + \left(x_{4i-2} - 2x_{4i-1}\right)^4 + 10\left(x_{4i-3} - x_{4i}\right)^4 \right) \right], \tag{36}$$

where $\theta$ is selected uniformly at random from the box $\Theta = [-\theta_0, \theta_0]^{n/4}$ with $\theta_0 = 0.5$. Indeed, observe that since $\mathrm{E}_\theta[\theta] = 0$, the average function $F(x)$ in (36) can be written as

$$F(x) = \sum_{i=1}^{n/4} \left( \left(x_{4i-3} + 10x_{4i-2}\right)^2 + 5\left(x_{4i-1} - x_{4i}\right)^2 + \left(x_{4i-2} - 2x_{4i-1}\right)^4 + 10\left(x_{4i-3} - x_{4i}\right)^4 \right). \tag{37}$$

Standard starting point is $x_0 = (3, -1, 0, 1, \ldots, 3, -1, 0, 1)^T$ and the Hessian matrix at the standard starting point is nonsingular. The average function $F(x)$ in (36) is convex and the unique unconstrained minimizer is $x^* = (0, 0, \ldots, 0)^T$ with $F(x^*) = 0$. In Figure 2, for $n = 40$, $L = 5$, specially $\gamma_1 = 0.5$, $\gamma_2 = 2$, and $\overline{h} = 5$, since the minimizer is $x^* = (0, 0, \ldots, 0)^T$, we plot the variation of distance to optimality instead of relative distance with the increase of the number of iterations.

As the number of iterations increases, we can observe that SGD achieves accuracy $\|x_k\| = 8.9 \times 10^{-3}$ after $k = 2000$ iterations and has a slow convergence to optimality. However, our method needs only $k = 82$ iterations to achieve the same accuracy. Furthermore, after $k = 285$ iterations, stochastic trust region method has $\|x_k\| = 5.2 \times 10^{-4}$ with




(a) 1 (b) 2

Figure 2: Convergence of the stochastic trust region method and SGD for function (37). Distance to opotimality $\|x_k - x^*\|$ is shown with respect to the number $k$ of stochastic function processed.

$F(x_k) = 4.8532 \times 10^{-7}$. The result indicates that our method has adequate accuracy and a faster convergence rate than SGD for severe convex problems.

*4.3. Example 3: Extended Rosenbrock Function.* In mathematical optimization, the Rosenbrock function is frequently used in nonconvex optimization problems as a performance test problem for optimization algorithms, introduced by Howard H. Rosenbrock in 1960. The global minimum lies in a narrow, long, parabolic shaped flat valley. It is trivial to find the valley; however, convergence to the global minimum is difficult. Here we use a stochastic version of the Extended Rosenbrock function as a test case [18]. Specifically, consider a random vector $\theta \in R^{n/2}$, a variable $x \in R^n$. The function $F(x)$ is defined as

$$F(x) \coloneqq \mathrm{E}_\theta [f(x,\theta)] \coloneqq \mathrm{E}_\theta \left[ \sum_{i=1}^{n/2} (1+\theta_i) \cdot \left( \left( 100 \left( {x_{2i-1}}^2 - x_{2i} \right)^2 + (1 - x_{2i-1})^2 \right) \right] . \tag{38}$$

Similarly, we choose $\theta$ uniformly at random from the box $\Theta = [-\theta_0, \theta_0]^{n/2}$ and set $\theta_0 = 0.5$. In fact, from the observation $\mathrm{E}_\theta[\theta] = 0$, we can write the average function as

$$F(x) = \sum_{i=1}^{n/2} \left( 100 \left( {x_{2i-1}}^2 - x_{2i} \right)^2 + (1 - x_{2i-1})^2 \right). \tag{39}$$

The function in (39) has a minimizer $x^* = [1, 1, \ldots, 1]^T$ with $F(x^*) = 0$. For the purpose of this test problem, we choose $n = 50$, $L = 5$. In particular, set $\gamma_1 = 0.5$, $\gamma_2 = 2$, and $\overline{h} = 5$ for our method. The relative distances in (38) after running of stochastic trust region method and SGD are provided in Figure 3. After $k = 99$ iterations, stochastic trust region method achieves accuracy $\|x_k - x^*\|/\|x^*\| = 1.94 \times 10^{-4}$, and correspondingly SGD achieves such accuracy after $k = 1523$ iterations.

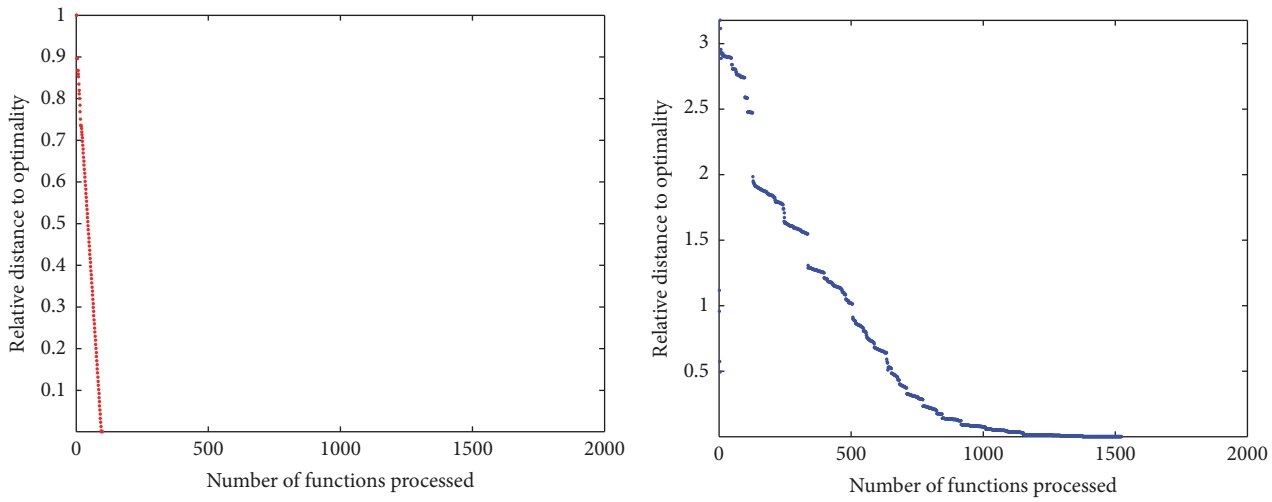
Since each iteration of stochastic trust region method is more complex than SGD, we also compare the performances in terms of central processing unit (CPU) runtime required to achieve accuracy $10^{-2}$. We report the average runtime of stochastic trust region method and SGD for problems above in Table 1; the value of parameters is the same as above. In particular, we call stochastic trust region method STR for short in the table. As we can see that our method enjoys a significant improvement in CPU runtime than SGD for both convex and nonconvex problems.

## 5. Conclusions

In this paper, we propose a stochastic trust region method and show that the new algorithm is convergent for solving unconstrained minimization problems with stochastic objectives. Based on the trust region framework and the BFGS update, our method can deal with convex optimization problems with ill-conditioned objective functions as well as nonconvex optimization problems. With careful analysis, we are able to show that our method is convergent. Numerical results illustrate that the method can efficiently solve the given test problems. Therefore, the new method is potentially efficient and thus paves the way towards developing concrete algorithms for specific applications.

## Data Availability

The date used to support the findings of this study are included within the article.




(a) 1 (b) 2

FIGURE 3: Convergence of the stochastic trust region method and SGD for function (39). Relative distance to opotimality $\|x_k - x^*\|/\|x^*\|$ is shown with respect to the number $k$ of the stochastic function processed. For our method, the number of iterations required to achieve a certain accuracy is smaller than the corresponding number for SGD.

TABLE 1: CPU runtime of STR and SGD.

| | STR | SGD |
|---|---|---|
| Standard Quadratic function | 0.22s | 0.7s |
| Extended Powell Singular function | 0.18s | 0.64s |
| Extended Rosenbrock function | 0.25s | 0.73s |

## Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication of this paper.

## Acknowledgments

The authors would like to thank the editor and the reviewers for their very careful reading and constructive comments which led to great improvement of the paper. This work was supported by National Natural Science Foundation of China [Grant nos. 11601252 and 11571178].

## References

[1] A. Mokhtari and A. Ribeiro, "A quasi-Newton method for large scale support vector machines," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2014*, pp. 8302–8306, Italy, May 2014.

[2] L. Bottou and Y. Le Cun, "On-line learning for very large data sets," *Applied Stochastic Models in Business and Industry*, vol. 21, no. 2, pp. 137–151, 2005.

[3] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proceedings of the COMPSTAT'2010*, pp. 177–186, Physica-Verlag/Springer, Berlin, Germnay, 2010.

[4] A. Mokhtari and A. Ribeiro, "A dual stochastic DFP algorithm for optimal resource allocation in wireless systems," in *Proceedings of the 14th Workshop on Signal Processing Advances in Wireless Communications (SPAWC 2013)*, pp. 21–25, Darmstadt, Germany, June 2013.

[5] A. Ribeiro, "Ergodic stochastic optimization algorithms for wireless communication and networking," *IEEE Transactions on Signal Processing*, vol. 58, no. 12, pp. 6369–6386, 2010.

[6] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, "Robust stochastic approximation approach to stochastic programming," *SIAM Journal on Optimization*, vol. 19, no. 4, pp. 1574–1609, 2009.

[7] S. Shalev-Shwartz and N. Srebro, "SVM optimization: Inverse dependence on training set size," in *Proceedings of the 25th International Conference on Machine Learning*, pp. 928–935, Finland, July 2008.

[8] S. Shalev-Shwartz, Y. Singer, and N. Srebro, "Pegasos: primal estimated sub-gradient solver for svm," in *Proceedings of the 24th International Conference on Machine Learning (ICML '07)*, pp. 807–814, June 2007.

[9] T. Zhang, "Solving large scale linear prediction problems using stochastic gradient descent algorithms," in *Proceedings of the twenty-first international conference on Machine learning*, pp. 919–926, Banff, Alberta, Canada, July 2004.

[10] M. Schmidt, N. Le Roux, and F. Bach, "Minimizing finite sums with the stochastic average gradient," *Mathematical Programming*, vol. 162, no. 5, pp. 1–30, 2017.

[11] J. Konecny, J. Liu, P. Richtarik, and M. Takac, "Mini-Batch Semi-Stochastic Gradient Descent in the Proximal Setting," *IEEE*

*Journal of Selected Topics in Signal Processing*, vol. 10, no. 2, pp. 242–255, 2016.

[12] L. Zhang, M. Mahdavi, and R. Jin, "Linear convergence with condition number independent access of full gradients," *Advances in Neural Information Processing Systems*, pp. 980–988, 2013.

[13] J. R. Birge, X. Chen, L. Qi, and Z. Wei, "A stochastic Newton method for stochastic quadratic programs with resource," Tech. Rep. 1, University of Michigan, Ann Arbor, Mich, USA, 1995, pp. 113-141.

[14] A. Mokhtari and A. Ribeiro, "Regularized Stochastic BFGS algorithm," *Global Conference on Signal and Information Processing IEEE*, vol. 62, no. 23, pp. 6089–6104, 2014.

[15] P. L. Toint, "Global convergence of a class of trust-region methods for nonconvex minimization in Hilbert space," *IMA Journal of Numerical Analysis (IMAJNA)*, vol. 8, no. 2, pp. 231–252, 1988.

[16] R. H. Byrd, R. B. Schnabel, and G. Shultz, "A trust region algorithm for nonlinearly constrained optimization," *SIAM Journal on Numerical Analysis*, vol. 24, no. 5, pp. 1152–1170, 1987.

[17] W. Sun and Y. Yuan, *Optimization Theory and Methods. Nonlinear Programming*, Springer, New York, NY, USA, 2006.

[18] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer, New York, NY, USA, 1999.