Journal of Computational Mathematics
Vol.41, No.6, 2023, 1192–1221.

http://www.global-sci.org/jcm
doi:10.4208/jcm.2112-m2021-0072

A STOCHASTIC NEWTON METHOD FOR NONLINEAR EQUATIONS*

Jiani Wang
School of Mathematical Sciences, Dalian University of Technology, Dalian, China
Email: jianiwang@163.com
Xiao Wang[1)]
School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing, China
Peng Cheng Laboratory, Shenzhen, China
Email: wangxiao@ucas.ac.cn; wangx07@pcl.ac.cn
Liwei Zhang
School of Mathematical Sciences, Dalian University of Technology, Dalian, China
Email: lwzhang@dlut.edu.cn

Abstract

In this paper, we study a stochastic Newton method for nonlinear equations, whose exact function information is difficult to obtain while only stochastic approximations are available. At each iteration of the proposed algorithm, an inexact Newton step is first computed based on stochastic zeroth- and first-order oracles. To encourage the possible reduction of the optimality error, we then take the unit step size if it is acceptable by an inexact Armijo line search condition. Otherwise, a small step size will be taken to help induce desired good properties. Then we investigate convergence properties of the proposed algorithm and obtain the almost sure global convergence under certain conditions. We also explore the computational complexities to find an approximate solution in terms of calls to stochastic zeroth- and first-order oracles, when the proposed algorithm returns a randomly chosen output. Furthermore, we analyze the local convergence properties of the algorithm and establish the local convergence rate in high probability. At last we present preliminary numerical tests and the results demonstrate the promising performances of the proposed algorithm.

Mathematics subject classification: 49M37, 65K05, 90C30.
Key words: Nonlinear equations, Stochastic approximation, Line search, Global convergence, Computational complexity, Local convergence rate.

1. Introduction

In this paper, we consider the following system of nonlinear equations:

$$F(x) = \mathbb{E}[f(x, \xi)] = 0, \tag{1.1}$$

where $\xi : \Omega \to \mathbb{W}$ is a random variable defined on a given probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and $f : \mathbb{R}^n \times \mathbb{W} \to \mathbb{R}^n$. Here $\mathbb{W}$ is a measurable space, and $\mathbb{E}$ represents the expectation with respect to the random variable. We assume that (1.1) has a solution and $F : \mathbb{R}^n \to \mathbb{R}^n$ is continuously differentiable. As in (1.1) the probability distribution function may not be

* Received March 9, 2021 / Revised version received June 28, 2021 / Accepted December 7, 2021 / Published online December 12, 2022 /
[1)] Corresponding author

available or the expectation with respect to ξ may be difficult to calculate, we assume in this paper that the exact function information such as function value and Jacobian matrix cannot be obtained. The problem (1.1) covers a wide range of applications, such as stochastic dynamic programming [27], and stochastic PDEs [16, 23]. Problem (1.1) can also be regarded as an extension of finding a stationary point of minimization problems. Consider the well-known expected risk minimization problem

$$\min_{x\in\mathbb{R}^n} \ H(x) = \mathbb{E}[h(x,\xi)],$$

where $H:\mathbb{R}^n\to\mathbb{R}$ is a smooth function. Referring to Theorem 7.44 [28], if the function $h(x,\xi)$ is differentiable at x with probability 1 and satisfies certain conditions, its local minimizer is also the solution of the following system of equations:

$$\mathbb{E}[\nabla_x h(x,\xi)] = 0.$$

The classic Newton's method, first presented by Newton in 1687 [34], for finding the root of $F(x)$ is to update iterates through

$$x^{k+1} = x^k - (\mathcal{J}F(x^k))^{-1}F(x^k),$$

if the Jacobian matrix $\mathcal{J}F(x^k)$ is non-singular. As is well-known, classic Newton's method owns a fast convergence rate when close to a solution of nonlinear equations. More specifically, locally q-quadratic convergence rate can be achieved under certain conditions if $\mathcal{J}F(x^*)$ is non-singular [7]. However, when solving nonlinear equations (1.1), as the exact function value $F(x)$ and its Jacobian $\mathcal{J}F(x)$ are not available, classic Newton's method is not applicable any more. Similar to [8,21,25], we assume that $F(x)$ and Jacobian $\mathcal{J}F(x)$ can be approximated via calls to stochastic zeroth-order oracles ($\mathcal{SZO}$) and stochastic first-order oracles ($\mathcal{SFO}$), respectively. In this paper, we will employ those stochastic information to propose a stochastic approximation method for solving (1.1).

Study on stochastic approximation (SA) methods for nonlinear optimization dates back to the pioneer work [26]. In the past decade, along with the development of complexity theory, profound research progress on SA methods for nonlinear optimization has been made, including but not limited to [2, 9, 11, 13, 14]. Quite recently, Milzarek *et al.* [18] propose a stochastic semismooth Newton method for nonsmooth nonconvex optimization by solving an equivalent nonsmooth fixed point-type equation and study global and local convergence properties of the proposed algorithm. However, to define proximal gradient steps and to set a growth condition for trial steps, the objective function of original optimization problem needs to be utilized, thus plays a crucial role in theoretical analysis. In [29], a stochastic Gauss-Newton method (SGN) was proposed for solving compositional optimization problems. This method can be used to solve stochastic nonlinear equations (1.1) by reformulating it into an optimization problem, namely, minimizing a given norm of $F(x)$. At each iteration, it solves an approximate prox-linear model which is constructed based on stochastic oracles. In contrast, SA methods directly designed for general nonlinear equations in the form of (1.1) are quite limited. In this paper, motivated by the success of classic Newton's method for deterministic nonlinear equations, we will propose a stochastic Newton method for (1.1) based on stochastic oracles and investigate its theoretical and numerical performances.

As is known, the step size has a great influence on both theoretical and numerical performances of an SA method. It is quite popular to set step sizes as either constants [18] or

a diminishing sequence [4]. However, in practical computation it is normally not easy to choose the best-tuned step sizes. In deterministic optimization line search strategy is usually incorporated to adaptively compute the step sizes by allowing a sufficient reduction of function values at each iteration and to ensure the global convergence of an algorithm. Recently, researchers consider to incorporate a line search strategy in SA methods for nonlinear optimization. For example, paper [24] studies a stochastic line search algorithm for stochastic optimization. As pointed in [24], due to the stochasticity when applying line search with stochastic estimates, it may lead to false steps which make the objective value at next iterate arbitrarily larger than current iterate. To deal with this challenge, in [24] it requires high probability with which the random gradient and function estimates (including the function value at next iterate) are representatives of their true counterparts. In [32], the use of the Armijo line search with stochastic gradient is investigated. However, the proposed algorithm requires the objective function satisfy the strong growth condition to achieve the desired convergence. Paper [12] studies SA methods with line search for stochastic variational inequalities where the operator is required to be pseudomonotone. Note that all aforementioned algorithms do not aim for general nonlinear equations of the form (1.1). Moreover, as the line search condition is designed based on stochastic oracles, it seems unnecessary trying to find a step size satisfying line search condition exactly.

Motivated by these points, we consider to incorporate an inexact Armijo line search strategy in our stochastic Newton method for nonlinear equations (1.1), allowing some inexact tolerance on the line search. In addition, to control the uncertainty of stochastic information, we only check the line search condition once at each iteration to determine if the unit step size is acceptable. Specifically, in this paper, we propose a stochastic Newton method for solving nonlinear equations of the form (1.1). Considering the problems with an unknown distribution or online data acquisition, sub-sampling strategy is applied to compute approximate function values and Jacobians. At each iteration, the method computes an inexact Newton step within a given tolerance, where the Newton's equation is built on stochastic zeroth- and first-order oracles. Then the unit step size is taken if it satisfies the inexact line search condition. Otherwise, a preset small step size is taken to secure good convergence properties of the proposed algorithm. We next explore the global convergence properties of the proposed algorithm and show the global convergence in expectation as well as almost-sure convergence under certain conditions. We also analyze its computational complexities when the proposed algorithm returns a randomly chosen iterate as the output. Furthermore, we study local convergence properties of proposed algorithm and prove the local convergence rates with high probability if the sample sizes and sampling rates are chosen appropriately and increase sufficiently fast. Numerical tests on some large data sets show the promising performance of the proposed algorithm.

We summarize contributions of this paper in the following table, by showing a comparison between our algorithm and those in related works. As we can always transform (1.1) into a minimization problem through a given norm, many SA methods for nonlinear minimization problems can be applied. Since the incorporation of an acceleration technique in our algorithm is out of the scope of the paper, we only consider the pure randomized stochastic gradient method (RSG) for nonconvex optimization and apply it to minimize $h(x) := \|F(x)\|_2^2$. By [9], RSG can find an ϵ-stationary point $\bar{x}$, i.e. $\mathbb{E}[\|\nabla h(\bar{x})\|_2] \leq \epsilon$, via $\mathcal{O}(\epsilon^{-4})$ calls to stochastic first-order oracles. Instead of directly solving (1.1), SGN [29] focuses on minimizing $\|F(x)\|$ through a given norm $\|\cdot\|$. Without any variance reduction technique, SGN can establish $\mathcal{SZO}$ complexity in $\mathcal{O}(\epsilon^{-6})$ and $\mathcal{SFO}$ complexity in $\mathcal{O}(\epsilon^{-4})$ to reach an ϵ-stationary point satisfying $\mathbb{E}[\|\tilde{G}_M(x^k)\|] \leq \epsilon$,

where $\tilde{G}_M(x) = M(x - \tilde{T}_M(x))$ with $\tilde{T}_M(x) = \text{argmin}_z\{\|\tilde{F}(x) + \tilde{J}(x)(z-x)\| + \frac{M}{2}\|z-x\|^2\}$ and $\tilde{F}(x)$ and $\tilde{J}(x)$ are stochastic zeroth- and first-order oracles. However, different from both RSG and SGN, our algorithm is designed directly for general nonlinear equations (1.1). Under certain conditions, our algorithm can find an ϵ-approximate solution $\bar{x}$, namely $\mathbb{E}[\|F(\bar{x})\|_2] \le \epsilon$. And it enjoys superlinear convergence rate in high probability. More comparison details are given in the following table. Specifically, $\mathcal{SFO}$ for RSG means the stochastic first-order oracle to the gradient of $\|F(x)\|_2^2$.

Table 1.1: Comparison between our algorithm with RSG and SGN.

Algorithm	Problem to solve	Criticality measure	Complexity	Local convergence
RSG [9]	min $\|F(x)\|_2^2$	$\mathbb{E}[\|\mathcal{J}F(\bar{x})F(\bar{x})\|_2] \le \epsilon$	$\mathcal{SFO} \sim \mathcal{O}(\epsilon^{-4})$	—
SGN [29]	min $\|F(x)\|$	$\mathbb{E}[\|\tilde{G}_M(\bar{x})\|] \le \epsilon$	$\mathcal{SZO} \sim \mathcal{O}(\epsilon^{-6})$ $\mathcal{SFO} \sim \mathcal{O}(\epsilon^{-4})$	—
Alg. 2.1 (this paper)	$F(x) = 0$	$\mathbb{E}[\|F(\bar{x})\|_2] \le \epsilon$	(Thm 4.1) $\mathcal{SZO} \sim \tilde{O}(\epsilon^{-6})$, $\mathcal{SFO} \sim \tilde{O}(\epsilon^{-4})$	Superlinear
			(Thm 4.2) $\mathcal{SZO} \sim \tilde{O}(\epsilon^{-6})$, $\mathcal{SFO} \sim \tilde{O}(\epsilon^{-2})$	

Organization The remainder of this paper is organized as follows. We present details of a stochastic Newton method for solving problem (1.1) in Section 2. Then we analyze the theoretical properties of the proposed algorithm and establish its global convergence in Section 3. In Section 4 we explore the computational complexity of the proposed algorithm with respect to stochastic zeroth- and first-order oracles to find an approximate solution. In Section 5, we analyze the local convergence and show the corresponding convergence rates under different conditions. Finally, we report some preliminary numerical results in Section 6.

Notations We use the following notations throughout this paper. $\mathbb{N}$ and $\mathbb{N}_+$ denote the set of nonnegative integers and positive integers, respectively. For any $n \in \mathbb{N}_+$, $[n]$ denotes the set $\{1, \cdots, n\}$ and for any $x \in \mathbb{R}$, $\lceil x \rceil$ denotes the smallest integer no less than x. We use $\|\cdot\|$ to represent the Euclidean norm and its induced matrix norm. The superscript ^{k} refers to an iteration number and $(\cdot^k)_k$ refers to a sequence. For a set $S \subseteq \mathbb{R}^n$, the mapping $\mathbf{1}_S : \mathbb{R}^n \to \{0, 1\}$ is the associated characteristic function of S. Given a mapping $\Phi : \Omega \to \mathbb{R}^n$, we set $\Phi^{-1}(S) = \{\omega \in \Omega : \Phi(\omega) \bigcap S \neq \emptyset\}$. The function Φ is called measurable if $\Phi^{-1}(S)$ is measurable for any closed set $S \subseteq \mathbb{R}^n$. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. We say $\xi \in \mathcal{F}$ if ξ is $\mathcal{F}$-measurable. We use $\mathcal{B}(\mathbb{R}^n)$ to denote the Borel σ-algebra of $\mathbb{R}^n$ and $\sigma(\xi^1, \cdots, \xi^k)$ to denote the σ-algebra generated by the family of random variables $\xi^1, \cdots, \xi^k$. We use $\mathbb{E}[\xi|\mathcal{H}]$ to denote the conditional expectation of a random variable $\xi \in L^1(\Omega)$ given a sub-σ-algebra $\mathcal{H} \subseteq \mathcal{F}$, where $L^1(\Omega) := L^1(\Omega, \mathbb{P})$ denotes the L^1 space in Ω. The conditional probability of $A \in \mathcal{F}$ given $\mathcal{H} \subseteq \mathcal{F}$ is defined as $\mathbb{P}(A|\mathcal{H}) := \mathbb{E}[\mathbf{1}_A|\mathcal{H}]$. The abbreviations "a.e." and "a.s." stand for "almost everywhere" and "almost surely", respectively. The space ℓ_+^1 consists of all sequences $(x^k)_{k\ge0}$ satisfying $0 \le x^k \in \mathbb{R}$ for any $k \ge 0$ and $\sum_{k\ge0} x^k < +\infty$.

2. A Stochastic Newton Method for (1.1)

Since computations of exact function value $F(x)$ and the Jacobian matrix $\mathcal{J}F(x)$ of (1.1) are difficult sometimes even impossible, traditional optimization methods for solving nonlinear

equations based on exact function information are not applicable any more. Motivated by this, we next propose an SA method based on stochastic function values and Jacobians to solve (1.1).

We assume in this paper that stochastic function values of F and its Jacobians can be obtained through calls to stochastic zeroth- and first-order oracles. Similar to [5, 18, 33], we compute the mini-batch approximate function values and Jacobians. More specifically, given an underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a measurable space $(\Xi, \mathcal{X})$, suppose that there exist a zeroth-order oracle $\digamma : \mathbb{R}^n \times \Xi \to \mathbb{R}^n$ and a first-order oracle $\mathcal{G} : \mathbb{R}^n \times \Xi \to \mathbb{R}^{n\times n}$, where both $\digamma$ and $\mathcal{G}$ are *Carathéodory* functions [28], namely both $\mathcal{F}$ and $\mathcal{G}$ are continuous with respect to $x \in \mathbb{R}^n$ for fixed $z \in \Xi$ and measurable with respect to $z \in \Xi$ for fixed $x \in \mathbb{R}^n$. Suppose that the space Ω is rich enough such that for each $k \in \mathbb{N}$, we can generate two mini-batches of random samples

$$t^k := \left\{t_1^k, \cdots, t_{n_F^k}^k\right\} \quad \text{and} \quad s^k := \left\{s_1^k, \cdots, s_{n_G^k}^k\right\}, \tag{2.1}$$

where $t_i^k, s_j^k : \Omega \to \Xi, i \in [n_F^k], j \in [n_G^k]$ are $(\mathcal{F}, \mathcal{X})$-measurable random mappings and mutually independent and n_F^k, n_G^k denote the sample size of t^k, s^k respectively. We can obtain the approximate function value $\digamma(x, t_i^k)$ and Jacobian matrix $\mathcal{G}(x, s_j^k)$ for each $i \in [n_F^k], j \in [n_G^k]$. Then we compute the mini-batch approximate function value $F_{t^k}(x)$ and Jacobian matrix $G_{s^k}(x)$ through

$$F_{t^k}(x) := \frac{1}{n_F^k}\sum_{i=1}^{n_F^k} \digamma(x, t_i^k), \quad G_{s^k}(x) := \frac{1}{n_G^k}\sum_{j=1}^{n_G^k} \mathcal{G}(x, s_j^k), \tag{2.2}$$

respectively. Based on these stochastic approximations we build the following approximate Newton's equation

$$G_{s^k}(x^k)d = -F_{t^k}(x^k). \tag{2.3}$$

Given a tolerance $\eta^k \in [0, 1)$, we look for a solution d^k of (2.3) satisfying

$$\|F_{t^k}(x^k) + G_{s^k}(x^k)d^k\| \le \eta^k \|F_{t^k}(x^k)\|. \tag{2.4}$$

With the search direction d^k, the next step is to determine the step size. As is well known, line search strategies have been widely used in deterministic optimization. They help to realize the global convergence of algorithms by forcing the objective function values decreasing sufficiently at each iteration. One of the most popular line search conditions is the Armijo line search [22]. Take general nonlinear equations $F(x) = 0$ for example. If $F(x^k) \neq 0$ and $\mathcal{J}F(x^k)$ is nonsingular, the Newton's direction is $d_N^k = -(\mathcal{J}F(x^k))^{-1}F(x^k)$. Then the classic Armijo line search condition (see also [3]) is to find the smallest nonnegative integer k_j such that

$$\|F(x_k + \rho^{k_j} d_N^k)\| \le (1 - 0.5c_1\rho^{k_j})\|F(x^k)\|, \tag{2.5}$$

where $c_1 \in (0, 1)$. However, we cannot apply (2.5) directly to problem (1.1) where only stochastic approximations are available. Hence in our stochastic Newton method, we adopt the following line search condition which is defined based on sub-sampled function values. Given $\varepsilon^k > 0$ and a randomly chosen sample set t^{k+1}, we check if the unit step size satisfies the condition

$$\|F_{t^{k+1}}(x^k + d^k)\| \le (1 - c)\|F_{t^k}(x^k)\| + \varepsilon^k, \tag{2.6}$$

where $c \in (0, 0.5)$. To better control the stochasticity of the function value at $x^k + d^k$, the randomly generated sample set t^{k+1} is independent of x^k and d^k, which plays a crucial role in

theoretical analysis later. Whenever (2.6) is satisfied, the unit step size is taken. Otherwise, we take a preset step size α^k to generate x^{k+1}.

We now summarize the stochastic Newton method for (1.1) in Algorithm 2.1 below.

Algorithm 2.1. A stochastic Newton method for (1.1)

Require: Parameters $c \in (0, 0.5)$ and $\eta \in [0, 1)$, initial iterate $x^0 \in \mathbb{R}^n$, initial sample sets t^0 and s^0, two sequences of sample sizes $(n_F^k)_k$ and $(n_G^k)_k$, non-negative sequences $(\eta^k)_k \subseteq [0, \eta]$ and $(\varepsilon^k)_k$, $(\alpha^k)_k \subseteq (0, 1)$. Let $k = 0$.

1: **for** $k = 0, 1, \cdots$ **do**
2: If x^k satisfies some termination criterion, then stop and return x^k.
3: Compute $F_{t^k}(x^k)$ and $G_{s^k}(x^k)$ through (2.2) and solve (2.3) obtaining d^k satisfying (2.4).
4: Randomly choose a sample set t^{k+1} with size n_F^{k+1} through (2.1) and compute $F_{t^{k+1}}(x^k + d^k)$ through (2.2). If (2.6) is satisfied, set $x^{k+1} = x^k + d^k$. Otherwise, set $x^{k+1} = x^k + \alpha^k d^k$.
5: Randomly choose the sample set s^{k+1} with sizes n_G^{k+1} through (2.1). Let $k = k + 1$.
6: **end for**

Note that in Algorithm 2.1 we only present the pseudocode of the stochastic Newton method. At the beginning of Algorithm 2.1, we first generate two sequences of sample sizes $(n_F^k)_k$ and $(n_G^k)_k$. For each $k \in \mathbb{N}$, the sample size n_F^{k+1} and the sampling set t^{k+1} is used to compute both $F_{t^{k+1}}(x^k + d^k)$ and $F_{t^{k+1}}(x^{k+1})$. We do not set the termination criterion. It will be specified in Section 6. Moreover, we do not specify the values of α^k here. We will investigate the role that step sizes play and analyze theoretical properties of Algorithm 2.1 with different settings of α^k in Section 3. Roughly speaking, we will discuss two cases where α^k is diminishing and α^k is constant, then establish the global convergence and computational complexity accordingly.

Define the filtration

$$\mathcal{F}^k := \sigma(t^0, s^0, \cdots, t^k, s^k) \quad \text{and} \quad \hat{\mathcal{F}}^k := \sigma(t^0, s^0, \cdots, t^k, s^k, t^{k+1}), \quad k \geq 0,$$

and $\hat{\mathcal{F}}^{-1} := \hat{\mathcal{F}}^0$. We next show by induction that

$$x^k \in \hat{\mathcal{F}}^{k-1} \quad \text{for any} \quad k \geq 0. \tag{2.7}$$

It is obvious that $x^0 \in \hat{\mathcal{F}}^{-1}$. Now assume that $x^k \in \hat{\mathcal{F}}^{k-1}$, then $x^k \in \mathcal{F}^k$. We next show that $x^{k+1} \in \hat{\mathcal{F}}^k$. Since the stochastic oracle $\mathcal{G}$ is a *Carathéodory* function, by Section 4.10 in [1], $\mathcal{G}$ is jointly measurable with $\mathcal{G} \in \mathcal{B}(\mathbb{R}^n) \otimes \mathcal{X}$, where $\mathcal{B}(\mathbb{R}^n) \otimes \mathcal{X}$ denotes the product σ-algebra of the product space $\mathbb{R}^n \times \Xi$. Hence the functions $\xi_i^k : \Omega \to \mathbb{R}^n \times \Xi$, $i \in [n_G^k]$, defined as $\xi_i^k(\omega) := (x^k(\omega), s_i^k(\omega))$, are $(\mathcal{F}^k, \mathcal{B}(\mathbb{R}^n) \otimes \mathcal{X})$-measurable. For simplicity, we omit ω in above notations. Hence, G_{s^k} is a $\mathcal{F}^k$-measurable mapping, which yields $G_{s^k}(x^k) \in \mathcal{F}^k$. Similarly, we can obtain that F_{t^k} is an $\hat{\mathcal{F}}^{k-1}$-measurable mapping thus $F_{t^k}(x^k) \in \hat{\mathcal{F}}^{k-1}$. Then from (2.4) it obviously holds that $d^k \in \mathcal{F}^k$. We now define

$$x_1^{k+1} = x^k + d^k, \quad x_2^{k+1} = x^k + \alpha^k d^k. \tag{2.8}$$

As $x^k \in \hat{\mathcal{F}}^{k-1}$ and $d^k \in \mathcal{F}^k$, we have

$$x_1^{k+1}, x_2^{k+1} \in \mathcal{F}^k. \tag{2.9}$$

Define Y^{k+1} by

$$Y^{k+1} = \begin{cases} 1, & \text{if (2.6) is satisfied,} \\ 0, & \text{otherwise.} \end{cases}$$

Then Y^{k+1} is a random variable and x^{k+1} can be expressed as

$$x^{k+1} = Y^{k+1}x_1^{k+1} + (1 - Y^{k+1})x_2^{k+1}. \tag{2.10}$$

As $F_{t^{k+1}}(x_1^{k+1}) \in \hat{\mathcal{F}}^k$, the indicator function Y^{k+1} is $\hat{\mathcal{F}}^k \otimes \mathcal{B}(\mathbb{R}^n)$-measurable. Therefore, it follows from (2.10) that $x^{k+1} \in \hat{\mathcal{F}}^k$.

We now give several assumptions used throughout this paper.

Assumption 2.1.

(A1) *The function $F : \mathbb{R}^n \to \mathbb{R}^n$ is continuously differentiable on $\mathbb{R}^n$.*

(A2) *The Jacobian matrix $\mathcal{J}F(\cdot) : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ is Lipschitz continuous with modulus $L' > 0$.*

(A3) *There exist positive constants μ_F^k, μ_G^k, $\bar{\sigma}_F$ and $\bar{\sigma}_G$ such that for all $k \in \mathbb{N}$, t_i^k, $s_j^k \in \Xi$, $i \in [n_F^k]$, $j \in [n_G^k]$, and for any $x \in \mathbb{R}^n$,*

$$\|\mathbb{E}[\digamma(x, t_i^k)] - F(x)\| \le \mu_F^k, \qquad \|\mathbb{E}[\mathcal{G}(x, s_j^k)] - \mathcal{J}F(x)\| \le \mu_G^k,$$
$$\mathbb{E}[\|\digamma(x, t_i^k) - \mathbb{E}[\digamma(x, t_i^k)]\|^2] \le \bar{\sigma}_F^2, \qquad \mathbb{E}[\|\mathcal{G}(x, s_j^k) - \mathbb{E}[\mathcal{G}(x, s_j^k)]\|^2] \le \bar{\sigma}_G^2.$$

It is worth noting that assumption (A1) is reasonable. Following Theorem 7.44 in [28], if $F(\cdot)$ is well defined and finite valued at point $x \in \mathbb{R}^n$ and $f(\cdot, \xi)$ is differentiable at x for almost every $\xi \in \Omega$ and Lipschitz continuous in a neighborhood of x, then $F(\cdot)$ is differentiable at x. Moreover, following Theorem 7.43 in [28], if $\nabla f(x, \xi)$ is bounded by a P-integeable function for all x in a neighborhood of x, then ∇F is continuous at x, thus F is continuously differentiable at x.

Note that by the Lipschitz continuity of $\mathcal{J}F$ and the integral mean value theorem, we obtain

$$\begin{aligned} &\|F(y) - F(x) - \mathcal{J}F(x)(y - x)\| \\ &= \left\| \int_0^1 \mathcal{J}F(x + t(y - x))(y - x)dt - \mathcal{J}F(x)(y - x) \right\| \\ &= \left\| \int_0^1 [\mathcal{J}F(x + t(y - x)) - \mathcal{J}F(x)](y - x)dt \right\| \\ &\le \int_0^1 L'\|y - x\|^2 t dt = \frac{L'}{2}\|y - x\|^2 \quad \text{for any} \quad x, y \in \mathbb{R}^n, \end{aligned}$$

which further yields that

$$\|F(y)\| \le \|F(x) + \mathcal{J}F(x)(y - x)\| + \frac{L'}{2}\|y - x\|^2, \quad \forall x, y \in \mathbb{R}^n. \tag{2.11}$$

It is popular that in the convergence analysis of SA methods, such as [9] and [18], stochastic oracles are assumed unbiased. Different from these works in our algorithm we allow deviations of stochastic oracles from the true values. That is, both stochastic zeroth- and first-order oracles can provide biased estimates, as assumed in assumption (A3). We now define

$$\mathcal{E}_F^k(x) = \|F_{t^k}(x) - F(x)\|, \quad \mathcal{E}_G^k(x) = \|G_{s^k}(x) - \mathcal{J}F(x)\|. \tag{2.12}$$

Then it follows from (2.9), (2.10) and assumptions (A2) and (A3) that

$$\begin{aligned}
&\mathbb{E}[(\mathcal{E}_F^k(x^k))^2|\mathcal{F}^{k-1}]\\
&=\mathbb{E}[Y^k\|F_{t^k}(x_1^k)-F(x_1^k)\|^2|\mathcal{F}^{k-1}]+\mathbb{E}[(1-Y^k)\|F_{t^k}(x_2^k)-F(x_2^k)\|^2|\mathcal{F}^{k-1}]\\
&\le\mathbb{E}[\|F_{t^k}(x_1^k)-F(x_1^k)\|^2|\mathcal{F}^{k-1}]+\mathbb{E}[\|F_{t^k}(x_2^k)-F(x_2^k)\|^2|\mathcal{F}^{k-1}]\\
&\le 2\mathbb{E}[\|F_{t^k}(x_1^k)-\mathbb{E}[F_{t^k}(x_1^k)]\|^2|\mathcal{F}^{k-1}]+2\mathbb{E}[\|\mathbb{E}[F_{t^k}(x_1^k)]-F(x_1^k)\|^2|\mathcal{F}^{k-1}]\\
&\quad+2\mathbb{E}[\|F_{t^k}(x_2^k)-\mathbb{E}[F_{t^k}(x_2^k)]\|^2|\mathcal{F}^{k-1}]+2\mathbb{E}[\|\mathbb{E}[F_{t^k}(x_2^k)]-F(x_2^k)\|^2|\mathcal{F}^{k-1}]\\
&\le\frac{4\bar{\sigma}_F^2}{n_F^k}+4(\mu_F^k)^2. \qquad (2.13)
\end{aligned}$$

Similarly, we can obtain

$$\mathbb{E}[(\mathcal{E}_G^k(x^k))^2|\mathcal{F}^{k-1}]\le\frac{4\bar{\sigma}_G^2}{n_G^k}+4(\mu_G^k)^2. \qquad (2.14)$$

By (2.13) and (2.14), we can see that the deviation of stochastic information from true values can be reduced via increasing batch sizes n_F^k and n_G^k, which plays an important role in theoretical analysis in next sections. Moreover, it implies from (2.13) and (2.14) together with Jensen's inequality that

$$\mathbb{E}[\mathcal{E}_F^k(x^k)|\mathcal{F}^{k-1}]\le(\mathbb{E}[(\mathcal{E}_F^k(x^k))^2|\mathcal{F}^{k-1}])^{1/2}\le\frac{2\bar{\sigma}_F}{(n_F^k)^{1/2}}+2\mu_F^k, \qquad (2.15)$$

$$\mathbb{E}[\mathcal{E}_G^k(x^k)|\mathcal{F}^{k-1}]\le(\mathbb{E}[(\mathcal{E}_G^k(x^k))^2|\mathcal{F}^{k-1}])^{1/2}\le\frac{2\bar{\sigma}_G}{(n_G^k)^{1/2}}+2\mu_G^k. \qquad (2.16)$$

3. Global Convergence

In this section, we will study global convergence properties of Algorithm 2.1. Let $(x^k)_k$ be generated by Algorithm 2.1. We first give two assumptions as follows.

Assumption 3.1.

(B1) *There exists a positive constant m_G such that $\|G_{s^k}(x^k)v\|\ge m_G\|v\|$ for any s^k, $k\ge 0$, and $v\in\mathbb{R}^n$.*

(B2) *There exists a positive constant M_F such that $\mathbb{E}[\|F_{t^k}(x^k)\|^2]\le M_F^2$ for t^k, $k\ge 0$.*

The following lemma provides an upper bound for the expectation of function value i.e. $\mathbb{E}[\|F(x^k)\|]$.

Lemma 3.1. *Suppose that assumptions* (A1)–(A3) *and* (B1)–(B2) *hold. Then for any $N\in\mathbb{N}_+$,*

$$\sum_{k=0}^{N}\min\{c,(1-\eta)\alpha^k\}\mathbb{E}[\|F(x^k)\|]\le\|F(x^0)\|+\sum_{k=0}^{N}(M_{c1}^k+M_{c2}^k), \qquad (3.1)$$

where

$$\begin{aligned}
M_{c1}^k&=(1-c)\mathbb{E}[\mathcal{E}_F^k(x^k)]+\mathbb{E}[\mathcal{E}_F^{k+1}(x^{k+1})]+\varepsilon^k,\\
M_{c2}^k&=(1+\eta)\alpha^k\mathbb{E}[\mathcal{E}_F^k(x^k)]+\bar{\eta}M_F\alpha^k(\mathbb{E}[(\mathcal{E}_G^k(x^k))^2])^{1/2}+\frac{1}{2}\bar{\eta}^2M_F^2L'(\alpha^k)^2
\end{aligned}$$

with $\bar{\eta}=(1+\eta)/m_G$.

Proof. Following from the algorithmic framework, we know that at kth iteration, the iterate x^{k+1} can be either $x^k + d^k$ satisfying (2.6) or otherwise, $x^k + \alpha^k d^k$. If the former case happens, by the definition of $\mathcal{E}_F^k(x)$ in (2.12), we have

$$\|F_{t^k}(x^k)\| \le \|F(x^k)\| + \mathcal{E}_F^k(x^k),$$

which implies that

$$\begin{aligned}\|F(x^{k+1})\| - (1-c)\|F(x^k)\| \le \|F_{t^{k+1}}(x^{k+1})\| + \mathcal{E}_F^{k+1}(x^{k+1}) \\ - (1-c)\|F_{t^k}(x^k)\| + (1-c)\mathcal{E}_F^k(x^k).\end{aligned}$$

Then by (2.6), it is easy to have

$$\|F(x^{k+1})\| - \|F(x^k)\| \le -c\|F(x^k)\| + (1-c)\mathcal{E}_F^k(x^k) + \mathcal{E}_F^{k+1}(x^{k+1}) + \varepsilon^k. \tag{3.2}$$

If (2.6) does not hold, $x^{k+1} = x^k + \alpha^k d^k$. Then it follows from (2.11) that

$$\begin{aligned}&\|F(x^{k+1})\| - \|F(x^k)\| \\ &\le \|F(x^k) + \alpha^k \mathcal{J}F(x^k)d^k\| + \frac{(\alpha^k)^2 L'}{2}\|d^k\|^2 - \|F(x^k)\| \\ &\le (1-\alpha^k)\|F(x^k)\| + \alpha^k \mathcal{E}_F^k(x^k) + \alpha^k\|F_{t^k}(x^k) + \mathcal{J}F(x^k)d^k\| + \frac{(\alpha^k)^2 L'}{2}\|d^k\|^2 - \|F(x^k)\| \\ &\le \alpha^k(1+\eta^k)\mathcal{E}_F^k(x^k) + \alpha^k(\eta^k - 1)\|F(x^k)\| + \alpha^k \mathcal{E}_G^k(x^k)\|d^k\| + \frac{(\alpha^k)^2 L'}{2}\|d^k\|^2,\end{aligned} \tag{3.3}$$

where we use (2.4) in the last inequality. Notice that by (2.4) and $\eta^k \le \eta$ it is easy to obtain

$$\begin{aligned}\|d^k\| &\le \|(G_{s^k}(x^k))^{-1}\|\|G_{s^k}(x^k)d^k\| \\ &\le (1+\eta^k)\|G_{s^k}(x^k)^{-1}\|\|F_{t^k}(x^k)\| \\ &\le \bar{\eta}\|F_{t^k}(x^k)\|.\end{aligned} \tag{3.4}$$

Then substituting (3.4) into (3.3) yields

$$\begin{aligned}\|F(x^{k+1})\| - \|F(x^k)\| \le \alpha^k(1+\eta)\mathcal{E}_F^k(x^k) - \alpha^k(1-\eta)\|F(x^k)\| \\ + \alpha^k\bar{\eta}\|F_{t^k}(x^k)\|\mathcal{E}_G^k(x^k) + \frac{1}{2}(\alpha^k)^2\bar{\eta}^2 L'\|F_{t^k}(x^k)\|^2.\end{aligned} \tag{3.5}$$

Now we combine above two cases together. It follows from (3.2) and (3.5) that

$$\|F(x^{k+1})\| - \|F(x^k)\| \le -\min\{c, \alpha^k(1-\eta)\}\|F(x^k)\| + \max\{M_1^k, M_2^k\}, \tag{3.6}$$

where

$$\begin{aligned}M_1^k &= (1-c)\mathcal{E}_F^k(x^k) + \mathcal{E}_F^{k+1}(x^{k+1}) + \varepsilon^k, \\ M_2^k &= \alpha^k(1+\eta)\mathcal{E}_F^k(x^k) + \alpha^k\bar{\eta}\|F_{t^k}(x^k)\|\mathcal{E}_G^k(x^k) + \frac{1}{2}(\alpha^k)^2\bar{\eta}^2 L'\|F_{t^k}(x^k)\|^2.\end{aligned} \tag{3.7}$$

By summing up (3.6) over $k = 0, \ldots, N$, we obtain

$$\|F(x^{N+1})\| - \|F(x^0)\| \le -\sum_{k=0}^{N} \min\{c, (1-\eta)\alpha^k\}\|F(x^k)\| + \sum_{k=0}^{N} \max\left\{M_1^k, M_2^k\right\},$$

which yields

$$\sum_{k=0}^{N}\min\{c,(1-\eta)\alpha^k\}\|F(x^k)\| \le \|F(x^0)\| - \|F(x^{N+1})\| + \sum_{k=0}^{N}\max\left\{M_1^k, M_2^k\right\}.$$

Taking expectation on both sides of above inequality indicates

$$\sum_{k=0}^{N}\min\{c,(1-\eta)\alpha^k\}\mathbb{E}[\|F(x^k)\|] \le \|F(x^0)\| + \sum_{k=0}^{N}(\mathbb{E}[M_1^k] + \mathbb{E}[M_2^k]),$$

where the expectation is taken w.r.t. all the random variables generated in all N iterations. Thus, (3.1) is obtained by $\mathbb{E}[M_1^k] = M_{c1}^k$ and $\mathbb{E}[M_2^k] \le M_{c2}^k$ from assumption (B2) and

$$\mathbb{E}[\|F_{t^k}(x^k)\|\mathcal{E}_k^G(x^k)] \le (\mathbb{E}[\|F_{t^k}(x^k)\|^2])^{1/2}(\mathbb{E}[(\mathcal{E}_k^G(x^k))^2])^{1/2} \le M_F(\mathbb{E}[(\mathcal{E}_k^G(x^k))^2])^{1/2}. \quad \square$$

The next theorem shows the global convergence of Algorithm 2.1, when α^k satisfies the small step size policy (3.8).

Theorem 3.1. *Suppose that assumptions* (A1)–(A3) *and* (B1)–(B2) *hold. If step sizes* $\{\alpha^k\}_k$ *satisfy conditions*

$$\sum_k \alpha^k = +\infty, \qquad \sum_k (\alpha^k)^2 < +\infty, \tag{3.8}$$

and $(\varepsilon^k)_k,\ (\mu_F^k)_k,\ ((\mu_G^k)^2)_k,\ ((n_F^k)^{-1/2})_k,\ ((n_G^k)^{-1})_k \in \ell_+^1$ *, then we have*

$$\liminf_{k\to\infty}\mathbb{E}[\|F(x^k)\|] = 0 \ \textit{and}\ \liminf_{k\to\infty} F(x^k) = 0 \ \textit{a.s.}.$$

Proof. By (2.13)–(2.14) and $(\mu_F^k)_k,\ (\mu_G^k)_k,\ ((n_F^k)^{-1/2})_k,\ ((n_G^k)^{-1})_k \in \ell_+^1$, we obtain

$$\sum_k \mathbb{E}[\mathcal{E}_F^k(x^k)] < +\infty \text{ and } \sum_k \mathbb{E}[(\mathcal{E}_k^G(x^k))^2] < +\infty.$$

Then it derives from (3.8) that

$$\sum_k \alpha_k\mathbb{E}[\mathcal{E}_k^F(x^k)] \le \left(\sum_k \alpha_k^2\right)^{1/2}\left(\sum_k \mathbb{E}[(\mathcal{E}_k^F(x^k))^2]\right)^{\frac{1}{2}} < +\infty,$$
$$\sum_k \alpha_k(\mathbb{E}[(\mathcal{E}_k^G(x^k))^2])^{1/2} \le \left(\sum_k \alpha_k^2\right)^{1/2}\left(\sum_k \mathbb{E}[(\mathcal{E}_k^G(x^k))^2]\right)^{1/2} < +\infty.$$

Hence, (3.1) implies

$$\sum_k \min\{c,(1-\eta)\alpha^k\}\mathbb{E}[\|F(x^k)\|] < +\infty,$$

which further yields from (3.8) that $\liminf_{k\to\infty}\mathbb{E}[\|F(x^k)\|] = 0$. Moreover, it follows from Fatou's lemma that

$$\mathbb{E}\left[\sum_{k=0}^{\infty}\alpha^k\|F(x^k)\|\right] \le \liminf_{N\to\infty}\mathbb{E}[\sum_{k=0}^{N}\alpha^k\|F(x^k)\|] < \infty.$$

It indicates $\sum_{k=1}^{\infty}\alpha^k\|F(x^k)\| < \infty$ with probability 1 thus $\liminf_{k\to\infty} F(x^k) = 0$ almost surely. $\square$

We next consider another assumption different from (B2).

Assumption 3.2.

(B2') *There exists a positive constant* $\bar{M}_F$ *such that* $\|F_{t^k}(x^k)\| \leq \bar{M}_F$ *holds almost surely for any* k *and* t^k.

Different from Lemma 3.1, under assumption (B2') we obtain the following lemma showing a different upper bound on expected function values.

Lemma 3.2. *Suppose that assumptions* (A1)–(A3), (B1) *and* (B2') *hold. If the step size* α^k *satisfies* $\alpha^k \leq c/(1-\eta)$ *for all* $k \geq 0$, *then for any* $N \in \mathbb{N}_+$,

$$\begin{aligned}&\sum_{k=0}^{N} \alpha^k (1-\eta-\bar{\eta}(\frac{2\bar{\sigma}_G}{(n_G^k)^{1/2}}+2\mu_G^k)-\frac{1}{2}\alpha^k L'\bar{M}_F\bar{\eta}^2)\mathbb{E}[\|F(x^k)\|]\\ \leq&\|F(x^0)\|+\sum_{k=0}^{N}(\mathbb{E}[M_{c1}^k]+\mathbb{E}[M_{c3}^k]),\end{aligned} \tag{3.9}$$

where M_{c1}^k *is defined in Lemma* 3.1 *and*

$$M_{c3}^k = \left[(1+\eta)\alpha^k + 2\alpha^k\bar{\eta}\left(\frac{\bar{\sigma}_G}{(n_G^k)^{1/2}}+\mu_G^k\right)+\frac{1}{2}(\alpha^k)^2\bar{\eta}^2 L'\bar{M}_F\right]\left(\mathbb{E}[(\mathcal{E}_F^k(x^k))^2]\right)^{1/2}.$$

Proof. From the proof of Lemma 3.1, (3.2) holds if the line search condition (2.6) is satisfied. Otherwise, we have (3.3). Note that it yields from (3.4) and the definition of $\mathcal{E}_F^k(x^k)$ that

$$\begin{aligned}&\|d^k\| \leq \bar{\eta}\|F_{t^k}(x^k)\| \leq \bar{\eta}(\|F(x^k)\|+\mathcal{E}_F^k(x^k)),\\ &\|d^k\|^2 \leq \bar{\eta}^2\|F_{t^k}(x^k)\|^2 \leq \bar{\eta}^2\|F_{t^k}(x^k)\|(\|F(x^k)\|+\mathcal{E}_F^k(x^k)).\end{aligned}$$

Then (3.3) together with $\eta^k \leq \eta$ yields

$$\begin{aligned}&\|F(x^{k+1})\|-\|F(x^k)\|\\ \leq&-\alpha^k\left(1-\eta-\bar{\eta}\mathcal{E}_G^k(x^k)-\frac{1}{2}\alpha^k L'\bar{\eta}^2\|F_{t^k}(x^k)\|\right)\|F(x^k)\|\\ &+\alpha^k(1+\eta)\mathcal{E}_F^k(x^k)+\alpha^k\bar{\eta}\mathcal{E}_F^k(x^k)\mathcal{E}_G^k(x^k)+\frac{(\alpha^k)^2L'\bar{\eta}^2}{2}\mathcal{E}_F^k(x^k)\|F_{t^k}(x^k)\|.\end{aligned} \tag{3.10}$$

Hence, it follows from (3.2), (3.10) and assumption (B2') that

$$\begin{aligned}&\|F(x^{k+1})\|-\|F(x^k)\|\\ \leq&\left(-\min\{c,\alpha^k(1-\eta)\}+\alpha^k\bar{\eta}\mathcal{E}_G^k(x^k)+\frac{1}{2}(\alpha^k)^2L'\bar{M}_F\bar{\eta}^2\right)\|F(x^k)\|+\max\{M_1^k,M_3^k\}\\ \leq&-\alpha^k\left(1-\eta-\bar{\eta}\mathcal{E}_G^k(x^k)-\frac{1}{2}\alpha^k L'\bar{M}_F\bar{\eta}^2\right)\|F(x^k)\|+\max\{M_1^k,M_3^k\}\end{aligned} \tag{3.11}$$

holds almost surely, where the second inequality is due to $\alpha^k \leq c/(1-\eta)$, M_1^k is defined in (3.7) and

$$M_3^k = \alpha^k(1+\eta)\mathcal{E}_F^k(x^k)+\alpha^k\bar{\eta}\mathcal{E}_F^k(x^k)\mathcal{E}_G^k(x^k)+\frac{1}{2}(\alpha^k)^2\bar{\eta}^2L'\bar{M}_F\mathcal{E}_F^k(x^k).$$

Then summing up (3.11) over $k=0,\cdots,N$ on both sides implies that

$$\sum_{k=0}^{N}\alpha^k\left(1-\eta-\bar{\eta}\mathcal{E}_G^k(x^k)-\frac{1}{2}\alpha^kL'\bar{M}_F\bar{\eta}^2\right)\|F(x^k)\| \leq \|F(x^0)\|+\sum_{k=0}^{N}\left(M_1^k+M_3^k\right)$$

holds almost surely, which leads to

$$\mathbb{P}\left(\sum_{k=0}^{N}\alpha^k\left(1-\eta-\bar{\eta}\mathcal{E}_G^k(x^k)-\frac{1}{2}\alpha^k L'\bar{M}_F\bar{\eta}^2\right)\|F(x^k)\|\le\|F(x^0)\|+\sum_{k=0}^{N}\left(M_1^k+M_3^k\right)\right)=1.$$

Note that for two random variables ξ_1 and ξ_2, if $\mathbb{P}(\xi_1-\xi_2\le 0)=1$, then $\mathbb{E}[\xi_1-\xi_2]\le 0$. Hence, we have

$$\sum_{k=0}^{N}\mathbb{E}[\alpha^k(1-\eta-\bar{\eta}\mathcal{E}_G^k(x^k)-\frac{1}{2}\alpha^k L'\bar{M}_F\bar{\eta}^2)\|F(x^k)\|]\le\|F(x^0)\|+\sum_{k=0}^{N}(\mathbb{E}[M_1^k]+\mathbb{E}[M_3^k]). \quad (3.12)$$

By (2.16) and $x^k\in\hat{\mathcal{F}}^{k-1}$ we have

$$\begin{aligned}\mathbb{E}[\mathcal{E}_G^k(x^k)\|F(x^k)\|] =&\mathbb{E}[\mathbb{E}[\mathcal{E}_G^k(x^k)\|F(x^k)\|\mid\hat{\mathcal{F}}^{k-1}]]\\ =&\mathbb{E}[\mathbb{E}[\mathcal{E}_G^k(x^k)\mid\hat{\mathcal{F}}^{k-1}]\|F(x^k)\|]\\ \le&\left(\frac{2\bar{\sigma}_G}{(n_G^k)^{1/2}}+2\mu_G^k\right)\mathbb{E}\left[\|F(x^k)\|\right].\end{aligned}$$

Then (3.12) yields

$$\begin{aligned}&\sum_{k=0}^{N}\alpha^k\left(1-\eta-\bar{\eta}\left(\frac{2\bar{\sigma}_G}{(n_G^k)^{1/2}}+2\mu_G^k\right)-\frac{1}{2}\alpha^k L'\bar{M}_F\bar{\eta}^2\right)\mathbb{E}[\|F(x^k)\|]\\ \le&\|F(x^0)\|+\sum_{k=0}^{N}\left(\mathbb{E}[M_1^k]+\mathbb{E}[M_3^k]\right).\end{aligned} \quad (3.13)$$

Notice that for any $k\ge 0$,

$$\begin{aligned}&\mathbb{E}[\mathcal{E}_F^k(x^k)\mathcal{E}_G^k(x^k)]\left(\mathbb{E}[(\mathcal{E}_F^k(x^k))^2]\right)^{1/2}\left(\mathbb{E}[(\mathcal{E}_G^k(x^k))^2]\right)^{1/2}\\ \le&\left(\frac{2\bar{\sigma}_G}{(n_G^k)^{1/2}}+2\mu_G^k\right)\left(\mathbb{E}[(\mathcal{E}_F^k(x^k))^2]\right)^{1/2},\end{aligned}$$

then

$$\begin{aligned}\mathbb{E}[M_3^k]=&(1+\eta)\alpha^k\mathbb{E}[\mathcal{E}_k^F(x^k)]+\alpha^k\bar{\eta}\mathbb{E}\left[\mathcal{E}_F^k(x^k)\mathcal{E}_G^k(x^k)\right]\\ &+\frac{1}{2}(\alpha^k)^2\bar{\eta}^2L'\bar{M}_F\mathbb{E}[\mathcal{E}_F^k(x^k)]\\ \le&\left((1+\eta)\alpha^k+2\alpha^k\bar{\eta}(\frac{\bar{\sigma}_G}{(n_G^k)^{1/2}}+\mu_G^k)+\frac{1}{2}(\alpha^k)^2\bar{\eta}^2L'\bar{M}_F\right)\\ &\times\left(\mathbb{E}[(\mathcal{E}_F^k(x^k))^2]\right)^{1/2}=M_{c3}^k.\end{aligned} \quad (3.14)$$

Consequently, plugging $\mathbb{E}[M_1^k]=M_{c1}^k$ and (3.14) into (3.13) yields the conclusion. □

Based on Lemma 3.2, Algorithm 2.1 can achieve global convergence when α^k is constant.

Theorem 3.2. *Suppose that assumptions* (A1)–(A3), (B1) *and* (B2') *hold. Further suppose that for any $k\ge 0$, the step size $(\alpha^k)_k$ satisfies*

$$\alpha^k=\bar{\alpha}\le\theta:=\min\left\{\frac{c}{1-\eta},\frac{2(1-\eta)}{3L'\bar{M}_F\bar{\eta}^2}\right\}, \quad (3.15)$$

and $\mu_G^k,\ n_G^k$ *satisfy*

$$\frac{2\bar{\sigma}_G}{(n_G^k)^{1/2}} + 2\mu_G^k \leq \frac{1-\eta}{3\bar{\eta}}. \tag{3.16}$$

If $(\varepsilon^k)_k, (\mu_F^k)_k, ((n_F^k)^{-1/2})_k \in \ell_+^1$, *then*

$$\lim_{k\to\infty} \mathbb{E}[\|F(x^k)\|] = 0 \quad \textit{and} \quad \lim_{k\to\infty} F(x^k) = 0 \quad \textit{a.s.}.$$

Proof. It follows from (3.9) and assumptions that

$$\sum_k \alpha^k \left(1 - \eta - \bar{\eta}\left(\frac{2\bar{\sigma}_G}{(n_G^k)^{1/2}} + 2\mu_G^k\right) - \frac{1}{2}\alpha^k L' \bar{M}_F \bar{\eta}^2\right) \mathbb{E}[\|F(x^k)\|] < +\infty.$$

Then it further yields from (3.15) and (3.16) that

$$\sum_k \frac{1}{3}\bar{\alpha}(1-\eta)\mathbb{E}[\|F(x^k)\|] < +\infty,$$

which implies $\lim_{k\to\infty} \mathbb{E}[\|F(x^k)\|] = 0$. Similar to the analysis of Theorem 3.1, we obtain that $\lim_{k\to\infty} F(x^k) = 0$ almost surely. □

Remark 3.1. From (3.4) we can see that Assumption B1 plays a crucial role in both Theorems 3.1 and 3.2 for proving the global convergence in expectation and almost-surely. Moreover, both theorems above show that the increase of sample size plays an important role in proving the almost sure convergence of $\{\|F(x^k)\|\}$. But as shown in Theorem 3.2, under assumption (B2'), only an increase in the sample size to compute the mini-batch approximate function value $F_{t^k}(x^k)$ is required.

4. Computational Complexity

In this section, we analyze the computational complexity of Algorithm 2.1 with respect to stochastic zeroth- and first-order oracles, when its output is chosen randomly from all iterates $x^k, k = 0, \ldots, N$, where N is the maximum iteration number. In this situation, we establish several complexity results under different choices of step sizes.

Theorem 4.1. *Suppose that assumptions* (A1)–(A3) *and* (B1)–(B2) *hold. Let Algorithm* 2.1 *return* x^R *as the output, where* $R \in \{0, \ldots, N\}$ *follows the distribution function*

$$\mathrm{P}(R = k) = \frac{\min\{c, \alpha^k(1-\eta)\}}{\sum_{k=0}^N \min\{c, \alpha^k(1-\eta)\}}, \quad k = 0, \ldots, N.$$

Then we have

$$\mathbb{E}[\|F(x^R)\|] \leq \frac{\|F(x^0)\| + \sum_{k=0}^N (M_{c1}^k + M_{c2}^k)}{\sum_{k=0}^N \min\{c, \alpha^k(1-\eta)\}}, \tag{4.1}$$

where M_{c1}^k, M_{c2}^k *are defined in Lemma* 3.1. *Furthermore, we have the following results:*

(i) *Suppose that* $c + \eta \geq 1$, $\alpha^0 = \gamma > 0$, $\alpha^k = k^{-\beta}$ *with* $\beta \in (0.5, 1)$, $k = 1, \ldots, N$ *and* ε^k, μ_F^k, μ_G^k, $(n_F^k)^{-1/2}$, $(n_G^k)^{-1} = \mathcal{O}(k^{-\delta})$ *with* $\delta > 1$. *Then,* $\mathbb{E}[\|F(x^R)\|] = \mathcal{O}(N^{\beta-1})$. *Thus to achieve* $\mathbb{E}[\|F(x^R)\|] < \epsilon$ *for given* $\epsilon > 0$, *the total number of evaluations of zeroth- and first-order oracles are in order of* $\mathcal{O}(\epsilon^{\frac{1+2\delta}{\beta-1}})$ *and* $\mathcal{O}(\epsilon^{\frac{1+\delta}{\beta-1}})$, *respectively.*

(ii) *Suppose that* $N > (\frac{1-\eta}{c})^2$, $\alpha^k = \alpha$, $k = 0, 1, \ldots, N$ *and* ε^k, μ_F^k, μ_G^k, $(n_F^k)^{-1/2}$, $(n_G^k)^{-1} = \mathcal{O}(k^{-\delta})$ *with* $\delta > 0$. *Then the following conclusions hold.*

(1) *If* $\delta > 1$ *and* $\alpha = N^{-1/2}$, *we obtain* $\mathbb{E}[\|F(x^R)\|] = \mathcal{O}(N^{-1/2})$. *Thus to achieve* $\mathbb{E}[\|F(x^R)\|] \leq \epsilon$ *for given* $\epsilon > 0$, *the total number of evaluations of zeroth- and first-order oracles are in order of* $\mathcal{O}(\epsilon^{-(2+4\delta)})$ *and* $\mathcal{O}(\epsilon^{-(2+2\delta)})$, *respectively.*

(2) *If* $\delta < 1$ *and* $\alpha = N^{-\delta/2}$, *we obtain* $\mathbb{E}[\|F(x^R)\|] = \mathcal{O}(N^{-\delta/2})$. *Thus to achieve* $\mathbb{E}[\|F(x^R)\|] \leq \epsilon$ *for given* $\epsilon > 0$, *the total number of evaluations of zeroth- and first-order oracles are in order of* $\mathcal{O}(\epsilon^{-(4+\frac{2}{\delta})})$ *and* $\mathcal{O}(\epsilon^{-(2+\frac{2}{\delta})})$, *respectively.*

(3) *If* $\delta = 1$ *and* $\alpha = (\frac{N}{\log N})^{-1/2}$, *we obtain* $\mathbb{E}[\|F(x^R)\|] = \mathcal{O}((\frac{N}{\log N})^{-1/2})$. *Thus to achieve* $\mathbb{E}[\|F(x^R)\|] \leq \epsilon$ *for given* $\epsilon > 0$, *the total number of evaluations of zeroth- and first-order oracles are in order of* $\tilde{\mathcal{O}}(\epsilon^{-6})$ *and* $\tilde{\mathcal{O}}(\epsilon^{-4})$, *respectively. Here we use* $\tilde{O}$ *to hide the dependence on logarithmic factors.*

Proof. Due to the randomness of R, it is straightforward to obtain (4.1) from Lemma 3.1 and

$$\mathbb{E}[\|F(x^R)\|] = \frac{\sum_{k=0}^N \min\{c, (1-\eta)\alpha^k\}\mathbb{E}[\|F(x^k)\|]}{\sum_{k=0}^N \min\{c, \alpha^k(1-\eta)\}}.$$

We first prove part (i). It implies from $c + \eta \geq 1$ and (4.1) that

$$\begin{aligned}
&\mathbb{E}[\|F(x^R)\|] \\
&\leq \frac{1}{(1-\eta)(\sum_{k=0}^N \alpha^k)}(\|F(x^0)\| + \sum_{k=0}^N((1-c)\mathbb{E}[\mathcal{E}_F^k(x^k)] + \mathbb{E}[\mathcal{E}_{k+1}^F(x^{k+1})] + \varepsilon^k) \\
&\quad + \sum_{k=0}^N \alpha^k[(1+\eta)\mathbb{E}[\mathcal{E}_F^k(x^k)] + \bar{\eta}M_F(\mathbb{E}[(\mathcal{E}_G^k(x^k))^2])^{1/2}] + \frac{1}{2}\bar{\eta}^2 M_F^2 L' \sum_{k=0}^N (\alpha^k)^2). \qquad (4.2)
\end{aligned}$$

Note that

$$\sum_{k=0}^N \alpha^k \geq \gamma + \frac{(N+1)^{1-\beta} - 1}{1-\beta} \quad \text{and} \quad \sum_{k=0}^N (\alpha^k)^2 \leq \frac{2\beta}{2\beta - 1} + \gamma^2.$$

Moreover, it follows from $(\mu_F^k)_k$, $((n_F^k)^{-1/2})_k \in \ell_+^1$ that $\sum_{k=0} \mathbb{E}[\mathcal{E}_F^k(x^k)] = \mathcal{O}(1)$. In addition, by $(\mu_G^k)_k^2$, $((n_G^k)^{-1})_k \in \ell_+^1$, we have

$$\sum_{k=0}^N \alpha^k \left(\mathbb{E}[(\mathcal{E}_k^G(x^k))^2]\right)^{1/2} \leq \left(\sum_{k=0}^N (\alpha^k)^2\right)^{1/2} \left(\sum_{k=0}^N \mathbb{E}[(\mathcal{E}_k^G(x^k))^2]\right)^{1/2} = \mathcal{O}(1),$$

which implies that $\mathbb{E}[\|F(x^R)\|]$ is in the order of $\mathcal{O}(N^{\beta-1})$. Therefore, to achieve $\mathbb{E}[\|F(x^R)\|] \leq \epsilon$, N should be in order of $\mathcal{O}(\epsilon^{\frac{1}{\beta-1}})$. Then the total number of evaluations of zeroth- and first-order oracles are

$$\sum_{k=0}^N n_F^k = \mathcal{O}(N^{1+2\delta}) = \mathcal{O}\left(\epsilon^{\frac{1+2\delta}{\beta-1}}\right) \quad \text{and} \quad \sum_{k=0}^N n_G^k = \mathcal{O}(N^{1+\delta}) = \mathcal{O}\left(\epsilon^{\frac{1+\delta}{\beta-1}}\right),$$

respectively.

Next we prove part (ii). For (1), it implies from $N > (\frac{1-\eta}{c})^2$, (4.1) and $\alpha = \frac{1}{\sqrt{N}}$ that

$$\begin{aligned}\mathbb{E}[\|F(x^R)\|] \leq \frac{1}{(1-\eta)N\alpha}\Bigg(\|F(x^0)\| + \sum_{k=0}^{N}\Bigg((1-c)\mathbb{E}[\mathcal{E}_F^k(x^k)] + \mathbb{E}[\mathcal{E}_{k+1}^F(x^{k+1})] + \varepsilon^k \\ + \alpha[(1+\eta)\mathbb{E}[\mathcal{E}_F^k(x^k)] + \bar{\eta}M_F(\mathbb{E}[(\mathcal{E}_G^k(x^k))^2])^{1/2}] + \frac{1}{2}\bar{\eta}^2M_F^2L'\alpha^2\Bigg)\Bigg) \\ = \frac{1}{(1-\eta)\sqrt{N}}\Bigg(\|F(x^0)\| + \sum_{k=0}^{N}\Bigg((1-c)\mathbb{E}[\mathcal{E}_F^k(x^k)] + \mathbb{E}[\mathcal{E}_{k+1}^F(x^{k+1})] + \varepsilon^k \\ + \frac{1}{\sqrt{N}}(1+\eta)\mathbb{E}[\mathcal{E}_F^k(x^k)] + \frac{1}{2}\mathbb{E}[(\mathcal{E}_G^k(x^k))^2] + \frac{\bar{\eta}^2M_F^2(L'+1)}{2N}\Bigg)\Bigg).\end{aligned}$$

Moreover, similar to part (i), we have $\sum_{k=0}\mathbb{E}[\mathcal{E}_F^k(x^k)] = \mathcal{O}(1)$ and $\sum_{k=0}\mathbb{E}[(\mathcal{E}_k^G(x^k))^2] = \mathcal{O}(1)$. Then $\mathbb{E}[\|F(x^R)\|]$ is in the order of $O(N^{-1/2})$. Therefore, to achieve $\mathbb{E}[\|F(x^R)\|] \leq \epsilon$, N should be in order of $\mathcal{O}(\epsilon^{-2})$. Then the total number of evaluations of zeroth- and first-order oracles are

$$\sum_{k=0}^{N} n_F^k = \mathcal{O}(N^{1+2\delta}) = \mathcal{O}\big(\epsilon^{-(2+4\delta)}\big) \quad \text{and} \quad \sum_{k=0}^{N} n_G^k = \mathcal{O}(N^{1+\delta}) = \mathcal{O}\big(\epsilon^{-(2+2\delta)}\big),$$

respectively.

To prove (2), we first derive from $\delta < 1$ that

$$\sum_{k=0}^{N}\mathbb{E}[\mathcal{E}_k^F(x^k)] = \mathcal{O}(N^{1-\delta}) \quad \text{and} \quad \sum_{k=0}^{N}\mathbb{E}[(\mathcal{E}_G^k(x^k))^2] = \mathcal{O}(N^{1-\delta}).$$

Then it yields from (4.2) and $\alpha = \mathcal{O}(N^{-\delta/2})$ that $\mathbb{E}[\|F(x^R)\|] = \mathcal{O}(N^{-\delta/2})$. Therefore, to achieve $\mathbb{E}[\|F(x^R)\|] \leq \epsilon$, N should be in order of $\mathcal{O}(\epsilon^{-2/\delta})$. Then the total number of evaluations of zeroth- and first-order oracles are

$$\sum_{k=0}^{N} n_F^k = \mathcal{O}(N^{1+2\delta}) = \mathcal{O}(\epsilon^{-(4+\frac{2}{\delta})}) \quad \text{and} \quad \sum_{k=0}^{N} n_G^k = \mathcal{O}(N^{1+\delta}) = \mathcal{O}(\epsilon^{-(2+\frac{2}{\delta})}),$$

respectively. We now prove (3). Notice that in this case

$$\sum_{k=0}^{N}\mathbb{E}[\mathcal{E}_k^F(x^k)] = \mathcal{O}(\log N) \quad \text{and} \quad \sum_{k=0}^{N}(\mathbb{E}[(\mathcal{E}_G^k(x^k))^2])^{1/2} = \mathcal{O}(\log N).$$

Then similar to previous two cases, we can obtain from (4.2) and $\alpha = (\frac{N}{\log N})^{-1/2}$ that $\mathbb{E}[\|F(x^R)\|] = \mathcal{O}((\frac{N}{\log N})^{-1/2})$. Therefore, to achieve $\mathbb{E}[\|F(x^R)\|] \leq \epsilon$, N should be in order of $\mathcal{O}(\epsilon^{-2}\log(\epsilon^{-1}))$, thus the total number of evaluations of zeroth- and first-order oracles are

$$\sum_{k=0}^{N} n_F^k = \mathcal{O}(N^3) = \tilde{\mathcal{O}}(\epsilon^{-6}) \quad \text{and} \quad \sum_{k=0}^{N} n_G^k = \mathcal{O}(N^2) = \tilde{\mathcal{O}}(\epsilon^{-4}),$$

respectively. □

Theorem 3.2 shows that the global convergence of Algorithm 2.1 can be guaranteed when step sizes are constant. Moreover, from (3.16) we can see that it is not necessary to require

$(n_G^k)^{-1}$ and μ_G^k approach zero, provided that they are small enough. It only requires stochastic zeroth-order oracles approach the real function values gradually. Accordingly, we can establish the computational complexity of Algorithm 2.1 as follows.

Theorem 4.2. *Suppose that assumptions* (A1)–(A3), (B1) *and* (B2') *hold, and Algorithm* 2.1 *returns* x^R *as the output, where* $R \in \{1, \cdots, N\}$ *follows the distribution function*

$$\mathrm{P}(R = k) = \frac{\alpha^k(1 - \eta - \phi^k)}{\sum_{k=0}^{N} \alpha^k(1 - \eta - \phi^k)}, \quad k = 0, \ldots, N,$$

with $\phi^k = \bar{\eta}(\frac{2\bar{\sigma}_G}{(n_G^k)^{1/2}} + 2\mu_G^k) + \frac{1}{2}\alpha^k L' \bar{M}_F \bar{\eta}^2$, $(\alpha^k)_k$ *satisfying* (3.15), $n_G^k \geq \frac{144\bar{\eta}^2\bar{\sigma}_G^2}{(1-\eta)^2}$ *and* $\mu_G^k \leq \frac{1-\eta}{12\bar{\eta}}$. *Then we have*

$$\mathbb{E}[\|F(x^R)\|] \leq \frac{3\|F(x^0)\| + 3\sum_{k=0}^{N}(M_{c1}^k + M_{c3}^k)}{(1-\eta)\sum_{k=0}^{N}\alpha^k}. \tag{4.3}$$

Furthermore, if ε^k, μ_F^k, $(n_F^k)^{-1/2} = \mathcal{O}(k^{-\delta})$ *with* $\delta > 0$, *the following conclusions hold.*

(1) *If* $\delta > 1$ *and* $\alpha^k = N^{-1/2}$ *with* $N > \theta^2$ *where* θ *is defined in* (3.15), *we have* $\mathbb{E}[\|F(x^R)\|] = \mathcal{O}(N^{-1/2})$. *Thus to achieve* $\mathbb{E}[\|F(x^R)\|] \leq \epsilon$ *for a given* $\epsilon > 0$, *the total number of evaluations of zeroth- and first-order oracles are in order of* $\mathcal{O}(\epsilon^{-(2+4\delta)})$ *and* $\mathcal{O}(\epsilon^{-2})$, *respectively.*

(2) *If* $\delta < 1$ *and* $\alpha^k = N^{-\delta/2}$ *with* $N > \theta^{\frac{2}{\delta}}$, *then* $\mathbb{E}[\|F(x^R)\|] = \mathcal{O}(N^{-\delta/2})$. *Thus to achieve* $\mathbb{E}[\|F(x^R)\|] \leq \epsilon$ *for given* $\epsilon > 0$, *the total number of evaluations of zeroth- and first-order oracles are in order of* $\mathcal{O}(\epsilon^{-(4+\frac{2}{\delta})})$ *and* $\mathcal{O}(\epsilon^{-2/\delta})$, *respectively.*

(3) *If* $\delta = 1$ *and* $\alpha^k = (\frac{N}{\log N})^{-1/2}$, *then* $\mathbb{E}[\|F(x^R)\|] = \mathcal{O}((\frac{N}{\log N})^{-1/2})$. *Thus, to achieve* $\mathbb{E}[\|F(x^R)\|] \leq \epsilon$ *for given* $\epsilon > 0$, *the total number of evaluations of zeroth- and first-order oracles are in order of* $\tilde{\mathcal{O}}(\epsilon^{-6})$ *and* $\mathcal{O}(\epsilon^{-2}\log(\epsilon^{-1}))$, *respectively.*

Proof. Note that the settings of n_G^k and μ_G^k satisfy (3.16). Then in all cases the number of evaluations of first-order oracles are in order of $\mathcal{O}(N)$. The remaining of the proof is similar to Theorem 4.1. □

Remark 4.1. Theorems 4.1 and 4.2 establish computational complexities of Algorithm 2.1 to achieve $\mathbb{E}[\|F(x^R)\|] \leq \epsilon$. Note that the criticality measure is stronger than those considered in RSG [9] and SGN [29], seen from Table 1.1. This is due to assumption B1 that we can guarantee the near feasibility of the approximate solution of (1.1).

5. Local Convergence

In this part we study the local convergence properties of Algorithm 2.1. In general, we consider a single trajectory of the stochastic process $(x^k)_k$, and show that local convergence and a fast convergence rate can be achieved with high probability if the sample sizes n_F^k and n_G^k are chosen appropriately. In the following, we denote $(x^k)_k$ as the underlying stochastic

process or a trajectory generated by a single run of Algorithm 2.1. Let x^* be an accumulation point of $(x^k)_k$. To establish the local convergence properties of Algorithm 2.1, we first refer to Lemma 4.3 in [18] and Theorem 1.6 in [30] about concentration inequalities for vector- and matrix-valued martingales, respectively. For completeness, we state them as Lemma A.1 in the appendix.

Throughout this section, we suppose that

$$(\delta^k)_k \subseteq (0, 1/2), \quad 0 < (\varepsilon^k)_k \in \ell^1_+,$$

$(\alpha^k)_k$ satisfy (3.15) and $(\mu_G^k)_k, (n_G^k)_k$ satisfy (3.16).

The following theorem shows the convergence of whole sequence $(x^k)_k$ to x^* in certain probability.

Theorem 5.1. *Under assumptions* (A1)–(A3), (B1) *and* (B2'), *suppose that $\mathcal{J}F(x^*)$ is nonsingular and there exists $\bar{l} \in \mathbb{N}$ such that $\mu_F^k = 0$ and $n_F^k \geq \frac{\bar{\sigma}_F^2}{(\varepsilon^k)^2\delta^k}$ for $k \geq \bar{l}$. Then with probability no less than $\delta_1^* = \prod_{k=\bar{l}}^{\infty}(1-2\delta^k)$, the whole sequence $(x^k)_k$ converges to x^* and x^* is a solution of* (1.1).

Proof. Recall that by (2.8) and framework of Algorithm 2.1, at kth iteration x^k is x_1^k when the line search condition is satisfied and x_2^k, otherwise. Given $\varepsilon > 0$, we define the the following events

$$\begin{aligned}
&\mathbf{E}_F^k(\varepsilon) = \{\omega \in \Omega : \mathcal{E}_F^k(x^k(\omega) \leq \varepsilon\},\\
&\mathbf{E}_1^k(\varepsilon) = \{\omega \in \Omega : \mathcal{E}_F^k(x_1^k(\omega) \leq \varepsilon\} \quad \text{and} \quad \mathbf{E}_2^k(\varepsilon) = \{\omega \in \Omega : \mathcal{E}_k^F(x_2^k(\omega)) \leq \varepsilon\}.
\end{aligned} \tag{5.1}$$

Notice that $x_1^k,\ x_2^k \in \mathcal{F}^{k-1}$ and $F(x_j^k, t_i^k) - F(x_j^k) \in \hat{\mathcal{F}}^{k-1}, i \in [n_F^k],\ j = 1, 2$. Letting

$$X_i^k = F(x_1^k, t_i^k) - F(x_1^k), \quad i = 1, \ldots, n_F^k,$$

we obtain from assumption (A3) and $\mu_F^k = 0$, $k \geq \bar{l}$, that for any $k \geq \bar{l}$,

$$\mathbb{E}[X_i^k | \mathcal{F}^{k-1}] = 0, \quad \mathbb{E}[\|X_i^k\|^2 | \mathcal{F}^{k-1}] \leq \bar{\sigma}_F^2.$$

Then it follows from Lemma A.1 (i) that

$$\begin{aligned}
&\mathbb{P}(\|F_{t^k}(x_1^k) - F(x_1^k))\| \geq (\delta^k n_F^k)^{-1/2}\bar{\sigma}_F \ |\mathcal{F}^{k-1})\\
=&\,\mathbb{P}\left(\left\|\frac{1}{n_F^k}\sum_{i=1}^{n_F^k} X_i^k\right\| \geq (\delta^k n_F^k)^{-1/2}\bar{\sigma}_F \ |\mathcal{F}^{k-1}\right)\\
=&\,\mathbb{P}\left(\left\|\sum_{i=1}^{n_F^k} X_i^k\right\| \geq (\delta^k)^{-1/2}(n_F^k)^{1/2}\bar{\sigma}_F \ |\mathcal{F}^{k-1}\right) \leq \delta^k,
\end{aligned}$$

which further yields from $n_F^k \geq \frac{\bar{\sigma}_F^2}{(\varepsilon^k)^2\delta^k}$ for all $k \geq \bar{l}$ that

$$\mathbb{P}(\|F_{t^k}(x_1^k) - F(x_1^k)\| \leq \varepsilon^k \ |\mathcal{F}^{k-1}) \geq 1 - \delta^k.$$

Similarly, we obtain

$$\mathbb{P}(\|F_{t^k}(x_2^k) - F(x_2^k)\| \leq \varepsilon^k \ |\mathcal{F}^{k-1}) \geq 1 - \delta^k$$

for all $k \geq \bar{l}$. Then by Bonferroni inequality it yields that for any $k \geq \bar{l}$,

$$\mathbb{P}(\mathbf{E}_1^k(\varepsilon^k) \cap \mathbf{E}_2^k(\varepsilon^k))|\mathcal{F}^{k-1}) \geq \mathbf{1}_{\mathbf{E}_1^k(\varepsilon^k)}(\omega) + \mathbf{1}_{\mathbf{E}_2^k(\varepsilon^k)}(\omega) - 1 \geq 1 - 2\delta^k.$$

Note that from $\mathcal{E}_F^k(x_j^k) = \|F_{t^k}(x^k) - F(x^k)\| \leq \varepsilon^k$, $j = 1, 2$, and (2.10) we have

$$\begin{aligned}
&\mathcal{E}_F^k(x^k) = \|F_{t^k}(x^k) - F(x^k)\| \\
=& \|Y^k(F_{t^k}(x_1^k) - F(x_1^k)) + (1 - Y^k)(F_{t^k}(x_2^k) - F(x_2^k))\| \\
\leq& Y^k \mathcal{E}_F^k(x_1^k) + (1 - Y^k)\mathcal{E}_F^k(x_2^k) \leq \varepsilon^k,
\end{aligned}$$

which yields

$$\mathbb{P}(\mathbf{E}_F^k(\varepsilon^k) \ |\mathcal{F}^{k-1}) \geq \mathbb{P}(\mathbf{E}_1^k(\varepsilon^k) \cap \mathbf{E}_2^k(\varepsilon^k)|\mathcal{F}^{k-1}) \geq 1 - 2\delta^k \tag{5.2}$$

for all $k \geq \bar{l}$. Thus it derives

$$\mathbb{P}\left(\bigcap_{k=\bar{l}}^{L} \mathbf{E}_F^k(\varepsilon^k)\right) \geq \prod_{k=\bar{l}}^{L}(1 - 2\delta^k).$$

Letting $L \to \infty$ and defining event $\mathbf{E} = \bigcap_{k=\bar{l}}^{\infty} \mathbf{E}_F^k(\varepsilon^k)$, we have

$$\mathbb{P}(\mathbf{E}) \geq \prod_{k=\bar{l}}^{\infty}(1 - 2\delta^k) = \delta_1^*.$$

Therefore, we can assume that the trajectory $(x^k)_k$ is generated by a sample point $\bar{\omega} \in \mathbf{E}$, which occurs with probability no less than δ_1^*. Note that the bound of n_F^k ensures that $(n_F^k)^{-1/2} \in \ell_+^1$. Then conditions of Theorem 3.2 are satisfied, which yields $F(x^k) \to 0$ in probability one conditioned on $\mathbf{E}$. Therefore, $F(x^k) \to 0$ and x^* is a solution of (1.1) with probability no less than δ^*. Moreover, as $\mathcal{J}F(x^*)$ is nonsingular, there exists a neighborhood of x^*, denoted by $B(x^*)$, and a positive number m_J, such that for $\|\mathcal{J}F(x)v\| \geq m_J\|v\|$ for any $x \in B(x^*)$ and $v \in \mathbb{R}^n$. Then for any $x \in B(x^*)$ and $x \neq x^*$, there exists $\xi_x \in (x^*, x) \subset B(x^*)$ such that

$$\|F(x)\| = \|F(x) - F(x^*)\| = \|\mathcal{J}F(\xi_x)(x - x^*)\| \geq m_J\|x - x^*\| > 0,$$

which indicates that x^* is an isolated solution of (1.1). Furthermore, it implies from (2.10) and

$$\begin{aligned}
&\|x^{k+1} - x^k\| = \|(Y^{k+1} + (1 - Y^{k+1})\alpha_k)d^k\| \\
\leq& (1 + \eta^k)\|G_{s^k}^{-1}(x^k)\|\|F_{t_k}(x^k)\| \leq \frac{1 + \eta^k}{m_G}(\|F(x^k)\| + \mathcal{E}_F^k(x^k))
\end{aligned}$$

that $\|x^{k+1} - x^k\| \to 0$ as $k \to \infty$. Therefore by Lemma 4.10 in [19] it derives the convergence of $(x^k)_k$ to x^* with probability no less than δ_1^*. This completes the proof. □

In the following theorem, we show the local convergence rate of Algorithm 2.1. Let $(\gamma^k)_k \subset (0, \infty)$ be a non-increasing sequence with $\gamma^k \to 0$ as $k \to \infty$.

Theorem 5.2. *Under same conditions as Theorem* 5.1*, suppose that there exist* $\hat{l} \geq \bar{l}$ *and* $\bar{\lambda} > 0$ *such that for all* $k \geq \hat{l}$, $\mu_G^k = 0$, $\|\mathcal{G}(x^k, s_j^k) - \mathcal{J}F(x^k)\| \leq \bar{\lambda}$ *for any* $s_j^k \in \Xi$, $j \in [n_G^k]$,

$$\eta^k < \min\left\{\frac{\min\{1, m_G\}}{3\|\mathcal{J}F(x^*)\|}, \frac{(1 - c)m_G}{4\|\mathcal{J}F(x^*)\|^2\|[\mathcal{J}F(x^*)]^{-1}\|}\right\},$$

and

$$n_F^k \geq \frac{\bar{\sigma}_F^2}{\delta^k(\lambda_F^k)^2}, \quad n_G^k \geq \log\left(\frac{2n}{\delta^k}\right)\left(\frac{2\bar{\sigma}_G^2}{(\lambda_G^k)^2} + \frac{2\bar{\lambda}}{3\lambda_G^k}\right),$$

where

$$\lambda_F^k = \min\{\lambda_k^\circ, \lambda_{k-1}^\circ\}, \quad \lambda_G^k < \min\left\{\frac{\varepsilon^k}{2}, \frac{(1-c)m_G}{8\|\mathcal{J}F(x^*)\|\|[\mathcal{J}F(x^*)]^{-1}\|}\right\}$$

with $\lambda_k^\circ = \min\{\frac{m_G^2\varepsilon^k}{(2-c)m_G^2+2m_G\|\mathcal{J}F(x^*)\|+6L'+2}, \frac{m_G(\gamma^k)^{k-\hat{l}}}{2}\}$. *Then there exists* $l_\circ \geq \hat{l}$ *such that for all* $k \geq l_\circ$,

$$\|x^k - x^*\| \leq \tau^k, \tag{5.3}$$

where with $\theta^k = \frac{1}{m_G}(\lambda_G^k + \eta^k\|\mathcal{J}F(x^*)\| + \frac{L'(1+\eta^k)}{2}\|x^k - x^*\|)$,

$$\tau^k = \begin{cases} \max\{\|x^{l_\circ} - x^*\|,\ (\gamma^{l_\circ})^{(l_\circ-\hat{l})/2}\}, & k = l_\circ, \\ \max\{(\theta^{k-1} + (\gamma^{k-1})^{(k-1-\hat{l})/2})\tau^{k-1},\ (\gamma^k)^{(k-\hat{l})/2}\}, & k > l_\circ. \end{cases}$$

Consequently, with probability $\delta_2^* = \prod_{k=\hat{l}}^\infty(1-2\delta^k)(1-\delta^k)$, $(x^k)_k$ *converges to* x^* *at least r-linearly. In addition, if* $\eta^k \to 0$ *and* $\lambda_G^k \to 0$ *as* $k \to \infty$, x^k *converges r-superlinearly to* x^* *with probability* δ_2^*.

Proof. Given $\varepsilon > 0$, define the event $\mathbf{E}_G^k(\varepsilon) = \{\omega \in \Omega : \mathcal{E}_G^k(x^k(\omega)) \leq \varepsilon\}$. Denote $X_j^k = \mathcal{G}(x^k, s_j^k) - \mathbb{E}[\mathcal{G}(x^k, s_j^k)], j = 1, \ldots, n_G^k$. As $\mu_G^k = 0$ for $k \geq \hat{l}$, we have

$$X_j^k = \mathcal{G}(x^k, s_j^k) - \mathcal{J}F(x^k), \quad j = 1, \ldots, n_G^k, \quad k \geq \hat{l},$$

and it yields from assumption (A3) that for any $k \geq \hat{l}$,

$$\max\left\{\left\|\sum_{j=1}^{n_G^k}\mathbb{E}[X_j(X_j^k)^T]\right\|, \left\|\sum_{j=1}^{n_G^k}\mathbb{E}[(X_j^k)^T X_j^k]\right\|\right\} \leq n_G^k\bar{\sigma}_G^2.$$

Recall that $x^k \in \hat{\mathcal{F}}^{k-1}$ and $G(x^k, s_j^k) \in \mathcal{F}^k,\ i \in [n_G^k]$. Then from Lemma A.1 (ii) and $\|X_j^k\| \leq \bar{\lambda}$, $j = 1, \ldots, n_G^k$ with $n_G^k \geq \log(2n/\delta^k)(2\bar{\sigma}_G^2(\lambda_G^k)^{-2} + 2\bar{\lambda}(\lambda_G^k)^{-1}/3)$ it implies that for any $k \geq \hat{l}$,

$$\begin{aligned} &\mathbb{P}(\|G_{s^k}(x^k) - \mathcal{J}F(x^k)\| \\ &\geq \lambda_G^k\ |\hat{\mathcal{F}}^{k-1}) = \mathbb{P}\left(\left\|\frac{1}{n_G^k}\sum_{i=1}^{n_G^k}X_i^k\right\| \geq \lambda_G^k\ |\hat{\mathcal{F}}^{k-1}\right) = \mathbb{P}\left(\left\|\sum_{i=1}^{n_G^k}X_i^k\right\| \geq n_G^k\lambda_G^k\ |\hat{\mathcal{F}}^{k-1}\right) \\ &\leq 2n\cdot\exp\left(\frac{-(n_G^k\lambda_G^k)^2/2}{n_G^k\bar{\sigma}_G^2 + \bar{\lambda}n_G^k\lambda_G^k/3}\right) \leq \delta^k, \end{aligned}$$

which yields

$$\mathbb{P}(\mathbf{E}_G^k(\lambda_G^k)\ |\hat{\mathcal{F}}^{k-1}) = \mathbb{P}(\|G_{s^k}(x^k) - \mathbb{E}[G_{s^k}(x^k)]\| \leq \lambda_G^k\ |\hat{\mathcal{F}}^{k-1}) \geq 1 - \delta^k \tag{5.4}$$

for all $k \geq \hat{l}$. Now consider the event $\mathbf{E} := \bigcap_{k=\hat{l}}^\infty \mathbf{E}_F^k(\lambda_F^k) \cap \mathbf{E}_G^k(\lambda_G^k)$. It follows from the tower property of the conditional expectation, (5.2) and (5.4) that for any $L > \hat{l}$,

$$\mathbb{P}\left(\bigcap_{k=\hat{l}}^L \mathbf{E}_F^k(\lambda_F^k) \cap \mathbf{E}_G^k(\lambda_G^k)\right) = \mathbb{E}\left[\prod_{k=\hat{l}}^{L-1}\mathbf{1}_{\mathbf{E}_F^k(\lambda_F^k)}\mathbf{1}_{\mathbf{E}_G^k(\lambda_G^k)}\{\mathbb{E}[\mathbf{1}_{\mathbf{E}_F^L(\lambda_F^L)}\mathbb{E}[\mathbf{1}_{\mathbf{E}_G^L(\lambda_G^L)}\ |\hat{\mathcal{F}}^{L-1}]\ |\mathcal{F}^{L-1}]\}\right]$$

$$\geq (1-2\delta^L)(1-\delta^L)\mathbb{E}\left[\prod_{k=\hat{l}}^{L-1}\mathbf{1}_{\mathbf{E}_F^k(\lambda_F^k)}\mathbf{1}_{\mathbf{E}_G^k(\lambda_G^k)}\right]$$
$$\geq \prod_{k=\hat{l}}^{L}(1-2\delta^k)(1-\delta^k)=\delta_2^*.$$

Taking the limit as $L\to\infty$, it yields $\mathbb{P}(\mathbf{E})\geq\delta_2^*$. Then we can assume that $(x^k)_k$ is generated by a sample point belonging to $\mathbb{E}$ such that $(x^k)_k$ converges to x^*, which thus occurs with probability no less than δ_2^* and by Theorem 5.1, x^* is a solution of (1.1).

Next, we will show that with probability at least δ_2^*, the line search condition (2.6) always holds whenever k is sufficiently large. In the following, if not specified, the event happens with probability at least δ_2^*. It follows from assumptions (A2) and (B1) that

$$\begin{aligned}
&\|x^k+d^k-x^*\|\\
\leq&\|G_{s^k}^{-1}(x^k)\|\|G_{s^k}(x^k)(x^k-x^*)+G_{s^k}(x^k)d^k\|\\
\leq&\frac{1}{m_G}\|(G_{s^k}(x^k)-\mathcal{J}F(x^k))(x^k-x^*)+(\mathcal{J}F(x^k)(x^k-x^*)+F(x^*)-F(x^k))\\
&-(F_{t^k}(x^k)-F(x^k))+G_{s^k}(x^k)d^k+F_{t^k}(x^k)\|\\
\leq&\frac{1}{m_G}\left(\mathcal{E}_G^k(x^k)+\frac{L'}{2}\|x^k-x^*\|\right)\|x^k-x^*\|+\frac{1}{m_G}\left(\mathcal{E}_F^k(x^k)+\eta^k\|F_{t^k}(x^k)\|\right)\\
\leq&\frac{1}{m_G}\left(\mathcal{E}_G^k(x^k)+\frac{L'}{2}\|x^k-x^*\|\right)\|x^k-x^*\|+\frac{1}{m_G}\mathcal{E}_F^k(x^k)\\
&+\frac{\eta^k}{m_G}\|F_{t^k}(x^k)-F(x^k)+(F(x^k)-F(x^*)+\mathcal{J}F(x^*)(x^k-x^*))-\mathcal{J}F(x^*)(x^k-x^*)\|\\
\leq&\frac{1}{m_G}\left(\mathcal{E}_G^k(x^k)+\eta^k\|\mathcal{J}F(x^*)\|+\Delta(x_k)\right)\|x^k-x^*\|+\frac{1+\eta^k}{m_G}\mathcal{E}_F^k(x^k)\\
\leq&\frac{1}{m_G}\left(\mathcal{E}_G^k(x^k)+\eta^k\|\mathcal{J}F(x^*)\|+\Delta(x_k)\right)\|x^k-x^*\|+\frac{2}{m_G}\mathcal{E}_F^k(x^k),
\end{aligned}\tag{5.5}$$

where $\Delta(x^k)=\frac{L'(1+\eta^k)}{2}\|x^k-x^*\|$. Notice that by assumptions on λ_G^k and η^k, we have that for any $k\geq\hat{l}$,

$$\mathcal{E}_G^k(x^k)\leq\lambda_G^k\leq\frac{1}{3},\quad \eta^k\leq 1,\quad \eta^k\|\mathcal{J}F(x^*)\|\leq\frac{1}{3},$$

and due to $x^k\to x^*$, there exists $l_1\geq\hat{l}$ such that for all $k\geq l_1$,

$$\|x^k-x^*\|\leq 1,\quad \Delta(x_k)\leq\frac{1}{3}.$$

Then we obtain that for any $k\geq l_1$,

$$\begin{aligned}
&\|x^k+d^k-x^*\|\leq\frac{1}{m_G}\|x^k-x^*\|+\frac{2}{m_G}\mathcal{E}_F^k(x^k),\\
&\Delta(x^k+d^k)=\frac{L'(1+\eta^k)}{2}\|x^k+d^k-x^*\|\leq\frac{1}{m_G}\Delta(x^k)+\frac{2L'}{m_G}\mathcal{E}_F^k(x^k),
\end{aligned}$$

which further indicates from $\Delta(x^k)\leq 1$, $\|x^k-x^*\|\leq 1$ and $\mathcal{E}_F^k(x^k)\leq 1$ for $k\geq l_1$ that

$$\Delta(x^k+d^k)\|x^k+d^k-x^*\|\leq\frac{\Delta(x^k)}{m_G^2}\|x^k-x^*\|+\frac{2}{m_G^2}\Delta(x^k)\mathcal{E}_F^k(x^k)$$

$$
\begin{aligned}
&+\frac{2L'}{m_G^2}\mathcal{E}_F^k(x^k)\|x^k-x^*\|+\frac{4L'}{m_G^2}(\mathcal{E}_F^k(x^k))^2\\
&\le\frac{\Delta(x^k)}{m_G^2}\|x^k-x^*\|+\frac{2+6L'}{m_G^2}\mathcal{E}_F^k(x^k). \qquad (5.6)
\end{aligned}
$$

By (2.11) and the definition of $\Delta(x)$, we have that for any $k\ge l_1$,

$$
\begin{aligned}
\|F(x^k+d^k)\|&\le\frac{L'}{2}\|x^k+d^k-x^*\|^2+\|\mathcal{J}F(x^*)\|\|x^k+d^k-x^*\|\\
&\le\Delta(x^k+d^k)\|x^k+d^k-x^*\|+\|\mathcal{J}F(x^*)\|\|x^k+d^k-x^*\|.
\end{aligned}
$$

Then it follows from (5.5) and (5.6) that

$$
\begin{aligned}
\|F(x^k+d^k)\|\le&\left[\left(\frac{1}{m_G}+\|\mathcal{J}F(x^*)\|\right)\frac{\Delta(x^k)}{m_G}+\frac{\eta^k}{m_G}\|\mathcal{J}F(x^*)\|^2+\frac{\mathcal{E}_G^k(x^k)}{m_G}\|\mathcal{J}F(x^*)\|\right]\|x^k-x^*\|\\
&+\left(\frac{2\|\mathcal{J}F(x^*)\|}{m_G}+\frac{6L'+2}{m_G^2}\right)\mathcal{E}_F^k(x^k).
\end{aligned}
$$

Notice that by the mean value theorem, for any x^k there exists $\xi^k\in(x^*,x^k)$ such that

$$
F(x^k)=F(x^k)-F(x^*)=\mathcal{J}F(\xi^k)(x^k-x^*). \qquad (5.7)
$$

As $x^k\to x^*$ and $\Delta(x^k)\to 0$, there exists $l_2\ge l_1$ such that for all $k\ge l_2$,

$$
\Delta(x^k)\le\frac{(1-c)m_G}{8(1/m_G+\|\mathcal{J}F(x^*)\|)\|[\mathcal{J}F(x^*)]^{-1}\|},
$$

and for any $x\in(x^*,x^k)$, $\mathcal{J}F(x)$ is nonsingular and $\|[\mathcal{J}F(x)]^{-1}\|\le 2\|[\mathcal{J}F(^*)]^{-1}\|$. Then it yields from (5.7) that

$$
\|x^k-x^*\|\le 2\|[\mathcal{J}F(x^*)]^{-1}\|\|F(x^k)\|.
$$

Therefore, it follows from

$$
\eta^k<\frac{(1-c)m_G}{4\|\mathcal{J}F(x^*)\|^2\|[\mathcal{J}F(x^*)]^{-1}\|},\quad \lambda_G^k<\frac{(1-c)m_G}{8\|\mathcal{J}F(x^*)\|\|[\mathcal{J}F(x^*)]^{-1}\|},\quad k\ge\hat{l},
$$

that for any $k\ge l_2$,

$$
\|F(x^k+d^k)\|\le(1-c)\|F(x^k)\|+\left(\frac{2\|\mathcal{J}F(x^*)\|}{m_G}+\frac{6L'+2}{m_G^2}\right)\mathcal{E}_F^k(x^k).
$$

In addition, according to the bound on λ_F^k, it indicates that for any $k\ge l_*:=\max\{l_2,\hat{l}\}$,

$$
\begin{aligned}
&\|F_{t^{k+1}}(x^k+d^k)\|-(1-c)\|F_{t^k}(x^k)\|\\
&\le\|F(x^k+d^k)\|+\mathcal{E}_F^{k+1}(x^k+d^k)-(1-c)\|F(x^k)\|+(1-c)\mathcal{E}_F^k(x^k)\\
&\le\mathcal{E}_F^{k+1}(x^k+d^k)+(1-c)\mathcal{E}_F^k(x^k)+\left(\frac{2\|\mathcal{J}F(x^*)\|}{m_G}+\frac{6L'+2}{m_G^2}\right)\mathcal{E}_F^k(x^k)\\
&\le\lambda_F^{k+1}+(1-c)\lambda_F^k+\left(\frac{2\|\mathcal{J}F(x^*)\|}{m_G}+\frac{6L'+2}{m_G^2}\right)\lambda_F^k\\
&\le\left(2-c+\frac{2\|\mathcal{J}F(x^*)\|}{m_G}+\frac{6L'+2}{m_G^2}\right)\lambda_k^\circ\le\varepsilon^k,
\end{aligned}
$$

which satisfies (2.6). Therefore, the line search condition is always satisfied for any $k \geq l_*$ with a probability no less than δ_2^*.

Finally we will derive the convergence rate of Algorithm 2.1. As $x^{k+1} = x^k + d^k$ for any $k \geq l_*$, it follows from (5.5) that

$$\begin{aligned}\|x^{k+1} - x^*\| &\leq \frac{1}{m_G}(\lambda_G^k + \eta^k \|\mathcal{J}F(x^*)\| + \Delta(x^k))\|x^k - x^*\| + \frac{1+\eta^k}{m_G}\lambda_F^k \\ &\leq \theta^k \|x^k - x^*\| + (\gamma^k)^{k-\hat{l}},\end{aligned}$$

where $\theta^k = \frac{1}{m_G}(\lambda_G^k + \eta^k\|\mathcal{J}F(x^*)\| + \Delta(x^k))$. By induction, from the definition of $(\tau^k)_{k\geq l_\circ}$, it is easy to obtain (5.3). Moreover, as $(\gamma^k)_k$ is a nonincreasing sequence, we obtain that for any $k \geq l_\circ$,

$$\begin{aligned}\frac{\tau^{k+1}}{\tau^k} &\leq \max\{\theta^k + (\gamma^k)^{(k-\hat{l})/2},\ (\gamma^{k+1})^{1/2}(\gamma^{k+1}/\gamma^k)^{(k-\hat{l})/2}\} \\ &\leq \max\{\theta^k + (\gamma^k)^{(k-\hat{l})/2},\ (\gamma^k)^{1/2}\}.\end{aligned}$$

Notice that the value of η^k plays a key role in determining the convergence rate of Algorithm 2.1. If $\eta^k < \frac{m_G}{3\|\mathcal{J}F(x^*)\|}$ for all sufficiently large k, then due to $\frac{\lambda_G^k}{m_G} < \frac{1}{8}$ and $\Delta(x^k), \gamma^k \to 0$ we obtain there exists $\bar{\theta} < 1$ such that $\theta^k \leq \bar{\theta}$ for any $k \geq l_\circ$. Hence $(\tau^k)_k$ converges q-linearly to 0, which indicates that x^k converges to x^* with r-linear rate. If $\eta^k \to 0$ and $\lambda_G^k \to 0$ as $k \to \infty$, then τ^k converges q-superlinerly to 0, thus x^k converges to x^* with r-superlinear rate. □

Remark 5.1. In Theorem 5.2, to guarantee the r-superlinear convergence rate, it requires the stochastic approximate function $F_{t^k}(\cdot)$ and the stochastic approximate Jacobian $G_{s^k}(\cdot)$ are unbiased estimates of $F(\cdot)$ and $\mathcal{J}F(\cdot)$, respectively, i.e.

$$\mathbb{E}[F_{t^k}(x)] = F(x), \quad \mathbb{E}[G_{s^k}(x)] = \mathcal{J}F(x).$$

Actually, the above two relations can be realized simultaneously. More specifically, with $s^k = t^k$ and $\mathcal{G}(x, s_j^k) = \mathcal{J}F(x, t_i^k)$ for any $x \in \mathbb{R}^n$, suppose that $\mathbb{E}[F_{t^k}(x)] = F(x)$. If $F(\cdot)$ is well defined and finite valued at any $x \in \mathbb{R}^n$, and for almost every $\xi \in \Xi$, $F_{t^k} = F_{t^k(\xi)}$ is differentiable and Lipschitz continuous in x, then by Theorem 7.44 in [28], we have

$$\mathbb{E}[G_{s^k}(x)] = \mathbb{E}[\mathcal{J}F_{t^k}(x)] = \mathcal{J}\mathbb{E}[F_{t^k}(x)] = \mathcal{J}F(x). \tag{5.8}$$

If $F_{t^k(\xi)}$ is random lower semicontinuous and convex in x, and F is proper, we can also obtain (5.8).

6. Numerical Experiments

In this section, we present some preliminary numerical results to illustrate the performance of Algorithm 2.1. All numerical experiments were implemented in MATLAB R2019a on a laptop with Intel(R) Core(TM) i5-6200U 2.30GHz and 8GB memory. The four data sets used in numerical experiments are displayed in the following Table 6.1.

We next list all the algorithms present in numerical experiments.

Table 6.1: Datasets used in the experiments.

Data Set	No. of Data Points: m	No. of Variables: n	Reference
Adult	1605	123	[20]
CINA	16033	132	[6]
gisette	6000	5000	[10]
rcv1	20242	47236	[17]

Alg. 2.1 Algorithm 2.1.

SN. Stochastic Newton method, obtained by removing the line search step in Algorithm 2.1 while only taking a constant step size.

LNM. Deterministic line search Newton method, obtained by replacing all the stochastic information in Algorithm 2.1 with exact function value and Jacobian, i.e. $F(x^k)$ and $\mathcal{J}F(x^k)$, respectively.

NM. Deterministic Newton method, obtained by removing the line search step in LNM while only taking a constant step size.

SGN. Stochastic Gauss-Newton algorithm proposed in [29] for nonconvex compositional optimization.

SGM. Stochastic gradient method combining the variance reduced stochastic oracle [15], using sub-sampled gradient information with 10% sample size of the training data size n.

In both Alg. 2.1 and SN, stochastic function values and Jacobian matrices are generated in the same way. More specifically, we first select the sub-samples $T_k \subseteq [m]$ uniformly at random and without replacement from the index set $\{1, \ldots, m\}$. Then compute the mini-batch stochastic oracles through (2.2), where the size of T_k is chosen increasingly by 5% from $\lceil 0.05m \rceil$ until it reaches m. In both Alg. 2.1 and LNM, we set $\varepsilon^k = k^{-4/3}$, $\eta^k = 10^{-5}$, $\alpha^k = 0.3$ and $c = 0.3$. In Alg. 2.1, SN, LNM and NM, Gauss-Seidel iterative method [31] is used to solve the approximate solution d^k of (2.4). In SGN and SGM, the parameter and step sizes are selected for best performance. Without further specification, we use the following termination criterion in numerical tests: $\|x^{k+1} - x^k\| \leq 10^{-9}$.

6.1. Logistic regression problem

We consider the following binary classification problem:

$$\min_{x \in \mathbb{R}^n} \; H(x) = \frac{1}{m} \sum_{i=1}^{m} h_i(x) + \frac{\lambda}{2} \|x\|^2, \tag{6.1}$$

where h_i is the logistic loss function, i.e., $h_i(x) := \log(1 + \exp(-b_i (a_i)^T x))$ and $\lambda = 0.01$. The vector $a^i \in \mathbb{R}^n$ represents the feature vector of the data and $b^i \in \{-1, 1\}$ represents the label of each data in the binary classification problem. As the objective function H is smooth and level bounded, it has stationary points which solve

$$0 = F(x) := \frac{1}{m} \sum_{i=1}^{m} \nabla h_i(x) + \lambda x. \tag{6.2}$$

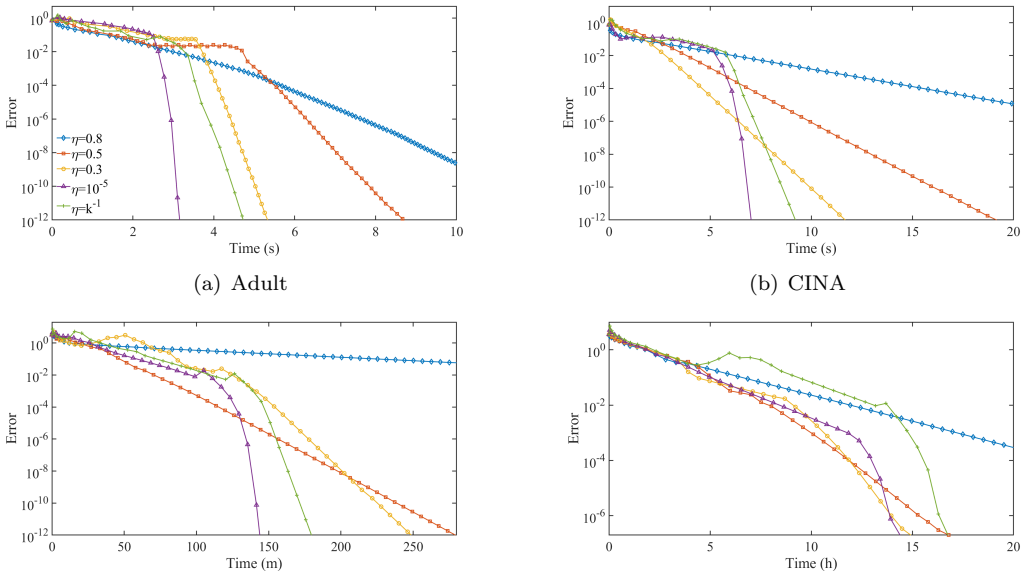
In the following we report numerical results by applying algorithms to solve the above nonlinear system of equations.

As shown in previous theoretical analysis, the parameter η^k has a very important influence on the convergence rate of Alg. 2.1. In Fig. 6.1 we report the numerical impact of η^k on the performance of Alg. 2.1, by recording changing curves of errors under different settings of η^k with respect to CPU time. Here, "Time (s)", "Time (m)" and "Time (h)" denote the CPU time in seconds, minutes and hours, respectively. The error at iterate x^k is defined as $\|F(x^k)\|$. We set η^k as 10^{-5}, 0.3, 0.5, 0.8 and $1/k$, respectively. We can see that Alg. 2.1 performs better as η^k is smaller and achieves best when $\eta^k = 10^{-5}$. This demonstrates that solving the approximate Newton's equation more accurately can improve the performance of Alg. 2.1.

Fig. 6.2 shows the influence of step size α^k on the performance of Alg. 2.1. In the experiments we test five different settings where $\alpha^k = 0.3,\ 1,\ 5k^{-1},\ k^{-1/2}$ and $k^{-1/4}$ for each data set. From this figure, we can see when $\alpha^k = 1$, the error barely changes. Comparatively, Alg. 2.1 performs better in other settings when α^k is smaller, particularly when $\alpha^k = 0.3$ and $5k^{-1}$. It shows that numerically simply unit step size is not enough to guarantee the global convergence, although it brings faster local convergence rate when close to solution.

In Fig. 6.3, we report the impact of the sample sizes on the performance of Alg. 2.1 by using different sampling rates. For all data sets, we choose the same initial sample set size as 5% of total sample size m, then increase it to 100% at the sampling rate of 1%, 2%, 5%, and 10%, respectively. From the figure we can see that the error decreases faster as the sampling rate increases. As more samples are used when higher sampling rate is set, it reveals that the use of more function information can help improve the algorithmic performance of Alg. 2.1.

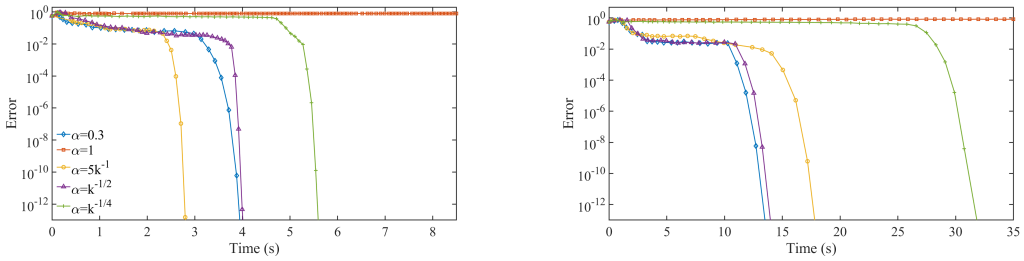
In local convergence analysis of previous Theorem 5.2 we have shown that when the iteration number k is sufficiently large, the line search condition can always hold, thus the step size of Alg. 2.1 is equal to 1. To show this numerically, we plot the values of step sizes along



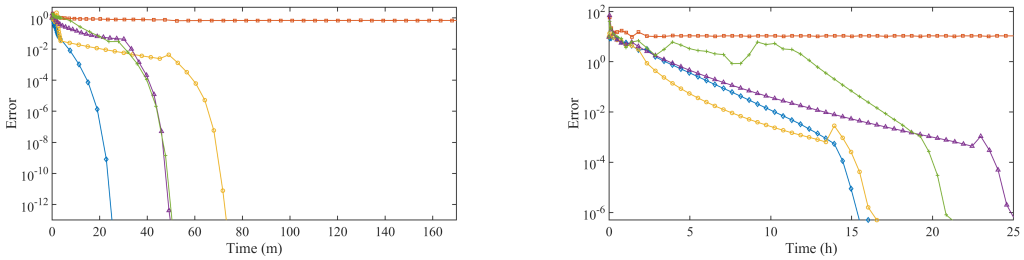
(a) Adult (b) CINA

(c) gisette (d) rcv1

Fig. 6.1. Performance profile of Alg. 2.1 associated with different η^k.

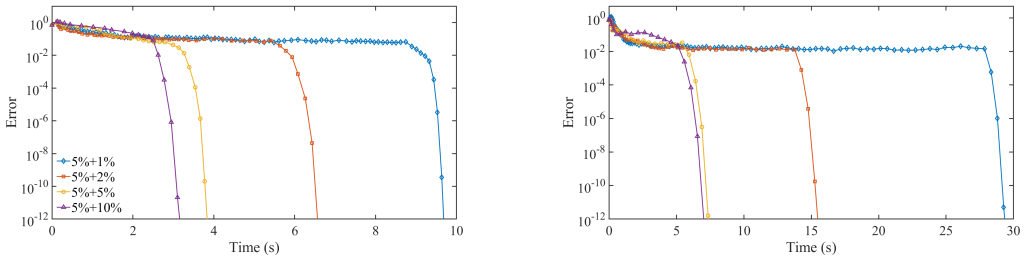


(a) Adult (b) CINA

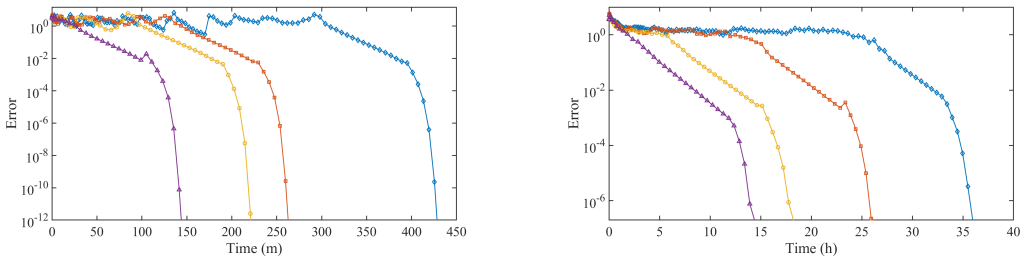


(c) gisette (d) rcv1

Fig. 6.2. Comparison of Alg. 2.1 associated with different α^k.



(a) Adult (b) CINA

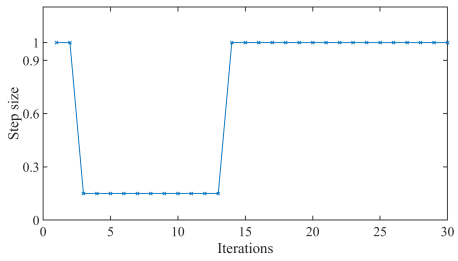


(c) gisette (d) rcv1

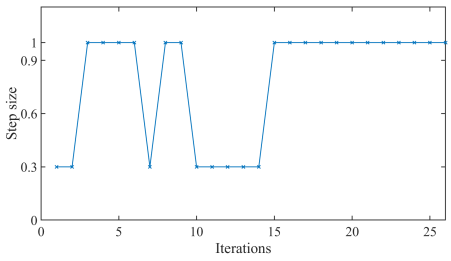
Fig. 6.3. Comparison on Alg. 2.1 associated with different sampling rates.

with the algorithm proceeds in Fig. 6.4. For each data set, we run Alg. 2.1 until it reaches $\|x^{k+1} - x^k\| \leq 10^{-15}$. We can see that in later stage of Alg. 2.1, the step size is always 1 which verifies our previous theoretical analysis.

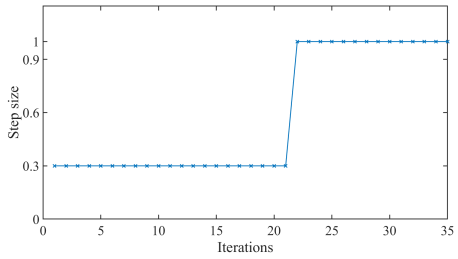
In Fig. 6.5, we show the comparison performances of all six algorithms aforementioned. The same initial point x^0 is chosen. From the numerical results, Alg. 2.1 and LNM converge more



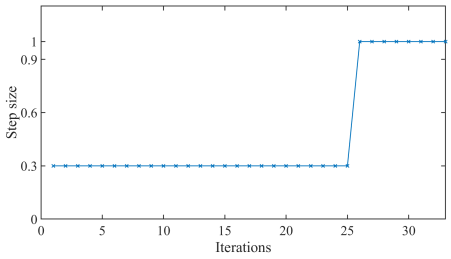
(a) Adult



(b) CINA

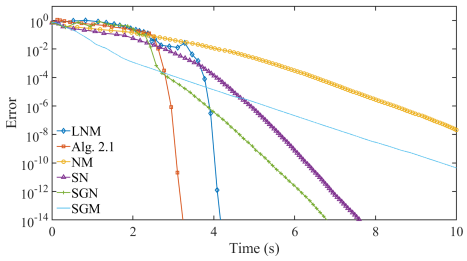


(c) gisette

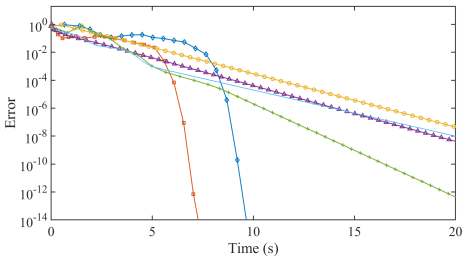


(d) rcv1

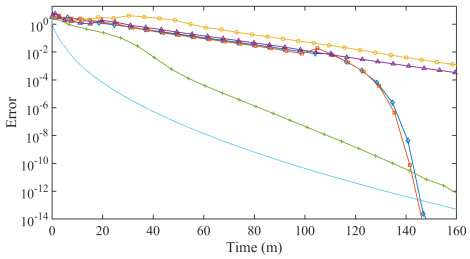
Fig. 6.4. The trend of the step size α^k in Alg. 2.1 for solving (6.2).



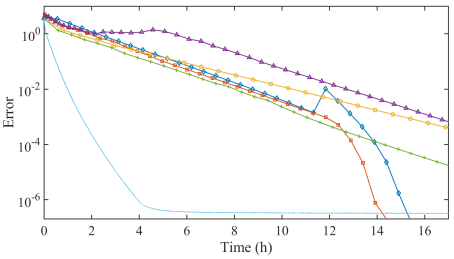
(a) Adult



(b) CINA



(c) gisette



(d) rcv1

Fig. 6.5. Comparison of six algorithms for solving (6.2).

rapidly than NM and SN in general for all data sets. After a few initial iterations the error returned by Alg. 2.1 and LNM decreases more dramatically. We believe that in these stages the iterates are approaching the solution thus the line search condition is always satisfied which brings faster convergence rate according to previous local convergence analysis. Moreover, Alg. 2.1 achieves lower error within the same CPU time on all data sets compared with LNM,

which reflects that the application of stochastic information can help to improve the algorithm within the same CPU time. This point can also been seen from the better performance of SN than NM. In addition, compared with SGN and SGM, Alg. 2.1 reveals faster local convergence rate in later stage of the algorithm process and shows superior performance, especially on data sets Adult and CINA.

6.2. Nonlinear equations

We consider the following nonlinear system of equations:

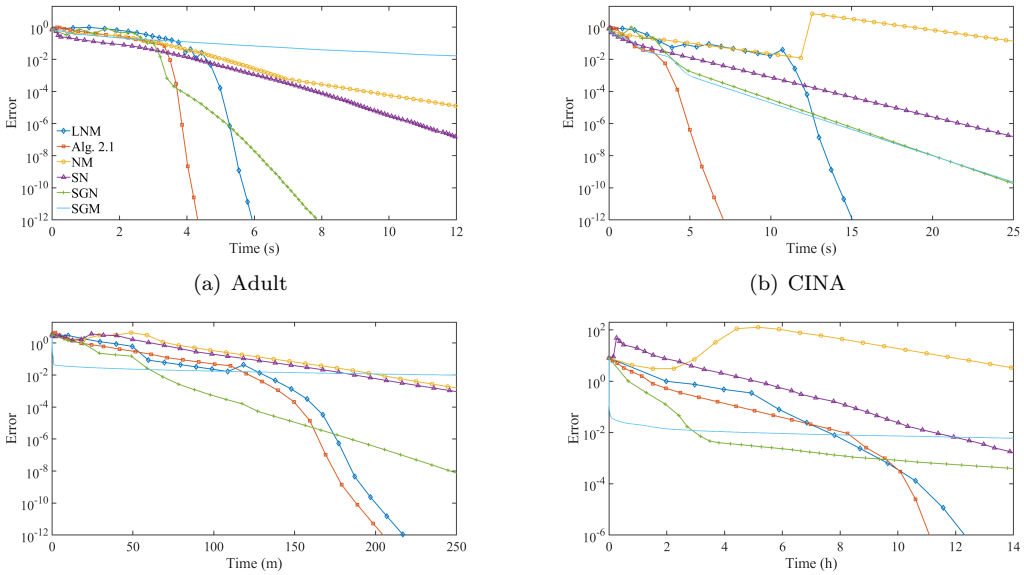
$$\frac{1}{m}\sum_{i=1}^{m} f_i(x) = 0, \tag{6.3}$$

where the mapping $f_i : \mathbb{R}^n \to \mathbb{R}^{2n}$ is defined as

$$f_i(x) = \begin{bmatrix} -b_i \cdot (1 - \tanh(b_i \cdot \langle a_i, x\rangle))^2 \cdot a_i \\ -b_i \cdot \dfrac{\exp(-b_i \cdot \langle a_i, x\rangle)}{1 + \exp(-b_i \cdot \langle a_i, x\rangle)} \cdot a_i + \lambda x \end{bmatrix}, \quad i = 1, 2, \ldots, m. \tag{6.4}$$

Here, $a_i \in \mathbb{R}^n$ and $b_i \in \{-1, 1\}$ are denoted same as those in problem (6.1) and $\lambda = 0.01$. Note that each f_i consists of gradients of previous logistic loss function and the sigmoid loss function which is defined as $1 - \tanh(b_i \cdot \langle a_i, x\rangle)$. Both loss functions are widely used in machine learning. As both functions are smooth and level bounded, it ensures that (6.3) has a solution. We compare all six algorithms on four data sets given in Table 6.1. Moreover, except for $\alpha = 0.15$ on the data set Adult, other parameters used in those algorithms are consistent with Section 6.1.

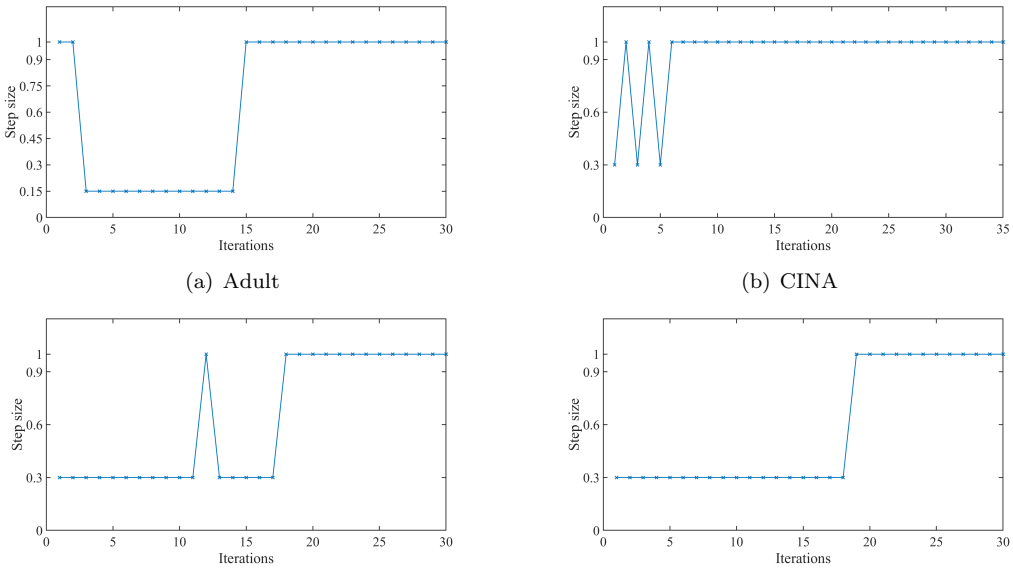
We apply all six algorithms to solve (6.3), (6.4) and report the comparison results in Fig. 6.6. We can see that after initial iterations and compared with SGN, SGM, NM and SN, both Alg. 2.1



(a) Adult (b) CINA

(c) gisette (d) rcv1

Fig. 6.6. Comparison of six algorithms for solving (6.4).



(a) Adult (b) CINA

(c) gisette (d) rcv1

Fig. 6.7. The trend of the step size α^k in Alg. 2.1 for solving (6.4).

and LNM reduce the errors much more rapidly. We believe that in these stages the line search condition is always satisfied thus faster local convergence rate can be achieved. Moreover, Alg. 2.1 performs better than LMN, which shows again that the use of stochastic information brings benefit to algorithmic performance within the same CPU time.

The trend of the step size α^k during the process of Alg. 2.1 until $\|x^{k+1} - x^k\| \leq 10^{-15}$ is drawn in Fig. 6.7 on four data sets. When $\alpha^k = 1$, it means that the line search condition (2.6) is satisfied. Otherwise, a constant step size is taken in Alg. 2.1. As can be seen from Fig. 6.7, Alg. 2.1 can guarantee that while the iteration number is increasing, the step size α^k is always equal to 1 after at most twenty iterations, which means that in these cases (2.6) is always satisfied.

7. Conclusions

In this paper, we propose a stochastic Newton method, Algorithm 2.1, for solving nonlinear equations which can only be accessed through stochastic oracles. At each iteration, we compute an inexact Newton direction by solving the approximate Newton's equation constructed based on stochastic zeroth- and first-order oracles. Then to determine the step size we consider an inexact backtracking line search condition which is relying on stochastic approximations. We take the unit step size if the line search condition is satisfied. Otherwise, a preset small step size is taken. We establish the global convergence of errors at iterates, i.e. $\|F(x^k)\|$, in expectation as well as its almost-sure convergence for Algorithm 2.1. Furthermore, we explore the computational complexities of Algorithm 2.1 with respect to calls to stochastic zeroth- and first-order oracles, when the algorithm returns a randomly chosen iterate as the output. Moreover, we analyze local convergence properties of Algorithm 2.1 and establish the local convergence rate in high probability. Finally, we report experimental results on some large data sets and the proposed algorithm shows very promising numerical performances.

Appendix

Lemma A.1. *Let $(\mathcal{U}_k)_{k=0}^m$ be a given filtration of the σ-algebra $\mathcal{F}$.*

(i) *Let $(X_k)_{k=1}^m, X_k : \Omega \to \mathbb{R}^n$, be a family of random vectors, satisfying $X_k \in \mathcal{U}_k$ and $\sigma \in \mathbb{R}^m$ be a given vector with $\sigma_k \neq 0$, $k = 1, \ldots, m$. Suppose that $\mathbb{E}[X_k \ |\mathcal{U}_{k-1}] = 0$, and $\mathbb{E}[\|X_k\|^2 \ |\mathcal{U}_{k-1}] \leq \sigma_k^2$ a.e. for all $k \in [m]$. Then it holds*

$$\mathbb{E}\left[\left\|\sum_{k=1}^m X_k\right\|^2 \middle\| \mathcal{U}_0\right] \leq \|\sigma\|^2, \quad \mathbb{P}\left(\left\|\sum_{k=1}^m X_k\right\| \geq \tau\|\sigma\| \ |\mathcal{U}_0\right) \leq \tau^{-2}, \quad \forall \tau > 0$$

almost everywhere.

(ii) *Let $(X_k)_{k=1}^m$, $X_k : \Omega \to \mathbb{R}^{d_1 \times d_2}$, be a sequence of random matrices satisfying $X_k \in \mathcal{U}_k$. Suppose that $\mathbb{E}[X_k \ |\mathcal{U}_{k-1}] = 0$, and there exists a positive constant R such that $\|X_k\| \leq R$ a.e. for all $k \in [m]$. Define $\nu^2 = \max\{\|\sum_{k=1}^m \mathbb{E}(X_k X_k^T)\|, \|\sum_{k=1}^m \mathbb{E}(X_k^T X_k)\|\}$. Then it holds*

$$\mathbb{P}\left(\left\|\sum_{k=1}^m X_k\right\| \geq t \ |\mathcal{U}_0\right) \leq (d_1 + d_2) \cdot \exp\left(\frac{-t^2/2}{\nu^2 + Rt/3}\right), \quad \forall t > 0$$

almost everywhere.

Acknowledgments. This research was partially supported by the National Natural Science Foundation of China (Nos. 11731013, 11871453 and 11971089), Young Elite Scientists Sponsorship Program by CAST (No. 2018QNRC001), Youth Innovation Promotion Association, CAS, and Fundamental Research Funds for the Central Universities, UCAS.

References

[1] C.D. Aliprantis and K.C. Border, *Infinite dimensional analysis: A hitchhiker's guide*, third ed, Springer, Berlin, Germany, 2006.

[2] P. Bianchi, Ergodic convergence of a stochastic proximal point algorithm, *SIAM J. Optim.*, **26**:4 (2016), 2235–2260.

[3] E.G. Birgin, N. Krejić and J.M. Marténez, Globally convergent inexact quasi-Newton methods for solving nonlinear equations, *Numer. Algorithms*, **32** (2003), 249–260.

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

[5] R. Bollapragada, R.H. Byrd and J. Nocedal, Exact and inexact subsampled Newton methods for optimization, *IMA J. Numer. Anal.*, **39**:3 (2019), 545–578.

[6] Causality workbench team, A marketing dataset, http://www.causality.inf.ethz.ch/data/CINA.html, 2008, September.

[7] J.E. Dennis and R.B. Schnabel, Numerical methods for unconstrained optimization and nonlinear equations, *SIAM J. Math. Data Sci.*, 1996, DOI: 10.1137/1.9781611971200.

[8] P. Dvurechensky and A. Gasnikov, Stochastic intermediate gradient method for convex problems with stochastic inexact oracle, *J. Optim. Theory Appl.*, **171**:1 (2016), 121–145.

[9] S. Ghadimi and G. Lan, Stochastic first-and zeroth-order methods for nonconvex stochastic programming, *SIAM J. Optim.*, **23**:4 (2013), 2341–2368.

[10] I. Guyon, S. Gunn, A. Ben-Hur and G. Dror, Result analysis of the NIPS 2003 feature selection challenge, *NIPS*, (2005), 545–552.

[11] A.N. Iusem, A. Jofré, R.I. Oliveira and P. Thompson, Extragradient method with variance reduction for stochastic variational inequalities, *SIAM J. Optim.*, **27**:2 (2017), 686–724.
[12] A.N. Iusem, A. Jofré, R.I. Oliveira, and P. Thompson, Variance-Based Extragradient Methods with Line Search for Stochastic Variational Inequalities, *SIAM J. Optim.*, **29**:1 (2019), 175–206.
[13] A.N. Iusem, A. Jofré and P. Thompson, Incremental constraint projection methods for monotone stochastic variational inequalities, *Math. Oper. Res.*, **44**:1 (2018), 236–263.
[14] H. Jiang and H. Xu, Stochastic approximation approaches to the stochastic variational inequality problem, *IEEE Trans. Automat. Contr.*, **53**:6 (2008), 1462–1475.
[15] R. Johnson and T. Zhang, Accelerating stochastic gradient descent using predictive variance reduction, *NIPS*, **1**:3 (2013), 315–323.
[16] P.E. Kloeden and E. Platen, Numerical Solution of Stochastic Differential Equations, Springer, 1992.
[17] D.D. Lewis, Y. Yang, T.G. Rose and F. Li, RCV1: A new benchmark collection for text categorization research, *J. Math. Learn. Res.*, **5** (2004), 361–397.
[18] A. Milzarek, X. Xiao, S. Cen, Z. Wen and M. Ulbrich, (2019). A Stochastic Semismooth Newton Method for Nonsmooth Nonconvex Optimization, *SIAM J. Optim.*, **29**:4 (2019), 2916–2948.
[19] J.J. Moré and D.C. Sorensen, Computing a trust region step, *SIAM J. Sci. Comput.*, **4**:3 (1983), DOI:10.1137/0904038.
[20] I. Mukherjee, K. Canini, R. Frongillo and Y. Singer, Parallel boosting with momentum, In *Joint European Conference on Mach. Learn. Knowl. Discov. Databases*, Springer, Berlin, Heidelberg, (2013), September, 17–32.
[21] A. Nemirovski, A. Juditsky, G. Lan and A. Shapiro, Robust stochastic approximation approach to stochastic programming, *SIAM J. Optim.*, **19**:4 (2009), 1574–1609.
[22] J. Nocedal and S. Wight, *Numerical optimization*, Springer, 2006.
[23] G.D. Nunno and T. Zhang, Approximations of stochastic partial differential equations, *Ann. Appl. Probab.*, **26**:3 (2016), 1443–1466.
[24] C. Paquette and K. Scheinberg, A stochastic line search method with convergence rate analysis, *SIAM J. Optim.*, **30**:1 (2020), 349–376.
[25] S.J. Reddi, S. Sra, B. Poczos and A.J. Smola, Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization, *NIPS*, (2016), 1145–1153.
[26] H. Robbins and S. Monro, A stochastic approximation method, *Ann. Math. Stat.*, (1951), 400–407.
[27] S.M. Ross, *Introduction to Stochastic Dynamic Programming*, Academic Press, Inc., USA, 1983.
[28] A. Shapiro, D. Dentcheva and A. Ruszczyński, Lectures on stochastic programming: modeling and theory, *MOS-SIAM Series on Optimization, SIAM*, Philadelphia, USA, **9**: (2009).
[29] Q. Tran-Dinh, N.H. Pham, and L. Nguyen, Stochastic Gauss-Newton Algorithms for Nonconvex Compositional Optimization, In *ICML 2020: 37th International Conference on Machine Learning,* 2020.
[30] J.A. Tropp, User-friendly tail bounds for sums of random matrices, *Found. Comut. Math.*, **12**:4 (2012), 389–434.
[31] R.S. Varga, *Matrix Iterative Analysis*, Springer series in Computational Mathematics, Prentice-Hall, Inc, Englewood Cliffs, New Jersey, 1962.
[32] S. Vaswani, A. Mishkin, I. Laradji, M. Schmidt, G. Gidel and S. Lacoste-Julien, Painless stochastic gradient: Interpolation, line search, and convergence rates, *NeurIPS*, **32** (2019), 3732–3745.
[33] Y. Xu and W. Yin, Block stochastic gradient iteration for convex and nonconvex optimization, *SIAM J. Optim.*, **25**:3 (2015), 1686–1716.
[34] T.J. Ypma, Historical development of the Newton-Raphson method, *SIAM Rev.*, **37**:4 (1995), 531–551.