Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online CONFERENCE PAPER ARCHIVE

IFAC-PapersOnLine 48-28 (2015) 229–233

On the Convergence of Smoothed Functional Stochastic Optimization Algorithms ⋆

Xing-Min Chen * **Chao Gao** **

* *School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, P. R. China (e-mail: xmchen@dlut.edu.cn).*
** *School of Information Science and Engineering, Dalian Polytechnic University, Dalian 116034, P. R. China (e-mail: gaochao198604@126.com)*

Abstract:
Smoothed functional gradient algorithm with perturbations distributed according to the Gaussian distribution is considered for stochastic optimization problem with additive noise. A stochastic approximation algorithm with expanding truncations that uses either one-sided or two-sided gradient estimate is given. At each iteration of the algorithm only two observations are required. The algorithm is shown to be convergent under only some mild conditions.

Keywords: Stochastic optimization, smoothed functional estimate, simultaneous perturbation stochastic approximation, strong consistency.

1. INTRODUCTION

Stochastic optimization algorithms have been found in a variety of applications spreading over many fields such as mathematical optimization, control theory, signal processing, and machine learning, see e.g. Chen (2002); Bhatnagar et al. (2012). One of the main problem of stochastic optimization is to find the extrema of an unknown function $f : \mathbb{R}^d \to \mathbb{R}$ which cannot be computed directly, but only estimated via noisy observations.

Kiefer and Wolfowitz (1952) published the first stochastic approximation algorithm for such a problem, which requires that for each gradient computation, $2d$ measurements are needed for every iteration of the algorithm. This means that when d is large, the Kiefer-Wolfowitz (KW) algorithm will require substantial computational effort per iteration, leading to slow convergence.

To address this problem, Spall (2000) proposed the use of simultaneous perturbations to estimate the gradient, which is known as the simultaneous perturbation stochastic approximation (SPSA) algorithm. This method would require only two simulations per iteration, regardless of the dimension d. Using the ordinary differential equation (ODE) method, the convergence and asymptotic normality of the algorithm are showed though the conditions are restrictive.

Chen et al. (1999) proposed a simultaneous perturbation algorithm that uses either one-sided or two-sided randomized differences and truncations at randomly varying bounds. At each iteration only two observations are required in contrast to $2d$ observations. A direct method rather than the classical probabilistic method or the ODE method to establish the convergence, the rate of convergence and asymptotic normality of the algorithm under only some mild conditions.

A remarkable feature of the simultaneous perturbation algorithm is that it estimates the gradient of the objective by simultaneously perturbing all parameter components and requires only two measurements of the objective function for this purpose. Smoothed functional (SF) algorithms also belong to the class of simultaneous perturbation methods, because they update the gradient of the objective using function measurements involving parameter updates that are perturbed simultaneously in all component directions. The SF gradient estimates were developed by Katkovnik and Kulchitsky (1972), the idea was to approximate the gradient of expected performance by its convolution with a multivariate Gaussian distribution. While the original SF algorithm in Katkovnik and Kulchitsky (1972) uses only one-sided simulation, Styblinski and Tang (1990) and Chin (1997) presented a related two-simulation SF algorithm based on a finite difference gradient estimate, which is shown in Styblinski and Tang (1990) to have lower variability compared to the one-sided one. By using the celebrated ODE method, it was shown that in Bhatnagar et al. (2012): Given $\varepsilon > 0$, there exists $\beta_0 > 0$, such that for all $\beta \in (0, \beta_0]$, the iterates converge to the open ε-neighborhood of the root set of f almost surely as time tends to infinity, where β is the scalar parameter used in SF gradient estimate. However, it *a prior* requires the state be bounded and the noise condition is difficult to be verified due to its state-dependence.

The SF gradient estimates are related to the derivative-free methods in mathematical optimization (see e.g. Nes-

⋆ This work is supported by the National Natural Science Foundation of China under Grant 61203118, the Fundamental Research Funds for Central Universities, and the Youth Foundation of Dalian Polytechnic University under Grant QNJJ201416.

2405-8963 © 2015, IFAC (International Federation of Automatic Control) Hosting by Elsevier Ltd. All rights reserved.
Peer review under responsibility of International Federation of Automatic Control.
10.1016/j.ifacol.2015.12.130

terov (2011)), where the derivative-free estimate rather than the gradient is used in development of optimization algorithms. An evident advantage of these methods is that the program of computation of the function value is always simpler than the one for computing the gradient. However, it has been realized that these methods are much more difficult for theoretical investigation.

In this paper, we focus on the case that the noise in the observation of the value of the objective f is additive. Based on smoothing the the objective using Gaussian density, we present a stochastic approximation algorithm with expanding truncations (SAAWET) Chen (2002) to solve such a stochastic optimization problem. The one-sided and two-sided randomized differences are used as the gradient estimate, and hence only two observations are required at each iteration of the algorithm. By using the powerful Trajectory-Subsequence (TS) method, we make the following improvements: 1) some restrict conditions on the noise process have been removed; 2) some boundness assumptions on the estimates have been removed; 3) the exact convergence result has been obtained.

The rest of the paper is organized as follows. The smoothed functional gradient algorithm with perturbations distributed according to the Gaussian distribution for stochastic optimization is proposed in Section 2. The main result, convergence analysis of the algorithm, is given in Section 3. Two numerical examples are presented in Section 4 to verify the validity of the algorithm. Finally, some concluding remarks are given in Section 5.

Notations: Let $\mathbb{R}^d$ be the d-dimensional real space, we write $\|x\|$ to denote the Euclidean norm of a vector x. Let $(\Omega, \mathscr{F}, \mathrm{P})$ be the basic probability space and $\mathrm{E}[\cdot]$ be the expectation operator.

2. SMOOTHED FUNCTIONAL STOCHASTIC OPTIMIZATION ALGORITHMS

Consider the following stochastic optimization problem

$$\min_{x\in\mathbb{R}^d} f(x) \tag{1}$$

i.e., to find the minimizer of an unknown function $f : \mathbb{R}^d \to \mathbb{R}$ for the case where the function $f(\cdot)$ itself rather than its gradient is observed and the observations are corrupted by noise. Here we consider the case that the noise in the observation of the value of the objective f is additive, i.e., the noisy observation of f at x_k is defined by

$$y_{k+1} = f(x_k) + \epsilon_{k+1}, \tag{2}$$

where $\{\epsilon_k, k \in \mathbb{N}\}$ is the observation noise process.

2.1 Motivation

For a continuous function $f : \mathbb{R}^d \to \mathbb{R}$, consider its *Gaussian approximation*

$$f_b(x) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} f(x+by)\, \mathrm{e}^{-\frac{1}{2}\|y\|^2}\, \mathrm{d}y, \tag{3}$$

where $b \geq 0$ is a smoothing parameter, which is also regarded as the bandwidth in nonparametric estimation. The probability density function of d-dimensional standard Gaussian distribution

$$K(x) = (2\pi)^{-d/2}\, \mathrm{e}^{-\|x\|/2}$$

plays the role of kernel function which is called the Gaussian kernel. Clearly, $\int_{\mathbb{R}^d} yK(y)\mathrm{d}y = 0$. For positive b the function f_b has better properties that f in general, for more details refer to Nesterov (2011).

For any positive b, the function f_b is differentiable. By introducing a new integration variable $z = x + by$, we can rewrite (3) in another form

$$f_b(x) = \frac{1}{b^d(2\pi)^{d/2}} \int_{\mathbb{R}^d} f(z)\, \mathrm{e}^{-\frac{1}{2b^2}\|z-x\|^2}\, \mathrm{d}z.$$

Then we obtain an expression for the gradient of f_b

$$\begin{aligned}\nabla f_b(x) &= \frac{1}{b^{d+2}(2\pi)^{d/2}} \int_{\mathbb{R}^d} f(z)\, \mathrm{e}^{-\frac{1}{2b^2}\|z-x\|^2}(z-x)\mathrm{d}z \\ &= \frac{1}{b(2\pi)^{d/2}} \int_{\mathbb{R}^d} f(x+by)\, \mathrm{e}^{-\frac{1}{2}\|y\|^2}\, y\mathrm{d}y \\ &= \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \frac{f(x+by)-f(x)}{b} y\, \mathrm{e}^{-\frac{1}{2}\|y\|^2}\, \mathrm{d}y\end{aligned} \tag{4}$$

Now it follows that

$$\nabla f_b(x) = \mathrm{E}_\eta\left[\frac{\eta}{b} f(x+b\eta)\right] = \mathrm{E}_\eta\left[\frac{\eta}{b}\big(f(x+b\eta) - f(x)\big)\right], \tag{5}$$

where the expectation above is taken with respect to the d-dimensional Gaussian kernel function $K(x)$ (i.e., the joint probability density function of d independent $\mathcal{N}(0,1)$-distributed random variables).

Now the one-sided gradient estimator suggested by is

$$\nabla f(x) \approx \frac{1}{n}\sum_{k=1}^{n} \frac{\eta_k}{b}\big(f(x+b\eta_k) - f(x)\big), \tag{6}$$

where $\eta_k = [\eta_{1,k}, \ldots, \eta_{d,k}]^T$ with $\eta_{i,k}, i = 1, \ldots, d$ being independent $\mathcal{N}(0,1)$-distributed random variables.

Note that the expression (4) can be rewritten into the following form

$$\begin{aligned}\nabla f_b(x) &= \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \frac{f(x)-f(x-by)}{b} y\, \mathrm{e}^{-\frac{1}{2}\|y\|^2}\, \mathrm{d}y \\ &= \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \frac{f(x+by)-f(x-by)}{2b} y\, \mathrm{e}^{-\frac{1}{2}\|y\|^2}\, \mathrm{d}y \\ &= \mathrm{E}_\eta\left[\frac{\eta}{2b}\big(f(x+b\eta) - f(x-b\eta)\big)\right]\end{aligned} \tag{7}$$

and subsequently the two-sided gradient estimator suggested by is

$$\nabla f(x) \approx \frac{1}{n}\sum_{k=1}^{n} \frac{\eta_k}{2b}\big(f(x+b\eta_k) - f(x-b\eta_k)\big). \tag{8}$$

See also Styblinski and Tang (1990) and Chin (1997).

2.2 SF Gradient Algorithm with SAAWET

Based on the one-sided smoothed function gradient estimator (6) or the two-sided one (8), we consider a stochastic approximation algorithm for the stochastic problem (1) by replacing the constant smoothing parameter b with a sequence of positive numbers b_k that tends to 0.

Suppose that the noisy observations of f at x_k, $x_k + b_k\eta_k$, and $x_k - b_k\eta_k$ are denoted by

$$y^0_{k+1} = f(x_k) + \epsilon^0_{k+1}, \tag{9}$$

$$y^+_{k+1} = f(x_k + b_k\eta_k) + \epsilon^+_{k+1}, \tag{10}$$

and

$$y_{k+1}^- = f(x_k - b_k\eta_k) + \epsilon_{k+1}^-, \tag{11}$$

respectively, where $\{\epsilon_k^0, k \in \mathbb{N}\}$, $\{\epsilon_k^+, k \in \mathbb{N}\}$, and $\{\epsilon_k^-, k \in \mathbb{N}\}$ are observation noise processes.

With a small abuse of notation, define y_{k+1} by the following equations

$$y_{k+1} = \frac{\eta_k(y_{k+1}^+ - y_{k+1}^0)}{b_k} \tag{12}$$

$$y_{k+1} = \frac{\eta_k(y_{k+1}^+ - y_{k+1}^-)}{2b_k} \tag{12'}$$

which can be regarded as estimates for the gradient of f at x_k. The noise ϵ_{k+1} is as follow

$$\epsilon_{k+1} = \epsilon_{k+1}^+ - \epsilon_{k+1}^0 \tag{13}$$

$$\epsilon_{k+1} = \epsilon_{k+1}^+ - \epsilon_{k+1}^-. \tag{13'}$$

It follows that

$$y_{k+1} = \frac{\eta_k(f(x_k + b_k\eta_k) - f(x_k))}{b_k} + \frac{\eta_k\epsilon_{k+1}}{b_k} \tag{14}$$

$$y_{k+1} = \frac{\eta_k(f(x_k + b_k\eta_k) - f(x_k - b_k\eta_k))}{2b_k} + \frac{\eta_k\epsilon_{k+1}}{2b_k}. \tag{14'}$$

Equation (12) is used for the one-sided gradient estimate and (12′) is used for the two-sided one. In the subsequent description the meaning of y_{k+1} is determined by whether the one-sided or the two-sided gradient estimate is used.

We now define a stochastic approximation algorithm with expanding truncations based on smoothed functional gradient estimate (6) or (8). Let $\{M_k\}$ be a sequence of positive numbers increasingly diverging to infinity and x^* be a fixed point in $\mathbb{R}^d$. Given any initial value x_0, the SF gradient algorithm with randomly varying truncations is defined by

$$\begin{aligned} x_{k+1} = &(x_k - a_k y_{k+1})\,\mathbb{1}_{\{\|x_k - a_k y_{k+1}\| \le M_{\sigma_k}\}} \\ &+ x^*\mathbb{1}_{\{\|x_k - a_k y_{k+1}\| > M_{\sigma_k}\}} \end{aligned} \tag{15}$$

$$\sigma_k = \sum_{i=0}^{k-1} \mathbb{1}_{\{\|x_i - a_i y_{i+1}\| > M_{\sigma_i}\}},\ \sigma_0 = 0 \tag{16}$$

where $\mathbb{1}_A$ denotes the indicator function of a set A. Clearly x_k is measurable with respective to $\mathscr{F}_k \triangleq \mathscr{F}_k^\epsilon \vee \mathscr{F}_{k-1}^\eta$, where $\mathscr{F}_k^\eta = \sigma(\eta_i, 1 \le i \le k)$ is the σ-algebra generated by the random variables $\{\eta_i, 1 \le i \le k\}$.

2.3 Assumptions

The following conditions are imposed on the algorithm.

A1. The objective function $f(\cdot)$ is continuously differentiable and Lipschitz continuous. There is an unique minimum of $f(\cdot)$ at x^0 that is the root of $\nabla f(\cdot)$ and $\nabla f(x) \neq 0$ for $x \neq x^0$. Further, x^* used in (15) is such that $f(x^*) > \sup_{\|x\|=c} f(x)$ for some c and $\|x^*\| \le c$.

A2. $a_k > 0$, $\sum_{k=1}^{\infty} a_k = \infty$, $\sum_{k=1}^{\infty} a_k^2 < \infty$;

A3. $b_k > 0$, $\lim_{k\to\infty} b_k = 0$;

A4. For any sufficiently large integer N

$$\lim_{T\to 0}\limsup_{k\to\infty}\frac{1}{T}\left\|\sum_{i=n_k}^{m(n_k,T_k)}\frac{a_i\eta_i\epsilon_{i+1}}{b_i}\mathbb{1}_{\{\|x_i\|\le L\}}\right\| = 0, \quad \forall T_k \in [0,T] \tag{17}$$

hold for any $\{n_k\}$ such that x_{n_k} converges, where

$$m(k,t) \triangleq \max\left\{m : \sum_{i=k}^{m} a_i \le t\right\}. \tag{18}$$

Remark 1. If the noise is such that

$$\sum_{k=1}^{\infty}\frac{a_k\eta_k\epsilon_{k+1}}{b_k} < \infty, \tag{19}$$

then A4 is satisfied. For example, if $\sum_{k=1}^{\infty} a_k^2/b_k^2 < \infty$ and the observation noise $\{\epsilon_k, k \in \mathbb{N}\}$ has the property that ϵ_{k+1} is independent of $\{\eta_i, 1 \le i \le k\}$ for each $k \in \mathbb{N}$ and satisfies one of the following two assumptions:

1) $\sup_k |\epsilon_k| \le \epsilon$ a.s., where ϵ is a random variable;

2) $\sup_k \mathrm{E}[\epsilon_k^2] < \infty$;

then condition A4 holds. For more details refer to Theorem 2 of Chen et al. (1999).

3. CONVERGENCE

Note that the observation y_{k+1} given by (12) or (12′) can be expressed as

$$y_{k+1} = \nabla f(x_k) + \xi_{k+1} \tag{20}$$

with

$$\xi_{k+1} = v_{k+1} + e_{k+1} + \varepsilon_{k+1} \tag{21}$$

where

$$\begin{aligned} v_{k+1} = &\frac{\eta_k}{b_k}\big(f(x_k + b_k\eta_k) - f(x_k)\big) \\ &- \mathrm{E}\left[\frac{\eta_k}{b_k}\big(f(x_k + b_k\eta_k) - f(x_k)\big)\middle|\mathscr{F}_k\right], \end{aligned} \tag{22}$$

$$e_{k+1} = \mathrm{E}\left[\frac{\eta_k}{b_k}\big(f(x_k + b_k\eta_k) - f(x_k)\big)\middle|\mathscr{F}_k\right] - \nabla f(x_k), \tag{23}$$

$$\varepsilon_{k+1} = \frac{\eta_k\epsilon_{k+1}}{b_k} \tag{24}$$

for y_{k+1} given by (12) and

$$\begin{aligned} v_{k+1} = &\frac{\eta_k}{2b_k}\big(f(x_k + b_k\eta_k) - f(x_k - b_k\eta_k)\big) \\ &-\mathrm{E}\left[\frac{\eta_k}{2b_k}\big(f(x_k+b_k\eta_k) - f(x_k-b_k\eta_k)\big)\middle|\mathscr{F}_k\right], \end{aligned} \tag{22'}$$

$$e_{k+1} = \mathrm{E}\left[\frac{\eta_k}{2b_k}\big(f(x_k+b_k\eta_k) - f(x_k-b_k\eta_k)\big)\middle|\mathscr{F}_k\right] - \nabla f(x_k), \tag{23'}$$

$$\varepsilon_{k+1} = \frac{\eta_k\epsilon_{k+1}}{2b_k} \tag{24'}$$

for y_{k+1} given by (12′).

Let the regression function be $g(x) = \nabla f(x)$, now the SF gradient algorithm given by (15)-(16) turns to be a standard stochastic approximation algorithm with expanding truncations, where the noise ξ_{k+1} expressed by (21) is composed of the structural error v_{k+1}, e_{k+1} and the random noise ε_{k+1} caused by inaccuracy of observations. The

major difficulty of analyzing algorithm (15)-(16) consists in that the noise v_{k+1} is state-dependent.

To establish the strong consistency of the algorithm, we need the following three lemmas.

Lemma 1. Assume A1-A4 hold. Then there is an Ω_0 with $P\{\Omega_0\} = 1$ such that for any $\omega \in \Omega_0$ and any bounded subsequence $\{x_{n_k}\}$ of $\{x_k\}$, there exists an integer k_0 such that for all $k \geq k_0$,

$$\|x_i - x_{n_k}\| \leq ct, \ \forall i : n_k \leq i \leq m(n_k, t), \ \forall t \in [0, T], \quad (25)$$

if T is small enough, where $m(n_k, t)$ is defined by (18).

Lemma 2. Assume A1-A4 hold. There is an $\Omega' \subset \Omega_0$ with $\mathrm{P}\{\Omega'\} = 1$, such that if $\omega \in \Omega'$ and if $\{x_{n_k}\}$ is a bounded subsequence of $\{x_k\}$, then

$$\lim_{T\to 0} \sup_{t\in[0,T]} \limsup_{k\to\infty} \frac{1}{T} \left\| \sum_{i=n_k}^{m(n_k,t)} a_i v_{i+1} \right\| = 0. \quad (26)$$

Lemma 3. Assume A1-A4 hold. Then almost surely

$$\lim_{k\to\infty} e_{k+1} = 0. \quad (27)$$

Proof. We prove (27) hold for e_{k+1} given by (23), while the case for e_{k+1} given by (23′) can be proven by a similar argument. Note that $\eta_k = [\eta_{1,k}, \ldots, \eta_{d,k}]^T$ is a vector of independent $\mathcal{N}(0,1)$-distributed random variates. By using the Taylor series expansion of $f(x_k + b_k\eta_k)$ around x_k, we have

$$f(x_k + b_k\eta_k) - f(x_k) = b_k\eta_k^T\nabla f(x_k) + o(b_k),$$

Thus,

$$\begin{aligned} &\mathrm{E}\left[\frac{\eta_k}{b_k}\big(f(x_k + b_k\eta_k) - f(x_k)\big)\Big|\, \mathscr{F}_k\right] \\ =&\mathrm{E}\left[\eta_k\eta_k^T\nabla f(x_k)\big|\, \mathscr{F}_k\right] + o(1). \end{aligned}$$

Now,

$$\mathrm{E}\left[\eta_k\eta_k^T\nabla f(x_k)\big|\, \mathscr{F}_k\right] = \mathrm{E}\left[\eta_k\eta_k^T\big|\, \mathscr{F}_k\right]\nabla f(x_k) = \nabla f(x_k).$$

The main result of the paper is the following convergence theorem.

Theorem 1. Assume A1-A4 hold. Let $\{x_k\}$ be given by (15)-(16) with any initial value. Then

$$\lim_{k\to\infty} x_k = x^0, \ \text{a.s.} \quad (28)$$

Proof. We apply Theorem 2.2.1 of Chen (2002) to prove the theorem.

Let us check Conditions A2.2.1-A2.2.4. Condition A2.2.1 holds by A2. By taking $v(x) = -\nabla f(x)$ and noticing that $J = \{x^0\}$, it follows that Condition A2.2.2 is automatically satisfied. By Lemma 1-3 and A4, we can verify that the noise ξ_{k+1} given by (21) satisfies the requirements.

4. NUMERICAL SIMULATION

Two numerical examples are given. First, let $f(x) = x^2 - 4x + 4$, which has a unique minimum at $x = 2$. The noise processes $\{\epsilon_k^0, k \in \mathbb{N}\}$, $\{\epsilon_k^+, k \in \mathbb{N}\}$, and $\{\epsilon_k^-, k \in \mathbb{N}\}$ are independent white Gaussian $\mathcal{N}(0, 0.01)$ processes. The sequences used in the algorithms are chosen as follow: $a_k = 1/k$, $b_k = 1/k^{0.2}$, $M_k = 2^k$. The graph of $\{x_k, 0 \leq k \leq 500\}$ that use one-sided or two-sided gradient estimate is showed in Fig. 1. As a comparison, the graph of $\{x_k, 0 \leq k \leq 1000\}$

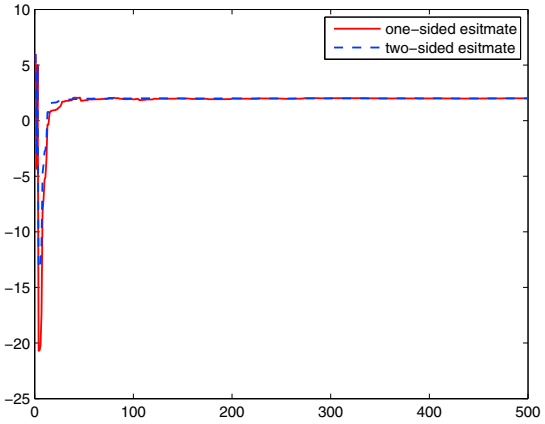

Fig. 1. The path of SF estimates $\{x_k\}$ with $x^* = 5$ and initial value $x_0 = 6$, where $f(x) = x^2 - 4x + 4$

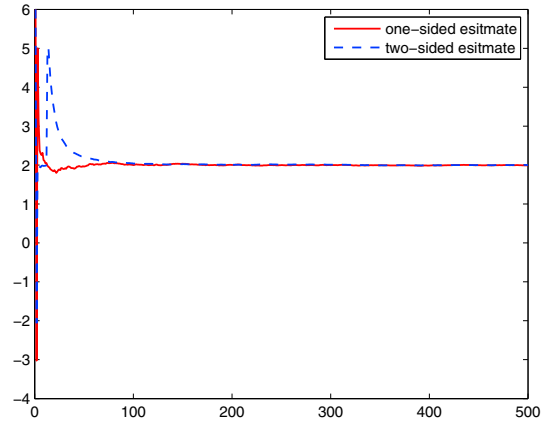

Fig. 2. The path of KW estimates $\{x_k\}$ with $x^* = 5$ and initial value $x_0 = 6$, where $f(x) = x^2 - 4x + 4$

that uses the KW algorithm with randomized differences proposed by Chen et al. (1999) is given in Fig. 2, where the random variable Δ_k^i used in the algorithm is uniformly distributed on $[-1, -0.5] \cup [0.5, 1]$, the other parameters $a_k = 1/k$, $c_k = 1/k^{0.2}$, $M_k = 2^k$ are the same as algorithm (15)-(16).

The second example uses the function $f(x) = (x^2-1)^3+1$, which have three minimal points $x = 0, \pm 1$ with $x = 0$ being its minimum point, see Fig. 3. The noise processes $\{\epsilon_k^0, k \in \mathbb{N}\}$, $\{\epsilon_k^+, k \in \mathbb{N}\}$ and $\{\epsilon_k^-, k \in \mathbb{N}\}$ are independent white Gaussian $\mathcal{N}(0, 0.01)$ processes. The sequences used in the algorithms are chosen as follow: $a_k = 1/k$, $b_k = 1/k^{0.2}$, $M_k = 2^k$. The graphs of $\{x_k, 0 \leq k \leq 500\}$ given by the proposed SF gradient algorithm that use one-sided or two-sided gradient estimate is showed in Fig. 4.

5. CONCLUSION

Based on stochastic approximation algorithm with expanding truncations, smoothed functional gradient algorithm with Gaussian perturbations is considered for stochastic optimization problem with additive noise. The one-sided and two-sided randomized differences are used

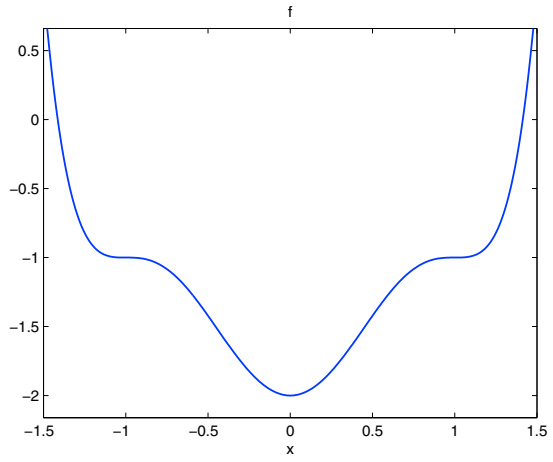

Fig. 3. Plot of function $f(x) = (x^2 - 1)^3 + 1$

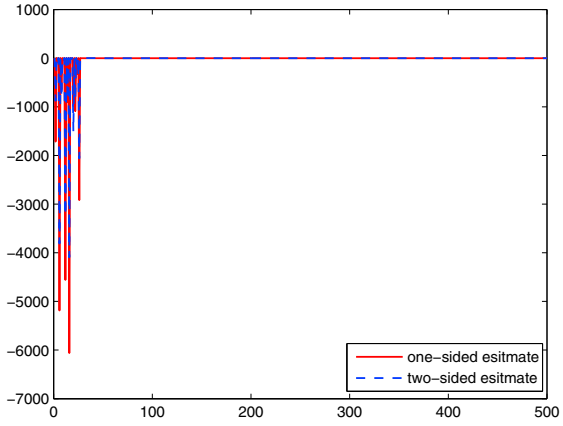

Fig. 4. The path of SF estimates $\{x_k\}$ with $x^* = 5$ and initial value $x_0 = 2$, where $f(x) = (x^2 - 1)^3 + 1$

as the gradient estimate, and hence only two observations are required at each iteration of the algorithm. For the convergence analysis, the TS method is used rather than the classical ODE method and hence some restrict conditions on the noise process and some boundness assumptions on the estimates have been removed, and finally the exact convergence result has been obtained.

It is of interest to consider the convergence rate of the SF gradient algorithm. Extension to the non-additive noise case is another interesting topic. The non-smooth stochastic optimization based on the SF gradient algorithm is of particular importance for further consideration.

REFERENCES

Bhatnagar, S., Prasad, H.L., and Prashanth, L.A. (2012). *Stochastic Recursive Algorithms for Optimization: Simultaneous Perturbation Methods*. Springer, London.

Chen, H.F., Duncan, T.E., and Pasik-Duncan, B. (1999). A Kiefer-Wolfowitz algorithm with randomized differences. *IEEE Transactions on Automatic Control*, 44(3), 442–453.

Chen, H.F. (2002). *Stochastic Approximation and Its Applications*. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Chin, D.C. (1997). Comparative study of stochastic algorithms for system optimization based on gradient approximations. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 27(2), 244–249.

Katkovnik, V. and Kulchitsky, Y. (1972). Convergence of a class of random search algorithms. *Automation and Remote Control*, 33(8), 1321–1326.

Kiefer, J. and Wolfowitz, J. (1952). Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3), 462–466.

Nesterov, Y. (2011). Random gradient-free minimization of convex functions. Technical report, Dept. Center Oper. Res. Econ., Univ. Catholique de Louvain, Louvain, Belgium.

Spall, J.C. (2000). Adaptive stochastic approximation by the simultaneous perturbation method. *IEEE Transactions on Automatic Control*, 45(10), 1839–1853.

Styblinski, M.A. and Tang, T.S. (1990). Experiments in nonconvex optimization: Stochastic approximation with function smoothing and simulated annealing. *Neural Networks*, 3(4), 467–483.