SIAM J. CONTROL AND OPTIMIZATION
Vol. 34, No. 2, pp. 437–472, March 1996

003

A DYNAMICAL SYSTEM APPROACH TO STOCHASTIC APPROXIMATIONS*

MICHEL BENAIM†

Abstract. It is known that some problems of almost sure convergence for stochastic approximation processes can be analyzed via an ordinary differential equation (ODE) obtained by suitable averaging. The goal of this paper is to show that the asymptotic behavior of such a process can be related to the asymptotic behavior of the ODE without any particular assumption concerning the dynamics of this ODE. The main results are as follows: a) The limit sets of trajectory solutions to the stochastic approximation recursion are, under classical assumptions, almost surely nonempty compact connected sets invariant under the flow of the ODE and contained in its set of chain-recurrence. b) If the gain parameter goes to zero at a suitable rate depending on the *expansion rate* of the ODE, any trajectory solution to the recursion is almost surely asymptotic to a forward trajectory solution to the ODE.

Key words. stochastic approximations, ordinary differential equations, chain-recurrence, neural networks

AMS subject classifications. 62L20, 34D05, 34C29

Introduction. The classical theory of stochastic approximations, born with the papers of Robbins and Monro (1951) and Kiefer and Wolfowitz (1952), concerns the study of stochastic algorithms whose general form can be written as

$$w_{n+1} - w_n = \gamma_n H(w_n, \xi_n), \tag{1}$$

where $H : \mathbf{R}^m \times \mathbf{R}^d \mapsto \mathbf{R}^m$ is a measurable function that characterizes the algorithm, $\{w_n\}_{n\geq 0} \in \mathbf{R}^m$ is the sequence of parameters to be recursively updated, $\{\xi_n\}_{n\geq 0} \in \mathbf{R}^d$ is a sequence of random inputs where $H(w_n, \xi_n)$ is observable, and $\{\gamma_n\}_{n\geq 0}$ is a sequence of "small" nonnegative scalar gains.

At each time step, the vector ξ_n is a new observation that causes w_n to be updated to take new information into account. The gain sequence $\{\gamma_n\}_{n\geq 0}$ can be chosen to be constant or decreasing. In this paper we restrict attention to algorithms with *decreasing gain sequence.* More precisely, we shall always assume that $\{\gamma_n\}_{n\geq 0}$ is a decreasing sequence of positive numbers which satisfies the classical relations

$$\lim_{n\to\infty} \gamma_n = 0$$

and

$$\sum_{n\geq 0} \gamma_n = +\infty.$$

To analyze the asymptotic behavior of the algorithm (1) it is convenient to introduce the averaged ordinary differential equation (ODE)

$$\frac{dw}{dt} = \overline{H}(w), \tag{2}$$

* Received by the editors August 9, 1993; accepted for publication (in revised form) October 18, 1994. This research was supported by a grant from the Centre National de la Recherche Scientifique (Programme Cogniscience).

† Department of Mathematics, University of California at Berkeley, Berkeley, CA 94720.

where

$$\overline{H}(w) = \lim_{n\to\infty} E(H(w, \xi_n))$$

and $E(.)$ denotes the mathematical expectation.

This method, called the *method of ordinary differential equation*, was introduced by Ljung (1977) and Kushner and Clark (1978) and widely studied thereafter. It has inspired a number of important works, such as the book by Kushner and Clark (1978), numerous articles by Kushner, and, more recently, the book by Benveniste, Métivier, and Priouret (1990). The main idea of the method is to describe the asymptotic behavior of the algorithm in terms of the behavior of the ODE. For stochastic algorithms having a decreasing gain sequence, the classical result stating the relationship between the algorithm (1) and the ODE (2) has the following form:

> *Let $w*$ be a stable equilibrium for the ODE. If $\{\gamma_n\}_{n\geq 0}$ goes to zero at a suitable rate and if the sequence $\{w_n\}_{n\geq 0}$ enters infinitely often a compact subset of the domain of attraction of $w*$, then $\{w_n\}_{n\geq 0}$ converges almost surely toward w^*.*

This kind of result has been obtained by Ljung (1977); Kushner and Clark (1978); Métivier and Priouret (1984, 1987); Benveniste, Métivier, and Priouret (1990); and Kuan and White (1992), among others, under fairly general conditions. It relies the asymptotic behavior of the algorithm with a strong notion of recurrence for the ODE: the notion of *fixed point.*

With increasing interest in *artificial neural networks* and due to some limitations of the standard *backpropagation* algorithm, "heuristic" learning rules for feedforward neural networks have been recently proposed and experimentally studied. The ODE associated with these algorithms is not given by a gradient vectorfield (as is the case for backpropagation), and the classical convergence results on stochastic gradient algorithms cannot be successfully applied. The consideration of these algorithms led us to formulate the following problem:

> *Without any particular assumption on the dynamics of $\overline{H}$, is it again possible to describe the asymptotic behavior of* (1) *in terms of the asymptotic behavior of* (2)*?*

The main goal of this paper is to address this question.

In §§1 and 2 we relate the behavior of the algorithm to a weak notion of recurrence for the ODE: the notion of *chain recurrence.* We state a theorem which asserts that under the assumptions of the Kushner and Clark lemma (1978) the limit sets of the trajectory solutions to (1) are nonempty compact connected sets invariant under the flow of the ODE and contained in its set of chain-recurrence.

This result shows that the limit sets of (1) look like the omega limit sets of (2), and we ask the question of their exact relationship. We address this question in §5. It is shown that it may happen that the limit set of a trajectory solution to (1) never coincides with an omega limit set of (2), but that it always does if the gain parameter goes to zero at a suitable rate depending on the vectorfield $\overline{H}$. Our approach, in this section, is essentially based on "shadowing" results recently proved by Morris W. Hirsch together with L^q estimates of the distance between the trajectory solutions to (1) and (2).

In §8 we apply the results of §§1–5 to prove some convergence theorems for the neural network learning algorithms mentioned above.

Main theorems are proved in §§4 and 7. Several applications are considered in §§3 and 6.

1. A deterministic theorem. In order to introduce the main result of this section we begin with a few notations and classical definitions from dynamical systems.

Notation and definitions. Let Γ be a topological space and $\Phi : \mathbf{R} \times \Gamma \mapsto \Gamma$ be a continous map denoted by $\Phi(t, x) = \Phi_t(x)$. The family $\{\Phi_t\}_{t \in \mathbf{R}}$ is called a *flow* on Γ if it satisfies the group property

$$\Phi_0 = \text{Identity},$$

$$\forall (t, s) \in \mathbf{R}^2, \ \Phi_t \circ \Phi_s = \Phi_{t+s}.$$

Let $\overline{H}$ denote a continous vectorfield defined on $\mathbf{R}^m$ with unique integral curves. The *flow* of $\overline{H}$ is the family of mappings defined on $\Gamma = \mathbf{R}^m$ by

$$\frac{d}{dt}\Phi_t(w) = \overline{H}(\Phi_t(w)).$$

A set X is said to be *invariant* (respectively, *positively invariant*) under the flow Φ if for all $t \in \mathbf{R}, \Phi_t(X) \subset X$ (respectively, for all $t \geq 0$). In this case we let $\Phi|X$ denote the restricted flow (respectively, semiflow).

A point x is an *equilibrium* if $\Phi_t(x) = x$ for all $t \in \mathbf{R}$. When Φ is induced by the vectorfield $\overline{H}$, equilibria coincide with zeros of $\overline{H}$. A point x is a *periodic point* if there exists $T > 0$ such that $\Phi_T(x) = x$. Equilibria and periodic points are clearly recurrent points. In general, we may say that a point is recurrent if it somehow returns near where it was under time evolution.

A notion of recurrence related to slightly perturbed orbits is the notion of *chain recurrence*. Suppose Γ is a metric space with a metric d. Let $\delta > 0$ and $T > 0$. A point x is said to be (δ, T) *recurrent* if there exist an integer k, some points y_i in Γ, and numbers $t_i, 0 \leq i \leq k-1$, such that

$$t_i \geq T; \quad d(y_0, x) < \delta; \quad d(\Phi_{t_i}(y_i), \ y_{i+1}) < \delta \quad \text{for } i = 0, \ldots, k-1; \quad x = y_k.$$

Intuitively (δ, T) recurrent points are points that one would take to be periodic if the position of points were only known with a finite accuracy δ. If x is (δ, T) recurrent for any $\delta > 0$ and $T > 0$, x is said to be *chain-recurrent*. We denote by $CR(\Phi)$ the set of chain-recurrent points. If Φ is induced by the vectorfield $\overline{H}$, we may also use the notation $CR(\overline{H})$ for $CR(\Phi)$. The set $CR(\Phi)$ has the property to be closed and invariant.

A subset $X \subset \Gamma$ is said *internally chain-recurrent* if X is a nonempty compact invariant set of which every point is chain-recurrent for the restricted flow $\Phi|X$ (i.e., $CR(\Phi|X) = X$).

For example, if Γ is compact, Conley (1978) proved that $CR(\Phi)$ is internally chain-recurrent.

The sets which describe the asymptotic behavior of the orbits of the flow Φ are the *omega limit sets*. The omega limit set of $w \in \Gamma$, denoted by $\omega(w)$, is the set of $x \in \Gamma$ such that $\lim_{k \to \infty} \Phi_{t_k}(w) = x$ for some sequence $t_k > 0$ with $\lim_{k \to \infty} t_k = +\infty$. If the forward trajectory $\{\Phi_t(w); t \geq 0\}$ has compact closure, $\omega(w)$ is a nonempty compact connected set internally chain-recurrent. The alpha limit set $\alpha(w)$ of w is defined as the omega limit set of w for the reversed flow $\{\Phi_{-t}\}_{t \geq 0}$.

To recapitulate, if we note $Per(\Phi)$ the set of periodic points (including the equilibria) and $\mathcal{L}^+(\Phi) = \bigcup_{w \in \Gamma} \omega(w)$, the following inclusions hold:

$$Per(\Phi) \subset \mathcal{L}^+(\Phi) \subset CR(\Phi).$$

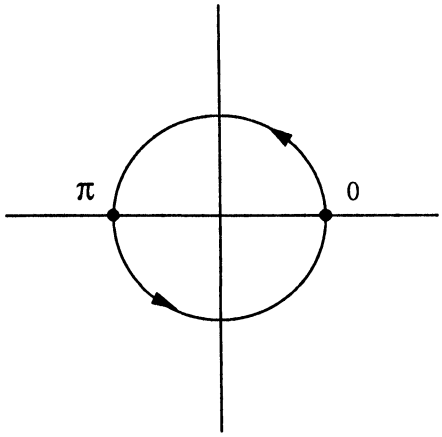

FIG. 1.

EXAMPLE 1.1. *Consider the flow on the unit circle* $S^1 = \mathbf{R}/2\pi\mathbf{Z}$ *induced by the differential equation*

$$\frac{d\theta}{dt} = f(\theta),$$

where f *is a* 2π*-periodic smooth nonnegative function such that*

$$f^{-1}(0) = \{k\pi : k \in \mathbf{Z}\}.$$

See Fig. 1.
We have

$$Per(\Phi) = \{0, \pi\} = \mathcal{L}^+(\Phi)$$

and

$$CR(\Phi) = S^1.$$

Internally chain-recurrent sets are $\{0\}, \{\pi\}$, *and* S^1. *Note that the set* $X = [0, \pi]$ *is a compact invariant set consisting of chain-recurrent points. However,* X *is not internally chain-recurrent.*

A deterministic theorem. To describe the asymptotic behavior of the algorithm (1) we introduce the limit set of the sequence $\{w_n\}_{n\geq 0}$. We denote this limit set by $L(\{w_n\}_{n\geq 0})$. It is the set of $x \in \mathbf{R}^m$ such that $\lim_{k\to\infty} w_{n_k} = x$ for some subsequence $\{n_k\}_{k\geq 0}$ with $\lim_{k\to\infty} n_k = +\infty$.

The following theorem is a deterministic result that will be applied in §2 to show that the limit sets of the trajectories solutions to the algorithm (1) have basically the same properties as the omega limit sets of the trajectories solution to the ODE (2). The assumptions A1, A2, and A3 of this theorem are the assumptions of the Kushner and Clark lemma (1978).

We use the following notation:

$$\tau_0 = 0,$$

$$\tau_n = \sum_{i=0}^{n-1} \gamma_i.$$

We let $\|.\|$ denote a norm on $\mathbf{R}^m$.

THEOREM 1.2. *Let $\overline{H} : \mathbf{R}^m \mapsto \mathbf{R}^m$ be a continuous vectorfield with unique integral curves. Let $\{w_n\}_{n\geq 0}$ be solution to the recursion*

$$w_{n+1} - w_n = \gamma_n(\overline{H}(w_n) + u_n + b_n), \tag{3}$$

where $\{\gamma_n\}_{n\geq 0}$ is a decreasing gain sequence. Assume that

- A1) *$\{w_n\}_{n\geq 0}$ is bounded.*
- A2) *$\lim_{n\to\infty} b_n = 0$.*
- A3) *For each $T > 0$,*

$$\lim_{n\to\infty} \left(\sup_{\{k\,;\, 0\leq \tau_k - \tau_n \leq T\}} \left\| \sum_{i=n}^{k-1} \gamma_i . u_i \right\| \right) = 0.$$

Then $L(\{w_n\}_{n\geq 0})$ is a connected set internally chain-recurrent for the flow Φ induced by $\overline{H}$.

The next theorem shows that Theorem 1.2 gives the best result that can be expected under the Kushner and Clark assumptions. It justifies the fact that the chain recurrence is a notion well suited to the description of the asymptotic behavior of (1).

Assume given a locally Lipschitz vectorfield $\overline{H} : \mathbf{R}^m \mapsto \mathbf{R}^m$ and a decreasing gain sequence $\{\gamma_n\}_{n\geq 0}$.

THEOREM 1.3. *Let $L \subset \mathbf{R}^m$ be a connected set internally chain-recurrent for the flow induced by $\overline{H}$. There exist sequences $\{b_n\}_{n\geq 0}$, $\{u_n\}_{n\geq 0}$, and $\{w_n\}_{n\geq 0}$ such that*

(a) *Conditions* A1, A2, *and* A3 *of Theorem* 1.2 *are satisfied.*

(b) *The sequence $\{w_n\}_{n\geq 0}$ is the solution to* (3) *and admits L as a limit set.*

Theorem 1.3 follows easily from the following proposition (see Benaim and Hirsch (1995b)).

PROPOSITION 1.4. *Let $L \subset \mathbf{R}^m$ be a connected set internally chain-recurrent for the flow induced by $\overline{H}$. There exists a continuous function $u : \mathbf{R}_+ \mapsto \mathbf{R}^m$ and a point $w_0 \in \mathbf{R}^m$ such that*

(a) *$\lim_{t\to\infty} u(t) = 0$.*

(b) *The solution to the nonautonomous system*

$$\frac{dw}{dt} = \overline{H}(w) + u(t)$$

with initial condition $w(0) = w_0$ is bounded and admits L as a limit set.

To prove Theorem 1.3 we let $w_n = w(\tau_n)$ and $u_n = u(\tau_n)$, where $w(.)$ and $u(.)$ are the functions of Proposition 1.4. Then we have

$$w_{n+1} - w_n = \gamma_n(\overline{H}(w_n) + u_n) + O(\gamma_n^2)$$

and Theorem 1.3 follows from Proposition 1.4.

REMARK 1.5. *Throughout this paper the process $\{w_n\}_{n\geq 0}$ will be assumed to be bounded. Several conditions ensuring that this assumption is fulfilled are discussed in the literature on stochastic approximations. They usually rely on the existence of*

some convergent supermartingale for the process (1) (*see, e.g., Theorem* 5.2, *chapter* 2, *of Nevel'son and Has'minskii* (1974) *or Theorem* 8 *of Fort and Pagès* (1994)).

In the spirit of this section, we give a simple condition which is purely deterministic.

PROPOSITION 1.6. *Assume that* $\overline{H}$ *is globally Lipschitz. Assume the existence of a function* $V : \mathbf{R}^m \to \mathbf{R}_+$ *uniformly continuous such that*

(i) $\lim_{||x||\to\infty} V(x) = \infty$.

(ii) *There exist positive numbers* δ, r, *and* T *such that*

$$\forall x \in \mathbf{R}^m,\ ||x|| \geq r \Longrightarrow V(\Phi_T(x)) - V(x) \leq -\delta.$$

Then conditions A2 *and* A3 *of Theorem* 1.2 *imply condition* A1.

The proof of this result follows easily from Lemma 4.4 and is left to the reader.

Note that if V is smooth, condition (ii) holds if the following more easily checked condition is satisfied: there exists $\delta' > 0$ such that for all $||x|| \geq r$

$$\langle \nabla V(x), \overline{H}(x) \rangle \leq -\delta',$$

where ∇ denotes the gradient.

2. Limit sets of stochastic approximation processes. In this section, we assume that $\{\xi_n\}_{n\geq 0}$ is a sequence of $\mathbf{R}^d$-valued random variables defined on a probability space $(\Omega, \mathcal{F}, \mathcal{P})$. We note $\mathcal{F}_n^m$, the σ field generated by $\{\xi_i;\ n \leq i \leq m\}$ for $m \geq n$. For $q \in [1, \infty]$ we let $\|.\|_q$ denote the $L^q(\Omega)$ norm for random variables ($\|X\|_q = E(\|X\|^q)^{1/q}$) and $\|.\|_\infty$ the $L^\infty(\Omega)$ norm ($\|X\|_\infty = \operatorname{ess\,sup}\|X\|$).

In applications of Theorem 1.2 to the stochastic approximation (1) one may choose

$$\overline{H}(w) = \lim_{n\to\infty} E(H(w, \xi_n)),$$

$$u_n = H(w_n, \xi_n) - \int H(w_n, \xi)\mu_n(d\xi),$$

and

$$b_n = \int H(w_n, \xi)\mu_n(d\xi) - \overline{H}(w_n),$$

where μ_n is the distribution of ξ_n. Then we try to verify assumptions A2 and A3 by use of some regularity properties of H and maximal inequalities for sum of random variables. Let us mention two examples.

Independent inputs. The first example is a classical Robbins–Monro algorithm in which the observations are assumed to be independent and identically distributed. This yields a simple martingale access to condition A3 as in Gladyshev (1965) and Hall and Heyde (1980).

We let M denote a given subset of $\mathbf{R}^m$ (not necessarily compact).

PROPOSITION 2.1. *Let* $\{w_n\}_{n\geq 0}$ *be the solution to* (1). *Assume that*

- A1) $\{\xi_n\}_{n\geq 0}$ *is a sequence of independent and identically distributed random variables.*
- A2) $P(\{w_n\}_{n\geq 0}$ *is bounded*$) = 1$ *and* $P(\forall n \in N, w_n \in M) = 1$.
- A3) $w \mapsto \overline{H}(w) = E(H(w, \xi_0))$ *is continuous with a unique flow.*

There exists $q \geq 2$ *such that*

- A4) $w \mapsto \|H(w,\xi_0)\|_q$ *is bounded on* M.
- A5) $\sum_{n=0}^{\infty} \gamma_n^{1+q/2} < +\infty$.

Then the conclusions of Theorem 1.2 *hold with probability one.*

Proof. To see that, we let $b_n = 0$ and $u_n = H(w_n,\xi_n) - \overline{H}(w_n)$. Then $E(u_n/\mathcal{F}_0^{n-1}) = 0$. For $q = 2$, assumptions A4 and A5 imply $\sum_n \gamma_n^2 . \|u_n\|_2^2 < +\infty$, and condition A3 of Theorem 1.2 is a direct consequence of the L^2-bounded martingale convergence theorem. For $q > 2$, it follows from a result of Métivier and Priouret (1987, Cor. 11). (See also Benveniste, Métivier, and Priouret (1990, Cor. 8, p. 297).) Note that in this case the sequence $\{\sum_n \gamma_n . u_n\}_{n\geq 0}$ is not necessarily convergent. □

Mixing inputs. The following example extends this result to situations in which the observable inputs are nonindependent and nonstationary random variables which satisfy a strong mixing condition. Such situations arise naturally in some applications of feedforward neural networks as forecasting, prediction of time series, or chaos modelling.

Here our approach is motivated by the work of Kuan and White (1992), who have proved some convergence results for stochastic approximation procedures by using the theory of *mixingales* developed by McLeish (1975). Conditions A1–A6 can be compared with conditions of Kuan and White's theorems (Thm. 2.2.1 and Cors. 2.2.3 and 2.3.5). The condition A6′ gives a generalization which allows a gain parameter of the order of $\frac{1}{n^\alpha}$ with $\alpha < 1$. The price for this is a strengthening of the boundness condition.

For $n \geq 0, m \geq 0$ define

$$\phi_{n,m} = \sup_{\{A\in\mathcal{F}_0^n, B\in\mathcal{F}_{n+m}^{n+m}\}} | P(B/A) - P(B) |,$$

$$\alpha_{n,m} = \sup_{\{A\in\mathcal{F}_0^n, B\in\mathcal{F}_{n+m}^{n+m}\}} | P(B\cap A) - P(B)P(A) |,$$

$$\phi_m = \sup_{n\geq 0} \phi_{n,m},$$

$$\alpha_m = \sup_{n\geq 0} \alpha_{n,m}.$$

We shall say that the process $\{\xi_n\}_{n\geq 0}$ is ϕ mixing (respectively, α mixing) if $\lim_{n\to\infty} \phi_n = 0$ (respectively, $\lim_{n\to\infty} \alpha_n = 0$). Observe, however, that this condition is a weakening of the classical ϕ mixing (respectively, α mixing) definition (see, for instance, Billingsley (1968, §20, p. 166)). It would be the same if $\mathcal{F}_{n+m}^{n+m}$ were replaced by $\mathcal{F}_{n+m}^{+\infty}$. This weaker definition is motivated by our use of McLeish's results (1975).

PROPOSITION 2.2. *Let* $\{w_n\}_{n\geq 0}$ *be the solution to* (1). *Assume that*

- A1) $\{\xi_n\}_{n\geq 0}$ *is a* ϕ *mixing (respectively,* α *mixing) process.*
- A2) $\{w_n\}_{n\geq 0}$ *is bounded with probability one.*
- A3) $\overline{H}(w) = \lim_{n\to\infty} E(H(w,\xi_n))$ *exists.*
- A4) *There exists a measurable function* $k(.)$ *such that*

$$\forall x, y \in \mathbf{R}^m \|H(x,\xi) - H(y,\xi)\| \leq k(\xi)\|x - y\|.$$

There exists $r \in [2,\infty]$ *such that*

- A5) *The map* $w \mapsto \sup_{n\geq 0} \|H(w,\xi_n)\|_r$ *is bounded on any bounded set and* $\sup_{n\geq 0} \|k(\xi_n)\|_r < +\infty$.
- A6) *(L^r case). If* $r < \infty$, $\phi_n = O(\frac{1}{n^\beta})$ *for some* $\beta > \frac{r}{2r-2}$ *(respectively,* $\alpha_n = O(\frac{1}{n^\beta})$ *for some* $\beta > \frac{r}{r-2}$*) and*

$$\sum_{n=0}^{\infty} \gamma_n^2 < +\infty.$$

- A6′) *(L^∞ case). If* $r = \infty$, $\phi_n = O(\frac{1}{n^\beta})$ *for some* $\beta > \frac{1}{2}$ *(respectively,* $\alpha_n = O(\frac{1}{n^\beta})$ *for some* $\beta > 1$*) and*

$$\sum_{n=0}^{\infty} \gamma_n^{1+q/2} < +\infty$$

for some $q \in [2, 2\beta + 1[$.

Then the conclusions of Theorem 1.2 *hold with probability one.*

The proof of this result is given in the appendix (§9).

In view of the fact that the assumptions of Theorem 1.2 are the assumptions of the Kushner and Clark lemma, several other examples of application can be found in the literature. We refer the reader to the book by Kushner and Clark (1978, Chap. II) for such examples. In the case where the input process $\{\xi_n\}_{n\geq 0}$ is a Markov process or, more generally, a Markov process controlled by the parameter w, condition A3 of Theorem 1.2 can be derived from the analysis provided in the articles by Ljung (1977) and Métivier and Priouret (1987, Cor. 11) (see also Benveniste, Métivier, and Priouret (1990, Cor. 8, p. 297)).

3. Applications. In this section we give a few examples to illustrate how results of §§1 and 2 can be used to describe the global asymptotic behavior of stochastic approximation processes.

In the remainder of this section $\overline{H}$ is a vectorfield on $\mathbf{R}^m$ with unique integral curves. The sequence $\{w_n\}_{n\geq 0}$ denotes either a deterministic sequence solution to (3) under assumptions of Theorem 1.2 or a random sequence solution to (1) under assumptions of Proposition 2.1 or 2.2. In this last case, all the properties stated below have to be understood as "almost sure" properties.

Local behavior. First, note that Theorem 1.2 generalizes the classical result mentioned in the introduction.

An equilibrium w^* of $\overline{H}$ is said *asymptotically stable* if there exists an open neighborhood U of w^* such that

$$\lim_{t\to\infty} \Phi_t(w) = w^*$$

uniformly in $w \in U$. The *domain of attraction* of w^* is the set of all points whose forward trajectories are attracted by w^*.

PROPOSITION 3.1. *Let* w^* *be an asymptotically stable equilibrium of* $\overline{H}$. *Assume that* $\{w_n\}_{n\geq 0}$ *enters infinitely often a compact subset—say,* Q*—of the domain of attraction of* w^*. *Then*

$$\lim_{n\to\infty} w_n = w^*.$$

Proof. According to Theorem 1.2, $L(\{w_n\}_{n\geq 0}) \cap Q$ is nonempty and is contained in $CR(\overline{H}) \cap Q$. On the other hand, it is not difficult[1] to show that $CR(\overline{H}) \cap Q = \{w^*\}$. Thus $\{w^*\} = L(\{w_n\}_{n\geq 0}) \cap Q$ and, as $L(\{w_n\}_{n\geq 0})$ is connected, $L(\{w_n\}_{n\geq 0}) = \{w^*\}$. □

Gradientlike systems. Let Φ be a flow on a metric space Γ and $\Lambda \subset \Gamma$ be an invariant set.

A C^0 map $V : \Gamma \mapsto \mathbf{R}$ is said to be a *Lyapunov function* for Λ if for all $x \in \Gamma$ the function $t \in \mathbf{R}_+ \mapsto V(\Phi_t(x))$ is constant for $x \in \Lambda$ and strictly decreasing for $x \notin \Lambda$.

If Λ equals the equilibria set, V is called a *strict Lyapunov function* and Φ is called a *gradientlike system.*

PROPOSITION 3.2. *Assume that Γ is compact. Let $\Lambda \subset \Gamma$ be a compact invariant set and $V : \Gamma \mapsto \mathbf{R}$ a Lyapounov function for Λ. Assume that the cardinal of $V(\Lambda)$ is finite. Then*

$$CR(\Phi) \subset \Lambda.$$

COROLLARY 3.3. *Assume that $\overline{H}$ admits a strict Lyapunov function and isolated equilibria. Then $\{w_n\}_{n\geq 0}$ converges toward an equilibrium.*

Proof. We apply Proposition 3.2 to the flow induced by $\overline{H}$ on $\Gamma = L(\{w_n\}_{n\geq 0})$. It follows from Theorem 1.2 that $L(\{w_n\}_{n\geq 0})$ consists of equilibria. As it is a connected set and equilibria are isolated, $L(\{w_n\}_{n\geq 0})$ is an equilibrium. □

REMARK 3.4. *Note that Corollary 3.3 applies to stochastic gradient algorithms for which $\overline{H}$ is the gradient of a cost function $C : \mathbf{R}^m \mapsto \mathbf{R}$,*

$$\overline{H}(w) = \nabla C(w).$$

In §8 we will give another application of Proposition 3.2 to a class of learning processes which are not given by a stochastic gradient.

Proof of Proposition 3.2. Let $V(\Lambda) = \{v_1, \ldots, v_l\}, v_1 < v_2 < \cdots < v_l$. Choose real numbers $v_1', v_2', \ldots, v_l'$ such that $v_1 < v_1' < v_2 < \cdots < v_{l-1}' < v_l < v_l'$, and define $M_i = \{x \in \Gamma / V(x) \leq v_i'\}$.

Let $\Lambda_i = \Lambda \cap V^{-1}(v_i)$; Λ_i is the largest invariant set contained in $M_i - M_{i-1}$. Indeed, let $A \subset M_i - M_{i-1}$ be an invariant set and let $x \in A$. By a standard theorem on Lyapunov functions, $\alpha(x) \cup \omega(x) \subset \Lambda$. So $V(\alpha(x)) = V(\omega(x)) = v_i$, and as V is strictly decreasing along any trajectory outside Λ, x is necessarily in Λ_i.

Let $T > 0$. By compactness of the sets M_j, there exists $\epsilon > 0$ such that

$$\forall x \in M_j, V(\Phi_T(x)) \leq v_j' - \epsilon.$$

Pick $\delta > 0$ such that

$$\forall (x, y) \in \Gamma \times \Gamma, d(x, y) \leq \delta \Rightarrow | V(x) - V(y) | \leq \epsilon.$$

It follows that any (δ, T) chain $\{y_0, y_1, \ldots, y_k\}$ (i.e., $d(\Phi_{t_i}(y_i), y_{i+1}) < \delta$ for some $t_i \geq T$) with $y_0 \in M_j$ is included in M_j. Therefore, the set $CR_j = CR(\Phi) \cap (M_j - M_{j-1})$ is invariant. Hence, $CR_j \subset \Lambda_j$ and $CR(\Phi) \subset \Lambda$. □

[1] This follows, for example, from Proposition 3.10.

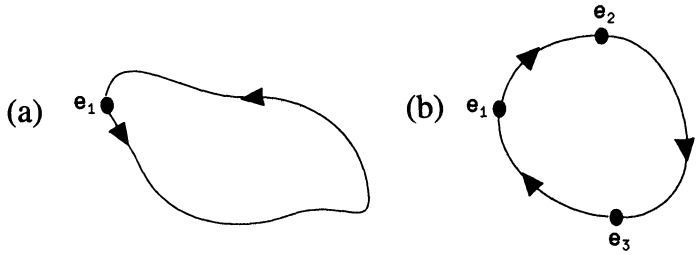

FIG. 2. (a) *A one-equilibrium cycle.* (b) *A three-equilibrium cycle.*

No-cycle systems. Let Φ be a flow on a metric space Γ. We say that Φ has *simple dynamics* if for every $x \in \Gamma$ the alpha and omega limit sets of x are equilibria. This means that every backward and forward trajectory converges toward an equilibrium. If Φ is induced by the vectorfield $\overline{H}$, we say that $\overline{H}$ has simple dynamics if $\Phi|L$ has simple dynamics for each compact invariant set $L \subset \mathbf{R}^m$.

For a flow with simple dynamics, we say that the equilibrium e_1 goes to the equilibrium e_2 if there exists a nonequilibrium orbit $\gamma \subset \Gamma$ such that $\alpha(\gamma) = e_1$ and $\omega(\gamma) = e_2$. γ is called a *connecting* orbit. To indicate that e_1 goes to e_2, we write $e_1 \rightsquigarrow e_2$. To indicate that γ is the connecting orbit from e_1 to e_2, we write $\gamma : e_1 \rightsquigarrow e_2$.

A *cycle of equilibria* is an union

$$A = \bigcup_{j=1}^{n} (\{e_j\} \cup \gamma_j)$$

consisting of equilibria e_j, $j = 1, \ldots, n$, and connecting orbits γ_j, $j = 1, \ldots, n$, such that

(i) $\gamma_j : e_j \rightsquigarrow e_{j+1}$, $j = 1, \ldots, n-1$.

(ii) $\gamma_n : e_n \rightsquigarrow e_1$.

REMARK 3.5. *A cycle of equilibria is connected internally chain-recurrent (Fig. 2).*

PROPOSITION 3.6. *Assume that Γ is compact and Φ has a finite number of equilibria, simple dynamics, and no cycle. Then $CR(\Phi)$ is the equilibria set.*

COROLLARY 3.7. *Assume that $\overline{H}$ has isolated equilibria, simple dynamics, and no cycle. Then $\{w_n\}_{n\geq 0}$ converges toward an equilibrium.*

Proof. We apply Proposition 3.6 to the flow induced by $\overline{H}$ on $\Gamma = L(\{w_n\}_{n\geq 0})$ and conclude exactly as in the proof of Corollary 3.3. □

Fort and Pagès (1994) recently proved a result similar to Corollary 3.7 by using the Kushner and Clark lemma. Systems with cycle of equilibria will be considered in §6.

The notion of *simple dynamics* and *no-cycle property* can be extended to nonconvergent situations. Denote by $\mathcal{L}(\Phi)$ the union of all alpha and omega limit sets of Φ. Assume that there exist nonempty compact disjoint invariant subsets $\Lambda_j \subset \Gamma$, $j = 1, \ldots, n$, such that

$$\mathcal{L}(\Phi) \subset \Lambda = \bigcup_{j=1}^{n} \Lambda_j.$$

If there exists $x \notin \Lambda$ such that $\alpha(x) \subset \Lambda_1$ and $\omega(x) \subset \Lambda_2$, we write $\Lambda_1 \rightsquigarrow \Lambda_2$ and define cycles among the Λ_j exactly as in the simple dynamics case.

PROPOSITION 3.8. *Assume Γ is compact and there is no cycle among the Λ_j. Then $CR(\Phi) \subset \Lambda$.*

Proof. Let $\hat{\Gamma}$ be the topological quotient space obtained by collapsing each Λ_i to a point. It is not difficult to check that $\hat{\Gamma}$ is a regular space with a countable basis. Therefore, by the Urysohn theorem, $\hat{\Gamma}$ is metrizable. Let π denotes the quotient map $\pi : \Gamma \to \hat{\Gamma}$. The flow Φ induces a flow $\hat{\Phi}$ on $\hat{\Gamma}$ defined by $\hat{\Phi} \circ \pi = \pi \circ \Phi$, which has simple dynamics, no cycle, and the Λ_j as equilibria. Therefore, by Proposition 3.6, chain-recurrent points of $\hat{\Phi}$ are equilibria. If $x \in \Gamma$ is chain-recurrent for Φ it is clear, by definition of chain-recurrence and uniform continuity of π, that $\pi(x)$ is chain-recurrent for $\hat{\Phi}$. Thus $CR(\Phi) \subset \Lambda$. □

COROLLARY 3.9. *Assume there exist nonempty compact disjoint subsets $\Lambda_j \subset \mathbf{R}^m$, $j = 1, \ldots, n$, invariant under the flow of $\overline{H}$ such that every alpha or omega limit point belongs to $\Lambda = \bigcup_{j=1}^n \Lambda_i$. Assume there is no cycle among the Λ_j. Then there exists $j \in \{1, \ldots, n\}$ such that $L(\{w_n\}_{n \geq 0}) \subset \Lambda_j$.*

Proof of Proposition 3.6. There are several ways to prove Proposition 3.6. For example it can be easily deduced from the "filtration theory" exposed in Shub (1986). Here, for simplicity we decided to deduce it from elementary properties of chain-recurrent sets. On the other hand these properties are very useful and give a good understanding of the notion of chain-recurrence.

Let $Y \subset \Gamma$. The forward trajectory of Y is the set $Y \cdot [0, \infty) = \Phi([0, \infty) \times Y) = \{\Phi_t(y);\ t \geq 0;\ y \in Y\}$.

The omega limit set (respectively, alpha limit set) of Y, denoted by $\omega(Y)$ (respectively, $\alpha(Y)$) is defined as the maximal invariant set in $\mathrm{clos}(Y \cdot [0, \infty))$ (respectively, $\mathrm{clos}(Y \cdot (-\infty, 0])$), where "clos" denotes closure.

A nonempty compact invariant set $A \subset X$ is an *attractor* if A has an open neighborhood U in X such that $\omega(U) = A$ or a *repeller* if $\alpha(U) = A$. An attractor or repeller is *proper* provided that it is not open in X.

The following proposition follows from §§5 and 6 of Conley (1978, Chap. 2).

PROPOSITION 3.10 (Conley (1978)).

(a) *Let $N \subset \Gamma$ be a compact set. Let $A \subset \Gamma$ be the maximal invariant set contained in N. If A is nonempty and* not *an attractor, there exists $p \in \partial N \subset \Gamma$ such that the backward orbit $\gamma_-(p) \subset N$ and $\alpha(p)$ is a nonempty subset of A.*

(b) *A internally chain-recurrent set has no proper attractor or repeller.*

(c) *The chain-recurrent set is internally chain-recurrent.*

Let us now prove Proposition 3.6. Let X be a connected component of $CR(\Phi)$. By assertion (c) of Proposition 3.10, X is internally chain-recurrent. Consider the flow $\Psi = \Phi | X$ and let $Equ(\Psi) = \{e_i,\ i = 1, \ldots, n\}$ denote the equilibria set of Ψ. Since Φ has simple dynamics and no cycle the relation $\rightsquigarrow$ induces a partial ordering on $Equ(\Psi)$.

Assume e_n is minimal for this partial ordering. We claim that e_n is an attractor for Ψ. It follows from assertion (b) of Proposition 3.10 that e_n is open and closed in X. Thus $X = \{e_n\}$.

It remains to prove that e_n is an attractor for Ψ. Let N be a compact neighborhood of e_n which separates e_n from other equilibria. The maximal invariant set in N is e_n; otherwise it would exist a entire orbit disjoint from e_n inside N. The dynamics being simple, this orbit would have to connect e_n to itself. Since we assume that there is no cycle, this is impossible.

Now we use assertion (a) of Proposition 3.10. If e_n is not an attractor, there

exists $p \in \partial N$ with $\alpha(p) = e_n$, but this contradicts the fact that e_n is minimal for the partial ordering $\rightsquigarrow$. □

Morse–Smale systems. In this subsection we mention briefly an application of the previous results to a class of stochastic approximation processes and urn models which have been recently considered by Benaim and Hirsch (1995a). For more details the reader is referred to that paper.

Assume $\overline{H}$ is C^r $(r \geq 1)$. $\overline{H}$ is called *Morse–Smale* if

(i) $\overline{H}$ has a global compact attractor (i.e., the point at infinity is a source);

(ii) all periodic orbits and equilibria are hyperbolic;

(iii) stable and unstable manifolds of periodic orbits (and equilibria) intersect only transversely;

(iv) every alpha or omega limit set is a periodic orbit or an equilibrium.

It is known that these conditions imply that there are only finitely many periodic orbits.

Suppose $\overline{H}$ is a Morse–Smale vector field. Denote by $\mathcal{L}(\overline{H})$ the union of all alpha and omega limit sets of $\overline{H}$, and by $Per(\overline{H})$ the union of all periodic orbits and equilibria. If $\overline{H}$ is Morse–Smale, $\mathcal{L}(\overline{H})$ decomposes as

$$\mathcal{L}(\overline{H}) = Per(\overline{H}) = \Lambda_1 \cup \cdots \cup \Lambda_n,$$

where the Λ_i are the distinct hyperbolic periodic orbits and equilibria. On the other hand, it follows from the transversal condition (iii) that there is no cycle among the Λ_i (see, e.g., Proposition 3.2 of Palis (1969)). Thus, we have the following corollary.

COROLLARY 3.11. *Assume $\overline{H}$ is Morse–Smale. Then $L(\{w_n\}_{n\geq 0})$ is an equilibrium or a periodic orbit.*

Proof. By Corollary 3.9, $L(\{w_n\}_{n\geq 0}) \subset \Lambda_i$ for some i. Since $L(\{w_n\}_{n\geq 0})$ is invariant and Λ_i is a periodic orbit or an equilibrium, we must have $L(\{w_n\}_{n\geq 0}) = \Lambda_i$. □

Nonconvergence toward unstable periodic orbits is considered in Benaim and Hirsch (1995a).

Planar systems. For planar systems it is possible to give a complete description of $L(\{w_n\}_{n\geq 0})$. A planar flow is a flow defined on a open subset of $\mathbf{R}^2$. The following theorem is proved in Benaim and Hirsch (1995c).

THEOREM 3.12. *Let Φ be a planar flow with isolated equilibria and L be an internally chain-recurrent set for Φ. Every point $x \in L$ satisfies one of the following conditions:*

(i) *x is an equilibrium.*

(ii) *x is a periodic point (i.e., x belongs to a periodic orbit).*

(iii) *There exists a cycle of equilibria in L which contains x.*

COROLLARY 3.13. *If $\overline{H}$ is a planar vectorfield with isolated equilibria, $L(\{w_n\}_{n\geq 0})$ is a connected union of equilibria, periodic orbits, and cycles of equilibria.*

Using the same kind of result, a Poincaré–Bendixson theorem for a class of stochastic differential equations is given in Benaim (1995b).

4. Proof of Theorem 1.2. We denote by 1_A the indicator function of the set A (i.e., $1_A(x) = 1$ if $x \in A$ and $1_A(x) = 0$ if $x \notin A$).

For any sequence $\{z_n\}_{n\geq 0} \in \mathbf{R}^m$ we denote by $Z(.)$ the function defined for all $t \geq 0$ by

$$Z(t) = \sum_{n\geq 0} z_n 1_{[\tau_n, \tau_{n+1}[}(t)$$

and by $Z^0(.)$ the interpolated process defined for all $t \geq 0$ by

$$Z^0(t) = \sum_{n\geq 0} \left[(z_{n+1} - z_n).\frac{(t-\tau_n)}{\gamma_n} + z_n \right] 1_{[\tau_n, \tau_{n+1}[}(t).$$

With these notations, the recursion satisfied by $\{w_n\}_{n\geq 0}$ can be rewritten as

$$W^0(t) - W^0(0) = \int_0^t \overline{H}(W(s))ds + \int_0^t U(s)ds + \int_0^t B(s)ds. \tag{4}$$

Remark that the assumptions A1, A2, and A3 are equivalent to

- A1′) $\{W^0(t), t \geq 0\}$ is bounded.
- A2′) $\lim_{t\to\infty} B(t) = 0$.
- A3′) For each $T > 0$,

$$\lim_{t\to\infty} \left(\sup_{h\in[0,T]} \left\| \int_t^{t+h} U(s)ds \right\| \right) = 0.$$

The function $t \mapsto W^0(t)$ is uniformly continous. This follows easily from the integral formula (4) and conditions A1′, A2′, A3′. This can also be deduced from the Kushner and Clark lemma (1978) (see Theorem 4.5).

We denote by $L(W^0)$ the limit set of $\{W^0(t), t \geq 0\}$ and let Q denote a compact subset of $\mathbf{R}^m$ which contains $\{W^0(t), t \geq 0\}$.

LEMMA 4.1. $L(\{w_n\}_{n\geq 0}) = L(W^0)$.

Proof. It is clear that $L(\{w_n\}_{n\geq 0}) \subset L(W^0)$. Conversely, let

$$w* = \lim_{t_k\to+\infty} W^0(t_k),$$

a limit point of W^0. Define the map $m : \mathbf{R}_+ \mapsto \mathbb{N}$ by

$$m(t) = \sup\{p \in \mathbb{N}/\tau_p \leq t\}. \tag{5}$$

One has $\lim_{t\to+\infty}(t - \tau_{m(t)}) = 0$ because $\lim_{n\to+\infty} \gamma_n = 0$. The uniform continuity of W^0 implies $\lim_{t_k\to+\infty} W^0(\tau_{m(t_k)}) = w*$. This proves the lemma. □

LEMMA 4.2. *For all* $T > 0$,

$$\lim_{t\to+\infty} \sup_{h\in[-T,T]} ||W^0(t+h) - \Phi_h(W^0(t))|| = 0.$$

For convenience, the proof of this lemma is postponed to the end of the section.

COROLLARY 4.3. $L(\{w_n\}_{n\geq 0})$ *is internally chain-recurrent.*

Proof. Since W^0 is continuous and bounded, $L(W^0)$ is a nonempty compact connected set.

Let us verify that $L(W^0)$ is invariant under Φ. Let $p \in L(W^0)$, $p = \lim_{t_i\to\infty} W^0(t_i)$ for some sequence $t_i \to \infty$. Let $T \in \mathbf{R}$. If $T > 0$, then

$$\lim_{t_i\to\infty} d(\Phi_T(W^0(t_i)), W^0(t_i + T)) = 0$$

by Lemma 4.2. Therefore $\Phi_T(p) = \lim_{t_i\to\infty} W^0(t_i + T) \in L(W^0)$. If $T < 0$, the proof is analogous.

It remains to prove that $L(W^0)$ is chain-recurrent for the restricted flow $\Phi|L(W^0)$. Here we adopt a method used by Robinson (1977) to show that a diffeomorphism on

a compact manifold is chain-recurrent on the set of chain-recurrence. Recall that $Q \subset \mathbf{R}^m$ denotes a compact set which contains $\{W^0(t), t \geq 0\}$.

Claim 1. Let $n \in \mathbf{N}$, $T > 0$, $p \in L(W^0)$. There exists a finite sequence

$$n \leq a_0^n \leq \cdots \leq a_{k(n)}^n$$

such that, with the notations

$$y_i^n = W^0(a_i^n), \qquad i = 0, \ldots, k(n),$$

and

$$t_i^n = a_{i+1}^n - a_i^n, \qquad i = 0, \ldots, k(n) - 1,$$

the following hold:

(a) $d(y_0^n, p) \leq \frac{1}{n}$ and $d(y_{k(n)}^n, p) \leq \frac{1}{n}$.

(b) $T \leq t_i^n \leq 2T$, $i = 0, \ldots, k(n) - 1$.

(c) $d(\Phi_{t_i^n}(y_i^n), y_{i+1}^n) \leq \frac{1}{n}$, $i = 0, \ldots, k(n) - 1$.

Proof. Let $n \in \mathbf{N}$. Lemma 4.2 shows that there exists $A_n > 0$ such that for any $t \geq A_n$ and for all $0 \leq h \leq 2T$, $d(\Phi_h(W^0(t)), W^0(t+h)) \leq \frac{1}{n}$.

As $p \in L(W^0)$ there exists $a_0^n \geq \sup(A_n, n)$ such that $d(W^0(a_0^n), p) \leq \frac{1}{n}$ and there exists $T' > T$ such that $d(W^0(a_0^n + T'), p) \leq \frac{1}{n}$. Write $a_0^n + T' = kT + r$, where $k \in \mathbf{N}$ and $0 \leq r < T$. Then define $a_i^n = a_0^n + i(T + \frac{r}{k})$, $i = 0, \ldots, k$. □

Let $C_n = \{y_i^n, i = 0, \ldots, k(n)\}$, where y_i^n is defined as in Claim 1. As C_n is a compact set, we may extract from $\{C_n\}_{n \geq 0}$ a subsequence which converges toward a compact set C for the Haussdorf metric in Q. It is clear that $C \subset L(W^0)$.

Claim 2. Let $\delta > 0$ and $T > 0$; then p is (δ, T) recurrent for the restricted flow $\Phi | L(W^0)$.

Proof. By uniform continuity of the flow on Q there exists $\alpha > 0$ such that $d(x, y) \leq \alpha$ implies $d(\Phi_t(x), \Phi_t(y)) \leq \delta/3$ uniformly in $t \in [0, 2T]$. We may always assume $\alpha \leq \delta/3$. Choose n large enough such that $1/n \leq \delta/3$ and $d(C_n, C) \leq \alpha$. Then we construct a finite sequence $Z_0, \ldots, Z_{k(n)} \in C$ such that $d(Z_i, y_i^n) \leq \alpha$ for $i = 0, \ldots, k(n)$. Then

$$d(Z_0, p) \leq \alpha + 1/n \leq \delta,$$

$$d(Z_{k(n)}, p) \leq \alpha + 1/n \leq \delta,$$

and

$$d(\Phi_{t_i^n}(Z_i), Z_{i+1}) \leq d(\Phi_{t_i^n}(Z_i), \Phi_{t_i^n}(y_i^n)) + d(\Phi_{t_i^n}(y_i^n), y_{i+1}^n) + d(y_{i+1}^n, Z_{i+1})$$

$$\leq \delta/3 + \frac{1}{n} + \alpha \leq \delta. \qquad \square$$

Proof of Lemma 4.2.

The Lipschitz case. Here we assume that $\overline{H}$ is locally Lipschitz. We let $L(Q)$ denote the Lipschitz constant of $\overline{H}$ on Q and $\|\overline{H}\|_Q$ the uniform norm of $\overline{H}$ on Q. The next lemma proves Lemma 4.2 with an estimate. This estimate will be useful to prove the main result of §5.

LEMMA 4.4. *For all* $T > 0$ *and all* $t \geq 0$,

$$\sup_{h\in[0,T]} \|W^0(t+h) - \Phi_h(W^0(t))\| \leq e^{L(Q)\,T}[2\epsilon(t,T)(1+TL(Q)) + TL(Q)\|\overline{H}\|_Q\gamma_{m(t)}],$$

where

$$\epsilon(t,T) = \sup_{\{k;\ 0\leq\tau_k-\tau_{m(t)}\leq T+1\}} \left\| \sum_{i=m(t)}^{k-1} \gamma_i.u_i \right\| + (T+1)\left[\sup_{\{k;0\leq\tau_k-\tau_{m(t)}\leq T+1\}} \|b_k\|\right].$$

Proof. We begin with a simple inequality:

$$\forall u, v \in [\tau_{m(t)}, \tau_{m(t+T)+1}[\ \left\| \int_u^v (U(s)+B(s))ds \right\| \leq 2\epsilon(t,T). \tag{6}$$

To prove (6) we note that for any $u \geq \tau_{m(t)}$ there exists $\alpha \in [0,1]$ for which

$$\int_{\tau_{m(t)}}^{u} (U(s)+B(s))ds = \alpha \int_{\tau_{m(t)}}^{\tau_{m(u)}} (U(s)+B(s))ds + (1-\alpha)\int_{\tau_{m(t)}}^{\tau(m(u)+1)} (U(s)+B(s))ds.$$

As for $u, v \in [\tau_{m(t)}, \tau_{m(t+T)+1}[$,

$$\int_u^v (U(s)+B(s))ds = -\int_{\tau_{m(t)}}^{u} (U(s)+B(s))ds + \int_{\tau_{m(t)}}^{v} (U(s)+B(s))ds.$$

Inequality (6) follows.

According to (4),

$$\begin{aligned} W^0(t+h) - \Phi_h(W^0(t)) = \int_0^h \overline{H}(W(t+s))ds - \int_0^h \overline{H}(\Phi_s(W^0(t))ds \\ + \int_t^{t+h} (U(s)+B(s))ds. \end{aligned} \tag{7}$$

Let

$$A(h) = \|W^0(t+h) - \Phi_h(W^0(t))\|.$$

Equation (7) implies

$$\begin{aligned} A(h) \leq L(Q)\int_0^h A(s)ds + \int_0^h \|\overline{H}(W^0(t+s)) - \overline{H}(W(t+s))\|ds \\ + \left\| \int_t^{t+h} (U(s)+B(s))ds \right\|. \end{aligned} \tag{8}$$

On the other hand, for any $h \in [0,T]$

$$\|W^0(t+h) - W(t+h)\| = \left\| \int_{\tau_{m(t+h)}}^{t+h} [\overline{H}(W(s)) + U(s) + B(s)]ds \right\|,$$

and inequality (6) implies

$$\|W^0(t+h) - W(t+h)\| \leq \gamma_{m(t+h)} \|\overline{H}\|_Q + 2\epsilon(t,T). \tag{9}$$

From inequalities (8) and (9), we deduce that for any $h \in [0,T]$

$$A(h) \leq L(Q) \int_0^h A(s)ds + 2\epsilon(t,T)(1 + TL(Q)) + TL(Q)\gamma_{m(t)} \|\overline{H}\|_Q,$$

and we conclude by using Gronwall's inequality. □

The non-Lipschitz case. Here we prove Lemma 4.2, assuming only that $\overline{H}$ is continous with unique integral curves. The key of the proof is to use the Kushner and Clark lemma (1978). Let $W^s()$ be the function defined for any $s \geq 0$ by

$$\forall t \geq -s, W^s(t) = W^0(t+s)$$

and

$$\forall t < -s, W^s(t) = w_0.$$

The Kushner and Clark lemma is the following.

THEOREM 4.5 (Kushner and Clark (1978)). *Under the assumptions* A1, A2, *and* A3 *of Theorem* 1.2, $\{W^s(.)\}_{s\geq 0}$ *is relatively compact in* $C^0(\mathbf{R}, \mathbf{R}^m)$ *with respect to the topology of uniform convergence on bounded intervals (i.e., from every sequence of the set* $\{W^s(.)\}_{s\geq 0}$ *it is possible to select a subsequence which converges uniformly on bounded intervals), and the limit of each convergent subsequence is the solution to the ODE.*

What we want to prove (i.e., Lemma 4.2) is equivalent to

$$\lim_{s\to\infty} sup_{h\in[-T,T]} \|W^s(h) - \Phi_h(W^s(0))\| = 0 \tag{10}$$

for all $T > 0$. Let D denote a distance on $C^0(\mathbf{R}, \mathbf{R}^m)$ induced by the topology of uniform convergence on bounded intervals; then (10) can be rewritten as

$$\lim_{s\to\infty} D(W^s(.), \Phi(., W^s(0))) = 0. \tag{11}$$

Let W^* be an arbitrary limit point of $\{W^s(.)\}_{s\geq 0}$. By Theorem 4.5 W^* is a solution to the ODE, and by uniqueness of integral curves $W^*(t) = \Phi(t, W^*(0))$ for all t. Thus, $W^*(.) = \Phi(., W^*(0))$. This proves (11). □

5. L^q estimates and shadowing. In this section we consider the following question:

Given $\{w_n\}_{n\geq 0}$, *a trajectory solution to* (1), *does there exist a solution to* (2) *whose omega limit set is* $L(\{w_n\}_{n\geq 0})$*?*

Theorem 1.3 shows that (at least under assumptions of Theorem 1.2) the answer is generally negative since $L(\{w_n\}_{n\geq 0})$ can be an arbitrary internally chain-recurrent set. However, it is useful to understand what kind of conditions ensure a positive answer to this question. A case of particular interest in applications is given by the following problem:

Assume that each solution to (2) *converges toward an equilibrium.*

Does every solution to (1) *converge also toward an equilibrium?*

We saw in §3 several examples for which $CR(\overline{H})$ is the set of equilibria and the theorems of §§1 and 2 were applied to answer positively. But it may happen that $CR(\overline{H})$ contains nonequilibrium points (see Example 6.3) and further conditions are required.

We begin with a simple example.

EXAMPLE 5.1. *Consider the recursion which is defined in polar coordinates* $\rho \geq 0$, $\theta \in \mathbf{R}/(2\pi\mathbf{Z})$ *by*

$$\rho_{n+1} - \rho_n = \gamma_n(g(\rho_n) + 1_{[0.5\,,3]}(\rho_n).\xi_n),$$

$$\theta_{n+1} - \theta_n = -\gamma_n,$$

where $\{\xi_n\}_{n\geq 0}$ *is a sequence of independently and identically distributed random variables with uniform distribution on* $[-\frac{1}{2}, \frac{1}{2}]$, $\gamma_n = \frac{1}{n^\alpha}$ *for some* $0 < \alpha \leq 1$, *and* $g : \mathbf{R}_+ \to \mathbf{R}$ *is a smooth function which is zero on* $\{0\} \cup [1, 2]$, *positive on* $]0, 1[$, *and negative on* $]2, \infty)$. *The ODE associated with this recursion is defined by*

$$\frac{d\rho}{dt} = g(\rho),$$

$$\frac{d\theta}{dt} = -1.$$

The phase portrait of this ODE is given by Fig. 3.

We see that any connected internally chain recurrent set of this ODE is either the equilibrium $0_{\mathbf{R}^2}$ *or a cylinder of periodic orbits*

$$C_{a,b} = \{\rho : a \leq \rho \leq b\} \times \{\theta \in \mathbf{R}/(2\pi\mathbf{Z})\},\ 1 \leq a \leq b \leq 2.$$

Assume that the initial condition of the process is not $0_{\mathbf{R}^2}$. *Therefore, according to Proposition* 2.1, *the limit set* $L(\{w_n\}_{n\geq 0})$ *of the process has to be a cylinder. In fact, it is not difficult to show that*

(a) *if* $\alpha > \frac{1}{2}$, $L(\{w_n\}_{n\geq 0})$ *is almost surely a periodic orbit* $L(\{w_n\}_{n\geq 0}) = C_{a,a}$ *for some* $1 \leq a \leq 2$;

(b) *if* $\alpha < \frac{1}{2}$, $L(\{w_n\}_{n\geq 0}) = C_{1,2}$.

The main reason is that the sum $\sum_n \gamma_n \xi_n$ *converges for* $\alpha > \frac{1}{2}$, *while*

$$\limsup_{n\to\infty} \sum_n \gamma_n \xi_n = -\liminf_{n\to\infty} \sum_n \gamma_n \xi_n = +\infty$$

for $\alpha < \frac{1}{2}$ (*see, e.g., Neveu* (1964, *p.* 138)).

In case (a) $L(\{w_n\}_{n\geq 0})$ *is an omega limit set of the ODE. Case* (b) *gives an example for which the asymptotic behavior of* (1) *is quite different from the asymptotic behavior of* (2).

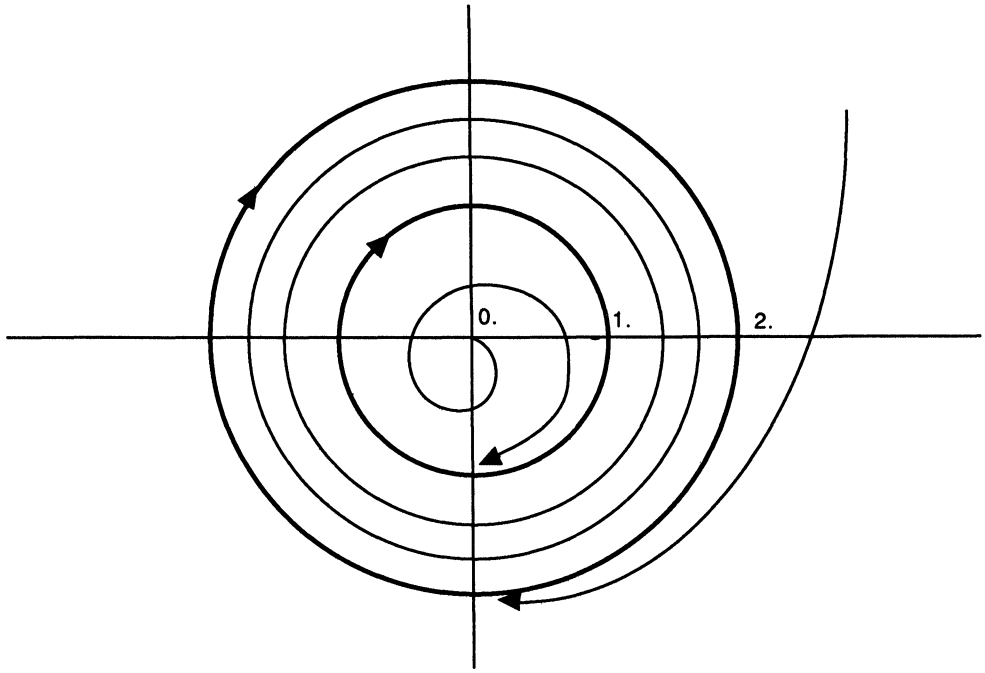

FIG. 3.

Expansion rate. In the previous example, the condition $\alpha < 1$ means, intuitively, that the convergence of $\{\gamma_n\}_{n\geq 0}$ to zero is not fast enough to ensure the convergence of $\{w_n\}_{n\geq 0}$ toward the omega limit sets of the ODE. We now formalize this idea and show that, conversely, if $\{\gamma_n\}_{n\geq 0}$ goes to zero at a suitable rate depending on the *expansivity* of the ODE, then $\{w_n\}_{n\geq 0}$ is in some sense asymptotic to a forward trajectory of (2).

Here we make crucial use of the ideas and methods introduced by Morris W. Hirsch in a recent paper (1993). The main idea of what follows is to use a shadowing theorem proved in Hirsch (1993) together with L^q estimates of the error which is made when (1) is replaced by (2).

To avoid technicalities, we will assume throughout the remainder of this section that $\overline{H}$ is a C^1 vectorfield on $\mathbf{R}^m$ with the point at infinity as a source. By ∞ as a source we mean that there exists a bounded nonempty open set $U \subset \mathbf{R}^m$ such that for all $w \in \mathbf{R}^m$

$$\lim_{t\to\infty} d(\Phi_t(w), \operatorname{clos}(U)) = 0$$

and for some $T > 0$

$$\Phi_T(\operatorname{clos}(U)) \subset U.$$

Let K denote a nonempty compact set positively invariant under the flow of $\overline{H}$. The *expansion rate* of $\overline{H}$ in K is defined in Hirsch (1993) (see also Hirsch and Pugh (1970)). For convenience, we introduce it in a logarithmic form:

$$l_{exp}(\overline{H}, K) = \lim_{t\to+\infty} \left[\min_{w\in K} \frac{\log(\|(D\Phi_t(w))^{-1}\|^{-1})}{t}\right],$$

where $D\Phi_t(w)$ denote the differential of Φ_t at w. The limit exists by subadditivity. We call $l_{exp}(\overline{H}, K)$ the "log-expansion rate" of $\overline{H}$ in K. This real number measures

the expansivity of the dynamical system induced by $\overline{H}$. It is zero if the flow is isometric, positive if the flow has a tendency to be expansive, and negative otherwise. Let $\mathrm{clos}(\mathcal{L}(\overline{H}))$ be the closure of all alpha and omega limit points of the trajectories solution to (2). As we assume that "∞" is a source, $\mathrm{clos}(\mathcal{L}(\overline{H}))$ is a compact nonempty invariant set. We define the "log-expansion rate" of $\overline{H}$ as

$$l_{exp}(\overline{H}) = l_{exp}(\overline{H}, \mathrm{clos}(\mathcal{L}(\overline{H}))).$$

This definition makes sense and is motivated by the following important property (Hirsch (1993)): If L is a compact invariant subset of K containing all alpha and omega limit points in K, then $l_{exp}(\overline{H}, K) = l_{exp}(\overline{H}, L)$. Some properties of $l_{exp}(\overline{H})$ are given in §6.

A shadowing theorem. Let $\{\alpha_n\}_{n\geq 0}$ denote a sequence of nonnegative real numbers.

Define the "log-convergence rate" of $\{\alpha_n\}_{n\geq 0}$ with respect to the time scale $\tau_n = \sum_{i=0}^{n-1} \gamma_i$ as

$$l_\tau(\alpha) = \limsup_{n\to+\infty} \frac{\log(\alpha_n)}{\tau_n}.$$

Now consider the same recursion as in Theorem 1.2 in a probabilistic framework:

$$w_{n+1} - w_n = \gamma_n \overline{H}(w_n) + \gamma_n u_n + \gamma_n b_n, \tag{12}$$

where $\{\gamma_n\}_{n\geq 0}$ is a decreasing gain sequence, $\{u_n\}_{n\geq 0}$ and $\{b_n\}_{n\geq 0}$ are two sequences of $\mathbf{R}^m$-valued random variables defined on the probability space $(\Omega, \mathcal{F}, \mathcal{P})$, and $\overline{H} : \mathbf{R}^m \mapsto \mathbf{R}^m$ is a C^1 vectorfield with ∞ as a source.

Recall that $\|.\|_q$ denotes the $L^q(\Omega)$ norm. For each $T > 0$ and each $q \in [1, +\infty[$ let

$$\alpha_n^{q,T} = \left\| \sup_{\{k\,;0\leq\tau_k-\tau_n\leq T\}} \left\| \sum_{i=n}^{k-1} \gamma_i u_i \right\| \right\|_q$$

and

$$\beta_n^{q,T} = \left\| \sup_{\{k\,;0\leq\tau_k-\tau_n\leq T\}} \|b_k\| \right\|_q .$$

THEOREM 5.2. *Let $\{w_n\}_{n\geq 0}$ be solution to the recursion* (12). *Assume that there exists $q \geq 1$ such that*

- A1) $E(\sup_{n\geq 0} \|\{w_n\}_{n\geq 0}\|^q) < +\infty$.
- A2) *For each $T > 0$,*

$$l_\tau(\beta^{q,T}) < \min(0, l_{exp}(\overline{H})),$$

$$l_\tau(\alpha^{q,T}) < \min(0, l_{exp}(\overline{H})),$$

$$l_\tau(\gamma) < \min(0, l_{exp}(\overline{H})).$$

Then

- a) *there exists a random vector* w' *such that*

$$\lim_{n\to+\infty} \|w_n - \Phi_{\tau_n}(w')\| = 0$$

almost surely.

- b) *If there exists a compact* $Q \subset \mathbf{R}^m$ *such that* $\{w_n\}_{n\geq 0}$ *remains in* Q *almost surely (in which case A1 is obviously satisfied), then the following estimate holds:*

$$l_\tau(\{\|w_n - \Phi_{\tau_n}(w')\|_q\}_{n\geq 0}) \leq \sup(l_\tau(\beta^{q,T}), l_\tau(\alpha^{q,T}), l_\tau(\gamma)).$$

REMARK 5.3. *Conclusion* (a) *of Theorem* 5.2 *implies that* $L(\{w_n\}_{n\geq 0}) = \omega(w')$ *almost surely.*

As in §2 we apply the previous result to the stochastic approximation (1), where $\{\xi_n\}_{n\geq 0}$ is a sequence of random variables defined on $(\Omega, \mathcal{F}, \mathcal{P})$. Maximal inequalities for sum of random variables reduce condition A2 to a simple condition on $l_\tau(\gamma)$. First of all, note that for any $\lambda > 0$,

$$l_{\lambda\tau}(\lambda\gamma) = \frac{1}{\lambda} l_\tau(\gamma).$$

If $\gamma_n = f(n)$ for some positive decreasing function f with $\int_1^{+\infty} f(s)ds = +\infty$, then

$$l_\tau(\gamma) = \limsup_{x\to+\infty} \frac{\log(f(x))}{\int_1^x f(s)ds}.$$

For example, if

$$\gamma_n = \frac{1}{n^\alpha \log(n)^\beta},$$

then $l_\tau(\gamma) = 0$ for $0 < \alpha < 1$ and $\beta \geq 0$, $l_\tau(\gamma) = -1$ for $\alpha = 1$ and $\beta = 0$, and $l_\tau(\gamma) = -\infty$ for $\alpha = 1$ and $0 < \beta \leq 1$.

Independent inputs. As in Proposition 2.1, we let M denote a subset of $\mathbf{R}^m$.

PROPOSITION 5.4. *Let* $\{w_n\}_{n\geq 0}$ *be the solution to* (1). *Assume that*

- A1) $\{\xi_n\}_{n\geq 0}$ *is a sequence of independent and identically distributed random variables.*
- A2) $P(\forall n \in N, w_n \in M) = 1$.
- A3) $w \mapsto \overline{H}(w) = E(H(w, \xi_0))$ *is* C^1 *with* ∞ *as a source.*

There exists $q \geq 2$ *such that*

- A4) $E(\sup_{n\geq 0} \|\{w_n\}_{n\geq 0}\|^q) < +\infty$ *and* $w \mapsto \|H(w, \xi_0)\|_q$ *is bounded on* M.
- A5) $l_\tau(\gamma) < 2\min(0, l_{exp}(\overline{H}))$.

Then

- a) *The conclusion* a) *of Theorem* 5.2 *holds.*
- b) *If* M *is compact,* $l_\tau(\{\|w_n - \Phi_{\tau_n}(w')\|_q\}_{n\geq 0}) \leq \frac{1}{2} l_\tau(\gamma)$.

Proof. Let $b_n = 0$ and $u_n = H(w_n, \xi_n) - \overline{H}(w_n)$. As already noted, $\{u_n\}_{n\geq 0}$ is a martingale difference. For $q = 2$, Doob's inequality for L^2 martingales gives

$$\alpha_n^{2,T} \leq \left[C(M) \sum_{i=n}^{m(\tau_n+T)-1} \gamma_i^2 \right]^{\frac{1}{2}},$$

where $C(M)$ is a positive constant and $m(T)$ is defined by (5). So

$$\alpha_n^{2,T} \leq [C(M)\gamma_n T]^{\frac{1}{2}}.$$

Then $l_\tau(\alpha^{2,T}) \leq \frac{1}{2} l_\tau(\gamma)$ and the condition A2 of Theorem 5.2 is satisfied. For $q > 2$,

$$\alpha_n^{q,T} \leq \left[C(M,T) \sum_{i=n}^{m(\tau_n+T)-1} \gamma_i^{1+q/2} \right]^{\frac{1}{2}}$$

for some constant $C(M,T)$. This inequality is proved, in a more general context, in Métivier and Priouret (1987, Prop. 8). Therefore,

$$\alpha_n^{q,T} \leq [C(M,T)T]^{\frac{1}{q}} \gamma_n^{1/2},$$

and the result follows. □

REMARK 5.5. *It is interesting to note that the condition* A5 *of Proposition* 5.4 *is always satisfied for* $\gamma_n = \frac{1}{n \log(n)}$. *For* $\gamma_n = \frac{\epsilon}{n+n_0}$, *it reduces to the condition* $\epsilon < -\frac{1}{2 l_{exp}(\overline{H})}$.

Mixing inputs. In the case corresponding to Proposition 2.2, in which the observations are given by a mixing process, our approach of condition A2 in Theorem 5.2 is based on some kind of uniform maximal inequalities. Unfortunately, these estimates depend on the dimension of the parameter space and the condition we obtain presents the "curse of dimensionality." Here we shall assume that $\{\xi_n\}_{n\geq 0}$ is stationary to facilitate the verification of assumption A2 of Theorem 5.2.

PROPOSITION 5.6. *Let* $\{w_n\}_{n\geq 0}$ *be the solution to* (1). *Assume that*

- A1) $\{\xi_n\}_{n\geq 0}$ *is a stationary* ϕ *mixing (respectively,* α *mixing) process.*
- A2) *There exists a compact set* $Q \subset \mathbf{R}^m$ *such that* $P(\forall n \geq 0, w_n \in Q) = 1$.
- A3) $\overline{H}(w) = \lim_{n\to\infty} E(H(w,\xi_n))$ *exists.*
- A4) *There exists a measurable function* $k(.)$ *such that*

$$\forall x, y \in \mathbf{R}^m \|H(x,\xi) - H(y,\xi)\| \leq k(\xi)\|x - y\|.$$

There exists $r \in [2,\infty]$ *such that*

- A5) *The map* $w \mapsto \sup_{n\geq 0} \|H(w,\xi_n)\|_r$ *is bounded on* Q *and* $\sup_{n\geq 0} \|k(\xi_n)\|_r < +\infty$.
- A6) *(L^r case). If* $r < \infty$, $\phi_n = O(\frac{1}{n^\beta})$ *for some* $\beta > \frac{r}{2r-2}$ *(respectively,* $\alpha_n = O(\frac{1}{n^\beta})$ *for some* $\beta > \frac{r}{r-2}$*).*
- A6′) *(L^∞ case). If* $r = \infty$, $\phi_n = O(\frac{1}{n^\beta})$ *for some* $\beta > \frac{1}{2}$ *(respectively,* $\alpha_n = O(\frac{1}{n^\beta})$ *for some* $\beta > 1$*).*
- A7) $l_\tau(\gamma) < 2(m+1)\min(0, l_{exp}(\overline{H}))$.

Then the conclusions of Theorem 5.2 *hold with probability one.*

The proof is given in appendix (§9).

6. Applications. Here again $\overline{H}$ denotes a C^1 vectorfield with ∞ as a source.

Convergent systems. We say that $\overline{H}$ is a *convergent system* if $\overline{H}$ admits a finite number of equilibria $\{e_1,\ldots,e_n\}$ and

$$\mathcal{L}(\overline{H}) = \{e_1,\ldots,e_n\}.$$

Equivalently, this means that the flow induced by $\overline{H}$ on $S^m = \mathbf{R}^m \cup \{\infty\}$ (the compactification of $\mathbf{R}^m$) has simple dynamics and finitely many equilibria.

Let $\{\lambda_i^j,\ j = 1, \ldots, m\}$ denote the set of eigenvalues of the matrix $D\overline{H}(e_i)$. Define

$$\beta(e_i) = \min\{\mathrm{Re}(\lambda_i^j) : j = 1, \ldots, m\},$$

where Re denotes the real part. Since e_i is a fixed point, $D\Phi_t(e_i) = exp(tD\overline{H}(e_i))$. Therefore

$$\lim_{t\to\infty} \frac{\log(\|(D\Phi_t(e_i))^{-1}\|^{-1})}{t} = \beta(e_i),$$

and by definition of the log-expansion rate we deduce the following proposition.

PROPOSITION 6.1. *If $\overline{H}$ is a convergent system with equilibria $\{e_1, \ldots, e_n\}$, then*

$$l_{exp}(\overline{H}) = \min\{\beta(e_i) : i = 1, \ldots, n\}.$$

COROLLARY 6.2. *Let $\{w_n\}_{n\geq 0}$ be the solution to* (1). *Assume that conditions* A1–A4 (*respectively,* A1–A6, A6′) *of Proposition* 5.4 (*respectively,* 5.6) *are satisfied. Assume that the averaged vectorfield $\overline{H}$ defined by* A3 *is convergent and that*

$$\forall i \in \{1, \ldots, n\},\ l_\tau(\gamma) < 2\min(0, \beta(e_i))$$

(respectively, $l_\tau(\gamma) < 2(m+1)\min(0, \beta(e_i))$). Then $\{w_n\}_{n\geq 0}$ converges almost surely toward an equilibrium.

EXAMPLE 6.3. *Consider the following stochastic approximation process defined on $\mathbf{R}^2$ by*

$$x_{n+1} - x_n = \frac{\epsilon}{n} H_1(x_n, y_n, \xi_n),$$

$$y_{n+1} - y_n = \frac{\epsilon}{n} H_2(x_n, y_n, \xi_n),$$

where $\{\xi_n\}_{n\geq 0}$ is a sequence of independently and identically distributed random variables uniformly distributed on $[-1, 1]$.

$$H_1(x, y, \xi) = (1 - (x^2 + y^2))x - yf(y) + \xi,$$

$$H_2(x, y, \xi) = (1 - (x^2 + y^2))y + xf(y) + \xi,$$

where $f(y) = y^2$. The phase portrait of the averaged ODE is given by Fig. 4.

We see that this system is convergent but admits $S^1 = \{(x, y) : x^2 + y^2 = 1\}$ as a cycle of equilibria. Therefore, the theorems of §§2 and 3 *are not sufficent to ensure the convergence of the process $\{x_n, y_n\}_{n\geq 0}$.*

The equilibria of this system are $e_1 = (0, 0)$, $e_2 = (1, 0)$, and $e_3 = (-1, 0)$. A simple computation shows that 0 *and* -2 *are the eigenvalues of the linearized ODE at points e_2 and e_3, and* 1 *is a double eigenvalue at point e_1. Thus,*

$$\beta(e_1) = 1,\ \beta(e_2) = \beta(e_3) = -2.$$

For $\gamma_n = \frac{\epsilon}{n}$ we have $l_\tau(\gamma) = -\frac{1}{\epsilon}$. Therefore, according to Corollary 6.2, *if $\epsilon < \frac{1}{4}$, the sequence $\{x_n, y_n\}_{n\geq 0}$ converges almost surely toward an equilibrium. Furthermore, a theorem of Pemantle* (1990) *can be used to show that this equilibrium cannot be the hyperbolic unstable equilibrium e_1.*

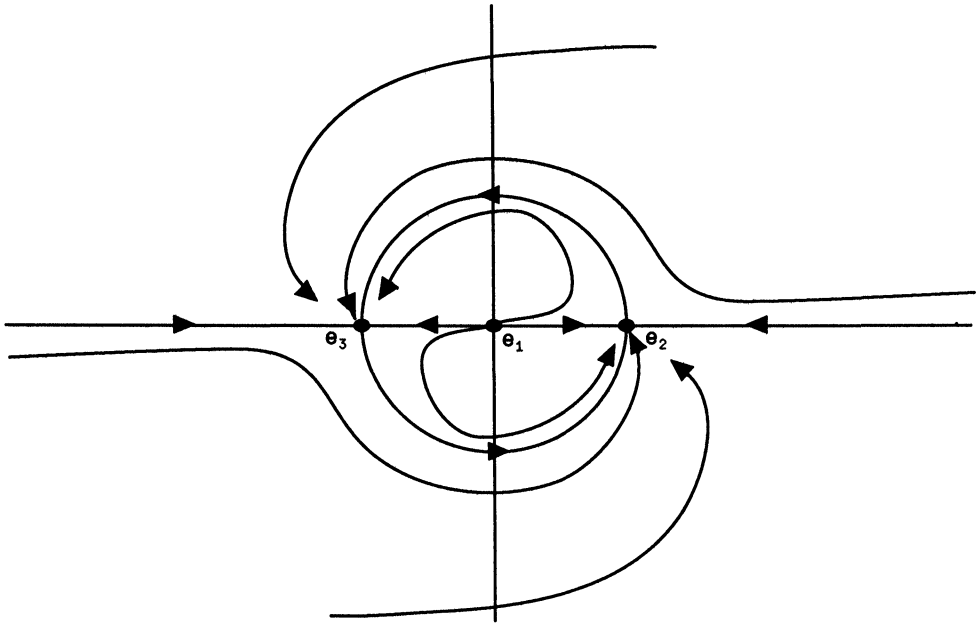

FIG. 4.

Globally convergent systems. A convergent system $\overline{H}$ with one unique equilibrium $\{e_1\}$ is said *globally convergent.*

The L^q estimate given by assertion (b) of Theorem 5.2 can be used to bound the L^q rate of convergence of algorithms associated to globally convergent ODEs. We mention here a corollary based on Proposition 5.4. Other estimates based on Proposition 5.6 or Theorem 5.2 are possible. Define

$$\rho(e_1) = \sup\{\mathrm{Re}(\lambda_1^j) : j = 1, \ldots, m\},$$

where $\{\lambda_1^j : j = 1, \ldots, m\}$ are the eigenvalues of $D\overline{H}(e_1)$.

Note that $\beta(e_1) \leq \rho(e_1) \leq 0$.

COROLLARY 6.4. *Let* $\{w_n\}_{n\geq 0}$ *be the solution to* (1). *Assume that conditions* A1–A4 *of Proposition* 5.4 *are satisfied and* M *is compact. Assume that the averaged vectorfield* $\overline{H}$ *defined by* A3 *is globally convergent and that* $l_\tau(\gamma) < 2\min(0, \beta(e_1))$. *Then*

$$l_\tau(||w_n - e_1||_q) \leq \rho(e_1).$$

Proof. Proposition 6.1 and (b) of Proposition 5.4 imply that $l_\tau(||w_n - \Phi_{\tau_n}(w')||_q) \leq \beta(e_1)$ for some random variable $w' \in M$. Since $\overline{H}$ is globally convergent and M is compact, we have the estimate $l_\tau(||\Phi_{\tau_n}(w') - e_1||) \leq \rho(e_1)$. Thus $l_\tau(||w_n - e_1||_q) \leq l_\tau(||\Phi_{\tau_n}(w') - e_1||_q + ||w_n - \Phi_{\tau_n}(w')||_q) \leq \sup(\rho(e_1), \beta(e_1)) = \rho(e_1)$. □

For $\gamma_n = \frac{\epsilon}{n+n_0}$ and $\epsilon < \frac{1}{2|\beta(e_1)|}$ this gives the following estimate: For all $\delta > 0$

there exists $n(\delta) \geq 0$ such that

$$\forall n \geq n(\delta),\ \|w_n - e_1\|_q \leq \frac{1}{n^{\epsilon(|\rho(e_1)|-\delta)}}.$$

This estimate can be compared with L^2 upper bounds given in Eweda and Macchi (1983) and Benveniste, Métivier, and Priouret (1990, Thms. 22 and 24, pp. 244, 246). It is slightly weaker but requires a weaker condition on the vectorfield.

Nonconvergent systems. For a general vectorfield $\overline{H}$ the log-expansion rate can be difficult to compute. The following proposition is useful to estimate it.

PROPOSITION 6.5 (Hirsch (1993)).

(a) *Let $\beta_s(w)$ be the smallest eigenvalue of the symmetric matrix*

$$\frac{1}{2}(D\overline{H}(w) + D\overline{H}(w)^T),$$

where ^{T} denotes the transpose operation. Then

$$l_{exp}(\overline{H}) \geq \min\{\beta_s(w) : w \in \operatorname{clos}(\mathcal{L}(\overline{H}))\}.$$

(b) *$l_{exp}(\overline{H})$ is invariant by C^1 change of coordinates.*

Assertion (a) is proved in Hirsch (1993). Assertion (b) is easy to check from the definition.

EXAMPLE 6.6. *Consider the stochastic approximation process defined in Example 6.3, where the function $f(.)$ which appears in the definition of H_1 and H_2 is now chosen to be the function $f(y) = 1$. The averaged ODE admits two internally chain-recurrent sets: the unstable equilibrium $0_{\mathbf{R}^2}$ and the stable limit cycle $S^1 = \{(x, y) : x^2 + y^2 = 1\}$. By a theorem of Pemantle already mentioned, $L(\{w_n\}_{n\geq 0})$ cannot be $0_{\mathbf{R}^2}$. Thus, according to Proposition 2.1, $L(\{w_n\}_{n\geq 0}) = S^1$.*

Note that this result is true for all values of $\epsilon > 0$. Let us now show how it can be sharpened by use of the log-expansion rate. To compute the log-expansion rate we use (b) of Proposition 6.5. In polar coordinates, the averaged ODE takes the simple form

$$\frac{d\rho}{dt} = \rho(1 - \rho^2), \qquad \frac{d\theta}{dt} = 1$$

from which it is easy to deduce that the log-expansion rate is given as $l_{exp}(\overline{H}) = \min\{-2, 1, 0\} = -2$. Let θ_n be the angular variable which measures the angle between the x-axis and the vector (x_n, y_n). If $\epsilon < \frac{1}{4}$, Proposition 5.4 applies and we deduce the "asymptotic phase property":

$$\lim_{n\to\infty} \theta_n - [\epsilon log(n)] \bmod 2\pi = \theta^*,$$

where θ^* is a random variable taking values in $[0, 2\pi]$.

7. Proof of Theorem 5.2.

LEMMA 7.1. *Under the assumptions of Theorem 5.2, the conditions of Theorem 1.2 are satisfied.*

Proof. Assumption A1 of Theorem 5.2 implies condition A1 of Theorem 1.2. Now check conditions A2 and A3. Let $\{n_j\}_{j\geq 0}$ be the sequence defined by $n_j = m(jT)$ for $j \geq 0$, where $m(.)$ is defined by (5). For any integer $n \in [n_j, n_{j+1}[$,

$$\|b_n\| \leq \sup_{\{k;\ 0\leq\tau_k-\tau_{n_j}\leq T\}} \|b_k\|.$$

Thus

$$E\left(\sup_{n\geq n_p}\|b_n\|^q\right)\leq E\left(\sup_{j\geq p}\sup_{\{k;\,0\leq\tau_k-\tau_{n_j}\leq T\}}\|b_k\|^q\right)\leq\sum_{j\geq p}\beta_{n_j}^{q,T}.$$

Now, assumption A2 of Theorem 1.2 implies

$$\beta_{n_j}^{q,T}\leq Ce^{-\lambda\tau_{n_j}}\leq C'e^{-(\lambda T)j}$$

for some constants $C, C', \lambda > 0$. It follows that

$$E\left(\sup_{n\geq n_p}\|b_n\|^q\right)\leq\sum_{j\geq p}C'e^{-(\lambda T)j}<+\infty$$

and the Cauchy criterion implies that condition A2 of Theorem 1.2 holds almost surely.

For A3, remark that for any integer $n\in[n_j,n_{j+1}[$,

$$\sum_{i=n_j}^{k-1}\gamma_iu_i=\sum_{i=n_j}^{n_{j+1}-1}\gamma_iu_i-\sum_{i=n_j}^{n-1}\gamma_iu_i+\sum_{i=n_{j+1}}^{k-1}\gamma_iu_i$$

with the convention $\sum_i^j=-\sum_j^i$. Working as previously, we deduce

$$E\left(\sup_{n\geq n_p}\sup_{\{k;\,0\leq\tau_k-\tau_n\leq T\}}\left\|\sum_{i=n}^{k-1}\gamma_iu_i\right\|^q\right)\leq 3^q\sum_{j\geq p}\alpha_{n_j}^{q,T}$$

and conclude exactly as for A2. □

The following definitions and theorem are due to Morris W. Hirsch. The main result of §5 will be derived from this theorem.

Let (E,d) be a metric space and $G: E\mapsto E$ be a map. Let $0\leq\lambda<1$. A sequence $\{Y_k\}_{k\geq 0}$ in E is called a *λ-pseudoorbit* for G if

$$\limsup_{k\to+\infty}d(G(Y_k);Y_{k+1})^{\frac{1}{k}}\leq\lambda.$$

A point $Z\in E$ is said to *λ-shadow* the sequence $\{Y_k\}_{k\geq 0}$ if

$$\limsup_{k\to+\infty}d(G^k(Z);Y_{k+m})^{\frac{1}{k}}\leq\lambda$$

for some integer m.

The following theorem is a consequence of Hirsch (1993, Thm. 3.2) (more precisely, a consequence of its proof).

THEOREM 7.2 (Hirsch (1993)). *Assume E is a complete metric space. Assume there exists $\rho_*>0$ and $\mu>0$ such that for all $0\leq\rho\leq\rho_*$*

$$\forall X\in E\ B(G(X),\rho\mu)\subset G(B(X,\rho)).$$

Let $\{Y_k\}_{k\geq 0}$ a λ-pseudoorbit *for G in E such that*

$$0<\lambda<\min(1,\mu).$$

Then

- a) *there exists* $Z \in E$ *which* λ*-shadows* $\{Y_k\}_{k\geq 0}$;
- b) *if* $Z, Z' \in E$ *both* λ*-shadow* $\{Y_k\}_{k\geq 0}$, *then there exists natural numbers* l, r *such that* $G^l(Z) = G^r(Z')$.

Now we prove Theorem 5.2. Consider the recursion

$$v_{n+1} - v_n = \gamma_n.(f(v_n)\overline{H}(v_n) + u_n + b_n), \tag{13}$$

where $\{u_n\}_{n\geq 0}$ and $\{b_n\}_{n\geq 0}$ are the sequences of recursion (12), v_0 is in $L^q(\Omega)$, and f is a smooth function which is 1 on a closed ball $\overline{B(0,r)}$ which contains $CR(\overline{H})$ and is zero outside $B(0, r+1)$.

The proof decomposes in two steps. The first step is to prove that, under the assumptions of Theorem 5.2, $\{v_n\}_{n\geq 0}$ is asymptotic to a trajectory solution to (2). The second step is to show that any trajectory solution to (12) is asymptotically a solution to (13).

Step 1. Let $\{\Psi\}_{t\in\mathbf{R}}$ be the flow of the vectorfield $f\overline{H}$. The set of all α and ω limit points for $f\overline{H}$ is the disjoint union of $\mathcal{L}(\overline{H})$ and $\{x \in \mathbf{R}^m; \|x\| \geq r+1\}$. Therefore,

$$l_{exp}(f\overline{H}) = \min(0, l_{exp}(\overline{H})).$$

Note $\nu = l_{exp}(f\overline{H})$. Assumption A2 of Theorem 5.2 allows us to choose two real numbers ν'', ν' such that $0 < \nu'' < \nu' < \nu$ and

$$\sup(l_\tau(\beta^{q,T}), l_\tau(\alpha^{q,T}), l_\tau(\gamma)) < \nu''. \tag{14}$$

Because $\nu' < \nu$, there exists $T > 0$ such that for all $x \in \mathbf{R}^m$ and all $t \geq T$

$$\|D\Psi_t(x)^{-1}\|^{-1} \geq e^{\nu' t}. \tag{15}$$

Let $E = L^q(\Omega)$ and G be the map defined by $G(X) = \Psi_T(X)$. First we note that G is well defined (G maps E into E). Indeed, for any random vector $X \in L^q(\Omega)$,

$$E(\|G(X)\|^q) = E(\|G(X)\|^q 1_{X\in B(0,r+1)}) + E(\|G(X)\|^q 1_{X\notin B(0,r+1)}).$$

The first term on the right of this equality is bounded by continuity of the flow and the second term is finite because $G(x) = x$ outside $B(0, r+1)$.

Let $\mu = e^{-\nu' T}$. Inequality (15) implies

$$\forall x, y \in \mathbf{R}^m \|\Psi_T^{-1}(y) - x\| \leq \mu\|y - \Psi_T(x)\|,$$

so, for all X, Y in E,

$$\|G^{-1}(Y) - X\|_q \leq \mu\|Y - G(X)\|_q.$$

Therefore,

$$B(G(X), \mu\rho) \subset G(B(X, \rho))$$

for any $\rho > 0$.

In order to apply Theorem 7.2, it remains to construct a λ-pseudoorbit for G in E with $0 < \lambda < \min(1, \mu)$. With this purpose let $V^0(.)$ be the interpolated process associated to the sequence $\{v_n\}_{n\geq 0}$ (see §4 for the definition of the interpolated process) and define $Y_n = V^0(nT)$ for all integers n. As v_0, u_n, and b_n are in $L^q(\Omega)$, a

simple induction shows that v_n is in $L^q(\Omega)$ for all n. Hence $\{Y_n\}_{n\geq 0}$ is a sequence of L^q.

As $\overline{H}$ is C^1, $f\overline{H}$ is Lipschitz and bounded. Let L denote a Lipschitz constant for $f\overline{H}$ and K denote a bound for $\|f\overline{H}(x)\|$. Lemma 4.4, applied to V^0 and $f\overline{H}$, gives

$$\|V^0((n+1)T) - \Psi_T(V^0(nT))\| \leq e^{L.T}[2\epsilon(nT,T)(1+TL) + TLK\gamma_{m(nT)}], \tag{16}$$

where $\epsilon(t,T)$ is defined as in Lemma 4.4.

Let $\lambda = e^{-\nu''T}$. From (14) and (16) $\{Y_n\}_{n\geq 0}$ is a λ-pseudoorbit for G. As $0 < \lambda < \min(1,\mu)$, the Theorem 7.2 applies. Thus, there exists $Z \in E$ such that for any $\lambda < \lambda_1 < \inf(1,\mu)$ and k large enough

$$\|G^k(Z) - Y_{k+m}\|_q \leq \lambda_1^k \tag{17}$$

for some integer m. The Borel–Cantelli lemma now implies

$$\lim_{k\to+\infty}(G^k(Z) - Y_{k+m}) = 0$$

almost surely. Let $Z' = \Psi_{-mT}(Z)$. We have

$$\lim_{k\to+\infty}(G^k(Z') - Y_k) = 0. \tag{18}$$

For any $t > 0$, write $t = kT + r$ with $k \in N$ and $0 \leq r < T$. Thus

$$\begin{aligned}\Psi_t(Z') - V^0(t) = [\Psi_r(\Psi_{kT}(Z')) - \Psi_r(V^0(kT))] \\ +[\Psi_r(V^0(kT)) - V^0(kT+r)].\end{aligned} \tag{19}$$

Uniform continuity of the flow on $[0,T]$ and relation (18) imply that the first term on the right of equality (19) goes to zero. The second term goes to zero by Lemma 4.2. Then

$$\lim_{t\to+\infty}\Psi_t(Z') - V^0(t) = 0$$

almost surely. This concludes step 1.

The equality (19) also proves part (b) of Theorem 5.2. Indeed, if $\{w_n\}_{n\geq 0}$ remains in a compact Q almost surely, the ball $B(0,r)$ can be chosen large enough to contain Q and $\{w_n\}_{n\geq 0}$ is solution to (13). The first term on the right side of equality (19) can be bounded in $L^q(\Omega)$, using (17) and the fact that $x \mapsto \Psi_r(x)$ is Lipschitz uniformly in $r \in [0,T]$. The second term can be bounded in $L^q(\Omega)$ by using (14) and (16). □

Step 2. Let x be a vector arbitrary chosen outside $B(0,r)$. For any integer k define

$$\Omega_k = \cap_{n\geq k}\{w_n \in B(0,r)\} \cap \{w_{k-1} \notin B(0,r)\}$$

with the convention $w_1 = x$. Let $\Omega' = \cup_{k\in N}\Omega_k$. Because $CR(\overline{H}) \subset B(0,r)$ and $L(\{w_n\}_{n\geq 0}) \subset CR(\overline{H})$ almost surely (Theorem 1.2), we have $P(\Omega') = 1$.

Let $\{v_n^k\}_{n\geq k}$ be the sequence solution to (13) defined by the initial condition $v_k^k = w_k$. As $w_k \in L^q(\Omega)$, $v_k^k \in L^q(\Omega)$. Therefore, we deduce from Step 1 the existence of a vector $Z_k \in L^q(\Omega)$ such that

$$\lim_{n\to+\infty}(v_n^k - \Psi_{\tau_n}(Z_k)) = 0$$

almost surely. Define $Z = \sum_{k\in N} 1_{\Omega_k} Z_k$. Because $v_n^k = w_n$ on Ω_k for $n \geq k$, we have

$$\lim_{n\to+\infty}(w_n - \Phi_{\tau_n}(Z)) = 0$$

almost surely. □

8. An application to neural network learning. In this section we show briefly how the previous result can be applied to prove the consistency of some "hybrid" learning rules recently proposed in the neural network area.

A feedforward neural network can be seen as a function $G : I \times W \mapsto O$ mapping the Cartesian product of an *input space* I and a *weight space* W into an *output space* O. The dimension of I and O are the number of input units and the number of output units, respectively. Without loss of generality we take $O \subset \mathbf{R}$, $I \subset \mathbf{R}^{d-1}$, and $W \subset \mathbf{R}^m$. The function G embodies the network architecture. Given an input $x \in I$ and a weight vector $w \in W$, the network's output is given as $G(x, w)$. At this level of description, the form of G is not of particular importance. We only assume that G is smooth enough. For more details and an in-depth presentation of feedforward neural networks in the framework of approximation theory we refer the reader to the excellent book by Halbert White and co-workers (1992).

The goal of learning is to adapt the weight vector w in such a way that the network realizes some specific relationship between the input space and the output space. This relationship is generally expressed by an "environmental" probability law μ defined on $I \times O$. A "training set" is a sequence $\{\xi_n\}_{n\geq 0} \subset I \times O$ asymptotically stationary with μ as limiting law. We let $\xi_n = (x_n, y_n)$, x_n is referred to as the "input vector" and y_n is referred to as the "desired output" or "target." A general learning rule for feedforward net can be written as (1). The gain γ_n is called the "learning rate" in the connectionist jargon.

The most popular example is the classical "backpropagation algorithm." Given a pair of input and target $\xi = (x, y)$ and weight w, the network error is given as $Er(w, \xi) = e(G(x, w), y)$ where $e : \mathbf{R} \times \mathbf{R} \mapsto \mathbf{R}^+$ is a smooth "error function" (usually, $e(o, y) = (y - o)^2$). The algorithm is given by (1) with $H(w, \xi) = -\nabla_w Er(w, \xi)$. Here $\nabla_w Er(w, \xi)$ denotes the gradient of the map $w \mapsto Er(w, \xi)$. Therefore, assuming that interchange of derivative and expectation is possible, the ODE associated with the backpropagation is a gradient vectorfield:

$$\frac{dw}{dt} = -\nabla \overline{Er}(w), \tag{20}$$

where

$$\overline{Er}(w) = \int Er(w, \xi)\mu(d\xi).$$

Convergence of the backpropagation can be analyzed by using classical results on stochastic gradients (see, e.g., Nevel'son and Has'minskii (1974). See also Benveniste, Métivier, and Priouret (1990, p. 91) for a presentation of the backpropagation as a stochastic gradient). It is also a direct application of Corollary 3.3 restated here for convenience.

PROPOSITION 8.1. *Let* $\{w_n\}_{n\geq 0}$ *be the solution to* (1). *Assume that the assumptions of Proposition* 2.1 *or* 2.2 *hold with* $\overline{H}$ *given by* (20). *Assume that critical points of* $\overline{Er}$ (*i.e., the zeros of* (20)) *are isolated. Then* $\{w_n\}_{n\geq 0}$ *converges almost surely toward a critical point of* $\overline{Er}$.

"Hybrid" learning rules have been considered by Moody and Darken (1989), Poggio and Girosi (1990), Nowlan (1990), Benaim and Tomasini (1991, 1992), and Benaim (1995c) among others for neural architectures with "nonsigmoid" units. The main idea of these algorithms is to train each layer according to different learning rules.

Consider, for simplicity, a single hidden-layer network. (Extension to multilayers is easy.) Formally, $G(x, w) = G_2(G_1(x, w_1), w_2)$, where $w = (w_1, w_2) \in \mathbf{R}^{m_1} \times \mathbf{R}^{m_2}$,

$m_1 + m_2 = m$, $G_1 : \mathbf{R}^{d-1} \times \mathbf{R}^{m_1} \mapsto \mathbf{R}^k$, $G_2 : \mathbf{R}^k \times \mathbf{R}^{m_2} \mapsto \mathbf{R}$. The integer k is the number of hidden units, G_1 embodies the architecture of the (input-layer, hidden-layer) subnet and G_2 the architecture of the (hidden-layer, output-unit) subnet. The subnet G_1 is trained according to an "unsupervised" learning rule (for example, a data clustering algorithm (Moody and Darken (1989)) or a maximum likelihood algorithm (Nowlan (1990), Benaim and Tomasini (1992)), and the subnet G_2 is trained according to a "supervised" algorithm (backpropagation). This leads to an ODE of the form

$$\frac{dw_1}{dt} = -\nabla \overline{E_1}(w_1), \tag{21}$$

$$\frac{dw_2}{dt} = -\nabla_{w_2} \overline{E_2}(w_1, w_2) \tag{22}$$

for some smooth functions $E_1 : \mathbf{R}^{m_1} \mapsto \mathbf{R}$, $E_2 : \mathbf{R}^{m_1} \times \mathbf{R}^{m_2} \mapsto \mathbf{R}$. Remark that such an ODE is not a gradient vectorfield. It is a *cascade* of gradients.

It is often assumed that the output unit is linear: $G_2(G_1, w_2) = \langle w_2, G_1 \rangle$, and the performance of the subnet G_2 is measured by the squared error function, $e(o, y) = (y - o)^2$. In that case the equation (22) has the particular form

$$\frac{dw_2}{dt} = -A(w_1)w_2 + B(w_1), \tag{23}$$

where $A(w_1)$ is the $k \times k$ matrix defined by

$$A(w_1) = \int G_1(x, w_1).G_1(x, w_1)^T \nu(dx)$$

with $\nu(.) = \int \mu(., dy)$ and $B(w_1)$ is the k-dimensional vector

$$B(w_1) = \int y G_1(x, w_1) \mu(dx, dy).$$

PROPOSITION 8.2. *Let $\{w_n\}_{n \geq 0}$ be the solution to* (1). *Assume that assumptions of Proposition* 2.1 *or* 2.2 *hold with $\overline{H}$ given by the system* (21), (23). *Assume that equilibria of* (21) *are isolated. Then $L(\{w_n\}_{n \geq 0})$ is almost surely a connected compact subset of the equilibria set of $\overline{H}$.*

Proof. Write $w_n = (w_{1,n}, w_{2,n}) \in \mathbf{R}^{m_1} \times \mathbf{R}^{m_2}$. Proposition 8.1 shows that $\{w_{1,n}\}_{n \geq 0}$ converges almost surely toward an equilibrium of (21), say, ${w_1}^*$. Thus, $L(\{w_n\}_{n \geq 0}) = \{{w_1}^*\} \times L'$ for some set $L' \subset \mathbf{R}^{m_2}$. According to Proposition 2.1 (or 2.2), L' is compact and invariant under the dynamics

$$\frac{dw_2}{dt} = A({w_1}^*)w_2 - B({w_1}^*). \tag{24}$$

Since $A({w_1}^*)$ is a symmetric matrix, any compact invariant set for (24) is contained in the equilibria set of (24). This concludes the proof. □

For the more general system (21), (22), we shall use the fact that $L(\{w_n\}_{n \geq 0})$ is *internally* chain-recurrent combined with Proposition 3.2.

PROPOSITION 8.3. *Let $\{w_n\}_{n \geq 0}$ be a solution to* (1). *Assume that assumptions of Proposition* 2.1 *or* 2.2 *hold with $\overline{H}$ given by the system* (21), (22). *Assume that equilibria of* (21) *and $\overline{H}$ are isolated. Then, $L(\{w_n\}_{n \geq 0})$ is almost surely a equilibrium of $\overline{H}$.*

Proof. We begin exactly as in the proof of Proposition 8.2. Write $w_n = (w_{1,n}, w_{2,n}) \in \mathbf{R}^{m_1} \times \mathbf{R}^{m_2}$. Using Proposition 8.1 we see that $L(\{w_n\}_{n\geq 0}) = \{w_1{}^*\} \times L'$, where $w_1{}^*$ is an equilibrium of (21) and $L' \subset \mathbf{R}^{m_2}$ is a compact connected set invariant under the dynamics of

$$\frac{dw_2}{dt} = -\nabla \overline{E_2}(w_1^*, w_2). \tag{25}$$

Since $L(\{w_n\}_{n\geq 0})$ is *internally* chain-recurrent, every point of L' has to be chain-recurrent for the flow induced by (25). Since (25) is a gradient vectorfield with isolated equilibria, it follows from Proposition 3.2 that L' consists of equilibria. By connectedness, L' is an equilibrium of (25). □

9. Appendix.

Proof of Proposition 2.2. We denote by $\mathcal{F}_n^m$ the σ field generated by $\{\xi_i; n \leq i \leq m\}$ for $m \geq n \geq 0$ and let $\mathcal{F}_0^n = \{\emptyset, \Omega\}$ for $n < 0$.

DEFINITION 9.1. *Let $\{X_n\}_{n\geq 0}$ be a sequence of random variables belonging to $L^2(\Omega)$. $\{X_n\}_{n\geq 0}$ is said to be a mixingale process if there are sequences of finite nonnegative constants $\{c_n\}_{n\geq 0}$ and $\{\psi_m\}_{m\geq 0}$, where $\lim_{n\to\infty} \psi_m = 0$, such that for all $n \geq 1$ and $m \geq 0$*

- a) $\|E(X_n/\mathcal{F}_0^{n-m})\|_2 \leq c_n.\psi_m$,
- b) $\|X_n - E(X_n/\mathcal{F}_0^{n+m})\|_2 \leq c_n.\psi_{m+1}$.

Throughout this section we will only consider sequence of random variables $\{X_n\}_{n\geq 0}$ such that each X_n is measurable $\mathcal{F}_0^n$ so that condition b) holds automatically.

The following lemma relates the concept of mixing process to that of mixingale. It is due to McLeish (1975, Lem. 2.1).

LEMMA 9.2 (McLeish (1975)). *Suppose that $\{\xi_n\}_{n\geq 0}$ is a ϕ mixing (respectively, α mixing) process. Let $\{X_n\}_{n\geq 0}$ be a sequence of random variables such that each X_n is measurable $\mathcal{F}_0^n$ and $E(X_n) = 0$.*

Then, for $2 \leq r \leq +\infty$, $n, m \geq 0$,

- a) $\|E(X_n/\mathcal{F}_0^{n-m}\|_2 \leq 2\phi_m^{1-1/r}\|X_n\|_r$,
- b) $\|E(X_n/\mathcal{F}_0^{n-m}\|_2 \leq 2(1+\sqrt{2})\alpha_m^{1/2-1/r}\|X_n\|_r$.

REMARK. *It follows from this lemma that $\{X_n\}_{n\geq 0}$ is a mixingale with $c_n = \|X_n\|_r$ and $\psi_m = 2\phi_m^{1-1/r}$ in the ϕ mixing case or $\psi_m = 2(1+\sqrt{2})\alpha_m^{1/2-1/r}$ in the α mixing case.*

The following lemma is the main result of this section. The proof of the lemma follows closely the proof of McLeish's Theorem 1.6 (1975), but instead of using Doob's inequality for martingales (as McLeish does) we use the Burkholder inequality together with the ideas involved in the proof of Métivier and Priouret's Proposition 8 (1987). Note that for $q = 2$ the lemma is a direct consequence of McLeish's Theorem 1.6 (1975).

LEMMA 9.3. *Let $\{X_n\}_{n\geq 0}$ be a sequence of real random variables. Let $S_n = \sum_{k=0}^n \gamma_k.X_k$. We assume that*

- *X_n is measurable $\mathcal{F}_0^n$;*
- *$X_n \in L^q(\Omega)$ for some $q \geq 2$;*
- *$\{X_n\}_{n\geq 0}$ is a mixingale with sequences $\{c_n\}_{n\geq 0}$ and $\{\psi_m\}_{m\geq 0}$.*

We assume that there exists a positive decreasing sequence $\{a_n\}_{n\geq -1}$ such that

- $\sum_{k\geq 0} \psi_k^2(a_k^{-1} - a_{k-1}^{-1}) < +\infty$;
- $\sum_{k\geq 0} a_k^{1/(q-1)} < +\infty$.

Then

$$E\left(\sup_{m\le n} \mid S_m \mid^q\right) \le \tau_{n+1}^{q/2-1} D(q) \sum_{i=0}^{n} \gamma_i^{1+q/2} c_i^2 \|X_i^{q-2}\|_\infty,$$

where $D(q) > 0$ *is a constant.*

Proof. Let $Z_{i,k} = E(X_i/\mathcal{F}_0^{i-k}) - E(X_i/\mathcal{F}_0^{i-k-1})$ for $i \ge 0, k \ge 0$. Let $Y_{n,k} = \sum_{i=0}^{n} \gamma_i Z_{i,k}$. Since $E(X_i/\mathcal{F}_0^i) = X_i$ and $E(X_i/\mathcal{F}_0^{i-k}) = E(X_i) = 0$ for $k > i$, it is clear that

$$X_i = \sum_{k=0}^{i} Z_{i,k} = \sum_{k\ge 0} Z_{i,k}.$$

Thus $S_n = \sum_{k\ge 0} Y_{n,k}$ and by Hölder's inequality we have

$$\mid S_n \mid^q \le \sum_{k\ge 0} \frac{\mid Y_{n,k} \mid^q}{a_k} \left(\sum_{k\ge 0} a_k{}^l\right)^{1/l}, \tag{26}$$

where $l = \frac{1}{q-1}$.

Observe that $E(Z_{n,k}/\mathcal{F}_0^{n-k}) = 0$. So by Burkholder and Rosenthal's inequalities (Hall and Heyde (1980, Thms. 2.10–2.12)):

$$E\left(\sup_{m\le n} \mid Y_{m,k} \mid^q\right) \le C(q) E\left(\left(\sum_{i=0}^{n} \gamma_i^2 Z_{i,k}^2\right)^{q/2}\right). \tag{27}$$

Now, exactly as in Métivier and Priouret (1987), we shall apply the following form of Hölder's inequality:

$$\left(\sum_i \mid \alpha_i \beta_i \mid\right)^u \le \left(\sum_i \alpha_i^{\delta u/(u-1)}\right)^{u-1} \left(\sum_i \alpha_i^{(1-\delta)u} \mid \beta_i \mid^u\right) \tag{28}$$

for $\alpha_i \ge 0, \beta_i \in \mathbf{R}, u > 1, 0 < \delta < 1$.

Applying (28) to (27)) with $\alpha_i = \gamma_i^2, \beta_i = \mid Z_{i,k} \mid^2, u = \frac{q}{2}$, and $\delta = \frac{q-2}{2q}$ we obtain

$$E\left(\sup_{m\le n} \mid Y_{m,k} \mid^q\right) \le C(q) \tau_{n+1}^{(q/2-1)} \sum_{i=0}^{n} E(\mid Z_{i,k} \mid^q) \gamma_i^{1+q/2}, \tag{29}$$

and by using (26), we deduce

$$E\left(\sup_{m\le n} \mid S_m \mid^q\right) \le \sum_{k\ge 0} (a_k^l)^{1/l} C(q) \tau_{n+1}^{(q/2-1)} \sum_{k\ge 0} a_k^{-1} \sum_{i=0}^{n} E(\mid Z_{i,k} \mid^q) \gamma_i^{1+q/2}. \tag{30}$$

On the other hand we have

$$E(\mid Z_{i,k} \mid^q) \le \|Z_{i,k}^{q-2}\|_\infty E(\mid Z_{i,k} \mid^2),$$

and from the Pythagorean theorem in $L^2(\Omega)$,

$$E(\mid Z_{i,k} \mid^2) = \|E(X_i/\mathcal{F}_0^{i-k})\|_2^2 - \|E(X_i/\mathcal{F}_0^{i-k-1})\|_2^2.$$

It follows that

$$\sum_{k\geq 0} a_k^{-1} \sum_{i=0}^{n} E(\mid Z_{i,k} \mid^q)\gamma_i^{1+q/2} \leq$$

$$\sum_{k\geq 0} \psi_k^2(a_k^{-1} - a_{k-1}^{-1}).\sum_{i=0}^{n} \gamma_i^{1+q/2} c_i^2 \|X_i^{q-2}\|_\infty. \tag{31}$$

Putting together (30) and (31) and letting $D(q) = \sum_{k\geq 0}(a_k^l)^{1/l} C(q) \sum_{k\geq 0} \psi_k^2(a_k^{-1} - a_{k-1}^{-1})$ conclude the proof. □

From this lemma we shall deduce the following lemma, which is analogous to Corollary 11 of Métivier and Priouret (1987).

LEMMA 9.4. *Let* $\{X_n\}_{n\geq 0}$ *be as in Lemma* 9.3. *Then*

- a) *for all* $T \geq 0$,

$$E\left(\sup_{\tau_n \geq \tau_p}\left(\sup_{0\leq \tau_k - \tau_n \leq T}\left|\sum_{i=n}^{k-1} \gamma_i X_i\right|^q\right)\right) \leq T^{q/2-1} D(q) \sum_{i\geq p} \gamma_i^{1+q/2} c_i^2 \|X_i^{q-2}\|_\infty;$$

- b) *if* $\sum_{i\geq 0} \gamma_i^{1+q/2} c_i^2 \|X_i^{q-2}\|_\infty < +\infty$, *then*

$$\lim_{n\to\infty}\left(\sup_{\{0\leq \tau_k - \tau_n \leq T\}}\left|\sum_{i=n}^{k-1} \gamma_i X_i\right|\right) = 0$$

with probability one.

The proof is exactly the same as the proof of Corollary 11 in Métivier and Priouret (1987).

COROLLARY 9.5. *Suppose that* $\{\xi_n\}_{n\geq 0}$ *is a* ϕ *mixing (respectively,* α *mixing) process. Let* $\{X_n\}_{n\geq 0}$ *be a sequence of random variables such that*

- *each* X_n *is measurable* $\mathcal{F}_0^n$ *and* $E(X_n) = 0$;
- $\sup_n \|X_n\|_r < +\infty$ *for some* $r \in [0, +\infty]$;
- *if* $r < \infty$, $\phi_n = O(\frac{1}{n^\beta})$ *for some* $\beta > \frac{r}{2r-2}$ *(respectively,* $\alpha_n = O(\frac{1}{n^\beta})$ *for some* $\beta > \frac{r}{r-2}$*) and* $\sum_{n=0}^{\infty} \gamma_n^2 < +\infty$;
- *if* $r = \infty$, $\phi_n = O(\frac{1}{n^\beta})$ *for some* $\beta > \frac{1}{2}$ *(respectively,* $\alpha_n = O(\frac{1}{n^\beta})$ *for some* $\beta > 1$*) and* $\sum_{n=0}^{\infty} \gamma_n^{1+q/2} < +\infty$ *for some* $q \in [2, 2\beta + 1[$.

Then

- a) $\lim_{n\to\infty}(\sup_{k;\, 0\leq \tau_k - \tau_n \leq T} \mid \sum_{i=n}^{k-1} \gamma_i X_i \mid) = 0$;
- b) *there exists a constant* $D(q, r)$ *such that*

$$E\left(\sup_{k;\, 0\leq \tau_k - \tau_n \leq T}\left|\sum_{i=n}^{k-1} \gamma_i X_i\right|^q\right) \leq T^{q/2-1} D(q, r) \sum_{i=n}^{m(\tau_n + T)-1} \gamma_i^{1+q/2},$$

where $q = 2$ *for* $r < +\infty$ *and* $q \in [2, 2\beta + 1[$ *for* $r = +\infty$.

Proof. The proof follows from Lemma 9.2, the remark which follows Lemma 9.2, and Lemma 9.3 for part (a) and the Lemma 9.4 for part (b). □

We are now able to prove Proposition 2.2. Let $\{w_n\}_{n\geq 0}$ be a solution to (1). Let

$$u_n(w) = H(w, \xi_n) - E(H(w, \xi_n)),$$

$$u_n = u_n(w_n),$$

$$b_n(w) = E(H(w, \xi_n)) - \overline{H}(w),$$

$$b_n = b_n(w_n).$$

We suppose that conditions A1 to A6 of Proposition 2.2 hold. Let $K = \sup_n \|k(\xi_n)\|_r$. We remark that $\overline{H}$ is K Lipschitz. Indeed, for any x, y,

$$\|\overline{H}(x) - \overline{H}(y)\| = \lim_{n\to\infty} \|E(H(x, \xi_n) - H(y, \xi_n))\|$$

but

$$\|E(H(x, \xi_n) - H(y, \xi_n))\| \leq E(k(\xi_n))\|x - y\| \leq K\|x - y\|$$

by application of Jensen's inequality. Using the same kind of argument we see that

$$\|b_n(x) - b_n(y)\| \leq 2K\|x - y\|, \tag{32}$$

$$\|u_n(x) - u_n(y)\| \leq K\|x - y\| + k(\xi_n)\|x - y\|. \tag{33}$$

Let us now verify the conditions A1, A2, and A3 of Theorem 1.2. A1 is true by assumption (assumption A2 of Proposition 2.2). The inequality (33) shows that the sequence $\{b_n(.)\}_{n\geq 0}$ is equicontinuous, and assumption A3 of the Proposition 2.2 means that $\{b_n(w)\}_{n\geq 0}$ converges to zero for any $w \in \mathbf{R}^m$. It follows that $\{b_n(.)\}_{n\geq 0}$ converges to zero uniformly on any compact set of $\mathbf{R}^m$, and as $\{w_n\}_{n\geq 0}$ is assumed to be bounded, $\{b_n\}_{n\geq 0}$ converges to zero. This proves assumption A2 of Theorem 1.2.

We now check the condition A3 of Theorem 1.2. Let $T \geq 0$ and let

$$S_n^T(w) = \sup_{\{k,\ 0\leq\tau_k-\tau_n\leq T\}} \left| \sum_{i=n}^{k} \gamma_i u_i(w) \right|,$$

$$Z_n^T = \sup_{\{k,\ 0\leq\tau_k-\tau_n\leq T\}} \sum_{i=n}^{k} \gamma_i k(\xi_i).$$

From inequalities (32) and (33) it follows that for any $w, w' \in \mathbf{R}^d$

$$\left| \sum_{i=n}^{k} \gamma_i u_i(w) \right| \leq \left| \sum_{i=n}^{k} \gamma_i u_i(w') \right| + K|w - w'| \sum_{i=n}^{k} \gamma_i + |w - w'| \sum_{i=n}^{k} \gamma_i k(\xi_i).$$

Thus

$$0 \leq S_n^T(w) \leq S_n^T(w') + TK|w - w'| + Z_n^T|w - w'|. \tag{34}$$

Lemma 9.4(b) shows that under the assumptions A1 and A6, A6′ of Proposition 2.2, $\lim_{n\to\infty} Z_n^T = 0$ and $\lim_{n\to\infty} S_n^T(w) = 0$ for any $w \in \mathbf{R}^m$. From inequality (34) it is easy to see that $\{S_n^T(w)\}_{n\geq 0}$ converges to zero uniformly on any compact set of $\mathbf{R}^m$ so that $\lim_{n\to\infty} S_n^T(w_n) = 0$. □

Proof of Proposition 5.6. Let Q be a compact set. Let $\epsilon_n > 0$ such that $\lim_{n\to+\infty} \epsilon_n = 0$ and let $\{B(x_i, \epsilon_n)\}_{i\in I_n}$ be a finite cover of Q by balls of radius ϵ_n. From (34), we deduce

$$\sup_{w\in Q} S_n^T(w) \leq \sum_{i\in I_n} S_n^T(x_i) + KT\epsilon_n + \epsilon_n Z_n^T.$$

Therefore, Corollary 9.5(b) implies

$$\left\| \sup_{w\in Q} S_n^T(w) \right\|_q \leq (A.\#(I_n) + B\epsilon_n) \left(\sum_{i=n}^{m(\tau_n+T)-1} \gamma_i^{1+q/2} \right)^{1/q} + KT\epsilon_n \tag{35}$$

for some constants A and B depending on $Q, q, T, \overline{H}$, and $\#(I_n)$ denotes the cardinal of I_n.

Since Q is a compact subset of $\mathbf{R}^m$, the family $\{x_i\}_{i\geq 0}$ can be chosen such that $\#(I_n) \leq C.\epsilon_n^{-m}$ for some $C > 0$. Noting that

$$\sum_{i=n}^{m(\tau_n+T)-1} \gamma_i^{1+q/2} \leq T\gamma_n^{q/2},$$

inequality (35) gives

$$\left\| \sup_{w\in Q} S_n^T(w) \right\|_q \leq (A.C.\epsilon_n^{-m} + B\epsilon_n)T\gamma_n^{q/2} + KT\epsilon_n.$$

Therefore, with α_q^T as defined in §5,

$$\begin{aligned} l_\tau(\alpha^{q,T}) &\leq \sup\left(\frac{1}{2}l_\tau(\gamma) - ml_\tau(\epsilon), l_\tau(\epsilon), \frac{1}{2}l_\tau(\gamma) + l_\tau(\epsilon)\right) \\ &\leq \sup\left(\frac{1}{2}l_\tau(\gamma) - ml_\tau(\epsilon), l_\tau(\epsilon)\right). \end{aligned} \tag{36}$$

Now we choose a sequence $\{\epsilon_n\}_{n\geq 0}$ such that

$$l_\tau(\epsilon) = \min_{\{l<0\}} \sup\left(l, \frac{1}{2}l_\tau(\gamma) - ml\right).$$

This easily gives

$$l_\tau(\epsilon) = \frac{l_\tau(\gamma)}{2(1+m)}.$$

For example, one may choose $\epsilon_n = \gamma_n^{1/(2+2m)}$.

Acknowledgment. I am particulary grateful to Morris W. Hirsch for many comments and fruitful discussions on the topics in this paper. A large part of this work makes a crucial use of some of his ideas and results on "shadowing."

REFERENCES

M. BENAIM (1994a), *Un théorème de Poincaré–Bendixson pour une classe d'equations différentielles stochastiques*, C. R. Acad. Sci. Paris Sér. I, 318, pp. 837–839.

——— (1994b), *On functional approximation with normalized Gaussian units*, Neural Computation, 6, pp. 319–333.

M. BENAIM AND M. W. HIRSCH (1995a), *Dynamics of Morse–Smale urns processes*, Ergodic Theory Dynamical Systems, to appear.

——— (1995b), *Asymptotic pseudo-trajectories, chain-recurrent flows and stochastic approximations*, J. Dynamics Differential Equations, to appear.

——— (1995c), *Chain recurrence in surface flows*, Discrete and Continuous Dynamics Systems, 1 (1995), pp. 1–17.

M. BENAIM AND L. TOMASINI (1991), *Competitive and self-organizing algorithms based on the minimization of an information criterion*, in Artificial Neural Networks I, Vol. 1, T. Kohonen, ed., North-Holland, Amsterdam, pp. 391–396.

——— (1992), *Approximating function and predicting time series with multisigmoidal basis functions*, in Artificial Neural Networks II, Vol. 1, I. Aleksander and J. Taylor, eds., Elsevier, Amsterdam, pp. 407–411.

A. BENVENISTE, M. MÉTIVIER, AND P. PRIOURET (1990), *Stochastic Approximations and Adaptive Algorithms*, Springer-Verlag, Berlin, Heidelberg, New York. Translated from *Algorithmes adaptatifs et approximations stochastiques*, Masson, Paris, 1987.

P. BILLINGSLEY (1968), *Convergence of Probability Measures*, Wiley, London, New York.

C. C. CONLEY (1978), *Isolated Invariant Sets and the Morse Index*, Regional Conference Series in Mathematics 38, American Mathematical Society, Providence, RI.

E. EWEDA AND O. MACCHI (1983), *Quadratic mean and almost sure convergence of unbounded stochastic approximation algorithms with correlated observations*, Ann. Inst. H. Poincaré, 19, pp. 235–255.

J. C. FORT AND G. PAGÈS (1994), *Convergences d'algorithmes stochastiques: le théorème de Kushner et Clark revisité*, pré-pub. labo. de proba., Univ. Paris 6.

E. G. GLADYSHEV (1965), *On stochastic approximation*, Theory Probab. Appl., 10, pp. 275–278.

D. HALL AND C. C. HEYDE (1980), *Martingale Limit Theory and Applications*, Academic Press, New York.

M. W. HIRSCH (1993), *Asymptotic phase, shadowing and reaction-diffusion systems*, in Control Theory, Dynamical Systems and Geometry of Dynamics, K. D. Elworthy, W. N. Everitt, and E. B. Lee, eds., Marcel Dekker, New York.

M. W. HIRSCH AND C. PUGH (1970), *Stable manifolds for hyperbolic sets*, in Global Analysis, Proceedings of Symposia in Pure Mathematics 14, American Mathematical Society, Providence, RI, pp. 133–163.

J. KIEFER AND J WOLFOWITZ (1952), *Stochastic estimation of the maximum of a regression function*, Ann. Math. Statist., 23, pp. 462–466.

C. M. KUAN AND H. WHITE (1992), *Artificial neural networks: An econometric perspective*, Econometric Reviews, to be published.

H. J. KUSHNER AND D. S. CLARK (1978), *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer-Verlag, Berlin, Heidelberg, New York.

L. LJUNG (1977), *Analysis of recursive stochastic algorithms*, IEEE Trans. Automat. Control, AC-22, pp. 551–575.

D. L. MCLEISH (1975), *A maximal inequality and dependent strong laws*, Ann. Probab., 3, pp. 829–839

M. MÉTIVIER AND P. PRIOURET (1984), *Application of a Kushner and Clark lemma to general classes of stochastic algorithms*, IEEE Trans. Inform. Theory, IT-30, pp. 140–150

——— (1987), *Théorèmes de convergence presque sure pour une classe d'algorithmes stochastiques à pas décroissant*, Probab. Theory Related Fields, 74, pp. 403–428.

J. MOODY AND C. DARKEN (1989), *Fast learning in networks of locally-tuned processing units*, Neural Computation, 1, pp. 281–582.

M. B. NEVEL'SON AND R. Z. HAS'MINSKII (1974), *Stochastic Approximation and Recursive Estimation*, Translation of Math. Monographs 47, American Mathematical Society, Providence, RI.

J. NEVEU (1964), *Bases mathematiques du calcul des probabilites*, Masson, Paris.

S. NOWLAN (1990), *Maximum likelihood competitive learning*, in Proceedings of Neural Information Processing Systems, pp. 574–582.

J. PALIS (1969), *On Morse–Smale dynamical systems*, Topology, 8, pp. 385–405.

R. PEMANTLE (1990), *Nonconvergence to unstable points in urn models and stochastic approximations*, Ann. Probab., 18, pp. 698–712.

T. POGGIO AND F. GIROSI (1990), *Regularization algorithms for learning that are equivalent to multilayer networks*, Science, 247, pp. 979–982.
H. ROBBINS AND S. MONRO (1951), *A stochastic approximation method*, Ann. Math. Statist., 22, pp. 400–407.
C. ROBINSON (1977), *Stability theorems and hyperbolicity in dynamical systems*, Rocky J. Math., 7, pp. 425–434.
M. SHUB (1986), *Global Stability of Dynamical Systems*, Springer-Verlag, New York.
H. WHITE (1992), *Artificial Neural Networks, Approximation and Learning Theory*, Blackwell, Oxford, Cambridge.