J Nonlinear Sci (2012) 22:169–185
DOI 10.1007/s00332-011-9112-3

Journal of Nonlinear Science

Nonautonomous Stochastic Search in Global Optimization

Jerzy Ombach · Dawid Tarłowski

Received: 6 January 2011 / Accepted: 29 November 2011 / Published online: 29 December 2011

Abstract We present a general method how to prove convergence of a sequence of random variables generated by a nonautonomous scheme of the form $X_t = T_t(X_{t-1}, Y_t)$, where Y_t represents randomness, used as an approximation of the set of solutions of the global optimization problem with a continuous cost function. We show some of its applications.

Keywords Global optimization · Stochastic algorithm · Random search · Foias operator · Lyapunov function

Mathematics Subject Classification (2000) 90C15 · 60J20

One of the most common problems in applied mathematics is finding approximations of optimal solutions of many specific cases. There exist a lot of numerical optimization procedures, and recent years have witnessed an explosion of heuristic stochastic algorithms. While the performance of some of these at specific instances looks good and is experimentally confirmed, theoretical background definitely lags behind. This is one of a few papers suggesting a theoretical framework to handle the problem.

Communicated by A. Stuart.

J. Ombach (✉) · D. Tarłowski
Instytut Matematyki, Wydział Matematyki i Informatyki, Uniwersytet Jagielloński,
ul. Łojasiewicza 6, 30 348 Kraków, Poland
e-mail: ombach@im.uj.edu.pl

D. Tarłowski
e-mail: dawid.tarlowski@im.uj.edu.pl

Many iterative optimization schemes can, in general, be described by an autonomous difference equation $x_t = T(x_{t-1})$, where x_t are successive approximation of the solution and T represents the method used. If we use a stochastic scheme, then the above equation takes another form: $x_t = T(x_{t-1}, y_t)$, where y_t represents some random factor. The simplest example is the pure random search (PRS): one draws a point y_t and if it is better (in the sense of the cost function) than x_{t-1}, then set $x_t := y_t$; otherwise $x_t := x_{t-1}$ and t is increased. In Ombach (2007, 2008) we established some general sufficient conditions for stochastic convergence of the autonomous stochastic scheme for the global optimization problem and we discussed some special cases. The essential assumption there was that $\int f(T(x, y))\,\mathrm{d}y < f(x)$, where f is the cost function to be minimized (note that $\mathrm{d}y$ means integration with respect to some probability measure).

Still, in more advanced schemes, the operator T itself is changing in time and we get a nonautonomous difference equation, $x_t = T_t(x_{t-1}, y_t)$. For example, the method T may depend on some parameters that can be adjusted in time. Also, the distributions used for generating points y_t may change. Besides, the above strong inequality is not always easy to verify or just does not hold in some cases, while the weak inequality $\int f(T(x, y))\,\mathrm{d}y \leq f(x)$ is an immediate consequence of the inequality $f(T(x, y)) \leq f(x)$, which is natural in many instances. A good illustration of such a case is the accelerated random search (ARS), established in Appel et al. (2003), or the grenade explosion method (GEM), established in Ahrari and Atai (2010), Ahrari et al. (2009).

We mention that the specific case, when the distribution for generating points y_t is constant over time but the methods T belong to a finite set and are changed cyclically, was analyzed in Radwański (2007). In this paper we extend the results of Ombach (2007, 2008), Radwański (2007) to cover some of the situations mentioned above. However, the main tool remains the same. We express the problem of convergence in terms of a sequence of Foias operators defined on the space of measures and apply an appropriate Lyapunov function to it. We believe that this approach may be useful in further proofs of convergence of various stochastic schemes, like particle swarm optimization (PS0) see Poli et al. (2007) for example, the simulated annealing algorithm (SA) (Yang 2000), or in the study of the convergence rates of stochastic optimization schemes. Foias operators have proved to be very useful in the theory of iterated function systems (IFS), see Lasota and Mackey (1994) and references in there, and we believe that they are also a good framework for examining stochastic optimization algorithms.

Since most stochastic search methods actually result in non-homogeneous Markov Chains, we just note that the Markov Chains and the Markov Operators have been already widely used while explaining some optimization procedures or for studying random iterative functions (iterated function systems, IFS) in various contexts, including fractals. We refer to Borovkov and Yurinsky (1998) and Meyn and Twedie (1993) for an extensive review of Markov Chains and Processes and their applications in exploring processes generated by autonomous equations of the form $x_t = T_t(x_{t-1}, y_t)$. In Lasota and Mackey (1994) the authors used the Foias Operator for studying IFS and fractals. More direct use of Markov Chains in fractals and other applications might be found in Diaconis and Freedman (1999). As an optimization tool, nonautonomous Markov Chains are discussed in Ljung et al. (1992). The above references discuss problems that in fact reduce to the classical question: how to prove the existence and stability of the unique stationary state. That goal is achieved by various methods, including the Lyapunov function method. We, however, consider a different situation. Namely, optimization problems naturally led to Markov operators for which we want to prove the existence of an attractor, which is uncountable when the cost function attains its global minimum at many points.

Our main result, Theorem 1, is established in Sect. 1. In Sects. 2 and 3 we prepare necessary tools from the theory of dynamical systems and the theory of Foias operators. Section 4 consists of the proof of Theorem 1. In Sect. 5 we apply Theorem 1 to a simple example and we mention the possibility of repeating the proof for some more advanced algorithms with a similar mechanism, like GEM. Section 6 shows an application of Theorem 1 to establish general criteria for the convergence of a broad family of stochastic algorithms. In Sect. 7 we mention the possibility of the IFS-type approach to the minimization problem, as it seems to fit nicely into our framework.

1

Let (A, d) be a compact metric space and let $f : A \longrightarrow \mathbb{R}$ be a continuous function. Let $A^\star \subset A$ be the set of all the solutions of the global minimization problem, i.e.

$$A^\star = \arg\min f = \big\{a \in A : f(a) \le f(x), \text{ for all } x \in A\big\}.$$

A vast amount of stochastic algorithms used for finding a solution of the global optimization problem have the following global form:

$$X_t = T_t(X_{t-1}, Y_t), \quad \text{for } t = 1, 2, 3, \dots, \tag{1}$$

where $T_t : A \times B \longrightarrow A$ are operators chosen from an available class of operators identifying the algorithm and Y_t are random vectors. We are interested in convergence of X_t to the set $A^\star$ of solutions of the global optimization problem.

Remark 1 Often, in practice there is no real need for finding an exact solution from $A^\star$, but rather one needs to find some x such that $f(x) \le \alpha$, where $\alpha > \min f$ is known. This is, nevertheless, equivalent to the problem of finding a global minimum of the continuous function f_α, defined by $f_\alpha(x) = \max(f(x), \alpha)$.

Let $(\Omega, \Sigma, \mathrm{Prob})$ be a probability space. Let μ_0 be a probability measure defined on $\mathcal{B}(A)$—the family of Borel subsets of the space A—and let $X_0 : \Omega \longrightarrow A$ be a random variable with distribution μ_0. Let B be a Polish (separable and complete) metric space and $\mathbf{Y}_t : \Omega \longrightarrow B$ be a sequence of independent random variables distributed according to some distributions ν_t. If the T_t are measurable, then the X_t, defined by (1), are random variables.

Let $\mathcal{T}$ be a family of all measurable operators $T\colon\ A \times B \longrightarrow A$ equipped with the uniform convergence topology. For any $T \in \mathcal{T}$ let $D_T \subset A \times B$ be the set of all discontinuities of T. Let M denote the space of all Borel probability measures defined on A and let N denote the family of all Borel probability measures on B, both equipped with the weak convergence topology; see Sect. 3 for the details. We consider the space $\mathcal{T} \times N$ as equipped with the product topology.

As usual, for $\mu \in M$ and $\nu \in N$, $\mu \times \nu$ denotes the Cartesian product of measures μ and ν, which is uniquely characterized by $(\mu \times \nu)(C \times D) = \mu(C) \cdot \nu(D)$, for all $C \in \mathcal{B}(\mathcal{A})$, $D \in \mathcal{B}(B)$. By $B(a, r)$ we denote a closed ball of radius r centered in a. If P_t is a sequence of Borel probability measures on some metric space, then the weak convergence of the sequence to some Borel probability measure P will be simply denoted by $P_t \longrightarrow P$.

The following theorem, to be proved in Sects. 2–4, provides general sufficient conditions for the stochastic convergence of X_t to the set $A^\star$.

Theorem 1 *Let $U \subset \mathcal{T} \times N$ be a compact set. Assume that for any $u = (T, \nu) \in U$:*

(A) *For any $x_0 \in A$, there is a Borel set $D_T(x_0) \subset B$ with $\nu(D_T(x_0)) = 0$, such that T is continuous in (x_0, y), for any $y \notin D_T(x_0)$.*
(B) *For any $x \in A^\star$ and $y \in B$, $T(x, y) \in A^\star$.*
(C1) *For any $x \in A \setminus A^\star$:*

$$\int_B f\big(T(x, y)\big)\nu(\mathrm{d}y) \leq f(x). \tag{2}$$

(C2) *There is a closed set $U_0 \subset U$ such that for any $(T, \nu) \in U_0$ and $x \in A \setminus A^\star$:*

$$\int_B f\big(T(x, y)\big)\nu(\mathrm{d}y) < f(x). \tag{3}$$

Let $\{u_t = (T_t, \nu_t) : t \geq 1\} \subset U$ satisfy the following:

(U0) *There is $t_0 \geq 1$ such that for any $t \geq 1$ there is $s \leq t_0$ with $u_{t+s} \in U_0$.*

Then, for every $\varepsilon > 0$:

$$\lim_{t \to \infty} \mathrm{Prob}\big(\mathrm{dist}\big(X_t, A^\star\big) < \varepsilon\big) = 1. \tag{4}$$

Assume additionally the following:

(D) *For any $t \geq 1$, $x \in A$ and $y \in B$: $f(T_t(x, y)) \leq f(x)$.*

Then

$$\mathrm{Prob}\big(X_t \longrightarrow A^\star,\ as\ t \longrightarrow \infty\big) = 1, \tag{5}$$

i.e.

$$\mathrm{Prob}\big(\big\{\omega \in \Omega\colon\ \mathrm{d}\big(X_t(\omega), A^\star\big) \longrightarrow 0,\ as\ t \longrightarrow \infty\big\}\big) = 1.$$

Remark 2 One can release the assumption of compactness of the set A assuming (D):

(E) There exists $r > \min f$ such that set $A_r := \{x \in A : f(x) \leq r\}$ is compact and $\operatorname{supp} \mu_0 \subset A_r$.

In fact, by (D) $T_t(A_r \times B) \subset A_r$. Clearly, μ_0 is a probability measure on A_r and $A^\star \subset A_r$. Hence, we may apply Theorem 1 to set A_r.

In Sects. 2–4 we present a detailed proof of Theorem 1. Its main idea is to view the algorithm as the nonautonomous semidynamical system defined on the space of all probability measures on the set A. The system is determined by a family of so-called Foias operators, which transport the probability measures by the pairs $(T_t, \nu_t) \in U$. Its trajectories $\{\mu_t\colon t=0,1,\dots\}$ are just sequences of the distributions of the algorithm X_t, $t=0,1,\dots$, and weakly converge to the set of all measures supported on the set $A^\star$. This implies stochastic convergence (4). The convergence with probability one (5) is a consequence of (4) and the monotonicity of $f(X_t)$.

2

Recall some definitions and concepts from the theory of dynamical systems. Let X be a metric space with a distance ϱ_X. Let $\varphi : X \longrightarrow X$ be a continuous map. For any $x \in X$, $o(x)$ denotes the orbit of φ, i.e. $o(x)=\{\varphi^t x : t \geq 0\}$. Here φ^t denotes the tth iterates of φ, i.e. $\varphi^0 = Id_X$ and $\varphi^t = \varphi \circ \varphi^{t-1}$, for $t = 1, 2, 3, \dots$. In other words, $o(x)=\{x_t : x_0 = x$, and $x_{t+1} = \varphi x_t$, for $t=0,1,2,\dots\}$. A compact set $\emptyset \neq K \subset X$ is invariant if $\varphi(K) \subset K$. For any $x \in X$, $\omega(x)$ denotes the ω-limit set of x: $\omega(x) = \{y \in X : \exists\, t_i \longrightarrow \infty, \varphi^{t_i} x \longrightarrow y\}$. It is easy to see that, if X is compact, then any ω-limit set is nonempty, compact and invariant. Also, for any invariant set $K \subset X$ and $x \in X$: $\varrho_X(\varphi^t x, K) \longrightarrow 0$ for $t \longrightarrow \infty$, if and only if $\omega(x) \subset K$.

The following theorem is a version of the well-known Lyapunov stability theorem. However, it is simpler than its classical counterpart, as we are not interested in stability here but in attractiveness. On the other hand, we do not assume strong monotonicity of the Lyapunov function along trajectories, but we use a weaker assumption instead which is quite natural in our context.

Theorem 2 *Let (X, ϱ_X) be a compact metric space, $\emptyset \neq K \subset X$ a compact and invariant set, $\varphi : X \longrightarrow X$ a continuous map. Let $W : X \longrightarrow \mathbb{R}$ be a Lyapunov function, i.e.:*

1. *W is continuous.*
2. *$W(x) = 0$, for $x \in K$.*
3. *$W(x) > 0$, for $x \in X \setminus K$.*
4. *For every $x \in X \setminus K$*
$$W(\varphi x) \leq W(x) \tag{6}$$
and there exists $s \geq 1$ such that
$$W\big(\varphi^s x\big) < W(x). \tag{7}$$

Then, for every $x \in X$,
$$\varrho_X\big(\varphi^t x, K\big) \longrightarrow 0, \quad \text{as } t \longrightarrow \infty. \tag{8}$$

Proof Let $x \in X$. As noted above, $\omega(x) \neq \emptyset$. We will show that W is constant on $\omega(x)$. In fact, choose two points $y, z \in \omega(x)$ and corresponding sequences $\{s_i\}$ and $\{t_i\}$ tending to infinity such that $\varphi^{s_i}(x) \longrightarrow y$ and $\varphi^{t_i}(x) \longrightarrow z$. Taking subsequences, if necessary, one can assume that for any i: $s_i < t_i < s_{i+1} < t_{i+1}$. Hence
$$W\big(\varphi^{s_i}(x)\big) \leq W\big(\varphi^{t_i}(x)\big) \leq W\big(\varphi^{s_{i+1}}(x)\big) \leq W\big(\varphi^{t_{i+1}}(x)\big).$$
Letting $i \longrightarrow \infty$, by continuity of W, we have
$$W(y) \leq W(z) \leq W(y) \leq W(z).$$

Therefore, as we have claimed, $W(y)=W(z)$. As $\omega(x)$ is invariant, then $o(y)\subset\omega(x)$ for any $y\in\omega(x)$. Hence, for any $y\in\omega(x)$, $W(\varphi^t(y))=W(y)$ for all $t\geq 1$. Using the assumption (7) we see that $y\in K$. Finally, $\omega(x)\subset K$ and x is attracted to K, as required. □

Let $(\mathcal{U},\varrho_{\mathcal{U}})$ be a compact metric space and let $\Pi:\mathcal{U}\times M\ni(u,m)\longrightarrow\Pi_u m\in M$ and $\theta:\mathcal{U}\longrightarrow\mathcal{U}$ be given continuous maps. For $t\geq 1$, denote by $\Pi^{[u,t]}$ the composition $\Pi^{[u,t]}=\Pi_{\theta^{t-1}u}\circ\cdots\circ\Pi_{\theta u}\circ\Pi_u$.

The main result of this section is stated in the following theorem.

Theorem 3 *Let $\emptyset\neq K\subset M$ be an invariant set for any Π_u, $u\in\mathcal{U}$. Let $V:M\longrightarrow\mathbb{R}$ be a Lyapunov function for any Π_u, $u\in U$, i.e.*

1. *V is continuous.*
2. *$V(m)=0$, for $m\in K$.*
3. *$V(m)>0$, for $m\in M\setminus K$.*
4. *For every $m\in M\setminus K$ and $u\in\mathcal{U}$*

$$V(\Pi_u m)\leq V(m) \tag{9}$$

and there exists $s\geq 1$ such that

$$V\big(\Pi^{[u.s]}m\big)<V(m). \tag{10}$$

Then, for each $m\in M$ and $u\in\mathcal{U}$,

$$\varrho_M\big(\Pi^{[u,t]}m,K\big)\longrightarrow 0,\quad \text{as } t\longrightarrow\infty. \tag{11}$$

Proof Let $X=\mathcal{U}\times M$ and $\varrho_X((u_1,m_1),(u_2,m_2))=\varrho_{\mathcal{U}}(u_1,u_2)+\varrho_M(m_1,m_2)$. Let $\varphi:X\longrightarrow X$ be given by

$$\varphi(u,m)=(\theta u,\Pi_u m). \tag{12}$$

Note first that for any $(u,m)\in\mathcal{U}\times M$ and $t\geq 1$, by simple induction,

$$\varphi^t(u,m)=\big(\theta^t u,\Pi^{[u,t]}m\big). \tag{13}$$

Now consider the set $K'=\mathcal{U}\times K\subset X$ and the function $W:X\ni(u,m)\longrightarrow V(m)$. It is obvious that K' is invariant under φ and the function W fulfils all the assumptions of Theorem 2. Then, for any $u\in\mathcal{U}$ and $m\in M$, we have $\varrho_X(\varphi^t(u,m),K')\longrightarrow 0$, as $t\longrightarrow\infty$. Hence $\varrho_M(\Pi^{[u,t]}m,K)\longrightarrow 0$, as $t\longrightarrow\infty$, as required. □

3

In this section we will use the notation from Sect. 1. Recall that M and N denote the spaces of all Borel probability measures on A and B, respectively. On both spaces we consider the topologies of weak convergence of measures. Let us recall then some facts about the weak convergence; for more details one may refer to Billingsley (1999) or Parthasarathy (2005). Let (S,d) be a Polish space and $M(S)$ be the family of Borel probability measures on S. The sequence $\mu_n \in M(S)$ weakly converges to $\mu \in M(S)$ if and only if for all continuous and bounded functions $h : S \longrightarrow \mathbb{R}$ we have

$$\int_S h \,\mathrm{d}\mu_n \longrightarrow \int_S h \,\mathrm{d}\mu, \quad \text{as } n \longrightarrow \infty. \tag{14}$$

Another, equivalent condition for weak convergence is

$$\mu_n(C) \longrightarrow \mu(C), \tag{15}$$

as $n \longrightarrow \infty$, for any $C \in \mathcal{B}(S)$ such that $\mu(\partial C) = 0$, where ∂C is the boundary of C. As S is a Polish space, $M(S)$ can be metrized as a complete space and several metrics can be used. One of them is the Fortet–Mourier metric:

$$\varrho_M(\mu_1, \mu_2) = \sup_{g \in \mathcal{G}} \left| \int_S g \,\mathrm{d}\mu_1 - \int_S g \,\mathrm{d}\mu_2 \right|,$$

where $\mathcal{G} = \{g : S \longrightarrow \mathbb{R};\ |g(x) - g(y)| \le |x - y|, |g(x)| \le 1, \text{for all } x, y \in S\}$.

If S is compact, so is $M(S)$. We have assumed that A is compact, thus $M = M(A)$ is compact. In this case any continuous function $h : A \longrightarrow \mathbb{R}$ is bounded, which makes the condition (14) easy to verify.

For any $T \in \mathcal{T}$ and $\nu \in M$ the Foias operator $P_{(T,\nu)} : M \longrightarrow M$ is defined as follows:

$$P_{(T,\nu)}\mu(C) = (\mu \times \nu)\big(T^{-1}(C)\big), \quad \text{for } \mu \in M,\ C \in \mathcal{B}(A), \tag{16}$$

where $T^{-1}(C) = \{(x, y) \in A \times B : T(x, y) \in C\}$ is the preimage of C. We will also write $P_{(T,\nu)}\mu =: (\mu \times \nu)T^{-1}$.

Definition (16) yields immediately:

Lemma 1 *Let $X : \Omega \longrightarrow A$ and $Y : \Omega \longrightarrow B$ be independent random variables with distributions μ and ν, respectively. Then $T(X, Y)$ is distributed according to $P_{(T,\nu)}\mu$. Furthermore, for any continuous function $h : A \longrightarrow \mathbb{R}$, by change of variables,*

$$\int_A h \,\mathrm{d}P_{(T,\nu)}\mu = \int_\Omega h\big(T(X, Y)\big) \,\mathrm{d}P = \int_{A \times B} (h \circ T) \,\mathrm{d}(\mu \times \nu). \tag{17}$$

We will take advantage of the following general result; see Theorems 2.8 and 2.7 in Billingsley (1999).

Lemma 2

1. *Let μ_n, ν_n be sequences of Borel probability measures on separable metric spaces S_1, S_2, respectively, with $\mu_n \longrightarrow \mu$ and $\nu_n \longrightarrow \nu$ for some Borel probability measures μ on S_1 and ν on S_2. Then $\mu_n \times \nu_n \longrightarrow \mu \times \nu$.*
2. *Assume that S_1, S_2 are metric spaces, μ is a Borel probability measure on S_1 and $T : S_1 \longrightarrow S_2$ is measurable with $\mu(D_T) = 0$, where D_T is the set of all discontinuities of T. Then, for any sequence μ_n of Borel probability measures on S_1, if $\mu_n \longrightarrow \mu$, then $\mu_n T^{-1} \longrightarrow \mu T^{-1}$.*

Lemma 2 leads to the following.

Proposition 1 *Assume that $U \subset \mathcal{T} \times M$ satisfy the assumption* (A) *of Theorem 1. Then the function $P : U \times M \ni (u, \mu) \longrightarrow P_u\mu \in M$ is continuous, where P_u denotes the Foias operator.*

Proof Let $U \times M \ni (T_n, \mu_n, \nu_n) \longrightarrow (T, \mu, \nu) \in U \times M$. We would like to show that $(\mu_n \times \nu_n)T_n^{-1} \longrightarrow (\mu \times \nu)T^{-1}$. First note that A is separable, being compact, and B is separable, as assumed. Hence, by statement 1. of Lemma 2, $\mu_n \times \nu_n \longrightarrow \mu \times \nu$. By Fubini's theorem, we have $(\mu \times \nu)(D_T) = \int_A \nu(D_T(x))\mu(\mathrm{d}x)$, where $B \supset D_T(x)$ is the set of $y \in B$ such that (x, y) are discontinuities of T. By assumption (A), $\nu(D_T(x)) = 0$ and $(\mu \times \nu)(D_T) = 0$. By statement 2. of Lemma 2, $(\mu_n \times \nu_n)T^{-1} \longrightarrow (\mu \times \nu)T^{-1}$. Equivalently, for any continuous function $h : A \longrightarrow \mathbb{R}$

$$\int_{A\times B} (h \circ T)\,\mathrm{d}(\mu_n \times \nu_n) \longrightarrow \int_{A\times B} (h \circ T)\,\mathrm{d}(\mu \times \nu). \tag{18}$$

We want to show

$$\int_{A\times B} (h \circ T_n)\,\mathrm{d}(\mu_n \times \nu_n) \longrightarrow \int_{A\times B} (h \circ T)\,\mathrm{d}(\mu \times \nu),$$

for any continuous $h : A \longrightarrow \mathbb{R}$. In fact, we have

$$\begin{aligned}
&\int_{A\times B} (h \circ T_n)\,\mathrm{d}(\mu_n \times \nu_n) - \int_{A\times B} (h \circ T)\,\mathrm{d}(\mu \times \nu) \\
&\quad = \left(\int_{A\times B} (h \circ T_n)\,\mathrm{d}(\mu_n \times \nu_n) - \int_{A\times B} (h \circ T)\,\mathrm{d}(\mu_n \times \nu_n)\right) \\
&\qquad + \left(\int_{A\times B} (h \circ T)\,\mathrm{d}(\mu_n \times \nu_n) - \int_{A\times B} (h \circ T)\,\mathrm{d}(\mu \times \nu)\right).
\end{aligned}$$

The second component tends to 0 by (18), while the first one satisfies

$$\begin{aligned}
&\left|\int_{A\times B} (h \circ T_n)\,\mathrm{d}(\mu_n \times \nu_n) - \int_{A\times B} (h \circ T)\,\mathrm{d}(\mu_n \times \nu_n)\right| \\
&\quad \le \sup_{(a,b)\in A\times B} \big|h\big(T_n(a, b)\big) - h\big(T(a, b)\big)\big|,
\end{aligned}$$

and tends to 0 as $T_n \longrightarrow T$ uniformly on $A \times B$ and h is uniformly continuous on A. □

Define

$$M^\star = \big\{\mu \in M : \operatorname{supp}\mu \subset A^\star\big\}.$$

Note that $\mu \in M^\star$ if and only if $\mu(A^\star) = 1$. It is easy to see that $M^\star$ is a compact subset of M as $A^\star$ is a compact subset of A. Also, it is obvious that condition (B) yields invariance of $M^\star$ under each $P_{(T,\nu)}$, $T \in \mathcal{T}$ and $\nu \in N$.

As a consequence of Theorem 3 we will prove the following.

Theorem 4 *Assume the conditions* (A), (B), (C1), (C2) *and* (U0) *of Theorem 1. Let $\mu_t = P_{(T_t,\nu_t)}\mu_{t-1}$, $t = 1, 2, 3, \ldots$. Then*

$$\varrho_M\big(\mu_t, M^\star\big) \longrightarrow 0, \quad \textit{as } t \longrightarrow \infty.$$

Proof As mentioned above the compactness of A implies compactness of M. Define the shift space Σ on the alphabet U, i.e.

$$\Sigma = \big\{(u_1, u_2, u_3, \dots) : u_i \in U, i = 1, 2, 3, \dots\big\},$$

and the shift map $\theta : \Sigma \longrightarrow \Sigma$,

$$\theta(u_1, u_2, u_3, \dots) = (u_2, u_3, u_4, \dots).$$

Let U_0 and $t_0 \geq 1$ be such as in Theorem 1. Define

$$\mathcal{U} = \{u \in \Sigma : \text{for any } i \text{ there exists } 0 \leq j \leq t_0 : u_{i+j} \in U_0\}. \tag{19}$$

Consider the distance d on Σ defined by

$$d(u, v) = \sum_{i=1}^{\infty} 2^{-i} d_U(u_i, v_i), \tag{20}$$

where d_U is a metric on U compatible with the product topology endowed from the topology of uniform convergence of $\mathcal{T}$ and the topology of weak convergence of N. As U was assumed to be compact, Σ is compact. Clearly θ is continuous; actually it is Lipschitz. Also it is evident that $\mathcal{U}$ is invariant under θ. Moreover, as a closed subset of Σ, $\mathcal{U}$ is compact. In fact, let $u^n \longrightarrow u$, as $n \longrightarrow \infty$ and $u^n \in \mathcal{U}$. Fix $i_0 \geq 1$. Fix $\varepsilon > 0$ and choose n such that $d(u^n, u) \leq \varepsilon 2^{-(i_0+t_0)}$, where t_0 is to satisfy assumption (U0) of Theorem 1. Then, for any $j \leq t_0$, we have $2^{-(i_0+j)} \mathrm{d}_U(u^n_{i_0+j}, u_{i_0+j}) \leq d(u^n, u) \leq \varepsilon 2^{-(i_0+t_0)}$, and hence $d_U(u^n_{i_0+j}, u_{i_0+j}) \leq \varepsilon 2^{-(i_0+t_0)} 2^{-(i_0+j)} \leq \varepsilon$. Assumption (U0) means that for some $j \leq t_0$, $u^n_{i_0+j} \in U_0$ and consequently $d(u_{i_0+j}, U_0) \leq d(u_{i_0+j}, u^n_{i_0+j}) \leq \varepsilon$. We have just proved that for any $\varepsilon > 0$ there exists $j \leq t_0$ with $d(u_{i_0+j}, U_0) \leq \varepsilon$. This means that there exists $j_0 \leq t_0$ such that $d(u_{i_0+j_0}, U_0) = 0$. As U_0 is closed, $u_{i_0+j_0} \in U_0$. Thus, $u \in \mathcal{U}$, and so $\mathcal{U}$ is closed and then compact, as required.

As the projection $\Sigma \ni u \longrightarrow u_1 \in U$ is continuous, by Proposition 1 the map $\Pi : \mathcal{U} \times M \ni ((u_1, u_2, u_3, \dots), \mu) \longrightarrow P_{u_1}\mu \in M$ is continuous.

Define the function $V : M \longrightarrow \mathbb{R}$:

$$V(\mu) = \int_A (f - \min f)\,\mathrm{d}\mu = \int_A f\,\mathrm{d}\mu - \min f$$

to be a Lyapunov function. We are going to show that the assumptions of Theorem 3 are satisfied with $K = M^\star$.

The continuity of V is an immediate consequence of the definition of the topology on M. Let $\mu_n \longrightarrow \mu$. We put $h = f - \min f$ in (14) to get

$$V(\mu_n) = \int_A h\,\mathrm{d}\mu_n \longrightarrow \int_A h\,\mathrm{d}\mu = V(\mu).$$

Clearly $V(\mu) \geq 0$ for all $\mu \in M$ and $V(\mu) = 0$ for all $\mu \in M^\star$. Let $V(\mu) = 0$ for some $\mu \in M$. Then we have $0 = V(\mu) = \int_A (f - \min f)\,\mathrm{d}\mu = \int_{A^\star} (f - \min f)\,\mathrm{d}\mu + \int_{A \setminus A^\star} (f - \min f)\,\mathrm{d}\mu = \int_{A \setminus A^\star} (f - \min f)\,\mathrm{d}\mu$. As $f - \min f$ is strictly positive on $A \setminus A^\star$ and $\operatorname{supp}\mu \subset A^\star$, we have $\mu \in M^\star$.

We need to verify condition 4 of Theorem 3. Let $\mu \in M \setminus M^\star$ and $u = ((T_1, \nu_1), (T_2, \nu_2), (T_3, \nu_3), \dots) \in \mathcal{U}$. Condition (C1) says that for any $x \in A \setminus A^\star$ we have $\int_B f(T_1(x, y))\nu_1(\mathrm{d}y) \leq f(x)$. Hence, by (17) and Fubini's theorem,

$$
\begin{aligned}
V(\Pi_u\mu) = V(P_{u_1}\mu) &= \int_A f\,\mathrm{d}P_{u_1}\mu - \min f = \int_{A\times B}(f\circ T_1)\,\mathrm{d}(\mu\times\nu_1) - \min f \\
&= \int_A\biggl(\int_B f\bigl(T_1(x,y)\bigr)\nu_1(\mathrm{d}y)\biggr)\mu(\mathrm{d}x) - \min f \\
&\le \int_A f\,\mathrm{d}\mu - \min f = V(\mu). \qquad (21)
\end{aligned}
$$

To prove the second assertion in 4, note first that $\Pi^{[u,t]} = \Pi_{\theta^{t-1}u}\circ\cdots\circ\Pi_u = P_{u_t}\circ\cdots\circ P_{u_1} = P_{u_t}\circ\Pi^{[u,t-1]}$, as $\theta^{t-1}u = (u_t, u_{t+1}, u_{t+2},\ldots)$. By what we have just proved, we get a sequence of inequalities: $V(\Pi^{[u,t]}\mu)\le V(\Pi^{[u,t-1]}\mu)\le\cdots\le V(\Pi^{[u,1]}\mu) = V(P_{u_1}\mu)\le V(\mu)$. Just note here that if some $\Pi^{[u,s]}\mu\in M^\star$, then also $P_{u_{s+1}}\circ\Pi^{[u,s]}\mu\in M^\star$ and the appropriate inequality is still clearly satisfied. Let j be such that $u_{1+j}\in U_0$. Then $V(\Pi^{[u,j+1]}\mu) = V(P_{u_{j+1}}\circ\Pi^{[u,j]}\mu) = V(P_{u_{j+1}}\bar\mu)$, with $\bar\mu = \Pi^{[u,j]}\mu$. If $\bar\mu\in M^\star$, then $V(\bar\mu) = 0 < V(\mu)$, as we have assumed $\mu\in M\setminus M^\star$. Then $V(\Pi^{[u,j+1]}\mu) < V(\mu)$. Assume now that $\bar\mu\in M\setminus M^\star$. We then have

$$
\begin{aligned}
V\bigl(\Pi^{[u,j+1]}\mu\bigr) = V(P_{u_{j+1}}\bar\mu) &= \int_A f\,\mathrm{d}P_{u_{1+j}}\bar\mu - \min f \\
&= \int_A\biggl(\int_B f\bigl(T_{1+j}(x,y)\bigr)\nu_{1+j}(\mathrm{d}y)\biggr)\bar\mu(\mathrm{d}x) - \min f \\
&< \int_A f\,\mathrm{d}\bar\mu - \min f = V(\bar\mu)\le V(\mu).
\end{aligned}
$$

Theorem 3 completes the proof, as the sequence $\{u_t\}$ specified in Theorem 1 belongs to $\mathcal{U}$ defined by (19) and $\mu_t = \Pi^{[u,t]}\mu_0$. □

4

Proof of Theorem 1 We interpret the above Theorem 4 in terms ofs random variables X_t. Note first that for any measure $\mu^\star\in M^\star$ and any set $C\in\mathcal{B}(A)$ such that $A^\star\subset\operatorname{int}C$ we have $\mu^\star(\delta C) = 0$ and $\mu^\star(C) = 1$. Thus, the condition (15) implies that for any sequence of measures $\mu_n\in M$ such that $\mu_n\longrightarrow\mu^\star$ we have

$$
\mu_n(C)\longrightarrow 1,\quad \text{as } n\longrightarrow\infty.
$$

Note now that the measures μ_t, defined by $\mu_t = P_{(T_t,\mu_t)}\mu_{t-1}$, $t = 1,2,3,\ldots$ are, by Lemma 1, the distributions of the random variables X_t.

Let $B(A^\star,\varepsilon) = \{a\in A : \operatorname{dist}(a,A^\star) < \varepsilon\}$. By Theorem 4 the sequence μ_t of the distributions of X_t tends to the compact set $M^\star$. Hence, the set of partial limits of $\{\mu_t\}$ is nonempty and is contained in $M^\star$. Then, for any sequence $t_n\longrightarrow\infty$, there exists a subsequence $t_{n_i}\longrightarrow\infty$ and a measure $\mu^\star\in M^\star$ such that $\mu_{t_{n_i}}\longrightarrow\mu^\star$. Hence $\mu_{t_{n_i}}(B(A^\star,\varepsilon))\longrightarrow\mu^\star(B(A^\star,\varepsilon)) = 1$. But this means that $\mu_t(B(A^\star,\varepsilon))\longrightarrow 1$, as $t\longrightarrow\infty$. As μ_t is the distribution of X_t, we have $\mu_t(B(A^\star,\varepsilon)) = \operatorname{Prob}(X_t\in$

$B(A^\star, \varepsilon))$. We have thus proved the condition (4), i.e. for every $\varepsilon > 0$:

$$\lim_{t\to\infty} \operatorname{Prob}\bigl(\operatorname{dist}\bigl(X_t, A^\star\bigr) < \varepsilon\bigr) = 1, \tag{22}$$

which completes the proof of the first part of Theorem 1.

We will prove the second part. First we prove the following.

Lemma 3 *For every* $\varepsilon > 0$:

$$\lim_{t\to\infty} \operatorname{Prob}\bigl(f(X_t) < \min f + \varepsilon\bigr) = 1. \tag{23}$$

Proof Fix $\varepsilon > 0$. As A is compact, f is uniformly continuous and one can find $\delta > 0$ such that $\operatorname{dist}(x, A^\star) < \delta$ implies $f(x) < \min f + \varepsilon$. Now we can use the condition (4), with $\varepsilon = \delta$. □

Proposition 2 (Folklore) *Let* $\xi_n : \Omega \longrightarrow \mathbb{R}$ *be a nonincreasing sequence of non-negative random variables stochastically convergent to* 0. *Then* ξ_n *tends to* 0 *with probability* 1.

Proof Fix $\varepsilon > 0$. For any natural N let $C_N = \{\xi_N < \varepsilon\}$. By monotonicity of ξ_n, $C_N = \{\xi_n \le \varepsilon, n \ge N\}$. Let $D_\varepsilon = \bigcup_N \{\xi_n \le \varepsilon, n \ge N\}$. As $C_N \subset C_{N+1}$ and $\operatorname{Prob}(C_N) \longrightarrow 1$, we have $\operatorname{Prob}(D_\varepsilon) = \operatorname{Prob}(\bigcup_N C_n) = \lim_{N\longrightarrow\infty} \operatorname{Prob}(C_N) = 1$. On the other hand, $\{\xi_n \longrightarrow 0\} = \bigcap_{m=1}^{\infty} D_{\frac{1}{m}}$, hence $\operatorname{Prob}(\{\xi_n \longrightarrow 0\}) = 1$. □

Now, one can see that the sequence $\xi_t = f(X_t) - \min f$ satisfies the assumptions of the above proposition. Hence, it tends to 0 with probability 1. The compactness of A and the continuity of f imply

$$\bigl\{\omega : f\bigl(X_t(\omega)\bigr) - \min f \longrightarrow 0,\ \text{as} \longrightarrow \infty\bigr\} \subset \bigl\{\omega : X_t(\omega) \longrightarrow A^\star,\ \text{as} \longrightarrow \infty\bigr\}.$$

The proof of Theorem 1 is thus completed. □

Remark 3 If we assume (A), (B) and replace (C1), (C2), (U0) with:

(C) For any $x \in A \setminus A^\star$:

$$\int_B f\bigl(T(x, y)\bigr)\nu(\mathrm{d}y) < f(x), \tag{24}$$

then the statements of Theorem 1 remain true.

5

In this section we illustrate the functionality of Theorem 1 by analyzing the following example.

Let $A = [0, 1]^n \subset \mathbb{R}^n$ and $\{r_t : t = 1, 2, 3, \ldots\} \subset [\varepsilon, 1]$ for some $0 < \varepsilon < 1$.

Algorithm

0. Set $t=0$. Sample X_0 uniformly from A.
1. Given $X_t \in A$, generate Q_{t+1} from the uniform distribution on $B(X_t, r_{t+1})$, where $B(x,r)=\{y \in A\colon |x^i - y^i| \le r,\ i=1,2,\dots,n\}$.
2. If $f(Q_{t+1}) < f(X_t)$, then let $X_{t+1}=Q_{t+1}$.
   Else if $f(Q_{t+1}) \ge f(X_t)$, then let $X_{t+1}=X_t$.
   Increment $t := t+1$ and go to Step 1.

Let d denote the maximum metric on A and let A' denote the set of local minimums $a \in A \setminus A^*$ of the function f. For any $a \in A$ let

$$A_{f(a)} := \bigl\{x \in A : f(x) < f(a)\bigr\}.$$

Theorem 5 *Let $R \in (0,1)$ be such that for any $a \in A'$ $d(a, A_{f(a)}) < R$. Assume that there exists $t_0 \ge 1$, such that for any t, $r_{t+i} \ge R$ for some $i \le t_0$. Assume, additionally, that for any $c \in \mathbb{R}$, the level curve $l_c = \{x \in A : f(x) = c\}$ has Lebesgue measure* 0. *Then*

$$\operatorname{Prob}\bigl(X_t \longrightarrow A^\star \text{ as } t \longrightarrow \infty\bigr) = 1. \tag{25}$$

Proof Note that for any $x \in A$, $B(x, r_t) = \prod_{i=1}^n [a^i(x,r_t), b^i(x,r_t)]$, where $a^i(x,r) = \max\{x^i - r, 0\}$, $b^i(x,r) = \min\{x^i + r, 1\}$. Define $B = [\varepsilon, 1] \times A$. Let

$$Q : A \times B \ni \bigl(x, (r,y)\bigr) \longrightarrow a(x,r) + y \otimes \bigl(b(x,r) - a(x,r)\bigr) \in A,$$

where $\otimes :\ A^2 \ni [(x^1, \dots, x^n), (y^1, \dots, y^n)] \longrightarrow (x^1y^1, \dots, x^ny^n) \in A$.

It is easy to see that if $Z\colon \Omega \longrightarrow [0,1]^n$ is uniformly distributed on $[0,1]^n$, then $Q(x,r,Z)$ is uniformly distributed on $B(x,r)$. Let $Z_t : \Omega \longrightarrow [0,1]^n$ be i.i.d. sequence of uniformly distributed random variables. We define $Y_t : \Omega \longrightarrow B$ as $Y_t = (r_t, Z_t)$. Obviously, for any t, $Q(x, Y_t)$ is uniformly distributed on $B(x, r_t)$.

Let $\tilde{T} : A \times A \longrightarrow A$ satisfy

$$\tilde{T}(x,y) = \begin{cases} x, & \text{if } f(y) \ge f(x), \\ y, & \text{if } f(y) < f(x). \end{cases}$$

Define

$$T : A \times B \ni (x,r,y) \longrightarrow \tilde{T}\bigl(x, Q(x,r,y)\bigr)$$

and

$$X_{t+1} := \tilde{T}\bigl(X_t, Q(X_t, Y_t)\bigr) = T(X_t, Y_t).$$

Let ν denote the uniform distribution on $[0,1]^n$. We consider U as the set $\{T\} \times \{\nu_r,\ r \in [\varepsilon, 1]\}$, where $\nu_r = \delta_r \times \nu$ is a distribution on B and δ_r denotes the Dirac measure on $[\varepsilon,1]$ for $r \in [\varepsilon,1]$. Obviously, for any t, ν_{r_t} is the distribution of Y_t. Let $U_0 := \{T\} \times \{\nu_r,\ r \in [R,1]\}$. It is clear that U is compact and $U_0 \subset U$ is closed. To check (A), fix $x_0 \in A$ and $(T, \nu_{r_0}) \in U$. Recall that $T(x_0, r, u) = \tilde{T}(x_0, Q(x_0, r, u))$. Note that since Q is continuous, then to verify that

$$\nu_{r_0}\bigl(\bigl\{(r,y) \in B\colon\ T \text{ is not continuous in } (x_0, r, y)\bigr\}\bigr) = 0,$$

it is enough to show that

$$\nu_{r_0}\big(\big\{(r,y)\in[0,1]\times[0,1]^n : \tilde{T} \text{ is not continuous in } \big(x_0, Q(x_0,r,y)\big)\big\}\big)=0.$$

Equivalently $\nu(\{y\in[0,1]^n : \tilde{T}$ is not continuous in $(x_0, Q(x_0,r_0,y))\})=0$. But the discontinuities (x,y) of the function $\tilde{T}$ must satisfy $f(x)=f(y)$, and consequently, it is enough to show that $\nu(\{y\colon\ Q(x_0,r_0,y)\in l_{f(x_0)}\})=0$. Equivalently, $P(Q(x_0,r_0,Z)\in l_{f(x_0)})=0$, where Z is distributed according to ν—but this is clear, since $l_{f(x_0)}$ has Lebesgue measure 0 and $Q(x_0,r_0,Z)$ is uniformly distributed on $B(x_0,r_0)$.

By the definition of T, (D) is satisfied, and hence (B) and (C1) are.

To prove (C2), note that by the definition of $\tilde{T}$, it is enough to show that for any $r_0\in[R,1]$ and $x\in A\setminus A^*$,

$$\int_A \min\big\{f(x), f\big(Q(x,r_0,y)\big)\big\}\nu(\mathrm{d}y) < f(x).$$

We need $\nu(\{y : Q(x,r_0,y)\in A_{f(x)}\})>0$. Equivalently, $\nu(B(x,r_0)\cap A_{f(x)})>0$—this is satisfied immediately for all x besides the set $A'\cup A^*$ by the continuity of f, while for $x\in A'$ it follows from $d(x,A_{f(x)})<R\le r_0$ and again by the continuity of f.

Condition (U0) follows from the description of the algorithm. □

Remark 4 For any continuous function f with $\nu(l_c)=0$, $c\in\mathbb{R}$, there exists $R<1$ such that $d(a,A_{f(a)})<R$, $a\in A'$. Furthermore, if $A'=\varnothing$, then for any sequence $\{r_t\}_{t=0}^{\infty}\subset[\varepsilon,1]$ the algorithm converges, since the assumptions of Theorem 5 are satisfied for $R=\varepsilon$.

Remark 5 Let the assumptions of Theorem 5 hold true. Assume that some algorithm $\tilde{X}_t$ satisfies the condition (A) of Theorem 1 and takes the following form: $\tilde{X}_t=\tilde{T}(\tilde{X}_{t-1},\tilde{Q}(\tilde{X}_{t-1},Y_t))$, where for any $x\in A$, the distribution of a random variable $Q(x,Y_t)$ is absolutely continuous according to Lebesgue measure and its density is positive on $B(x,r_t)$. Then the algorithm $\tilde{X}_t$ converges, since the conditions (D) and (C2) are satisfied for reasons analogous to the above. For an interesting example of such an algorithm $\tilde{X}_t$, see the GEM algorithm established in Ahrari and Atai (2010), Ahrari et al. (2009).

Remark 6 The assumption that the level curves have zero Lebesgue measure is a technical one, and perhaps it is not necessary. We need it here to get condition (A) in Theorem 1. On the other hand, in most practical cases this assumption is fulfilled.

6

Now we show an application of Theorem 1 to a broad class of numerical methods used for an approximation of the set $A^\star$, which are sometimes called multistart algorithms, and which can be described as follows.

A map $\varphi : A \longrightarrow A$ will be called a *local method* if $f(\varphi(x))\le f(x)$, for all $x\in A$, where f is the cost function to be minimized.

Let $\mu_0 \in M$ and let k, m be natural numbers. Let Φ denote the set of local methods, let $N \subset M$ be compact and let N_0 be a closed subset of N, such that for any $\nu \in N_0$, $\nu(G) > 0$ for any open neighborhood G of the set $A^\star$.

Algorithm

1. Let $t = 0$. Choose an initial population, i.e. a simple sample of points from A distributed according to μ_0:
$$x = (x^1, \dots, x^m) \in A^m.$$
2. Let $t = t + 1$.
3. Draw independently k points $y^i \in A$ according to a distribution $\nu^{t_i} \in N$ each, $i = 1, \dots, k$. Let $y = (y^1, \dots, y^k) \in A^k$.
4. Apply $\varphi^{t_i} \in \Phi$ to x^i, $i = 1, \dots, m$.
5. Sort the sequence
$$(\varphi^{t_i}(x^1), \dots, \varphi^{t_m}(x^m), y^1, \dots, y^k)$$
using f as a criterion to get
$$(\bar{x}^1, \dots, \bar{x}^{m+k}) \quad \text{with } f(\bar{x}^1) \le \dots \le f(\bar{x}^{m+k}).$$
6. Form the next population with the first m points
$$\bar{x} = (\bar{x}^1, \dots, \bar{x}^m)$$
and go to point 2 with $x = \bar{x}$.

Repeat 2–6 according to a stopping rule.

There is a number of local methods available. For example, a classical one is the gradient method. It requires differentiability of the objective function f and still it is quite effective in finding local minima attained at interior points of the set A. If f is not a smooth function or its local minimum point is at the boundary of A, then more sophisticated methods can be used, see Robert and Casella (2004) and for a survey Wright (2005). The algorithm above admits use of various methods at the same time or just one method with various parameters (like a step size or a number of steps taken). Obviously, the identity map is a local method.

For $t = 1, 2, 3, \dots$ define the map
$$T_{\varphi^1,\dots,\varphi^m} : A^m \times A^k \longrightarrow A \quad \text{as } T_{\varphi^1,\dots,\varphi^m}(x, y) = \bar{x}.$$
Let $\widehat{f} : A^m \longrightarrow \mathbb{R}$ be defined as $\widehat{f}(x) = f(x^1)$. Let us note that $\widehat{A^\star} = A^\star \times A^{m-1}$ is the set of global minima of $\widehat{f}$.

The following theorem gives sufficient conditions for almost sure convergence of the above algorithm to the set of solutions of the global minimization problem.

Theorem 6 *Let $\{X_t : t = 1, 2, 3, \dots\}$ be the sequence generated by (1), where $X_0 = (X_0^1, \dots, X_0^m)$ is a random vector with distribution $(\mu_0)^m$. Let for each $t = 1, 2, 3, \dots,$ $Y_t = (Y_t^1, \dots, Y_t^k)$ be independent random vectors, and independent of X_0, distributed according to $\nu^{t_1} \times \dots \times \nu^{t_k}$ with $\nu^{t_i} \in N$. Assume that:*

(z1) *For any* $c \in \mathbb{R}$ *and* $\nu \in N$, $\nu(l_c) = 0$.
(z2) *There exists* t_0 *such that for any* $t \geq 1$ *there is* $0 \leq s \leq t_0$ *and some* $1 \leq j \leq k$ *with* $\nu^{(t+s)_j} \in N_0$.

Then

$$\mathrm{Prob}\big(X_t \longrightarrow \widehat{A}^\star,\ as\ t \longrightarrow \infty\big) = 1. \tag{26}$$

Proof

Step 1. Define the algorithm $\tilde{X}_t : \Omega \longrightarrow A$ as

$$\tilde{X}_0 := X_0^1 \quad \text{and} \quad \tilde{X}_t = T\big(\tilde{X}_{t-1}, Y_t^1, \ldots, Y_t^k\big),$$

where $T(x^1, \ldots, x^{k+1}) = x^i$ with $f(x^i) = \min_{j=1,\ldots,k+1} f(x^j)$. We now apply Theorem 1 to $\tilde{X}_t$. Here, $B = A^k$ and $U = \{(T, \nu) : \nu \in N^k\}$, $U_0 = \{(T, \nu) \in U : \nu^j \in N_0$ for some $j \leq k\}$. Note that U is compact and U_0 is its closed subset.
Now, since the condition (D) implies (B) and (C1), it is enough to show that the conditions (z1), (z2) imply all the conditions (A), (C2), (U0) and (D). Let $u = (T, \nu) \in U$ be fixed.
We prove (A). Fix $x_0 \in A$. Let

$$D_T'(x_0) := \left\{ y \in A^k : \forall i\ y^i \notin \left(l_{f(x_0)} \cup \bigcup_{j \neq i} l_{f(y^j)} \right) \right\} \quad \text{and}$$

$$D_T(x_0) := A^k \setminus D_T'(x_0).$$

We will show $\nu(D_T'(x_0)) = 1$. For $k = 1$ it is immediate by (z1). For $k > 1$, let $D_T'(x_0, y^1, \ldots, y^{k-1}) := A \setminus (l_{f(x_0)} \cup \bigcup_{j=1,\ldots,k-1} l_{f(y^j)})$. The condition (z1) implies that $\nu^k(D_T'(x_0, y^1, \ldots, y^{k-1})) = 1$ for any $y^1, \ldots, y^{k-1}$, and consequently, by Fubini's theorem, $(\nu^1 \times \cdots \times \nu^k)(D_T'(x_0)) = 1$. Fix $y_0 \in D_T'(x_0)$. Let $A \times A^k \ni (x_n, y_n) \longrightarrow (x_0, y_0)$. By the continuity of f, there is $n_0 > 0$, such that for all $n > n_0$, in the vector $(x_n, y_n^1, \ldots, y_n^k) \in A^{k+1}$, the point with the smallest value of f remains at the same position, as the point $T(x_0, y_0)$ in the vector $(x_0, y_0^1, \ldots, y_0^k)$. Consequently, $T(x_n, y_n) \longrightarrow T(x_0, y_0)$.
The condition (D) follows from the definition of T. We prove (C2). Recall that $U_0 = \{(T, \nu) \in U : \nu^j \in N_0$ for some $1 \leq j \leq k\}$ is a closed subset of U. Let $(T, \nu) \in U_0$ and $x \notin A^*$. By continuity of f and compactness of $A^\star$, there is an open neighborhood G of A^* with $f(z) < f(x)$ for all $z \in G$. By the definition of U_0, we have $\nu^j(G) > 0$ for some $j \leq k$. Recall that $f(T(x, y)) \leq f(x)$ for any $y \in A^k$, and note that $f(T(x, y)) < f(x)$ for any y from the set $\{y \in A^k\colon\ y^j \in G\}$ of positive measure, since $\nu^j(G) > 0$. Hence, (C2) is satisfied.
The condition (U0) follows immediately from (z2) and the definition of U_0.
Hence, by Theorem 1,

$$\mathrm{Prob}\big(\tilde{X}_t \longrightarrow A^\star,\ t \longrightarrow \infty\big) = 1. \tag{27}$$

Step 2. By simple induction, it is easy to see that

$$\widehat{f}(X_t) \leq f(\tilde{X}_t), \tag{28}$$

for any t. By the compactness of $A^\star$ and the continuity of f, (27) implies that $\text{Prob}(f(X_t) \searrow 0) = 1$. Hence, by (28), $\text{Prob}(\widehat{f}(X_t) \searrow 0) = 1$ and equivalently, by the compactness of A^m and the continuity of $\widehat{f}$, $\text{Prob}(X_t \longrightarrow \widehat{A}^\star,\ t \longrightarrow \infty) = 1$. □

7

In this section we suggest an algorithm working according to the iterated function system scheme. The advantage is that we may also admit methods which do not satisfy the condition (D), like mutations, but still the optimization process would be stochastically convergent to $A^\star$. The proposed scheme might be as follows. Let $U = \{(T_i, \nu_i), i = 1, \dots, m\} \subset \mathcal{T} \times M$ be finite and let $p_1, \dots, p_m$, $p_i > 0$, $\sum_{i=1}^m p_i = 1$, be a distribution on $\{1, \dots, m\}$. Let μ_0 be a probability measure on A:

1. Generate X_0 from the distribution μ_0. Let $t = 1$.
2. Draw i according to the distribution $p_1, \dots, p_m$.
3. Generate Z_i from the distribution ν_i.
4. Put $X_t = T_i(X_{t-1}, Z_i)$.
5. Increase t, $t := t + 1$ and go to 2.

Theorem 7 *Assume that for any $u = (T, \nu) \in U$ the conditions* (A), (B) *hold true and*

(C) *for any $x \in A \setminus A^\star$:*

$$\sum_{i=1}^{m} p_i \int_B f\bigl(T_i(x, y)\bigr)\nu_i(\mathrm{d}y) < f(x). \tag{29}$$

Let $\{X_t\}$ be the sequence defined by the above algorithm. Then, for every $\varepsilon > 0$:

$$\lim_{t\to\infty} \text{Prob}\bigl(\text{dist}\bigl(X_t, A^\star\bigr) < \varepsilon\bigr) = 1. \tag{30}$$

Proof Define $\bar{B} = \{1, \dots, m\} \times A$,

$$\bar{T} : A \times \bar{B} \ni \bigl(x, (i, y)\bigr) \longrightarrow T_i(x, y) \in A$$

and let $\bar{\nu}$ be the distribution on $\bar{B}$ defined by $\bar{\nu}(i, C) = p_i \cdot \nu_i(C)$, for any $i \in \{1, \dots, m\}$ and any Borel set $C \subset A$. Let $\bar{U} = \{(\bar{T}, \bar{\nu})\}$. Now it is enough to note that $\bar{U}$ satisfies conditions (A) and (B) of Theorem 1 and condition (C) of Remark 3. □

If one wants to use an algorithm like the one above, and one knows that some operators T and measures ν satisfy the strong inequality $\int_A f(T(x, y))\nu(\mathrm{d}y) < f(x)$ for all $x \in A \setminus A^\star$ (for example, the pure random search satisfies the inequality for any continuous function), then one can take advantage of the fact that all the integrals $\int_A f(T(x, y))\nu(\mathrm{d}y)$ are bounded from above by $M_0 = \sup_A |f|$ and then choose a distribution $p_1, \dots, p_m$ in such a manner that (29) is satisfied, and so X_t converges.

Let us note that in the case that the T_i do not depend on y, i.e. we do not perform Step 3 in the algorithm, then the condition (29) takes the form

$$\sum_{i=1}^{m} p_i f\big(T_i(x)\big) < f(x),$$

which is even easier to interpret.

References

Ahrari, A., Atai, A.A.: Grenade explosion method—a novel tool for optimization of multimodal functions. Appl. Soft Comput. **10**, 1132–1140 (2010)
Ahrari, A., Shariat-Panahi, M., Atai, A.A.: GEM method: a novel evolutionary optimization with improved neighborhood search. Appl. Math. Comput. **210**, 376–386 (2009)
Appel, M.J., Labarre, R., Radulovic, D.: On accelerated random search. SIAM J. Optim. **14**(3), 708–731 (2003)
Billingsley, P.: Convergence of Probability Measures, second edn. Wiley-Interscience, New York (1999)
Borovkov, A.A., Yurinsky, V.: Ergodicity and Stability of Stochastic Processes. Wiley, New York (1998)
Diaconis, P., Freedman, D.: Iterated random functions. SIAM Rev. **41**(3), 45–76 (1999)
Lasota, A., Mackey, M.: Chaos, Fractals and Noise. Springer, Berlin (1994)
Ljung, L., Pflug, G.C., Walk, H.: Stochastic Approximation and Optimization of Random Systems. DMV Seminar, vol. 17. Birkhäuser, Basel (1992)
Meyn, S., Twedie, R.: Markov Chains: Stochastic Stability. Springer, Berlin (1993)
Ombach, J.: A proof of convergence of general stochastic search for global minimum. J. Differ. Equ. Appl. **13**, 795–802 (2007)
Ombach, J.: Stability of evolutionary algorithms. J. Math. Anal. Appl. **342**, 326–333 (2008)
Parthasarathy, K.: Probability Measures on Metric Spaces. AMS, New York (2005)
Poli, R., Kennedy, J., Blackwell, T.: Particle swarm optimization. Swarm Intell. **1**, 33–57 (2007)
Radwański, M.: Convergence of nonautonomous evolutionary algorithm. Univ. Iagell. Acta Math. **45**, 197–206 (2007)
Robert, Ch., Casella, G.: Monte Carlo Statistical Methods. Springer, Heidelberg (2004)
Wright, M.: The interior-point revolution in optimization: history, recent developments, and lasting consequences. Bull. Am. Math. Soc. **42**, 39–56 (2005)
Yang, R.L.: Convergence of the simulated annealing algorithm for continuous global optimization. J. Optim. Theory Appl. **104**, 691–716 (2000)