Control
and Cybernetics

VOL. 16 (1987) No. 1

Stochastic approximation method with subgradient filtering and on-line stepsize rules for nonsmooth, nonconvex and unconstrained problems*)

by

WOJCIECH SYSKI

Systems Research Institute
Polish Academy of Sciences
ul. Newelska 6
01-447 Warszawa, Poland

A practical stochastic approximation algorithm for finding unconstrained minima of nonsmooth and nonconvex functions is described. It uses an auxiliary filter which averages stochastic subgradient estimates observed, thus producing directions for subsequent iterations. Stepsize coefficients and filter gains are controlled on-line on the basis of information gathered in the course of computations according to the rules derived from the concept of the regularized improvement function. Convergence of the method with probability 1 is proved, asymptotic properties are studied and a numerical example is described.

1. Introduction

The purpose of this paper is to analyse properies of a certain stochastic subgradient algorithm for solving the problem

$$\text{minimize } F(x) \text{ over } x \in R^n, \tag{1}$$

where $F: R^n \to R^1$ is a lower—C^2 function (see Remark 3 in the next section and [10]). We assume that neither the values of F nor its subgradients are available. Instead of those, at any point x^k one can only obtain a random vector $\xi^k = g^k + r^k$, where: $g^k \in \partial F(x^k)$ ($\partial F(x^k)$ denotes the subdifferential of F at x^k, cf. [10]) and r^k is a random noise of zero expectation. We shall call ξ^k a stochastic subgradient of F at x^k. Such a situation is typical in stochastic programming problems with objectives of the form $F(x) = Ef(x, \theta)$, where

*) This work was supported by the Research Program CPBP 02-15.

θ is a random parameter and E denotes the expected value. Then it is hard to evaluate F or its subgradients, but stochastic subgradients can be calculated with less effort (cf. [1], [7]).

In [1] a stochastic subgradient method for solving problem (1) in the convex case was suggested, which consists in the following iterations

$$x^{k+1} = x^k - \tau_k \xi^k, \qquad k = 0, 1, ... \tag{2}$$

where τ_k is a nonnegative stepsize coefficient. Since then the method has been extended to nonconvex problems (cf. [1], [3], [5], [7]) and various improvements consisting in the application of the averaging of directions have been suggested (cf. [1], [3], [4], [11]). But still one of the crucial questions connected with applications of method (2) and other recursive stochastic algorithms is the choice of the sequence of stepsizes $\{\tau_k\}$. The general theoretical rules: τ_k measurable with respect to $\{x^0, x^1, ..., x^k\}$, $\sum_{k=0}^{\infty} \tau_k = \infty$ w.p.1, $\sum_{k=0}^{\infty} E\tau_k^2 < \infty$, are insufficient in practice. Obviously, the sequence $\tau_k = \tau_0/(k+1)$, $k = 0, 1, ...$, satisfies these conditions, but with these stepsizes practical convergence of method (2) is very slow (see the example in Section 6). Therefore, in order to enhance convergence far from the solution of problem (1), some on-line rules are needed to determine stepsizes depending on the behavior of method (2).

In [2] and [11] a constant stepsize was applied and using some heuristic tests after a series of iterations it was checked whether the stepsize was too large or too small. Another approach (cf. [6], [12], [13], [14], [15]) is based on the ideas borrowed from the deterministic concept of directional minimization. We discuss it in more detail in Section 2 (see Remark 1), where we also describe our algorithm. In Section 3 properties of the stepsizes are analysed. In Section 4 we establish convergence of the method. Section 5 concerns its asymptotic properties. Finally, in Section 6 some modifications of our algorithm are proposed and a numerical example is studied.

We use $\langle \cdot, \cdot \rangle$ and $|\cdot|$ to denote the usual inner product and norm in n-dimensional Enclidean space. For a set X we denote by diam X its diameter, i.e. diam $X = \sup_{x,y \in X} |x-y|$. Abbreviation w.p.1 is used for "with probability 1".

2. The algorithm and assumptions

The algorithm generates sequences of random directions $\{d^k\}$ and points $\{x^k\}$ in R^n, $k = 0, 1, ...$, according to the following recursive formulae

$$d^k = (\xi^k + I_k \gamma_k d^{k-1})/(1+\gamma_k), \tag{3}$$

$$x^{k+1} = \begin{cases} x^k - \min\{\tau_k(1+\gamma_k),\ t/|d^k|\}\, d^k, & \text{if } x^k \in X, \\ x^0, & \text{if } x^k \notin X, \end{cases} \tag{4}$$

where X is a certain compact set such that $\arg\min_{x \in R^n} F(x) \subset X$, and ξ^k is a stochastic subgradient of F at x^k i.e. $\xi^k = g^k + r^k$, where $g^k \in \partial F(x^k)$ and r^k is a random noise. In (3) and (4) τ_k is a positive stepsize coefficient, γ_k is a nonnegative aggregation coefficient, $I_k \in \{0, 1\}$ is a reset coefficient and $t > 0$. At the starting point $x^0 \in X$, we set $d^{-1} = 0$ and thus it follows from (3) that the direction d^k is a convex combination of the null vector and the previous stochastic subgradients ξ^i, $i = 0, 1, \ldots, k$. We shall call it the aggregate stochastic subgradient. From (4) we deduce that each time the algorithm exceeds set X, we return to the starting point. This concept allow us to stabilize the whole method.

The stepsizes $\{\tau_k\}$ are computed recursively as follows:

$$\begin{aligned} &\tau_0 > 0, \\ &\tau_k = \min\{\bar{\tau}, \tau_{k-1}[\exp\min(\eta, -N_k \alpha u_k - J_k \delta \tau_{k-1})]\}, \\ &\qquad\qquad k = 1, 2, \ldots, \end{aligned} \tag{5}$$

where

$$u_k = \langle \xi^k, \Delta x^k \rangle + \lambda |\Delta x^k|^2, \tag{6}$$

$\Delta x^k = x^k - x^{k-1}$, and $\bar{\tau} > 0$, $\eta > 0$, $\alpha > 0$, $\delta > 0$, λ are fixed parameters. The coefficients N_k, J_k in (5) are binary multipliers satysfying the relations:

$$\begin{aligned} N_k &= 1, \quad \text{if} \quad x^{k-1} \in X, \\ N_k &= 0, \quad \text{if} \quad x^{k-1} \notin X, \end{aligned} \tag{7}$$

$$\begin{aligned} J_k &\in \{0, 1\}, \quad \text{if} \quad |\Delta x^k| \geqslant A\sqrt{\tau_{k-1}}, \\ J_k &= 1, \quad \text{if} \quad |\Delta x^k| < A\sqrt{\tau_{k-1}}, \end{aligned} \tag{8}$$

where A is a small positive constant.

Similar rules are used for determining the aggregation coefficients $\{\gamma_k\}$:

$$\begin{aligned} &\gamma_0 = \gamma_1 \geqslant 0, \\ &\gamma_k = \min\{\bar{\gamma}, \gamma_{k-1} \exp(-N_k \beta v_k - I_{k-1} J_{k-1} \varkappa \gamma_{k-1})\}, \qquad k = 2, 3, \ldots, \end{aligned} \tag{9}$$

with

$$v_k = I_{k-1}(\langle \xi^k, \Delta x^{k-1} \rangle + \lambda \langle \Delta x^k, \Delta x^{k-1} \rangle) \tag{10}$$

and some parameters $\bar{\gamma} > 0$, $\beta > 0$, $\varkappa > 0$.

Finally the reset coefficients $\{I_k\}$ are defined as follows:

$$\begin{aligned} I_k &\in \{0, N_k\}, \quad \text{if} \quad |\xi^{k-1}| \leqslant \bar{\xi}, \\ I_k &= 0, \quad \text{if} \quad |\xi^{k-1}| > \bar{\xi}, \end{aligned} \tag{11}$$

where $\bar{\xi} > 0$ is a fixed threshold.

In further considerations we denote by $\mathscr{F}_k$ the σ-subfield generated by $\{x^0, x^1, ..., x^k, \xi^0, \xi^1, ..., \xi^{k-1}\}$ and by E_k the conditional expectation with respect to $\mathscr{F}_k$.

REMARK 1. To motivate the rules (5) and (9) suppose that the algorithm operates in the interior of X ($N_k = 1$) and $t = \infty$. For given x^{k-1} and d^{k-2} consider the regularized improvement function

$$\varphi_k(\tau, \gamma, I) = E_{k-1}\left[F\left(x^k(\tau, \gamma, I, \xi^{k-1})\right) - F(x^k) + \right.$$
$$\left. + \frac{1}{2}\lambda |x^k(\tau, \gamma, I, \xi^{k-1}) - x^{k-1}|^2\right], \tag{12}$$

where $x^k(\tau, \gamma, I, \xi^{k-1})$ is defined by (3) and (4) i.e.

$$x^k(\tau, \gamma, I, \xi^{k-1}) = x^{k-1} - \tau(\xi^{k-1} + I\gamma d^{k-2}).$$

A natural and the most convenient solution would be to choose τ_{k-1} and γ_{k-1} so as to minimize (12). This is however exteremely difficult to realize. Therefore let us use some values of τ_{k-1} and γ_{k-1}. After simple calculations one obtains

$$E_{k-1}\left(\frac{1}{\tau_{k-1}} u_k, \frac{\tau_{k-1}}{\tau_{k-2}(1+\gamma_{k-2})} v_k\right) \in \partial_{(\tau,\gamma)} \varphi_k(\tau_{k-1}, \gamma_{k-1}, I_{k-1}), \tag{13}$$

provided that $E_k r^k = 0$. Thus the vector (u_k, v_k) may be interpreted as a stochastic subgradient of $\varphi_k(\cdot, \cdot, I_{k-1})$ at $(\tau_{k-1}, \gamma_{k-1})$. It is used in (5) and (9) to correct the coefficients τ_{k-1} and γ_{k-1} for the next iteration.

The additional terms $J_k \delta\tau_{k-1}$ and $I_{k-1} J_{k-1} \varkappa\gamma_{k-1}$ in (5) and (9) are to force a slow decrease of $\{\tau_k\}$ and $\{\gamma_k\}$ in the case of u_k and v_k being close to zero.

REMARK 2. From (4) we deduce that the sequence $\{x^k\}$ is bounded and $\{x^k\} \subset$ $\subset X_t = \{y \in R^n : |y - x| \leqslant t \text{ for some } x \in X\}$. Moreover $|\Delta x^k| \leqslant T = t + \operatorname{diam} X$.

Similar rules for determining sequences $\{\tau_k\}$, $\{\gamma_k\}$ and $\{I_k\}$ but applied to other algorithms were considered in [12], [13], [14]. The main difference is that in our method after each escape from the set X the rules stop for one iteration and the direction d^k is refreshed. In [13], [14] an aggregate subgradient method with projection was analysed. The method consists in the following iterations

$$x^{k+1} = \pi_X\left[x^k - \min\{\tau_k(1+\gamma_k), t/|d^k|\} d^k\right], \quad k = 0, 1, ..., \tag{14}$$

(in [14] $t = \infty$) where π_X is the orthogonal projection onto a certain compact set X. In [12] an unconstrained version of [14] was considered

$$x^{k+1} = x^k - \min\{\tau_k(1+\gamma_k), t/|d^k|\}\, d^k, \qquad k = 0, 1, \ldots \tag{15}$$

(in both (14) and (15) the directions $\{d^k\}$ are computed according to formula (3)). As proved in [13] for a convex objective F all accumulation points of the sequence $\{x^k\}$ generated by method (14) belong to the set $\arg\min_{x\in X} F(x)$ w.p.1. In [12] it was shown that for algorithm (15)

$$\liminf_{k\to\infty} |\nabla F(x^k)| = 0 \qquad \text{w.p.1},$$

$$\limsup_{k\to\infty} F(x^k) = \limsup_{\substack{k\to\infty \\ \nabla F(x^k)\to 0}} F(x^k) \qquad \text{w.p.1},$$

provided that F is differentiable, and there exist constants: $L > 0$, $\mu > 0$, $m > 0$ and $M > 0$ such, that

$$|\nabla F(y) - \nabla F(x)| \leqslant L|y-x| \qquad \text{for all} \qquad x, y \in R^n, \tag{16}$$

$$F(x) \geqslant m|\nabla F(x)|^{\mu} - M \qquad \text{for all} \qquad x \in R^n. \tag{17}$$

In this paper our aim is to weaken those rather strong assumptions imposed on the objective F. To this end we apply another technique of proving convergence properties of algorithm (3)–(11).

Let us formulate the following assumptions.

(H1) The set X is compact.

(H2) There exist a constant ν and a convex, open set $\mathcal{X}_t \supset X_t = \{y \in R^n: |y-x| \leqslant t \text{ for some } x \in X\}$ such that the function $G(x) = F(x) - \nu|x|^2$ is convex on $\mathcal{X}_t$.

(H3) $\lambda + \nu > 0$.

(H4) $\inf_{x\notin X} F(x) > F(x^0)$.

(H5) The set $F(X^*)$, where $X^* = \{x^* \in X: 0 \in \partial F(x^*)\}$ does not contain any segment of nonzero length.

(H6) $\xi^k = q^k + r^k$, where $q^k \in \partial F(x^k)$ and $E_k r^k = 0$ w.p.1 for all $k \geqslant 0$.

(H7) There exist constants $\bar{z} > 0$ and $S > 0$ such that for any $z \in R^n$ with $|z| \leqslant \bar{z}$, one has $E_k \exp(\langle z, r^k\rangle) \leqslant S$ w.p.1 for all $k \geqslant 0$.

REMARK 3. Repeating the argumentation from [10, Theorem 6] we get that condition (H2) is equivalent to the following: for each $x \in \mathcal{X}_t$ there exists some open neighbourhood $\mathcal{X}$ of x such that the objective F has a representation

$$F(x) = \max_{y\in Y} g(x, y),$$

where Y is a compact set, $g: \mathcal{X} \times Y \to R^1$ is a function which has partial derivatives up to the second order with respect to x and which are jointly

continuous on the set $\mathcal{X} \times Y$ (in this case F is lower $-$ C^2 on the set $\mathcal{X}_t$). (H2) is also equivalent to the following condition: there exists a constant v such that for all $x, y \in \mathcal{X}_t$, $g \in \partial F(x)$:

$$F(y) - F(x) \geq \langle g, y-x \rangle + v|y-x|^2. \tag{18}$$

Hence (18) implies that F belongs to the class of weakly convex functions on $\mathcal{X}_t$ (see [7], [8]).

REMARK 4. Assumption (H5) is purely technical; one can hardly imagine a function F for which (H5) doesn't hold.

REMARK 5. (H7) is closely related to the stepsize rules (5) and (9). This assumption is similar in a sense to the Cramer's condition for scalar variables. It holds for each uniformly bounded distributions of $\{r^k\}$, as well as for many unbounded distributions.

From (H7) we obtain that there exists a sequence of constants $\{R_j\}$ such that

$$E_k |r^k|^j \leq R_j \qquad \text{w.p.1 for all} \qquad k \geq 0, \quad j \geq 1. \tag{19}$$

3. Properties of stepsizes

In this section we prove that the sequences $\{\tau_k\}$ and $\{\gamma_k\}$, although determined on-line in a sophisticated way, possess some of the properties usually required from the coefficients in stochastic approximations algorithms. Our argumentation extends and modifies the results obtained in [13].

We start from a property of the noises $\{r^k\}$.

LEMMA 1. *For each $z_0 > 0$, $\varepsilon > 0$ one can find $s_0 > 0$ such that for any $|z| \leq z_0$, $0 \leq s \leq s_0$ and every $k \geq 1$ one has*

$$E_k \exp\left[-s\left(\langle r^k, z\rangle + \varepsilon |z|^2\right)\right] \leq 1 \qquad \text{w.p.1.} \tag{20}$$

Proof: From (H6) follows that the left-hand side of (20) exists for all $0 \leq s \leq \bar{z}/z_0$. Let us assume that $z \neq 0$ (for $z = 0$ (20) is obvious) and use the inequality

$$\exp(-ay) + \exp(ay) \leq 2 + a^2[\exp(-y) + \exp y],$$

which holds for every $|a| \leq 1$ and each $y \in R^1$. Setting $a = s|z|/\bar{z}$ and $y = \bar{z}\langle r^k, z\rangle/|z|$ we obtain the relation

$$\exp\left(-s\langle r^k, z\rangle\right) + \exp\left(s\langle r^k, z\rangle\right) \leq 2 + (s|z|/\bar{z})^2\left[\exp\left(-\bar{z}\langle r^k, z\rangle/|z|\right) + \exp\left(\bar{z}\langle r^k, z\rangle/|z|\right)\right].$$

Let us apply the operator E_k to both sides of the above inequality. By (H7) the conditional expectation of the right-hand side does not

exceed $2+Cs^2|z|^2$, where $C = 2S/\bar{z}^2$. From Jensen's inequality it follows that $E_k \exp(s\langle r^k, z\rangle) \geqslant 1$. Therefore

$$E_k \exp(-s\langle r^k, z\rangle) \leqslant 1+Cs^2|z|^2 \leqslant \exp(Cs^2|z|^2). \tag{21}$$

If $0 \leqslant s \leqslant s_0 = \min\{\bar{z}/z_0, \varepsilon/C\}$, then $Cs^2|z|^2 \leqslant s\varepsilon|z|^2$ and from (21) we obtain (20) as required. ■

Let us define an auxiliary sequence of random variables

$$p_k = \tau_k^{1/\alpha} \exp[F(x^k)], \qquad k = 0, 1, \ldots \tag{22}$$

LEMMA 2. *There exists $\varepsilon > 0$ such that for all $k \geqslant 1$*

$$p_k \leqslant p_{k-1} \exp[-\langle r^k, N_k \Delta x^k\rangle - \varepsilon(|\Delta x^k|^2 + \tau_{k-1})] \text{ w.p.1} \tag{23}$$

Proof: From (5) and (22) we deduce that for $k \geqslant 1$

$$p_k \leqslant p_{k-1} \exp[F(x^k) - F(x^{k-1}) - N_k u_k - J_k \delta\tau_{k-1}/\alpha].$$

By (6) and (18) one has

$$F(x^k) - F(x^{k-1}) - u_k - J_k \delta\tau_{k-1}/\alpha \leqslant$$
$$\leqslant -\langle r^k, \Delta x^k\rangle - (\lambda+\nu)|\Delta x^k|^2 - J_k \delta\tau_{k-1}/\alpha. \tag{24}$$

If $N_k = J_k = 1$, then (23) is satisfied with $\varepsilon_1 = \min\{\lambda+\nu, \delta/\alpha\}$. If $N_k = 1$ and $J_k = 0$ from (8) follows that $|\Delta x^k|^2 \geqslant A^2 \tau_{k-1}$. Then we get inequality (23) with $\varepsilon_2 = \min\{(\lambda+\nu)/2, A^2(\lambda+\nu)/2\}$. In the case of $N_k = 0$ we have $\tau_k \leqslant \tau_{k-1}$, $x^{k-1} \notin X$ and $x^k = x^0$ (see (7), (5) and (4)). By definitions (5) and (4) we obtain: $0 < \tau_{k-1} \leqslant \bar{\tau}$, $|\Delta x^k| \leqslant T$ (see Remark 2) and $p_k \leqslant p_{k-1} \exp[F(x^0) - F(x^{k-1})] \leqslant p_{k-1} \exp[-\varepsilon_3(|\Delta x^k|^2 + \tau_{k-1})]$, where $\varepsilon_3 = \inf_{x \notin X} [F(x) - F(x^0)]/(T^2+\bar{\tau})$. Choosing $\varepsilon = \min\{\varepsilon_1, \varepsilon_2, \varepsilon_3\}$ ($\varepsilon > 0$ by (H3) and (H4)) we get the required result. ■

We are now ready to derive the first important property of step-sizes $\{\tau_k\}$.

LEMMA 3. *For any $s > 0$ one has*

$$\sum_{k=0}^{\infty} E\tau_k^{1+s} < \infty.$$

Proof: We have $|\Delta x^k| \leqslant T$. By Lemmas 1 (with $z = N_k \Delta x^k$) and 2 for all sufficiently small $s > 0$ we obtain

$$E_k p_k^s \leqslant p_{k-1}^s \exp(-s\varepsilon\tau_{k-1}), \qquad k = 1, 2, \ldots.$$

Since $0 < \tau_{k-1} \leqslant \bar{\tau}$ one has $\exp(-s\varepsilon\tau_{k-1}) \leqslant 1 - C\tau_{k-1}$, where $C = [1 - \exp(-s\varepsilon\bar{\tau})]/\bar{\tau}$. Thus $E_k p_k^s \leqslant p_{k-1}^s - C p_{k-1}^s \tau_{k-1}$. Taking the expectation of both sides of this inequality and noting that $p_k > 0$, for all $k \geqslant 0$, we

conclude that $\sum_{k=0}^{\infty} Ep_k^s \tau_k < \infty$. Recalling the definition of $\{p_k\}$ (22) we get $\sum_{k=0}^{\infty} E\tau_k^{1+s} \exp[sF(x^k)] < \infty$ for all sufficiently small $s > 0$. Since $\{x^k\} \subset X_t$, where X_t is compact, the sequence $\{\exp[sF(x^k)]\}$ is bounded from below by some positive constant. Thus $\sum_{k=0}^{\infty} E\tau_k^{1+s} < \infty$ for all sufficiently small $s > 0$. But $0 < \tau_k \leqslant \bar{\tau}$ and hence s may be an arbitrary positive number, which completes the proof. ■

REMARK 6. It is clear from the proof of the above Lemma why the additional term $J_k \delta\tau_{k-1}$ has been inserted into the exponent in (5). Without it (with $\delta = 0$) one can only show that $\sum_{k=1}^{\infty} E(\tau_{k-1}^s |\Delta x^k|^2) < \infty$, for $s > 0$, but this is insufficient for convergence w.p.1. But as proved in [14] in the convex case this condition ensures that for the algorithm with projection (16) (with $\gamma_0 = 0$, $t = \infty$) the sequence of weighted averages:

$$\bar{x}^k = \sum_{i=0}^{k} \tau_i x^i / \sum_{i=0}^{k} \tau_i, \quad k = 0, 1, \ldots,$$

converges to a solution of the problem $\min_{x \in X} F(x)$ w.p.1.

Let us pass to the analysis of the directions $\{d^k\}$.

LEMMA 4. *For all $k \geqslant 0$ one has*

$$|(1+\gamma_k) d^k - \xi^k| \leqslant I_k \gamma_k \bar{\xi} \qquad \text{w.p.1.} \tag{25}$$

Proof: By (3) for $k \geqslant 1$ we have

$$(1+\gamma_k) d^k - \xi^k = \frac{I_k \gamma_k}{1+\gamma_{k-1}} [(1+\gamma_{k-1}) d^{k-1} - \xi^{k-1}] + \frac{I_k \gamma_k}{1+\gamma_{k-1}} \xi^{k-1}.$$

From (11) we obtain that $I_k = I_k^2$ and $I_k |\xi^{k-1}| \leqslant \bar{\xi}$. Hence from the above inequality follows the relation

$$|(1+\gamma_k) d^k - \xi^k| \leqslant \frac{I_k \gamma_k}{1+\gamma_{k-1}} |(1+\gamma_{k-1}) d^{k-1} - \xi^{k-1}| + \frac{I_k \gamma_k}{1+\gamma_{k-1}} \bar{\xi}.$$

Since $(1+\gamma_0) d^0 - \xi^0 = 0$ we get by induction assertion (25). ■

From Lemmas 3 and 4 we deduce the following useful results:

LEMMA 5. *For any integer $j \geqslant 0$ one has:*

$$\sum_{k=0}^{\infty} E(\tau_k^2 |\xi^k|^j) < \infty, \quad \sum_{k=0}^{\infty} E(\tau_k^2 |d^k|^j) < \infty \quad \text{and} \quad \sum_{k=1}^{\infty} E(N_k \tau_{k-1}^{2-j} |\Delta x^k|^j) < \infty.$$

Proof: We have $|\xi^k|^j \leqslant 2^j |g^k|^j + 2^j |r^k|^j$. The series $\sum_{k=0}^{\infty} E(\tau_k^2 |g^k|^j)$ is convergent by Lemma 3 ($s = 1$) and by the boundedness of subgradients $\{g^k\}$ in the compact set X_t. Next it follows from (5) that $\tau_k \leqslant \tau_{k-1} \exp \eta$ and $E(\tau_k^2 |r^k|^j) \leqslant$ $\leqslant \exp(2\eta) E(\tau_{k-1}^2 E_k |r^k|^j) \leqslant R_j \exp(2\eta) E\tau_{k-1}^2$, since τ_{k-1} is $\mathscr{F}_k$ — measurable and $E_k |r^k|^j \leqslant R_j$ (see inequality (19)). Using again Lemma 3 we obtain $\sum_{k=0}^{\infty} E(\tau_k^2 |r^k|^j) < \infty$ and $\sum_{k=0}^{\infty} E(\tau_k^2 |\xi^k|^j) < \infty$. Next from Lemma 4 we deduce that $|d^k|^j \leqslant 2^j |d^k - \xi^k/(1+\gamma_k)|^j + 2^j |\xi^k|^j \leqslant 2^j \bar{\xi}^j + 2^j |\xi^k|^j$. This proves our second assertion. The third assertion is a simple corollary of the second one and of definitions (4) and (7). ■

In the following two lemmas we prove that the rule (5) does not reduce the stepsizes too rapidly.

LEMMA 6. $\lim_{k\to\infty} \tau_{k-1}/\tau_k = 1$ w.p.1 *and* $\lim_{k\to\infty} (1 - \tau_{k-1}/\tau_k) r^k = 0$ w.p.1.

Proof: Consider the exponent in (5). From Lemmas 3 and 5 ($j = 2$) we see that $\tau_k \to 0$ w.p.1 and $N_k |\Delta x^k| \to 0$ w.p.1. We shall prove that $N_k \langle \xi^k, \Delta x^k \rangle \to 0$ w.p.1. We have $N_k \langle \xi^k, \Delta x^k \rangle = N_k \langle g^k, \Delta x^k \rangle + N_k (r^k, \Delta x^k \rangle$. The sequence $\{g^k\}$ is bounded, hence $N_k \langle g^k, \Delta x^k \rangle \to 0$ w.p.1. Next, by (19), (H6) and Lemma 5 ($j = 2$) the series $\sum_{k=1}^{\infty} N_k \langle r^k, \Delta x^k \rangle$ is a convergent martingale and thus $N_k \langle r^k, \Delta x^k \rangle \to 0$ w.p.1. Consequently, the exponent in (5) tends to 0 w.p.1 and $\tau_{k-1}/\tau_k \to 0$ w.p.1, as required. Moreover, we also see that there exists a random index m ($m < \infty$ w.p.1) such that for all $k \geqslant m$ one has both $\tau_{k-1}/\tau_k = \exp(N_k \alpha u_k + J_k \delta \tau_{k-1})$ and $N_k \alpha u_k + J_k \delta \tau_{k-1} \leqslant 1$. Since $\exp(\cdot)$ is convex and increasing, the two preceding relations imply that for $k \geqslant m$ we have

$$|1 - \tau_{k-1}/\tau_k| \leqslant e (N_k \alpha |u_k| + \delta \tau_{k-1})$$

and thus

$$|1 - \tau_{k-1}/\tau_k| |r^k| \leqslant e (N_k \alpha |\xi^k| |\Delta x^k| |r^k| + N_k \alpha |\lambda| |\Delta x^k|^2 |r^k| + \delta \tau_{k-1} |r^k|). \tag{26}$$

We have $N_k |\xi^k| |\Delta x^k| |r^k| \leqslant N_k |\Delta x^k| |r^k|^2 + N_k |g^k| |\Delta x^k| |r^k|$. By (19) $E_k |r^k|^4 \leqslant R_4$, for all $k \geqslant 0$. Therefore $E(N_k |\Delta x^k|^2 |r^k|^4) \leqslant R_4 E(N_k |\Delta x^k|^2)$. From Lemma 5 ($j = 2$) we deduce that $\sum_{k=1}^{\infty} E(N_k |\Delta x^k|^2 |r^k|^4) < \infty$, which implies that $N_k |\Delta x^k| \times$ $\times |r^k|^2 \to 0$ w.p.1 and $N_k |\Delta x^k| |r^k| \to 0$ w.p.1. Hence $N_k |\xi^k| |\Delta x^k| |r^k| \to 0$ w.p.1. In a similar fashion we treat the other components of the right-hand side of (26) and obtain the second assertion of the lemma. The proof is complete. ■

LEMMA 7. $\sum_{k=0}^{\infty} \tau_k = \infty$ w.p.1.

Proof: From Lemma 3 we deduce that $\tau_k \to 0$ w.p.1. By Lemma 6 $\tau_k/\tau_{k-1} = \exp(-N_k \alpha u_k + J_k \delta \tau_{k-1})$ for large indices k. Therefore one must have $\sum_{k=1}^{\infty} (N_k u_k + \tau_{k-1}) = \infty$ w.p.1. Consider the series $\sum_{k=0}^{\infty} N_k u_k = \sum_{k=1}^{\infty} (N_k \langle g^k, \Delta x^k \rangle + N_k \langle r^k, \Delta x^k \rangle + N_k \lambda |\Delta x^k|^2)$. Since $\sum_{k=1}^{\infty} N_k |\Delta x^k|^2 < \infty$ w.p.1 by Lemma 5 ($j = 2$), these components may be left out of account. Next, by (H6), (19) and Lemma 5 ($j = 2$) the series $\sum_{k=1}^{\infty} N_k \langle r^k, \Delta x^k \rangle$ is a convergent martingale and hence does not matter for $\sum_{k=1}^{\infty} (N_k u_k + \tau_{k-1})$ being infinite. Therefore

$$\sum_{k=1}^{\infty} (N_k \langle g^k, \Delta x^k \rangle + \tau_{k-1}) = \infty \qquad \text{w.p.1.} \tag{27}$$

By the compactness of X_t, there exists $\bar{q}$ such that $\langle g^k, \Delta x^k \rangle \leqslant \bar{q} |\Delta x^k|$ for all $k \geqslant 1$. Therefore in view of (7), (4), (5) and Lemma 4

$$N_k \langle g^k, \Delta x^k \rangle \leqslant N_k \bar{q} |\Delta x^k| \leqslant \bar{q} \tau_{k-1} \left[|(1+\gamma_{k-1}) d^{k-1} - \xi^{k-1}| + |\xi^{k-1}| \right] \leqslant \bar{q} \exp \eta \tau_{k-2} (I_{k-1} \gamma_{k-1} \bar{\xi} + |g^{k-1}| + |r^{k-1}|) \leqslant C_1 \tau_{k-2} (C_2 + |r^{k-1}|),$$

with some constants $C_1 > 0$ and $C_2 > 0$. By (19) we have $E_{k-1} |r^{k-1}| \leqslant R_1$. Thus we obtain the inequality $N_k \langle g^k, \Delta x^k \rangle \leqslant C_1 \tau_{k-2} (C_3 + |r^{k-1}| - E_{k-1} |r^{k-1}|)$, where $C_3 = C_2 + R_1$. Since τ_{k-2} is $\mathcal{F}_{k-1}$—measurable and $E_{k-1} (|r^{k-1}| - E_{k-1} |r^{k-1}|)^2 \leqslant R_2$, the series $\sum_{k=2}^{\infty} \tau_{k-2} (|r^{k-1}| - E_{k-1} |r^{k-1}|)$ is a convergent martingale. Therefore (27) implies that

$$\sum_{k=2}^{\infty} (C_1 C_3 \tau_{k-2} + \tau_{k-1}) = \infty \qquad \text{w.p.1,}$$

which yields the required result. ■

Let us now pass to the analysis of the aggregation coefficients $\{\gamma_k\}$.

LEMMA 8. $\lim_{k \to \infty} I_k \gamma_k = 0$ w.p.1.

Proof: Obviously, it is enough to consider the case when $I_k = 1$ infinitely often. From Lemma 5 ($j = 4$) we deduce that $N_k |\Delta x^k| / \sqrt{\tau_{k-1}} \to 0$ w.p.1. Hence by (8) one can choose a random index m ($m < \infty$ w.p.1) such, that $J_k \geqslant N_k$ for all $k \geqslant m$. Define $\mathcal{K} = \{k : I_{k-1} J_{k-1} = 1\}$. By (11) this set is

infinite and $I_{k-1}\, J_{k-1} = I_{k-1}$, for $k > m$. Hence, from (9) and (10) we obtain

$$\limsup_{k\to\infty} \gamma_k = \limsup_{\substack{k\to\infty \\ k\in\mathcal{M}}} \gamma_k.$$

Suppose that $\limsup_{k\to\infty} \gamma_k = \varepsilon > 0$. Proceeding as in the proof of Lemma 6 we obtain $N_k\, v_k \to 0$ w.p.1. Let $\mathcal{K} \subset \mathcal{M}$ be such that $\gamma_k \to \varepsilon$ for $k \to \infty$, $k \in \mathcal{K}$. From (9) we then get $\varepsilon = \lim_{\substack{k\to\infty \\ k\in\mathcal{K}}} \gamma_k \leqslant \limsup_{\substack{k\to\infty \\ k\in\mathcal{K}}} \gamma_{k-1} \exp(-\varkappa\gamma_{k-1}) \leqslant$ $\leqslant \limsup_{k\to\infty} \gamma_k \exp(-\varkappa\gamma_k) < \varepsilon$. We have arrived at a contradiction, which completes the proof. ■

4. Convergence

Having established useful properties of stepsizes and aggregation coefficients we shall prove that our method is convergent to a stationary point of problem (1) w.p.1. Define

$$X^* = \{x^* \in X : 0 \in \partial F(x^*)\}.$$

We start from the following lemma.

LEMMA 9. *There exist sequences of random vectors* $\{s^k\}$ *and* $\{w^k\}$ *such that for all* $k \geqslant 0$ *one has*:

$$x^{k+1} = \begin{cases} x^k - \tau_k (g^k + s^k) + w^k, & \text{if} \quad x^k \in X, \\ x^0, & \text{if} \quad x^k \notin X, \end{cases} \tag{28}$$

where: $g^k \in \partial F(x^k)$, $\lim_{k\to\infty} |s^k| = 0$ w.p.1 *and* $\left|\sum_{k=0}^{\infty} w^k\right| < \infty$ w.p.1.

Proof: Let $x^k \in X$. Denote $t_k = \min\{\tau_k(1+\gamma_k), t/|d^k|\}$.
Then

$$x^{k+1} = x^k - t_k\, d^k. \tag{29}$$

We have:

$$\begin{aligned} t_k\, d^k &= \tau_k(1+\gamma_k)\, d^k + [t_k - \tau_k(1+\gamma_k)]\, d^k = \\ &= \tau_k\, g^k + \tau_k\, r^k + \tau_k [(1+\gamma_k)\, d^k - \xi^k] + [t_k - \tau_k(1+\gamma_k)]\, d^k = \\ &= \tau_k\, g^k + \tau_{k-1}\, r^k + \tau_k [(1+\gamma_k)\, d^k - \xi^k + (1 - \tau_{k-1}/\tau_k)\, r^k] + \\ &\quad + [t_k - \tau_k(1+\gamma_k)]\, d^k. \end{aligned} \tag{30}$$

Using this identity in (29) we get

$$\begin{aligned} s^k &= (1+\gamma_k)\, d^k - \xi^k + (1 - \tau_{k-1}/\tau_k)\, r^k, \\ w^k &= \tau_{k-1}\, r^k + [t_k - \tau_k(1+\gamma_k)]\, d^k. \end{aligned}$$

Directly from Lemmas 8, 4 and 6 we see that $|s^k| \to 0$ w.p.1. By (H6), (19) and Lemma 3 the series $\sum_{k=1}^{\infty} \tau_{k-1} r^k$ is a convergent martingale. Finally, the series $\sum_{k=1}^{\infty} [t_k - \tau_k (1+\gamma_k)] d^k$ is convergent w.p.1 since $t_k = \tau_k (1+\gamma_k)$ for all sufficiently large k. The proof is complete. ■

We are now ready to prove our main result. In [7] for solving problem (1), the following algorithm with deterministic stepsizes $\{\tau_k\}$ was proposed

$$x^{k+1} = \begin{cases} x^k - \tau_k \xi^k = x^k - \tau_k g^k + w^k, & \text{if} \quad x^k \in X, \\ x^0, & \text{if} \quad x^k \notin X, \end{cases} \tag{31}$$

where: $g^k \in \partial F(x^k)$, $\left|\sum_{k=0}^{\infty} w^k\right| < \infty$ w.p.1,

$$\tau_k \geqslant 0, \ \sum_{k=0}^{\infty} \tau_k = \infty, \ \lim_{k\to\infty} \tau_k = 0 \quad \text{and} \quad \lim_{k\to\infty} \frac{\tau_{k-1}}{\tau_k} = 1. \tag{32}$$

By Lemma 9 our algorithm differs from (31) only by the existence of the sequence $\{s^k\}$. One can easily verify that this sequence does not affect its convergence properties (see [7, Theorem 1, pp. 94–100]). Next, from Lemmas 3, 6 and 7 we deduce that conditions (32) are satisfied w.p.1. Hence following the argumentation from [7, Theorem 1, pp. 94–100] or [8, Theorem 1, pp. 109–116] (for a deterministic algorithm, slightly different form (31)) we get our convergence theorem (since we work on paths it doesn't matter that in (28) the sequence $\{\tau_k\}$ is random).

THEOREM 1. *Assume* (H1) *to* (H7). *Then almost surely the sequence* $\{x^k\}$ *generated by algorithm* (3)–(11) *only finitely many times leaves the set* X. *Moreover the sequence* $\{F(x^k)\}$ *is convergent w.p.1 and all accumulation points of the sequence* $\{x^k\}$ *belong to* X^* *w.p.1.*

5. Some asymptotic properties

Although our aim is to accelerate convergence far from the solution of (1), it could be interesting to verify whether our stepsize rules change asymptotic properties of the method when compared with the classical approach (cf. [5], [9]). Clearly the crucial question here is the asymptotic behavior of stepsizes $\{\tau_k\}$ and $\{\gamma_k\}$. It follows from Theorem 1, that for large k equations (4) and (15) are equivalent. Thus $N_k = 1$ for large k. Next, by Lemma 5 ($j = 4$) $|\Delta x^k|/\sqrt{\tau_{k-1}} \to 0$ w.p.1 and $J_k = 1$ for sufficiently large k. Hence we can follow the argumentation from [13, Theorem 2].

THEOREM 2. *Assume* (H1) *to* (H7). *Additionally suppose that* F *is continuously differentiable in an open set* $\mathcal{X}^*$ *containing* X^* *and* $|\nabla F(x)-\nabla F(y)| \leqslant L|x-y|$ *for all* $x, y \in \mathcal{X}^*$ *and some constant* L. *Then*

$$\lim_{k\to\infty} (k+1)\,\tau_k = 1/\delta \qquad \text{w.p.1.} \tag{32}$$

Moreover if there exists a random index k_0 ($k_0 < \infty$ *w.p.1*) *such that* $I_k = 1$ *for all* $k \geqslant k_0$, *then*

$$\lim_{k\to\infty} (k+1)\,\gamma_k = 1/\varkappa \qquad \text{w.p.1.} \tag{33}$$

By Lemma 5 ($j = 4$) and Theorem 1 the conditions used in the second part of Theorem 2 hold if the noises $\{r^k\}$ are uniformly bounded and the reset coefficients are defined as follows (compare with (8) and (11)):

$$I_k \in \{0, N_k\}, \qquad \text{if } |\Delta x^k| \geqslant A\sqrt{\tau_{k-1}},$$

$$I_k = 1 \qquad \text{if } |\Delta x^k| < A\sqrt{\tau_{k-1}},$$

Using the results of Theorem 2 we estimate the convergence rate of our method in the smooth case. We start form the following definition.

DEFINITION 1. We say that the essential supremum in the sense of the expected value of a random value does not exceed c (we write $\limsup_{k\to\infty} \bar{E} z_k \leqslant c$), if there exists $\varepsilon_0 > 0$, a function $h: [0, \varepsilon_0] \to R^1$, right continuous at zero such that $h(0) = c$ and for any $\varepsilon_0 \geqslant \varepsilon > 0$ one can choose random sequences $\{a_k\}$ and $\{b_k\}$ satysfying the following conditions:

$$\begin{aligned} z_k &= a_k + b_k, \qquad \text{for all} \qquad k \geqslant 0, \\ \limsup_{k\to\infty} a_k &\leqslant 0 \qquad \text{w.p.1}, \end{aligned} \tag{34}$$

$$\limsup_{k\to\infty} E b_k \leqslant h(\varepsilon). \tag{35}$$

Similar definition was considered in [9, p. 103]. Having established the definition of the essential supremum we can estimate the asymptotic properties of our algorithm. These properties are stated in the following theorem.

THEOREM 3. *Let the conditions of Theorem 2 be satisfied. Assume that the noises* $\{r^k\}$ *are uniformly bounded and* $\delta < \nu$ (*see* (H2)). *Then*

$$\limsup_{k\to\infty} \bar{E} k\,[F(x^k) - F^*] \leqslant \frac{L R_2}{\delta(\nu-\delta)}, \tag{36}$$

where: $F^* = \min_{x\in R^n} F(x)$ *and* $R_2 \geqslant E_k |r^k|^2$ w.p.1 *for all* $k \geqslant 0$.

Proof: Let Ω be the sample space on which the process $\{x^k\}$ is defined and let Ω_0 be the null set excluded in Theorem 1. Let $\omega \notin \Omega_0$ and consider the path $\{x^k(\omega)\}$. Henceforth we shall for brevity omit the argument ω.

From (18) and the fact that $v > 0$ we obtain

$$F(x) - F^* \leqslant \frac{1}{v} |\nabla F(x)|^2 \qquad \text{for all} \quad x \in \mathcal{X}_t \tag{37}$$

Next, by Theorem 1 $\nabla F(x^k) \to 0$, $x^k \to x^*$ w.p.1, as $k \to \infty$ (from (18) follows that $X^* = \{x^*\}$ in this case) and for all sufficiently large k (see (29) and (30))

$$\Delta x^{k+1} = -\tau_k (1+\gamma_k) d^k = -\tau_k \nabla F(x^k) - \tau_{k-1} r^k - \\ -\tau_{k-1} \left(\frac{\tau_k}{\tau_{k-1}} - 1 \right) r^k - \tau_k [(1+\gamma_k) d^k - \xi^k]. \tag{38}$$

By (5) and (6) for $C_1 = (\exp \eta - 1)/\eta$ we have

$$\left| \frac{\tau_k}{\tau_{k-1}} - 1 \right| \leqslant C_1 |N_k \alpha u_k + J_k \delta \tau_{k-1}| \leqslant C_2 \tau_{k-1} \tag{39}$$

for some constant C_2. From Lemma 4 we deduce that

$$|(1+\gamma_k) d^k - \xi^k| \leqslant I_k \gamma_k \bar{\xi}. \tag{40}$$

Using (32) and (33) in (38)–(40) we get

$$\Delta x^{k+1} = -\tau_k \nabla F(x^k) - \tau_{k-1} r^k + o\left(\frac{1}{k^2}\right), \tag{41}$$

where $\limsup_{k \to \infty} k^2 \left| o\left(\frac{1}{k^2}\right) \right| < \infty$ w.p.1. Hence we obtain

$$|\Delta x^{k+1}|^2 = \frac{|r^k|^2}{\delta^2 k (k+1)} + o\left(\frac{1}{k^2}\right) \qquad \text{w.p.1} \tag{42}$$

and $\lim_{k \to \infty} k^2 o\left(\frac{1}{k^2}\right) = 0$ w.p.1. By (41) and (42)

$$F(x^{k+1}) \leqslant F(x^k) + \langle \nabla F(x^k), \Delta x^{k+1} \rangle + L |\Delta x^{k+1}|^2 = \\ = F(x^k) - \tau_k |\nabla F(x^k)|^2 - \tau_{k-1} \langle \nabla F(x^k), r^k \rangle + \frac{L |r^k|^2}{\delta^2 k (k+1)} + o\left(\frac{1}{k^2}\right). \tag{43}$$

Introducing a new variable $z_k = k [F(x^k) - F^*]$ (43), (37) and (32) yield

$$z_{k+1} \leqslant \left[1 - \frac{v \tau_k (k+1) - 1}{k} \right] z_k + \frac{L |r^k|^2}{\delta^2 k} - (k+1) \tau_{k-1} \langle r^k, \nabla F(x^k) \rangle + \\ + o\left(\frac{1}{k}\right) = \left[1 - \frac{v - \delta}{\delta k} + o\left(\frac{1}{k}\right) \right] z_k + \frac{L |r^k|^2}{\delta^2 k} - \\ - (k+1) \tau_{k-1} \langle r^k, \nabla F(x^k) \rangle + o\left(\frac{1}{k}\right). \tag{44}$$

Let $\varepsilon > 0$ be any constant such that $\nu - \delta - \varepsilon\delta > 0$. Then $\{z_k\}$ has the following representation

$$z_{k+1} \leqslant \left(1 - \frac{\nu - \delta - \varepsilon\delta}{\delta k}\right) z_k + \frac{L|r^k|^2}{\delta^2 k} - (k+1)\,\tau_{k-1}\,\langle r^k, \nabla F(x^k)\rangle + \\ + \frac{\varepsilon}{k} + \frac{p_k - \varepsilon}{k}\, z_k + \frac{s_k - \varepsilon}{k}, \qquad (45)$$

where $p_k \to 0$ and $s_k \to 0$, as $k \to \infty$ w.p.1. Hence from (45) follows that $z_k = a_k + b_k$, $k = 1, 2, \ldots$, where

$$a_{k+1} = \left(1 - \frac{\nu - \delta - \varepsilon\delta}{\delta k}\right) a_k + \frac{p_k - \varepsilon}{k}\, z_k + \frac{s_k - \varepsilon}{k},$$

and

$$b_{k+1} \leqslant \left(1 - \frac{\nu - \delta - \varepsilon\delta}{\delta k}\right) b_k + \frac{L|r^k|^2}{\delta k} - (k+1)\,\tau_{k-1}\,\langle r^k, \nabla F(x^k)\rangle + \frac{\varepsilon}{k}.$$

One can easily check that since for sufficiently large k $p_k - \varepsilon < 0$ and $s_k - \varepsilon < 0$ w.p.1, $\limsup_{k\to\infty} a_k \leqslant 0$ w.p.1. Moreover, since τ_{k-1} is $\mathcal{F}_k$-measurable $E_k r^k = 0$ and $E_k|r^k|^2 \leqslant R_2$, $Eb_{k+1} \leqslant \left(1 - \frac{\nu - \delta - \varepsilon\delta}{\delta k}\right) Eb_k + \frac{LR_2}{\delta^2 k} + \frac{\varepsilon}{k}$. Therefore $\limsup_{k\to\infty} Eb_k \leqslant$ $\leqslant \frac{LR_2 + \varepsilon\delta^2}{\delta(\nu - \delta - \varepsilon\delta)}$ and the theorem follows. ■

Following the above argumentation one can easily prove that in the case of $\tau_k = 1/[\delta(k+1)]$, $\gamma_k = 0$, for $k = 0, 1, \ldots$ (the classical approach) the estimation (36) holds as well. Thus our rules for determining stepsizes and aggregation coefficients do not improve the rate of convergence of the stochastic approximation algorithms. It should be stressed however, that relation (36) can be observed after a very large number of iterations (which are often impossible to perform due to time limitations) and practically the most important is the behavior of the algorithm in the phase when its asymptotic properties do not manifest themselves (see the example in Section 6).

It is worth mentioning that the estimate (36) attains its optimal value for $\delta = \nu/2$. Then $\limsup_{k\to\infty} \bar{E}k\,[F(x^k) - F^*] \leqslant 4LR_2/\nu^2$. But this result has only theoretical importance. Numerical experiments indicate that the value of δ is irrelevant for practical computations (in our example $\delta = 10^{-10}$).

6. Modifications of the method

The basic model (3)–(11) may be modified in various ways so as to improve its practical efficiency while preserving theoretical convergence properties.

Although we assume for simplicity that $E_k \xi^k = g^k$ (see (H6)), similar results may by derived for biased subgradient estimates i.e. if $E_k \xi^k = g^k + b^k$. Then some additional conditions on bias terms are required. For example, instead of (H3) and (H6) one can demand

(H3a) $\sup_{\{k:|\Delta x^k|\neq 0\}} |b^k|/|\Delta x^k| < \lambda + \nu$.

(H6a) $\xi^k = g^k + b^k + r^k$, where $g^k \in \partial F(x^k)$, b^k is $\mathcal{F}_k$-measurable and $E_k r^k = 0$ w.p.1 for all $k \geqslant 0$.

One can easily verify that in this case all results obtained in this paper (except Theorem 3) are in force.

REMARK 7. Assumptions (H3a) and (H6a) are satisfied if the objective $F(x) = Ef(x, \theta)$ is differentiable on R^n, its gradients satisfy condition (16) and we use finite difference stochastic gradient estimation formulae with difference intervals proportional to $|\Delta x^k|$ (cf. [1, pp. 107–112]).

Crucial from the practical point of view are the values of parameters α and β in (5) and (9). With constant values of these parameters, there is a danger of rapid changes of stepsizes and aggregation coefficients due to a wide range of changes of stochastic subgradients $\{\xi^k\}$. To avoid it one can replace α, β, δ and $\varkappa$ with varying coefficients $\{\alpha_k\}$, $\{\beta_k\}$, $\{\delta_k\}$ and $\{\varkappa_k\}$, provided that the following conditions are satisfied (see [13]):

(H8) For all k the coefficients α_k, β_k, δ_k and $\varkappa_k$ are $\mathcal{F}_{k+1}$-measurable.

(H9) $\underline{\alpha} \leqslant \alpha_k \leqslant \bar{\alpha}$, $\underline{\alpha} \leqslant \beta_k \leqslant \bar{\alpha}$, $\underline{\delta} \leqslant \delta_k \leqslant \bar{\delta}$ and $\underline{\delta} \leqslant \varkappa_k \leqslant \bar{\delta}$ w.p.1 for all k and some positive constants $\underline{\alpha}$, $\bar{\alpha}$, $\underline{\delta}$ and $\bar{\delta}$.

(H10) There exist constants $T_1 \geqslant 0$ and $T_2 \geqslant 0$ such that

$$\sum_{i=1}^{k} \frac{1}{\alpha_i} (\ln \tau_i - \ln \tau_{i-1}) \geqslant -T_1 - T_2 \ln \tau_k \text{ w.p.1}, \quad k = 1, 2, \ldots.$$

Under (H1)–(H10) algorithm (3)–(11) remains convergent, i.e. Theorem 1 is still true.

Table 1

Results of computations — adaptive stepsizes and aggregation coefficients

k	τ_k	γ_k	x_1^k	x_2^k	$F(x^k) - F(x^*)$
0	$8.8 \cdot 10^{-4}$	1.0	−1.000	2.000	$1.0 \cdot 10^{2}$
50	$3.3 \cdot 10^{-4}$	$9.6 \cdot 10^{-1}$	−1.316	1.739	5.4
100	$5.2 \cdot 10^{-4}$	1.3	−1.291	1.676	5.3
200	$3.0 \cdot 10^{-3}$	8.2	−0.526	0.287	2.3
300	$1.3 \cdot 10^{-3}$	7.0	1.074	1.150	$7.3 \cdot 10^{-3}$
400	$5.5 \cdot 10^{-4}$	2.5	1.030	1.060	$8.7 \cdot 10^{-4}$
500	$3.0 \cdot 10^{-4}$	$9.9 \cdot 10^{-1}$	1.022	1.047	$8.1 \cdot 10^{-4}$
700	$9.8 \cdot 10^{-5}$	$6.5 \cdot 10^{-1}$	1.021	1.042	$4.5 \cdot 10^{-4}$
1000	$3.1 \cdot 10^{-5}$	$4.7 \cdot 10^{-1}$	1.020	1.041	$4.4 \cdot 10^{-4}$

Table 2

Results of computations — the classical approach: $\tau_k = \tau_0/(k+1)$, $\gamma_k = 0$, $k = 0, 1, ...$

k	τ_k	x_1^k	x_2^k	$F(x^k) - F(x^*)$
0	$8.8 \cdot 10^{-4}$	−1.000	2.000	$1.0 \cdot 10^2$
50	$1.8 \cdot 10^{-5}$	−1.346	1.820	5.5
100	$8.8 \cdot 10^{-6}$	−1.346	1.819	5.5
200	$4.4 \cdot 10^{-6}$	−1.346	1.818	5.5
300	$2.9 \cdot 10^{-6}$	−1.345	1.818	5.5
400	$2.2 \cdot 10^{-6}$	−1.345	1.818	5.5
500	$1.8 \cdot 10^{-6}$	−1.345	1.817	5.5
700	$1.3 \cdot 10^{-6}$	−1.345	1.817	5.5
1000	$8.8 \cdot 10^{-7}$	−1.345	1.816	5.5

On the basis of these assumptions a practical algorithm using adaptively choosen values of $\{\alpha_k\}$ and $\{\beta_k\}$ (different from that described in [13]) was constructed. Below we present a simple numerical example.

EXAMPLE (Rosenbrock's "banana valley").

Consider the problem of minimizing over R^2 the function

$$F(x) = Ef(x, \theta) = E\,[100\,(x_1^2 - x_2)^2 + (x_1 - 1)^2 + \theta_1\,x_1 + \theta_2\,x_2],$$

where θ_1 and θ_2 are independent Gaussian variables with $E\theta_i = 0$, $E\theta_i^2 = 1$, $i = 1, 2$. F attains its minimum at $x^* = (1, 1)$, but is hard to minimize numerically because of ill conditioning. For the purpose of testing the algorithm, at each point x^k the stochastic gradient ξ^k was constructed as $\xi^k = \nabla_x f(x^k, \theta^k)$, where θ^k was drawn from a pseudorandom number generator. The following values of the algorithm parameters were used: $t = 10^{10}$, $\bar{\tau} = 10^{10}$, $\bar{\gamma} = 10^{10}$, $\eta = 1$, $\lambda = 0$, $\delta = 10^{-10}$, $\varkappa = 10^{-10}$ and $\bar{\xi} = 10^{10}$. The results of computations are collected in Table 1. For comparison in Table 2 the outcome of a classical approach ($\tau_k = \tau_0/(k+1)$, $\gamma_0 = 0$, $k = 0, 1, ...$) is presented. In all the above cases τ_0 is chosen so as to minimize $f\,[x^0 - \tau \nabla_x f(x^0, E\theta), E\theta]$ over $\tau \geqslant 0$.

7. Conclusions

The method described in this paper appears to be an efficient tool for solving stochastic, unconstrained optimization problems. Its efficiency is due to the well-known trick of averaging stochastic subgradients and to adaptive on-line rules for determining stepsizes. The computational results indicate that the coefficients are rapidly adjusted to proper values providing a significant progress towards minimum. Although our algorithm has the same

asymptotic properties as the classical method based on the harmonical choice of stepsizes its practical efficiency is much better.

It seems that similar rules may be inserted into many other stochastic approximation algorithms.

References

[1] ERMOLIEV Ju. M. Metody stochastičeskogo programmirovanija. Moskva, Nauka, 1976.

[2] ERMOLIEV Y. M., LEONARDI G., VIRA J. The stochastic quasi-gradient method applied to a facility location problem. Raport W.P. 81/14, IIASA, Laxenburg, 1981.

[3] GUPAL A. M. Stochastičeskie metody rešenija negladkich ekstremalnych zadač. Kiev, Naukova Dumka, 1979.

[4] KOROSTELEV F. P. O mnogošagovych procedurach stochastičeskoj optimizacii. *Avtomatika i Telemechanika*, **5** (1981), 82–90.

[5] KUSHNER H. J., CLARK D. S. Stochastic Approximation Methods for Constrained and Unconstrained Systems. New York, Springer, 1978.

[6] MIRZOACHMEDOV F., URJASEV S. P. Adaptivnaja regulirovka šaga dlja algoritma stochastičeskoj optimizacii. *Žurnal vyčislitelnoj matematiki i matematičeskoj fiziki*, **6** (1983), 1314–1325.

[7] NURMINSKIJ E. A. Čislennye metody rešenija determinirovannych i stochastičeskich minimaksnych zadač. Naukova Dumka, 1979.

[8] NURMINSKI E. A. Subgradient method for minimizing weakly convex functions and ε-subgradient methods of convex optimization. Progress in Nondifferentiable Optimization E. A. Nurminski, ed., CP-82-58, Laxenburg, Int. Institute for Applied Systems Analysis, 1982.

[9] POLJAK B. T., CYPKIN Ja. Kriterjalnyje algoritmy stochastičeskoj optimizacii. *Avtomatika i Telemechanika*, **6** (1984), 95–104.

[10] ROCKAFELLAR R. T. Favorable classes of Lipschitz-continuous functions in subgradient optimization. Progress in Nondifferentiable Optimization, E. A. Nurminski, ed., CP-88-58, Laxenburg. Int. Institute for Applied Systems Analysis, 1982.

[11] RUSZCZYŃSKI A., SYSKI W. Stochastic approximation method with gradient averaging for unconstrained problems. *IEEE Transactions on Automatic Control*, **28** (1983), 1097–1105.

[12] RUSZCZYŃSKI A., SYSKI W. Stochastic approximation algorithm with gradient averaging and on line stepsize rules. Preprints of the 9th Congress of IFAC. J. Gertler, L. Kevicky, eds. Vol. 7 (1984), 230–234.

[13] RUSZCZYŃSKI A., SYSKI W. A method of aggregate stochastic subgradients with on-line stepsize rules for convex stochastic programming problems. *Math. Programming Study on Stochastic Programming*, **28** (1986), 113–131.

[14] RUSZCZYŃSKI A., SYSKI W. On convergence of the stochastic subgradient method with on line stepsize rules. *J. of Math. Analysis and Applications*, **2** (1986), 512–527.

[15] URJASEV S. P. Regulirovka šaga dlja prjamych metodov stochastičeskogo programmirovanija. *Kibernetika*, **6** (1980), 85–87.

Received, November 1985.

Algorytm aproksymacji stochastycznej z filtracją subgradientu i z wyborem współczynników kroku na bieżąco dla zadań niegładkich, niewypukłych i bez ograniczeń

W niniejszej pracy przedstawiono praktyczny algorytm aproksymacji stochastycznej dla znajdowania minimum niegładkiej i niewypukłej funkcji celu w przypadku zadania bez ograniczeń. Do wyznaczania kierunków poszukiwań algorytm wykorzystuje pomocniczy filtr, który uśrednia stochastyczne subgradienty funkcji celu. Współczynniki kroku i agregacji są określane on-line na bazie informacji zebranej w czasie obliczeń, zgodnie z regułami wynikającymi z koncepcji zregularyzowanej funkcji poprawy. Udowodniono zbieżność metody z prawdopodobieństwem 1 oraz zbadano asymptotyczne własności metody. Działanie algorytmu zilustrowano przykładem obliczeniowym.

Алгоритм стохастической аппроксимации с фильтрацией субградиента и с текущим выбором коэффициентов для негладких, невыпуклых задач без ограничений

В данной работе представлен практический алгоритм стохастической аппроксимации для нахождения минимума негладкой и невыпуклой функции цели в случае задачи без ограничений. Для определения направления поиска алгоритм использует вспомогательный фильтр, который усредняет стохастические субградиенты функции цели. Коэффициенты шага и агрегирования определяются непосредственно на основе информаций накопленных по ходу вычислений, согласно правилам, вытекающим из идеи регуляризованной функции улучшения. Доказана сходимость метода с вероятностью 1, а также исследованы асимптотические свойства метода. Действие алгоритма иллюстрируется на численном примере.