J Glob Optim (2017) 69:869–888
DOI 10.1007/s10898-017-0544-7

On the convergence rate issues of general Markov search for global minimum

Dawid Tarłowski[1]

Received: 26 November 2016 / Accepted: 29 June 2017 / Published online: 7 July 2017

Abstract This paper focuses on the convergence rate problem of general Markov search for global minimum. Many of existing methods are designed for overcoming a very hard problem which is how to efficiently localize and approximate the global minimum of the multimodal function f while all information which can be used are the f-values evaluated for generated points. Because such methods use poor information on f, the following problem may occur: the closer to the optimum, the harder to generate a "better" (in sense of the cost function) state. This paper explores this issue on theoretical basis. To do so the concept of lazy convergence for a globally convergent method is introduced: a globally convergent method is called lazy if the probability of generating a better state from one step to another goes to zero with time. Such issue is the cause of very undesired convergence properties. This paper shows when an optimization method has to be lazy and the presented general results cover, in particular, the class of simulated annealing algorithms and monotone random search. Furthermore, some attention is put on accelerated random search and evolution strategies.

Keywords Global optimization · Convergence rate · Markov search · Simulated annealing · Accelerated random search · Self-adaptation

Mathematics Subject Classification Primary 60J05; Secondary 60J2 · 93D05 · 93D20

1 Introduction

Let (A, d) be a separable metric space and let $f\colon A \to [0, \infty)$ be a Borel measurable function having its global minimum $f^\star = \min f(A)$. There is a great number of iterative numerical methods which are used for finding a global minimum of f in case the global minimization problem

✉ Dawid Tarłowski
dawid.tarlowski@im.uj.edu.pl; dawid.tarlowski@gmail.com

[1] Faculty of Mathematics and Computer Science, Institute of Mathematics, Jagiellonian University, Łojasiewicza 6, 30 348 Kraków, Poland

$$\min_{x\in A} f(x)$$

cannot be solved analytically. Many of those iterative techniques are designed for solving difficult, irregular or multimodal, real world problems. This paper focuses on the class of Markov methods which, as A is assumed to be separable, admit the following general representation

$$X_{t+1} = T_t(X_t, Y_t),$$

where Y_t is an independent sequence and independent of X_0, see [8]. We will say that X_t is globally convergent if it converges stochastically to $A^\star$. It is often an easy task to examine the global convergence property of such methods on theoretical basis (of course there are exceptions, especially in case of self-adaptive methods) and general techniques are based, in particular, on Borel–Cantelli lemma, classical probability theory [24,31], Lyapunov functions [2,22,30,32,33], and Markov chains [1,4,12]. At the same time, the theoretical convergence rate analysis is usually extremely difficult. The convergence rate analysis must take into account the optimization scheme, the initial parameters of the given procedure and the appropriate properties of the given problem function (in general, the function can be multimodal and complex which strongly determines the algorithm efficiency, [23,37]). While it is justified theoretically that gradient based local search methods are fast [21], the existing theoretical results regarding derivative-free global random search techniques usually indicate slow convergence rate or concern some special cases—for instance, in many cases of convex optimization the derivative-free methods may efficiently use gradient-estimates, see [10,11]. However, many global random search methods are designed for overcoming a general and almost impossible problem which is how to efficiently localize and approximate the global minimum of a multimodal function while all information which can be used are the f-values evaluated for generated points. Furthermore, the global minimal value is usually unknown. The derivative may not exist or may be unavailable (for instance, in case of so called "black box " problems, usually all one have is the possibility of compute the value $f(x)$ at given state $x \in A$ and this computation often requires much effort), and hence many methods belong to the class of derivative-free algorithms, [27]. Because the given method uses poor information on f, its convergence may have very undesired properties based on the following issue: the closer to the optimum, the harder to generate a "better" (in sense of the cost function) state. This paper explores this issue on theoretical basis. To do so the concept of lazy convergence for a globally convergent method is introduced: a globally convergent method is called lazy if the probability of generating a better state from one step to another goes to zero with time. It is shown, in particular, that a monotonic method X_t [in sense $f(X_{t+1}) \leq f(X_t)$] is lazy iff for any $k \in \mathbb{N}$ we have $P(X_{t+k} = X_{t+k-1} = \cdots = X_t) \overset{t\to\infty}{\longrightarrow} 1$ and the expected length k of constant finite subsequences $X_t = X_{t+1} = \cdots = X_{t+k}$ goes to infinity with time t. The above property is extended to the case of nonmonotonic methods as the property of the corresponding best iterate sequence. This paper shows when an optimization method is lazy and the presented general results cover, in particular, the class of simulated annealing algorithms and monotone random search. To provide an application example from the class of methods which are based on parameters' self-adaptation it is shown that the finite descent Accelerated Random Search [3], converges lazily. As it is discussed further, the undesired lazy convergence property appears to be the property of an optimization method rather than the property of the problem function f. The author of this paper believes that the methods based on parameters' self-adaptation may be a good way to overcame the convergence issues presented here and the last, additional, chapter of this paper focuses on this class of methods. Finally, it is worth to mention that this paper is about

the convergence behaviour of the given optimization method X_t as the method approaches the global minimum. The alternative convergence problem is also the object of analysis in literature: given $\varepsilon > 0$, how to analyze the expected time of hitting the ε- neigbhourhood of global minimum by the given optimization method X_t and, in particular, how this time changes as ε goes to zero, see [35]. Those two convergence research aspects are based on different approaches and one of the reasons for that is that in the first case the target (the global minimum) usually has a zero Lebeasgue'a measure and in the second case the target (the ε-ball) has positive Lebesgue'a measure.

This paper is organized as follows. Section 2 introduces the general assumptions and corresponding notation, and next it introduces and discusses the concept of lazy convergence. Section 3 presents general results for the class of comparison based monotone homogeneous Markov search. Section 4 successively develops the framework of Sect. 3 up to the full generality. Both sections discuss the results and present the corresponding illustrative examples. The main result of Sect. 4 is Theorem 5 which cover, in particular, the class of Simulated Annealing algorithms and monotone random search. Additionally, the lazy convergence of finite descent ARS is provided as the conclusion of Theorem 3. The last section is an additional chapter which shortly indicates that the self-adaptive methods may be a good way to overcome the issues analysed here.

2 General assumptions and lazy convergence

This section presents general assumptions and notation which will hold throughout the paper. Next it introduces and discusses the concept of lazy convergence.

2.1 General assumptions

We assume that $A \subset \mathbb{R}^d$. The presented methodology can be extended to more general spaces however the full generality is not a purpose of this paper as the clarity of the presented ideas is more important. We will assume that the metric d on A either is a metric for the Euclidean topology or, in the case $A = I_d := (0, 1]^d$, is the d-dimensional torus metric d_T given by:

$$d_T(x, y) = \max_{i=1,\ldots,d} \min\{|x_i - y_i|, 1 - |x_i - y_i|\}.$$

We will always assume that

$$f^\star = \min f(A) = 0.$$

Now we introduce general notation:

(1) $A^\star = \{x \in A\colon\ f(x) = 0\}$,
(2) $A_\delta = \{x \in A\colon f(x) < \delta\}$, where $\delta > 0$,
(3) $A^\star(\varepsilon) = B(A^\star, \varepsilon) = \bigcup_{a \in A^\star} B(a, \varepsilon)$, where $\varepsilon > 0$

We will always assume that the measurable problem function $f\colon A \to [0, \infty)$ satisfies the following natural conditions:

(A1) $\forall \varepsilon > 0\ \exists \delta > 0\ A_\delta \subset A^\star(\varepsilon)$,
(A2) $\forall \delta > 0\ \exists \varepsilon > 0\ A^\star(\varepsilon) \subset A_\delta$.

Condition $(A1)$ means that for any $\varepsilon > 0$ we have

$$\inf_{x \notin A^\star(\varepsilon)} f(x) > 0.$$

Condition $(A2)$ is satisfied, for example, if the set of global minimums $A^\star$ is finite and the function f is continuous at points from $A^\star$. If $A^\star$ is infinite, condition $(A2)$ still holds true if for some $\varepsilon > 0$ the set $\overline{A^\star(\varepsilon)}$ is compact and f is continuous on $\overline{A^\star(\varepsilon)}$. Under conditions $(A1)$, $(A2)$, for any sequence $x_t \in A$ we have:

$$x_t \to A^\star \iff f(x_t) \to 0.$$

Let (Ω, Σ, P) be a probability space and let $\{X_t\}_{t=0}^{\infty}$ be a measurable sequence which represents the successive states of the given optimization method. Under (A1) and (A2) different types of basic global convergence modes are equivalent, see Observation 1 and Theorem 1 in [33]. For instance, it is easy to see that under (A1) and (A2) the following conditions are equivalent:

(B1) $f(X_t) \to 0$ in probability
(B2) $d(X_t, A^\star) \to 0$ in probability

In this paper we will say that an optimization method X_t is globally convergent (has a global convergence property) if it satisfies conditions (B1), (B2). Condition (B2) represents, in fact, the stochastic convergence of X_t to $A^\star$ and thus the global convergence of X_t will be denoted by

$$X_t \xrightarrow{s} A^\star$$

2.2 Lazy convergence

The aim of this paper is to explain on the theoretical basis why many global search methods cannot be convergent quickly. The main attention is paid on methods with global convergence property. Still, the general results of next sections cover the case of methods which are not necessarily convergent towards $A^\star$. Below we introduce the definition of lazy convergence which expresses the undesired convergence behaviour of many random search techniques.

Definition 1 We will say that a globally convergent sequence X_t *converges lazily towards* $A^\star$, or shortly that X_t *is lazy*, if it satisfies $\lim_{t\to\infty} P(f(X_{t+1}) < f(X_t)) = 0$. It will be denoted by $X_t \overset{l-s}{\to} A^\star$

Proposition 1 presents some rather straightforward consequences of this definition. Theorem 1 provides proper intuition behind this concept and we believe it explains the use of term "lazy" for this convergence type.

Proposition 1 *Assume that* $X_t \overset{l-s}{\to} A^\star$. *We have:*

(C1) *for any* $k \in \mathbb{N}$, $\lim_{t\to\infty} P(f(X_{t+k}) < f(X_t)) = 0$
(C2) *for any* $k \in \mathbb{N}$, $\liminf_{t\to\infty} E\left(\frac{f(X_{t+k})}{f(X_t)}\right) \geq 1$.
(C3) $E\tau_{X_n} \longrightarrow \infty$, *where* $\tau_{X_n} = \inf\{k \in \mathbb{N}:\ f(X_{n+k}) < f(X_n)\}$.

Proof Let $\Omega \supset C_t := \{f(X_{t+1}) \geq f(X_t)\}$, $t \in \mathbb{N}$. We have $P(C_t) \to 1$ and thus for any $k \in \mathbb{N}$,

$$\lim_{t\to\infty} P(C_{t+k} \cap C_{t+k-1} \cap \cdots \cap C_t) = 1. \tag{2.1}$$

To see condition (C1) it remains to note that

$$\{f(X_{t+k}) \geq f(X_t)\} \supset C_{t+k} \cap C_{t+k-1} \cap \cdots \cap C_t, \quad k \in \mathbb{N}.$$

To see (C2) note that for any $k \in \mathbb{N}$, based on condition (C1),

$$E\left(\frac{f(X_{t+k})}{f(X_t)}\right) \geq \int\limits_{\{f(X_{t+k}) \geq f(X_t)\}} \frac{f(X_{t+k})}{f(X_t)} dP \geq P\left[f(X_{t+k}) \geq f(X_t)\right] \overset{t\to\infty}{\longrightarrow} 1.$$

To see (C3) note that for any $n \in \mathbb{N}$ and $M \in \mathbb{N} \setminus \{0\}$, from the definition of τ_{X_n} we have $\{\tau_{X_n} > M\} \supset \{C_{n+M-1} \cap \cdots \cap C_n\}$, and thus

$$E\tau_{X_n} \geq M \cdot P(\tau_{X_n} > M) \geq M \cdot P(C_{n+M-1} \cap \cdots \cap C_n).$$

Hence, from (2.1), $\lim\limits_{n\to\infty} E\tau_{X_n} \geq M$. This finishes the proof as M can be arbitrarily big. □

From (C1) it follows, in particular, that for a monotonic sequence (in sense $f(X_{t+1}) \leq f(X_t)$), we have that for any $k \in \mathbb{N}$,

$$P(X_{t+k} = X_{t+k-1} = \cdots = X_t) \overset{t\to\infty}{\longrightarrow} 1,$$

and that the expected length of constant finite subsequences goes to infinity with time t (condition (C3)). If the method X_t is not monotonic then we can consider the associated current best iterate sequence $\hat{X}_t$ given by

$$\hat{X}_t = X_{k_t}, \text{ where } k_t = \min\{i \leq t\colon f(X_i) = \min_{k=0,\ldots,t} f(X_k)\}.$$

It is an easy observation that if the sequence X_t converges lazily towards $A^\star$ then the current best iterate $\hat{X}_t$ is a monotonic sequence which converges lazily towards $A^\star$. In fact, we have that if $f(\hat{X}_{t+1}) < f(\hat{X}_t)$ then $f(X_{t+1}) < f(X_t)$ and thus $P[f(\hat{X}_{t+1}) < f(\hat{X}_t)] \leq P[f(X_{t+1}) < f(X_t)]$, $t \in \mathbb{N}$. The below theorem presents the properties of lazy convergence which provides the proper intuition behind this notion.

Theorem 1 *If the sequence X_t converges lazily towards $A^\star$ then the associated best iterate sequence $\hat{X}_t$ satisfies*

$$P(\hat{X}_{t+k} = \hat{X}_{t+k-1} = \cdots = \hat{X}_t) \overset{t\to\infty}{\longrightarrow} 1, \quad \textit{for any } k \in \mathbb{N},$$

and the expected lenght k of constant finite subsequences $\hat{X}_t = \hat{X}_{t+1} = \cdots = \hat{X}_{t+k}$ goes to infinity with time t.

The above rather simple result gives some insight into the properties of the stopping conditions for the class of lazy methods. Fix $k, n \in \mathbb{N}$ with $n \geq k$ and let $h\colon A \to [0, \infty)$ be the given function. Consider the stopping criterion $\tau_{(h,n,k)}$ given by $\tau_{(h,n,k)} = \tau_{(h,k)}(X_n) = \inf\{m \geq n\colon f(\hat{X}_m) \geq f(\hat{X}_{m-k}) - h(\hat{X}_{m-k})\}$, so the variable $X_{\tau_{(h,n,k)}}$ represents the outcome of the process which performs at least n iterations and next it stops when the value of the improvement during the last k steps does not exceed the value determined by the function h. The stopping condition will take the form $f(\hat{X}_m) \geq f(\hat{X}_{m-k}) - \varepsilon$ for $h = \varepsilon$ or $\frac{f(\hat{X}_m) - f(\hat{X}_{m-k})}{f(\hat{X}_{m-k})} \geq \varepsilon$ for $h(x) = \varepsilon \cdot f(x) + f(x)$, where $\varepsilon > 0$. Theorem 1 immediately implies thae following observation.

Observation 1 *If we have $X_t \overset{l-s}{\to} 0$, then for any $k \in \mathbb{N}$ and $h\colon A \to [0, \infty)$ we have* $\lim\limits_{n\to\infty} P(X_{\tau_{(h,n,k)}} = X_n) = 1$.

3 Monotone homogeneous Markov search

This section presents the general result for comparison based monotone homogeneous Markov search. This class of methods was an initial motivation for the research presented in this paper. The methodology of this section is extended to the general case of inhomogeneous Markov search techniques in next section.

3.1 Illustrative examples

First we will discuss some illustrative examples. For now, to clarify the presentation, we will assume the most natural case when $A^\star = \{a\}$ is a singleton. We will focus on the class of monotonic homogeneous random search methods which can be described as follows. Given the current state x_t the algorithm samples a candidate for the next step q_t from the probability kernel $P_Q(x, \cdot)$ which depends on the point $x = x_t$. The new candidate is chosen as the next state x_{t+1} if it is "better" than the current state so we have $f(x_{t+1}) = \min\{f(x_t), f(q_t)\}$. From the theoretical perspective this scheme admits the following general representation:

$$X_{t+1} = \begin{cases} Q(X_t, Y_t) & \text{if } f(Q(X_t, Y_t)) < f(X_t) \\ X_t & \text{if } f(Q(X_t, Y_t)) \geq f(X_t) \end{cases}, \tag{3.1}$$

where:

- $Q\colon A \times B \longrightarrow A$ is Borel measurable and (B, d_B) is a separable metric space
- $Y_t\colon \Omega \to B, t \in \mathbb{N}$ are i.i.d. random variables and independent of the initial state X_0

We will sometimes use the following more compact form

$$X_{t+1} = T(X_t, Y_t), \tag{3.2}$$

where the mapping T is uniquely defined by the equation (3.1). We also denote

$$P_Q(x, C) := P(Q(x, Y_1) \in C) \quad \text{, for any Borel set } C \subset A. \tag{3.3}$$

To give some simple examples: if P_Q does not depend on x then $P_Q(x, C) = P_Q(C)$ is a probability measure on A and a method (3.1) represents PRS algorithm (if A is bounded then P_Q is usually defined as uniform distribution on A). In case $A = \mathbb{R}^d$ another simple example of P_Q is normal distribution centered at x with some covariance matrix Σ: $P_Q(x, \cdot) = N(x, \Sigma)$.

As we will show later, the following natural property of the probability kernel P_Q is the cause of insufficient convergence behaviour of methods (3.1):

$$\sup_{x \in B(A^\star, \varepsilon)} P_Q(x, (B(A^\star, \varepsilon)) \searrow 0 \quad \text{as } \varepsilon \searrow 0. \tag{$\star$}$$

In the present case $A = \{a\}$ the ($\star$) condition states for $\sup_{x \in B(a,\varepsilon)} P_Q(x, (B(a, \epsilon)) \searrow 0$ as $\varepsilon \searrow 0$, of course. To provide some intuition for the commonness of the ($\star$) property we will shortly discuss some examples. Consider for a moment the class of Markov monotone symmetric search methods which was analysed in papers [34–36]. Methods from this class are natural for spaces $(A, d) = (\mathbb{R}^d, |\cdot|)$ and $(A, d) = (I_d, d_T)$ which exclude boundary issues connected to defining symmetric densities. Those methods satisfy the general scheme (3.1) and the candidate points are sampled from some density $p(x_t, \cdot)$ on $(B, d_B) = (A, d)$, where x_t is the current state, and the $p(x_t, y)$ is a nonincreasing function of the distance between x_t and y. We thus have

$$p(x_t, y) = h(d(x_t, y)) \tag{3.4}$$

for some nonincreasing function $h\colon (0,\infty) \to [0,\infty)$ which satisfies the normalization condition $\int_A h(d(x_t, y))dy = 1$. Assume for now that the algorithm (3.1) satisfies the above symmetry condition (3.4). In case $A = \mathbb{R}^d$ one can consider, for example: sampling from the normal distribution $P_Q(x,\cdot) = N(x, \sigma \cdot I)$ or from the uniform distribution $P_Q(x,\cdot) = U(B(x,R))$ [after little modifications those two examples work also for $(A,d) = (I_d, d_T)$]. Condition (3.4) implies that for any $\varepsilon > 0$ and $x \in A$,

$$P_Q(x, B(x,\epsilon)) = \varphi(\varepsilon) \tag{3.5}$$

for some function $\varphi : (0,\infty) \to [0,1]$. From the continuity property of a probability measure it follows immediately that $\varphi(\varepsilon) \searrow 0$ as $\varepsilon \searrow 0$. This implies that $(\star)$ condition is satisfied. In fact, for any $x \in B(a,\epsilon)$, if $Q(x, Y_t) \in B(a,\epsilon)$ then $Q(x, Y_t) \in B(x, 2\epsilon)$ and thus that we have

$$\sup_{x\in B(a,\epsilon)} P_Q(x, (B(a,\epsilon)) \le \sup_{x\in B(a,\epsilon)} P_Q(x, (B(x,2\epsilon))) \le \varphi(2\varepsilon) \searrow 0 \text{ as } \varepsilon \searrow 0.$$

Note that we did not put any assumptions on the extremum $a \in A$ in the above case so the analogous condition holds true for any point. Now we are going back to the situation (3.1) (no symmetries assumed). Assume for now that the A is a closed subset of $\mathbb{R}^d$ with the induced euclidean metric. Property (3.8) is too strong to be satisfied for a bounded domain situation because of the issues of the boundary regions. Still, some modifications of it will hold true. For example, consider the class of methods that generate a candidate point from some distribution on $\mathbb{R}^n$ around the current position x_t and next if the candidate is created outside of the set of admissible solutions then it is taken back to the boundary of this set according to some procedure. This mechanism causes an efficient search of the boundary of the domain. To see this assume for a moment that the algorithm (3.1) satisfies:

(1) $Y_t\colon \Omega \to A$ are such that $P(Y_t = 0) = 0$
(2) $Q(x,y) = x + y$ if $x + y \in A$
(3) $Q(x,y) \in \partial A$ if $x + y \notin A$

If Y_t is centered at $0 \in \mathbb{R}^d$ then the above almost explicit form of Q is natural. The only assumption on Y_t is very natural too as there is not any sens in generating a candidate equal to the current state. Note that there is a nonnegative valued function φ with $\varphi(\varepsilon) \searrow 0$ as $\varepsilon \searrow 0$ such that for any x from the interior of A and for any $\varepsilon > 0$ we have:

$$B(x,\varepsilon) \subset A \Longrightarrow P_Q(x, B(x,\epsilon)) = \varphi(\varepsilon).$$

In fact, we have

$$K_{\mathbb{R}^n}(x,\epsilon) \subset A \Longrightarrow \mathbb{P}(Q(x, Y_t) \in K_{\mathbb{R}^n}(x,\epsilon)) = \mathbb{P}(Y_t \in B(0,\epsilon)) =: \varphi(\epsilon).$$

Now we can repeat the previous argumentation to obtain that if the global minimum a belongs to the interior of A then it satisfies condition $(\star)$. Methods with more sophisticated rules of taking back a candidate to the admissible domain also satisfy the $(\star)$ condition under natural assumptions and proving that would be based more or less on the same idea regarding the algorithm behaviour: the closer to the optimum, the harder to sample an appropriate candidate. Below we start the general theoretical justification of this issue.

3.2 Theory

From now on we release the assumption that $A^\star$ is a singleton and we assume that the set $A^\star$ is compact instead. Recall that condition $(\star)$ takes the following form:

$$\sup_{x\in B(A^\star,\epsilon)} P_Q(x, B(A^\star,\epsilon)) \searrow 0 \quad \text{as } \varepsilon \searrow 0. \qquad (\star)$$

and note that under conditions (A1), (A2) this is equivalent to the following condition:

$$\sup_{x\in A_\delta} P_Q(x, A_\delta) \searrow 0 \quad \text{as } \delta \searrow 0.$$

Let $B(A)$ denote the family of Borel subsets of A and let $\mathcal{M}^1(A)$ denote the topological space of Borel probability measures on A with the weak convergence topology, see [7] or [13] for the general theory. Let us recall that

$$\mathcal{M}^1(A) \ni \mu_t \to \mu \in \mathcal{M}^1(A) \text{ weakly iff } \limsup_{t\to\infty} \mu_t(\overline{C}) \le \mu(\overline{C}),\ C \in B(A). \qquad (3.6)$$

As a direct consequence of Proposition 5 from the next section (i.e. from Conclusion 2 stated there) we will have that if P_Q satisfies two conditions:

- $\sup_{a\in A^\star} P_Q(a, A^\star) = 0$,
- there is an open neigbourhood U of $A^\star$ such the function $U \ni x \longrightarrow P_Q(x,) \in \mathcal{M}^1(A)$ is continuous,

then P_Q satisfies $(\star)$ condition. Note that the assumption $P_Q(x, A^\star) = 0, x \in A$, is satisfied, for example, if $A^\star$ has zero Lebesgue'a measure and the distributions $P_Q(x, \cdot)$ are absolutely continuous.

Below we present the main result of this section. For any $\delta > 0$, if $P(X_t \in A_\delta) = 0$ then we simply put $P(f(X_{t+1}) < f(X_t)|X_t \in A_\delta) := 0$.

Theorem 2 *Assume that X_t is a method of the form (3.1) such that condition $(\star)$ is satisfied. Then, for any $0 < C < 1$ there is $\delta > 0$ such that for any $x \in A_\delta$ we have*

$$P(f(T(x, Y_t)) < f(x)) < 1 - C. \qquad (3.7)$$

Furthermore, we have

$$\lim_{\delta\to 0} \sup_{t\in\mathbb{N}} P(f(X_{t+1}) < f(X_t)|X_t \in A_\delta) = 0$$

and thus, if $X_t \xrightarrow{s} A^\star$ then $X_t \xrightarrow{l-s} 0$.

Proof First, note that from the construction of the algorithm we have that X_t and Y_t are independent and hence, from the Fubini's theorem:

- $P[f(X_{t+1}) < f(X_t)|X_t = x] = P[f(T(x, Y_t)) < f(x)]$
- $P[f(X_{t+1} < f(X_t)] = \int_A P[f(T(x, Y_t)) < f(x)] P_{X_t}(dx)$,
- $P[\{f(X_{t+1} < f(X_t)\} \cap X_t \in D] = \int_D P[f(T(x, Y_t)) < f(x)] P_{X_t}(dx)$.

Fix $C \in (0, 1)$. Let

$$\varphi(\varepsilon) := \sup_{x\in B(A^\star,\varepsilon)} P_Q(x, B(A^\star, \varepsilon)).$$

From ($\star$) it follows that there is $\varepsilon > 0$ with $\varphi(\varepsilon) < 1 - C$ and from ($A1$) it follows that there is $\delta > 0$ with $A_\delta \subset B(A^\star, \varepsilon)$. For any $x \in A_\delta$,

$$P(f(T(x, Y_t)) < f(x)) = P(f(Q(x, Y_t)) < f(x)) \leq P(f(Q(x, Y_t) < \delta) = P_Q(x, A_\delta).$$

The constants $\delta > 0$ and $\varepsilon > 0$ are chosen in such a way that for any $x \in A_\delta$ we have:

$$P_Q(x, A_\delta) \leq P_Q(x, B(x, \varepsilon)) \leq \varphi(\varepsilon) < 1 - C. \tag{3.8}$$

The above proves (3.7). Now, fix $C \in (0, 1)$ and let $\delta > 0$ be small enough to have condition (3.7) satisfied. Using Fubini's theorem we obtain that for any $t \in \mathbb{N}$ with $P_{X_t}(A_\delta) > 0$ we have:

$$P[f(X_{t+1}) < f(X_t) | X_t \in A_\delta] = \frac{\int_{A_\delta} P[f(T(x, Y_t) < f(x)] P_{X_t}(dx)}{P_{X_t}(A_\delta)} < \\ < \frac{(1 - C) \cdot P_{X_t}[A_\delta]}{P_{X_t}[A_\delta]} = 1 - C.$$

The above proves $\lim_{\delta \to 0} \sup_{t \in \mathbb{N}} P(f(X_{t+1}) < f(X_t) | X_t \in A_\delta) = 0$. Hence, if X_t is globally convergent and thus for any $\delta > 0$ it satisfies

$$P[f(X_{t+1}) < f(X_t)] - P[f(X_{t+1}) < f(X_t) | X_t \in A_\delta] \overset{t \to \infty}{\longrightarrow} 0,$$

then we must have $\lim_{t \to \infty} P[f(X_{t+1}) < f(X_t)] = 0$ (alternatively, condition $\lim_{t \to \infty} P[f(X_{t+1}) < f(X_t)] = 0$ can be nicely derived from (3.7) for a globally convergent method). □

Note that the undesired convergence properties expressed by the lazy convergence notion are in fact consequences of the algorithm general scheme and the ($\star$) property of the probability kernel P_Q. In fact, we practically did not put any assumptions on the problem function properties. Thus the methods of the form (3.1) are in some sense condemned for the "lazy convergence"–the information on the function f which is used by method (3.1) is insufficient to keep "good" convergence behaviour as the method approaches the extremum. The next section extends Theorem 2 to cover the general inhomogeneous case. More advanced examples will be presented.

4 General inhomogeneous Markov search

4.1 General case

From now we assume that the sequence X_t is given by the general recursive equation:

$$X_{t+1} = T_t(X_t, Y_t), \tag{4.1}$$

where the mappings $T_t \colon A \times B \to A$ and the distributions of $Y_t \colon \Omega \to B$ can change over time. Naturally, the sequence $X_0, Y_1, Y_2, \ldots$ is assumed to be independent.

We will say that a sequence X_t satisfies ($\diamond$) condition if it satisfies:

$$\sup_{t \in \mathbb{N}} \sup_{x \in A_\delta} P\left[f(T_t(x, Y_t)) < f(x)\right] \searrow 0 \text{ as } \delta \searrow 0. \tag{$\diamond$}$$

Proposition 2 *Assume that X_t is given by (4.1) and that condition ($\diamond$) is satisfied. We have*

$$\lim_{\delta\to 0}\sup_{t\in\mathbb{N}} P(f(X_{t+1}) < f(X_t)|X_t \in A_\delta) = 0.$$

In particular, if X_t is globally convergent, then it converges lazily towards $A^\star$.

Proof The proof follows from the proof of Theorem 2. In fact, from ($\diamond$) we have that for any $0 < C < 1$ there is $\delta_C > 0$ with

$$P(f(T_t(x, Y_t)) < f(x)) < 1 - C,\ t \in \mathbb{N}, x \in A_{\delta_C}.$$

The above is the inhomogeneous analogy of the inequality (3.7) from the proof of Theorem 2 and it has been shown in that proof that (3.7) implies the thesis in the homogeneous case–this part of the proof can be directly repeated in the present inhomogeneous case. □

4.2 Example: simulated annealing

Consider for a moment nonmonotone generalization of scheme (3.1). Assume that we have $\{Y_t^1 \colon \Omega \to B_1\}_{t\in\mathbb{N}}$ (B_1 is assumed to be a separable metric space) and $\{Y_t^2 \colon \Omega \to [0, 1]\}_{t\in\mathbb{N}}$ such that $\{(Y_t^1, Y_t^2)\}_{t\in\mathbb{N}}$ is i.i.d and all the variables Y_1^1, Y_1^2,Y_2^1, Y_2^2, ... are independent. Assume that the method X_t is given by the equation

$$X_{t+1} = T_t(X_t, Y_t^1, Y_t^2) \tag{4.2}$$

and such that for some $Q \colon A \times B_1 \to A$ we have:

$$f(T_t(X_t, Y_t)) \geq \min\{f(X_t), f(Q(X_t, Y_t^1))\}. \tag{4.3}$$

A good illustration of the above inequality is a method X_t which at every step t samples a candidate $Q(X_t, Y_t^1)$ and next this candidate is accepted with some probability $p_t(X_t, Y_t^1)$. Assume that Y_t^2 are uniformly distributed on [0, 1] and that we have measurable functions functions $p_t \colon A \times B_1 \to [0, 1]$, $t \in \mathbb{N}$. We see that the following scheme

$$X_{t+1} = \begin{cases} Q(X_t, Y_t^1) & \text{if } Y_t^2 \leq p_t(X_t, Y_t^1) \\ X_t & \text{if } Y_t^2 > p_t(X_t, Y_t^1) \end{cases}, \tag{4.4}$$

satisfies equation (4.2) and inequality (4.3). This scheme describes the well known Simulated Anenaling method (although the candidate distributions are constant over time, for now).

Proposition 3 *Assume that X_t is a method given by (4.2) which satisfies condition (4.3) for some mapping Q and such that the probability kernel P_Q (given by (3.3)) satisfies ($\star$) condition. Then X_t satisfies condition ($\diamond$) and thus, if X_t is globally convergent, then X_t converges lazily towards $A^\star$.*

Proof In fact, fix $\delta > 0$ and $x \in A_\delta$. We have

$$P[f(T_t(x, Y_t)) < f(x)] \leq P[f(Q(x, Y_1^1)) < f(x)] \leq P[f(Q(x, Y_1^1)) < \delta], \quad t \in \mathbb{N}.$$

Hence, to show that X_t satisfies ($\diamond$) it is enough to show that

$$\sup_{x\in A_\delta} P[f(Q(x, Y_1)) < \delta] \searrow 0 \text{ as } \delta \searrow 0.$$

This is equivalent to ($\star$) as f satisfies conditions $(A1)$ and $(A2)$. Proposition 2 finishes the proof. □

The conclusion below is a direct consequence of Proposition 3.

Conclusion 1 *Assume that X_t is a Simulated Annealing method given by (4.4) and that the candidate distribution $P_Q(x,\cdot)$ satisfies $(\star)$ condition. If $X_t \xrightarrow{s} A^\star$, then $X_t \xrightarrow{l-s} 0$.*

The most common acceptance probability p_t for the Simulated Annealing Algorithm given by (4.4) depends on the value of the difference $\Delta_t = f(X_t) - f(Q(X_t, Y_t^1))$ and on time t and is given by the Metropolis formula

$$p_t(\Delta_t) = \min\{1, \exp(-\frac{1}{\beta_t}\cdot\Delta_t)\},$$

where β_t is a sequence (so called "cooling schedule") with $\beta_t \to 0$. This formula causes that the good candidate Q_t [in sense: $f(Q_t) \le f(X_t)$] is always accepted while the candidate with $f(Q_t) > f(X_t)$ still have the positive acceptance probability equal to $\exp(-\frac{1}{\beta_t}\cdot\Delta_t)$. Various acceptance probability formulas (the most frequently analysed aspect is the convergence rate of β_t towards zero) have been analysed in the context of global convergence property. The various choices for p_t formula can help the SA method to avoid local minima. Many papers focus on methods (4.4) and the global convergence is achieved by various techniques, see [1,12] for applications of Markov chains theory or [16] for more classical approach. Theorem 5 in [33] gives the condition on the probability kernel P_Q under which the condition $\beta_t \to 0$ ensures the global convergence regardless of the convergence rate of β_t. However, as stated in Conclusion 1, regardless of the acceptance probabilities, the convergence of this method cannot be quick if the probability kernel for sampling a candidate is constant over time. The assumptions of this chapter allow the p_t to be dependent only on time t, X_t, and Y_t [and thus on $Q(X_t, Y_t^1)$, off course], however the extension of the presented results to more general case $p_t = p_t(X_0, \ldots, X_t, Y_t)$ is obvious.

Further in this section we will present a general result, Theorem 4, which cover, in particular, the class of SA algorithms for which the probability kernel for sampling a candidate may change in time.

4.3 General case

In order to provide the final result for the general case (4.1) we need to extend the notion of $(\star)$ condition to the case of family of probability kernels (Markov kernels). We say that a mapping $K\colon A \times B(A) \longrightarrow [0,1]$ is a probability kernel on A if it satisfies the following conditions:

(1) for any Borel set $B \in B(A)$ the function $x \longrightarrow B(x,B)$ is measurable,
(2) for any $x \in A$ the function $B \longrightarrow K(x,B)$ is a probability measure on Borel sets $B(A)$.

Let $\mathcal{K}(A)$ denote the set of probability kernels on A.

Definition 2 We will say that a family $\mathcal{C} \subset \mathcal{K}(A)$ satisfies $(\star\star)$ condition if any of the following equivalent (under (A1) and (A2)) conditions is satisfied:

$$\sup_{P_Q\in\mathcal{C}} \sup_{x\in B(A^\star,\varepsilon)} P_Q(x, B(A^\star,\varepsilon)) \searrow 0 \text{ as } \varepsilon \searrow 0, \tag{$\star\star$}$$

$$\sup_{P_Q\in\mathcal{C}} \sup_{x\in A_\delta} P_Q(x, A_\delta) \searrow 0 \text{ as } \delta \searrow 0. \tag{4.5}$$

If C is some metric space and the following mapping is given:

$$C \ni Q \longrightarrow P_Q \in \mathcal{K}(A)$$

then will say that the set C satisfies the $(\star\star)$ condition if the family of probability kernels $\{P_Q\}_{Q\in C} \subset \mathcal{K}(A)$ satisfies the $(\star\star)$ condition.

Set C can represent, for example, the set of parameters of a given method. It can be also the subset of $\mathcal{M}(A \times B, A) \times \mathcal{M}^1(B)$, where $\mathcal{M}(A \times B, A)$ denotes measurable functions $H\colon A \times B \to A$. Note that if the optimization method produces a candidate $Q_t(x_t, Y_t)$ at step t then the pair $Q = (Q_t, P_{Y_t}) \in \mathcal{M}(A \times B, A) \times \mathcal{M}^1(B)$ uniquely defines the corresponding probability kernel P_Q. We assume that the space $\mathcal{M}(A \times B, A)$ is equipped with the topology of uniform convergence of functions and $\mathcal{M}^1(B)$ has the topology of weak convergence of probability measures. Let

$$\mathcal{U} = \mathcal{M}(A \times B, A) \times \mathcal{M}^1(A) \tag{4.6}$$

denote the topological space equipped with the product topology. Let

$$P\colon \mathcal{U} \ni (Q, \nu) \longrightarrow P_{(Q,\nu)} \in \mathcal{K}(A)$$

be given by

$$P_{(Q,\nu)}(x, Z) = \nu\{y \in B\colon Q(x, y) \in Z\}, \quad x \in A, \quad Z \in \mathcal{B}(A). \tag{4.7}$$

Note that we have $P(Q_t(x, Y_t) \in Z) = P_{(Q_t, P_{Y_t})}(x, Z)$. The following characterization of continuity in this case is the consequence of Proposition 1 stated in [22].

Proposition 4 *Let $C \subset \mathcal{U}$ be such that for any pair $(Q, \nu) \in C$ and any $x \in A$*

$$\nu\{y \in B\colon Q \text{ is not continuous at } (x,y)\} = 0.$$

Then, the mapping

$$A \times C \ni (x, Q, \nu) \longrightarrow P_{(Q,\nu)}(x, \cdot) \in \mathcal{M}^1(A)$$

is continuous.

The following proposition provides exemplary sufficient conditions for $(\star\star)$ which are verifiable in practical cases. Some examples will be given further.

Proposition 5 *Let C be a compact metric space and $P\colon C \ni Q \longrightarrow P_Q \in \mathcal{K}(A)$ be given mapping such that for any $Q \in C$ we have $\sup_{a\in A^\star} P_Q(a, A^\star) = 0$. Assume that for some neighbourhood U of $A^\star$ the function $U \times C \ni (x, Q) \to P_Q(x, \cdot) \in \mathcal{M}^1(A)$ is continuous. Then the set C satisfies $(\star\star)$ condition.*

Proof We will show that

$$\sup_{(Q,x)\in C\times\overline{B}(A^\star,\varepsilon)} P_Q(x, \overline{B}(A^\star, \varepsilon)) \searrow 0 \text{ as } \varepsilon \searrow 0.$$

Equivalently

$$\lim_{t\to\infty} \sup_{(Q,x)\in C\times\overline{B}(A^\star,\frac{1}{t})} P_Q(x, \overline{B}(A^\star, \frac{1}{t})) = 0.$$

Recall that from the Wierestrass theorem we have that an upper semi-continuous function attains its upper bound on a compact set. We will use the Wierestrass theorem several times in this proof and in order to simplify the argument presentation we will not always be explicit about that.

Fix $n > 0$. The continuity of $(Q, x) \to P_Q(x, \cdot)$ and the upper semicontinuity of $\mathcal{M}^1(A) \ni \mu \to \mu(\overline{D}) \in [0, 1]$ [where $D \in B(A)$, recall (3.6)] imply that for any t with $\overline{B}(A^\star, \frac{1}{t}) \subset U$ the function

$$C \times \overline{B}(A^\star, \frac{1}{t}) \ni (Q, x) \longrightarrow P_Q(x, \overline{B}(A^\star, \frac{1}{n})) \in [0, 1]$$

is upper semi-continuous and thus there is $(Q_t, x_t) \in C \times \overline{B}(A^\star, \frac{1}{t})$ such that

$$P_{Q_t}(x_t, \overline{B}(A^\star, \frac{1}{n})) = \sup_{(Q,x)\in C\times\overline{K}(A^\star, \frac{1}{t})} P_Q(x, \overline{B}(A^\star, \frac{1}{n})), \quad t \in \mathbb{N}.$$

It is easy to see that:

$$\lim_{t\to\infty} \sup_{(Q,x)\in C\times\overline{K}(A^\star, \frac{1}{t})} P_Q(x, \overline{B}(A^\star, \frac{1}{t})) \le \lim_{t\to\infty} P_{Q_t}(x_t, \overline{B}(A^\star, \frac{1}{n})) \tag{4.8}$$

[note that the sequence $P_{Q_t}(x_t, \overline{B}(A^\star, \frac{1}{n}))$ is decreasing and thus has a limit]. Let $\{\hat{Q}_m\}_{m\in\mathbb{N}} \subset C$ be such that

$$\sup_{a\in A^\star} P_{\hat{Q}_m}(a, \overline{B}(A^\star, \frac{1}{m})) = \sup_{a\in A^\star} \sup_{Q\in C} P_Q(a, \overline{B}(A^\star, \frac{1}{m})), \quad m \in \mathbb{N} \tag{4.9}$$

[the existence of $\hat{Q}_m$ follows from the Wierestrass theorem which can be applied to the upper semicontinuous function $C \ni Q \to \sup_{a\in A^\star} P_Q(a, \overline{B}(A^\star, \frac{1}{m})) \in \mathbb{R}$].

First we will show that

$$\lim_{t\to\infty} P_{Q_t}(x_t, \overline{B}(A^\star, \frac{1}{n})) \le \sup_{a\in A^\star} P_{\hat{Q}_n}(a, \overline{B}(A^\star, \frac{1}{n})) \tag{4.10}$$

which, by (4.8), will give us

$$\lim_{t\to\infty} \sup_{(Q,x)\in C\times\overline{K}(a, \frac{1}{t})} P_Q(x, \overline{B}(A^\star, \frac{1}{t})) \le \sup_{a\in A^\star} P_{\hat{Q}_n}(a, \overline{B}(A^\star, \frac{1}{n})).$$

As $n \in \mathbb{N}$ can be arbitrarilly big, to finish the proof it will remain to show:

$$\lim_{n\to\infty} \sup_{a\in A^\star} P_{\hat{Q}_n}(a, \overline{B}(A^\star, \frac{1}{n})) = 0. \tag{4.11}$$

To show (4.10) assume for a contradiction that there is a subsequence $t_k \in \mathbb{N}$ with $P_{Q_{t_k}}(x_{t_k}, \overline{B}(A^\star, \frac{1}{n})) \ge \sup_{a\in A^\star} P_{\hat{Q}_n}(a, \overline{B}(A^\star, \frac{1}{n})) + \varepsilon$ for some $\varepsilon > 0$. As $A^\star$ is compact and C is compact we can additionally assume that $Q_{t_k} \to \hat{Q}$ for some $\hat{Q} \in C$ and $x_{t_k} \to a_0$ for some $a_0 \in A^\star$ [recall that $x_t \in \overline{B}(A^\star, \frac{1}{t})$]. But this leads to $P_{\hat{Q}}(a_0, \overline{B}(A^\star, \frac{1}{n})) \ge \sup_{a\in A^\star} P_{\hat{Q}_n}(a, \overline{B}(A^\star, \frac{1}{n})) + \varepsilon$, a contradiction with the definition of $\hat{Q}_n$ given by (4.9). We thus proved that we have

$$\lim_{t\to\infty} \sup_{(Q,x)\in C\times\overline{K}(A^\star, \frac{1}{t})} P_Q(x, \overline{B}(A^\star, \frac{1}{t})) \le \sup_{a\in A^\star} P_{\hat{Q}_n}(a, \overline{B}(a, \frac{1}{n})).$$

As $n \in \mathbb{N}$ was chosen arbitrarily, it remains to note that condition (4.11) is satisfied. In fact, if for some $\varepsilon > 0$ there is a subsequence $(\hat{Q}_{n_k}, a_{n_k}) \in C \times A^\star$ with $P_{\hat{Q}_{n_k}}(a_{n_k}, \overline{B}(A^\star, \frac{1}{n_k})) > \varepsilon$

and $(\hat{Q}_{n_k}, a_{n_k}) \to (\hat{Q}, \hat{a})$ for some $(\hat{Q}, \hat{a}) \in C \times A^\star$ (compactness of $C \times A^\star$ again) then it is easy to see that for any N we have

$$P_{\hat{Q}}(\hat{a}, \overline{B}(A^\star, \frac{1}{N})) \geq \limsup_{k\to\infty} P_{\hat{Q}_{n_k}}(a_{n_k}, \overline{B}(A^\star, \frac{1}{N})) \geq$$
$$\geq \limsup_{k\to\infty} P_{\hat{Q}_{n_k}}(a_{n_k}, \overline{B}(A^\star, \frac{1}{n_k})) \geq \varepsilon$$

and thus $P_{\hat{Q}}(\hat{a}, A^\star) \geq \varepsilon > 0$ which contradicts the basic assumption $\sup_{a\in A^\star} P_Q(a, A^\star) = 0$, $Q \in C$. □

Conclusion 2 *Assume that $P_Q \in \mathcal{K}(A)$ is such that $\sup_{a\in A^\star} P_Q(a, A^\star) = 0$ and that for some open neighbourhood U of $A^\star$ the function $U \ni x \longrightarrow P_Q(x,) \in \mathcal{M}^1(A)$ is continuous. Then condition $(\star)$ is satisfied.*

Proof To see that it is enough to note that for set $C := \{P_Q\}$ and the identity mapping $P\colon C \to C$ the assumptions of Proposition 5 are satisfied. □

The following result is a simple consequence of Proposition 2.

Theorem 3 *Let C be a metric space and let $P\colon C \longrightarrow \mathcal{K}(A)$ be given mapping. Assume that for some neighbourhood U of $A^\star$ a method (4.1) satisfies*

$$P(f(T_t(x, Y_t)) < f(x)) \leq \sup_{Q\in C} P_Q(x, A_{f(x)}), \quad x \in U, \quad t \in \mathbb{N}. \tag{4.12}$$

If the family $\{P_Q\}_{Q\in C}$ satisfies condition $(\star\star)$ then $\lim_{\delta\to 0} \sup_{t\in\mathbb{N}} P(f(X_{t+1}) < f(X_t) | X_t \in A_\delta) = 0$. In particular, if $X_t \overset{s}{\to} A^\star$, then $X_t \overset{l-s}{\to} 0$.

The above and Proposition 5 immediately yields the following conclusion.

Conclusion 3 *Assume that C is compact, the function $U \times C \ni (x, Q) \to P_Q(x, \cdot) \in \mathcal{M}^1(A)$ is continuous and $\sup_{a\in A^\star} P_Q(a, A^\star) = 0$ for any $Q \in C$. Let X_t be a method (4.1) such that (4.12) holds true. If X_t is globally convergent, then X_t converges lazilly towards $A^\star$.*

Proof of Theorem 3 Based on Proposition 2, to prove the theorem it will be enough to show $(\diamond)$ condition, i.e. we will show that:

$$\sup_{t\in\mathbb{N}} \sup_{x\in A_\delta} P[f(T_t(x, Y_t)) < f(x)] \searrow 0 \text{ as } \delta \searrow 0.$$

For any $\delta > 0$ and $t \in \mathbb{N}$, from (4.12) we already have:

$$\sup_{x\in A_\delta} P[f(T_t(x, Y_t)) < f(x)] \leq \sup_{x\in A_\delta} \sup_{Q\in C} P_Q(x, A_{f(x)}) = \sup_{Q\in C} \sup_{x\in A_\delta} P_Q(x, A_{f(x)}),$$

and thus

$$\sup_{t\in\mathbb{N}} \sup_{x\in A_\delta} P[f(T_t(x, Y_t)) < f(x)] \leq \sup_{Q\in C} \sup_{x\in A_\delta} P_Q(x, A_{f(x)}) \leq \sup_{Q\in C} \sup_{x\in A_\delta} P_Q(x, A_\delta).$$

As condition $(\star\star)$ is satisfied, the latter goes to 0 as δ goes to zero. □

4.4 The main result

Consider the following nonhomogeneous generalization of scheme (4.4):

$$X_{t+1} = \begin{cases} Q_t(X_t, Y_t^1) & \text{if } Y_t^2 \leq p_t(X_t, Y_t^1) \\ X_t & \text{if } Y_t^2 > p_t(X_t, Y_t^1) \end{cases}, \tag{4.13}$$

Here, the measurable mappings $Q_t : A \times B \to A$ and the distributions of $Y_t^1 : \Omega \to B$ can change over time. The appropriate acceptance probabilities p_t can represent Simulated Annealing or monotone random search. Naturally, we have

$$P[f(X_{t+1}) < f(X_t) | X_t = x] \leq \sup_{t \in \mathbb{N}} P_{(Q_t, P_{Y_t})}(x, A_{\delta(x)}),$$

where $P_{(Q_t, P_{Y_t})}$ is given by (4.7). Thus, based on Theorem 3, the following result, Theorem 4, is immediate. The extension of Theorem 4 to the case $p_t = p_t(X_0, \ldots, X_t, Y_t^1)$ is straightforward.

Theorem 4 *Assume that X_t is given by (4.13) and that the family $\{P_{(Q_t, P_{Y_t})}\}_{t \in \mathbb{N}}$ satisfies $(\star\star)$ condition. Then,*

$$\lim_{\delta \to 0} \sup_{t \in \mathbb{N}} P(f(X_{t+1}) < f(X_t) | X_t \in A_\delta) = 0.$$

In particular, if $X_t \xrightarrow{s} A^\star$ then $X_t \xrightarrow{l-s} A^\star$.

The above, based on Propositions 4 and 5, immediately leads to the following result.

Theorem 5 *If X_t of the form (4.13) is globally convergent and any of the following conditions is satisfied:*

(1) *for some compact set C and neighbourhood U of $A^\star$ there is a continuous function $P : U \times C \ni (x, M) \to P_M(x, \cdot) \in \mathcal{M}^1(A)$ which is such that $\sup_{a \in A^\star} P_M(a, A^\star) = 0$ for any $M \in C$ and such that $P_{(Q_t, P_{Y_t})} \in \{P_M(\cdot, \cdot)\}_{M \in C}$*
(2) *the closure C of the family $\{(Q_t, P_{Y_t})\}_{t \in \mathbb{N}} \subset \mathcal{U}$ (in the topology of U given by (4.6)) is compact and such that for any $(Q, \nu) \in C = \overline{\{(Q_t, P_{Y_t})\}_{t \in \mathbb{N}}}$ we have the following two conditions:*

   (a) *$\nu\{y \in B : Q$ is not continuous at $(x,y)\} = 0$, $x \in A$,*
   (b) *$\sup_{a \in A^\star} P_{(Q,\nu)}(a, A^\star) = 0$, where $P_{(Q,\nu)}$ is given by (4.7),*

then X_t converges lazily towards $A^\star$.

Proof Under assumption (1), the thesis follows directly from Theorem 4 and Proposition 5. If we have (2) then Proposition 4 implies that the mapping $C \ni (x, Q, \nu) \to P_{Q,\nu}(x, \cdot) \in M^1(A)$ is continuous and thus the assmuption (1) holds true for the compact set $C = \overline{\{(Q_t, P_{Y_t})\}_{t \in \mathbb{N}}}$. □

Assume for simplicty the natural case $A^\star = \{a\}$. Roughly speaking, from the above theorem it follows that a globally convergent method (4.13) which does not converge lazily towards $A^\star$ must have a subsequence of probability kernels $P_{(Q_{t_k}, P_{Y_{t_k}})} \in \mathcal{K}(A)$ which converges (in some sense) to a $P_{(Q,\nu)} \in \mathcal{K}(A)$ with $P_{(Q,\nu)}(a, \{a\}) > 0$. To give a simple example, assume that the candidate $Q_t(x, Y_t)$ is uniformly distributed on the ball centered at the current state x according to $P_{(Q_t, Y_t)}(x, \cdot) = U[B(x, R_t)]$. If we have

$\liminf_{t\to\infty} R_t = \hat{R} > 0$ then the algorithm cannot converge fast towards a. Another simple example is $P_{(Q_t,Y_t)}(x,\cdot) = N(x, \sigma_t \cdot I_d)$, where I_d is the identity matrix. Naturally, the condition $\liminf_{t\to\infty} \sigma_t = \sigma > 0$ implies that condition $(\star\star)$ is satisfied.

Various formulas for $Q_t(\cdot, Y_t)$ have been analysed in the literature, especially in the context of Simulated Annealing one can see, for instance, [1,5,14–18,38]. The existing results usually are limited to the analysis of the global convergence property or indicate slow convergence rate. Some of the analysed cases satisfy our assumptions, some of them are based on the use of additional information like gradient or the value of $f^\star$. Many algorithms use the previous history $X_0, X_1, \ldots, X_t$. Still, they often can be analysed as Markov Chains after considering the appropriate associated sequences. Sometimes it is enough to consider (X_t, β_t), where β_t is the cooling schedule or $(X_t, \beta_t, \hat{X}_t)$, where $\hat{X}_t$ is the best found point. We also mention that after some effort, the results of this section can be extended to cover the general case $X_{t+1} = T_t(X_t, \ldots, X_0, Y_t)$.

4.5 Accelerated random search

Below we present a simple example from the class of self-adaptive methods, so called Accelerated Random Search, as it fits nicely into our framework. This method was analysed in [3] and it was shown that this algorithm outperforms the simple PRS. As follows from below the finite descent version of ARS (the case $\rho > 0$) converges lazily towards $A^\star$. It is worth to mention that after some modifications regarding the restart mechanism, paper [26] proved that the infinite descent version of ARS (the case $\rho = 0$) has "subexponential" convergence rate under general assumptions regarding the problem function.

The algorithm. Let $A = [0,1]^n \subset \mathbb{R}^n$. Let a contraction factor $c > 1$ and a precision threshold $\rho \geq 0$ be given.

0 Set $t = 0$ and $r_0 = 1$. Generate X_0 from a uniform distribution on A.
1 Given $X_t \in A$ and $r_t \in (0, 1]$, generate Q_t from the uniform distribution on $\overline{B}(X_t, r_t) \cap A$, where $\overline{B}(x, r)$ is the closed ball of radius r centered at x.
2 If $f(Q_t) < f(X_t)$, then let $X_{t+1} = Q_t$ and $r_{t+1} = 1$.
Else if $f(Q_t) \geq f(X_t)$, then let $X_{t+1} = X_t$ and $r_{t+1} = r_t/c$.
If $r_{t+1} < \rho$, then $r_{t+1} = 1$.
Increment $t := t + 1$ and go to Step 1.

Theorem 6 *If $\rho > 0$ then ARS converges lazily towards $A^\star$.*

Proof The global convergence property of X_t is straightforward as in the case $\rho > 0$ the algorithm samples infinitely many times from the whole cube $[-1, 1]^n$. Define $\hat{A} = A \times [\frac{\rho}{c}, 1]$ and let $\hat{f}(x, r) = f(x)$. Note that the set $A^\star \times [\frac{\rho}{c}, 1]$ is the (compact) set of global minimums of $\hat{f}$. The optimization method is of the form $\hat{X}_t = (X_t, r_t) \colon \Omega \to \hat{A}$. We do not need to introduce the formal model (4.1) precisely We just assume that we have some measurable functions $T \colon A \times [\frac{\rho}{c}, 1] \times [0,1]^n \to A \times [\frac{\rho}{c}, 1]$ and $Q \colon A \times [\frac{\rho}{c}, 1] \times [0,1]^n \to A$ such that for some independent sequence $Y_t \colon \Omega \to [0,1]^n$ we have that $Q(x, r, Y_t)$ is uniformly distributed on $B(x, r)$ and the sequence (X_t, r_t) satisfies

$$(X_{t+1}, r_{t+1}) = T(X_t, r_t, Y_t) \text{ and } f(X_{t+1}) = \min\{f(X_t), f(Q(X_t, r_t, Y_t))\}.$$

Now it is enough to note that for any $(x, r) \in \hat{A}$

$$P[\hat{f}(T(x, r, Y_t)) < \hat{f}(x, r)] \leq \max_{i=0,1,\ldots,N} P_{Q_i}(x, A_{\delta(x)}),$$

where $P_{Q_i}(x,\cdot) = U(K(x, \frac{1}{c^i}))$ and $N > \log_c \frac{1}{\rho}$. Theorem 3 finishes the proof as the family $\{P_{Q_i}\}_{i\in\{0,1,\ldots,N\}}$ satisfies condition $(\star\star)$. □

5 Self-adaptation

As discussed earlier, a basic reason for the slow convergence rate of many techniques is related to the following issue: sampling a better candidate point q_t goes to zero as the current state x_t approximates to the optimum. One approach to overcome this difficulty is to set the parameters of nonhomogeneous search in such a way that the optimization method gradually moves from the global search to the local search. The proper procedure for changing parameters' values should cause that the method does not lose global convergence property and performs the reasonable local search at the same time. Going back to the example $P_{(Q_t,Y_t)}(x,\cdot) = N(x, \sigma_t \cdot I_d)$, it is easy to show that the condition $\sigma_t \to 0$ exclude the $(\star\star)$ condition but, on the other hand, finding a proper pattern for σ_t (for the given class of problems) is a very difficult open problem.

5.1 Self-adaptation

An important class of methods which partially avoid the difficulties mentioned above are methods which use self-adaptive mechanisms for parameters' changes. The numerical experiments indicate that the methods based on self-adaptation can perform very fast convergence towards global minimum. The existing theoretical results, see [4,9], show that in some special cases many of such methods converge towards minimum with very fast (exponential) convergence rate. Those results are based on very restrictive assumptions on the problem function but still they indicate that this fast convergence mode can be satisfied for more natural cases. This section is an additional chapter and presents a simple explanation how the "proper" self-adaptation overcomes the problems analysed in the previous sections.

Self-adaptive methods can be, in general, written in the following form:

$$(X_{t+1}, \sigma_{t+1}) = T(X_t, \sigma_t, Y_t), \quad t \in \mathbb{N},$$

where:

(1) the sequence $X_t : \Omega \to A^n$ represents the successive states of the algorithm,
(2) $\sigma_t : \Omega \to C$, where $C \subset (\mathbb{R}^k)^n$, represents the successive parameter's values
(3) $Y_t : \Omega \to B$ is i.i.d. and independent of the (X_0, σ_0)
(4) $T : A \times C \times B \longrightarrow A$ is measurable function

To simplify further analysis we will consider the case $n = 1, k = 1$ and $A = \mathbb{R}^d$, and we will focus on the class methods satisfying:

$$X_{t+1} = \begin{cases} X_t + \sigma_t \cdot Y_t & \text{if } f(X_t + \sigma_t \cdot Y_t) < f(X_t) \\ X_t & \text{if } f(X_t + \sigma_t \cdot Y_t) \geq f(X_t) \end{cases}, \tag{5.1}$$

where Y_t is uniformly distributed on $[-1, 1]$ and $\sigma_t : \Omega \to (0, \infty) > 0$. Given the current values $(X_t, \sigma_t) = (x, \sigma)$ we see that the candidate $x_{t+1} = Q(x, \sigma, Y_t)$ for the next state is uniformly distributed on the ball $B(x, \sigma)$. We will write

$$Q : A \times (0,\infty) \times [-1,1] \ni (x,\sigma,y) \longrightarrow x + \sigma \cdot y \in A,$$
$$P_Q(x,\sigma,C) := \frac{\mu(C \cap K(x,\sigma))}{\mu(K(x,\sigma))}, \quad C \in \mathcal{B}(A).$$

The σ_t parameter is adjusted according to some general procedure

$$\sigma_{t+1} = T_2(X_t, \sigma_t, Y_t).$$

This procedure may naturally use the values of X_t, $Q(X_t, \sigma_t, Y_t)$, $f(Q(X_t, \sigma_t, Y_t))$, $f(X_t)$, Y_t. This method is an example of evolutionary algorithm (1+1). Recall that evolution strategies $(\mu + \lambda)$ are methods which at every step of evolution (every time-step) transform the population of μ individuals by producing λ descendents (candidates) and next choosing μ the best fitted individuals among the population of $(\mu + \lambda)$ parents and descendents, see [6].

If the self-adaptive mechanism of σ_t keeps the proper balance between $d(x_t, a)$ and σ_t then the local search capabilities are adjusted to the current algorithm position in such a way that the issues analysed in the previous sections do not occur. This is expressed in Theorem 7. Below we present the exemplary condition for the above mentioned proper balance:

$$\exists M_1 > 0 \exists\, M_2 > M_1 \;\; \liminf_{t\to\infty} P(M_1 < \frac{\sigma_t}{d(X_t, a)} < M_2) =: P_0 > 0. \tag{5.2}$$

From now, we will assume that $A^\star = \{0\}$ and we will assume that d is the maximum metric $d(x, y) = |x - y|$ so we will write

$$|X_t| = \max\{|X_i| \colon i = 1, \ldots, d\} = d(X_t, a).$$

For any $\delta > 0, 0 < M_1 < M_2 < \infty$ we denote:

$$A(\delta, M_1, M_2) := \{(x, \sigma) \in A \times (0, \infty) \colon \; |x| < \delta \wedge M_1 < \frac{\sigma}{|x|} < M_2\}.$$

Theorem 7 *Let X_t be a method (5.1). We have*

(1) $\forall M_2 > M_1 > 0\, \forall \varepsilon > 0 \exists \delta > 0 \inf_{(x,\sigma)\in A(\delta, M_1, M_2)} P_Q(x, \sigma, K(0, \varepsilon)) \geq (\frac{M_1}{M_2})^d$

(2) *If condition (5.2) is satisfied and X_t is globally convergent then:*

$$\forall \varepsilon > 0 \liminf_{t\to\infty} P\left(Q(X_t, \sigma_t, Y_t) \in K(0, \varepsilon)\right) \geq P_0 \cdot (\frac{M_1}{M_2})^d$$

Proof To prove (1) fix $M_2 > M_1 > 0, \varepsilon > 0$ and choose $\delta > 0$ such that for any x from the ball $K(0, \delta)$ we have $K(x, M_1 \cdot |x|) \subset K(0, \varepsilon)$. For any (x, σ) from the set $A(\delta, M_1, M_2)$ we have

$$\begin{aligned} P_Q((x, \sigma), K(0, \varepsilon)) &= \frac{\mu(K(x, \sigma) \cap K(x, \varepsilon))}{\mu(K(x, \sigma)} \\ &\geq \frac{\mu(K(x, M_1|x|) \cap K(x, \varepsilon))}{\mu(K(x, M_2|x|)} = \frac{\mu(K(x, M_1|x|))}{\mu(K(x, M_2|x|)} = \left(\frac{M_1}{M_2}\right)^d. \end{aligned}$$

To see (2) note that for fixed $\varepsilon > 0$ and $\delta > 0$ as above we have

$$\begin{aligned} P\left(Q(X_t, \sigma_t, Y_t) \in K(0, \varepsilon)\right) &= \int\limits_{A\times\mathbb{R}^+} P\left(Q(x, \sigma, Y_t) \in K(0, \varepsilon)\right) P_{(X_t, \sigma_t)}(d(x, \sigma)) \\ &\geq \int\limits_{A(\delta, M_1, M_2)} P\left(Q(x, \sigma, Y_t) \in K(0, \varepsilon)\right) P_{(X_t, \sigma_t)}(d(x, \sigma)) \\ &\geq (\frac{M_1}{M_2})^d \cdot P_{(X_t, \sigma_t)}(A(\delta, M_1, M_2)). \end{aligned}$$

As $P(|X_t| < \delta) \to 1$ we have that

$$\liminf_{t\to\infty} P_{(X_t,\sigma_t)}(A(\delta, M_1, M_2)) = \liminf_{t\to\infty} P_{(X_t,\sigma_t)}\left(\{(x,\sigma)\colon M_1 < \frac{\sigma}{|x|} < M_2\}\right) = P_0,$$

which finishes the proof. □

We mention that proving condition (5.2) and some weak form of the "asymptotic independence" between the behaviour of sequences $\frac{X_t}{\sigma_t}$ and X_t may be a good base for proving the geometric convergence rate for some class of self-adaptive methods and problem functions. While in general proving such a result will be a difficult task, it is already proved that in some special cases the $\frac{X_t}{\sigma_t}$ is a Markov chain which converges to some stationary distribution Π supported on A, see [4]. This off course implies that condition (5.2) is satisfied for any $0 < M_1 < M_2$. While in general situation the sequence $\frac{X_t}{\sigma_t}$ is not a Markov chain still the sequence $(\frac{X_t}{\sigma_t}, X_t)$ is Markov and we believe that the analysis of $(\frac{X_t}{\sigma_t}, X_t)$ based on Markov chains theory may be a good direction for the development of theoretical tools for the convergence rate analysis of self-adaptive methods.

Acknowledgements The project was supported by National Science Centre (Poland) Grant based on Decision DEC-2013/09/N/ST/04262.

References

1. Andrieu, C., Breyer, L.A., Doucet, A.: Convergence of simulated annealing using FosterLyapunov criteria. J. Appl. Probab. **38**, 975–994 (2001)
2. Aoki, M.: Optimization of Stochastic Systems, Topics in Dicrete-Time Systems. Academic Press, London (1967)
3. Appel, M.J., Labarre, R., Radulovic, D.: On Accelerated Random Search. SIAM J. Optim. **14**, 708–731 (2003)
4. Auger, A., Hansen, N.: Linear convergence of comparison-based step-size adaptive randomized search via stability of Markov chains. SIAM J. Optim. **26**(3), 1589–1624 (2016)
5. Bélisle, C.J.: Convergence theorems for a class of simulated annealing algorithms. J. Appl. Probab. **29**, 885–895 (1992)
6. Beyer, H.G., Schwefel, H.P.: Evolution strategies—a comprehensive introduction. Nat. Comput. **1**, 3–52 (2002)
7. Billingsley, P.: Convergence of Probability Measures, 2nd edn. A Wiley-Interscience Publication, New York (1999)
8. Borovkov, A.A., Yurinsky, V.: Ergodicity and Stability of Stochastic Processes. Wiley, Chichester (1998)
9. Correa, C.R., Wanner, E.F., Fonseca, C.M.: Lyapunov design of a simple step-size adaptation strategy based on success. In: International Conference on Parallel Problem Solving from Nature, pp. 101–110. Springer International Publishing (2016)
10. Duchi, J.C., Bartlett, P.L., Wainwright, M.J.: Randomized smoothing for stochastic optimization. SIAM J. Optim. **22**(2), 674–701 (2012)
11. Duchi, J.C., Jordan, M.I., Wainwright, M.J., Wibisono, A.: Optimal rates for zero-order convex optimization: the power of two function evaluations. IEEE Trans. Inf. Theory **61**(5), 2788–2806 (2015)
12. Douc, R., Moulines, E., Rosenthal, J.S.: Quantitative bounds on convergence of time-inhomogeneous Markov chains. Ann. Appl. Probab. **14**(4), 1643–1665 (2004)
13. Dudley, R.M.: Real Analysis and Probability. Cambridge University Press, Cambridge (2004)
14. Gerber, M., Bornn, L.: Improving simulated annealing through derandomization. J. Glob. Optim. **68**(1), 189–217 (2017)

15. Locatelli, M.: Convergence properties of simulated annealing for continuous global optimization. J. Appl. Probab. **33**(04), 1127–1140 (1996)
16. Locatelli, M.: Convergence of a simulated annealing algorithm for continuous global optimization. J. Global Optim. **18**, 219–233 (2000)
17. Locatelli, M.: Convergence and first hitting time of simulated annealing algorithms for continuous global optimization. Math. Methods Oper. Res. **54**(2), 171–199 (2001)
18. Locatelli, M.: Simulated annealing algorithms for continuous global optimization. In: Pardalos, P.M., Romeijn, H.E. (eds.) Handbook of Global Optimization. Nonconvex Optimization and Its Applications, vol. 62. Springer, Boston, MA (2002). doi:10.1007/978-1-4757-5362-2_6
19. Locatelli, M., Schoen, F.: Random linkage: a family of acceptance/rejection algorithms for global optimization. Math. Program. **85**(2), 379–396 (1999)
20. Meyn, S., Tweedie, R.: Markov Chains and Stochastic Stability. Springer, London (1993)
21. Nocedal, J., Wright, S.: Numerical Optimization. Springer, New York (1999)
22. Ombach, J., Tarłowski, D.: Nonautonomous stochastic search in global optimization. J. Nonlinear Sci. **22**, 169–185 (2012)
23. Peng, Z., Wu, D., Zhu, W.: The robust constant and its applications in random global search for unconstrained global optimization. J. Global Optim. **64**(3), 469–482 (2016)
24. Pintér, J.: Convergence Properties of Stochastic Optimization Procedures, Math. Operationsforsch. u. Statist., ser. Optimization **15**, 405–427 (1984)
25. Pinter, J.D.: Global Optimization in Action. Kluwer Academic Publishers, Dordrecht (1996)
26. Radulović, D.: Pure random search with exponential rate of convergency. Optimization **59**, 289–303 (2010)
27. Rios, L., Sahinidis, N.: Derivative-free optimization: a review of algorithms and comparison of software implementations. J. Global Optim. **56**(3), 1247–1293 (2013)
28. Rudolph, G.: Convergence Properties of Evolutionary Algorithms. Kovac, Hamburg (1997)
29. Schmitt, L.M.: Fundamental study: theory of genetic algorithm. Theor. Comput. Sci. **259**, 1–61 (2001)
30. Semenov, M.A., Terkel, D.A.: Analysis of convergence of an evolutionary algorithm with self-adaptation using a stochastic Lyapunov function. Evol. Comput. **11**(4), 363–379 (2003)
31. Solis, F.J., Wets, R.J.-B.: Minimization by random search techniques. Math. Oper. Res. **6**, 19–30 (1981)
32. Tarłowski, D.: Nonautonomous stochastic search for global minimum in continuous optimization. J. Math. Anal. Appl. **412**(2), 631–645 (2014)
33. Tarłowski, D.: Global convergence of discrete-time inhomogeneous Markov processes from dynamical systems perspective. J. Math. Anal. Appl. **448**(2), 1489–1512 (2017)
34. Tikhomirov, A.S.: On the Markov homogeneous optimization method. Comput. Math. Math. Phys. **46**, 361–375 (2006)
35. Tikhomirov, A.S.: On the convergence rate of the Markov homogeneous monotone optimization method. Comput. Math. Math. Phys. **47**, 780–790 (2007)
36. Tikhomirov, A., Stojunina, T., Nekrutkin, V.: Monotonous random search on a torus: integral upper bounds for the complexity. J. Stat. Plann. Inference **137**, 4031–4047 (2007)
37. Vavasis, S.A.: Complexity issues in global optimization: a survey. In: Horst, R., Pardalos, P.M. (eds.) Handbook of Global Optimization, p. 2741. Kluwer, Dordrecht (1995)
38. Yang, R.L.: Convergence of the simulated annealing algorithm for continuous global optimization. J. Optim. Theory Appl. **104**, 691–716 (2000)
39. Zhigljavsky, A.A.: Theory of Global Random Search. Kluwer Academic Publishers, Dordrecht (1991)
40. Zhigljavsky, A., Žilinskas, A.: Stochastic Global Optimization. Springer, New York (2008)