A stochastic approximation algorithm with multiplicative step size modification

Alexander Plakhov
plakhov@ua.pt
Pedro Cruz
pedrocruz@ua.pt
Department of Mathematics
University of Aveiro — Portugal

Abstract

An algorithm of searching a zero of an unknown function $\varphi : \mathbb{R} \to \mathbb{R}$ is considered: $x_t = x_{t-1} - \gamma_{t-1} y_t$, $t = 1,\ 2, \ldots$, where $y_t = \varphi(x_{t-1}) + \xi_t$ is the value of φ measured at x_{t-1} and ξ_t is the measurement error. The step sizes $\gamma_t > 0$ are modified in the course of the algorithm according to the rule: $\gamma_t = \min\{u\,\gamma_{t-1},\ \bar{g}\}$ if $y_{t-1} y_t > 0$, and $\gamma_t = d\,\gamma_{t-1}$, otherwise, where $0 < d < 1 < u$, $\bar{g} > 0$. That is, at each iteration γ_t is multiplied either by u or by d, provided that the resulting value does not exceed the predetermined value $\bar{g}$. The function φ may have one or several zeros; the random values ξ_t are independent and identically distributed, with zero mean and finite variance. Under some additional assumptions on φ, ξ_t, and $\bar{g}$, the conditions on u and d guaranteeing a.s. convergence of the sequence $\{x_t\}$, as well as a.s. divergence, are determined. In particular, if $\mathrm{P}(\xi_1 > 0) = \mathrm{P}(\xi_1 < 0) = 1/2$ and $\mathrm{P}(\xi_1 = x) = 0$ for any $x \in \mathbb{R}$, one has convergence for $ud < 1$ and divergence for $ud > 1$. Due to the multiplicative updating rule for γ_t, the sequence $\{x_t\}$ converges rapidly: like a geometric progression (if convergence takes place), but the limit value may not coincide with, but instead, approximates one of the zeros of φ. By adjusting the parameters u and d, one can reach arbitrarily high precision of the approximation; higher precision is obtained at the expense of lower convergence rate.

Key words: stochastic approximation, accelerated convergence, step size adaptation.

AMS subject classification: 62L20 (Stochastic approximation), 90C15 (Stochastic programming), 93B30 (System identification)

1 Introduction

Suppose that we are given a function $\varphi : \mathbb{R} \to \mathbb{R}$; it is required to find a zero of φ. The function φ can be measured at any point x with some error, so that the measured value is $y = \varphi(x) + \xi$; the measurement error ξ is a random value with zero mean. The standard stochastic approximation algorithm consists in calculating successive approximations of the required value, $x_0,\ x_1,\ x_2, \ldots$, according to the rule

$$x_t = x_{t-1} - \gamma_{t-1} y_t, \quad t = 1,\ 2, \ldots, \tag{1}$$

where

$$y_t = \varphi(x_{t-1}) + \xi_t. \tag{2}$$

Usually it is assumed that the step sizes of the algorithm, $\gamma_0,\ \gamma_1,\ \gamma_2, \ldots$, are positive real numbers satisfying the relations $\sum \gamma_t = \infty,\ \sum \gamma_t^2 < \infty$. Then, under some additional assumptions on the function φ and the sequence $\{\xi_t\}$, the algorithm a.s. converges to a zero point x_* of φ (see, e.g., [1, 2]).

It is also known how to choose the coefficients $\gamma_0,\ \gamma_1,\ \gamma_2, \ldots$, in order to ensure the highest possible convergence rate [1, 2]. Unfortunately, to make this choice, one needs to know the derivative $\varphi'(x_*)$ at the required point. This difficulty was overcome in the papers [3, 4], where a modification of the basic algorithm was proposed (the Polyak-Ruppert algorithm with averaging of iterates). This new algorithm does not need any *a priori* information of φ and has the best asymptotic convergence rate. There were also obtained generalizations of these results to the case of many dimensions, which is more important for applications [3, 5].

Asymptotical optimality implies that

$$E\left(x_t - x_*\right)^2 = c\, t^{-1/2}\left(1 + o(1)\right), \tag{3}$$

where c is a positive constant (a positive definite matrix in the multidimensional case) and cannot be diminished. The problem, however, is that when solving practical tasks, $o(1)$ in the right hand side of (3) may be very large, and it may take very much time until this value becomes comparable with 1. Therefore, an asymptotically optimal algorithm can be unsatisfactory on any reasonable time interval.

There were proposed various stochastic approximation algorithms, aimed at increasing the efficiency of algorithm on reasonable time scales. In particular, there was used the idea that the step size values γ_t should be random, rather than deterministic, and should be modified in the course of algorithm, in accordance with the current data. (See [6, 9, 12, 14] for heuristic algorithms utilizing this idea.) In this way, Kesten [10] studied an algorithm using the rule (1), (2) and the modification rule for γ_t:

$$\gamma_t = \gamma(s_t), \qquad s_t = \begin{cases} s_{t-1}, & \text{if} \quad y_{t-1} y_t > 0 \\ s_{t-1} + 1, & \text{if} \quad y_{t-1} y_t \le 0, \end{cases} \qquad t = 2, 3, \ldots, \tag{4}$$

where $s_0 = 0,\ s_1 = 1;\ \gamma(0),\ \gamma(1),\ \gamma(2), \ldots$ is a decreasing sequence of positive numbers such that $\sum \gamma(m) = \infty,\ \sum \gamma^2(m) < \infty$. Thus, the step size γ_t cannot increase in the course of the algorithm; it can only decrease or remain unchanged. If the sign of several consecutive increments $\Delta x_t = x_t - x_{t-1}$ remains unchanged, one can admit that the algorithm is still far from the required solution x_*; in this case, according to (4) and (1), γ_t gets "frozen". On the other hand, if the sign of Δx_t changes frequently, it seems probable that x_t oscillate around the solution x_*, and, according to (4) and (1), γ_t decreases. Kesten proved that if there is a unique zero of φ then x_t converges to this value with probability 1. A multidimensional version of this algorithm was studied in [11].

Yet, the step size "adaptation" of this algorithm is not rapid enough. On the other hand, there are heuristical algorithms (in particular, in artificial neural networks) utilizing a multiplicative step size modification rule: depending on the current data, the step size is multiplied either by a constant greater than 1, or by a positive constant less than 1 [7, 8, 12, 13].

The step size of these algorithms is modified very rapidly, and based on this rule, one can reach reasonably fast convergence. However, the sequence of step sizes may converge like a geometric progression, and therefore the limit value of the algorithm needs not to coincide with the true solution x_*. Nevertheless, the utilization of such algorithms may be justified if they produce an output value close enough to the true solution.

In this paper, we study analytically a stochastic approximation algorithm utilizing a multiplicative rule of step size modification. The algorithm consists in the rule (1), (2) combined with the following rule

$$\gamma_t = \begin{cases} \min\{u\,\gamma_{t-1}, \bar{\mathrm{g}}\}, & \text{if} \quad y_{t-1}y_t > 0, \\ d\,\gamma_{t-1}, & \text{if} \quad y_{t-1}y_t \le 0, \end{cases} \qquad t = 2, 3, \ldots. \tag{5}$$

Here $0 < d < 1 < u$, $0 < \gamma_0, \gamma_1 \le \bar{\mathrm{g}}$, $\bar{\mathrm{g}}$ is a positive constant. The main differences between (5) and Kesten's rule (4) are the following. First, in our algorithm the step size may both decrease and increase. Second, in Kesten's algorithm one always has $\sum \gamma_t = \infty$, while in our algorithm (1), (2), (5) it looks likely that (in the case of convergence of the algorithm) $\{\gamma_t\}$ converges like a geometric progression (this conjecture will be justified in section 3), therefore the limit of $\{x_t\}$ may not be a zero point of φ.

Let us consider a simple illustrative example. Take the function $\varphi(x) = x/\sqrt{x^2+1}$ and consider the problem of computing the zero of φ (which is obviously $x_* = 0$). We compare convergence properties for different algorithms with the same initial state $x_0 = 20$ and initial step $\gamma_0 = 1$. The variables of noise ξ_t are taken to be i.i.d. $\mathcal{N}(0, 1)$. The step sizes for the standard stochastic approximation algorithm (SA) are $\gamma_t = 1/(1 + t)$; this choice ensures asymptotically optimal convergence of the algorithm. For Kesten algorithm it was taken $\gamma(s) = 1/(1 + s)$. For Polyak-Ruppert algorithm (PR) we chose $\gamma_t = 1/\sqrt{1+t}$. For the multiplicative step size algorithm (MUL), the parameter values $\gamma_0 = \gamma_1 = 1$, $\bar{\mathrm{g}} = 1$, and $d = 0.95$, with three successively increasing values of u, (a) $u = 1.01$, (b) $u = 1.03$, and (c) $u = 1.05$, were chosen. Note that we always have $ud < 1$, and the value ud becomes successively closer to 1 in the cases (a), (b), and (c).

For a fixed precision ε, we calculated the average time (average number of iterations) needed to reach this precision, $|x_t - x_*| < \varepsilon$. The obtained diagrams, for the Kesten and MUL (a), (b), (c) algorithms, are shown on the figure. At each point, the algorithms were repeated 10 times. The precision varies between 10^{-2} and 10^{-5}. The corresponding values for SA and PR algorithms are larger than 10^5 and are beyond the scope of the figure.

The number t of iterations needed for MUL to reach a given precision ε gradually increases when ε decreases, and jumps sharply exceeding the limiting value 10^5 adopted by us when ε exceeds a certain value, indicating that better precision cannot be attained.

For $d = 0.95$ and $u = 1.01$, the best possible precision is $\varepsilon = 10^{-3}$ and the average number of iterations needed to reach it is approximately $t = 200$. For $d = 0.95$ and $u = 1.03$, the corresponding values are $\varepsilon = 4 \cdot 10^{-4}$ and $t = 300$, and for $d = 0.95$ and $u = 1.05$, they are $\varepsilon = 2.5 \cdot 10^{-5}$ and $t = 2400$. For the Kesten algorithm, the number of iterations increases from approximately 8000 to 16000, when ε decreases from 10^{-2} to 10^{-5}.

This example points out a characteristic feature of our algorithm: it can *rapidly* give an *approximate* answer. By adjusting the parameters (getting u closer to $1/d$), the quality of answer can be improved at the expense of rapidity of the response.

The mathematical formulation of this feature is given by the convergence theorem, which is stated in section 2 and proved in section 3. In short, the result is as follows. Under the assumptions A1–A6 on φ, ξ_t, and $\bar{\mathrm{g}}$, stated below, the process (1), (2), (5) is proved to a. s. converge (not necessarily to a zero of φ) if some relation between u and d holds, and diverge if another relation between u and d holds. In the particular case, where

$$\mathrm{P}(\xi_t > 0) = \mathrm{P}(\xi_t < 0) = 1/2, \tag{6}$$

these relations take an especially simple form: the process a. s. converges if $ud < 1$ and diverges if $ud > 1$. Moreover, a monotone decreasing family of closed sets $\mathcal{U}(\lambda) \subset \mathbb{R},\ 0 < \lambda < 1$ is determined, such that for any λ, $\mathcal{U}(\lambda)$ contains the set Z of zeros of φ and $\cap_\lambda \mathcal{U}(\lambda) = \mathrm{Z}$. It is proved that, in the case of convergence, the limit of $\{x_t\}$ belongs to $\mathcal{U}\left(\frac{\ln u}{\ln(1/d)}\right)$.

Thus, by adjusting the parameters u and d (for example, if (6) holds, one can fix d and let $u \to 1/d - 0$), one can reach arbitrarily high precision of the algorithm; higher precision is obtained at the expense of lower convergence rate.

The algorithm is stated and studied only in the one-dimensional case while, from the applications viewpoint, the multidimensional case is more interesting. In fact, as will be seen from the following, even the study of the one-dimensional case is quite complicated; moreover, at the moment it is not completely clear how to generalize the algorithm to the multidimensional case. Therefore, the multidimensional case is postponed to the future.

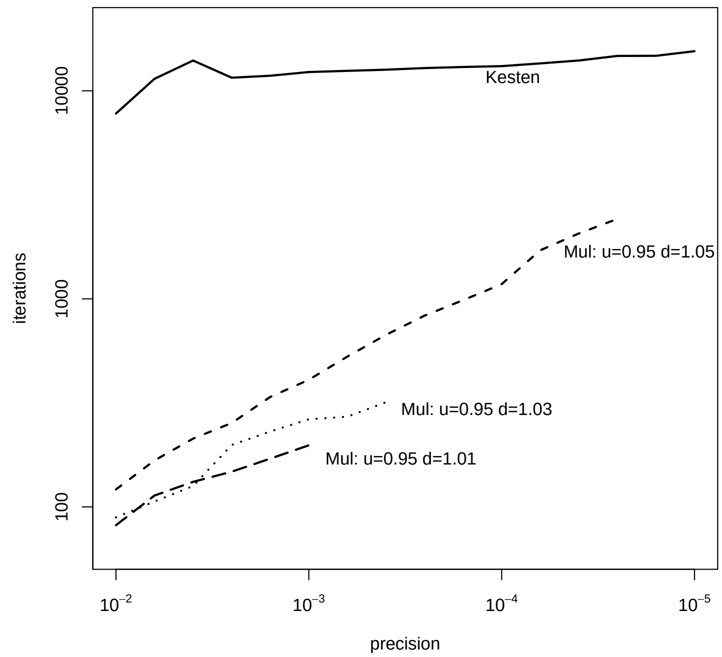

Figure 1: The average number of iterations needed to reach a given precision is shown. The results for the MUL algorithm with several combinations of parameters and for the Kesten algorithm are given. It is seen that the MUL algorithm is generally more rapid that the Kesten one, but cannot reach precision better than a fixed value.

2 Definition of the algorithm and statement of the main result

Consider the algorithm given by (1), (2), (5). The rule (5) means that at each instant t, step size is multiplied by u or by d, if the result of multiplication is less than $\bar{\mathrm{g}}$; otherwise, step size is set to be $\bar{\mathrm{g}}$. Thus, the maximal possible value of step size equals $\bar{\mathrm{g}}$.

The rule (5) can be written in the form

$$\begin{aligned} \ln \tilde{\gamma}_t &= \ln \gamma_{t-1} + \ln u \cdot \ \mathbb{I}(y_{t-1} y_t > 0) + \ln d \cdot \ \mathbb{I}(y_{t-1} y_t \le 0), \\ \ln \gamma_t &= \min\{\ln \tilde{\gamma}_t, \ln \bar{\mathrm{g}}\}. \end{aligned} \tag{7}$$

Let us take the following assumptions:

A1 Denote $\mathcal{F}_t$, $t = 0, 1, 2, \ldots$ the σ-algebra generated by x_i, γ_i, and ξ_i, $0 \le i \le$

t; then ξ_{t+1} does not depend on $\mathcal{F}_t$.

A2 The values ξ_t are identically distributed, with zero mean and finite variance: $\mathrm{E}\xi_t = 0$, $\mathrm{Var}\xi_t =: S < +\infty$.

A3 (a) There exists $L > 0$ such that for any interval $I \subset [-L, L]$, $\mathrm{P}(\xi_1 \in I) > 0$; (b) $\mathrm{P}(\xi_1 = 0) = 0$.

A4 $\varphi \in \mathbb{C}^1(\mathbb{R})$ and $\sup_x |\varphi'(x)| =: M < \infty$.

A5 $\bar{\mathrm{g}} < 2/M$.

A6 There exists $R > 0$ such that

(a) $x\varphi(x) > 0$ as $|x| \geq R$, and

(b) $\displaystyle\inf_{|x|\geq R} \varphi^2(x) > \frac{\bar{\mathrm{g}}MS}{2 - \bar{\mathrm{g}}M}$.

Remark 1 *From A4 and A6 (a) it follows that the set* Z *is non-empty and is contained in* $(-R, R)$.

Remark 2 *Note that assumptions A4–A6 guarantee convergence of the deterministic counterpart of algorithm (1), (2), (5) (that is, of the algorithm with* $\xi_t \equiv 0$*). Moreover, under these conditions, any deterministic algorithm* $x_t = x_{t-1} - \gamma_{t-1}\varphi(x_{t-1})$ *converges, whatever the sequence* $\{\gamma_t\}$ *satisfying* $\gamma_t \leq \bar{\mathrm{g}}$.

Introduce the functions:

$$k_+(z) := \lim_{\epsilon\to 0^+} \sup_{\varphi_1,\varphi_2} \{\mathrm{P}(\varphi_1 + \xi_1)(\varphi_2 + \xi_2) > 0);\ \ |\varphi_1 - z| < \epsilon,\ |\varphi_2 - z| < \epsilon\},\quad (8)$$

$$k_-(z) := \lim_{\epsilon\to 0^+} \inf_{\varphi_1,\varphi_2} \{\mathrm{P}((\varphi_1 + \xi_1)(\varphi_2 + \xi_2) > 0):\ \ |\varphi_1 - z| < \epsilon,\ |\varphi_2 - z| < \epsilon\};\quad (9)$$

one has $k_+(z) \geq 1/2,\ 0 \leq k_\pm(z) \leq 1,\ \lim_{z\to\infty} k_\pm(z) = 1$.

Further, define the sets of real numbers

$$V_\pm^{(a)} := \{x : k_\pm(\varphi(x)) < a\},\quad V_\pm^{[a]} := \{x : k_\pm(\varphi(x)) \leq a\};\qquad (10)$$

obviously, $V_+^{(a)} \subset V_-^{(a)},\ V_\pm^{(a)} \subset V_\pm^{[a]}$ for any a.

Note that $V_+^{(a)}$ is open. Indeed, let $x \in V_+^{(a)}$, then there exists $\epsilon > 0$ such that

$$\sup_{\varphi_1,\varphi_2} \{\mathrm{P}((\varphi_1 + \xi_1)(\varphi_2 + \xi_2) > 0):\ \ |\varphi_1 - \varphi(x)| < \epsilon,\ |\varphi_2 - \varphi(x)| < \epsilon\} =: c < a.$$

Then for x' close enough to x one has $|\varphi(x') - \varphi(x)| < \varepsilon/2$, hence

$$\sup_{\varphi_1,\varphi_2} \{\mathrm{P}((\varphi_1 + \xi_1)(\varphi_2 + \xi_2) > 0):\ \ |\varphi_1 - \varphi(x')| < \epsilon/2,\ |\varphi_2 - \varphi(x')| < \epsilon/2\} \leq c < a.$$

This implies that $k_+(\varphi(x')) < a$, hence $x' \in V_+^{(a)}$.

Denote also

$$\mathrm{k} := \frac{\ln(1/d)}{\ln(u/d)}. \tag{11}$$

Denote by Z the set of zeros of φ, i.e., $\mathrm{Z} := \{x : \varphi(x) = 0\}$. Suppose that $x \in V_+^{(\mathrm{k})}$, $x_{t-2} \in (x - \epsilon, x + \epsilon) \subset V_+^{(\mathrm{k})}$, and $\gamma_{t-2} < \epsilon$, where ϵ is a small positive number. Then, with a probability close to 1, x_{t-1} also belongs to a small (possibly larger) neighborhood of x contained in $V_+^{(\mathrm{k})}$, and taking into account (8) and (10), one gets

$$\begin{aligned}&\mathrm{P}(y_{t-1}y_t > 0 \,\Big|\, |x_{t-2} - x| < \epsilon, \gamma_{t-2} < \epsilon) = \\ &= \mathrm{P}((\varphi(x_{t-2}) + \xi_{t-1})(\varphi(x_{t-1}) + \xi_t) > 0 \,\Big|\, |x_{t-2} - x| < \epsilon, \gamma_{t-2} < \epsilon) < \mathrm{k}.\end{aligned}$$

Then, using (7) and (11), one obtains

$$\begin{aligned}&\mathrm{E}[\ln\gamma_t - \ln\gamma_{t-1} \,\Big|\, |x_{t-2} - x| < \epsilon, \gamma_{t-2} < \epsilon] \leq \\ &\ln u \cdot \mathrm{P}(y_{t-1}y_t > 0 \,\Big|\, |x_{t-2} - x| < \epsilon, \gamma_{t-2} < \epsilon) + \ln d \cdot \mathrm{P}(y_{t-1}y_t \leq 0 \,\Big|\, |x_{t-2} - x| < \epsilon, \gamma_{t-2} < \epsilon) \\ &< \ln u \cdot \mathrm{k} + \ln d \cdot (1 - \mathrm{k}) = 0.\end{aligned}$$

Thus, in a sense, the set $V_+^{(\mathrm{k})}$ can be regarded to be a *domain of decrease of step size*: if several consecutive values of x_t belong to $V_+^{(\mathrm{k})}$ and are close enough to each other, and if the first term of the sequence of corresponding step sizes γ_t is small enough, then the sequence of mean values $E[\ln\gamma_t]$ decreases.

Now, suppose that $x \in \mathbb{R} \setminus V_-^{[\mathrm{k}]}$, $x_{t-2} \in (x - \epsilon, x + \epsilon) \subset \mathbb{R} \setminus V_-^{[\mathrm{k}]}$, and that $\gamma_{t-2} < \epsilon$. Analogously, for ϵ small enough, one has

$$\mathrm{P}(y_{t-1}y_t > 0 \,\Big|\, |x_{t-2} - x| < \epsilon, \gamma_{t-2} < \epsilon) > \mathrm{k},$$

and then, using again (7) and (11) and taking into account that for $\epsilon < \bar{\mathrm{g}}/u^2$, $\tilde{\gamma}_t = \gamma_t$, one obtains

$$\begin{aligned}&\mathrm{E}[\ln\gamma_t - \ln\gamma_{t-1} \,\Big|\, |x_{t-2} - x| < \epsilon, \gamma_{t-2} < \epsilon] = \\ &\ln u \cdot \mathrm{P}(y_{t-1}y_t > 0 \,\Big|\, |x_{t-2} - x| < \epsilon, \gamma_{t-2} < \epsilon]) + \ln d \cdot \mathrm{P}(y_{t-1}y_t \leq 0 \,\Big|\, |x_{t-2} - x| < \epsilon, \gamma_{t-2} < \epsilon]) \\ &> \ln u \cdot \mathrm{k} + \ln d \cdot (1 - \mathrm{k}) = 0.\end{aligned}$$

Thus, the set $\mathbb{R} \setminus V_-^{[\mathrm{k}]}$ can be regarded as a *domain of increase of step size*: if several consecutive values of x_t belong to $\mathbb{R} \setminus V_-^{[\mathrm{k}]}$ and are close enough to each other, and if the first of the corresponding values of γ_t is small enough, then the sequence of mean values $E[\ln\gamma_t]$ increases.

Note that if $\mathrm{k} > k_+(0)$ then, by virtue of (10), $\mathrm{Z} \subset V_+^{(\mathrm{k})}$, that is, all the zeros of φ belong to the region of decrease of step size. On the other hand, if $\mathrm{k} < \inf_z k_-(z)$ then $V_-^{[\mathrm{k}]} = \emptyset$, which means that the region of increase of step size coincides with $\mathbb{R}$.

It seems likely that in the first case the algorithm can converge, and in the second one, cannot. This conjecture is confirmed by the following theorem, which is the main result of the paper.

Theorem *Let the assumptions A1–A6 be satisfied; consider the process* $\{x_t,\ \gamma_t\}$ *defined by (1), (2), (5). Recall that* $\mathrm{k} = \frac{\ln(1/d)}{\ln(u/d)}$*. Then*
(a) If $\mathrm{k} > k_+(0)$ *then* $\{x_t\}$ *a.s. converges to a point from* $V_-^{[\mathrm{k}]}$*.*
(b) If $\mathrm{k} < \inf_z k_-(z)$ *then* $\{x_t\}$ *a.s. diverges.*

Suppose that $\mathrm{P}(\xi_1 = x) = 0$ for any real x and that $\mathrm{P}(\xi_1 > 0) = \mathrm{P}(\xi_1 < 0)$. Then the function $k(\cdot) := k_+(\cdot)$ coincides with $k_-(\cdot)$, is continuous, and is given by

$$k(z) = \mathrm{P}((z + \xi_1)(z + \xi_2) > 0);$$

$z = 0$ is the unique minimum of $k(\cdot)$, and $k(0) = \inf_z k(z) = 1/2$. After a simple algebra, one can rewrite the hypotheses of theorem in the form (a) $ud < 1$, (b) $ud > 1$. Denote $\mathcal{U}(\lambda) := V^{[\frac{1}{1+\lambda}]} = \{x:\ k(\varphi(x)) \le \frac{1}{1+\lambda}\}$; $\mathcal{U}(\lambda),\ 0 < \lambda < 1$ is a monotone decreasing family of sets containing Z and tending to Z as $\lambda \to 1^-$.

Thus, one comes to

Corollary *Let, in addition to assumptions A1–A6,* $\mathrm{P}(\xi_1 = x) = 0$ *for any* $x \in \mathbb{R}$*, and* $\mathrm{P}(\xi_1 > 0) = \mathrm{P}(\xi_1 < 0) = 1/2$*. Consider the process defined by (1), (2), (5). Then there exists a monotone decreasing family of sets* $\mathcal{U}(\lambda)$*,* $0 < \lambda < 1$ *such that* $\mathcal{U}(\lambda) \supset \mathrm{Z}$*,* $\partial(\mathcal{U}(\lambda), \mathrm{Z}) \to 0$ *as* $\lambda \to 1^-$*, and*
(a) if $ud < 1$ *then* $\{x_t\}$ *a.s. converges to a point from* $\mathcal{U}(\frac{\ln u}{-\ln d})$*;*
(b) if $ud > 1$ *then* $\{x_t\}$ *a.s. diverges.*

Remark 3 *Theorem does not give any information about behavior of the algorithm for the values* $u,\ d$ *such that*

$$\inf_z k_-(z) \le \frac{\ln(1/d)}{\ln(u/d)} \le k_+(0).$$

In particular, under the hypotheses of corollary, the case $ud = 1$ *remains unexplored. These issues will be addressed elsewhere.*

3 Proof of theorem

First we prove 10 auxiliary lemmas, and then, basing on them, we prove theorem.

In the sequel, we shall mainly designate random values by Greek letters, and real numbers and functions from $\mathbb{R}$ to $\mathbb{R}$, by Latin ones; the letters $t,\ i,\ j,\ s$ will denote integer non-negative numbers. The function φ and the random values $x_t,\ y_t$ are exceptions; also, traditional notation $\epsilon,\ \delta$ for small positive numbers will be used.

In what follows, all statements about random variables are supposed to be true almost surely.

Lemma 1 *If* $\sum_t \gamma_t < \infty$ *then the sequence* $\{x_t\}$ *converges.*

Proof. Note that without loss of generality one can assume that x_0 is bounded. Indeed, replacing x_0 by $\tilde{x}_0 = x_0 \cdot \mathbb{I}(|x_0| < X)$ changes the process only with probability $\mathrm{P}(|x_0| > X)$. By taking X large enough, one can make this probability arbitrarily small.

Let $C > 0$; define the stopping time $\tau_C = \inf\{t : \sum_{i=0}^{t} \gamma_i > C\}$ and introduce the new process x_t^C, γ_t^C by

$$\begin{array}{ll} x_t^C = x_t, & \gamma_t^C = \gamma_t \text{ as } t < \tau_c, \text{ and} \\ x_t^C = x_{\tau_C}, & \gamma_t^C = 0 \text{ as } t \geq \tau_c. \end{array}$$

First, let us prove that the sequence $\{x_t^C\}$ is bounded. Designate $M_R := \sup_{|x|\geq R} \frac{\varphi(x)}{x}$; from A4 it follows that $M_R < \infty$. One has

$$|x_t^C| \leq |x_{t-1}^C - \gamma_{t-1}^C \varphi(x_{t-1}^C)| + \gamma_{t-1}^C |\xi_t|. \tag{12}$$

Using that $\gamma_{t-1}^C \leq C$ and $|\varphi(x_{t-1})^C| \leq |\varphi(0)| + M|x_{t-1}^C|$, one obtains

$$|x_t^C| \leq |x_{t-1}^C|(1 + CM) + \gamma_{t-1}^C(|\varphi(0)| + |\xi_t|). \tag{13}$$

If $\gamma_{t-1}^C \leq 2/M_R$, an even more precise estimate for x_t^C can be obtained. We shall distinguish between two cases: (i) $|x_{t-1}| \leq R$ and (ii) $|x_{t-1}^C| > R$.

In case (i), designating $\bar{b} := \sup_{|x|\leq R} |\varphi(x)|$, one has

$$|x_{t-1}^C - \gamma_{t-1}^C \varphi(x_{t-1}^C)| \leq |x_{t-1}^C| + \gamma_{t-1}^C \bar{b}. \tag{14}$$

In the case (ii) one has

$$0 \leq \gamma_{t-1}^C \frac{\varphi(x_{t-1}^C)}{x_{t-1}^C} \leq \frac{2}{M_R} M_R = 2,$$

hence

$$|x_{t-1}^C - \gamma_{t-1}^C \varphi(x_{t-1}^C)| \leq |x_{t-1}^C|. \tag{15}$$

Thus, in both cases (i) and (ii), from (12), (14), and (15) one gets

$$|x_t^C| \leq |x_{t-1}^C| + \gamma_{t-1}^C(\bar{b} + |\xi_t|). \tag{16}$$

The overall number of values of t such that $\gamma_{t-1}^C \leq 2/M_R$ is less than $CM_R/2$; therefore, using (13) and (16), one concludes that

$$|x_t^C| \leq \left(|x_0| + \sum_{i=1}^{t} \gamma_{i-1}^C(\bar{b} + |\varphi(0)| + |\xi_i|) \right) \cdot (1 + CM)^{CM_R/2}. \tag{17}$$

Denote $c_0 := \bar{b} + |\varphi(0)| + \mathrm{E}|\xi_1|$ and $\zeta_t := |\xi_t| - \mathrm{E}|\xi_t|$; using that $\sum_1^\infty \gamma_{i-1}^C \leq C$ one gets

$$|x_t^C| \leq \left(|x_0| + C\, c_0 + \sum_{i=1}^{t} \gamma_{i-1}^C \zeta_i \right) \cdot (1 + CM)^{CM_R/2}. \tag{18}$$

Using that $\sum_1^\infty \mathrm{E}(\gamma_{t-1}^C \zeta_t)^2 = \mathrm{E}\zeta_1^2 \cdot \sum_1^\infty \mathrm{E}(\gamma_{t-1}^C)^2 < \infty$, one obtains that the martingale $\sum_1^t \gamma_{i-1}^C \zeta_i$ is bounded; the value x_0 is also bounded, so, by (18), one concludes that the sequence $\{x_t^C\}$ is bounded.

Now, let us show that $\{x_t^C\}$ converges. From the definition of x_t^C and γ_t^C it follows that

$$x_t^C = x_0 - \sum_1^t \gamma_{i-1}^C \varphi(x_{i-1}^C) - \sum_1^t \gamma_{i-1}^C \xi_i.$$

Using that the sequence $\{\varphi(x_{i-1}^C)\}$ is bounded and that $\sum_1^\infty \gamma_{i-1}^C \le C$, one gets that the series $\sum_1^\infty \gamma_{i-1}^C \varphi(x_{i-1}^C)$ converges. Further, one has

$$\sum_1^\infty \mathrm{E}(\gamma_{t-1}^C \xi_t)^2 = S \cdot \sum_1^\infty \mathrm{E}(\gamma_{t-1}^C)^2 < \infty,$$

hence the martingale $\sum_1^t \gamma_{i-1}^C \xi_i$ converges. This implies that $\{x_t^C\}$ also converges.

Define the events $A_C = \{\sum_t \gamma_t \le C\}$ and $A_\infty = \{\sum_t \gamma_t < \infty\}$. One has $A_\infty = \cup_C A_C$. If $\sum_t \gamma_t \le C$ then $x_t^C = x_t$ for any t; this means that $\mathbb{I}(A_C) \cdot (x_t^C - x_t) = 0$ for any t and C. The sequence $\{\mathbb{I}(A_C) x_t^C\}$ converges, therefore the sequence $\{\mathbb{I}(A_C) x_t\}$ also converges, and passing to the limit $C \to \infty$ one obtains that $\{\mathbb{I}(A_\infty) x_t\}$ converges. This means exactly that if $\sum_t \gamma_t < \infty$ then $\{x_t\}$ converges. □

Lemma 2 *If $\lim_{t\to\infty} x_t = x$ then $x \in V_-^{[\mathrm{k}]}$.*

Proof. Note that, using A3 (a), it is easy to show that there exists $\delta_0 > 0$ such that $\mathrm{P}(\xi_1 \notin [x - L/2,\, x + L/2]) > \delta_0$, whatever $x \in \mathbb{R}$.

Next, for any $x \notin V_-^{([\mathrm{k}])}$ there exist $w(x) > 0$ and $0 < \epsilon(x) < L/4$ such that the following holds: for any two random variables ϕ_1 and ϕ_2 satisfying the relations $|\phi_l - \varphi(x)| \le \epsilon(x),\ l = 1, 2$ one has

$$\mathrm{P}((\phi_1 + \xi_1)(\phi_2 + \xi_2) > 0) > \frac{\ln(1/d) + w(x)}{\ln u + \ln(1/d)}.$$

Choose a countable set of intervals $U_i = (\varphi(x_i) - \epsilon(x_i),\ \varphi(x_i) + \epsilon(x_i))$ covering the set $\varphi(\mathbb{R} \setminus V_-^{[\mathrm{k}]})$, and denote $w_i := w(x_i)$. Fix i and $s \in \{0, 1, 2, \ldots\}$, and define the auxiliary process $x_t^{(is)}$, $\gamma_t^{(is)}$ by formulas:

if $t < s$ then $x_t^{(is)} = x_t$, and if $t \ge s$ then

$$x_t^{(is)} = \begin{cases} x_{t-1}^{(is)} - \gamma_{t-1}^{(is)}\, y_t^{(is)} & \text{if } \ \varphi(x_{t-1}^{(is)} - \gamma_{t-1}^{(is)}\, y_t^{(is)}) \in U_i, \\ x_i & \text{elsewhere;} \end{cases} \tag{19}$$

$$y_t^{(is)} = \varphi(x_{t-1}^{(is)}) + \xi_t, \tag{20}$$

$$\gamma_t^{(is)} = \begin{cases} \min\{u\gamma_{t-1}^{(is)},\, \bar{\mathrm{g}}\} & \text{if } \ y_{t-1}^{(is)}\, y_t^{(is)} > 0, \\ d\gamma_{t-1}^{(is)} & \text{if } \ y_{t-1}^{(is)}\, y_t^{(is)} \le 0. \end{cases} \tag{21}$$

So, as $t \geq s$, $\varphi(x_t^{(is)})$ is forced to be contained in U_i.

For $t \geq s+2$, using that $y_{t-1}^{(is)} = \varphi(x_{t-2}^{(is)}) + \xi_{t-1}$, $y_t^{(is)} = \varphi(x_{t-1}^{(is)}) + \xi_t$, $\varphi(x_{t-2}^{(is)}) \in U_i$, one obtains that

$$\mathrm{P}(y_{t-1}^{(is)}\, y_t^{(is)} > 0) > \frac{\ln(1/d) + w_i}{\ln u + \ln(1/d)}$$

and

$$\mathrm{P}(y_{t-1}^{(is)}\, y_t^{(is)} \leq 0) < \frac{\ln u - w_i}{\ln u + \ln(1/d)},$$

hence

$$\mathrm{E}[\ln u \cdot\ \mathbb{I}(y_{t-1}^{(is)}\, y_t^{(is)} > 0)\ +\ \ln d \cdot\ \mathbb{I}(y_{t-1}^{(is)}\, y_t^{(is)} \leq 0)] >$$
$$> \ln u \cdot \frac{\ln(1/d) + w_i}{\ln u + \ln(1/d)}\ +\ \ln d \cdot \frac{\ln u - w_i}{\ln u + \ln(1/d)} = w_i.$$

Consider variables $\phi_1 = f_1(\xi_1, \xi_2)$ and $\phi_2 = f_2(\xi_1, \xi_2)$ providing a solution of the (deterministic) minimization problem:

$$(\phi_1 + \xi_1)(\phi_2 + \xi_2) \to \min,$$

subject to

$$|\phi_1 - \varphi(x_i)| \leq \epsilon(x_i)$$
$$|\phi_2 - \varphi(x_i)| \leq \epsilon(x_i),$$

and denote $Y_{t-1}^1 = f_1(\xi_{t-1}, \xi_t) + \xi_{t-1}$, $Y_t^2 = f_2(\xi_{t-1}, \xi_t) + \xi_t$, $\eta_t = \ln u \cdot \mathbb{I}(Y_{t-1}^1 Y_{t-1}^2 > 0) + \ln d \cdot\ \mathbb{I}(Y_{t-1}^1 Y_{t-1}^2 \leq 0)$. One has

(i) $\eta_t \leq \ln u \cdot\ \mathbb{I}(y_{t-1}^{(is)}\, y_t^{(is)} > 0) + \ln d \cdot\ \mathbb{I}(y_{t-1}^{(is)}\, y_t^{(is)} \leq 0)$;

(ii) η_t are identically distributed, and $\mathrm{E}\eta_t \geq w_i$;

(iii) the set of random variables $\{\eta_t,\ t \text{ even},\ t \geq s+2\}$ as well as the set $\{\eta_t,\ t \text{ odd},\ t \geq s+2\}$, are mutually independent.

From (ii)–(iii) it follows that almost surely $\sum_t \eta_t = +\infty$, and from (i) it follows that

$$\sum_t [\ln u \cdot\ \mathbb{I}(y_{t-1}^{(is)}\, y_t^{(is)} > 0) + \ln d \cdot\ \mathbb{I}(y_{t-1}^{(is)}\, y_t^{(is)} \leq 0)] = +\infty,$$

so, by virtue of (21), $\gamma^{(is)}$ does not go to zero.

Thus, there exists a random value $\chi > 0$ such that for infinitely many values of t, $\gamma_t^{(is)} \geq \chi$.

Define a sequence of stopping times τ_0, τ_1, $\tau_2, \ldots$ inductively, letting $\tau_0 = 0$ and $\tau_j = \inf\{t > \tau_{j-1} : \gamma_t^{(is)} \geq \chi\}$ for $j \geq 1$. The events $B_j = \{|\xi_{\tau_j+1} + \varphi(x_i)| > L/2\}$ happen with probability more that δ_0 (recall the remark done in the beginning of proof), and every event B_j, $j \geq 2$ does not depend on the set of events $\{B_1, \ldots, B_{j-1}\}$. Therefore, for infinitely many values of j, B_j, takes place, i.e., $|\xi_{\tau_j+1} + \varphi(x_i)| > L/2$, and hence, taking into account that

$|y_{\tau_j+1}| \ge |\xi_{\tau_j+1} + \varphi(x_i)| - |\varphi(x_{\tau_j}) - \varphi(x_i)|$ and $|\varphi(x_{\tau_j}) - \varphi(x_i)| < \epsilon(x_i) < L/4$, for these values of j one has $|y_{\tau_j+1}| \ge L/4$. Thus, one concludes that

$$\text{for infinitely many values of } j, \quad |\gamma_{\tau_j} y_{\tau_j+1}| \ge \chi\, L/4. \tag{22}$$

Suppose that x_t converges to a point from $\mathbb{R} \setminus V_-^{[\mathrm{k}]}$, then for some i and s one has $x_t \in U_i$ as $t \ge s$, hence the process $x_t^{(is)}$, $\gamma_t^{(is)}$ coincides with x_t, γ_t, and therefore $\gamma_t\, y_{t+1} \to 0$ as $t \to \infty$. The last relation contradicts (22), thus Lemma 2 is proved. □

Lemma 3 *Let $\sum_t \gamma_t = \infty$. Then for any open set $\mathcal{O}$ containing* Z *there exists a positive constant $g = g(\mathcal{O})$ such that either (i) for some t, $x_t \in \mathcal{O}$, or (ii) for some t, $|x_t| < R$ and $\gamma_t > g$.*

Proof. Designate by f the primitive of φ such that $\inf_x f(x) = 0$. Define the stopping time

$$\tau = \tau(\mathcal{O}, g) := \inf\{t : \text{ either (i) } x_t \in \mathcal{O}, \text{ or (ii) } |x_t| < R \text{ and } \gamma_t \ge g\}.$$

The value of $g \in (0, \bar{\mathrm{g}})$ will be specified below.

Consider the sequence $\mathrm{E}_t = \mathrm{E}[f(x_t)\, \mathbb{I}(t < \tau)]$. Introducing shorthand notation $f(x_t) =: f_t$, $\mathbb{I}(t < \tau) =: I_t$, $f'(x_t) =: f'_t = \varphi_t$, and using that $I_t \le I_{t-1}$, one gets

$$E_t - E_{t-1} = \mathrm{E}[f_t\, \mathbb{I}_t - f_{t-1}\, \mathbb{I}_{t-1}] \le \mathrm{E}[(f_t - f_{t-1})\, \mathbb{I}_{t-1}]. \tag{23}$$

Next, we utilize the Taylor decomposition

$$f_t = f(x_{t-1} - \gamma_{t-1} y_t) = f_{t-1} - f'_{t-1}\, \gamma_{t-1} y_t + \frac{1}{2} f''(x')\, \gamma_{t-1}^2 y_t^2,$$

x' being some point between x_{t-1} and x_t. Substituting $y_t = \varphi_{t-1} + \xi_t$ and recalling that $f'_{t-1} = \varphi_{t-1}$ and $f''(x') = \varphi'(x') \le M$, one obtains

$$f_t - f_{t-1} \le -\gamma_{t-1}\, \varphi_{t-1}(\varphi_{t-1} + \xi_t) + \frac{M}{2}\, \gamma_{t-1}^2\, (\varphi_{t-1} + \xi_t)^2. \tag{24}$$

Using (23) and (24) and taking into account that each of the values γ_{t-1}, φ_{t-1}, $\mathbb{I}_{t-1}$ is mutually independent with ξ_t (see A1), one gets

$$\begin{aligned} &E_t - E_{t-1} \le \mathrm{E}[(-\gamma_{t-1}\, \varphi_{t-1}^2 - \gamma_{t-1}\, \varphi_{t-1}\xi_t + \tfrac{M}{2}\gamma_{t-1}^2\, \varphi_{t-1}^2 + M\gamma_{t-1}^2\, \varphi_{t-1}\xi_t + \tfrac{M}{2}\gamma_{t-1}^2\, \xi_t^2)\, \mathbb{I}_{t-1}] = \\ &= \mathrm{E}[(-\varphi_{t-1}^2 + \tfrac{M}{2}\gamma_{t-1}\, \varphi_{t-1}^2 + \tfrac{M}{2}\gamma_{t-1} S)\gamma_{t-1}\, \mathbb{I}_{t-1}] = \\ &= \mathrm{E}[(-\varphi_{t-1}^2(1 - M\gamma_{t-1}/2) + M\gamma_{t-1} S/2)\gamma_{t-1}\, \mathbb{I}_{t-1}]. \end{aligned} \tag{25}$$

If $\mathbb{I}_{t-1} = 1$ then either (i) $x_{t-1} \in [-R, R] \backslash \mathcal{O}$ and $\gamma_{t-1} < g$, or (ii) $|x_{t-1}| \ge R$.

In the case (i) one has

$$-\varphi_{t-1}^2(1 - M\gamma_{t-1}/2) + M\gamma_{t-1} S/2 \le -c_0(1 - Mg/2) + MgS/2 =: -c'_g, \tag{26}$$

where $c_0 := \inf\{|\varphi(x)| : x \in [-R, R] \setminus \mathcal{O}\}$; obviously, $c_0 > 0$. Let us fix a $g \in (0, \bar{\mathrm{g}})$ such that $c'_g > 0$.

In the case (ii), designating $b_0 := \inf_{|x| \geq R} \varphi^2(x)$, one has

$$-\varphi_{t-1}^2(1 - M\gamma_{t-1}/2) + M\gamma_{t-1}S/2 \leq -b_0(1 - M\bar{\mathrm{g}}/2) + M\bar{\mathrm{g}}S/2 =: -c''. \quad (27)$$

Using A6, one gets that $c'' > 0$.

Denote $c = \min\{c'_g, c''\}$. The relations (26) and (27) imply that if $\mathbb{I}_{t-1} = 1$ then $-\varphi_{t-1}^2(1 - M\gamma_{t-1}/2) + M\gamma_{t-1}S/2 \leq -c < 0$, hence, by virtue of (25),

$$E_t - E_{t-1} \leq -c \cdot \mathrm{E}[\gamma_{t-1}\, \mathbb{I}_{t-1}]. \quad (28)$$

Summing up both sides of (28) over $t = 1, \ldots, s$ and denoting $\mathbb{I}_\infty = \mathbb{I}(\tau = \infty) = \min_t \mathbb{I}_t$, one obtains

$$\mathrm{E}_s - \mathrm{E}_0 \leq -c \cdot \mathrm{E}\left[\sum_{i=0}^{s-1} \gamma_i \cdot \mathbb{I}_\infty\right].$$

One has $\mathrm{E}_s \geq 0$, and x_0 is bounded, hence $E_0 < \infty$. Thus, for arbitrary s

$$\mathrm{E}\left[\sum_{i=0}^{s-1} \gamma_i \cdot \mathbb{I}_\infty\right] \leq \frac{\mathrm{E}_0}{c} < \infty.$$

This implies that a.s. either $\sum_0^\infty \gamma_i < \infty$, or $\tau = \infty$. Lemma 3 is proved. □

Denote $c_1 := 1 - M\bar{\mathrm{g}}/2$. Recall that f is the primitive of φ such that $\inf_x f(x) = 0$; the assumption A6 implies that $\lim_{x\to\pm\infty} f(x) = +\infty$. Denote $H := \sup_{|x|\leq R} f(x)$. Denote also $c_3 := \bar{\mathrm{g}} \cdot \sup\{|\varphi(x)| : f(x) \leq H\} + 1$, $z^l := \inf\{x : f(x) \leq H\} - c_3$, $z^r := \sup\{x : f(x) \leq H\} + c_3$, $c_2 := \inf\{|\varphi(x)| : x \in [z^l, z^r] \setminus \mathcal{O}\}$, and $\mathrm{K} := \sup\{|\varphi(x)| : x \in [z^l, z^r]\}$. Obviously, $c_1 > 0$ and $\mathrm{K} \geq c_2 > 0$.

Fix an open set $\mathcal{O}$ containing Z. Let $g > 0$, $0 < w < 1$. We shall say that a (finite or infinite) deterministic sequence $\{z_0, z_1, z_2, \ldots\}$ is (g, w)-admissible if $|z_0| \leq R$ and there exist deterministic sequences $\{q_t\}$, $\{h_t\}$ such that

1) $|h_t| \leq w$;

2) if $\{z_0, z_1, \ldots, z_t\} \subset [z^l, z^r] \setminus \mathcal{O}$ then $gd^2 \leq q_s \leq \bar{\mathrm{g}}$, $s = 0, 1, \ldots, t$;

3) $z_t = z_{t-1} - q_{t-1}\,\varphi(z_{t-1}) - h_t$, $t = 1, 2, \ldots$.

Proposition 1 *There exists constants t_0 and w such that any (g, w)-admissible sequence $\{z_t,\ t = 0, 1, \ldots, t_0\}$ has non-empty intersection with $\mathcal{O}$.*

Proof. Let $w := \min\{1, gd^2c_2^2c_1/(2\mathrm{K})\}$. Designate $\tilde{t} = \inf\{t : z_t \in \mathcal{O}\}$; $\tilde{t}$ takes values from $\{0, 1, \ldots, t_0, +\infty\}$. We shall use shorthand notation $f_t := f(z_t)$, $f'_t = \varphi_t := \varphi(z_t)$. One has

$$f_t = f(z_{t-1} - q_{t-1}\varphi_{t-1} - h_t) = f(z_{t-1} - q_{t-1}\varphi_{t-1}) - f'(\tilde{z}).h_t, \quad (29)$$

where $\tilde{z}$ is a point between $z_{t-1} - q_{t-1}\varphi_{t-1}$ and $z_{t-1} - q_{t-1}\varphi_{t-1} - h_t$.

Next, one has

$$f(z_{t-1} - q_{t-1}\varphi_{t-1}) = f_{t-1} - f'_{t-1}q_{t-1}\varphi_{t-1} + \frac{1}{2}f''(\hat{z})\, q_{t-1}^2\varphi_{t-1}^2, \qquad (30)$$

where $\hat{z}$ is a point between z_{t-1} and $z_{t-1} - q_{t-1}\varphi_{t-1}$.

We are going to prove by induction that

$$\text{if } 0 \le s \le \tilde{t} \ \text{ then } \ f_s \le H - s \cdot gd^2c_2^2c_1/2. \qquad (31)$$

For $s = 0$, (31) follows from the condition $|z_0| \le R$ and the definition of H. Now, let $1 \le t \le \tilde{t}$; suppose that formula (31) is true for $0 \le s \le t-1$ and prove it for $s = t$. For $0 \le s \le t-1$, one has $f(z_s) \le H$, $z_s \notin \mathcal{O}$, therefore $z_s \in [z^l, z^r] \setminus \mathcal{O}$; hence, by virtue of 2), $gd^2 \le q_s \le \bar{\text{g}}$ for $0 \le s \le t-1$. One has $f(z_{t-1}) \le H$, $|q_{t-1}\varphi_{t-1}| \le \bar{\text{g}} \cdot \sup\{|\varphi(x)| : f(x) \le H\}$, and $|h_t| \le w \le 1$, hence $|q_{t-1}\varphi_{t-1}| \le c_3$, $|q_{t-1}\varphi_{t-1} + h_t| \le c_3$, and so, $z_{t-1} - q_{t-1}\varphi_{t-1} \in [z^l, z^r]$, $z_{t-1} - q_{t-1}\varphi_{t-1} - h_t \in [z^l, z^r]$, thus $\tilde{z}$ also belongs to $[z^l, z^r]$. This implies that $|\varphi(\tilde{z})| = |f'(\tilde{z})| \le \text{K}$. Then, combining (29) and (30) and using that $|h_t| \le w$ and $|f''(\hat{z})| = |\varphi'(\hat{z})| \le M$, one obtains

$$f_t \le f_{t-1} - q_{t-1}\varphi_{t-1}^2(1 - \frac{1}{2}q_{t-1}M) + w\text{K}. \qquad (32)$$

One has $z_{t-1} \in [z^l, z^r] \setminus \mathcal{O}$, hence $|\varphi(z_{t-1})| = |\varphi_{t-1}| \ge c_2$. Using also that $q_{t-1} \ge gd^2$, $1 - \frac{1}{2}q_{t-1}M \ge c_1$, and $w\text{K} \le gd^2c_2^2c_1/2$, one gets from (32) that

$$f_t \le f_{t-1} - gd^2c_2^2c_1/2,$$

and using the induction hypothesis, one concludes that

$$f_t \le H - t \cdot gd^2c_2^2c_1/2.$$

Formula (31) is proved.

Let $t_0 := \lfloor 2H/(gd^2c_2^2c_1) \rfloor + 1$; here $\lfloor z \rfloor$ stands for the integral part of z. Then, taking into account that $f_s \ge 0$, from (31) one concludes that $\tilde{t} < t_0$, thus Proposition 1 is proved. □.

Proposition 2 *If $\gamma_{t-1} < 1/(3M)$, $|\xi_t| < c_2$, $|\xi_{t+1}| < c_2$, x_{t-1} and x_t belong to $[z^l, z^r] \setminus \mathcal{O}$, then $\gamma_{t+1} \ge \gamma_t$.*

Proof. Using notation $\varphi_t := \varphi(x_t)$, one gets

$$\varphi_t = \varphi(x_{t-1} - \gamma_{t-1}(\varphi_{t-1} + \xi_t)) = \varphi_{t-1} - \varphi'(\tilde{x}) \cdot \gamma_{t-1}(\varphi_{t-1} + \xi_t),$$

where $\tilde{x}$ is a point between x_{t-1} and x_t. Therefore,

$$\varphi_{t-1}\varphi_t = \varphi_{t-1}^2 \cdot [1 - \varphi'(\tilde{x})\gamma_{t-1} \cdot (1 + \xi_t/\varphi_{t-1})].$$

Using that $|\varphi'(\tilde{x})| \le M$, $\gamma_{t-1} < 1/(3M)$, $|\xi_t| < c_2$, $|\varphi_{t-1}| \ge c_2$, one obtains $1 - \varphi'(\tilde{x})\,\gamma_{t-1} \cdot (1 + \xi_t/\varphi_{t-1}) \ge 1/3$, hence $\varphi_{t-1}\varphi_t > 0$. Further, using that $|\xi_t| < c_2$, $|\xi_{t+1}| < c_2$, $|\varphi_{t-1}| \ge c_2$, $|\varphi_t| \ge c_2$, one gets

$$y_t\, y_{t+1} = \varphi_{t-1}\varphi_t \cdot (1 + \xi_t/\varphi_{t-1})(1 + \xi_{t+1}/\varphi_t) > 0.$$

This implies that $\gamma_{t+1} = \min\{u\gamma_t, \bar{\text{g}}\} \ge \gamma_t$. □

Lemma 4 *For any open set $\mathcal{O}$, containing* Z*, and any $g > 0$ there exists $\delta = \delta(\mathcal{O}, g) > 0$ such that*

$$\textit{if} \ \ |x_0| \le R, \ \gamma_0 \ge g \ \ \textit{then} \quad \mathrm{P}(\textit{for some } t, \ x_t \in \mathcal{O}) \ge \delta.$$

Proof. Without loss of generality suppose that $g < 1/(3M)$. Define the event

$$A := \{|\xi_i| < \min\{c_2, \, w/\bar{\mathrm{g}}\}, \ i = 1, 2, \dots, t_0\},$$

where w and t_0 are the same as in the proof of Proposition 1: $w = \min\{1, \, gd^2c_2^2c_1/(2\mathrm{K})\}$, $t_0 = \lfloor 2H/(gd^2c_2^2c_1) \rfloor + 1$.

Denote

$$\delta := P(A) = (\mathrm{P}(|\xi_1| < \min\{c_2, \, w/\bar{\mathrm{g}}\}))^{t_0};$$

by virtue of A3 (a), $\delta > 0$. Let us show that for any elementary event $\omega \in A$, the sequence $\{z_t = x_t(\omega), \ t = 0, 1, \dots, t_0\}$ is $(g, \, w)$-admissible.

One has $|z_0| = |x_0(\omega)| < R$. Further, one has $z_t = z_{t-1} - q_{t-1}\varphi(z_{t-1}) - h_t$, with $q_{t-1} = \gamma_{t-1}(\omega)$, $h_t = \gamma_{t-1}(\omega)\, \xi_t(\omega)$, and using that $\gamma_{t-1}(\omega) \le \bar{\mathrm{g}}$ and $|\xi_t(\omega)| < \omega/\bar{\mathrm{g}}$, one gets $|h_t| \le w$. Thus, conditions 1) and 3) are verified.

Now, let $\{z_0, z_1, \dots, z_t\} \subset [z^l, \, z^r] \setminus \mathcal{O}$, $t \le t_0$. Let $s_0 \in \{0, 1, 2, \dots, t\}$ be the minimal value such that $q_{s_0} = \min\{q_0, q_1, \dots, q_t\}$. If $s_0 = 0$ then $\min\{q_0, q_1, \dots, q_t\} = q_0 = \gamma_0(\omega) \ge g \ge gd^2$. If $s_0 = 1$ then $\min\{q_0, q_1, \dots, q_t\} = q_1 = \gamma_1(\omega) \ge gd \ge gd^2$. If $s_0 \ge 2$ then $\gamma_{s_0-2}(\omega) \ge 1/(3M)$; otherwise, using that $|\xi_{s_0-1}| < c_2$, $|\xi_{s_0}| < c_2$, $x_{s_0-2}(\omega)$ and $x_{s_0-1}(\omega)$ belong to $[z^l, \, z^r] \setminus \mathcal{O}$, and applying Proposition 2, one would conclude that $\gamma_{s_0}(\omega) \ge \gamma_{s_0-1}(\omega)$, which contradicts the definition of s_0.

Thus, $\gamma_{s_0}(\omega) \ge 1/(3M) \cdot d^2 \ge gd^2$, and therefore, $\min\{q_0, q_1, \dots, q_t\} = \gamma_{s_0}(\omega) \ge gd^2$. So, the condition 2) is also verified.

Now, applying Proposition 1 to the $(g, \, w)$-admissible sequence $\{z_t\}$, one concludes that there exists a non-negative $\tau \le t_0$ such that $z_\tau = x_\tau(\omega) \in \mathcal{O}$. This implies that

$$\mathrm{P}(\text{for some } t, \ x_t \in \mathcal{O}) \ge \mathrm{P}(A) = \delta.$$

□

Lemma 5 *If $\sum_t \gamma_t = \infty$ then for any open set $\mathcal{O}$ containing* Z *there exists t such that $x_t \in \mathcal{O}$.*

Proof. Let us fix an open set $\mathcal{O} \supset \mathrm{Z}$, and denote $\delta = \delta(\mathcal{O}, g(\mathcal{O}))$. Combining Lemma 3 and Lemma 4, one concludes that for any $\mathcal{O} \supset \mathrm{Z}$ there exists $\delta > 0$ such that whatever the initial conditions $x_0, \, \gamma_0, \, \gamma_1$,

$$\mathrm{P}(\text{for some } t, \ x_t \in \mathcal{O} \,\Big|\, \sum_t \gamma_t = \infty) > \delta.$$

Then one can choose a measurable integer-valued function $n(\cdot, \cdot, \cdot)$ defined on $\mathbb{R} \times (0, \bar{\mathrm{g}}] \times (0, \bar{\mathrm{g}}]$ such that for $\nu = n(x_0, \gamma_0, \gamma_1)$ one will have

$$\mathrm{P}(\text{for some } t \le \nu, \ x_t \in \mathcal{O} \,\Big|\, \sum_t \gamma_t = \infty) > \delta/2.$$

Designate
$$\bar{p} = \sup \mathrm{P}(\text{for all } t,\ x_t \notin \mathcal{O} \Big| \sum_t \gamma_t = \infty),$$
the supremum being taken over all the initial conditions x_0, γ_0, γ_1. Fix x_0, γ_0, γ_1, then

$$\begin{aligned} &\mathrm{P}(\text{for all } t,\ x_t \notin \mathcal{O} \Big| \textstyle\sum_t \gamma_t = \infty) = \\ &= \mathrm{P}(\text{for all } t > \nu,\ x_t \notin \mathcal{O} \Big| \text{for all } t \le \nu,\ x_t \notin \mathcal{O} \text{ and } \textstyle\sum_t \gamma_t = \infty)\cdot \\ &\cdot \mathrm{P}(\text{for all } t \le \nu,\ x_t \notin \mathcal{O} \,|\, \textstyle\sum_t \gamma_t = \infty) \le \bar{p}\,(1 - \delta/2). \end{aligned} \tag{33}$$

Taking supremum of the left hand side of (33) over all $(x_0, \gamma_0, \gamma_1) \in \mathbb{R} \times (0, \bar{\mathrm{g}}] \times (0, \bar{\mathrm{g}}]$, one obtains $\bar{p} \le \bar{p}\,(1 - \delta/2)$, hence $\bar{p} = 0$. Lemma 5 is proved. □.

Denote $\mathcal{O}_* = \{x : |\varphi(x)| < L/2\}$.

Lemma 6 *For any open bounded sets $\mathcal{O}$, $\mathcal{O}_1$ such that $\bar{\mathcal{O}} \subset \mathcal{O}_1 \subset \mathcal{O}_*$ and for any $w > 0$ there exists $\delta = \delta(\mathcal{O}, \mathcal{O}_1, w) > 0$ such that*

$$\textit{if } \ x_0 \in \mathcal{O} \ \textit{then } \ \mathrm{P}(\textit{for some } n,\ x_n \in \mathcal{O}_1 \ \textit{and } \gamma_n < w) \ge \delta.$$

Proof. Denote $n = \lfloor \frac{\ln \bar{\mathrm{g}} - \ln w}{\ln(1/d)} \rfloor + 2$, where $\lfloor \cdot \rfloor$ means the integer part of a number. Denote also
$$\varepsilon = \min\left\{ \frac{L}{2},\ \frac{\mathrm{dist}(\mathcal{O},\, \mathbb{R} \setminus \mathcal{O}_1)}{n\bar{\mathrm{g}}} \right\},$$
where $\mathrm{dist}(A, B) := \inf_{x \in A, y \in B} |x - y|$ for arbitrary sets of real numbers A, B (in particular, $\mathrm{dist}(x, B) := \inf_{y \in B} |x - y|$)). Using assumption A3 (a), one obtains that there exists $\delta_1 > 0$ such that for any $x \in \mathcal{O}_1$ and for any integer t,
$$\mathrm{P}\left((-1)^{t-1}\varphi(x) < (-1)^t \xi_1 < (-1)^{t-1}\varphi(x) + \varepsilon\right) \ge \delta_1.$$
This implies that if $x_0 \in \mathcal{O}$ then
$$\mathrm{P}(0 < (-1)^t y_t < \varepsilon,\ \mathrm{dist}(x_{t-1},\, \mathcal{O}) < (t-1)\bar{\mathrm{g}}\varepsilon,\ t = 1, 2, \ldots, n+1) \ge \delta_1^{n+1}.$$
Denoting $\delta = \delta_1^{n+1}$, one concludes that the following statements (i) and (ii) hold with probability at least δ:

(i) $\mathrm{dist}(x_n,\, \mathcal{O}) < n\bar{\mathrm{g}}\varepsilon \le \mathrm{dist}(\mathcal{O},\, \mathbb{R} \setminus \mathcal{O}_1)$, hence $x_n \in \mathcal{O}_1$;

(ii) as $t = 2, 3, \ldots, n+1$, one has $y_{t-1} y_t < 0$, hence $\gamma_t = d\gamma_{t-1}$, therefore $\gamma_n = d^{n-1}\gamma_1 \le d^{n-1}\bar{\mathrm{g}} < w$.

Lemma 6 is proved. □

Lemma 7 *If $\sum_t \gamma_t = \infty$, $\mathcal{O}$ is an open set containing* Z*, and $w > 0$ then for some t, $x_t \in \mathcal{O}$ and $\gamma_t < w$.*

Proof. Without loss of generality, suppose that $\mathcal{O}$ is bounded and $\mathcal{O} \subset \mathcal{O}_*$. Choose an open set $\mathcal{O}_1$ such that $\mathrm{Z} \subset \mathcal{O}_1,\ \bar{\mathcal{O}}_1 \subset \mathcal{O}$; applying Lemmas 5 and 6, one gets that for $\delta = \delta(\mathcal{O}_1, \mathcal{O}, w)$ and for arbitrary initial conditions,

$$\mathrm{P}(\text{for some } t,\ x_t \in \mathcal{O} \text{ and } \gamma_t < w) > \delta.$$

Repeating the argument of Lemma 5, one concludes that there exists t such that $x_t \in \mathcal{O}$ and $\gamma_t < w$. □

From now on we suppose that $\mathrm{k} > k_+(0)$. Choose k' such that $k_+(0) < k' < \mathrm{k}$; using A3 (b), one obtains that for some $\varepsilon_0 > 0$, $\mathrm{P}(\xi_1\xi_2 > 0,\text{ or } |\xi_1| < \varepsilon_0,\text{ or } |\xi_2| < \varepsilon_0) \le k'$. Denote $\mathcal{O}_0 = \{x : |\varphi(x)| < \varepsilon_0\}$ and $\tau = \inf\{t : x_t \notin \mathcal{O}_0\}$. Without loss of generality, suppose that $\mathcal{O}_0$ is bounded.

Lemma 8 *Suppose that* $\mathrm{k} > k_+(0)$, *then there exist a constant* $b > 0$ *and a monotone decreasing function* $p(\cdot)$ *such that* $\lim_{a\to+\infty} p(a) = 0$ *and*

$$\textit{if}\ \ \gamma_0 < w\ \textit{then}\ \ \mathrm{P}(\ln\gamma_t < \ln v - bt \textit{ for all } t < \tau) > 1 - p(v/w).$$

Proof. Define the sequences $\{\rho_t\}$ and $\{\sigma_t\}$ by

$$\begin{aligned}\rho_t &= \ln u \cdot \mathbb{I}(\xi_{t-1}\xi_t > 0,\ \text{or } |\xi_{t-1}| < \varepsilon_0,\ \text{or } |\xi_t| < \varepsilon_0) + \\ &+ \ln d \cdot \mathbb{I}(\xi_{t-1}\xi_t \le 0\ \&\ |\xi_{t-1}| \ge \varepsilon_0\ \&\ |\xi_t| \ge \varepsilon_0),\end{aligned}$$

$$\sigma_t = \ln w + \sum_{i=1}^{t} \rho_i.$$

Using (7) and definition of τ, one obtains that for all $t < \tau$, $\gamma_t \le \sigma_t$. The variables ρ_t are identically distributed, take the values $\ln u$ and $\ln d$, and

$$\begin{aligned}E\rho_t &= \ln u \cdot \mathrm{P}(\xi_{t-1}\xi_t > 0,\ \text{or } |\xi_{t-1}| < \varepsilon_0,\ \text{or } |\xi_t| < \varepsilon_0) + \\ &+ \ln d \cdot \mathrm{P}(\xi_{t-1}\xi_t \le 0\ \&\ |\xi_{t-1}| \ge \varepsilon_0\ \&\ |\xi_t| \ge \varepsilon_0) \le \\ &\le \ln u \cdot k' + \ln d \cdot (1 - k') < \ln u \cdot \mathrm{k} + \ln d \cdot (1 - \mathrm{k}) = 0.\end{aligned}$$

Moreover, the variables in the set $\{\rho_t,\ t \text{ even}\}$, as well as the variables in the set $\{\rho_t,\ t \text{ odd}\}$, are independent.

Denote $b = -E\rho_t/2$. One has

$$\mathrm{P}(\ln\gamma_t < \ln v - bt\ \text{ for all } t < \tau) \ge \mathrm{P}(\sigma_t < \ln v - bt\ \text{ for all } t) =$$

$$= \mathrm{P}(\sum_{i=1}^{t}(\rho_i + 2b) < \ln v - \ln w + bt\ \text{ for all } t) \ge 1 - p(v/w),$$

where $p(a) = p_1(a) + p_2(a)$,

$$p_1(a) = \mathrm{P}\left(\sum_{1\le i\le t}{}'(\rho_i + 2b) \ge \frac{\ln a}{2} + \frac{b}{2}t\ \text{ for all } t\right),$$

$$p_2(a) = \mathrm{P}\left(\sum_{1\le i\le t}{}'' (\rho_i + 2b) \ge \frac{\ln a}{2} + \frac{b}{2}\, t \ \text{ for all } t\right);$$

the sum $\sum'$ ($\sum''$) is taken over the even (odd) values of i. Both $\sum'$ and $\sum''$ are sums of i.i.d.r.v. with zero mean, hence both $p_1(a)$ and $p_2(a)$ tend to zero as $a \to +\infty$. Lemma 8 is proved. □

Define the stopping times $\tau_v = \inf\{t : \ x_t \notin \mathcal{O}_0 \text{ or } \ln \gamma_t \ge \ln v - bt\}$. Recall that f is the primitive of φ such that $\inf_x f(x) = 0$. Fix an open set $\mathcal{O}'$ such that $\mathrm{Z} \subset \mathcal{O}' \subset \mathcal{O}_0$ and $\sup_{x\in\mathcal{O}'} f(x) < \inf_{x\notin\mathcal{O}_0} f(x)$, and denote $\delta = \inf_{x\notin\mathcal{O}_0} f(x) - \sup_{x\in\mathcal{O}'} f(x)$.

Lemma 9 *Let* $\mathrm{k} > k_+(0)$*,* $x_0 \in \mathcal{O}'$*, and* $\gamma_0 < w$*, then*

$$\mathrm{P}(\tau_v < \infty) \le K\, v^2 + p(v/w);$$

here K *is a positive constant, and* $p(\cdot)$ *satisfies the statement of lemma 8.*

Proof. We shall use shorthand notation of Lemma 3: $f_t := f(x_t)$ and $\varphi_t := \varphi(x_t)$. According to (24), one has

$$f_t - f_{t-1} \le -\gamma_{t-1}\varphi_{t-1}(\varphi_{t-1} + \xi_t) + \frac{M}{2}\,\gamma_{t-1}^2(\varphi_{t-1} + \xi_t)^2 \le$$

$$\le -\gamma_{t-1}\varphi_{t-1}\xi_t + M\gamma_{t-1}^2(\varphi_{t-1}^2 + \xi_t^2).$$

This implies that $f_t - f_1 \le Q_t' + Q_t''$, with

$$Q_t' = \Big|\sum_{i=2}^{t} \gamma_{i-1}\varphi_{i-1}\xi_i\Big|, \qquad Q_t'' = M\sum_{i=2}^{t} \gamma_{i-1}^2(\varphi_{i-1}^2 + \xi_i^2).$$

Using Lemma 8, one gets

$$\mathrm{P}(\tau_v < \infty) \ \le\ p(v/w) + P' + P'',$$

where

$$P' = \mathrm{P}(Q_{\tau_v}' \ge \delta/2) \quad \text{and} \quad P'' = \mathrm{P}(Q_{\tau_v}'' \ge \delta/2).$$

According to the Chebyshev inequality,

$$P' \ \le\ \frac{4}{\delta^2}\, EQ_{\tau_v}'^2 = \frac{4}{\delta^2}\sum_{i,j=1}^{\infty} E_{ij},$$

where

$$E_{ij} \ =\ E\left[\gamma_{i-1}\varphi_{i-1}\xi_i\ \mathbb{I}(i-1<\tau_v)\cdot\gamma_{j-1}\varphi_{j-1}\xi_j\ \mathbb{I}(j-1<\tau_v)\right].$$

Using that the values $\gamma_i,\ \varphi_i,\ \xi_i$, and $\mathbb{I}(i<\tau_v)$ are $\mathcal{F}_i$-measurable, and using assumptions A1 and A2, one obtains that for $i \ne j,\ E_{ij} = 0$, and for $i = j$,

$$E_{ii} = E\left[\gamma_{i-1}^2\varphi_{i-1}^2\ \mathbb{I}(i-1<\tau_v)\cdot\xi_i^2\right] \le v^2 e^{-2bi} \sup_{x\in\mathcal{O}_0}\varphi^2(x)\cdot S.$$

Therefore,

$$P' \leq \frac{4}{\delta^2} \sum_{i=2}^{\infty} E_{ii} \leq \frac{4v^2 S}{\delta^2} \; \frac{e^{-4b}}{1-e^{-2b}} \; \sup_{x \in \mathcal{O}_0} \varphi^2(x).$$

Similarly,

$$P'' \leq \frac{2}{\delta} EQ''_{\tau_v} = \frac{2M}{\delta} \sum_{i=2}^{\infty} E\left[\gamma_{i-1}^2 (\varphi_{i-1}^2 + \xi_i^2)\, \mathbb{I}(i-1 < \tau_v)\right] \leq$$

$$\leq \frac{2Mv^2}{\delta} \sum_{i=2}^{\infty} e^{-2bi} \left(\sup_{x \in \mathcal{O}_0} \varphi^2(x) + S \right) = \frac{2Mv^2}{\delta} \; \frac{e^{-4b}}{1-e^{-2b}} \left(\sup_{x \in \mathcal{O}_0} \varphi^2(x) + S \right).$$

Taking

$$K = \left[\frac{4S}{\delta^2} \sup_{x \in \mathcal{O}_0} \varphi^2(x) + \frac{2M}{\delta} \left(\sup_{x \in \mathcal{O}_0} \varphi^2(x) + S \right) \right] \frac{e^{-4b}}{1-e^{-2b}},$$

one gets that $P' + P'' \leq K\, v^2$. Lemma 9 is proved. □

Lemma 10 *If* $\mathrm{k} > k_+(0)$ *then* $\sum_t \gamma_t < \infty$.

Proof. From the definition of τ_v one easily sees that if $\tau_v = \infty$ for some $v > 0$, then $\sum_t \gamma_t < \infty$. This implies that for any $v > 0$

$$\mathrm{P}\left(\sum \gamma_t = \infty \right) \leq \mathrm{P}(\tau_v < \infty). \tag{34}$$

Further, by virtue of Lemma 9, if $x_0 \in \mathcal{O}'$ and $\gamma_0 < w$ then

$$\mathrm{P}(\tau_{\sqrt{w}} < \infty) \leq Kw + p(1/\sqrt{w}). \tag{35}$$

Combining (34) and (35), one gets that for any $w > 0$

$$\mathrm{P}\left(\sum \gamma_t = \infty \mid x_0 \in \mathcal{O}' \text{ and } \gamma_0 < w \right) \leq Kw + p(1/\sqrt{w}). \tag{36}$$

Define the event $\mathcal{A}_w = \{$ for some $t,\ x_t \in \mathcal{O}'$ and $\gamma_t < w\}$, then by virtue of (36),

$$\mathrm{P}\left(\sum \gamma_t = \infty \mid \mathcal{A}_w \right) \leq Kw + p(1/\sqrt{w}). \tag{37}$$

Denote by $\bar{\mathcal{A}}_w$ the complementary event, $\bar{\mathcal{A}}_w = \{$ for any $t,\ x_t \notin \mathcal{O}'$ or $\gamma_t \geq w\}$. By virtue of Lemma 7,

$$\mathrm{P}\left(\sum \gamma_t = \infty \ \&\ \bar{\mathcal{A}}_w \right) = 0. \tag{38}$$

Using (37) and (38), one gets

$$\mathrm{P}\left(\sum \gamma_t = \infty \right) = \mathrm{P}\left(\sum \gamma_t = \infty \ \&\ \mathcal{A}_w \right) + \mathrm{P}\left(\sum \gamma_t = \infty \ \&\ \bar{\mathcal{A}}_w \right) \leq$$

$$\leq (Kw + p(1/\sqrt{w})) \cdot \mathrm{P}(\mathcal{A}_w).$$

Taking into account that w can be chosen arbitrarily small and that $Kw + p(1/\sqrt{w}) \to 0$ as $w \to 0^+$, one concludes that $\mathrm{P}\left(\sum_t \gamma_t = \infty\right) = 0$. □

Now, we are in a position to prove the theorem. Suppose that $\mathrm{k} < \inf_z k_-(z)$, then $V_-^{[\mathrm{k}]} = \emptyset$, and by Lemma 2, $\{x_t\}$ diverges. So, the statement (b) of Theorem is proved.

On the other hand, according to Lemma 10, if $\mathrm{k} > k_+(0)$ then $\sum_t \gamma_t < \infty$, and by Lemmas 1 and 2, the sequence $\{x_t\}$ converges to a point from $V_-^{[\mathrm{k}]}$. Thus, the statement (a) of theorem is also established.

Acknowledgements

This work was supported by *Centre for Research on Optimization and Control* (CEOC) from the "*Fundação para a Ciência e a Tecnologia*" (FCT), cofinanced by the European Community Fund FEDER/POCTI, and by FCT: research project PTDC/MAT/72840/2006. The second author (PC) also gratefully acknowledges the financial support by the Portuguese program PRODEP 'Medida 5 - Acção 5.3 - Formação Avançada de Docentes do Ensino Superior - Concurso nr. 2/5.3/PRODEP/2001'.

References

[1] M. Nevel'son and R. Has'minskii, *Stochastic approximation and recursive estimation. Translated from the Russian by Israel Program for Scientific Translations. Translation edited by B. Silver.*, Translations of Mathematical Monographs. Vol. 47. (1976) Providence, R.I.: American Mathematical Society.

[2] Harold J. Kushner and G.George Yin, *Stochastic approximation algorithms and applications.*, Applications of Mathematics. 35. (1997) Berlin: Springer.

[3] B. T. Polyak, *New stochastic approximation type procedures*, Avtomat. i Telemekh. No 7 (1990), pp. 98-107.

[4] D. Ruppert, *Efficient estimators from a slowly convergent Robbins-Monroe process*, Tech. Report No 781, School of Operations Research and Industrial Engineering, Cornell University (1988).

[5] B. T. Polyak and A. Juditsky, *Acceleration of stochastic approximation by averaging*, SIAM J. Control Optim., 30 (1992), pp. 838-855.

[6] Y. Fang and T. J. Sejnowski, *Faster learning for dynamic recurrent backpropagation*, Neural Computation, 2 (1990), pp. 270–273.

[7] Fernando M. Silva and Luís B. Almeida, *Speeding up backpropagation*, Advanced Neural Computers, (1990), R. Eckmiller (Ed.), Elsevier Science Publishers, Amsterdam pp. 151-158.

[8] L. B. Almeida, T. Langlois, J. D. Amaral, and A. Plakhov, *Parameter adaptation in stochastic optimization*, Online Learning in Neural Networks, Saad, D. (Ed.), (1998), pp. 111–134.

[9] P. J. Werbos, *Neurocontrol and supervised learning: an overview and evaluation*, in Handbook of Intelligent Control (Neural, Fuzzy, and Adaptive approaches), D. A. White and D. A. Sofge, eds., Van Nostrand Reinhold, New York, (1992), pp. 65–89.

[10] H. Kesten, *Accelerated stochastic approximation.*, Ann. Math. Stat., 29 (1958), pp. 41–59.

[11] B. Delyon and A. Juditsky, *Accelerated stochastic approximation.*, SIAM J. Optim., 3 (1993), pp. 868–881.

[12] R. Salomon and van J. L. Hemmen, *Accelerating Backpropagation through Dynamic Self-Adaptation*, Neural Networks, 9:4, (1996) Elsevier Science Publishers, pp 589–601.

[13] Roberto Battiti, *Accelerated Backpropagation Learning: Two Optimization Methods*, Complex Systems, Inc., (1989) pp. 331-342.

[14] Marcus Frean, *A "thermal" perceptron learning rule*, Neural Computation, 4:6 (1992), The MIT Press, Cambridge–Massachusetts, pp 946–957.