Available online at www.sciencedirect.com
ScienceDirect
European Journal of Operational Research 181 (2007) 1174–1188
EUROPEAN JOURNAL OF OPERATIONAL RESEARCH
www.elsevier.com/locate/ejor

Simultaneous perturbation stochastic approximation of nonsmooth functions

Vaida Bartkutė [a,*], Leonidas Sakalauskas [b]

[a] *Institute of Mathematics and Informatics, Akademijos 4, Vilnius, Lithuania*
[b] *Vilnius Gediminas Technical University, Sauletekio al 11, Vilnius, LT-10223, Lithuania*

Received 2 November 2004; accepted 2 September 2005
Available online 3 May 2006

Abstract

A simultaneous perturbation stochastic approximation (SPSA) method has been developed in this paper, using the operators of perturbation with the Lipschitz density function. This model enables us to use the approximation of the objective function by twice differentiable functions and to present their gradients by volume integrals. The calculus of the stochastic gradient by means of this presentation and likelihood ratios method is proposed, that can be applied to create SPSA algorithms for a wide class of perturbation densities. The convergence of the SPSA algorithms is proved for Lipschitz objective functions under quite general conditions. The rate of convergence $O\left(\frac{1}{k^{\gamma}}\right)$, $1 < \gamma < 2$ of the developed algorithms has been established for functions with a sharp minimum, as well as the dependence of the rate of convergence is explored theoretically as well as by computer simulation. The applicability of the presented algorithm is demonstrated by applying it to minimization of the mean absolute pricing error for the calibration of the Heston stochastic volatility model.

Keywords: Simultaneous perturbation; Stochastic approximation; Lipschitz function; Stochastic gradient; Monte-Carlo method

1. Introduction

Application of stochastic approximation (SA) to nonsmooth optimization is both a theoretical and practical problem. Computational properties of SA algorithms are mainly determined by the approximation approach to the stochastic gradient (Robins and Monro, 1951; Kiefer and Wolfowitz, 1952; Blum, 1954; Dvoretzky, 1956; Yudin, 1965; Wasan, 1969; Ermoliev, 1976; Michalevitch et al., 1987; Ermoliev et al., 1995; Kushner and Yin, 2003, etc.). Thus, it is of interest to consider simultaneous perturbation stochastic approximation (SPSA) methods, in which values of the function for estimating the stochastic gradient are required only at one or several points. The SPSA algorithms were considered by several authors who used various smoothing operators. SPSA methods, uniformly smoothing the function in an n-dimensional hypercube, are described in Michalevitch et al. (1987). Spall (1992) proposed the SPSA algorithm with the Bernoulli perturbation model and indicated the computational efficiency of SPSA as compared with the standard finite

* Corresponding author. Tel.: +370 5 2109323; fax: +370 5 2729209.
E-mail addresses: vaidaba@one.lt (V. Bartkutė), sakal@ktl.mii.lt (L. Sakalauskas).

0377-2217/$ - see front matter
doi:10.1016/j.ejor.2005.09.052

difference approximation. The convergence and asymptotic behaviour of this algorithm were established in the class of differentiable functions.

Application of the SPSA algorithms to nondifferentiable functions is of particular theoretical and practical interest. In this paper, we confine ourselves to objective functions from the Lipschitz class. The class of these functions is broad enough involving many cases of continuous functions, considered in mathematical programming. In addition, Lipschitz functions have many essential properties allowing us to define a stochastic gradient and create converging numerical procedures. We consider the SPSA algorithms that introduce perturbation models described in terms of density functions, which are also Lipschitzian. This model enables us to extend a set of perturbation operators involving a lot of practical cases, for instance, perturbation densities of a piecewise linear shape, presented by polygons, etc. Besides, this assumption makes it possible to approximate the objective function by twice differentiable smoothed functions, as well as applying the sampling and likelihood ratios method (Rubinstein and Shapiro, 1993).

The paper is arranged in the following way: Section 2 presents general assumptions and definitions necessary for the convergence of the SPSA algorithms. The convergence proof for the algorithm is given in Section 3, while the convergence rate is estimated in Section 4. Section 5 provides computer modelling results with test functions, while Section 6 presents an example of volatility estimation by the SPSA algorithm.

2. General assumptions and definitions

We consider the multidimensional continuous optimization problem:

$$f(x) \to \min, \tag{1}$$

where the objective function $f : \Re^n \to \Re$ is Lipschitzian. Let $\partial f(x)$ be the generalized gradient (GG), i.e., the Clarke subdifferential (Clarke, 1983) of this function.

Assume X^* to be the set of stationary points:

$$X^* = \{x | 0 \in \partial f(x)\}$$

and F^* the set of stationary function values:

$$F^* = \{z | z = f(x),\ x \in X^*\}.$$

Denote the gradient of function f at the point $x \in \Re^n$ by $\frac{\partial f(x)}{\partial x}$, if it is differentiable in the usual sense at this point.

Let the objective function be satisfying the following assumptions:

(A) it is Lipschitzian with constant K,
(B) it is bounded from below,
(C) $\underline{\lim}_{x\to\infty} \inf_{g\in\partial f(x)} \|g\| > 0$, i.e., the generalized gradient cannot be zero at infinity,
(D) for all $\varepsilon > 0$ we can find $\delta > 0$ such that for all y, $\|y - x\| \leqslant 2\delta$, where the function $f(x)$ is differentiable in the usual sense, the inequality $\min_{z\in\partial f(x)} \left\| \frac{\partial f(y)}{\partial y} - z \right\| \leqslant \frac{\varepsilon}{2}$ holds uniformly in x,
(E) the set of stationary function values F^* does not contain inner points.

Remark 1. One can see that the sets X^* (and F^*, respectively) are bounded according to the assumption on the gradient behaviour at infinity.

It follows by virtue of the Lipschitz condition that $\sup_{\substack{g\in\partial f(x)\\ x\in\Re^n}} \|g\| \leqslant K$.

Regularization concept is a well-known technique for constructing optimization algorithms which is considered by many authors (see Yudin, 1965; Donoghue, 1969; Spall, 1992; Rubinstein and Shapiro, 1993; Ermoliev et al., 1995; Kushner and Yin, 2003, etc.). In this case, we succeed in approximating gradients of smoothed functions by stochastic estimators and creating the methods where the calculation of gradients is not required. Sometimes, regularized gradients are called mollifier subgradients (Ermoliev et al., 1995).

In general, this idea may be expressed as follows: a sequence of smooth functions $f(x,\sigma)$ is introduced which converges to $f(x)$, as $\sigma \to 0$. Then, the solution of (1) is obtained by minimizing this sequence, if the smoothing parameter σ is changed in an appropriate way.

The sequence of smoothed functions may usually be introduced using the expectation:

$$\bar{f}(x,\sigma) = Ef(x+\sigma\xi),$$

where $\xi \in \Omega$ is a random vector from the probability space (Ω,Σ,P), while $\sigma \geqslant 0$ is the value of the smoothing parameter. The properties of regularized gradients of Lipschitz functions obtained in this way are exhaustively described in Michalevitch et al. (1987), while the mollifier subgradients of semicontinuous functions are explored by Ermoliev et al. (1995). We assume the measure P to be absolutely continuous, i.e., it can be defined by a certain density function, which is also Lipschitzian. Thus, assume $\xi \in \Omega$ to be a random vector from the probability space (Ω,Σ,P), where the measure P is defined by the density function $p:\Omega \to \Re_+$, which satisfies the Lipschitz condition with a certain constant. We consider that Ω is actually identical to $\Re^n$ and Σ to the Borel σ-algebra in this space. Let us denote the support of measure P as follows: $\mathrm{dom}(P) = \{y|p(y) > 0\}$. Let $\partial p(y)$ be a generalized gradient of p.

Let us express the smoothed function through multivariate integrals (see also Rubinstein and Shapiro, 1993; Ermoliev et al., 1995; Sakalauskas, 2002):

$$\bar{f}(x,\sigma) = \int_\Omega f(x+\sigma y)\cdot p(y)\,\mathrm{d}y = \frac{1}{\sigma^n}\int_\Omega f(y)\cdot p\Big(\frac{y-x}{\sigma}\Big)\,\mathrm{d}y. \tag{2}$$

It can be easily shown that $\bar{f}(x,\sigma)$ is also a Lipschitz function in σ:

$$|\bar{f}(x,\sigma_1) - \bar{f}(x,\sigma_2)| \leqslant E|f(x+\sigma_1\xi) - f(x+\sigma_2\xi)| \leqslant C\cdot K\cdot|\sigma_1-\sigma_2|, \tag{3}$$

where $C = E\|\xi\|$.

Remark 2. Generalized gradients ∂p and ∂f are set-valued mappings, defined in the whole space $\Re^n$, whose values, in general, are uniformly bounded convex closed sets. It is well known from the Rademacher theorem that Lipschitz functions are differentiable in the usual sense almost everywhere, and the Borel measure of the set is zero, where the generalized gradients are set-valued (see references in Michalevitch et al., 1987). Thus, the integrals over the absolutely continuous measure, where GG is included, can be defined unambiguously as a.s. point valued expressions.

Thus, formal differentiation of the first integral in (2) gives us the expression

$$\bar{g}(x,\sigma) = \frac{\partial\bar{f}(x,\sigma)}{\partial x} = E\partial f(x+\sigma\xi), \tag{4}$$

where some value of the GG of ∂f is taken when it is set-valued.

Let us consider a gradient approximation by the likelihood ratios (LR) method (Rubinstein and Shapiro, 1993; Ermoliev et al., 1995) that allows us to introduce twice differentiable smoothed functions.

Lemma 1. *Assume that Ψ is an absolutely continuous measure with the bounded density function $\psi:\Omega\to\Re_+$, the support of which is identical to that of P: $dom(\Psi) = \{y|\psi(y)>0\} \equiv dom(P)$. Assume that $\int_{\mathrm{dom}(P)}\|y\|^2\cdot\frac{\|\partial p(y)\|^2}{\psi(y)}\,\mathrm{d}y = A < \infty$, $\int_{\mathrm{dom}(P)}\|\partial p(y)\|\cdot\mathrm{d}y = L < \infty$. Then, the gradient of the smoothed function (2) is a.s. point valued mapping which can be expressed as the expectation*:

$$\bar{g}(x,\sigma) = E(g(x,\sigma,\xi)), \tag{5}$$

$g(x,\sigma,\xi)$ is the stochastic gradient expressed by

$$g(x,\sigma,\xi) = \frac{(f(x+\sigma\xi)-f(x))}{\sigma}\cdot\frac{\partial p(\xi)}{\psi(\xi)}, \tag{6}$$

where some value of the generalized gradient $\partial p(\xi)$ is taken when it is set-valued (*according to Remark 2*),

$$E(\|g(x,\sigma,\xi)\|)^2 \leqslant K^2\cdot A. \tag{7}$$

Hessian of the smoothed function (2) is

$$V(x,\sigma)=\frac{\partial^2\bar{f}(x,\sigma)}{\partial x^2}=\frac{1}{\sigma}\cdot E\left(\frac{\partial f(x+\sigma\xi)\cdot(\partial p(\xi))^{\mathrm{T}}}{\psi(\xi)}\right), \tag{8}$$

where $\|V(x,\sigma)\|\leqslant\frac{K\cdot L}{\sigma}$.

The proof is given in Appendix A.

Thus, the smoothed function (2) is an a.s. point valued differentiable in the usual sense in general. In the case where $\Psi=P$, we have

$$\bar{g}(x,\sigma)=E\left(\frac{(f(x+\sigma\xi)-f(x))}{\sigma}\cdot\partial\ln(p(\xi))\right), \tag{9}$$

where the GG of $\ln(p(y))$ is defined as follows:

$$\partial\ln(p(y))=\begin{cases}\frac{\partial p(y)}{p(y)}, & \text{if } y\in\mathrm{dom}(P),\\ 0, & \text{if } y\notin\mathrm{dom}(P).\end{cases}$$

Example 1. If the perturbation operator is expressed in terms of the Gaussian density:

$$p(y)=\left(\frac{1}{2\pi}\right)^{\frac{n}{2}}\exp\left(-\frac{\|y\|^2}{2}\right),$$

then, according to (9), we have the stochastic gradient as follows:

$$g(x,\sigma,\xi)=\frac{(f(x+\sigma\xi)-f(x))\cdot\xi}{\sigma}.$$

Example 2. Let a perturbation operator be defined as density distributed in the unit ball:

$$p(y)=\begin{cases}W\cdot(1-\|y\|), & \text{if } \|y\|\leqslant 1,\\ 0, & \text{if } \|y\|>1,\end{cases}$$

where $W=\frac{1}{\int_{\|y\|\leqslant 1}(1-\|y\|)\,\mathrm{d}y}=\frac{n\cdot(n+1)\cdot\Gamma\left(\frac{n}{2}\right)}{(2\pi)^{\frac{n}{2}}}$. Undoubtedly, this function is Lipschitzian. It is easy to see that the stochastic gradient can be expressed as

$$g(x,\sigma,\xi)=\frac{(f(x+\sigma\xi)-f(x))\cdot\xi}{\sigma\cdot\|\xi\|},$$

when the likelihood ratios density of ξ is uniformly distributed in the unit ball:

$$\psi(y)=\begin{cases}\frac{1}{V_n}, & \text{if } \|y\|\leqslant 1,\\ 0, & \text{if } \|y\|>1,\end{cases}$$

where V_n is the volume of the n-dimensional ball.

The next lemma considers the approximation of the function $f(x)$ by the smoothed functions $\bar{f}(x,\sigma)$, as $\sigma\to 0$.

Lemma 2. *Let $f(x)$ be Lipschitzian with constant K, and, besides, for all $\varepsilon>0$ we can find $\delta>0$ such that for all y, $\|y-x\|\leqslant 2\delta$, where the function $f(x)$ is differentiable, the inequality $\min\limits_{z\in\partial f(x)}\left\|\frac{\partial f(y)}{\partial y}-z\right\|\leqslant\frac{\varepsilon}{2}$ holds uniformly with respect to x. Let $\sigma_i\to 0$ and assume $E\|\xi\|^2<\infty$. Then for any $\varepsilon>0$ we can find k such that for all $i\geqslant k$ and for all y, $\|y-x\|\leqslant\delta$, the inequality holds uniformly with respect to x: $\min\limits_{z\in\partial f(x)}\|\bar{g}(y,\sigma_i)-z\|\leqslant\varepsilon$.*

Corollary. *If $\{x^k\}\to x$, $\bar{g}(x^k,\sigma_k)\to g$, $\sigma_k\to 0$, $k\to\infty$, then $g\in\partial f(x)$.*

The proof is given in Appendix A.

Lemmas 1 and 2 make it possible to express smoothed gradients by volume integrals which can be numerically implemented by the Monte-Carlo method. Note that, as it follows from (7), formula (6) requires only two function values to be computed for the estimation of the gradient with bounded variance that does not depend on σ.

3. Convergence of the method

Let us introduce a sequence

$$x^{k+1} = x^k - \rho_k \cdot g^k, \quad k = 1, 2, \ldots, \tag{10}$$

where $g^k = g(x^k, \sigma_k, \xi_{k+1})$ is the value of the stochastic gradient estimated by (6) at the point $x^k, \xi_1, \xi_2, \ldots$ are independent copies of ξ, ρ_k is a scalar multiplier, σ_k is the value of the perturbation parameter in the iteration k, and x^0 is the initial point. As we see from the following theorem, the sequence (10) converges almost surely (a.s.) to the set X^* under certain conditions.

Theorem 1. *If the function $f : \Re^n \to \Re$ satisfies conditions (A)–(E) and*

$$\int_{\mathrm{dom}(\Psi)} \|y\|^4 \cdot \frac{\|\partial p(y)\|^4}{(\psi(y))^3}\, \mathrm{d}y < \infty,$$

$$\int_{\mathrm{dom}(P)} \|\partial p(y)\| \cdot \mathrm{d}y < \infty, \quad E\|\xi\|^2 < \infty,$$

$\{\rho_k\}$, $\{\sigma_k\}$ are sets of non-negative numbers, such that

$$\sum_{k=1}^{\infty} \rho_k = \infty, \quad \sum_{k=1}^{\infty} \rho_k^2 < \infty, \quad \sigma_k \to 0, \quad |\sigma_k - \sigma_{k-1}|/\rho_k \to 0, \quad \frac{\rho_k}{\sigma_k} \to 0,$$

then $\lim_{k\to\infty} x^k \in X^*$ a.s.

The proof of the theorem is based on the facts of SA convergence (see Wasan, 1969; Ermoliev, 1976; Gupal and Norkin, 1977; Nurminski, 1979; Michalevitch et al., 1987; Dupac, 1988; Kushner and Yin, 2003, etc.). Since it contains the peculiarities related with the considered SPSA model, we present it in detail in Appendix A. Examples of sequences $\{\rho_k\}$, $\{\sigma_k\}$ satisfying the condition G of Theorem 1, are worth mentioning:

$$\rho_k = \min\left(c, \frac{a}{k}\right), \quad \sigma_k = \min\left(d, \frac{b}{k^\beta}\right),$$

where $a, b, c, d > 0$ are certain constants (see Sections 4 and 5 and Wasan, 1969; Michalevitch et al., 1987).

4. Study of the rate of convergence

The rate of convergence of stochastic approximation for differentiable objective functions has been considered by many authors (see Wasan, 1969; Nurminski, 1979; Michalevitch et al., 1987; Kushner and Yin, 2003, etc.). The rate of convergence $\mathrm{O}\left(\frac{1}{k}\right)$ has been established for twice continuously differentiable functions computed without noise (see e.g., Dupac, 1988). In the case where the objective function is differentiable and computed with a stochastically distributed error, the rate of convergence decreases: $\mathrm{O}\left(\frac{1}{k^\gamma}\right)$, $0 < \gamma < 1$ (see Poliak, 1987; Granichin and Poliak, 2003; Kushner and Yin, 2003, etc.). On the other hand, it is known that, in certain cases, the rate of convergence for functions with a sharp minimum can be higher than that for smoothed functions (Poliak, 1987). Similarly, we show further that the upper bound of the convergence rate with $1 \leqslant \gamma < 2$ can be achieved in SPSA algorithms. Our consideration of the rate of convergence of the SPSA methods is based on the study of two processes: first, on the convergence to the minimum of a smoothed function (2) of the optimizing sequence (10), second, on tending of the smoothed function minimum to the minimum of the objective function, as $\sigma_k \to 0$.

Let us consider the Lipschitz function $f(x)$ having a sharp minimum at the point x^* with a certain constant $\mu > 0$ (see e.g., Poliak, 1987):

$$f(x) - f(x^*) \geqslant \mu\|x - x^*\| \quad \text{for each } x \in \Re^n. \tag{11}$$

Denote by $h : \Re^n \to \Re$ the Clarke generalized directional derivative at the point $x^* \in \Re^n$, i.e., the mapping such that (Clarke, 1983; Rockafellar, 1979)

$$h(y) = \max_{z \in \partial f(x^*)} (z \cdot y), \quad y \in \Re^n. \tag{12}$$

The next function is introduced to describe properties of the objective function in the neighbourhood of the sharp minimum:

$$\bar{h}(x) = E(h(x + \xi)). \tag{13}$$

Remark 3. Since $h(\cdot)$ is a convex Lipschitz function (see Lemma 4 in Appendix A), the function $\bar{h}(x)$, being an expectation of the latter, is twice differentiable and, thus, its derivatives can be evaluated, using the results from the previous sections.

Denote the functions $h(x,\sigma) = \frac{f(x^* + \sigma\cdot(x - x^*)) - f(x^*)}{\sigma}$ and $\bar{h}(x,\sigma) = E(h(x + \xi, \sigma))$.

In general, the minimum point x^* of the objective function $f(x)$ differs from the minimum point y^* of the function $\bar{h}(x)$. The relation between the minimum point x^*_σ of the smoothed function and x^* of the objective function can be established: $x^*_\sigma = x^* + \sigma(y^* - x^*) + \mathrm{O}(\sigma)$ (Lemmas 5 and 6 in Appendix A).

The rate of convergence of the SPSA methods for functions with a sharp minimum is studied in the next theorem.

Theorem 2. *Let the conditions of Theorem 1 be valid and, besides, the function $f(x)$ is semismooth and has a single sharp minimum with the constant $\mu > 0$ at the point x^*. Assume the condition $\lim_{\sigma\to 0} \frac{\partial^2 \bar{h}(y,\sigma)}{\partial y^2} = \frac{\partial^2 \bar{h}(y)}{\partial y^2} > 0$ be valid for all y taken from the certain small neighbourhood of the point y^*. If*

$$\rho_k = \frac{a}{k}, \ a > 0, \quad \sigma_k = \frac{b}{k^\beta}, \ b > 0, \ 0 < \beta < 1, \quad \frac{a}{b} > \frac{1+\beta}{2 \cdot H},$$

then

$$E\left(\|x^{k+1} - x^*_{\sigma_{k+1}}\|^2\right) \leqslant \frac{A \cdot K^2 \cdot a \cdot b}{H} \cdot \frac{1}{k^{1+\beta}} + \mathrm{O}\left(\frac{1}{k^{\frac{2aH}{b}}}\right), \tag{14}$$

as $k \to \infty$, where x^k is defined according to (10), y^* is the minimum point of $\bar{h}(x)$,

$$H = \left\| \left(\frac{\partial^2 \bar{h}(y)}{\partial y^2}\bigg|_{y=y^*} \right)^{-1} \right\|^{-1}, \quad A = \int_{\mathrm{dom}(P)} \|y\|^2 \cdot \frac{\|\partial p(y)\|^2}{\psi(y)}\,\mathrm{d}y.$$

The proof is given in Appendix A.

Let us compare the latter constants for different perturbation operators in the next examples.

Assume $f(x) = \mu \cdot \|x\|$.

Example 3. If the smoothing operator is Gaussian, then the constant is

$$\frac{A \cdot K^2}{H} = \frac{K^2}{\mu^2} \cdot \frac{\int_{\Re^n} \|y\|^2 \cdot \|\partial \ln(p(y))\|^2 \cdot p(y)\,\mathrm{d}y}{\int_{\Re^n} \|y\| \cdot \left(\frac{\partial^2 p(y)}{\partial y^2} - \left(\frac{\partial p(y)}{\partial y}\right) \cdot \left(\frac{\partial p(y)}{\partial y}\right)^{\mathrm{T}} \right)_{1,1} \cdot p(y)\,\mathrm{d}y} = \frac{K^2 \cdot E\|\xi\|^4}{\mu^2 \cdot E(\|\xi\| \cdot (\xi_1^2 - 1))}$$

$$= \frac{K^2 \cdot \Gamma\left(\frac{n}{2}\right)}{\mu^2 \cdot \Gamma\left(\frac{n+1}{2}\right)} \cdot \frac{n^2 \cdot (n+2)}{\sqrt{2}} = \frac{K^2 \cdot n^{\frac{5}{2}}}{\mu^2} + \mathrm{O}(1). \tag{15}$$

For the smoothing operator with the density from Example 2 the constant is

$$\frac{A\cdot K^2}{H}=\frac{K^2}{\mu^2}\cdot\frac{\int_{\|y\|\leqslant 1}\|y\|^2\,\mathrm{d}y}{\int_{\|y\|\leqslant 1}\frac{y_1^2}{\|y\|^2}\,\mathrm{d}y}=\frac{K^2\cdot n^2}{\mu^2\cdot(n+2)}=\frac{K^2\cdot n}{\mu^2}+\mathrm{O}(1). \tag{16}$$

Indeed, the rate of convergence of the SPSA method largely depends on the smoothing operator. The question about the Lipschitzian smoothing operator with optimal properties has not been studied yet and, thus, it may be a subject of the future research. Note that the convergence rate in (14) follows from the stochastic gradient expression (6), when there are no noises in computation of the objective function. If the function is computed with stochastically distributed errors, then the rate of convergence should essentially decrease (see Poliak, 1987; Granichin and Poliak, 2003, etc.).

5. Computer modelling

The proposed method was studied by computer modelling. We considered a class of test functions $f=\sum_{k=1}^{n}a_k|x_k|+M$, where a_k were randomly and uniformly generated in the interval $[\mu,K]$, $K>\mu>0$. The samples of $T=500$ test functions were generated. The test functions were minimized by the SPSA algorithm, using the stochastic gradient (6), which requires only two function values to be computed at each iteration, and comparing it with that obtained by applying the standard difference method to formula (4), which, in its turn, requires that $n+1$ function values be computed (Michalevitch et al., 1987; Granichin and Poliak, 2003, etc.). The multipliers were chosen in SPSA algorithm according to the condition G of Theorem 1 as follows: $\rho_k=\min\left(c,\frac{a}{k}\right)$, $\sigma_k=\min\left(d,\frac{b}{k^\beta}\right)$, where $a=0.25$, $b=1$, $c=n\cdot 0.005$, $d=d_0\cdot 0.01$, $d_0=\sqrt{\frac{(n+2)\cdot(n+3)}{n\cdot(n+1)}}$ are certain empirically chosen constants. These constants were chosen for the difference method as follows: $a=0.25$, $b=1$, $c=0.005$, $d=d_0\cdot 0.01$. The parameters of the simulated function class were as follows: $\mu=2$, $K=5$. In Fig. 1 we can see sampling dependences in the logarithmic scale of the Monte-Carlo estimate of $\Delta_k=E\|x^k-x^*\|^2$ on the number of computation of function values. The SPSA algorithm presented here appeared to be more efficient for small n than the standard finite difference approach. When the dimensionality of the task increases, the issue of the efficiency of SPSA compared to the finite difference approach requires some additional investigation. Besides, the theoretical and empirical least squares estimates of the rate of convergence by the Monte-Carlo method are presented in Table 1.

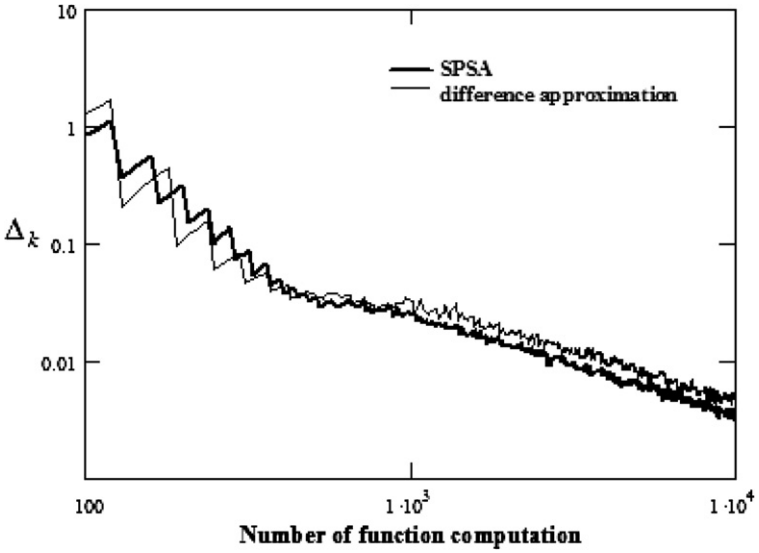

Fig. 1. The rate of convergence for SPSA and the standard difference approximation methods, $n=2$.

Table 1
Empirical and theoretical rates of convergence

	$\beta = 0.5$	$\beta = 0.75$	$\beta = 0.9$
Theoretical	1.5	1.75	1.9
Empirical			
$n = 2$	1.45509	1.72013	1.892668
$n = 4$	1.41801	1.74426	1.958998

As we can see from the figure and the table, computer simulation corroborates the theoretically defined convergence rates.

6. Volatility estimation by the SPSA algorithm

Financial data analysis, as well as risk analysis in the market research and management is often related to the implied and realized volatility. Let us consider the application of SPSA to the minimization of the mean absolute pricing error for the parameter estimation in the Heston stochastic volatility model. The Heston stochastic volatility model is a direct expansion of the classical Black–Scholes case and provides a natural framework for theoretical option pricing because a closed-form solution can be derived by means of Fourier inversion techniques for a wide class of models (Heston, 1993). In this model option pricing biases can be compared to the observed market prices, based on the latter solution and pricing error. We consider the mean absolute pricing error (MAE) defined as

$$\mathrm{MAE}(\kappa, \sigma, \rho, r, q, \theta) = \frac{1}{N} \sum_{i=1}^{N} \left| C_i^{\mathrm{H}}(\kappa, \sigma, \rho, r, q, \theta) - C_i \right|, \tag{17}$$

where N is the total number of options, C_i and C_i^{H} represent the realized market price and the implied theoretical model price, respectively, while κ, σ, ρ, r, q, θ ($n = 6$) are the parameters of the Heston model to be estimated (κ-mean denotes the reverting speed, σ is "volatility of volatility", ρ is the correlation coefficient between the asset return and its volatility, r, q are the interest rate and dividend yield, respectively, and θ is the long run mean level).

To compute option prices by the Heston model, one needs input parameters that can hardly be found from the market data. We need to estimate the above parameters by an appropriate calibration procedure. The estimates of the Heston model parameters are obtained by minimizing MAE:

$$\mathrm{MAE}(\kappa, \sigma, \rho, r, q, \theta) \to \min. \tag{18}$$

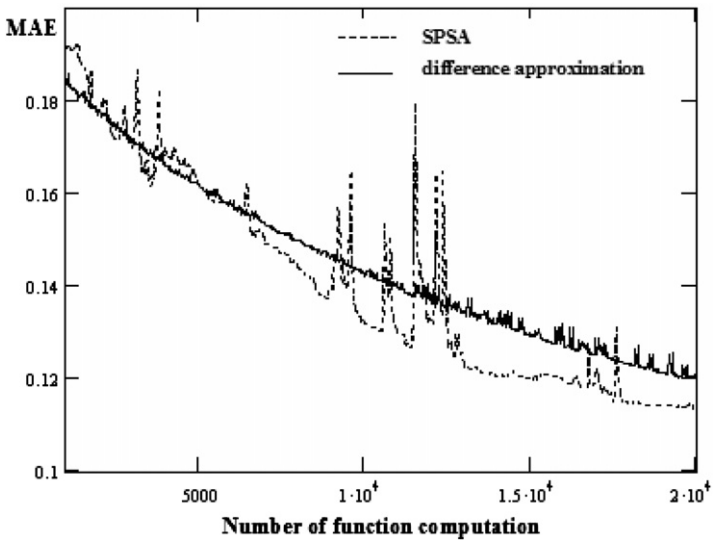

Fig. 2. Minimization of MAE by SPSA and difference approximation methods.

The implied theoretical model price C_i^{H} can be described as follows:

$$C_i^{\mathrm{H}} = S \cdot \mathrm{e}^{-q\cdot\tau} P_1 - K \cdot \mathrm{e}^{-r\cdot\tau} \cdot P_2,$$

where, P_j, $j = 1, 2$, are two probability functions. The details of this formula's derivation can be found in Heston (1993) and S is the assets price, while K is a given strike price.

The Heston model was implemented for the Call option on SPX (29 May 2002). Heston's stochastic volatility model explains many effects in financial stock markets; however, the implementation of this model is associated with the model calibration, which needs a more sophisticated optimization algorithm. The SPSA algorithm with perturbation Example 2 was applied to the calibration of the Heston model. Usually, SPSA requires that MAE be computed several hundred times that is reasonable for interactive Heston model calibration. In Fig. 2, we can see the dependence of MAE on the number of computations of function values, where the same dependence obtained by the difference approximation method is given for comparison. Fig. 2 illustrates the applicability of the SPSA algorithm in practice. The multipliers were chosen in the same way as in the previous Section with the following constants: $a = 0.5$, $b = 0.002$, $c = n \cdot 0.0005$, $d = d_0 \cdot 0.0002$, $d_0 = \sqrt{\frac{(n+2)\cdot(n+3)}{n\cdot(n+1)}}$ for SPSA method and $a = 0.5$, $b = 0.1$, $c = 0.0005$, $d = d_0 \cdot 0.0005$ for the finite difference method.

7. Conclusions and future research

The method for SPSA has been developed, using the operators of perturbation with the Lipschitz density function. This SA model enables us to use the approximation of the objective function by twice differentiable functions and to present their gradients by volume integrals. Using this presentation and the likelihood ratios method, we have proposed the calculus of the stochastic gradient which is applied to create SPSA algorithms for various perturbation densities. The convergence of the created SPSA algorithms was established for Lipschitz functions under general conditions. However, the problem relating to the Lipschitzian smoothing operator with optimal properties has not been studied and, thus, it is a subject for further investigation.

The rate of convergence of the developed approach was explored for the functions with a sharp minimum. The rate of convergence was studied by analysing two processes: the convergence of the optimizing sequence to the minimum of the smoothed function, and the convergence of the smoothed function to the objective function. We have proved that the rate of convergence $E(\|x^{k+1} - x^*_{\sigma_{k+1}}\|^2) = \mathrm{O}\left(\frac{1}{k^\gamma}\right)$, $1 < \gamma < 2$. Note that this convergence rate follows from the stochastic gradient expression (6), when there are no noises in computation of the objective function. The results obtained prove that the rate of convergence for the functions with a sharp minimum can be higher than that for the smoothed functions. Theoretical results were validated by computer simulation. During the Monte-Carlo simulation the empirical values of the estimated rate of convergence corroborated the theoretical estimation of the convergence order $1 < \gamma < 2$. The SPSA algorithm presented here has appeared to be more efficient for small n than the standard finite difference approach. When the dimensionality of the task increases, the issue of efficiency of SPSA in comparison with the finite difference approach should be investigated more thoroughly. Finally, the developed algorithm was applied to the minimization of the mean absolute pricing error for parameter estimation in the Heston stochastic volatility model to demonstrate its applicability for practical purposes.

Appendix A

Proof of Lemma 1. The formal differentiation of the second integral in (2) gives us an expression of the smoothed function gradient:

$$\begin{aligned}\frac{\partial \bar{f}(x,\sigma)}{\partial x} &= \frac{1}{\sigma}\cdot\int_{\Omega} f(x+\sigma y)\cdot\partial p(y)\,\mathrm{d}y = \frac{1}{\sigma}\cdot\int_{\mathrm{dom}(P)} f(x+\sigma y)\cdot\frac{\partial p(y)}{\psi(y)}\cdot\psi(y)\,\mathrm{d}y \\ &= \frac{1}{\sigma}\cdot E\left(f(x+\sigma\xi)\cdot\frac{\partial p(\xi)}{\psi(\xi)}\right), \qquad (1\mathrm{A})\end{aligned}$$

where some value of the generalized gradient is taken when it is set-valued as well as the appropriated importance density ψ introduced. Since the measure of the set is zero, where GG's are set-valued (see Remark 2), besides these GG's are bounded and the corresponding measure is absolutely continuous, expression (1A) is defined unambiguously as a.s. expectation.

On the other hand, by differentiating the identity $\int_\Omega p(y)\,dy \equiv \int_\Omega p(y+x)\,dy = 1$ with respect to x, the next identity

$$\int_\Omega \partial p(y+x)\,dy \equiv \int_\Omega \partial p(y)\,dy = 0$$

follows, and, thus, we have

$$\int_\Omega \partial p(y)\,dy = \int_{\mathrm{dom}(P)} \frac{\partial p(y)}{\psi(y)} \cdot \psi(y)\,dy = E\left(\frac{\partial p(\xi)}{\psi(\xi)}\right) = 0. \tag{2A}$$

Then, (5) follows from (1A) and (2A).

Note that the gradient of smoothed function is bounded $\|\bar{g}(x,\sigma)\| \leqslant K$. Next,

$$E(g(x,\sigma,\xi))^2 = E\left(\frac{(f(x+\sigma\xi)-f(x))}{\sigma}\cdot\frac{\partial p(\xi)}{\psi(\xi)}\right)^2 = \int_{\mathrm{dom}(\Psi)} \frac{(f(x+\sigma y)-f(x))^2}{\sigma^2}\cdot\frac{\|\partial p(y)\|^2}{\psi(y)}\,dy$$
$$\leqslant K^2 \cdot \int_{\mathrm{dom}(\Psi)} \|y\|^2 \cdot \frac{\|\partial p(y)\|^2}{\psi(y)}\,dy \leqslant K^2 \cdot A.$$

The Hessian (8) of the smoothed function is obtained by differentiating with respect to x the last expression in (1A) and the estimate is presented as follows:

$$\|V(x,\sigma)\| = \frac{1}{\sigma}\cdot\left\|E\left(\frac{\partial f(x+\sigma\xi)\cdot(\partial p(\xi))^{\mathrm{T}}}{\psi(\xi)}\right)\right\|$$
$$\leqslant \frac{1}{\sigma}\cdot E\frac{\|\partial f(x+\sigma\xi)\cdot(\partial p(\xi))^{\mathrm{T}}\|}{\psi(\xi)}\,\frac{1}{\sigma}\cdot E\frac{\|\partial f(x+\sigma\xi)\|\cdot\|\partial(p(\xi))\|}{\psi(\xi)} \leqslant \frac{K\cdot L}{\sigma}. \qquad \square \tag{3A}$$

Proof of Lemma 2. Denote the indicator of the set A by $I(A)$. Let $z_1 \in \partial f(x)$ be such that, for some y:

$$\left\|E\left((\partial f(y+\sigma_i\xi)-z_1)\cdot I\left(\|\xi\| \leqslant \frac{\delta}{\sigma_i}\right)\right)\right\| = \min_{z\in\partial f(x)}\left\|E\left((\partial f(y+\sigma_i\xi)-z)\cdot I\left(\|\xi\| \leqslant \frac{\delta}{\sigma_i}\right)\right)\right\|$$

and k be such that $\sigma_i \leqslant \frac{\delta}{2}\cdot\sqrt{\frac{\varepsilon}{K\cdot E\|\xi\|^2}}$ for all $i \geqslant k$. Hence, we have by virtue of Remark 2 and the Chebyshev inequality:

$$E\left(\|\partial f(y+\sigma_i\xi)-z_1\|\cdot I\left(\|\xi\| > \frac{\delta}{\sigma_i}\right)\right) \leqslant 2K\cdot P\left(\|\xi\| > \frac{\delta}{\sigma_i}\right) \leqslant \frac{2\cdot K\cdot\sigma_i^2\cdot E\|\xi\|^2}{\delta^2} \leqslant \frac{\varepsilon}{2}. \tag{4A}$$

Let $H:\Re^n \to \partial f(x)$ be a selector, i.e., a measurable function such that $\left\|\frac{\partial f(y)}{\partial y} - H(y)\right\| = \min_{z\in\partial f(x)}\left\|\frac{\partial f(y)}{\partial y} - z\right\|$ (see the theorem on selectors in Kuratowski, 2003), when $f(y)$ is differentiable in the usual sense at the point y, and otherwise, $H(y)$ obtains any finite value, if the gradient of the $f(y)$ is multivalued. Then, taking into account Remark 2, if $\|y-x\| \leqslant \delta$, we have

$$\min_{z\in\partial f(x)}\|\bar{g}(y,\sigma_i)-z\| \leqslant \|E(\partial f(y+\sigma_i\xi)-z_1)\|$$
$$= \left\|E\left((\partial f(y+\sigma_i\xi)-z_1)\cdot I\left(\|\xi\| \leqslant \frac{\delta}{\sigma_i}\right)\right) + E\left((\partial f(y+\sigma_i\xi)-z_1)\cdot I\left(\|\xi\| > \frac{\delta}{\sigma_i}\right)\right)\right\|$$
$$\leqslant \min_{z\in\partial f(x)}\left\|E\left((\partial f(y+\sigma_i\xi)-z)\cdot I\left(\|\xi\| \leqslant \frac{\delta}{\sigma_i}\right)\right)\right\|$$
$$+ \left\|E\left((\partial f(y+\sigma_i\xi)-z_1)\cdot I\left(\|\xi\| > \frac{\delta}{\sigma_i}\right)\right)\right\| \leqslant \varepsilon$$

by virtue of (4A) and

$$\begin{aligned}\frac{\varepsilon}{2} &\geqslant E\left(\left(\min_{z\in\partial f(x)}\|(\partial f(y+\sigma_i\xi)-z)\|\right)\cdot I\left(\|\xi\|\leqslant\frac{\delta}{\sigma_i}\right)\right)=E\left\|(\partial f(y+\sigma_i\xi)-H(y+\sigma_i\xi))\cdot I\left(\|\xi\|\leqslant\frac{\delta}{\sigma_i}\right)\right\|\\ &\geqslant\left\|E\left(\left(\partial f(y+\sigma_i\xi)-E\left(H(y+\sigma_i\xi)\middle|\|\xi\|\leqslant\frac{\delta}{\sigma_i}\right)\right)\cdot I\left(\|\xi\|\leqslant\frac{\delta}{\sigma_i}\right)\right)\right\|\\ &\geqslant\min_{z\in\partial f(x)}\left(\left\|E\left((\partial f(y+\sigma_i\xi)-z)\cdot I\left(\|\xi\|\leqslant\frac{\delta}{\sigma_i}\right)\right)\right\|\right)\end{aligned}$$

because $E\left(H(y+\sigma_i\xi)\middle|\|\xi\|\leqslant\frac{\delta}{\sigma_i}\right)\in\partial f(x)$ due to convexity of the GG mapping. □

We need several lemmas for proving the theorems.

Denote by $\{\Theta_k\}_{k=0}^{\infty}$ a sequence of σ-algebras generated by the sequence $\{x^k\}_{k=0}^{\infty}$.

Lemma 3. *If $\{\varphi_k\}$ is a sequence of random variables measurable with respect to $\{\Theta_k\}_{k=0}^{\infty}$, $\sum_{k=0}^{\infty}E\varphi_k^2<\infty$, then $\sum_{k=0}^{\infty}\varphi_k-E(\varphi_k|\Theta_{k-1})$ converges a.s.*

The proof of this lemma can be found in Wasan (1969, Lemma 1, Appendix 2, Chapter 4).

Note, the function $v(x)$ is conical at the point $x^*\in\Re^n$ if $v(\lambda\cdot(x-x^*)+x^*)=\lambda\cdot v(x)$ for any $\lambda\geqslant 0$.

Lemma 4. *Let $f(x)$ be Lipschitzian. Then the function $h:\Re^n\to\Re$, defined in (12), is convex, conical and Lipschitzian. If the function $f(x)$ has a sharp minimum with constant μ, then*

$$h(y)\geqslant\mu\cdot\|y\|,\quad y\in\Re^n. \tag{5A}$$

Proof. The convexity and cone property of the generalized directional derivative mapping h are proved by Clarke (1983). By definition (12) we have $h(y_1)-h(y_2)=\max_{z\in\partial f(x^*)}(z\cdot y_1)-\max_{z\in\partial f(x^*)}(z\cdot y_2)=(z_1\cdot y_1)-(z_2\cdot y_2)\leqslant(z_1\cdot(y_1-y_2))\leqslant K\cdot\|y_2-y_1\|$, where $(z_1\cdot y_1)=\max_{z\in\partial f(x^*)}(z\cdot y_1)$ and $(z_2\cdot y_2)=\max_{z\in\partial f(x^*)}(z\cdot y_2)$. In the same way we prove $h(y_2)-h(y_1)\leqslant K\cdot\|y_2-y_1\|$. Thus, the Lipschitz property follows with constant K:

$$|h(y_2)-h(y_1)|\leqslant K\cdot\|y_2-y_1\|.$$

The proof of (5A) easily follows from definition of generalized directional derivative (see Clarke, 1983; Michalevitch et al., 1987) and sharp property (11):

$$h(y)=\lim_{\delta\to 0_+}\sup_{\substack{\|v\|\leqslant\delta\\0<\sigma\leqslant\delta}}\frac{f(x^*+v+\sigma\cdot y)-f(x^*+v)}{\sigma}\geqslant\lim_{\sigma\to 0_+}\frac{f(x^*+\sigma\cdot y)-f(x^*)}{\sigma}\geqslant\mu\cdot\|y\|.\qquad\square$$

Lemma 5. *Let $f(x)$ be the semismooth function. Then $\lim_{\sigma\to 0}\bar{h}(x,\sigma)=\bar{h}(x)$.*

The proof follows from the definition of generalized directional derivative, definition of $\bar{h}(x,\sigma)$ and the Lebesgue Convergency Theorem:

$$\begin{aligned}\lim_{\sigma\to 0}\bar{h}(x,\sigma)&=\lim_{\sigma\to 0}E\left(\frac{f(x^*+\sigma\cdot(x-x^*)+\sigma\cdot\xi)-f(x^*)}{\sigma}\right)=E\left(\lim_{\sigma\to 0}\left(\frac{f(x^*+\sigma\cdot(x-x^*)+\sigma\cdot\xi)-f(x^*)}{\sigma}\right)\right)\\&=E(h(x+\xi))=\bar{h}(x).\qquad\square\end{aligned}$$

Lemma 6. *Let x_σ^* be a minimum point of the smoothed function (2), where f is semismooth. Then*

$$\lim_{\sigma\to 0}\frac{x_\sigma^*-x^*}{\sigma}=y^*-x^*,$$

where y^* is a minimum point of the function $\bar{h}(x)$.

Proof. It is easy to see that the minimum points y_σ^* and x_σ^* of the functions $\bar{h}(x,\sigma)$ and $\bar{f}(x,\sigma)$, respectively, are related as follows: $x_\sigma^* = x^* + \sigma(y_\sigma^* - x^*)$. Note that, the function $\bar{h}(x,\sigma)$ is continuous, bounded from below, and infinitely increasing, as $x \to \infty$. It means that this function has its minimum at the finite point y_σ^*.

The Lemma is proved, because $y^* = \lim_{\sigma\to 0} y_\sigma^*$ according to Lemma 5. □

Lemma 7. *Assume that* $\{u_k\}_1^\infty$ *is the sequence of non-negative numbers such that for certain* $k \geqslant k_0$ *we have*

$$u_{k+1}^2 \leqslant \left(u_k \cdot \left(1 - \frac{c_k}{k^p}\right) + \frac{l_k}{k^s}\right)^2 + \frac{d}{k^t},$$

where $c_k > 0$, $l_k > 0$, $c_k \to c$, $l_k \to l$, d, $l > 0$, $c > \frac{t-p}{2} > 0$, $s > \frac{t+p}{2}$. *Then*

$$u_k^2 \leqslant \frac{d}{2c} \cdot \frac{1}{k^{t-p}} + \mathrm{O}\left(\frac{1}{k^{2c}}\right).$$

Proof. Assume, $c_k \geqslant c' = c - \frac{\varepsilon}{2}$, $l_k \leqslant l' = l + \frac{\varepsilon}{2}$ for a certain $k \geqslant k_0$ and a certain small $\varepsilon > 0$. We have due to (Wasan, 1969, Appendix 3, Lemma 6) that

$$\frac{d}{k^t} \leqslant \frac{d}{2\cdot(c-\varepsilon)}\left(\frac{1}{(k+1)^{t-p}} - \left(1 - \frac{2\cdot(c-\varepsilon)}{k^p}\right)\frac{1}{k^{t-p}}\right). \tag{6A}$$

Let k_0 is such that $u_k^2 \leqslant \frac{d}{2\cdot(c-\varepsilon)} \cdot \frac{1}{k^{t-p}}$ for $k \geqslant k_0$.

Then, by virtue of (6A), lemma and the latter assumption:

$$\begin{aligned} u_{k+1}^2 - \frac{d}{2(c-\varepsilon)} \cdot \frac{1}{(k+1)^{t-p}} &\leqslant u_k^2\left(1 - \frac{c-\frac{\varepsilon}{2}}{k^p}\right)^2 + 2\cdot u_k\left(1 - \frac{c-\frac{\varepsilon}{2}}{k^p}\right)\cdot\frac{l+\frac{\varepsilon}{2}}{k^s} + \frac{\left(l+\frac{\varepsilon}{2}\right)^2}{k^{2s}} \\ &\quad - \frac{d}{2(c-\varepsilon)}\cdot\left(1 - \frac{2(c-\varepsilon)}{k^p}\right)\frac{1}{k^{t-p}} \\ &\leqslant \left(1 - \frac{c-\frac{\varepsilon}{2}}{k^p}\right)^2\left(u_k^2 - \frac{d}{2(c-\varepsilon)}\cdot\frac{1}{k^{t-p}}\right) - \frac{d}{2\cdot(c-\varepsilon)}\cdot\frac{\varepsilon}{k^t} + \mathrm{O}\left(\frac{1}{k^t}\right) \leqslant 0. \end{aligned}$$

Hence, $u_k^2 \leqslant \frac{d}{2(c-\varepsilon)} \cdot \frac{1}{k^{t-p}}$ would be true for $k \geqslant k_0$, if this is true for some sufficiently large k_0. Next, let, on the contrary, for all sufficiently large k:

$$u_k^2 \geqslant \frac{d}{2(c-\varepsilon)} \cdot \frac{1}{k^{t-p}} + \mathrm{O}\left(\frac{1}{k^{2(c-\varepsilon)}}\right).$$

Then, using (6A) after simple manipulations, for sufficiently large k we get

$$\begin{aligned} u_{k+1}^2 - \frac{d}{2(c-\varepsilon)} \cdot \frac{1}{(k+1)^{t-p}} &\leqslant \left(1 - \frac{c-\varepsilon}{k^p}\right)^2\left(u_k^2 - \frac{d}{2(c-\varepsilon)}\cdot\frac{1}{k^{t-p}}\right) \\ &\quad - \frac{u_k}{k^p}\left(u_k\cdot\varepsilon\cdot\left(1 - \frac{c-\frac{3\varepsilon}{4}}{k^p}\right) - 2\left(1 - \frac{c}{k^p}\right)\cdot\frac{l+\frac{\varepsilon}{2}}{k^{s-p}}\right) + \frac{\left(l+\frac{\varepsilon}{2}\right)^2}{k^{2s}} + \frac{d(c-\varepsilon)}{2\cdot k^{t+p}} \\ &\leqslant \left(1 - \frac{c-\varepsilon}{k^p}\right)^2\left(u_k^2 - \frac{d}{2(c-\varepsilon)}\cdot\frac{1}{k^{t-p}}\right) - \varepsilon\cdot\frac{d}{2(c-\varepsilon)}\cdot\frac{1}{k^t} + \mathrm{O}\left(\frac{1}{k^t}\right) \\ &\leqslant \left(1 - \frac{c-\varepsilon}{k^p}\right)^2\left(u_k^2 - \frac{d}{2(c-\varepsilon)}\cdot\frac{1}{k^{t-p}}\right). \end{aligned}$$

We denote as u'_{k+1} the left side of the last inequality and the right side as $u'_k\left(1 - \frac{c-\varepsilon}{k^p}\right)^2$. Then $0 \leqslant u'_{k+1} \leqslant u'_k\left(1 - \frac{c-\varepsilon}{k^p}\right)^2$. This sequence is lower bounded and monotonous, and therefore convergent. Thus, $u'_{k+1} \leqslant u'_m \cdot \prod_{i=m}^k \left(1 - \frac{c-\varepsilon}{i}\right)^2 \leqslant u'_m \cdot \mathrm{e}^{-2(c-\varepsilon)\cdot\sum_{i=m}^k \frac{1}{i}} = \mathrm{O}\left(\frac{1}{k^{2(c-\varepsilon)}}\right)$, $m > c$. Consequently, in both cases

$$u_k^2 \leqslant \inf_{\varepsilon>0}\left(\frac{d}{2(c-\varepsilon)}\cdot\frac{1}{k^{t-p}} + \mathrm{O}\left(\frac{1}{k^{2(c-\varepsilon)}}\right)\right) \leqslant \frac{d}{2c}\cdot\frac{1}{k^{t-p}} + \mathrm{O}\left(\frac{1}{k^{2c}}\right). \qquad \square$$

Proof of Theorem 1. According to (3), we have

$$\bar{f}(x^{k+1},\sigma_{k+1}) \leqslant \bar{f}(x^{k+1},\sigma_k) + C\cdot K\cdot|\sigma_{k+1}-\sigma_k|. \tag{7A}$$

From the Lagrange formula (Dieudonné, 1960), (7), (8), (10) and Lemma 1 we obtain that

$$\begin{aligned}
\bar{f}\left(x^{k+1},\sigma_k\right) &= \bar{f}(x^k,\sigma_k) + \left(\bar{g}(x^k+\tau(x^{k+1}-x^k),\sigma_k) - \bar{g}(x^k,\sigma_k)\right)^{\mathrm{T}}\cdot(x^{k+1}-x^k) + \bar{g}(x^k,\sigma_k)^{\mathrm{T}}\cdot(x^{k+1}-x^k)\\
&\leqslant \bar{f}(x^k,\sigma_k) + \bar{g}(x^k,\sigma_k)^{\mathrm{T}}\cdot(x^{k+1}-x^k) + \frac{K\cdot L}{\sigma_k}\cdot\|x^{k+1}-x^k\|^2\\
&= \bar{f}(x^k,\sigma_k) - \rho_k g(x^k,\sigma_k,\xi_{k+1})^{\mathrm{T}}\bar{g}(x^k,\sigma_k) + \frac{K\cdot L\cdot\rho_k^2}{\sigma_k}\|g(x^k,\sigma_k,\xi_{k+1})\|^2\\
&= \bar{f}(x^k,\sigma_k) + \frac{K\cdot L}{\sigma_k}\rho_k^2\|g^k\|^2 - \rho_k\|\bar{g}^k\|^2 + \rho_k(\bar{g}^k)^{\mathrm{T}}(\bar{g}^k-g^k)\\
&\leqslant \bar{f}(x^k,\sigma_k) + \frac{K^3\cdot L\cdot A}{\sigma_k}\rho_k^2 - \rho_k\|\bar{g}^k\|^2 + \frac{K\cdot L}{\sigma_k}\rho_k^2\cdot\left(\|g^k\|^2 - E\|g^k\|^2\right) + \rho_k(\bar{g}^k)^{\mathrm{T}}(\bar{g}^k-g^k),
\end{aligned} \tag{8A}$$

where $0\leqslant\tau\leqslant 1$, and, for simplicity, we denote $\bar{g}^k=\bar{g}(x^k,\sigma_k)$, $g^k=g(x^k,\sigma_k,\xi_{k+1})$.

First of all, we show that there exists an infinite subsequence that converges to X^*, which is bounded with respect to Remark 1. In the opposite case, we see that we can find $\bar{s}$ and $\bar{\delta}$, such that a sequence of sets $\{x\,|\,\|x^s-x\|\leqslant 2\bar{\delta}\}$ has no intersection with X^*, as $s\geqslant\bar{s}$. By virtue of condition C and Lemma 2 we can find $\varepsilon>0$ such that $\|\bar{g}^s\|\geqslant\varepsilon>0$, if $s\geqslant\bar{s}$, and if $\bar{s}$ is large enough. Let $\bar{s}$ be such that, for $i\geqslant\bar{s}$, the inequality

$$\frac{C\cdot K\cdot|\sigma_{i+1}-\sigma_i|}{\rho_i} + \frac{\rho_i\cdot K^3\cdot L\cdot A}{\sigma_i} \leqslant \frac{\varepsilon^2}{2}$$

is true. Then, according to (7A) and (8A), we get

$$\begin{aligned}
\bar{f}(x^{k+1},\sigma_{k+1}) &\leqslant \bar{f}(x^s,\sigma_s) - \frac{\varepsilon^2}{2}\sum_{i=s}^{k}\rho_i - \sum_{i=s}^{k}\rho_i\left(\frac{\varepsilon^2}{2} - \frac{C\cdot K\cdot|\sigma_{i+1}-\sigma_i|}{\rho_i} - \frac{\rho_i\cdot K^3\cdot L\cdot A}{\sigma_i}\right)\\
&\quad + \sum_{i=s}^{k}\rho_i^2\cdot\frac{K\cdot L}{\sigma_i}\cdot\left(\|g^i\|^2 - E\left(\|g^i\|^2|\Theta_i\right)\right) + \sum_{i=s}^{k}\rho_i(\bar{g}^i)^{\mathrm{T}}(\bar{g}^i-g^i)\\
&\leqslant \bar{f}(x^s,\sigma_s) - \frac{\varepsilon^2}{2}\sum_{i=s}^{k}\rho_i + \sum_{i=s}^{k}\rho_i(\bar{g}^i)^{\mathrm{T}}(\bar{g}^i-g^i) + \sum_{i=s}^{k}\rho_i^2\cdot\frac{K\cdot L}{\sigma_i}\cdot\left(\|g^i\|^2 - E\left(\|g^i\|^2|\Theta_i\right)\right).
\end{aligned} \tag{9A}$$

Since $\sum_{i=0}^{k}\rho_i^2\left(E((g^i)^{\mathrm{T}}|\Theta_i)\cdot\bar{g}^i\right)^2 \leqslant \sum_{i=0}^{k}\rho_i^2\left(E\left(\|g^i\|^2|\Theta_i\right)\right)^2 \leqslant K^4\cdot A^2\cdot\sum_{i=0}^{k}\rho_i^2$ and $\sum_{k=1}^{\infty}\rho_k^2<\infty$, Lemma 3 implies that $\sum_{i=0}^{\infty}\rho_i(\bar{g}^i)^{\mathrm{T}}\cdot(\bar{g}^i-g^i)$ is converging a.s. Similarly we can show that $\sum_{i=0}^{\infty}\rho_i^2\cdot\frac{K\cdot L}{\sigma_i}\cdot\left(\|g^i\|^2 - E\left(\|g^i\|^2|\Theta_i\right)\right)$ is converging a.s., because $E\|g^k\|^4 = E\left(\left\|\frac{f(x^k+\sigma_k\xi)-f(x^k)}{\sigma_k}\cdot\frac{\partial\ln p(\xi)}{\psi(\xi)}\right\|^4\right) \leqslant K^4\cdot\int_{\mathrm{dom}(P)}\|y\|^4\cdot\frac{\|\partial\ln p(y)\|^4}{(\psi(y))^3}\,\mathrm{d}y<\infty$.

So the two last components in (9A) are bounded a.s. Hence, we get a contradiction in (9A), because $\sum_{i=s}^{\infty}\rho_i=\infty$ and $\bar{f}(x,\sigma)>-\infty$. Consequently, there must exist an infinite subsequence converging to X^*. Say that such a subsequence does exist in sequence (10), which either converges to $x'\notin X^*$ or to infinity.

Now we will show that if in the sequence $\{x^k\}_{k=0}^{\infty}$, there exists such a bounded subsequence convergent to the finite point x', where $\inf_{g\in\partial f(x')}\|g\|>0$, then there exists δ_0 such that, for all $\delta\in(0,\delta_0]$, we can find such sequences of indices $\{l_s\}_{s=0}^{\infty}$, $\{k_s\}_{s=0}^{\infty}$, $l_s<k_s$, so that $\|x^i-x'\|\leqslant 2\delta$ for all $i\in[l_s,k_s-1]$ and

$$\overline{\lim_{s\to\infty}} f(x^{k_s}) < \lim_{s\to\infty} f(x^{l_s}). \tag{10A}$$

According to our assumptions on subsequences that converge to $x'\notin X^*$ and $x\in X^*$, there exist sequences of indices l_s, k_s, $l_s<k_s$ such that, for sufficiently small δ we have $\|x^{l_s}-x'\|<\delta$, $\|x^{k_s}-x'\|>2\delta$ and $\|x^i-x'\|\leqslant 2\delta$, as $i\in[l_s,k_s-1]$. Now we show that (10A) follows from the existence of the sequences. We see that

$$x^{k_s} = x^{l_s} - \sum_{i=l_s}^{k_s-1} \rho_i g^i = x^{l_s} - \sum_{i=l_s}^{k_s-1} \rho_i \bar{g}^i - \sum_{i=l_s}^{k_s-1} \rho_i (g^i - \bar{g}^i).$$

According to Lemma 3, $\sum_{i=0}^{\infty} \rho_i \cdot (g^i - \bar{g}^i)$ converges a.s. It means that $\left\| \sum_{i=l_s}^{k_s-1} \rho_i (g^i - \bar{g}^i) \right\| < \frac{\delta}{2}$, when $s \geqslant \bar{s}$, if $\bar{s}$ is large enough. Therefore

$$\delta \leqslant \|x^{k_s} - x^{l_s}\| \leqslant \left\| \sum_{i=l_s}^{k_s-1} \rho_i \bar{g}^i \right\| + \left\| \sum_{i=l_s}^{k_s-1} \rho_i (g^i - \bar{g}^i) \right\| \leqslant \left\| \sum_{i=l_s}^{k_s-1} \rho_i \bar{g}^i \right\| + \frac{\delta}{2}.$$

Thus, $\left\| \sum_{i=l_s}^{k_s-1} \rho_i \bar{g}^i \right\| \geqslant \frac{\delta}{2}$ and $\sum_{i=s}^{k_s-1} \rho_i \geqslant \frac{\delta}{2 \cdot \sqrt{K}}$, because $\|\bar{g}^i\|^2 < K$. Then an estimate follows from (9A):

$$\bar{f}(x^{k_s}, \sigma_{k_s}) \leqslant \bar{f}(x^{l_s}, \sigma_{l_s}) - \frac{\varepsilon^2 \cdot \delta}{8 \cdot \sqrt{K}}.$$

According to the uniform convergence with respect to x $\bar{f}(x, \sigma) \to f(x)$, as $\sigma \to 0$, we have that (10A) is true. It means that for all numbers f' and f'', such that $\overline{\lim_{s\to\infty}} f(x^{k_s}) < f' < f'' < \lim_{s\to\infty} f(x^{l_s})$, the sequence $f(x^k)$ crosses the interval (f', f'') many times. Then we can find two subsequences $\{x^{r_i}\}$ and $\{x^{p_i}\}$, for which

$$f(x^{r_i}) \leqslant f', \quad f(x^{r_i+1}) > f', \tag{11A}$$

$$f(x^{p_i}) > f'', \quad f(x^k) > f', \quad r_s < k < p_s. \tag{12A}$$

Without loss of generality, assume the sequence $\{x^{r_i}\}$ to be convergent; in the opposite case, instead of this sequence, let us take its converging subsequence. It follows from the convergence of the sequence $\{x^{r_i}\}$, the continuity of f, (11A) and $\rho_i \to 0$ that $\lim_{i\to\infty} f(x^{r_i}) = f'$. Further, by virtue of (11A) and (12A), we obtain

$$\overline{\lim_{i\to\infty}} f(x^{p_i}) \geqslant \lim_{i\to\infty} f(x^{r_i}). \tag{13A}$$

Since F^* does not contain inner points, f' can be chosen so that $f' \notin F^*$. So (10A) can be derived when $\{r_s\}$ corresponds to $\{l_s\}$. But in this way, (13A) contradicts to (10A).

If such a subsequence converging to infinity exists in the sequence $\{x^k\}_{k=0}^{\infty}$, analogously it is shown, that there exists δ_0 such that, as $\delta \in (0, \delta_0]$, we can find such a sequence of indices $\{l_s\}_{s=0}^{\infty}$, $\{k_s\}_{s=0}^{\infty}$, that $\inf_{x\in X^*} \|x^{l_s} - x\| > 2 \cdot \delta$, $\inf_{x\in X^*} \|x^{k_s} - x\| < \delta$, and $\inf_{x\in X^*} \|x^i - x\| \geqslant \delta$ for all $i \in [l_s, k_s - 1]$, and $\overline{\lim_{s\to\infty}} f(x^{k_s}) < \lim_{s\to\infty} f(x^{l_s})$ are derived. Further, the existence of the corresponding indices r_s, p_s is established leading to a contradiction. □

Proof of Theorem 2. Since the objective function is assumed having only one sharp minimum, the sequence (10) converges a.s. to the point of this minimum. According to (10) and the optimality condition $\bar{g}(x^*_{\sigma_k}, \sigma_k) = 0$, we have that

$$\begin{aligned}
\left\| x^{k+1} - x^*_{\sigma_{k+1}} \right\|^2 &= \left\| x^k - x^*_{\sigma_{k+1}} - \rho_k \cdot (\bar{g}(x^k, \sigma_k) - \bar{g}(x^*_{\sigma_k}, \sigma_k)) - \rho_k \cdot (g(x^k, \sigma_k, \xi_{k+1}) - \bar{g}(x^k, \sigma_k)) \right\|^2 \\
&= \left\| x^k - x^*_{\sigma_{k+1}} - \rho_k \cdot (\bar{g}(x^k, \sigma_k) - \bar{g}(x^*_{\sigma_k}, \sigma_k)) \right\|^2 - 2\rho_k \cdot (g(x^k, \sigma_k, \xi_{k+1}) \\
&\quad - \bar{g}(x^k, \sigma_k))^{\mathrm{T}} \cdot (x^k - x^*_{\sigma_{k+1}} - \rho_k \cdot (\bar{g}(x^k, \sigma_k) - \bar{g}(x^*_{\sigma_k}, \sigma_k))) \\
&\quad + \rho_k^2 \cdot \left\| g(x^k, \sigma_k, \xi_{k+1}) - \bar{g}(x^k, \sigma_k) \right\|^2.
\end{aligned}$$

By averaging both sides of this equality, we obtain by virtue of (7)

$$\begin{aligned}
E\left(\left\| x^{k+1} - x^*_{\sigma_{k+1}} \right\|^2 | \Theta_k \right) &= \left\| x^k - x^*_{\sigma_{k+1}} - \rho_k (\bar{g}(x^k, \sigma_k) - \bar{g}(x^*_{\sigma_k}, \sigma_k)) \right\|^2 \\
&\quad + \rho_k^2 E\left(\left\| g(x^k, \sigma_k, \xi_{k+1}) - \bar{g}(x^k, \sigma_k) \right\|^2 | \Theta_k \right) \\
&\leqslant \left\| x^k - x^*_{\sigma_{k+1}} - \rho_k (\bar{g}(x^k, \sigma_k) - \bar{g}(x^*_{\sigma_k}, \sigma_k)) \right\|^2 + \rho_k^2 \cdot K^2 \cdot A.
\end{aligned}$$

Note that, $\frac{\partial^2 \bar{f}(x,\sigma)}{\partial x^2} = \frac{1}{\sigma} \cdot \frac{\partial^2 \bar{h}\left(\frac{(x-x^*)}{\sigma} + x^*, \sigma\right)}{\partial x^2}$.

Further, from the Lagrange formula and the latter expression the estimate follows: $\left\|\bar{g}(x^k,\sigma_k)-\bar{g}(x^*_{\sigma_k},\sigma_k)\right\|^2 \geqslant \frac{H_k^2}{\sigma_k^2}\cdot\left\|x^k-x^*_{\sigma_k}\right\|^2$, where $H_k=\min\limits_{0\leqslant\tau\leqslant 1}\left\|\left(\left.\frac{\partial^2\bar{h}(y,\sigma)}{\partial y^2}\right|_{y=\frac{x^k-x^*+\tau\cdot(x^k-x^*_{\sigma_k})}{\sigma_k}+x^*}\right)^{-1}\right\|^{-1}$. By the latter estimates and Lemma 6, we get

$$E\left(\|x^{k+1}-x^*_{\sigma_{k+1}}\|^2|\Theta_k\right)\leqslant\left(\|x^k-x^*_{\sigma_k}\|\cdot\left(1-\frac{\rho_k}{\sigma_k}\cdot H_k\right)+(\sigma_k-\sigma_{k+1})\cdot(\|y^*-x^*\|+\mathrm{O}(\sigma_k))\right)^2+\rho_k^2\cdot K^2\cdot A, \tag{14A}$$

where $H_k\to H=\left\|\left(\left.\frac{\partial^2\bar{h}(y)}{\partial y^2}\right|_{y=y^*}\right)^{-1}\right\|^{-1}$, $k\to\infty$, according to theorem condition.

After simple manipulations averaging both sides of (14A), the inequality follows:

$$E\left(\left\|x^{k+1}-x^*_{\sigma_{k+1}}\right\|^2\right)\leqslant\left(\sqrt{E\|x^k-x^*_{\sigma_k}\|^2}\cdot\left(1-\frac{\rho_k}{\sigma_k}\cdot H_k\right)+(\sigma_k-\sigma_{k+1})\cdot(\|y^*-x^*\|+\mathrm{O}(\sigma_k))\right)^2+\rho_k^2\cdot K^2\cdot A.$$

Finally by virtue of Lemma 7, we obtain the estimate:

$$E\left(\left\|x^{k+1}-x^*_{\sigma_{k+1}}\right\|^2\right)\leqslant\frac{A\cdot K^2\cdot a\cdot b}{H}\cdot\frac{1}{k^{1+\beta}}+\mathrm{O}\left(\frac{1}{k^{\frac{2\alpha H}{b}}}\right).\qquad\square$$

References

Blum, J., 1954. Multidimensional stochastic approximation procedures. Annals of Mathematical Statistics 25 (4).

Clarke, F.H., 1983. Optimization and Nonsmooth Analysis. John Wiley, New York.

Dieudonné, J., 1960. Foundations of Modern Analysis. Academic Press, NY, London.

Donoghue, W.F., 1969. Distributions and Fourier Transforms. Academic Press, NY, London.

Dupac, V., 1988. Stochastic approximation. In: Krishnaiah, P.R., Sen, P.K. (Eds.), Handbook of Statistics. Nonparametric Methods. Nord Holand, NY.

Dvoretzky, A., 1956. On stochastic approximation. In: Neumann, J. (Ed.), Proceedings of the 3rd Berkeley Symposium of Mathematical Statistics and Probability, vol. I. University of California Press, Berkeley, pp. 39–55.

Ermoliev, Yu.M., 1976. Methods of Stochastic Programming. Nauka, Moscow (in Russian).

Ermoliev, Yu.M., Norkin, V.I., Wets, R.J.-B., 1995. The minimization of semicontinuous functions: Mollifier subgradients. Control and optimization 3 (1), 149–167.

Granichin, O.N., Poliak, B.T., 2003. Randomized Algorithms for Estimation and Optimization with Almost Arbitrary Errors. Nauka, Moskow (in Russian).

Gupal, A.M., Norkin, V.I., 1977. An algorithm for minimization of discontinuous functions. Kibernetika, 73–75.

Heston, S.L., 1993. A closed-form solution for options with stochastic volatility with applications to bond and currency options. The Review of Financial Studies 6 (2), 327–343.

Kiefer, J., Wolfowitz, J., 1952. A stochastic estimation of the maximum of a regression function. Annals of Mathematical Statistics 23 (3), 462–466.

Kushner, H.J., Yin, G.G., 2003. Stochastic Approximation and Recursive Algorithms and Applications. Springer, NY, Heidelberg, Berlin.

Kuratowski, K., 2003. Topology, vol. II. Academic Press, NY.

Michalevitch, V.S., Gupal, A.M., Norkin, V.I., 1987. Methods of Nonconvex Optimization. Nauka, Moscow (in Russian).

Nurminski, E.A., 1979. Numerical Methods for Solving Deterministic and Stochastic Minimax Problems. Naukova Dumka, Kiev (in Russian).

Poliak, B.T., 1987. Introduction to Optimization. Translations Series in Mathematics and Engineering. Optimization Software, Inc., Publications Division, New York.

Robins, H., Monro, S., 1951. A stochastic approximation method. Annals of Mathematical Statistics 22 (3), 400–407.

Rockafellar, R.T., 1979. Directionally Lipschitzian functions and subdifferential calculus. Proceedings of the London Mathematical Society 39, 331–355.

Rubinstein, R., Shapiro, A., 1993. Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method. Wiley, New York, NY.

Sakalauskas, L., 2002. Nonlinear stochastic programming by Monte-Carlo estimators. Informatica 137, 558–573.

Spall, J.C., 1992. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. IEEE Transactions on Automatic Control 37, 332–341.

Wasan, M.T., 1969. Stochastic approximation. Transactions in Mathematics and Mathematical Physics. Cambridge University Press, Cambridge.

Yudin, D.B., 1965. Qualitative methods for analysis of complex systems. Izv. AN SSSR, Ser. "Technicheskaya. Kibernetika", No. 1, pp. 3–13 (in Russian).