GLOBAL CONVERGENCE OF DISCRETE-TIME INHOMOGENEOUS MARKOV PROCESSES FROM DYNAMICAL SYSTEMS PERSPECTIVE

DAWID TARŁOWSKI (†)
(†) INSTITUTE OF MATHEMATICS, FACULTY OF MATHEMATICS AND COMPUTER SCIENCE, JAGIELLONIAN UNIVERSITY, ŁOJASIEWICZA 6, 30 348 KRAKÓW, POLAND
FAX: (12) 664-66-74
(†)E-MAIL ADRESSES: DAWID.TARLOWSKI@IM.UJ.EDU.PL, DAWID.TARLOWSKI@GMAIL.COM
TEL: (48) 605-548-532

Abstract. Given the continuous real-valued objective function f and the discrete time inhomogeneous Markov process X_t defined by the recursive equation of the form $X_{t+1} = T_t(X_t, Y_t)$, where Y_t is an independent sequence, we target the problem of finding conditions under which the X_t converges towards the set of global minimums of f. Our methodology is based on the Lyapunov function technique and extends the previous results to cover the case in which the sequence $f(X_t)$ is not assumed to be a supermartingale. We provide a general convergence theorem. An application example is presented: the general result is applied to the Simulated Annealing algorithm.

1. Introduction

Let (A, d) be a compact metric space. Assume that $f\colon A \to \mathbb{R}$ is the continuous problem function with global minimum $\min f = 0$ and $A^\star = \{x \in A\colon f(x) = 0\}$ is the set of the solutions of the global minimization problem. The last decades have witnessed the great development of iterative numerical techniques designed for finding an element from $A^\star$. The most popular methods are: genetic and evolutionary algorithms [35, 34, 5, 33], inspired by the mechanisms of biological evolution, Simulated Annenaling algorithm (SA) [4, 39, 20, 21, 22, 1], which is based on analogy with the physical process of annealing, and methods based on the swarm intelligence of individuals [16] like Particle Swarm Optimization (PSO) [10, 9] or Ant Colony Optimization (ACO) [13]. Those methods, and many other iterative

Key words and phrases. Global optimization, global convergence, inhomogeneous Markov process, Foias operator, Lyapunov function
MSC 2000 subject classifications. Primary: 60J05. Secondary-60J2, 93D05,93D20,
Please cite the article in press as: D. Tarłowski, Global convergence of discrete-time inhomogeneous Markov processes from dynamical systems perspective, J. Math. Anal. Appl. (2017), http://dx.doi.org/10.1016/j.jmaa.2016.11.076 .

heuristics, [3, 15, 23, 32, 40, 41] are used in practice for solving difficult real world problems for which analytical methods fail. The corresponding literature puts great attention to the numerical aspect of the subject. From the theoretical perspective, majority of such optimization techniques can be represented as discrete–time inhomogeneous Markov processes of the form

$$x_{t+1} = T_t(x_t, y_t), \text{ for } t \in \mathbb{N}, \tag{1.1}$$

where the sequence x_t represents the successive states of the algorithm, y_t represents the probability distributions of the algorithm and T_t stands for the deterministic "mechanisms" of the algorithm. Recursions of the form (1.1) have been studied in many contexts, including control theory, iterated function systems (IFS), fractals, and other applications. Various examples can be found, for instance, in [24, 11, 14, 18]. Generally speaking, the standard analysis of processes (1.1) concerns the problem of the existence and the convergence to the unique stationary distribution. This paper continues the research taken in the series of papers [25, 26, 27, 28, 29] which aim at the problem how to prove that the process given by equation (1.1) converges towards $A^\star$ under conditions that can be verified in practical cases.

General results on global convergence are often based on the classical probability theory [38], [31]. Markov chains theory is used to prove the convergence towards $A^\star$ in some cases, see for example [35] or [1]. An important class of global optimization methods are methods with the supermatingale property - we shortly say that an optimization method X_t is a supermartingale if the corresponding sequence of record values $f(X_t)$ is a supermatingale. In this case stochastic Lyapunov functions arise quite naturally as a tool ensuring stability of the process X_t, see Chapter VIII in [2] for general framework or [36] for an example from evolutionary optimization. Previous papers [25]-[29] work under assumption $E(f(T_t(x, Y_t))|X_t = x) \le f(x)$, $x \in A$, which implies that they also aim at the supermartingale class. The general methodology used there was to consider the nonautonomous dynamical system on the set $\mathcal{M}(A)$ of Borel probability measures on A (the system is induced by equation (1.1)) and next to prove the asymptotic stability of the set $M^\star = \{\mu \in \mathcal{M}(A) \colon \mu(A^\star) = 1\}$. One of the basic tools used in the proof was the Lyapunov function given by $V \colon \mathcal{M}(A) \ni \mu \to \int\limits_A f d\mu \in [0, \infty)$. This paper extends this methodology to cover the case of non-supermartingales. For instance, Simulated Annealing algorithm and Evolution Algorithms with non-elitist selection strategies belong to the class of non-supermatingales. The main result of this paper is Theorem 2. Theorem 3, which is the conclusion of Theorem 2, is less general but easier to use and still covers some important practical cases like Simulated Annealing and many non-elitist evolutionary methods. To present how the general results work in practice Theorem 2 is applied to the SA algorithm. The SA convergence result is expressed in Theorem 4.

This paper is organized as follows. Section 2 presents some general equivalences between basic modes of stochastic global convergence. These general results are rather easy to prove but according to the author's knowledge such general statements are not formulated in literature (special cases are proved separately in various papers). Section 3 presents and discusses the main results of this paper, Theorem 2, and its conclusion, Theorem 3. Section 4 applies Theorem 3 to the Simulated Annealing algorithm. Section 5 presents some facts on the weak convergence of Borel probability measures and Section 6 presents some ideas expressed in the language of dynamical systems and necessary for the proof of the main result. Finally, Section 7 uses the results of Sections 5 and 6 to prove Theorem 2. Appendix presents the proofs of results from Section 2.

2. Some Equivalences for Stochastic Global Convergence

This section presents some general equivalences for stochastic global convergence and introduces corresponding notation which will be used in further sections. Although this paper targets the compact case situation, the general results of this section are presented under assumption that the metric space A is separable. The results are rather simple but they generalize many partial observations stated in the literature and will be used further in this paper. The corresponding proofs can be found in Appendix.

We denote:

(1) $A^\star = \{x \in A\colon f(x) = 0\}$,
(2) $A_\delta = \{x \in A : f(x) \leq \delta\}$, where $\delta > 0$,
(3) $A(\delta) = \{x \in A : f(x) < \delta\}$, where $\delta > 0$,
(4) $A^\star(\varepsilon) = \{x \in A\colon d(x, A^\star) < \varepsilon\}$, where $\varepsilon > 0$ and $d(x, A^\star) = \inf\limits_{a \in A^\star} d(x, a)$.

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. The global minimization task usually stands either for generating a sequence $x_t \in A$ which converge towards the set $A^\star$ of solutions of the global minimization problem $f(x) = 0$ or for generating a sequence $x_t \in A$ which satisfies $f(x_t) \to 0$. We will say that a sequence $X_t\colon \Omega \to A$ stochastically converges to $A^\star \subset A$, which we will denote by $X_t \xrightarrow{s} A^\star$, iff

$$\forall \varepsilon > 0 \ \lim_{t \to \infty} P(d(X_t, A^\star) < \varepsilon) = 1. \tag{2.1}$$

Naturally, the stronger condition $P(d(X_t, A^\star) \to 0) = 1$ is sometimes an object of analysis but for some cases it is to strong. Recall that $f(x) < \delta$ iff $x \in A(\delta)$. Thus, in the context of the $f(X_t)$ convergence, the convergence in probability takes the form:

$$\forall \delta > 0 \ P(X_t \in A(\delta)) \to 1. \tag{2.2}$$

For instance, Theorems 1 and 2 in [4] say that under appropriate assumptions the sequence $f(X_t)$ generated by the SA algorithm satisfies (2.2) but does not converge surely. Additionally, from Observation 1 it follows that under those assumptions SA does not converge to $A^\star$ surely but still it is convergent stochastically to it. This paper works on sufficient conditions for an optimization method to satisfy (2.1) and (2.2) which under assumptions of next sections are equivalent to $E(d(X_t, A^\star)) \longrightarrow 0$ and $E(f(X_t)) \longrightarrow 0$. This section shows, in particular, that under some general assumptions various notions of global convergence reduce to the weak convergence of probability distributions $\mu_t = P_{X_t}$ of X_t towards the set $M^\star = \{\mu \in \mathcal{M}(A) \colon \mu(A^\star) = 1\}$.

Assume that the measurable function f is given and that it satisfies the following, rather natural, conditions:

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

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

For example, conditions $A1)$ and $A2)$ are satisfied if for some $\delta_0 > 0$ the underlevel set A_{δ_0} is compact and the function f is continuous on A_{δ_0}. Under those conditions both previous mentioned interpretations of global minimization problem ($x_t \to A^\star$ and $f(x_t) \to 0$) are equivalent which is expressed in Observation 1. This simple observation generalizes many special cases existing in literature. In particular, in the case where $A^\star$ is a singleton some equivalences from statement (1) were proved and used in [4] and [32]. Under assumptions related to $A1)$, $A2)$, the general equivalence from statement (2) was presented in [31]. The precise meaning of condition (1)(b) will be given in Section 5.

Observation 1. *Assume that the space (A, d) is separable and that the function $f \colon A \to \mathbb{R}$ satisfies conditions $A1)$ and $A2)$. We have:*

1. *The following conditions are equivalent:*
   - (a) *X_t converges stochastically to $A^\star$,*
   - (b) *probability distributions of X_t converge towards the set of distributions $\mathcal{M}^\star = \{\mu \in \mathcal{M}(A) \colon \mu(A^\star) = 1\}$ in the weak convergence topology,*
   - (c) *$f(X_t)$ converges in probability to 0,*
   - (d) *$f(X_t)$ converge to 0 in distributions*
2. *The following conditions are equivalent:*
   - (a) *$f(X_t) \to 0$ almost sure*
   - (b) *$d(X_t, A^\star) \to 0$ almost sure.*
3. *Assume additionally that the measurable functions $f(X_t)$ and $d(X_t, A^\star)$ are bounded from the above by some measurable function $Z \colon \Omega \to [0, +\infty)$ with $E(Z) < \infty$. Then the following conditions are equivalent:*
   - (a) *$E(f(X_t)) \longrightarrow 0$,*
   - (b) *$E(d(X_t, A^\star)) \longrightarrow 0$.*

   Additionally, under the above boundedness assumption, they are equivalent to conditions 1(a),1(b),1(c),1(d).

Warning 1. *Conditions $E(f(X_t)) \to 0$ and $E(d(X_t, A^\star)) \to 0$ are not equivalent to each other in general unbounded situation.*

We will focus for a moment on methods which satisfy the following supermartingale inequality:

$$E(f(X_{t+1})|f(X_t), f(X_{t-1}), \dots, f(X_0)) \le f(X_t) \text{ a. s.}$$

The above inequality follows from the stronger condition

$$E(f(X_{t+1})|X_t, X_{t-1}, \dots, X_0) \le f(X_t) \text{ a. s.}$$

which is easier to verify in practice. In particular, if the sequence X_t is a Markov chain, then the above supermartingale–type inequalities follow from the following inequality:

$$E(f(X_{t+1}))|X_t = x) \le f(x), \ x \in A, \tag{2.3}$$

which was one of the convergence assumptions of the previous paper [28]. Thus the following theorem, a consequence of Observation 1, can be applied to some results of [28]. In particular, this result covers the class of monotonic methods (in sense $f(X_{t+1}) \le f(X_t)$), i.e. methods which always remember the best found problem solution candidate.

Theorem 1. *Assume that the space (A, d) is separable and that the function $f : A \to \mathbb{R}$ satisfies A1), A2). If $f(X_t)$ is a supermartingale then the following conditions are equivalent:*

(1) *probability distributions of X_t converge towards the set of distributions $\mathcal{M}^\star = \{\mu \in \mathcal{M}(A) \colon \mu(A^\star) = 1\}$ in the weak convergence topology,*
(2) *$X_t \to A^\star$ stochastically,*
(3) *$d(X_t, A^\star) \to 0$ with probability one,*
(4) *$f(X_t)$ converges to 0 in distribution,*
(5) *$f(X_t)$ converges to 0 in probability,*
(6) *$f(X_t)$ converges to 0 with probability one,*

If we assume additionally that the measurable functions $f(X_t)$ and $d(X_t, A^\star)$ are bounded from the above by some measurable function $Z \colon \Omega \to [0, +\infty)$ with $E(Z) < \infty$ then the above conditions (1),(2),(3),(4),(5),(6) are equivalent to the following conditions:

(7) $E(d(X_t, A^\star)) \longrightarrow 0$,
(8) $E(f(X_t)) \searrow 0$.

Remark 1. *For any Borel probability measures μ_1 and μ_2 on A let*

$$\|\mu_1 - \mu_2\| = \sup_{B \in \mathcal{B}(A)} |\mu_1(B) - \mu_2(B)|.$$

denote the total variation distance. In a standard situation $A \subset \mathbb{R}^n$ the probability distributions μ_t of X_t are absolutely continuous with respect to the Lebesgue Measure μ and usually we have $\mu(A^\star) = 0$ which leads to $||\mu_t - m|| = 1$ for any $m \in M^\star$. The total variaton convergence is thus to strong to be analysed in the context of this paper.

3. Main result and conclusions

In this section we present the main result of this paper, Theorem 2, and next we discuss some of the implications. In particular, from Theorem 2 we conclude Theorem 3 and the latter is less general but also less technical and still covers some important cases. Also we show that the main result of the previous paper [28] which concerns the supermartingale case is a special case of Theorem 2. First we will introduce further assumptions and notation which will hold throughout the paper.

From now on we assume that (A, d) is a compact metric space and that the function $f\colon A \to \mathbb{R}$ is continuous. Recall that this function obtains its minimal value and the set $A^\star = f^{-1}(0)$ is the set of global minima. Let (Ω, Σ, P) be a probability space and let (B, d) be a separable metric space. We assume that the sequence of random variables $X_t\colon \Omega \to A$, $t \in \mathbb{N}$, is defined by the following nonautonomous equation:

$$X_{t+1} = T_t(X_t, Y_t), \tag{3.1}$$

where

- $Y_t\colon \Omega \to B$, $t \in \mathbb{N}$, are random variables
- $T_t\colon A \times B \longrightarrow A$, $t \in \mathbb{N}$, are Borel measurable
- the random variables $X_0, Y_0, Y_1, \cdots$ are independent.

For a Markov chain X_t given on a separable metric space there exists a representation of the form $X_{t+1} = T(X_t, Y_t)$, where $B = [0, 1]$ and $T_t = T$, $t \in \mathbb{N}$, see [8]. However, in cases of many stochastic iterative optimization methods the corresponding theoretical representations arise naturally in more general form (3.1). This paper thus concerns this general situation. Let $\mathcal{T} = \mathcal{M}(A \times B, A)$ denote a topological space of all measurable operators $T\colon A \times B \longrightarrow A$ equipped with the topology of uniform convergence: a sequence $\{T_n\}_{n\in\mathbb{N}} \subset \mathcal{T}$ converges to a limit $T \in \mathcal{T}$ iff

$$\sup_{(a,b)\in A\times B} d(T_n(a, b), T(a, b)) \overset{n\to\infty}{\longrightarrow} 0.$$

As the space A is assumed to be compact (and thus bounded), the above topology is induced by the uniform convergence metric. Let $\mathcal{N} = \mathcal{M}(B)$ denote the topological space of Borel probability measures on B equipped with the weak convergence topology. The space $\mathcal{T} \times \mathcal{N}$ is endowed with the product topology.

For any $\delta > 0$ we define sets $U(\delta) \subset \mathcal{T} \times \mathcal{N}$ and $U_0(\delta) \subset \mathcal{T} \times \mathcal{N}$ as follows:

$$\mathcal{T} \times \mathcal{N} \supset U(\delta) \ni (T, \nu) \overset{def}{\iff} \begin{cases} \int\limits_B f(T(x, y))v(dy) \le f(x) & \text{for } x \notin A(\delta) \\ \int\limits_B f(T(x, y))v(dy) \le \delta & \text{for } x \in A(\delta) \end{cases}$$

$$\mathcal{T} \times \mathcal{N} \supset U_0(\delta) \ni (T,\nu) \overset{def}{\iff} \begin{cases} \int\limits_B f(T(x,y))v(dy) < f(x) & \text{for } x \notin A(\delta) \\ \int\limits_B f(T(x,y))v(dy) \le \delta & \text{for } x \in A(\delta). \end{cases}$$

Recall that a function $F\colon S \to \mathbb{R}$ given on a metric space S is upper semi–continuous (lower semi–continuous) at $x_0 \in S$ iff for any sequence $x_n \in S$, if $x_n \to x_0$, then $\limsup\limits_{n\to\infty} F(x_n) \le F(x_0)$ ($\liminf\limits_{n\to\infty} F(x_n) \ge F(x_0)$). We will also say that a family of sets $\{U_n\}_{n\in\mathbb{N}}$ is a decreasing family iff $U_{n+1} \subset U_n$, $n \in \mathbb{N}$. Finally, to simplify the notation for any sequence of real numbers $\{\delta_i\}_{i\in\mathbb{N}}$ and $t \in \mathbb{N}$ we will denote $\sum\limits_{i=t}^{t-1} \delta_i := 0$ and similarly $\sum\limits_{i\in\varnothing} \delta_i := 0$.

The following theorem is the most general result of this paper. As we will show later, it directly implies both the supermartingale case analysed in [28] and Theorem 3 which is simpler but still covers many important cases of non-supermartingale methods (including SA method).

Theorem 2. *Assume that we have a decreasing family of compact sets $\{U_0^k\}_{k\in\mathbb{N}}$ with $U_0^k \subset U_0(\frac{1}{k})$, $k \in \mathbb{N}$, and a sequence $\{\delta_t\}_{t\in\mathbb{N}} \subset \mathbb{R}$ such that the following conditions are satisfied:*

- **(A1)** *for any $k \in \mathbb{N}$, any pair $(T,\nu) \in U_0^k$ and $x \in A$, the $f \circ T$ is upper semi–continuous at (x,y) for ν-almost any y from B,*
- **(B1)** *$\forall t \in \mathbb{N}$ $(T_t,\nu_t) \in U(\delta_t)$ and $\lim\limits_{t\to\infty} \delta_t = 0$,*
- **(C1)** *for any $k \in \mathbb{N}$ the sequence (T_t,ν_t) contains a subsequence $(T_{t_n^k},\nu_{t_n^k}) \in U_0^k$ such that $\lim\limits_{n\to\infty} S_n^k = 0$, where:*

$$S_n^k = \sum_{i=t_n^k+1}^{t_{n+1}^k-1} \delta_i.$$

Then

$$Ed(X_t, A^\star) \overset{t\to\infty}{\longrightarrow} 0 \quad \textit{and} \quad Ef(X_t) \overset{t\to\infty}{\longrightarrow} 0.$$

Note that the variables X_t and Y_t are independent, and thus we have

$$E(f(X_{t+1})|X_t) = E(f(T_t(X_t,Y_t))|X_t) = \int\limits_B f(T_t(X_t,y))v_t(dy).$$

Hence, for $t \in \mathbb{N}$ and $x \in A$, we have

$$\int\limits_B f(T_t(x,y))v_t(dy) = E(f(X_{t+1})|X_t = x). \tag{3.2}$$

This makes assumptions of Theorem 2 more intuitive. In particular,

$$U(\delta) \ni (T_t, \nu_t) \iff \begin{cases} E(f(X_{t+1})|X_t = x) \le f(x) & \text{for } x \notin A(\delta), \\ E(f(X_{t+1})|X_t = x) \le \delta & \text{for } x \in A(\delta). \end{cases}$$

Note that under assumptions of Theorem 2 for $C_n^k = \{i \in \mathbb{N} \colon t_n^k < i < t_{n+1}^k\}$ we have $S_n^k = \sum\limits_{i \in C_n^k} \delta_i$. The case in which the sets C_n^k are empty strongly simplifies the theorem formulation - this situation is expressed in Theorem 3.

Theorem 3. *Let $\{U_0^k\}_{k\in\mathbb{N}}$ be a decreasing family of compact sets with $U_0^k \subset U_0(\frac{1}{k})$ and such that the following conditions are satisfied:*

- **(A1)** *for any $k \in \mathbb{N}$, $(T, \nu) \in U_0^k$ and $x \in A$, $f \circ T$ is upper semi–continuous at (x, y) for ν-almost any $y \in B$,*
- **(C2)** *for any $t \in \mathbb{N}$, (T_t, ν_t) belongs to $U_0^{k_t}$, where k_t is a sequence with $k_t \to \infty$.*

Then

$$E(d(X_t, A^*) \overset{t\to\infty}{\longrightarrow} 0 \quad \textit{and} \quad Ef(X_t) \overset{t\to\infty}{\longrightarrow} 0.$$

Proof. Define $\delta_t = \frac{1}{k_t}$. The thesis of the theorem follows from Theorem 2. In fact, as $U_0^{k_t} \subset U(\frac{1}{k_t})$, condition **(B1)** of Theorem 2 follows directly from condition **(C2)**. Furthermore, condition **(C1)** of Theorem 2 also follows from condition **(C2)** - to see this it is enough to note that for any $k \in \mathbb{N}$ we can simply put $t_n^k = n$ to have that almost all elements of the sequence $\{S_n^k\}_{n\in\mathbb{N}}$ are equal to 0. In fact, for a fixed $k_0 \in \mathbb{N}$, as the family $\{U_0^k\}_{k\in\mathbb{N}}$ is decreasing and $(T_t, \nu_t) \in U_0^{k_t}$ with $k_t \to \infty$, we have that $(T_t, \nu_t) \in U_0^{k_0}$ for all t big enough. Thus the corresponding sets $\{C_n^{k_0}\}_n$ are empty for n big enough (as $C_n^{k_0}$ denote the indexes between the n-th and the $(n+1)$-th visit of the sequence (T_t, ν_t) in $U_0^{k_0}$). □

Before we recall the result which concerns the supermartingale situation as a special case of Theorem 2.

Theorem 4 ([28])**.** *Assume that $U_0 \subset \mathcal{T} \times \mathcal{N}$ is a compact set such that the following conditions are satisfied:*

- **(A)** *for any $(T, \nu) \in U_0$ and $x \in A$, $f \circ T$ is upper semi–continuous in (x, y) for any y from some set of full measure ν,*
- **(B)** *for any $x \in A$ and $t \in \mathbb{N}$,*

$$\int_B f(T_t(x, y)) v_t(dy) \le f(x), \tag{3.3}$$

- **(C)** *for any $(T, \nu) \in U_0$ and $x \in A \setminus A^*$*

$$\int_B f(T(x, y)) v(dy) < f(x) \tag{3.4}$$

If the sequence (T_t, ν_t) contains a subsequence $(T_{t_n}, \nu_{t_n}) \in U_0$, then

$$d(X_t, A^\star) \longrightarrow 0 \text{ and } f(X_t) \longrightarrow 0 \text{ almost sure.}$$

Proof. To show that assumptions of Theorem 2 hold true it is enough to define the decreasing family of compact sets $\{U_0^k\}_{k\in\mathbb{N}}$ by $U_0^k = U_0$. Condition **(B1)** is satisfied as $(T_t, \nu_t) \in \bigcap_{\delta>0} U(\delta)$, $t \in \mathbb{N}$. Condition **(A1)** thus follows from **(A)**. Condition **(C1)** is also easy to verify as the set U_0 contains a subsequence (T_{t_k}, ν_{t_k}) and again $(T_t, \nu_t) \in \bigcap_{\delta>0} U(\delta)$, $t \in \mathbb{N}$. The assumptions of Theorem 2 are thus satisfied which leads to $Ed(X_t, A^\star) \overset{t\to\infty}{\to} 0$. Theorem 1 finishes the proof as from equation (3.2) it follows that under condition **(B)** the sequence $f(X_t)$ is a supermartingale. □

4. Simulated Annealing.

This section applies Theorem 3 to the Simulated Annealing algorithm which illustrates the functionality of the presented methodology. This optimization method is inspired by the physical procedure called annealing which is used to remove defects from metals and crystals by heating and slow re-cooling the materials so they could lower the energy configuration. According to the physical interpretation, a point $x \in A$ corresponds to the configuration of the atoms of a substance and $f(x)$ determines the energy of the configuration. At every step t, the algorithm which currently is at state x_t generates a candidate $Q(x_t, \cdot)$ for the next state. The candidate is accepted with the probability $p_t(\Delta)$ which depends on the value $\Delta = f(x_t) - f(Q(x_t, \cdot))$. The better points ($\Delta \leq 0$) are accepted with probability one but in case $\Delta > 0$ the acceptance probability is still positive. One of the crucial convergence assumptions is that for any $\Delta < 0$ the acceptance probabilities $p_t(\Delta)$ converge to 0 as the iteration number t goes to infinity. A standard formula for the acceptance probabilities p_t is given by the Metropolis function $p_t(\Delta) = \min\{1, \exp(-\frac{1}{\beta_t} \cdot \Delta)\}$, where the sequence $\beta_t > 0$ is called a "cooling schedule" and converges to zero.

The method is very popular and has been widely analyzed from both numerical and theoretical perspective. This section presents the convergence result for an algorithm with standard Metropolis formula for acceptance probabilities, deterministic cooling schedule and a general regularity assumption on the candidate function Q. One can compare the presented methodology and the obtained convergence result with the results and techniques of many papers regarding SA, including papers [19], [20], [4],[21],[22], [1]. As one can see in those papers various relations between the cooling schedule β_t and candidate function Q determine (or exclude) convergence. Theorem 5 presented here provides general conditions under which the condition $\beta_t \to 0$ is enough to guarantee global convergence (more common approach makes some assumptions on the β_t convergence rate to ensure the global convergence). The crucial convergence assumption of Theorem 5 is $P(f(Q(x,Y_t) < f(x)) > 0$ for $x \notin A^\star$.

Below we construct a formal model of the algorithm which satisfies equation (3.1). Let $A \subset \mathbb{R}^n$ be a compact metric space and let $\bar{B}$ be a separable metric space. Let $M > 0$ and let $B = \bar{B} \times [0,1] \times [0,M]$. Let $\xi_t \colon \Omega \to \bar{B}$, $t \in \mathbb{N}$, be an i.i.d. sequence distributed according to some $\bar{\nu} \in \mathcal{M}(\bar{B})$ and let $r_t \colon \Omega \to [0,1]$, $t \in \mathbb{N}$, be an i.i.d. sequence with uniform distribution. Let $[0,M] \ni \beta_t$ be a sequence with $\lim\limits_{t\to\infty} \beta_t = 0$ and let $Q \colon A \times \bar{B} \to A$ be measurable. Now we can define the mapping $T \colon A \times B \to A$ as:

$$T(x,\xi,r,\beta) = \begin{cases} Q(x,y), \text{if } f(Q(x,y)) \le f(x) \\ Q(x,y), \text{if } f(Q(x,y)) > f(x) \wedge r \le \exp(-\frac{1}{\beta} \cdot |f(Q(x,y)) - f(x)|) \\ x, \text{otherwise} \end{cases}$$

If $\beta = 0$, then in the above formula we put $-\frac{1}{\beta} = -\infty$ and $\exp(-\infty) = 0$. The Simulated Annealing algorithm $X_t \colon \Omega \to A$ is defined by:

$$X_{t+1} = T(X_t, \xi_t, r_t, \beta_t),$$

where X_0 is a random variable independent of the sequence $\{\xi_t, r_t\}_{t\in\mathbb{N}}$.

Theorem 5. *Let $f \colon A \to \mathbb{R}$ be a continuous function given on a compact metric space A. Assume that for any $x \in A$, $\bar{\nu}(D_{f\circ Q}(x)) = 0$, where $D_{f\circ Q}(x)$ is the set of $y \in \bar{B}$ such that $f \circ Q$ is not continuous at (x,y). Assume additionally that for any $x \in A \setminus A^\star$,*

$$\bar{\nu}(\{y \in \bar{B} \colon f(Q(x,y)) < f(x)\}) > 0. \tag{4.1}$$

Then,

$$Ed(X_t, A^*) \overset{t\to\infty}{\longrightarrow} 0 \text{ and } Ef(X_t) \overset{t\to\infty}{\to} 0.$$

Proof. First note that the probability distribution ν_t of the random vector $Y_t = (\xi_t, r_t, \beta_t)$ satisfies $\nu_t = \bar{\nu} \times u_d \times \delta_{\beta_t} \in \mathcal{M}(B)$, where $\bar{\nu}$ is the distribution of ξ_t and $u_d = U(0,1)$. For $k \in \mathbb{N}$ we define

$$\mathcal{M}(B) \supset U_k^{(2)} = \{\bar{\nu} \times u_d \times \delta_\beta \colon \beta \in [0,\beta_k]\} \text{ and } U_k = \{T\} \times U_k^{(2)}.$$

We will show that there is an increasing sequence $n_k \in \mathbb{N}$ with $U_{n_k} \subset U_0(\frac{1}{k})$, $k \in \mathbb{N}$ such that the sets

$$U_0^k := U_{n_k} \subset U_0(\frac{1}{k})$$

satisfy the assumptions of Theorem 3.

First note that for any $k \in \mathbb{N}$ the set U_k is compact as the continuous image of an interval $[0,\beta_k]$ and any family of the form $\{U_{n_k}\}_{k\in\mathbb{N}}$ is decreasing if the sequence n_k satisfies $\beta_{n_k} \searrow 0$. We will show now that any family of the form $\{U_{n_k}\}_{k\in\mathbb{N}}$ satisfies condition **(A1)**, i.e. we will prove that for any $\beta \in [0,M]$ we have the following:

$$(\diamond)\ (\bar{\nu} \times u_d \times \delta_\beta)(D^u_{f\circ T}(x)) = 0 \text{ for any } x \in A.$$

Fix $x \in A$ and $\beta \in [0, M]$. We need to show that the function $f \circ T$ is continuous at (x, ξ, r, β) for $(\bar{\nu} \times u_d)$−almost any (ξ, r). Fix $\xi \in \bar{B}$ such that $f \circ Q$ is continuous at (x, ξ). We have $f \circ T(x, \xi, \cdot, \beta) \in \{f(x), f(Q(x, \xi))\}$. Fix $r \in [0, 1]$ and let $z_n = (x_n, \xi_n, r_n, \beta_n) \to (x, \xi, r, \beta) = z$. If $f(x) > f(Q(x, \xi))$ then condition $(\diamond)$ is satisfied by the continuity of f at x and $f \circ Q$ at (x, ξ) as for n big enough we will have $f(x_n) > f(Q(x_n, \xi_n))$ and $f(T(z_n)) = f(Q(x_n, \xi_n)) \to f(Q(x, \xi)) = f(T(z))$. The case $f(x) = f(Q(x, \xi))$ is even more straightforward. It remains to consider the inequality $f(x) < f(Q(x, \xi))$. For simpler notation, define:

$$\Delta f(x, \xi) = |f(Q(x, \xi)) - f(x)|, \quad x \in A, \xi \in B.$$

It is enough to consider two cases: $r < \exp(-\frac{1}{\beta}\Delta f(x, \xi))$ and $r > \exp(-\frac{1}{\beta}\Delta f(x, \xi))$. In both situations the required condition $(\diamond)$ is satisfied because of the continuity of functions: $(x, \xi, \beta) \longrightarrow \exp(-\frac{1}{\beta}\Delta f(x, \xi))$, f and the continuity of $f \circ Q$ at point (x, ξ) (the argument is very similar to the the case $f(x) > f(Q(x, \xi))$). Thus, for any $x \in A$ we have that for any $\beta \in [0, M]$, $\xi \in \bar{B}$ such that $f \circ Q$ is continuous at (x, ξ) and any $r \in [0, 1]$ with $r \neq \exp(-\frac{1}{\beta}\Delta f(x, \xi))$ the function $f \circ T$ is continuous at point (x, ξ, r, β). Thus those ξ and r, which satisfy the desired continuity assumption, form a set of full measure.

It remains to prove condition **(C2)**. Fix $k \in \mathbb{N}$. Recall that $U_n = \{(T, \bar{\nu} \times u_d \times \delta_\beta) \colon \beta \leq \beta_n\}$. As $\beta_n \to 0$, to prove condition **(C2)** it will be enough to show that for any $k \in \mathbb{N}$ and any $n \geq \bar{n}_k \in \mathbb{N}$ big enough we have

$$\begin{cases} (\star) \int\limits_{\bar{B}} f(T(x_0, y))\nu(dy) < f(x_0), \text{ for } x_0 \in A(\frac{1}{k})' \\ (\star\star) \int\limits_{\bar{B}} f(T(x_0, y))\nu(dy) \leq \frac{1}{k}, \text{ for } x_0 \in A(\frac{1}{k}). \end{cases}$$

for all (T, ν) from U_n. For any $x \in A$ and $\beta \in [0, M]$ let $A_1(x), A_2(x, \beta), A_3(x, \beta) \subset \bar{B} \times [0, 1]$ be defined as:

- $A_1(x) = \{(\xi, r) \colon f(Q(x, \xi)) \leq f(x)\}$,
- $A_2(x, \beta) = \{(\xi, r) \colon f(Q(x, \xi)) > f(x) \wedge r > \exp(-\frac{1}{\beta} \cdot \Delta f(x, \xi))\}$,
- $A_3(x, \beta) = \{(\xi, r) \colon f(Q(x, \xi)) > f(x) \wedge r \leq \exp(-\frac{1}{\beta} \cdot \Delta f(x, \xi))\}$.

For $x \in A$, $\beta \in [0, M]$ and $n \in \mathbb{N}$ let:

- $A_3^1(x, \beta, n) = A_3(x, \beta) \cap \{\Delta f(x, \xi) \leq \frac{1}{n}\}$,
- $A_3^2(x, \beta, n) = A_3(x, \beta) \cap \{\Delta f(x, \xi) > \frac{1}{n}\}$.

Note that for any $x \in A$, $\beta \in [0, M]$ and $n \in \mathbb{N}$, we have:

$$A_3(x, \beta) = A_3^1(x, \beta, n) \cup A_3^2(x, \beta, n).$$

Let:

$$I(x) = \int\limits_{A_1(x)} f(Q(x, \xi))(\bar{\nu} \times u_d)(d\xi, dr) = \int\limits_{\{\xi \colon f(Q(x, \xi)) \leq f(x)\}} f(Q(x, \xi))\bar{\nu}(d\xi),$$

$$II_\beta(x) = \int\limits_{A_2(x,\beta)} f(x)(\bar{\nu}\times u_d)(d\xi,dr),$$

$$III_\beta(x) = \int\limits_{A_3(x,\beta)} f(Q(x,\xi))(\bar{\nu}\times u_d)(d\xi,dr).$$

$$III_3^1(x,\beta,n) = \int\limits_{A_3^1(x,\beta,n)} f(T(x,\xi,r,\beta))(\bar{\nu}\times u_d)(\xi,r),$$

$$III_3^2(x,\beta,n) = \int\limits_{A_3^2(x,\beta,n)} f(T(x,\xi,r,\beta))(\bar{\nu}\times u_d)(\xi,r).$$

For $\nu = (\bar{\nu}, u_d, \delta_\beta) \in M(B)$, $n \in \mathbb{N}$ and any $x \in A$ we have

$$\int\limits_B f(T(x,y))\nu(dy) = \int\limits_{[0,1]}\int\limits_{\bar{B}} f(T(x,\xi,r,\beta))\bar{\nu}(d\xi)dr =$$

$$I(x) + II_\beta(x) + III_\beta(x) = I(x) + II_\beta(x) + III_3^1(x,\beta,n) + III_3^2(x,\beta,n).$$

Note that for any x, β and n we have

$$I(x) + II_\beta(x) + III_3^1(x,\beta,n) \le I(x) + \left(f(x) + \frac{1}{n}\right)\cdot(\nu\times u_d)(A_1(x)').$$

We will show now that for any $n \in \mathbb{N}$ and $x \in A(\frac{1}{n})'$ there is $\varepsilon_n > 0$ with

$$I(x) + f(x)\cdot(\nu\times u_d)(A_1(x)') < f(x) - \varepsilon_n. \tag{4.2}$$

To see this consider $\tilde{T}\colon A\times B \longrightarrow A$ defined by

$$\tilde{T}(x,\xi,r,\beta) = \begin{cases} Q(x,\xi), \text{if } f(Q(x,\xi)) \le f(x) \\ x, \text{otherwise.}\end{cases}$$

From assumption (4.1) it follows that

$$\int\limits_{\bar{B}} f(\tilde{T}(x,y))\nu(dy) = I(x) + f(x)\cdot(\nu\times u_d)(A_1(x)') < f(x). \tag{4.3}$$

Since the function $f\circ\tilde{T}(x,\xi,r,\beta) = \max\{f(x), f(Q(x,\xi))\}$ is continuous at $(x,y) \in A\times B$ for $\nu-$ almost any $y = (\xi,r,\beta)$ then from Proposition 1 it follows that the function $x \longrightarrow \int\limits_B f(\tilde{T}(x,y))\nu(dy)$ upper semi-continuous and hence the function $x \longrightarrow f(x) - \int\limits_{\bar{B}} f(\tilde{T}(x,y))\nu(dy)$ is bounded from zero on a compact set $A(\frac{1}{n})'$ for any $n \in \mathbb{N}$. From (4.3) we thus have

$I(x)+f(x)\cdot(\nu\times u_d)(A_1(x)') < f(x)-\varepsilon_n$ for any $x \in A(\frac{1}{n})'$ and some $\varepsilon_n > 0$,

which proves inequality (4.2).

Fix $k \in \mathbb{N}$ and let ε_k satisfies (4.2). We have that for $n_k \in \mathbb{N}$ big enough and any $x \in A(\frac{1}{k})'$:

$$I(x) + \left(f(x) + \frac{1}{n_k}\right)\cdot(\nu\times u_d)(A_1(x)') < f(x) - \frac{\varepsilon_k}{2}.$$

For such $n_k \in \mathbb{N}$, any $x \in A(\frac{1}{k})'$ and any $\beta \in [0, M]$ we thus have

$$I(x)+II_\beta(x)+III_3^1(x,\beta,n) \leq I(x)+\left(f(x)+\frac{1}{n}\right)\cdot(\nu\times u_d)(A_1(x)') < f(x)-\frac{\varepsilon_k}{2}.$$

To show condition $(\star)$ it is enough to show that for any β small enough we have:

$$III_3^2(x,\beta,n_k) = \int\limits_{A_3^2(x,\beta,n_k)} f(T(x,\xi,r,\beta))(\nu\times u_d)(\xi,r) \leq \frac{\varepsilon_k}{2} \tag{4.4}$$

As the function f is bounded from above it will be enough to show that for any β small enough we have $(\nu\times u_d)(A_3^2(x,\beta,n_k)) \leq \frac{1}{\max\limits_{x\in A} f(x)} \cdot \frac{\varepsilon_k}{2}$ for any x. We have

$$A_3^2(x,\beta,n_k) = \{(\xi,r)\colon r \leq \exp(-\frac{1}{\beta}\cdot\Delta f(x,\xi))\} \cap \{\Delta f(x,\xi) > \frac{1}{n_k}\}$$

and

$$\{(\xi,r)\colon r \leq \exp(-\frac{1}{\beta}\cdot\Delta f(x,\xi))\}\cap\{\Delta f(x,\xi) > \frac{1}{n_k}\} \subset \{(\xi,r)\colon r \leq \exp(-\frac{1}{\beta}\cdot\frac{1}{n_k})\}$$

which leads to $(\nu\times u_d)(A_3^2(x,\beta,n_k)) \leq \exp(-\frac{1}{\beta}\cdot\frac{1}{n_k})$ and proves that conditions $(\star)$ is satisfied as $\exp(-\frac{1}{\beta}\cdot\frac{1}{n_k}) \to 0$ with $\beta \to 0^+$.

We have thus shown that:

(4.5)

$$\forall k \in \mathbb{N}\exists \bar{n}_k \in \mathbb{N}\forall n > \bar{n}_k\forall (T,\nu) \in U_n\forall x \in A(\frac{1}{k})' \int\limits_{\bar{B}} f(T(x,y))\nu dy < f(x).$$

Now we will prove $(\star\star)$. To be more specific, we will show that:

(4.6)

$$\forall k \in \mathbb{N}\exists \bar{n}_k \in \mathbb{N}\forall n > \bar{n}_k\forall (T,\bar{\nu}) \in U_n\forall x \in A(\frac{1}{k}) \int\limits_{\bar{B}} f(T(x,y))\bar{\nu} dy \leq \frac{1}{k}.$$

Fix $k \in \mathbb{N}$. First we will show that for β small enough we have $III_\beta(x) \leq \frac{3}{4k}$ for any $x \in A(\frac{1}{4k})$. Fix $n_k \in \mathbb{N}$ such that $\frac{1}{n_k} < \frac{1}{4k}$. From (4.4) for $\beta \leq \beta_{t_k}$ small enough we have $III_3^2(x,\beta,n_k) < \frac{1}{4k}$. From the definition of $A_3^1(x,\beta,n_k)$ for any $x \in A(\frac{1}{4k})$ we have $III_3^1(x,\beta,n_k) \leq \frac{1}{4k} + \frac{1}{n_k} \leq \frac{1}{2k}$. Hence, for any $x \in A(\frac{1}{4k})$, $III_\beta(x) \leq \frac{3}{4k}$ and

$$\int\limits_{\bar{B}} f(T(x,y))\bar{\nu} dy \leq f(x) + III_\beta(x) \leq \frac{1}{4k} + \frac{3}{4k} = \frac{1}{k}.$$

To prove (4.6) it is enough to assume additionally that β_{t_k} is small enough to have equation (4.5) satisfied for any $x \notin A(\frac{1}{4k})$. We have thus shown that condition **(C2)** of Theorem 3 holds true.

$\square$

5. Weak Convergence, Foias Operators and Optimization

In this section we recall some facts about weak convergence of Borel probability measures, [7, 12], and some properties of Foias operators which are induced by equation (3.1) (the use of Poias operators in other contexts one can see in [18, 37]). This section also introduces further notation.

Let (S, d_S) be a separable metric space and let $\mathcal{B}(S)$ denote the sigma-algebra of Borel subsets of S. By $\mathcal{M}(S)$ we will denote the space of Borel probability measures on S. A sequence $\mu_n \in \mathcal{M}(S)$ weakly converges to some $\mu \in \mathcal{M}(S)$ iff for any bounded continuous function $h\colon S \to \mathbb{R}$ we have

$$\int_S h\, d\mu_n \to \int_S h\, d\mu, \text{ as } n \to \infty. \tag{5.1}$$

Another equivalent condition for the weak convergence is that the sequence $\mu_n \in \mathcal{M}(S)$ converges to some $\mu \in \mathcal{M}(S)$ iff for any upper semi–continuous function $h\colon S \to \mathbb{R}$ bounded from above we have

$$\limsup_{n\to\infty} \int_S h d\mu_n \le \int_S h d\mu, \text{ as } n \to \infty. \tag{5.2}$$

The mapping $S \ni s \longrightarrow \delta_s \in M(S)$, where δ_s denotes a Dirac measure concentrated on the point $s \in S$, is continuous and injective. The weak topology on $M(S)$ is separable and additionally, if we assume that the metric space S is compact then $\mathcal{M}(S)$ is compact. Let S_1 and S_2 be separable metric spaces. For any $h\colon S_1 \to S_2$, we will write $D_h = \{x \in S_1\colon \text{ } h \text{ is not continuous in } x\}$. If $S_2 = \mathbb{R}$, then we write $D_h^u = \{x \in S_1\colon \text{ } h \text{ is not upper semi–continuous in } x\}$. Weak convergence condition given by (5.2) can be rather easily strengthened to the form presented in the following lemma (standard argument can be found, for example, in [28], Lemma 2).

Lemma 1. *Assume that μ_n is a sequence of Borel probability measures on a metric space S with $\mu_n \to \mu$ for some $\mu \in M(S)$. Then for any bounded from above and measurable function $h\colon S \to \mathbb{R}$, if $\mu(D_h^u) = 0$, then* $\limsup\limits_{n\to\infty} \int_S h d\mu_n \le \int_S h d\mu$.

If $h\colon S_1 \to S_2$ is a Borel function, then for any $\mu \in M(S_1)$, μh^{-1} denotes a Borel probability measure on S_2, defined by $\mu h^{-1}(C) = \mu\left(h^{-1}(C)\right)$, for any $C \in \mathcal{B}(S_2)$. As S_1 and S_2 are separable, we have $\mathcal{B}(S_1 \times S_2) = \mathcal{B}(S_1) \otimes \mathcal{B}(S_2) = \Sigma(A_1 \times A_2\colon A_1 \in S_1, A_2 \in S_2)$. For $\mu \in M(S_1)$ and $\nu \in M(S_2)$, $\mu \times \nu$ 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 $C \in \mathcal{B}(S_1), D \in \mathcal{B}(S_2)$. Finally, as S is separable, the topology of weak convergence on $M(S)$ is metrizable and one of available metrics is Prohorov metric:

$$d_M(\nu_1, \nu_2) = \inf\{\varepsilon > 0\colon \nu_1(D) \le \nu_2(D(\varepsilon)) + \varepsilon \text{ for any Borel set D}\},$$

where $D(\varepsilon) = \{x \in S\colon d_S(x, y) < \varepsilon \text{ for some } y \in D\}$.

Now we recall assumptions of Section 3: it is assumed that A is a compact metric space, the spaces $\mathcal{M} = \mathcal{M}(A)$ and $\mathcal{N} = \mathcal{M}(B)$ are equipped with the Prohorov metric and $\mathcal{T} = \mathcal{M}(A \times B; A)$ are measurable operators with the uniform convergence topology induced by the uniform convergence metric. The $\mathcal{T} \times \mathcal{N}$ is considered as equipped with the product metric.

For any $u = (T, \nu) \in \mathcal{T} \times \mathcal{N}$, by $P_u \colon \mathcal{M} \ni \mu \to P_u\mu \in \mathcal{M}$ we denote a Foias operator which transforms probability measures on A according to the following relation:

$$(P_\nu\mu)(C) = (\mu \times \nu)(T^{-1}(C)), \text{ for any Borel set } C \subset A. \tag{5.3}$$

The following lemma is the immediate consequence of the above definition.

Lemma 2. *Let $X : \Omega \to A$ and $Y : \Omega \to B$ 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 $h : A \to \mathbb{R}$, by change of variables,*

$$\int_A h dP_{(T,\nu)}\mu = \int_\Omega h\left(T(X, Y)\right) dP = \int_{A\times B} (h \circ T) d(\mu \times \nu). \tag{5.4}$$

In consequence, if the sequence $\{\mu_t\}_{t\in\mathbb{N}} \subset \mathcal{M}$ corresponds to probability distrubutions of the process $\{X_t\}_{t\in\mathbb{N}}$ given by equation (3.1), then we have $\mu_{t+1} = P_{u_t}\mu_t$, $t \in \mathbb{N}$, where $u_t = (T_t, \nu_t)$ and ν_t is the distribution Y_t. Thus the algorithm's distrubutions are the trajectories of dynamical system given by the family of Foias operators determined by (3.1).

Lemma 3 (Theorem 2.8 in [7])**.** *Let μ_n, ν_n be sequences of Borel probability measures on separable metric spaces S_1, S_2 respectively, with $\mu_n \to \mu$ and $\nu_n \to \nu$ for some $\mu \in M(S_1)$ and $\nu \in M(S_2)$. Then $\mu_n \times \nu_n \to \mu \times \nu$.*

Lemma 4 (Theorem 2.7 in [7])**.** *Assume that S_1, S_2 are metric spaces, $\mu \in M(S_1)$ and $T \colon S_1 \to S_2$ is measurable with $\mu(D_T) = 0$. Then, for any sequence μ_n of Borel probability measures on S_1, if $\mu_n \to \mu$, then $\mu_n T^{-1} \to \mu T^{-1}$.*

Lemmas 1, 3 and 4 lead to the Proposition 1 which explains the regularity assumption **(A)** (and thus **(A1)**) of Section 3 (the detailed proof is presented in [28], Proposition 2).

Proposition 1. *Assume that $U_0 \subset \mathcal{T} \times N$ satisfies the assumption* ***(A)*** *of Theorem 4. Consider $U_0 \times \mathcal{M}$ as equipped with the product topology. The function*

$$U_0 \times \mathcal{M} \ni (u, \mu) \longrightarrow \int\limits_A f dP_u\mu \in \mathbb{R}$$

is upper semi–continuous.

Finally, for a $D \in \mathcal{B}(\mathcal{S})$ define $M^\star(D) = \{\mu \in M |\ \mu(D) = 1\}$ and $D(\varepsilon) = \bigcup\limits_{d\in D} K(d, \varepsilon)$. The following observation is rather easy to prove.

Observation 2 (Folklore)**.** *For any sequence $\mu_n \in \mathcal{M}(S)$ and a set $D \in \mathcal{B}(\mathcal{S})$ we have*

$$d_M(\mu_n, \mathcal{M}^\star(D)) \to 0 \Longleftrightarrow \forall \varepsilon > 0 \ \mu_n(D(\varepsilon)) \to 1.$$

Note that if μ_t is the sequence of distributions of X_t and $D = A^\star$ then the above observation expresses the stochastic global convergence of the sequence X_t in terms of the distributions' convergence towards the set $\mathcal{M}^\star = \mathcal{M}^\star(A^\star)$. The next section will prepare necessary tools for showing that under assumptions of Theorem 2 the set $\mathcal{M}^\star$ is a global attractor for the dynamical system induced by the Foias operators corresponding to equation (3.1). In other words, Section 7 will show that if the sequence $u_t = (T_t, \nu_t)$ determined by (3.1) satisfies assumptions of Theorem 2 for some decreasing family of compact sets $\{U_0^k\}, k \in \mathbb{N}$, then any sequence $\mu_t \in \mathcal{M}$ defined by $\mu_{t+1} = P_{u_t}\mu_t$ converges towards $\mathcal{M}^\star$.

6. Some Concepts of Dynamical Systems in Metric Spaces

The ideas used for proving Theorem 2 are presented in this section and are formulated in the language of the discrete time nonautonomous systems, see [17]. For the general stability concepts of (continuous time) topological dynamical systems we refer to [6].

Let (X, d) be a metric space and let $\varphi \colon X \to X$ be a given map. We do **not** assume that φ is continuous or invertible. According to [30], the mapping φ induces a discrete–time pseudo–dynamical semi–system on X. For any $x \in X$ its orbit is given by $o(x) = \{\varphi^0(x), \varphi(x), \varphi^2(x), \ldots\}$, where $\varphi^0(x) = x$ and $\varphi^{t+1}(x) = \varphi(\varphi^t(x))$. Let $\varnothing \neq K \subset X$ be a closed set. We say that a point $x \in X$ is attracted to K iff $d(\varphi^t(x), K) \longrightarrow 0$. We will shortly say that a set $D \subset X$ is attracted to K if it satisfies

$$d(\varphi^t(x), K) \overset{t\to\infty}{\longrightarrow} 0, \text{ for any } x \in D,$$

i.e. any point $x \in D$ is attracted to K.

Note that we do **not** assume in the above definitions that K is positively invariant under φ (we do not assume that $\varphi(K)$ is a subset of K) or stable. Recall that K is stable iff

$$\forall \varepsilon > 0 \exists \delta > 0 \forall x \in K(\delta), \ o(x) \subset K(\varepsilon),$$

where $K(\varepsilon) = \{x \in X \colon \ d(x, K) < \varepsilon\}$. The tools for examining the attractiveness, presented in this section, are based on the Lyapunov function technique. Untypically, we will not assume that the Lyapunov function is monotonically decreasing along trajectories as a more general concept is necessary for further use.

Assume that the function $W \colon X \to [0, \infty)$ satisfies:

(1) $W(x) = 0$, for $x \in K$,
(2) $W(x) > 0$, for $x \in X \setminus K$.

For any $\varepsilon > 0$ we write

$$W(\varepsilon) = \{x \in X \colon\ W(x) < \varepsilon\}.$$

It is a simple observation that if the following conditions:

$$\forall \varepsilon > 0\ \exists \delta > 0\ \ W(\delta) \subset K(\varepsilon), \tag{6.1}$$

$$\forall \delta > 0\ \exists \varepsilon > 0\ \ K(\varepsilon) \subset W(\delta), \tag{6.2}$$

are satisfied, then

$$d(\varphi^t(x), K) \to 0 \Longleftrightarrow W(\varphi^t(x)) \to 0.$$

Note that if X is compact and W is continuous then conditions (6.1) and (6.2) are satisfied - this follows from the uniform continuity of W and the sequential compactness of K.

The observation below concerns a property of a closed set K weaker than a standard stability concept.

Observation 3. *Let the conditions* (6.1) *and* (6.2) *be satisfied and let $D_0 \subset X$. Assume additionally that for some sequence $\varepsilon_s \to 0^+$ and any $x \in D_0$ the function W satisfies*

$$\forall t \in \mathbb{N}\ \forall s \in \mathbb{N},\ \ W(\varphi^t(\varphi^s(x))) \le W(\varphi^s(x)) + \varepsilon_s. \tag{6.3}$$

Then,

$$\forall \varepsilon > 0\ \exists \varepsilon_0 > 0\ \exists s_0 > 0\ \forall s > s_0\ \forall x \in K(\varepsilon_0) \cap \varphi^s(D_0),\ \ \ o(x) \subset K(\varepsilon).$$

If the function W decreases along trajectories, i.e. $\varepsilon_s = 0$ for $s \in \mathbb{N}$ and $D_0 = X$,then the set K is positively invariant and stable.

Proof. If $\varepsilon > 0$ then from (6.1) and (6.2) it follows that there are $\delta > 0$ and $\varepsilon_0 > 0$ with $K(\varepsilon_0) \subset W(\frac{\delta}{2}) \subset W(\delta) \subset K(\varepsilon)$, which implies $o(x) \subset W(\delta) \subset K(\varepsilon)$ for any $x \in K(\varepsilon_0) \cap \varphi^s(D_0)$, where s is big enough to have $\varepsilon_s < \frac{\delta}{2}$. Under the additional assumption $\varepsilon_s = 0$ and $D_0 = X$ we have $\varepsilon_s < \frac{\delta}{2}$ for any s which proves the stability of K. The invariantness is also a straightforward consequence of the monotonicity of the sequence $W(\varphi^s(x))$, $x \in X$, as K is the set of global minimums of W. $\square$

Lemma 5. *Let $W \colon X \to [0, +\infty)$ be a continuous function with $W^{-1}(\{0\}) = K$ and let $D \subset X$, $\varepsilon > 0$ be such that:*

$$W(\varphi^s(x)) \le W(x) + \varepsilon \text{ for any } x \in D \text{ and } s \in \mathbb{N}. \tag{6.4}$$

Let $S \subset X$ be such that for any $x \in X$, $o(x) \cap S \neq \varnothing$. Assume that for some $\delta > 0$ and $\varepsilon_0 > \varepsilon$ we have

$$\forall x \in S \setminus W(\delta)\ \ W(\varphi(x)) < W(x) - \varepsilon_0. \tag{6.5}$$

Then

$$\limsup_{t \to \infty} W(\varphi^t(x)) \le \delta + \epsilon \text{ for any } x \in D.$$

Proof. From (6.4) it is enough to show that for any $x \in D$ there is $s \in \mathbb{N}$ with $\varphi^s(x) \in W(\delta)$. Assume for a contradiction that $W(\varphi^s(x_0)) \geq \delta$ for some $x_0 \in D$ and any $s \in \mathbb{N}$. Hence, as $o(x) \cap S \neq \varnothing$, $x \in X$, there exists a susbequence $\varphi^{k_n}(x_0)$ of $\{\varphi^k(x_0)\}_{k\in\mathbb{N}}$ with $\varphi^{k_n}(x_0) \in S \setminus W(\delta)$. But for $s \geq 1$ we have

$$W(\varphi^s(x_0)) = W(x_0) - \sum_{i=0}^{s-1} \left(W(\varphi^i(x_0)) - W(\varphi^{i+1}(x_0)) \right).$$

From (6.4) and (6.5) it follows that $W(\varphi^{k_1}(x)) \leq W(x) - (\varepsilon_0 - \varepsilon)$ and, by induction, $W(\varphi^{k_n}(x_0)) \leq W(x_0) - n \cdot (\varepsilon_0 - \varepsilon)$. As $\varepsilon_0 > \varepsilon$, it leads to $W(\varphi^{k_n}(x_0)) \overset{k\to\infty}{\to} -\infty$, a contradiction. □

Theorem 6. *Let $W \colon X \to [0, +\infty)$ be a continuous function with $W_{\min} = 0$ and let $D_0 \subset X$ be such that for any $x \in D_0$ and $s \in \mathbb{N},\ t \in \mathbb{N}$*

$$W(\varphi^t(\varphi^s(x))) \leq W(\varphi^s(x)) + \varepsilon_s, \tag{6.6}$$

where $\varepsilon_s > 0$ is a sequence with $\varepsilon_s \to 0$. Let $\delta > 0$, $S \subset X$, $S' \subset X$ be such that S' is compact and

(1) *$o(x) \cap S \neq \varnothing$ for any $x \in X$,*
(2) *$W(\varphi(s)) < W(s)$ for any $s \in S' \setminus W(\delta)$,*
(3) *$W \circ \varphi \colon X \to \mathbb{R}$ is upper semi–continuous on S',*
(4) *there exists a function $\psi \colon S \to S'$ with $W|_S = W \circ \psi$ and $W \circ \varphi|_S = W \circ \varphi \circ \psi$*

Then $\limsup_{t\to\infty} W(\varphi^t(x)) \leq \delta$ for any $x \in D_0$.

Proof. Fix $\delta_1 > 0$ and $x \in D_0$. We will show that $\limsup_{t\to\infty} W(\varphi^t(x)) \leq \delta + \delta_1$. First note that since for any $s \in S$, $W(s) = W(\psi(s))$, we have $\psi(S \setminus W(\delta)) \subset S' \setminus W(\delta)$. The compactness of S' and the continuity of W imply that $S' \setminus W(\delta)$ is compact. Furthermore, by (2) and (3), the function $W - W \circ \varphi$ is lower semi–continuous and positive on $S' \setminus W(\delta)$. In consequence, it is bounded from zero on $S' \setminus W(\delta)$ (from Weierstrass theorem a lower semi-continuous function attains its infimum on a compact set). Thus, for some positive $\varepsilon_0 > 0$ we have

$$W(x) - W(\varphi(x)) > \varepsilon_0 \text{ for any } x \in S \setminus W(\delta).$$

From (6.6) it follows that for $s \in \mathbb{N}$ greater from some s_0 big enough we have

$$W(\varphi^t(\varphi^s(x))) \leq W(\varphi^s(x)) + \min\{\delta_1, \frac{\varepsilon_0}{2}\}, \quad x \in D_0.$$

Now we can use Lemma 5 as set S, constant $\varepsilon_0 > 0$, the set $D = \varphi^{s_0}(D_0)$ and $\varepsilon = \min\{\delta_1, \frac{\varepsilon_0}{2}\}$ satisfy the assumptions. □

Let $(\mathcal{U}, d_{\mathcal{U}})$ be a metric space and let $(\mathcal{M}, d)$ be a compact metric space. Let $\theta \colon \mathcal{U} \ni u \to \theta u \in \mathcal{U}$ and $\Pi \colon \mathcal{U} \times \mathcal{M} \colon (u, m) \to \Pi_u m \in \mathcal{M}$ be given maps. We define $\Pi^t\colon \mathcal{U}\times\mathcal{M}\ni(u,m)\to\Pi^t_u m\in\mathcal{M}$, $t\in\mathbb{N}$ by the following formula:

$$\Pi^0(u,m)=m \text{ and } \Pi^{t+1}_u m=\Pi_{\theta^t u}\Pi^t_u m, \text{ where } \theta^0 u=u. \tag{6.7}$$

In other words: $\Pi^t_u m=(\Pi_{\theta^{t-1}u}\circ\Pi_{\theta^{t-2}u}\circ\ldots\circ\Pi_u)(m)$, $t\geq 1$. Fix $u\in\mathcal{U}$. By equation (6.7), Π and θ determine a non–autonomous pseudo–dynamical system on $\mathcal{M}$. Space $\mathcal{U}$ can be called a base space and space X can be called a state space. The orbit of an element $m\in\mathcal{M}$ is given by $o_u(m)=\{\Pi^t_u m\colon\ t=0,1,2,\ldots\}$. The pair (θ,Π) forms the skew-product flow $\varphi\colon U\times\mathcal{M}\longrightarrow U\times\mathcal{M}$ according to the relation $\varphi(u,m)=(\theta u,\Pi_u m)$, this general construction is featured, for example, in [17]. The following theorem provides the sufficient conditions for the attractiveness and a stability type property of a closed set $\mathcal{M}^\star\subset\mathcal{M}$ in this context.

Theorem 7. *Let $\mathcal{U}_0\subset\mathcal{U}$, $\varnothing\neq\mathcal{M}^\star\subset\mathcal{M}$ and let $V\colon\mathcal{M}\to\mathbb{R}$ be a continuous function such that:*

1. $V(m)=0$ *for* $m\in\mathcal{M}^*$,
2. $V(m)>0$ *for* $m\in\mathcal{M}\setminus\mathcal{M}^*$,
3. $V(\Pi^t_{\phi^s u}\Pi^s_u m)\leq V(\Pi^s_u m)+\varepsilon_s$ *for any* $t\in\mathbb{N}$, $s\in\mathbb{N}$, $u\in\mathcal{U}_0$ *and* $m\in\mathcal{M}$, *where* $\varepsilon_s\to 0^+$.

Assume that for any $k\in\mathbb{N}$ there are sets $\mathcal{U}_0(\frac{1}{k})\subset\mathcal{U}$ and $\mathcal{U}_0'(\frac{1}{k})\subset\mathcal{U}$ such that $\mathcal{U}_0'(\frac{1}{k})$ is compact and:

(a) *for any $u\in\mathcal{U}$ there is $s\geq 0$ with $\theta^s u\in\mathcal{U}_0(\frac{1}{k})$,*
(b) *for any $u\in\mathcal{U}_0'(\frac{1}{k})$ and $m\in\mathcal{M}\setminus V(\frac{1}{k})$, $V(\Pi_u m)<V(m)$,*
(c) *$V\circ\Pi\colon\mathcal{U}\times\mathcal{M}\to\mathbb{R}$ is upper semi–continuous on $\mathcal{U}_0'(\frac{1}{k})\times\mathcal{M}$,*
(d) *there is a function $\zeta_k\colon\mathcal{U}_0(\frac{1}{k})\to\mathcal{U}_0'(\frac{1}{k})$ such that for $u\in\mathcal{U}_0(\frac{1}{k})$ and $m\in\mathcal{M}$*

$$V(\Pi_u m)=V(\Pi_{\zeta_k(u)}m) \tag{6.8}$$

Then, for any $u\in\mathcal{U}_0$ and $m\in\mathcal{M}$,

$$d(\Pi^t_u m,\mathcal{M}^*)\to 0,\ \text{as } t\to\infty,\quad\text{and}\quad V(\Pi^t_u m)\to 0,\ \text{as } t\to\infty.$$

Furthermore, for any $\varepsilon>0$ there is $\delta>0$ and s_0 such that for any $s>s_0$ and $u\in\mathcal{U}_0$ we have

$$d(\Pi^t_{\theta^s u}m,\mathcal{M}^\star)<\varepsilon,\ t\in\mathbb{N},\ \text{for any } m\in\mathcal{M}^\star(\delta)\cap\Pi_{\theta^s u}(\mathcal{M}).$$

Proof. Let $X=\mathcal{U}\times\mathcal{M}$ be the product metric space and let the product mapping $\varphi\colon X\to X$ be as follows

$$\varphi(u,m)=(\theta u,\Pi_u m).$$

Clearly $\varphi^t(u,m)=(\theta^t u,\Pi^t_u m)$ and $\varphi^{t+s}(u,m)=(\theta^{t+s}u,\Pi^t_{\theta^s u}\Pi^s_u m)$. Let $K=\mathcal{U}\times\mathcal{M}^*$ and let

$$W\colon X\ni(u,m)\to V(m)\in\mathbb{R}.$$

We will take advantage of Theorem 6. It is clear that $W^{-1}(0) = K$. From condition (3) it follows that equation (6.6) is satisfied with $D_0 = \mathcal{U}_0 \times \mathcal{M}$. Fix $k \in \mathbb{N}$. Define $S = \mathcal{U}_0(\frac{1}{k}) \times \mathcal{M}$ and $S' = \mathcal{U}_0'(\frac{1}{k}) \times \mathcal{M}$. Clearly S' is compact and, by (a), $o(x) \cap S \neq \varnothing$ for any $x = (u, m) \in X$. From (b), $W(\varphi(u, m)) < W(u, m)$ for any $(u, m) \in S' \setminus W(\frac{1}{k}) = \mathcal{U}_0'(\frac{1}{k}) \times (\mathcal{M} \setminus V(\frac{1}{k}))$. Furthermore, since $W \circ \varphi = V \circ \Pi$, then, by (c), $W \circ \varphi$ is upper semi–continuous on S'. Let

$$\psi \colon S \ni (u, m) \to (\zeta_k(u), m) \in S'.$$

For any $(u, m) \in S$, by the definitions of W and ψ, and (d),

$$W(u, m) = V(m) = W(\zeta_k(u), m) = W(\psi(u, m))$$

and

$$W(\varphi(u, m)) = V(\Pi_u m) = V(\Pi_{\zeta_k(u)} m) = W(\varphi(\zeta_k(u), m)) = W(\varphi(\psi(u, m))).$$

The assumptions of Theorem 6 are satisfied and hence $\limsup_{t\to\infty} W(\varphi^t(x)) \leq \frac{1}{k}$. We thus have $\limsup_{t\to\infty} V(\Pi_u^t m) \leq \frac{1}{k}$. Letting $k \to \infty$ we obtain that $V(\Pi_u^t m) \to 0$. To prove the remaining part ot the theorem, note that since $\mathcal{M}$ is compact and V is continuous, the function W and the system φ on X satisfy equations (6.1) and (6.2). Now we apply Observation 3 to the function W, the system φ and the set D_0, which finishes the proof. $\square$

7. The Proof

In this section we prove Theorem 2. We will work under notation of Section 3. First recall that for any $\delta > 0$ the sets $U(\delta) \subset \mathcal{T} \times \mathcal{N}$ and $U_0(\delta) \subset \mathcal{T} \times \mathcal{N}$ are defined as follows:

$$\mathcal{T} \times \mathcal{N} \supset U(\delta) \ni (T, \nu) \overset{def}{\Longleftrightarrow} \begin{cases} \int_B f(T(x, y)) v(dy) \leq f(x) & \text{for } x \notin A(\delta) \\ \int_B f(T(x, y)) v(dy) \leq \delta & \text{for } x \in A(\delta) \end{cases}$$

$$\mathcal{T} \times \mathcal{N} \supset U_0(\delta) \ni (T, \nu) \overset{def}{\Longleftrightarrow} \begin{cases} \int_B f(T(x, y)) v(dy) < f(x) & \text{for } x \notin A(\delta) \\ \int_B f(T(x, y)) v(dy) \leq \delta & \text{for } x \in A(\delta). \end{cases}$$

Now we recall the theorem.

Theorem 8. *Assume that we have a decreasing family of compact sets $\{U_0^k\}_{k\in\mathbb{N}}$ with $U_0^k \subset U_0(\frac{1}{k})$, $k \in \mathbb{N}$, and a sequence $\{\delta_t\}_{t\in\mathbb{N}} \subset \mathbb{R}$ such that the following conditions are satisfied:*

- **(A1)** *for any $k \in \mathbb{N}$, any pair $(T, \nu) \in U_0^k$ and $x \in A$, the $f \circ T$ is upper semi–continuous at (x, y) for ν-almost any y from B,*
- **(B1)** $\forall t \in \mathbb{N}$ $(T_t, \nu_t) \in U(\delta_t)$ *and* $\lim_{t\to\infty} \delta_t = 0$,

(C1) *for any $k \in \mathbb{N}$ the sequence (T_t, ν_t) contains a subsequence $(T_{t_n^k}, \nu_{t_n^k}) \in U_0^k$ such that $\lim_{n\to\infty} S_n^k = 0$, where:*

$$S_n^k = \sum_{i=t_n^k+1}^{t_{n+1}^k-1} \delta_i.$$

Then

$$Ed(X_t, A^*) \overset{t\to\infty}{\longrightarrow} 0 \quad \textit{and} \quad Ef(X_t) \overset{t\to\infty}{\longrightarrow} 0.$$

We will start the proof with Lemma 6. Let

$$V \colon \mathcal{M} \ni \mu \longrightarrow \int_A f d\mu \in [0,\infty), \text{ where } \mathcal{M} = \mathcal{M}(A),$$

and let

$$V(\frac{1}{k}) = \{\mu \in \mathcal{M} \colon V(\mu) < \frac{1}{k}\} \text{ and } V(\frac{1}{k})' = \mathcal{M} \setminus V(\frac{1}{k}).$$

Lemma 6. *We have*

$$\forall k \in \mathbb{N}\ \exists n_k \in \mathbb{N}\ \exists \bar{\varepsilon}_k > 0\ \forall u \in U_0^{n_k}\ \forall \mu \in V(\frac{1}{k})', \quad V(P_u\mu) < V(\mu) - \bar{\varepsilon}_k,$$

where $P_u \colon \mathcal{M} \to \mathcal{M}$ denotes the Foias operator.

Proof. Fix $k \in \mathbb{N}$. At first note that for any $n \in \mathbb{N}$, $u = (T,\nu) \in U_0(\frac{1}{n})$ and $\mu \in \mathcal{M}$ we have:

$$(7.1)\quad V(P_u\mu) = \int_A f dP_u\mu = \int_{A(\frac{1}{n})} f dP_u\mu + \int_{A(\frac{1}{n})'} f dP_u\mu \le \frac{1}{n} + \int_{A(\frac{1}{n})'} f dP_u\mu.$$

Now, note that the function

$$\overline{V} \colon U_0^k \times A \ni ((T,\nu),x) \to \int_B f(T(x,y))v(dy) \in \mathbb{R}$$

is upper semi–continuous. In fact, the function $X \ni x \to \delta_x \in \mathcal{M}$ is continuous and

$$\overline{V}((T,\nu),x) = \int_B f(T(x,y))v(dy) = \int_A\int_B f(T(z,y))v(dy)\delta_x(dz) = \int_A f dP_{(T,\nu)}\delta_x.$$

Thus, the upper semi-continuity follows from Proposition 1. Now, since $U_0^k \subset U_0(\frac{1}{k})$ and $U_0^k \times A(\frac{1}{k})'$ is compact, and $\overline{V}(u,x) = \int_A f dP_u\delta_x = \int_B f(T(x,y))\nu(dy) < f(x)$ for any $x \in A(\frac{1}{k})'$ and $u = (T,\nu) \in U_0^k$, we have that the function

$$(T,\nu,x) \longrightarrow f(x) - \int_B f(T(x,y))\nu(dy)$$

is bounded from zero on the set $U_0^k \times A(\frac{1}{k})'$. In other words, for any $x \in A(\frac{1}{k})'$ and $u \in U_0^k$, we have $f(x) - \int_B f(T(x,y))\nu(dy) > \varepsilon_k$ for some $\varepsilon_k > 0$ and, in consequence, for any $\mu \in \mathcal{M}$,

$$\int_{A(\frac{1}{k})'} f d\mu - \int_{A(\frac{1}{k})'} f dP_u\mu = \int_{A(\frac{1}{k})'} f(x)\mu(dx) - \int_{A(\frac{1}{k})'} \int_B f(T(x,y))\nu(dy)\mu(dx) \geq \varepsilon_k \mu(A(\frac{1}{k})').$$

Now, note that there is $\varepsilon > 0$ such that for any $\mu \in V(\frac{2}{k})'$ we have $\mu(A(\frac{1}{k})') > \varepsilon > 0$. In fact, if $\mu_n \in V(\frac{2}{k})'$ is a sequence with $\mu_n(A(\frac{1}{k})') \overset{n\to\infty}{\to} 0$, then

$$\limsup_{n\to\infty} \int_A f d\mu_n = \limsup_{n\to\infty} \int_{A(\frac{1}{k})} f d\mu_n \leq \frac{1}{k},$$

a contradiction with $\mu_n \in V(\frac{2}{k})'$. Furthermore, for any $n \geq k$ we have $U_0^n \subset U_0^k$ and, for any $\mu \in V(\frac{2}{k})'$, $\mu(A(\frac{1}{n})') \geq \mu(A(\frac{1}{k})') > \varepsilon$. Thus for any $n \geq k$, $u = (T,\nu) \in U_0^n$ and $\mu \in V(\frac{2}{k})'$,

$$\int_{A(\frac{1}{k})'} f d\mu - \int_{A(\frac{1}{k})'} f dP_u\mu \geq \varepsilon_k \cdot \varepsilon > 0.$$

Hence, by (7.1), for $n \geq k$, $u \in U_0^n$ and $\mu \in V(\frac{2}{k})'$, we have

$$V(P_u\mu) \leq \frac{1}{n} + \int_{A(\frac{1}{n})'} f dP_u\mu \leq \frac{1}{n} + \int_{A(\frac{1}{n})'} f d\mu - \varepsilon_k \cdot \varepsilon \leq V(\mu) + \frac{1}{n} - \varepsilon_k \cdot \varepsilon.$$

Thus, for $\bar{n}_k \in \mathbb{N}$ big enough we have that $V(P_u\mu) < V(\mu) - \bar{\delta}_k$ for any $\mu \notin V(\frac{2}{k})$, $u \in U_0^{\bar{n}_k}$ and some $\bar{\delta}_k > 0$. This finishes the proof of the lemma as constant $k \in \mathbb{N}$ was chosen arbitrarily and constants $n_k = \bar{n}_{2k}$ and $\bar{\varepsilon}_k = \bar{\delta}_{2k}$ satisfy the lemma' thesis. □

Proof of Theorem 2. . The proof will be based on Theorem 7. Recall that the space $\mathcal{M} = \mathcal{M}(A)$, equipped with the weak convergence topology, is compact. Let $U = \mathcal{T} \times \mathcal{N}$ and let

$$\mathcal{U} = \{u \in U^{\mathbb{N}}:\ \exists \{t_k\}_{k\in\mathbb{N}}\ \ u_{t_k} \in U_0^k\}$$

be a metric space with the product metric $d_{\mathcal{U}}$, which is defined by

$$d_{\mathcal{U}}(u,v) = \sum_{i=0}^{\infty} 2^{-i} d_U(u_i, v_i),$$

where $d_U \colon U \times U \to \mathbb{R}$ is a metric compatible with the topology on U. For $u = (u_0, u_1, \dots) \in \mathcal{U}$ we will use the following notation:

(B1)$(u) \equiv \forall t \in \mathbb{N}\ u_t \in U(\delta_t)$,
(C1)$(u) \equiv \forall k \in \mathbb{N}\ \forall n \in \mathbb{N}\ u_{t_n^k} \in U_0^k$,

where $\{\delta_t\}_{t\in\mathbb{N}}$ and $\{t_n^k\}_{n\in\mathbb{N}}$ are sequences from assumptions of Theorem 2. Let

$$\mathcal{U}_0 = \{u = (u_0, u_1, \ldots) \in \mathcal{U}\colon \text{ conditions } \mathbf{(B1)}(u) \text{ and } \mathbf{(C1)}(u) \text{ are satisfied } \}$$

Obviously assumptions of Theorem 1 imply that the sequence $u_t = (T_t, \nu_t)$ belongs to $\mathcal{U}_0$. Let $\theta\colon \mathcal{U} \to \mathcal{U}$ be a shift map, defined by

$$\theta(u_0, u_1, \ldots) = (u_1, u_2, \ldots)$$

and let $\Pi\colon \mathcal{U} \times \mathcal{M} \to \mathcal{M}$ be as follows

$$\Pi\colon \mathcal{U} \times \mathcal{M} \ni ((u_0, u_1, \ldots), \mu) \longrightarrow P_{u_0}\mu \in \mathcal{M}.$$

Recall that Π^t, defined by equation (6.7), satisfies $\Pi_u^{t+1} = \Pi_{\theta^t u} \circ \Pi_u^t = P_{u_t} \circ \Pi_u^t$. Note that if $\mu_t = P_{X_t}$ denote the distributions of the sequence X_t then from Lemma 2 it follows that $\Pi_u^t(\mu_0) = \mu_t$, where $u = (T_t, \nu_t)_{t=0}^{\infty}$. Recall that $V\colon \mathcal{M} \to \mathbb{R}$ and $\mathcal{M}^\star \subset \mathcal{M}$ are defined as

$$V(\mu) = \int_A f d\mu \quad \text{and} \quad \mathcal{M}^\star = \mathcal{M}^\star(A^\star).$$

We assumed that f is continuous (and bounded as A is compact), therefore the continuity of V follows directly from the definition of weak convergence. It is easy to see that V satisfies assumptions (1),(2) of Theorem 7. In fact, $\mu \in \mathcal{M}^\star(A^\star)$ iff $\mu(A^\star) = 1$. For any $x \in A$, $f(x) \geq 0$ and $f(x) = 0 \Leftrightarrow x \in A^\star$, and hence for any μ from $\mathcal{M}$, $V(\mu) \geq 0$ and $V(\mu) = 0 \Leftrightarrow \mu(A^\star) = 1 \Leftrightarrow \mu \in \mathcal{M}^\star$.

Now we define

$$\mathcal{U}_0(\frac{1}{k}) = U_0^{n_k} \times U^{\mathbb{N}} \text{ and } \mathcal{U}_0'(\frac{1}{k}) = U_0^{n_k} \times U_0^{n_{k+1}} \times U_0^{n_{k+2}} \times \ldots,$$

where n_k is a sequence as in Lemma 6 and $\{U_0^k\}_{k\in\mathbb{N}}$ is a family from the assumptions of Theorem 2. Recall that compactness of sets $U_0^{n_k}$ implies that $\mathcal{U}_0'(\frac{1}{k})$ is compact. We will show now that for any $k \in \mathbb{N}$, $\mathcal{U}_0(\frac{1}{k})$ and $\mathcal{U}_0'(\frac{1}{k})$ satisfy the assumptions (a),(b),(c),(d) of Theorem 7. Assumption (a) follows directly from the definitions of $\mathcal{U}$, $\mathcal{U}_0(\frac{1}{k})$ and θ. Condition (b) is a consequence of Lemma 6 as for any $u = (u_0, u_1, \cdots) \in \mathcal{U}$ and $\mu \in \mathcal{M}$ we have $V(\Pi_u\mu) = V(P_{u_0}\mu)$. As the projection $\mathcal{U} \ni u \to u_0 \in U$ is continuous, to check (c) it is enough to note that $(V \circ \Pi)(u, \mu) = \int_A f dP_{u_0}\mu$, $u \in \mathcal{U}$, $\mu \in \mathcal{M}$ and that from Proposition 1 it follows that the function $(u_0, \mu) \longrightarrow \int_A f dP_{u_0}\mu$ is upper semi-continuous on $U_0^{n_k} \times \mathcal{M}$ for any $k \in \mathbb{N}$.
To show condition d) fix $k \in \mathbb{N}$ and a sequence $w = (w_0, w_1, \ldots) \in \mathcal{U}_0'(\frac{1}{k})$ and define

$$\zeta_k\colon \mathcal{U}_0(\frac{1}{k}) \ni u \longrightarrow (u_0, w_1, w_2, \ldots) \in \mathcal{U}_0'(\frac{1}{k}).$$

To see d) it is enough to note that for any $u \in \mathcal{U}$ and $\mu \in \mathcal{M}$ we have:

$$V(\Pi_u \mu) = \int_A f dP_{u_0} \mu = V(\Pi_{\zeta_k(u)} \mu).$$

It remains to show condition (3). Fix $\varepsilon > 0$. We need to show that for all $s > s_0$ big enough we have $V(\Pi^t_{\phi^s u} \Pi^s_u \mu) \leq V(\Pi^s_u \mu) + \varepsilon$, for any $\mu \in \mathcal{M}$, $u \in \mathcal{U}_0$ and $t \in \mathbb{N}$. It will be enough to show that for all $s > s_0$ big enough,$u \in \mathcal{U}_0$ and $\mu \in \mathcal{M}$ we have

$$V(P_{u_{s+t}} \dots P_{u_{s+2}} P_{u_{s+1}} \mu) < V(\mu) + \varepsilon, \ t \in \mathbb{N}. \tag{7.2}$$

Let k_0 be such that $\frac{1}{k_0} < \frac{\varepsilon}{4}$ and let $\varepsilon_0 > 0$ satisfy

$$\varepsilon_0 < \inf\{\int f d\mu - \int f dP_u \mu \colon \ u \in U_0^{n_{k_0}}, \ \mu \in V(\frac{1}{k_0})'\}$$

(for example we can take $\varepsilon_0 = \bar{\varepsilon}_{k_0}$ from Lemma 6). As $\{u_t\}_t \subset \mathcal{U}_0$ there is a subsequence u_{t_s} of u_t such that $u_{t_s} \in U_0^{n_{k_0}}$ and that $S_s = \sum_{i=t_s+1}^{t_{s+1}-1} \delta_i \to 0$ as $s \to \infty$. We can assume that t_0 is big enough to have

$$S_s = \sum_{i=t_s+1}^{t_{s+1}-1} \delta_i < \min\{\frac{\varepsilon}{2}, \frac{\varepsilon_0}{2}\}, s \in \mathbb{N}.$$

Note that to show (7.2) it is enough to show that for any $s \in \mathbb{N}$ and any $0 < i < t_{s+1} - t_s$ we have

$$V(P_{u_{(t_s+i)}} \dots P_{u_{(t_s+2)}} P_{u_{(t_s+1)}} \mu) < V(\mu) + \frac{\varepsilon}{2}, \ \mu \in \mathcal{M} \tag{7.3}$$

and

$$V(P_{u_{t_{(s+1)}}} \dots P_{u_{t_{(s)}+2}} P_{u_{t_{(s)}+1}} \mu) < \max\{V(\mu), \frac{\varepsilon}{2}\}, \ \mu \in \mathcal{M}. \tag{7.4}$$

In fact, by simple induction, from (7.4) it follows that for any $s \in \mathbb{N}$ and $k \in \mathbb{N}$ we have

$$V(P_{u_{t_{(s+k)}}} \dots P_{u_{t_s+2}} P_{u_{t_s+1}} \mu) < \max\{V(\mu), \frac{\varepsilon}{2}\}$$

and thus from (7.4) and (7.3) easily follows (7.2). To see (7.3) note that if $0 < i < t_{s+1} - t_s$ then, as $u_t \in U(\delta_t)$, from (7.1) we easily conclude that

$$V(P_{u_{t_s+i}} \dots P_{u_{t_s+1}} \mu) < V(\mu) + \delta_{t_s+1} + \delta_{t_s+2} + \dots + \delta_{t_s+i} \leq V(\mu) + S_s < V(\mu) + \frac{\varepsilon}{2}.$$

Now, for $\mu \in \mathcal{M}$ and $s \in \mathbb{N}$, let

$$\mu_s = P_{u_{t_{(s+1)}}} \dots P_{u_{t_s+1}} \mu \text{ and } \mu_{s-1} = P_{u_{t_{(s+1)}-1}} \dots P_{u_{t_s+1}} \mu$$

so we have

$$\mu_s = P_{u_{t_{(s+1)}}} \mu_{(s-1)}$$

(we put $\mu_{s-1} = \mu$ if $t_{(s+1)} = t_{(s)} + 1$). Recall that $u_{t_{(s+1)}} \in U_0^{n_{k_0}}$. We have either $\mu_{s-1} \in V(\frac{1}{k_0})$ or $\mu_{s-1} \in \mathcal{M} \setminus V(\frac{1}{k_0})$. In the first case we have

$$V(\mu_s) \leq V(\mu_{s-1}) + \frac{1}{k_0} \leq \frac{2}{k_0} < \frac{\varepsilon}{2}.$$

In the second case, we have

$$V(\mu_s) < V(\mu_{s-1}) + S_s - \varepsilon_0 < V(\mu_{s-1}) - \frac{\varepsilon_0}{2}.$$

We thus have that

$$V(\mu_s) = V(P_{u_{t_{(s+1)}}} \ldots P_{u_{t_s+2}} P_{u_{t_s+1}} \mu) < \max\{V(\mu) - \frac{\varepsilon_0}{2}, \frac{\varepsilon}{2}\},$$

which proves (7.4). We thus have that all the assumptions of Theorem 7 is satisfied. Therefore, the probability distributions μ_t of X_t weakly converge towards $\mathcal{M}^\star$. Observation 1 finishes the proof. □

Below we state an additional stability result which follows from Theorem 7 and the above proof. Let

$$\mathcal{M}^\star(\varepsilon) = \{\mu \in \mathcal{M} \colon d_{\mathcal{M}}(\mu, \mathcal{M}^\star) < \varepsilon\},$$

where $\varepsilon > 0$ and $d_{\mathcal{M}}$ states for Prohorov metric.

Theorem 9. *Let X_t be a process defined by* (3.1)*. Under assumptions of Thereom 2 we have:*

$$\forall \varepsilon > 0 \; \exists \delta > 0 \; \exists t_0 \; \forall t > t_0 \; P(d(X_t, A^\star) < \delta) \geq 1 - \delta \Longrightarrow$$

$$\Longrightarrow P(d(X_{t+s}, A^\star) < \varepsilon) \geq 1 - \varepsilon, \quad s \in \mathbb{N}.$$

Proof. The objects constructed in the previous proof satisfy the assumptions of Theorem 7. Hence, as $u = (T_t, \nu_t) \in \mathcal{U}_0$, the sequence $\mu_t \in \mathcal{M}$, defined by $\mu_{t+1} = P_{(T_t,\nu_t)}\mu_t$ and $\mu_0 = P_{X_0}$, is such that for any $\varepsilon > 0$ there is $\delta > 0$ such that for any s big enough,

$$\mu_s \in \mathcal{M}^\star(\delta) \Longrightarrow \mu_{s+t} M^\star(\varepsilon), \; t \in \mathbb{N}.$$

The above, expressed in terms of Prohorov metric and the process X_t, takes the form:

$$P(d(X_s, A^\star) < \delta) \geq 1 - \delta \Longrightarrow P(d(X_{s+t}, A^\star) < \varepsilon) \geq 1 - \varepsilon.$$

□

Acknowledgments. The research was supported by National Science Centre grant DEC-2013/09/N/ST/04262.

8. APPENDIX

Proof of Observation 1. To prove statement (1) note that conditions $A1$) and $A2$) imply that

$$\forall \varepsilon > 0 \; \exists \delta > 0 \; P(X_t \in A^\star(\varepsilon)) \geq P(X_t \in A(\delta))$$

and

$$\forall \delta > 0 \; \exists \varepsilon > 0 \; P(X_t \in A(\delta)) \geq P(X_t \in A^\star(\varepsilon)).$$

This proves the equivalence between conditions $1a$) and $1c$). Condition $1c$) means that $f(X_t)$ goes in probability to a one point distributed limit 0 and thus this is equivalent to the weak convergence expressed in condition 1d). The equivalence $1a) \Leftrightarrow 1b)$ is a straightforward conclusion of Observation 2 from Section 5. To show the second statement it is enough to note that for any sequence $x_t \in A$ we have $d(x_t, A^\star) \to 0 \Leftrightarrow f(x_t) \to 0$, which follows directly from $A1$), $A2$). To see the third statement it is enough to notice that under the boundedness assumption the sequences $f(X_t)$ and $d(X_t, A^\star)$ are uniformly integrable and hence the expected value convergence is equivalent to the convergence in probability in case of both sequences $f(X_t)$ and $d(X_t, A^\star)$. $\square$

Proof of Theorem 1. Based on Observation 1 it is enough to show equivalence (5) $\leftrightarrow$ (6). This follows from the Doob's martingale convergence theorem as $f(X_t)$ is a supermartignale with $0 \leq f(X_t)$ and thus the almost sure limit $\lim_{t\to\infty} f(X_t)$ exists and equals to the stochastic limit. $\square$

References

[1] Andrieu, C. Breyer, L.A. Doucet, A., "Convergence of simulated annealing using FosterLyapunov criteria", J. Appl. Probab., vol. 38, 2001, p.975-994

[2] M. Aoki, Optimization of Stochastic Systems, Topics in Dicrete-Time Systems, Academic Press, New York, London, 1967.

[3] M. J. Appel, R. Labarre, D. Radulovic, On Accelerated Random Search, SIAM J. Optim., 14(2003), pp. 708 - 731

[4] Claude J.P. Bélisle, Convergence Theorems for a Class of Simulated Annealing Algorithms, J. Appl. Probab. 29 (1992), pp. 885-895

[5] H.G. Beyer, H.P. Schwefel, Evolution Strategies – A Comprehensive Introduction, Nat. Comput. 1 (2002),pp. 3-52.

[6] N.P. Bhatia, G.P. Szegö, Dynamical Systems: Stability Theory and Applications, Springer-Verlag, Berlin, 1967

[7] P. Billingsley, Convergence of Probability Measures, Second Edition, A Wiley-Interscience Publication, New York, 1999.

[8] A.A. Borovkov, V. Yurinsky, Ergodicity and stability of stochastic processes, Wiley, Chichester, 1998

[9] M. Clerc, Particle Swarm Optimization, ISTE Ltd, London, 2006

[10] M. Clerc, J. Kennedy, The Particle Swarm – Explosion, Stability, and Convergence in a Multidimensional Complex Space, IEEE Trans. Evol. Comput., 6 (2002), pp. 58–73

[11] P. Diaconis, D. Friedman: Iterated Random Functions, SIAM Rev., 41 (1999), 45–76

[12] R.M. Dudley, Real Analysis and Probability, Cambridge University Press,Cambridge, 2004

[13] W. J. Gutjahr , ACO algorithms with guaranteed convergence to the optimal solution, Inform. Process. Lett. 82 (2002), pp. 145–153

[14] S.F. Jarner,R.L. Tweedie, Locally contracting iterated functions and stability of Markov chains, J. Appl. Probab. Volume 38 (2001), 494-507.
[15] D. Karaboga, C. Ozturk, A novel clusterring aproach: Artificial Bee Colony (ABC) algorithm, Appl. Soft Comput., 11 (2011), pp. 652–657
[16] E. Kanso, D. Saintillan, Special Issue Editorial: Emergent Collective Behavior: From Fish Schools to Bacterial Colonies, Journal of Nonlinear Science vol 25 (2015)
[17] P.E. Kloeden, M. Rasmussen, Nonautonomous Dynamical Systems, American Mathematical Society, Providence, 2011.
[18] A. Lasota, M. Mackey, Chaos, Fractals and Noise, Springer Verlag, New York, 1994.
[19] M. Locatelli, Convergence Properties of Simulated Annealing for Continuous Global Optimization, Journal of Applied Probability, Vol. 33, No. 4 (Dec., 1996), pp. 1127-1140
[20] M. Locatelli, Convergence of a Simulated Annealing Algorithm for Continuous Global Optimization, J. Global Optim., 18 (2000), pp. 219-233
[21] M. Locatelli, Convergence and first hitting time of simulated annealing algorithms for continuous global optimization. Math. Methods Oper. Res. 54 (2001), no. 2, 171199.
[22] M. Locatelli, Simulated annealing algorithms for continuous global optimization. Handbook of global optimization, Vol. 2, 179229, Nonconvex Optim. Appl., 62, Kluwer Acad. Publ., Dordrecht, 2002.
[23] M. Locatelli, F. Schoen, Random Linkage: a Family of Acceptance/Rejection Algorithms for Global Optimization, Math. Program. 85(2) (1999), 379–396
[24] S. Meyn , R. Tweedie, Markov Chains and Stochastic Stability, Springer-Verlag, London, 1993
[25] J. Ombach, Stability of evolutionary algorithms, Journal Math Anal Appl. 342(2008), 326 - 333.
[26] J. Ombach, A Proof of Convergence of General Stochastic Search for Global Minimum, Difference Equ. Appl. 13 (2007), pp. 795 – 802.
[27] J. Ombach, D. Tarłowski, Nonautonomous Stochastic Search in Global Optimization, J NONLINEAR SCI vol. 22(2012) (2012), 169 - 185
[28] D. Tarłowski, Nonautonomous stochastic search for global minimum in continuous optimization, J MATH ANAL APPL vol. Volume 412 Issue 2 (2014), 631-645
[29] D. Tarłowski, Sufficient conditions for the convergence of non-autonomous stochastic search for a global minimum, UIAM (2011), 73-83
[30] A. Pelczar, Stability of sets in Pseudo-dynamical systems II, Bull. Acad. Polon. Sci., Ser. Sci. Math., Astronom. et Phys. 19 (1971), 951–957
[31] J. Pintér: Convergence Properties of Stochastic Optimization Procedures, Math. Operationsforsch. u. Statist., ser: Optimization, 15 (1984),pp. 405 - 427
[32] R.G.Regis, Convergence guarantees for generalized adaptive stochastic search methods for continuous global optimization, European Journal of Operational Research 207 (2010) 1187-1202
[33] G. Rudolph: Convergence Properties of Evolutionary Algorithms, Kovac, Hamburg, 1997
[34] R. Schaefer, Podstawy genetycznej optymalizacji globalnej (In English: Introduction to Genetic Global Optimization), WUJ, Kraków, 2002
[35] L.M.Schmitt, Fundamental Study: Theory of Genetic Algorithm, Theoret. Comput. Sci., 259(2001), pp. 1–61
[36] M.A. Semenov, D.A. Terkel, Analysis of convergence of an evolutionary algorithm with self-adaptation using a stochastic Lyapunov function
[37] W. Słomczyński, Dynamical Entropy, Markov Operators, and Iterated Function Systems, WUJ, Kraków 2003
[38] F.J. Solis, R.J-B. Wets,: Minimization by Random Search Techniques, Math. Oper. Res., 6 (1981), pp 19–30

[39] R.L. Yang, Convergence of the Simulated Annealing Algorithm for Continuous Global Optimization, J. Optim. Theory Appl., 104 (2000),pp. 691–716

[40] A.A. Zhigljavsky: Theory of Global Random Search, Kluwer Academic Publishers, Dordrecht, 1991

[41] A. Zhigljavsky , A. Žilinskas, Stochastic Global Optimization, Springer, New York, 2008