Stochastics and Dynamics, Vol. 3, No. 3 (2003) 409–417

World Scientific
www.worldscientific.com

A LAW OF LARGE NUMBERS FOR RESCALED RANDOM DIFFERENCE EQUATIONS

ROBERT M. BURTON

Department of Mathematics, Oregon State University,
368 Kidder Hall, Corvallis, OR 97331, USA
bob@orst.edu

HEROLD G. DEHLING

Fakultät für Mathematik, Ruhr-Universität Bochum,
Universitätsstraße 150, 44780 Bochum, Germany
herold.dehling@ruhr-uni-bochum.de

UWE RÖSLER

Mathematisches Seminar, Christian-Albrechts-Universität Kiel,
Ludewig-Meyn-Strasse 4, 24118 Kiel, Germany

Received 28 October 2002
Revised 29 May 2003

We study the behavior of stochastic processes defined as an iterated function system

$$X_{n+1} = X_n + af(X_n, U_{n+1})$$

with initial value $X_0 = x_0$ and a stationary ergodic input signal $(U_n)_{n\geq 0}$ for small values of the parameter a. We obtain almost sure convergence of the path to the solution of the corresponding deterministic dynamical system defined by $\dot{y} = F(y)$, where $F(y) = E(f(y, U))$. The results have applications in the study of neural network learning algorithms.

Keywords: Ergodic theorem; random difference equations; learning algorithms.

AMS Subject Classification: Primary 60F15, 60J05; secondary 60J20

1. Introduction

In this paper we study the behavior of the $\mathbb{R}^d$-valued stochastic process $(X_n^a)_{n\geq 0}$ defined by the stochastic recursion scheme

$$X_{n+1}^a = X_n^a + af(X_n^a, U_{n+1}) \tag{1.1}$$

with fixed initial value $X_0^a = x_0$ and stationary ergodic input signal $(U_n)_{n\geq 1}$. Moreover $f : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ is assumed to be measurable, with further smoothness assumptions made later.

When $(U_n)_{n\geq 1}$ is an i.i.d. process taking finitely many values $a_1, \ldots, a_K$ with associated probabilities $p_1, \ldots, p_K$, the process defined by (1.1) is also called an Iterated Function System with probabilities. In such a system, the next state is a function of the present state, where the function to be applied is chosen at random from a finite set $f_1, \ldots, f_K$. Iterated function systems have been studied in connection with fractal image encoding, see e.g. [2]. Roughly speaking, an IFS with probabilities encodes the invariant distribution of the Markov process $(X_n)_{n\geq 0}$, which can in turn be recovered by running a simulation of $(X_n)_{n\geq 0}$.

Here we are particularly interested in the behavior of the process $(X_n^a)_{n\geq 0}$ for small values of the parameter a. We will show that as $a \to 0$, a properly time-scaled version of $(X_n^a)_{n\geq 0}$ converges to a solution of the deterministic differential equation

$$\frac{d}{dt}y(t) = F(y(t)) \tag{1.2}$$

with the same initial value $y(0) = x_0$. Here $F(y)$ denotes the mean vector field

$$F(y) = Ef(y, U) = \int f(y, u)d\mu(u)\,, \tag{1.3}$$

where μ is the marginal distribution of U. More precisely, we define the continuous-time stochastic processes

$$X^a(t) := X^a_{\lfloor t/a \rfloor}\,, \quad t \geq 0 \tag{1.4}$$

where $\lfloor s \rfloor$ denotes the integer part of s. Then we obtain

Theorem 1.1. *Let $(X^a(t))_{t\geq 0}$ be defined as above, and assume that $(U_n)_{n\geq 1}$ is a stationary ergodic stochastic process with marginal distribution μ. Moreover, let $f : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ be uniformly Lipschitz-continuous in the first coordinate and μ-integrable in the second coordinate. I.e., there exists a constant K such that*

$$|f(x, u) - f(y, u)| \leq K|x - y| \tag{1.5}$$

for all $x, y \in \mathbb{R}^d$, $u \in \mathbb{R}$, and in addition, $\int |f(x, u)|d\mu(u) < \infty$, for all $x \in \mathbb{R}$. Then, as $a \to 0$, the process $(X^a(t))_{t\geq 0}$ converges uniformly on compact sets to $(y(t))_{t\geq 0}$, almost everywhere. More precisely, for all $T \geq 0$,

$$\sup_{0\leq t\leq T} |X^a(t) - y(t)| \to 0 \quad \textit{as } a \to 0,\ \textit{a.e.}$$

Results of this type have been obtained before, e.g. [4]. In contrast with earlier work, we make minimal assumptions concerning the dependence structure of the input process $(U_n)_{n\geq 0}$, only requiring ergodicity.

Remark 1.2. Theorem 1.1 can also be applied to a time-dependent recursion process

$$X^a_{n+1} = X^a_n + af(an, X^a_n, U_{n+1})\,,$$

where $f : \mathbb{R} \times \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$. The corresponding limit is the solution to the time-dependent differential equation $\frac{d}{dt}y(t) = F(t, y(t))$ where $F(t, y) = Ef(t, y, U) = \int f(t, y, u)d\mu(u)$. For the proof we introduce the spacetime process $V_n^a = (an, X_n^a)$ and note that it satisfies

$$V_{n+1}^a = V_n^a + ag(V_n^a, U_{n+1}) ,$$

where $g(v, u) := (1, f(v, u))$. Let $G(v) := Eg(v, U) = (1, F(v))$. Now, if $y(t)$ is a solution to $\frac{d}{dt}y(t) = F(t, y(t))$, then $w(t) = (t, y(t))$ solves the differential equation

$$\frac{d}{dt}w(t) = (1, F(t, y(t))) = G(w(t)) .$$

Thus Theorem 1.1 is applicable and provides the desired result

$$\sup_{0\le t\le T} |X^a(t) - y(t)| \underset{a\to 0}{\to} 0 .$$

Example 1.3. To illustrate the result of the theorem, we take $f(x, u) = u - x$ and let $(U_n)_{n\ge 1}$ be i.i.d. symmetric Bernoulli, i.e. $P(U_n = -1) = P(U_n = 1) = 1/2$. Then

$$X_n^a = X_{n-1}^a + a(U_n - X_{n-1}) = (1 - a)X_{n-1} + aU_n ,$$

a process known in time series analysis as AR(1)-process. In this case, the recursion with initial value $X_0^a = x_0$ can be solved explicitly to yield

$$X_n^a = (1 - a)^n x_0 + \sum_{k=1}^{n} a(1 - a)^{n-k}U_k ,$$

and thus

$$X^a(t) = (1 - a)^{\lfloor t/a\rfloor} x_0 + \sum_{k=1}^{\lfloor t/a\rfloor} a(1 - a)^{\lfloor t/a\rfloor - k}U_k .$$

As $a \to 0$, the first term on the R.H.S. converges to e^{-t}, uniformly in $t \ge 0$. Using some standard, but lengthy, calculations one can show that $\sum_{k=1}^{\lfloor t/a\rfloor} a(1-a)^{\lfloor t/a\rfloor - k}U_k$ converges to 0, uniformly on compact sets. Thus $(X^a(t))_{t\ge 0}$ converges uniformly on compact sets to $y(t) := x_0e^{-t}$, which is indeed the solution to the mean differential equation $\dot{y} = F(y) = Ef(y, U) = -y$ with initial value $y(0) = x_0$. For this special case, one can thus directly verify the conclusions of Theorem 1.1.

The motivation for this research arises in part from a study of learning algorithms for artificial neural networks. A neural network assigns to an input vector $x \in \mathbb{R}^p$ an output $y = f_w(x) \in \mathbb{R}$, where $w \in \mathbb{R}^d$ is a vector of parameters of the network, the weights and thresholds of the neurons and the synapses in the network. It is the goal of neural learning to have the actual output of the network be as close as possible to a given teacher $T : \mathbb{R}^p \to \mathbb{R}$. We define the local error at x by

$$\Phi(x, w) = |f_w(x) - T(x)|^2 .$$

Now suppose that a stationary input signal $(\xi_k)_k$ is presented to the network. Then we can define the global error $\Phi(w)$ as the average local error

$$\Phi(w) := \int \Phi(x, w) d\mu(x) \,,$$

where μ is the distribution of ξ_k. One would like to choose the weight w in such a way that $\Phi(w)$ gets minimized. To achieve this one can use a gradient descent algorithm

$$w_{n+1} = w_n + a\nabla\Phi(w_n) \,,$$

$n \geq 1$, with given initial value w_0. However, it may be that $\Phi(w)$ is unknown. For this and other reasons we take a local gradient descent algorithm, defined by

$$W_{n+1} = W_n + a\nabla\Phi(W_n, \xi_n) \,,$$

$n \geq 1$, again with initial value w_0. Besides being simpler to implement this algorithm has closer analogy with biological neural learning. Moreover, it may enhance learning by exploring more of the error surface. Note that the random inputs make the weight sequence now a stochastic process. Simulations indicate that as $a \to 0$, the path of (W_n) closely follows the global gradient flow. Intuitively this is clear, because there are many small displacements $a\nabla\Phi(w, X_k)$ which by the Law of Large Numbers should be close to $\nabla\Phi(w)$.

2. Preliminary Results

The following proposition provides an extension of Birkhoff's pointwise ergodic theorem to certain functions of two variables. This result will play a crucial role in the proof of our main theorem.

Proposition 2.1. *Let $(U_n)_{n\geq 1}$ be a stationary, ergodic process with marginal distribution μ. Suppose that $g : [0,1] \times \mathbb{R} \to \mathbb{R}^d$ is μ-integrable in the second coordinate and uniformly Lipschitz-continuous in the first coordinate, i.e. satisfying*

$$|g(s,u) - g(t,u)| \leq K|s-t|$$

for all s, t, u and for some positive constant K. Then, as $a \to 0$

$$\sup_{0\leq t\leq 1} \left| a \sum_{k=1}^{\lfloor t/a \rfloor} g(ka, U_k) - \int \left(\int_0^t g(s,u) ds \right) d\mu(u) \right| \to 0 \,, \tag{2.1}$$

almost everywhere.

Proof. It suffices to consider the case $d = 1$. We first prove convergence in (2.1) for simple functions of the type $g(s,u) = 1_{[0,b]}(s) \cdot h(u)$ with $b \in [0,1]$ and $h : \mathbb{R} \to \mathbb{R}$ integrable. Then

$$a \sum_{k=1}^{\lfloor t/a \rfloor} g(ka, U_k) = a \sum_{k=1}^{\lfloor (t\wedge b)/a \rfloor} h(U_k) = (t \wedge b) \frac{1}{(t\wedge b)/a} \sum_{k=1}^{\lfloor (t\wedge b)/a \rfloor} h(U_k) \,.$$

By Birkhoff's ergodic theorem, $\frac{1}{s}\sum_{k=1}^{s} h(U_k) \to \int h(u)d\mu(u)$, as $s \to \infty$, almost everywhere. For a moment, we consider a fixed ω for which convergence in the ergodic theorem holds. Then, given $\varepsilon > 0$, there exists n_0 such that

$$\left| \frac{1}{s} \sum_{k=1}^{\lfloor s \rfloor} h(U_k) - \int h(u)d\mu(u) \right| \leq \varepsilon$$

for all $s \geq n_0$. Define $M = \max_{n=1,\ldots,n_0} \frac{1}{n} |\sum_{k=1}^{n} h(U_k)| + |\int h(u)d\mu(u)|$ and choose a_0 so that $n_0 M a_0 \leq \varepsilon$. Then

$$\left| a \sum_{k=1}^{\lfloor t/a \rfloor} g(ka, U_k) - \int \int_0^t g(s,u)dsd\mu(u) \right|$$

$$= (t \wedge b) \left| \frac{1}{(t \wedge b)/a} \sum_{k=1}^{\lfloor (t \wedge b)/a \rfloor} h(U_k) - \int h(u)d\mu(u) \right| .$$

If $(t \wedge b)/a \geq n_0$, the R.H.S. is bounded by ε. Otherwise $t \wedge b \leq n_0 a \leq \varepsilon/M$ and hence the R.H.S. is again bounded by ε. Together we obtain

$$\sup_{0 \leq t \leq 1} \left| a \sum_{k=1}^{\lfloor t/a \rfloor} g(ka, U_k) - \int \int_0^t g(s,u)dsd\mu(u) \right| \leq \varepsilon ,$$

thus establishing (2.1) for simple functions. By linearity, we can extend this to finite linear combinations of simple functions, i.e. to functions of the form

$$g(s,u) = \sum_{i=1}^{m} 1_{(a_i, b_i]}(s) h_i(u) . \tag{2.2}$$

Now let $g(s,u)$ be an arbitrary function satisfying the conditions of the proposition. Let $0 = a_0 \leq a_1 \leq \cdots \leq a_m = 1$ be a partition of $[0,1]$ with mesh $\Delta := \max_{i=1,\ldots,m} |a_i - a_{i-1}|$ satisfying $\Delta \leq \varepsilon/K$, and define

$$g_\varepsilon(s,u) = \sum_{i=1}^{m} 1_{(a_{i-1}, a_i]}(s) g(a_{i-1}, u) .$$

Then, for $a_{i-1} < s \leq a_i$,

$$|g_\varepsilon(s,u) - g(s,u)| = |g(a_{i-1},u) - g(s,u)| \leq K\Delta \leq \varepsilon ,$$

by the Lipschitz property of g. Hence $\sup_{s,u} |g_\varepsilon(s,u) - g(s,u)| \leq \varepsilon$ and thus

$$\sup_{0 \leq t \leq 1} \left| a \sum_{j=1}^{\lfloor t/a \rfloor} g(ka, U_k) - a \sum_{j=1}^{\lfloor t/a \rfloor} g_\varepsilon(ka, U_k) \right| \leq \varepsilon , \tag{2.3}$$

$$\sup_{0 \leq t \leq 1} \left| \int \int_0^t g(s,u)dsd\mu(u) - \int \int_0^t g_\varepsilon(s,u)dsd\mu(u) \right| \leq \varepsilon . \tag{2.4}$$

Moreover, $g_\varepsilon(s,u)$ is of the form (2.2) and hence (2.1) holds, except on a set Ω_ε of measure 0. Restricting ε to rational numbers, we obtain convergence in (2.1) for all

g_ε, except on the null set $\Omega_0 = \bigcup_{\varepsilon \text{ rational}} \Omega_\varepsilon$. Thus, except possibly on Ω_0, we have for N large enough

$$\sup_{0\le t\le 1} \left| a \sum_{k=1}^{\lfloor t/a \rfloor} g_\varepsilon(ka, U_k) - \int \int_0^t g_\varepsilon(s,u) ds d\mu(u) \right| \le \varepsilon \,. \tag{2.5}$$

Now, (2.3)–(2.5) together prove the statement of the proposition. □

It is interesting to note two special cases of Proposition 2.1, namely when $g(s,u) = g(s)$ is a function of s only, and when $g(s,u) = g(u)$ is a function of u only. In the second case, our proposition is just the Birkhoff ergodic theorem, actually under optimal conditions. In the first case, we obtain the convergence of Riemann sums to the integral. Here, however, the condition on our proposition, Lipschitz continuity of $g(s)$ is unnecessarily restrictive. These considerations also suggest that the conditions of our proposition are not sharp.

Corollary 2.2. *Let $(U_n)_{n\ge 1}$ be a stationary ergodic process of bounded random variables, and let $f : [0,1] \to \mathbb{R}$ be Lipschitz-continuous. Then, as $N \to \infty$,*

$$\frac{1}{N} \sum_{j=1}^{N} f\left(\frac{j}{N}\right) U_j \to E(U_1) \int_0^1 f(s) ds \,,$$

almost everywhere.

Proof. The proof follows directly from Proposition 2.1 with $g(s,u) = f(s) \cdot u$. □

The following lemma is a simplified and discrete version of Gronwall's lemma (for the continuous version see, e.g., [5]). For the sake of completeness we also provide a proof here.

Lemma 2.3. *Let $(b_n)_{n\ge 0}$ be a sequence of real numbers satisfying the recursive inequalities*

$$b_n \le \alpha + \sum_{k=0}^{n-1} \gamma_k b_k \tag{2.6}$$

for $n \ge 0$ and non-negative constants α, γ_k. Then

$$b_n \le \alpha \exp\left(\sum_{k=0}^{n-1} \gamma_k \right) \tag{2.7}$$

holds for all $n \ge 0$.

Proof. We will actually show the stronger statement that

$$\alpha + \sum_{k=0}^{n-1} \gamma_k b_k \le \alpha \exp\left(\sum_{k=0}^{n-1} \gamma_k \right) \tag{2.8}$$

holds for all $n \geq 0$. These inequalities will be established by induction on n. For $n = 0$, (2.8) holds trivially. Then, using (2.6), the induction hypothesis and $1 + x \leq e^x$, we get

$$\begin{aligned}\alpha + \sum_{k=0}^{n} \gamma_k b_k = \alpha + \sum_{k=0}^{n-1} \gamma_k b_k + \gamma_n b_n &\leq \alpha + \sum_{k=0}^{n-1} \gamma_k b_k + \gamma_n \left(\alpha + \sum_{k=0}^{n-1} \gamma_k b_k \right) \\ &= (1 + \gamma_n) \left(\alpha + \sum_{k=0}^{n-1} \gamma_k b_k \right) \\ &\leq e^{\gamma_n} \alpha \exp \sum_{k=0}^{n-1} \gamma_k \,,\end{aligned}$$

thus establishing the induction step in the proof of (2.8). Finally, we combine (2.6) and (2.8) to obtain the statement of the lemma. □

3. Proof of Theorem 1.1

By (1.5), the mean vector field $F : \mathbb{R}^d \to \mathbb{R}^d$, as defined in (1.3), is Lipschitz-continuous with Lipschitz constant K. Hence the differential equation (1.2) has a uniquely defined, continuously differentiable solution $y : [0, \infty) \to \mathbb{R}^d$ with $y(0) = x_0$. To show convergence of $X^a(t)$ to $y(t)$ uniformly on all compact sets in $[0, \infty)$, we have to show that, as $a \to 0$,

$$\sup_{0 \leq t \leq T} |X^a(t) - y(t)| \to 0 \,,$$

almost everywhere. Without loss of generality, we take $T = 1$. By continuity of y and by definition of X^a, it suffices to show that

$$\max_{0 \leq n \leq \frac{1}{a}} |X^a(na) - y(na)| \to 0 \,,$$

almost everywhere. To this end, we define the difference $Z_n^a := X^a(na) - y(na)$, $n \geq 0$. From (1.1) we obtain the following recursion formula for $(Z_n^a)_{n \geq 0}$:

$$\begin{aligned}Z_{n+1}^a &= X^a((n+1)a) - y((n+1)a) \\ &= X^a(na) + af(X^a(na), U_{n+1}) - y(na) - \int_{na}^{(n+1)a} \dot{y}(t)dt \\ &= Z_n^a + a[f(X^a(na), U_{n+1}) - f(y(na), U_{n+1})] \\ &\quad + a\left[f(y(na), U_{n+1}) - \frac{1}{a} \int_{na}^{(n+1)a} F(y(t))dt \right],\end{aligned}$$

where we have used that $\dot{y}(t) = F(y(t))$. Iterating this recursion and noting that $Z_0^a = 0$, we obtain

$$\begin{aligned} Z_{n+1}^a &= a\sum_{k=0}^{n}[f(X^a(ka), U_{k+1}) - f(y(ka), U_{k+1})] \\ &\quad + a\sum_{k=0}^{n} f(y(ka), U_{k+1}) - \int_0^{(n+1)a} F(y(t))dt \\ &= a\sum_{k=0}^{n}[f(X^a(ka), U_{k+1}) - f(y(ka), U_{k+1})] \\ &\quad + a\sum_{k=0}^{n} f(y(ka), U_{k+1}) - \int_0^{(n+1)a} \int f(y(t), u)d\mu(u)dt\,. \end{aligned} \tag{3.1}$$

To the last difference on the R.H.S., we can apply Proposition 2.1 with $g(t, u) = f(y(t), u)$. Thus we obtain

$$\varepsilon = \varepsilon(\omega, a) := \sup_{0\le n\le \frac{1}{a}} \left| a\sum_{k=0}^{n} f(y(ka), U_{k+1}) - \int_0^{na} \int f(y(t), u)d\mu(u)dt \right| \to 0$$

as $a \to 0$, almost everywhere. Regarding the first term, we make use of the Lipschitz property of f, yielding

$$|f(X^a(ka), U_{k+1}) - f(y(ka), U_{k+1})| \le K|X^a(ka) - y(ka)| = K|Z_k^a|\,.$$

In total we obtain

$$|Z_{n+1}^a| \le aK\sum_{k=0}^{n} |Z_k^a| + \varepsilon$$

for all $n = 1, 2, \ldots, 1/a$. We can now finish the proof by using Lemma 2.3 with $\alpha = \varepsilon(\omega, a)$ and $\gamma_k \equiv aK$. Then we obtain

$$|Z_n^a| \le \varepsilon e^{aKn} \le \varepsilon e^K\,,$$

proving almost everywhere convergence of Z_n^a to zero as $a \to 0$.

Acknowledgments

Research supported by The Netherlands Organization for Scientific Research (NWO) grant B 61-404, by NATO collaborative research grant CRG 930819 and NSF grant DMS 96-26575.

References

1. T. M. Apostol, *Calculus* (J. Wiley & Sons, 1969), Vol. II.
2. M. Barnsley, *Fractals Everywhere* (Academic Press, 1988).

3. Y. S. Chow and H. Teicher, *Probability Theory*: *Independence*, *Interchangeability*, *Martingales* (Springer Verlag, 1978).
4. D. P. Derevitskii and A. L. Fradkov, *Two models for analyzing the dynamics of adaptation algorithms*, *Automation Remote Control* **35** (1974) 59–67.
5. M. W. Hirsch and S. Smale, *Differential Equations*, *Dynamical Systems and Linear Algebra* (Academic Press, 1974).
6. M. L. Minski and S. A. Papert, *Perceptrons* (The MIT Press, 1988).
7. D. E. Rumelhart, J. L. McClelland, and the PDP research group, *Parallel Distributed Processing* (The MIT Press, 1986).